High-Dimensional Variable Selection for Multivariate and Survival Data with Applications to Brain Imaging and Genetic Association Studies

by

Yanming Li

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2014

Doctoral Committee:

Professor Bin Nan, Chair
Professor Timothy D. Johnson
Assistant Professor Xiaoquan William Wen
Professor Ji Zhu

To my parants

ACKNOWLEDGEMENTS

First and foremost, I would like to take this opportunity to express my deepest gratitude to my advisor Dr. Bin Nan, who supervised me during my Ph.D. study. My appreciation also goes to Dr. Ji Zhu, who together with Dr. Nan provided invaluable advice and guidance. Without their enormous support, patience and enlightening suggestions, this work could not be completed. I benefited so much from their knowledge and insights in statistics and conducting scientific research. I also appreciate them for providing me enough freedom in doing research and for their constant encouragement.

I would also like to give my sincere thanks to Dr. Timothy D. Johnson and Dr. Xiaoquan William Wen for being my committee members, providing insightful comments on my dissertation work and sharing their expertise. Their encouragement and assistance have been invaluable for me.

In addition to my dissertation committee, I am also deeply indebted to Dr. Kerby Shedden and Dr. Brenda Gillespie, who have been great mentors and supervisors when I worked as a graduate student consultant at the Center for Statistical Consultation and Research (CSCAR) at the University of Michigan. I would also like to thank all my CSCAR colleagues and friends for everything I have learnt from them and for the good time and memories they shared with me.

I would like to thank Dr. Goncalo Abecasis and Dr. Michael Boehnke for providing supervision when I worked as a graduate student research assistant at the Center for Statistical Genetics at the University of Michigan, and for introducing me into the world of statistical genetics.

My special gratitude also goes to Dr. Yi Li, Dr. Zhi Kevin He and Dr. Min

Zhang for their advice on pursuing an academic career.

Finally, I am especially grateful to my parents. Their boundless love and unconditional support have always been with me through the time.

To all of those who have helped, I extend my heartfelt thanks.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

High-Dimensional Variable Selection for Multivariate and Survival Data with Applications to Brain Imaging and Genetic Association Studies

by

Yanming Li

Chair: Bin Nan

In this dissertation, we aim to solve important high-dimensional variable selection problems with either structured multivariate or discrete survival outcomes, with applications to brain imaging and genetic association studies.

In the first project, we introduce the multivariate sparse group lasso for variable selection in multivariate multiple regressions with both grouped covariates and responses. We propose an efficient mixed coordinate descent algorithm for the penalized least square estimation. The method is able to effectively remove unimportant groups and unimportant individual coefficients within important groups, particularly for large p small n problems. It is flexible in handling various complex group structures such as overlapping, nested, or multilevel hierarchical structures. The finite sample oracle properties of the proposed method are established and the method is applied to an eQTL association study.

In the second project, we propose a multi-stage method for conducting structured brain-wide-genome-wide association studies via the multivariate sparse group lasso. Compared to conventional single-voxel-to-single-SNP approaches, our multi-stage approach is more efficient in selecting the important signals and can avoid large number of multiple comparisons while effectively control the false discoveries by

using the stability selection. We apply the proposed method to a brain-wide GWAS using ADNI PET imaging and genotype data. Our method can handle the ultra-high dimensionalities of both 3D images and genetic markers, while considering the anatomic brain structure and the gene structure in the human genome. We confirm several previously reported and also find some novel genes that are either associated with brain glucose metabolism or with their associations significantly modified by Alzheimer's disease status.

In the third project, we propose a full-likelihood based variable selection method for a discrete-time and cure-rate survival model with high-dimensional time-varying predictors. The method is motivated by the ADNI longitudinal brain imaging study to predict conversions from mild-cognitive-impairment (MCI) to Alzheimers-disease (AD). The conversion time was only observed on discrete time intervals and the studied sample consists of a mixture of a non-cure group and a cure group. The proposed method uses the full likelihood to jointly model the cure rate and the survival probabilities of the non-cure subjects. Both models involve high-dimensional predictors. Variable selection is carried out using the elastic net penalties. The method can efficiently and effectively select the important predictors in both models. And it can be applied to many biomedical studies for analyzing grouped failure time data with a cure portion and high-dimensional predictors. We evaluate the method through extensive simulations and apply it to the ADNI PET brain imaging data to predict MCI-to-AD conversions.

CHAPTER I

Introduction

Statistical methods for analyzing high-dimensional data have attracted much attention in recent years. The word "high-dimension" refers to the case when the number of unknown parameters p is larger than the sample size n. Accompanied by the arrival of the "big data" era, it comes the need for more reliable and feasible variable selection methods, from either theoretical or computational perspective, in high-dimensional settings. Traditional variable selection procedures, such as the best-subset selection are known to be unstable (Breiman, 1996) and have poor prediction accuracy. They are also computationally prohibitive when the number of variables is large. To overcome such drawbacks, many modern variable-selection techniques have been proposed, such as shrinkage estimation together with stability selection (Meinshausen and Bühlmann, 2010) or sure independent screening (Fan and Lv, 2008). Followed by the the lasso method introduced by R. Tibshirani (1996), various other penalization methods have been introduced and widely studied, such as the adaptive lasso (Zou, 2006), the elastic net (Zou and Hastie, 2005), the group lasso (Yuan and Lin, 2006) and the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), just to name a few.

Many efficient algorithms were developed for solving the lasso and its variations such as the least angle regressions (LARs) (Efron et al., 2004), the "shooting" algorithm and the coordinate descent algorithms (Fu, 1998; Friedman et al., 2007; Wu and Lange, 2008; Tseng, 2001). In the high-dimensional cases, especially the ultrahigh-dimensional cases where the number of parameters is of exponential order

of the sample size, an efficient algorithm is rather crucial. A comprehensive coverage of algorithms for penalization (constrained) optimization can be found in Boyd and Vandenberghe (2004); Lange (2004) and Sra et al. (2012).

Oftentimes, the high-dimensional predictors are correlated or having group structures. Elastic net (Zou and Hastie, 2005; Zou and Zhang, 2009) can be used to select both independent and highly correlated variables simultaneously without requiring prior knowledge of grouping structures. When the grouping structure is known, the group lasso (Yuan and Lin, 2006) is a popular tool on selecting important groups. It does not, however, encourage sparsity within a group. More interests have been focused on simultaneously selecting important groups and important individual variables within a group, given the group structure. Simon et al. (2013) proposed using a mixture of l_1 and l_2 penalties to achieve this goal. Some other most recent developments can be found in Wang et al. (2009); Huang et al. (2009); Zhou and Zhu (2010); Bunea et al. (2011) and some of the references therein.

Despite their popularity, most of the group-and-within-group variable selection methods have been only focusing on univariate outcome cases. In the first part of this dissertation, we develop the multivariate sparse group lasso to handle the case where both the responses and the predictors are of high dimensions and both have known grouping structures. For the case that responses are of high-dimensional, the number of parameters are of much higher order than that in the univariate cases. Therefore it required a algorithm with much faster speed. Similar to Simon et al. (2013), our method employed both l_1 and l_2 penalties. In the presence of the l_2 component, the penalty term is no longer separable (Tseng, 2001) and therefore the coordinate descent algorithm does not have a closed form solution for each coordinate in each step. We propose a mixed coordinate descent algorithm which avoids solving for the fixed point solution in each iteration but still converges to the global optimizer. Our proposed algorithm fasten the convergence speed by at least thousand of times. We also show that the multivariate sparse group lasso enjoy the finite sample oracle properties.

The proposed multivariate sparse group lasso is motivated by real-world applications. Remarkable technologies nowadays such as microarrays, high-throughput sequencing, high-resolution imaging scan techniques have provided not only high-dimensional predictor data, but also high-dimensional phenotypic data. One primary motivation for this dissertation is to find associations between the human genome and the human brain and to understand how their associations are affected by, for example, the Alzheimer's disease (AD) status. AD is the most common type of dementia, It accounts for 60%-80% of dementia cases and affects more than 35 million people worldwide, with the number expected to be more than tripled by the year of 2050 (Alzheimer's Association, 2013). Both the association between the human genome and AD (http://www.alzgene.org) and the association between the human brain and AD (Kukull et al., 1996; Biffi et al., 2010) have been widely studied. Genetic markers such as apolipoprotein E gene and many brain regions have been identified to be associated with AD. But association studies between the human brain and the human genome have just become feasible recently(Stein et al., 2010a; Hibar et al., 2011). However, most of the current brain-wide and genome-wide association studies are still using pairwise approaches by looking at one pair of voxel and genetic marker at a time (Stein et al., 2010a). This kind of approaches have the limitation of controlling for the false discoveries from a huge number of multiple comparisons. Given the number of the imaging voxels in a human brain and the number of the genetic markers in the human genome, the true signals, if there were any, are not much likely to survive from any multiple comparison adjustment criterion. Therefore the power of such approaches is limited. The multivariate sparse group lasso provides a remedy.

In the second part of this dissertation, we introduce a multi-stage method for conducting structured brain-wide-genome-wide association studies (brain-GWAS) via the multivariate sparse group lasso. The data used in our brain-GWAS are Fluorine-fluorodeoxiglucose positron emission tomography (PDG-PET) brain images and DNA genotypes obtained from the Alzheimer's Disease Neuroimaging Ini-

tiative (ADNI) database. Compared to the pairwise approaches, our multi-stage approach has several advantages. First, it takes advantage of intrinsic biological structures of the brain and the genome and therefore is more efficient for selecting important signals by first ruling out unimportant group-level (brain-region-to-gene) signals and only focusing on the selected groups of signals in the subsequent stages. Thus it avoids the large number of multiple comparisons. Secondly, our method can effectively control for the number of false positive discoveries (FD) by employing the stability selection (Meinshausen and Bühlmann, 2010). Thirdly, it handles the within-group correlation or multicollinearity by imposing group level penalties and therefore increases the power for detecting true signals. We also consider interactions between the genotypes the AD status. From our brain-GWAS, we are able to confirm several previously reported genes and also find some novel genes that are either associated with brain glucose metabolism or with their associations significantly modified by Alzheimer's disease status.

The PDG-PET images used in our brain-GWAS were scanned at baseline during the first phase of ADNI study launched in 2003. There were 236 subjects diagnosed as mild cognitive impairment (MCI) patients at their baseline visits. Through the three phases of ADNI study, the 236 MCI patients were followed at 6th, 12th, 18th, 24th, 36th, 48th, 60th and 72th month after their first visits. Their disease status and PET image scans were recorded at each follow-up visit. Some of those MCI patients converted to AD during the follow-up. We are interested in seeing which parts of the brain are associated with the MCI-to-AD conversion and which can predict conversion probability based on patients historical brain scans. This is a variable selection problem for survival data with high-dimensional time-varying imaging predictors, with the event of interest being MCI-to-AD conversion.

Regularization methods have also been widely used for survival data (Tibshirabi, 1997; Yang and Zou, 2013; Engler and Li, 2009; Hastie and Tibshirabi, 1990). Tibshirabi (1997) first proposed the lasso for the Cox PH model. Fan and Li (2002) used SCAD for variable selections in the Cox PH models. Wang et al. (2008) consider a

doubly penalized Buckley-James method for accelerate-failure-time (AFT) models. Engler and Li (2009) proposed a Cox PH based and an AFT based adaptations of the elastic net. For a literature review on model selections for high-dimensional survival data, see Nan (2010) and Meijer and Goeman (2014). However, most of variable selection methods for survival models have focused on continuous survival time. While in our longitudinal PET imaging and MCI-to-AD conversion study, the events of interest were only observed within a few discrete time intervals. In such a case the partial likelihood is not applicable. Moreover, some MCI patients will never convert to AD from a clinical point of view, therefore a latent cure population (Boag, 1949; Berkson and Gage, 1952) may exist. To achieve the goal for predicting MCI-to-AD conversion while taking consideration of the above issues, we propose a full-likelihood (in difference to the partial likelihood) based variable selection method for a discrete-time and cure-rate survival model with high-dimensional time-varying predictors. The model integrates the mixture cure models and discrete-time survival models in a high-dimensional variable selection setting. It also utilizes computational algorithms such as quadratic approximation and majorization optimization to speedup the variable selection process. Use the proposed model, we are able to select several brain regions associated with the MCI-to-AD conversion which were confirmed by the existing AD literature and also some novel signals.

The rest of this dissertation consists of four chapters. In Chapter II, we introduce the multivariate sparse group lasso (MSGLasso) for a multivariate multiple linear regression with an arbitrary group structure. We propose a fast mixed coordinate descent algorithm for solving the MSGlasso. Its finite sample oracle properties are established and its performance under various settings are evaluated by extensive simulations. The results show that the MSGLasso can effectively remove unimportant groups and select important variables within the important groups. It has a better prediction performance than other competing shrinkage methods. Chapter III concerns a real-world application of the MSGLasso. In this chapter, we conduct a brain-wide and genome-wide association study modified by the AD status via the

MSGLasso. We are able to detect several genes who are either associated with the brain metabolic functions or with their associations significantly modified by the AD status. Chapter IV aims to use longitudinal PET images to predict MCI-to-AD conversions. We build a cure-rate and discrete time survival model for selecting the predictive imaging voxels for either the cure rate or the non-cure survival. We are able to select some important predictive brain regions supported by existing clinical studies. Finally, Chapter V contains a few concluding remarks and several potential directions for future work.

CHAPTER II

Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure

2.1 Introduction

Genomic association studies with a single phenotype have been widely studied. Such association studies often encounter high dimensional predictors with sparsity, i.e., only a small number of predictors are associated with the response variable. To select truly associated predictors, it is necessary to use regularization penalties to shrink the coefficients of irrelevant predictors to exactly zero. Popular penalties for regression models with a univariate response include the lasso (Tibshirani, 1996), the adaptive lasso (Zou, 2006), the elastic net (Zou and Hastie, 2005) and the smoothly clipped absolute deviation (Fan and Li, 2001), among many others.

An important characteristic of high-dimensional genomic predictors is the intrinsic group structures. For example, the DNA marker predictors, also known as single nucleotide polymorphisms (SNPs), can often be grouped into genes, and genes can be grouped into biological pathways. Such grouping strategies have been applied successfully to genomic studies in rare variant detection (Zhou et al., 2010; Biswas and Lin, 2012). For group variable selection, Yuan and Lin (2006) proposed the group lasso method for the univariate response case. It penalizes the L_2 norm of each predictor group and selects important groups in an "all-in-all-out" fashion.

That is, all the predictors in a group must be included or excluded simultaneously. However, in real applications, this is rarely the case. Oftentimes, not all the variables in an important group are important. For example, a gene associated with a certain complex trait does not mean that all the variants within the gene are causal, and a pathway that regulates certain gene expressions does not necessarily indicate that all its components have regulatory effects. Recent efforts have been made to select both important groups and important within-group signals simultaneously. Huang et al. (2009) and Zhou and Zhu (2010) adopted a L_γ, $0 < \gamma < 1$, penalty to select important groups while removing unimportant variables within them; Zhou et al. (2010) used a penalized logistic regression with a mixed L_1/L_2 penalty to select both common and rare variants in a genome-wide association study; and Simon et al. (2013) proposed the sparse group lasso for selecting both important groups and within group predictors. However, all the above methods concern a univariate response.

Many other genomic data analyses focus on investigating the associations between high dimensional response variables and high-dimensional covariates, such as gene-gene associations (Park and Hastie, 2008; Zhang et al., 2010), protein-DNA associations (Zamdborg and Ma, 2009) and brain fMRI-DNA (or gene) associations (Stein et al., 2010a). Oftentimes pairwise associations are calculated in such studies. For example, many multivariate genome-wide association studies nowadays still look for one association at a time between a single marker and a single trait, and then correct for multiple hypothesis testing (Dudoit et al., 2003; Stein et al., 2010a). However, when both responses and predictors are of high dimensions, most of the family-wise type I error controlling procedures are usually too conservative and yield poor performance (Stein et al., 2010a), and oftentimes adjusted analysis considering multiple variables simultaneously is more appropriate.

High dimensional responses also have natural group structures very often, for example, pathway group structures for gene expression responses and brain functional regions for fMRI intensity responses. For multivariate responses, Peng et al.

(2010) adopted the mixed L_1/L_2 penalty in an orthonormal setting for identifying hub covariates in a gene regulation network; Obozinski et al. (2011) and Bunea et al. (2011) studied joint support union and joint rank selections; Lounici et al. (2011) proved oracle inequalities for multitask learning. Despite all the efforts, little focus, to our knowledge, has been put on the cases where the responses also have a group structure, whereas such cases are commonly encountered in biological studies. A possible strategy for multivariate-response analysis is to perform covariate selection for one response variable at a time. In such analysis the predictor group structure can be considered but the response group structure is overlooked.

In this article, we propose a regularization method for making a good use of the intrinsic biological group structures on both covariates and responses to facilitate a better variable selection on multivariate-response and multiple-predictor data by effectively removing unimportant blocks of regression coefficients. Both the predictor and response group structures, or more generally, the block structures of the regression coefficient matrix, are assumed known. Information of many biologically confirmed group structures can be achieved from publicly available repositories, for example, RefSeq gene files from NCBI Reference Sequence Database (http://www.ncbi.nlm.nih.gov/refseq/), KEGG pathway maps from Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg/), and Brodmann brain anatomic region atlas from https://surfer.nmr.mgh.harvard.edu/fswiki/. The proposed method can handle cases where the number of variables in either responses or predictors is much greater than the sample size, and complex group structures such as overlapping groups where a variable belongs to multiple groups. The estimators enjoy finite sample oracle bounds for the prediction error, the estimation error, and the estimated sparsity of the regression coefficient matrix. Extensive simulations show that the proposed method outperforms competitive regularization methods. We applied the proposed method to a yeast gene expression quantitative loci (eQTL) study, where the numbers of gene expression responses and genetic marker predictors are both much larger than the sample size. The gene expression respons-

es are grouped into biological pathways and the genetic markers are grouped into genes. We demonstrate by considering both group structures that the proposed method generates a much more interpretable and predictive eQTL network between the gene expressions and genetic markers, comparing with several other commonly used regularized approaches.

2.2 Multivariate linear model with arbitrary grouping

We consider the multivariate linear model

$$Y = XB + W, \tag{2.1}$$

where $Y = (y_1, \cdots, y_q) \in \mathbb{R}^{n\times q}$ is the response matrix of n samples and q variables, $X = (x_1, \cdots, x_p) \in \mathbb{R}^{n\times p}$ is the covariate matrix of n samples and p variables, $B = (\beta_{jk})_{p\times q} \in \mathbb{R}^{p\times q}$ is the coefficient matrix and $W = (w_1, \cdots, w_q) \in \mathbb{R}^{n\times q}$ is the matrix of error terms with each $w_k \sim N(0, \sigma_k^2 I_{n\times n})$, $k = 1, \cdots, q$. Assume Y and X are centered so that there is no intercept in B. We adopt the notational convention that the column vectors of X are indexed by j, the column vectors of Y and W are indexed by k, and the samples are indexed by i.

Assume B contains G groups, and each group, denoted as B_g where $g \in \{1, \cdots, G\}$, is a subset of two or more elements in B. We denote the group structure by $\mathcal{G} = \{B_1, \cdots, B_G\}$. We use B or B_g to denote either the set of all their elements or the numerical values of all their elements, depending on the context, which should not cause any confusion. Note that the union of all the groups in $\mathcal{G}$ does not need to contain all the elements of B, in other words, some β_{jk} may not belong to any group. We say B_{g_1} is *nested* in B_{g_2} if $B_{g_1} \subset B_{g_2}$; B_{g_1} and B_{g_2} are *overlapping* if $B_{g_1} \cap B_{g_2}$ is not empty. Obviously, nested groups are a special case of overlapping. A group structure with overlapping groups is common in biological studies. For example, when grouping genetic variants according to genes or pathways, different genes or pathways can overlap with each other.

Though the proposed method works for an arbitrary group structure $\mathcal{G}$ on B, in real applications, a biologically meaningful group structure on B is usually introduced from the group structures of both predictors and responses. Specifically, suppose X has m_1 column groups and Y has m_2 column groups, then they yield $m_1 \times m_2$ intersection block groups on B. We denote this intersection block group structure by $\mathcal{G}_{XY}$, the row block group structure only determined by the predictor groups by $\mathcal{G}_X$, and the nested group structure containing all groups in $\mathcal{G}_{XY}$ and $\mathcal{G}_X$ by $\mathcal{G}_{XY} \cup \mathcal{G}_X$. In the eQTL association study, a nonzero group in $\mathcal{G}_{XY}$ indicates that the corresponding gene group has SNPs associated with expressions in the corresponding pathway group. A nonzero group in $\mathcal{G}_X$ indicates that the corresponding gene group has an effect on some or all of the expressions.

For an arbitrary group structure $\mathcal{G}$ with G groups, let $\sum_{g=1}^{G} \|B_g\|_2$ be the total sum of L_2 norms of every group in $\mathcal{G}$, where $\|B_g\|_2^2 = \sum_{\beta_{jk} \in B_g} \beta_{jk}^2$. The group L_2 norm reduces to the Frobenius norm $\|A\|_2 = \{\mathrm{tr}(A^{\mathrm{T}}A)\}^{1/2}$ for a matrix group A and to the vector L_2 norm $\|a\|_2 = \{a^{\mathrm{T}}a\}^{1/2}$ for a vector group a.

2.3 The regularization method and its properties

2.3.1 The multivariate sparse group lasso

For an arbitrary group structure $\mathcal{G}$ on B, to simplify the notation, we denote $\{g : B_g \in \mathcal{G}\}$ by $\{g \in \mathcal{G}\}$ as long as it does not cause any confusion. For $j = 1, \ldots, p$ and $k = 1, \ldots, q$, let $\lambda_{jk} \geq 0$ be the adaptive lasso tuning parameter for β_{jk}, with $\lambda_{jk} = 0$ if β_{jk} is not penalized. Let $\lambda_g \geq 0$ be the adaptive tuning parameter for group $B_g \in \mathcal{G}$, with $\lambda_g = 0$ if group B_g is not penalized. We consider the following penalized optimization problem for a general regularized multivariate multiple linear regression:

$$\arg\min_{B} \frac{1}{2n}\|Y - XB\|_2^2 + \sum_{1 \leq j \leq p, 1 \leq k \leq q} \lambda_{jk}|\beta_{jk}| + \sum_{g \in \mathcal{G}} \lambda_g \|B_g\|_2, \tag{2.2}$$

where the L_2 penalty term aims to shrink unimportant groups to zero and the L_1 penalty term aims to shrink unimportant entries within an important group to zero. We call it the multivariate sparse group lasso (MSGLasso). We exclude the trivial case that $\lambda_g = 0$ for all $g \in \mathcal{G}$ and $\lambda_{jk} = 0$ for all j, k. To better understand the solution to (2.2), we develop the following theorem for β_{jk} when all other elements in B are fixed.

Theorem II.1. *For an arbitrary group structure $\mathcal{G}$ on B, let $\hat{B}$ be the solution to (2.2) and $\hat{\beta}_{jk}$ be its jk-th element. If for some group $B_{g0} \in \mathcal{G}$ with a tuning parameter λ_{g0},*

$$\sqrt{\sum_{\{jk:\ \beta_{jk} \in B_{g0}\}} (|S_{jk}|/n - \lambda_{jk})_+^2} \leq \lambda_{g0}, \tag{2.3}$$

then $\hat{\beta}_{jk} = 0$ for every $\beta_{jk} \in B_{g0}$. Otherwise, $\hat{\beta}_{jk}$ satisfies

$$\hat{\beta}_{jk} = \frac{\operatorname{sgn}(S_{jk}) \left(|S_{jk}| - n\lambda_{jk}\right)_+}{\|x_j\|_2^2 + n \sum_{\{g \in \mathcal{G}:\ \beta_{jk} \in B_g\}} \lambda_g / \|\hat{B}_g\|_2}, \tag{2.4}$$

where $S_{jk} = x_j^{\mathrm{T}}(Y - X\hat{B}_{(-j)})_{\cdot k}$ with $\hat{B}_{(-j)}$ being the j-th row of $\hat{B}$ replaced by zeros, the subscript $\cdot k$ refers to the k-th column of a matrix, and $a_+ = a$ if $a > 0$ and 0 otherwise.

Note that Theorem II.1 is a general solution form and applies to arbitrary group structures. If there is no group structure assigned on B, then $\mathcal{G}$ becomes an empty set and (2.4) reduces to the lasso solution; If $\lambda_{jk} = 0$ for all j, k, then (2.4) and (2.3) provide the group lasso solution. It is of interest to consider certain special group structures that are intuitive and commonly used in many applications. Specifically, we consider model (2.2) with the following four group structures: (I) $\mathcal{G} = \emptyset$, no group structure assigned on B; (II) $\mathcal{G}_X$; (III) $\mathcal{G}_{XY}$; (IV) $\mathcal{G}_{XY} \cup \mathcal{G}_X$. The corresponding

optimization problems become

$$\arg\min_B \frac{1}{2n}\|Y - XB\|_2^2 + \lambda|B|_1, \tag{2.5}$$

$$\arg\min_B \frac{1}{2n}\|Y - XB\|_2^2 + \lambda|B|_1 + \lambda_1 \sum_{g_1 \in \mathcal{G}_X} \omega_{g_1}^{1/2}\|B_{g_1}\|_2, \tag{2.6}$$

$$\arg\min_B \frac{1}{2n}\|Y - XB\|_2^2 + \lambda|B|_1 + \lambda_2 \sum_{g_2 \in \mathcal{G}_{XY}} \omega_{g_2}^{1/2}\|B_{g_2}\|_2, \tag{2.7}$$

$$\arg\min_B \frac{1}{2n}\|Y - XB\|_2^2 + \lambda|B|_1 + \lambda_1 \sum_{g_1 \in \mathcal{G}_X} \omega_{g_1}^{1/2}\|B_{g_1}\|_2 \tag{2.8}$$

$$+\lambda_2 \sum_{g_2 \in \mathcal{G}_{XY}} \omega_{g_2}^{1/2}\|B_{g_2}\|_2, \tag{2.9}$$

where $|B|_1 = \sum_{jk} |\beta_{jk}|$ is the L_1 norm of B, and ω_{g_1} and ω_{g_2} are some weights, in particular, the group sizes. The tuning parameter $\lambda_{jk} = \lambda$ for all lasso penalties, $\lambda_g = \lambda_1 \omega_{g_1}^{1/2}$ if $g \in \mathcal{G}_X$, and $\lambda_g = \lambda_2 \omega_{g_2}^{1/2}$ if $g \in \mathcal{G}_{XY}$.

In the remaining of this article, we call (2.5) the *Lasso* model, (2.6) the *Lasso+X* model, (2.7) the *Lasso+XY* model, and (2.8) the *Lasso+X+XY* model.

Let $\hat{B}_L$, $\hat{B}_{LX}$, $\hat{B}_{LXY}$ and $\hat{B}_{LXXY}$ be the solutions to (2.5), (2.6), (2.7) and (2.8), respectively. Their corresponding expressions from Theorem II.1 further reduce to some interesting simpler forms under the othornormal design, in particular, $\hat{B}_{LX}$ and $\hat{B}_{LXY}$ are just further shrinkages of $\hat{B}_L$, and $\hat{B}_{LXXY}$ is a further shrinkage of either $\hat{B}_{LX}$ or $\hat{B}_{LXY}$. We are also interested in the group lasso cases where $\lambda = 0$ in (2.6), (2.7) and (2.8), with their solutions denoted by $\hat{B}_{GX}$, $\hat{B}_{GXY}$ and $\hat{B}_{GXXY}$, respectively. Then the main theorems in Yuan and Lin (2006) and Peng et al. (2010) become special cases.

In the eQTL example that we will analyze later, method (2.5) does not take the advantage of knowing the group structure. Method (2.6) only concerns the predictor group structure, therefore can select important gene groups. However, it ignores which pathways those genes are associated with. Method (2.7) considers both predictor and response group structures, therefore can select gene-to-pathway association blocks. Method (2.8) pertains advantages of both (2.6) and (2.7) and is more robust to misspecified group structures.

2.3.2 Oracle inequalities

The lasso method has been shown to achieve the oracle bounds for both prediction and estimation in the multiple linear regression model, which are the error bounds one would obtain if the true model were given, see for example, Bickel et al. (2009). Similar bounds also hold for a total of pq regression coefficients in the multivariate multiple linear regression model with a multivariate mixed L_1/L_2 penalty. For notational simplicity, we consider the following special case of (2.2) with $\lambda_{jk} = \lambda$ for all j, k:

$$\arg\min_B \frac{1}{2n}\|Y - XB\|_2^2 + \lambda|B|_1 + \sum_{g\in\mathcal{G}} \lambda_g \|B_g\|_2. \tag{2.10}$$

We follow the method of Bickel et al. (2009). Let $J_1(B) = \{jk : |\beta_{jk}| \neq 0\}$ be the index set of nonzero elements in B, and $J_2(B) = \{g \in \mathcal{G}, \|B_g\|_2 \neq 0\}$ be the index set of nonzero groups in $\mathcal{G}$. Define $M_1(B) = \sum_{jk} I(\beta_{jk} \neq 0) = |J_1(B)|$ and $M_2(B) = \sum_{g\in\mathcal{G}} I(\|B_g\|_2 \neq 0) = |J_2(B)|$. For any matrix $\Delta \in \mathbb{R}^{p\times q}$ and any given index set $J_1 \subseteq \{jk : 1 \leq j \leq p, 1 \leq k \leq q\}$, denote Δ_{J_1} the projection of Δ on the index set J_1, that is the matrix with the same elements of Δ on coordinates J_1 and zeros on the complementary coordinates J_1^c. Also for any group index set $J_2 \subseteq \{1, \cdots, |\mathcal{G}|\}$, denote Δ_{J_2} the set of projection of Δ on each of $\{B_g : g \in J_2\}$, that is $\Delta_{J_2} = \{\Delta_{B_g} : g \in J_2\}$. Denote $M_1(B) = r$ and $M_2(B) = s$. We then impose a restricted eigenvalue assumption for the multivariate linear regression model with a multivariate mixed L_1/L_2 penalty, which leads to the desirable oracle inequalities.

Assumption II.2. *Let $J_1 \subseteq \{jk : 1 \leq j \leq p, 1 \leq k \leq q\}$ and $J_2 \subseteq \{1, \cdots, |\mathcal{G}|\}$ be any index sets that satisfy $|J_1| \leq r$ and $|J_2| \leq s$. Let $\tilde{\rho} = \{\rho_g : g \in \mathcal{G}\}$ be a set of positive numbers. Then for any nontrivial matrix $\Delta \in \mathbb{R}^{p\times q}$ that satisfies*

$$|\Delta_{J_1^c}|_1 + 2\sum_{g\in J_2^c} \rho_g \|\Delta_{B_g}\|_2 \leq 3|\Delta_{J_1}|_1 + 2\sum_{g\in J_2} \rho_g \|\Delta_{B_g}\|_2,$$

the following minimums exist and are positive:

$$\kappa_1(r,s,\tilde{\rho}) = \min_{J_1,J_2,\Delta\neq 0} \frac{\|X\Delta\|_2}{n^{1/2}\|\Delta_{J_1}\|_2} > 0, \quad \kappa_2(r,s,\tilde{\rho}) = \min_{J_1,J_2,\Delta\neq 0} \frac{\|X\Delta\|_2}{n^{1/2}\|\Delta_{J_2}\|_2} > 0.$$

Theorem II.3. *Consider model (2.10). Let B^* be the true coefficient matrix. Assume each column of the error matrix, w_k, follows a multivariate normal distribution $N(0,\sigma_k I_n)$, and all the diagonal elements of the matrix $X^{\mathrm{T}}X/n$ are equal to 1. Suppose $M_1(B^*) = r$ and $M_2(B^*) = s$. Let $\psi_{\max}$ be the largest eigenvalue of $X^{\mathrm{T}}X/n$, $\sigma = \max\{\sigma_1,\cdots,\sigma_q\}$, $\lambda_g = \rho_g\lambda$ for $g \in \mathcal{G}$, $\rho = \min\{1,\rho_g; g\in\mathcal{G}\}$, c be the maximum number of duplicates of a coefficient in overlapping groups in $\mathcal{G}$, and*

$$\lambda = 2\sigma A\{\log(pq)/n\}^{1/2}$$

for some constant $A > 2^{1/2}$. Furthermore, assume Assumption II.2 holds with $\kappa_1 = \kappa_1(r,s,\tilde{\rho})$ and $\kappa_2 = \kappa_2(r,s,\tilde{\rho})$. Then with probability at least $1-(pq)^{1-A^2/2}$, we have the following oracle bounds for the prediction error, the estimation error and the order of sparsity:

$$\frac{1}{n}\|X(\hat{B}-B^*)\|_2^2 \le 16\lambda^2\left(\frac{r^{1/2}}{\kappa_1} + \frac{\left(\sum_{g\in J_2(B^*)}\rho_g^2\right)^{1/2}}{\kappa_2}\right)^2,$$

$$|\hat{B}-B^*|_1 \le \frac{32(c+2)\sigma A}{1+\rho}\left(\frac{\log(pq)}{n}\right)^{1/2}\left(\frac{r^{1/2}}{\kappa_1} + \frac{\left(\sum_{g\in J_2(B^*)}\rho_g^2\right)^{1/2}}{\kappa_2}\right)^2,$$

$$M_1(\hat{B}) \le 64\psi_{\max}\left(\frac{r^{1/2}}{\kappa_1} + \frac{\left(\sum_{g\in J_2(B^*)}\rho_g^2\right)^{1/2}}{\kappa_2}\right)^2.$$

The mean square prediction error is bounded by a factor of order $\lambda^2 \sim \log(pq)/n$, the l_1 norm of the estimation error is bounded by a factor of order $\sqrt{\log(pq)/n}$, and the estimated order of sparsity is bounded by a constant related to Assumption II.2. These results are similar to those in Bickel et al. (2009). Note that Theorem II.3

will still hold for flexible λ_{jk} in (2.2), as long as $\lambda_{jk} > 0$ for all j, k.

2.4 The mixed coordinate descent algorithm

Based on Theorem II.1, the zero groups can be determined according to (2.3) and the entries in a nonzero group can be determined by solving for the fixed point solution of (2.4) using a coordinate descent algorithm. The coordinate algorithm updates each coefficient coordinate β_{jk} at a step while fixing all the other coefficients at their current values. Theoretically, the coordinate descent algorithm would work if one can solve (2.4) for $\hat{\beta}_{jk}$ exactly. Practically, since $\hat{\beta}_{jk}$ also appears in the term $\sum_{\{g\in\mathcal{G}:\ \beta_{jk}\in B_g,\ \|\hat{B}_g\|_2>0\}} \lambda_g/\|\hat{B}_g\|_2$ on the right hand side of (2.4), unlike lasso, a closed form solution is usually not available and numerically solving for $\hat{\beta}_{jk}$ requires iteratively updating (2.4), which can be time consuming. Here we propose a mixed coordinate descent algorithm, which only updates $\hat{\beta}_{jk}$ once from $\hat{\beta}_{jk}^{(m-1)}$ to $\hat{\beta}_{jk}^{(m)}$ according to (2.4) without iteratively solving (2.4). In particular, the algorithm updates $\hat{\beta}_{jk}$ according to the following.

(I) If any of the groups $B_g \in \mathcal{G}$ containing β_{jk} satisfies (2.3), then the entire group is estimated at zero. Otherwise $\hat{\beta}_{jk}$ will be updated according to one of the following situations (II)-(IV):

(II) If all the groups containing β_{jk} satisfy $\|\hat{B}_{g-(jk)}^{(m-1)}\|_2 = 0$ at the current step, where $\hat{B}_{g-(jk)}^{(m-1)}$ is $\hat{B}_g^{(m-1)}$ with its jkth element replaced by zero, then $\hat{\beta}_{jk}$ is updated by

$$\hat{\beta}_{jk}^{(m)} = \frac{\mathrm{sgn}(S_{jk}^{(m-1)})\left(|S_{jk}^{(m-1)}| - n\sum_{\{g\in\mathcal{G}:\ \beta_{jk}\in B_g,\ \|\hat{B}_{g-(jk)}^{(m-1)}\|_2=0\}} \lambda_g - n\lambda_{jk}\right)_+}{\|x_j\|_2^2}.$$

Notice that in this case, (2.4) becomes a closed form lasso solution.

(III) If all the groups containing β_{jk} satisfy $\|\hat{B}_{g-(jk)}^{(m-1)}\|_2 > 0$ at the current step

and $\lambda_{jk} = 0$, then $\hat{\beta}_{jk}^{(m-1)}$ is updated by the group lasso formulation

$$\hat{\beta}_{jk}^{(m)} = \frac{S_{jk}^{(m-1)}}{\|x_j\|_2^2 + n \sum_{\{g \in \mathcal{G}:\ \beta_{jk} \in B_g,\ \|\hat{B}_{g-(jk)}^{(m-1)}\|_2 > 0\}} \lambda_g / \|\hat{B}_g^{(m-1)}\|_2}.$$

Notice in this case, all the entries in B_g with $\|\hat{B}_{g-(jk)}\|_2 > 0$ will enter as nonzero entries, or in other words, the whole group B_g will be selected as an important group.

(IV) If some but not all groups containing β_{jk} satisfy $\|\hat{B}_{g-(jk)}\|_2 = 0$ at the current step, then $\hat{\beta}_{jk}^{(m-1)}$ belongs to a mixture of the lasso case (for groups with $\|\hat{B}_{g-(jk)}^{(m-1)}\|_2 = 0$) and the group lasso case (for groups with $\|\hat{B}_{g-(jk)}^{(m-1)}\|_2 > 0$), and it is updated as if by a mixture of the lasso and the group lasso through

$$\hat{\beta}_{jk}^{(m)} = \frac{\mathrm{sgn}(S_{jk}^{(m-1)}) \left(|S_{jk}^{(m-1)}| - n \sum_{\{g \in \mathcal{G}:\ \beta_{jk} \in B_g,\ \|\hat{B}_{g-(jk)}^{(m-1)}\|_2 = 0\}} \lambda_g - n\lambda_{jk} \right)_+}{\|x_j\|_2^2 + n \sum_{\{g \in \mathcal{G}:\ \beta_{jk} \in B_g\ \|\hat{B}_{g-(jk)}^{(m-1)}\|_2 > 0\}} \lambda_g / \|\hat{B}_g^{(m-1)}\|_2}.$$

Specifically, the algorithm is given in the following for a fixed set of values of all the tuning parameters.

Step 1. Standardize the data such that

$$\sum_{i=1}^n y_{ik} = 0, \quad \sum_{i=1}^n x_{ij} = 0, \quad \sum_{i=1}^n x_{ij}^2 = 1, \quad \text{for all } j \in \{1, \cdots, p\},\ k \in \{1, \cdots, q\}.$$

In our numerical examples, we also standardize y_k such that $\sum_{i=1}^n y_{ik}^2 = 1$ to minimize the impact of different scales of variations across y_k on the regression coefficients for all $k \in \{1, \cdots q\}$.

Step 2. Set initial values for all $\hat{\beta}_{jk}$ and the iteration index $m = 1$. We use initial values $\hat{\beta}_{jk}^{(0)} = 0$ in our numerical examples.

Step 3. For a given pair (j, k), fix $\beta_{j'k'}$ at $\hat{\beta}_{j'k'}^{(m-1)}$ for all $j' \neq j$ or $k' \neq k$. Then update $\hat{\beta}_{jk}^{(m-1)}$ to $\hat{\beta}_{jk}^{(m)}$ by (I) to (IV) accordingly.

Step 4. Repeat Step 3 for all $j \in \{1, \cdots, p\}$ and $k \in \{1, \cdots, q\}$, and iterate until $\|\hat{B}^{(m)} - \hat{B}^{(m-1)}\|$ reaches a prespecified precision level for some norm $\|\cdot\|$. We use

infinity norm in our numerical examples.

Convergence of different types of coordinate descent algorithms have been studied in the literature. Tseng (2001) provided conditions for convergence of cyclic coordinate descent algorithm with general separable objective functions. Wu and Lange (2008) proved the convergence of greedy coordinate descent algorithm with a L_2 loss and the lasso penalty. Based on Wu and Lange (2008), we show the convergence of our mixed coordinate descent algorithm which is given in the following proposition. Details are provided in the supplemental materials, where we also illustrate that the speed of convergence of our mixed coordinate descent algorithm is much faster than the coordinate descent algorithm that solves the fixed point solution to (2.4) with inner iterations.

Proposition II.4. *A sequence of coordinate estimates iteratively updated by the mixed coordinate descent algorithm converge to a global minimizer of the objective function.*

We implemented the MSGLasso and the mixed coordinate descent algorithm with C/C++ language and wrapped into an `R` package. It is available upon request and will soon be upload to CRAN repository.

2.5 Numerical studies

2.5.1 Simulations

In this section, we first investigate the numerical performances of *Lasso*, *Lasso+X*, *Lasso+XY*, *Lasso+X+XY* methods and their group lasso counterparts when the true coefficient matrix B^* takes a group structure of either $\mathcal{G}_X$, $\mathcal{G}_{XY}$ or $\mathcal{G}_{XY} \cup \mathcal{G}_X$. We also compare the proposed MSGLasso method with lasso and group lasso for an overlapping group structure.

All the true group structures considered in our simulations are given in Fig.2.1 (a)-(d). For each group structure, we consider two scenarios: (i) "all-in-all-out",

where all the coefficients in an important group are important, and (ii) "not-all-in-all-out", where only a subset of coefficients in an important group are important. Specifically, we generate B^* by setting $\beta^*_{jk} = 0$ if it is from an unimportant group, and drawing its value from a uniform distribution on $[-5, -1] \cup [1, 5]$ and fixing it for the simulations if it is from an important group. The sparsity of an important group in the "not all in all out" setting is randomly set between 1/4 and 1/6.

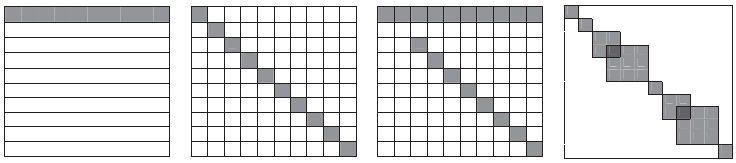


Figure 2.1: B^* group structures. Important groups are shaded. (a) X group structure, (b) XY group structure, (c) X+XY group structure (nesting group structure) and (d) overlapping group structure.

Each B^* is of dimension 200×200. For a nonoverlapping group structure, each X row group is of dimension 20×200; each XY block group is of dimension 20×20. For the overlapping group structure, the groups start on coordinates $(1, 21, 41, 61, 101, 121, 141, 181)$ and end on coordinates $(20, 40, 70, 100, 120, 150, 180, 200)$, for both X and Y variables.

Covariates $X_{i\cdot}^{\mathrm{T}}$, $i = 1, \cdots, n$, are generated from a multivariate normal distribution $N_p(0, \Sigma_X)$, where $\Sigma_X = diag(\Sigma_{g_1}, \cdots, \Sigma_{g_{10}})$ is block diagonal and each block corresponds to each group of X which has the first order autoregressive structure. Specifically, $\Sigma_{g_i}(j, k) = \rho^{|j-k|}$ for any j, k pair from the same group, $i = 1, \cdots, 10$. The error terms w_{ik} are generated from a normal distribution $N(0, \sigma^2)$, where σ^2 is to yield a signal to noise ratio of 2. Finally, the responses are generated from $Y = XB^* + W$.

The optimal values of tuning parameters may be selected by different criteria. Since the degrees of freedom are difficult to determine for a penalty with multiple tuning parameters, we search for the optimal tuning parameter values using a 5-fold

cross-validation over a wide range of candidate values. The searching process starts with the largest candidate tuning parameter values with each by itself shrinking all the coefficients to zero. The converged estimates $\hat{B}$ obtained from the previous searching step are used as the initial values for B in the next searching step with a new set of tuning parameter values. We find it is very effective in reducing the computational cost.

For each simulation setup, we run a hundred replications and calculate the averages of the following quantities:

$$\begin{aligned}
&\text{false positives} = |\{ij \text{ pairs} : \hat{\beta}_{ij} \neq 0 \text{ and } \beta^*_{ij} = 0\}|,\\
&\text{false negatives} = |\{ij \text{ pairs} : \hat{\beta}_{ij} = 0 \text{ and } \beta^*_{ij} \neq 0\}|,\\
&\text{sensitivity} = \frac{|\{ij \text{ pairs} : \hat{\beta}_{ij} \neq 0 \text{ and } \beta^*_{ij} \neq 0\}|}{|\{ij \text{ pairs} : \beta^*_{ij} \neq 0\}|},\\
&\text{specificity} = \frac{|\{ij \text{ pairs} : \hat{\beta}_{ij} = 0 \text{ and } \beta^*_{ij} = 0\}|}{|\{ij \text{ pairs} : \beta^*_{ij} = 0\}|},\\
&\text{prediction error} = \|Y_{\text{test}} - X_{\text{test}}\hat{B}\|_2^2\,,
\end{aligned}$$

where $|\cdot|$ is the number of elements in a set and $(Y_{\text{test}}, X_{\text{test}})$ is an independently generated testing set of 100 samples.

Figure 2.2 summarizes these quantities for simulation setups with "not all in and all out" for all the group structures in Fig.2.1 at $p = q = 200$, $n = 150$, and $\rho = 0.5$. The proposed method using *Lasso+X+XY* for the nonoverlapping group structures $\mathcal{G}_X$, $\mathcal{G}_{XY}$ and $\mathcal{G}_{XY} \cup \mathcal{G}_X$ as well as for the overlapping group structure are highlighted in black. The methods for the correctly specified group structures are highlighted in grey except in Fig.2.2 and Fig.2.2, where the implemented group structures are by themselves the correctly specified group structures. From Fig.2.2 we see that correctly incorporating group structure improves both variable selection and prediction, and our proposed method *Lasso+X+XY*, or the MSGLasso, performs at least the same as, if not better than, the methods for the correct group structures and yields the lowest prediction errors.

Figure 2.3 illustrates fitted results for a data set randomly chosen from one hun-

dred replications, where B^* has a "not all in all out" either $\mathcal{G}_{XY} \cup \mathcal{G}_X$ or overlapping group structure with $p = 200$, $q = 200$ and $\rho = 0.5$. It clearly shows that the MSGLasso results for correctly specified group structure, both in Fig.2.11 and in Fig.2.3, yield the most desirable estimates. Methods without lasso penalty yield too many false positives inside the important groups for the "not all in all out" case even when the groups are correctly specified, while methods with lasso penalty but incorrectly specified groups yield too many false positives outside the important groups.

Figures 2.4 and 2.5 illustrate the comparisons between univariate approaches and multivariate approaches. The true regression coefficient matrix takes a $\mathcal{G}_{XY} \cup \mathcal{G}_X$ group structure. It can be seen that when different response variables have a similar sparsity to the predictors, the multiple univariate lasso (using different λ values for different response variables) and the multivariate lasso (using the same λ value for all response variables) have similar performance on variable selection. The multiple univariate sparse group lasso approach has a slightly better variable selection performance than the multiple univariate lasso. The proposed multivariate sparse group lasso yields the best variable selection result by borrowing information from other response variables within the same group. It also has the smallest prediction error.

Figures 2.7 to 2.11 at the end of this chapter show the variable selection and prediction effects in some other simulation settings, such as with different autocorrelation coefficient values or with a true "all-in-all-out" group structure.

2.5.2 Yeast eQTL data analysis

In this section, we demonstrate our method by analyzing a yeast eQTL data set generated by Brem and Kruglyak (2005), see also Yin and Li (2011), where gene expressions are grouped into, possibly overlapping, pathways and the genetic markers are grouped into genes.

The data set contains 6216 yeast genes assayed for 112 individual segregant.

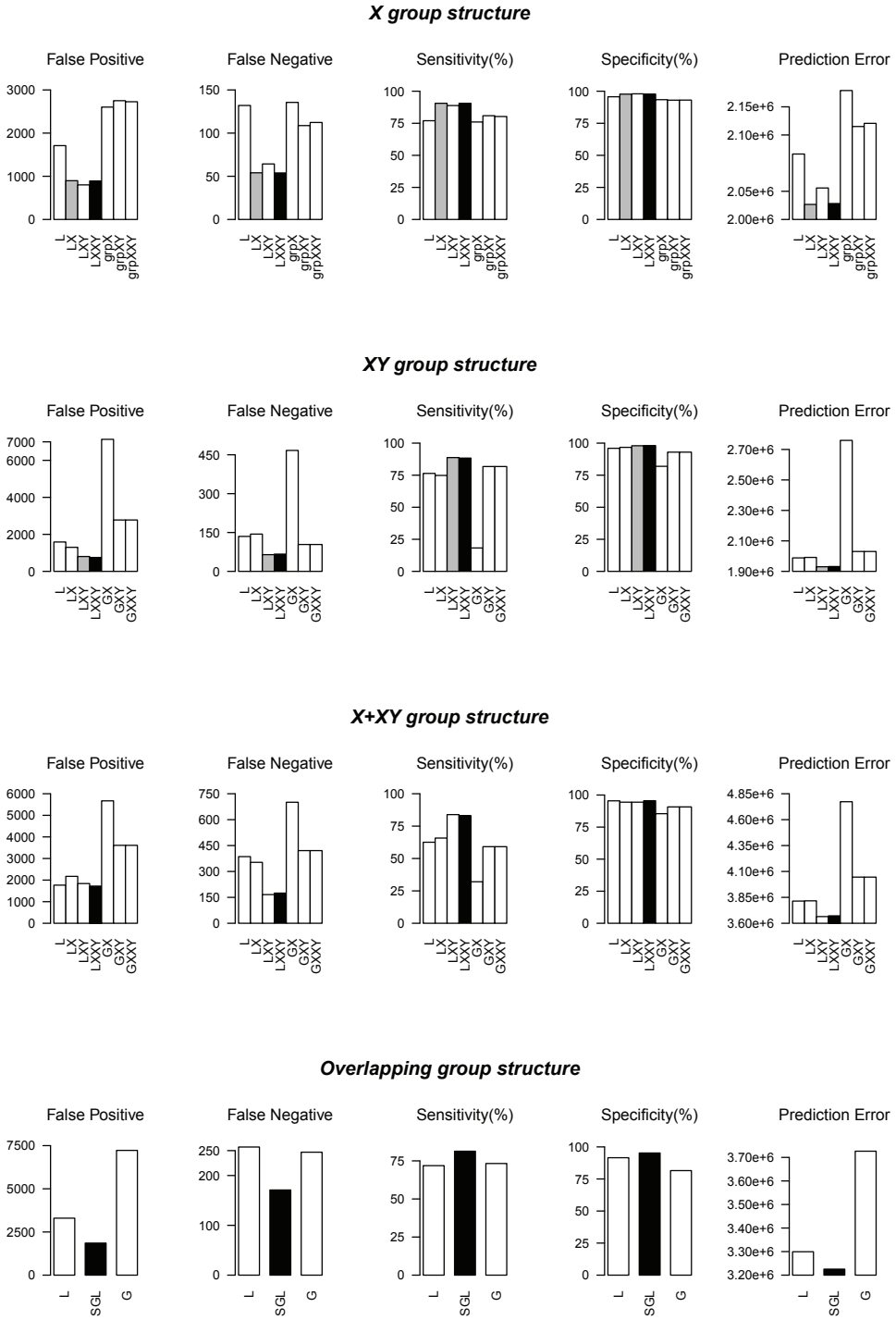


Figure 2.2: Simulation results, large p small n, "not all in all out" cases with $n = 100$, $p = q = 200$ and $\rho = 0.5$. SGL: the multivariate sparse group lasso; G: the multivariate group lasso.

Genotypes of these 112 segregant at 2956 marker positions were also collected using GeneChip Yeast Genome S98 microarrays. The 6216 expressed genes are grouped

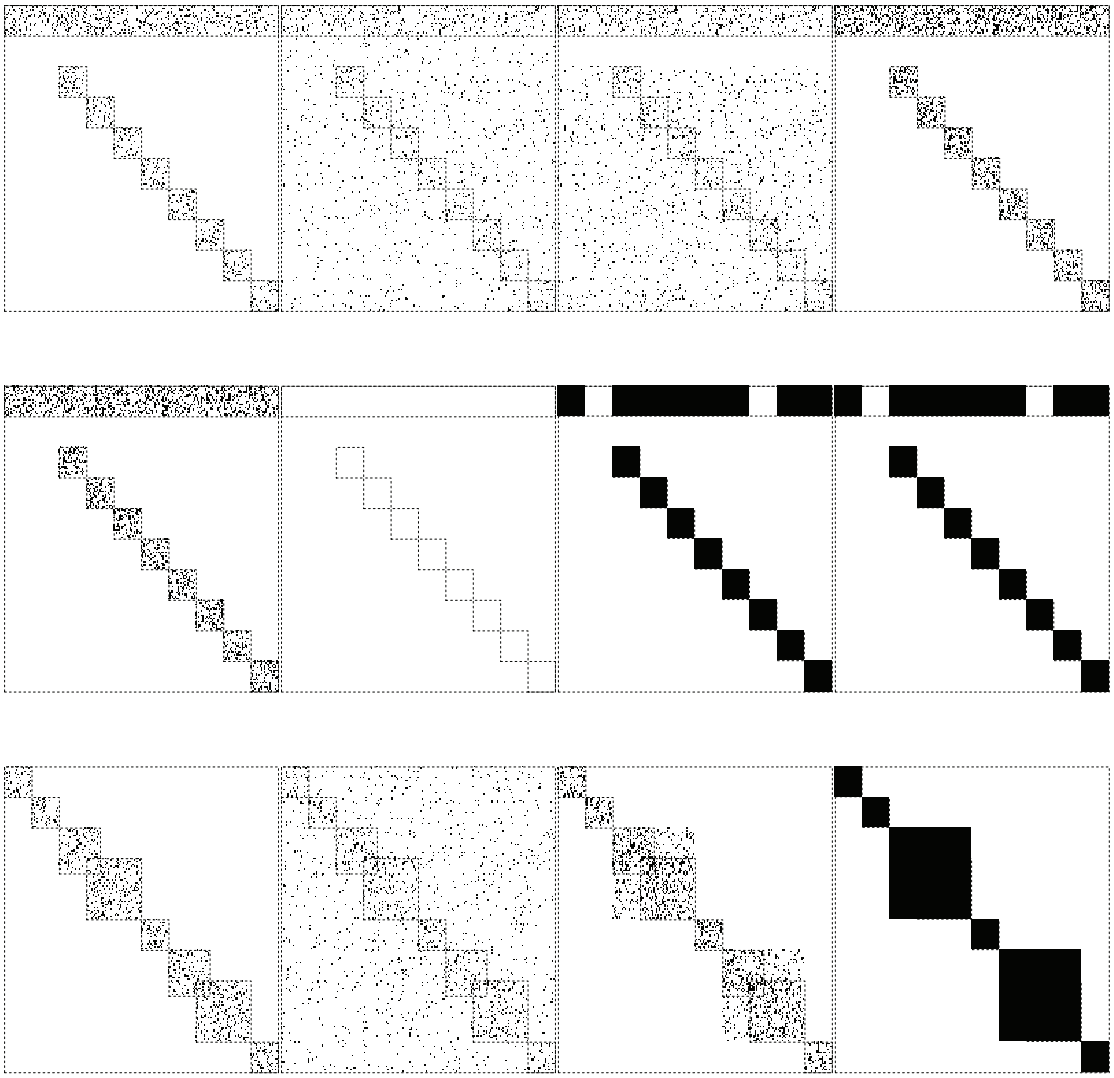


Figure 2.3: Heatmaps of coefficient matrices, selection effects. (a)-(h): "Not all in all out" $X{+}XY$ nonoverlapping group structure with $n = 100$, $p = 200$, $q = 200$, and $\rho = 0.5$. (a) B^*; (b) $\hat{B}_L$; (c) $\hat{B}_{LX}$; (d) $\hat{B}_{LXY}$; (e) $\hat{B}_{LXXY}$; (f) $\hat{B}_{GX}$; (g) $\hat{B}_{GXY}$; (h) $\hat{B}_{GXXY}$. (i)-(l): "Not all in all out" overlapping group structure with $n = 100$, $p = 200$, $q = 200$, and $\rho = 0.5$. (i) B^*; (j) $\hat{B}_L$; (k) $\hat{B}_{SGL}$; (l) $\hat{B}_G$.

by Kyoto Encyclopedia of Genes and Genomes pathways and the 2956 markers are grouped by genes, taking isoform genes as the same gene. To illustrate the method, in the reported analysis we only include genes from the following four pathways: the *mitogen-activated protein kinases (MAPK)* pathway containing 54 genes, the *cell cycle* pathway containing 116 genes, the *cancer* pathway containing 20 genes and the *ribosome* pathway containing 137 genes. There are in total 315 distinct expressed

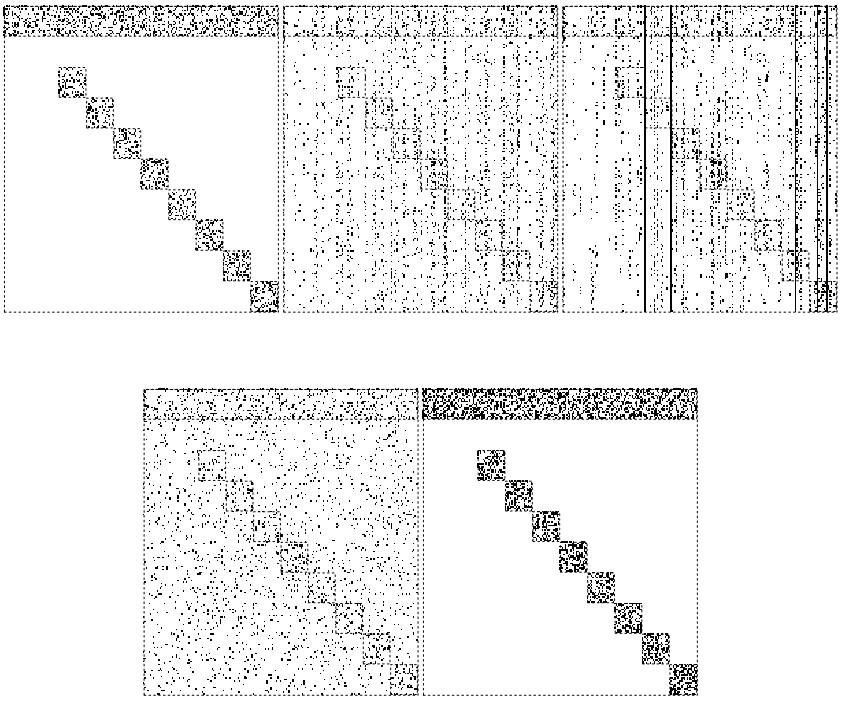


Figure 2.4: Heatmaps of coefficient matrices. (a) True B^*; (b) The multiple univariate lasso; (c) The multiple univariate sparse group lasso (d) The multivariate lasso; (e) The multivariate sparse group lasso; The true B^* has a "not all in all out" and $X{+}XY$ group structure with $p = q = 200$, $n = 100$, $\rho = 0.5$.

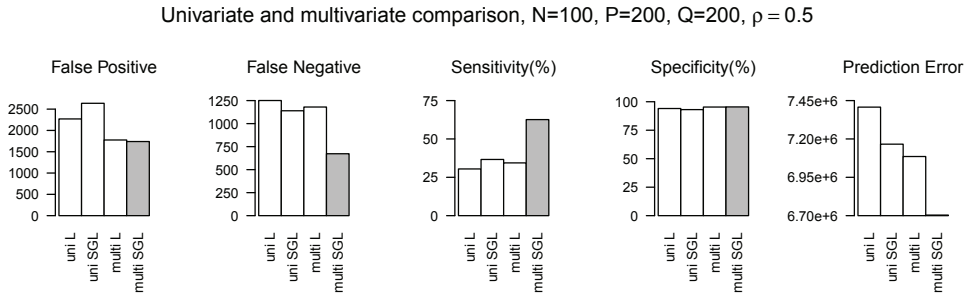


Figure 2.5: Comparison between multiple-univariate and multivariate approaches from 100 simulated data sets. "uni L" – the multiple univariate lasso; "uni SGL" – the multiple univariate group lasso; "multi L" – the multivariate lasso; "multi SGL" – the multivariate sparse group lasso with an XY group structure on the coefficient matrix.

genes in these pathways, with 5 genes overlapping between *MAPK* and *cell cycle*, 5 genes overlapping between *MAPK* and *cancer*, 3 genes overlapping between *cell cycle* and *cancer*, and 1 gene overlapping between *MAPK*, *cell cycle* and *cancer*.

Ribosome does not contain overlapping genes with the other three pathways.

We follow a similar procedure of Yin and Li (2011) for prescreening genotype markers by performing univariate linear regressions across all the 315 gene expressions and 2956 markers, and include the 395 markers with p-value of 0.01 or smaller into the final analysis. These 395 markers are embedded in 45 distinct genes.

Since some marker within a gene is associated with some gene expression in a pathway does not necessarily imply the gene must be associated with all four pathways, we exclude the $\mathcal{G}_X$ group structure and only apply an overlapping $\mathcal{G}_{XY}$ group structure in the data analysis. We cross-validate the performance of the multivariate sparse group lasso, the multivariate lasso, the multivariate group lasso and the univariate lasso. In particular, we randomly divide the 112 samples into five approximately equal sized subsets, set one subset aside as the test set, and use the remaining four subsets as the training set. Then for each model, we run 5-fold cross-validation on the training set to estimate the coefficient matrix, and use the estimated model to compute the prediction error on the test set. We repeat the above procedures until each of the five subsets has been used as the test set once. The overall cross-validated prediction errors, the sum of squares, are reported in Table 2.1. The univariate lasso is conducted by first selecting variables on the training set using 315 separate lasso regressions, each for a single gene expression variable, and then implementing multivariate linear regression on only the selected set of covariates to obtain $\hat{B}$. Our proposed method has the best performance. The univariate lasso gives the highest prediction error, this is expected as the relations among responses are totaly overlooked. And this leads to high variability and overfitting (Peng et al., 2010). The proposed method shows roughly a 10% decrease of the cross-validated prediction error over the multivariate lasso method, the second best approach among all four compared methods.

We then apply the multivariate sparse group lasso to the entire data set with 315 gene expressions and 395 markers. The final tuning parameters are $\lambda = 7 \times 10^{-2}$ and $\lambda_1 = 2 \times 10^{-4}$ determined by a 5-fold cross-validation. We also investigate the

Table 2.1: Comparison of prediction errors between different methods

Method	MSG lasso	M lasso	MG lasso	lasso
Prediction error	3094.5	3396.8	3557.4	3683.3

MSG lasso = multivariate sparse group lasso, M lasso = multivariate lasso, MG lasso = multivariate group lasso, lasso = univariate lassos.

selection stability following Meinshausen and Bühlmann (2010) by calculating the selection frequencies of the top selected associations using one hundred bootstrap datasets. The top associations in terms of size, with selection frequency no less than 95%, are given in Table 2.2. The p-values in the last column are obtained from marginal simple linear regressions. Overall there are 1422 nonzero elements in the estimated coefficient matrix, which gives an overall estimated sparsity of about 1%. There are 235 markers with nonzero coefficients related to genes in the *MAPK* pathway, 135 markers related to genes in the *cell cycle* pathway, 65 markers related to genes in the *cancer* pathway, and 65 markers related to genes in the *ribosome* pathway. Among those, 34 markers are related to genes in the overlap of *MAPK* and *cell cycle* pathways, 23 markers are related to genes in the overlap of *MAPK* and *cancer* pathways, and 5 markers is related to a gene in the overlap of *MAPK*, *cell cycle* and *cancer* pathways.

Table 2.3 lists the top pathway-gene groupwise associations in terms of the group L_2 norms with a 100% group-wise selection frequency. Out of 180 block groups, 89 groups contain nonzero coefficients. Several top selected genes have been reported in the literature. For example, one of the isoforms of *YCR* gene, *YCR073C*/*SSK22* is *MAPK* cascade involved in osmosensory signaling pathway. Gene groups *YJL* and *YGR* in the Scr homology 3 domains are interacting with gene *Pbs2* in one of the three kinase components in the *MAPK* pathway (Zarrinpar et al., 2003). The top association signals detected between the gene expressions in the joint of *MAPK*, *cell cycle* and *cansor* pathways and markers in *NHR* gene group also confirm the regulation effects of *NHR* genes on *cell cycle* pathway and other autophagy-related genes (Nicole, 2011).

It worth noting that none of the association p-values from marginal simple linear regressions between gene *YJL* and pathway *MAPK* survives the Bonferroni correction for multiple comparisons. For example, the 14^{th} signal in Table 2.2 has a univariate marginal p-value of 0.044, therefore it is very unlikely to be picked up by the pairwise analysis. However, the MSGLasso successfully selected this signal in an adjusted analysis with high individual and group selection frequencies given in Tables 2.2 and 2.3. This finding is supported by Zarrinpar et al. (2003). It demonstrates that besides the advantage of dimension reduction, the MSGLasso can also pick out important signals that would be missed by the pairwise method.

Figure 2.6 shows the eQTL network between the gene expressions and the genetic markers constructed from the multivariate sparse group lasso method.

The stability selection results show that the first 40 selected top signals do not contain zero within their 2.5%-97.5% bootstrap percentile band, and the bootstrap Q1-Q3 band of the top 100 selected signals do not contain zero, indicating that the top selected signals using proposed method have high selection frequencies from bootstrap samples.

Table 2.2: Top selected expression-marker associations

Id	$\hat{\beta}_{jk}$	Sel. Freq.*	Expr.** name	Expr. pathways	Marker Chr:BP***	Marker gene	p-value
1	-1.481	100	*YKL178C*	*MAPK*	3:201166	*YCR041W*	2.4e-51
2	1.465	100	*YFL026W*	*MAPK*	3:201166	*YCR041W*	2.8e-55
3	-1.264	100	*YPL187W*	*MAPK*	3:201166	*YCR041W*	7.1e-45
4	1.061	100	*YNL145W*	*MAPK*	3:201166	*YCR041W*	5.5e-39
5	-0.735	100	*YGL089C*	*MAPK*	3:201166	*YCR041W*	8.5e-20
6	0.650	100	*YFL026W*	*MAPK*	3:201167	*YCR041W*	2.8e-55
7	-0.649	100	*YKL178C*	*MAPK*	3:201167	*YCR041W*	2.4e-51
8	-0.554	98	*YPL187W*	*MAPK*	3:201167	*YCR041W*	7.1e-45
9	0.452	100	*YDR461W*	*MAPK*	3:201166	*YCR041W*	8.4e-14
10	-0.385	98	*YPL187W*	*MAPK*	3:177850	*gCR02*	1.7e-33
11	0.352	100	*YGR088W*	*MAPK*	15:170945	*gOL02*	1.5e-10
12	0.346	100	*YGR088W*	*MAPK*	15:174364	*gOL02*	1.5e-10
13	-0.318	97	*YKL178C*	*MAPK*	3:177850	*gCR02*	2.4e-37
14	0.257	98	*YGR088W*	*MAPK*	10:51003	*YJL204C*	0.044
15	-0.175	95	*YGL089C*	*MAPK*	2:681361	*YML056C*	0.66

* Sel. Freq. = Selection Frequency in %. ** expr. = gene expression. *** Marker is denoted by its physical position in the format of "chromosome:basepair".

Table 2.3: Top selected pathway-gene associations (with 100% selection frequency)

Id	Pathway	Gene	$\|\hat{B}_g\|_2$	$\|\hat{B}_g\|_0^*$	Top expr.** in pathway	Top marker*** in gene	Top $\hat{\beta}_{jk}$ in group
1	*MAPK*	*YCR*	3.06	23	*YKL178C*	3:201166	-1.48
2	*MAPK*	*gOL*	0.508	10	*YGR088W*	15:170945	0.35
3	*MAPK*	*gCR*	0.499	3	*YPL187W*	3:177850	-0.39
4	*MAPK*	*YJL*	0.424	23	*YGR088W*	10:51003	0.26
5	*MAPK*	*NHR*	0.420	49	*YCL027W*	8:111686	-0.18
6	*MAPK*	*NBR*	0.382	15	*YGL089C*	2:681361	0.21
7	*MAPK*	*YBR*	0.372	81	*YGR088W*	2:368060	0.17
8	*ribosome*	*YER*	0.342	119	*YER102W*	5:350744	-0.06
9	*cancer*	*YLR*	0.286	14	*YJR048W*	12:674651	0.16
10	*MAPK*	*YGR*	0.275	3	*YGL089C*	7:916471	-0.17
11	*MAPK*	*YPL*	0.274	18	*YGR088W*	12:428612	0.24
12	*MAPK*	*YLR*	0.252	62	*YCL027W*	12:957108	0.09
13	*MAPK*	*YER*	0.229	23	*YPL187W*	7:321714	0.14
14	*MAPK*	*YML*	0.214	23	*YGL098C*	13:164026	-0.18
15	*MAPK*	*YHL*	0.205	15	*YKL178C*	8:98513	-0.13
16	*MAPK*	*YNL*	0.183	23	*YGL089C*	14:418269	-0.08
17	*MAPK*	*YCL*	0.176	27	*YCL027W*	3:64311	0.14
18	*MAPK; cell cycle*	*NHR*	0.175	44	*YJL157C*	8:111686	-0.061
19	*MAPK*	*gJL*	0.131	9	*YFL026W*	10:259991	0.098
20	*MAPK*	*YOL*	0.125	26	*YPL187W*	15:193911	0.084
21	*MAPK; cell cycle; cancer*	*NHR*	0.098	5	*YBL016W*	8:111686	-0.044
22	*cell cycle*	*YCR*	0.067	5	*YLR288C*	3:201166	0.046
23	*cell cycle*	*YCL*	0.063	16	*YDL003W*	3:64311	-0.035
24	*cell cycle*	*YLR*	0.029	37	*YBR093C*	12:674651	0.012

* $\|\hat{B}_g\|_0$ = number of nonzero $\hat{\beta}_{jk}$ in group. ** expr. = gene expression. *** Top marker in gene is denoted by its physical position in the format of "chromosome:basepair".

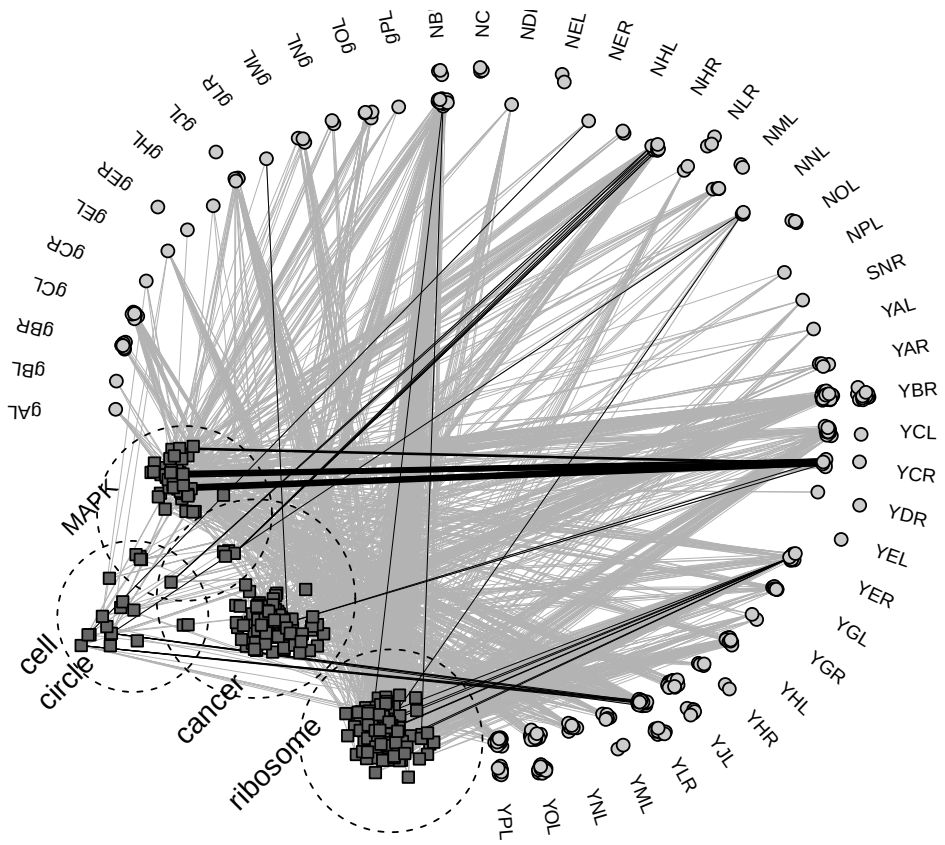


Figure 2.6: Network constructed from the multivariate sparse group lasso method. Network structure is between gene expressions grouped in *mitogen-activated protein kinases (MAPK)*, *cell cycle*, *cancer*, *ribosome* pathways and markers grouped in 45 gene groups. Gray lines connect expression-marker pairs with non-zero $\hat{\beta}_{jk}$. Dark lines are for the top 10 associations in each pathways. The strength of these top associations are indicated by the width of the dark lines. The dotted circles indicate the overlapping pathway group structure.

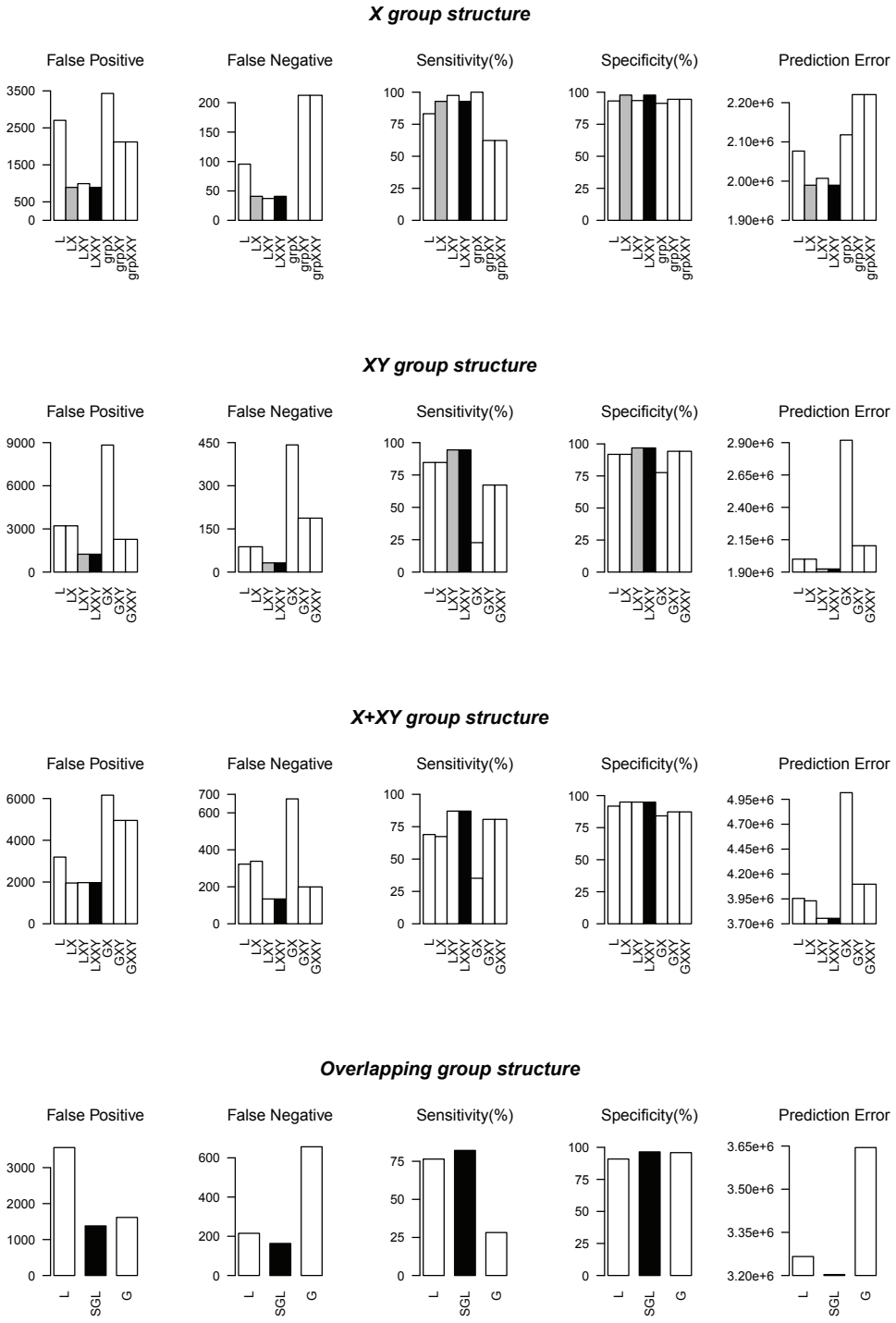


Figure 2.7: More simulation results, "not all in all out" cases with $n = 150$, $p = q = 200$ and $\rho = 0.2$.

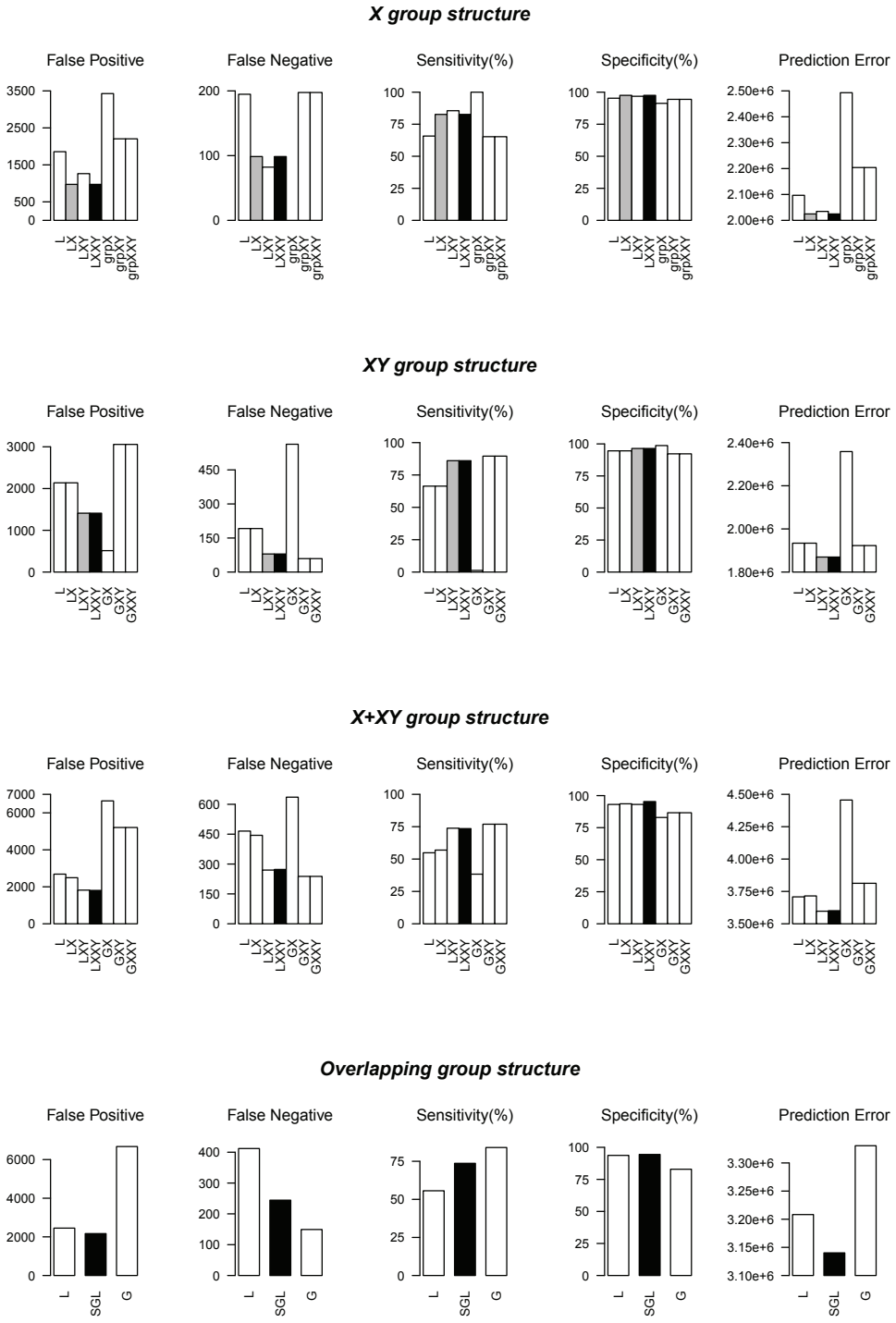


Figure 2.8: More simulation results, "not all in all out" cases with $n = 150$, $p = q = 200$ and $\rho = 0.8$.

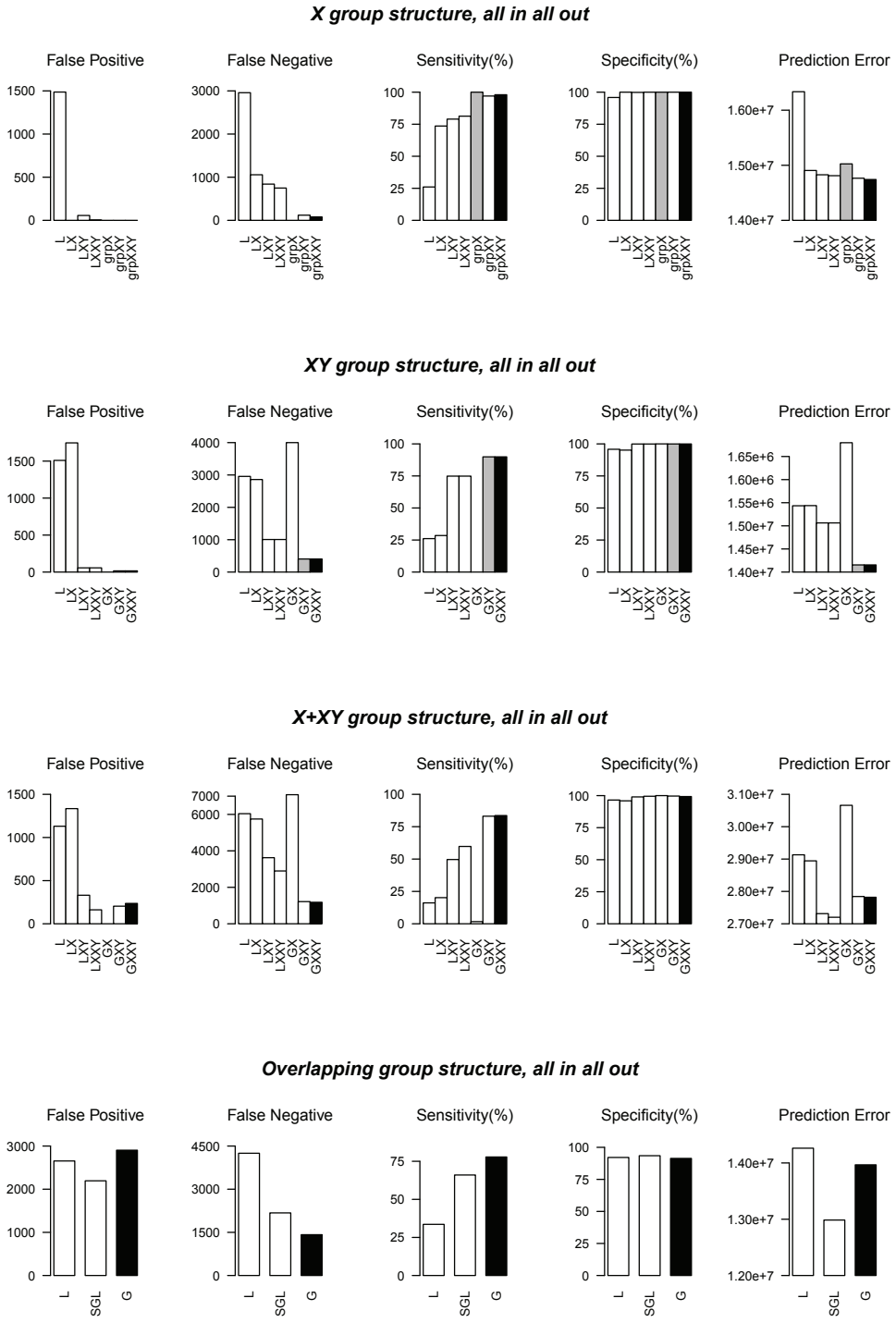


Figure 2.9: More simulation results, "all in all out" cases with $n = 150$, $p = q = 200$ and $\rho = 0.5$.

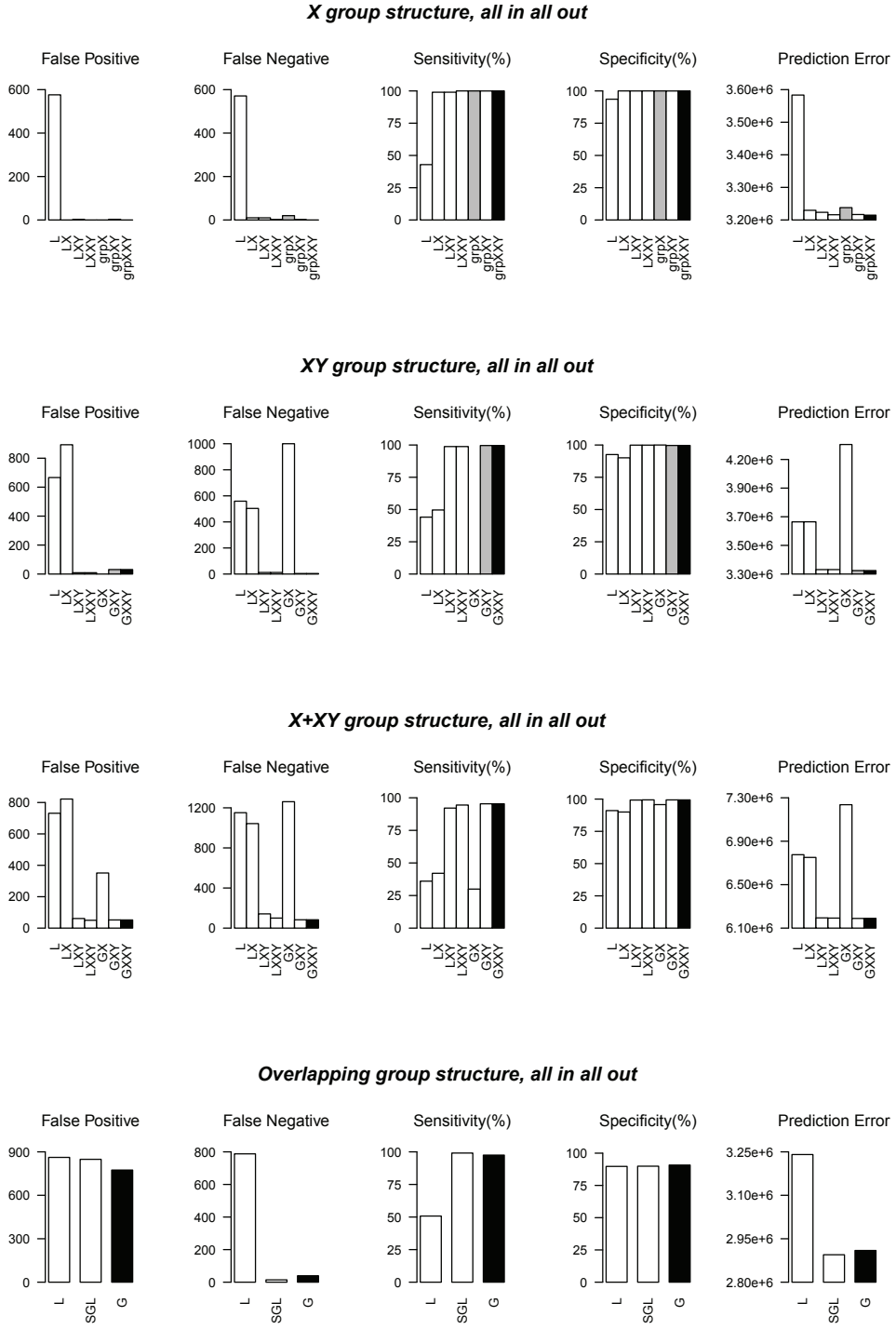


Figure 2.10: More simulation results, "all in all out" cases with $n = 150$, $p = q = 100$ and $\rho = 0.5$.

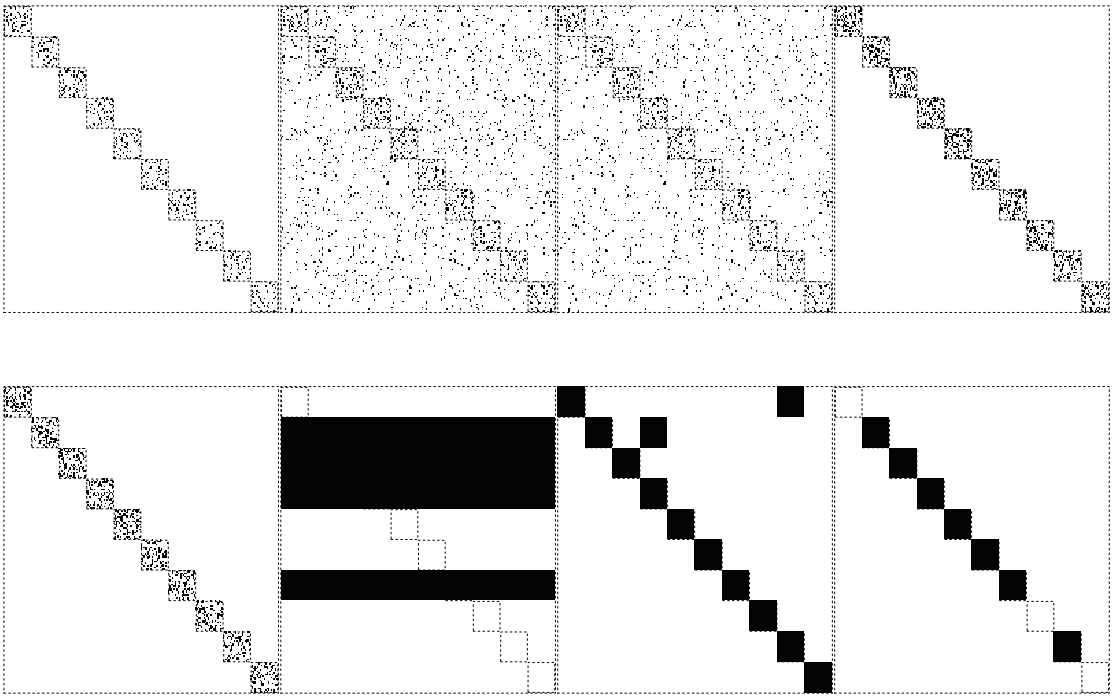


Figure 2.11: Heatmaps of coefficient matrices, selection effects. "Not all in all out" XY group structure with $n = 100$, $p = 200$, $q = 200$, and $\rho = 0.5$. (a) B^*; (b) $\hat{B}_L$; (c) $\hat{B}_{LX}$; (d) $\hat{B}_{LXY}$; (e) $\hat{B}_{LXXY}$; (f) $\hat{B}_{GX}$; (g) $\hat{B}_{GXY}$; (h) $\hat{B}_{GXXY}$.

CHAPTER III

A structured brain-wide and genome-wide association study via multivariate sparse group lasso using ADNI PET images

3.1 Introduction

Human brain structures are highly heritable (Braber et al., 2013; Peper et al., 2007). The modern technologies of neuroimage scans and next generation sequencing of human genomes have both provided powerful phenotypic and explanatory resolutions to detect associations between common genetic variants and human brain structures. However, given the enormous numbers of variables in both the imaging data and genotype data, it is an extremely huge computational burden to jointly analyze the brain-wide and the genome-wide data. It is well recognized that single-response-to-single-predictor type of approaches, followed by multiple comparison adjustment, have limited power to detect true signals, especially in ultrahigh-dimensional settings or when both responses and predictors have complicated structures. Many of the current brain-wide genome-wide association studies (GWAS's) analyze one single-voxel to single-genetic variant at a time and ignore the intrinsic grouping (or functional correlation) structures embedded in either the human brain or the genome (Stein et al., 2010a).

As pointed out, such approaches are especially powerless to detect signals from ultrahigh dimensional data, due to the fact that they usually need to correct for huge

numbers of multiple comparisons, as the number of either the imaging variables or the genetic variables can easily exceed hundreds of thousands while the sample sizes are usually of hundreds to thousands. Some other approaches, including shrinkage estimators in multiple regressions with high dimensional predictors (Kohannim et al., 2012) or gene based regressions (Hibar et al., 2011), have been implemented for a single response variable. However, to the best of our knowledge, there is no current method that can jointly analyze the entire human brain and the genome, and take into consideration of both the brain and genome functional structures at the same time.

In this chapter, we propose a multi-stage method for selecting and estimating important association signals from ultra-high dimensional multivariate multiple structured data. We analyze brain-wide and genome-wide-association-study (GWAS) data with Fluorine-uorodeoxiglucose positron emission tomography (PDG-PET) images and DNA genotypes from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. PDG-PET images measure brain glucose metabolism, and can reflect changes of glucose metabolism as diagnostics of disease progresses, such as the Alzheimer's disease (AD) (Mosconi, 2005). The proposed method can efficiently select important associations between brain image and genetic variants. It advances the univariate approaches in at least two aspects. First, it is a joint multivariate multiple regression approach that avoids the huge number of multiple comparisons of the univariate approach. Secondly, it is a structured approach, which takes into consideration the intrinsic functional grouping structures in both the brains and the genomes, specifically, the anatomical brain functional regions and gene or pathway structures in the genome. Taking such group structures into consideration will implicitly increase the strength of modeling the correlation and multicollinearity among the responses and predictors, and therefore reduces the number of false discoveries at either the group level or within group level (Li et al., 2013).

In the considered brain wide GWAS, each response image consists of $349,182$ voxels and each genome consists of about $560,000$ single nucleotide polymorphisms

(SNPs). Both the image and the genotype data are available for 373 subjects in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database.

The proposed brain wide GWAS method consists of two selection stages and one estimation stage. The first selection stage aims to select the important brain-region-to-gene association signals. To reduce the ultra-high dimensions in both the image data and genetic data, we summarize the voxel level measures within each brain region and genotypes within each gene respectively using major principal components, then apply the multivariate sparse group lasso (MSGLasso) (Li et al., 2013) to select important brain region to gene associations. Then in the second selection stage, the multivariate lasso is applied to the original (normalized) voxel level image measures and genotypes in each selected brain region and gene pair. The stability selection (Meinshausen and Bühlmann, 2010) is used in both selection stages. In the final stage, we estimate the effects of selected SNPs (from the second stage selection) on each selected voxel and infer their significance by the standard multiple linear regressions. We regress only the relevant predictor SNPs selected from the selection stages on each selected voxel. Compared to the conventional single-voxel-to-single-SNP approaches with about $349,182 \times 560,000$ multiple hypothesis tests, our approach significantly reduces the number of hypothesis tests by only making inference on the selected voxel-to-SNP signals. Since the proposed method is a joint multivariate -response-multiple-predictor approach and it also incorporates the intrinsic grouping structures in both responses and predictors, such as biological anatomic brain and genetic grouping structures, it has more power to detect the true association signals compared to the univariate-response-multiple-predictor or univariate-response-single-predictor approaches (Li et al., 2013). It worth pointing out here that despite their prevalence, the post-selection estimation and inference could potentially provide biased estimation or invalid inference results (Berk et al., 2013).

Computationally, the proposed method is in general more efficient compared to the single-voxel-single-SNP approaches (Stein et al., 2010a). The major computa-

tional cost saving comes from the dimension reduction in the first selection stage and the fact that we only make inference on the selected signals.

3.2 ADNI data

The data set used in the brain-GWAS analysis contains two parts: the imaging data and the genetic data, both from the ADNI database. Samples with both imaging and genotype data are included in the analysis, which result in a data set with 373 samples including 86 Alzheimer's disease (AD) patients, 188 mild cognitive impairment (MCI) patients and 99 normal controls (NC).

3.2.1 PET images and ROI's

Images used in our analysis are FDG-PET images, which have been widely used in neuroimage studies for over 20 years. FDG-PET images measure cerebral glucose metabolic activity. Since year 2003, ADNI has acquired in total 403 FDG-PET scans at approximately 50 different participating sites, including 95 subjects with AD, 206 subjects with MCI and 102 NC subjects. Due to missing genetic information, only 373 individuals are included in our study. Each image used in our study contains $349,182$ voxels embedded in a $160 \times 160 \times 96$ 3D array. Those images were preprocessed to produce a uniform isotropic resolution.

In many brain image analyses, the voxel level data can be grouped into region-of-interests (ROI) based on brain anatomic structures. PET images used in our analysis were segmented by Brodmann atlas (Brodmann, 2010). As a result, the voxels in each image were grouped into 106 Brodmann ROI areas, which constitutes the brain anatomic group structure. Figure 3.1 illustrates these regions of interest and their positions in brain. The voxels not indexed by the Brodmann atlas are not used in the analysis. The regions on the left hemisphere brain are symmetric mirror reflection of the ones on the right hemisphere. In the following context, we use "(L)" to denote the regions on the left brain hemisphere and "(R)" to denote the regions on the right brain hemisphere. For example, "Temporal cortex_BA20(L)"

refers the temporal cortex region "BA20" on the left hemisphere and "Temporal cortex_BA20(R)" refers the corresponding symmetric region on the right hemisphere.

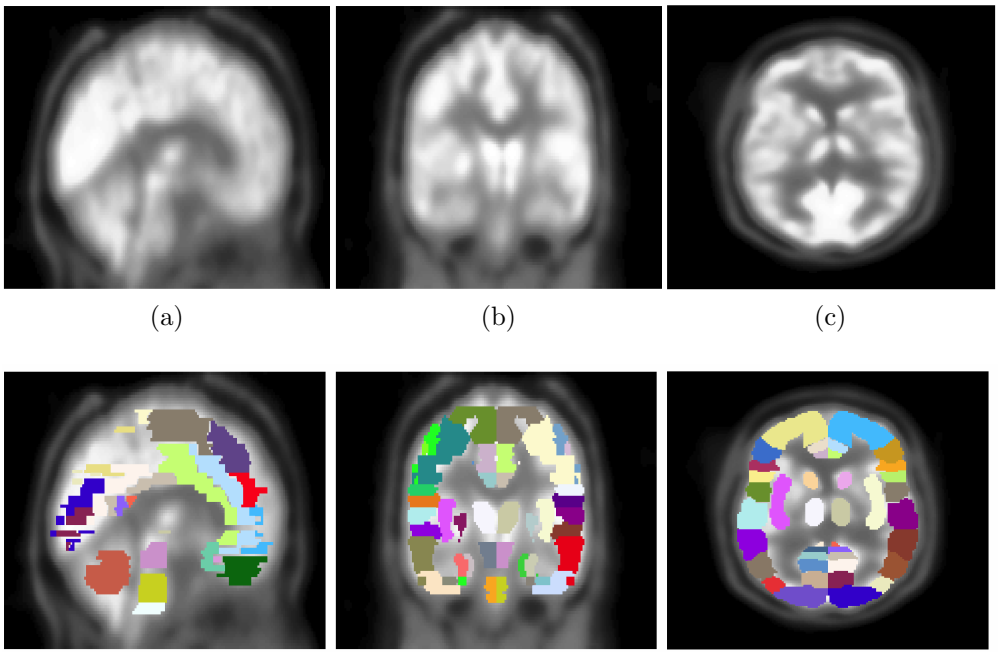


Figure 3.1: Illustration of mapping Brodmann atlas of ROI's onto segmented PET images. ROI's are highlighted with colors. The upper panel are images for a normal sample, the lower panel are the same images overlapped with Brodmann atlas ROI's. (a,d) Sagittal slice at midline. (b,e) Coronal slice at midline. (c,f) Axial slice at midline.

3.2.2 Genotypes

The ADNI DNA data were genotyped using Illumina 610 Quad array with more than $620,000$ tag SNPs. Genotyping was performed by Polymorphic DNA Technologies. We grouped SNP genotypes into genes using the UCSC known genes list of NCBI36 assembly (http://genome.ucsc.edu), with each gene containing the SNPs within its physical range plus a flanking region of 100 KB up- and down- streams. This resulted in total of $29,458$ genes in the 22 autosomes. For isoform genes, we took the join regions of all the isoforms to be the same gene.

The raw genotypes were screened by a series of quality control procedures. SNPs with missing rates greater than 1%, hetrozygous haploid and markers with Hardy-

Weinberg equilibrium p-value less than 10^{-6} were removed, which left in total of $564,636$ SNPs in the analysis. The missing genotypes with missing rate under 1% were imputed by the average genotype scores from the nearby markers.

3.3 Models and methods

The proposed method consists of two selection stages and a post-selection estimation stage. The ultimate goal is to efficiently and jointly select and estimate the important associations between ultra-high dimensional responses and predictors.

Within each selection stage, we use the following multivariate linear regression to model the associations between responses and predictors.

$$Y = \boldsymbol{\beta}_0 + XB_X + \mathbf{I}_{ad}\boldsymbol{\beta}_{ad}^t + \mathbf{I}_{mci}\boldsymbol{\beta}_{mci}^t + (X \times \mathbf{I}_{ad})B_{Xad} + (X \times \mathbf{I}_{mci})B_{Xmci} + \boldsymbol{\varepsilon}, \quad (3.1)$$

where Y is the response matrix, X is the matrix of genetic predictors, $\mathbf{I}_{ad}$ and $\mathbf{I}_{mci}$ are indicators for AD and MCI subjects, $X \times \mathbf{I}_{ad}$ and $X \times \mathbf{I}_{mci}$ are the interaction terms between the genetic predictors and the disease status, B, $\boldsymbol{\beta}_{ad}$, $\boldsymbol{\beta}_{mci}$, B_{Xad} and B_{Xmci} are the regression coefficient vectors or matrices with respect to corresponding predictor terms, $\boldsymbol{\beta}_0$ is the grand intercept and $\boldsymbol{\varepsilon}$ is the matrix of noise terms.

When variables in X and Y are centralized, $\boldsymbol{\beta}_0$ is zero and model (3.1) reduces to

$$Y = \mathbb{X}\mathbb{B} + \boldsymbol{\varepsilon} \quad (3.2)$$

with $\mathbb{X} = (X, X \times \mathbf{I}_{ad}, X \times \mathbf{I}_{mci}, \mathbf{I}_{ad}, \mathbf{I}_{mci})$ and $\mathbb{B} = (B_X^t, B_{Xad}^t, B_{Xmci}^t, \boldsymbol{\beta}_{ad}, \boldsymbol{\beta}_{mci})^t$.

Here we assume that the association signals are sparse, i.e., each voxel is only associated with few number of SNPs compared to the sample size. To select the sparse signals, we aim to solve for the regularized solution of (3.2). Regularization methods are widely used in regression settings when the number of predictors is much greater than the sample size, the so-called large-p-small-n problems.

For our brain-GWAS data, both responses and predictors are of ultrahigh dimensions and both of them have intrinsic biological meaningful group structures.

In order to efficiently select the important association signals by taking into consideration the group structures, we propose to use the multivariate sparse group lasso (MSGLasso) introduced by Li et al. (2013) in the first selection stage to select the important brain-region-to-gene association groups. The MSGLasso solves the penalized least square estimators of the regression coefficients in model (3.2) and is capable in particular of handling the ultrahigh dimensionalities and the complex group structures in both responses and predictors.

Then in the second selection stage, we use the multivariate lasso to further select the important voxel-to-SNP associations within the selected group signals from the first stage. With group level sparsity, many unimportant region-to-gene groups are shrunken to zero in the first selection stage and therefore are ruled out from the following analysis. Similarly, with individual level sparsity, the number of variables passing into the estimation stage are further reduced in the second stage selection. At last, in the estimation stage, we only focus on estimating the effects of the selected SNPs on each selected voxel. Hence, with this hierarchical selection and estimation procedure, we efficiently remove the unimportant voxel-to-SNP association pairs and avoid huge number of multiple comparisons.

3.3.1 First selection stage: region-wise-gene-based mapping using the multivariate sparse group lasso

Li et al. (2013) introduced the multivariate sparse group lasso (MSGLasso) for high dimensional variable selections in multivariate-multiple settings when both responses and predictors have some known group structures. The MSGLasso is an appropriate tool for the association signal selection in the brain-wide GWAS setting. The MSGLasso minimizes the following objective function

$$\arg\min_{\mathbb{B}} \frac{1}{2n}\|Y - \mathbb{X}\mathbb{B}\|_2^2 + \lambda \sum_{\beta\in\mathbb{B}} |\beta|_1 + \lambda_1 \sum_{g\in\mathcal{G}^2} \omega_g^{1/2} \|\mathbf{B}_g\|_2, \tag{3.3}$$

where the lasso penalty is to encourage the within group voxel-to-SNP level sparsity and the group lasso penalty aims to shrink the unimportant region-to-gene groups to

zero. Here $\mathcal{G}^2$ denotes the set of region-to-gene interaction blocks on the regression coefficient matrix $\mathbb{B}$, $\mathbf{B}_g$ denotes one such group block, and ω_g is adaptive weight assigned to g'th group block. In our brain-wide GWAS, we use $\omega_g = \sqrt{v \times s}$, the square root of the group block size (Yuan and Lin, 2006; Silver et al., 2012), where v is the number of voxels in the corresponding region and s is the number of SNPs in the corresponding gene.

Li et al. (2013) proposed a mixed coordinate descent (MCD) algorithm for iteratively solving for the global minimizer of (3.3). They proved that the MCD algorithm converges to the global minimizer. Even though we can benefit from the computational efficiency of the MCD algorithm, the performance of the MSGLasso is governed by a factor of order $\sqrt{PQ/n}$ (Li et al. (2013)), where P is the number of predictors, Q is the number of responses and n is the sample size, which is a huge number given the fact that $P \approx 560,000$ and a $Q \approx 350,000$, but $n = 373$. We need to effectively control the complexity to achieve a satisfactory selection performance.

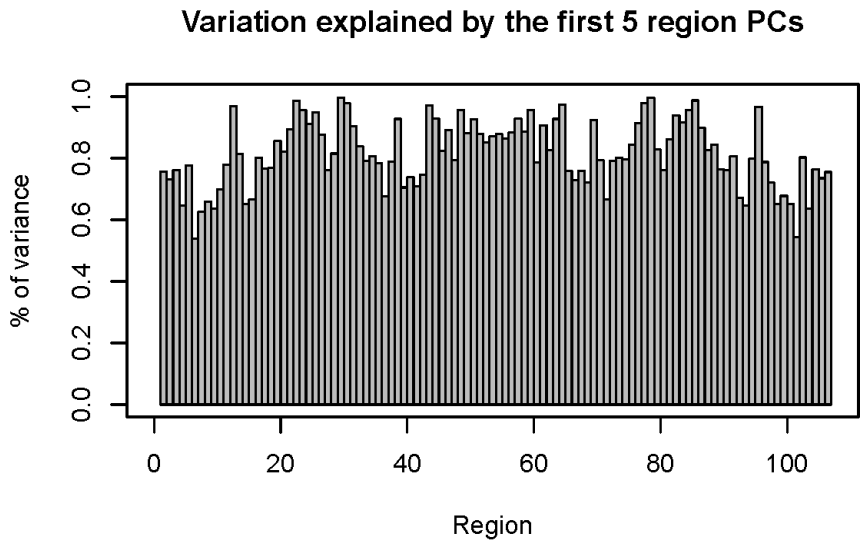
So to reduce the scale of the problem, we use a few major principle components (PCs) of each brain region and of each gene, respectively, in MSGLasso instead of using the raw voxel level scores and SNP genotypes. We interpret the selected PCs as the evidence of associations between their representative brain regions and genes. The advantage of using the principle components analysis (PCA) is two-fold. First, it helps reduce the dimensions and improve the efficiency of group selection. We emphasize that the first stage group selection serves only as a screening procedure. It aims to rule out the unimportant region-to-gene group signals. Secondly, the major PCs explain the largest variations across the sample, so they catch the essential information contained in the data. Also since PCs are orthogonal (independent) to each other, they avoid the complications from collinearity in the model.

We use the first five PCs for each brain region and up to the first twenty PCs for each gene in the first selection stage. Figure 3.2 (a) shows the percentage of total variation explained by the first five PCs in each brain region. Most of the regions have more than 70% of their variations explained by their first five PCs. As an

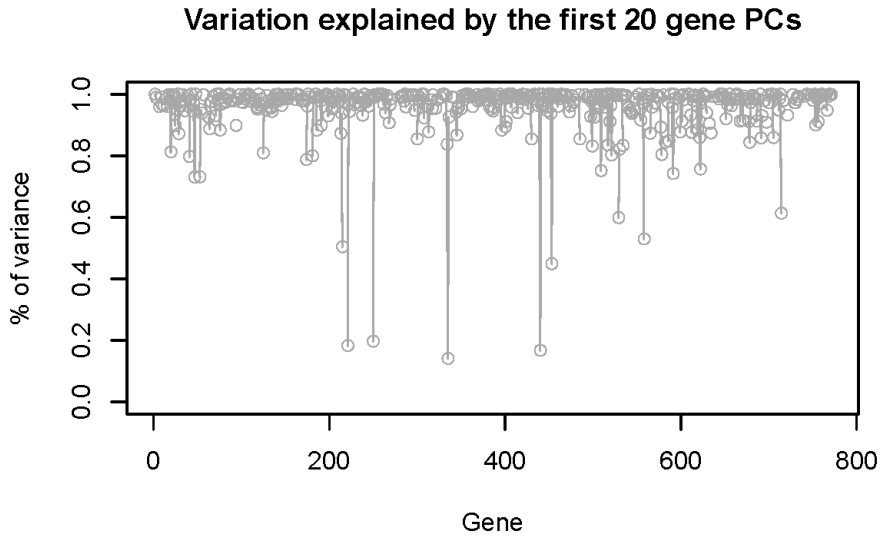
illustration, Figure 3.2 (b) shows the percentage of variation explained by up to the first 20 PCs in each gene on chromosome 20. Most of the genes have more than 80% of variations explained by their first up to 20 PCs. Seven out of 800 genes on chromosome 20 have less than 60% of variations explained by their first 20 PCs. For those genes, we still only include their first 20 PCs in the analysis, to keep the total number of gene PCs on each chromosome in a handleable scale. It could limit the power of detecting the effects of such genes and further separate investigation focusing only on those genes might be needed.

We then run the MSGLasso using the major PCs to select the important region-to-gene associations. To avoid the complexity and reduce the computational burden from cross validation in selecting the optimal tuning parameters, we instead use the method of stability selection (Meinshausen and Bühlmann, 2010) from 100 bootstrapped data sets. Meinshausen and Bühlmann (2010) suggested to use a fixed set of tuning parameter values on re-randomized data sets. As long as the proposed tuning parameter values are from a reasonable range, the variable selection results are quite stable. We select those region-to-gene pairs with at least 75% stability selection frequency in the second stage analysis.

Figure 3.3 shows stability selection frequencies for several brain regions across the whole genome. For example, (a) is for the gene PC effect selection frequency for region CERHEM(L), where gene *RIN2* has two independent PCs with selection frequencies more than 75% and is therefore selected into the second stage analysis for within voxel-to-SNP level selection, among a few others. In (b) and (c), gene×AD interaction effect selection frequency on regions BA39(L) and BA39 is plotted. where genes *MED1* and *COL9A3* are selected, among a few others. In (d), gene×MCI interaction effect on region BA17 is plotted and gene *PRDM15* is selected into the second stage.



(a)

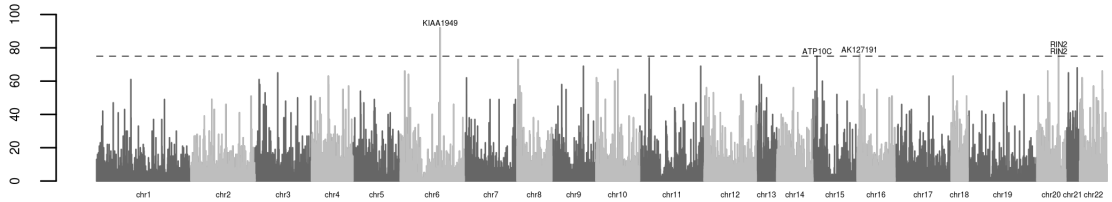


(b)

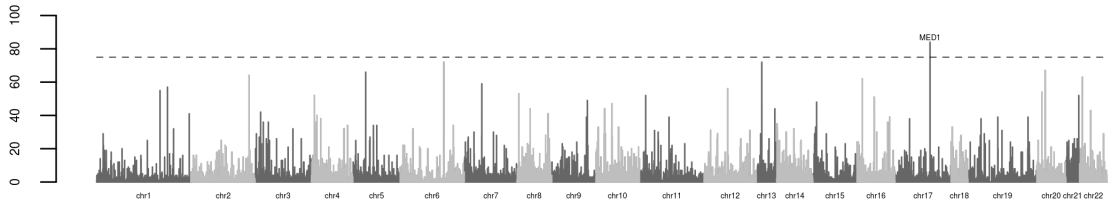
Figure 3.2: (a): percent of variation explained by the first 5 region PCs for the 106 regions. (b): percent of variation explained by up to the first 20 gene PCs for chromosome 20.

3.3.2 Second selection stage: voxel-wise-SNP-based fine mapping on the selected region-gene pairs using the multivariate lasso

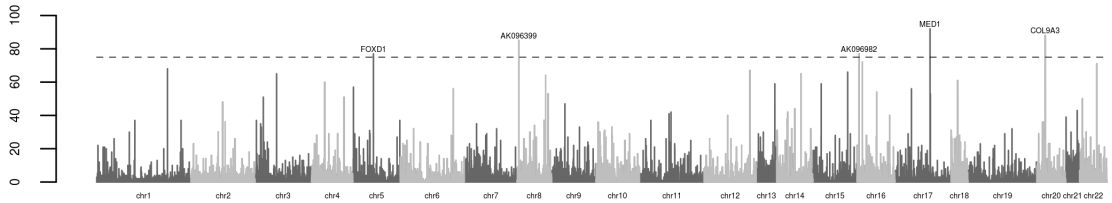
In the second stage, for each region-gene pairs selected from the first stage, we further zoom in to look at the associations at voxle-SNP level. For each such region-gene pair, we fit a multivariate multiple linear regression model regularized with only the lasso penalty, where the response variables are the original voxel intensity scores



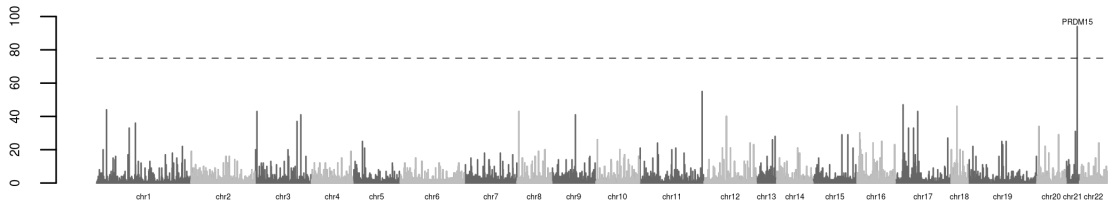
(a) Region: CERHEM(L)



(b) Region: Superior parietal cortex BA39(L)



(c) Region: Superior parietal cortex BA39



(d) Region: Occipital cortex BA17

Figure 3.3: Example Manhattan plots of region-gene select frequencies for each example region across the genome. (a) gene PC effects. (b) and (c) gene PC×AD interaction effect. (d) gene PC×MCI interaction effect.

and predictors are the union of SNP genotypes of all the selected genes (w.r.t. the selected region) on a single chromosome. The objective function of the multivariate lasso is simply that of the MSGLasso without the group lasso penalty:

$$\arg\min_{\mathbb{B}} \frac{1}{2n}\|Y - \mathbb{X}\mathbb{B}\|_2^2 + \lambda \sum_{\beta \in \mathbb{B}} |\beta|_1. \tag{3.4}$$

To run the multivariate lasso using MSGLasso, just set the whole regression coefficient matrix as one single group and set the group lasso tuning parameter for that group to be zero.

Figure 3.4 shows some gene effects on the selected regions. The effect types are represented in color: blue for gene effect, green for gene-to-AD interaction effect and yellow for gene-to-MCI interaction effect. Voxels with more than 80% selection frequency on the most significant SNP (with the most significant p-value obtained from the later estimation stage) are highlighted in red, so to indicate the top SNP's effect regime on the region.

3.3.3 Stability selection and control for false discoveries

Stability selection was used in both selection stages. The reason that we adopt the stability selection is two fold. First, it reduces computing cost in choosing tuning parameters for such large data sets. Secondly, stability selection provides a quantitative way to govern the number of false discoveries (NFD).

Illustrated in Figure 3.5, stability selection is very robust against different tuning parameter values in terms of consistently selecting the important variables. In the figure, we plot the selection path of region CERHEM(L) and region occipital cortex BA19(R) on chromosome 20 in the first stage selection. The vertical axis is the select frequency out of 100 bootstrap data sets and the horizontal axis is for the different settings of the tuning parameters. We fixed the ratio of the individual level tuning parameter to the group level tuning parameter at 10. The highlighted selection paths are for the top regression coefficients with the largest absolute values using the proposed tuning parameter values in the actual analysis. It can be seen that as long as the proposed tuning parameters are not ridiculously off, i.e., neither too large which shrink almost everything to zeros or too small which barely shrink anything, the top signals are consistently selected from the bootstrap data sets. The stability selection can be easily implemented parallelly on a multi-core computing cluster, therefore saves more computation time.

Meinshausen and Bühlmann (2010) showed that the expected number V of falsely

selected variables is bounded from above by

$$E(V) \leq \frac{1}{2\pi_{thr} - 1} \frac{q^2}{P}, \tag{3.5}$$

where π_{thr} is the thresholding frequency used for the selection, which in our case is 75% for the first stage selection and 80% for the second stage selection, and q is the average number of selected variables. In our study, the typical numbers of selected variables are from tens to hundreds out of tens of thousands of variables in total, which yield $q^2/P < 1$. Therefore the error number per chromosome is controlled by $< 1/(2 \times 0.75 - 1) = 2$, i.e., in the first stage selection, for each brain region PET PC, we will select at most 1 falsely discovered gene PC per chromosome on average.

3.3.4 Estimation stage: post-selection inference on the selected signals

Using stability selection in the above second selection stage, for each voxel we assign its important predictors to be the SNPs with selection frequency greater than 80%. Then we apply a multiple linear regression for each voxel with its important predictor SNPs. If for some voxel, the number of selected important SNPs exceeded the sample size, a more stringent selection frequency threshold could be applied. In fact, the numbers of important SNPs for all selected voxels are much smaller than the sample size in our brain-GWAS, so only ordinary multiple linear regressions are used for post-selection inference, and we obtain the usual p-values for each selected voxel-to-SNP pair. Table 3.1 reports all the voxel-to-SNP level signals that satisfy both the criteria of having more significant p-values than 10^{-6} and greater selection frequencies than 80%. Since there is no any SNP-MCI interaction effect satisfying both criteria, we provide a list of a few top MCI interactions in Table 3.2. Figure 3.7 gives the most significant SNP's p-value(s) and its selection frequency pattern on the selected region for each of those signals in Table 3.1 and Table 3.2.

Compared to the conventional single-voxel-to-single-SNP approaches with multiple comparisons, our two-stage selection approach is less conservative. We illustrate this advantage in Figure 3.6 with an example signal between *BIN2* gene (with 90

SNPs) and region CERHEM(L) (with more than 5000 voxels). The single-voxel-to-single-SNP approach is carried out by simple linear regressions on each voxel-to-SNP pair along with Bonferroni criterion for multiple comparison corrections. None of the signals survives the multiple comparison justification. In contrast, our two-stage-MSGLasso-plus-stability-selection is much less conservative since the number of hypothesis tests in the post-selection inference is significantly reduced.

3.4 Results

Table 3.1 provides a list of top signals that meet both criteria of p-value$\leq 10^{-6}$ and selection frequency$\geq 80\%$. Table 3.2 provides some other top gene-to-AD and gene-to-MCI interaction signals that meet both criteria of p-value$\leq 10^{-5}$ and selection frequency $\geq 80\%$. The selected brain regions and strength of the gene effects listed in Table 3.1 and 3.2 are also illustrated in Figure 3.4 and 3.7, respectively.

Our brain-wide GWA study identifies some brain regions that have either significant gene effects or gene-AD interaction effects. For example, many regions, such as BA40(L), BA39(R), BA39(L), BA7(R) and BA7(L), in superior parietal cortex are found associated with some genes or have their associations significantly modified by the AD status. Mills et al. (2013) reported associations between lipid metabolism in superior parietal cortex and alternatively spliced isoforms in RNA transcriptome. Other identified regions include BA18(R), BA18(L), BA19(R), BA19(L) in occipital cortex (Braskie et al., 2011) and BA20(R), BA20(L), BA21(R), BA21(L), BA22(R) and BA22(L) in temporal cortex (Stein et al., 2010b; Risacher et al., 2009; Braskie et al., 2011).

We also confirm some of the genetic findings in the literature. For example, Wang et al. (2013) found that inhibiting *IL8RB* (*CXCR2*) can turn down anyloid-β production and protect neural cells. Nakamura et al. (2006) found a similar effect of *COLEC12* (*SRCL*) gene in AD samples. Other direct supports on AD interactions include Burns et al. (2011) with *SAKCA* (*KCNMA1*) gene, Xie et al. (2010) with *PRIMA* gene, Nakamura et al. (2006) with *COLEC12* gene and Broer et al. (2011)

with *HSPA13* gene.

Some gene-to-AD interaction effects are found in the literature to be associated with other cognitive-related disease phenotypes such as autism and hearing impairment. Such cases include *AK096399* gene in Cannon et al. (2010), *GJB2* gene in Lingala et al. (2009), *SNX29* gene in Teasdale and Collins (2012), *MED1* gene in Giordano and Macaluso (2011) and Wong et al. (2013), and *COL9A3* gene in Solovieva et al. (2006) and Asamura et al. (2005).

We also confirm some gene effects on brain metabolizing. For example, *CDC42EP3* gene encodes certain family of guanosine triphosphate metabolizing proteins and the gene is weakly expressed in brain (provided by RefSeq, Jul 2012); *PACS2* plays a role in membrane traffic with tumour-necrosis-factor-related apoptosis-inducing-ligand (TRAIL) induced apophasis (Aslan et al., 2009), which in turn can cause human brain cell death (Nitsch et al., 2000).

The other interesting findings are about indirect effects of genes on certain chemical compound or protein translocation, which are in turn associated with AD. For example, Dai et al. (2013), Sakamoto and Holman (2008) demonstrate that *TBC1D4* plays an role in regulation of GluT4 traffic, which, on the other hand is associated with AD (Talbot et al. (2012), Yang et al. (2013)). Nolte et al. (2006), Lu et al. (2007) together give a chain of relationships of *HOXD4* gene to Pax6 protein to AD. Heikaus et al. (2002) states that *Raloxifene* results in increased expression of *GJB2* (*CX26*) mRNA and Yaffe et al. (2005) talked about the prevention effect of *Raloxifene* on cognitive impairment.

There are also some novel signals for those we didn't find trace of evidence support in the literature, such as associations between *BC007399* gene and BA39(R) in the superior parietal cortex, between *GALNT4* gene and BA19(L) in occipital cortex and between *RIN2* gene and CERHEM(L).

3.5 Discussion

Most of the previous brain-wide GWAS's (Stein et al., 2010a; Hibar et al., 2011) focus on fMRI images. Since PET image measures brain metabolism, our study features on different perspective of human brain than those conveyed from fMRI images. To the best of our knowledge, this is the first structured brain-wide GWAS using PET images.

The overall computational cost of our two-stage approach is lower than most voxel-to-SNP approaches (Stein et al., 2010a) on a similar data scale, because our approach rules out the unimportant region-to-gene signals in its first stage and only focus on the selected region-to-gene pairs in the later analysis. The most computational burden of our approach comes from the stability selections performed on large amount of bootstrap samples and from running MSGLasso on either vary large regions (with tens of thousands of voxels) or very large genes (with hundreds or more SNPs). To save the total computational time, we parallelized the computational jobs on multi-core UNIX clusters.

It is well known that some genes are associated with AD status (Kukull et al., 1996; Biffi et al., 2010). The genes associated with AD are not necessarily associated with brain metabolism. In our analysis, we did not find that any top AD-associated genes reported on http://www.alzgene.org has direct effects on brain metabolism.

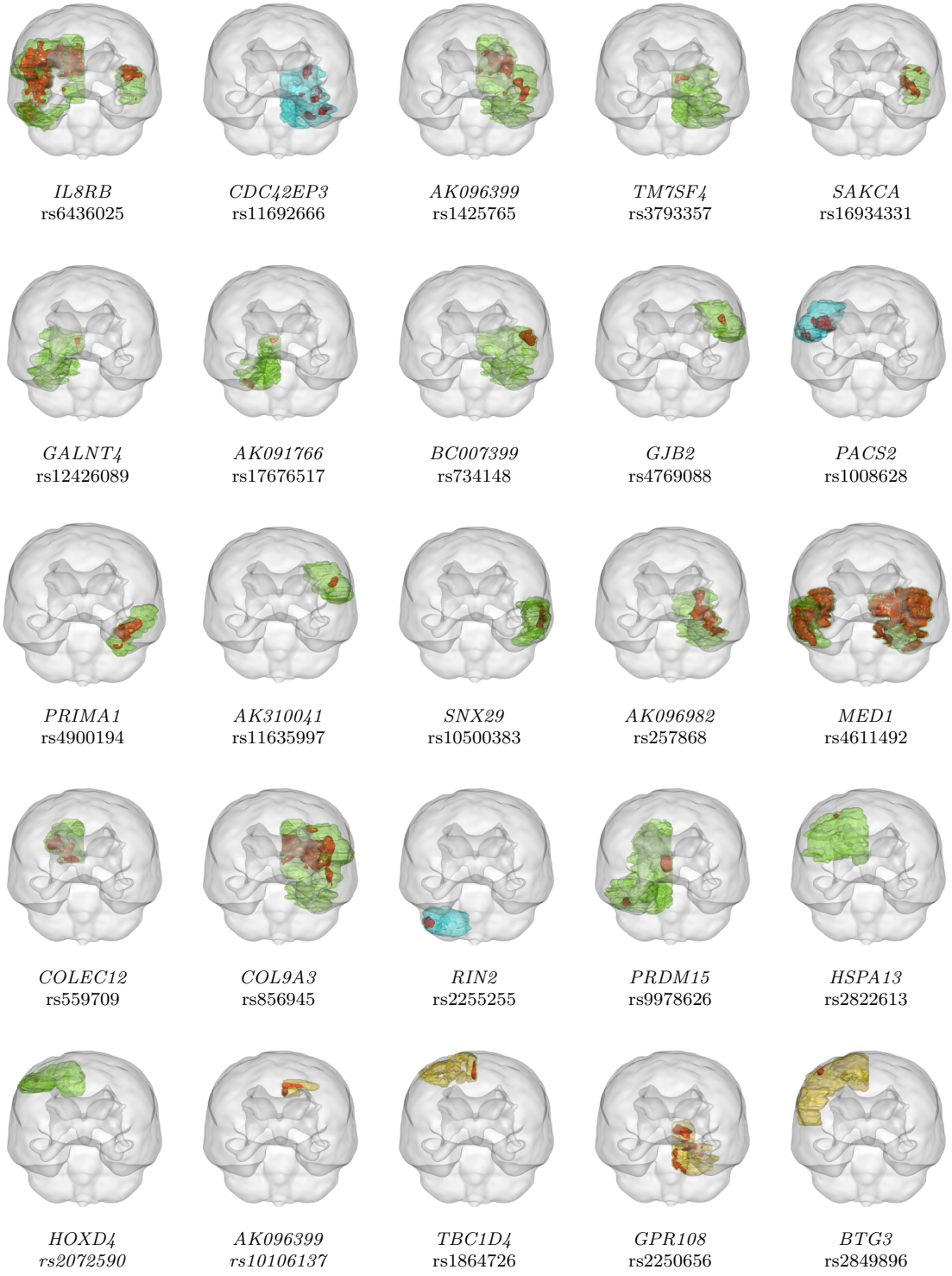


Figure 3.4: Gene effect on regions for signals with top SNP's p-value less than 10^{-6} in Table 3.1 and the signals in Table 3.2. The red highlighted voxels in each region are voxels with at least 80% stability selection frequencies on the listed SNP within the listed gene, the darker the red color the higher the stability select frequency. Regions are highlighted in the colors: (i) blue - the gene effects; (ii) green - the gene-to-AD interaction effects; (iii) yellow - the gene-to-MCI interaction effects.

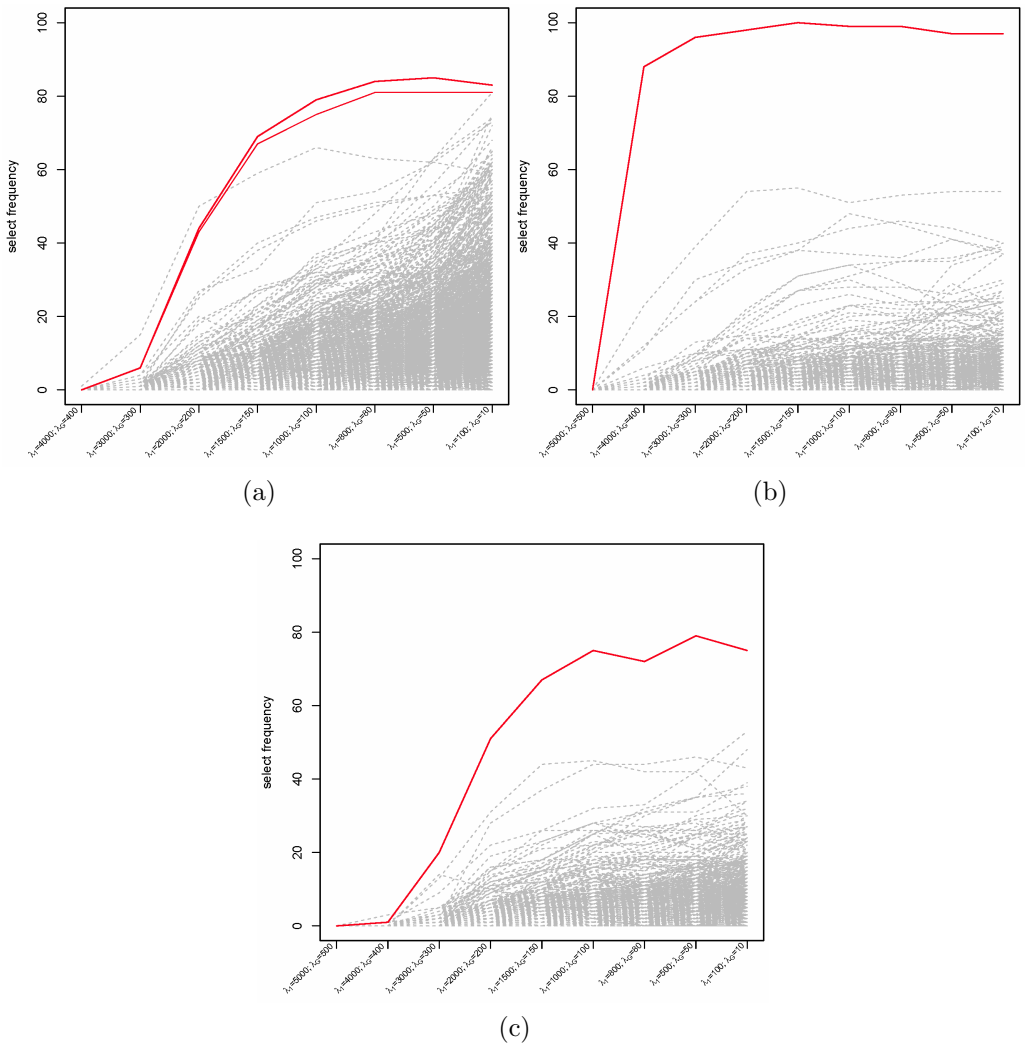


Figure 3.5: Illustration of robustness to the tuning parameters of stability selection (a) Gene effect of region CERHEM(L) PCs on chr20 gene PCs. (b) Gene AD interaction effect of region occipital cortex BA19(R) PCs on chr20 gene PCs. (c) Gene MCI interaction effect of region occipital cortex BA19(R) PCs on chr20 gene PCs. Each curve is for the selection path of a regression coefficient. The red curves are election paths for the top regression coefficients with the largest absolute values using the proposed tuning parameter values in the actual analysis.

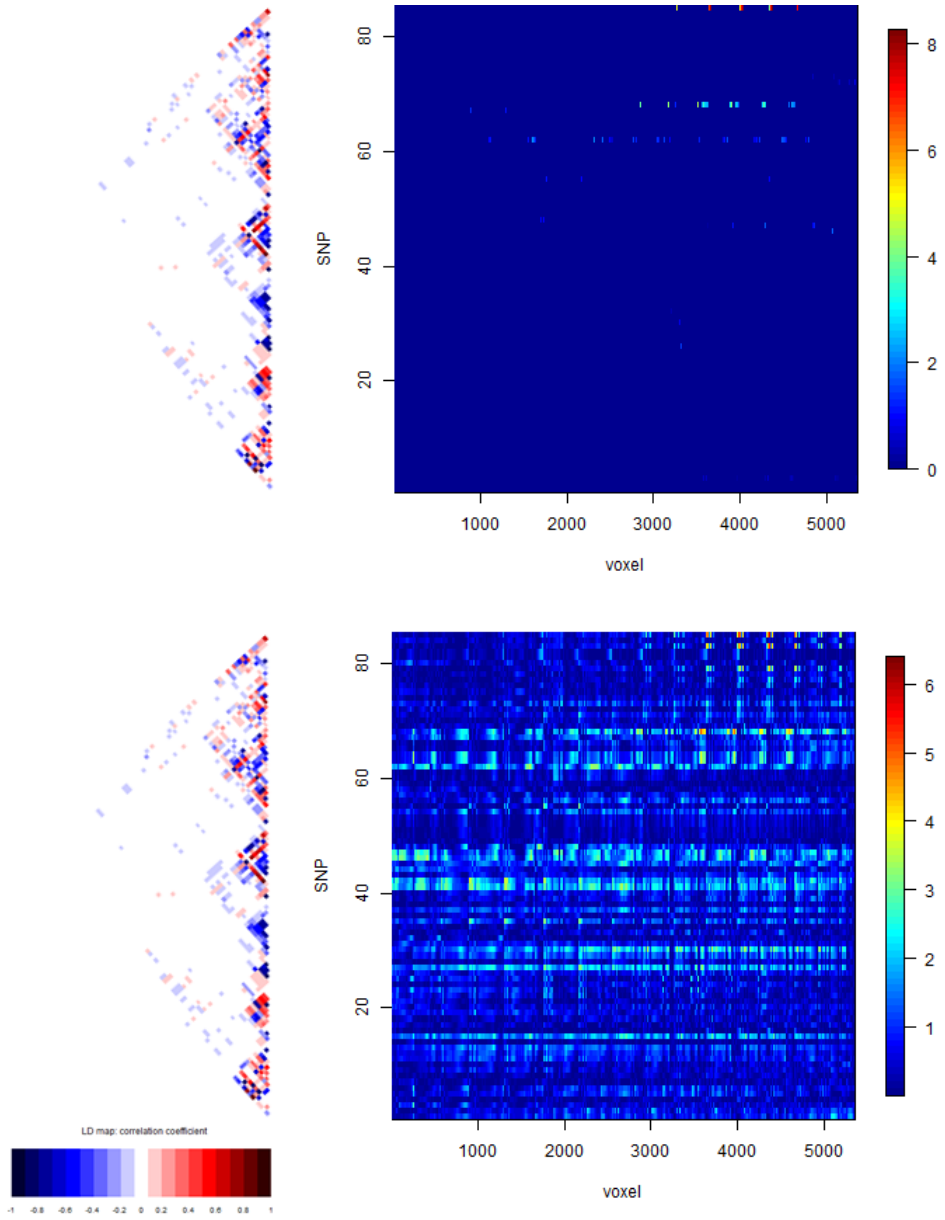
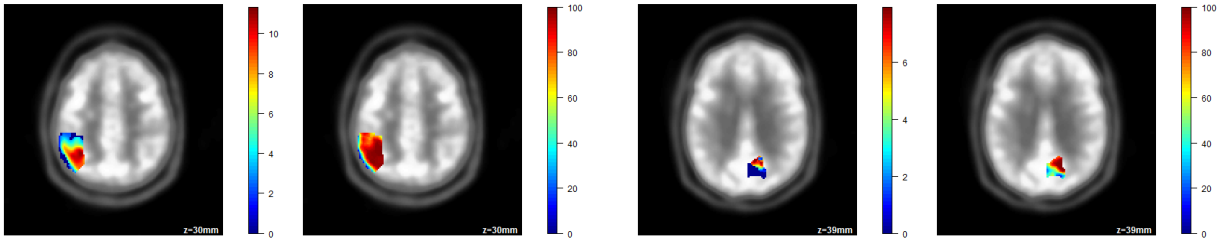


Figure 3.6: Comparing the MSGLasso to the conventional approach of looking at each voxel-to-SNP pair at a time. The upper panel is the selection and estimation results ($-\log_{10} p$-values) from our two stage approach, focusing on region CEREHEM(L) and gene *RIN2*. The lower panel is the result ($-\log_{10} p$-values) from simple linear regressions on each voxel-to-SNP pair on the same region and gene. The triangles on the left side indicating the correlation pattern across the 85 SNPs in gene *RIN2*.

-log$_{10}$(p-value) *select frequency* *-log$_{10}$(p-value)* *select frequenc*

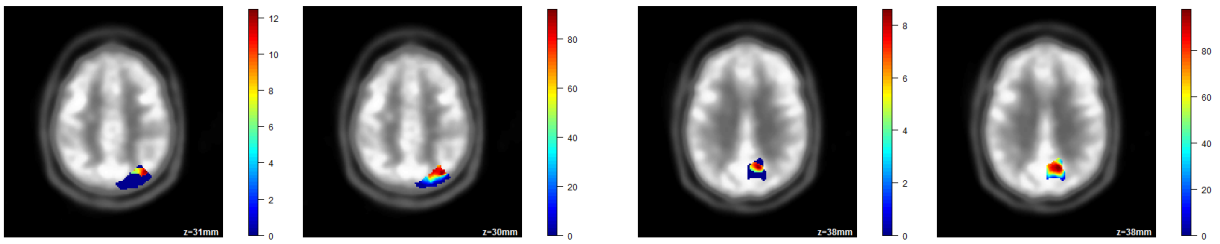


IL8RB::rs6436025
BA40(L)

IL8RB::rs6436025
BA40(L)

CDC42EP3::rs11692666
BA18(R)

CDC42EP3::rs11692666
BA18(R)

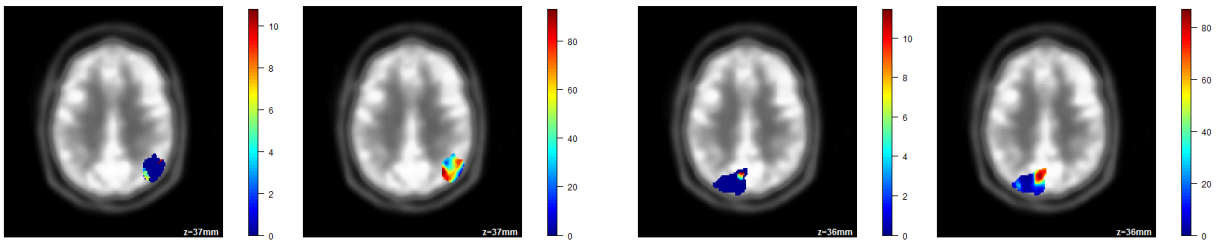


AK096399::rs1425765
BA19(R)

AK096399::rs1425765
BA19(R)

TM7SF4::rs3793357
BA18(R)

TM7SF4::rs3793357
BA18(R)

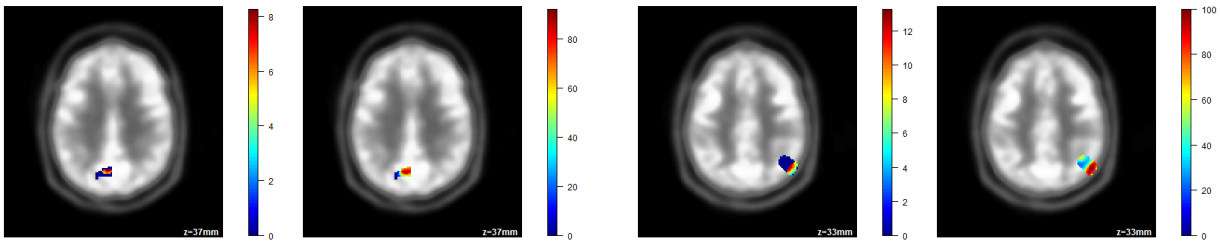


SAKCA::rs16934331
BA39(R)

SAKCA::rs16934331
BA39(R)

GALNT4::rs12426089
BA19(L)

GALNT4::rs12426089
BA19(L)



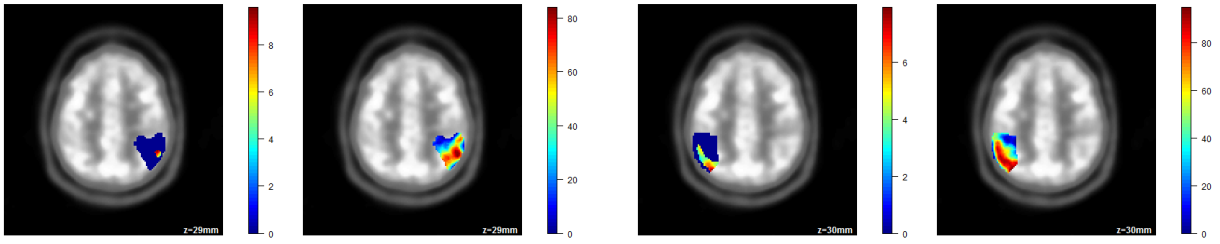
AK091766::rs17676517

BA18(L)

AK091766::rs17676517
BA18(L)

BC007399::rs734148
BA39(R)

BC007399::rs734148
BA39(R)

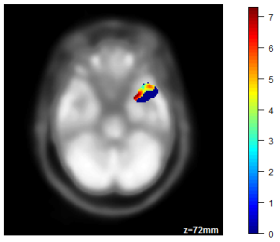
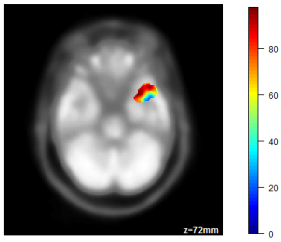
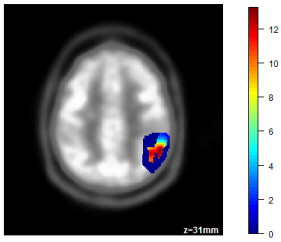
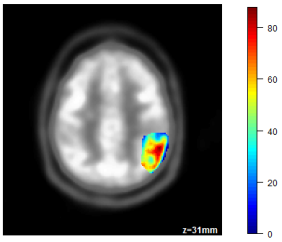
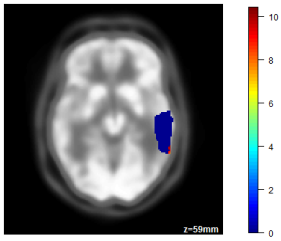
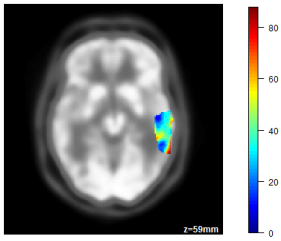
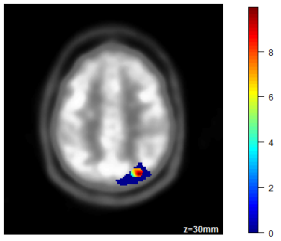
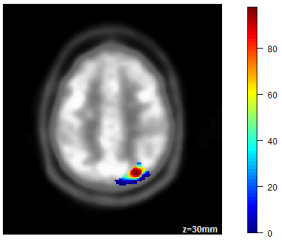
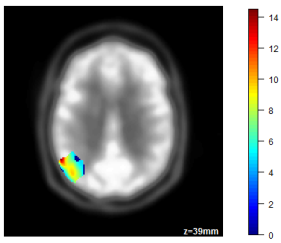
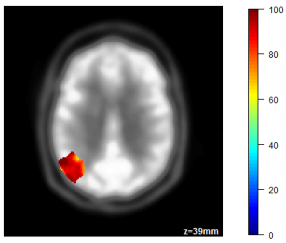
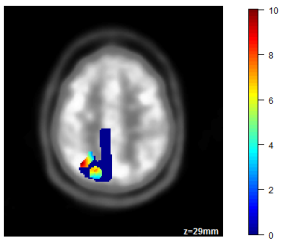
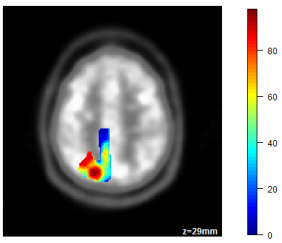
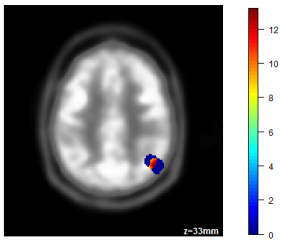
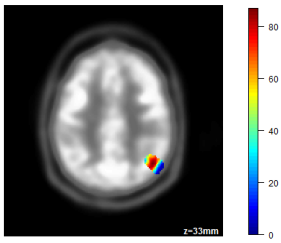
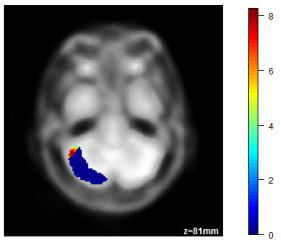
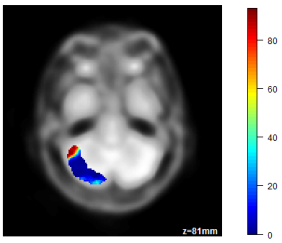
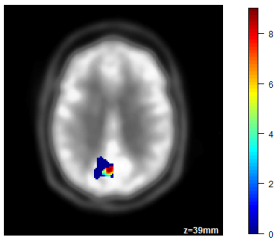
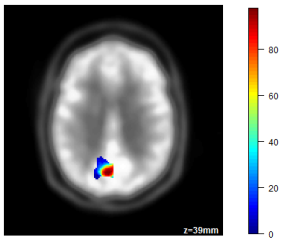
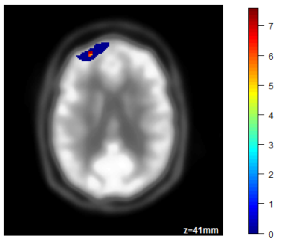
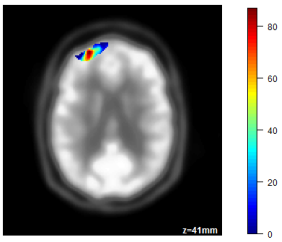


GJB2::rs4769088
BA40(R)

GJB2::rs4769088
BA40(R)

PACS2::rs1008628
BA40(L)

PACS2::rs1008628
BA40(L)

-log$_{10}$(p-value)	select frequency	-log$_{10}$(p-value)	select frequenc
			
PRIMA1::rs4900194 BA38(R)	*PRIMA1*::rs4900194 BA38(R)	*AK310041*::rs11635997 BA40(R)	*AK310041*::rs11635997 BA40(R)
			
SNX29::rs10500383 BA21(R)	*SNX29*::rs10500383 BA21(R)	*AK096982*::rs257868 BA19(R)	*AK096982*::rs257868 BA19(R)
			
MED1::rs4611492 BA39(L)	*MED1*::rs4611492 BA39(L)	*COLEC12*::rs559709 BA7(L)	*COLEC12*::rs559709 BA7(L)
			
COL9A3::rs856945 BA40(R)	*COL9A3*::rs856945 BA40(R)	*RIN2*::rs2255255 CERHEM(L)	*RIN2*::rs2255255 CERHEM(L)
			
PRDM15::rs9978626 BA18(L)	*PRDM15*::rs9978626 BA18(L)	*HSPA13*::rs2822613 BA10(L)	*HSPA13*::rs2822613 BA10(L)

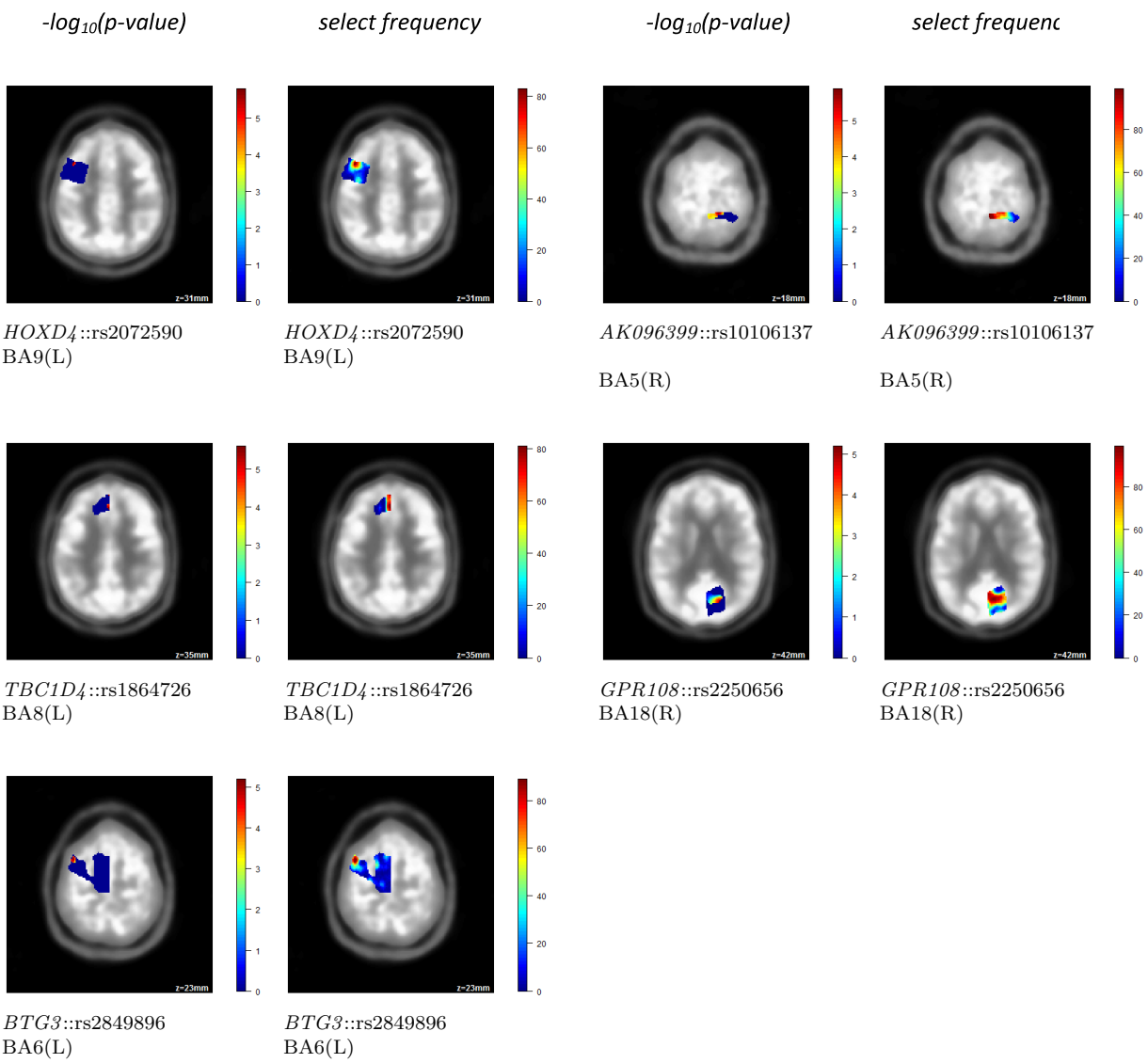


Figure 3.7: The most significant SNPs' effects, their $-\log_{10}(p-\text{values})$ on voxels across the associated region, and their selective frequency pattern on the region.

Table 3.1: Top selected genes, their associated regions and within them the top SNPs those are with p-values more significant than 10^{-6} and selection frequencies more than 80%

gene information				top selective SNP in gene		asso.	effect type	reference
name	chr	num. SNP in gene	% var. by 20 PCs	SNP name	most sig. p-value	region		
IL8RB	2	11	100%	rs6436025	4.8e-12	Superior parietal cortex_BA40(L)	G×AD	Liu et al. (2010), Vallès et al. (2006), Horuk et al. (1997)
					3.8e-11	Superior parietal cortex_BA39(L)	G×AD	Desbaillets et al. (1997), Tsai et al. (2002), Xia et al. (1996)
					2.9e-10	Superior parietal cortex_BA7(L)	G×AD	Wang et al. (2013)
					7.0e-09	Superior parietal cortex_BA39(R)	G×AD	
					1.9e-08	Posterior cingulated_BA31(L)	G×AD	
					7.8e-08	Inferior parietal cortex_BA37(L)	G×AD	
					1.9e-07	Temporal cortex_BA20(L)	G×AD	
				rs4674246	1.2e-07	Medial frontal cortex_BA9(R)	G×AD	
					1.1e-07	Temporal cortex_BA20(R)	G×AD	

Continued on next page

gene information				top selective SNP in gene		asso.	effect type	reference
name	chr	num. SNP in gene	% var. by 20 PCs	SNP name	most sig. p-value	region		
CDC42EP3	2	36	98.6%	rs11692666	1.2e-08	Occipital cortex_BA18(R)	G	–
					2.0e-7	Occipital cortex_BA19(R)	G	
AK096399	8	40	98.6%	rs1425765	3.3e-13	Occipital cortex_BA18(R)	G×AD	Cannon et al. (2010)
					9.3e-13	Superior parietal cortex_BA39(R)	G×AD	
					3.0e-10	Superior parietal cortex_BA7(R)	G×AD	
				rs269197	5.8e-07	Superior parietal cortex_BA7(L)	G×AD	
TM7SF4	8	47	97%	rs3793357	2.5e-09	Occipital cortex_BA18(R)	G×AD	Chung et al. (2010), Korac et al. (2013)
					1.5e-07	Occipital cortex_BA19(R)	G×AD	
SAKCA	10	109	81%	rs16934331	1.6e-11	Superior parietal cortex_BA39(R)	G×AD	Burns et al. (2011), Ertekin-Taner (2010)
				rs1871066	1.0e-07	Posterior cingulated_BA31(R)	G	
				rs3781141	6.9e-07	Superior parietal cortex_BA39(R)	G×AD	
				rs2247557	8.3e-07	Primary somatosensory cortex_BA2(R)	G×AD	
GALNT4	12	9	100%	rs12426089	3.5e-12	Occipital cortex_BA19(L)	G×AD	–

Continued on next page

gene information				top selective SNP in gene		asso.	effect type	reference
name	chr	num. SNP in gene	% var. by 20 PCs	SNP name	most sig. p-value	region		
AK091766	12	39	96.3%	rs17676517	5.5e-09	Occipital cortex_BA18(L)	G×AD	–
BC007399	12	36	97%	rs734148	5.1e-14	Superior parietal cortex_BA39(R)	G×AD	–
					1.4e-8	Occipital cortex_BA19(R)	G	
				rs12423428	7.1e-09	Superior parietal cortex_BA39(R)	G	
GJB2	13	25	98.1%	rs4769088	2.6e-10	Superior parietal cortex_BA40(R)	G×AD	Lingala et al. (2009), Yaffe et al. (2005), Heikaus et al. (2002)
				rs10870680	1.3e-08	Superior parietal cortex_BA40(R)	G×AD	
				rs945373	2.7e-08	Superior parietal cortex_BA40(R)	G	
DCT	13	44	97.2%	rs7336995	6.3e-07	Medial frontal cortex_BA8(L)	G	–
PACS2	14	8	100%	rs1008628	1.2e-08	Superior parietal cortex_BA40(L)	G	Aslan et al. (2009), Nitsch et al. (2000)
PRIMA1	14	48	90.3%	rs4900194	4.9e-08	BA38(R)	G×AD	Xie et al. (2010)
				rs12895346	2.9e-07	BA38(R)	G×AD	
				rs2064930	3.4e-07	BA38(R)	G×AD	

Continued on next page

gene information				top selective SNP in gene		asso.	effect type	reference
name	chr	num. SNP in gene	% var. by 20 PCs	SNP name	most sig. p-value	region		
AK310041	15	2	100%	rs11635997	2.1e-14	Superior parietal cortex_BA40(R)	G×AD	–
SNX29	16	249	76.2%	rs10500383	3.6e-11	Temporal cortex_BA21(R)	G×AD	Teasdale and Collins (2012)
				rs10500383	3.0e-09	Temporal cortex_BA22(R)	G×AD	
				rs11859327	8.4e-07	Occipital cortex_BA19(R)	G	
AK096982	16	1	100%	rs257868	1.1e-10	Occipital cortex_BA19(R)	G×AD	–
MED1	17	4	100%	rs4611492	3.1e-15	Superior parietal cortex_BA39(L)	G×AD	Giordano and Macaluso (2011) Wong et al. (2013)
					8.2e-14	Superior parietal cortex_BA39(R)	G×AD	
					5.6e-13	Temporal cortex_BA22(L)	G×AD	
					5.9e-11	Occipital cortex_BA19(R)	G×AD	
					1.9e-10	Temporal cortex_BA21(L)	G×AD	
COLEC12	18	127	75%	rs559709	9.4e-11	Superior parietal cortex_BA7(L)	G×AD	Nakamura et al. (2006)
				rs12960602	2.5e-09	Superior parietal cortex_BA39(L)	G×AD	
ELP2	18	127	75%	rs7235689	9.4e-7	Medial frontal cortex_BA11(L)	G×AD	–
KISS1R	19	16	100%	rs2306718	3.3e-07	Pre-motor cortex_BA6(R)	G×AD	Chilumuri and Milton (2013)

Continued on next page

gene information				top selective SNP in gene		asso.	effect type	reference
name	chr	num. SNP in gene	% var. by 20 PCs	SNP name	most sig. p-value	region		
COL9A3	20	29	96.6%	rs856945	4.6e-14	Superior parietal cortex_BA40(R)	G×AD	Solovieva et al. (2006), Asamura et al. (2005)
					6.1e-14	Superior parietal cortex_BA39(R)	G×AD	
					1.2e-10	Superior parietal cortex_BA7(R)	G×AD	
					1.7e-9	Occipital cortex_BA19(R)	G×AD	
C20orf186	20	39	98.5%	rs378098	8.4e-08	Superior parietal cortex_BA7(R)	G×AD	–
RIN2	20	85	80.4%	rs2255255	5.3e-9	CERHEM(L)	G	–
PRDM15	21	93	82.7%	rs9978626	9.6e-10	Occipital cortex_BA18(L)	G×AD	–
					3.2e-9	Occipital cortex_BA19(L)	G×AD	
					6.2e-9	Superior parietal cortex_BA7(L)	G×AD	
					6.0e-8	Inferior parietal cortex_BA37(L)	G×AD	
				rs13049896	1.5e-07	Inferior parietal cortex_BA37(L)	G×AD	
HSPA13	21	35	99.2%	rs2822613	2.6e-08	Medial frontal cortex_BA10(L)	G×AD	Broer et al. (2011)

Table 3.2: Some other gene×AD and gene×MCI interaction effects of top SNPs with p-values more significant than 10^{-5} and selection frequencies more than 80%

gene information				top selective SNP in gene		asso.	effect type	reference
name	chr	num. SNP in gene	% var. by 20 PCs	SNP name	most sig. p-value	region		
HOXD4	2	11	100%	rs2072590	1.6e-06	Medial frontal cortex_BA9(L)	G×AD	Nolte et al. (2006), Nolte et al. (2006)
AK096399	8	40	98.6%	rs6436025	1.3e-06	Primary somatosensory cortex_BA5(R)	G×MCI	Cannon et al. (2010)
TBC1D4	13	86	87.1%	rs1864726	2.4e-06	Medial frontal cortex_BA8(L)	G×MCI	Talbot et al. (2012), Yang et al. (2013), Dai et al. (2013), Sakamoto and Holman (2008)
GPR108	19	27	98%	rs2250656	6.4e-06	Occipital cortex_BA18(L)	G×MCI	U Brüggemeier (2004)
BTG3	21	26	99.5%	rs2849896	6.3e-06	Pre-motor cortex_BA6(L)	G×MCI	Carson (2007)

CHAPTER IV

A cure model for analyzing longitudinal brain PET images and MCI conversions

4.1 Introduction

This chapter is motivated by the ADNI longitudinal positron emission tomography (PET) brain imaging study on mild cognitive impairment (MCI) patients. MCI is a brain function syndrome that involves cognitive impairments (Petersen et al., 1999). MCI patients may not meet neuropathologic criteria for Alzheimer's disease (AD), but they may be in a transitional stage of evolving AD. The MCI disease diagnostics and high-resolutional 3D PET brain imaging scans for each of the 236 MCI subjects were recorded at registration time (baseline) and then were followed up at month 6, 12, 18, 24, 36, 48, 60 or 72 thereafter. During the followup, some of the MCI patients converted to AD patients. Some subjects had missed certain follow-up visits and some were censored at certain follow-up time point. According to the AD clinical literature, it is believed that part of the MCI population will never convert to AD (Petersen et al., 2009). Therefore, it is natural to assume that the MCI population consists of a mixture of a non-cure portion with people who, given long enough follow-up, will eventually (but may not be observed in the study period) convert from MCI to AD, and a cure portion with people who will never experience the conversion. The cure portion has also been addressed as long-term survival portion, or immune portion in survival literature. Hereafter non-cure refers to AD conversion and cure refers to MCI without AD conversion ever.

The goals of the study were: (1) to select and estimate important imaging predictor voxels associated with AD conversion status, (2) to select and estimate the effect sizes of the important imaging predictors whose longitudinal profile patterns are associated with the time to AD from MCI diagnosis in non-cure group, and (3) to predict whether a new MCI subject will eventually convert to AD and when if it were to happen based on patient's PET scans. To accomplish these goals, we introduce in this chapter a variable selection method for a cure-rate discrete-time survival model in high-dimensional settings.

The studies of cure-rate survival models can be traced back to 60 years ago. Boag (1949) and Berkson and Gage (1952) introduced mixture cure rate models. Yakovlev et al. (1993) developed the so-called bounded cumulative hazard (BCH) model, which models the latent number of metastasis competent tumor cells that were left active after the initial treatment for a cancer patient using the Poisson distribution. Parametric and semi-parametric versions of both types of models had been widely studied (Mendenhall and Hader, 1958; Farewell, 1982; Hougaard, 1986; Kuk and Chen, 1992; Laska and Meisner, 1992; Taylor, 1995; Yakovlev and Tsodikov, 1996; Maller and Zhou, 1990; Ibrahim et al., 2001; Law et al., 2002). Some effort has also been focused on variable selection for cure-rate models (Liu et al., 2012). However, very little attention has been paid to variable selection for cure-rate models in high-dimensional settings. In our study, both the cure status and the non-cure survival can depend on high-dimensional imaging predictors. We adopt a mixture-model approach, which allows either one of the components or both to be high-dimensional. We use a logistic link function for the cure rate component, which is one of the most commonly used link functions in mixture cure-rate model (Taylor, 1995; Ibrahim et al., 2001).

In our study, the AD conversion status were observed only in a set of fixed discrete time intervals determined by the study design. Therefore the commonly used partial likelihood (Cox, 1975) approach for the proportional hazards model is not applicable as it was developed for the continuous survival times. This requires building a

discrete time survival model for grouped interval-censored survival data. A brief literature review of the discrete-time survival models can be found in, for example, Kalbfleisch and Prentice (2002); Allison (1982) and Singer and Willett (1993). We use a discrete time Cox proportional hazards (PH) model, and use the full likelihood for the non-cure discrete survival time based on multinomial distributions (Prentice and Gloeckler, 1978; Li et al., 2008).

At the end, the logistic model for cure-rate and the PH model for non-cure AD conversion probability in each time interval are integrated into building a complete likelihood, given the latent non-cure indicator variable. Thus, an EM algorithm can be employed to search for the penalized maximum-likelihood estimator (MLE). Variable selection for either the cure-rate or the non-cure survival is carried out via imposing the elastic net penalty that shrinks the effects of unimportant predictors to zero, and meanwhile takes into account the spatial correlations. Coordinate descent and majorization minimization algorithms are integrated to speedup the optimization procedure.

4.2 ADNI longitudinal PET imaging data

Uorodeoxyglucose (FDG) PET imaging data used in this study were obtained from the ADNI database (adni.loni.ucla.edu). When the ADNI project was first launched in 2003, one of the primary goals was to test whether serial magnetic resonance imaging (MRI) or PET images, together with other clinical and neuropsychological assessments can be used to help diagnose and measure the progression of MCI to AD.

Through the three phases of ADNI study, FDG PET scans and clinical diagnosis were collected longitudinally for each participant. Subjects were classified into three groups: AD, MCI and NC based on their disease status diagnosed at their initial visits. During the first phase of ADNI study from year 2003 to 2010, image scans and diagnosis were performed at months 6, 12, 18, 24, 36 for MCI subjects. With additional funding, the ADNI study moved into the second phase, the ADNI GO

study, in year 2010 for an additional 2-year period. Moreover, while the ADNI GO project continues, ADNI launched its third phase in 2011, known as ADNI 2, to further investigate MCI-to-AD conversions. As a result, the MCI subjects continued to receive follow-up at months 48, 60 and 72 till March 2013, when our analytical data were acquired. Among the 236 individuals who were diagnosed as MCI at their baseline visits, 106 converted to AD at certain point during their follow up.

Table 4.1 lists the number at risk and the number of MCI-to-AD conversions diagnosed at each followup time point. The followup image scans of six individuals who belong to the converter group according to their clinical diagnosis were all missing, thus these individuals are excluded from the analysis.

At each visit of a subject in the study, a FDG dose of 5.0 ± 0.5 mCi was first injected, and then a subsequent of post-injection PET scans were performed from 30 to 60 minutes acquiring 6 five minute frames. The PET scans were preprocessed by being co-registered to the first frame image file and the six co-registered frames were averaged to create one single 30 minute PET image, then the averaged PET scan was reoriented into a standard 160 by 160 by 96 voxel image grid with 1.5 mm cubic voxels. In our analysis, we further re-scaled the voxels in each image by dividing the average values in "Pons" and "Cerebellar vermis" so to achieve a desirable contrast between the AD and MCI images.

Table 4.1: Follow up status

Time (month)	baseline	6	12	18	24	36	48	60	72
# at risk	236	229	206	188	165	145	134	128	127
# of converters	0	7	23	18	23	20	11	6	1

Total # of subjects: 236

4.3 Methods

4.3.1 Mixture cure-rate models

As suggested by the AD clinical literature, we assume that the study sample contains two sub-samples, one with people who will eventually experience the event given long enough followup time (none-cure portion) and one with people who never experience the event (cure portion). This assumption is also suggested by the estimated survival curve (Figure 4.1). Samples from the none-cure population can either experience the event or be censored. Samples from the cure population are all censored. The none-cure proportion π is a population parameter to be inferred from data, which can depend on baseline covariates $\mathbf{Z}$ and hence is denoted as $\pi(\mathbf{Z})$. The population survival function then can be written as

$$S_{pop}(t|\mathbf{X},\mathbf{Z}) = \pi(\mathbf{Z})S(t|\mathbf{X}) + 1 - \pi(\mathbf{Z}),$$

where $\mathbf{X}$ are longitudinal imaging predictors. We assume a logit model $\pi(\mathbf{Z}) = \exp(\gamma'\mathbf{Z})/(1+\exp(\gamma'\mathbf{Z}))$ for the non-cure rate, which is commonly used in the mixture cure-rate models along with other popular link functions such as loglog and probit links (Cai et al., 2012; Peng, 2009). Here the regression coefficient vector γ contains an intercept for the average cure rate when all baseline covariates are valued at zaro. Mixture cure models have the advantage that the cure portion and the none-cure portion survival can be modeled separately as we will demonstrate in details shortly.

4.3.2 Discrete-time survival models

For the non-cure portion, suppose the event is observed in one of J fixed discrete time intervals: $(t_{j-1}, t_j]$, $j = 1, \cdots, J$. Assume that $t_0 = 0$ and $t_J = +\infty$. Let $S(r) = P(T > t_r)$ be the survival probability beyond the rth time interval. Set $S(t_J) = S(+\infty) = 0$. Let $X(t)$ be the generic notation for time varying covariate process.

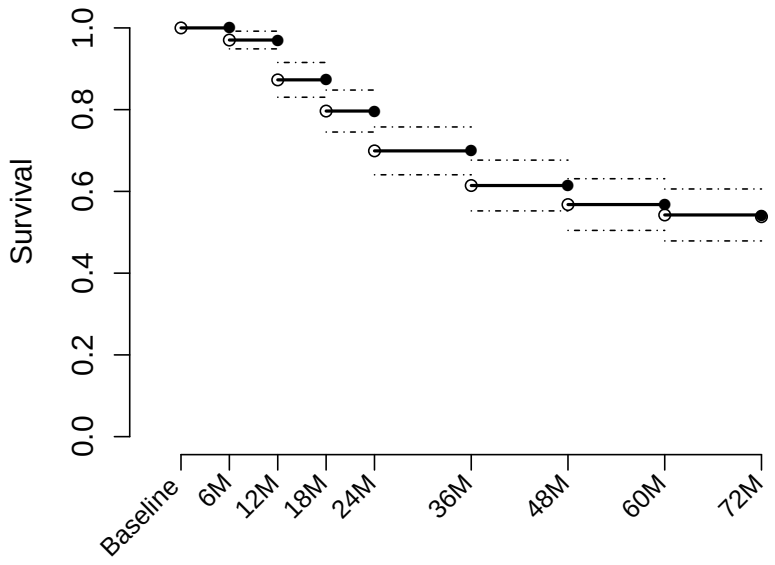


Figure 4.1: Discrete KaplanMeier estimated survival curve (solid line) and its 95% confidence interval (dotted lines).

We assume a Cox proportional hazards model in each time interval, i.e.

$$\lambda(t|\mathbf{X}(t)) = \lambda_0(t)\exp\{\beta'\mathbf{X}(t)\} \quad \text{for } t_{j-1} < t \leq t_j,\ j = 1, \cdots, J,$$

where β is the regression coefficient vector also with an intercept. Further assume that the time varying covaraites are predictable process with constant values in each time interval, $\mathbf{X}(t) = \mathbf{X}(t_{r-1})$ for $t_{r-1} \leq t < t_r$. Then the cumulative hazard function at t_r given the covariate history up to time t_r, $\bar{\mathbf{X}}(t_r)$, is

$$\begin{aligned}
\Lambda(r|\bar{\mathbf{X}}(t)) &= \int_0^{t_r} \lambda_0(u)\exp\{\beta' X(u)\}du \\
&= \sum_{j=1}^{r}\int_{t_{j-1}}^{t_j} \lambda_0(u)\exp\{\beta' X(u)\}du = \sum_{j=1}^{r}\int_{t_{j-1}}^{t_j} \lambda_0(u)\exp\{\beta' X(t_{j-1})\}du \\
&= \sum_{j=1}^{r}\exp\{\beta' X(t_{j-1})\}D\Lambda_{0j} = \sum_{j=1}^{r}\exp\{\lambda_j + \beta' X(t_{j-1})\},
\end{aligned}$$

for $r = 1, \cdots, J$, where $D\Lambda_{0j} = \Lambda_0(t_j) - \Lambda_0(t_{j-1}) = \int_{t_{j-1}}^{t_j} \lambda_0(u)du$ is the change of baseline cumulative hazard function in the jth interval for $j = 1, \cdots, J-1$, $D\Lambda_{0J} = +\infty$, and $\lambda_j = \log(D\Lambda_{0j})$. The survival function at t_r is

$$S(r|\bar{\mathbf{X}}(t_r)) = \exp\{-\Lambda(r|\bar{\mathbf{X}}(t_r))\} = \exp\left\{-\sum_{j=1}^{r} \exp\{\lambda_j + \beta' X(t_{j-1})\}\right\}. \quad (4.1)$$

Let Γ_i indicate the time interval that either event or censoring happens for subject i, and Δ_i indicates whether subject i experienced the event ($\Delta_i = 1$) or was censored ($\Delta_i = 0$) in the interval Γ_i. Assume the censoring mechanism is independent of the converting time. Then given Γ_i, Δ_i and $\mathbf{X}_i(t)$, the non-cure likelihood function for subject i can be written as

$$\begin{aligned}
&L_i(\theta; \Gamma_i = r_i, \Delta_i = \delta_i, \mathbf{X}_i(t)) = \{P(t_{r_i-1} < T_i \leq t_{r_i})\}^{\delta_i} \{P(T_i > t_{r_i})\}^{1-\delta_i} \\
&= \{S(r_i - 1) - S(r_i)\}^{\delta_i} \{S(r_i)\}^{1-\delta_i} \\
&= \left\{1 - \frac{S(r_i)}{S(r_i - 1)}\right\}^{\delta_i} \left\{\frac{S(r_i)}{S(r_i - 1)}\right\}^{1-\delta_i} \{S(r_i - 1)\} \\
&= (1 - \exp\{-\exp\{\lambda_{r_i} + \beta' X_i(t_{r_i-1})\}\})^{\delta_i} (\exp\{-\exp\{\lambda_{r_i} + \beta' X_i(t_{r_i-1})\}\})^{1-\delta_i} \\
&\times \exp\left\{-\sum_{j=1}^{r_i-1} \exp\{\lambda_j + \beta' X_i(t_{j-1})\}\right\},
\end{aligned}$$

with $\theta = (\lambda_1, \cdots, \lambda_{J-1}, \beta')'$ being the vector of parameters.

4.3.3 Variable selection for discrete-time cure-rate survival models using the full likehood

Let $Y = (Y_1, \cdots, Y_n)$ be a vector of latent non-cure indicator variables with Y_i indicating whether the ith individual is in the non-cure portion ($Y_i = 1$) or in the cure portion ($Y_i = 0$). Then given values of $Y = (y_1, \cdots, y_n)$ and the observed data $\mathbf{Obs} = (\mathbf{Z}, \mathbf{X}(t), \Gamma, \Delta)$, the complete likelihood can be written as

$$L_C(\lambda, \beta, \gamma|\mathbf{Obs}, y) = \prod_{i=1}^{n} (1 - \pi(\mathbf{Z}_i))^{1-y_i} \pi(\mathbf{Z}_i)^{y_i} \{L_i(\theta; \Gamma_i = r_i, \Delta_i = \delta_i, \mathbf{X}_i(t))\}^{y_i}, \quad (4.2)$$

and the observed likelihood given **Obs** is

$$
\begin{aligned}
L_O(\lambda,\beta,\gamma|\mathbf{Obs}) &= \prod_{i=1}^{n}\Big[\Big\{\pi(\mathbf{Z}_i)L_i(\theta;\Gamma_i=r_i,\Delta_i=1,\mathbf{X}_i(t))\Big\}^{\delta_i} \\
&\quad\times\Big\{1-\pi(\mathbf{Z}_i)+\pi(\mathbf{Z}_i)L_i(\theta;\Gamma_i=r_i,\Delta_i=0,\mathbf{X}_i(t))\Big\}^{(1-\delta_i)}\Big] \\
&= \prod_{i=1}^{n}\Big[\Big\{\pi(\mathbf{Z}_i)\Big(1-\exp\{-\exp\{\lambda_{r_i}+\beta' X_i(t_{r_i-1})\}\}\Big) \\
&\times\exp\Big\{-\sum_{j=1}^{r_i-1}\exp\{\lambda_j+\beta' X_i(t_{j-1})\}\Big\}\Big\}^{\delta_i} \\
&\times\Big\{1-\pi(\mathbf{Z}_i)+\pi(\mathbf{Z}_i)\Big\{(\exp\{-\exp\{\lambda_{r_i}+\beta' X_i(t_{r_i-1})\}\}) \\
&\times\exp\{-\sum_{j=1}^{r_i-1}\exp\{\lambda_j+\beta' X_i(t_{j-1})\}\}\Big\}\Big\}^{(1-\delta_i)}\Big]. \qquad (4.3)
\end{aligned}
$$

The derivation of (4.3) is given in detail in the appendix. EM algorithm can be conveniently employed to find the penalized MLE, the maximizer of (4.3) with an additional penalty term. It involves maximizing the penalized expected complete likelihood (4.2) conditioning on estimated y_i in each iteration of EM steps. A nice property of the M step is that it can be undertaken with respect to γ and β separately, thus simplifies the maximization steps in the EM. The logarithm of the expected complete likelihood can be written as $l = l_1 + l_2$, with

$$
\begin{aligned}
l_1 &= \sum_{i=1}^{n}\hat{y}_i\log[\pi(\mathbf{Z}_i)]+(1-\hat{y}_i)\log[1-\pi(\mathbf{Z}_i)], \qquad (4.4)\\
l_2 &= \sum_{i=1}^{n}\hat{y}_i\delta_i\log\left(1-\exp\left\{-\exp\left\{\lambda_{r_i}+\beta'\mathbf{X}_i(t_{r_i-1})\right\}\right\}\right) \qquad (4.5)\\
&\quad+\sum_{i=1}^{n}\hat{y}_i(1-\delta_i)\left\{-\exp\left\{\lambda_{r_i}+\beta'\mathbf{X}_i(t_{r_i-1})\right\}\right\} \\
&\quad+\sum_{i=1}^{n}\hat{y}_i\left\{-\sum_{j=1}^{r_i-1}\exp\{\lambda_j+\beta'\mathbf{X}_i(t_{j-1})\}\right\},
\end{aligned}
$$

where parameters γ's are only involved in l_1 and β's are only involved in l_2, and $\hat{y}_i$ is conditional expectation of Y_i.

For the case that both predictors for the cure rate and the non-cure survival are

of high-dimensional, penalized maximization is applied in M-steps for the purpose of variable selection. We provide the details of each EM step in the following.

1. E-steps.

Given the observations $\boldsymbol{O} = (\boldsymbol{r}, \boldsymbol{\delta}, \mathbf{X}, \mathbf{Z})$ and parameter estimates from the kth step $\boldsymbol{\Theta}^{(m)} = (\beta, \gamma; \lambda)$ with $\lambda = (\lambda_1, \cdots, \lambda_{J-1})$, the conditional expectations of the latent variables Y_i, $i = 1, \cdots, n$, are

$$\begin{aligned} \hat{y}_i^{(m)} &= E(Y_i|\boldsymbol{O}, \boldsymbol{\Theta}^{(m)}) \\ &= \delta_i + (1-\delta_i)\frac{\pi(\mathbf{Z}_i)S(r_i|Y_i=1, \bar{\mathbf{X}}_i(t_{r_i}))}{1-\pi(\mathbf{Z}_i)+\pi(\mathbf{Z}_i)S(r_i|Y_i=1, \bar{\mathbf{X}}_i(t_{r_i}))} \end{aligned} \tag{4.6}$$

with $S(r_i|Y_i = 1, \mathbf{X}_i(t))$ as given in (4.1) (Cai et al., 2012).

2a. M-steps: updating the nuisance parameters.

When fix β, the estimates of the nuisance parameters $\lambda_1, \cdots, \lambda_{J-1}$ in each EM step can be profiled out and need to satisfy the score equations:

$$\partial l_2/\partial \lambda_s = 0, \ s = 1, \cdots, J-1.$$

The Newton-Raphson method can be used to solve the estimating equations. The first order partial derivative of l_2 with respect to each λ_s, $s = 1, \cdots, J-1$, is:

$$\frac{\partial l_2}{\partial \lambda_s} = \sum_{i=1}^{n} \hat{y}_i \delta_i \frac{\exp\{-h_{is}\}h_{is}}{1-\exp\{-h_{is}\}} I(s = r_i) - \sum_{i=1}^{n} \hat{y}_i (1-\delta_i) h_{is} I(s = r_i) - \sum_{i=1}^{n} \hat{y}_i h_{is} I(s < r_i),$$

where $h_{is} = \exp\{\lambda_s + \beta' \mathbf{X}_i(t_{s-1})\}$. Let

$$b_{is} = \frac{h_{is}\exp\{-h_{is}\}}{1-\exp\{-h_{is}\}}\left(1 - \frac{h_{is}}{1-\exp\{-h_{is}\}}\right), \quad 1 \le i \le n, 1 \le s \le J-1.$$

Then the second partial derivatives are

$$\begin{aligned} &\frac{\partial^2 l_2}{\partial \lambda_s^2} = \sum_{i=1}^{n} \hat{y}_i \delta_i b_{is} I(s = r_i) - \hat{y}_i (1-\delta_i) h_{is} I(s = r_i) - \hat{y}_i h_{is} I(s < r_i) \\ &\frac{\partial^2 l_2}{\partial \lambda_s \partial \lambda_t} = 0, \ s \neq t. \end{aligned} \tag{4.7}$$

Let I_λ be the $(J-1)\times(J-1)$ matrix of $-\partial^2 l_2/\partial\lambda_s\partial\lambda_t$, $1\le s,t\le J-1$, then at each M-step, the nuisance parameters $\lambda=(\lambda_1,\cdots,\lambda_{J-1})$ can be updated by the Newton-Raphson method till convergence through

$$(\lambda_1^{(k)},\cdots,\lambda_{J-1}^{(k)})' = (\lambda_1^{(k-1)},\cdots,\lambda_{J-1}^{(k-1)})' + \left\{ I_\lambda^{-1}\frac{\partial l_2}{\partial\lambda}\right\}\Big|_{\lambda=\lambda^{(k-1)}}. \tag{4.8}$$

In real applications, even in the low-dimensional cases, time intervals that contains very few observations could yield very biased estimates. We suggest congregating adjacent such intervals into one bigger interval in such cases.

2b. M-steps: updating γ and β.

In the maximization steps, the following expectations are maximized, plus elastic-net penalty terms given later, with respect to γ and β, respectively,

$$\begin{aligned} l_1(\gamma) &= \sum_{i=1}^n \hat{y}_i \log[\pi(\mathbf{Z}_i)] + (1-\hat{y}_i)\log[1-\pi(\mathbf{Z}_i)], \qquad (4.9)\\ &= \sum_{i=1}^n \hat{y}_i(\gamma'\mathbf{Z}_i) - \log(1+\exp(\gamma'\mathbf{Z}_i)),\\ l_2(\beta) &= \sum_{i=1}^n \hat{y}_i\delta_i \log\left(1-\exp\left\{-\exp\left\{\hat{\lambda}_{r_i}+\beta'\mathbf{X}_i(t_{r_i-1})\right\}\right\}\right) \qquad (4.10)\\ &\quad +\hat{y}_i(1-\delta_i)\left\{-\exp\left\{\hat{\lambda}_{r_i}+\beta'\mathbf{X}_i(t_{r_i-1})\right\}\right\}\\ &\quad +\hat{y}_i\left\{-\sum_{j=1}^{r_i-1}\exp\{\hat{\lambda}_j+\beta'\mathbf{X}_i(t_{j-1})\}\right\}. \end{aligned}$$

The first and the second order partial derivatives of l_1 with respect to γ are respectively:

$$\begin{aligned} \frac{\partial l_1}{\partial\gamma_j} &= \sum_{i=1}^n \hat{y}_i Z_{ij} - \frac{Z_{ij}\exp\{\gamma'\mathbf{Z}_i\}}{1+\exp\{\gamma'\mathbf{Z}_i\}},\\ \frac{\partial^2 l_1}{\partial\gamma_j\partial\gamma_{j'}} &= \sum_{i=1}^n Z_{ij}Z_{ij'}\left[-\frac{\exp\{\gamma'\mathbf{Z}_i\}}{1+\exp\{\gamma'\mathbf{Z}_i\}}+\left(\frac{\exp\{\gamma'\mathbf{Z}_i\}}{1+\exp\{\gamma'\mathbf{Z}_i\}}\right)^2\right], \end{aligned}$$

with $1\le j,j'\le q$, q is the number of variables in Z. And the first and the second order partial derivatives of l_2 with repsect to β are respectively:

$$
\begin{aligned}
\frac{\partial l_2}{\partial \beta_k} &= \sum_{i=1}^{n}\left[\hat{y}_i\delta_i\frac{\exp\{-\hat{h}_{ir_i}\}\hat{h}_{ir_i}X_{ik}(t_{r_i-1})}{1-\exp\{-\hat{h}_{ir_i}\}} - \hat{y}_i(1-\delta_i)\hat{h}_{ir_i}X_{ik}(t_{r_i-1})\right. \\
&\quad \left. - \sum_{j=1}^{r_i-1}\hat{y}_i\hat{h}_{ij}X_{ik}(t_{j-1})\right], \\
\frac{\partial^2 l_2}{\partial \beta_k \partial \beta_{k'}} &= \sum_{i=1}^{n}\left[\hat{y}_i\delta_i\hat{b}_{ir_i}X_{ik}(t_{r_i-1})X_{ik'}(t_{r_i-1}) - \hat{y}_i(1-\delta_i)\hat{h}_{ir_i}X_{ik}(t_{r_i-1})X_{ik'}(t_{r_i-1})\right. \\
&\quad \left. - \sum_{j=1}^{r_i-1}\hat{y}_i\hat{h}_{ij}X_{ik}(t_{j-1})X_{ik'}(t_{j-1})\right], \qquad (4.11)
\end{aligned}
$$

with $\hat{h}_{ij} = \exp\{\hat{\lambda}_{r_i} + \beta'\mathbf{X}_i(t_{j-1})\}$, $\hat{b}_{ij} = b_{ij}(\hat{h}_{ij})$ and $1 \le j \le J-1$. And $1 \le k, k' \le p$, p is the number of variables in X.

We propose an elastic-net (Zou and Hastie, 2005; Zou and Zhang, 2009) penalized maximization procedure for the purpose of variable selection, which takes care of the spatial correlations of the voxel measures. For β, the penalty takes the form (Yang and Zou, 2013; Zou and Hastie, 2005):

$$
Pen(\beta) = \sum_k \left\{\alpha v_k|\beta_k| + \frac{1}{2}(1-\alpha)\beta_k^2\right\},\ 0 < \alpha \le 1.
$$

Here v_k are adaptive weights and α is a tuning parameter for a balance between the lasso and ridge penalties. The lasso penalty tends to select independent variables while the ridge penalty tends to select correlated variables together. The same form of penalty applies to γ. Specifically, in the high-dimensional setting, we aim to minimize the following objective functions in the M-steps:

$$
\begin{aligned}
F_1^{obj}(\gamma) &= -\frac{1}{n}l_1(\gamma) + \lambda_\gamma \sum_j \left\{\alpha v_j|\gamma_j| + \frac{1}{2}(1-\alpha)\gamma_j^2\right\}, \text{ and} \qquad (4.12) \\
F_2^{obj}(\beta) &= -\frac{1}{n}l_2(\beta) + \lambda_\beta \sum_k \left\{\alpha v_k|\beta_k| + \frac{1}{2}(1-\alpha)\beta_k^2\right\}. \qquad (4.13)
\end{aligned}
$$

Minimizing $F_1^{obj}(\gamma)$ and $F_2^{obj}(\beta)$ are equivalent to optimization problems of maximizing $l_1(\gamma)$ subject to $(1-a_1)\|v*\gamma\|_1 + a_1\|\gamma\|_2^2 \le R_1$ and maximizing $l_2(\beta)$ subject

to $(1-a_2)\|v*\beta\|_1 + a_2\|\beta\|_2^2 \leq R_2$ with $a_1 = a_2 = \frac{1-\alpha}{1+\alpha}$ and some positive real numbers R_1 and R_2 (Zou and Hastie, 2005; Yang and Zou, 2013), where $*$ denotes component-wise multiplication of two vectors.

The following theorem states that the expectations $-l_1(\gamma)$ and $-l_2(\beta)$ in the above equations are both smooth convex functions. This, together with the fact that the Elastic-net penalty function is convex and separable in terms of either γ or β, suggests the use of coordinate descent algorithm in optimizing (4.12) and (4.13).

Theorem IV.1. *The expectation functions $l_1(\gamma)$ and $l_2(\beta)$ in equations (4.9) and (4.10) are concave functions with respect to γ and β respectively. Further more, the conditional expectation of the complete likelihood function (4.2) is a joint concave function with respect to (γ, β).*

The following theorem IV.2 states that the second partial derivatives of $-l_1(\gamma)$ with respect to each γ_j are uniformly bounded from above. With this property, (4.12) can be minimized using the majorization-minimization coordinate descent algorithm (Yang and Zou, 2013), which does not require inner iterations to solve for the fixed point solution to the Karush-Kuhn-Tucker score equation in each step of the coordinate descent and therefore much faster than the conventional coordinate descent algorithm.

Theorem IV.2. *The second partial derivative terms $-\partial^2 l_1(\gamma)/\partial\gamma_j^2$ are uniformly bounded from above for all $\gamma \in \mathbb{R}^q$.*

We will briefly summarize majorization-minimization coordinate descent algorithm in the following. Suppose at the mth step, we need to update the jth coordinate by minimizing the following objective function with the other coordinates fixed at their values obtained in the $(m-1)$th step:

$$g(\gamma_j) = -\frac{1}{n} l_1(\gamma_j | \gamma_{-j} = \hat{\gamma}_{-j}^{(m-1)}) + \lambda_\gamma \left\{ \alpha v_j |\gamma_j| + \frac{1}{2}(1-\alpha)\gamma_j^2 \right\}.$$

1. Set the majorization function to be

$$\begin{aligned} q(\gamma_j|\hat{\gamma}^{(m-1)}) &= -\frac{1}{n}l_1(\hat{\gamma}^{(m-1)}) - \frac{1}{n}\partial_j l_1(\hat{\gamma}^{(m-1)})(\gamma_j - \hat{\gamma}_j^{(m-1)}) + \frac{D_j}{2n}(\gamma_j - \hat{\gamma}_j^{(m-1)})^2 \\ &\quad +\lambda_\gamma \left\{ \alpha v_j|\gamma_j| + \frac{1}{2}(1-\alpha)\gamma_j^2 \right\}, \end{aligned}$$

where $|\partial^2 E l_1(\gamma)/\partial\gamma_j^2| \leq D_j$ is the uniform bound of the second partial derivative with respect to jth coordinate.

2. At each updating step, we optimize the majorization function by

$$\hat{\gamma}_j^{(m)} = \frac{S\left[D_j\hat{\gamma}^{(m-1)} + \partial_j l_1(\hat{\gamma}^{(m-1)}),\, n\lambda_\gamma\alpha v_j\right]}{D_j + n\lambda_\gamma(1-\alpha)}, \tag{4.14}$$

where $S[\cdot,\diamond] = (|\cdot| - \diamond)_+\text{sgn}(\cdot)$ is the soft-thresholding function.

3. Then from the majorization and minimization, we have

$$g(\hat{\gamma}_j^{(m)}) \leq q(\hat{\gamma}_j^{(m)}|\gamma^{(m-1)}) \leq q(\hat{\gamma}_j^{(m-1}|\gamma^{(m-1)}) = g(\hat{\gamma}_j^{(m-1)}).$$

However $-l_2$ does not have the property of bounded second derivatives. We use a quadratic programming approach (Tibshirabi, 1997; Engler and Li, 2009; Hastie and Tibshirabi, 1990) to optimize (4.13). Specifically, let $\eta = \beta'\mathbf{X}$, $\mu = \partial l_2/\partial\eta$, $A = -\partial^2 l_2/\partial\eta\partial\eta'$ and $z = \eta + A^{-1}\mu$. Notice that λ and η always show together in the exponential terms in l_2, similar to the calculation of $\partial^2 l_2/\partial\lambda\partial\lambda'$ in (4.7), it is easy to see that A is a diagonal matrix. The first order Taylor series expansion of $-l_2$ can then be expressed as $(z-\eta)'A(z-\eta)$. Minimizing $F_2^{obj}(\beta)$ can be approximated by solving the penalized weighted least square problem

$$\arg\min_\beta \frac{1}{n}(z-\eta)'A(z-\eta) + \lambda_\beta \sum_k \left\{ \alpha v_k|\beta_k| + \frac{1}{2}(1-\alpha)\beta_k^2 \right\}. \tag{4.15}$$

Set $Q = A^{1/2}$ and let $\tilde{z} = Qz$, $\tilde{\mathbf{X}} = Q\mathbf{X}$. Then (4.15) can be replaced by the

unweighted elastic net least square problem

$$\arg\min_{\beta} \frac{1}{n}(\tilde{z} - \beta'\tilde{\mathbf{X}})'(\tilde{z} - \beta'\tilde{\mathbf{X}}) + \lambda_\beta \sum_k \left\{ \alpha v_k |\beta_k| + \frac{1}{2}(1-\alpha)\beta_k^2 \right\},$$

which yields

$$\hat{\beta}_k = \frac{S\left[(\tilde{z} - \beta'_{-k}\tilde{\mathbf{X}})'\tilde{\mathbf{X}}_{\cdot k},\, n\lambda_\beta \alpha v_k\right]}{\|\tilde{\mathbf{X}}_{\cdot k}\|_2 + n\lambda_\beta(1-\alpha)}, \tag{4.16}$$

where $S[\cdot, \diamond] = (|\cdot| - \diamond)_+ \text{sgn}(\cdot)$ is the soft-thresholding function and β'_{-k} is β with its kth entry replaced by zero.

Notice that A, and therefore $\tilde{z}$ and $\tilde{\mathbf{X}}$, also depend on β, so an inner loop of iteration is needed for solving the fixed point solution of (4.16) when updating the kth entry. Empirically, we found it still converges to the global minimizer without the inner loops. Similar to the idea of the mixed coordinate descent algorithm (Li et al., 2013), update of $\hat{\beta}$ in one step without inner loop decreases the objective function. Even though the amount of decrease is less than that from the fixed point solution update with the inner iteration, but overall, updating (4.16) without inner loops converges to the same global minimizer and it saves the computational time significantly.

4.4 Computational algorithms

When both $\mathbf{Z}$ and $\mathbf{X}$ are of high-dimensional, the algorithm for variable selection goes as: (i) Initialize $\hat{\beta}$, $\hat{\gamma}$, $\hat{\lambda}$, y. We use zeros as initial values in our numerical studies. (ii) In the E-step, update the conditional expectation of y by (4.6). (iii) In the M-step, update $\hat{\lambda}$ by iterating (4.8), update $\hat{\beta}$ by iterating (4.16) and update $\hat{\gamma}$ by iterating (4.14). Recall the algorithms used for those three updates are Newton-Rhapson, quadratic programming and majorization-minimization. (iv) Repeat (ii) and (iii) until $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\lambda}$ all converge.

In the high-dimensional settings, we use 5-fold cross validation to select the tunning parameters λ_β and λ_γ based on predicted observed log likelihood values. For comparison, we also tried variable selection via the stability selection (Meinshausen

and Bühlmann, 2010), which gives similar results on variable selection. Since the stability selection does not provide estimation of the regression coefficients, which are needed for prediction, therefore its results are not reported here. The value of α should be determined based on how strong the image voxels are correlated. In our numerical studies, we used a fixed α value in each setting according to the prior knowledge on cross-predictor correlations (Zhou et al., 2010).

The above algorithm for high-dimensional and together with algorithms for cases when either or both of $\mathbf{Z}$ and $\mathbf{X}$ are of low-dimensional using Newton-Raphson algorithm are programmed into an R package "SeCuredSurv" (SElection for CURE-rate Discrete-time SURVival models), and will be uploaded on CRAN.

4.5 Simulation studies

In this section, we investigate the performance of the proposed method on simulated data sets mimicking the structure of the longitudinal PET imaging data. We assume that the cure rate only depends on baseline image. Also we assume that α value to be the same for (4.12) and (4.13) since the cross-voxel correlation level for images at different time points would likely be similar for the same individual.

Three high-dimensional simulation settings are studied, one with independent imaging predictors and the other two with spatially correlated predictors with an AR(1) correlation coefficient $\rho = 0.5$ and $\rho = 0.9$ respectively. Images are also longitudinally autocorrelated following an AR(1) model with a serial correlation coefficient $\rho = 0.3$. Each data set contains 500 subjects observed or censored on 6 discrete time intervals. There are 1000 voxels in each image, with the true parameter values $\beta_1 = \cdots = \beta_{20} = 2$, $\beta_{21} = \cdots = \beta_{40} = -2$, $\beta_{41} = \cdots = \beta_{1000} = 0$; $\gamma_1 = \cdots = \gamma_{100} = 0$, $\gamma_{101} = \cdots = \gamma_{120} = 2$, $\gamma_{121} = \cdots = \gamma_{140} = -2$, $\gamma_{141} = \cdots = \gamma_{1000} = 0$; $(\lambda_1, \cdots, \lambda_5) = (-2, -1, 0, 0, 0)$. The non-cure indicator variable y_i for subject i was randomly drawn from a Bernoulli distribution with probabilities of success $\pi(\mathbf{Z}_i)$. For a non-cured subject, the probability of conversion being observed in each time interval is calculated from (4.1), then the event time interval is randomly drawn from

Table 4.2: Variable selection results out of 100 replicated data sets

	FN	FP	SE(%)	SP(%)
Independent variables:				
γ	13.18(3.13)	62.60(7.91)	67.05(7.82)	93.48(0.82)
β	12.08(3.40)	89.60(9.24)	69.80(8.50)	90.67(0.96)
Correlated variables($\rho = 0.5$):				
γ	6.46(3.52)	1.91(1.79)	83.85(8.79)	99.80(0.19)
β	7.96(2.40)	73.66(10.82)	80.10(6.00)	92.34(1.13)
Correlated variables($\rho = 0.9$):				
γ	0.00(0.00)	0.66(0.94)	100.00(0.00)	99.93(0.10)
β	0.01(0.10)	3.09(2.12)	99.98(0.25)	99.68(0.22)

• FN=false negative, FP=false positive, SE(%)=sensitivity in percentage, SP(%)=specificity in percentage.
• Numbers in parentheses are stand errors of the estimates.

a multinomial distribution with six cells. Censoring time intervals are generated from a discrete uniform subdistribution at the first $J-1$ time intervals combined with a truncation at the last time interval, which yield a censoring rate about $15\% - 20\%$.

For each setting, we generate 100 replicated data sets. Model fitting is conducted on each replicated data set with the optimal tuning parameters selected by a 5-fold cross validation to give the largest observed likelihood value. We simply use a fixed α value in each setting. For the independent settings, we set $\alpha = 0.9$, i.e., assign more weight to the lasso penalty over the ridge penalty, and for the correlated settings, we set $\alpha = 0.5$ and 0.9 for the cases $\rho = 0.5$ and 0.9 respectively. Analysis results are summarized over the 100 replicates in each setting. Table 4.2 shows the selection results. Averages for false positive number, false negative number, sensitivity (in percentage) and specificity (in percentage) for variable selection are reported. In the correlated data cases, since we put more weight on the ridge penalty, the correlated important variables tend to be selected together. As a result, there are much less false negatives compared to the independent case.

We then investigate the prediction performance of the proposed model by looking at its predictive power on a separate test data set independently generated with 1000 subjects. We specifically looked at two types of quantities: the individual level predicted non-cure rate calculated by $\check{\pi}_i = \exp(\hat{\gamma}'\mathbf{Z}_i)/(1+\exp(\hat{\gamma}'\mathbf{Z}_i))$ and the individual level predicted probabilities $\check{Pr}_i(r) = \check{S}(r|\mathbf{X}(t)) - \check{S}(r-1|\mathbf{X}(t))$ with $\check{S}(r|\mathbf{X}(t))$ calculated by (4.1). We applied model fitting results from 100 independently generated training sets to the separate test set. Table 4.3 gives the prediction results on the individual-level non-cure indicators. The predicted non-cure indicator $\check{y}_i$ for subject i is set to be 1 if $\check{\pi}_i \geq 0.5$ and 0 otherwise. The receiver operating characteristic (ROC) curve for each case was achieved by varying the cut-off point of the predicted $\check{\pi}_i$'s. Area under the ROC curve (AUC) and the oracle AUC are also reported. The oracle AUC are calculated from prediction with $\hat{\beta}$ and $\hat{\gamma}$ estimated by Newton-Raphson algorithm given the knowledge of what coefficients are nonzero. Table 4.4 reflects the prediction results of the survival probabilities. We set the predicted event time interval for a subject to be the interval with the largest predicted interval probability. Table 4.4 lists the frequency distribution of the difference between the predicted and observed event time intervals for converted subjects in the separate test set.

4.6 Analyzing ADNI data

First, we group the survival time into $(0,6]$, $(6,12]$, $(12,18]$, $(18,24]$, $(24,36]$, $(36,48]$, $(48,60]$ and $(60,+\infty]$ (in month) eight disjoint discrete time intervals. Assuming that the cure rate is only relevant to the early stage brain imaging, we set $\mathbf{Z}$ to be the baseline image for everyone. If an image at the starting point of certain time interval is missing, then the image values in that time interval are assigned to be the same as the image in the previous interval. To be able to compute with the imaging data, we reduce the resolution of each image to $40 \times 40 \times 24$ in Brodmann functional regions (Brodmann, 2010), i.e., each voxel in the new image contains $4 \times 4 \times 4$ nearby voxels in the original image. The image value of a new voxel is

Table 4.3: Prediction results on individual non-cure status

	FN	FP	SE(%)	SP(%)	AUC	Oracle AUC
• Independent variables:						
mean	162.08	148.95	68.89	68.90	0.753	0.962
(s.e.)	(15.71)	(13.09)	(3.12)	(2.64)	(0.032)	(0.006)
• Correlated variables($\rho = 0.5$):						
mean	45.47	38.02	90.89	92.41	0.948	0.966
(s.e.)	(13.64)	(15.04)	(2.73)	(3.00)	(0.021)	(0.010)
• Correlated variables($\rho = 0.9$):						
mean	23.84	21.81	95.39	95.48	0.967	0.968
(s.e.)	(25.35)	(23.61)	(4.90)	(4.89)	(0.032)	(0.012)

• s.e.=standard error.
• There are 479, 501 and 483 non-cure subjects in the separate testing sets for the independent, correlated with $\rho = 0.5$ and correlated with $\rho = 0.9$ cases, respectively.

the average of the original $4 \times 4 \times 4$ voxel values. As a result, each image contains 6582 congregated voxels. We standardized each image by subtracting its mean and dividing by its standard deviation. We also include patient baseline age and gender as covariates and intercept terms for both γ and β in the model.

We then run 5-fold cross-validation to select the optimal λ_γ and λ_β. We fixed α value at 0.5 to achieve a relative balance between the lasso and the ridge penalties. The selected voxels for non-cure survival and the cure rate with their regression coefficients are depicted in Figure 4.2 and 4.3 respectively. There are 43 selected nonzero β regression coefficients that are associated with non-cure survival and 12 selected nonzero γ regression coefficients that are associated with cure rate. Some findings are consistent with the current literature on early stage AD progression. For examples, voxels in region BA31 in posterior cingulate (highlighted at frontal-center position in the slide at -10mm in Figure 4.2) are identified to be associated with the non-cure survival ($\hat{\beta} = -6.92 \times 10^{-2}$). Huang et al. (2002) reported that reduced relative blood flow of the posterior cingulate gyrus were found about two years before the MCI-to-AD conversion. Association effect of posterior cingulate with AD was also reported in Jacobs et al. (2012). The strongest signal is in the thamalus region with $\hat{\beta} = 2.06 \times 10^{-1}$ (highlighted in the slide at +2mm in Figure 4.2). The

association effect of thamalus with AD is supported by Moretti et al. (2012) and Vogt (2009). Other regions associated with non-cure survival, along with their top voxel's effect sizes and literature supports, include region BA7 in superior parietal cortex and BA18 (Okello et al. (2009)) in occipital cortex both with $\hat{\beta} = -1.23\times10^{-1}$ (negative effect voxels highlighted in the slides at -16mm and -10mm in Figure 4.2); region BA32 in anterior cingulate cortex with $\hat{\beta} = -1.10\times10^{-1}$ (negative effect voxels highlighted in the lower half of slide at -10mm and -4mm in Figure 4.2, Ye et al. (2012); Jones et al. (2006) and Fennema-Notestine et al. (2009)); region BA10 in medial frontal cortex with $\hat{\beta} = -1.10 \times 10^{-1}$ (negative effect voxels highlighted in the lower half of slide at -4mm in Figure 4.2, Sun et al. (2013)); region BA4 in primary motor cortex with $\hat{\beta} = 1.10 \times 10^{-1}$ (voxels highlighted in the right half of slide at -28mm in Figure 4.2, Suva et al. (1999)); region BA6 in pre-motor cortex with $\hat{\beta} = 9.40 \times 10^{-2}$ (voxels highlighted in the right half of slide at -28mm in Figure 4.2, Annweiler et al. (2013)); and region BA20 in temporal cortex with $\hat{\beta} = -7.70\times10^{-2}$ (Risacher et al. (2009)). Almost all the selected signals associated with cure rate are in either region BA18 or BA31, both of which are found to be associated with the non-cure survival as well. Neither baseline age nor gender is selected for either non-cure survival or cure rate.

Figure 4.4 gives the cross-validated frequency of the difference between predicted and observed event time intervals for the observed cases. In particular, we split the data randomly into five groups of about equal sizes, then take one of them as the test set and the other four combined as the training set at a time. We use the training set to fit the model and then apply the estimated values of selected parameters to the test set to predict cure rate and event time for each individual as what we did in the simulations. We then take a different group as the test set and repeat the above steps till all the five groups have been taken as the test set. More than 75% of the time, the predicted event time interval is within two adjacent time intervals of the observed one. Since the true non-cure statuses are unknown, we can not calculate the non-cure prediction error here as we did in the simulations. However,

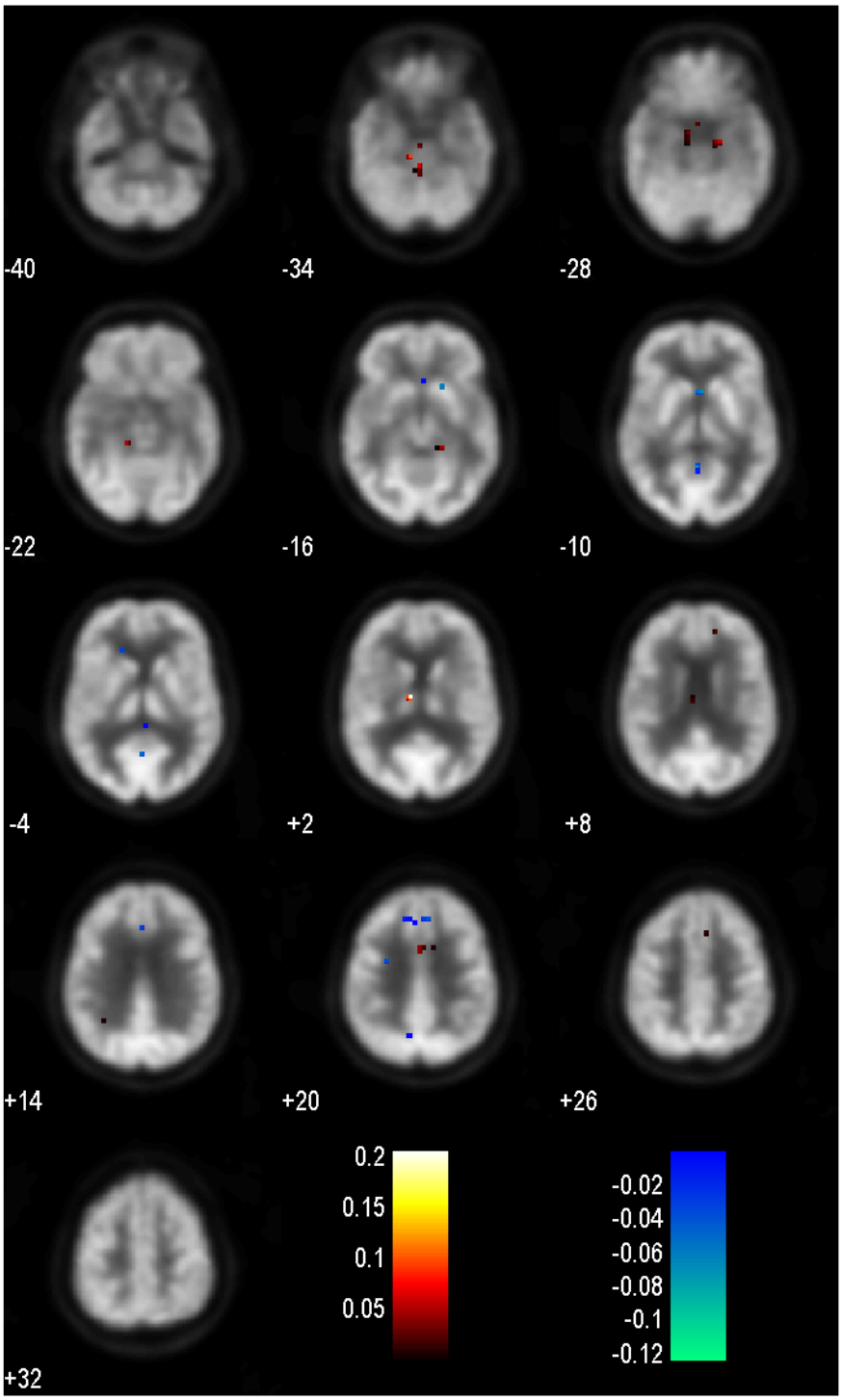


Figure 4.2: Selected collapsed voxels associated with non-cure survival. Viewed on the axial axis. Warm color for positive effects and cold color for negative effects.

the average of individual level predicted non-cure rate is 0.491 ± 0.07, which can be used as an estimate for the population level non-cure rate, and it is consistent with what suggested by the discrete Kaplan-Meier estimated survival curve in Figure 4.1.

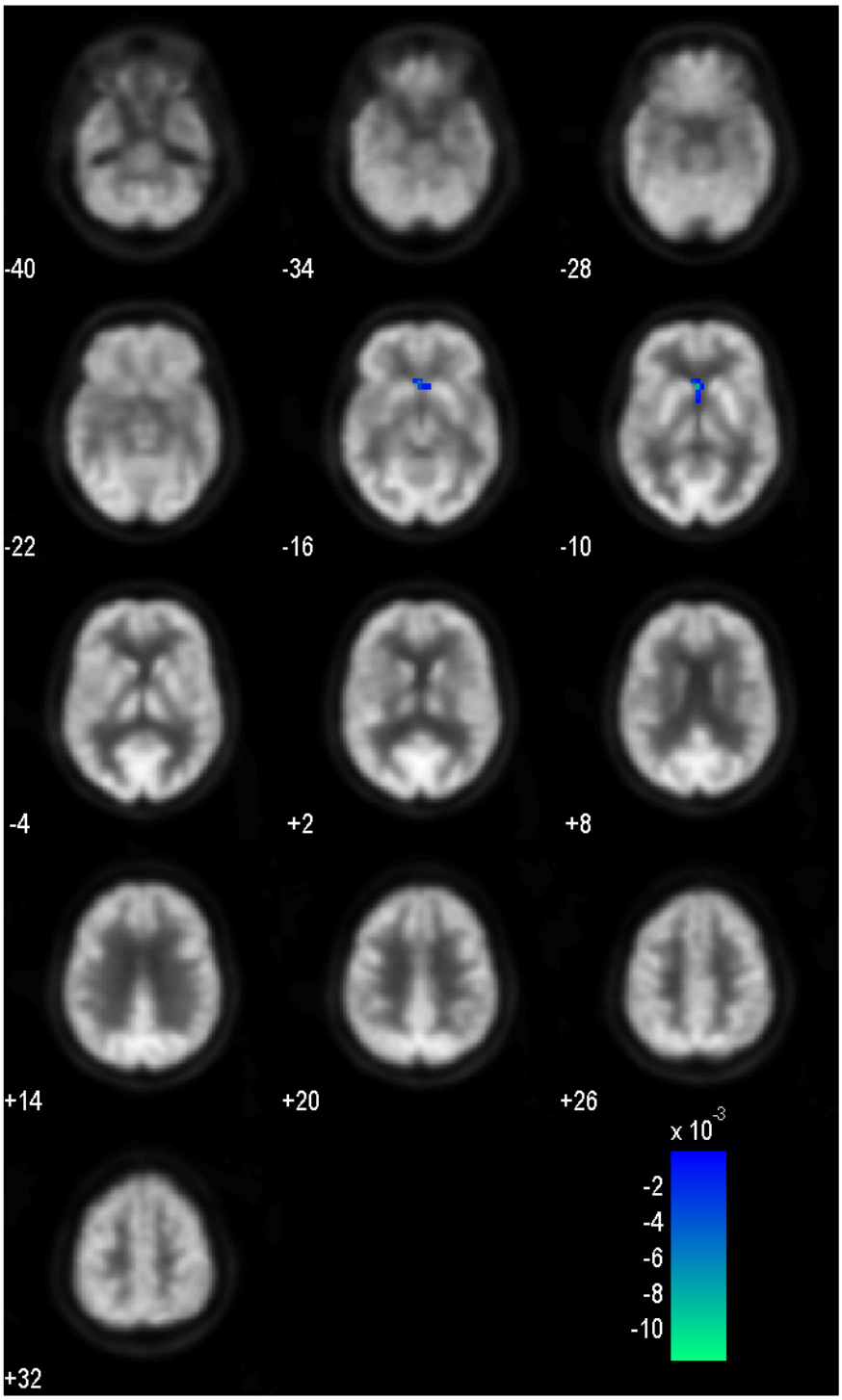


Figure 4.3: Selected collapsed voxels associated with cure rate. Viewed on the axial axis.

4.7 Discussion

It is known that the cure models sometimes face the identifiability issues (Farewell, 1982), which means it is impossible to distinguish a censored non-cured subject from

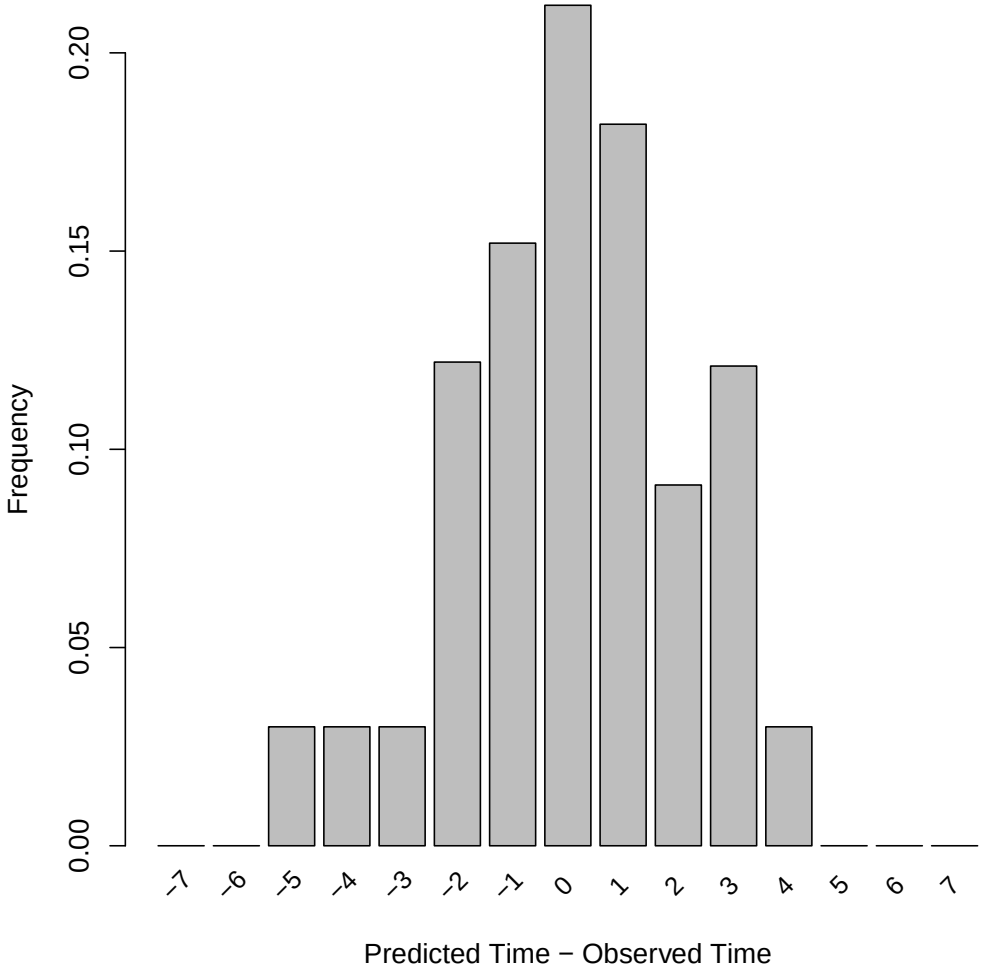


Figure 4.4: Frequencies of difference between predicted and observed event times for ADNI data. Total number of observed cases is 109.

a cured subject. Lia et al. (2001) showed that for the mixed cure model with a logistic linked cure rate and PH non-cure survival, the model is identifiable. The main theorem in Lia et al. (2001) can be applied to our discrete time survival model with some minor modification.

Table 4.4: Distribution of difference between predicted and observed event times

$T_{Pred} - T_{Obs}$	-5	-4	-3	-2	-1	0	1	2	3	4	5
Independent variables:											
Mean (%)	0.21	1.17	3.33	8.24	15.50	56.64	11.20	3.02	0.51	0.15	0.04
(s.e. %)	(0.26)	(0.46)	(0.87)	(1.41)	(2.88)	(3.70)	(3.49)	(1.57)	(0.61)	(0.32)	(0.22)
Correlated variables($\rho = 0.5$):											
Mean (%)	0.20	0.13	1.10	1.65	4.23	80.31	5.99	3.21	1.64	0.98	0.57
(s.e. %)	(0.20)	(0.15)	(0.42)	(0.56)	(1.60)	(3.85)	(2.39)	(1.20)	(1.08)	(0.56)	(0.42)
Correlated variables($\rho = 0.9$):											
Mean (%)	0.01	0.73	1.62	4.45	3.93	87.83	1.17	0.21	0.06	0.005	0.00
(s.e. %)	(0.04)	(0.30)	(0.29)	(0.84)	(1.32)	(2.12)	(0.90)	(0.26)	(0.14)	(0.03)	(0.00)

- s.e.=standard error.
- The mean and s.e. are taken over 100 independent predictions.
- There are 395, 405 and 405 converted subjects in the separate testing sets for the independent, correlated with $\rho = 0.5$ and correlated with $\rho = 0.9$ cases, respectively.

CHAPTER V

Future work

In this dissertation, we have developed high-dimensional variable selection methodologies for either structured multivariate or discrete time survival settings. We have applied the methodologies to a yeast gene network eQTL study, a brain-wide and genome-wide association study and a longitudinal brain imaging study for predicting MCI-to-AD conversions.

For multivariate data, oftentimes, the group structure for the responses is unknown. Structure learning techniques, such as cluster analysis or factor analysis, can be applied to explore the response grouping structure first before applying the MSGLasso. Since the MSGLasso requires the group structure to be pre-determined, it is not immune to the mis-specification of the group structure. Some interests have been shown on learning the response group structure and selecting the important variables simultaneously. Yin and Li (2011) proposed a conditional Guassian graphical model to select nonzero entries in the precision matrix conditioning on simultaneously selected predictors. It is still of interest to see how one can select important predictors via the MSGlasso based on a simultaneously learned response group structure.

Grouping techniques had also been widely used in genetics to detect rare variants (Li and Leal, 2008). Zhou et al. (2010) used a sparse-group lasso penalized logistic regression to jointly select common and rare genetic variants that are associated with the risk of familial breast cancer. Biswas and Lin (2012) used a logistic Bayesian lasso method to detect common and rare haplotypes in association with

age-related macular degeneration (AMD). In practice, the MSGLasso can be used to simultaneously detect rare and common variants associated to the human brain functions in brain-wide GWA studies.

In our brain-wide GWA study, the quality of the post-selection estimation and inference on the selected voxel-to-SNP set depends on the performance of the selection stages. In recent years, many studies have been focusing on improving the performance of post-selection inference (Berk et al., 2013; Taylor et al., 2014). Theoretical properties of the post-selection inference remain challenging for the MSGlasso, especially in ultra-high dimensional cases.

Besides the fact that the variable selection for discrete time survival models has a wide range of applications, its theoretical finite sample properties are also very interesting. Kong and Nan (2013) and Huang et al. (2013) considered the non-asymptotic oracle properties for the Cox PH model with the lasso penalty. The oracle properties for survival models with grouped interval censored data are worth further investigation.

APPENDICES

APPENDIX A

Appendix for Chapter II

Proofs of technical results

Some matrix algebra

Lemma A.1. *Let*

$$L^0(B) = \frac{1}{2}\|Y - XB\|_2^2 = \frac{1}{2}tr(Y - XB)^{\mathrm{T}}(Y - XB) = \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{q}(y_{ik} - \sum_{j=1}^{p} x_{ij}\beta_{jk})^2.$$

Then

$$\partial L^0(B)/\partial\beta_{jk} = -x_j^{\mathrm{T}}(Y - XB)_{\cdot k} = -S_{jk} + \|x_j\|_2^2\beta_{jk},$$

where $S_{jk} = x_j^{\mathrm{T}}(Y - XB^{(-j)})_{\cdot k}$.

Proof of Theorem 3.1

Following Lemma A.1,

$$\begin{aligned} L(B) &= \frac{1}{2n}\|Y - XB\|_2^2 + \sum_{g=1}^{|\mathcal{G}|}\lambda_g\|B_g\|_2 \\ &= \frac{1}{n}L^0(B) + \sum_{g\in\mathcal{G}^1}\lambda_g|\beta_g| + \sum_{g\in\mathcal{G}^2}\lambda_g\|B_g\|_2. \end{aligned}$$

For a coordinate β_{jk} in B, denote $\mathcal{G}^1_{jk} = \{g : \beta_{jk} \in B_g \in \mathcal{G}^1\}$ and $\mathcal{G}^2_{jk} = \{g : \beta_{jk} \in B_g \in \mathcal{G}^2\}$, then

$$\frac{\partial L(B)}{\partial \beta_{jk}} = -S_{jk}/n + \|x_j\|_2^2 \beta_{jk}/n + \sum_{\mathcal{G}^1_{jk}} \lambda_g \text{sign}(\beta_{jk}) + \sum_{\mathcal{G}^2_{jk}} \lambda_g \beta_{jk}/\|B_g\|_2.$$

If $\beta_{jk} > 0$, we have

$$\frac{\partial L(B)}{\partial \beta_{jk}} = -S_{jk}/n + \|x_j\|_2^2 \beta_{jk}/n + \sum_{\mathcal{G}^1_{jk}} \lambda_g + \sum_{\mathcal{G}^2_{jk}} \lambda_g \beta_{jk}/\|B_g\|_2.$$

Notice that $\partial L(B)/\partial \beta_{jk} \geq 0$ if and only if

$$\beta_{jk} \geq \frac{S_{jk} - n \sum_{\mathcal{G}^1_{jk}} \lambda_g}{\|x_j\|_2^2 + n \sum_{\mathcal{G}^2_{jk}} \lambda_g / \|B_g\|_2} \triangleq \tilde{\beta}^+_{jk},$$

and $\partial L(B)/\partial \beta_{jk} < 0$ if and only if $\beta_{jk} < \tilde{\beta}^+_{jk}$. So fixing all other coordinates of B, if $\beta_{jk} > 0$, then $L(B)$ is monotone increasing with respect to β_{jk} when $\beta_{jk} > \tilde{\beta}^+_{jk}$ and decreasing when $\beta_{jk} < \tilde{\beta}^+_{jk}$. Therefore, if $\hat{\beta}^+_{jk,min}$ minimizes $L(B)$ with respect to β_{jk} when $\beta_{jk} > 0$, then

$$\hat{\beta}^+_{jk,min} = \begin{cases} \frac{S_{jk} - n \sum_{\mathcal{G}^1_{jk}} \lambda_g}{\|x_j\|_2^2 + n \sum_{\mathcal{G}^2_{jk}} \lambda_g / \|\hat{B}_g\|_2}, & \text{if } S_{jk} > n \sum_{\mathcal{G}^1_{jk}} \lambda_g \\ 0, & \text{otherwise.} \end{cases} \tag{A.1}$$

Similarly, if $\beta_{jk} \leq 0$, $\partial L(B)/\partial \beta_{jk} \geq 0$ if and only if

$$\beta_{jk} \geq \frac{S_{jk} + n \sum_{\mathcal{G}^1_{jk}} \lambda_g}{\|x_j\|_2^2 + n \sum_{\mathcal{G}^2_{jk}} \lambda_g / \|B_g\|_2} \triangleq \tilde{\beta}^-_{jk},$$

and $\partial L(B)/\partial \beta_{jk} < 0$ if and only if $\beta_{jk} < \tilde{\beta}^-_{jk}$. So fixing all other coordinates of B, if $\hat{\beta}^-_{jk,min}$ minimizes $L(B)$ with respect to β_{jk} when $\beta_{jk} \leq 0$, then

$$\hat{\beta}^-_{jk,min} = \begin{cases} \frac{S_{jk} + n \sum_{\mathcal{G}^1_{jk}} \lambda_g}{\|x_j\|_2^2 + n \sum_{\mathcal{G}^2_{jk}} \lambda_g / \|\hat{B}_g\|_2}, & \text{if } S_{jk} < -n \sum_{\mathcal{G}^1_{jk}} \lambda_g \\ 0, & \text{otherwise.} \end{cases} \tag{A.2}$$

Let $\hat{\beta}_{jk}$ be the minimizer of $L(B)$ with respect to β_{jk} with all other coordinates fixed at $\hat{B}$. Then from (A.1) and (A.2), we have

$$\hat{\beta}_{jk} = \frac{\mathrm{sgn}(S_{jk})\left(|S_{jk}| - n\sum_{\mathcal{G}^1_{jk}} \lambda_g\right)_+}{\|x_j\|_2^2 + n\sum_{\mathcal{G}^2_{jk}} \lambda_g / \|\hat{B}_g\|_2}.$$

□

Proof of Theorem 3.4

Lemma A.2. *Under the assumptions in Theorem 2, for any $B \in \mathbb{R}^{p\times q}$, with probability at least $1-(pq)^{1-A^2/2}$,*

$$\begin{aligned} &\frac{1}{n}\|X(B^*-\hat{B})\|_2^2 + \lambda|\hat{B}-B|_1 + 2\sum_{g\in\mathcal{G}^2}\lambda_g\|\hat{B}_g - B_g\|_2 \qquad (A.3)\\ &\quad\le \frac{1}{n}\|X(B^*-B)\|_2^2 + 4\lambda\sum_{jk\in J_1(B)}|\hat{\beta}_{jk}-\beta_{jk}| + 4\sum_{g\in J_2(B)}\lambda_g\|\hat{B}_g - B_g\|_2,\\ &M(\hat{B}) \le \frac{4}{\lambda^2 n^2}\sum_{jk\in J_1(\hat{B})}|[X^{\mathrm{T}}X(\hat{B}-B^*)]_{jk}|^2 \le \frac{4}{\lambda^2 n^2}\|X^{\mathrm{T}}X(\hat{B}-B^*)\|_2^2. \qquad (A.4)\end{aligned}$$

Proof of Lemma A.2.

For any $B \in \mathbb{R}^{p\times q}$, we have

$$\frac{1}{n}\|Y - X\hat{B}\|_2^2 + 2\lambda|\hat{B}|_1 + \sum_{g\in\mathcal{G}^2}2\lambda_g\|\hat{B}_g\|_2 \le \frac{1}{n}\|Y-XB\|_2^2 + 2\lambda|B|_1 + \sum_{g\in\mathcal{G}^2}2\lambda_g\|B_g\|_2.$$

Plugging $Y = XB^* + W$ into the above inequality, we obtain

$$\begin{aligned}\frac{1}{n}\|X(B^*-\hat{B})\|_2^2 \;\le\; &\frac{1}{n}\|X(B^*-B)\|_2^2 + \frac{2}{n}\sum_{i=1}^{n}\sum_{k=1}^{q}[X(\hat{B}-B)]_{ik}\omega_{ik}\\ &+2\lambda(|B|_1 - |\hat{B}|_1) + \sum_{g\in\mathcal{G}^2}2\lambda_g(\|B_g\|_2 - \|\hat{B}_g\|_2),\end{aligned}$$

where $[X(\hat{B}-B)]_{ik}$ denotes the ik^{th} element of the product matrix $X(\hat{B}-B)$ and

ω_{ik} is the ik^{th} element of W. Notice that

$$\sum_{i=1}^{n}\sum_{k=1}^{q}[X(\hat{B}-B)]_{ik}\omega_{ik} = \sum_{i=1}^{n}\left\{\sum_{k=1}^{q}\left[\sum_{j=1}^{p}x_{ij}(\hat{\beta}_{jk}-\beta_{jk})\right]\omega_{ik}\right\}$$
$$\leq \max_{1\leq k\leq q, 1\leq j\leq p}\left|\sum_{i=1}^{n}x_{ij}\omega_{ik}\right|\sum_{k=1}^{q}\sum_{j=1}^{p}|\hat{\beta}_{jk}-\beta_{jk}| = |X^{\mathrm{T}}W|_{\infty}|\hat{B}-B|_{1}$$

where $|X^{\mathrm{T}}W|_{\infty} = \max_{1\leq k\leq q,1\leq j\leq p}|\sum_{i=1}^{n}x_{ij}\omega_{ik}|$ is the maximum absolute value of entries of $X^{\mathrm{T}}W$.

Let $V_{jk} = x_j^{\mathrm{T}}\cdot w_k$, $1\leq j\leq p$, $1\leq k\leq q$. Since $w_k\sim N(0,\sigma_k^2 I_q)$ for $1\leq q\leq Q$, then $\mathrm{var}(V_{jk}) = x_p^{\mathrm{T}}\mathrm{cov}(w_q)x_p = n\sigma_q^2$. Therefore $(n\sigma_q^2)^{-1/2}V_{jk}$ are standard normal random variables. Consider the random event

$$\mathcal{A} = \left\{\frac{2}{n}|X^{\mathrm{T}}W|_{\infty}\leq\lambda\right\}.$$

It is easy to see that the complement of $\mathcal{A}$ can be expressed as

$$\mathcal{A}^c = \left\{\text{At least one } |V_{jk}| > \frac{\lambda n}{2}, 1\leq j\leq p, 1\leq k\leq q\right\}.$$

Denote $B(0,\lambda n/2)$ to be a 1-dimensional ball centered at 0 and with radius $\lambda n/2$, then

$$\begin{aligned}
Pr\{\mathcal{A}^c\} &\leq \sum_{j=1}^{p}\sum_{k=1}^{q}Pr\left\{V_{jk}\notin B\left(0,\frac{\lambda n}{2}\right)\right\}\\
&= p\sum_{k=1}^{q}Pr\left\{(n\sigma_k^2)^{-1/2}V_{jk}\notin B\left(0,\frac{\lambda n^{1/2}}{2\sigma_k}\right)\right\}\\
&\leq pq\times Pr\left\{|Z|\geq\frac{\lambda n^{1/2}}{2\sigma}\right\}\\
&\leq pq\exp\left(\frac{-\lambda^2 n}{8\sigma^2}\right)\\
&= (pq)^{1-A^2/2},
\end{aligned}$$

where Z is a standard noraml random variable, and the last inequality is obtained

by $Pr\{|Z| > a\} \leq \exp(-a^2/2)$. But on event $\mathcal{A}$, we have

$$\begin{aligned}
&\frac{1}{n}\|X(B^* - \hat{B})\|_2^2 + \lambda|\hat{B} - B|_1 + 2\sum_{g\in\mathcal{G}^2}\lambda_g\|\hat{B}_g - B_g\|_2 \\
&\quad\leq \frac{1}{n}\|X(B^* - B)\|_2^2 + 2\lambda(|\hat{B} - B|_1 + |B|_1 - |\hat{B}|_1) \\
&\qquad +2\sum_{g\in\mathcal{G}^2}\lambda_g(\|\hat{B}_g - B_g\|_2 + \|B_g\|_2 - \|\hat{B}_g\|_2) \\
&\quad\leq \frac{1}{n}\|X(B^* - B)\|_2^2 + 4\lambda\sum_{jk\in J_1(B)}|\hat{\beta}_{jk} - \beta_{jk}| + 4\sum_{g\in J_2(B)}\lambda_g(\|\hat{B}_g - B_g\|_2.
\end{aligned}$$

This completes the proof of the first inequality in Lemma A.2.

To prove the second inequality, we use the KKT conditions and obtain

$$\begin{cases}
(1/n)[X^{\mathrm{T}}(Y - X\hat{B})]_{jk} = 2\lambda\mathrm{sgn}(\hat{\beta}_{jk}) + 2\sum_{g\in\mathcal{G}^2}\lambda_g\hat{\beta}_{jk}/\|\hat{B}_g\|_2, & \hat{\beta}_{jk} \neq 0; \\
(1/n)|[X^{\mathrm{T}}(Y - X\hat{B})]_{jk}| \leq 2\lambda + 2\sum_{g\in\mathcal{G}^2}\lambda_g, & \hat{\beta}_{jk} = 0.
\end{cases}$$

From the first condition we can see that $\forall\hat{\beta}_{jk} \neq 0$,

$$\lambda \leq \frac{1}{n}|[X^{\mathrm{T}}(Y - X\hat{B})]_{jk}|.$$

On the other hand, we have on $\mathcal{A}$

$$\begin{aligned}
\frac{1}{n}|[X^{\mathrm{T}}(Y - X\hat{B})]_{jk}| &\leq \frac{1}{n}|[X^{\mathrm{T}}X(B^* - \hat{B})]_{jk} + [X^{\mathrm{T}}W]_{jk}| \\
&\leq \frac{1}{n}|[X^{\mathrm{T}}X(B^* - \hat{B})]_{jk}| + \frac{1}{n}|X^{\mathrm{T}}W|_\infty \\
&\leq \frac{1}{n}|[X^{\mathrm{T}}X(B^* - \hat{B})]_{jk}| + \frac{\lambda}{2}.
\end{aligned}$$

Then combine the above two inequalities, we have

$$\frac{\lambda}{2} \leq \frac{1}{n}|[X^{\mathrm{T}}X(\hat{B} - B^*)]_{jk}|.$$

Therefore

$$M(\hat{B}) = |J_1(\hat{B})| \leq \frac{4}{\lambda^2 n^2} \sum_{jk \in J_1(\hat{B})} |[X^{\mathrm{T}} X(\hat{B} - B^*)]_{jk}|^2 \leq \frac{4}{\lambda^2 n^2} \|X^{\mathrm{T}} X(\hat{B} - B^*)\|_2^2.$$

This completes the proof of Lemma A.2. □

Proof of Theorem 3.4.

By setting $B = B^*$ in (A.3) in Lemma A.2, we have that on event $\mathcal{A}$,

$$\begin{aligned} \frac{1}{n}\|X(\hat{B} - B^*)\|_2^2 &\leq 4\lambda \sum_{jk \in J_1(B^*)} |\hat{\beta}_{jk} - \beta^*_{jk}| + 4 \sum_{g \in J_2(B^*)} \lambda_g \|\hat{B}_g - B^*_g\|_2 \quad \text{(A.5)} \\ &\leq 4\lambda r^{1/2} \|(\hat{B} - B^*)_{J_1(B^*)}\|_2 \\ &\quad + 4 \left(\sum_{g \in J_2(B^*)} \lambda_g^2 \right)^{1/2} \|(\hat{B} - B^*)_{J_2(B^*)}\|_2. \end{aligned}$$

The last inequality is by Cauchy-Schwarz. Specifically, we have

$$\begin{aligned} \left(\sum_{jk \in J_1(B^*)} |\hat{\beta}_{jk} - \beta^*_{jk}| \right)^2 &= \left(\sum_{jk \in J_1(B^*)} 1 \times |\hat{\beta}_{jk} - \beta^*_{jk}| \right)^2 \\ &\leq \left(\sum_{jk \in J_1(B^*)} 1^2 \right) \left(\sum_{jk \in J_1(B^*)} |\hat{\beta}_{jk} - \beta^*_{jk}|^2 \right) \\ &= r \|(\hat{B} - B^*)_{J_1(B^*)}\|_2^2, \end{aligned}$$

and

$$\left(\sum_{g \in J_2(B^*)} \lambda_g \|\hat{B}_g - B^*_g\|_2 \right)^2 \leq \left(\sum_{g \in J_2(B^*)} \lambda_g^2 \right) \left(\sum_{g \in J_2(B^*)} \|\hat{B}_g - B^*_g\|_2^2 \right).$$

Also by inequality (A.3), on event $\mathcal{A}$, we have

$$\begin{aligned} &\lambda |\hat{B} - B^*|_1 + 2 \sum_{g \in \mathcal{G}^2} \lambda_g \|\hat{B}_g - B^*_g\|_2 \quad \text{(A.6)} \\ &\leq 4\lambda \sum_{jk \in J_1(B^*)} |\hat{\beta}_{jk} - \beta^*_{jk}| + 4 \sum_{g \in J_2(B^*)} \lambda_g \|\hat{B}_g - B^*_g\|_2. \end{aligned}$$

This is equivalent to

$$\lambda \sum_{jk \in J_1^c(B^*)} |\hat{\beta}_{jk} - \beta^*_{jk}| + 2 \sum_{g \in J_2^c(B^*)} \lambda_g \|\hat{B}_g - B^*_g\|_2$$
$$\leq 3\lambda \sum_{jk \in J_1(B^*)} |\hat{\beta}_{jk} - \beta^*_{jk}| + 2 \sum_{g \in J_2(B^*)} \lambda_g \|\hat{B}_g - B^*_g\|_2.$$

Thus the condition in Assumption 1 holds with $\Delta = \hat{B} - B^*$ and $\rho_g = \lambda_g/\lambda$. Therefore,

$$\|(\hat{B} - B^*)_{J_1(B^*)}\|_2 \leq \frac{\|X(\hat{B} - B^*)\|_2}{\kappa_1 n^{1/2}}, \quad \|(\hat{B} - B^*)_{J_2(B^*)}\|_2 \leq \frac{\|X(\hat{B} - B^*)\|_2}{\kappa_2 n^{1/2}}.$$

Plugging the above two inequalities into (A.5), we have

$$\begin{aligned} \frac{1}{n}\|X(\hat{B} - B^*)\|_2^2 &\leq \left(\frac{4\lambda r^{1/2}}{\kappa_1 n^{1/2}} + \frac{4\left(\sum_{g \in J_2(B^*)} \lambda_g^2\right)^{1/2}}{\kappa_2 n^{1/2}} \right) \|X(\hat{B} - B^*)\|_2 \\ &= \left(\frac{4\lambda r^{1/2}}{\kappa_1 n^{1/2}} + \frac{4\lambda\left(\sum_{g \in J_2(B^*)} \rho_g^2\right)^{1/2}}{\kappa_2 \sqrt{n}} \right) \|X(\hat{B} - B^*)\|_2, \end{aligned}$$

which gives

$$\frac{1}{n}\|X(\hat{B} - B^*)\|_2^2 \leq 16\lambda^2 \left(\frac{r^{1/2}}{\kappa_1} + \frac{\left(\sum_{g \in J_2(B^*)} \rho_g^2\right)^{1/2}}{\kappa_2} \right)^2.$$

Define $\|A\|_{2,1} = \sum_{g \in \mathcal{G}^2 \cup \mathcal{G}^1} \|A\|_2$, where each coefficient in $\mathcal{G}^1 = \mathcal{G}_L$ forms a group. Hence

$$\|\hat{B} - B^*\|_{2,1} = |\hat{B} - B^*|_1 + \sum_{g \in \mathcal{G}^2} \|\hat{B}_g - B^*_g\|_2 \tag{A.7}$$
$$\leq (c+1)|\hat{B} - B^*|_1. \tag{A.8}$$

Then we have

$$
\begin{aligned}
(\lambda+\rho\lambda)\|\hat{B}-B^*\|_{2,1} &= \lambda\|\hat{B}-B^*\|_{2,1}+\rho\lambda\|\hat{B}-B^*\|_{2,1} \\
&\le (c+1)\lambda|\hat{B}-B^*|_1+\rho\lambda\|\hat{B}-B^*\|_{2,1} \qquad \text{by (A.8)} \\
&= (c+1)\lambda|\hat{B}-B^*|_1+\rho\lambda|\hat{B}-B^*|_1+\sum_{g\in\mathcal{G}^2}\rho\lambda\|\hat{B}_g-B^*_g\|_2 \\
&\le (c+2)\lambda|\hat{B}-B^*|_1+\sum_{g\in\mathcal{G}^2}\lambda_g\|\hat{B}_g-B^*_g\|_2 \\
&\le (c+2)\lambda|\hat{B}-B^*|_1+2\sum_{g\in\mathcal{G}^2}\lambda_g\|\hat{B}_g-B^*_g\|_2 \\
&\le (c+2)\left[\lambda|\hat{B}-B^*|_1+2\sum_{g\in\mathcal{G}^2}\lambda_g\|\hat{B}_g-B^*_g\|_2\right].
\end{aligned}
$$

By (A.6) and the last inequality in (A.5) we obtain

$$
\begin{aligned}
&\frac{1+\rho}{c+2}\lambda\|\hat{B}-B^*\|_{2,1} \\
&\quad\le \lambda|\hat{B}-B^*|_1+2\sum_{g\in\mathcal{G}^2}\lambda_g\|\hat{B}_g-B^*_g\|_2 \\
&\quad\le 4\lambda\sum_{jk\in J_1(B^*)}|\hat{\beta}_{jk}-\beta^*_{jk}|+4\sum_{g\in J_2(B^*)}\lambda_g\|\hat{B}_g-B^*_g\|_2 \\
&\quad\le 4\lambda r^{1/2}\|(\hat{B}-B^*)_{J_1(B^*)}\|_2+4\left(\sum_{g\in J_2(B^*)}\lambda_g^2\right)^{1/2}\|(\hat{B}-B^*)_{J_2(B^*)}\|_2 \\
&\quad\le \left(\frac{4\lambda r^{1/2}}{\kappa_1 n^{1/2}}+\frac{4\left(\sum_{g\in J_2(B^*)}\lambda_g^2\right)^{1/2}}{\kappa_2 n^{1/2}}\right)\|X(\hat{B}-B^*)\|_2 \\
&\quad\le \left(\frac{4\lambda r^{1/2}}{\kappa_1 n^{1/2}}+\frac{4\lambda\left(\sum_{g\in J_2(B^*)}\rho_g^2\right)^{1/2}}{\kappa_2 n^{1/2}}\right)4n^{1/2}\lambda\left(\frac{r^{1/2}}{\kappa_1}+\frac{\left(\sum_{g\in J_2(B^*)}\rho_g^2\right)^{1/2}}{\kappa_2}\right) \\
&\quad= 16\lambda^2\left(\frac{r^{1/2}}{\kappa_1}+\frac{\left(\sum_{g\in J_2(B^*)}\rho_g^2\right)^{1/2}}{\kappa_2}\right)^2.
\end{aligned}
$$

Therefore,

$$\begin{aligned}\|\hat{B} - B^*\|_{2,1} &\leq \frac{16(c+2)\lambda}{1+\rho}\left(\frac{r^{1/2}}{\kappa_1} + \frac{\left(\sum_{g\in J_2(B^*)}\rho_g^2\right)^{1/2}}{\kappa_2}\right)^2 \\ &= \frac{32(c+2)\sigma A}{1+\rho}\left(\frac{\log(pq)}{n}\right)^{1/2}\left(\frac{r^{1/2}}{\kappa_1} + \frac{\left(\sum_{g\in J_2(B^*)}\rho_g^2\right)^{1/2}}{\kappa_2}\right)^2.\end{aligned}$$

It is trivial that $|\hat{B} - B^*|_1 \leq \|\hat{B} - B^*\|_{2,1}$.

From (A.4) in Lemma A.2, we obtain

$$M(\hat{B}) \leq \frac{4}{\lambda^2 n^2}\|X^{\mathrm{T}}X(\hat{B} - B^*)\|_2^2 \leq \frac{4\psi_{\max}}{\lambda^2 n}\|X(\hat{B} - B^*)\|_2^2,$$

where the second inequality is from

$$\begin{aligned}\|[X^{\mathrm{T}}X(\hat{B} - B^*)]_{\cdot k}\|_2^2 &= (\hat{B} - B^*)_{\cdot k}^{\mathrm{T}}X^{\mathrm{T}}(XX^{\mathrm{T}})X(\hat{B} - B^*)_{\cdot k} \\ &\leq n\psi_{\max}\|X(\hat{B} - B^*)_{\cdot k}\|_2^2\end{aligned}$$

for each $1 \leq k \leq q$. By the upper bound of $\|X(\hat{B} - B^*)\|_2^2$ we have

$$M(\hat{B}) \leq 64\psi_{\max}\left(\frac{r^{1/2}}{\kappa_1} + \frac{\left(\sum_{g\in J_2(B^*)}\rho_g^2\right)^{1/2}}{\kappa_2}\right)^2.$$

□

Proof of Proposition 4.1

To prove Proposition 4.1, we first show the following lemma.

Lemma A.3. *A sequence of coordinate estimates by iteratively solving the exact*

solution of

$$\hat{\beta}_{jk} = \frac{\text{sgn}(S_{jk})\left(|S_{jk}| - n\sum_{\{g\in\mathcal{G}^1:\ \beta_{jk}\in B_g\}}\lambda_g\right)_+}{\|x_j\|_2^2 + n\sum_{\{g\in\mathcal{G}^2:\ \beta_{jk}\in B_g\}}\lambda_g/(\|\hat{B}_{g-(jk)}\|_2^2 + |\hat{\beta}_{jk}|^2)^{1/2}}, \tag{T9}$$

converge to a minimum point of the objective function.

First, it is easy to see that the exact solution of (T9) exists. If $\|\hat{B}_{g-(jk)}\|_2 = 0$, the close form solution of (T9) is just the lasso solution. If $\|\hat{B}_{g-(jk)}\|_2 \neq 0$, then the right hand side of (T9) is a continuous function of $\hat{\beta}_{jk}$, which is monotonic when $\hat{\beta}_{jk} > 0$ or $\hat{\beta}_{jk} < 0$, bounded away from zero when $\hat{\beta}_{jk} = 0$, and bounded away from $\pm\infty$ when $\hat{\beta}_{jk}$ goes to $\pm\infty$, therefore must intersect with either $y = \hat{\beta}_{jk}$ or $y = -\hat{\beta}_{jk}$. Therefore an exact solution of (T9) must exist.

Wu and Lange (2008) proved the convergence to a minimal point of the lasso objective function for the greedy coordinate descent algorithm. In a similar way, we can extend the proof to our sparse group lasso objective function. For the completeness of the context, we elaborate the proof of Lemma A.3 as following.

Proof of Lemma A.3.

From proof of Theorem 3.1, one can see that the algorithm the multivariate sparse group lasso objective function decreases at each iteration step of solving (T9). Since the objective function is convex and bounded from below, so a global minimum exists.

Next, we show that the mixed coordinate descent algorithm converges to a stationary point. A stationary point is a point such that any directional derivative of the objective function at this point is nonnegative. Based on convexity, a stationary point is also a global minimizer.

Define

$$h(B) = \min_{jk}\min\left\{\frac{\partial L(B)}{\partial\beta_{jk}}, \frac{\partial L(B)}{\partial(-\beta_{jk})}\right\}$$

as the minimum directional derivative magnitude along each coordinate direction and the opposite direction. Similar to Wu and Lange (2008), one can show that h is upper semi-continuous, which by definition means that $\limsup_{m\to\infty} h(B_m) \leq h(B)$

given that B_m converges to B. In fact, $\|Y-XB\|_2^2/(2n)$ and $\sum_{\{g\in\mathcal{G}^2:\|B_g\|_2\neq 0\}}\lambda_g\|B_g\|_2$ are differentiable parts of $L(B)$ and hence having continuous directional derivatives. The non-differentiable parts $\sum_{\{g\in\mathcal{G}^1\}}\lambda_g|\beta|$ and $\sum_{\{g\in\mathcal{G}^2:\|B_g\|_2=0\}}\lambda_g\|B_g\|_2$ both have directional derivative of each of their summands w.r.t. β_{jk} equals to $\text{sgn}(\beta_{jk})\lambda_g$ and w.r.t. $-\beta_{jk}$ equals to $-\text{sgn}(\beta_{jk})\lambda_g$, hence both have directional derivatives as finite sum of constants. Therefore, $h(B)$ is upper semi-continuous since upper semi-continuous functions includes continuous functions and is closed under operations of finite sum and minima.

Now suppose that there is a sequence of $B^{(m)}$ generated by iteratively solving (T9) has a subsequence $B^{(m_l)}$ that converges to a none-stationary point B^*, then

$$h(B^{(m_l)}) \leq h(B^*) < 0.$$

Considering the current updating coordinate β_{jk}. Without causing confusion, denote $L(\beta)$ as a function of jk'th coordinate with all other coordinates' value fixed. Let $\beta_{jk}^{(m_l)}$ be the solution to (T9) from m'th iteration and β_{jk}^* be the jk'th coordinate of B^*. Using Taylor expansion we have

$$L(\beta) = L(\beta_{jk}^{(m_l)}) + L^{'}(\beta_{jk}^{(m_l)})(\beta-\beta_{jk}^{(m_l)}) + \frac{1}{2}L^{''}(\tilde{\beta})(\beta-\beta_{jk}^{(m_l)})^2 \qquad \text{(A.9)}$$

for some $\tilde{\beta}$ lying between β and $\beta_{jk}^{(m_l)}$. Notice that

$$L^{''}(\beta) = \frac{\|x_j\|_2^2}{n} + \sum_{\{g\in\mathcal{G}^2:\|B_g\|_2\neq 0\}} \lambda_g\omega_g^{1/2}\left(\frac{1-\beta_{jk}^*}{\|B_g^*\|_2}\right)$$

is a finite sum of finite numbers and hence is bounded away from $\pm\infty$. Assume it pertains an upper bound c, then (A.9)

$$\leq L(\beta_{jk}^{(m_l)}) + L^{'}(\beta_{jk}^{(m_l)})(\beta-\beta_{jk}^{(m_l)}) + \frac{c}{2}(\beta-\beta_{jk}^{(m_l)})^2.$$

The right hand side of the above equation is a quadratic majorization function of the objective function $L(\beta)$. The majorization function is minimized at point

$\omega = \beta_{jk}^{(m_l)} - L'(\beta_{jk}^{(m_l)})/c$ with a minimum value $L(\beta_{jk}^{(m_l)}) - L'(\beta_{jk}^{(m_l)})^2/(2c)$.

Then the minimizer of $L(\beta)$ at the m'th iteration, also the solution to (9), $\beta_{jk}^{(m_l+1)}$ satisfies

$$L(\beta_{jk}^{(m_l+1)}) \leq L(\omega) = L(\beta_{jk}^{(m_l)}) - \frac{L'(\beta_{jk}^{(m_l)})^2}{2c} \leq L(\beta_{jk}^{(m_l)}) - \frac{h(B^*)^2}{2c}.$$

And the above inequality will be satisfied by all subsequent m_l, therefore will force $L(\beta_{jk}^{(m_l)})$ going to $-\infty$, which contradicts the fact that $L(\beta)$ is bounded from below by zero. Therefore, any sequences $B^{(m)}$ generated by iteratively solving (T9) converge to a stationary point and also a minimizer. □

Proof of Proposition 4.1.

Denote $\hat{\beta}_{jk}^{(m)}$ a sequence of estimates of jk'th coordinate by iteratively solving (T9). Starting from $\hat{B}^{(m-1)}$, denote $\hat{\beta}_{jk}^{\mathrm{MCD}(m-1)}$ the next step update of the jk'th coordinate by the mixed coordinate descent algorithm. We prove in the following that

$$|\hat{\beta}_{jk}^{(m)}| \leq |\hat{\beta}_{jk}^{\mathrm{MCD}(m)}| \leq |\hat{\beta}_{jk}^{(m-1)}| \tag{A1}$$

with equalities hold only when $|\hat{\beta}_{jk}^{(m)}| = |\hat{\beta}_{jk}^{(m-1)}|$.

(i) If

$$\hat{\beta}_{jk}^{(m-1)} < \hat{\beta}_{jk}^{(m)} = \frac{-\left(|S_{jk}| - n\sum_{\{g\in\mathcal{G}^1:\ \beta_{jk}\in B_g\}}\lambda_g\right)_+}{\|x_j\|_2^2 + n\sum_{\{g\in\mathcal{G}^2:\ \beta_{jk}\in B_g\}}\lambda_g/(\|\hat{B}_{g-(jk)}^{(m-1)}\|_2^2 + |\hat{\beta}_{jk}^{(m)}|^2)^{1/2}} < 0,$$

then

$$\hat{\beta}_{jk}^{\mathrm{MCD}(m)} = \frac{-\left(|S_{jk}| - n\sum_{\{g\in\mathcal{G}^1:\ \beta_{jk}\in B_g\}}\lambda_g\right)_+}{\|x_j\|_2^2 + n\sum_{\{g\in\mathcal{G}^2:\ \beta_{jk}\in B_g\}}\lambda_g/(\|\hat{B}_{g-(jk)}^{(m-1)}\|_2^2 + |\hat{\beta}_{jk}^{(m-1)}|^2)^{1/2}} < \hat{\beta}_{jk}^{(m)}.$$

From the proof of Theorem 3.1, $\hat{\beta}_{jk}^{(m-1)} < \hat{\beta}_{jk}^{(m)}$ if and only if

$$\left.\frac{\partial L(B)}{\partial \beta_{jk}}\right|_{\hat{\beta}_{jk}^{(m-1)}} = -S_{jk}/n + \|x_j\|_2^2\hat{\beta}_{jk}^{(m-1)}/n - \sum_{\mathcal{G}_{jk}^1}\lambda_g + \sum_{\mathcal{G}_{jk}^2}\lambda_g\hat{\beta}_{jk}^{(m-1)}/\|\hat{B}_g^{(m-1)}\|_2 < 0.$$

Notice that above is also partial derivative of $L^{\text{net}}(B)$ w.r.t. β_{jk} taking value at $\hat{\beta}_{jk}^{(m-1)}$, with $L^{\text{net}}(B)$ the elastic net objective function

$$L^{\text{net}}(B) = \frac{1}{2n}\|Y - XB\|_2^2 + \sum_{g\in\mathcal{G}^1} \lambda_g |B_g|_1 + \sum_{g\in\mathcal{G}^2} \lambda_g \|B_g\|_2^2/(2\|\hat{B}_g^{(m-1)}\|_2)$$

holding $\|\hat{B}_g^{(m-1)}\|_2$ as constants and constraining that $\beta_{jk} < 0$.

Following exactly the same argument as in the proof of Theorem 3.1, we can prove that $\frac{\partial L^{\text{net}}(B)}{\partial \beta_{jk}}\Big|_{\hat{\beta}_{jk}^{(m-1)}} < 0$ if and only if $\hat{\beta}_{jk}^{(m-1)}$ is less than the solution of $\partial L(B)/\partial\beta_{jk} = 0$ with constrain $\beta_{jk} < 0$. And that is the solution of

$$-S_{jk}/n + \|x_j\|_2^2\beta_{jk}/n - \sum_{\mathcal{G}_{jk}^1} \lambda_g + \sum_{\mathcal{G}_{jk}^2} \lambda_g \beta_{jk}/\|\hat{B}_g^{(m-1)}\|_2 = 0$$

constraining on $\beta_{jk} < 0$, which, using the same calculation as in the proof of Theroem 3.1, is

$$\frac{-\left(|S_{jk}| - n\sum_{\{g\in\mathcal{G}^1:\ \beta_{jk}\in B_g\}} \lambda_g\right)_+}{\|x_j\|_2^2 + n\sum_{\{g\in\mathcal{G}^2:\ \beta_{jk}\in B_g\}} \lambda_g/\|\hat{B}^{(m-1)}\|^2} = \hat{\beta}_{jk}^{\text{MCD}(m)}.$$

Therefore, we have

$$\hat{\beta}_{jk}^{(m-1)} < \hat{\beta}_{jk}^{\text{MCD}(m)} < \hat{\beta}_{jk}^{(m)} < 0.$$

(ii) If

$$\hat{\beta}_{jk}^{(m-1)} > \hat{\beta}_{jk}^{(m)} = \frac{\left(|S_{jk}| - n\sum_{\{g\in\mathcal{G}^1:\ \beta_{jk}\in B_g\}} \lambda_g\right)_+}{\|x_j\|_2^2 + n\sum_{\{g\in\mathcal{G}^2:\ \beta_{jk}\in B_g\}} \lambda_g/(\|\hat{B}_{g-(jk)}^{(m-1)}\|_2^2 + |\hat{\beta}_{jk}^{(m)}|^2)^{1/2}} \geq 0,$$

with similar argument, we have that

$$\hat{\beta}_{jk}^{(m-1)} > \hat{\beta}_{jk}^{\text{MCD}(m)} > \hat{\beta}_{jk}^{(m)} \geq 0.$$

(iii) If $\hat{\beta}_{jk}^{(m-1)} = \hat{\beta}_{jk}^{(m)}$, the mixed coordinate descent algorithm will be exact update and we will have

$$\hat{\beta}_{jk}^{(m-1)} = \hat{\beta}_{jk}^{\text{MCD}(m)} = \hat{\beta}_{jk}^{(m)}.$$

In summary, we have (A1).

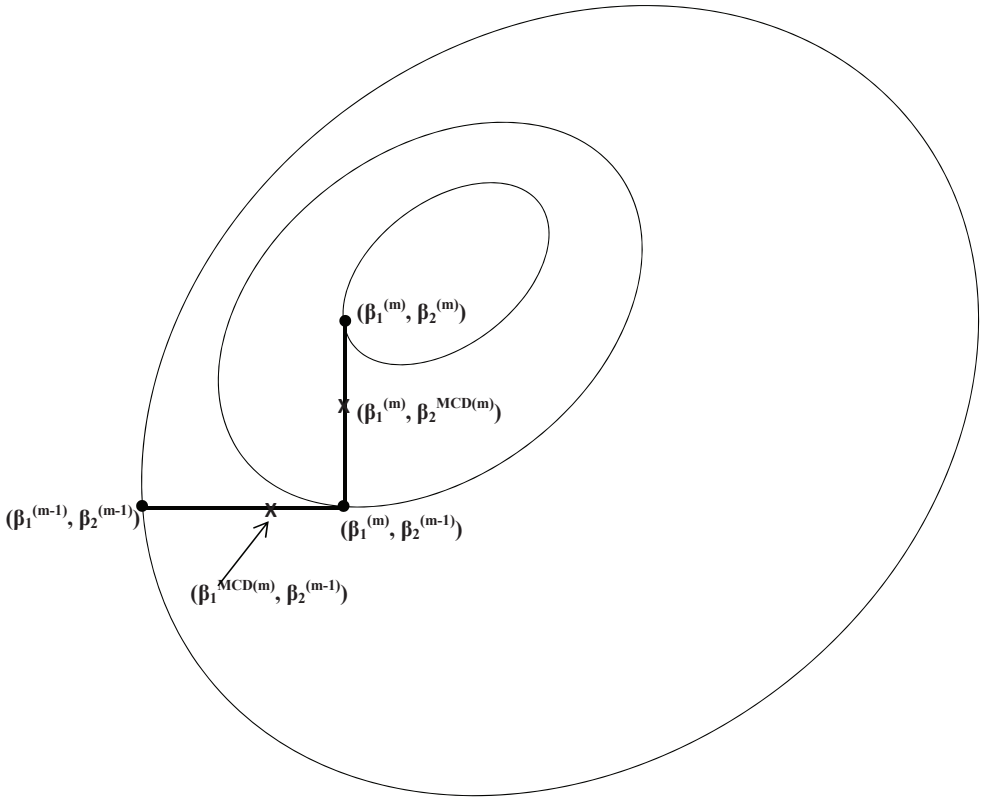


Figure A.1: Illustration of coordinate updates by the cyclical coordinate descent and the mixed coordinate descent algorithms on a contour surface of a two-dimensional objective function.

Lemma A.3 shows that the sequence of estimates of jk'th coordinate $\{\hat{\beta}_{jk}^{(m)}\}$ iteratively updated from solving (T9) converges to a minimizer regardless the value of the starting point. And since the objective function is convex, it is easy to show that the minimum point is unique. For each term in the sequence $\{\hat{\beta}_{jk}^{\mathrm{MCD}(l)}\}$, suppose one can construct a sequence of $\{\hat{\beta}_{jk}^{(m)}\}$ starting from $\hat{\beta}_{jk}^{\mathrm{MCD}(l)}$, then those sequences all converge to minimizers (if the minimizer is not unique, e.g. for not strictly convex objective function) giving the same minimum value. By (A1), those sequences will gauge $\{\hat{\beta}_{jk}^{\mathrm{MCD}(l)}\}$ converge to a minimizer that also gives the minimum point. Therefore $\{\hat{\beta}_{jk}^{\mathrm{MCD}(l)}\}$ converges to a global minimizer of the objective function.
□

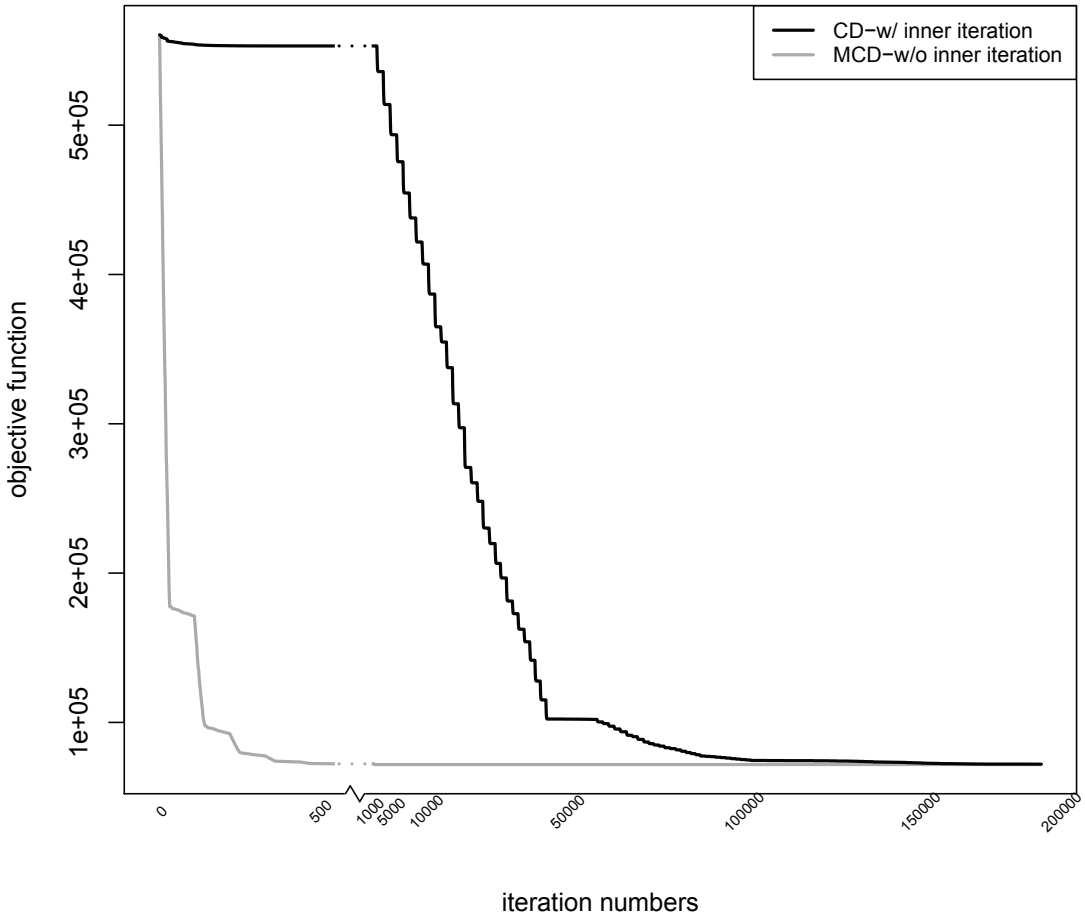


Figure A.2: Decreasing of the objective function. Dark gray line for using the mixed coordinate descent (MCD) algorithm with out inner iterations of updating (T9) and black line using the coordinate descent (CD) algorithm with inner iterations.

In Figure A.1, $\hat{\beta}_{jk}^{(m)}$'s are the exact solutions of (T9) at each step of iteration. Such exact solutions can be achieved by adding an inner loop of iterating

$$\hat{\beta}_{jk}^{(m_l)} = \frac{\mathrm{sgn}(S_{jk}^{(m-1)}) \left(|S_{jk}^{(m-1)}| - n \sum_{\{g \in \mathcal{G}^1:\ \beta_{jk} \in B_g\}} \lambda_g \right)_+}{\|x_j\|_2^2 + n \sum_{\{g \in \mathcal{G}^2:\ \beta_{jk} \in B_g\}} \lambda_g / (\|\hat{B}_{g-(jk)}^{(m-1)}\|_2^2 + |\hat{\beta}_{jk}^{(m_l-1)}|^2)^{1/2}}$$

till convergence and take the convergent point as $\hat{\beta}_{jk}^{(m)}$.

The computational cost of coordinate descent algorithm with inner iterations is much more than our mixed coordinate descent algorithm. Figure A.2 shows speed comparison between the coordinate descent algorithm with inner iterations and the mixed coordinate descent algorithm. The group structure of the regression coefficient matrix used is set to be (b) in Figure 1 in the main text. In Figure A.2, the mixed

coordinate descent algorithm converges to a minimizer after 500 iterations while the coordinate descent with inner iterations converges after 150000 iteration steps.

APPENDIX B

Appendix for Chapter IV

Proofs of Theorems

Proof of Theorem IV.1

The concavity of the logistic likelihood function $l_1(\gamma)$ has been established in the literature (Pratt, 1981; Meier et al., 2008).

To show that $l_2(\beta)$ is a concave function of β, it suffices to show that the $p \times p$ matrix $D = [\partial^2 l_2/\partial\beta_k^2\partial\beta_{k'}^2]_{1\geq k\leq p, 1\geq k'\leq p}$ is negative definite. From (4.11) It is easy to see that for any vector $v = (v_1, \cdots, v_p)' \in \mathbb{R}^p$,

$$\begin{aligned} v'Dv &= \sum_{i=1}^{n} \hat{y}_i\delta_i\hat{b}_{ir_i}\left(\sum_{k=1}^{p} v_k X_{ik}(t_{r_i-1})\right)^2 - \hat{y}_i(1-\delta_i)\hat{h}_{ir_i}\left(\sum_{k=1}^{p} v_k X_{ik}(t_{r_i-1})\right)^2 \\ &\quad - \sum_{j=1}^{r_i-1} \hat{y}_i\hat{h}_{ij}\left(\sum_{k=1}^{p} v_k X_{ik}(t_{j-1})\right)^2. \end{aligned}$$

Notice that $\hat{h}_{ij} \geq 0$, $\hat{y}_i \geq 0$ and $1-\delta_i \geq 0$, so the second and third term in each summand of $v'Dv$ are less than or equal to zero. To show that the first term is also less than or equal to zero, it suffices to show $\hat{b}_{ij} \leq 0$. Notice that $\hat{b}_{ij} = \frac{\hat{h}_{ij}\exp\{-\hat{h}_{ij}\}}{1-\exp\{-\hat{h}_{ij}\}}(1 - \frac{\hat{h}_{ij}}{1-\exp\{-\hat{h}_{ij}\}})$ and $\frac{\hat{h}_{ij}\exp\{-\hat{h}_{ij}\}}{1-\exp\{-\hat{h}_{ij}\}} \geq 0$, so we only need to show that $(1 - \hat{h}_{ij}/(1-\exp\{-\hat{h}_{ij}\})) \leq 0$. Or equivalently $f(h_{ij}) := 1 - \exp\{-h_{ij}\} - h_{ij} \leq 0$ for all $h_{ij} \geq 0$. The fact $f'(h_{ij}) \leq 0$ gives that $f(h_{ij})$ is non-increasing w.r.p. to h_{ij}

when $h_{ij} \geq 0$. Notice that $f(0) = 0$, therefore $f(h_{ij}) \leq 0$ when $h_{ij} \geq 0$. Therefore $v'Dv \leq 0$.

And since the conditional expectation of the complete likelihood function can be written as a separable sum of functions $l_1(\gamma)$ and $l_2(\beta)$, the last statement of the theorem follows. □

Proof of Theorem IV.2

$$\begin{aligned} -\frac{\partial^2 E(l_1)}{\partial \gamma_j^2} &= \sum_{i=1}^{n} \left[\frac{Z_{ij}^2 \exp\{\gamma' \mathbf{Z}_i\}}{1+\exp\{\gamma' \mathbf{Z}_i\}} - \left(\frac{Z_{ij} \exp\{\gamma' \mathbf{Z}_i\}}{1+\exp\{\gamma' \mathbf{Z}_i\}} \right)^2 \right] \\ &\leq n \left(\max_{1 \leq i \leq n} Z_{ij} - \min_{1 \leq i \leq n} Z_{ij} \right)^2 /4. \end{aligned} \tag{B.1}$$

The last inequality holds because that (B.1) can be treated as n times the variance of a discrete random variable Z with distribution $P(Z = Z_{ij}) = \exp\{\gamma' \mathbf{Z}_i\}/(1 + \exp\{\gamma' \mathbf{Z}_i\})$. The variance is maximized when Z has a two-point distribution on $\max_{1 \leq i \leq n} Z_{ij}$ and $\min_{1 \leq i \leq n} Z_{ij}$ (Yang and Zou, 2013). □

Derivation of the observed likelihood (4.3)

The observed likelihood (4.3) is obtained by integrating out the latent variable Y_i on each term in the complete likelihood (4.2) under the constrain that when $\delta_i = 1$, y_i has to be 1. Therefore when $\delta_i = 1$,

$$\begin{aligned} & l_C^{(i)}(\lambda, \beta, \gamma | \mathbf{Z}_i, \mathbf{X}_i(t), \Gamma_i, \delta_i = 1; y_i = 1) \\ &= \pi(\mathbf{Z}_i) \left\{ L_i(\theta; \Gamma_i = r_i, \Delta_i = 1, \mathbf{X}_i(t)) \right\} \\ &= \pi(\mathbf{Z}_i) \left(1 - \exp\{-\exp\{\lambda_{r_i} + \beta' X_i(t_{r_i - 1})\}\} \right) \exp\left\{ -\sum_{j=1}^{r_i - 1} \exp\{\lambda_j + \beta' X_i(t_{j-1})\} \right\}. \end{aligned}$$

When $\delta_i = 0$,

$$
\begin{aligned}
& l_C^{(i)}(\lambda, \beta, \gamma | \mathbf{Z}_i, \mathbf{X}_i(t), \Gamma_i, \delta_i = 0; y_i) \\
& = (1 - \pi(\mathbf{Z}_i))^{1-y_i} \, \pi(\mathbf{Z}_i)^{y_i} \, \{L_i(\theta; \Gamma_i = r_i, \Delta_i = 0, \mathbf{X}_i(t))\}^{y_i} \\
& = (1 - \pi(\mathbf{Z}_i))^{1-y_i} \, \pi(\mathbf{Z}_i)^{y_i} \, (\exp\{-\exp\{\lambda_{r_i} + \beta' X_i(t_{r_i - 1})\}\})^{y_i} \\
& \times \exp\left\{ -\sum_{j=1}^{r_i - 1} \exp\{\lambda_j + \beta' X_i(t_{j-1})\} \right\}^{y_i} .
\end{aligned}
$$

Then corresponding term in the observed likelihood is

$$
\begin{aligned}
& l_O^{(i)}(\lambda, \beta, \gamma | \mathbf{Z}_i, \mathbf{X}_i(t), \Gamma_i, \delta_i = 0) \\
& \quad = \sum_{y_i} l_C^{(i)}(\lambda, \beta, \gamma | \mathbf{Z}_i, \mathbf{X}_i(t), \Gamma_i, \delta_i = 1, y_i) \\
& \quad = 1 - \pi(\mathbf{Z}_i) + \pi(\mathbf{Z}_i) \\
& \quad \times \left\{ \left(\exp\left\{ -\exp\{\lambda_{r_i} + \beta' X_i(t_{r_i - 1})\} \right\} \right) \exp\left\{ -\sum_{j=1}^{r_i - 1} \exp\{\lambda_j + \beta' X_i(t_{j-1})\} \right\} \right\} .
\end{aligned}
$$

Then

$$
\begin{aligned}
& L_O(\lambda, \beta, \gamma | \mathbf{Obs}) = \\
& \prod_{i=1}^{n} l_O^{(i)}(\lambda, \beta, \gamma | \mathbf{Z}_i, \mathbf{X}_i(t), \Gamma_i, \delta_i = 1)^{\delta_i} l_O^{(i)}(\lambda, \beta, \gamma | \mathbf{Z}_i, \mathbf{X}_i(t), \Gamma_i, \delta_i = 0)^{(1-\delta_i)}
\end{aligned}
$$

gives the observed likelihood.

BIBLIOGRAPHY

BIBLIOGRAPHY

Allison, P. (1982). Discrete time methods for the analysis of event histories. *Sociological Methodology*, page 61?98.

Annweiler, C., Beauchet, O., Bartha, R., Wells, J., Borrie, M., Hachinski, V., and Montero-Odasso, M. (2013). Motor cortex and gait in mild cognitive impairment: a magnetic resonance spectroscopy and volumetric imaging study. *Brain*, 136:85971.

Asamura, K., Abe, S., Fukuoka, H., Nakamura, Y., and Usami, S. (2005). Mutation analysis of *COL9A3*, a gene highly expressed in the cochlea, in hearing loss patients. *Auris Nasus Larynx.*, 32(2):113–7.

Aslan, J., You, H., Williamson, D. M., Endig, J., amd L Thomas, R. Y., Shu, H., Du, Y., Milewski, R., Brush, M., Possemato, A., Sprott, K., Fu, H., Greis, K., Runckel, D., Vogel, A., and Thomas, G. (2009). Akt and 14-3-3 control a *PACS-2* homeostatic switch that integrates membrane traffic with trail-induced apoptosis. *Mol Cell*, 34(4):497–509.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *Ann. Statist.*, 41(2):802–37.

Berkson, J. and Gage, R. P. (1952). Survival curves for cancer patients following treatment. *Journal of the American Statistical Association*, 47:501–15.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *Ann. Stat.*, 37:1705–32.

Biffi, A., Anderson, C., Desikan, R., Sabuncu, M., Cortellini, L., Schmansky, N., Salat, D., Rosand, J., and ADNI (2010). Genetic variation and neuroimaging measures in alzheimer disease. *Arch Neurol.*, 67(6):677–85.

Biswas, S. and Lin, S. (2012). Logistic Bayesian lasso for identifying association with rare haplotypes and application to age-related macular degeneration. *Biometrics*, 68:587–97.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society*, 11:15–53.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Braber, A., Bohlken, M., Brouwer, R., Ent, D., Kanai, R., Kahn, R., Geus, E., Pol, H., and Boomsm, D. (2013). Heritability of subcortical brain measures: A perspective for future genome-wide association studies. *NeuroImage*, 83:98–102.

Braskie, N., Ringman, J., and Thompson, P. (2011). Neuroimaging measures as endophenotypes in alzheimer's disease. *International Journal of Alzheimer's Disease*, page http://dx.doi.org/10.4061/2011/490140.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24:235083.

Brem, R. B. and Kruglyak, L. (2005). The landscape of genetic complexity across 5700 gene expression traits in yeast. *Proceddings of National Academy of Sciences*, 102:1572–77.

Brodmann, K. (2010). *Brodmann's Localisation in the Cerebral Cortex*. Springer.

Broer, L., Ikram, M., Schuur, M., DeStefano, A., Bis, J., Liu, F., Rivadeneira, F., Uitterlinden, A., Beiser, A., Longstreth, W., Hofman, A., Aulchenko, Y., Seshadri, S., Fitzpatrick, A., Oostra, B., Breteler, M., and van Duijn, C. (2011). Association of *HSP70* and its co-chaperones with alzheimer's disease. *J Alzheimers Dis.*, 25(1):93–102.

Bunea, F., She, Y., and Wegkamp, M. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Stat.*, 39:1282–1309.

Burns, L., Minster, R., Demirci, F., Barmada, M., Ganguli, M., Lopez, O. L., DeKosky, S., and Kamboha, M. (2011). Replication study of genome-wide associated SNPs with late-onset Alzheimers disease. *Am J Med Genet B Neuropsychiatr Genet*, 156(4):507–12.

Cai, C., Zou, Y., Peng, Y., and Zhang, J. (2012). smcure: An r-package for estimating semiparametric mixture cure models. *Cimput Methods Programs Biomed.*, 108(3):1255–60.

Cannon, D., Miller, J., Robison, R., Villalobos, M., Wahmhoff, N., Allen-Brady, K., McMahon, W., and Coon, H. (2010). Genome-wide linkage analyses of two repetitive behavior phenotypes in utah pedigrees with autism spectrum disorders. *Molecular Autism*, 1:3.

Carson, M. I. (2007). *Focus on Mental Retardation Research*. Nova Publishers.

Chilumuri, A. and Milton, N. (2013). The role of neurotransmitters in protection against amyloid-β toxicity by kiss-1 overexpression in SH-SY5Y neurons. *Neuroscience*, http://dx.doi.org/10.1155/2013/253210.

Chung, P., Beyens, G., Boonen, S., Papapoulos, S., Geusens, P., Karperien, M., Vanhoenacker, F., Verbruggen, L., Fransen, E., Offel, J. V., Goemaere, S., Zmierczak, H., Westhovens, R., Devogelaer, J., and Hul, W. V. (2010). The majority of the genetic risk for pagets disease of bone is explained by genetic variants close to the *CSF1*, *OPTN*, *TM7SF4*, and *TNFRSF11A* genes. *Human genetics.*, 128:615–26.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–76.

Dai, M., Freeman, B., Shikani, H. J., Bruno, F. P., Collado, J. E., Macias, R., Reznik, S. E., Davies, P., Spray, D. C., Tanowitz, H. B., Weiss, L. M., and Desruisseaux,

M. S. (2013). Altered regulation of akt signaling with murine cerebral malaria, effects on long-term neuro-cognitive function, restoration with lithium treatment. *PLoS ONE*, 10:e44117.

Desbaillets, I., Diserens, A., Tribolet, N., Hamou, M., and Meir, E. V. (1997). Upregulation of interleukin 8 by oxygen-deprived cells in glioblastoma suggests a role in leukocyte activation, chemotaxis, and angiogenesis. *J Exp Med.*, 186(8):1201–12.

Dudoit, S., SHaffer, J., and Boldrick, J. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18):71–103.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32 (2):40799.

Engler, D. and Li, Y. (2009). Survival analysis with high-dimensional covariates: An application in microarray studies. *Stat Appl Genet Mol Biol*, 8(1):1–22.

Ertekin-Taner, N. (2010). Genetics of Alzheimer disease in the pre- and post-GWAS era. *Alzheimer's Research & Therapy*, 2:3.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.*, 96:1348–60.

Fan, J. and Li, R. (2002). Variable selection for cox's proportional hazards model and frailty model. *Ann. Statist.*, 30(1):74–99.

Fan, J. and Lv, J. (2008). Sure independent screening for ultrahigh dimensional feature space. *J. R. Statist. Sco. B*, 70(5):849911.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38:104146.

Fennema-Notestine, C., Hagler, D., McEvoy, L., Fleisher, A., Wu, E., Karow, D., A.Dale, and ADNI (2009). Structural mri biomarkers for preclinical and mild alzheimers disease. *Hum Brain Mapping*, 30(10):3238–53.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332.

Fu, W. (1998). Penalized regressions: the bridge vs the lasso. *JCGS*, 7(3):397416.

Giordano, A. and Macaluso, M. (2011). *Cancer Epigenetics: Biomolecular Therapeutics in Human Cancer*. John Wiley & Sons.

Hastie, T. and Tibshirabi, R. (1990). *Generalized Additive Models*. Chapman & Hall.

Heikaus, S., Winterhager, E., Traub, O., and Grümmer, R. (2002). Responsiveness of endometrial genes *Connexin26*, *Connexin43*, *C3* and clusterin to primary estrogen, selective estrogen receptor modulators, phyto- and xenoestrogens. heikaus s, winterhager. *J Mol Endocrinol*, 29(2):239–49.

Hibar, D., Stein, J., Kohannim, O., Jahanshad, N., Saykin, A., Shen, L., Kim, S., Pankratz, N., Foroud, T., Huentelman, M., Potkin, S., Jack, C. J., Weiner, M., Toga, A., Thompson, P., and ADNI (2011). Voxelwise gene-wide association study (vgenewas): multivariate gene-based association testing in 731 elderly subjects. *Neuroimage*, 56(4):1875–91.

Horuk, R., Martin, A., Wang, Z., Schweitzer, L., Gerassimides, A., Guo, H., Lu, Z., Hesselgesser, J., Perez, H., Kim, J., Parker, J., Hadley, T., and Peiper, S. (1997). Expression of chemokine receptors by subsets of neurons in the central nervous system. *J Immunol. 1997*, 158(6):2882–90.

Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73:387–96.

Huang, C., Wahlund, L., Svensson, L., Winblad, B., and Julin, P. (2002). Cingulate cortex hypoperfusion predicts alzheimer's disease in mild cognitive impairment. *BMC Neurol*, 12:2–9.

Huang, J., Ma, S., Xie, H., and Zhang, C. (2009). A group bridge approach for variable selection. *Biometrika*, 2:339–55.

Huang, J., Sun, T., Ying, Z., Yu, Y., and Zhang, C. (2013). Oracle inequalities for the lasso in the cox model. *The Annals of Statistics*, 41(3):1142–65.

Ibrahim, J. G., Chen, M. H., and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer-Verlag New York, Inc.

Jacobs, H., Boxtel, M., Jolles, J., Verhey, F., and Uylings, H. (2012). Parietal cortex matters in alzheimers disease: An overview of structural, functional and metabolic findings. *Neuroscience and Biobehavioral Reviews*, 36:297309.

Jones, B., Barnes, J., Uylings, H., Fox, N., Frost, C., Witter, M., and Scheltens, P. (2006). Differential regional atrophy of the cingulate gyrus in alzheimer disease: a volumetric mri study. *Cereb Cortex*, 16(12):1701–08.

Kalbfleisch, J. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons.

Kohannim, O., Hibar, D., Stein, J., Jahanshad, N., Hua, X., Rajagopalan, P., Toga, A., Jack, C. J., Weiner, M., de Zubicaray, G., McMahon, K., Hansell, N., Martin, N., Wright, M., Thompson, P., and ADNI (2012). Discovery and replication of gene influences on brain structure using lasso regression. *Front Neurosci.*, 6:115.

Kong, S. and Nan, B. (2013). Non-asymptotic oracle ineqaulities for the high-dimensional cox regression via lasso. *Statistica Sinica*, Inpress.

Korac, J., Schaeffer, V., Kovacevic, I., Clement, A., Jungblut, B., Behl, C., Terzic, J., and Dikic, I. (2013). Ubiquitin-independent function of optineurin in autophagic clearance of protein aggregates. *J Cell Sci.*, 126(02):580–92.

Kuk, A. Y. C. and Chen, C. H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79:531–41.

Kukull, W., Schellenberg, G., Bowen, J., McCormick, W., Yu, C., Teri, L., Thompson, J., O'Meara, E., and Larson, E. (1996). Apolipoprotein e in alzheimer's disease risk and case detection: a case-control study. *J Clin Epidemiol.*, 49(10):1143–8.

Lange, K. (2004). *Optimization.* Springer London, Inc.

Laska, E. M. and Meisner, M. J. (1992). Nonparametric estimation and testing in a cure rate model. *Biometrics*, 48:1223–34.

Law, N. J., Taylor, J. M. G., and Sandler, H. (2002). The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure. *Biostatistics*, 3:547–63.

Li and Leal (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet*, 83:311–21.

Li, Y., Nan, B., and Zhu, J. (2013). *Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure.*

Li, Z., Gilbert, P., and Nan, B. (2008). Weighted likelihood method for grouped survival data in case-cohort studies with application to hiv vaccine trials. *Biometrics*, 64(4):1247–55.

Lia, C., Taylor, J., and Syc, J. (2001). Identifiability of cure models. *Statistics & Probability Letters*, 54(4):38995.

Lingala, H. B., Sankarathi, and Penagaluru, P. R. (2009). Role of connexin 26 (*GJB2*) & mitochondrial small ribosomal RNA (mt 12S rRNA) genes in sporadic & aminoglycoside-induced non syndromic hearing impairment. *Indian J Med Res*, 130:369–78.

Liu, X., Peng, Y., Tu, D., and Liang, H. (2012). Variable selection in semiparametric cure models based on penalized likelihood, with application to breast cancer clinical trials. *Statistics in Medicine*, 31(24):2882–91.

Liu, Y., Guo, D., Tian, L., Shang, D., Zhao, W., Li, B., Fang, W., Zhu, L., and Chen, Y. (2010). Peripheral t cells derived from Alzheimer's disease patients overexpress *CXCR2* contributing to its transendothelial migration, which is microglial TNF-alpha-dependent. *Neurobiol Aging*, 31(2):175–88.

Lounici, K., Pontil, M., Tsybakov, A. B., and van de Geer, S. (2011). Oracle inequalities and optimal inference under group sparsity. *Annal of Statistics*, 39:2164–2204.

Lu, Y., He, X., and Zhong, S. (2007). Cross-species microarray analysis with the OSCAR system suggests an *INSR*→*Pax6*→*NQO1* neuro-protective pathway in aging and Alzheimer's disease. *Nucleic Acids Res.*, 35:W105–14.

Maller, R. A. and Zhou, S. (1990). *Survival Analysis with Long-Term Survivors.* New York: Wiley.

Meier, L., van̈de Geer, S., and Bëhlmann, P. (2008). The group lasso for logistic regression. *J. R. Statist. Soc. B*, 70(1):53–71.

Meijer, R. J. and Goeman, J. J. (2014). Model selection for high-dimensional models. *Handbooks of Survival Analysis*.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *J. R. Statist. Soc. B.*, 72:417–73.

Mendenhall, W. and Hader, R. J. (1958). Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data. *Biometrika*, 45:504–20.

Mills, J., Nalpathamkalam, T., Jacobs, H., Merico, C. J. D., Hu, P., and Janitz, M. (2013). Rna-seq analysis of the parietal cortex in alzheimer's disease reveals alternatively spliced isoforms related to lipid metabolism. *Neurosci Lett.*, 536:90–5.

Moretti, D., Paternicò, D., Binetti, G., Zanetti, O., and Frisoni, G. (2012). Analysis of grey matter in thalamus and basal ganglia based on eeg a3/a2 frequency ratio reveals specific changes in subjects with mild cognitive impairment. *ASN Neuro*, 4(7):e00103.

Mosconi, L. (2005). Brain glucose metabolism in the early and speci c diagnosis of alzheimer's disease. fdg-pet studies in mci and ad. *European Journal of Nuclear Medicine and Molecular Imaging*, 32:466–510.

Nakamura, K., Ohya, W., Funakoshi, H., Sakaguchi, G., Kato, A., Takeda, M., Kudo, T., and Nakamura, T. (2006). Possible role of scavenger receptor *SRCL* in the clearance of amyloid-beta in alzheimer's disease. *J Neurosci Res.*, 84(4):874–90.

Nan, B. (2010). Survival analysis with high-dimensional covariates. *High-Dimensional Data Analysis - Frontiers of Statoistics. Edited by T. Cai and X. Shen*, pages 223–54.

Nicole, A. (2011). Integration of nutritional status with germline proliferation: characterizing the roles of nhr-88 and nhr-49 in the c. elegans gonad.

Nitsch, R., Bechmann, I., Deisz, R., Haas, D., Lehmann, T., Wendling, U., and Zipp, F. (2000). Human brain-cell death induced by tumour-necrosis-factor-related apoptosis-inducing ligand (trail). *Lancet*, 356(9232):827–8.

Nolte, C., Rastegar, M., Amores, A., Bouchard, M., Grote, D., Maas, R., Kovacs, E., Postlethwait, J., Rambaldi, I., Rowan, S., Yan, Y., Zhang, F., and Featherstone, M. (2006). Stereospecificity and *PAX6* function direct *Hoxd4* neural enhancer activity along the antero-posterior axis developmental biology. *Developmental Biology*, 299(2):582–93.

Obozinski, G., Wainwright, M., and Jordan, M. (2011). Support union recovery in high-dimensional multivariate regression. *Ann. Stat.*, 39:1–47.

Okello, A., Koivunen, J., Edison, P., Archer, H., Turkheimer, F., Någren, K., Bullock, R., Walker, Z., Kennedy, A., Fox, N., Rossor, M., Rinne, J., and Brooks, D. (2009). Conversion of amyloid positive and negative mci to ad over 3 years: an 11c-pib pet study. *Neurology*, 73(10):754–60.

Park, M. Y. and Hastie, T. (2008). Penalized logistic regression for detecting gene-interactions. *Biostat*, 9:30–50.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., D-Y. Noh, J. P., and Wang, P. (2010). Newblock regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.*, 4:53–77.

Peng, J. Z. Y. (2009). Accelerated hazards mixture cure model. *Lifetime Data Analysis*, 15(4):455–467.

Peper, J., Brouwer, R., Boomsma, D., Kahn, R., and Pol, H. (2007). Genetic influences on human brain structure: A review of brain imaging studies in twins. *Human Brain Mapping*, 28(6):464–73.

Petersen, R., Roberts, R., Knopman, D., Boeve, B., Geda, Y., Ivnik, R., Smith, G., and Jack, C. (2009). Mild cognitive impairment: ten years later. *Arch Neurol.*, 66(12):1447–55.

Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., and Kokmen, E. (1999). Mild cognitive impairment: clinical characterization and outcome. *Arch. Neurol.*, 56(3):303–8.

Pratt, J. (1981). Concavity of the log likelihood. *Journal of the American Statistical Association*, 76(373):103–106.

Prentice, R. L. and Gloeckler, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, 34:5767.

Risacher, S., West, A. S. J., Shen, L., Firpi, H., and McDonald, B. (2009). Baseline mri predictors of conversion from mci to probable ad in the adni cohort. *Curr Alzheimer Res.*, 6(4):347–61.

Sakamoto, K. and Holman, G. D. (2008). Emerging role for *AS160/TBC1D4* and *TBC1D1* in the regulation of *GLUT4* traffic. *Am J Physiol Endocrinol Metab*, 295:e29–37.

Silver, M., Montana, G., and Alzheimer's-Disease-Neuroimaging-Initiative (2012). Fast identification of biological pathways associated with a quantitative trait using group lasso with overlaps. *Stat Appl Genet Mol Biol*, 11(1):1–43.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22:231–45.

Singer, J. D. and Willett, J. B. (1993). It's about time: using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics*, 18:155–95.

Solovieva, S., Lohiniva, J., Leino-Arjas, P., Raininko, R., Luoma, K., Ala-Kokko, L., and Riihimäki, H. (2006). Intervertebral disc degeneration in relation to the *COL9A3* and the *IL-1ss* gene polymorphisms. *Eur Spine J.*, 15(5):613–9.

Sra, S., Nowozin, S., and Wright, S. J. (2012). *Optimization for Machine Learning*. The MIT Press.

Stein, J., Hua, X., Lee, S., Ho, A., Leow, A., Toga, A., Saykin, A., Shen, L., Foroud, T., Pankratz, N., Huentelman, M., Craig, D., Gerber, J., Allen, A., Corneveaux, J., Dechairo, B., Potkin, S., Weiner, M., Thompson, P., and Initiative, A. D. N. (2010a). Voxelwise genome-wide association study (vgwas). *Neuroimage*, 53(3):1160–74.

Stein, J., Hua, X., Morra, J., Lee, S., Hibar, D., Ho, A., Leow, A., Toga, A., Sul, J., Kang, H., Eskin, E., AJ, A. S., Shen, L., Foroud, T., Pankratz, N., Huentelman, M., Craig, D., Gerber, J., Allen, A., Corneveaux, J., DA, D. S., Webster, J., DeChairo, B., Potkin, S., Jack, C., Weiner, M., Thompson, P., and ANDI (2010b). Genome-wide analysis reveals novel genes influencing temporal lobe structure with relevance to neurodegeneration in alzheimer's disease. *Neuroimage*, 51(2):542–54.

Sun, H., Hu, B., Yao, Z., and Jackson, M. (2013). A pet study of discrimination of cerebral glucose metabolism in alzheimer's disease and mild cognitive impairment. *Biomedical Engineering and Informatics (BMEI), 2013 6th International Conference on*, pages 6–11.

Suva, D., Favre, I., Kraftsik, R., Esteban, M., Lobrinus, A., and Miklossy, J. (1999). Primary motor cortex involvement in alzheimer disease. *J Neuropathol Exp Neurol.*, 58:112534.

Talbot, K., Wang, H., Kazi, H., Han, L., Bakshi, K. P., Stucky, A., Fuino, R. L., Kawaguchi, K. R., Samoyedny, A. J., Wilson, R. S., Arvanitakis, Z., Schneider, J. A., Wolf, B. A., Bennett, D. A., Trojanowski, J. Q., and Arnold1, S. E. (2012). Demonstrated brain insulin resistance in alzheimers disease patients is associated with IGF-1 resistance, IRS-1 dysregulation, and cognitive decline. *Journal of Clinical Investigation*, 122(4):1316–38.

Taylor, J., Lockhart, R., Tibshirani, R., and Tibshirani, R. (2014). Post-selection adaptive inference for least angle regression and the lasso. *arXiv:1401.3889*.

Taylor, J. M. G. (1995). Semi-parametric estimation in failure time mixture models. *Biometrics*, 51:899–907.

Teasdale, R. D. and Collins, B. M. (2012). Insights into the PX (phox-homology) domain and *SNX* (sorting nexin) protein families: structures, functions and roles in disease. *Biochem. J.*, 441:39–59.

Tibshirabi, R. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16:385–95.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B.*, 58:267–88.

Tsai, H., Frost, E., To, V., Ffrench-Constant, S. R. C., Geertman, R., Ransohoff, R., and Miller, R. (2002). The chemokine receptor *CXCR2* controls positioning of oligodendrocyte precursors in developing spinal cord by arresting their migration. *Cell*, 110(3):373–83.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization: Theory and Applications*, 109:275–94.

U Brüggemeier, A Geerts, S. G. (2004). G-protein coupled receptor lustr2 and uses thereof.

Vallès, A., Grijpink-Ongering, L., de Bree, F., Tuinstra, T., and Ronken, E. (2006). Differential regulation of the *CXCR2* chemokine network in rat brain trauma: implications for neuroimmune interactions and neuronal survival. *Neurobiol Dis.*, 22(2):312–22.

Vogt, B. (2009). *Cingulate Neurobiology and Disease*. Oxford University Press.

Wang, J., Shi, Z., Xu, X., Xin, G., Chen, J., Qi, L., and Li, P. (2013). Triptolide inhibits amyloid-β production and protects neural cells by inhibiting *CXCR2* activity. *J Alzheimers Dis.*, 33(1):217–29.

Wang, S., Nan, B., Zhu, J., and Beer, D. G. (2008). Doubly penalized buckley-james method for survival data with high-dimensional covariates. *Biometrics*, 64:132–40.

Wang, S., Zhou, N., Nan, B., and Zhu, J. (2009). Hierarchically penalized cox model for survival data with grouped variables and its oracle property. *Biometrika*, 96:322.

Wong, C. C. Y., Meaburn, E. L., Ronald, A., Price, T. S., Jeffries, A. R., Schalkwyk, L. C., Plomin, R., and Mill, J. (2013). Methylomic analysis of monozygotic twins discordant for autism spectrum disorder and related behavioural traits. *Molecular Psychiatry*, doi: 10.1038/mp.2013.41.

Wu, T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Annal of Applied Statistics*, 2:224–44.

Xia, M., Qin, S., McNamara, M., and Hyman, B. T. (1996). Type b *IL8* receptor (*IL8RB*) in neuritic plaques of Alzheimer's disease. *Journal of Neuropathology and Experimental Neurology*, 55(5).

Xie, H., D, D. L., Leung, K., Chen, V., Zhu, K., Chan, W., Choi, R., Massoulié, J., and Tsim, K. (2010). argeting acetylcholinesterase to membrane rafts: a function mediated by the proline-rich membrane anchor (*PRiMA*) in neurons. *J Biol Chem.*, 285(15):11537–46.

Yaffe, K., Krueger, K., Cummings, S. R., Blackwell, T., Henderson, V. W., Sarkar, S., Ensrud, K., and Grady, D. (2005). Effect of raloxifene on prevention of dementia and cognitive impairment in older women: The multiple outcomes of raloxifene evaluation (MORE) randomized trial. *Am J Psychiatry*, 162:683–90.

Yakovlev, A. Y., Asselain, B., Bardou, V. J., Fourquet, A., Hoang, T., Rochefediere, A., and Tsodikov, A. D. (1993). A simple stochastic model of tumor recurrence and its applications to data on pre-menopausal breast cancer. *Biometrie et Analyse deDormees Spatio Temporelles*, 12:66–82.

Yakovlev, A. Y. and Tsodikov, A. D. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. New Jersey: World Scientific.

Yang, J., Li, S., and Liu, Y. (2013). Systematic analysis of diabetes- and glucose metabolism-related proteins and its application to Alzheimers disease. *J. Biomedical Science and Engineering*, 6:615–44.

Yang, Y. and Zou, H. (2013). A cocktail algorithm for solving the elastic net penalized cox's regression in high dimensions. *Statistics and Its Inference*, 6:167–73.

Ye, J., Farnum, M., Yang, E., Verbeeck, R., Lobanov, V., Raghavan, N., Novak, G., DiBernardo, A., Narayan, V., and ADNI (2012). Sparse learning and stability selection for predicting mci to ad conversion using baseline adni data. *BMC Neurology*, 12(46):1–12.

Yin, J. and Li, H. (2011). A sparse conditional gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Stat.*, 4:2630–50.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B.*, 68:49–67.

Zamdborg, L. and Ma, P. (2009). Discovery of protein-dna interactions by penalized multivariate regression. *Nucl. Acids Res.*, 37:5246–54.

Zarrinpar, A., Park, S. H., and Lim, W. A. (2003). Optimization of specificity in a cellular protein interaction network by negative selection. *Nature*, 426:676–80.

Zhang, S., Ching, W., Tsing, N., Leung, H., and Guo, D. (2010). A new multiple regression approach for the construction of genetic regulatory networks. *Artificial Intelligence in Medicine*, 48:153–60.

Zhou, H., Sehl, M. E., Sinsheimer, J. S., and Lange, K. (2010). Association screening of common and rare genetic variants by penalized regression. *Nucl. Acids Res.*, 26:2375–82.

Zhou, N. and Zhu, J. (2010). Group variable selection via a hierarchical lasso and its oracle property. *Statistics and Its Interface*, 4:557–74.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Statist. Assoc.*, 101:1418–29.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B.*, 67:301–20.

Zou, H. and Zhang, H. (2009). On the adaptive elastic net with a diverging number of parameters. *Annals of Statistics*, 37 (4):173351.