The Pennsylvania State University
The Graduate School

THE BAYESIAN LASSO, BAYESIAN SCAD AND BAYESIAN GROUP LASSO WITH APPLICATIONS TO GENOME-WIDE ASSOCIATION STUDIES

A Dissertation in
Statistics
by
Jiahan Li

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2011

The dissertation of Jiahan Li was reviewed and approved* by the following:

Rongling Wu
Professor of Public Health Sciences, Statistics
Dissertation Co-Advisor, Co-Chair of Committee

Runze Li
Professor of Statistics, Public Health Sciences
Dissertation Co-Advisor, Co-Chair of Committee

Bruce Lindsay
Professor of Statistics, Head of the Department of Statistics

Tao Yao
Assistant Professor of Industrial and Manufacturing Engineering

*Signatures are on file in the Graduate School.

Abstract

Recently, genome-wide association studies (GWAS) have successfully identified genes that may affect complex traits or diseases. However, the standard statistical tests for each single-nucleotide polymorphism (SNP) separately are too simple to elucidate a comprehensive picture of the genetic architecture of phenotypes. A simultaneous analysis of a large number of SNPs, although statistically challenging, especially with a small number of samples, is crucial for genetic modeling. This is a variable selection problem for high-dimensional data, with SNPs as the predictors and phenotypes as the responses in our statistical model.

In genome-wide association studies, phenotypical values are either collected at a single time point for each subject, or collected repeatedly over a period at subject-specific time points. When the response variable is univariate, we present two-stage procedures designed for the problems where the number of predictors greatly exceeds the number of observations. At the first stage, we preprocess the data such that variable selection procedure can be proceeded in an accurate and efficient manner. At the second stage, variable selection techniques based on penalized linear regression are applied to the preprocessed data.

When the longitudinal phenotype of interest is measured at irregularly spaced time points, we develop a Bayesian regularized estimation procedure for the variable selection of nonparametric varying-coefficient models. Our method could simultaneously selection important predictors and estimate their time-varying effects. We approximate time-varying effects by Legendre polynomials, and present a Bayesian hierarchical model with group lasso penalties that encourage sparse solutions at the group level.

In both scenarios, our models obviate the choice of the tuning parameters

by imposing diffuse hyperpriors on them and estimating them along with other parameters, and provide not only point estimates but also interval estimates of all parameters. Markov chain Monte Carlo (MCMC) algorithms are developed to simulate the parameters from their posterior distributions. The proposed methods are illustrated with numerical examples and a real data set from the Framingham Heart Study.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my deepest gratitude to my advisor, Dr. Rongling Wu, for his excellent guidance, encouragement and help for the past several years at The University of Florida and The Pennsylvania State University. His infectious enthusiasm and immense talent have been major driving forces through my graduate studies. His mentorship is paramount in building the solid foundation for my future academic endeavors.

I would like to deeply thank my co-advisor, Dr. Runze Li. As an expert in statistics especially in variable selection and functional data analysis, his courses, publications and brilliant ideas inspired me from time to time. Without him my research could not be carried out in the manner it has been done.

I would also like to express my gratitude to Dr. Bruce Lindsay. His remarkable contributions to statistical theory and, in particular, to mixture models deepened my understanding in this field. Besides, the useful discussions we had and his valuable feedback helped me to improve this dissertation in many ways.

Special thanks go to Dr. Tao Yao, who is from the Department of Industrial and Manufacturing Engineering and willing to serve on my committee. His insights and suggestions helped me to shape my research skills and gain a better understanding of other disciplines.

At last but not at least I want to thank my family. They taught me the value of hard work and perseverance. I thank my family for their unconditioned love and support, and for their patience and encouragement that carried me on through difficult times.

Chapter 1

Introduction

1.1 Genome-wide Association Studies

Phenotypic variance among individuals is in general attributed to two factors: genetic factor and environmental factor. In particular, genetic factor refers to variations across human genome. Researchers in genetics, statistics and bioinformatics have attempted to develop various molecular tools and statistical models to explain genetic and environmental influences.

In 1918, R. A. Fisher established quantitative genetics, by which the proportion of total phenotypic variation in a population that is due to genetic factors can be estimated. Later in 1980s, with the aid of genetic markers and quantitative trait loci mapping models, regions of the genome that contain genes responsible for a quantitative trait can be identified. Recently, genotyping technologies allow the fast and accurate collection of genotype data throughout the entire genome for many subjects. By genome-wide association studies (GWASs), the genetic variants associated with a complex disease or trait, their chromosomal distribution and individual effects, can be identified. Moreover, the interactions between difference genetic variants and the interplay between genetic factors and environmental factors can be revealed.

With the rapid development of efficient and inexpensive high-throughput genotyping techniques, GWAS has been increasingly used to study the genetic control of complex human diseases and biochemical or anthropometric traits (Shuldiner et al. 2009; Takeuchi et al. 2009; Teichert et al. 2009; Yang et al. 2010), gaining a

growing body of novel findings with potential clinical relevance (Daly 2010). Since 2005, there have been nearly 500 GWAS published aiming to improve our understanding of human diseases or traits, such as type 1 and type 2 diabetes (Sladek et al. 2007; Steinthorsdottir et al. 2007), inflammatory bowel disease (Duerr et al. 2006; Hampe et al. 2007), prostate cancer (Thomas et al. 2008, Gudmundsson et al. 2008), asthma (Moffatt et al. 2007), height (Weedon et al. 2007) and fat mass (Frayling 2007), among which there are two kinds of cohorts being studied: case-control cohorts or population cohorts.

GWAS based on case-control cohorts want to test the associations between single-nucleotide polymorphisms (SNPs) and diseases, while those based on population cohorts would like to estimate genetic effects of SNPs on human traits. In both cases, usually there are hundreds of thousands of SNPs genotyped on samples involving thousands of subjects. This typical problem, having the number of predictors far exceeding the number of observations, makes it impossible to analyze the data using traditional multivariate regression. In current GWAS, simple univariate linear regression that analyzes one SNP at a time is usually used. By adjusting for multiple comparisons using Bonferroni correction or FDR technique, the significance level of the detected SNPs is then calculated (McCarthy et al. 2008), and those SNPs whose p-values are less than the threshold are claimed to be significant and responsible for the trait we are interested in.

These single SNP-based GWAS analysis have been instrumental for reproducibly detecting significants genes for various complex diseases or traits (Donnelly 2008; Hindorff et al. 2009). However, such strategies have four major disadvantages, limiting the future applications of GWAS. First, a single SNP analysis can only detect a very small portion of genetic variation of most complex traits, and may not be powerful for identifying weaker associations (Hoggart et al. 2008) because it subjects to severe multiple comparisons adjustment. Second, different genes may interact with each other to form a complex network of genetic interactions, which cannot be characterized from a single SNP analysis. Third, many GWAS analyze genetic associations separately for different environments, such as males and females, and then make an across-environment comparison in genetic effects. This analysis is neither powerful nor precise for the identification of gene-environment interactions. Fourth, when longitudinal or functional phenotypical

data are presented in many clinical trials and social science research, the statistical analysis using a single measurement from each subject is not capable of revealing the dynamic pattern of genetic control over a time course. Because of these limitations, many authors have developed alternative approaches for simultaneously analyzing multiple SNPs for GWASs (Wu et al. 2009; Yang et al. 2010; Logsdon et al. 2010), although most approaches focus on case-control cohorts.

There is a daunting need on the development of statistical models to identify SNPs with significant effects on quantitative traits in population cohorts and estimate the effects of all selected SNPs simultaneously. This is a variable selection problem for high-dimensional data, with SNPs as the predictors and phenotypes as the responses. Successful variable selection models for GWAS could not only reduce the computational cost dramatically, but assess the risk of disease-susceptible loci more accurately to guide future biomedical research.

1.2 Penalized Least Squares

Variable selection plays a central role in high-dimensional statistical modeling. Given a large number of predictors in the data set, it retains only a subset of the predictors in a regression model, and eliminates the rest from the final model. By doing this, prediction accuracy can be improved, and the estimated final model is more interpretable.

Traditionally, the selection of a subset of predictors for a final regression model is implemented by some intuitive procedures, including forward addition, backward elimination, and stepwise selection. These procedures are based on different variable selection criteria, such as Mallows' C_p (Mallows, 1973), AIC (Akaike,1973), and BIC (Schwarz, 1978), which involve optimizing a fit criterion (a least squares or likelihood function) modified by a model complexity penalty. This reflects our intuition that a good model should fit well while using few parameters. However, these intuitive approaches are computationally expensive and unstable even when the number of predictors is not large. When the number of potential predictors is moderate or large, these methods become computationally expensive or infeasible.

Recently, alternative approaches have been developed to overcome the theoretical and computational disadvantages of classical variable selection procedures. The

new development includes, among others, ridge regression, bridge regression (Frank and Friedman, 1993), least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), elastic net (Zou and Hastie, 2005) and the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001).

It has been shown that all variable selection techniques mentioned above can be unified in a penalized least squares framework. Consider a linear regression model

$$\mathbf{y} = X\boldsymbol{\beta} + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I_n), \tag{1.1}$$

where $\mathbf{y}$ is the vector of response variable, $X = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$ is a $n \times p$ design matrix, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is the vector of regression coefficients. The penalized least squares can be expressed as

$$Q(\beta) = \frac{1}{2}||y - X\boldsymbol{\beta}||^2 + \sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{1.2}$$

where $p_\lambda(.)$ is a penalty function indexed by the regularization parameter $\lambda > 0$, and λ balances the accuracy of in-sample fit and the parsimony of final model. By minimizing the penalized least squares (1.2), we hope to simultaneously select important variables and estimate their associated regression coefficients. In other words, those predictors whose regression coefficients are estimated as zero will not be considered in the final model.

In particular, the best subset selection that drops all small predictors takes the hard thresholding penalty function $p_\lambda(\theta) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| - \lambda)$. The AIC of Akaike (1973), which estimates the Kullback-Leibler distance of a model from the true likelihood function, is asymptotically equivalent to the following penalized least squares

$$\frac{1}{2}||y - X\boldsymbol{\beta}||^2 + n\frac{(\sigma\sqrt{2/n})^2}{2}\sum_{j=1}^{d} I(|\beta_j| \neq 0). \tag{1.3}$$

BIC is asymptotically equivalent to the following penalized least squares

$$\frac{1}{2}||y - X\boldsymbol{\beta}||^2 + n\frac{(\sigma\sqrt{\log(n)/n})^2}{2}\sum_{j=1}^{d} I(|\beta_j| \neq 0). \tag{1.4}$$

It can be seen that these classical variable selection criteria evaluate the complexity of final model based on the number of predictors, but more sophisticated criteria may lead to better performance and nice theoretical properties. This gives rise to penalized least squares with continuous penalty function, which takes into account both the number of predictors in the final model and the size of estimated regression coefficients. The continuous penalty function could be either convex or nonconvex, leading to different theoretical properties and empirical variable selection performance.

For example, ridge regression does a proportional shrink by imposing the convex L_2 penalty $p_\lambda(\theta) = \lambda\theta^2$, and thus fails to produce a sparse model. Lasso, proposed by Tibshirani (1996), is the penalized least square estimates with the L_1 penalty $p_\lambda(\theta) = \lambda|\theta|$. It is effective for continuous model selection, and is advantageous when the number of predictors p is greater than the number of observations n (Chen, 1998). Zou and Hastie (2005) suggested a compromise between the ridge regression and lasso by the elastic net penalty, which is a linear combination of the L_1 penalty and L_2 penalty.

However, it has been shown that L_2 penalized regression could not set small estimated coefficients to zero to achieve variable selection, since convex L_q penalty with $q > 1$ does not satisfy the sparsity condition. Moreover, the convex L_1 penalty does not satisfy the unbiasedness condition, and the resulting estimator is biased when the true regression coefficient is large. For this reason, Fan and Li (2001) provided deep insights into how penalty function should be chosen, and proposed the smoothly clipped absolute deviation (SCAD) penalty, whose derivative is defined as

$$p'_\lambda(|\beta_j|) = \lambda I(|\beta_j| \leq \lambda) + \frac{(s\lambda - |\beta_j|)^+}{(s-1)\lambda} I(|\beta_j| > \lambda), \tag{1.5}$$

for $\beta_j \neq 0$ and $p_\lambda(0) = 0$. Usually $s = 3.7$ from a Bayesian perspective. SCAD estimator performs as well as the oracle procedure and possesses three desirable properties: sparsity, continuity and unbiasedness. The regularization parameter λ in the penalty functions controls the amount of shrinkage toward zero: the larger the value of λ, the greater the amount of shrinkage. It should be adaptively chosen to minimize an estimate of expected prediction error. Although it is challenging to minimize the SCAD penalized least squares due to the nonconvexity of SCAD

penalty function, various optimization methods have been proposed (see Fan and Li, 2001; Hunter and Li, 2005; Zou and Li, 2008).

In genome-wide association studies, the number of subjects is usually on the order of thousands while the number of SNPs is usually on the order of millions. The same order of magnitude occurs in microarray gene expression data and many other applications. In such problems, the noise level is very high, and it is more difficult to identify a subset of important predictors and build reliable predictive models. For variable selection problems where the number of predictor variables is much larger than that of subjects, Fan and Lv (2008) proposed a two-stage procedure by first suppressing the high dimensionality of response into its low-dimensional representation and then finding a subset of predictors that can predict the suppressed response. This procedure enjoys the sure screening property, in the sense that all the important predictor variables are retained with asymptotic probability one. Paul et al. (2008) also proposed a two-stage variable selection approach based on supervised principle components. Specifically, based on those predictors whose marginal correlations with the response are large enough, a denoised version of the response variable is obtained. In the following analysis, the new response will replace the original response in hopes of enhancing the variable selection performance.

1.3 Contributions of This Dissertation

This dissertation developed the Bayesian SCAD and Bayesian group lasso, and applied these new developments to the analysis of genome-wide association studies. Aiming at specific characteristics of high-dimensional data sets in GWAS, two-stage variable selection procedures and the variable selection for nonparametric varying-coefficient models are presented respectively. In the Bayesian framework, these procedures overcome several disadvantages of their frequentist counterparts, and expand the literature on variable selection and genome-wide association studies. Genome-wide association studies equipped with these sophisticated statistical models will play a more important role in identifying genetic associations for complex traits and diseases.

This dissertation is organized as follows. Chapter 2 reviews some theoretical

properties and computational techniques of lasso penalized regression, and shows how lasso penalize regression and preconditioning can be applied to the high-dimensional data analysis of genome-wide association studies. A GWAS model that bridges phenotypic values and genotypes is given, and the corresponding variable selection algorithms are presented in a Bayesian framework. Simulation studies demonstrate its advantages over traditional single SNP analysis and the merit of preconditioning step.

Chapter 3 goes beyond Chapter 2 by considering SCAD penalized regression instead of lasso penalized regression, in order to enhance the variable selection performance and reduce the estimation bias. Moreover, sure independence screening is employed to reduce the dimensionality of feature space entering the variable selection procedure. This step largely decreases the computational burden as well as the false positive rate in the following variable selection step. We developed the Bayesian SCAD algorithm based on the local linear approximation of penalty functions, and applied this model and corresponding algorithms to both simulated data sets and a real data set.

In Chapter 4, we extend models in previous chapters by accommodating longitudinal phenotypes measured at irregularly spaced time points, as commonly seen in clinical trials and biomedical research. This leads to a nonparametric varying-coefficient model where the genetic effects of SNPs may vary over time. We developed the variable selection procedure for varying-coefficient models through group lasso penalized regression, and presented MCMC algorithms for statistical inference. Both simulations and real data analysis are performed.

Finally, Chapter 5 provides some discussion for our findings and future research.

Chapter 2

The Preconditioned Bayesian Lasso for Genome-wide Association Studies

2.1 Introduction

Variable selection plays a central role in high-dimensional data analysis. It could identify predictors with nonzero effects and enhance the predictive power of the final model for the high-dimensional statistical problem. Among various penalized methods, the penalized least squares method with L_1 penalty is well studied and fundamental to the computation of other penalized least square estimators. It does a kind of continuous subset selection, since making λ sufficiently large will force a subset of the coefficients to be exactly zero. By convex duality, Rosset and Zhu (2007) showed that when $p > n$, the number of non-zero coefficients is at most n for all values of λ, and thus the lasso provides a severe form of feature selection.

When the L_1 penalty is used, the objective function is convex and hence convex optimization algorithms can be applied. Efron et al. (2004) proposed a fast and efficient least angle regression (LARS) algorithm for lasso regression, a simple modification of which produces the entire LASSO solution path. The computation is based on the fact that the lasso solution path is piecewise linear in λ. Fu (1998) and Daubechies et al. (2004) proposed a coordinate descent algorithm,

which iteratively optimizes one parameter at a time, holding other parameters fixed. Furthermore, Park and Casella (2008) introduced a Bayesian lasso approach where Laplace prior gives each regression coefficient a high probability of being near zero and in the meantime gives each coefficient a chance to be large. In this chapter, we focus on the Bayesian lasso which in practical could handle large problems, and give posterior probabilities that are easy to interpret.

In genome-wide association studies, the number of SNPs is usually much larger than the number of observations. When the dimensionality p greatly exceeds the sample size n, variable selection should be implemented with caution. Fan and Li (2001) showed that, when the dimension p is fixed or diverges more slowly than n, the SCAD estimator has three desirable properties. Kim et al. (2008) proved that for high-dimensional problems, the SCAD estimator still has the oracle property and can achieve model selection consistency, but they argued that when signal variables are highly correlated or the true model is sparse, the lasso outperforms the SCAD. However, in this case, some necessary conditions for the covariance matrix of observations have to be satisfied, see Zhao and Yu (2006) and Zou (2005), without which lasso variable selection process is not consistent.

Recently, two-stage methods for the variable selection problems with $p > n$ are recommended and have proved powerful. Fan and Lv (2008) suggested to reduce the dimensionality p from a large scale to a smaller scale d by an efficient and reliable method, and then apply well-developed variable selection techniques to the reduced feature space. Paul et al. (2008) also suggested a two-stage approach, which first finds a consistent predictor of the true response and then finds a subset of predictors that can predict the "preconditioned" response variable by a standard variable selection procedure. They showed that under a certain Gaussian latent variable model, application of the lasso to the preconditioned response variable is consistent as p and n increase, and the procedure gives a more accurate estimate than lasso when the observational noise is rather large.

In this chapter, we formulate a two-stage variable selection procedure for identifying important SNPs in GWAS. In step one, we find a linear combination of predictors that are strongly correlated with the response by a supervised principle component analysis and get a consistent "preconditioned" estimate of response variable. In step two, we implement the Bayesian lasso for variable selection based

on the "preconditioned" response that mitigates the observational noise. The Markov chain Monte Carlo (MCMC) algorithm is used to estimate all the parameters. Our model shows a great flexibility to fit many SNPs and many covariates at the same time. The statistical properties of the model were investigated through simulation studies. We use a real GWAS data set from the Framingham Heart Study to validate the usefulness and utilization of the new model.

2.2 Bayesian GWAS Model

2.2.1 Preconditioning

Preconditioning is a technique that is implemented before model selection in order to mitigate the effects of noise on the selection process, when the number of predictors far exceeds the number of observations (Paul et al., 2008). It finds a consistent estimate $\tilde{\mathbf{y}}$ of response variable using supervised principal components, which will replace $\mathbf{y}$ in the following variable selection method.

Suppose we have p variables, principle component analysis is an unsupervised learning algorithm to find linear combinations of variables that exhibit large variation in the dataset. Note that in the context of linear regression model (1.1), the leading component of the result of applying standard principal component analysis to X may not be highly correlated with the response variable $\mathbf{y}$. However, what we are more interested in here are linear combinations with both high variation and significant correlations with the response. To encourage principal component analysis to find linear combinations of predictors that have high correlations with the response, we want to restrict our attention to predictors which individually have sizable correlations with the outcome. This method is summarized as supervised principal components (SPC) (Bair et al., 2006).

Supervised principal components first finds the estimates the standardized regression coefficients $\hat{\beta}_1, \cdots, \hat{\beta}_p$ for the univariate linear regression model

$$y = X_j\beta_j + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I_n), \quad j = 1, \cdots, p, \tag{2.1}$$

where $\mathbf{y}$ is the vector of response variable, X_j is the j-th column of the $n \times p$ design matrix X, and β_j is the regression coefficient for the j-th predictor. Then

for a threshold θ, it forms a reduced design matrix consisting of only those predictors whose estimated regression coefficients exceed θ in absolute value, that is, the reduced design matrix $X_{reduced}$ consists of the j'-th column of X, where $j' \in \{j : |\hat{\beta}_j| > \theta\}$. Finally, it computes the principal components of $X_{reduced}$, which are called supervised principal components. The first m supervised principal components can serve as independent variables in a linear regression model, based on which a consistent predictor $\tilde{\mathbf{y}}$ of the true response is obtained. In practice, θ and m could be selected by cross-validation. Now we are ready to apply a standard variable selection procedure to the preconditioned response variable $\tilde{\mathbf{y}}$.

For illustration, we generate data on $p = 3000$ and $n = 100$ according to the model

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \sigma Z_i, \quad i = 1, \cdots, n, \tag{2.2}$$

where $\sigma = 5$, $Z_i \sim N(0,1)$, $\beta_j = 1$ for $1 \leq j \leq 20$ and $\beta_j = 0$ for $j > 20$. We consider the additive model in the genome-wide association studies where the range of the predictors x_{ij} is restricted to the three values 1, 0 and -1, corresponding to the three genotypes AA, Aa and aa, respectively. For ease of simulation, x_{ij} is derived from u_{ij}, where each u_{ij} has a standard normal distribution marginally, and $cov(u_{ij}, u_{ik}) = \rho$ for $j, k \leq 20, j \neq k$. Then, to mimic a SNP with equal allele frequencies, we set

$$x_{ij} = \begin{cases} 1, & u_{ij} > c \\ 0, & -c \leq u_{ij} \leq c \\ -1, & u_{ij} < -c, \end{cases}$$

where $-c$ is the first quartile of a standard normal distribution. Therefore, the first 20 predictors are pairwisely correlated, while the remainder are uncorrelated. In every simulation, we set $\rho = 0.5$.

We implemented two selection methods to the 50 simulated datasets from this model: standard lasso and lasso applied to the preconditioned response from supervised principal components. Let N_1 be the number of nonzero estimates for the first 20 predictors, or $N_1 = \sum_{j=1}^{20} I(\hat{\beta}_j \neq 0)$, and let $N_2 = \sum_{j=21}^{p} I(\hat{\beta}_j \neq 0)$ be the number of nonzero estimates for the remaining predictors. The average numbers of N_1 and N_2 over 50 simulations are shown in Table 2.1. We see that when $p \gg n$, the preconditioned lasso is able to identify nonzero coefficients and zero coefficients

correctly in almost every simulation.

Table 2.1. Performance of lasso and preconditioned lasso in a simulation example.

Method	N_1	N_2
Lasso	16.04	5.84
Preconditioned lasso	19.96	0.38

2.2.2 Lasso Penalized Regression

Given phenotypical measurements and genotype information, we could obtain the preconditioned response $\tilde{y}$ based on the generic form of linear regression (1.1). However in genome-wide association studies, a number of covariates, which are either discrete or continuous, may be measured for each subject. In order to estimate genetic effects precisely by adjusting for these covariates, a GWAS model that takes into account the effects of important covariates would be more appropriate. Therefore, we describe the preconditioned value $\tilde{y}_i$ of a quantitative trait for subject i as

$$\tilde{y}_i = \mu + \mathbf{X}_i^T\boldsymbol{\alpha} + \mathbf{Z}_i^T\boldsymbol{\beta} + \boldsymbol{\xi}_i^T\mathbf{a} + \boldsymbol{\zeta}_i^T\mathbf{d} + \epsilon_i, \quad i = 1, \cdots, n, \tag{2.3}$$

where μ is the overall mean, $\mathbf{X}_i$ is the d_1-dimensional vector of discrete covariates for subject i, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_{d_1})^T$ is the vector of regression coefficients for discrete covariates, $\mathbf{Z}_i$ is the d_2-dimensional vector of continuous covariates for subject i, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_{d_2})^T$ is the vector of regression coefficients for continuous covariates, $\mathbf{a} = (a_1, \cdots, a_p)^T$ and $\mathbf{d} = (d_1, \cdots, d_p)^T$ are the p-dimensional vectors of the additive and dominant effects of SNPs, respectively, $\boldsymbol{\xi}_i$ and $\boldsymbol{\zeta}_i$ are the indicator vectors of the additive and dominant effects of SNPs for subject i, and ϵ_i is the residual error assumed to follow a $N(0, \sigma^2)$ distribution. The j-th elements of $\boldsymbol{\xi}_i$ and $\boldsymbol{\zeta}_i$ are defined as

$$\xi_{ij} = \begin{cases} 1, & \text{if the genotype of SNP } j \text{ is } AA \\ 0, & \text{if the genotype of SNP } j \text{ is } Aa \\ -1, & \text{if the genotype of SNP } j \text{ is } aa, \end{cases}$$

$$\zeta_{ij} = \begin{cases} 1, & \text{if the genotype of SNP } j \text{ is } Aa \\ 0, & \text{if the genotype of SNP } j\text{is } AA \text{ or } aa. \end{cases}$$

Despite $p \gg n$ in the GWAS, most of the regression coefficients in (2.3) are expected to have no or only weak effects on the phenotype. To identify a few SNPs that may have notable effects and enhance prediction performance, we put L_1 lasso penalties on the sizes of additive effects and the dominant effects and encourage sparse solutions using

$$\sum_{j=1}^{p} |a_j| \leq t, \quad \sum_{j=1}^{p} |d_j| \leq t^*, \quad \text{for} t \geq 0, t^* \geq 0, \tag{2.4}$$

where t and t^* are a certain value chosen to penalize the additive and dominant effects, respectively. Thus, parameters in equation (2.3) are estimated by the penalized least squares

$$\frac{1}{2}||\tilde{\mathbf{y}} - \boldsymbol{\mu} - X\boldsymbol{\alpha} - Z\boldsymbol{\beta} - \xi\mathbf{a} - \zeta\mathbf{d}||^2 + \lambda \sum_{j=1}^{p} |a_j| + \lambda^* \sum_{j=1}^{p} |d_j|, \tag{2.5}$$

where $\tilde{\mathbf{y}} = (\tilde{y}_1, \cdots, \tilde{y}_n)^T$, $\boldsymbol{\mu} = (\mu, \cdots, \mu)^T$, $\xi = (\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_n)^T$, $\zeta = (\boldsymbol{\zeta}_1, \cdots, \boldsymbol{\zeta}_n)^T$, $X = (X_1, \cdots, X_n)^T$, $Z = (Z_1, \cdots, Z_n)^T$, and λ and λ^* are tuning parameters or lasso parameters that control the degrees of shrinkage in the estimate of the genetic effects.

2.2.3 Bayesian Hierarchical Representation

Noting the form of the L_1-penalty term in (2.5), Tibshirani (1996) suggested that lasso estimates can be interpreted as posterior mode estimates when the regression parameters have independent and identical Laplace (i.e., double-exponential) priors. Therefore, when lasso penalties are imposed on the additive and dominant effects of SNPs, the conditional prior for a_j is a Laplace distribution with the scale parameter σ/λ:

$$\pi(\mathbf{a}|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|a_j|/\sqrt{\sigma^2}}, \tag{2.6}$$

Similarly, the conditional Laplace prior for d_j is

$$\pi(\mathbf{d}|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda^*}{2\sqrt{\sigma^2}} e^{-\lambda^*|d_j|/\sqrt{\sigma^2}}. \tag{2.7}$$

Since the Laplace distribution can be represented as a scale mixture of a normal distribution with an exponential distribution (Andrews and Mallows, 1974), we have the following hierarchical representation of the penalized regression model:

$$\begin{aligned}
&\tilde{\mathbf{y}}|\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{a}, \mathbf{d}, \sigma^2 \sim N_n(\boldsymbol{\mu} + X\boldsymbol{\alpha} + Z\boldsymbol{\beta} + \xi\mathbf{a} + \zeta\mathbf{d}, \sigma^2 I_n),\\
&\boldsymbol{\alpha} \sim N_{d_1}(0, \Sigma_\alpha),\\
&\boldsymbol{\beta} \sim N_{d_2}(0, \Sigma_\beta),\\
&\mathbf{a}|\sigma^2, \tau_1^2, \cdots, \tau_p^2 \sim N_p(0, \sigma^2 \mathrm{diag}(\tau_1^2, \cdots, \tau_p^2)),\\
&\tau_1^2, \cdots, \tau_p^2|\lambda \sim \prod_{j=1}^{p} \exp(\frac{\lambda^2}{2}),\\
&\mathbf{d}|\sigma^2, \tau_1^{*2}, \cdots, \tau_p^{*2} \sim N_p(0, \sigma^2 \mathrm{diag}(\tau_1^{*2}, \cdots, \tau_p^{*2})),\\
&\tau_1^{*2}, \cdots, \tau_p^{*2}|\lambda^* \sim \prod_{j=1}^{p} \exp(\frac{\lambda^{*2}}{2}),\\
&\sigma^2 \sim \pi(\sigma^2),\\
&\sigma^2, \tau_1^2, \cdots, \tau_p^2, \tau_1^{*2}, \cdots, \tau_p^{*2} > 0.
\end{aligned}$$

After integrating out $\tau_1^2, \cdots, \tau_p^2$ and $\tau_1^{*2}, \cdots, \tau_p^{*2}$, the conditional priors on $\mathbf{a}$ and $\mathbf{d}$ have the desired forms (2.6) and (2.7), respectively. We assign conjugate normal priors to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ because they are low-dimensional and not the parameters of interest. Finally, since the data are usually sufficient to estimate μ and σ, we can use an independent, flat prior $\pi(\mu) = 1$ for μ and a noninformative scale-invariant prior $\pi(\sigma^2) = 1/\sigma^2$ for σ^2.

The tuning parameters of the ordinary lasso can be prespecified by cross-validation, generalized cross-validation, or the idea based on Stein's unbiased risk estimate. However, in the Bayesian lasso, λ and λ^* can be estimated along with other parameters by assigning appropriate hyperpriors to them. This procedure avoids the choice of lasso parameters and allows us to determine the amount of

shrinkage from the data. In particular, we consider the conjugate gamma priors on $\lambda^2/2$ and $\lambda^{*2}/2$,

$$\pi\left(\frac{\lambda^2}{2}\right) \sim \text{Gamma}(a, b),$$

$$\pi\left(\frac{\lambda^{*2}}{2}\right) \sim \text{Gamma}(a^*, b^*).$$

where a, b, a^*, and b^* are small values so that the priors are essentially noninformative. With this specification, lasso parameters can be treated similar to the other parameters and estimated by the Gibbs sampler.

2.3 Posterior computation and interpretation

2.3.1 MCMC Algorithm

We estimate the parameters by sampling from their conditional posterior distributions through the MCMC algorithm. The joint posterior distribution can be expressed as:

$$\begin{aligned} & \pi(\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{a}, \tau_1^2, \cdots, \tau_p^2, \lambda, \mathbf{d}, \tau_1^{*2}, \cdots, \tau_p^{*2}, \lambda^*, \sigma^2 | \tilde{\mathbf{y}}) \\ \propto & \quad \prod_{i=1}^n \pi(\tilde{y}_i|\cdot)\pi(\mu)\pi(\sigma^2)\pi_(\boldsymbol{\alpha})\pi_(\boldsymbol{\beta}) \\ & \prod_{j=1}^p \pi(a_j|\tau_j^2)\pi(\tau_j^2|\lambda)\pi(\lambda)\pi(d_j|\tau_j^{*2})\pi(\tau_j^{*2}|\lambda^*)\pi(\lambda^*) \end{aligned}$$

Two-level hierarchical modeling allows us to easily derive the conditional posterior distributions of parameters and hyperparameters, from which the Gibbs sampler draws posterior samples. Conditional on the parameters of additive effects, dominant effects and $(\tau_1^2, \cdots, \tau_p^2, \tau_1^{*2}, \cdots, \tau_p^{*2})$, the model is the standard linear regression and, thus, the conditional posterior distributions of $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2)$ are

$$\boldsymbol{\alpha}|\cdot \sim N_{d_1}\left(\Sigma'\left(\frac{\sum_{i=1}^n X_i(\tilde{y}_i - \mu - \mathbf{Z}_i^T\boldsymbol{\beta} - \boldsymbol{\xi}_i^T\mathbf{a} - \boldsymbol{\zeta}_i^T\mathbf{d})}{\sigma^2}\right), \Sigma'\right),$$

$$\text{with } \Sigma' = \left(\frac{\sum_{i=1}^n X_i X_i^T}{\sigma^2} + \Sigma_\alpha^{-1}\right)^{-1},$$

$$\boldsymbol{\beta}|\cdot \sim N_{d_2}\left(\Sigma''\left(\frac{\sum_{i=1}^{n} Z_i(\tilde{y}_i - \mu - \mathbf{X}_i^T\boldsymbol{\alpha} - \boldsymbol{\xi}_i^T\mathbf{a} - \boldsymbol{\zeta}_i^T\mathbf{d})}{\sigma^2}\right), \Sigma''\right),$$
$$\text{with } \Sigma'' = \left(\frac{\sum_{i=1}^{n} Z_i Z_i^T}{\sigma^2} + \Sigma_\beta^{-1}\right)^{-1},$$

$$\sigma^2|\cdot \sim Inv-\chi^2\left(n, \frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - \mu - \mathbf{X}_i^T\boldsymbol{\alpha} - \mathbf{Z}_i^T\boldsymbol{\beta} - \boldsymbol{\xi}_i^T\mathbf{a} - \boldsymbol{\zeta}_i^T\mathbf{d})^2\right).$$

Conditional on the parameters $(\tau_1^2, \cdots, \tau_p^2, \tau_1^{*2}, \cdots, \tau_p^{*2}, \boldsymbol{\alpha}, \boldsymbol{\beta})$, the model becomes the weighted linear regression, and thus the conditional posterior distributions of $(\mathbf{a}, \mathbf{d})$ are

$$\mathbf{a}|\cdot \sim N\left(A_a^{-1}\xi(\tilde{y}_i - \mu - \mathbf{X}_i^T\boldsymbol{\alpha} - \mathbf{Z}_i^T\boldsymbol{\beta} - \boldsymbol{\zeta}_i^T\mathbf{d}), \sigma^2 A_a^{-1}\right),$$
$$\text{with } A_a^{-1} = \left(\xi\xi^T + \text{diag}(\tau_1^2, \cdots, \tau_p^2)^{-1}\right)^{-1},$$

$$\mathbf{d}|\cdot \sim N\left(A_d^{-1}\zeta(\tilde{y}_i - \mu - \mathbf{X}_i^T\boldsymbol{\alpha} - \mathbf{Z}_i^T\boldsymbol{\beta} - \boldsymbol{\xi}_i^T\mathbf{a}), \sigma^2 A_d^{-1}\right),$$
$$\text{with } A_d^{-1} = \left(\zeta\zeta^T + \text{diag}(\tau_1^{*2}, \cdots, \tau_p^{*2})^{-1}\right)^{-1}.$$

Moreover, the full conditional for $\tau_1^2, \cdots, \tau_p^2, \tau_1^{*2}, \cdots, \tau_p^{*2}$ are conditionally independent, with

$$\frac{1}{\tau_j^2}|\cdot \sim \text{Inverse-Gaussian}\left(\sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}, \lambda^2\right), \quad j = 1, \cdots, p,$$

and

$$\frac{1}{\tau_j^{*2}}|\cdot \sim \text{Inverse-Gaussian}\left(\sqrt{\frac{\lambda^{*2}\sigma^2}{\beta_j^2}}, \lambda^{*2}\right), \quad j = 1, \cdots, p.$$

Finally, with the conjugate priors Gamma(a, b) and Gamma(a^*, b^*), the conditional posterior distributions of the hyperparameters are gammas

$$\lambda^2|\cdot \sim \text{Gamma}\left(p + a, \sum_{j=1}^{p}\frac{\tau_j^2}{2} + b\right),$$

and

$$\lambda^{*2}|\cdot \sim \text{Gamma}\left(p + a^*, \sum_{j=1}^{p} \frac{\tau_j^{*2}}{2} + b^*\right).$$

An efficient Gibbs sampler based on these full conditionals proceeds to draw posterior samples from each full conditional posterior distribution, given the current values of all other parameters and the observed data. This process continues until all chains converge. We use the potential scale reduction factor $\hat{R}$ to access the convergence (Gelman and Rubin, 1992). Once $\hat{R} < 1.1$ for all scalar estimands of interest, we continue to draw $15,000$ iterations to obtain samples from the joint posterior distribution.

2.3.2 Posterior interpretation

The proposed MCMC algorithm for our Bayesian lasso model can provide posterior median estimates of the additive effects and dominant effects of individual SNPs, while adjusting for the effects of all other SNPs and covariates. Furthermore, using the posterior samples of **a**, **d**, and the observed genotypes, we can calculate the proportion of the phenotypic variance explained by a particular SNP, i.e., heritability, by

$$h_j^2 = \frac{2\hat{p}_1\hat{p}_0(\hat{a}_j + (\hat{p}_1 - \hat{p}_0)\hat{d}_j)^2 + 4\hat{p}_1^2\hat{p}_0^2\hat{d}_j^2}{\text{var}(\tilde{y})}, \quad j = 1, \cdots, p, \tag{2.8}$$

where $\hat{p}_1$ is the estimated allele frequency for A, and $\hat{p}_0$ is the estimated allele frequency for a, $\hat{a}_j$ is the median estimate of the additive effect for SNP j, and $\hat{d}_j$ is the median estimate of the dominant effect for SNP j. Since heritability estimates are unitless, they could guide variable selection and identify SNPs that have relatively large effects on the phenotype.

2.4 Results

2.4.1 Real Data Analysis

We used the newly developed model to analyze a real GWAS data set from the Framingham Heart Study (FHS), a cardiovascular study based in Framingham,

Massachusetts, supported by the National Heart, Lung, and Blood Institute, in collaboration with Boston University (Dawber et al. 1951). Recently, 550,000 SNPs have been genotyped for the entire Framingham cohort (Jaquish 2007), from which 418 males and 559 females were chosen for our data analysis. These subjects were measured for body mass index (BMI) at different ages from 29 and 61 years. As is standard practice, SNPs with minor allele frequency $< 10\%$ were excluded from data analysis. The numbers and percentages of non-rare allele SNPs vary among different chromosomes and ranges from 4,417 to 28,771 and from 64% to 72%, respectively.

In principle, our approach can handle an extremely large number of SNPs at the same time. To save our computing time, however, we use those SNPs that cannot be neglected according to a simple single SNP analysis. We chose the phenotypic data of BMI in a middle measure age of each subject for a single SNP analysis, separately for males and females. Figure 2.1 gives $-log_{10}$ p-values for each SNP in the two sexes from which 1837 SNPs with a $-log_{10}$ p-value of > 3.5 in at least one sex were selected for Bayesian lassso analysis. Before this analysis, we imputed missing genotypes for a small proportion of SNPs (5.16%) according to the distribution of genotypes in the population. A preconditional analysis with $m = 3$ and $\theta = 0.426$ was used to mitigate observational noise, leading to the preconditioned phenotypes. Like original measures, the preconditioned BMI also displays a normal distribution (Figure 2.2), which meets the normality assumption required by the new approach.

By treating the sex as a discrete covariate and age as a continuous covariate, we imposed lasso penalties on the additive effects $a_1, \cdots, a_p$ and dominant effects $d_1, \cdots, d_p$ to identify those SNPs with notable effects on BMI. We employ the proposed MCMC algorithms to estimate all parameters and implement variable selection, where $\Sigma_\alpha = 1$, $\Sigma_\beta = 1$, and all parameters in the conjugate gamma hyperpriors are 0.1. In unreported tests, we find that the posteriors are not sensitive to these prior specifications, as long as a and b are small values so that the priors are relatively flat (Park and Casella, 2008; Yi and Xu, 2008). Figure 2.3 plots the estimated additive and dominant effects of each SNP after adjusting for the effects of other SNPs and covariates. The heritability explained by each SNP is shown in Figure 2.4. The Bayesian hierarchical model automatically shrinks small coeffi-

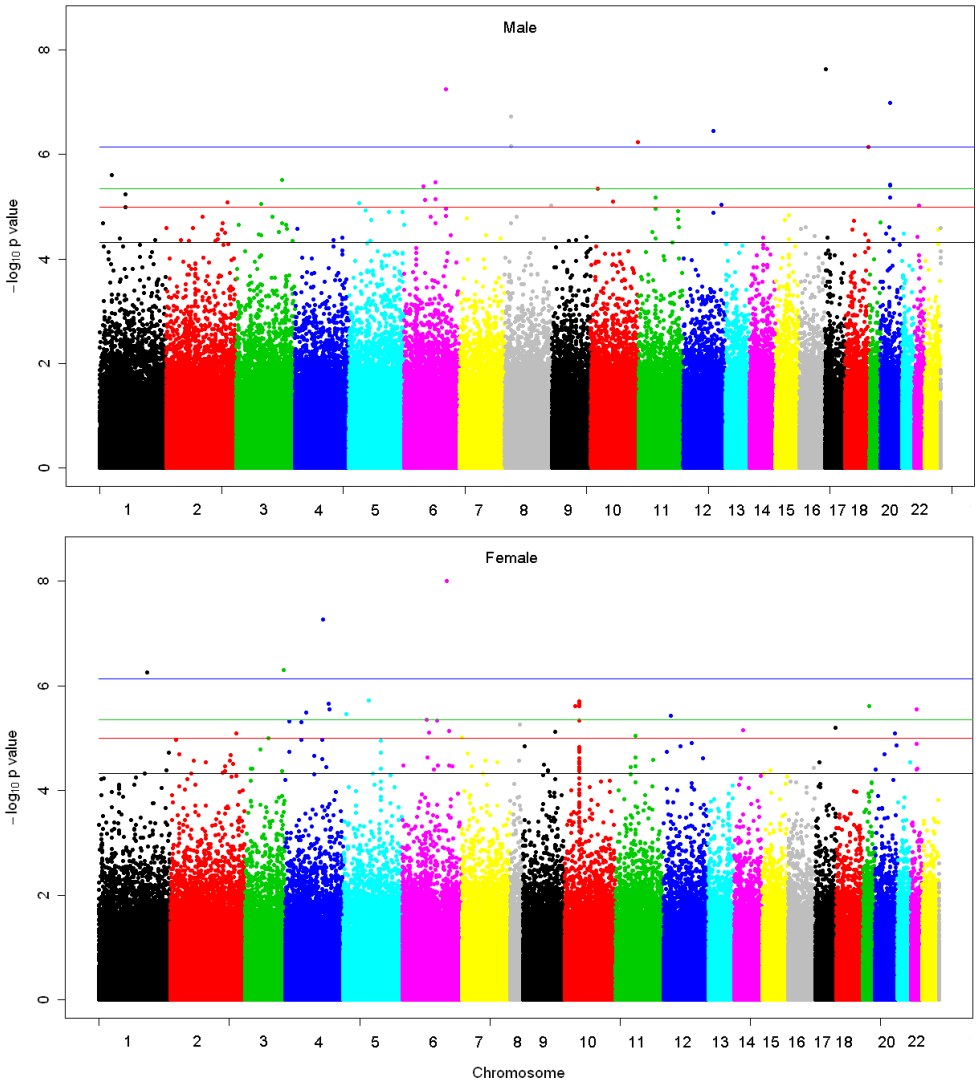

Figure 2.1. Single SNP analysis for the Framingham genome-wide association study

cients to zero, and hence the posterior estimates of **a**, **d** and h_j^2 can guide variable selection. We claim that a genetic effect is significant if its 95% posterior credible interval does not contain zero. Alternatively, Hoti and Sillanpaa (2006) suggested to preset a threshold value, c, such that one SNP is included into the final model if the heritability explained by this SNP is greater than c. We usually report the SNPs with high heritabilities, and in general this threshold can be chosen on more subjective grounds.

Table 2.2 tabulates the names and positions of SNPs with the heritability (h_j^2)

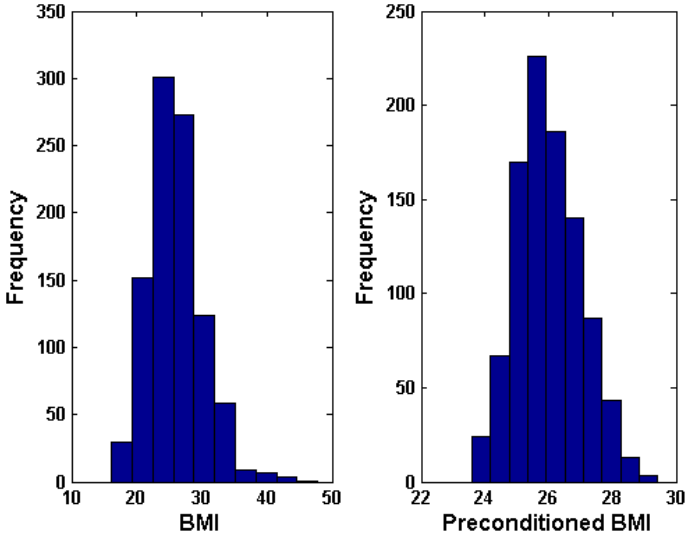

Figure 2.2. The histograms of original and preconditioned BMI

greater than 0.5, as well as the estimated additive effects and heritabilities. We do not report the estimated dominant effects since they are relatively low in this example. The Bayesian lasso tends to shrink small effects of genes into zero. Assuming that $a = d = 0.4$ for a SNP with allele frequencies of 0.5 in a population, the additive and dominant variances explained by this SNP is $\frac{1}{2}a^2 = 0.08$ and $\frac{1}{4}d^2 = 0.04$, respectively. Thus, there is a possibility that the dominant effects are shrunk to a greater extent than the additive effects if they are of similar size. This may partly explain why the dominant effects estimated for significant SNPs are much smaller than the additive effects.

The amount of shrinkage in the estimates of additive and dominant effects are quantified by two hyperparameters λ and λ^* determined from the data. The posterior medians for λ and λ^* are 54.474 and 54.523, respectively, with the 95% posterior intervals being [53.325, 55.626] and [53.359, 55.678], respectively. These suggest that the tuning parameters for the additive and dominant effects can be estimated precisely.

Since five significant SNPs are selected from chromosome 1, and four from chromosome 10, we will further examine the correlations among the significant SNPs from the same chromosome, as suggested by a referee. The correlation

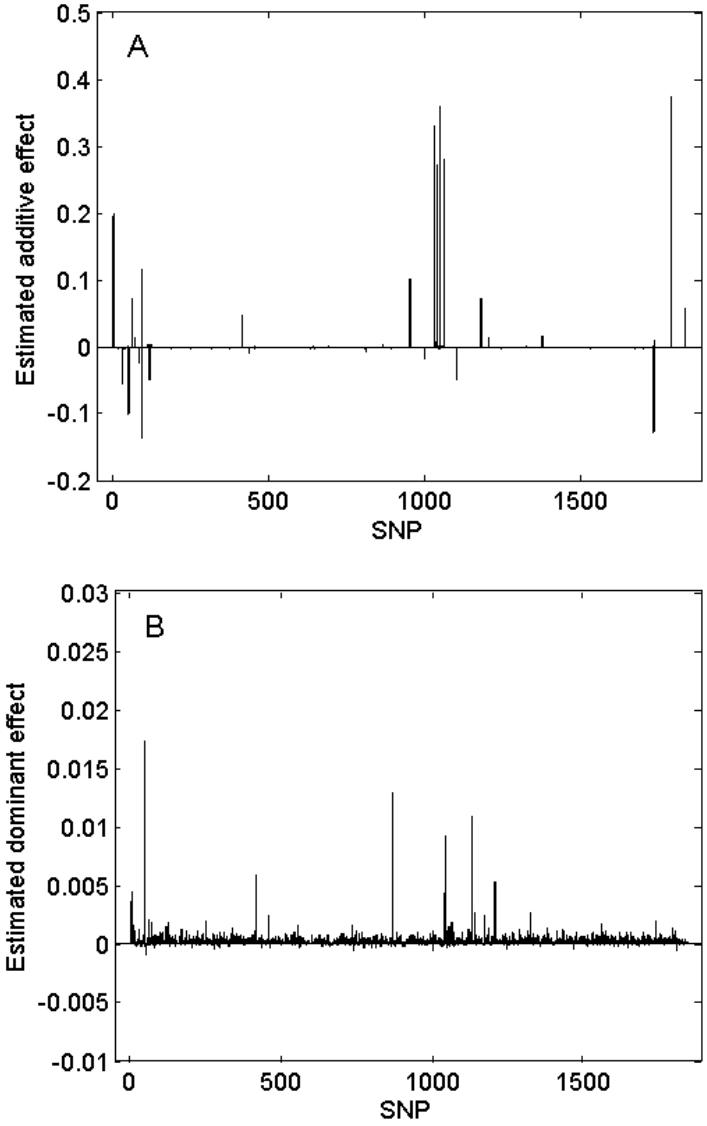

Figure 2.3. Estimated additive (A) and dominant effects (B) of each SNP on BMI by the Bayesian lasso.

matrix of five significant SNPs from chromosome 1 is given by

$$\begin{pmatrix} 1.00 & 0.85^* & 0.86^* & -0.01 & 0.01 \\ 0.85^* & 1.00 & 0.78^* & -0.01 & 0.02 \\ 0.86^* & 0.78^* & 1.00 & -0.04 & 0.02 \\ -0.01 & -0.01 & -0.04 & 1.00 & -0.84^* \\ 0.01 & 0.02 & 0.02 & -0.84^* & 1.00 \end{pmatrix},$$

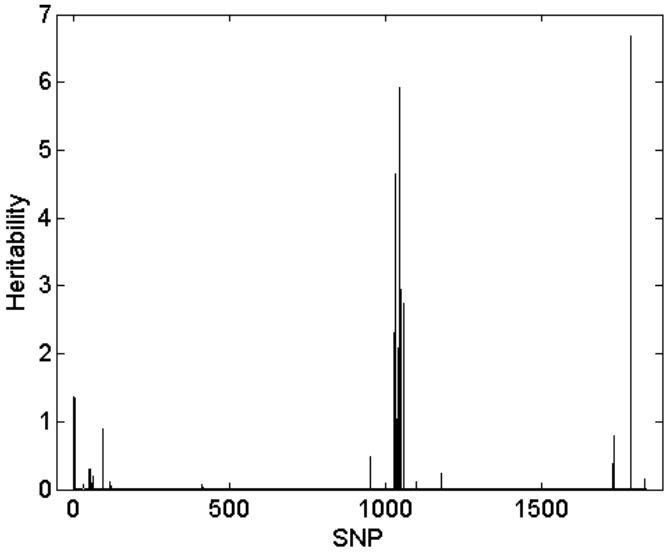

Figure 2.4. Estimated heritability for BMI explained by each SNP.

Table 2.2. The estimates of additive and dominant effects triggered by each significant SNP.

Chr	Name	Position	Additive	Heritability (%)
1	ss66185476	8445140	0.15	0.74
1	ss66374301	8451728	0.19	1.36
1	ss66295856	8578082	0.20	1.35
1	ss66516012	198313489	-0.13	0.89
1	ss66364251	198321700	0.12	0.66
10	ss66311679	32719838	0.33	4.65
10	ss66293192	32903593	0.27	2.08
10	ss66303064	32995111	0.36	5.93
10	ss66128868	33407810	0.28	2.75
20	ss66171460	22580931	-0.13	0.78
22	ss66055592	23420006	0.33	5.13
22	ss66164329	23420370	0.37	6.70

where star denotes significant correlations at the significance level 1%. Clearly, these SNPs can be classified into two groups, and within each group, SNPs are highly correlated. The correlation matrix of four significant SNPs from chromo-

some 10

$$\begin{pmatrix} 1.00 & 0.53^* & 0.48^* & 0.30^* \\ 0.53^* & 1.00 & 0.52^* & 0.53^* \\ 0.48^* & 0.52^* & 1.00 & 0.45^* \\ 0.30^* & 0.53^* & 0.45^* & 1.00 \end{pmatrix}$$

also suggested that these SNPs are closely linked to each other.

2.4.2 Computer Simulation

The new approach is investigated through simulation studies. We generate data according to the model (2.3) with $\mu = 0$, $\sigma^2 = 10$ and $n = 500$. For ease of simulation, ξ_{ij} is derived from u_{ij}, where each u_{ij} has a standard normal distribution marginally, and $\rho = cov(u_{ij}, u_{ik}) = 0.1$. Then, to mimic a SNP with equal allele frequencies, we set

$$\xi_{ij} = \begin{cases} 1, & u_{ij} > c \\ 0, & -c \leq u_{ij} \leq c \\ -1, & u_{ij} < -c, \end{cases}$$

where $-c$ is the first quartile of a standard normal distribution. Finally, ζ_{ij} is derived from ξ_{ij}. We assume that there are 1000 SNPs from which 20 are significant for a phenotypic trait. The positions and additive and dominant effects of individuals are given in Table 2.3. It is assumed that the trait is measured at a subject-specific age, following the data structure of the Framingham Heart Study.

Figure 2.5 gives the estimated additive and dominant genetic effects of different SNPs over 50 simulations, and Figure 2.6 plots the heritability explained by each SNP. It is clear that lasso penalties shrink small genetic effects to zeros, resulting in sparse solutions of the regression coefficients. In general, the 20 assumed SNPs can be well identified and their additive and dominant effects well estimated. Also, two hyperparameters λ and λ^* whose influence the degree of shrinkage can be well estimated. In Figure 2.7, the histograms of these two hyperparameters are shown.

Then, we carry out another simulation study to compare the performance of preconditioned Bayesian lasso, Bayesian lasso without preconditioning, and the traditional single SNP analysis. Without loss of generality, only the additive model is considered. Specifically, we generate data on $n = 200$ and $p = 500$ or 1000

Table 2.3. Genetic effects of 20 assumed SNPs for data simulation.

Position	Additive	Position	Dominant
100	1.2	50	1.2
200	1.2	150	1.2
300	1.2	250	1.2
400	0.8	350	0.8
500	0.8	450	0.8
600	0.8	550	0.8
700	0.4	650	0.4
800	0.4	750	0.4
850	1.2	850	1.2
900	0.8	900	1.2
950	1.2	950	0.8
1000	0.8	1000	0.8

according to the model (2.3), with $\mu = 0$, $\sigma^2 = 10$, $\rho = 0.1$, $a_j = 1$ for $1 \leq j \leq 20$ and $a_j = 0$ for $j > 20$.

We apply three methods to the 100 simulated datasets: single SNP analysis (SSA), standard Bayesian lasso (B-lasso), and the Bayesian lasso applied to the preconditioned response from supervised principal components (PB-lasso). In single SNP analysis, we reject the null hypothesis that the genetic effect of an individual SNP equals to zero at the significance level of 5% with the FDR adjustment. For the Bayesian lasso and preconditioned Bayesian lasso, we reject the null hypothesis based on 95% Bayesian credible intervals. To ameliorate the bias of the parameter estimates introduced by lasso penalties, we always refit the linear regression model without the penalty term using only those SNPs selected by the model selection procedure.

For each estimated genetic effect obtained from each method, we calculate the average bias and empirical standard error over 100 simulations. Since the first 20 genetic effects are nonzeros with the same true value, in Table 2.4 we report the average values over the first 20 SNPs and over the rest of the SNPs separately. The standard error of each average is in parentheses. In the column labeled "Aver. Nonzeros", we present the average number of nonzero coefficients correctly identified to be nonzero, or the average number of zero coefficients incorrectly estimated to be nonzero in 100 replications. In the column "Proportion of Correct-fit", we

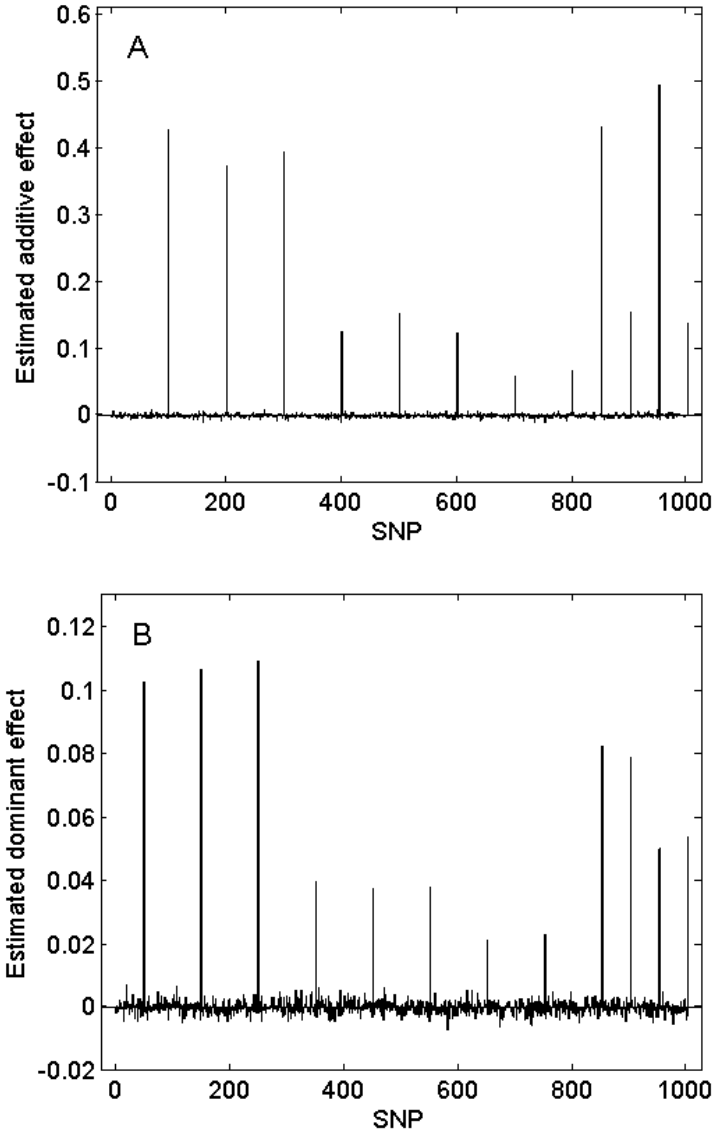

Figure 2.5. Estimated additive (A) and dominant effects (B) based on 50 simulations.

present the proportion of replications that the exact true model was identified.

As can be seen from Table 2.4, the single SNP analysis tend to overestimate the genetic effect, since when we test a SNP for the association with the phenotype, we assume the genetic variation is solely due to this particular SNP, and ignore the effects from all other SNPs. Therefore, in terms of parameter estimates, model selection methods that simultaneously estimate the genetic effects associated with

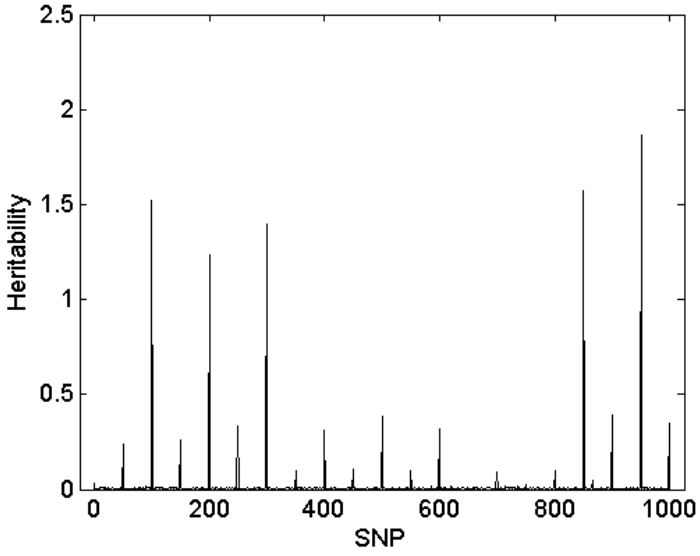

Figure 2.6. Estimated heritability explained by each SNP based on 50 simulations.

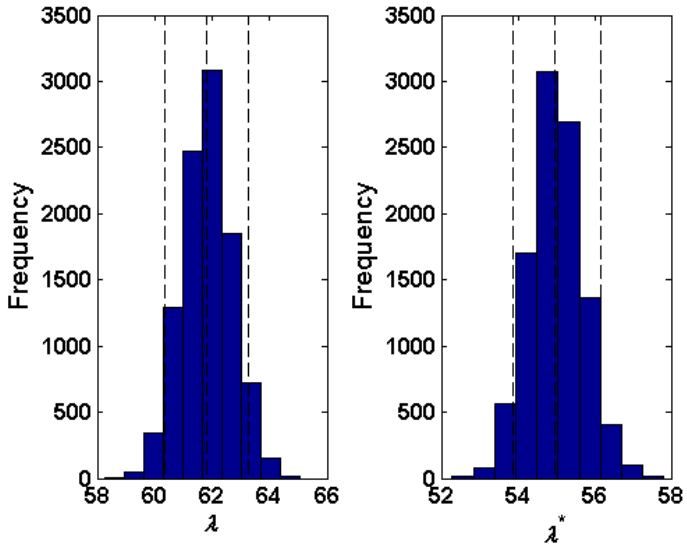

Figure 2.7. Histograms of two hyperparameters in the first simulation study.

all SNPs outperform the traditional single SNP analysis. In terms of variable selection, although preconditioned Bayesian lasso has a slightly higher false positive rate due to the preconditioning step, it greatly improves the probability of correctly identifying regression coefficients with nonzero effects. Moreover, as the

number of SNPs gets larger, single SNP analysis detected fewer important SNPs, since this method subjects to severe multiple comparison adjustment. However, preconditioned Bayesian lasso is still able to identify nonzero coefficients and zero coefficients correctly in almost every simulation. Table 2.5 displays the simulation results when $\rho = 0.5$, which are consistent with our findings.

Table 2.4. Simulation results for three methods based on 100 simulations when $\rho = 0.1$.

Method	Bias	Empirical SE	Aver. Nonzeros	Proportion of Correct-fit
$n = 200, p = 500$, β_1 - β_{20}				
SSA	4.17	1.99	16.62	0.08
	(0.21)	(0.19)	(1.51)	
B-lasso	0.07	0.34	18.28	0.18
	(0.04)	(0.02)	(1.36)	
PB-lasso	0.00	0.35	19.68	0.63
	(0.03)	(0.03)	(0.65)	
$n = 200, p = 500$, β_{21} - β_{500}				
SSA	0.44	0.43	0.78	0.46
	(0.05)	(0.07)	(1.09)	
B-lasso	0.00	0.04	0.95	0.42
	(0.01)	(0.01)	(0.94)	
PB-lasso	0.00	0.03	1.25	0.30
	(0.01)	(0.04)	(0.73)	
$n = 200, p = 1000$, β_1 - β_{20}				
SSA	4.13	1.96	15.71	0.04
	(0.18)	(0.17)	(2.73)	
B-lasso	0.36	0.38	17.11	0.08
	(0.06)	(0.07)	(2.69)	
PB-lasso	0.00	0.36	19.24	0.51
	(0.04)	(0.03)	(1.81)	
$n = 200, p = 1000$, β_{21} - β_{1000}				
SSA	0.43	0.43	0.42	0.69
	(0.04)	(0.06)	(0.84)	
B-lasso	0.00	0.02	0.33	0.76
	(0.01)	(0.01)	(0.18)	
PB-lasso	0.00	0.02	1.17	0.56
	(0.00)	(0.01)	(1.38)	

Table 2.5. Simulation results for three methods based on 100 simulations when $\rho = 0.5$.

Method	Bias	Empirical SE	Aver. Nonzeros	Proportion of Correct-fit
$n = 200, p = 500$, β_1 - β_{20}				
SSA	21.82	3.70	19.37	0.38
	(0.32)	(0.25)	(0.49)	
B-lasso	0.01	0.41	19.73	0.79
	(0.04)	(0.03)	(0.55)	
PB-lasso	0.00	0.42	19.99	0.99
	(0.03)	(0.03)	(0.11)	
$n = 200, p = 500$, β_{21} - β_{500}				
SSA	0.46	0.44	0.05	0.95
	(0.05)	(0.07)	(0.22)	
B-lasso	0.00	0.03	0.16	0.94
	(0.00)	(0.01)	(0.24)	
PB-lasso	0.00	0.01	0.29	0.78
	(0.00)	(0.01)	(0.38)	
$n = 200, p = 1000$, β_1 - β_{20}				
SSA	21.94	3.88	13.32	0.33
	(0.37)	(0.32)	(0.47)	
B-lasso	0.03	0.40	19.50	0.58
	(0.04)	(0.03)	(0.64)	
PB-lasso	0.00	0.42	19.99	0.98
	(0.03)	(0.03)	(0.11)	
$n = 200, p = 1000$, β_{21} - β_{1000}				
SSA	0.43	0.43	0.03	0.97
	(0.05)	(0.06)	(0.11)	
B-lasso	0.00	0.02	0.00	0.99
	(0.01)	(0.01)	(0.00)	
PB-lasso	0.00	0.00	0.10	0.92
	(0.00)	(0.01)	(0.20)	

Chapter 3

The Independence Screening and Bayesian SCAD for Genome-wide Association Studies

3.1 Introduction

Variable selection methods unified in the penalized linear regression framework are crucial for high dimensional data analysis. They provide not only computational efficiency but statistical accuracy. As explained in Chapter 1, many popular variable selection procedures are special cases when penalty functions are specified in different ways. In particular, Lasso (Tibishrani, 1996) considers the L_1 penalty, ridge regression takes the L_2 penalty and elastic net (Zou and Hastie, 2005) makes a compromise between lasso regression and ridge regression, in hopes of retaining advantages from each side.

However, although the convexity of penalty functions of these procedures reduces the computational cost, each of them has some undesirable theoretical properties from different aspects. For example, by imposing the L_2 penalty, ridge regression does a proportional shrink of all regression coefficients and thus fails to produce a sparse model. On the other hand, although the L_1 penalty in lasso regression yields sparse solutions and the model sparsity varies according to the size of tuning parameter λ, the estimates can be biased for large regression coefficients

since large penalties are imposed on these coefficients. Other traditional model selection procedures such as best subset selection drop all small predictors, which is equivalent to imposing the hard thresholding penalty function. It increases the computational expense, and the solution is not robust even when the number of predictors is not large.

Fan and Li (2001) provided deep insights into how penalty function should be chosen, and argued that the following three statistical properties should be considered when the penalized least square estimators are evaluated: sparsity, unbiasedness and continuity. First, the goal of variable selection is guaranteed by the sparsity of solutions. Second, we may wish to obtain unbiased estimates for large regression coefficients. Finally, the variable selection procedure should be carried on in a continuous way, in the sense that if the sample we observed changes slightly, the selection of significant predictors should not be affected too much. Based on these principles, Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) estimator that enjoys all these desirable properties. Moreover, it performs as well as the oracle procedure in terms of selecting the correct model and estimating the true nonzero regression coefficients.

Although SCAD penalized regression procedure has promising theoretical properties and in general outperforms the other prevailing variable selection procedures, its penalty function has to be singular at the origin and nonconvex over $(0, \infty)$. The singularity and nonconvexity of SCAD penalty function challenge the statistical inference since our goal is to minimize a nondifferentiable nonconcave function.

The first algorithm was proposed by Fan and Li (2001), which suggested the local quadratic approximation (LQA) of penalty functions such that Newton - Raphson algorithm can be directly used to optimize the penalized least squares. However, if a predictor is deleted from the model at any iteration in the LQA algorithm, it cannot be included again in a later step. Hunter and Li (2005) addressed this drawback by adopting a perturbed version of LQA, which alleviates the aforementioned drawback, but introduces another tuning parameter. To address these issues, Zou and Li (2008) proposed a new unified algorithm based on the local linear approximation (LLA) of SCAD penalty functions, where one-step LLA estimator from the LLA algorithm are used as the final estimates. Compared with the LQA estimator, the proposed LLA estimator naturally adopts a sparse

representation, and enjoys the oracle properties provided that the initial estimators are good enough.

When the number of predictors p is much larger than the number of observations n, two stage procedures for variable selection are recommended. In the first step, data is preprocessed such that noises are separated and removed as much as possible. In the second step, highly regularized approaches, such as penalized regression models, are considered to identify nonzero coefficients, enhance model predictability, and avoid over-fitting (Hastie el al., 2009).

In this chapter, we develop a two-stage variable selection procedure including the Bayesian SCAD for high dimensional problems, and apply this procedure to genome-wide association studies analysis. For that purpose, we first employ the independence screening technique to reduce the dimensionality of feature space. Then we consider the SCAD penalized least squares for the variable selection of predictors retained in the reduced feature space. We approximate the SCAD penalty function based on the local linear approximation, and derive the Markov chain Monte Carlo (MCMC) algorithm to estimate all model parameters. Our model is efficient and accurate to select important predictors and estimate their effects at the same time. The statistical properties of the model were investigated through simulation studies.

3.2 Sure Independence Screening

Aimed at several concerns when the standard variable selection techniques are directly applied to high or ultrahigh dimensional data set, Fan and Lv (2008) proposed sure independence screening (SIS) for high dimensional data analysis. This variable screening technique reduces the dimensionality of the problem from a very large scale $\tilde{p}$ to a relatively large scale p before the implementation of an existing variable selection method. Specifically, SIS ranks the importance of predictors according to their marginal utility measures and retains those predictors whose marginal correlations with the response variable are strong enough. If the linear relationship is assumed, the correlation coefficient between response and each predictor are considered as the marginal utility measure.

Under some technical conditions, it can be shown that sure independence

screening enjoys the sure screening property. Due to the sure screening property, the reduced p-dimensional model is capable of retaining all the important variables with asymptotic probability one. Moreover, since p is usually much smaller than $\tilde{p}$, the following variable selection procedure such as L_1 regularization method or SCAD regularization method could be performed in a more efficient manner.

Recall that for preconditioning based variable selection procedures, we have to prespecify the dimensionality of the reduced design matrix and the number of principle components that we use to generate the preconditioned response variable. Therefore, compared with preconditioning, SIS has less tuning parameters to be determined. Moreover, after SIS we only need to deal with p dimensional problem instead of the original $\tilde{p}$ dimensional problem. However, preconditioning fails to reduce the feature space and thus the following variable selection methods still face the curse of dimensionality.

Sure independence screening first finds the estimates the standardized regression coefficients $\hat{\beta}_1, \cdots, \hat{\beta}_{\tilde{p}}$ for the univariate linear regression model

$$y = X_j \beta_j + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I_n), \quad j = 1, \cdots, \tilde{p}, \tag{3.1}$$

where $\mathbf{y}$ is the vector of response variable, X_j is the j-th column of the $n \times \tilde{p}$ design matrix X, and β_j is the regression coefficient for the j-th predictor. Then for a given p, we keep those predictors whose estimated regression coefficients are among the first p largest of all. In practice, p is usually set to be $\frac{n}{log(n)}$ as suggested by Fan and Lv (2008). After this screening step, we are ready to apply a standard variable selection method to select important predictors from these p candidate predictors.

3.3 SCAD Penalized Regression

In genome-wide association studies, a number of covariates, which are either discrete or continuous, may be measured for each subject. In order to estimate genetic effects as well as the effects of these covariates, a GWAS model that takes into account the effects of important covariates will be considered. Therefore, we describe

the phenotypical value y of a continuous trait for subject i as

$$y_i = \mu + \mathbf{X}_i^T\boldsymbol{\alpha} + \mathbf{Z}_i^T\boldsymbol{\beta} + \boldsymbol{\xi}_i^T\mathbf{a} + \boldsymbol{\zeta}_i^T\mathbf{d} + \epsilon_i, \quad i = 1, \cdots, n, \tag{3.2}$$

where μ is the overall mean, $\mathbf{X}_i$ is the d_1-dimensional vector of discrete covariates for subject i, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_{d_1})^T$ is the vector of regression coefficients for discrete covariates, $\mathbf{Z}_i$ is the d_2-dimensional vector of continuous covariates for subject i, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_{d_2})^T$ is the vector of regression coefficients for continuous covariates, $\mathbf{a} = (a_1, \cdots, a_p)^T$ and $\mathbf{d} = (d_1, \cdots, d_p)^T$ are the p-dimensional vectors of the additive and dominant effects of SNPs, respectively, $\boldsymbol{\xi}_i$ and $\boldsymbol{\zeta}_i$ are the indicator vectors of the additive and dominant effects of SNPs for subject i, and ϵ_i is the residual error assumed to follow a $N(0, \sigma^2)$ distribution. The j-th elements of $\boldsymbol{\xi}_i$ and $\boldsymbol{\zeta}_i$ are defined as

$$\xi_{ij} = \begin{cases} 1, & \text{if the genotype of SNP } j \text{ is } AA \\ 0, & \text{if the genotype of SNP } j \text{ is } Aa \\ -1, & \text{if the genotype of SNP } j \text{ is } aa, \end{cases}$$

$$\zeta_{ij} = \begin{cases} 1, & \text{if the genotype of SNP } j \text{ is } Aa \\ 0, & \text{if the genotype of SNP } j\text{is } AA \text{ or } aa. \end{cases}$$

In genome-wide association studies, $\tilde{p}$ is usually much larger than n. Although SIS greatly reduced the model dimensionality, penalized regressions are still preferred to identify SNPs from the whole genome with notable genetic effects and enhance the model predictive performance. Specifically, we put penalties on the sizes of additive effects and the dominant effects and minimize the penalized least squares

$$\frac{1}{2}||\tilde{\mathbf{y}} - \boldsymbol{\mu} - X\boldsymbol{\alpha} - Z\boldsymbol{\beta} - \xi\mathbf{a} - \zeta\mathbf{d}||^2 + \sum_{j=1}^{p} p_\lambda(|a_j|) + \sum_{j=1}^{p} p_{\lambda^*}(|d_j|), \tag{3.3}$$

where $\mathbf{y} = (y_1, \cdots, y_n)^T$, $\boldsymbol{\mu} = (\mu, \cdots, \mu)^T$, $\xi = (\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_n)^T$, $\zeta = (\boldsymbol{\zeta}_1, \cdots, \boldsymbol{\zeta}_n)^T$, $X = (X_1, \cdots, X_n)^T$, $Z = (Z_1, \cdots, Z_n)^T$ and λ and λ^* are tuning parameters in penalty functions that control the degrees of shrinkage in the estimate of the genetic effects.

Here we consider the smoothly clipped absolute deviation (SCAD) penalty function, which is a convex function defined by $p_\lambda(0) = 0$ and

$$p'_\lambda(|a_j|) = \lambda I(|a_j| \leq \lambda) + \frac{(s\lambda - |a_j|)^+}{(s-1)\lambda} I(|a_j| > \lambda), \tag{3.4}$$

$$p'_{\lambda^*}(|d_j|) = \lambda^* I(|d_j| \leq \lambda^*) + \frac{(s\lambda^* - |d_j|)^+}{(s-1)\lambda} I(|d_j| > \lambda^*) \tag{3.5}$$

for $a_j \neq 0$, $d_j \neq 0$ and some $s > 2$. The notation x^+ represents the positive part of x. That is, x^+ is x if $x > 0$ and zero otherwise. We fix $s = 3.7$, as recommended by Fan and Li (2001). This is a specially designed penalty function proposed by by Fan and Li (2001). It can be shown that SCAD penalized regression could produce sparse solutions where the estimates for large regression coefficients are unbiased. Under appropriate conditions, the SCAD penalized least squares estimator is consistent for variable selection and enjoys the oracle property, in the sense that the estimators of nonzero coefficients have the same asymptotic distributions as they would have if the true model were known in advance. Because of these desirable statistical properties, it is suitable for variable selection after the initial independence screening step.

3.4 Estimation

3.4.1 Local Linear Approximation

When we consider SCAD regularization method for selecting important SNPs in genome-wide association studies, our goal is to minimize the SCAD penalized least squares (3.3). However, since the penalty function p_λ is nonconvex, it is challenging to minimize the objective function through traditional optimization procedures. Fan and Li (2001) propose a unified and effective algorithm for optimizing nonconcave penalized likelihood. The algorithm locally approximates the objective function by a quadratic function, and hence is named as local quadratic approximation (LQA). Since LQA plays the same role as the E-step in the EM algorithm (Hunter and Li, 2005), LQA has similar behavior to EM algorithm and has the quadratic convergence rate.

Zou and Li (2008) argued that a better approximation can be achieved through local linear apporximation (LLA). Specifically, for a given initial value $\boldsymbol{a}^{(0)} = (a_1^{(0)}, \cdot, a_p^{(0)})^T$, the penalty function p_λ can be locally approximated by a linear function as

$$p_\lambda(|a_j|) \approx p_\lambda(|a_j^{(0)}|) + p'_\lambda(|a_j^{(0)}|)(|a_j| - |a_j^{(0)}|) \quad \text{for} \quad |a_j| \approx |a_j^{(0)}|. \tag{3.6}$$

Similarly, the penalty function p_{λ^*} can be locally approximated by

$$p_{\lambda^*}(|d_j|) \approx p_{\lambda^*}(|d_j^{(0)}|) + p'_{\lambda^*}(|d_j^{(0)}|)(|d_j| - |d_j^{(0)}|) \quad \text{for} \quad |d_j| \approx |d_j^{(0)}|. \tag{3.7}$$

Since LLA is the tightest convex majorant of the concave penalty function p_λ on $[0, \infty)$, it is better than LQA. Zou and Li (2008) showed that with LLA, the estimates of regression coefficients in SCAD penalized least squares (3.3) can be obtained by

$$(\boldsymbol{a}^{(1)^T}, \boldsymbol{d}^{(1)^T})^T = \arg\min\{\frac{1}{2}||y - Ey||^2 + n\sum_{j=1}^{p} p'_\lambda(|a_j^{(0)}|)|a_j| + n\sum_{j=1}^{p} p'_{\lambda^*}(|d_j^{(0)}|)|d_j|\}. \tag{3.8}$$

The SCAD estimators enjoy the oracle properties and naturally adopt a sparse representation. Moreover, (3.8) indicates that the nonconvex penalized least squares can be minimized based on L_1 penalized regression, which motivated us to incorporate Bayesian lasso algorithm to the SCAD regression, and formulate Bayesian SCAD algorithm.

3.4.2 Bayesian Implementation

In this section we will describe the estimation procedure for SCAD penalized linear regression in the Bayesian framework. For illustrative purpose, we will only focus on the additive genetic model, that is, dominant effects $d_j = 0$ for $j = 1, \cdots, p$. The extension to a full GWAS model including both additive effects and dominant effects is straightforward.

From a Bayesian perspective, all parameters are estimated by sampling from their conditional posterior distributions. Like Bayesian lasso, we introduce a set of new parameters $\tau_j, j = 1, \cdots, p$ which guarantees the close-forms of all pos-

terior distributions and efficient Gibbs samplers. Therefore, the joint posterior distribution can be expressed as:

$$\begin{aligned} & \pi(\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{a}, \tau_1^2, \cdots, \tau_p^2, \lambda, \sigma^2 | \mathbf{y}) \\ \propto & \prod_{i=1}^n \pi(y_i|\cdot)\pi(\mu)\pi(\sigma^2)\pi_(\boldsymbol{\alpha})\pi_(\boldsymbol{\beta}) \prod_{j=1}^p \pi(a_j|\tau_j^2)\pi(\tau_j^2|\lambda)\pi(\lambda) \end{aligned}$$

Conditional on genetic effects $\boldsymbol{a} = (a_1, \cdots, a_p)^T$ that we want to penalize, the model is a standard linear regression, and thus the conditional posterior distribution of $\boldsymbol{\alpha}$ is

$$\begin{aligned} \boldsymbol{\alpha}|\cdot \sim N_{d_1}\left(\Sigma'\left(\frac{\sum_{i=1}^n X_i(y_i - \mu - \mathbf{Z}_i^T\boldsymbol{\beta} - \boldsymbol{\xi}_i^T\mathbf{a})}{\sigma^2}\right), \Sigma'\right), \\ \text{with } \Sigma' = \left(\frac{\sum_{i=1}^n X_i X_i^T}{\sigma^2} + \Sigma_\alpha^{-1}\right)^{-1}. \end{aligned} \tag{3.9}$$

Similarly,

$$\begin{aligned} \boldsymbol{\beta}|\cdot \sim N_{d_2}\left(\Sigma''\left(\frac{\sum_{i=1}^n Z_i(y_i - \mu - \mathbf{X}_i^T\boldsymbol{\alpha} - \boldsymbol{\xi}_i^T\mathbf{a})}{\sigma^2}\right), \Sigma''\right), \\ \text{with } \Sigma'' = \left(\frac{\sum_{i=1}^n Z_i Z_i^T}{\sigma^2} + \Sigma_\beta^{-1}\right)^{-1}. \end{aligned} \tag{3.10}$$

Given $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$, our goal is to minimize the SCAD penalized least squares

$$\frac{1}{2}||\mathbf{y} - \boldsymbol{\mu} - X\boldsymbol{\alpha} - Z\boldsymbol{\beta} - \xi\mathbf{a}||^2 + \sum_{j=1}^p p_\lambda(|a_j|), \tag{3.11}$$

where genetic effects $\boldsymbol{a} = (a_1, \cdots, a_p)^T$ are decision variables of this objective function. Based on the local linear approximation (3.6), minimizing (3.11) is equivalent to minimizing

$$\frac{1}{2}||\mathbf{y} - \boldsymbol{\mu} - X\boldsymbol{\alpha} - Z\boldsymbol{\beta} - \xi\mathbf{a}||^2 + \sum_{j=1}^p p'_\lambda(|a_j^{(0)}|)(|a_j|), \tag{3.12}$$

and thus the estimation can be accomplished relatively easily using Bayesian lasso.

Specifically, let us define

$$U = \{j : p'_\lambda(|a_j^{(0)}|) = 0\} \quad \text{and} \quad V = \{j : p'_\lambda(|a_j^{(0)}|) > 0\}, \tag{3.13}$$

and partition the design matrix ξ and one-step LLA estimator $\boldsymbol{a}^{(1)}$ in the following way

$$\xi = [\xi_U, \xi_V] \quad \text{and} \quad \boldsymbol{a}^{(1)} = (\boldsymbol{a}^{(1)T}{}_U, \boldsymbol{a}^{(1)T}{}_V)^T. \tag{3.14}$$

We create working data by setting $\boldsymbol{\xi}_i^* = \sqrt{D_{ii}}\boldsymbol{\xi}_i$, $y_i^* = \sqrt{D_{ii}}\boldsymbol{\xi}_i\boldsymbol{a}^{(0)}$ for $i = 1, \cdots, n$, where $D_{ii} = 2$ in linear regression models. Then we transform $\boldsymbol{\xi}_j^*$ by setting $\boldsymbol{\xi}_j^* = \boldsymbol{\xi}_j^* \frac{\lambda}{p'_\lambda(a_j^{(0)})}$ for $j \in V$. Let H_U be the projection matrix in the space of $\{\boldsymbol{\xi}_j^*, j \in U\}$. We compute $\xi_V^{**} = \xi_V^* - H_U\xi_V^*$ and $y^{**} = y^* - H_U y^*$. Now the problem has been formulated as the lasso penalized least squares and our goal is to find the minimizer $\hat{a}_V^*$ of

$$\frac{1}{2}||y^{**} - \xi_V^{**}\boldsymbol{a}||^2 + \lambda \sum_{j \in V} |a_j|. \tag{3.15}$$

Therefore, the Bayesian lasso in chapter 1 can be directly applied to minimize (3.15). In particular, conditional on parameters $(\tau_1^2, \cdots, \tau_v^2, \boldsymbol{\alpha}, \boldsymbol{\beta})$ and the transformed data, the model becomes the weighted linear regression, and thus the conditional posterior distribution of $\mathbf{a}_V$ is

$$\begin{aligned} \mathbf{a}_V|\cdot &\sim N\left(A_a^{-1}\xi_V^{**}y_i^{**}, \sigma^2 A_a^{-1}\right), \\ \text{with } A_a^{-1} &= \left(\xi_V^{**}\xi_V^{**T} + \text{diag}(\tau_1^2, \cdots, \tau_v^2)^{-1}\right)^{-1}. \end{aligned} \tag{3.16}$$

Moreover, the full conditional for $\tau_1^2, \cdots, \tau_v^2$ are conditionally independent, with

$$\frac{1}{\tau_j^2}|\cdot \sim \text{Inverse-Gaussian}\left(\sqrt{\frac{\lambda^2\sigma^2}{a_j^2}}, \lambda^2\right), \quad j = 1, \cdots, v. \tag{3.17}$$

Due to the conjugate Gamma(a, b) prior, the conditional posterior distribution of the tuning parameter λ is

$$\lambda^2|\cdot \sim \text{Gamma}\left(v + a, \sum_{j=1}^{v} \frac{\tau_j^2}{2} + b\right). \tag{3.18}$$

Finally we sample σ_V^2 from

$$\sigma_V^2|\cdot \sim Inv-\chi^2\left(n, \frac{1}{n}\sum_{i=1}^{n}(y_i^{**}-\xi_{Vi}^{**}\boldsymbol{a}_V)^2\right). \tag{3.19}$$

Now an efficient Gibbs sampler based on equations (3.16)-(3.19) proceeds to draw posterior samples from each full conditional posterior distribution, given the current values of all other parameters and the observed data. We use the potential scale reduction factor $\hat{R}$ to access the convergence (Gelman and Rubin, 1992), and usually all chains converge very fast. We take the median as Bayesian estimates $\mathbf{a}_V^*$. Lastly, we compute $\hat{\mathbf{a}}_U^* = (\xi_U^{*T}\xi_U^*)^{-1}\xi_U^{*T}(y^* - \xi_V^*\hat{\mathbf{a}}_V^*)$. For $j \in U$, the SCAD estimator $a_j^{(1)} = \hat{a}_j^*$ and for $j \in V$, $a_j^{(1)} = \hat{a}_j^* \frac{\lambda}{p'_\lambda(|a_j^{(0)}|)}$.

Conditioning on the estimated genetic effects, we go back and sample covariates according to (3.9) and (3.10). We also need to sample σ^2 from

$$\sigma^2|\cdot \sim Inv-\chi^2\left(n, \frac{1}{n}\sum_{i=1}^{n}(y_i-\mu_i-\mathbf{X_i}^T\boldsymbol{\alpha}-\mathbf{Z_i}^T\boldsymbol{\beta}-\boldsymbol{\xi_i}^T\mathbf{a})^2\right). \tag{3.20}$$

This process continues until all chains converge. Again we use the potential scale reduction factor $\hat{R}$ to access the convergence. Once $\hat{R} < 1.1$ for all scalar estimands of interest, we continue to draw $15,000$ iterations to obtain samples from the joint posterior distribution.

3.4.3 Posterior interpretation

The proposed MCMC algorithm for our Bayesian SCAD can provide posterior median estimates of the additive effects and dominant effects of individual SNPs, while adjusting for the effects of all other SNPs and covariates. Furthermore, using the posterior samples of $\mathbf{a}$, $\mathbf{d}$, and the observed genotypes, we can calculate the proportion of the phenotypic variance explained by a particular SNP, i.e., heritability, by

$$h_j^2 = \frac{2\hat{p}_1\hat{p}_0(\hat{a}_j + (\hat{p}_1-\hat{p}_0)\hat{d}_j)^2 + 4\hat{p}_1^2\hat{p}_0^2\hat{d}_j^2}{\mathrm{var}(\tilde{y})}, \quad j = 1, \cdots, p,$$

where $\hat{p}_1$ is the estimated allele frequency for A, and $\hat{p}_0$ is the estimated allele frequency for a, $\hat{a}_j$ is the median estimate of the additive effect for SNP j, and $\hat{d}_j$ is the median estimate of the dominant effect for SNP j. Since heritability estimates are unitless, they could guide variable selection and identify SNPs that have relatively large effects on the phenotype.

3.5 Examples

3.5.1 Computer Simulation

The new approach is investigated through simulation studies. We generate data according to the model (3.2) with $\mu = 0$, $\sigma^2 = 4$ or 16 and $n = 200$. For ease of simulation, ξ_{ij} is derived from u_{ij}, where each u_{ij} has a standard normal distribution marginally, and $\rho = cov(u_{ij}, u_{ik})$. Then, to mimic a SNP with equal allele frequencies, we set

$$\xi_{ij} = \begin{cases} 1, & u_{ij} > c \\ 0, & -c \leq u_{ij} \leq c \\ -1, & u_{ij} < -c, \end{cases}$$

where $-c$ is the first quartile of a standard normal distribution. For simplicity, we only consider additive genetic model. We assume that there are 500 or 1000 SNPs from which 20 have significant additive effects for a phenotypic trait. Specifically, we generate data on $p = 500$ or 1000, with $\rho = 0.1$, $a_j = 1$ for $1 \leq j \leq 20$ and $a_j = 0$ for $j > 20$.

We apply three methods to the 100 simulated datasets: single SNP analysis (SSA), the Bayesian lasso applied to the preconditioned response from supervised principal components (PB-lasso) and sure independence screening followed by Bayesian SCAD (SIS-SCAD). In single SNP analysis, we reject the null hypothesis that the genetic effect of an individual SNP equals to zero at the significance level of 5% with the FDR adjustment. For the Bayesian lasso and Bayesian SCAD, we reject the null hypothesis if the Bayesian estimate is greater than 0.1.

For each estimated genetic effect obtained from each method, we calculate the average bias and empirical standard error over 100 simulations. Since the first 20 genetic effects are nonzeros with the same true value, in Table 3.1 we

report the average bias and empirical standard error over the first 20 SNPs and over the rest of the SNPs separately. The standard error of each average is in parentheses. In the column labeled "Aver. Nonzeros", we present the average number of nonzero coefficients correctly identified to be nonzero, or the average number of zero coefficients incorrectly estimated to be nonzero in 100 replications. In the column "Standardized CPU Time", we report the amount of time that each method took. For the ease of comparison, the computational time is standardized according to that of single SNP analysis.

As can be seen from Table 3.1, the single SNP analysis tends to overestimate the genetic effect, since when we test a SNP for the association with the phenotype, we assume the genetic variation is solely due to this particular SNP, and ignore the effects from all other SNPs. Therefore, in terms of parameter estimation, model selection methods that simultaneously estimate the genetic effects associated with all SNPs outperform the traditional single SNP analysis. In terms of variable selection, preconditioned Bayesian lasso and SIS based Bayesian SCAD greatly improve the probability of correctly identifying regression coefficients with nonzero effects. Moreover, as the number of SNPs gets larger, single SNP analysis detects fewer important SNPs, since this method subjects to severe multiple comparison adjustment. By contrast, two-stage variable selection methods are still able to identify nonzero coefficients and zero coefficients correctly in almost every simulation.

Moreover, compared with preconditioned Bayesian lasso, SIS based Bayesian SCAD has higher statistical power and lower false positive rate. The reason is that although preconditioning denoises the response variable by supervised principle component analysis, the number of predictors entering variable selection remains. The large amount of potential predictors increases the probability of incorrectly identifying predictors whose true effects are zero. Moreover, due to the high dimensional feature space, preconditioned Bayesian lasso takes longer in the variable selection step. Finally, SIS based Bayesian SCAD yields estimates with much lower bias, due to the specific structure of SCAD penalty function. Table 3.2 displays the simulation results when $\sigma = 4$, which consistent with our findings.

Table 3.1. Simulation results for three methods based on 100 simulations when $\sigma = 2$.

Method	Bias	Empirical SE	Aver. Nonzeros	Standardized CPU time
$n = 200, p = 500$, β_1 - β_{20}				
SSA	4.81	2.04	16.74	1.00
	(0.31)	(0.27)	(1.98)	
PB-lasso	0.23	0.28	19.66	473.22
	(0.04)	(0.03)	(0.51)	
SIS-SCAD	0.05	0.37	19.84	14.62
	(0.02)	(0.03)	(0.64)	
$n = 200, p = 500$, β_{21} - β_{500}				
SSA	0.44	0.43	0.14	
	(0.06)	(0.08)	(0.40)	
PB-lasso	0.01	0.03	2.40	
	(0.01)	(0.04)	(0.47)	
SIS-SCAD	0.01	0.01	0.44	
	(0.01)	(0.01)	(0.61)	
$n = 200, p = 1000$, β_1 - β_{20}				
SSA	5.02	2.07	16.18	1.00
	(0.34)	(0.23)	(2.47)	
PB-lasso	0.58	0.31	18.56	568.76
	(0.05)	(0.03)	(1.43)	
SIS-SCAD	0.06	0.39	19.46	9.14
	(0.04)	(0.03)	(0.89)	
$n = 200, p = 1000$, β_{21} - β_{1000}				
SSA	0.43	0.43	0.02	
	(0.06)	(0.08)	(0.14)	
PB-lasso	0.00	0.01	2.82	
	(0.01)	(0.01)	(0.89)	
SIS-SCAD	0.00	0.00	0.07	
	(0.01)	(0.01)	(0.89)	

3.5.2 Real Data Analysis

We applied the sure independence screening followed by Bayesian SCAD to analyze a real GWAS data set from the Framingham Heart Study (FHS), a cardiovascular study based in Framingham, Massachusetts, supported by the National Heart,

Table 3.2. Simulation results for three methods based on 100 simulations when $\sigma = 4$.

Method	Bias	Empirical SE	Aver. Nonzeros	Standardized CPU Time
$n = 200, p = 500$, β_1 - β_{20}				
SSA	3.23	1.75	12.02	1.00
	(0.30)	(0.18)	(2.64)	
PB-lasso	0.22	0.32	18.64	277.71
	(0.04)	(0.03)	(1.19)	
SIS-SCAD	0.05	0.39	18.80	13.35
	(0.02)	(0.03)	(1.04)	
$n = 200, p = 500$, β_{21} - β_{500}				
SSA	0.43	0.43	0.04	
	(0.06)	(0.08)	(0.19)	
PB-lasso	0.01	0.03	4.24	
	(0.01)	(0.04)	(1.15)	
SIS-SCAD	0.00	0.01	1.62	
	(0.01)	(0.01)	(0.98)	
$n = 200, p = 1000$, β_1 - β_{20}				
SSA	3.50	1.79	10.22	1.00
	(0.26)	(0.19)	(3.89)	
PB-lasso	0.57	0.33	16.68	728.15
	(0.06)	(0.04)	(1.66)	
SIS-SCAD	0.07	0.42	17.52	10.53
	(0.04)	(0.03)	(1.27)	
$n = 200, p = 1000$, β_{21} - β_{1000}				
SSA	0.43	0.43	0.04	
	(0.06)	(0.08)	(0.20)	
PB-lasso	0.00	0.02	3.14	
	(0.01)	(0.01)	(1.36)	
SIS-SCAD	0.01	0.00	2.24	
	(0.01)	(0.01)	(1.20)	

Lung, and Blood Institute, in collaboration with Boston University (Dawber et al. 1951). Recently, 550,000 SNPs have been genotyped for the entire Framingham cohort (Jaquish 2007), from which 418 males and 559 females were chosen for our data analysis. These subjects were measured for body mass index (BMI) at

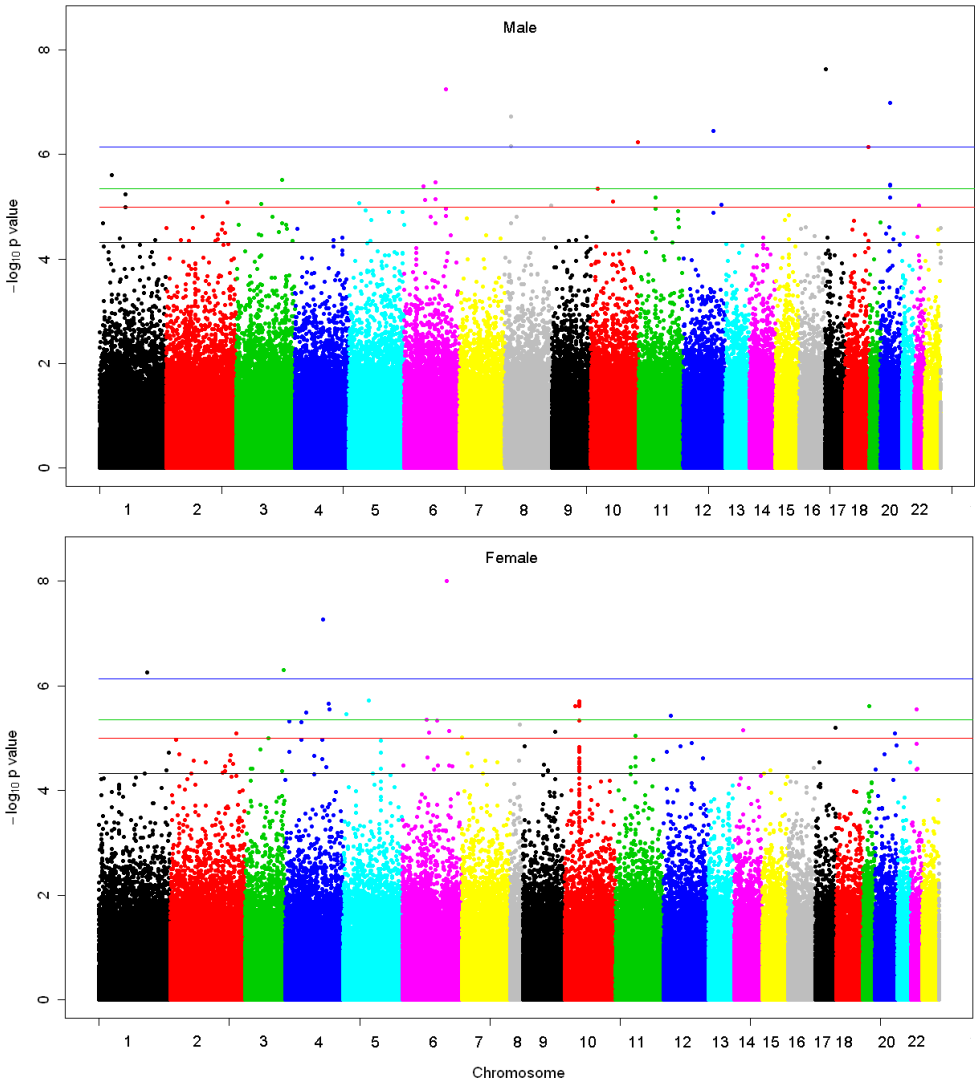

Figure 3.1. Single SNP analysis for the Framingham genome-wide association study

different ages from 29 and 61 years. As is standard practice, SNPs with minor allele frequency $< 10\%$ were excluded from data analysis. The numbers and percentages of non-rare allele SNPs vary among different chromosomes and ranges from 4,417 to 28,771 and from 64% to 72%, respectively.

Figure 3.1 gives $-log_{10}$ p-values for each SNP in the two sexes, from which we selected 1837 SNPs for Bayesian lasso analysis in Chapter 2. However in this Chapter, we will consider all SNPs across the entire genome simultaneously, since SIS is designed for ultra-high dimensional problem. It will dramatically decrease

the dimensionality of the variable selection problem, which expedites the following Bayesian SCAD procedure. Before the statistical analysis, we imputed missing genotypes for a small proportion of SNPs (5.16%) according to the distribution of genotypes in the population. Then the sure independence screening was used to select potential predictors. From more than half million SNPs, it selected 141 SNPs from 21 chromosomes for the Bayesian SCAD.

By treating the sex as a discrete covariate and age as a continuous covariate, we imposed SCAD penalties on the additive effects $a_1, \cdots, a_p$ to identify those SNPs who display notable effects on BMI. Dominant effects are not considered here since they are pretty low as shown in the previous chapter. We employ the proposed MCMC algorithms to estimate all parameters and implement variable selection, where $\Sigma_\alpha = 1$, $\Sigma_\beta = 1$, and all parameters in the conjugate gamma hyperpriors are 0.1. In unreported tests, we find that the posteriors are not sensitive to these prior specifications, as long as the prior distributions are relatively flat (Park and Casella, 2008; Yi and Xu, 2008).

Table 3.3 tabulates the names and positions of 21 identified SNPs, as well as their minor alleles and estimated additive genetic effects. The Bayesian SCAD automatically shrinks all small coefficients to zero, while estimates large genetic effects unbiasedly. Note that although the preconditioned Bayesian lasso in the previous chapter only considers 1837 SNPs, SNP ss66364251 was also reported in its result. However, since lasso penalty functions shrink large regression coefficients towards zero, the additive effect for this SNP estimated by the Bayesian lasso is much smaller than that obtained by the Bayesian SCAD.

Table 3.3. The estimates of additive and dominant effects triggered by each significant SNP. The heritability of each SNP is also given.

Chr	Name	Position	Minor Allele	Additive Effect
1	ss66123680	18922202	A	-1.2253
1	ss66349851	150858886	G	0.4037
1	ss66482440	158368880	A	0.7253
1	ss66364251	198321700	G	0.4841
1	ss66452206	244896239	T	0.4865
5	ss66535637	141394644	T	0.5149
6	ss66274176	33761089	T	0.4340
6	ss66127246	44184758	T	-0.4700
7	ss66206374	45181437	G	-0.4065
7	ss66396358	111410732	A	0.4448
7	ss66053198	146798866	T	-0.4542
8	ss66222552	133743208	C	0.4882
10	ss66444366	21172593	C	0.4046
12	ss66282946	42033770	T	0.4425
12	ss66295249	78074072	A	0.5296
14	ss66518701	75414324	A	-0.4022
15	ss66491162	86995786	T	-0.5662
16	ss66297378	88096716	A	0.5200
17	ss66219268	27897975	C	-0.4162
17	ss66386743	62148632	G	-0.4707
20	ss66193993	17644239	T	-1.1298

Chapter 4

Bayesian Group Lasso for Nonparametric Varying-Coefficient Models with Application to Functional Genome-Wide Association Studies

4.1 Introduction

In the presence of longitudinal and functional phenotypical data in many clinical trials, measurements of a complex trait are often collected repeatedly over a period at subject-specific time points. The statistical analysis of GWAS using single measurements at one time point is not capable of revealing the dynamic pattern of genetic control over a time course, and the reported SNPs can only explain a small proportion of genetic variance. Therefore, there is a daunting need on the development of a variable selection model to accommodate irregular longitudinal data.

If we assume that the effects of SNPs are smooth functions of time and could be estimated nonparametrically, variable selection in nonparametrical setting is equivalent to selecting a subset of predictors with nonzero functional coefficients.

Lin and Zhang (2006) developed COSSO for model selection in smoothing spline ANOVA model, with the penalty term being the sum of component norms. Zhang and Lin (2006) further extended it to nonparametric regression in exponential family. Wang et al. (2008) estimated time-varying effects using basis expansion and selected significant predictors by imposing SCAD penalty functions on the L_2-norm of these basis expansions.

In this chapter, we develop novel statistical models and algorithms that can analyze multiple SNPs simultaneously, and integrate the developmental mechanisms of trait formation into a general GWAS framework through mathematical functions. Specifically, we consider orthogonal polynomials to approximate time-varying effects in functional GWAS model, and propose Bayesian group lasso approach for variable selection in nonparametric setting.

Group lasso was first proposed by Yuan and Lin (2006). They considered the problem of selecting important groups of independent variables in linear regression models, and generalized lasso by encouraging sparsity at the group level. However, since the Hessian is not defined at the optimal solution, they did not provide variance estimates for the regression coefficients. Here, we express time-varying effects as a linear combination of legendre polynomials, and in such case, the selection of important predictors corresponds to the selection of groups of polynomials. We develop a Bayesian hierarchical model for group variable selection, and estimate all parameters by MCMC algorithms. Our method provides not only point estimates but also interval estimates of all parameters, and the traditional Bayesian lasso (Park and Casella, 2008) is its special case in which response variable is univariate.

The rest of the chapter is organized as follows. In section 2 we introduce functional GWAS model that connects genotypes and irregular longitudinal phenotypical data. In section 3 we propose the Bayesian hierarchical model for functional GWAS where group lasso penalties are applied to the genetic effects of all SNPs. In section 4 we provide the posterior computations as well as the interpretation of the results. In section 5 and section 6 we give examples using simulated data and real data, respectively. We provide concluding remarks in section 7.

4.2 Functional GWAS Model

The model for functional genome-wide association studies (*f*GWAS) are the integration of functional data analysis and genome-wide association studies, with the primary goal being to study the dynamic pattern of genetic actions and interactions triggered by significant SNPs throughout the entire genome. Beyond traditional GWAS, *f*GWAS targets phenotypic traits that are measured longitudinally at repeated time points. Suppose in a genome-wide association study involving n subjects, a continuous longitudinal trait of interest is measured at irregularly spaced time points, which is not common to all subjects. Let $\mathbf{y}_i = (y_i(t_{i1}), \cdots, y_i(t_{iT_i}))^T$ be the T_i-dimensional vector of measurements on subject i where $\mathbf{t}_i = (t_{i1}, \cdots, t_{iT_i})^T$ is the corresponding vector of measurement time points after standardization. $\mathbf{y}_i$ can be described as

$$\begin{aligned}
\begin{pmatrix} y_i(t_{i1}) \\ \vdots \\ y_i(t_{iT_i}) \end{pmatrix} &= \begin{pmatrix} \mu(t_{i1}) \\ \vdots \\ \mu(t_{iT_i}) \end{pmatrix} + \begin{pmatrix} \alpha_1(t_{i1}) & \cdots & \alpha_q(t_{i1}) \\ \vdots & & \vdots \\ \alpha_1(t_{iT_i}) & \cdots & \alpha_q(t_{iT_i}) \end{pmatrix} \begin{pmatrix} X_{i1} \\ \vdots \\ X_{iq} \end{pmatrix} \\
&+ \begin{pmatrix} a_1(t_{i1}) & \cdots & a_p(t_{i1}) \\ \vdots & & \vdots \\ a_1(t_{iT_i}) & \cdots & a_p(t_{iT_i}) \end{pmatrix} \begin{pmatrix} \xi_{i1} \\ \vdots \\ \xi_{ip} \end{pmatrix} \\
&+ \begin{pmatrix} d_1(t_{i1}) & \cdots & d_p(t_{i1}) \\ \vdots & & \vdots \\ d_1(t_{iT_i}) & \cdots & d_p(t_{iT_i}) \end{pmatrix} \begin{pmatrix} \zeta_{i1} \\ \vdots \\ \zeta_{ip} \end{pmatrix} + \begin{pmatrix} e_i(t_{i1}) \\ \vdots \\ e_i(t_{iT_i}) \end{pmatrix}, \quad (4.1)
\end{aligned}$$

or the observed phenotypical value of the ith subject at discrete time point t_{il} can be described by

$$\begin{aligned}
y_i(t_{i\ell}) &= \mu(t_{i\ell}) + \boldsymbol{\alpha}(t_{i\ell})^T \mathbf{X}_i + \mathbf{a}(t_{i\ell})^T \boldsymbol{\xi}_i + \mathbf{d}(t_{i\ell})^T \boldsymbol{\zeta}_i + e_i(t_{i\ell}), \\
&\quad i = 1, \cdots, n, \quad \ell = 1, \cdots, T_i, \qquad (4.2)
\end{aligned}$$

where $\mu(t_{i\ell})$ is the overall mean, $\boldsymbol{\alpha}(t_{i\ell}) = (\alpha_1(t_{i\ell}), \cdots, \alpha_q(t_{i\ell}))^T$ is the q-dimensional vector of covariate effects, $\mathbf{X}_i = (X_{i1}, \cdots, X_{iq})^T$ is the observed covariate vector for subject i, $\mathbf{a}(t_{i\ell}) = (a_1(t_{i\ell}), \cdots, a_p(t_{i\ell}))^T$ and $\mathbf{d}(t_{i\ell}) = (d_1(t_{i\ell}), \cdots, d_p(t_{i\ell}))^T$ are the

p-dimensional vectors of the additive and dominant effects of SNPs, respectively, $\boldsymbol{\xi}_i = (\xi_{i1}, \cdots, \xi_{ip})^T$ and $\boldsymbol{\zeta}_i = (\zeta_{i1}, \cdots, \zeta_{ip})^T$ are the indicator vectors of the additive and dominant effects of SNPs for subject i, and $e_i(t_{i\ell})$ is the residual error assumed to follow a $\mathrm{N}(0, \sigma^2(t_{i\ell}))$ distribution. The j-th elements of $\boldsymbol{\xi}_i$ and $\boldsymbol{\zeta}_i$ are defined as

$$\xi_{i\ell} = \begin{cases} 1, & \text{if the genotype of SNP } j \text{ is } AA \\ 0, & \text{if the genotype of SNP } j \text{ is } Aa \\ -1, & \text{if the genotype of SNP } j \text{ is } aa, \end{cases}$$

$$\zeta_{i\ell} = \begin{cases} 1, & \text{if the genotype of SNP } j \text{ is } Aa \\ 0, & \text{if the genotype of SNP } j \text{is } AA \text{ or } aa. \end{cases}$$

In the *f*GWAS model, the effects of covariates and SNPs are assumed to be functions of time. Many methods of estimating time-varying coefficients of a linear model in longitudinal data setting have been proposed and studied, including basis expansion methods, local polynomial kernel methods, and smoothing spline methods. In particular, if we consider orthogonal polynomials and approximate the effect of k-th covariate by a Legendre polynomial of order $v-1$, we have the following representation

$$(\alpha_k(t_{i1}), \cdots, \alpha_k(t_{iT_i}))^T = U_i \mathbf{r}_k, \quad k = 1, \cdots, q, \tag{4.3}$$

where $\mathbf{r}_k = (r_{k0}, \cdots, r_{k(v-1)})^T$ are the Legendre polynomial coefficients, and

$$U_i = \begin{pmatrix} \mathbf{u}_{i1}^T \\ \vdots \\ \mathbf{u}_{iT_i}^T \end{pmatrix} = \begin{pmatrix} 1 & t_{i1} & \frac{1}{2}(3t_{i1}^2 - 1) & \frac{1}{2}(5t_{i1}^3 - 3t_{i1}) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_{iT_i} & \frac{1}{2}(3t_{iT_i}^2 - 1) & \frac{1}{2}(5t_{iT_i}^3 - 3t_{iT_i}) & \cdots \end{pmatrix}, \tag{4.4}$$

are Legendre polynomial functions. Similarly, other time-varying effects can be represented as

$$(a_j(t_{i1}), \cdots, a_j(t_{iT_i}))^T = U_i \mathbf{b}_j, \quad j = 1, \cdots, p, \tag{4.5}$$

$$(d_j(t_{i1}), \cdots, d_j(t_{iT_i}))^T = U_i \mathbf{c}_j, \quad j = 1, \cdots, p, \tag{4.6}$$

$$(\mu(t_{i1}), \cdots, \mu(t_{iT_i}))^T = U_i \mathbf{m}, \tag{4.7}$$

where $\mathbf{b}_j = (b_{j0}, \cdots, b_{j(v-1)})^T$ are the Legendre polynomial coefficients for the additive effect of the j-th SNP, $\mathbf{c}_j = (c_{j0}, \cdots, c_{j(v-1)})^T$ are the Legendre polynomial coefficients for the dominant effect of the j-th SNP, and $\mathbf{m} = (m_0, \cdots, m_{v-1})^T$ are the Legendre polynomial coefficients for the overall mean function.

After introducing Legendre polynomials to approximate time-varying effects of covariates and SNPs, the full model of *f*GWAS becomes

$$\begin{aligned} y_i(t_{i\ell}) &= \mathbf{u}_{il}^T\mathbf{m} + (\mathbf{u}_{il}^T\mathbf{r}_1, \cdots, \mathbf{u}_{il}^T\mathbf{r}_q)\mathbf{X}_i + (\mathbf{u}_{il}^T\mathbf{b}_1, \cdots, \mathbf{u}_{il}^T\mathbf{b}_p)\boldsymbol{\xi}_i \\ &\quad + (\mathbf{u}_{il}^T\mathbf{c}_1, \cdots, \mathbf{u}_{il}^T\mathbf{c}_p)\boldsymbol{\zeta}_i + e_i(t_{i\ell}), \quad i = 1, \cdots, n, \quad \ell = 1, \cdots, T_i. \end{aligned} \tag{4.8}$$

Lastly, since measurements within each subject are possibly correlated with one another, we assume that $\mathbf{e}_i = (e_i(t_{i1}), \cdots, e_i(t_{iT_i}))^T$ follows a multivariate normal distribution with zero mean and covariance matrix Σ_i. Both parametric and nonparametric methods have been developed to model the structure of covariance between longitudinal measurements (Ma et al. 2002; Zhao et al. 2005; Yap et al. 2009). Since the focus of this article is to estimate time-varying coefficients nonparametrically, and select significant predictors by penalized methods, we will employ the first-order autoregressive (AR(1)) model as the within-subject covariance matrix for illustrative purpose. The extension to the model with other covariance structures is straightforward.

4.3 Bayesian Hierarchical Representation

In high-dimensional regression problems, such as GWAS, a regularized approach is preferred to identify predictors with nonzero effects and achieve better out-of-sample predictive performance. When the parameters that we would like to penalize are Euclidean, we may apply different penalty functions to them to perform variable selection. However, when these parameters are nonparametric smooth functions, a traditional regularization procedure cannot be directly applied. In this situation, regularized estimation for selecting important predictors is equivalent to selecting functional coefficients that are not identically zero.

Let $||\mathbf{b}_j||$ be the L_2 norm of the vector $\mathbf{b}_j$. The time-varying additive effect of jth SNP is identically zero if and only if $||\mathbf{b}_j|| = 0$. Therefore, if we estimate

additive effects by a Legednre polynomial of order v, and would like to identify significant additive effects via penalized methods, we may partition all parameters of additive effects $(\mathbf{b}_1^T, \cdots, \mathbf{b}_p^T)$ into p groups of size v according to p SNPs, and encourage sparse solution at the group level or select a subset of groups with nonzero L_2 norms. That is, group lasso minimizes the following penalized least square

$$\frac{1}{2}||\mathbf{y} - \boldsymbol{\mu}||^2 + \lambda \sum_{j=1}^{p} ||\mathbf{b}_j|| + \lambda^* \sum_{j=1}^{p} ||\mathbf{c}_j||, \tag{4.9}$$

where $\mathbf{y}^T = (\mathbf{y}_1^T, \cdots, \mathbf{y}_n^T)$, $\boldsymbol{\mu}^T = E\mathbf{y}^T = (\boldsymbol{\mu}_1^T, \cdots, \boldsymbol{\mu}_n^T)$ and λ and λ^* are two regularization parameters. λ and λ^* control the amount of shrinkage toward zero: the larger their values, the greater the amount of shrinkage. It should be adaptively determined from the data to minimize an estimate of expected prediction error.

From a Bayesian perspective, group lasso estimates can be interpreted as posterior mode estimates when the regression parameters have multivariate independent and identical Laplace priors. Therefore, when group lasso penalties are imposed on the Legendre coefficients of additive and dominant effects, the conditional prior for $\mathbf{b}_j$ is a multivariate Laplace distribution with the scale parameter $(v\lambda^2/\sigma^2)^{-\frac{1}{2}}$:

$$\pi(\mathbf{b}_j|\sigma^2) = (v\lambda^2/\sigma^2)^{\frac{v}{2}} \exp(-(v\lambda^2/\sigma^2)^{-\frac{1}{2}}||\mathbf{b}_j||), \tag{4.10}$$

and the conditional multivariate Laplace prior for dominant effects $\mathbf{c}_j$ is

$$\pi(\mathbf{c}_j|\sigma^2) = (v\lambda^{*2}/\sigma^2)^{\frac{v}{2}} \exp(-(v\lambda^{*2}/\sigma^2)^{-\frac{1}{2}}||\mathbf{c}_j||). \tag{4.11}$$

To ensure the derived conditional distribution of $\mathbf{b}_j$ having a standard form, we may rewrite the multivariate Laplace prior distribution as a scale mixture of a Normal distribution with a Gamma distribution, that is

$$\begin{aligned} & \text{M-Laplace}(\mathbf{b}_j|0, (v\lambda^2/\sigma^2)^{-\frac{1}{2}}) \\ \propto\ & (v\lambda^2/\sigma^2)^{\frac{v}{2}} \exp(-(v\lambda^2/\sigma^2)^{\frac{1}{2}}||\mathbf{b}_j||_2) \\ \propto\ & \int_0^\infty \text{N}(\mathbf{b}_j|0, \sigma^2\tau_j^2)\text{Gamma}\left(\tau_j^2|\frac{v+1}{2}, \frac{2}{v\lambda^2}\right) d\tau_j^2, \end{aligned} \tag{4.12}$$

where $(v\lambda^2/\sigma^2)^{-\frac{1}{2}}$ is the scale parameter of multivariate Laplace distribution, $\frac{v+1}{2}$

is the shape parameter of Gamma distribution, and $\frac{2}{v\lambda^2}$ is the scale parameter of Gamma distribution. After integrating out τ_j^2, the conditional priors on $\mathbf{b}_j$ has the desired form (4.10). Then, in a Bayesian hierarchical model, we can rewrite the multivariate Laplace priors on $\mathbf{b}_j$ as

$$\begin{aligned} \mathbf{b}_j|\tau_j^2, \sigma^2 &\sim N(0, \sigma^2\tau_j^2), \\ \tau_j^2|\lambda &\sim \text{Gamma}\left(\frac{v+1}{2}, \frac{2}{v\lambda^2}\right). \end{aligned} \tag{4.13}$$

Likewise, the multivariate-Laplacian priors on $\mathbf{c}_j$ can be replaced by

$$\begin{aligned} \mathbf{c}_j|\tau_j^{*2}, \sigma^2 &\sim N(0, \sigma^2\tau_j^{*2}), \\ \tau_j^{*2}|\lambda &\sim \text{Gamma}\left(\frac{v+1}{2}, \frac{2}{v\lambda^{*2}}\right). \end{aligned} \tag{4.14}$$

Then, given λ and λ^*, we have the following hierarchical representation of the penalized regression model:

$$\begin{aligned} \mathbf{y}|\cdot &\propto (2\pi)^{-\frac{\sum_i^n T_i}{2}}(\prod_i^n |\Sigma_i|^{-\frac{1}{2}})e^{-\frac{1}{2}\sum_i^n(\mathbf{y}_i-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)}, \\ \mathbf{m} &\sim N_v(0, \Sigma_{m0}), \\ \mathbf{r}_k &\sim N_v(0, \Sigma_{r0}), \quad k = 1, \cdots, q, \\ \mathbf{b}_j|\tau_j^2, \sigma^2 &\sim N(0, \sigma^2\tau_j^2), \quad j = 1, \cdots, p, \\ \tau_j^2|\lambda &\sim \text{Gamma}\left(\frac{v+1}{2}, \frac{2}{v\lambda^2}\right), \quad j = 1, \cdots, p, \\ \mathbf{c}_j|\tau_j^{*2}, \sigma^2 &\sim N(0, \sigma^2\tau_j^{*2}), \quad j = 1, \cdots, p, \\ \tau_j^{*2}|\lambda^* &\sim \text{Gamma}\left(\frac{v+1}{2}, \frac{2}{v\lambda^{*2}}\right), \quad j = 1, \cdots, p, \\ \rho &\sim \text{U}(-1, 1), \\ \sigma^2 &\sim \pi(\sigma^2), \\ \sigma^2, \lambda, \lambda^* &> 0. \end{aligned} \tag{4.15}$$

where λ and λ^* are regularization parameters or group lasso parameters that control the degrees of shrinkage in the estimate of the genetic effects. We assign conjugate Normal priors to $\mathbf{m}$ when estimating the overall mean function. We also assign

conjugate Normal priors to the Legendre coefficients of covariates $\mathbf{r}_k, k = 1, \cdots, q$, because covariates in GWAS are usually low-dimensional and not the parameters of interest. ρ is the autoregressive parameter in the assumed AR(1) covariance structure, and we use Uniform distribution on $[-1, 1]$ as a prior for ρ. Finally, since the data are usually sufficient to estimate σ, we can use a noninformative prior such as $\pi(\sigma^2) = 1/\sigma^2$ for σ^2.

Note that there are two lasso parameters λ and λ^* in the Bayesian hierarchical representation of group lasso. Traditionally, they can be prespecified by cross-validation or generalized cross-validation. However, in the Bayesian group lasso setting, λ and λ^* can be estimated along with other parameters by assigning appropriate hyperpriors to them. This procedure allows us to determine the amount of regularization from the data, and avoids refitting the model repeatedly. In particular, we consider the conjugate gamma priors on $\lambda^2/2$ and $\lambda^{*2}/2$,

$$\pi\left(\frac{\lambda^2}{2}\right) \sim \text{Gamma}(a, b), \tag{4.16}$$

$$\pi\left(\frac{\lambda^{*2}}{2}\right) \sim \text{Gamma}(a^*, b^*), \tag{4.17}$$

where a, b, a^*, and b^* are small values so that the priors are essentially noninformative. With this specification, group lasso parameters can simply join the other parameters in the Gibbs sampler.

4.4 Posterior Computation and Interpretation

We estimate the unknown parameters and hyperparameters by sampling from their conditional posterior distributions through MCMC algorithms. Given the data likelihood and prior distributions, the posterior distributions of all unknowns can be obtained by Bayes' theorem. For most of the parameters, the conditional posterior distributions have closed forms by conjugacy, which facilitates drawing posterior samples.

Assuming that priors for different predictors are independent, we can express

the joint posterior distribution of all parameters as:

$$\begin{aligned}
& \pi(\mathbf{m}, \mathbf{r}_k, \mathbf{b}_j, \tau_j^2, \lambda, \mathbf{c}_j, \tau_j^{*2}, \lambda^*, \sigma^2, \rho|\mathbf{y}) \\
\propto\ & \pi(\mathbf{y}|\cdot)\pi(\mathbf{m})\pi(\sigma^2)\pi(\rho)\prod_{k=1}^{q}\pi(\mathbf{r}_k) \\
& \prod_{j=1}^{p}\pi(\mathbf{b}_j|\tau_j^2)\pi(\tau_j^2|\lambda)\pi(\lambda)\pi(\mathbf{c}_j|\tau_j^{*2})\pi(\tau_j^{*2}|\lambda^*)\pi(\lambda^*). \qquad (4.18)
\end{aligned}$$

Conditional on the parameters $(\mathbf{r}_k, \mathbf{b}_j, \tau_j^2, \lambda, \mathbf{c}_j, \tau_j^{*2}, \lambda^*, \sigma^2, \rho)$, we derive the conditional posterior distribution of $\mathbf{m}$ as

$$\begin{aligned}
& \pi(\mathbf{m}|\mathbf{y}, \mathbf{r}_k, \mathbf{b}_j, \tau_j^2, \lambda, \mathbf{c}_j, \tau_j^{*2}, \lambda^*, \sigma^2, \rho) \\
\propto\ & \pi(\mathbf{m})\pi(\mathbf{y}|\cdot) \\
\propto\ & \exp\Big(-\frac{1}{2}\mathbf{m}^T\Sigma_{m0}^{-1}\mathbf{m} - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu}_{i(-m)} - U_i\mathbf{m})^T\Sigma_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_{i(-m)} - U_i\mathbf{m})\Big) \\
\propto\ & \exp\Big(\mathbf{m}^T\Sigma_{m0}^{-1}\mathbf{m} + \sum_{i=1}^{n}(U_i\mathbf{m})^T\Sigma_i^{-1}(U_i\mathbf{m}) - 2\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu}_{i(-m)})^T\Sigma_i^{-1}(U_i\mathbf{m})\Big) \\
\propto\ & \exp\Big(\mathbf{m}^T(\Sigma_{m0}^{-1} + \sum_{i=1}^{n}U_i^T\Sigma_i^{-1}U_i)\mathbf{m} - 2\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu}_{i(-m)})^T\Sigma_i^{-1}(U_i\mathbf{m})\Big). \qquad (4.19)
\end{aligned}$$

Hence, the conditional posterior distribution of $\mathbf{m}$ is $\mathrm{MVN}_v(\boldsymbol{\mu}_m, \Sigma_m)$, where

$$\boldsymbol{\mu}_m = \left(\Sigma_{m0}^{-1} + \sum_{i=1}^{n}U_i^T\Sigma_i^{-1}U_i\right)^{-1}\left(\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu}_{i(-m)})^T\Sigma_i^{-1}U_i\right)^T,$$

and

$$\Sigma_m = \left(\Sigma_{m0}^{-1} + \sum_{i=1}^{n}U_i^T\Sigma_i^{-1}U_i\right)^{-1}.$$

Similarly, since $\mathbf{r}_k$, $\mathbf{b}_j$ and $\mathbf{c}_j$ have conjugate Normal priors, the posterior distribution for $\mathbf{r}_k$ is $\mathrm{MVN}_v(\boldsymbol{\mu}_{r_k}, \Sigma_{r_k})$, with

$$\boldsymbol{\mu}_{r_k} = \left(\Sigma_{r0}^{-1} + \sum_{i=1}^{n}(X_{ik}U_i)^T\Sigma_i^{-1}(X_{ik}U_i)\right)^{-1}\left(\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu}_{i(-r_k)})^T\Sigma_i^{-1}(X_{ik}U_i)\right)^T,$$

and

$$\Sigma_{r_k} = \left(\Sigma_{r0}^{-1} + \sum_{i=1}^{n} (X_{ik}U_i)^T \Sigma_i^{-1} (X_{ik}U_i) \right)^{-1},$$

the posterior distribution for $\mathbf{b}_j$ is $\text{MVN}_v(\boldsymbol{\mu}_{b_j}, \Sigma_{b_j})$, with

$$\boldsymbol{\mu}_{b_j} = \left((\sigma^2\tau_j^2)^{-1} + \sum_{i=1}^{n} (\xi_{ij}U_i)^T \Sigma_i^{-1} (\xi_{ij}U_i) \right)^{-1} \left(\sum_{i=1}^{n} (\mathbf{y}_i - \boldsymbol{\mu}_{i(-b_j)})^T \Sigma_i^{-1} (\xi_{ij}U_i) \right)^T,$$

and

$$\Sigma_{b_j} = \left((\sigma^2\tau_j^2)^{-1} + \sum_{i=1}^{n} (\xi_{ij}U_i)^T \Sigma_i^{-1} (\xi_{ij}U_i) \right)^{-1},$$

and the posterior distribution for $\mathbf{c}_j$ is $\text{MVN}_v(\boldsymbol{\mu}_{c_j}, \Sigma_{c_j})$, with

$$\boldsymbol{\mu}_{c_j} = \left((\sigma^2\tau_j^{*2})^{-1} + \sum_{i=1}^{n} (\zeta_{ij}U_i)^T \Sigma_i^{-1} (\zeta_{ij}U_i) \right)^{-1} \left(\sum_{i=1}^{n} (\mathbf{y}_i - \boldsymbol{\mu}_{i(-c_j)})^T \Sigma_i^{-1} (\zeta_{ij}U_i) \right)^T,$$

and

$$\Sigma_{c_j} = \left((\sigma^2\tau_j^{*2})^{-1} + \sum_{i=1}^{n} (\zeta_{ij}U_i)^T \Sigma_i^{-1} (\zeta_{ij}U_i) \right)^{-1}.$$

Now, we derive the conditional posterior distribution for τ_j^2 and λ^2 from the joint posterior distribution. Since

$$\begin{aligned}
& \pi(\tau_j^2 | \mathbf{y}, \mathbf{m}, \mathbf{r}_k, \mathbf{b}_j, \lambda, \mathbf{c}_j, \tau_j^{*2}, \lambda^*, \sigma^2, \rho) \\
\propto\ & \pi(\tau_j^2|\lambda)\pi(\mathbf{b}_j|\tau_j^2, \sigma^2) \\
\propto\ & (\tau_j^2)^{\frac{v+1}{2}-1} \exp\Big(-\tau_j^2 \frac{v\lambda^2}{2} \Big) (\tau_j^2)^{-\frac{v}{2}} \exp\Big(-\frac{1}{2} \mathbf{b}_j^T (\sigma^2 \text{diag}(\tau_j^2, \cdots, \tau_j^2))^{-1} \mathbf{b}_j \Big) \\
\propto\ & \exp\Big(-\tau_j^2 \frac{v\lambda^2}{2} - \frac{1}{2\sigma^2\tau_j^2} ||\mathbf{b}_j||^2 \Big) (\tau_j^2)^{-\frac{1}{2}}, \qquad (4.20)
\end{aligned}$$

and

$$\begin{aligned}
& \pi(\lambda^2 | \mathbf{y}, \mathbf{m}, \mathbf{r}_k, \mathbf{b}_j, \tau_j^2, \mathbf{c}_j, \tau_j^{*2}, \lambda^*, \sigma^2, \rho) \\
\propto\ & \pi(\lambda^2) \prod_{j=1}^{p} \pi(\tau_j^2|\lambda)
\end{aligned}$$

$$\propto \quad (\lambda^2)^{a-1}\exp\Big(-b\lambda^2\Big)\prod_{j=1}^{p}\Big(\frac{v\lambda^2}{2}\Big)^{\frac{v+1}{2}}\exp\Big(-\frac{v\lambda^2}{2}\tau_j^2\Big), \tag{4.21}$$

the posterior distribution for $\frac{1}{\tau_j^2}$ is inverse-Gaussian$\Big(v\lambda^2, \sqrt{\frac{v\lambda^2\sigma^2}{||\mathbf{b}_j||^2}}\Big)$, and the posterior distribution for λ^2 is Gamma$\Big(a+\frac{pv+p}{2}, b+\frac{v\sum_{j=1}^{p}\tau_j^2}{2}\Big)$.

Similarly, the posterior distribution for $\frac{1}{\tau_j^{*2}}$ is inverse-Gaussian$(v\lambda^{*2}, \sqrt{\frac{v\lambda^{*2}\sigma^2}{||\mathbf{b}_j||^2}})$, and the posterior distribution for λ^{*2} is Gamma$(a^*+\frac{pv+p}{2}, b^*+\frac{v\sum_{j=1}^{p}\tau_j^{*2}}{2})$. From these posteriors, we can see that the hierarchical expansion of Multivariate Laplace prior indeed gives closed forms of posterior distributions for efficient Gibbs sampling.

Lastly, if we assume a stationary AR(1) covariance structure, i.e.,

$$\Sigma_i = \sigma^2\Gamma_i = \sigma^2\begin{pmatrix} 1 & \rho^{|t_{i1}-t_{i2}|} & \dots & \rho^{|t_{iT_i}-t_{i1}|} \\ \rho^{|t_{i2}-t_{i1}|} & 1 & \dots & \rho^{|t_{iT_i}-t_{i2}|} \\ \vdots & \vdots & \vdots & \vdots \\ \rho^{|t_{i1}-t_{iT_i}|} & \rho^{|t_{i2}-t_{iT_i}|} & \dots & 1 \end{pmatrix}, \tag{4.22}$$

the posterior distribution for σ^2 is inverse chi-square distribution, or,

$$\pi(\sigma^2|\cdot) \sim \text{Inv-}\chi^2\left(\sum_{i=1}^{n}T_i, \frac{\sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\mu}_i)^T\Gamma_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)}{\sum_{i=1}^{n}T_i}\right), \tag{4.23}$$

where the first parameter is the degree of the freedom parameter, and the second one is the scale parameter, and

$$\begin{aligned}\pi(\rho|\cdot) &\propto \pi(\mathbf{y}|\cdot)\pi(\rho) \\ &\propto \prod_{i=1}^{n}(|\Gamma_i|^{-\frac{1}{2}})\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\mu}_i)^T\Gamma_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)\right).\end{aligned} \tag{4.24}$$

Based on this expression, the corresponding Metropolis-Hastings algorithm can be developed to update ρ.

We use MCMC algorithms to estimate the posterior distribution of each parameter by drawing posterior samples from corresponding conditional posterior distribution, given the current values of all other parameters and the observed

data. This process continues until all chains converge. We use the potential scale reduction factor $\hat{R}$ to access the convergence.

4.5 Examples

4.5.1 Computer Simulation

We first investigate the new Bayesian group lasso approach for selecting significant time-varying effects through simulation studies. We generate data in the *f*GWAS setting according to the model (4.8) with $n = 200$, $q = 1$ and $p = 30$ or 300. For ease of simulation, genotypical data ξ_{ij} is derived from u_{ij} for $i = 1, \cdots, n$ and $j = 1, \cdots, p$, where each u_{ij} has a standard normal distribution marginally, and $cov(u_{ij}, u_{ik}) = 0$. We set

$$\xi_{ij} = \begin{cases} 1, & u_{ij} > c \\ 0, & -c \leq u_{ij} \leq c \\ -1, & u_{ij} < -c \end{cases}$$

to mimic a SNP with equal allele frequencies, where $-c$ is the first quartile of a standard normal distribution. Then, we form indicator matrix ζ_{ij} of dominant effects from ξ_{ij}.

We assume that the dynamic pattern of the trait is controlled by 5 SNPs and 1 covariate. In particular, we set $\mathbf{b}_j = \mathbf{0}$ for $j = 4, \cdots, p$, and $\mathbf{c}_j = \mathbf{0}$ for $j = 1, 2, 6, \cdots, p$. Sex is included as a covariate and is generated by randomly assigning a sex to each subject. The time-varying effects of overall mean, covariate, and significant SNPs are generated by Legendre polynomial, with Legendre coefficients listed in Table 4.1. The order of the Legendre polynomial was set at $v = 3$.

To simulate irregular longitudinal phenotypical data, we assume that the number of measurements for each subject is between 5 and 12, and all subjects are in the age range of 30 to 80 years. For each subject with a specific number of measurements, trait of interest are observed at ages randomly drawn from 30 to 80. The residual covariance matrix among different time points was assumed to be AR(1) with $\sigma = 4$ and $\rho = 0.4$. The phenotypes observed at subject-specific time points and genotypes of all subjects are collected for Bayesian analysis.

Table 4.1. Parameters used in the simulated example.

		Legendre Coefficients			
Time-Varying Effect	Parameter	0	1	2	3
Mean effect	$\mathbf{m}$	13.40	-3.08	1.88	-3.20
Covariate Effect	$\mathbf{r}_1$	3.00	0.15	-2.67	3.25
Additive Effect	$\mathbf{b}_1$	1.04	0.88	-2.05	0.55
	$\mathbf{b}_2$	1.17	-0.22	0.74	-4.72
	$\mathbf{b}_3$	1.40	-2.25	1.00	0.00
Dominant Effect	$\mathbf{c}_3$	1.49	-2.13	4.82	1.42
	$\mathbf{c}_4$	1.00	1.32	1.90	1.50
	$\mathbf{c}_5$	1.26	-1.22	2.70	-1.96

For each simulated data set, we minimized the penalized least square (4.9) by implementing MCMC algorithms as described in Section 4.4. Table 4.2 summarizes the average estimates and the empirical standard errors of Legendre coefficients $\mathbf{m}$, $\mathbf{r}_1$, $\mathbf{b}_j$, $j = 1, 2, 3$ and $\mathbf{c}_j$, $j = 3, 4, 5$ over 100 replications. As can be seen from this table, when $p = 30$, the estimates produced by Bayesian group lasso resemble the least-squares estimates with no shrinkage. When $p = 300$, our method shrinks time-varying effects towards zero, where the amount of shrinkage is automatically determined from the data. To confirm this, Figure 4.1 compares the histograms of the regularization parameters λ for $p = 30$ and $p = 300$, respectively. Clearly, Bayesian group lasso prefers a larger penalty when p is larger, because in high-dimensional problems, the penalties imposed on the size of regression coefficients could avoid overfitting and thus archive superior prediction performance.

To evaluate the variable selection performance of the proposed procedure, we calculate several measures of model sparsity for the final model obtained by the Bayesian group lasso. We say that a SNP is significant and is included in the final model if its L_2 norm is smaller than a cutoff value ϵ. In our implementation, we consider models with different noise levels and set ϵ to 0.1. Simulation results are summarized in Table 4.3. Column "C" shows the average number of SNPs with nonzero varying-coefficients correctly included in the final model, and column

Table 4.2. Parameter estimates in the simulated example.

	Estimated Legendre Coefficients				Empirical SE			
Order	0	1	2	3	0	1	2	3
$n = 200, p = 30$								
$\mathbf{m}$	10.333	-2.045	1.821	-1.463	0.345	0.367	0.296	0.350
$\mathbf{r}_1$	3.056	0.075	-1.983	2.244	0.137	0.361	0.306	0.415
$\mathbf{b}_1$	0.957	0.663	-1.556	0.329	0.270	0.221	0.304	0.235
$\mathbf{b}_2$	1.187	-0.342	0.639	-3.776	0.138	0.300	0.150	0.312
$\mathbf{b}_3$	1.402	-1.910	0.871	-0.024	0.110	0.347	0.264	0.308
$\mathbf{c}_3$	1.637	-2.107	4.030	0.724	0.219	0.328	0.432	0.403
$\mathbf{c}_4$	1.226	0.775	1.362	1.043	0.274	0.264	0.377	0.351
$\mathbf{c}_5$	1.448	-1.012	2.118	-1.280	0.267	0.229	0.578	0.370
$n = 200, p = 300$								
$\mathbf{m}$	6.910	-1.622	2.266	-0.889	0.364	0.249	0.291	0.245
$\mathbf{r}_1$	3.028	-0.030	-1.785	2.073	0.265	0.343	0.693	0.659
$\mathbf{b}_1$	0.445	0.282	-0.507	0.107	0.122	0.165	0.219	0.065
$\mathbf{b}_2$	0.803	-0.043	0.300	-1.589	0.148	0.091	0.176	0.382
$\mathbf{b}_3$	0.750	-0.897	0.322	0.000	0.132	0.231	0.112	0.090
$\mathbf{c}_3$	0.922	-0.829	1.324	0.286	0.190	0.179	0.239	0.156
$\mathbf{c}_4$	0.306	0.147	0.176	0.083	0.120	0.066	0.095	0.064
$\mathbf{c}_5$	0.467	-0.238	0.377	-0.244	0.160	0.100	0.180	0.120

"IC" is the average number of SNPs with no genetic effect incorrectly included in the final model. Column "Under-fit" represents the proportion of excluding any relevant SNP in the final model. Similarly, column "Correct-fit" represents the proportion that the extract true model was selected and column "Over-fit" gives the proportion of including all relevant SNPs as well as one or more irrelevant SNPs. Clearly, the noise level plays a significant role in how well our procedure could select the exact correct model. When the noise level is low, the Bayesian group lasso could select the exact correct model with high probabilities.

Finally, we examine how well the proposed method estimates the time-varying effects. To ameliorate the bias of the parameter estimates introduced by group lasso penalties, we always refit the *f*GWAS model using selected variables in the

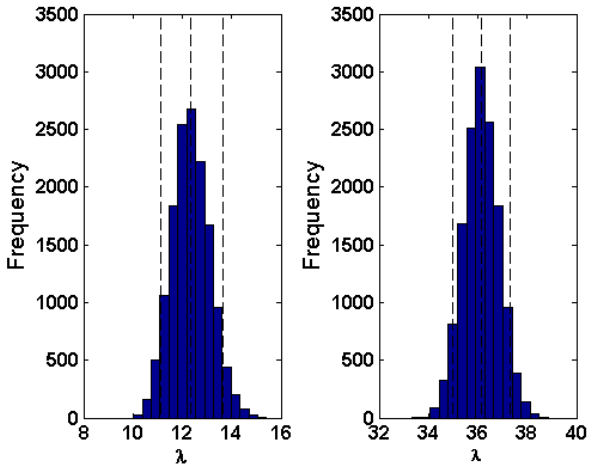

Figure 4.1. Histograms of the posterior samples for λ when $p = 30$ (left) and $p = 300$ (right). The dashed lines represent the posterior 5, 50, and 95% quantiles.

Table 4.3. Variable selection performance in the simulated example.

	No. of Nonzeros		Proportion of		
Noise Level (σ)	C	IC	Under-fit	Correct-fit	Over-fit
$n = 200, p = 30$					
3.00	5.00	0.00	0.00	1.00	0.00
4.00	5.00	0.00	0.00	1.00	0.00
6.00	4.92	0.87	0.11	0.75	0.14
$n = 200, p = 300$					
3.00	5.00	0.00	0.00	1.00	0.00
4.00	4.94	0.06	0.06	0.88	0.06
6.00	4.62	0.65	0.28	0.35	0.37

final model with all regularization parameters being zero. Figure 4.2 shows the estimated time-varying additive effects and dominant effects for the model with $p = 30, \sigma = 4$.

4.5.2 Real Data Analysis

We use the newly developed model to analyze a real GWAS data set from the Framingham Heart Study (FHS), a cardiovascular study based in Framingham,

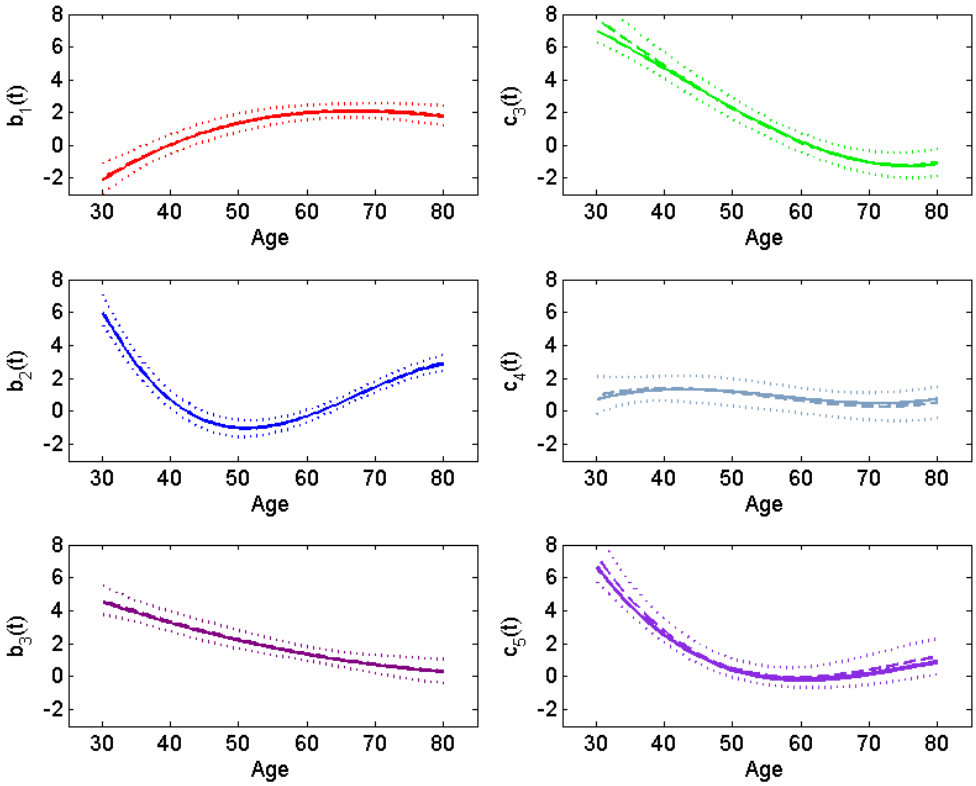

Figure 4.2. True (solid line) and the average of estimated (dashed line) time-varying effects ($\pm$ 2$\times$pointwise standard deviation) over 100 simulations.

Massachusetts, supported by the National Heart, Lung, and Blood Institute, in collaboration with Boston University (Dawber et al. 1951). Recently, 550,000 SNPs have been genotyped for the entire Framingham cohort (Jaquish 2007), from which 493 males and 372 females were chosen for our data analysis. These subjects were measured for body mass index (BMI) at multiple time points from age 29 to age 61. The number of measurements for a subject ranges from 2 to 18, and the intervals of measurement are also highly variable among subjects. As is standard practice, SNPs with rare allele frequency $< 10\%$ were excluded from data analysis. The numbers and percentages of non-rare allele SNPs vary among different chromosomes and ranges from 4,417 to 28,771 and from 0.64 to 0.72, respectively.

A single-SNP analysis was used to analyze the phenotypic data of BMI for males and females separately. Figure 4.3 gives $-log_{10}$ p-values for each SNP in the two sexes, from which 756 SNPs with p-values greater than 4.0 in at least one

sex were selected. Before applying Bayesian group lasso analysis to this irregular longitudinal data set, we imputed missing genotypes for a small proportion of SNPs (5.16%) according to the distribution of genotypes in the population. Then, by treating the sex as a covariate, we imposed group lasso penalties on both additive effects and dominant effects in hopes of identifying SNPs with notable effects on BMI, when all effects are possibly the functions of time. We find the best order of Legendre polynomial is $v = 3$ to fit the longitudinal data set based on the Bayesian information criterion.

In total, there are 11 significant SNPs selected by the Bayesian group lasso, which are located on chromosomes 2, 4, 5, 6, 12, 13, 14, 15, and 19. Table 4.4 tabulates the names, positions, alleles, the posterior median estimates of the Legendre coefficients of these SNPs, as well as their standard errors. The first allele in column "Alleles" represents the minor allele. Using the refitted Legendre coefficients of significant SNPs, we plot the age-specific changes of additive and dominant effects of these SNPs after adjusting for the covariate effect (Figure 4.4). While most SNPs display significant additive genetic effects in a time course, only a couple of SNPs exert significant dominant effects. Some of these detected SNPs are detected to locate in a similar region of candidate genes for obesity. For example, the detected SNPs ion chromosomes 4, 6, and 12 are close to candidate genes for BMI-related type 2 diabetes (Frayling 2007).

In the proposed two-level Bayesian hierarchical model, the amount of shrinkage in the estimates of time-varying effects depends on two hyperparameters λ and λ^* in the priors, and the values of λ and λ^* are determined by the data. Figure 4.5 shows the histograms of hyperparameters λ and λ^*, where the posterior medians for λ and λ^* are 55.745 and 55.575, respectively, with the 95% posterior intervals being [54.636, 56.938] and [54.360, 56.807], respectively. We can see that these two lasso parameters for the additive effects and dominant effects can be estimated with quite high precision.

Table 4.4. Information about significant SNPs in the real data example.

				Estimated Legendre Coefficients			
chr	Name	Position	Alleles	0	1	2	3
Additive Effect							
2	ss66418612	131490656	C/T	-1.418	0.170	0.381	0.301
				(0.475)	(1.020)	(1.009)	(0.660)
5	ss66055735	180435974	G/A	-7.212	-4.667	-4.397	-2.577
				(1.195)	(2.590)	(2.447)	(1.424)
6	ss66478784	82559928	C/G	-1.172	0.533	0.424	-0.059
				(0.430)	(0.899)	(0.873)	(0.588)
12	ss66340721	9779820	A/G	-0.916	0.851	1.540	0.509
				(0.541)	(1.115)	(1.076)	(0.737)
13	ss66176305	33454138	C/G	-0.364	0.392	0.440	0.033
				(0.284)	(0.606)	(0.592)	(0.395)
14	ss66500898	26550323	T/C	-0.833	-0.980	-0.517	0.133
				(0.262)	(0.559)	(0.555)	(0.387)
15	ss66418740	50993515	C/G	-0.699	-0.203	-0.468	-0.400
				(0.334)	(0.718)	(0.705)	(0.476)
15	ss66329929	78402022	A/G	-2.390	1.548	1.733	3.010
				(0.699)	(1.314)	(1.259	(0.982)
19	ss66397067	7528744	G/A	-0.774	0.232	0.091	0.101
				(0.293)	(0.617)	(0.608)	(0.413)
Dominant Effect							
4	ss66058920	95891432	T/C	1.332	0.903	0.915	0.613
				(0.483)	(1.074)	(1.040)	(0.656)
19	ss66264538	33301987	T/C	0.928	0.124	-0.480	-0.518
				(0.600)	(1.299)	(1.303)	(0.834)

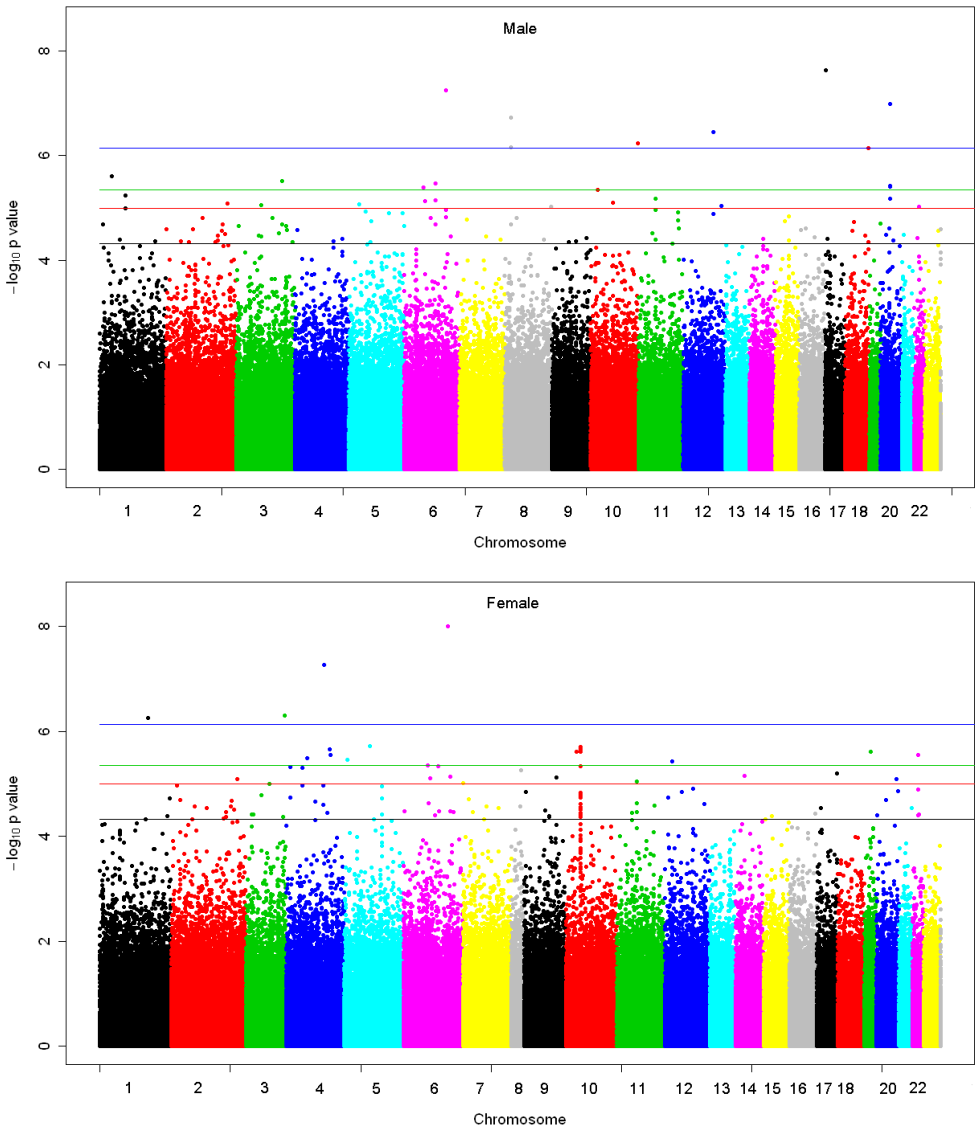

Figure 4.3. Manhattan plot of p-values for association by genomic position for different sexes.

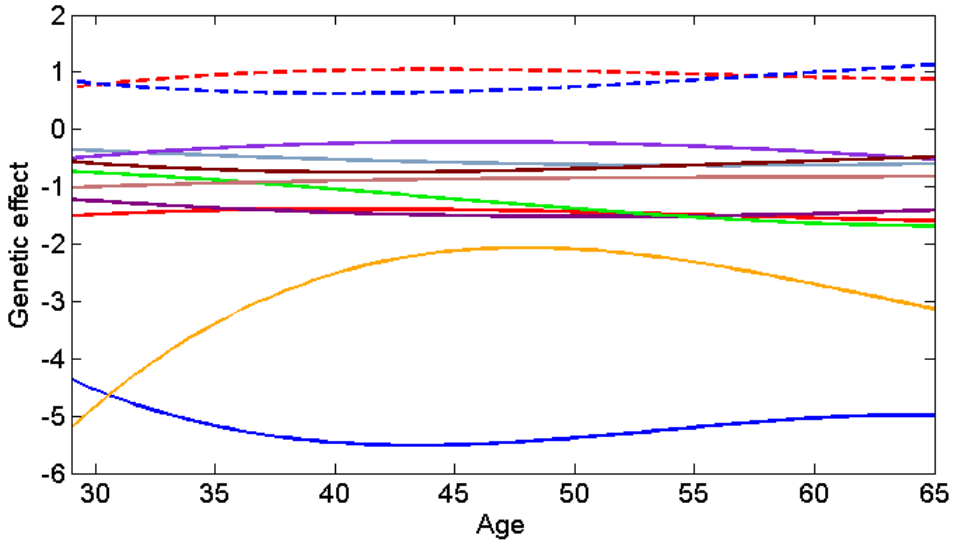

Figure 4.4. Additive (solid line) and dominant effects (dashed line) of significant SNPs in the real data example.

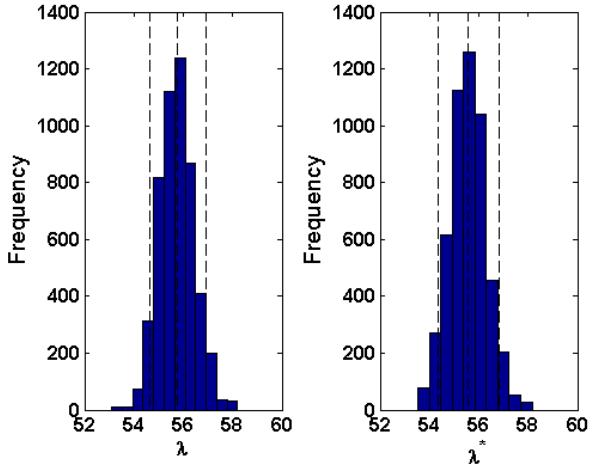

Figure 4.5. Histograms of the posterior samples for group lasso parameters. The dashed lines represent the posterior 5, 50, and 95% quantiles.

Chapter 5

Discussion

5.1 Summary of Findings

When the number of predictors p is much larger than the number of observations n, penalized regression models are favorable to identify nonzero coefficients, enhance model predictability, and avoid overfitting (Hastie 2009). The L_1 penalized regression or lasso is one of the most popular techniques, and is fundamental to the computations of other penalized regressions. This dissertation developed various variable selection models and computational algorithms, and applied these procedures to genome-wide association studies.

In Chapter 2, we presented a Bayesian hierarchical model with lasso penalties to simultaneously fit and estimate all possible genetic effects associated with all SNPs in a GWAS, adjusting for both discrete and continuous covariates. Lasso penalties are imposed on the additive and dominant effects, and implemented by assigning double-exponential priors to their regression coefficients. It shrinks small effects towards zero and produces sparse solutions. In this framework, SNPs with significant genetic effects can be selected accurately.

We fit the model in a fully Bayesian setting, employing the MCMC algorithm to generate posterior samples from the joint posterior distribution, which can be used to make various posterior inferences. Although computationally intensive, it is easy to implement and provides not only point estimates but also interval estimates of all parameters. The Bayesian lasso treats tuning parameters as unknown hyperparameters and generates their posterior samples when estimating

other parameters. This technique avoids the choice of tuning parameters, and automatically accounts for the uncertainty in its selection that affects the estimation of the final model. By contrast, standard lasso algorithms usually select tuning parameters by K-fold cross-validation, which involves partitioning the whole data set and refitting the model many times. This process may result in unstable tuning parameter estimates.

In order to improve the performance of lasso when p is greater than n in the context of GWAS analysis, preconditioning is considered before variable selection. Preconditioning encourages the principal components of a reduced design matrix to be highly correlated with the response, and thus in most cases only the first or first few components tend to be useful for prediction. It denonises the response variable so that variable selection becomes more efficient. Our simulation demonstrated that when p greatly exceeds n, preconditioned Bayesian lasso could successfully identify almost all the SNPs with true genetic effects.

By analyzing real data, the preconditioned Bayesian lasso is shown to produce biologically relevant results. For example, the approach detected a significant SNP ss66171460 at position 22580931 of chromosome 20 associated with BMI. It is interesting to note that this SNP is within 500Kb of the FOXA2 (Forkhead Box A2) gene, an important genetic variant that regulates obesity (Wolfrum et al., 2003).

One simulation example of Paul et al. (2008) implies that, in the context of genome-wide association studies, SNPs that are marginally independent of the phenotype could be screened out by preconditioning, but can be identified by standard variable selection techniques such as lasso or Bayesian lasso. However, since most of the SNPs are correlated with the phenotype through marginal correlations, we believe preconditioning step is worthwhile to identify more important SNPs. What is more, the preconditioned Bayesian lasso is always advantageous over single SNP analysis, since when one SNP is analyzed at a time, we are always testing the marginal correlation between the predictor and response.

In Chapter 3, we proposed another two-stage variable selection procedure and applied it to the same GWAS data set. We first employed sure independence screening to reduce the dimensionality of feature space. This step largely decreases the computational burden as well as the false positive rate in the following

variable selection step. Since predictors with true nonzero regression coefficients are included in the reduced feature space with an overwhelming probability, the performance of variable selection is also guaranteed. In the second step, we considered SCAD regularization method to select important SNPs in a GWAS and estimate all possible genetic effects simultaneously, adjusting for both discrete and continuous covariates. Specifically, SCAD penalties are imposed on the additive genetic effects. The SCAD penalized regression shrinks small effects towards zero and produces sparse unbiased estimates. Therefore in this framework, SNPs with significant genetic effects can be identified more efficiently and accurately.

The optimization of SCAD penalized least squares is implemented based on the local linear approximation of SCAD penalty, which was developed by Zou and Li (2008) in order to overcome several drawbacks of local quadratic approximation. Based on the local linear approximation of penalty functions, we reformulated this problem as the Bayesian lasso and solved it by Gibbs samplers presented in Chapter 2. The MCMC-implemented Bayesian SCAD algorithm could estimate all parameters in SCAD penalized least squares, from which a subset of significant predictors could be selected. The Bayesian SCAD also treats tuning parameters as unknown hyperparameters and generates their posterior samples when estimating other parameters.

Our simulation studies imply that, in the context of genome-wide association studies, SIS based Bayesian SCAD outperforms single SNP analysis and preconditioned Bayesian lasso in terms of parameter estimation, variable selection and computational efficiency. Since preconditioning gives a new version of response variable but retains all potential predictors, all predictors will enter the variable selection step and thus the computational cost remains after preconditioning. On the contrary, SIS provides a subset of potential predictors for the following variable selection step, and thus the computational burden is significantly reduced. Moreover, since lasso regression gives biased estimates, we may want to refit the linear regression model without the penalty terms using only selected SNPs. This extra step incurs additional computational cost.

In Chapter 4, we proposed a Bayesian regularized estimation procedure for nonparametric varying-coefficient models that could simultaneously estimate time-varying effects and implement variable selection. The procedure extends the stan-

dard Bayesian lasso (Park and Casella 2008) and standard group lasso (Yuan and Lin 2006) to a nonparametric setting, and is applicable to irregular longitudinal data.

We approximated time-varying effects by Legendre polynomials, and presented a Bayesian hierarchical model with group lasso penalties that encourage sparse solutions at the group level. The group lasso penalties are introduced by assigning multivariate Laplace priors to regression coefficients, and are implemented on the basis of its hierarchical expansion which yields an efficient Gibbs sampler in the MCMC estimation. The MCMC algorithms generate posterior samples from the joint posterior distribution of all parameters, which can be used to make various posterior inferences. Although computationally intensive, it outperforms the standard group lasso in the sense that it provides not only point estimates but also interval estimates of time-varying effects.

In one of the most powerful but challenging areas in genetics, we incorporated our new procedure to genome-wide association studies (GWAS) by testing a large number of SNPs simultaneously, particularly with $p \gg n$, based on the dynamic pattern of genetic effects triggered by each SNP. In this application, traditional GWAS is integrated into functional data analysis, leading to a new analytical framework, functional genome-wide association studies or *f*GWAS. We embedded Bayesian group lasso within the *f*GWAS setting, facilitating the test and characterization of time-varying effects of genetic variants on complex phenotypes or diseases. We first applied the new approach to *f*GWAS for age-specific changes of BMI and successfully identified several significant SNPs, some of which are confirmed by empirical genetic studies (Frayling 2007). Simulation studies indicated that the proposed procedure is very effective in estimating the smooth regression coefficient functions and selecting significant predictors. When $p > n$, it could identify nonzero time-varying effects by highly regularized approach and automatically determine the amount of regularization from the data. Results from the application to functional genome-wide association studies indicate that the procedure could effectively select predictors that exhibit notable effects on the response over time.

5.2 Challenges for Future Research

So far we have concentrated on two-stage variable selection methods for continuous traits in GWAS. The proposed procedures and MCMC algorithms can be readily extended to survival data analysis and penalized logistic regression in case−control disease gene mapping. Moreover, when two or more traits are collected for each subject, say blood pressure and body mass index, these traits may interact with each other or be controlled by one or more SNPs. For example, Gudmundsson et al (2007) found that the same variants in TCF2 influence risk to both type 2 diabetes and prostate cancer. Therefore, it is interesting to jointly model these two traits by introducing a covariance structure in our variable selection procedures and test for gene pleiotropy.

Recently, epistatic interactions among multiple genes are believed to be responsible for many common diseases, instead of single genetic variant in one gene. Wu et al. (2009) suggested to look for gene-gene interaction effects after identifying main effects through variable selection procedures. However, this strategy may overlook important interactions between loci that did not display noticeable marginal effects. With our two-stage variable selection procedures for high-dimensional problems, we may conduct exhaustive search to find significant epistatic interactions more efficiently, or develop other procedures without exhaustive search. The model with a capacity to identify epistatic interactions will enables geneticists to decipher a detailed picture of the genetic architecture of a complex trait.

What is more, our current sure independence screening is based on a linear assumption between each SNP and the response. In the future, we may relax this model assumption by incorporating nonlinear independence screening techniques in the first stage.

On Bayesian group lasso method for irregular longitudinal data, we have approximated nonparametric time-varying effects by Legendre polynomials. From a theoretical point of view, the proposed method can also approximate the varying-coefficients by other nonparametric techniques including basis expansions, and model the within-subject correlation by other parametric or nonparametric covariance structures, such as ARMA(1,1), compound symmetry and ante-dependence covariance structures. Given its potential influence, an optimal model for longi-

tudinal covariance structure should be chosen in terms of the nature of practical data (Zhao et al. 2005; Yap et al. 2009).

More generally, the Bayesian group lasso can be easily extended to problems where the number of variables in each group varies, such as the multi-factor ANOVA with each factor having several levels. Practically, the proposed procedure and MCMC algorithms can also be readily extended to multivariate longitudinal data with more than one trait of interest, or high-dimensional phenotypes such as human face and tumor growth.

Bibliography

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle, *Second International Symposium on Information Theory* (Edited by B. N. Petrov and F. Csaki), 267-281. Akademiai Kiado, Budapest.

Andrews,D.F., and Mallows,C.L. (1974) Scale mixture of normal distributions, *J. R. Stat. Soc. Ser. B*, 36, 99-102.

Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006) Prediction by supervised principal components, *J. Amer. Statist. Assoc.*, 101, 119-137.

Chen,S.S. (1998) Atomic decomposition by basis persuit, *SIAM J. Sci. Comput.*, 20, 33-61.

Daly, A. K. (2010) Genome-wide association studies in pharmacogenetics, *Nat. Rev. Genet.*, 11, 241-246.

Daubechies, I., Defrise, M. and De Mol, C. (2004) An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, *Comm. Pure Appl. Math.*, 57, 1413-1457.

Dawber, D., Meadors, G. and Moore F. (1951) Epidemiological approaches to heart disease: the Framingham Study, *Am J Public Health*, 41, 279-293.

Donnelly, P. et al. (2008) Progress and challenges in genome-wide association studies in humans, *Nature*, 465, 728-731.

Duerr, R. H. et al. (2006) A genomewide association study identifies IL23R as an inflammatory bowel disease gene, *Science*, 314, 1461-1463.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression (with discussion), *Ann. Statist.*, 32, 407-499.

Fan,J., and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Amer. Statist. Assoc.*, 96, 1348-1360.

Fan, J. and Lv, J. (2008) Sure independence screening for ultrahigh dimensional feature space (with discussion), *J. R. Stat. Soc. Ser. B*, 70, 849-911.

Frank, I. E., and Friedman, J. H. (1993) A statistical view of some chemometrics regression tools, *Technometrics*, 35, 109-148.

Frayling, T. (2007), Genome-wide Association Studies Provide New Insights into Type 2 Diabetes Aetiology, *Nat. Rev. Genet.*, 8, 657-662.

Fu, W. J. (1998) Penalized regression: the bridge versus the LASSO, *J. Comput. Graph. Statist.*, 7, 397-416.

Gelman, A. and Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences, *Stat. Sci.*, 7, 457-511.

Gudmundsson, J. et al. (2007) Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes, *Nature Genet.*, 39, 977-983.

Gudmundsson, J. et al. (2008) Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer, *Nature Genet.*, 40, 281-283.

Hampe, J. et al. (2007) A genome-wide association scan of non-synonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1, *Nature Genet.*, 39, 207-211.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) High-Dimensional Problems: $p > N$. *The elements of statistical learning*. 2nd edn. Springer, New York.

Hindorff, L., Sethupathyb, P., Junkinsa, H., Ramosa, E., Mehtac, J., Collinsb, F. and Manolioa, T. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits, *PNAS*, 106, 9362-9367.

Hoggart, C., Whittaker, J., De lorio, M. and Balding, D. (2008) Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies, *PLoS Genet.*, 4(7), e1000130.

Hoti, F., and Sillanpaa, M. J. (2006) Bayesian mapping of genotype $\times$ expression interactions in quantitative and qualitative traits, *Heredity*, 97, 4-8.

Hunter, D. and Li, R. (2005) Variable selection using mm algorithms, *Ann. Statist.*, 33, 1617-1642.

Jaquish, C. (2007) The Framingham Heart Study, on its way to becoming the gold standard for Cardiovascular Genetic Epidemiology, *BMC Med. Genet.*, 8, 63.

Kim, Y., Choi, H. and Oh., H. (2008) Smoothly clipped absolute deviation on high dimensions, *J. Amer. Statist. Assoc.*, 103, 1665-1673.

Lin, Y. and Zhang, H. H. (2006) Conponent selection and smoothing in smoothing spline analysis of variance models - COSSO, *Ann. Statist.*, 34, 2272-2297.

Logsdon, B.A., Hoffman, G.E. and Mezey, J.G. (2010) A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis, *BMC Bioinformatics*, 27, 11-58.

Ma, C., Casella, G. and Wu, R. (2002) Functional Mapping of Quantitative Trait Loci Underlying the Character Process: a Theoretical Framework, *Genetics*, 161, 1751-1762.

Mallows, C. L. (1973) Some comments on C_p, *Technometrics*, 15, 661-675.

McCarthy, M., Abecasis, G., Cardon, L., Goldstein, D., Little, J., Ioannidis, J. and Hirschhorn, J. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges, *Nat. Rev. Genet.*, 9(5), 356-369.

Moffatt, M. F. et al. (2007) Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma, *Nature*, 448, 470-473.

Park,T., and Casella, G. (2008) The Bayesian lasso, *J. Amer. Statist. Assoc.*, 103, 681-686.

Paul, D., Bair, E., Hastie, T. and Tibshirani, R. (2008) Preconditioning for feature selection and regression in high-dimensional problems, *Ann. Statist.*, 36, 1595-1618.

Rosset, S., and Zhu, J. (2007) Piecewise Linear Regularized Solution Paths, *Ann. Statist.*, 35, 1012-1030.

Schwartz, G. (1978) Estimating the dimension of a model, *Ann. Statist.*, 6, 461-464.

Shuldiner, A. R. et al. (2009) Association of cytochrome P450 2C19 genotype with the antiplatelet effect and clinical efficacy of clopidogrel therapy, *J. Amer. Med. Assoc.*, 302, 849-857.

Sladek, R. et al. (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes, *Nature*, 445, 881-885.

Steinthorsdottir, V. et al. (2007) A variant in CDKAL1 influences insulin response and risk of type 2 diabetes, *Nature Genet.*, 39, 770-775.

Takeuchi, F. et al. (2009) A genome-wide association stud confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose, *PLoS Genet.*, e1000433.

Teichert, M. et al. (2009) A genome-wide association study of acenocoumarol maintenance dosage, *Hum. Mol. Genet.*, 18, 3758-3768.

Thomas, G. et al. (2008) Multiple loci identified in a genome-wide association study of prostate cancer, *Nature Genet.*, 40, 310-315.

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

Wang, L., Li, H. and Huang, J. (2008) Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements, *J. Amer. Statist. Assoc.*, 103, 1556-1569.

Weedon, M. N. et al. (2007) A common variant of HMGA2 is associated with adult and childhood height in the general population, *Nature Genet.*, 39, 1245-1250.

Wolfrum, C., Shih, D. Q., Kuwajima, S., Norris, A. W., Kahn, C. R. and Stoffel, M. (2003) Role of Foxa-2 in adipocyte metabolism and differentiation, *J. Clin. Invest.*, 112, 345-356.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E. and Lange, K. (2009) Genome-wide association analysis by lasso penalized logistic regression, *Bioinformatics*, 25, 714-721.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S. et al. (2010) Common SNPs explain a large proportion of the heritability for human height, *Nature Genet.*, 42, 565-569.

Yap, J., Fan, J. and Wu, R. (2009), Nonparametric Modeling of Covariance Structure in Functional Mapping of Quantitative Trait Loci, *Biometrics*, 65, 1068-1077.

Yi,N., and Xu, S. (2008) Bayesian lasso for quantitative trait loci mapping, *Genetics*, 179, 1045-1055.

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

Zhang, H. H. and Lin, Y. (2006) Conponent selection and smoothing for nonparametric regression in exponential families, *Statistica Sinica*, 16, 1021-1042.

Zhao, W., Chen, Y., Casella, G, Cheverud, J. and Wu, R. (2005) A Nonstationary Model for Functional Mapping of Complex Traits, *Bioinformatics*, 21, 2469-2477.

Zhao, P. and Yu, B. (2006) On model selection consistency of Lasso, *J. Mach. Learn. Res.*, 7, 2541-2563.

Zou, H. (2005) The adaptive lasso and its oracle properties, *J. Amer. Statist. Assoc.*, 101, 1418-1429.

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net, *J. R. Stat. Soc. Ser. B*, 67, 301-320.

Zou, H. and Li, R. (2008) One-step sparse estimates in nonconcave penalized likelihood models (with discussion), *Ann. Statist.*, 36, 1509-1566.

JIAHAN LI

330 Thomas Building
Department of Statistics
Pennsylvania State University
University Park, PA 16802

Phone: (814) 441-0080
E-mail: jiahanli@psu.edu
http://www.personal.psu.edu/jzl185

Education

Ph.D. in Statistics, 2008 (relocated with my advisor) - 2011
Department of Statistics, Pennsylvania State University, PA
Advisors (joint): Professor Rongling Wu and Professor Runze Li

M.S. in Statistics, 2006 - 2008
Department of Statistics, University of Florida, FL

B.S. 2002 - 2006
Shanghai Jiao Tong University, Shanghai, China

Experience

- **Research Assistant**
  Penn State University, 01/2009 –
  Research assistant of the Statistics Department working on statistical decision theory, high-dimensional predictive models, nonlinear predictive models, and panel data analysis.

  University of Florida, 08/2006 – 12/2008
  Research assistant of the Statistics Department working on computational biology, Bayesian modeling and panel data analysis.

- **Teaching**
  Penn State University, 01/2010 – 05/2010
  Instructor of STAT 460: Intermediate Applied Statistics. Topics covered include basic statistics and probability, linear regression model, logistic regression model and nonparametric statistics.

- **Statistical Consultant**
  Penn State University, 08/2009 – 05/2010
  Work with PhD students and faculty members outside the Department of Statistics, design experiments or survey, analyze data, visualize and interpret statistical results.

  University of Florida, 08/2008 – 12/2008
  Statistical consultant for *2009 State of the Field Report: Arts in Healthcare*, Society for the Arts in Healthcare. Provide a statistical assessment of the benefits of literary, performing, and visual arts in healthcare services.

Awards

- Student Travel Grant for Joint Statistical Meetings, 2010
- The Finalist of Best Paper Competition Award of the service science section, INFORMS, 2009