HIGH DIMENSIONAL STATISTICAL METHODS FOR GENE-ENVIRONMENT INTERACTIONS

By

Cen Wu

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Statistics – Doctor of Philosophy

2013

ABSTRACT

HIGH DIMENSIONAL STATISTICAL METHODS FOR GENE-ENVIRONMENT INTERACTIONS

By

Cen Wu

The genetic influences on complex disease traits generally depends on the joint effects of multiple genetic variants, environment factors, as well as their interplays. Gene×environment (G×E) interactions play vital roles in determining an individual's disease risk, but the underlying genetic machinery is poorly understood. Traditional analysis assuming linear relationship between genetic and environmental factors, along with their interactions, is commonly pursued under the regression-based framework to examine G×E interactions. This assumption, however, could be violated due to nonlinear responses of genetic variants to environmental stimuli. As an extension to our previous work on continuous traits, we proposed a flexible varying-coefficient model for the detection of nonlinear G×E interaction with binary disease traits. Varying coefficients were approximated by a non-parametric regression function through which one can assess the nonlinear response of genetic factors to environmental changes. A group of statistical tests were proposed to elucidate various mechanisms of G×E interaction. The utility of the proposed method was illustrated via simulation and real data analysis with application to Type 2 Diabetes.

It has been increasingly recognized the power of genetic variant set based association analysis over the single variant based approach. We develop a variant set based approach to examine how variants in a genetic system mediated by a common environment factor to affect the phenotype response. The problem can be approached from a high dimensional variable selection perspective. In particular, we can select genetic variants with varying, non-zero

constant and zero coefficients, which are corresponding to cases of G×E interactions, no G×E interactions and no genetic effects, correspondingly. The procedure was implemented in a two stage iterative framework via Smoothly Clipped Absolute Deviation (SCAD) penalty. With proper regularity conditions, we can establish the consistency in variable selection and effect separation of our two stage iterative estimator, as well as the optimal convergence rates of the estimates for varying effect. In addition, it can be shown that the estimate of non-zero constant coefficient enjoys the oracle property. The utility of our procedure will be demonstrated through extensive simulation study and real data analysis.

Due to the drawback of local quadratic approximations in the aforementioned two-stage framework, the approach is not efficient in handling cases when the dimension p is very large. A group coordinate descent (GCD) based approach was proposed within the framework, which is computationally efficient particularly for high dimensional problems where $p > n$, because the computational complexity increases only linearly with the number of predictor groups after basis expansion. The advantage of our method is demonstrated through extensive simulation study and real data analysis.

I dedicate this thesis to my parents, Guizhi Ye and Liyang Wu.

ACKNOWLEDGMENTS

First and foremost, I am deeply grateful to my advisor, Dr. Yuehua Cui, for his tremendous assistance, constant support and extreme patience. He is always approachable and ready to help in every facet of life. Dr. Cui led me into the area of statistical genetics and inspired me through numerous discussions. His training in the past five years helped me not only grasp the skills needed for the interdisciplinary research in statistics and biology, but also gradually develop the capability of thinking independently. Without his excellent guidance and advisement, this thesis would not have been possible.

I would like to thank Dr. Robin Buell, my Quantitative Biology Ph.D program co-advisor, for her valuable suggestions and guidance that significantly improve my background in biology and broaden my training in this interdisciplinary area. My sincere thanks also goes to my thesis committee members: Dr. Lifeng Wang, Dr. Ping-Shou Zhong and Dr. Qing Lu. My transition to the research area of penalized regressions cannot be so smooth without frequent discussions with Dr. Wang. I am also indebted to Dr. Zhong's help in picking up skills necessary for the technical proof in this thesis. I appreciate all the insightful comments they made on this work.

I would also like to extend my sincere gratitude to all the faculty and staff in our department. In particular, my thanks goes to Dr. Dennis Gilliland for his constant support and encouragement, especially for my job hunting. Statistical consulting is a crucial part of my Ph.D training, and I am grateful to Dr. Robert J. Tempelman and Dr. Sasha Kravchenko for accepting me as a consultant at CANR Statistical Consulting Center. I benefited enormously from our weekly group meetings. I also thank their assistance for my job applications.

My special thanks goes to Qi Yan, now at Minnesota, for her tremendous help, especially

during my preparation for the qualify exams. I am also grateful to the assistance and support from Drs. Gengxin Li, Shaoyu li, Xiaoqin Tang, Wei Wang and Ming Gu. So lucky to have Gengxin and Shaoyu as my academic sisters. They offered me generous help and precious suggestions during each stage of my Ph.D life. I appreciate Tao He, Honglang Wang, Bin Gao and Ran Cao, who are the current members in Dr. Cui's group, for creating the stimulating research atmosphere. The memories of heated discussions we had in our weekly journal clubs will never fade away. I also want to thank Shaoyu and Tao for providing me a cozy office environment.

Last, but definitely not least, I would like to express my deepest gratitude to my parents, Guizhi Ye and Liyang Wu, for their endless love and undying support. I was not able to overcome insurmountable obstacles without them, since they never gave me up under any circumstances. Besides, I thank Xin Tan and Ling Gong who made the winter in Michigan no longer tough for me. I spent so many weekends and holidays with them in the last five years, which turned driving on I-96 one of the most pleasant things for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 A review of basic genetics

Genes are the basic functional units where the biological characteristics can be passed from parents to offspring. The genes of each cell are arranged in linear order on chromosomes. Majority of the multicellular organisms have duplicate copies of each gene, hence they are diploid. The number of paired chromosomes varies across different species. For instance, *Brassica Oleracea* has 9 pairs of chromosomes, *Zea Mays* has 10 pairs, Mus musculus has 20 pairs and human beings have 23 pairs. The entire set of chromosomes form the genome of a particular organism and organelles.

Each gene resides on certain site, or locus of the chromosomes. For diploid organisms, the genes corresponding to the locus take two forms (say A and a), called alleles. The three genetic compositions of the two alleles, AA, Aa, and aa, are genotypes. The pair of identical alleles (AA or aa) and different alleles (Aa) are called homozygous and heterozygous respectively. One or several loci may determine the observable characteristics or phenotypic traits, such as eye color, body weight, blood pressure, and so on.

The traits that can be continuously measured are defined as quantitative traits, like the aforementioned body weight and blood pressure. The investigation of the genetic basis of a quantitative trait is the major task of quantitative genetics. In classical quantitative genetics,

the phenotypic value (P) is due to genetic factors (G) and environment factors (E):

$$P = G + E \tag{1.1}$$

By assuming the independence of the terms in (1.1), the phenotypic variance of a quantitative trait (V_P) can be decomposed into its genetic (V_G) and environmental (V_E) variance components:

$$V_P = V_G + V_E \tag{1.2}$$

Based on the three modes of gene actions(additive, dominance and epistasis), we can further partition V_G as

$$V_G = V_{Ga} + V_{Gd} + V_{Ge} \tag{1.3}$$

where V_{Ga}, V_{Gd}, V_{Ge} are additive, dominance and epistatic(or interaction) genetic variance respectively. V_{Gd} and V_{Ge} are referred to as nonadditive genetic variance. In quantitative genetics, heritability characterizes the relative importance of the role genetic variance plays in determining the phenotypic variance. The two types of heritability, broad-sense heritability (H^2) and narrow-sense heritability (h^2) , are defined separately as:

$$H^2 = \frac{V_G}{V_G + V_E} \tag{1.4}$$

and

$$h^2 = \frac{V_{Ga}}{V_G + V_E} \tag{1.5}$$

The degree of overall genetic control over the quantitative trait can be measured by the two heritability parameters H^2 and h^2 [1, 2].

1.2 Genetic mapping of complex traits

The past decades witnessed waves of breakthroughs in genetic mapping of complex diseases (traits). The family-based linkage study prevails as conventional means of locating disease genes. Transmission disequilibrium test (TDT)[3], the most popular association test in family based study, was proposed to test the linkage between a genetic marker and disease susceptibility. However, large pedigrees are needed for fine-mapping in linkage study, so the utility of the study is confined when such information is not available.

While the linkage study predominated in discovering disease variants with major effects, the population based association study gained advantage in detecting genes with modest effects [4], which is of particular significance for complex human diseases[5]. The rapid progress in the high-throughput genotyping technologies made possible the large scale, highly dense genome-wide association studies (GWAS) for millions of genetic variants, such as single-nucleotide polymorphism (SNP) across the entire human genome[6].

The GWAS has thus identified a large number of susceptibility variants associated with the complex human diseases, including asthma[7], breast cancer [8, 9], coronary heart disease [10, 11] and Type 2 Diabetes[12, 13]. A detailed list is given in [14]. Despite the huge success achieved in GWAS, the disease etiology is still not clearly elucidated since a substantial proportion of the heritability remains obscure. Therefore, there are pressing needs to investigate the part of unexplained heritability.

There are a number of potential sources of missing heritability, such as rare variants, structural variation, gene-gene (G×G) interactions and gene-environment (G×E) interactions [15]. For example, the 'Common Disease, Common Variants' hypothesis is commonly adopted in GWAS, assuming that majority of the genetic risk of common complex dis-

eases can be attributed, at least partially, to common disease variants which is of more than 5% minor allele frequency (MAF) [16, 17]. The rare (MAF<0.5%) and low frequency (0.5%<MAF<5%) variants cannot be well captured by GWAS commercial genotyping arrays which are aimed at covering common genetic variants [18]. Detection of such rare and low frequency variants using next generation sequencing technologies [19, 20, 21, 22]on either the genome regions of interest or the whole genome will lead to a better interpretation of heritability.

Compared to the detection of main effects of genetic variants in GWAS, the discovery of high order effects, like G×G and G×E interactions, requires much higher statistical power and thus impinges on the effort to better understand the missing heritability not attributable to the identified disease variants [15, 23]. Therefore investigations on G×G, G×E interactions will help make the best use of GWAS, improving disease prediction, prevention and treatment.

1.3 Gene-environment (G×E) interactions

1.3.1 Basics of G×E interactions

G×E interactions can be traced back to [24, 25] and were formulated by Falconer in [26], which refers to how phenotypes are reactive to different genotypes under various environmental conditions. It has been widely acknowledged that not only the genetic and environmental factors themselves, but also the interactions between them, are involved in the genetic basis of complex diseases [27]. A growing number of evidence of G×E interactions have been found in a wide range of complex diseases, such as colorectal cancer[28], chronic beryllium lung disease[29], skin cancer[30], cardiovascular diseases [31]and psychiatry diseases [32].

The environment factor in a G×E interaction study design can be defined either continuously or discretely. For instance, in a study of G×E interaction on skin cancer, the intensity of sunlight to which the skins are exposed could be defined as an environment condition with a continuous measure, and the risk of developing skin cancer triggered by a specific gene might be quite different given various amount of sunlight. For a G×E interaction study design related to myocardial infarction, drinking status can be defined as an environment factor with 2 categories, coded as 1 (drinking alcohol) and 0 (not drinking alcohol).

Study designs for G×E interaction mainly fall into two categories: the family-based study and the population-based association study[27]. In family-based studies, direct estimations on particular G×E interaction are possible if the environmental information is integrated into designs, such as the pedigree or sib-pair designs. Compared to the association studies in unrelated subjects, the family-based study may demand more effort to gather information needed in the design. Depending on the timing of information collection on environmental,dietary and lifestyle variables, the typical case-control association studies could be further categorized into retrospective (data collection after disease diagnosis) and prospective studies (data collection at the beginning of study). Refer to [27] for a detailed discussion on pros and cons for all these designs.

1.3.2 Challenges and issues in G×E interactions

The major challenges in the study of G×E interactions are how to appropriately model and test G×E effects. The multifactor dimensionality reduction (MDR)-based approach, a data mining method for identifying interactions among independent variables, was proposed in [33] to examine gene-environment interactions. While MDR can be considered as a model-free approach, other statistical methods were developed within the traditional regression

framework, such as the parametric [34] and semi-parametric methods [35, 36].

A common issue in current parametric models for G×E interactions, as pointed out in [37], is that a strong linear assumption on G×E interactions is demanded. Recall that the linear regression model below is the starting point to examine G×E interactions:

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 G + \alpha_3 GX + \varepsilon \tag{1.6}$$

Here Y is the continuous response, α_0 is the overall mean, $\alpha_2, \alpha_3, \alpha_4$ are the effects of environmental factor(X), genetic factor(G), and their interactions ($G \times X$) respectively. The error term ε is distributed with mean 0 and finite variance σ^2.

The G×E interaction is modeled as a product term between G and X in (1.6). A rearrangement of (1.6) results in

$$Y = \alpha_0 + \alpha_1 X + (\alpha_2 + \alpha_3 X)G + \varepsilon \tag{1.7}$$

(1.7) explicitly conveys the message that the variation in response Y caused by the genetic effect G is a linear function of environment X. However, the genetic effect may not necessarily take a linear format in practice. Positing such linear form might lead to model mis-specification and inflated bias. A varying coefficient model approach, together with a group of goodness-of-fit tests, were thus proposed in [37] to investigate the non-linear machinery of G×E interactions for continuous phenotypic response. Extensions of [37] to binary response is urgently needed as majority of complex human diseases are casted in the case-control association study framework where the response are of two categories, disease(case) or no disease(control).

The recent success of set-based GWAS, such as in the pathway-based [38] and gene-centric study [39, 40, 41], in bringing novel interpretations to the disease signals, motivates us to coin G×E interaction within a set-based association framework. When multiple variants within a genetic system, such as the pathway or gene set, are involved, we can jointly model their effect with an environment factor, especially if any non-linear effects are present, by proposing an additive varying-coefficient model:

$$Y = \alpha_0(X) + \alpha_1(X)G_1 + \ldots + \alpha_d(X)G_d + \varepsilon \tag{1.8}$$

where Y is the phenotypic response, d is the total number of SNP variants in a genetic feature and G_j refers to the jth SNP, and ε is the random error. The model has particular power to help us understand how multiple genetic variants in a system are mediated by a common mediator X to affect disease risk. When d is relatively large, the problem can be approached from a high dimensional variable selection perspective.

1.4 Objectives and organization of the dissertation

Due to the crucial roles G×E interactions played in elucidating the genetic basis of complex diseases, the objective of this dissertation will be on developing novel statistical methodology and powerful computational tools to tackle the challenges originated from high dimensional G×E interaction analysis.

The dissertation is organized as follows. In chapter 2, the dissection of the nonlinear penetrance effect of the genetic variants with regard to the environmental factor will be extended from continuous phenotypic response in [37] to case-control association study within

the varying coefficient model framework with an application to Type 2 Diabetes. The varying coefficients are estimated through nonparametric regression spline techniques. A group of statistical tests was proposed to elucidate the machinery of G×E interactions. In chapter 3, a set-based method examining G×E interactions will be developed. We propose a penalized additive varying coefficient model to select genetic variants with varying effects (the presence of G×E interactions), constant effects (no G×E interactions), and zero effects (no association with phenotype). The selection consistency of this approach and the oracle property of the corresponding penalized estimator will be rigorously established. In chapter 4, we further extend the framework in chapter 3 to the scenario where the number of genetic variants exceeds the sample size, the so called "large p,small n", via group coordinate descend (GCD) algorithms. A thorough investigation on the performance of different penalty functions within this framework will be conducted. Concluding remarks and outline of the future work will be given in chapter 5.

Chapter 2

A Novel Method For Identifying Nonlinear Gene-Environment Interactions In Case-Control Association Studies

2.1 Introduction

It has been increasingly recognized that the predisposition of many complex diseases is not purely triggered by genetic factors. They are also influenced by environmental exposures, due to potential gene-environment interactions. For example, Type 2 Diabetes mellitus is a typical complex human disease whose incidence is heavily contingent on the environmental exposures such as behavioral and dietary factors, in addition to genetic susceptibility [42, 43]. Studies on gene×environment (G×E) interactions will shed novel light on the genetic responses to environment dynamics and how environment changes mediate gene expression to increase disease risks. Such phenomenon that disease risk or genetic expression varies under different environment conditions is also termed phenotypic plasticity [44].

G×E interactions were historically pursued by evaluating the gene effect under different

environment conditions. Figure 2.1A shows the case of G×E interaction under two discrete environment conditions, protective and predisposing such as non-smoking and smoking. When environment conditions are measured in a continuous scale, more information is available to assess the gradient/dynamic change of genetic effect under subtle environment changes. For example, adult bone mineral density changes with age and vitamin D intake [45]. Figure 2.1B-D display several scenarios where the environment mediator is measured in a continuous scale. Example of continuous environment could be age for age related diseases such as Alzheimer, or body mass index for Type 2 Diabetes or hypertension. In Figure 2.1B, no G×E interaction is observed since the genetic effects of the three genotypes are parallel to each other. Figure 2.1C shows a typical example of linear G×E interaction, while Figure 2.1D displays a non-linear G×E interaction pattern assuming the *Aa* genotype is the baseline. As will seen in the following section, most current G×E interaction model assumes the case displayed in Figure 2.1C. Few statistical analysis has considered the case shown in Figure 2.1D.

In fact, many literature work supports the view of nonlinear G×E interaction. Sparrow et al. [46] found that mutations in gene HES7 and MESP2 caused congenital scoliosis, and the risk was highly related to transient hypoxia during mice pregnancy. The rate of risk increase was non-linearly correlated with increasing hypoxic levels. Laitala et al. [47] reported that the reaction of personal genetic effects on coffee consumption showed a non-linear relationship with age. Martinez et al. [48] found that women carrying Gln27Glu genotype in ADRB2 gene had higher probability for obese and the obesity rate was nonlinearly correlated with the amount of carbohydrate intake. Even though these empirical evidences are limited to small-scale observational studies, they underscore the importance of further exploration on non-linear G×E interaction when searching for genetic roots of complex diseases.

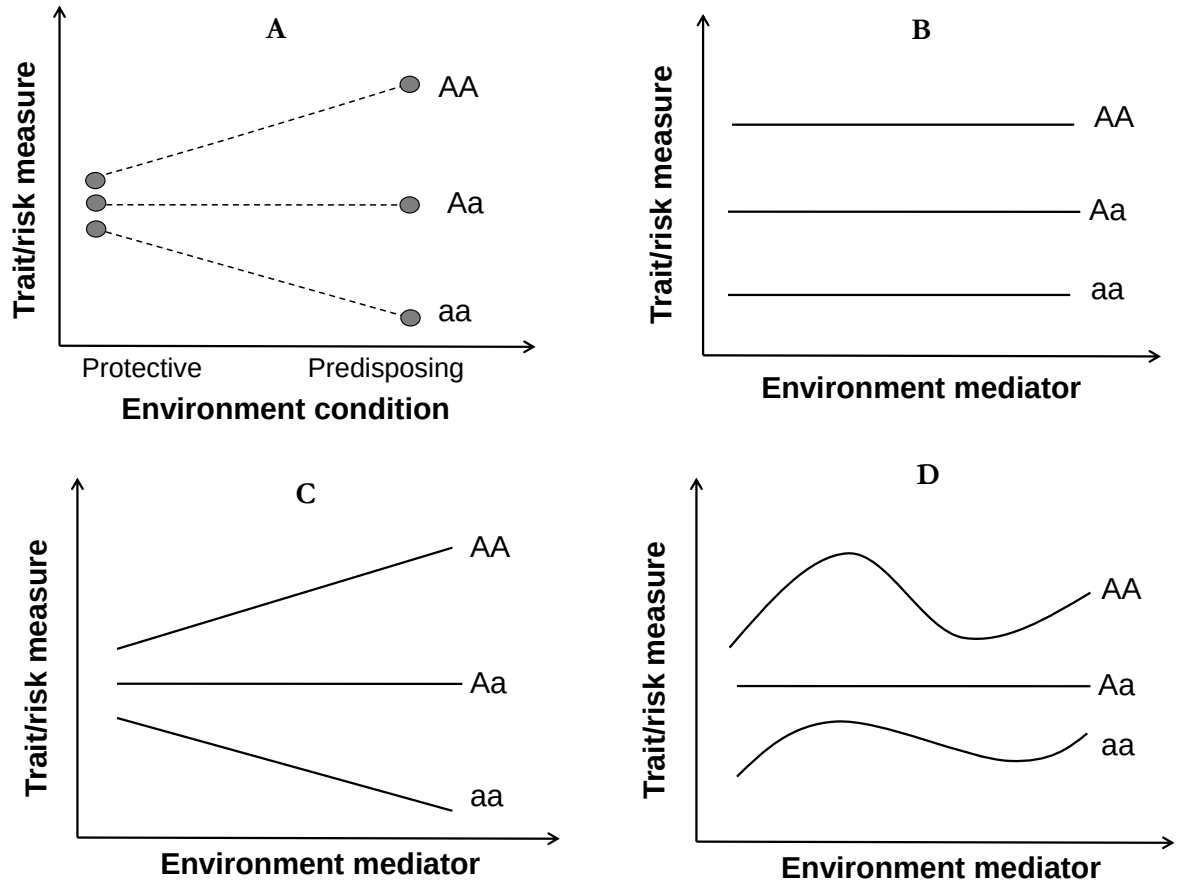

Figure 2.1: Different models of gene-environment interaction: (A) the interaction of gene and environment in discrete environmental conditions; cases with (B) no G×E interaction; and (C) linear and (D) non-linear G×E interactions. AA, Aa and aa represent three different genotypes in a gene, and environment mediator represents a continuous environment variable.

Within the statistical framework, G×E interactions in human diseases have been investigated mainly through model-based approaches, ranging from the standard linear model with interaction in diverse design settings, such as the case-control design, the case only design and the two-stage screening design, to more sophisticated models, such as profile likelihood-based semi-parametric models, empirical Bayesian models and Bayesian model average (reviewed in Mukherjee et al. [48]). However, as pointed out in Ma et al. [37], the model-based regression framework generally needs strong model assumptions between genetic effects and environmental influences, which cannot be directly applied to the above

mentioned empirical studies in which non-linear interaction exists.

In this chapter, we extended the varying-coefficient (VC) model proposed in Ma et al. [37] for continuous quantitative responses to binary disease responses. We first laid out the VC modeling framework for binary responses, with details on parameter estimation and hypothesis testing. The utility of our approach was demonstrated through extensive simulations. Finally, we applied our method to two case-control Type 2 Diabetes cohorts data sets, followed by discussions.

2.2 Statistical method

For a sample of n unrelated individuals collected from a population, let n_1 and n_2 be the number of affected (cases) and unaffected (controls) individuals, respectively with $n = n_1 + n_2$. All individuals in the sample could be genotyped either based on candidate genes or on a whole genome-wide scale. Let $Y_i = 1$ if the ith individual is affected and 0 otherwise. Let G be the genetic variable which is coded as 0, 1, 2 corresponding to genotype aa, Aa and AA where allele A is the minor allele. This coding scheme assumes an additive disease model, although a genetic variant may show dominant or recessive action mode. In reality we can do a model selection to choose which one is the optimal one using AIC or BIC criterion.

Suppose in addition to the genetic variables, the disease risk is also affected by environmental factors as well as the interaction between them. Let X be the environmental variable which is measured in a continuous scale. Throughout this chapter, we are only interested in environment changes that display in a continuous scale (e.g., geographical location or temporal changes). Traditional analysis for G×E was commonly pursued by discretizing an environmental variable into different groups (e.g., old vs young), as shown in Figure 2.1.

However, we can have more information to assess the G×E relationship when a continuously measured environment factor is treated in a continuous scale. Thus, the purpose of the work is to model the genetic responses under different environmental stimuli, and further assess in what form genes respond to these changes.

For a continuous phenotype Y, the general form of an additive VC model to investigate the non-linear G×E interaction between X and G can be expressed as,

$$Y = \alpha(X) + \beta(X)G + \sigma(X)\varepsilon \tag{2.1}$$

where the error term ε satisfies $E(\varepsilon|X, G) = 0$ and $Var(\varepsilon|X, G) = 1$. Ma et al. [37] evaluated the performance of the model by assuming $\alpha(X) = \alpha_0 + \alpha_1 X$. The key components of the VC model lie in proper estimation of the smoothing function $\beta(X)$ and the variance function $\sigma^2(X)$, through which the effect of the genetic variant can be evaluated as a function of environment exposures. Various tests have been proposed to assess the linear or non-linear mechanisms via likelihood ratio test. When inhomogeneous variance (i.e., $\sigma(X)$ varies with X) and no parametric distribution are assumed for the error term, wild bootstrap is a common choice to assess the significance of the likelihood ratio statistic.

In human genetics, many diseases are displayed as discrete qualitative traits. The focus of this work is to extend the above model to responses that do not follow continuous distribution. In a generalized linear model set up, the relationship between the mean of a response variable Y and the independent variables (X, G) under the varying-coefficient model can be expressed as

$$E(Y|X, G) = \mu = g^{-1}\{\alpha(X) + \beta(X)G\}$$

where g is a link function. When Y is measured as counts (e.g., tumor numbers), a *log* link

function can be assumed. When Y is a binary variable (i.e., affected vs unaffected), then a *logit* link function is commonly applied. In the later case, the *logit* varying-coefficient model is given by,

$$\text{logit}(p) = \alpha(X) + \beta(X)G \tag{2.2}$$

where p=Pr$(Y = 1|X, G)$. In this work, we allow the intercept function $\alpha(X)$ varies with X instead of assuming a linear structure, to make it more flexible to capture the underlying mean function when there is no genetic contribution (i.e., $\beta(X) = 0$).

If we allow $\beta(X) = \beta_1$, the logistic VC model is reduced to a logistic linear predictor model without G×E interaction (denoted as LM). If we allow $\beta(X) = \beta_1 + \beta_2 X$, the logistic VC model is reduced to a logistic linear predictor model with linear G×E interaction (denoted as LM-I), i.e.,

$$\begin{aligned} \text{logit}(p) =& \alpha(X) + (\beta_1 + \beta_2 X)G \\ =& \alpha(X) + \beta_1 G + \beta_2 XG \end{aligned} \tag{2.3}$$

One can also put structures on the function of $\alpha(X)$. For example, we can let $\alpha(X) = \alpha_0 + \alpha_1 X$. Such a model like $\text{logit}(p) = \alpha_0 + \alpha_1 X + \beta_1 G + \beta_2 XG$ is often applied in assessing G×E interactions in a typical logistic regression analysis by testing $H_0 : \beta_2 = 0$. It can also be seen that this model assumes a linear G×E interaction structure, that is, the function $\beta(X)$ is linear in X. Thus, without assuming specific structure on the linear predictors, the VC model has much flexibility to capture the underlying interaction mechanism via fitting $\beta(X)$ using smoothed nonparametric functions. The VC interaction model can be considered as a generalization to the linear interaction model.

2.3 Estimating $\beta(X)$ function

The nonparametric estimation of varying coefficients has undergone intensive investigations in the last two decades and falls generally into three categories: the local kernel polynomial smoothing, polynomial spline, and smoothing spline (Fan and Zhang [49]; Huang et al. [50]). Huang et al. [50] approximated the varying-coefficient functions via B-spline basis expansion. Using the B-spline technique, the authors further established the relevant asymptotic properties of the estimators, such as consistency, convergence rates and asymptotic normality. In addition, the estimation of B-spline estimators is computationally fast and numerically stable. These merits are especially important in the context of high-dimensional genetic data analysis, which make it a natural choice for us to choose when estimating the varying-coefficient functions $\alpha(X)$ and $\beta(X)$.

Let h be the degree of B-splines and N be the corresponding interior knots. Further assume that the knots are equally distributed for the B-spline basis matrix $\{\mathbf{B}_s : 1 \leq s \leq (N+h+1)\}$. Ideally we can select h and N for $\alpha(X)$ and $\beta(X)$ separately using the B-spline technique when fitting each SNP variant. This process involves a search of optimal degree and knots through a list of possible combinations for both functions. This, however, could incur heavy computation burden when the estimation is done for each SNP given that the number of SNP variants to be tested could be huge. Thus, the degree h_0 and knots N_0 for $\alpha(X)$ are selected first by fitting a logistic VC model without the genetic components. Once the degree and knots for function $\alpha(X)$ are selected, they will be fixed when estimating degree and knots for function $\beta(X)$ for each SNP. The selection is done by using the Bayesian

Information Criterion (BIC) criteria defined as,

$$\arg\min_{N,h} BIC(N,h) = \arg\min_{N,h} \ell(\boldsymbol{\gamma}_1) + (N+h)\log(n)/n,$$

where $\ell(\boldsymbol{\gamma}_1)$ refers to the log-likelihood function. A grid search for possible combinations of N and h can be done and the values corresponding to the minimum BIC are the "optimal" ones.

Once the degree and knots for $\alpha(X)$ are determined, the function $\alpha(X)$ can be estimated by $\hat{\alpha}(X) = \hat{\boldsymbol{\gamma}}_{\mathbf{1}}{}^T \mathbf{B_1}(\boldsymbol{X}) = \sum_{k=1}^{N_0+h_0+1} \hat{\gamma}_{1k} B_{1k}(x)$. The degree h_1 and the number of knots N_1 for $\beta(X)$ are also selected using the same BIC criterion defined above. The estimator for $\beta(X)$ is given by $\hat{\beta}(X) = \boldsymbol{\gamma}_{\mathbf{2}}^T \mathbf{B_2}(\boldsymbol{X}) = \sum_{k=1}^{N_1+h_1+1} \gamma_{2k} B_{2k}(x)$. Regular Newton-Raphson or Fisher scoring algorithm can be applied to estimate the parameters.

2.4 Assessing G×E interaction

Our goal is to assess if a genetic variant is sensitive to environment changes. If it does, then in what form, linear or nonlinear. For this purpose, we first propose to assess if the genetic effect is a constant by testing

$$\begin{cases} H_0^C: & \beta(\cdot) = \beta \\ H_a^C: & \beta(\cdot) \neq \beta \end{cases} \tag{2.4}$$

where β is an unknown constant and $\text{logit}(p) = \alpha(X) + \beta G$ is the corresponding reduced model under the null hypothesis. Under the H_0, the genetic effect is a constant and its contribution to disease risk has nothing to do with environmental changes. If we fail to reject the null, then association can be assessed via testing $H_0 : \beta = 0$ by fitting the reduced model. Rejecting the null hypothesis leads to the conclusion that the G×E interaction exists.

We next test the linear effect of G×E interaction by formulating,

$$\begin{cases} H_0^L: & \beta(\cdot) = \beta_1 + \beta_2 X \\ H_a^L: & \beta(\cdot) \neq \beta_1 + \beta_2 X \end{cases} \tag{2.5}$$

where β_1 and β_2 are unknown constants. Under the H_0, the reduced model is given by $\text{logit}(p) = \alpha(X) + \beta_1 G + \beta_2 XG$. If we fail to reject the null, then association can be assessed via testing $H_0 : \beta_1 = \beta_2 = 0$ by fitting the reduced model. If the null is rejected, it indicates nonlinear G×E interaction effect and next we fit model 2.2 to assess genetic association.

The above tests are sequential. At each step if we fail to reject the null, we stop and fit the null model and assess the genetic effect by a likelihood ratio test. When H_0^L is rejected, a nonlinear G×E interaction effect is implied and we allow the data tell the shape of the effect by fitting the above described nonparametric B-spline functions. The nonlinear effect is then assessed by testing $H_0 : \beta(X) = 0$ using a likelihood ratio test which asymptotically follows a chi-square distribution with the degrees of freedom equal the number of fitted B-spline coefficients for function $\beta(X)$.

2.5 Simulation

The statistical behavior of the proposed approach was evaluated through extensive Monte Carlo simulations. When using B-spline functions to estimate the varying-coefficients, a uniform distribution on X is generally assumed. In real application, the environment measure (X) may not be uniformly distributed as in the Type 2 Diabetes data analyzed later in the chapter. Instead, it is often normally distributed. To mimic real situations, we generated a continuous environment measure X^* from a normal distribution, and subsequently trans-

formed it by $X = \Phi(\frac{X^* - \bar{X^*}}{S_{X^*}})$, to make X^* evenly distributed on the B-spline subintervals, where $\Phi(\cdot)$ is the standard normal cumulative distribution function and $\bar{X^*}$ and S_{X^*} are the sample mean and sample standard deviation of X^*, respectively. The B-spline basis matrix was constructed on the transformed values. For a given minor allele frequency (MAF) p_A and assuming Hardy-Weinberg equilibrium, SNP genotypes AA, Aa and aa were simulated from a multinomial distribution with frequencies p_A^2, $2p_A(1-p_A)$ and $(1-p_A)^2$ for the three genotypes, respectively. We coded the genetic variable G_i as (2, 1, 0) corresponding to genotypes (AA, Aa, aa).

2.5.1 False positive control

We fist evaluated the false positive control for the VC model at the nominal 0.05 level. For comparison purpose, we also reported the error rate for the linear predictor model with and without interaction. Under the null of no genetic effects, the disease phenotypes were simulated with $\text{logit}(p) = \alpha_0 + \alpha(x)$ where $\alpha(x)$ was generated via the B-spline basis function, i.e., $\alpha(x) = \sum_{k=1}^{4} \gamma_k B(x)$ for given spline coefficients $\gamma_1 = 6.162$, $\gamma_2 = 5.948$, $\gamma_3 = 3.858$, $\gamma_4 = 3.640$. The spline coefficients were obtained by fitting the real data (described later) without fitting the genetic effect. We added a constant α_0 in order to control the simulated proportion of case:control ratio to approximately 1:1 (by varying the size of α_0). A total of 10,000 simulation replicates were taken under all the combinations of sample size ($n = 500, 1000, 2000$) and MAF ($p_A = 0.1, 0.3, 0.5$).

The results were summarized in Figure 2.2. As we can observe, the false positive rates were estimated sensibly from the simulated data. The VC model slightly overestimated the false positive rate under low allele frequency (p_A=0.1). But the performance improved as MAF increases for a fixed sample size. In addition, the performance improved as sample

size increased under a fixed MAF. In general, there were no significant deviations from the nominal 0.05 level for all the 3 models, except in some cases under low MAF and small sample size.

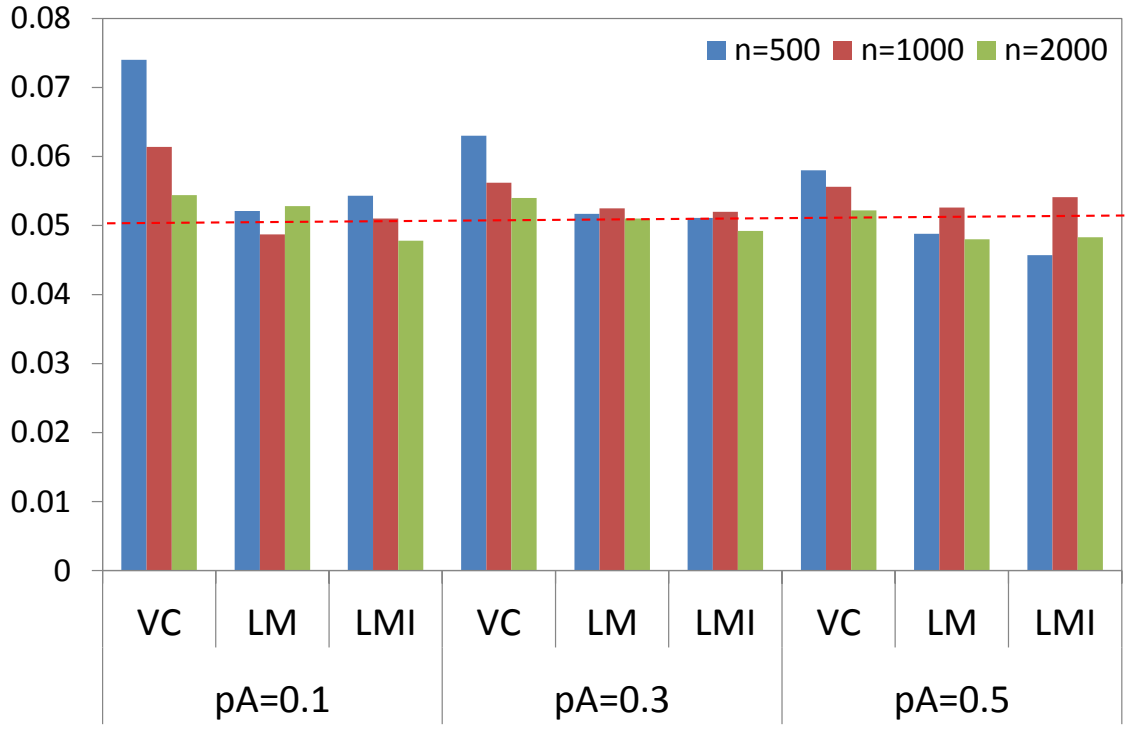

Figure 2.2: The false positive rate of different models at the 0.05 level.(For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.)

2.5.2 Power evaluation

For given genetic effects, the disease status was simulated from a Bernoulli trial. The varying-coefficient function $\beta(\cdot)$ was estimated through $\hat{\beta}(x) \equiv \sum_{s=1}^{N_1+h_1+1} \hat{\gamma}_s B(x)$. In a typical simulation study with VC models, people generally simulate data assuming a nonlinear function such as a sin or exponential function. As SNPs do not function in such form, we simulated data according to the fit calculated from the real data to make it more realistic. Three scenarios were considered. Scenario 1 assumed that the true G×E interaction was

nonlinear and the data were generated with the VC model. In scenario 2, we assumed there was no G×E interaction, while in scenario 3 we assumed linear G×E interaction. The simulated data were then analyzed using the VC, LM-I and LM models, to compare the performance under model mis-specification.

For a given MAF, the data assuming nonlinear G×E interaction were generated with the following VC model,

$$\text{logit}(p_i) = \alpha_0 + \alpha(X_i) + \beta(X_i)G_i$$

where $p_i = p(Y = 1|X, G)$, and α_0 was a constant used to control the case:control ratio to make it close to 1. The varying coefficient functions $\alpha(X)$ and $\beta(X)$ were computed based upon the quadratic B-spline basis matrix with $\alpha(X) = \boldsymbol{\gamma_1'}\mathbf{B}_1(X)$ and $\beta(X) = \boldsymbol{\gamma_2'}\mathbf{B}_2(X)$, where $\boldsymbol{\gamma}_1 = (7.287, 7.146, 3.917, 3.413)^T$ and $\boldsymbol{\gamma}_2 = (0.080, -0.460, -0.201, 0.465)^T$ were obtained from real data fit, namely SNP rs4506565 on chromosome 10 of the Nurses' Health Study (NHS) data in GENEVA consortium (described later). The binary responses were then generated from a Bernoulli trial with case probability p_i.

The likelihood ratio test was applied to assess the significance of each test illustrated in previous section. The comparison results were shown in Figure 2.3. As we expected, a common trend for the three models is that the power increases as MAF and sample size increase. Under the same sample size or MAF, the VC model always has the best power among the three, which is not surprising since the phenotypes were generated from a VC model. In addition, the LM-I model performs better than the LM model since structurally it is more close to the VC model.

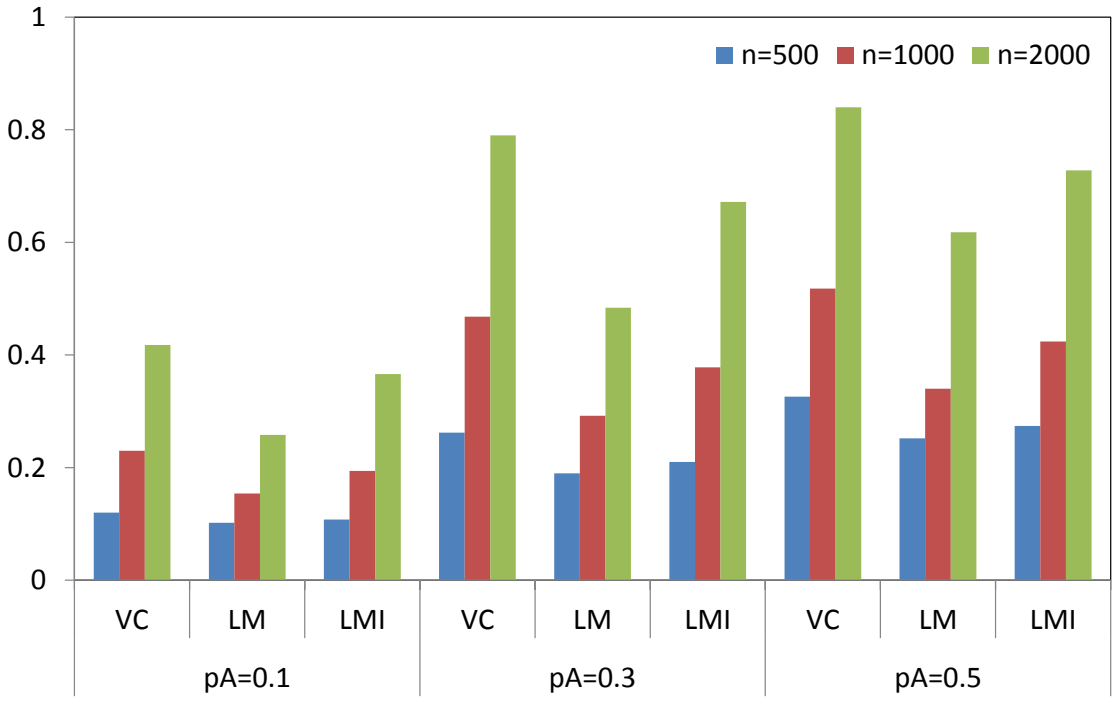

Figure 2.3: The power of different models under different MAFs and sample sizes when data were generated with the VC model.

We also simulated data assuming no G×E interaction using the following model,

$$\text{logit}(\text{p}_i) = \alpha_0 + \alpha(X_i) + \beta_1 G_i$$

where $\alpha(x)$ was generated from the B-spline basis function with $\alpha(X) = \boldsymbol{\gamma_0'}\mathbf{B}_0(X)$. The spline coefficient vector was given by $\hat{\boldsymbol{\gamma}}_0 = (5.977, 6.011, 3.843, 3.668)^T$, and the genetic coefficient was set as $\beta_1 = 0.271$ (corresponding to an odds ratio of 1.3). These coefficients were obtained by fitting the real data with a linear predictor without interaction for SNP rs12255372 on chromosome 10. α_0 was used to adjust the case:control ratio as described before under different sample sizes and MAFs. The results shown in Figure 2.4 demonstrate that the LM model outperforms the other two models in all the scenarios, since data were analyzed with the true data generating model. As MAF increases, the power differences

among the three models diminishes for larger sample sizes. For example, the power difference among the three models is very small under $p_A = 0.5$ and $n = 2000$.

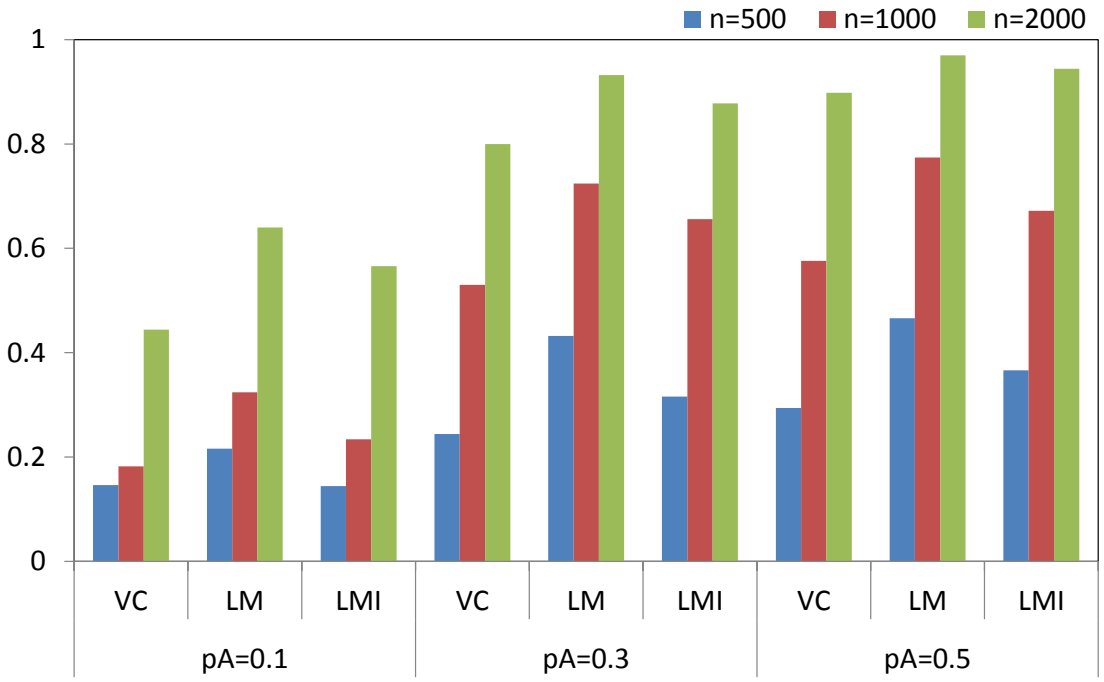

Figure 2.4: The power of different models under different MAFs and sample sizes when data were generated with the LM model.

The following linear interaction model (LM-I) was assumed to generate the linear G×E interaction data,

$$\text{logit}(\text{p}_i) = \alpha_0 + \alpha(X_i) + \beta_1 G_i + \beta_2 X_i G_i$$

where $\alpha(X) = \boldsymbol{\gamma}_{\mathbf{0}}^{'}\mathbf{B}_0(X)$ with spline coefficients $\hat{\boldsymbol{\gamma}}_0 = (6.358, 6.481, 4.232, 4.113)^T$. The genetic coefficient β_1=0.226 and interaction coefficient β_2=-0.787. All the coefficients were obtained by fitting model 2.3 to SNP rs17537178 on chromosome 10. Figure 2.5 shows that the linear interaction model has the best performance among the three. In addition, the power of the VC model is more close to the linear interaction model since it is more structurally close to the linear interaction model. As sample size and MAF increase, the

power difference between the VC and LM-I model vanishes quickly.

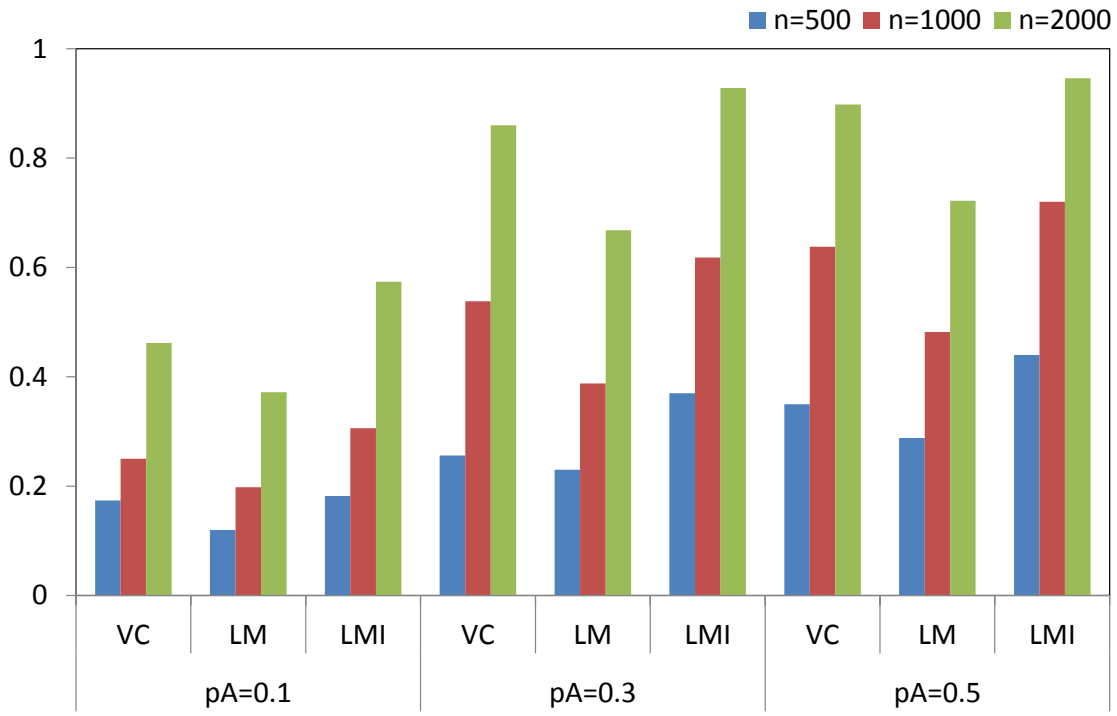

Figure 2.5: The power of different models under different MAFs and sample sizes when data were generated with the LM-I model.

In summary, when the true G×E interaction is linear or when there is no interaction at all, the model assuming linear or constant coefficient outperforms the VC model. However, the VC model outperforms the other two when the true interaction in nonlinear. In addition, the LM or LM-I models suffer more from power loss when the underly true interaction is nonlinear in comparison to the case when the underlying truth is linear or no interaction. This is not surprising since the B-spline estimator is consistent for large samples. Under large sample sizes, the VC model should perform similar to the LM and LM-I model. However, one has to be careful in finite samples. The simulation results suggest that one should assess the function $\beta(X)$ first before testing $\beta(X) = 0$. In practice, one can test if $\beta(X) = \beta$ or $\beta(X) = \beta_1 + \beta_2 X$, then fit the appropriate model depending on the test result.

2.6 Real data analysis

The fast increase in global prevalence of Type 2 Diabetes draws worldwide attentions for the disease. About 50 novel loci have been reported in association with Type 2 Diabetes so far (Perry et al. [52]. However, only a small proportion of disease heritability has been explained by these loci, leaving the question of how to effectively accounting for gene-environment interaction in the search of T2D susceptibility variants with the hope to capture the missing heritability. We applied our model to two nested case-control cohort studies of Type 2 Diabetes, the Nurses' Health Study (NHS) and the Health Professionals Follow-up Study (HPFS), from the Gene, Environment Association Studies Consortium (GENVEA) (Cornelis et al. [53]). The two data sets are well-characterized cohorts of genome-wide association studies investigating a set of hypotheses about the dietary and lifestyle factors to the triggering of a series of diseases, including Type 2 Diabetes, for women and men. Details of the two cohorts can be found from Colditz et al. [54] and Rimm et al. [55]. The data sets from the two cohort studies originally contain 3,391 females (NHS) and 2,599 males (HPFS) with European ancestry. After data cleaning by removing subjects with unmatched phenotypes and genotypes, excluding SNPs with MAF< 0.05 and deviation from Hardy-Weinberg equilibrium, the final data contain 3,391 females (1,646 cases and 1,745 controls) with 635,748 SNPs in the NHS set and 2,570 males (1,300 cases and 1,270 controls) with 636,764 SNPs in the HPFS set.

Body mass index (BMI), calculated as the quotient between an individual's mass (kg) and the square of height (m^2), is an indicator of human obesity. It is widely recognized that the risk of Type 2 Diabetes could be potentially influenced by obesity condition evidenced by strong association between them for both women and men (Holbrook et al. [56]; Carey

et al. [57]; Chan et al. [58]). Therefore, individual's BMI can be regarded as a type of environmental condition pivotal in evaluating the incidence of Type 2 Diabetes. Individuals carrying the same gene may have different risks of Type 2 Diabetes under different obese conditions. The phenomenon could be elucidated, at least partially, by the complicated interaction mechanism between the carrier's gene and the environment (measured by BMI). Thus, we can treat the genetic sensitivity to obese as a dynamic process which can be captured by the proposed VC model, if any.

We analyzed the male and female data sets separately in order to find sex-specific genes responsible for T2D risk. Figures 2.6–2.8 showed the Manhattan plot of the -log10(p-values) for the male data. To compare the performance of the three models (VC, LM, LM-I), we plotted all the signals at each SNP locus. It can be seen that the overall signals for the three models are quite consistent. The dashed red line corresponds to the genome-wide Bonferroni threshold (7.9E-8) and the dotted blue line corresponds to the suggestive threshold (5E-6). Table 2.1 tabulated SNPs that passed both threshold. Seven SNPs passed the Bonferroni threshold are marked by $*$. Testing constant coefficient showed that the majority of SNPs has constant coefficients, which indicated they are not sensitive to obese condition. This also explained why the LM model gives relatively stronger signals than the other two models. Columns with P_CON and P_LIN showed the p-values for testing $H_0 : \beta(X) = \beta$ and $H_0 : \beta(X) = \beta_1 + \beta_2 X$. The smaller p-values for testing constant and linear coefficients in the top panel showed that the effects of those SNPs were neither constant nor linear, thus the VC model gave the strongest signals evidenced by smaller P_VC than P_LM and P_LMI. For example, SNP rs4635456 had P_CON=9.5E-07 and P_LIN=0.0117 which indicated the coefficient of this SNP is varying over BMI. Thus, fitting a VC model gave the strongest signal (P_VC=3.05E-06 vs P_LM=0.6299 and P_LMI=1.58E-05).

Table 2.1: List of SNPs with p-value < 5E-06 in the HPFS (Male) data set

SNP ID	GeneName	Chr	P_VC	P_CON	P_LIN
			fitted with VC model		
rs4635456	SEMA6B	19	**3.05E-06**	9.49E-07	0.0117
rs4972250	Unknown	2	**3.99E-06**	2.21E-06	1.65E-06
rs4842244	RXRA	9	**4.18E-06**	1.25E-06	2.91E-06
			fitted with LM model		
rs2371765	ADAMTS9-AS2	3	6.82E-09	0.2909	-
rs7901695	TCF7L2	10	1.49E-06	0.8638	-
rs7991210	PCCA	13	2.80E-07	0.2234	-
rs12243326	TCF7L2	10	1.14E-06	0.6896	-
rs4132670	TCF7L2	10	1.64E-06	0.8560	-
rs12255372	TCF7L2	10	1.89E-06	0.7372	-
rs4506565	TCF7L2	10	2.66E-06	0.8546	-
rs11013381	C10orf67	10	7.83E-05	0.8865	-
rs6893115	Unknown	5	9.19E-05	0.8287	-
			fitted with LMI model		
rs699253	PDE4B	1	3.93E-05	0.0108	0.6792

SNP ID	GeneName	Chr	P_LM	P_LMI	P_I
			fitted with VC model		
rs4635456	SEMA6B	19	0.6299	1.58E-05	2.91E-06
rs4972250	Unknown	2	0.2772	0.1982	0.1516
rs4842244	RXRA	9	0.7146	0.0886	0.0299
			fitted with LM model		
rs2371765	ADAMTS9-AS2	3	**2.38E-10**∗	1.88E-09	0.8140
rs7901695	TCF7L2	10	**1.72E-08**∗	1.06E-07	0.5633
rs7991210	PCCA	13	**1.81E-08**∗	7.07E-08	0.2655
rs12243326	TCF7L2	10	**1.87E-08**∗	8.88E-08	0.3570
rs4132670	TCF7L2	10	**1.94E-08**∗	1.07E-07	0.4632
rs12255372	TCF7L2	10	**2.93E-08**∗	1.49E-07	0.4076
rs4506565	TCF7L2	10	**3.31E-08**∗	1.87E-07	0.4967
rs11013381	C10orf67	10	**1.32E-06**	7.74E-06	0.7106
rs6893115	Unknown	5	**1.79E-06**	1.11E-05	0.9266
			fitted with LMI model		
rs699253	PDE4B	1	1.5E-04	**4.21E-06**	0.00125

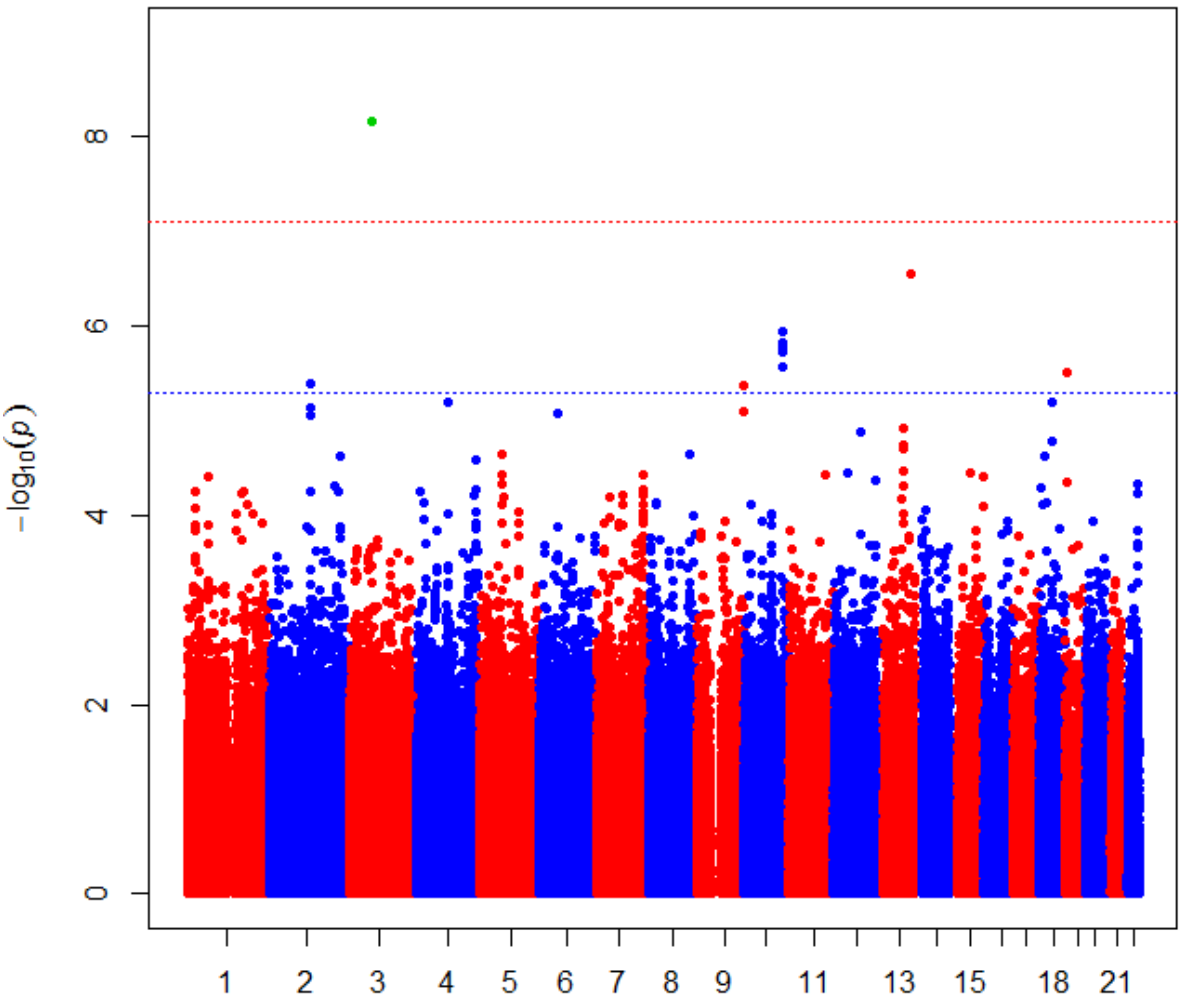

Figure 2.6: The Manhattan plot of -log10(p-values) for testing $H_0 : \beta(X) = 0$ when fitting the VC model to the male data set.

The mid-panel in the table listed SNPs with the strongest signals fitted with the LM model. The P_{const} values for the SNPs were all large (>0.05), which suggests that $\beta(X)$ was a constant and there was no G×E interaction for these SNPs. Hence the LM model assuming no interaction gave the strongest signals. The bottom SNP in the table had the strongest signal when data were fitted with the LM-I model since we rejected constant coefficient

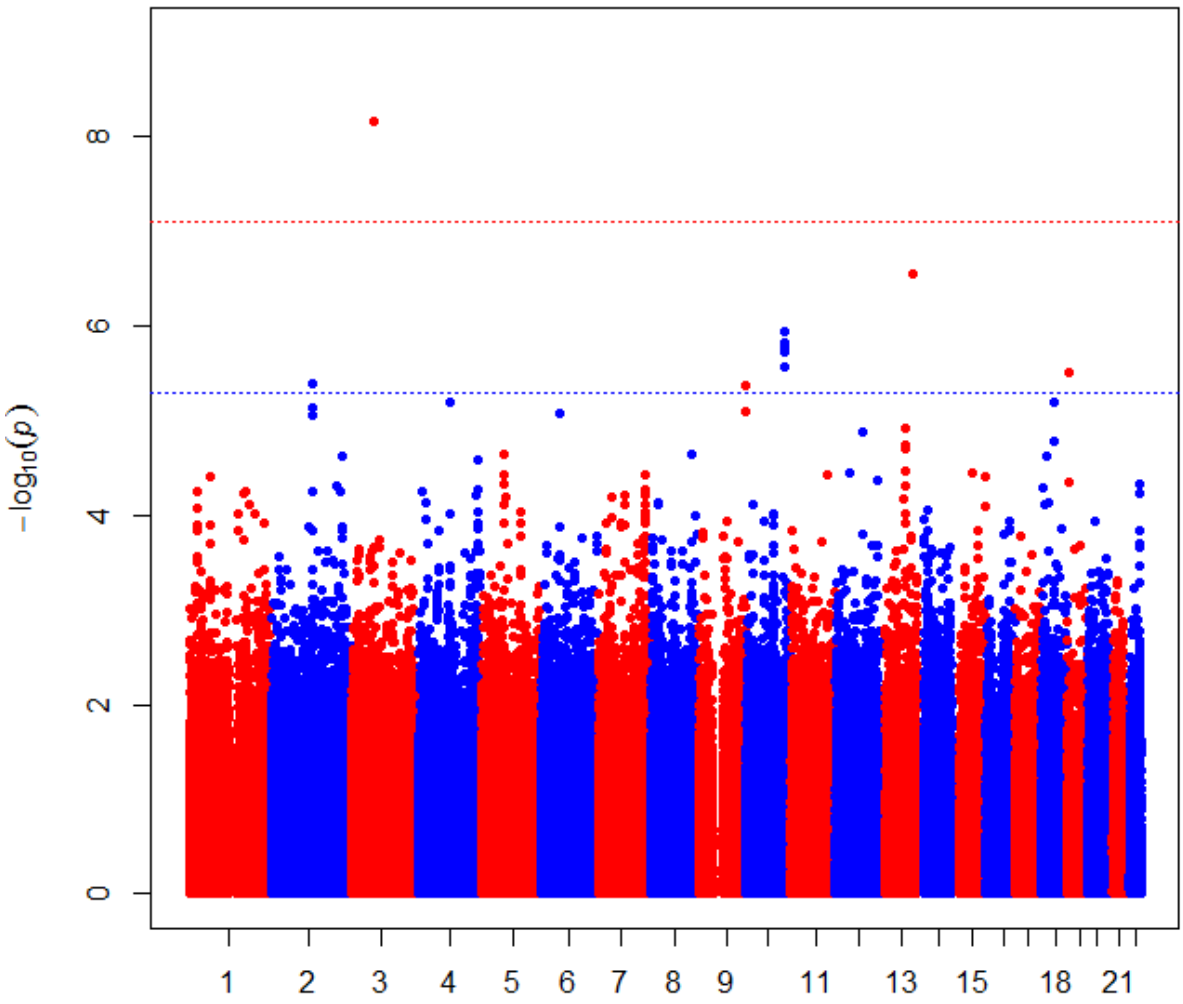

Figure 2.7: The Manhattan plot of -log10(p-values) for testing $H_0 : \beta = 0$ when fitting the LM model to the male data set.

(P_CON=0.0108) but failed to reject linear coefficient (P_LIN=0.6792).

Among the SNPs listed in the table, some have been reported in other studies. For example, transcription factor 7-like 2 (TCF7L2) is an intensively examined gene associated with a broad categories of diseases, including Type 2 Diabetes. The causal genetic association between SNPs of the gene and the Type 2 Diabetes was first reported in Grant et al. [59] and

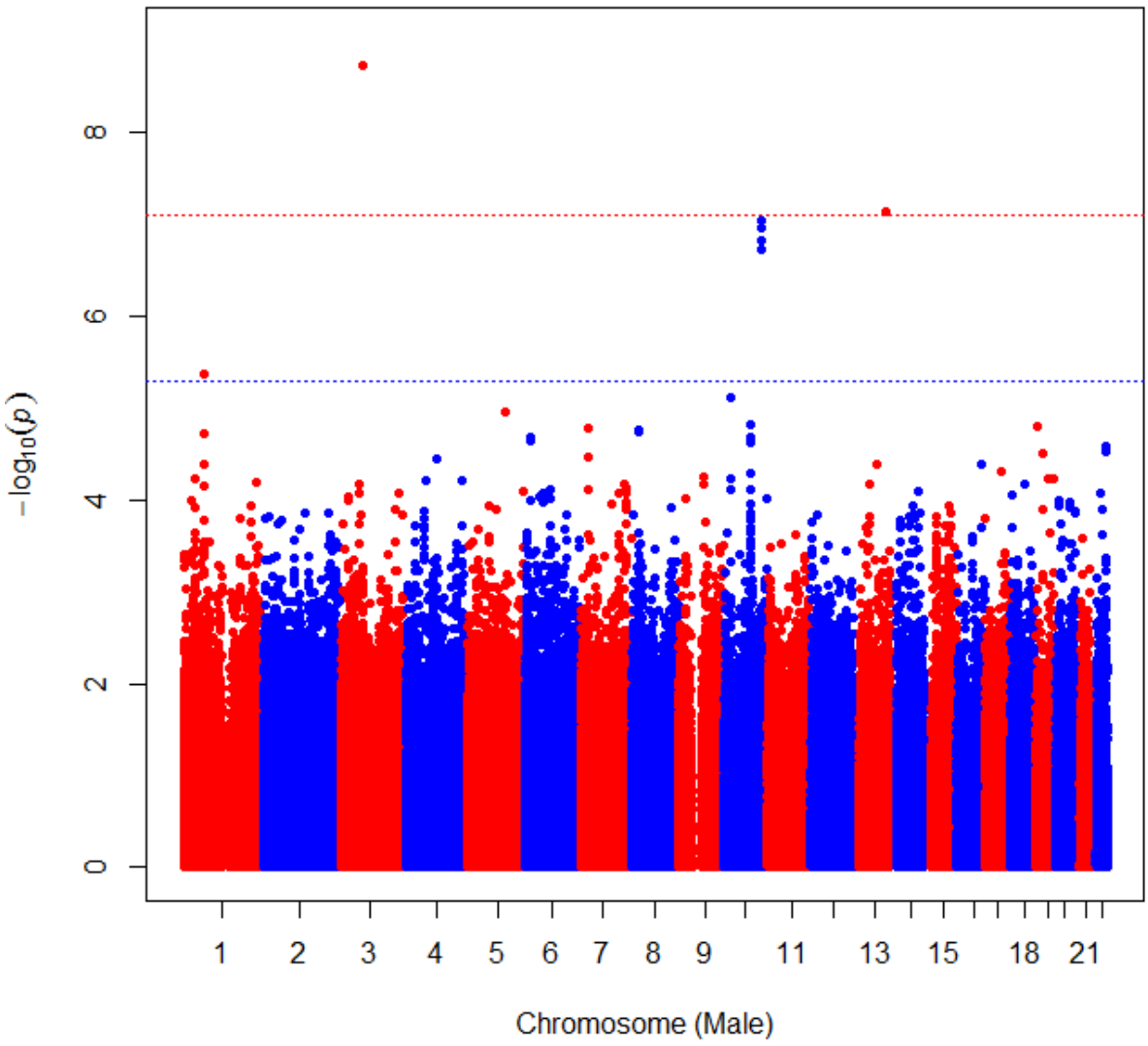

Figure 2.8: The Manhattan plot of -log10(p-values) for testing $H_0 : \beta_1 = \beta_2 = 0$ when fitting the LM-I model to the male data set.

was subsequently replicated in many ethnic groups (Jin and Liu [60]). As the SNPs in this gene are not sensitive to obesity, it is not surprise that they can be identified in other studies by using methods assuming a linear relationship. But our method identified three more that show nonlinear G×E relationship, even though they did not pass the genome-wide Bonferroni threshold. We also did QQ plot of the p-values for data fitted with the three models. The

p-values are quite uniformly distributed and only a few showing departure from the expected values (see the QQ plot). This indicates that the models have no serious inflation of false positives and the strong signals are likely to be true.

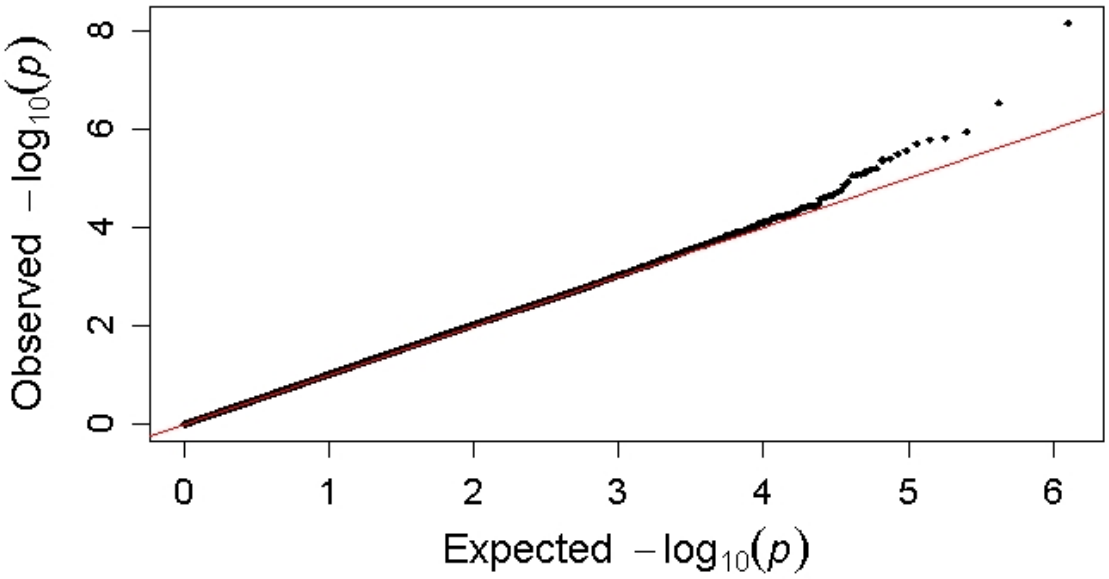

Figure 2.9: The QQ plot of genome-wide p-values for the male data, when data are fitted with the VC model

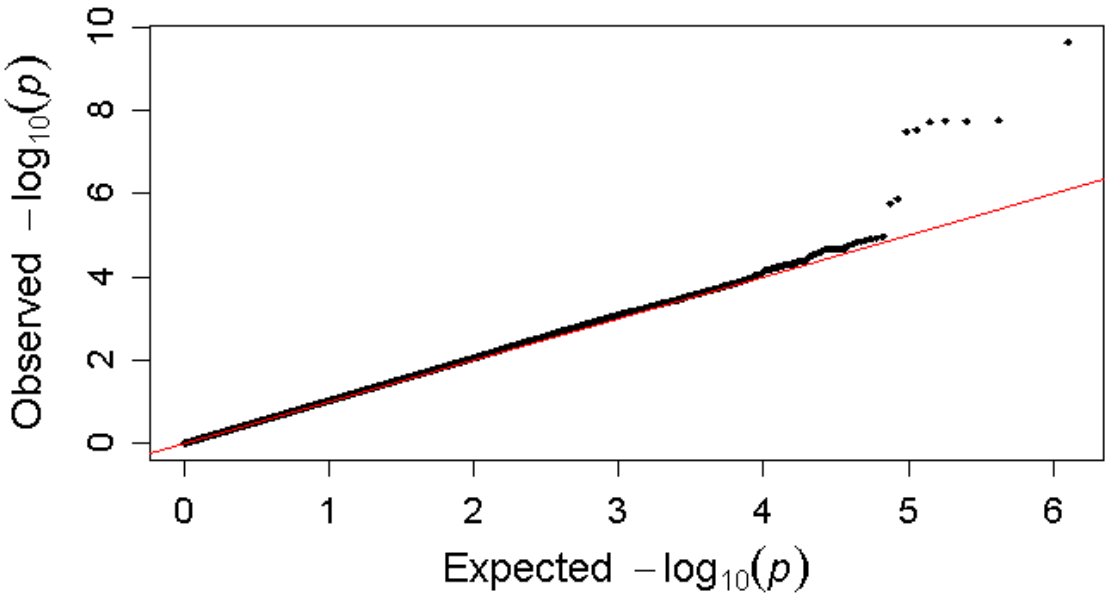

Figure 2.10: The QQ plot of genome-wide p-values for the male data, when data are fitted with the LM model

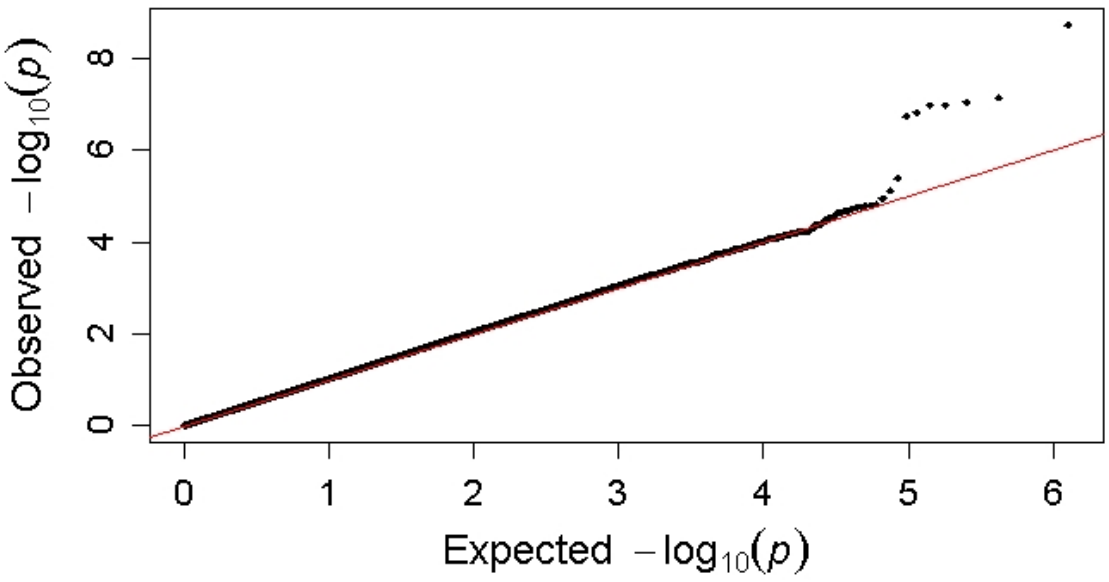

Figure 2.11: The QQ plot of genome-wide p-values for the male data, when data are fitted with the LM-I model

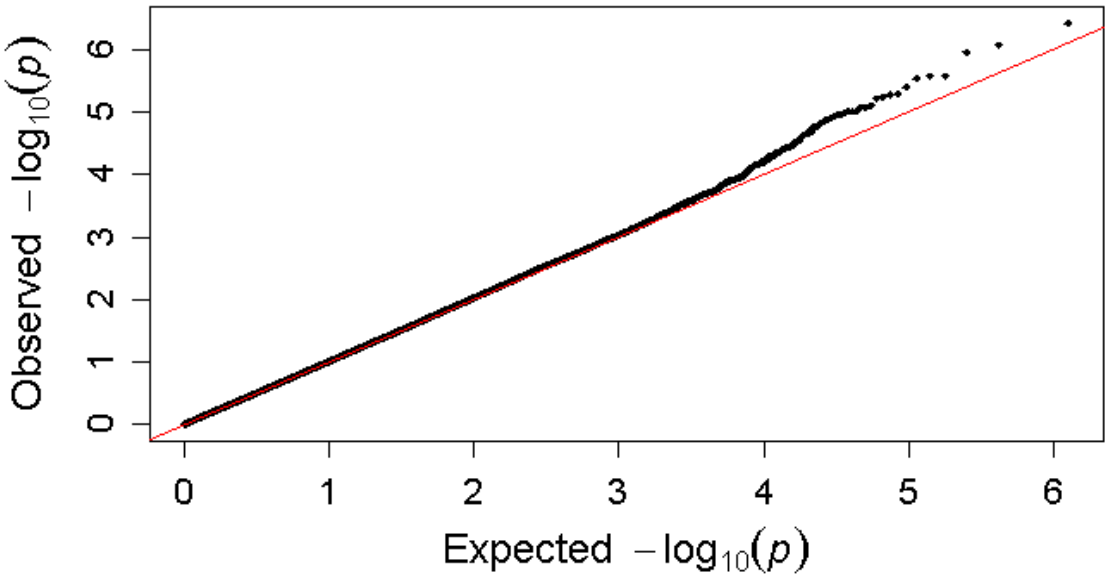

Figure 2.12: The QQ plot of genome-wide p-values for the female data, when data are fitted with the VC model

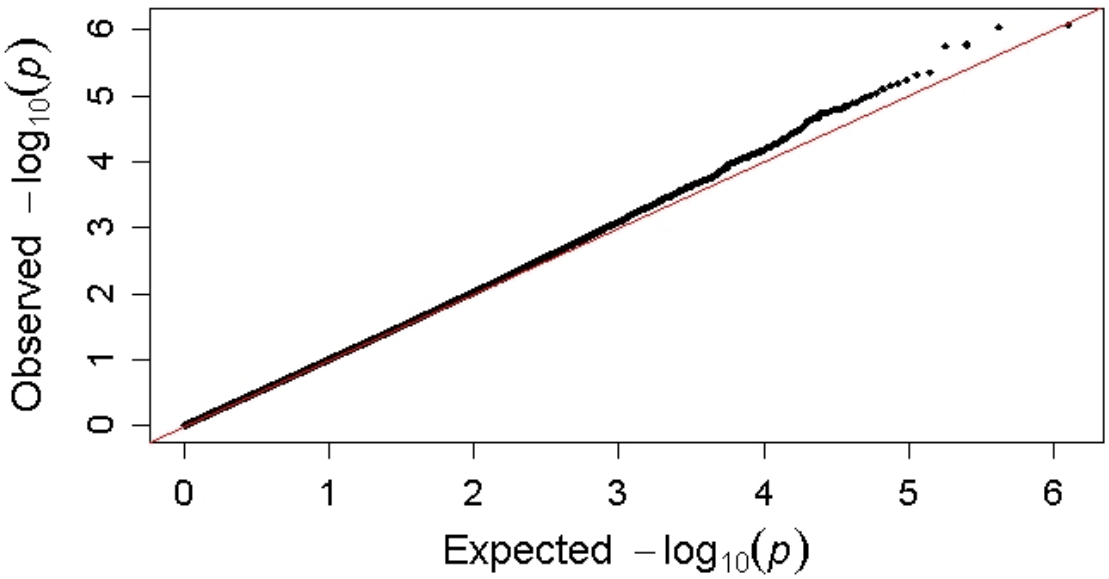

Figure 2.13: The QQ plot of genome-wide p-values for the female data, when data are fitted with the LM model

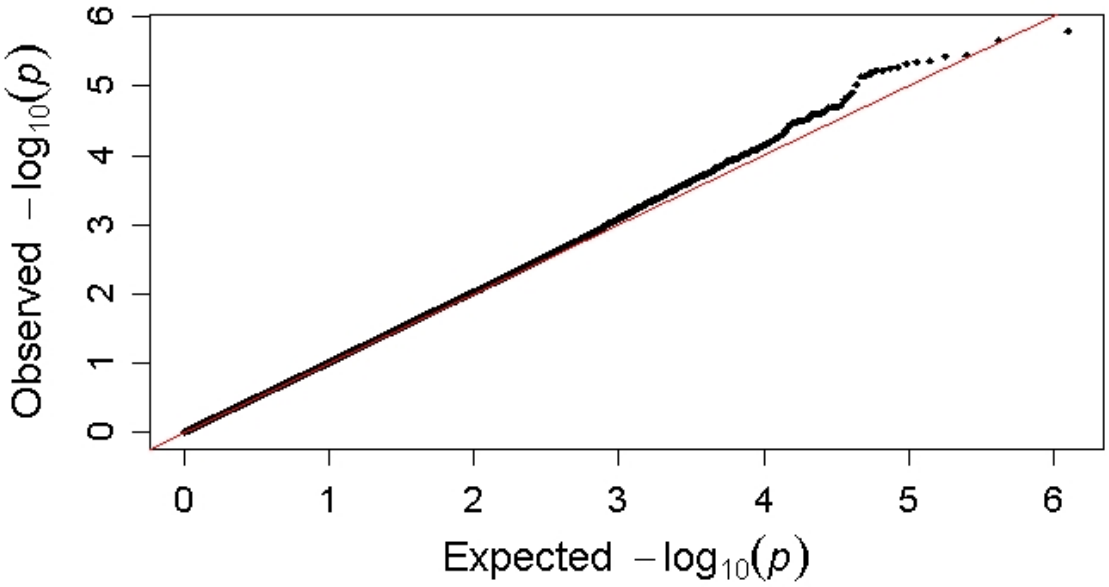

Figure 2.14: The QQ plot of genome-wide p-values for the female data, when data are fitted with the LM-I model

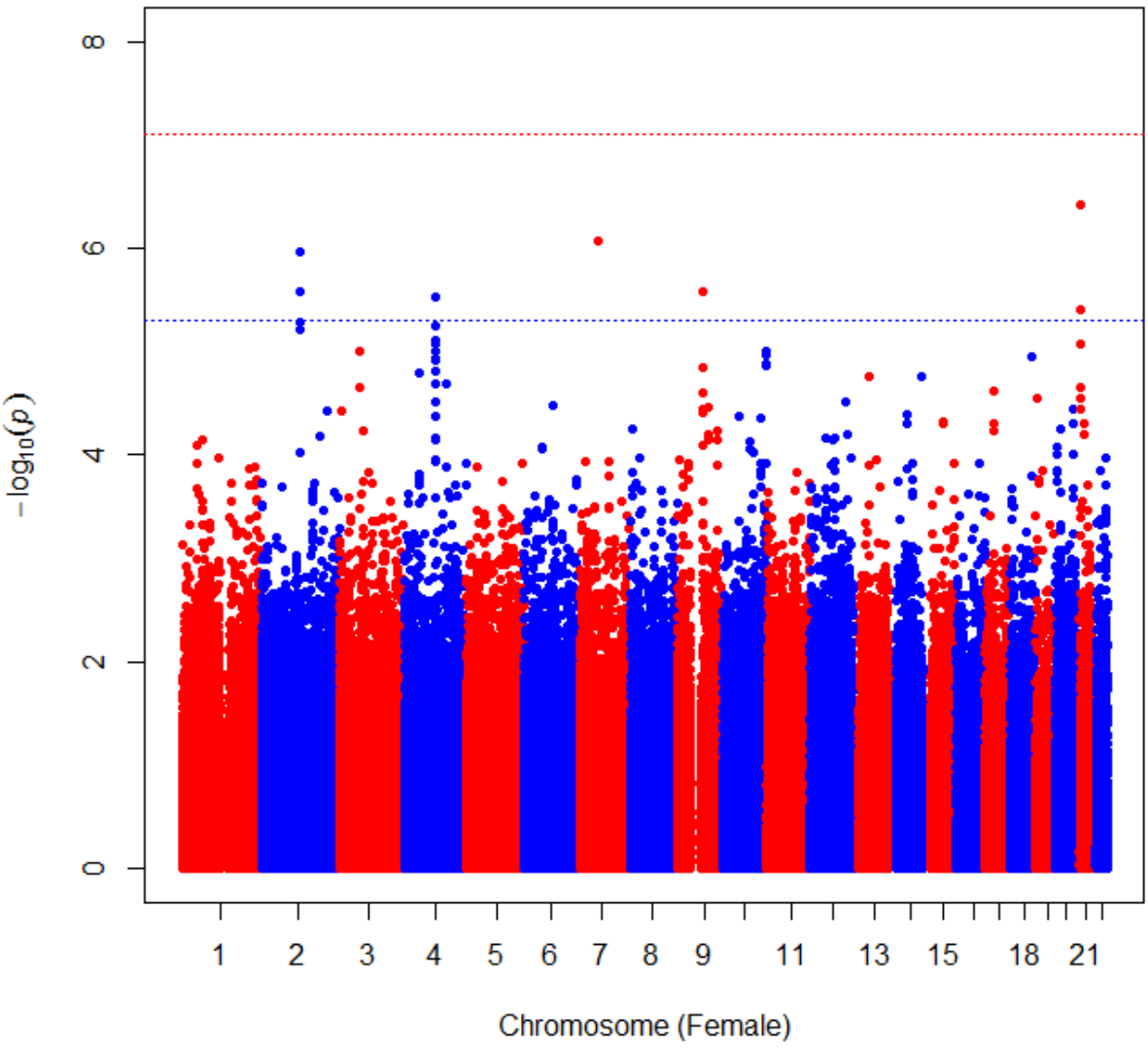

Figure 2.15: The Manhattan plot of -log10(p-values) for testing $H_0 : \beta(X) = 0$ when fitting the VC model to the female data set.

Figures 2.15–2.17 showed the Manhattan plot of the -log10(p-values) for the female data. Even though no SNPs passed the genome-wide Bonferroni threshold, we did see stronger signals fitted by the VC model. Those SNPs that passed the suggestive threshold are listed in Table 2.2. Again, gene TCF7L2 does not show sign of sensitivity to obesity to affect

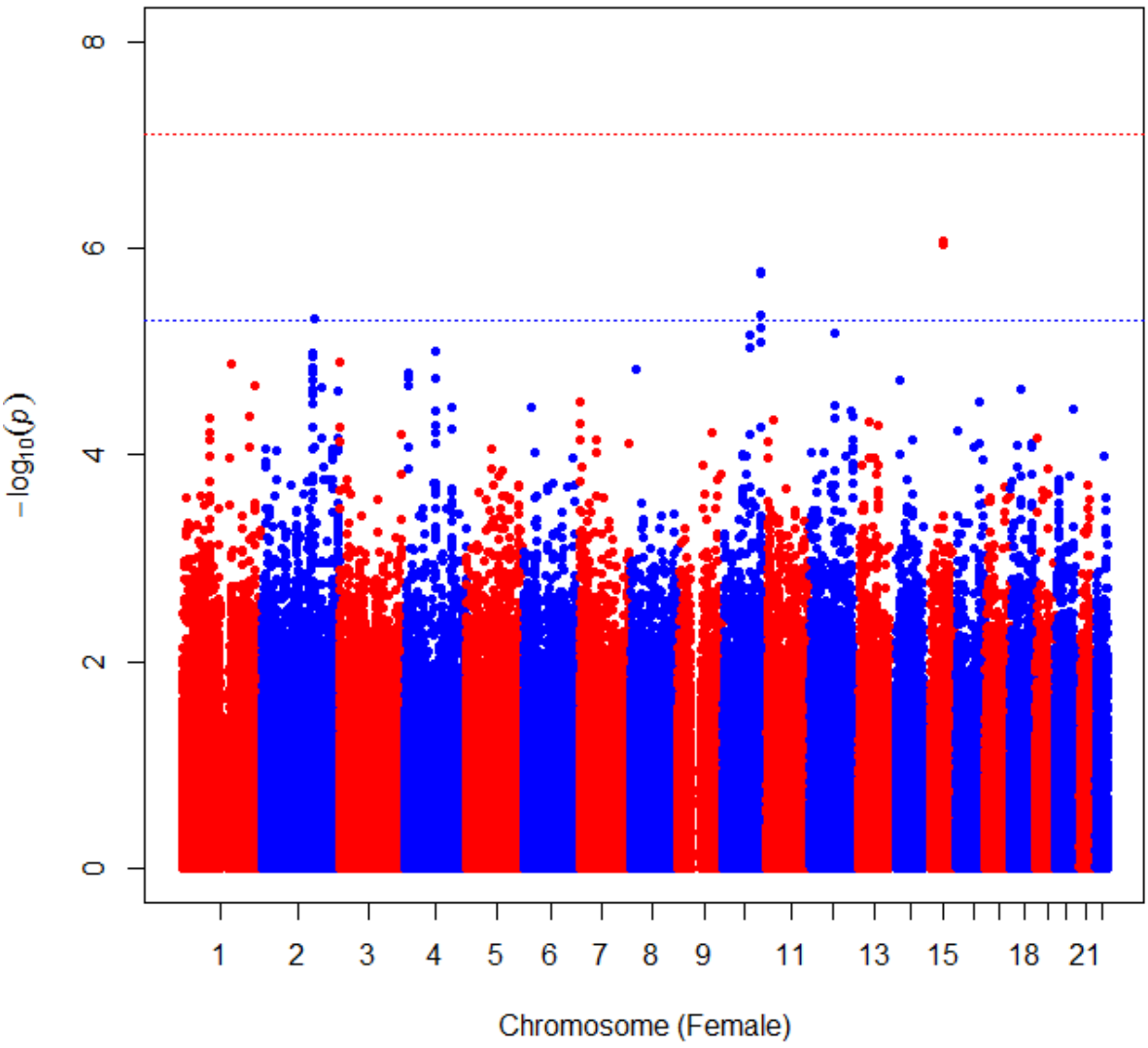

Figure 2.16: The Manhattan plot of -log10(p-values) for testing $H_0 : \beta = 0$ when fitting the LM model to the female data set.

T2D risk. Gene GLI2 shows sign of interaction with obese to affect T2D risk. Two SNPs in gene NRIP1 located on chromosome 21 show sign of nonlinear interaction with obese to affect T2D risk. In comparison to the male data, it is clear that SNP effects are stronger in the male population than in the female population. Moreover, the genetic effects in females are relatively more sensitive to obesity to affect T2D risk. In summary, strong sex-specific

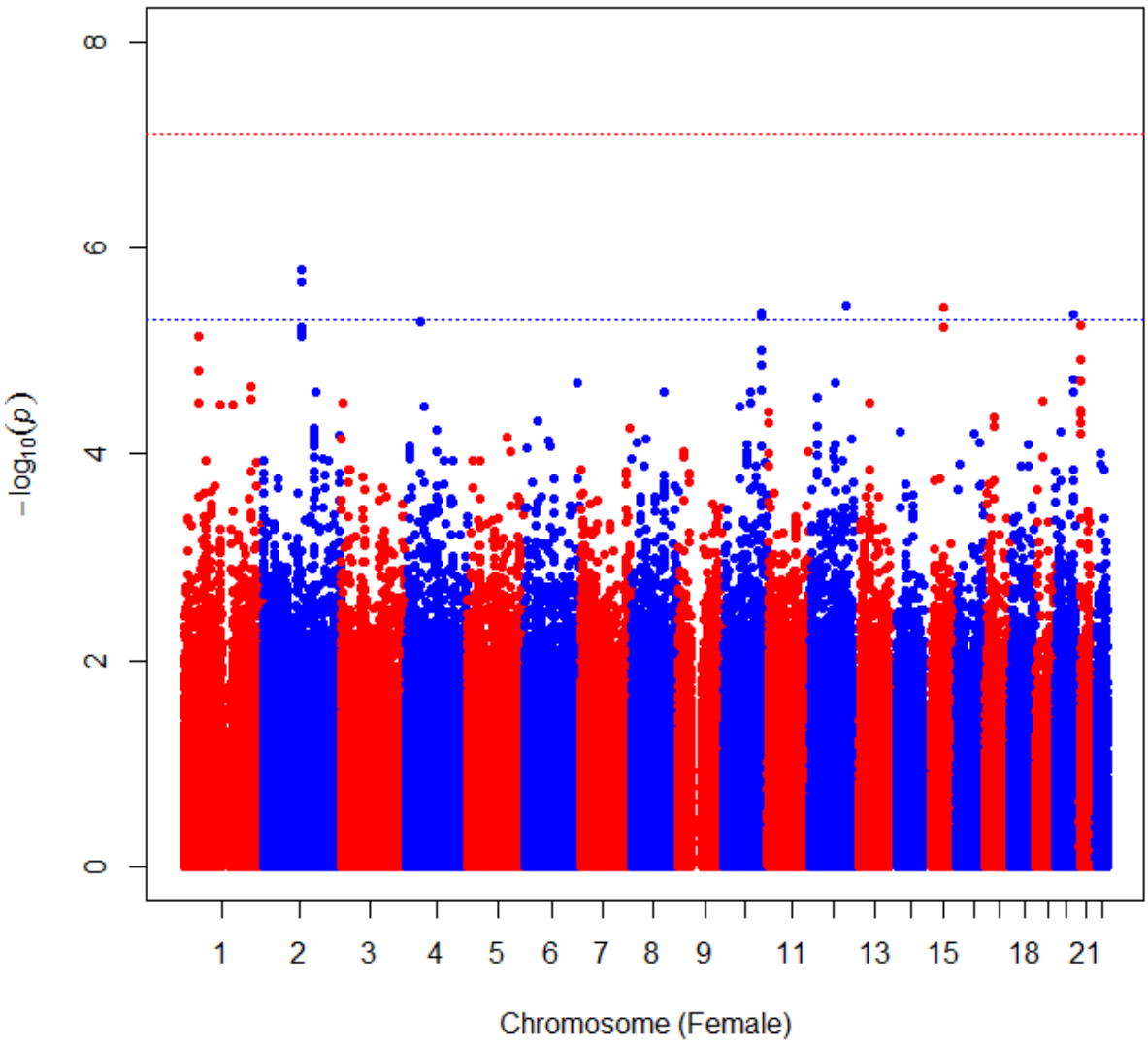

Figure 2.17: The Manhattan plot of -log10(p-values) for testing $H_0 : \beta_1 = \beta_2 = 0$ when fitting the LM-I model to the female data set.

genetic effects were observed, for example, those SNPs on chromosome 2, 3, 4, and 21.

To further demonstrate the utility of the method, we plotted the dynamic effect of SNP rs13050325 on chromosome 21 from the female data (upper panel) and SNP rs4635456 on chromosome 19 from the male data (lower panel). The two curves on the left side of Figure 2.18 showed the estimated dynamic genetic effect as a function of BMI fitted with the B-

Table 2.2: List of SNPs with p-value < 5E-06 in the NHS (Female) data set

SNP ID	GeneName	Chr	P_VC	P_CON	P_LIN
			fitted with VC model		
rs13050325	NRIP1	21	**3.79E-07**	3.77E-06	0.0016
rs2331061	LANCL2	7	**8.60E-07**	1.30E-06	5.26E-07
rs1466042	GLI2	2	**1.10E-06**	3.48E-06	0.0389
rs11145373	VPS13A	9	**2.63E-06**	8.41E-04	2.56E-04
rs3775043	UNC5C	4	**2.95E-06**	0.0018	6.96E-04
rs12627409	NRIP1	21	**4.00E-06**	2.38E-05	0.0441
			fitted with LM model		
rs10519107	RORA	15	4.84E-05	0.8381	-
rs809736	RORA	15	4.96E-05	0.8145	-
rs4506565	TCF7L2	10	4.35E-05	0.4953	-
rs7901695	TCF7L2	10	4.42E-05	0.4895	-
rs12255372	TCF7L2	10	1.2E-04	0.5576	-
rs4368343	Unknown	2	1.88E-04	0.7537	-
			fitted with LMI model		
rs2677528	GLI2	2	2.63E-06	2.62E-05	0.0732
rs7978946	Unknown	12	3.09E-05	0.0117	0.6078
rs887370	TSHZ2	20	3.63E-05	1.45E-05	0.5868
SNP ID	GeneName	Chr	P_LM	P_LMI	P_I
			fitted with VC model		
rs13050325	NRIP1	21	0.0062	1.23E-05	1.0E-04
rs2331061	LANCL2	7	0.0703	0.1456	0.4471
rs1466042	GLI2	2	0.0241	1.61E-06	3.37E-06
rs11145373	VPS13A	9	1.27E-04	6.2E-04	0.7679
rs3775043	UNC5C	4	6.11E-05	2.56E-04	0.4929
rs12627409	NRIP1	21	0.0119	5.60E-06	2.38E-05
			fitted with LM model		
rs10519107	RORA	15	**8.52E-07**	3.72E-06	0.3802
rs809736	RORA	15	**9.22E-07**	5.84E-06	0.8961
rs4506565	TCF7L2	10	**1.69E-06**	4.66E-06	0.2018
rs7901695	TCF7L2	10	**1.75E-06**	4.30E-06	0.1729
rs12255372	TCF7L2	10	**4.47E-06**	9.73E-06	0.1543
rs4368343	Unknown	2	**4.75E-06**	2.53E-05	0.6320
			fitted with LMI model		
rs2677528	GLI2	2	0.0064	**2.16E-06**	1.55E-05
rs7978946	Unknown	12	1.04E-04	**3.62E-06**	0.0016
rs887370	TSHZ2	20	0.4492	**4.46E-06**	9.30E-07

spline function. We can see clear nonlinear genetic effects over BMI, which indicates nonlinear interaction between BMI and the variants. The figures in the right panel show the plot of fitted probabilities against individual BMI values corresponding to different genotypes. We coded the heterozygote as 0 in our model. This implies that the green curves in the two plots correspond to the mean fitted probability when $G = 0$. In general, the risk of T2D increases as BMI increases. This is consistent with our prior knowledge that the disease prevalence is strongly associated with body weight (McCarthy[61]).

For SNP rs13050325 on chromosome 21, the allele frequency for the minor allele G is 0.2587. For SNP rs4635456 on chromosome 19, the allele frequency for the minor allele G is 0.3771. In both cases, the overall trend for T2D risk for the baseline (corresponding to genotype AG) increased as BMI level increases (green curve). However, individuals carrying AA genotype had much higher chance to develop T2D than those carrying AG or GG genotype. Man with genotype AA had the lowest risk of conferring T2D susceptibility when BMI level was below 28 in male and below 33 in female. After the transition points, the AA genotype triggers larger effect, resulting in higher risk of T2D. The association signals for both LM and LM-I model are weaker than the one fitted with the VC model, leading to potential mis-identification of these variants. The results offered personalized preventive suggestions based upon our findings fitted with the VC model. For example, man carrying genotype AA at this SNP locus should pay more attention to control their body weight if their BMI level is above 28 to avoid the risk of T2D.

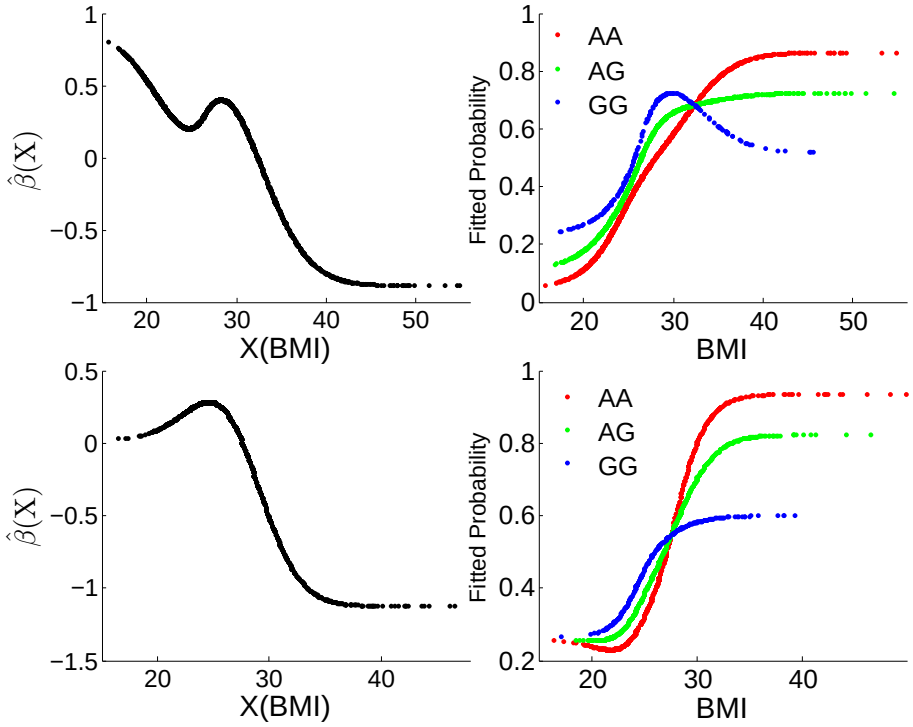

Figure 2.18: The estimated varying-coefficient function and fitted probability of SNP rs13050325 (upper panel) on chromosome 21 of female population and SNP rs4635456 (lower panel) on chromosome 19 of male population

2.7 Discussion

It is broadly recognized that naturally occurring variations in most complex disease traits have a genetic basis. However, the degree of variability is believed to have a strong environmental component in addition to genetic causes for many disease traits such as obesity and Type 2 Diabetes (Qi and Cho [62]). Recent efforts on epigenetics study reveals the importance of epigenetic modification on complex diseases (Liu et al. [63]). These epige-

netic changes involve major chromatin remodeling processes such as DNA methylation and histone modification. Being the environmentally driven plasticity at the DNA level, these structural changes at the DNA level reveal the interplay of gene-environment interaction in the regulation of phenotype, which is increasingly recognized as the epigenetic basis of many complex diseases (Liu et al. [63]). Large efforts have been devoted to the exploration of epigenetic mechanisms for a better understanding of the molecular machinery underlying complex diseases (Feinberg and Irizarry [64]). However, how environment mediates epigenetic changes to affect phenotypic plasticity is still poorly understood, largely due to the lack of powerful statistical methods to dissect this complicated process.

In this chapter, we proposed a novel statistical method by modeling the genotypic effect on disease risk as a dynamic function of environment mediators. Our model is built upon well-studied statistical varying-coefficient model implemented with the nonparametric spline technique to estimate the varying coefficients. The model extends out previously developed method on continuous traits to a case-control population-based design. Simulation studies show dramatically improved power when the underlying genetic penetrance behaves nonlinearly under certain environmental stimulus. Our model can capture the dynamic changes of the gene functions over environmental changes, hence has particular power to tackle long-standing genetic questions regarding gene action and phenotypic plasticity (Feinberg [44]).

Our simulation studies indicate that model mis-specification is an issue in G×E study. The power to detect genetic signals is heavily dependent upon the models to fit the data. Simple models are always the first choice due to their simplicity to interpret. However, if they cannot capture the underlying functional mechanism, they suffer tremendously from power loss. For example, if the true genetic effect does vary nonlinear across environmental changes, fitting a simple linear model would loss power (Figure 2.3). On the other hand,

complex models always suffer from large degrees of freedom for testing. We proposed a sequential testing procedure to assess if a simpler model fits the data better. The real data analysis confirms that this strategy works. For example, when testing constant shows that there is no G×E interaction, the model with linear predictor and without interaction term gives the smallest testing p-values (see Table 2.1 and Table 2.2). In real data analysis, one should always start by assessing constant coefficient first, then move to test linear or varying coefficients.

We applied our model to two Type 2 Diabetes data sets. Cornelis et al. [65] evaluated seven statistical models to dissect G×E interactions using the same data sets. Both Cornelis et al. [65] and our work treated BMI as the environmental factor. Cornelis et al. [65] claimed that specifying BMI as a continuous covariate will lead to inflated type 1 error, which has consequence in detecting increased number of false positives as the true signal. They converted the continuous environment factor BMI into a binary variable prior to further comparisons of all the 7 models. However, this conversion will result in information loss, which might be the reason that no G×E interaction signals passed the genome-wide significance levels for all the seven models in both data sets in their analysis (Cornelis et al. [65]). In our approach, we allowed the nonlinear effect of BMI on Type 2 Diabetes (modeled by function $\alpha(X)$) rather than treating it as a linear function (i.e., $\alpha_0 + \alpha_1 X$). This greatly alleviated the type 1 error inflation compared to a model fitted with a linear function in BMI (data not shown). In our analysis, several signals reached the genome-wide significance level, which is a piece of convincing evidence for keeping the continuous BMI measure as an environmental variable.

In the real data analysis, we observed strong sex-specific variants associated with T2D. There were not much overlap between genes identified in both data sets except for SNPs

in gene TCF7L2. Identification of SNPs in gene TCF7L2 on the pathogenesis of Type 2 Diabetes has been successfully replicated from different populations (Grant et al. [59]). This information indicates the robustness of our model. In addition, we observed stronger signals in the male data evidenced by seven SNPs from three genes reaching the genome-wide significance threshold (cutoff=7.9E-8), as shown in Table (2.1). However, we observed stronger BMI×G interaction to affect T2D in females than in males evidenced by more nonlinear G×E interaction in the female data set (Table 2.2). We could miss these signals if we only focused on linear predictor models. In a recent investigation of a Italian population, Vaccaro et al. [66] found a significantly higher average BMI levels in diabetes women. So possibly certain genes may be sensitive to high BMI level to increase T2D risk. Our model provides a testable framework to identify the underlying genetic blueprint sensitive to obese changes to affect T2D risk. The results obtained by our model can be applied to pathway or gene-set enrichment analysis to identify potential sex-specific pathways for T2D.

In this chapter, we generalized the VC model for continuous quantitative response to the case-control binary response. There is several ongoing work worthy of further investigation. First, the model can be easily extended to other types of phenotype data, such as count data or survival data by applying different link functions. Second, more replication studies are needed by applying our approach to Type 2 Diabetes of different ethnic groups to further confirm the robustness of the method. Third, it is worth noting that the interesting result reported by Perry et al. [52] that stratification on the Type 2 Diabetes patients based on BMI might help enrich the significance of potential susceptibility loci. We could also try to carry out analysis to test if this hypothesis leads to any new discoveries based on the VC model. Finally, our model can easily incorporate population stratification (PS) effect by first doing a principle component analysis using software such as EIGENSTRAT (Price et

al. [67]), then incorporate those PCs as covariates into the model to account for the effect of PS.

Chapter 3

High Dimensional Variable Selection In Gene-Environment Interactions

3.1 Introduction

Gene-environment (G×E) interaction has been traditionally examined by assessing genetic responses corresponding to various environmental stimuli, which provides novel insight in elucidating the genetic basis of complex diseases, because the disease risk is not only contingent on genetic risk factors, but also on the environmental pressures, as well as the interplay between them. The environmental pressure could be either discrete or continuous. When it comes to a G×E interaction study related to asthma, the environmental factor could be discrete, such as smoking status (smoking v.s. non-smoking). A much more clear picture on the interaction will be tangible if the environmental factor is evaluated on a continuous scale, since we can trace the varying patterns of genetic effect responsive to changes in environment.

Conventional statistical modelling of G×E interaction often requires a linear relationship assumption between genetic and environmental factors, which could be violated in practice, as pointed out in Ma et al [37] and Wu and Cui [68]. Consequently, a varying coefficient (VC) model framework, together with a sequence of goodness-of-fit tests, were proposed in

[37] and [68] for continuous and binary responses respectively, to track down the dynamic features of genetic responses to environmental pressures. Because of the particular power and flexibility of VC models to capture the variations in regression coefficients, the framework demonstrated significant advantage over the conventional methods especially in the presence of non-linear G×E interaction

Unlike the predominant single genetic variant based approaches dissecting G×E interactions, such as the parametric methods in Guo [34], non-parametric methods in Ma et al [37] and Wu and Cui [68], and semi-parametric methods in Chatterjee et al [69] and Maity et al [70], we propose a variant set based framework to investigate how variants in a set are mediated by a common environment factor to affect the phenotypic response, since it has been increasing acknowledged the merit of set based association analysis, such as in the gene-centric analysis in Cui et al [39] and Wu and Cui [40], gene-set analysis in Schaid et al [71] and Efron and Tibshirani [72], as well as the pathway based analysis in Wang et al [38]. When the number of variants within the genetic system is large, the problem can be approached from the entry point of high dimensional variable selection. In particular, we can select genetic variants with varying, non-zero constant and zero coefficients, which are corresponding to scenarios of G×E interactions, no G×E interactions and no genetic effects, respectively. To the best of our knowledge, this is the first time that the problem is tackled from the angle of high dimension variable, on the contrary to the popular single genetic variant based approaches coined in a hypothesis testing framework.

Through B spline basis expansion, the varying coefficient function can be separated into constant and varying portions, respectively. Then the distinction of the 3 effects could be achieved by penalizing the 2 portions in a two-stage iterative framework, as shown in Tang et al.[73]. Though the asymptotic properties of the two stage estimator with adaptive

LASSO penalty were established in [73], the finite sample performance still has a large margin to improve. Therefore we proposed a Smoothly Clipped Absolute Deviation (SCAD) based approach to examine the separation of varying, non-zero constant and zero coefficient functions. Our approach has significantly improved percentages of choosing the exact true model and reduced error in parameter estimation. Assuming suitable regularity conditions, we can establish the consistency in variable selection and effect separation of our estimator, as well as the optimal convergence rates of the estimates for varying effect. Furthermore, it can be shown that the estimate of non-zero constant coefficient enjoys the oracle property, that is, the asymptotic distribution of the non-zero constant coefficient function is the same as that when the true model is known in priori.

In this chapter, we describe the penalized least square estimation procedure via basis expansion and SCAD penalty, as well as the computational algorithms. Next we present the theoretical results including consistency in variable selection and oracle property. The merit of the proposed approaches were demonstrated through extensive simulation study and real data analysis. Discussions will be given at the end of the chapter. We relegate technical proofs to the Appendix.

3.2 The proposed variable selection method

3.2.1 The penalized estimation via SCAD

Let $(\mathbf{X}_i, Y_i, Z_i)$, $i = 1, \ldots, n$ be independent and identically distributed (i.i.d.) random vectors, then the varying coefficient (VC) model, proposed by Hastie and Tibshirani [74],

has the form

$$Y_i = \sum_{j=0}^{d} \beta_j(Z_i) X_{ij} + \varepsilon_i \tag{3.1}$$

where X_{ij} is the jth component of $(d+1)$-dimensional vector $\mathbf{X}_i$ with the first component X_{i0} being 1, $\beta_j(\cdot)$'s are unknown varying-coefficient functions, Z_i's are the scalar index variable, and ε_i is the random error such that $E(\varepsilon|X,Z) = 0$ and $Var(\varepsilon|X,Z) = \sigma^2 < \infty$.

The smooth functions $\{\beta_j(\cdot)\}_{j=0}^{d}$ in (3.1) can be approximated by polynomial splines. Without loss of generality, suppose that $Z \in [0,1]$. Let w_k be a partition of the interval [0,1], with k_n uniform interior knots

$$w_k = \{0 = w_{k,0} < w_{k,1} < \ldots < w_{k,k_n} < w_{k,k_n+1} = 1\}$$

Let $\mathcal{F}_n$ be the collection of functions on [0,1] satisfying (3.1) the function is a polynomial of degree p or less on subintervals $I_s = [w_{k,s}, w_{k,s+1}), s = 0, \ldots, N_n - 1$ and $I_{N_n} = [w_{j,N_n}, w_{j,N_n+1})$. (2) the functions are $p-1$ times continuous differentiable on [0,1]. Let $\bar{B}(\cdot) = \{\bar{B}_{jl}(\cdot)\}_{l=1}^{L_j}$ be a set of normalized B spline basis of $\mathcal{F}_n$. Then for $j = 0, \ldots, d$, the VC functions can be approximated by basis functions $\beta_j(Z) \approx \sum_{l=1}^{L_j} \bar{\gamma}_{jl} \bar{B}_{jl}(Z)$, where L_j is the number of basis functions in approximating the functions $\beta_j(Z)$. By changing of equivalent basis, the basis expansion can be reexpressed as

$$\beta_j(\cdot) \approx \sum_{l=1}^{L_j} \gamma_{jl} B_{jl}(\cdot) \doteq \gamma_{j1} + \tilde{B}_j^T(\cdot)\gamma_{j,*}$$

the spline coefficient vector $\boldsymbol{\gamma}_j \doteq (\gamma_{j,1}, \boldsymbol{\gamma}_{j*}^T)^T$, and $\tilde{B}_j(\cdot) = (B_{j2}(\cdot), \ldots, B_{jL_j}(\cdot))^T$. $\gamma_{j,1}$ and $\boldsymbol{\gamma}_{j*}$ correspond to the constant and varying part of the coefficient function respectively. We treat $\boldsymbol{\gamma}_{j*}$ as a group. If $\|\boldsymbol{\gamma}_{j*}\|_2$=0, then the jth predictor only has a non-zero constant effect

and moreover, if $\gamma_{j,1}$=0, then the predictor is redundant.

To carry out variable selection separating the varying, non-zero constant, and zero effects, we minimize the penalized least square function

$$\begin{aligned} Q(\boldsymbol{\gamma}) =& \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \sum_{j=0}^{d}\sum_{l=1}^{L}\gamma_{jl}X_{ij}B_{jl}(Z_i)\right]^2 + \sum_{j=1}^{d} p_{\lambda_1}(\|\boldsymbol{\gamma}_{j*}\|_2) \\ &+ \sum_{j=1}^{d} p_{\lambda_2}(|\gamma_{j1}|)I(\|\boldsymbol{\gamma}_{j*}\|_2 = 0) \end{aligned} \tag{3.2}$$

where λ_1 and λ_2 are the penalization parameters, $p_\lambda(\cdot)$ is the SCAD penalty function, defined as

$$p_\lambda(x) = \begin{cases} \lambda x & \text{if } 0 \le x \le \lambda \\ -\frac{(x^2 - 2a\lambda x + \lambda^2)}{2(a-1)} & \text{if } \lambda < x \le a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } x > a\lambda \end{cases} \tag{3.3}$$

To express (3.2) by vectors and matrices, we redefine

$$\begin{aligned} Q(\boldsymbol{\gamma}) =& (\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{\gamma})^T(\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{\gamma}) + n\sum_{j=1}^{d} p_{\lambda 1}(\|\gamma_{j*}\|_2) \\ &+ n\sum_{j=1}^{d} p_{\lambda 2}(|\gamma_{j1}|)I(\|\boldsymbol{\gamma}_{j*}\|_2 = 0) \end{aligned} \tag{3.4}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_0^T, \ldots, \boldsymbol{\gamma}_d^T)^T$, $U_i = (X_{i0}B(Z_i)^T, \ldots, X_{id}B(Z_i)^T)^T$, and $\boldsymbol{U} = (U_1^T, \ldots, U_n^T)^T$. The function of the 2nd term of $Q(\boldsymbol{\gamma})$ is to separate the varying and constant effects by penalizing the L_2 norm of the varying part of the coefficient functional. The indicator functions in the 3rd term helps to penalize the variables of the constant effects. Both $\gamma_{j,1}$ and $\boldsymbol{\gamma}_{j,*}$ will be shrunk to zero if the predictor $\boldsymbol{X}_j$ has no effect.

3.2.2 The computational algorithms

The SCAD penalty function is singular at the origin, and do not have continuous 2nd order derivatives, therefore the regular gradient-based optimization cannot be applied. In this section, we develop an iterative two-stage algorithm to minimize the penalized loss function using local quadratic approximation to the SCAD penalty. Following Fan and Li [75], in a neighbourhood of a given positive $x_0 \in \mathbb{R}^+$,

$$p_\lambda(x) \approx p_\lambda(x_0) + \frac{p_\lambda^{'}(x_0)}{2x_0}(x^2 - x_0^2)$$

where $p_\lambda^{'}(x) = \lambda\{I(x \leqslant \lambda) + \frac{(a\lambda - x)_+}{(a-1)\lambda} I(x > \lambda)\}$ for a=3.7 and x >0. Here we use a similar quadratic approximation by substituting x with $\|\boldsymbol{\gamma}_{j*}\|_2$ and $|\gamma_{k1}|$ in LQA, for $k = 0, ..., d$. Therefore we have

$$p_\lambda(\|\boldsymbol{\gamma}_{j*}\|_2) \approx p_\lambda(\|\boldsymbol{\gamma}_{j*}^0\|_2) + \frac{p_\lambda^{'}(\|\boldsymbol{\gamma}_{j*}^0\|_2)}{2\|\boldsymbol{\gamma}_{j*}^0\|_2}(\|\boldsymbol{\gamma}_{j*}\|_2^2 - \|\boldsymbol{\gamma}_{j*}^0\|_2^2) \tag{3.5}$$

and

$$p_\lambda(|\gamma_{j,1}|) \approx p_\lambda(|\gamma_{j,1}^0|) + \frac{p_\lambda^{'}(|\gamma_{j,1}^0|)}{2|\gamma_{j,1}^0|}(|\gamma_{j,1}|^2 - |\gamma_{j,1}^0|^2) \tag{3.6}$$

The sets of predictors with varying, non-zero constant, and zero effects are termed as $\mathcal{V}$, $\mathcal{C}$ and $\mathcal{Z}$ respectively. We implement the iterative algorithm in the following two-stage procedure. At stage 1, using the LQA (3.5) and dropping the irrelevant constant terms, we minimize

$$Q_1(\boldsymbol{\gamma}) = (\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{\gamma})^T (\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{\gamma}) + \frac{n}{2}\boldsymbol{\gamma}^T \boldsymbol{\Omega}_{\lambda_1}(\boldsymbol{\gamma}_0)\boldsymbol{\gamma} \tag{3.7}$$

where the initial spline vector $\boldsymbol{\gamma}_0$ is the unpenalized estimator, $\boldsymbol{\Omega}_{\lambda_1}(\boldsymbol{\gamma}_0)$=diag$\{\boldsymbol{\Omega}_0, \boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_d\}$,

where $\boldsymbol{\Omega}_0 = \mathbf{0}_L$, $\boldsymbol{\Omega}_j = \left\{0, \frac{p'_{\lambda 1}(\|\boldsymbol{\gamma}^0_{j*}\|_2)}{\|\boldsymbol{\gamma}^0_{j*}\|_2}, \ldots, \frac{p'_{\lambda 1}(\|\boldsymbol{\gamma}^0_{j*}\|_2)}{\|\boldsymbol{\gamma}^0_{j*}\|_2}\right\}_L$ for $j = 1, \ldots, d$. Hence the estimator can be iteratively obtained as

$$\hat{\boldsymbol{\gamma}}^{(m)}_{\mathcal{VC}} = \left\{\boldsymbol{U}^T\boldsymbol{U} + \frac{n}{2}\boldsymbol{\Omega}_{\lambda_1}(\hat{\boldsymbol{\gamma}}^{(m-1)}_{\mathcal{VC}})\right\}^{-1}\boldsymbol{U}^T\boldsymbol{Y} \tag{3.8}$$

Suppose that all the predictors are in $\mathcal{V}$ at the beginning. The jth predictor will be moved to $\mathcal{C}$ if $\|\hat{\gamma}^{\mathcal{VC}}_{j*}\|_2$=0, otherwise it will stay in $\mathcal{V}$.

At stage 2, using the LQA (3.6) and dropping the irrelevant constant terms, we minimize the penalized loss only for the predictors in $\mathcal{C}$:

$$Q_2(\boldsymbol{\gamma}) = (\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{\gamma})^T(\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{\gamma}) + \frac{n}{2}\boldsymbol{\gamma}^T\boldsymbol{\Omega}_{\lambda_2}(\hat{\boldsymbol{\gamma}}_{\mathcal{VC}})\boldsymbol{\gamma} \tag{3.9}$$

where $\boldsymbol{\Omega}_{\lambda_2}(\hat{\boldsymbol{\gamma}}_{\mathcal{VC}})$=diag$\{\boldsymbol{\Omega}_0, \boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_d\}$ with $\boldsymbol{\Omega}_0 = \mathbf{0}_L$,
$\boldsymbol{\Omega}_j = \left\{\frac{p'_{\lambda 2}(|\hat{\gamma}^{\mathcal{VC}}_{j,1}|)}{|\hat{\gamma}^{\mathcal{VC}}_{j,1}|} I(\|\hat{\boldsymbol{\gamma}}^{\mathcal{VC}}_{j*}\|_{L_2} = 0), 0, \ldots, 0\right\}_L$. The estimator can be iteratively obtained as

$$\hat{\boldsymbol{\gamma}}^{(m)}_{\mathcal{CZ}} = \left\{\boldsymbol{U}^T\boldsymbol{U} + \frac{n}{2}\boldsymbol{\Omega}_{\lambda_2}(\hat{\boldsymbol{\gamma}}^{(m-1)}_{\mathcal{CZ}})\right\}^{-1}\boldsymbol{U}^T\boldsymbol{Y} \tag{3.10}$$

If the jth predictor is in $\mathcal{C}$, then it will be moved to $\mathcal{Z}$ if $|\hat{\gamma}^{\mathcal{CZ}}_{k,1}|$=0, otherwise it stays in $\mathcal{C}$.

We can obtain the estimator $\hat{\boldsymbol{\gamma}}$ at convergence from the iterative procedure between the above two stages, and the estimated coefficient function in (3.1) as $\hat{\beta}_j(z) = B^T(z)\hat{\boldsymbol{\gamma}}_j$. $\hat{\boldsymbol{\beta}}_j(z)$ will be a varying function in z, non-zero constant and zero if $\hat{\boldsymbol{\gamma}}_j$ is in $\mathcal{V}$, $\mathcal{C}$ and $\mathcal{Z}$ correspondingly.

3.2.3 Selection of tuning parameters

In this section, we choose the tuning parameters N,p, λ_1 and λ_2 from a data driven procedure. N is the number of interior knots uniformly spaced on [0,1], p is the degree of the spline basis. here p and N control the smoothness of the coefficient functions, while λ_1 and λ_2 determine the threshold for variable selection.

At the beginning, we use BIC in Schwarz [76] to choose N and p. The range for N is $[\max(\lfloor 0.5n^{\frac{1}{(2p+3)}} \rfloor, 1), \lfloor 1.5n^{\frac{1}{(2p+3)}} \rfloor]$, where $\lfloor x \rfloor$ denotes the integer part of x. The optimal pair of N and p can be achieved via a two-dimensional grid search, according to the following criterion:

$$\mathrm{BIC}_{N,p} = \log(\mathrm{RSS}_{N,p}) + \frac{(N+p+1)}{n}\log(n)$$

where $\mathrm{RSS}_{N,p} = (\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{\gamma}})^T(\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{\gamma}})/n$, $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\gamma}}_0^T, \mathbf{0}^T, \ldots, \mathbf{0}^T)^T$. Conditional on the selected N and p, λ_1 is the minimizer of

$$\mathrm{BIC}_{\lambda_1} = \log(\mathrm{RSS}_{\lambda_1}) + \frac{df_{\lambda_1}}{n}\log(n)$$

where $\mathrm{RSS}_{\lambda_1} = (\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{\gamma}}_{\lambda_1})^T(\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{\gamma}}_{\lambda_1})/n$, $\hat{\boldsymbol{\gamma}}_{\lambda 1}$ is the minimizer of (3.7), and df_{λ_1} is the effective degree of freedom, defined as the total number of predictors in $\mathcal{V}$ and $\mathcal{C}$.

Conditional on $\hat{\boldsymbol{\gamma}}_{\lambda_1}$, λ_2 is the minimizer of

$$\mathrm{BIC}_{\lambda_2} = \log(\mathrm{RSS}_{\lambda_2}) + \frac{df_{\lambda_2}}{n}\log(n)$$

where $\mathrm{RSS}_{\lambda_2} = (\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{\gamma}}_{\lambda_2})^T(\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{\gamma}}_{\lambda_2})/n$, , $\hat{\boldsymbol{\gamma}}_{\lambda_2}$ is the minimizer of (3.8), and df_{λ_2} is the effective degree of freedom, defined similarly as df_{λ_1}.

3.3 Asymptotic results

Here we establish the asymptotic properties of the penalized least square estimators. Without loss of generality, we assume there are v varying coefficients as $\beta_j(\cdot) \equiv \beta_j(z)$,$j = 1, \ldots, v$, $(c - v)$ non-zero constant coefficients as $\beta_j(\cdot) \equiv \beta_j > 0$, $j = v + 1, \ldots, c$, and $(d - c)$ zero coefficients as $\beta_j(\cdot) \equiv 0$, $j = (c+1), \ldots, d$. Our asymptotic results are based on the following assumptions.

(A1) Let $\mathcal{H}_r$ be the collection of all functions on the compact support [0,1] such that the r_1th order derivatives of the functions are Hölder of order b with $r = r_1 + r_2$, i.e., $|h^{r1}(z_1) - h^{r1}(z_2)| \leq C_0|z_1 - z_2|^{r2}$ where $0 \leq z_1, z_2 \leq 1$ and C_0 is a finite positive constant. Then $\beta_j(z) \in \mathcal{H}_r$, $j = 0, 1, \ldots, v$, for some $r \geq \frac{3}{2}$.

(A2) The density function of the index variable Z, $f(z)$, is continuous and bounded away from 0 and infinity on $[0, 1]$, i.e., there exist finite positive constants C_1 and C_2 such that $C_1 \leq f(z) \leq C_2$ for all $z \in [0, 1]$.

(A3) Let $\lambda_0 \leq \ldots \leq \lambda_d$ be the eigenvalues of $E[\mathbf{X}\mathbf{X}^T|Z = z]$. Then λ_j $(k = 0, \ldots, d)$ are uniformly bounded away from 0 and infinity in probability. In addition, the random design vector are bounded in probability.

(A4) For w_j, the partition of the compact interval [0,1] defined as $\{0 = w_{j,0} < w_{j,1} < \ldots < w_{j,k_n} < w_{j,k_n+1} = 1\}$, $j = 0, \ldots, d$, there exists finite positive constant C_3 such that

$$\frac{\max(w_{j,k+1} - w_{j,k}, k = 0, \ldots, k_n)}{\min(w_{j,k+1} - w_{j,k}, k = 0, \ldots, k_n)} \leq C_3$$

(A5) The tuning parameters satisfy $k_n^{\frac{1}{2}} \max\{\lambda_1, \lambda_2\} \to 0$ and $n^{\frac{1}{2}} k_n^{-1} \min\{\lambda_1, \lambda_2\} \to \infty$.

(A6) $\max_j\{|p''_{\lambda_1}(|\gamma_{j*}|)| : \gamma_{j*} \neq 0\} \to 0$ as $n \to \infty$ and $\max_j\{|p''_{\lambda_2}(|\gamma_{j1}|)| : \gamma_{j1} \neq 0\} \to 0$

as $n \to \infty$

(A7) $\liminf_{n\to\infty} \liminf_{\theta\to 0^+} \lambda_1^{-1} p'_{\lambda_1}(\theta) > 0$ and $\liminf_{n\to\infty} \liminf_{\theta\to 0^+} \lambda_2^{-1} p'_{\lambda_2}(\theta) > 0$

The above assumptions are commonly used in literature of polynomial splines and variable selections. The assumption similar to (A1) could be found in Kim [77] and Tang et al [73]. (A1) guarantees certain degrees of smoothness of the true coefficient function in order to improve goodness of approximation. (A2) and (A3) are similar to those in Huang et al [50, 51] and Wang et al [78]. (A4) suggests that the knot sequence is quasi-uniform on [0,1], by Schumaker [79]. (A5-A7) are conditions on tuning parameters, of which (A5) could be found in Tang et al [73]; (A6) and (A7) are similar to those in Fan and Li [75] and Wang et al [78].

Theorem 1. Under the assumptions (A1-A7) and suppose $k_n = O_p\left(n^{\frac{1}{2r+1}}\right)$, then we have

(1) $\hat{\beta}_j(z)$ are nonzero constant, $j = v+1, \ldots, c$ and $\hat{\beta}_j(z) = 0$, $j = c+1, \ldots, d$, with probability approaching 1;

(2) $\|\hat{\beta}_j - \beta_j\|_2 = O_p(n^{\frac{-r}{2r+1}})$, $j = 0, \ldots, v$.

Denote $\beta^* = (\beta_{v+1}, \ldots, \beta_c)^T$ as the vector of true nonzero constant coefficients.

Theorem 2. Under the assumptions (A1-A7) and suppose $k_n = O_p(n^{\frac{1}{2r+1}})$, then with $n \to \infty$,

$$\sqrt{n}(\hat{\beta}^* - \beta^*) \xrightarrow{d} N(0, \sigma^2 \Sigma^{-1})$$

where Σ is defined in the Appendix.

3.4 Simulation

The performance of our proposed approach is demonstrated through extensive simulation study in this section. We use the percentage of choosing the true model out of total R replicates, or oracle percentage, to evaluate the accuracy of variable selection by identifying varying, non-zero constant and zero effects. The precision of estimation is assessed by integrated mean squared error (IMSE).

Let $\hat{\beta}_j^{(r)}$ be the estimator of a nonparametric function β_j in the rth ($1 \leqslant r \leqslant R$) replication, and $\{z_m\}_{m=1}^{n_{\text{grid}}}$ be the grid points where $\hat{\beta}_j^{(r)}$ is evaluated. We use the integrated mean squared error (IMSE) of $\hat{\beta}_k(x)$, defined as IMSE$(\hat{\beta}_j(z))$=$\frac{1}{R}\sum_{r=1}^{R}\frac{1}{n_{\text{grid}}}\sum_{m=1}^{n_{\text{grid}}}\{\hat{\beta}_k^{(r)}(z_m) - \beta_j(z_m)\}^2$, to evaluate the estimation accuracy of coefficient β_j, and the total integrated mean squared error (IMSE) of all the d coefficients (TIMSE), defined as TIMSE=$\sum_{j=1}^{d}\hat{\beta}_j(z)$, is used to evaluate the overall estimation accuracy. Note that IMSE$(\hat{\beta}_j)$ will be reduced to MSE$(\hat{\beta}_j)$ when $\hat{\beta}_j$ is a constant.

Example 3.1. We simulate data from the following VC model

$$Y_i = \beta_0(Z_i) + \sum_{j=1}^{d}\beta_j(Z_i)X_{ij} + \varepsilon_i$$

where the index variable Z_i $\sim$Uniform(0,1), and the predictors X_i are generated from a multivariate normal distribution with mean $\mathbf{0}$ and $Cov(Z_{ij}, Z_{ij'}) = 0.5^{|j-j'|}$ for $0 \leq j, j' \leq d$. The performance is evaluated under both d=10 and 50. X_i^j, $j = 0, 1, 2$ are of varying effects, X_i^j, $j = 3, 4$ are of non-zero constant effects, and the rest variables are redundant. The random error ε_i were generated from standard normal distributions and t distribution with 3 degrees of freedom respectively. The coefficients were set as: $\beta_0(z) = \sin(2\pi z)$, $\beta_1(z) = 2 - 3\cos\{(6z-5)\pi/3\}$, $\beta_2(z) = 3(2z-1)^3$, $\beta_3(z) = 2$, $\beta_4(z) = 2.5$, and $\beta_j(z) = 0$

for $j = 5, \ldots, 10$. The results are listed in Figure 3.1 and Table 3.1.

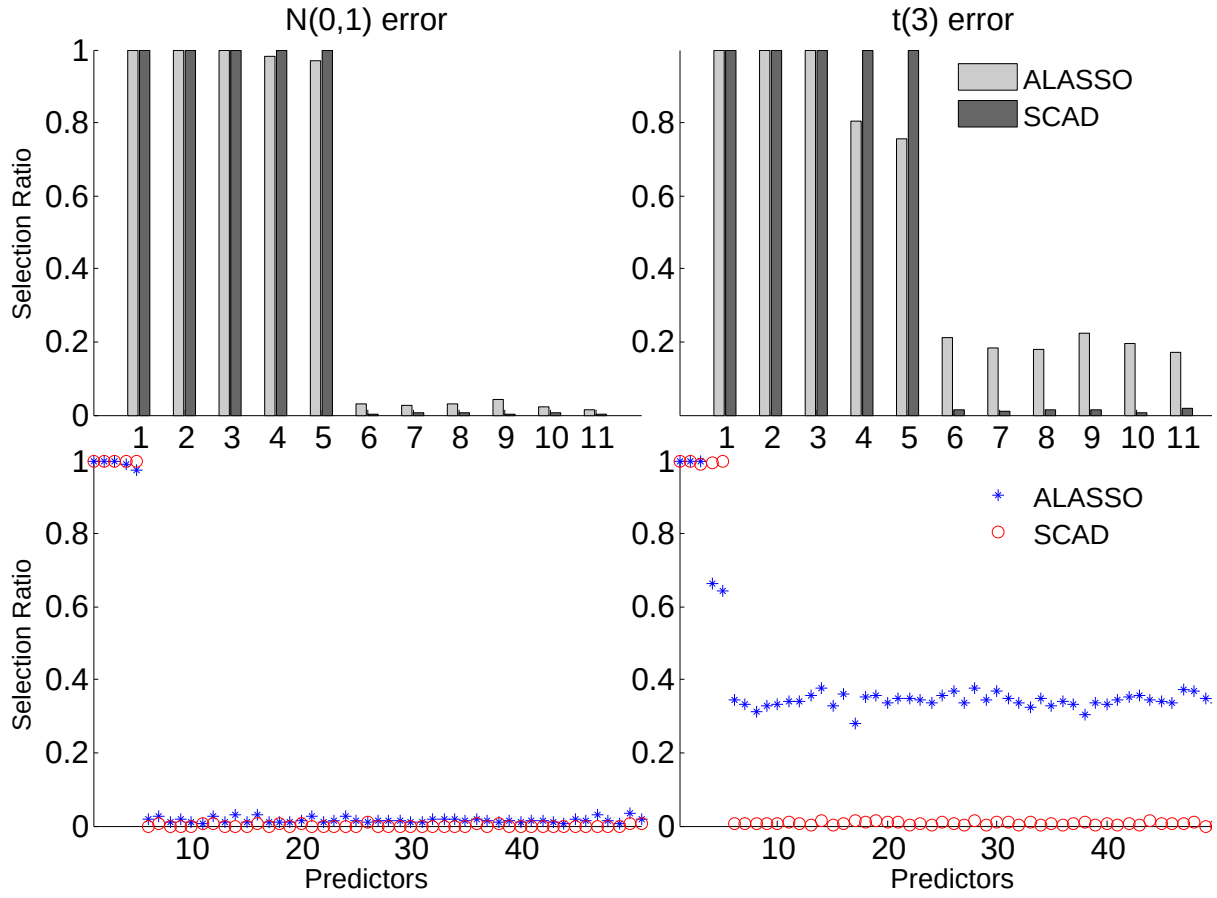

Figure 3.1: The selection ratio of Example 3.1

Figure 3.1 shows the selection ratio for predictors under different groups of d and error distributions. The height of bars on the top panel of Figure 3.1 denote the selection ratio for true positives for the first 5 predictors, and false positives for the rest predictors. Under both standard normal and and $t(3)$ error, compared with the method based on adaptive LASSO, our method is capable of correctly identifying significant effect with high percentages, and has kept a very small percentages of choosing false positive predictors. In addition, our method has relative stable performance in terms of correct selection ratio when dimension grows.

The oracle percentage and parameter estimation results are summarized in Table 3.1. In Tang et al [73], MSE was computed for constant coefficients only when the corresponding predictor is chosen with non-zero constant effect. To reflect the overall estimation precision, we compute IMSEs for all predictors, including β_4 and β_5. When β_j $(j = 4, 5)$ is selected as non-zero constant, IMSE reduces to MSE. The IMSEs will be calculated if β_j $(j = 4, 5)$ incorrectly identified as varying. In all the pairs of d and error distribution type, the SCAD approach demonstrates superior performance over the adaptive LASSO approach.

Example 3.2. Now we consider the simulation in genetics settings from the following VC model

$$Y_i = \beta_0(Z_i) + \sum_{j=1}^{d} \beta_j(Z_i) X_{ij} + \varepsilon_i$$

where the SNP X_i was coded with 3 categories (1,0,-1) for genotypes (AA,Aa,aa) respectively. We simulate the SNP genotype data based on the pairwise linkage disequilibrium(LD) structure. Suppose the two risk alleles A and B of two adjacent SNPs have the minor allele frequencies (MAFs) q_A and q_B, respectively, with LD denoted as δ. Then the frequencies of four haplotypes can be expressed as $q_{ab} = (1 - q_A)(1 - q_B) + \delta$, $q_{Ab} = q_A(1 - q_B) - \delta$,

Table 3.1: Simulation results of Example 3.1

		$N(0,1)$ error			$t(3)$ error		
		SCAD	ALASSO	Oracle	SCAD	ALASSO	Oracle
d=10	Oracle Perc.	0.972	0.82	1	0.92	0.315	1
	IMSE						
	$\beta_0(u)$	0.0214	0.0243	0.0216	0.0398	0.0448	0.1929
	$\beta_1(u)$	0.0902	0.0930	0.0951	0.1166	0.1254	0.3392
	$\beta_2(u)$	0.0365	0.1018	0.0431	0.0764	0.2211	0.5859
	$\beta_3(u)$	0.0122	0.2405	0.0032	0.0753	0.6248	0.1775
	$\beta_4(u)$	0.0045	0.0405	0.0031	0.0183	0.1713	0.1100
	TIMSE	0.1648	0.5075	0.1661	0.3282	1.3000	0.4017
d=50	Oracle Perc.	0.945	0.635	1	0.8	0.012	1
	IMSE						
	$\beta_0(u)$	0.0221	0.0230	0.0219	0.0431	0.0612	0.0426
	$\beta_1(u)$	0.0878	0.0896	0.0927	0.1230	0.1477	0.1253
	$\beta_2(u)$	0.0404	0.0551	0.0428	0.1042	0.0969	0.0751
	$\beta_3(u)$	0.0478	0.0776	0.0027	0.1727	0.0771	0.0105
	$\beta_4(u)$	0.0101	0.0165	0.0029	0.0239	0.0608	0.0083
	TIMSE	0.2086	0.2966	0.1631	0.5146	2.4926	0.2619

$q_{aB} = (1-q_A)q_B - \delta$, and $q_{AB} = p_A p_B + \delta$. With the Hardy-Weinberg equilibrium assumption, the SNP genotype at locus A can be simulated assuming a multi-nomial distribution with frequencies p_A^2,$2p_A(1-p_A)$ and $(1-p_A)^2$ for genotypes (AA,Aa,aa) correspondingly. We can subsequently generate the SNP2 genotypes conditional on SNP1 can be simulated based on the conditional probability matrix in Cui et al. [39]. The non-zero coefficients of the model are the same as those in Example 1. The simulation with sample size 500 were performed 500 replicates.

Figure 3.2 show the selection ratio when d=10, under different combinations of MAF and error distributions. The height of bars is defined similarly as that in Example 1. When the random error is standard normal, our approach has higher proportions to choose true positive SNPs, especially those with no effect on G×E interactions, and lower proportions to choose

false positive SNPs. When MAF increases, both approaches lead to higher selection ratios for true positive SNPs and lower selection ratios for false positive SNPs. A similar pattern can be observed when the error distribution follows a $t(3)$ distribution. The performance of both approaches is better when the error is normal. An analogous conclusion can be reached in Figure 3.3 when d=50.

Table 3.2 presents the oracle proportions and estimation results for d=10. Under the standard normal error, we observe the superior performance of our approach over the adaptive LASSO based approach in terms of both oracle percentage and estimation precision. The estimation accuracy of our approach is pretty close to that of the true model. The accuracy of all the 3 methods improves as MAF increases from 0.1 to 0.5. The performance of our method under $t(3)$ error are still comparable to that under the standard normal error, while the ALASSO method did much worse for the $t(3)$ error. A similar pattern can be observed for the high dimensional case (d=50) in Table 3.3. Our approach is still powerful in high dimensional scenario, especially under the $N(0,1)$ random error, while the ALASSO approach barely select the correct model.

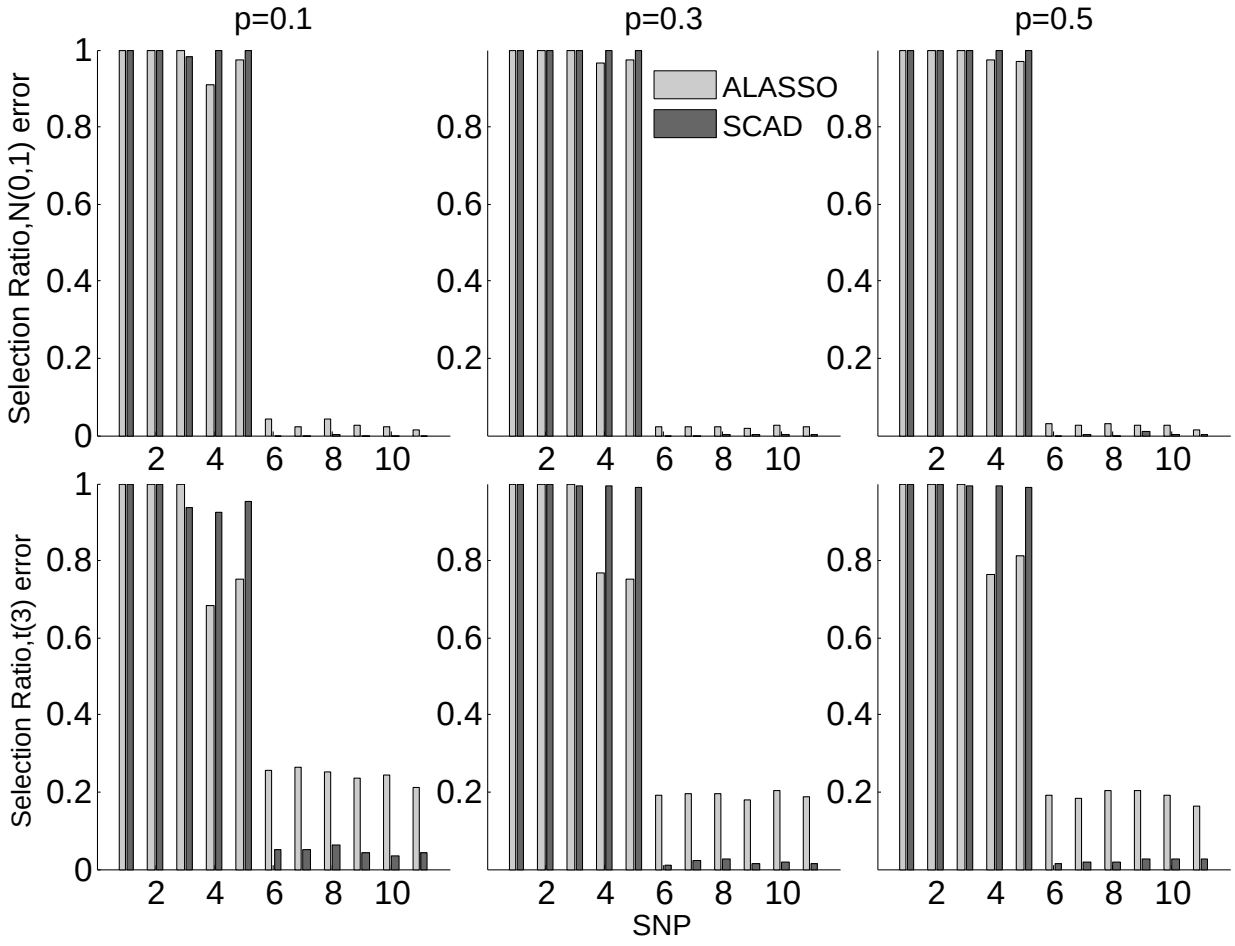

Figure 3.2: The selection ratio of Example 3.2, $d = 10$

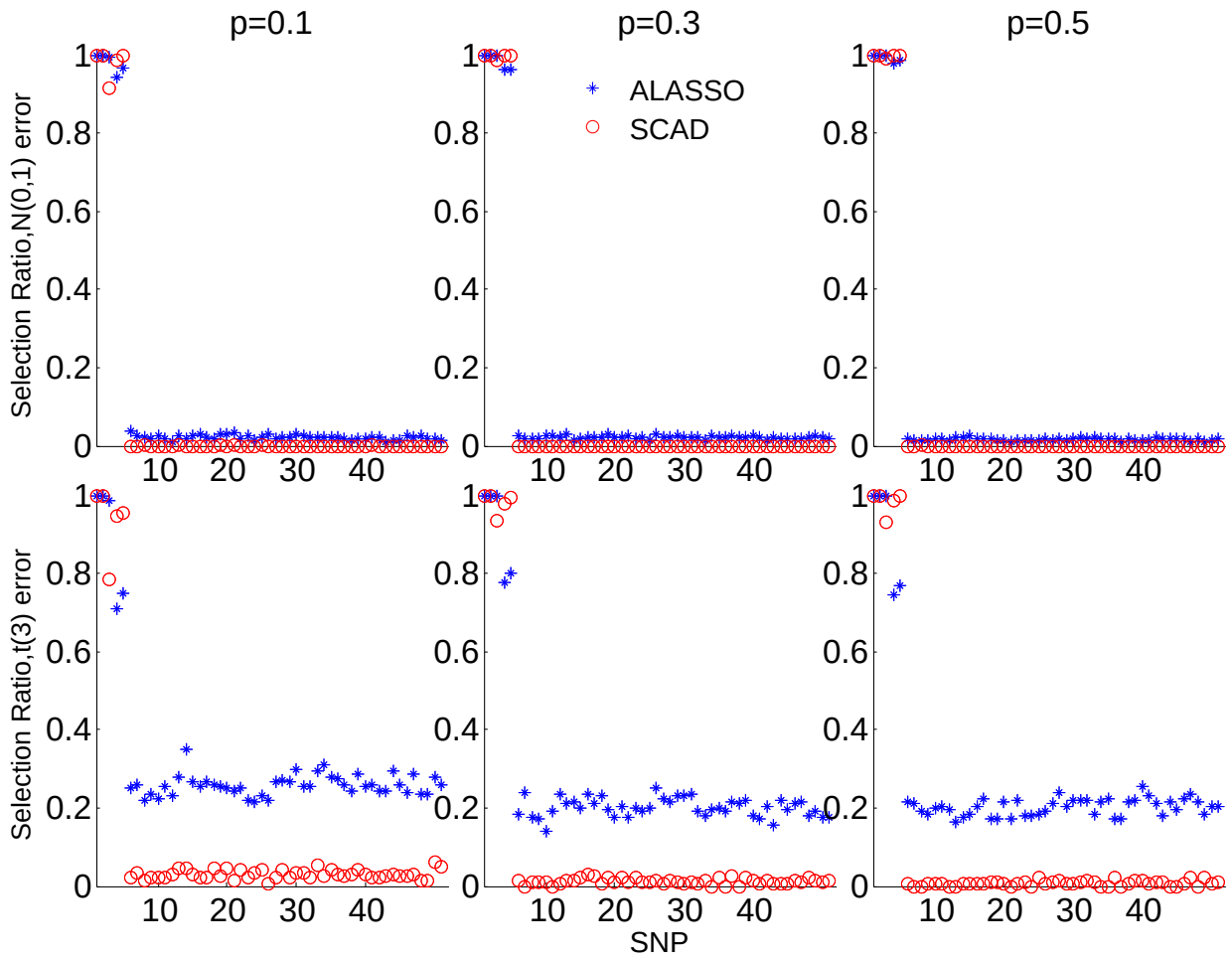

Figure 3.3: The selection ratio of Example 3.2, $d = 50$

Table 3.2: Simulation results of Example 3.2, $d = 10$

		$N(0,1)$ error			$t(3)$ error		
		SCAD	ALASSO	Oracle	SCAD	ALASSO	Oracle
pA=0.1	Oracle Perc.	0.976	0.784	1	0.72	0.268	1
	IMSE						
	$\beta_0(u)$	0.0863	0.1250	0.0891	0.3078	1.6608	0.2247
	$\beta_1(u)$	0.1611	0.1601	0.1667	0.3285	0.3947	0.3557
	$\beta_2(u)$	0.1264	0.1358	0.1238	0.4890	1.2776	0.2932
	$\beta_3(u)$	0.0270	0.1183	0.0192	1.3307	2.8155	0.0643
	$\beta_4(u)$	0.0191	0.0433	0.0174	0.2943	2.1633	0.0475
	TIMSE	0.4205	0.6106	0.4162	2.9342	9.2044	0.9855
pA=0.3	Oracle Perc.	0.992	0.84	1	0.91	0.33	1
	IMSE						
	$\beta_0(u)$	0.0268	0.0297	0.0273	0.0607	0.0975	0.0601
	$\beta_1(u)$	0.1071	0.1074	0.1174	0.1600	0.2065	0.1746
	$\beta_2(u)$	0.0561	0.0551	0.0637	0.1360	0.1373	0.1320
	$\beta_3(u)$	0.0086	0.0271	0.0084	0.1111	0.1216	0.0237
	$\beta_4(u)$	0.0066	0.0118	0.0065	0.0443	0.1125	0.0222
	TIMSE	0.2007	0.2404	0.2233	0.5311	1.3069	0.4126
pA=0.5	Oracle Perc.	0.98	0.846	1	0.894	0.34	1
	IMSE						
	$\beta_0(u)$	0.0213	0.0214	0.0214	0.0431	0.0485	0.0451
	$\beta_1(u)$	0.1044	0.1043	0.1106	0.1581	0.1721	0.1725
	$\beta_2(u)$	0.0497	0.0507	0.0604	0.1101	0.3270	0.1170
	$\beta_3(u)$	0.0077	0.0210	0.0077	0.0439	0.7984	0.0192
	$\beta_4(u)$	0.0063	0.0103	0.0063	0.0240	0.3082	0.0135
	TIMSE	0.1895	0.2177	0.2065	0.4072	1.8768	0.3673

Table 3.3: Simulation results of Example 3.2, $d = 50$

		$N(0,1)$ error			$t(3)$ error		
		SCAD	ALASSO	Oracle	SCAD	ALASSO	Oracle
pA=0.1	Oracle Perc.	0.908	0.542	1	0.435	0.025	1
	IMSE						
	$\beta_0(u)$	0.1929	0.9911	0.0884	0.5687	1.2335	0.2209
	$\beta_1(u)$	0.2064	0.1988	0.1684	0.3851	0.3484	0.3340
	$\beta_2(u)$	0.5235	0.8382	0.1218	0.6934	0.4432	0.2614
	$\beta_3(u)$	2.0918	2.0345	0.0196	2.4522	0.7892	0.0484
	$\beta_4(u)$	0.3475	0.4798	0.0158	0.5996	0.4671	0.0445
	TIMSE	3.3644	4.7239	0.4140	5.7021	8.9145	0.9092
pA=0.3	Oracle Perc.	0.986	0.642	1	0.745	0.06	1
	IMSE						
	$\beta_0(u)$	0.0289	0.0732	0.0278	0.0860	0.1970	0.0599
	$\beta_1(u)$	0.1107	0.1124	0.1137	0.1858	0.1974	0.1742
	$\beta_2(u)$	0.0817	0.1834	0.0646	0.2205	0.1768	0.1301
	$\beta_3(u)$	0.1083	0.4072	0.0075	0.3865	0.2018	0.0254
	$\beta_4(u)$	0.0229	0.0748	0.0068	0.0840	0.1099	0.0220
	TIMSE	0.3526	0.9334	0.2204	1.2288	3.3013	0.4117
pA=0.5	Oracle Perc.	0.988	0.706	1	0.8	0.07	1
	IMSE						
	$\beta_0(u)$	0.0215	0.0232	0.0216	0.0450	0.0560	0.0434
	$\beta_1(u)$	0.1048	0.1073	0.1123	0.1551	0.1716	0.1608
	$\beta_2(u)$	0.0608	0.1269	0.0579	0.1754	0.1525	0.1085
	$\beta_3(u)$	0.0470	0.2846	0.0078	0.1681	0.1501	0.0167
	$\beta_4(u)$	0.0120	0.0444	0.0053	0.0480	0.0889	0.0190
	TIMSE	0.2461	0.6426	0.2050	0.6492	2.8755	0.3484

3.5 Real data analysis

We applied the method to a real dataset from a study conducted at Department of Obstetrics and Gynecology at Sotero del Rio Hospital in Puente Alto, Chile. The initial objective of the study was to pinpoint genetic variants associated with a binary response indicating large for gestational age (LGA) or small for gestational age (SGA) depending on new born babies' weight and mother's gestational age. After data cleaning by removing SNPs with MAF less than 0.05 or deviation from Hardy-Weinberg equilibrium, the dataset contains 1536 new born babies with 189 genes covered by 660 single nucleotide polymorphisms (SNPs).

Mother's body mass index (MBMI), defined as mother's body mass (kg) divided by the square of their height (m^2), is a measure for mothers' body shape and obesity condition. The environment factor for a baby inside mother's body is defined through the mother, such as mother's obesity condition (MBMI) or age. Due to the complicated interaction between fetus's genes and mother's obesity level, the birth weight might be different for a fetus with the same gene but under different environment conditions. The phenomenon of regular variation in birth weight could be explained by corresponding genetic variants and how they respond to different MBMI.

We applied both methods to the Janus kinase/signal transducers and activators of transcription (JAK/STAT) signaling pathway, which has 68 SNPs covering 24 genes in our real data. JAK/STAT signaling pathway is the main signaling mechanism for a broad range of cytokines and growth factors in mammals [80]. Our method select the model

$$Y = \beta_0(z) + \beta_k X_k + \varepsilon$$

where X_k corresponds to SNP 2069762, and β_k is a constant, while the ALASOO method

only identifies the varying intercept. SNP 2069762 is of non-zero constant effect, therefore this one is a genetic risk factor associated with birthright but not sensitive to MBMI to influence birth weight.

To further validate our result, we conducted the single SNP based analysis in Ma et al [37] and tabulate SNPs with p-value less than 0.001 when fitting the candidate models (LM, LMI and VC). The p-values for the overall genetic association test with the LM, LMI and VC model are denoted as P_CON, P_LIN and P_VC. It follows from the test on constant coefficient that SNP 2069885 does not vary across MBMI (PP_CON¡ 0.05). Consequently, the p-value obtained from test with LM model is less than those obtained from VC and LM model.

Table 3.4: List of SNPs with p-value < 0.001 from the Jak-STAT signaling pathway

SNP ID	GeneName	Location	P_VC	P_CON	P_LIN	P_LM	P_LMI	P_I
			LM model					
2069885	IL9	exon	0.0014	0.0913	-	**7.32E-05**	8.93E-05	0.0875

3.6 Discussion

The significance of G×E interactions in complex traits has stimulated waves of discussion. A diversity of statistical models have been proposed to assess the gene effect under different environmental exposures, as reviewed in Cornelis et al [65]. The success of genetic variant set based association analysis, as shown in Wang et al [38], Cui et al [39], Wu and Cui [40] and Schaid et al [71], motivates us to propose a high dimensional variable selection approach to understand the mechanism of G×E interactions associated with complex traits. We adopt a penalized regression method within the VC model framework to investigate how

multiple variants within a genetic system, like the pathway, were mediated by a common environmental factor to influence the phenotypic response.

Variable selection and parameter estimation can be achieved simultaneously within the framework. The varying coefficient function are divided into 2 parts after B spline basis expansion, for the non-zero constant and varying effect. We can determine if a particular genetic variant is sensitive to environmental stimuli by examining the status of the coefficient function. Specifically, the presence of G×E interactions, no G×E interactions and no association with the phenotype are corresponding to varying, non zero constant and zero effects of the coefficients. A two-stage iterative procedure was developed in Tang et al [73] to distinguish different effects. Here, we adopted the framework and carried out the separation with SCAD penalty. Asymptotic properties of the two-stage estimator were established under suitable regularity conditions.

A comprehensive comparison between our method and that in Tang et al [73] was conducted in the simulation session, in terms of two criteria, the percentage of choosing the exact true model and the precision in parameter estimation. The estimation accuracy was calculated as IMSE for varying coefficients and MSE for constant coefficients. In Tang et al [73], for the predictors with non-zero constant coefficients, MSE was calculated when the predictor is corrected identified. However, this won't reveal the error caused by failure to classify the coefficient as non-zero constant. Instead, we suggest calculating IMSE for all the predictors, since IMSE reduces to MSE when the coefficient is a constant. A much more accurate assessment on assessing the performance of the model can thus be achieved. The advantage of our approach has been endorsed by the extensive simulation study and real data analysis.

Both algorithms are based on local quadratic approximations (LQA) to the penalty func-

tions, which suffers from the efficiency loss caused by repeated factorizations of large matrices. LQA limits the power of the framework to dissect G×E interactions when the dimension is large, especially in cases where $p > n$. We will integrate group coordinate descent (GCD) approach into the current framework and demonstrate the merit of the new scheme in next chapter.

3.7 Technical proofs

3.7.1 Useful notations and lemmas

For convenience, the following notations are adopted :

$\bar{Y} = E(Y|\boldsymbol{X}, T)$, $\bar{\boldsymbol{\gamma}} = (\boldsymbol{U}^T\boldsymbol{U})^{-1}\boldsymbol{U}^T\bar{Y}$, $\bar{\boldsymbol{\beta}} = \boldsymbol{B}\bar{\boldsymbol{\gamma}}$

$\boldsymbol{\gamma}_{(v)} = (\boldsymbol{\gamma}_0^T, \ldots, \boldsymbol{\gamma}_v^T)^T$, $\boldsymbol{\gamma}_{(c)} = (\boldsymbol{\gamma}_{v+1}^T, \ldots, \boldsymbol{\gamma}_c^T)^T$, $\boldsymbol{\gamma}_{(d)} = (\gamma_{v+1,1}^T, \ldots, \gamma_{d,1}^T)^T$,

$\tilde{\boldsymbol{\gamma}}_{(v)} = (\tilde{\boldsymbol{\gamma}}_0^T, \ldots, \tilde{\boldsymbol{\gamma}}_v^T)^T$, $\tilde{\boldsymbol{\gamma}}_{(c)} = (\tilde{\boldsymbol{\gamma}}_{v+1}^T, \ldots, \tilde{\boldsymbol{\gamma}}_c^T)^T$, $\tilde{\boldsymbol{\gamma}}_{(d)} = (\gamma_{v+1,1}, \ldots, \gamma_{d,1})^T$,

$\boldsymbol{G}_n = (B(z_1), \ldots, B(z_n))(B(z_1), \ldots, B(z_n)^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$

$\Phi_n = n^{-1}\sum_{i=1}^n \boldsymbol{U}_{(v)i}\boldsymbol{U}_{(v)i}^T$, $\Psi_n = n^{-1}\sum_{i=1}^n \boldsymbol{U}_{(v)i}\boldsymbol{U}_{(c)i}^T$, $\Lambda_i = \boldsymbol{U}_{(c)i} - \Psi_n^T\Phi_n^{-1}\boldsymbol{U}_{(c)i}$

First we provide several lemmas to facilitate the proofs of Theorems 1 and 2.

Lemma A.1. Under assumptions (A1-A3), there exists finite positive constants C_1 and C_2 such that all the eigenvalues of $(k_n/n)\boldsymbol{G}_n$ fall between C_1 and C_2, and therefore, $\boldsymbol{G}_n$ is invertible.

Lemma A.2. Under assumptions (A1-A3), for some finite constant C_0, there exists $\tilde{\boldsymbol{\gamma}} = (\tilde{\boldsymbol{\gamma}}_0^T, \ldots, \tilde{\boldsymbol{\gamma}}_d^T)^T$ satisfying

(1) $\|\tilde{\gamma}_{j*}\|_{L_2} > C_0$, $j = 0, \ldots, v$; $\tilde{\gamma}_{j1} = \beta_j$, $\|\tilde{\boldsymbol{\gamma}}_{j*}\|_{L_2} = 0$, $j = v+1, \ldots, c$; $\tilde{\gamma}_j = \mathbf{0}$, $j = c+1, \ldots, d$

(2) $\sup_{t\in[0,1]}|\beta_j(z) - B(z)^T\tilde{\boldsymbol{\gamma}}_j| = O_p(k_n^{-r})$, $j = 0, \ldots, d$, where $\tilde{\boldsymbol{\gamma}}_j = (\tilde{\gamma}_{j,1}, \tilde{\boldsymbol{\gamma}}_{j*}^T)^T$

(3) $\sup_{(t,x)\in[0,1]\times R^{d+1}} |\boldsymbol{X}^T\beta(z) - \boldsymbol{U}(\boldsymbol{X})'\tilde{\boldsymbol{\gamma}}| = O_p(k_n^{-r})$

3.7.2 Proofs of Theorem 1.

(A) Proof of Theorem 1(1) (Part 1)

Here we first show $\hat{\beta}_j(z)$ is constant for $j = v+1, \ldots, d$ in probability, which amounts to demonstrating $\|\hat{\boldsymbol{\gamma}}_{j*}^{vc}\|_j = \mathbf{0}$, $j = v+1, \ldots, d$ with probability tending to 1, as $n \to \infty$. For

$$Q_1(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \left(Y_i - \boldsymbol{U}_i^T\boldsymbol{\gamma}\right)^2 + n\sum_{j=1}^{d} p_{\lambda_1}(\|\gamma_{j*}\|) \tag{3.11}$$

Let $\alpha_n = n^{-\frac{1}{2}}k_n + a_n$ and $\hat{\boldsymbol{\gamma}}^{vc} = \tilde{\boldsymbol{\gamma}} + \alpha_n\boldsymbol{\delta}$. We want to show that for any given $\varepsilon > 0$, there exists a large constant C such that

$$P\left\{\inf_{\|\boldsymbol{\delta}\|=C} Q_1(\hat{\boldsymbol{\gamma}}^{vc}) \geq Q_1(\tilde{\boldsymbol{\gamma}})\right\} \geq 1 - \varepsilon \tag{3.12}$$

This suggests that with probability at least $1-\varepsilon$ there exists a local minimum in the ball $\{\tilde{\boldsymbol{\gamma}} + \alpha_n\boldsymbol{\delta} : \|\boldsymbol{\delta}\| \leq C\}$. Hence, there exists a local minimizer such that $\|\hat{\boldsymbol{\gamma}}^{vc} - \tilde{\boldsymbol{\gamma}}\| = O_p(\alpha_n)$. A direct computation yields

$$\begin{aligned} D_n(\boldsymbol{\delta}) &= Q_1(\hat{\boldsymbol{\gamma}}^{vc}) - Q_1(\tilde{\boldsymbol{\gamma}}) \\ &= -2\alpha_n\boldsymbol{\delta}\sum_{i=1}^{n}\left[\varepsilon_i + X_1^T r(z_i)\right]\boldsymbol{U}_i^T + \alpha_n^2\boldsymbol{\delta}^2\sum_{i=1}^{n}\boldsymbol{U}_i^T\boldsymbol{U}_i \\ &\quad + n\sum_{j=1}^{d}\left[p_{\lambda_1}(\|\hat{\gamma}_{j*}^{vc}\|) - p_{\lambda_1}(\|\tilde{\gamma}_{j*}\|)\right] \\ &\equiv \Delta_1 + \Delta_2 + \Delta_3 \end{aligned} \tag{3.13}$$

where $r_j(z) = B(z)^T\tilde{\gamma}_j - \beta_j(z)$, $j = 1, \ldots, d$ and $r(z) = (r_1(z), \ldots, r_d(z))^T$. By the fact $E(\varepsilon_i|\boldsymbol{U}, z_i) = 0$, we obtain that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\boldsymbol{U}_i^T\boldsymbol{\delta} = O_p(\|\boldsymbol{\delta}\|) \tag{3.14}$$

Recall Lemma A.1, then

$$\frac{1}{n}\sum_{i=1}^{n}X_i^T r(z_i)\boldsymbol{U}\boldsymbol{\delta} = O_p(k_n^{-r}\|\boldsymbol{\delta}\|) \tag{3.15}$$

Therefore

$$\Delta_1 = O_p(\sqrt{n}\alpha_n\|\boldsymbol{\delta}\|) + O_p(nk_n^{-r}\alpha_n\|\boldsymbol{\delta}\|) = O_p(nk_n^{-r}\alpha_n)\|\boldsymbol{\delta}\|$$

We can also show that $\Delta_2 = O_p(n\alpha_n^2)\|\boldsymbol{\delta}\|^2$. Then, by choosing a sufficiently large C, Δ_1 is dominated by Δ_2 uniformly in $\|\boldsymbol{\delta}\| = C$. It follows from Taylor expansion that

$$\begin{aligned}\Delta_3 &\leq n\sum_{j=1}^{d}\left[\alpha_n p'_{\lambda 1}(\|\tilde{\gamma}_{j*}\|)\frac{\tilde{\gamma}_{j*}}{\|\tilde{\gamma}_{j*}\|}\|\boldsymbol{\delta}_{j*}\| + \alpha_n^2 p''_{\lambda 2}(\|\tilde{\gamma}_{j*}\|)\|\boldsymbol{\delta}_{j*}\|^2(1 + o_p(1))\right] \\ &\leq n\sqrt{d}\alpha_n f_n\|\boldsymbol{\delta}\| + b_n\alpha_n^2\|\boldsymbol{\delta}\|^2\end{aligned}$$

where $f_n = \max_j\{|\tilde{\gamma}_{j*}| : \tilde{\gamma}_{j*} \neq 0\}$. With assumption (A6), we can prove that Δ_2 dominates Δ_3 uniformly in $\|\boldsymbol{\delta}\| = C$. Therefore, (3.12) holds for sufficiently large C, and we have $\|\hat{\boldsymbol{\gamma}}^{vc} - \tilde{\boldsymbol{\gamma}}\| = O_p(\alpha_n)$.

In order to prove $\hat{\beta}_j(z) = 0$ for $j = v+1, \ldots, d$ in probability, it is sufficient to demonstrate that $\hat{\boldsymbol{\gamma}}_{j*}^{vc} = \mathbf{0}$,$j = v+1, \ldots, d$. It follows from the definition that when $\max(\lambda_1, \lambda_2) \to 0$, $a_n = 0$ for large n. Then we need to show that with probability approaching 1 as $n \to \infty$,

for any $\hat{\boldsymbol{\gamma}}^{vc}$ satisfying $\|\hat{\boldsymbol{\gamma}}^{vc} - \tilde{\boldsymbol{\gamma}}\| = O_p(n^{-\frac{1}{2}}k_n)$ and some small $\varepsilon_n = Cn^{-\frac{1}{2}}k_n$, we have

$$\begin{aligned} \frac{\partial Q_1(\boldsymbol{\gamma})}{\partial \gamma_{j,*}} &< 0, \quad \text{for} \quad -\varepsilon_n < \gamma_{j,*} < 0, \quad j = v+1, \ldots, d \\ &> 0, \quad \text{for} \quad 0 < \gamma_{j,*} < \varepsilon_n, \qquad j = v+1, \ldots, d \end{aligned} \tag{3.16}$$

where $\gamma_{j,*}$ denotes the individual component of γ_{j*}. It's not hard to show that

$$\begin{aligned} \frac{\partial Q_1(\hat{\boldsymbol{\gamma}}^{vc})}{\partial \hat{\gamma}^{vc}_{j,*}} &= -2\sum_{i=1}^{n} \boldsymbol{U}_{ij}\left[Y_i - \boldsymbol{U}_i^T\hat{\boldsymbol{\gamma}}^{vc}\right] + np'_{\lambda_1}(|\hat{\gamma}_{j,*}|)\mathrm{sgn}(\hat{\gamma}_{j,*}) \\ &= -2\sum_{i=1}^{n} \boldsymbol{U}_{ij}[\varepsilon_i + \boldsymbol{X}_i^T r(z_i)] - 2\sum_{i=1}^{n} \boldsymbol{U}_{ij}\boldsymbol{U}_i^T[\tilde{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}^{vc}] \\ &\quad + np'_{\lambda_1}(|\hat{\gamma}_{j,*}|)\mathrm{sgn}(\hat{\gamma}^{vc}_{j,*}) \\ &= n\lambda_1\left[O_p(\lambda_1^{-1}n^{\frac{-r+1/2}{2r+1}}) + \lambda_1^{-1}p'_{\lambda}(|\hat{\gamma}_{j,*}|)\mathrm{sgn}(\hat{\gamma}^{vc}_{j,*})\right] \end{aligned} \tag{3.17}$$

By assumption (A5), $\lambda_1^{-1}n^{\frac{-r+1/2}{2r+1}} \to 0$. Then it follows from assumption (A7) that the sign of the derivative is completely determined by that of $\hat{\gamma}^{vc}_{j,*}$. Therefore, $\hat{\boldsymbol{\gamma}}^{vc}$, the minimizer of Q_1, is achieved at $\hat{\boldsymbol{\gamma}}^{vc}_{j*} = \mathbf{0}$, $j = v+1, \ldots, d$. This completes the proof of Theorem 1(1), part 1. □

(B) Proof of Theorem 1 (2)

Next we establish the consistency of the varying coefficient estimator. Let $\alpha_n = n^{-\frac{1}{2}}k_n + a_n$, $\hat{\boldsymbol{\gamma}}_{(v)} = \tilde{\boldsymbol{\gamma}}_{(v)} + \alpha_n\boldsymbol{\delta}_v$, $\hat{\boldsymbol{\gamma}}_{(d)} = \tilde{\boldsymbol{\gamma}}_{(d)} + \alpha_n\boldsymbol{\delta}_d$, and $\boldsymbol{\delta} = (\boldsymbol{\delta}_v^T, \boldsymbol{\delta}_d^T)^T$

$$Q_2(\boldsymbol{\gamma}_{(v)}, \boldsymbol{\gamma}_{(d)}) = \sum_{i=1}^{n}\left(Y_i - \boldsymbol{U}_{(v)i}^T\boldsymbol{\gamma}_{(v)} - \boldsymbol{U}_{(d)i}^T\boldsymbol{\gamma}_{(d)}\right)^2 + n\sum_{j=v+1}^{d} p_{\lambda_2}(|\gamma_{j,1}|) \tag{3.18}$$

We need to show that for any given $\varepsilon > 0$, there exists a large constant C_ε such that

$$P\left\{\inf_{\|\boldsymbol{\delta}\|=C} Q_2(\hat{\boldsymbol{\gamma}}_{(v)}, \hat{\boldsymbol{\gamma}}_{(d)}) \geq Q_2(\tilde{\boldsymbol{\gamma}}_{(v)}, \tilde{\boldsymbol{\gamma}}_{(d)})\right\} \geq 1 - \varepsilon \tag{3.19}$$

which implies that with probability at least $1 - \varepsilon$ there exists a local minimum in the ball $\{\tilde{\boldsymbol{\gamma}}_{(v)} + \alpha_n \boldsymbol{\delta}_v : \|\boldsymbol{\delta}_v\| \leq C\}$ and $\{\tilde{\boldsymbol{\gamma}}_{(d)} + \alpha_n \boldsymbol{\delta}_d : \|\boldsymbol{\delta}_d\| \leq C\}$ respectively. Therefore, there exists local minimizers such that $\|\hat{\boldsymbol{\gamma}}_{(v)} - \tilde{\boldsymbol{\gamma}}_{(v)}\| = O_p(\alpha_n)$ and $\|\hat{\boldsymbol{\gamma}}_{(d)} - \tilde{\boldsymbol{\gamma}}_{(d)}\| = O_p(\alpha_n)$. We have

$$\begin{aligned}
D_n(\boldsymbol{\delta}_v, \boldsymbol{\delta}_d) &= Q_2(\hat{\boldsymbol{\gamma}}_{(v)}, \hat{\boldsymbol{\gamma}}_{(d)}) - Q_2(\tilde{\boldsymbol{\gamma}}_{(v)}, \tilde{\boldsymbol{\gamma}}_{(d)}) \\
&= -2\alpha_n \sum_{i=1}^{n} \left[\varepsilon_i + X_1^T R(Z_i)\right] \left[\boldsymbol{U}_{(v)i}^T \boldsymbol{\delta}_{(v)} + \boldsymbol{U}_{(d)i}^T \boldsymbol{\delta}_{(d)}\right] \\
&\quad + \alpha_n^2 \sum_{i=1}^{n} \left[\boldsymbol{U}_{(v)i}^T \boldsymbol{\delta}_{(v)} + \boldsymbol{U}_{(d)i}^T \boldsymbol{\delta}_{(d)}\right]^2 + n \sum_{j=v+1}^{d} \left[p_{\lambda_2}(|\hat{\gamma}_{j,1}|) - p_{\lambda_2}(|\tilde{\gamma}_{j,1}|)\right] \\
&\equiv \Delta_1 + \Delta_2 + \Delta_3
\end{aligned} \tag{3.20}$$

where $r(z) = (r_1(z), \ldots, r_d(z))^T$ and $r_j(z) = B(z)^T \tilde{\gamma}_j - \beta_j(z)$, $j = 1, \ldots, d$.

Since $E(\varepsilon_i | \boldsymbol{U}_{(v)}, \boldsymbol{U}_{(d)}, z_i) = 0$, we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i [\boldsymbol{U}_{(v)i}^T \boldsymbol{\delta}_{(v)} + \boldsymbol{U}_{(d)i}^T \boldsymbol{\delta}_{(d)}] = O_p(\|\boldsymbol{\delta}\|) \tag{3.21}$$

With Lemma A.1 we can show

$$\frac{1}{n} \sum_{i=1}^{n} X_i^T r(z_i) \left[\boldsymbol{U}_{(v)i}^T \boldsymbol{\delta}_{(v)} + \boldsymbol{U}_{(d)i}^T \boldsymbol{\delta}_{(d)}\right] = O_p\left(k_n^{-r} \|\boldsymbol{\delta}\|\right) \tag{3.22}$$

Combine the above two equations, we can obtain that

$$\Delta_1 = O_p(n^{\frac{1}{2}}\alpha_n\|\boldsymbol{\delta}\|) + O_p(nk_n^{-r}\alpha_n\|\boldsymbol{\delta}\|) = O_p(nk_n^{-r}\alpha_n)\|\boldsymbol{\delta}\|$$

Since $\Delta_2 = O_p(n\alpha_n^2)\|\boldsymbol{\delta}\|^2$, it's easy to show that by choosing a sufficiently large C, Δ_1 is dominated by Δ_2 uniformly in $\|\boldsymbol{\delta}\| = C$. By Taylor expansion,

$$\begin{aligned}\Delta_3 &\leq n\sum_{j=v+1}^{d}\left[\alpha_n p'_{\lambda 2}(|\tilde{\gamma}_{j,1}|)\mathrm{sgn}(\tilde{\gamma}_{j,1})|\delta_{dj}| + \alpha_n^2 p''_{\lambda 2}(|\tilde{\gamma}_{j,1}|)\delta_{dj}^2(1+o(1))\right] \\ &\leq (p-v)^{\frac{1}{2}} n\alpha_n f_n\|\boldsymbol{\delta}\| + b_n\alpha_n^2\|\boldsymbol{\delta}\|^2\end{aligned}$$

where $f_n = \max_j\{|\tilde{\gamma}_{j,1}| : \tilde{\gamma}_{j,1} \neq 0\}$. Recall (A6), then it follows that, by choosing an enough large C, Δ_2 dominates Δ_1 uniformly in $\|\boldsymbol{\delta}\| = C$. Consequently (3.19) holds for sufficiently large C, and we have $\|\hat{\boldsymbol{\gamma}}_v - \tilde{\boldsymbol{\gamma}}_v\| = O_p(\alpha_n)$ and $\|\hat{\boldsymbol{\gamma}}_d - \tilde{\boldsymbol{\gamma}}_d\| = O_p(\alpha_n)$. By the definition of $\boldsymbol{\gamma}^{cz}$, we have $\hat{\boldsymbol{\gamma}}^{cz}_{(d)} - \tilde{\boldsymbol{\gamma}}_{(d)} = O_p(\alpha_n)$. Then for $j = 0, \ldots, d$

$$\begin{aligned}\|\hat{\beta}_j(z_i) - \beta_j(z)\|^2 &= \int_0^1 \left[\hat{\beta}_j(z) - \beta_j(z)\right]^2 \mathrm{d}t \\ &\leq \int_0^1 \left[\boldsymbol{B}(z)^T\hat{\boldsymbol{\gamma}}_j^{cz}(z) - \boldsymbol{B}(z)^T\tilde{\boldsymbol{\gamma}}_j + r_j(z)\right]^2 \mathrm{d}t \\ &= \frac{2}{n}(\hat{\boldsymbol{\gamma}}_j^{cz} - \tilde{\boldsymbol{\gamma}}_j)^T\boldsymbol{G}_n(\hat{\boldsymbol{\gamma}}_j^{cz} - \tilde{\boldsymbol{\gamma}}_j) + 2\int_0^1 r_j(z)^2\mathrm{d}t \\ &= \Delta_1 + \Delta_2\end{aligned}$$

Recall Lemma A.1, A.2 and $k_n = O_p\left(n^{\frac{1}{2r+1}}\right)$, we can demonstrate that $\Delta_1 = O_p\left(k_n^{-1}\alpha_n^2\right)$, $\Delta_2 = O_p\left(k_n^{-2r}\right)$. Δ_1 is dominated by Δ_2, thus we finish the proof of Theorem 1(b). □

(C) Proof of Theorem 1(1) (Part 2)

To show $\hat{\beta}_j(z) = 0$ for $j = c+1, \ldots, d$, it is sufficient to demonstrate that $\hat{\boldsymbol{\gamma}}_{j,1}^{cz} = 0$, since the constancy of $\beta_j(z)$, $j = v+1, \ldots, d$ was already established in (A). It follows from the definition that when $\max(\lambda_1, \lambda_2) \to 0$, $a_n = 0$ for large n. Then we need to prove that with probability approaching 1 as $n \to \infty$, for any $\hat{\boldsymbol{\gamma}}_{(v)}$ and $\hat{\boldsymbol{\gamma}}_{(d)}$ satisfying $\|\hat{\boldsymbol{\gamma}}_{(v)} - \tilde{\boldsymbol{\gamma}}_{(v)}\| = O_p(n^{-\frac{1}{2}} k_n)$, $\|\hat{\boldsymbol{\gamma}}_{(d)} - \tilde{\boldsymbol{\gamma}}_{(d)}\| = O_p(n^{-\frac{1}{2}} k_n)$ respectively, as well as some small $\varepsilon_n = C n^{-\frac{1}{2}} k_n$, we have

$$\begin{aligned} \frac{\partial Q_2(\boldsymbol{\gamma}_{(v)}, \boldsymbol{\gamma}_{(d)})}{\partial \gamma_{j,1}} &< 0, \quad \text{for} \quad -\varepsilon_n < \gamma_{j,1} < 0, \quad j = c+1, \ldots, d \\ &> 0, \quad \text{for} \quad 0 < \gamma_{j,1} < \varepsilon_n, \quad j = c+1, \ldots, d \end{aligned} \tag{3.23}$$

We can prove that

$$\begin{aligned} \frac{\partial Q_2(\hat{\boldsymbol{\gamma}}_{(v)}, \hat{\boldsymbol{\gamma}}_{(d)})}{\partial \hat{\gamma}_{j,1}} &= -2 \sum_{i=1}^{n} \boldsymbol{U}_{(d)ij} \left[Y_i - \boldsymbol{U}_{(v)i}^T \hat{\boldsymbol{\gamma}}_{(v)} - \boldsymbol{U}_{(d)i}^T \hat{\boldsymbol{\gamma}}_{(d)} \right] + n p_\lambda'(|\hat{\gamma}_{j,1}|) \mathrm{sgn}(\hat{\gamma}_{j,1}) \\ &= -2 \sum_{i=1}^{n} \boldsymbol{U}_{(d)ij} \left[\varepsilon_i + \boldsymbol{X}_i^T r(z_i) \right] - 2 \sum_{i=1}^{n} \boldsymbol{U}_{(d)ij} \boldsymbol{U}_{(v)i}^T \left[\tilde{\boldsymbol{\gamma}}_v - \hat{\boldsymbol{\gamma}}_v \right] \\ &\quad - 2 \sum_{i=1}^{n} \boldsymbol{U}_{(d)ij} \boldsymbol{U}_{(d)i}^T \left[\tilde{\boldsymbol{\gamma}}_d - \hat{\boldsymbol{\gamma}}_d \right] + n p_\lambda'(|\hat{\gamma}_{j,1}|) \mathrm{sgn}(\hat{\gamma}_{j,1}) \\ &= n\lambda_2 \left[O_p \left(\lambda_2^{-1} n^{\frac{-r+1/2}{2r+1}} \right) + \lambda_2^{-1} p_\lambda'(|\hat{\gamma}_{j,1}|) \mathrm{sgn}(\hat{\gamma}_{j,1}) \right] \end{aligned} \tag{3.24}$$

By assumption (A5), $\lambda_2^{-1} n^{\frac{-r+1/2}{2r+1}} \to 0$. Then it follows from assumption (A7) that the sign of the derivative is completely determined by that of $\hat{\gamma}_{j,1}$. Therefore, $\hat{\boldsymbol{\gamma}}^{cz}$, the minimizer of Q_2, is achieved at $\hat{\boldsymbol{\gamma}}_{j,1}^{cz} = 0$, $j = c+1, \ldots, d$. This completes the proof of Theorem 1(1). $\square$

3.7.3 Proofs of Theorem 2.

In Theorem 1, we showed that both $\hat{\boldsymbol{\gamma}}_{j*} = \mathbf{0}$, $j = v+1, \ldots, c$ and $\hat{\gamma}_j = 0$, $j = c+1, \ldots, d$, hold in probability. Then Q_2 reduces to

$$\begin{aligned} Q_2(\boldsymbol{\gamma}_{(v)}, \boldsymbol{\gamma}_{(d)}) &= \sum_{i=1}^{n} \left(Y_i - \boldsymbol{U}_{(v)i}^T \boldsymbol{\gamma}_{(v)} - \boldsymbol{U}_{(c)i}^T \boldsymbol{\gamma}_{(c)} \right)^2 + n \sum_{j=v+1}^{c} p_{\lambda_2}(|\gamma_{j,1}|) \\ &\equiv Q_2(\boldsymbol{\gamma}_{(v)}, \boldsymbol{\gamma}_{(c)}) \end{aligned} \tag{3.25}$$

Since $(\hat{\boldsymbol{\gamma}}_{(v)}, \hat{\boldsymbol{\gamma}}_{(c)})$ is the minimal value of $Q_2(\boldsymbol{\gamma}_{(v)}, \boldsymbol{\gamma}_{(c)})$, we obtain

$$\frac{\partial Q_2(\hat{\boldsymbol{\gamma}}_{(v)}, \hat{\boldsymbol{\gamma}}_{(c)})}{\partial \hat{\boldsymbol{\gamma}}_{(v)}} = -2 \sum_{i=1}^{n} \boldsymbol{U}_{(v)i} \left[Y_i - \boldsymbol{U}_{(v)i}^T \hat{\boldsymbol{\gamma}}_{(v)} - \boldsymbol{U}_{(d)i}^T \hat{\boldsymbol{\gamma}}_{(d)} \right] = 0 \tag{3.26}$$

$$\begin{aligned} \frac{\partial Q_2(\hat{\boldsymbol{\gamma}}_{(v)}, \hat{\boldsymbol{\gamma}}_{(c)})}{\partial \hat{\boldsymbol{\gamma}}_{(c)}} &= -2 \sum_{i=1}^{n} \boldsymbol{U}_{(c)i} \left[Y_i - \boldsymbol{U}_{(v)i}^T \hat{\boldsymbol{\gamma}}_{(v)} - \boldsymbol{U}_{(c)i}^T \hat{\boldsymbol{\gamma}}_{(c)} \right] \\ &\quad + n \sum_{j=v+1}^{c} p_{\lambda 2}^{'}(|\hat{\gamma}_{j,1}|) \mathrm{sgn}(\hat{\gamma}_{j,1}) = 0 \end{aligned} \tag{3.27}$$

By applying Taylor expansion on $p_{\lambda 2}^{'}(|\hat{\gamma}_{j,1}|)$ in (3.27), we have

$$p_{\lambda 2}^{'}(|\hat{\gamma}_{j,1}|) = p_{\lambda 2}^{'}(|\gamma_{j,1}|) + p_{\lambda 2}^{''}(|\gamma_{j,1}|)(\hat{\gamma}_{j,1} - \gamma_{j,1})[1 + o_p(1)]$$

By the fact that $p_{\lambda 2}^{'}(|\hat{\gamma}_{j,1}|) = 0$ as $\lambda_2 \to 0$, and $p_{\lambda 2}^{''}(|\gamma_{j,1}|) = o_p(1)$ from the assumption, it follows that $\sum_{j=v+1}^{c} p_{\lambda 2}^{'}(|\hat{\gamma}_{j,1}|)\mathrm{sgn}(\hat{\gamma}_{j,1}) = o_p(\hat{\gamma}_{j,1} - \gamma_{j,1}) = o_p(\hat{\boldsymbol{\gamma}}_{(c)} - \boldsymbol{\gamma}_{(c)})$. Consequently, we have

$$\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{U}_{(c)i} \left[Y_i - \boldsymbol{U}_{(v)i}^T \hat{\boldsymbol{\gamma}}_{(v)} - \boldsymbol{U}_{(c)i}^T \hat{\boldsymbol{\gamma}}_{(c)} \right] + o_p(\hat{\boldsymbol{\gamma}}_{(c)} - \boldsymbol{\gamma}_{(c)}) = 0$$

Following similar lines of arguments in Theorem 1, we can show

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{U}_{(c)i}\left[\varepsilon_i + X_i^T r(z_i) + \boldsymbol{U}_{(v)i}^T(\boldsymbol{\gamma}_{(v)} - \hat{\boldsymbol{\gamma}}_{(v)}) + \boldsymbol{U}_{(c)i}^T(\boldsymbol{\gamma}_{(c)} - \hat{\boldsymbol{\gamma}}_{(c)})\right] + o_p(\hat{\boldsymbol{\gamma}}_{(c)} - \boldsymbol{\gamma}_{(c)}) = 0 \tag{3.28}$$

Meanwhile, a straightforward calculation yields

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{U}_{(v)i}\left[\varepsilon_i + \boldsymbol{X}_i^T r(u_i) + \boldsymbol{U}_{(v)i}^T(\boldsymbol{\gamma}_{(v)} - \hat{\boldsymbol{\gamma}}_{(v)}) + \boldsymbol{U}_{(c)i}^T(\boldsymbol{\gamma}_{(c)} - \hat{\boldsymbol{\gamma}}_{(c)})\right] = 0 \tag{3.29}$$

Recall the definition of Φ_n and Ψ_n, (3.29) is equivalent to

$$\hat{\boldsymbol{\gamma}}_{(v)} - \boldsymbol{\gamma}_{(v)} = \Phi_n^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{U}_{(v)i}\left[\varepsilon_i + \boldsymbol{X}_i^T r(z_i)\right] + \Psi_n[\boldsymbol{\gamma}_{(c)} - \hat{\boldsymbol{\gamma}}_{(c)}]\right\} \tag{3.30}$$

Plugging (3.30) into (3.28) results in

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{U}_{(c)i}\left\{\varepsilon_i + X_i^T r(z_i) - \boldsymbol{U}_{(v)i}^T\Phi_n^{-1}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{U}_{(v)i}\left[\varepsilon_i + \boldsymbol{X}_i^T r(z_i)\right]\right\}\\ &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{U}_{(c)i}\left[\boldsymbol{U}_{(c)i} - \Psi_n^T\Phi_n^{-1}\boldsymbol{U}_{(v)i}\right]^T(\hat{\boldsymbol{\gamma}}_{(c)} - \boldsymbol{\gamma}_{(c)}) + o_p(\hat{\boldsymbol{\gamma}}_{(c)} - \boldsymbol{\gamma}_{(c)})\end{aligned} \tag{3.31}$$

Together with the facts that

$$\frac{1}{n}\sum_{i=1}^{n}\Psi_n^T\Phi_n^{-1}\boldsymbol{U}_{(v)i}\left[\varepsilon_i + X_i^T r(z_i) - \boldsymbol{U}_{(v)i}^T\Phi_n^{-1}\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{U}_{(v)k}[\varepsilon_k + \boldsymbol{X}_k^T r(t_k)]\right] = 0$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\Psi_n^T\Phi_n^{-1}\boldsymbol{U}_{(v)i}\left[\boldsymbol{U}_{(c)i}^T - \Psi_n^T\Phi_n^{-1}\boldsymbol{U}_{(v)i}\right]^T = 0$$

and recall the definition of Λ_i, a direct computation from (3.31) leads to

$$\begin{aligned}\left[\frac{1}{n}\sum_{i=1}^{n}\Lambda_i\Lambda_i^T+o_p(1)\right]\sqrt{n}(\boldsymbol{\gamma}_{(c)}-\hat{\boldsymbol{\gamma}}_{(c)}) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Lambda_i\varepsilon_i+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Lambda_i\boldsymbol{X}_i^T r(z_i)\\ &\quad+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Lambda_i\boldsymbol{U}_{(v)i}^T\Phi_n^{-1}\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{U}_{(v)k}\left[\varepsilon_k+\boldsymbol{X}_k^T r(t_k)\right]\\ &= \Delta_1+\Delta_2+\Delta_3\end{aligned}$$

It follows from law of large numbers that

$$\frac{1}{n}\sum_{i=1}^{n}\Lambda_i\Lambda_i^T\xrightarrow{\ p\ }N(0,\Sigma)$$

where $\Sigma=E\left(\boldsymbol{U}_{(c)}\boldsymbol{U}_{(c)}^T\right)-E\left\{E(\Psi_n^T|T)E(\Phi_n|T)^{-1}E(\Psi_n|T)\right\}$. Consequently,

$$\Delta_2\xrightarrow{\ d\ }N(0,\sigma^2\Sigma)$$

follows from central limit theorem. Because $\boldsymbol{X}_i$ is bounded and $\|r(z)\|=o_p(1)$, we have $\Delta_2=o_p(1)$.Besides, $\sum_{i=1}^{n}\Lambda_i\boldsymbol{U}_{(v)i}^T$=0 implies that $\Delta_3=0$. Therefore, by Slutsky theorem, we complete the proof of Theorem 2. □

Chapter 4

A Group Coordinate Descent Approach For High Dimensional Variable Selection In Gene-Environment Interactions

4.1 Introduction

High dimensional data arises in a diversity of scientific areas, especially in the study of human genetics as tons of data covering the entire human genome are brought by the advancement of high-throughput genotyping technologies. Gene-environment (G×E) interaction draws our interest due to the crucial roles it plays in elucidating the etiology of complex disease and tracking down the disease variants. The risk of complex diseases is triggered not merely by genetic factors, but also by the environmental exposures, as well as their interactions.

The varying coefficient (VC) model framework, initially proposed in Hastie and Tibshirani [74], lends us significant flexibility in investigating genetic responses to environmental stimuli and how gene expressions are mediated by environmental influences to increase disease predispositions, for both continuous phenotype response in Ma et al [37] and binary

disease response in Wu and Cui [68]. The merit of the framework is especially prominent when the penetrance effect of genetic variants are non-linear, as the VC model is powerful to capture the dynamic fluctuations of regression coefficients.

Current methodologies on G×E interactions are mainly developed within single variant-based framework, using either parametric methods as reviewed in Mukherjee et al [81], semi-parametric methods as in N. Chatterjee et al [35] and Maity et al [36], non-parametric methods as in Ma et al [37] and Wu and Cui [68], or data mining approaches such as multifactor dimension reduction (MDR) in Hahn et al [33]. Accumulation of evidence showing the advantage of set based association analysis, such as in Neal and Sham [82], Cui et al [39], Wu and Cui [40] and Wang et al [38], motivated us to consider the joint modeling of a number of variants (p) within the genetic system given a common environment mediator. When the dimension p is large, the problem can be approached from a high dimensional variable selection perspective, within the VC model framework.

In recent years, much progress has been made on penalized regression methods for VC models. Wang et al [78] developed SCAD penalty based method for longitudinal response. Wang and Xia [83] considered variable selection for VC model via local constant kernel estimation with LASSO penalty, while the penalization method in Leng [84] was proposed in the framework of smoothing spline ANOVA models with component selection and smoothing operator (COSSO). The number of candidate predictors for selection in those models is finite and less than the sample size. A diversity of penalized group coordinate approaches have been developed for high dimensional case where the dimension of model p significantly exceeds the sample size n. See Yuan and Lin [85] for group LASSO approach, Wei et al [86] for group adaptive LASSO approach and Breheny and Huang [87] for group SCAD approach.

In this chapter, we develop a general framework, based on the group coordinate descent

(GCD) algorithm, to carry out variable selection for set-based G×E interactions in VC models. GCD was generalized to group case from coordinate-wise descent algorithms which are demonstrated to be effective in fitting penalized models such as in Friedman et al [88], Wu and Lange [89] and Friedman et al [90]. Our framework is implemented through a two-stage iterative procedure to separate the varying, non-zero constant and zero effect of the predictors. It is computationally efficient especially for high dimensional problems where $p > n$, since the computational complexity increases only linearly with the number of predictor groups.

The rest of the chapter is organized as follows. First, we describe the B spline basis expansion to the VC model. The computation algorithm via GCD with both convex and non-convex penalties will be proposed and the convergence of the algorithm will be discussed. Selection of the tuning parameters are examined subsequently. We demonstrate the utility of our approach through extensive simulation studies and real data analysis. Discussions and concluding remarks are given at the end of this chapter.

4.2 Statistical methods

Let $(\mathbf{X}_i, Y_i, Z_i)$, $i = 1, \ldots, n$ be random vectors which are independent and identically distributed (i.i.d.). Consider a varying coefficient (VC) model with p predictors

$$Y_i = \sum_{j=0}^{p} \beta_j(Z_i) X_{ij} + \varepsilon_i \tag{4.1}$$

where $\beta_j(\cdot)$ is the smooth varying-coefficient function, X_{ij} is the jth component of $(p+1)$-dimensional vector $\mathbf{X}_i$ with the first component X_{i0} being 1, Z_i's are the scalar index variable,

and ε_i is the model error.

The varying coefficient model helps us gain particular power to evaluate the nonlinear responses of genetic variants to environmental incentives in gene-environment (G×E) interactions [37, 68].We are interested in dissecting the penetrance of multiple genetic variants within a system, such as gene set or pathway, under various environmental stimuli, especially when those variants are mediated by a common mediator Z to affect phenotype. The effect of the variants can be determined by investigating the status of the coefficient function $\beta_j(\cdot)$ in (4.1). If $\beta_j(\cdot)$ is a varying function of the environmental factor Z, then the G×E interaction exists. The nonzero constancy of $\beta_j(\cdot)$ indicates that the G×E interaction is not present. The genetic variant is not associated with the response (phenotype) if $\beta_j(\cdot) = 0$. In such a set-based G×E interaction study design, the total number of variants in the system (p) can far exceed the sample size (n).

4.2.1 Basis expansion and penalized regression

Supposed that the coefficient function $\beta_j(z)$ $(j = 0, \ldots, p)$ in (4.1) can be approximated by basis expansion such that

$$\beta_j(Z) \approx \sum_{l=1}^{L} \gamma_{jl} B_{jl}(Z)$$

where L is the number of basis functions to approximate the coefficient function, $B(\cdot) = \{B_{jl}(\cdot)\}_{l=1}^{L}$ is a set of normalized B spline basis, and $\boldsymbol{\gamma}_j$ is the corresponding spline coefficient vector. It follows from change of basis [79] that the above basis expansion is equivalent to

$$\beta_j(\cdot) \approx \sum_{l=1}^{L} \gamma_{jl} B_{jl}(\cdot) \doteq \gamma_{j1} + \tilde{B}_j^T(\cdot)\gamma_{j*}$$

where the spline coefficient vector $\boldsymbol{\gamma}_j \doteq (\gamma_{j,1}, \boldsymbol{\gamma}_{j*}^T)^T$, and $\tilde{B}_j(\cdot) = (B_{k2}(\cdot), \ldots, B_{jL}(\cdot))^T$. $\gamma_{j,1}$ and $\boldsymbol{\gamma}_{j*}$ correspond to the constant and varying part of the coefficient functional respectively. Denote $\|\boldsymbol{\gamma}_j\|_j^2 = \boldsymbol{\gamma}_j^T \boldsymbol{R}_j \boldsymbol{\gamma}_j$, where $\boldsymbol{R}_j = \frac{1}{n}\sum_{i=1}^n \left[B_j(Z_i) X_{ij} X_{ij}^T B_j^T(Z_i)\right]$. Note that $\boldsymbol{R}_j$ is a $L \times L$ positive definite matrix.If $\|\boldsymbol{\gamma}_{j*}\|_j$=0, then the jth predictor only has a constant effect. Furthermore, if $\gamma_{j,1}$=0, then the predictor is not associated with the response. Therefore $\boldsymbol{\gamma}_{j*}$ can be treated as a group.

To separate the varying, constant and zero effect in the procedure of simultaneous variable selection and parameter estimation, we minimized the penalized loss function

$$\begin{aligned} Q(\boldsymbol{\gamma}) =& \frac{1}{n}\sum_{i=1}^n \left[Y_i - \sum_{j=0}^p \sum_{l=1}^L \gamma_{jl} X_{ij} B_{jl}(Z_i) \right]^2 + \sum_{j=1}^p p_{\lambda_1}(\|\boldsymbol{\gamma}_{j*}\|_j) \\ &+ \sum_{j=1}^p p_{\lambda_2}(|\gamma_{j1}|) I(\|\boldsymbol{\gamma}_{j*}\|_j = 0) \end{aligned} \tag{4.2}$$

where λ_1 and λ_2 are the penalization parameters. The penalty function p_{λ_k} (k=1,2) in (4.2) can be concave, such as in LASSO [91] or adaptive LASSO [92], or non-cave, such as the smoothly clipped absolute deviation (SCAD) penalty function [75]. Tang et al [73] first proposed the framework and established the asymptotic results for adaptive LASSO penalty function, while Wu et al [93] demonstrated improved finite sample performance of SCAD penalty over adaptive LASSO as well as the corresponding oracle property. Both approaches are based on local quadratic approximations (LQA) to the penalty function p_{λ_k} (k=1,2). However, LQA leads to a ridge-type solution dependent on repeated large matrix inversions, which renders the algorithm not efficient for large scale regression problems. Furthermore, as pointed out in Breheny and Huang [87], the quadratic approximation cannot benefit from the sparsity since this approach will not yield naturally sparse solutions.

Zou and Li [94] developed local linear approximation (LLA) to the non-convex penalty function and demonstrated its advantage over LQA. Breheny and Huang [95] further enhanced the LLA by showing penalized models with non-convex penalty can be fitted effectively by coordinate descent method. The main idea of GCD for (4.2) is to minimize the penalized loss function Q with respect to an individual predictor group after basis expansion at each step, and then cycle through all the predictor groups till convergence.

4.2.2 Computational algorithms

In this section, we extend the two-step iterative framework in [73, 93] with GCD approach. (4.2) can be rewritten using matrix notations as

$$\begin{aligned} Q(\boldsymbol{\gamma}) =& (\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma})^T(\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma}) + n\sum_{j=1}^{p} p_{\lambda 1}(\|\gamma_{j*}\|_j) \\ &+ n\sum_{j=1}^{p} p_{\lambda 2}(|\gamma_{j1}|)I(\|\boldsymbol{\gamma}_{j*}\|_j = 0) \end{aligned} \tag{4.3}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, $W_i = (X_{i0}B(Z_i)^T, \ldots, X_{ip}B(Z_i)^T)^T$, $\boldsymbol{W} = (W_1^T, \ldots, W_n^T)^T$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_0^T, \ldots, \boldsymbol{\gamma}_p^T)^T$.

We denote the subsets of predictors of varying, non-zero constant and zero effects by $\mathcal{V}$, $\mathcal{C}$, and $\mathcal{Z}$ respectively. The iterative algorithm is carried out in the following two steps.

At step 1, to separate the varying and nonzero constant effects, we minimize the following penalized loss:

$$Q_1(\boldsymbol{\gamma}) = (\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma})^T (\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma}) + n\sum_{j=1}^{p} p_{\lambda_1}(\|\gamma_{j*}\|_j) \tag{4.4}$$

to obtain $\boldsymbol{\gamma}^{\mathcal{VC}}$. All the predictors are assumed to be in the subset of varying effects $\mathcal{V}$ initially.

The penalization is conducted in a group manner. The jth predictor will be moved from subset $\mathcal{V}$ to $\mathcal{C}$ if $\|\hat{\gamma}_{j*}^{\mathcal{VC}}\|_j$=0, or it will stay in $\mathcal{V}$.

At step 2, the penalized criterion

$$Q_2(\boldsymbol{\gamma}) = (\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma})^T(\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma}) + n\sum_{j=1}^{p} p_{\lambda_2}(|\gamma_{j1}|)I(\|\boldsymbol{\gamma}_{j*}\|_j = 0) \tag{4.5}$$

is minimized only with regard to the predictors in set $\mathcal{C}$. If $\boldsymbol{X}_j$ is in $\mathcal{C}$, then it will be moved to $\mathcal{Z}$ if $|\hat{\gamma}_{j,1}^{\mathcal{CZ}}|$=0, otherwise it will stay in $\mathcal{C}$.

The above two steps will be iterated till convergence and we can obtain the estimator $\hat{\boldsymbol{\gamma}}$ at convergence. The coefficient function $\beta_j(z)$ $(j = 0, \ldots, p)$ in (4.1) can be estimated as $\hat{\beta}_j(z) = B^T(z)\hat{\boldsymbol{\gamma}}_j$. $\hat{\beta}_j(z)$ will be a varying function in z, non-zero constant and zero if $\hat{\boldsymbol{\gamma}}_j$ is in $\mathcal{V}$, $\mathcal{C}$ and $\mathcal{Z}$ respectively.

4.2.2.1 Group LASSO and group adaptive LASSO

In this section, we investigate the two-step iterative algorithm based on GCD approach using the convex penalty functions, such as LASSO and adaptive LASSO. Yuan and Lin [85] extended LASSO to the selection of variables group-wisely. By setting $p_{\lambda_i}(x) = \lambda_i\sqrt{D_{ij}}x$ for $i = 1, 2$ in (4.3), where D_{ij} is the corresponding dimension of predictor group j. Let $\widetilde{\boldsymbol{\gamma}}_j = \boldsymbol{R}_j^{\frac{1}{2}}\boldsymbol{\gamma}_j$ and $\widetilde{\boldsymbol{W}}_{ij} = \boldsymbol{R}_j^{-\frac{1}{2}}\boldsymbol{W}_{ij}$. Then (4.4) and (4.5) can be reexpressed as

$$Q_1(\widetilde{\boldsymbol{\gamma}}) = \left(\boldsymbol{Y} - \widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{\gamma}}\right)^T\left(\boldsymbol{Y} - \widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{\gamma}}\right) + n\sum_{j=1}^{p}\lambda_1\sqrt{D_{1j}}\|\widetilde{\boldsymbol{\gamma}}_{j*}\|_2$$

and

$$Q_2(\widetilde{\boldsymbol{\gamma}}) = \left(\boldsymbol{Y} - \widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{\gamma}}\right)^T\left(\boldsymbol{Y} - \widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{\gamma}}\right) + n\sum_{j=1}^{p}\lambda_1\sqrt{D_{2j}}|\widetilde{\boldsymbol{\gamma}}_{j1}|I(\|\widetilde{\boldsymbol{\gamma}}_{j*}\|_2 = 0)$$

respectively. From now on, we drop $\tilde{}$ from the formula for simplification of notation. The dimension of predictor group, D_{ij}, was included in the optimization, as in [85], to guarantee that the amount of penalization is consistent with the group size, so the small groups won't be dominated by large groups. However, note that the number of basis functions is the same for all the predictors in approximating the coefficient function, and the size of all the groups in Q_2 is 1, D_{ij} can be dropped from both Q_1 and Q_2.

By setting the partial derivative of Q_1 with respective to $\boldsymbol{\gamma}_{j*}$ to zero, we have

$$\hat{\boldsymbol{\gamma}}_{j*}^{\mathcal{VC}} = \left(1 - \frac{\lambda_1}{\|S_{1j}\|_2}\right)_+ S_{1j} \tag{4.6}$$

where $S_{1j} = \boldsymbol{W}_{j*}^T \left(\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma}_{-j*}\right)$ with $\boldsymbol{\gamma}_{-j*} = \left(\boldsymbol{\gamma}_0^T, \ldots, \boldsymbol{\gamma}_{j-1}^T, (\boldsymbol{\gamma}_{j1}, \mathbf{0}_{L-1})^T, \boldsymbol{\gamma}_{j+1}^T, \ldots, \boldsymbol{\gamma}_p^T\right)^T$, $\boldsymbol{W}_{j*}$ is the part of $\boldsymbol{W}_j$ corresponding to $\boldsymbol{\gamma}_{j*}$ and $(x)_+ = xI\{x > 0\}$. Similarly, setting the partial derivative of Q_2 with respective to $\boldsymbol{\gamma}_{j1}$ for those j such that $\|\boldsymbol{\gamma}_{j*}\|_2 = 0$ will lead to

$$\hat{\boldsymbol{\gamma}}_{j1}^{\mathcal{CZ}} = \left(1 - \frac{\lambda_2}{\|S_{2j}\|_2}\right)_+ S_{2j} \tag{4.7}$$

where $S_{2j} = \boldsymbol{W}_{j1}^T \left(\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma}_{-j1}\right)$ with $\boldsymbol{\gamma}_{-j1} = \left(\boldsymbol{\gamma}_0^T, \ldots, \boldsymbol{\gamma}_{j-1}^T, (0, \boldsymbol{\gamma}_{j*})^T, \boldsymbol{\gamma}_{j+1}^T, \ldots, \boldsymbol{\gamma}_p^T\right)^T$, and $\boldsymbol{W}_{j1}$ is the part of $\boldsymbol{W}_j$ corresponding to $\boldsymbol{\gamma}_{j1}$.

The group adaptive LASSO estimator could be obtained by simply replacing λ_1 and λ_2 with $\lambda_1\|S_{1j}\|_2^{-1}$ and $\lambda_2\|S_{2j}\|_2^{-1}$, in (4.6) and (4.7) respectively. (4.6) and (4.7) are essentially the multivariate version of the soft-thresholding operator in [96]. They will be updated solely for predictors with varying and constant effect, respectively. All the rest components of $\hat{\boldsymbol{\gamma}}_j^{\mathcal{VC}}$ and $\hat{\boldsymbol{\gamma}}_j^{\mathcal{CZ}}$ will remain as the updates in previous iteration. Both solutions have closed forms and are exempt from any sort of approximations. Furthermore, the GCD algorithm is

piecewise linear and therefore exceptionally well suited for high dimensional scenarios.

4.2.2.2 Group TLP and group SCAD

In group settings, the closed form solutions (4.6) and (4.7) exist not only for convex penalties such as LASSO and adaptive LASSO, but also for non-convex penalties like truncated L_1-function penalty, TLP, in Shen et al [97], and smoothly clipped absolute deviation penalty, SCAD, as in Fan and Li [75].

The TLP penalty function is given as

$$p_{\tau,\lambda}(x) = \min(\frac{|x|}{\tau}, 1)\lambda \tag{4.8}$$

where positive tuning parameters λ and τ control adaptive model selection and the degree of sparsity, correspondingly. As pointed out in Shen et al [97], with properly tuned τ, the approximation error of TLP to L_0 function reduces to 0, and low resolution coefficients can be taken good care of. By resorting to difference convex (DC) approach, Shen et al [97] turned non-convex optimization problems into its convex counterpart, thus significantly improved the computational efficiency and stability.

Following the similar lines of derivations in Shen et al [97] and Xue and Qu [98], the closed form solutions for (4.4) and (4.5) with group TLP penalty can be obtained by replacing λ_1 and λ_2 in (4.6) and (4.7) with $\frac{\lambda_1}{\tau_1} I\left(\|\hat{\boldsymbol{\gamma}}_{j*}\|_2 \leq \tau_1\right)$ and $\frac{\lambda_2}{\tau_2} I\left(\|\hat{\boldsymbol{\gamma}}_{j1}\|_2 \leq \tau_2\right)$, respectively, where $\hat{\boldsymbol{\gamma}}_{j*}$ and $\hat{\boldsymbol{\gamma}}_{j1}$ are taken as their most recent updates.

Before finding the solutions to (4.4) and (4.5) under group SCAD penalty, we briefly

revisit the SCAD penalty function, defined in Fan and Li [75] as

$$p_{\lambda,a}(x) = \begin{cases} \lambda x & \text{if } 0 \leq x \leq \lambda \\ -\frac{(x^2 - 2a\lambda x + \lambda^2)}{2(a-1)} & \text{if } \lambda < x \leq a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } x > a\lambda \end{cases} \tag{4.9}$$

where $\lambda > 0$ and $a > 2$, and its derivative

$$p'_{\lambda,a}(x) = \begin{cases} \lambda & \text{if } 0 \leq x \leq \lambda \\ -\frac{(x - a\lambda)}{a-1} & \text{if } \lambda < x \leq a\lambda \\ 0 & \text{if } x > a\lambda \end{cases} \tag{4.10}$$

The spirit of SCAD penalty could be explicitly conveyed by (4.10). SCAD starts its penalization with the same rate as LASSO till the point $x = \lambda$. Then the rate continuously reduces to 0 and will stay as 0 from the point $x = a\lambda$. Therefore SCAD is capable of correcting the bias introduced by LASSO while still performing variable selection. The group SCAD estimator for $\lambda > 0$ and $a_i > 2$ (i=1,2) can be derived as

$$\hat{\boldsymbol{\gamma}}_{j*}^{\mathcal{VC}} = \begin{cases} \left(1 - \frac{\lambda_1}{\|S_{1j}\|_2}\right)_+ S_{1j} & \text{if } \|S_{1j}\|_2 \leq 2\lambda \\ \frac{a_1 - 1}{a_1 - 2}\left(1 - \frac{a_1\lambda_1}{(a_1-1)\|S_{1j}\|_2}\right)_+ S_{1j} & \text{if } 2\lambda < \|S_{1j}\|_2 \leq a_1\lambda \\ S_{1j} & \text{if } \|S_{1j}\|_2 > a_1\lambda \end{cases} \tag{4.11}$$

and

$$\hat{\boldsymbol{\gamma}}_{j1}^{\mathcal{CZ}} = \begin{cases} \left(1 - \frac{\lambda_1}{\|S_{2j}\|_2}\right)_+ S_{2j} & \text{if } \|S_{2j}\|_2 \leq 2\lambda \\ \frac{a_2 - 1}{a_2 - 2}\left(1 - \frac{a_2\lambda_1}{(a_2-1)\|S_{2j}\|_2}\right)_+ S_{2j} & \text{if } 2\lambda < \|S_{2j}\|_2 \leq a_2\lambda \\ S_{2j} & \text{if } \|S_{2j}\|_2 > a_2\lambda \end{cases} \tag{4.12}$$

with notations defined the same as in (4.6) and (4.7), correspondingly. (4.11) and (4.12) are of the soft-thresholding property as $a_1 \longrightarrow \infty$ and $a_2 \longrightarrow \infty$ respectively.

4.2.2.3 Convergence of the GCD algorithm

In the previous two sections, we obtained the closed form updates for all the 4 penalty functions (LASSO, adaptive LASSO, TLP and SCAD) within the group settings. At each cycle of the GCD algorithm, the minimization of Q with respect to γ_j $(j = 1, \ldots, p)$ is achieved by iterating the minimization of Q_1 with respect to γ_{j*}, and Q_2 with respect to γ_{j1} correspondingly, where Q, Q_1 and Q_2 are defined in (4.3), (4.4) and (4.5) respectively. Hence, the two-stage iterative algorithm is of the descent property. Meanwhile, the value of Q at each cycle is nonnegative. So we have the following proposition:

Proposition 1. Let $\hat{\boldsymbol{\gamma}}^{(k)}$ be the estimated spline coefficient vector at the convergence of kth iteration. Then for the group penalties of LASSO, Adaptive LASSO, TLP and SCAD, GCD has the property such that

$$Q(\hat{\boldsymbol{\gamma}}^{(k)}) \leq Q(\hat{\boldsymbol{\gamma}}^{(k-1)})$$

Moreover, given an initial value $\hat{\boldsymbol{\gamma}}^{(0)}$, the sequence $\{Q(\hat{\boldsymbol{\gamma}}^{(k)}), k \geq 1\}$ converges to a local minimal of Q.

It follows from the above proposition that the two-stage iterative algorithm always converges. The penalized loss function Q is not convex in general, otherwise the convergence to the global optimum will be guaranteed by Proposition 1. Given that dimension p is fixed and $p < n$, the asymptotic properties of the two-stage estimator were established in Tang et al [73] with adaptive LASSO penalty, and Wu and Cui [93] with SCAD penalty. The global

optimality of the two stage estimator for large p small n is beyond the scope of this chapter and will not be discussed here.

4.2.3 Selection of tuning parameters

Here we propose a data-driven procedure to choose the proper tuning parameters. For group LASSO and group adaptive LASSO, there are 4 tuning parameters, O, L, λ_1 and λ_2, where the degree of the B spline basis O and the number of interior knots L uniformly spaced on [0,1] govern the smoothness of the varying coefficient functions, while λ_1 and λ_2 control the sparsity of the estimator.

At the beginning, we apply BIC in Schwarz [79] to choose O and L. The range for L can be determined by $[\max(\lfloor 0.5n^{\frac{1}{(2O+3)}} \rfloor, 1), \lfloor 1.5n^{\frac{1}{(2O+3)}} \rfloor]$, with $\lfloor x \rfloor$ denoting the integer part of x. The optimal combination of O and L can be reached by searching the corresponding two-dimensional grid, according to the following criterion

$$\text{BIC}_{O,L} = \log\left(\boldsymbol{Y} - \boldsymbol{W}\hat{\boldsymbol{\gamma}}\right)^T \left(\boldsymbol{Y} - \boldsymbol{W}\hat{\boldsymbol{\gamma}}\right) + \frac{(O+L+1)}{n}\log(n)$$

for $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\gamma}}_0^T, \mathbf{0}^T, \ldots, \mathbf{0}^T)^T$. Conditional on the chosen N and p, we can determine optimal λ_1 by minimizing

$$\text{BIC}_{\lambda_1} = \log\left(\boldsymbol{Y} - \boldsymbol{W}\hat{\boldsymbol{\gamma}}_{\lambda_1}\right)^T \left(\boldsymbol{Y} - \boldsymbol{W}\hat{\boldsymbol{\gamma}}_{\lambda_1}\right) + \frac{df_{\lambda_1}}{n}\log(n) \tag{4.13}$$

where $\hat{\boldsymbol{\gamma}}_{\lambda_1}$ is the minimizer of (4.4) given λ_1, and df_{λ_1} is the effective degree of freedom, defined as the total number of varying and non-zero constant predictors. Once we selected

N, p and λ_1, the optimal λ_2 can be determined by minimizing

$$\text{BIC}_{\lambda_2} = \log\left(\boldsymbol{Y} - \boldsymbol{W}\hat{\boldsymbol{\gamma}}_{\lambda_2}\right)^T\left(\boldsymbol{Y} - \boldsymbol{W}\hat{\boldsymbol{\gamma}}_{\lambda_2}\right) + \frac{df_{\lambda_2}}{n}\log(n) \tag{4.14}$$

where $\hat{\boldsymbol{\gamma}}_{\lambda_2}$ is the minimizer of (4.5) given λ_2, and df_{λ_2} is defined similarly as df_{λ_2}.

We need to choose additional tuning parameters τ_1 and τ_2 for group TLP algorithm. Slight modifications can be made to (4.13) so we can carry out a two-dimensional search for the best pair of λ_1 and τ_1. Optimal λ_2 and τ_2 can be taken similarly.

Note that in the group SCAD algorithm, we also have two more tuning parameters a_1 and a_2 than those in group LASSO and group adaptive LASSO. A search for the best combination of λ_i and a_i (i=1,2) over a two dimensional grid will be computationally intensive. It was pointed out in Fan and Li [75] that as a function of a in (4.9), the Bayesian risks are not sensitive to the choice of a. Here a is fixed as 2.2 in the subsequent analysis. We can tune λ_1 and λ_2 according to (4.13) and (4.14) correspondingly.

4.3 Simulation study

In this section, we carry out Monte Carlo simulations to assess the finite sample performance of group LASSO, group adaptive LASSO, group TLP and group SCAD in our framework through the GCD algorithm. The accuracy of variable selection and successful separation of the varying, non-zero constant and zero effects can be assessed by the percentage of choosing the correct model out of the total R replicates. Integrated mean squared error (IMSE), is adopted for evaluation of the estimation precision.

Let $\hat{\boldsymbol{\beta}}_j^{(r)}$ be the estimator of the coefficient function $\boldsymbol{\beta}_j$ in the rth ($1 \leq r \leq R$) replication,

and $\{z_k\}_{k=1}^{n_{\text{grid}}}$ be the grid points where $\hat{\boldsymbol{\beta}}_j^{(r)}$ is evaluated. The integrated mean squared error (IMSE) of $\hat{\boldsymbol{\beta}}_j(z)$ is defined as

$$\text{IMSE}(\hat{\boldsymbol{\beta}}_j(z)) = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{n_{\text{grid}}}\sum_{k=1}^{n_{\text{grid}}}\{\hat{\boldsymbol{\beta}}_j^{(r)}(z_k) - \boldsymbol{\beta}_j(z_k)\}^2 \tag{4.15}$$

(4.15) can be used to measure the estimation precision for the jth predictor. The overall estimation accuracy is reflected by the total integrated mean squared error (TIMSE) of all the coefficients, defined as TIMSE=$\sum_{j=0}^{p}\hat{\boldsymbol{\beta}}_j(z)$. Note that IMSE($\hat{\beta}_j$) reduces to MSE($\hat{\beta}_j$) when $\hat{\boldsymbol{\beta}}_j$ is a constant.

We rewrite model (4.1) here as

$$Y_i = \sum_{j=0}^{p}\beta_j(Z_i)X_{ij} + \varepsilon_i$$

($i = 1, \ldots, n$) for describing our simulation schemes. The predictors were generated from both continuous and discrete distributions. Two types of distribution for ε_i, a standard normal distribution and a t distribution with 3 degrees of freedom, were taken. Without loss of generality, assume the first 3 coefficients are varying, next two are constant and all the rest coefficients are redundant. The coefficients were set as: $\beta_0(z) = 5 + 3\sin(2\pi z)$, $\beta_1(z) = 2 - 3\cos\{(6z-5)\pi/3\}$, $\beta_2(z) = 7 - 7z$, $\beta_3(z) = 4.5$, $\beta_4(z) = 3$, and $\beta_j(z) = 0$ for $j = 5, \ldots, p$. All the simulations were carried out with sample size $n = 200$, p =10,100,200,400 and a total of R =200 replicates. All the approaches use LASSO as the initial estimate.

Example 4.1. We simulate the responses from model (4.1), where the index variable $X_i \sim$ Uniform(0,1), and the predictors were generated from a multivariate normal distribution with mean vector $\mathbf{0}$ and $Cov(X_{ij}, X_{ij'}) = 0.5^{|j-j'|}$ for $0 \le j, j' \le p$. The IMSE of the

estimator obtained from all the 4 approaches and the true model were calculated. Under the standard normal error, as p increases, group LASSO has a systematic less satisfactory performance than the others which have relative stable and comparable performances, with group TLP gradually establishing its advantage over the rest in terms of both oracle percentage and estimation precision, as shown in Table 4.1. A pretty much similar trend can be observed in Table 4.2 when the random error was generated from $t(3)$ distribution, though the performance of all the procedures are slightly worse than that under the standard normal error.

Table 4.1: Simulation results of Example 4.1, $N(0,1)$ error

		$N(0,1)$ error				
		gLASSO	gALASSO	gSCAD	gTLP	Oracle
p=10	Oracle Perc.	0.615	0.955	0.97	1	1
	IMSE					
	$\beta_0(u)$	0.1393	0.1380	0.1388	0.1492	0.1492
	$\beta_1(u)$	0.3199	0.1678	0.1427	0.1641	0.1669
	$\beta_2(u)$	0.4777	0.1539	0.1213	0.0666	0.0674
	$\beta_3(u)$	0.1771	0.0175	0.0122	0.0101	0.0101
	$\beta_4(u)$	0.0784	0.0231	0.0103	0.0081	0.0081
	TIMSE	1.2022	0.5019	0.4339	0.3980	0.4017
p=100	Oracle Perc.	0.515	0.995	0.995	1	1
	IMSE					
	$\beta_0(u)$	0.1650	0.1657	0.1648	0.1587	0.1506
	$\beta_1(u)$	0.2302	0.1541	0.1477	0.1607	0.1374
	$\beta_2(u)$	0.7810	0.6678	0.6515	0.0713	0.0855
	$\beta_3(u)$	0.0792	0.0140	0.0176	0.0096	0.0093
	$\beta_4(u)$	0.0711	0.0216	0.0128	0.0082	0.0081
	TIMSE	1.3308	1.0232	0.9945	0.4085	0.3909
p=200	Oracle Perc.	0.355	0.985	1	1	1
	IMSE					
	$\beta_0(u)$	0.1771	0.1761	0.1781	0.1612	0.1523
	$\beta_1(u)$	0.2259	0.1599	0.1595	0.1720	0.1480
	$\beta_2(u)$	0.8094	0.6731	0.6484	0.0764	0.0845
	$\beta_3(u)$	0.0812	0.0193	0.0202	0.0087	0.0089
	$\beta_4(u)$	0.0804	0.0271	0.0144	0.0081	0.0085
	TIMSE	1.3825	1.0554	1.0206	0.4265	0.4022
p=400	Oracle Perc.	0.27	0.99	1	1	1
	IMSE					
	$\beta_0(u)$	0.2104	0.2102	0.2099	0.1654	0.1654
	$\beta_1(u)$	0.2514	0.1744	0.1697	0.1729	0.1748
	$\beta_2(u)$	0.8967	0.7930	0.7840	0.0746	0.0756
	$\beta_3(u)$	0.0705	0.0256	0.0273	0.0108	0.0108
	$\beta_4(u)$	0.0749	0.0257	0.0124	0.0072	0.0072
	TIMSE	1.5188	1.2290	1.2033	0.4310	0.4339

Table 4.2: Simulation results of Example 4.1, $t(3)$ error

		$t(3)$ error				
		gLASSO	gALASSO	gSCAD	gTLP	Oracle
p=10	Oracle Perc.	0.29	0.83	0.945	0.97	1
	IMSE					
	$\beta_0(u)$	0.1929	0.1921	0.1913	0.2063	0.2066
	$\beta_1(u)$	0.3392	0.2198	0.2050	0.2343	0.2386
	$\beta_2(u)$	0.5859	0.3676	0.3358	0.1445	0.1464
	$\beta_3(u)$	0.1775	0.0457	0.0400	0.0333	0.0294
	$\beta_4(u)$	0.1100	0.0463	0.0358	0.0233	0.0211
	TIMSE	1.4520	0.8856	0.8170	0.6526	0.6421
p=100	Oracle Perc.	0.01	0.74	0.965	0.98	1
	IMSE					
	$\beta_0(u)$	0.2557	0.2537	0.2551	0.2109	0.2002
	$\beta_1(u)$	0.2834	0.2173	0.2220	0.2370	0.2134
	$\beta_2(u)$	0.9739	0.8729	0.8721	0.1386	0.1544
	$\beta_3(u)$	0.1044	0.0538	0.0510	0.0273	0.0293
	$\beta_4(u)$	0.1036	0.0549	0.0402	0.0260	0.0230
	TIMSE	1.8882	1.5170	1.4614	0.7184	0.6204
p=200	Oracle Perc.	0.005	0.67	0.915	0.955	1
	IMSE					
	$\beta_0(u)$	0.3046	0.2963	0.3026	0.2004	0.1999
	$\beta_1(u)$	0.2857	0.2077	0.2103	0.2405	0.2515
	$\beta_2(u)$	1.1327	1.0282	1.0327	0.1441	0.1577
	$\beta_3(u)$	0.0833	0.0484	0.0523	0.0256	0.0261
	$\beta_4(u)$	0.1156	0.0501	0.0340	0.0237	0.0212
	TIMSE	2.2144	1.8028	1.6943	0.7474	0.6564
p=400	Oracle Perc.	0	0.57	0.905	0.96	1
	IMSE					
	$\beta_0(u)$	0.3481	0.3428	0.3398	0.2097	0.1955
	$\beta_1(u)$	0.2902	0.2219	0.2279	0.2522	0.2433
	$\beta_2(u)$	1.2438	1.1509	1.1543	0.1555	0.1369
	$\beta_3(u)$	0.0886	0.0513	0.0491	0.0282	0.0239
	$\beta_4(u)$	0.1170	0.0482	0.0346	0.0227	0.0223
	TIMSE	2.3943	1.9293	1.8288	0.7677	0.6220

Example 4.2. Now the simulation in a genetic setup is considered. The responses were generated from model (4.1) with SNP X_j ($j = 0, \ldots, p$) coded as 3 categories (1,0,-1) for genotypes (AA,Aa,aa) respectively.

The LD based simulation scheme was adopted to generate the genotype data. Let p_A and p_B be the minor allele frequencies (MAFs) of two risk alleles A and B for two adjacent SNPs, with linkage disequilibrium denoted as δ. The frequencies of four haplotypes are $p_{ab} = (1-p_A)(1-p_B)+\delta$, $p_{Ab} = p_A(1-p_B)-\delta$, $p_{aB} = (1-p_A)p_B-\delta$, and $p_{AB} = p_Ap_B+\delta$. Assuming the Hardy-Weinberg equilibrium, we can simulate the SNP genotype at locus A from a multi-nomial distribution with frequencies p_A^2,$2p_A(1-p_A)$ and $(1-p_A)^2$ for genotypes (AA,Aa,aa) respectively. SNP genotype at locus B conditional on that at locus A can be simulated based on the conditional probability matrix in Cui et al. (2008)[39].

Table **??**–**??** summarize the estimation results in example 2. For all the four approaches, as the minor allele frequency (MAF) pA goes up from 0.1 to 0.5, the proportion of choosing the correct model generally increases, and the estimation error reduces. Group LASSO consistently performed worse than its counterparts in terms of the two criteria. Under the standard normal error, when p =10, the difference in performance among the 3 different MAFs, especially between MAF 0.3 and 0.5, is not significant for all the four methods. However, as p increases, dramatic differences are observed. Starting from p=100, all the approaches can barely choose the exact true model as pA=0.1, and the corresponding TIMSEs are quite large. Given MAF 0.3, the performance of all the approaches except group TLP in very high dimensions,such as p=200 and 400, fall way behind that given MAF 0.5. The performance of group TLP is relative stable compared to the other 3 methods, especially when the dimension is extremely high, as p=400.

We observe similar patterns when the procedures are assessed under t error with 3 de-

grees of freedom. In general, the performances of all the procedures under standard normal error are superior. Group TLP outperforms the others in all the scenarios.

Table 4.3: Simulation results of Example 4.2, $p = 10$, $n = 200$, $N(0,1)$ error

		$N(0,1)$ error				
		gLASSO	gALASSO	gSCAD	gTLP	Oracle
pA=0.1	Oracle Perc.	0.44	0.795	0.935	0.96	1
	IMSE					
	$\beta_0(u)$	0.7266	0.7252	0.7574	0.3981	0.3895
	$\beta_1(u)$	1.1823	1.5831	0.6841	0.4124	0.4861
	$\beta_2(u)$	1.8650	1.3619	0.7610	0.3528	0.3777
	$\beta_3(u)$	0.4954	0.2529	0.1010	0.0615	0.0573
	$\beta_4(u)$	0.2002	0.2767	0.0650	0.0415	0.0416
	TIMSE	4.5192	4.2036	2.3704	1.2934	1.3521
pA=0.3	Oracle Perc.	0.615	0.98	0.99	0.985	1
	IMSE					
	$\beta_0(u)$	0.2152	0.2131	0.2122	0.1785	0.1783
	$\beta_1(u)$	0.4975	0.2814	0.2064	0.2204	0.2277
	$\beta_2(u)$	1.0013	0.3706	0.2660	0.1237	0.1259
	$\beta_3(u)$	0.4007	0.0467	0.0268	0.0245	0.0237
	$\beta_4(u)$	0.1559	0.0665	0.0254	0.0195	0.0195
	TIMSE	2.2862	0.9808	0.7424	0.5699	0.5751
pA=0.5	Oracle Perc.	0.73	0.97	0.99	0.99	1
	IMSE					
	$\beta_0(u)$	0.2152	0.2131	0.2122	0.1785	0.1783
	$\beta_1(u)$	0.4975	0.2814	0.2064	0.2204	0.2277
	$\beta_2(u)$	1.0013	0.3706	0.2660	0.1237	0.1259
	$\beta_3(u)$	0.4007	0.0467	0.0268	0.0245	0.0237
	$\beta_4(u)$	0.1559	0.0665	0.0254	0.0195	0.0195
	TIMSE	2.2862	0.9808	0.7424	0.5699	0.5751

Table 4.4: Simulation results of Example 4.2, $p = 10$, $n = 200$, $t(3)$ error

		$t(3)$ error				
		gLASSO	gALASSO	gSCAD	gTLP	Oracle
pA=0.1	Oracle Perc.	0.33	0.54	0.705	0.79	1
	IMSE					
	$\beta_0(u)$	1.6609	1.6954	1.7427	1.0722	0.9165
	$\beta_1(u)$	1.4901	2.3887	1.7687	0.9836	1.0244
	$\beta_2(u)$	1.9043	2.2271	1.6040	0.9273	0.9033
	$\beta_3(u)$	0.5425	0.6404	0.3444	0.2340	0.1560
	$\beta_4(u)$	0.4238	0.6674	0.2963	0.1649	0.1261
	TIMSE	6.1169	7.6370	5.8026	3.5783	3.1263
pA=0.3	Oracle Perc.	0.25	0.895	0.935	0.955	1
	IMSE					
	$\beta_0(u)$	0.3432	0.3406	0.3432	0.2534	0.2491
	$\beta_1(u)$	0.5782	0.4111	0.4861	0.3823	0.3835
	$\beta_2(u)$	1.1033	0.6824	0.7795	0.3340	0.3098
	$\beta_3(u)$	0.3668	0.0949	0.0828	0.0825	0.0706
	$\beta_4(u)$	0.2144	0.1224	0.0648	0.0592	0.0522
	TIMSE	2.6777	1.6856	1.7235	1.1496	1.0652
pA=0.5	Oracle Perc.	0.23	0.845	0.96	0.965	1
	IMSE					
	$\beta_0(u)$	0.3432	0.3406	0.3432	0.2534	0.2491
	$\beta_1(u)$	0.5782	0.4111	0.4861	0.3823	0.3835
	$\beta_2(u)$	1.1033	0.6824	0.7795	0.3340	0.3098
	$\beta_3(u)$	0.3668	0.0949	0.0828	0.0825	0.0706
	$\beta_4(u)$	0.2144	0.1224	0.0648	0.0592	0.0522
	TIMSE	2.6777	1.6856	1.7235	1.1496	1.0652

Table 4.5: Simulation results of Example 4.2, $p = 100$, $n = 200$, $N(0,1)$ error

		$N(0,1)$ error				
		gLASSO	gALASSO	gSCAD	gTLP	Oracle
pA=0.1	Oracle Perc.	0	0	0.005	0.015	1
	IMSE					
	$\beta_0(u)$	5.9486	5.9577	5.9433	4.8575	0.4180
	$\beta_1(u)$	2.1664	4.2821	5.0012	1.0952	0.5239
	$\beta_2(u)$	2.5871	2.6322	2.6478	0.9247	0.5210
	$\beta_3(u)$	7.4511	0.5889	3.4381	0.0965	0.0624
	$\beta_4(u)$	4.1242	0.6425	2.0292	0.0693	0.0508
	TIMSE	11.978	14.4393	14.2111	8.7643	1.5761
pA=0.3	Oracle Perc.	0.01	0.565	0.82	0.99	1
	IMSE					
	$\beta_0(u)$	1.0107	1.0204	1.0199	0.1690	0.1666
	$\beta_1(u)$	0.5794	0.4570	0.5103	0.2204	0.2202
	$\beta_2(u)$	1.1485	0.9477	1.5749	0.1567	0.1237
	$\beta_3(u)$	0.2246	0.0596	0.0674	0.0233	0.0232
	$\beta_4(u)$	0.1854	0.1118	0.0457	0.0153	0.0153
	TIMSE	3.2486	2.6329	3.2207	0.5847	0.5489
pA=0.5	Oracle Perc.	0.55	1	1	1	1
	IMSE					
	$\beta_0(u)$	0.1593	0.1588	0.1606	0.1561	0.1561
	$\beta_1(u)$	0.3242	0.2071	0.2637	0.1953	0.2017
	$\beta_2(u)$	0.9660	0.8267	1.1570	0.1088	0.1145
	$\beta_3(u)$	0.1357	0.0330	0.0360	0.0197	0.0197
	$\beta_4(u)$	0.1328	0.0588	0.0198	0.0142	0.0143
	TIMSE	1.7231	1.2843	1.6370	0.4942	0.5063

Table 4.6: Simulation results of Example 4.2, $p = 100$, $n = 200$, $t(3)$ error

		$t(3)$ error				
		gLASSO	gALASSO	gSCAD	gTLP	Oracle
pA=0.1	Oracle Perc.	0	0	0	0	1
	IMSE					
	$\beta_0(u)$	6.4423	6.4941	6.4759	5.3221	0.8603
	$\beta_1(u)$	2.2632	3.5782	4.2976	1.9710	0.9942
	$\beta_2(u)$	2.6220	3.3151	3.2927	1.9261	0.9143
	$\beta_3(u)$	0.6904	1.0206	0.3533	0.2403	0.1557
	$\beta_4(u)$	0.6988	1.5284	0.6974	0.1799	0.1369
	TIMSE	13.0771	16.6350	15.6337	12.0716	3.0614
pA=0.3	Oracle Perc.	0	0.065	0.27	0.86	1
	IMSE					
	$\beta_0(u)$	1.9184	1.9154	1.9258	0.3391	0.2511
	$\beta_1(u)$	0.6734	0.6233	0.8530	0.4326	0.3638
	$\beta_2(u)$	1.3878	1.4093	2.1560	0.5030	0.3218
	$\beta_3(u)$	0.2408	0.1414	0.1447	0.0698	0.0650
	$\beta_4(u)$	0.2472	0.1723	0.0857	0.0474	0.0439
	TIMSE	5.1465	5.2540	5.9097	2.0213	1.0456
pA=0.5	Oracle Perc.	0.035	0.725	0.935	0.985	1
	IMSE					
	$\beta_0(u)$	0.2407	0.2443	0.2407	0.2057	0.2088
	$\beta_1(u)$	0.3715	0.2827	0.3359	0.3258	0.3332
	$\beta_2(u)$	1.3128	1.2550	1.4448	0.2307	0.2403
	$\beta_3(u)$	0.1391	0.1130	0.1130	0.0507	0.0523
	$\beta_4(u)$	0.1893	0.1120	0.0783	0.0361	0.0346
	TIMSE	2.5354	2.1107	2.2240	0.9181	0.8692

Table 4.7: Simulation results of Example 4.2, $p = 200$, $n = 200$, $N(0,1)$ error

		$N(0,1)$ error				
		gLASSO	gALASSO	gSCAD	gTLP	Oracle
pA=0.1	Oracle Perc.	0	0	0	0	1
	IMSE					
	$\beta_0(u)$	6.4035	6.4500	6.4518	5.0816	0.3961
	$\beta_1(u)$	2.1184	4.3209	4.9639	1.1121	0.4109
	$\beta_2(u)$	2.3654	2.6648	2.8046	0.8625	0.3500
	$\beta_3(u)$	0.7732	0.4838	0.3425	0.0938	0.0586
	$\beta_4(u)$	0.4501	0.5315	0.1985	0.5025	0.0426
	TIMSE	12.2412	14.8546	14.7897	8.9683	1.2582
pA=0.3	Oracle Perc.	0	0.075	0.345	0.98	1
	IMSE					
	$\beta_0(u)$	1.8631	1.8954	1.8833	0.1987	0.1703
	$\beta_1(u)$	0.6342	5.5156	0.8991	0.2307	0.2113
	$\beta_2(u)$	1.3076	1.2343	2.3473	0.1570	0.1374
	$\beta_3(u)$	0.2365	0.0549	0.0628	0.0238	0.0230
	$\beta_4(u)$	0.2280	0.1338	0.0575	0.0192	0.0193
	TIMSE	4.4824	4.0975	5.2837	0.6483	0.5613
pA=0.5	Oracle Perc.	0.44	0.985	1	1	1
	IMSE					
	$\beta_0(u)$	0.1588	0.1594	0.1583	0.1525	0.1525
	$\beta_1(u)$	0.2979	0.2000	0.2639	0.2003	0.2064
	$\beta_2(u)$	1.0129	0.9003	1.2411	0.1141	0.1157
	$\beta_3(u)$	0.1180	0.0343	0.0433	0.0178	0.0178
	$\beta_4(u)$	0.1310	0.0676	0.0216	0.0158	0.0159
	TIMSE	1.7286	1.3615	1.7282	0.5006	0.5083

Table 4.8: Simulation results of Example 4.2, $p = 200$, $n = 200$, $t(3)$ error

		$t(3)$ error				
		gLASSO	gALASSO	gSCAD	gTLP	Oracle
pA=0.1	Oracle Perc.	0	0	0	0	1
	IMSE					
	$\beta_0(u)$	7.8770	7.8787	7.8986	5.628	0.8906
	$\beta_1(u)$	2.237	3.5923	4.4315	2.0438	1.0194
	$\beta_2(u)$	2.542	3.4581	3.4735	2.2584	0.8541
	$\beta_3(u)$	0.8399	0.8507	0.5609	0.2691	0.1614
	$\beta_4(u)$	0.7117	1.4850	0.5707	0.1665	0.1282
	TIMSE	14.7035	18.344	17.356	12.8456	3.0536
pA=0.3	Oracle Perc.	0	0	0.015	0.705	1
	IMSE					
	$\beta_0(u)$	3.0900	3.1150	3.1170	0.7914	0.2705
	$\beta_1(u)$	0.7765	0.7070	1.0440	0.5144	0.4026
	$\beta_2(u)$	1.4769	1.6380	2.6080	0.5512	0.3126
	$\beta_3(u)$	0.2164	0.1412	0.1220	0.0745	0.0715
	$\beta_4(u)$	0.3116	0.2076	0.1040	0.0536	0.0480
	TIMSE	6.5667	6.8139	7.2260	2.5457	1.1053
pA=0.5	Oracle Perc.	0	0.695	0.915	0.96	1
	IMSE					
	$\beta_0(u)$	0.2872	0.2928	0.3075	0.2178	0.2089
	$\beta_1(u)$	0.4168	0.3450	0.3626	0.3954	0.3450
	$\beta_2(u)$	1.4403	1.4106	1.5519	0.2830	0.3179
	$\beta_3(u)$	0.1521	0.0960	0.1116	0.0578	0.0553
	$\beta_4(u)$	0.2092	0.1303	0.0670	0.0557	0.0461
	TIMSE	2.9052	2.5833	2.4056	1.0997	0.9732

Table 4.9: Simulation results of Example 4.2, $p = 400$, $n = 200$, $N(0,1)$ error

		$N(0,1)$ error				
		gLASSO	gALASSO	gSCAD	gTLP	Oracle
pA=0.1	Oracle Perc.	0	0	0	0	1
	IMSE					
	$\beta_0(u)$	6.9606	6.8309	6.8046	5.1098	0.3685
	$\beta_1(u)$	2.1276	4.0791	4.8938	1.2151	0.4558
	$\beta_2(u)$	2.4348	2.8612	2.8865	1.1964	0.3100
	$\beta_3(u)$	0.6920	0.6841	0.4083	0.1016	0.0510
	$\beta_4(u)$	0.5384	0.9365	0.2659	0.0713	0.0392
	TIMSE	12.8890	15.8640	15.3025	9.3504	1.2246
pA=0.3	Oracle Perc.	0	0	0.095	0.88	1
	IMSE					
	$\beta_0(u)$	2.783	2.792	2.793	0.421	0.181
	$\beta_1(u)$	0.675	0.572	0.810	0.313	0.243
	$\beta_2(u)$	1.422	1.467	2.753	0.364	0.135
	$\beta_3(u)$	0.236	0.081	0.090	0.030	0.023
	$\beta_4(u)$	0.278	1.763	0.090	0.021	0.017
	TIMSE	5.713	5.596	6.643	1.258	0.600
pA=0.5	Oracle Perc.	0.29	0.985	1	0.995	1
	IMSE					
	$\beta_0(u)$	0.1817	0.1825	0.1831	0.1558	0.1551
	$\beta_1(u)$	0.2926	0.2092	0.2928	0.2169	0.2232
	$\beta_2(u)$	1.1169	1.0209	1.3567	0.1325	0.1162
	$\beta_3(u)$	0.1110	0.0401	0.0426	0.0205	0.0204
	$\beta_4(u)$	0.1392	0.0676	0.0250	0.0152	0.0152
	TIMSE	1.8554	1.5204	1.9003	0.5409	0.5300

Table 4.10: Simulation results of Example 4.2, $p = 400$, $n = 200$, $t(3)$ error

		$t(3)$ error				
		gLASSO	gALASSO	gSCAD	gTLP	Oracle
pA=0.1	Oracle Perc.	0	0	0	0	1
	IMSE					
	$\beta_0(u)$	8.3537	8.1152	8.2099	6.0230	0.9324
	$\beta_1(u)$	2.1540	3.5756	4.5341	2.1497	0.9725
	$\beta_2(u)$	2.5676	3.3866	3.4152	2.5080	0.9093
	$\beta_3(u)$	0.8431	0.6975	0.4919	0.2443	0.1773
	$\beta_4(u)$	0.7683	1.3577	0.6356	0.1642	0.1549
	TIMSE	15.1398	18.0344	17.5993	13.2892	3.1463
pA=0.3	Oracle Perc.	0	0	0	0.425	1
	IMSE					
	$\beta_0(u)$	4.127	4.114	4.123	1.968	0.279
	$\beta_1(u)$	0.812	0.681	0.940	0.701	0.427
	$\beta_2(u)$	1.562	1.815	2.985	0.823	0.305
	$\beta_3(u)$	0.264	0.187	0.162	0.114	0.071
	$\beta_4(u)$	0.359	0.244	0.143	0.076	0.057
	TIMSE	7.973	8.427	8.711	5.132	1.146
pA=0.5	Oracle Perc.	0	0.61	0.935	0.97	1
	IMSE					
	$\beta_0(u)$	0.3306	0.3289	0.3317	0.1991	0.2016
	$\beta_1(u)$	0.4178	0.3265	0.3698	0.3515	0.3564
	$\beta_2(u)$	1.4987	1.4649	1.6647	0.3334	0.2950
	$\beta_3(u)$	0.1294	0.0916	0.0888	0.0485	0.0468
	$\beta_4(u)$	0.2179	0.1178	0.0518	0.0457	0.2438
	TIMSE	3.084	2.5707	2.5263	1.1916	0.9435

4.4 Real data analysis

We used the genome-wide association data from Genetic Analysis Workshop (GAW) 18 for 142 unrelated subjects to illustrate the utility of our approach. It is widely recognized that individual's blood pressure is related to age. The aim of our study is to track down the genetic variants that can interpret the variation in Diastolic Blood Pressure (DBP) caused by non-linear penetrance effect over time (age). The environment condition is defined as age, and the blood pressure, measured as DBP, might not be the same for an individual with the same gene but of different environmental exposures. This is triggered by the complex interplay between the age and gene effects.

The dataset was cleaned by removing SNPs with MAF less than 0.05 or departure from Hardy-Weinberg equilibrium. Subjects with missing DBP or age, as well as with more than 1/3 missing genotypes, were excluded from the dataset before final analysis. We take a subset of 250 SNPs from chromosome 9 and imputed the missing value before the final analysis.

We applied our approach to the data and select the model

$$Y = \beta_0(z) + \beta_{j_1} X_{j_1} + \beta_{j_2} X_{j_2} + \varepsilon$$

where X_{j_1} and X_{j_2} correspond to SNP rs723877 and rs10972462, respectively. Both coefficients β_{j_1} and β_{j_1} are varying. Thus we identified two genetic risk variants that are sensitive to age to affect blood pressure.

A cross validation examination, the single SNP based analysis in Ma et al [37], was performed for SNPs on chromosome 9. From the over all association test corresponding to LM, LMI and VC models, we calculated p-values denoted as P_CON, P_LIN and P_VC. SNPs with at least one of the three p-values less than 5E-06 were tabulated in Table 4.11. For SNP

rs10972462, testing on the constant coefficients implies that the coefficient function of this SNP does change across age, as P_CON¡ 0.05. Further test on linear structure indicates that rs10972462 has a linear interaction with age (P_LIN¿ 0.05). Hence it makes sense that the p-value obtained from the test corresponding to LMI model is smaller than its counterpart with LM and VC model. A similar observation can be concluded for SNP rs723877.

Table 4.11: List of SNPs with p-value < 5E-06 on Chromosome 9

SNP ID	GeneName	Location	P_VC	P_CON	P_LIN	P_LM	P_LMI	P_I
			LMI model					
rs723877	UNC13B	Chr9	6.52E-06	0.00649	0.12311	2.83E-05	**2.24E-06**	0.00357
rs10972462	UNC13B	Chr9	2.13E-06	0.00208	0.07052	3.01E-05	**1.24E-06**	0.00175

4.5 Discussion

A growing number of pieces of evidence have demonstrated the importance of G×E interactions in complex traits, as the responses of genetic factors to environmental exposures play a crucial role in affecting disease risks and variations of quantitative traits. Many statistical methods have been developed to explore G×E interactions. The linear assumption of gene effect under environmental stimuli on which these methods rely is often violated in practice. Ma et al [37] and Wu and Cui [68] relaxed the linear assumption to allow for a non-linear genetic penetrance effect for continuous and binary disease phenotype, respectively. The true effect of G×E interactions is captured by the model itself, through a sequence of likelihood ratio tests.

A common feature of these methods, including Ma et al [37] and Wu and Cui [68], and those reviewed in Cornelis et al [65], is that the identification of the presence of G×E interactions is casted in a single genetic variant based hypothesis testing framework. To

our best knowledge, Wu and Cui [93] for the first approached the problem from a high dimensional variable selection perspective. Specifically, within the VC model framework, the presence of G×E interactions, no G×E interactions and no association with the phenotype can be determined by separating the VC coefficient functions after B spline basis expansion approximations as varying, non-zero constant and zero, correspondingly.

The effect separation and variable selection in Wu and Cui [93] is attained by penalized estimation on the model parameters so some varying coefficients are continuously shrunk to a non-zero constant or zero. The procedure is dependent on local quadratic approximations (LQA) to SCAD penalty function. However, LQA leads to loss in efficiency and accuracy, especially when dimension p is large, due to frequent factorizations of large matrices. The local linear approximation (LLA) to penalty functions proposed in Zou and Li [94] improved LQA but then was shown outperformed by coordinate descent method (CD), as in Breheny and Huang [87]. In this work, we integrate the group version of CD, or GCD, in the two-stage iterative procedure to exploit G×E interactions in a high dimensional setting.

The most prominent character of GCD is the penalized objective function is optimized over one individual predictor group at a time, so the computational complexity only increase linearly with the dimension p. This attribute ensures the superior performance of our framework. We examined both convex (LASSO, adaptive LASSO) and non-convex (TLP, SCAD) penalty functions. Extensive simulation results manifest that all the penalty except LASSO perform satisfactorily, and TLP excels in all the scenarios. As p grows from low to very high dimensions, the performance of approaches with all penalty functions remain relative stable, though slight drops was observed. The phenomenon indicates the advantage of our framework from another perspective.

Chapter 5

Concluding Remarks

It has been widely recognized that the naturally occurring variations in most complex disease traits are not merely explained by genetic factors, but also can be understood from the mechanism of genetic responses to environmental mediators. G×E interactions, the genetic control of the pattern to environmental stimuli, shed novel light on examining the trait variations. This dissertation focuses mainly on developing powerful statistical methods to tackle the challenges originated from G×E interactions.

In chapter 2, we developed a new statistical approach to extend the varying coefficient model for continuous quantitative response in our previous work to the binary disease response. The varying coefficients were estimated by the nonparametric B spline method due to its computational expediency and nice asymptotic properties. Our scheme has particular advantage in hunting down the fluctuation in gene functions across environmental changes. A significant boost in power were indicated in the simulation study when underlying penetrance effect of genetic variants is non-linear.

The simulation results also show that when the model for underlying mechanism of G×E interactions is misspecified, VC model may not have the higher power than LM and LMI models, since it is much more complex and needs large degrees of freedom for hypothesis testing. To determine which model fits the data more appropriately, we developed a sequential testing scheme. Our scheme is applied to two Type 2 Diabetes studies by first evaluating

constant coefficients, then linear and varying coefficients. The novel disease signals captured by VC model in our framework could be missed if our focus is restricted to linear model only.

A broad spectrum of available methodologies in exploring G×E interactions are coined within single genetic variant based hypothesis testing framework. Set based association study, such as the gene-centric, gene-set and pathway based analysis, has continued to soar in popularity as its advantage has increasingly been acknowledged. In chapter 3, we proposed a variant set based framework to examine how variants in a genetic system are mediated by a common environment factor to influence quantitative phenotypic response. We tackle the issue from a high dimensional variable selection standpoint. Specifically, we can identify the sensitivity of genetic variants to environment stimuli, which is tantamount to determine the coefficient function as varying, non-zero constant and zero, corresponding to cases of existence of G×E interactions, no G×E interactions and no genetic effects.

The procedure was implemented in a two stage iterative framework. We established the selection consistency and oracle properties of the penalized estimator with SCAD penalty, and demonstrated dramatic improvement in finite sample performance over the adaptive LASSO in simulation study, in terms of oracle percentage and estimation accuracy. The application of our approach to the JAK/STAT signaling pathway in LGA/SGA study correctly select the risk SNP without G×E effect. This framework is critically dependent on local quadratic approximations (LQA). Because of repeated factorizations of matrices, significance loss in computational efficiency and estimation precision will be caused when dimension is large, especially in scenarios of $p > n$.

To overcome this issue, we developed the group coordinate descent (GCD) approach within the two-stage iterative framework in chapter 4. After basis expansion, the penalized

loss function is minimized with regard to single predictor group at a time, and all the predictor groups are cycled until convergence of the algorithm. Therefore, the computational complexity only grows linearly with dimension p. Through extensive simulation study with different penalty functions (LASSO, adaptive LASSO, TLP and SCAD), we demonstrated the merit of this framework. Our approach yields high oracle percentages and estimation precision even when dimension p is much larger than sample size n.

The main objective of this dissertation is to develop novel statistical methodology for the elucidation of complicated machinery in G×E interactions. By implementing different link functions, our framework on investigating the non-linear G×E interactions can be readily extended to different types of phenotypic responses such as count or survival outcomes. We can also try to test if the novel hypothesis in Perry et al [52] that the significance of potential risk loci could be enhanced by the stratification of patients with Type 2 Diabetes based on BMI will result in any new findings within our framework. To the best of our knowledge, this dissertation first proposed the scheme of approaching G×E interactions from a high dimensional variable selection perspective. Generalizations of our framework to binary disease response in case control association study and survival response are currently undergoing. The selection consistency and oracle property when dimension of predictors p exceeds the sample size n for the procedure will also be examined. It is worth noting that to explore G×E interactions in ultra-high dimensional feature space, we can integrate the sure independence screen (SIS) procedure in Fan and Lv [99] into our framework. Relevant investigations will be carried out in the near future.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] M. Lynch and B. Walsh. *Genetics and Analysis of Quantitative Traits.* Sinauer Associates,Sunderland, MA, 1998.

[2] R. Wu, C.X. Ma and G. Casella. *Statistical Genetics of Quantitative traits– Linkage, Maps and QTL.* Springer, 2007.

[3] R.S. Spielman, R.E. McGinnis, W.J. Ewens. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet.* 52 (3): 506-516, 1993.

[4] N. Risch and K. Merikangas. The future of genetics studies of complex human diseases. *Science* 273: 1516–1517, 1996.

[5] J. Altmuller, L.J. Palmer, G. Fischer, H. Scherb and M. Wjst. Genome-wide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet.* 69(3): 936–950, 2001.

[6] International HapMap Consortium. The haplotype map of the human genome. *Nature* 437: 1299-1320, 2005.

[7] M.F. Moffatt, M. Kabesch, L. Liang, A.L. Dixon, D. Strachan, et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* 448(7152):470–473, 2007.

[8] D.F. Easton, K.A. Pooley, A.M. Dunning, P.D.P. Pharoah, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 447: 1087–1093, 2007.

[9] D.J. Hunter, P. Kraft, K.B. Jacobs, D.G. Cox, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet.* 39(7): 870–874, 2007.

[10] A. Helgadottir, G. Thorleifsson, A. Manolescu, S. Gretarsdottir, et al. A Common Variant on Chromosome 9p21 Affects the Risk of Myocardial Infarction. *Science* 316(5830): 1491–1493, 2007.

[11] R. McPherson, A. Pertsemlidis, N. Kavaslar, A. Stewart, et al. A common allele on chromosome 9 associated with coronary heart disease. *Science* 316(5830): 1488–1491, 2007.

[12] R. Sladek, G. Rocheleau, J.Rung, C. Dina, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 445: 881–885, 2007.

[13] L.J. Scott, K.L. Mohlke, L.L. Bonnycastle, C.J. Willer, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 316(5829): 1341–1345, 2007.

[14] M.I. McCarthy, G.R. Abecasis, L.R. Cardon, D.B. Goldstein et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 9(5): 356–369, 2008.

[15] T.A. Manolio, F.S. Collins, N.J. Cox, D.B. Goldstein, et al. Finding the missing heritability of complex diseases. *Nature* 461(7265): 747–753,2009.

[16] D.E. Reich and E.S. Lander. On the allelic spectrum of human disease. *Trends Genet.* 17(9): 502–510, 2001.

[17] J.K. Pritchard and N.J. Cox. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum. Mol. Genet.* 11(20): 2417–2423, 2002.

[18] E. Zeggini, W. Rayner, A.P. Morris, A.T. Hattersley, et al. An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets. *Nat. Genet.* 37: 1320–1322, 2005.

[19] E.R. Mardis. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* 9:387–402, 2008.

[20] E.R. Mardis. The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24(3): 133–141, 2008.

[21] J. Shendure and H. L. Lee. Next-generation DNA sequencing. *Nature Biotech.* 26: 1135–1145, 2008.

[22] M.L. Metzker. Sequencing technologies - the next generation. *Nat Rev Genet.* 11(1):31–46 ,2009

[23] M.I. McCarthy and J.N. Hirschhorn. Genome-wide association studies: potential next steps on a genetic journey. *Hum. Mol. Genet.* 17(R2): R156–65, 2008.

[24] J.B.S. Haldane. Heredity and Politics. New York, NY: W W Norton, 1938.

[25] J.B.S. Haldane. The interaction of nature and nurture. *Ann. Eugen.* 13: 197–205, 1947.

[26] D.S. Falconer. The problem of environment and selection. *Amer. Natural.* 86: 293–298, 1952.

[27] D.J. Hunter. Gene-environment interactions in human diseases. *Nat. Rev. Genet.* 6(4): 287–298, 2005.

[28] C.M. Ulrich, E. Kampman, J. Bigler, C. Chen, et al. Colorectal adenomas and the C677T MTHFR polymorphism: evidence for gene-environment interaction? *Cancer Epidemiol. Biomarkers Prev.* 8(8): 659–668, 1999.

[29] L.A. Mai. Genetic and exposure risks for chronic beryllium disease. *Clin. Chest Med.* 23: 827–839, 2002.

[30] J.L. Ree. The genetics of sun sensivity in humances. *Am. J. Hum. Genet.* 75: 739–751, 2004.

[31] L.G. Costa and D.L. Eaton. *Gene-Environment Interactions: Fundamentals of Ecogenetics.* Hoboken, NJ: John Wiley & Sons, 2006.

[32] A. Caspi and T.E. Moffitt. Gene-environment interactions in psychiatry: joining forces with neuroscience. *Nat. Rev. Neurosci.* 7(7): 583–590, 2006.

[33] L.W. Hahn, M.D. Ritchie and J.H. Moore. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* 19(3): 376–382, 2003.

[34] S.W. Guo. Gene-environment interaction and the mapping of complex traits: some statistical models and their implications. *Hum. Hered.* 50(5): 286–303, 2000.

[35] N. Chatterjee and R.J. Carroll. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* 92: 399–418, 2005.

[36] A. Maity, R.J. Carroll, E. Mammen and N. Chatterjee. Testing in semiparametric models with interaction, with applications to gene-environment interactions. *J. R. Statist. Soc. B* 71: 75–96, 2009.

[37] S.J. Ma, L.J. Yang, R. Romero and Y.H. Cui. Varying coefficient models for gene-environment interaction: a non-linear look. *Bioinformatics* 27(15): 2119–2126, 2011.

[38] K. Wang, M.Y. Li and H. Hakonarson. Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.* 11, 843–854, 2010.

[39] Y.H. Cui, G.L. Kang, K.L. Sun, R. Romero, M. Qian, and W.J. Fu. Gene-centric genomewide association study via entropy. *Genetics* 179: 637–650, 2008.

[40] C. Wu and Y.H. Cui. Boosting signals in gene-based association studies via efficient SNP selection. *Brief. Bioinform.* in press, 2013.

[41] J.Z. Liu, A.F. McRae, D.R. Nyholt, S.E. Medland, et al. A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.* 87(1):139–145, 2010.

[42] P. Zimmet, K. G. M. M. Alberti and J. Shaw. Global and societal implications of the diabetes epidemic. *Nature* 414: 782–787, 2001.

[43] C.J. Patel, R. Chen, K. Kodama, J.P. Ioannidis and A.J. Butte. Systematic identification of interaction effects between genome- and environment-wide associations in type 2 diabetes mellitus. *Hum. Genet.* 132: 495–508, 2013.

[44] A.P. Feinberg. Phenotypic plasticity and the epigenetics of human disease. *Nature* 447: 433–440, 2004.

[45] M. Peacock, C.H. Turner, M.J. Econs and T. Foroud. Genetics of osteoporosis. *Endocr. Rev.* 23:303-326. 2002

[46] D.B. Sparrow, G. Chapman, A.J.Smith, M.Z. Matter, J.A. Major, et al. A mechanism for gene-environment interaction in the etiology of congenital scoliosis. *Cell* 149: 295-306, 2012.

[47] V.S. Laitala, J. Kaprio and K. Silventoinen. Genetics of coffee consumption and its stability. *Addiction* 103: 2054–2061, 2008.

[48] J.A. Martinez, M.S. Corbalan, A. Sanchez-Villegas, L. Forga, et al. Obesity risk is associated with carbohydrate intake in women carrying the Gln27Glu beta2-adrenoceptor polymorphism. *J. Nutr.* 133: 2549–2554, 2003.

[48] B. Mukherjee, J. Ahn, S.B. Gruber and N. Chatterjee. Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. *Am. J. Epidemiol.* 175: 177–190, 2012.

[49] J.Q.Fan and W. Zhang. Statistical methods with varying coefficient models. *Stat. Interface* 1: 179–195, 2008.

[50] J.H. Huang, C. Wu and L. Zhou. Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica* 14: 763–788, 2004.

[51] J.Z. Huang, C.O. Wu and L. Zhou. Varying-coefficient models and basis function approximation for the analysis of repeated measurements. *Biometrika* 89: 111–128, 2002.

[52] J.R.B. Perry, B.F. Voight, L. Yengo, N. Amin, J. Dupuis, et al. Stratifying type 2 diabetes cases by BMI identifies genetic risk variants in LAMA1 and enrichment for risk variants in lean compared to obese cases. *PLoS Genet.* 8(5):e1002741. doi:10.1371/journal.pgen.1002741. 2012.

[53] M.C. Cornelis, A. Agrawal, J.W. Cole, N.N. Hansel, K.C. Barnes, et al. The Gene, Environment Association Studies consortium (GENEVA): maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. *Genet. Epidemiol.* 34: 364–372, 2010.

[54] G.A. Colditz and S.E. Hankinson. The Nurse's Health Study: lifestyle and health among women. *Nat. Rev. Cancer* 5: 388–396, 2005.

[55] E.B. Rimm, E.L. Giovannucci, W.C. Willett, G.A. Colditz, A. Ascherio, B. Rosner and M.J. Stampfer. Prospective study of alcohol consumption and risk of coronary disease in men. *Lancet* 338: 464–468, 1991.

[56] T.L. Holbrook, E. Barrett-Connor and D.L. Wingard. The association of lifetime weight and weight control patterns with diabetes among men and women in an adult community. *Int. J. Obes.* 13: 723–729, 1989.

[57] V.J. Carey, E.E. Walters, G.A. Colditz, C.G. Solomon et al. Body fat distribution and risk of non-insulin-dependent diabetes mellitus in women. The Nurses' Health Study. *Am. J. Epidemiol.* 145: 614–619, 1997.

[58] J.M. Chan, E.B. Rimm, G.A. Colditz, M.J. Stampfer, W.C. Willett. Obesity, fat distribution, and weight gain as risk factors for clinical diabetes in men. *Diabetes Care* 17: 961–969, 1994.

[59] S.F. Grant, G. Thorleifsson, I. Reynisdottir, R. Benediktsson, A. Manolescu, et al. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat. Genet.* 38: 320–323, 2006.

[60] T. Jin and L. Liu. The Wnt signaling pathway effector TCF7L2 and type 2 diabetes mellitus. *Mol. Endocrinol.* 22: 2383–2392, 2008.

[61] M. I. McCarthy. Genomics, Type 2 Diabetes, and Obesity.*N. Engl. J. Med.* 363: 2339–2350, 2010.

[62] L. Qi and Y.A. Cho. Gene-environment interaction and obesity. *Nutr. Rev.* 66: 684–694, 2008.

[63] L. Liu, Y. Li, T.O. Tollefsbol. Gene-environment interactions and epigenetic basis of human diseases. *Curr. Issues Mol. Biol.* 10: 25–36, 2008.

[64] A.P. Feinberg and R.A. Irizarry. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc. Natl. Acad. Sci. USA* 107: 1757–1764, 2010.

[65] M.C. Cornelis, E.J. Tchetgen,L.M. Liang, L. Qi,N. Chatterjee, F.B. Hu, and P. Kraft. Gene-environment interactions in genome-wide association studies: A comparative study of tests applied to empirical studies of type 2 diabetes. *Am. J. Epidemiol.* 175(3):191–202, 2011.

[66] O. Vaccaro, M. Boemi, F. Cavalot, P. De Feo, R. Miccoli, et al. The clinical reality of guidelines for primary prevention of cardiovascular disease in type 2 diabetes in Italy. *Atherosclerosis* 198: 396–402, 2008.

[67] A.L. Price, N.J. Patterson, R.M. Plenge, M.E. Weinblatt, et al. Principal components analysis corrects for stratification in genome-wide association. *Nat. Genet.* 38: 904–909, 2006.

[68] C. Wu and Y.H. Cui. A novel method for identifying nonlinear gene-environment interactions in case-control association studies. (Under review) 2013.

[69] N. Chatterjee and R.J. Carroll. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* 92:399–418, 2005.

[70] A. Maity, R.J. Carrol, E. Mammen and N. Chatterjee. Testing in semiparametric models with interaction, with applications to gene-environment interactions. *J. R. Statist. Soc. B* 71: 75–96, 2009.

[71] D.J. Schaid, J.P. Sinnwell, G.D. Jenkins, et al. Using the gene ontology to scan multilevel gene sets for associations in genome wide association studies. *Genet. Epidemiol.* 36: 3-16, 2012.

[72] B. Efron, R. Tibshirani. On testing the significance of sets of genes. *Ann. Appl. Stat.* 1: 107-129, 2007.

[73] Y.L. Tang, H.X. Wang, Z.Y. Zhu and X.Y. Song. A unified variable selection approach for varying coefficient models. *Statist. Sinica* 22: 601–628, 2012.

[74] T. Hastie and R. Tibshirani. Varying-coefficient models. *J. R. Statist. Soc. B* 55: 757–796, 1993.

[75] J.Q. Fan and R.Z. Li. Variable selection via nonconcave penzlied likelihood and its oracle properties. *J. Amer. Stat. Assoc.* 96: 1348–1360, 2001.

[76] G. Schwarz. Estimating the dimension of a model. *Ann. Stat.* 6: 461–464, 1978.

[77] M.O. Kim. Quantile regression with varying coefficients. *Ann. Stat.* 35: 92-108, 2007.

[78] L.F. Wang, H.Z. Li and J.Z. Huang. Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Amer. Stat. Assoc.* 103: 1556–1569, 2008.

[79] L.L. Schumaker. Spline Functions: basic theory. Wiley, New York. 1981.

[80] J. S. Rawlings, K. M. Rosler and D.A. Harrison. The JAK/STAT signaling pathway. *J. Cell Sci.* 117: 1281–1283, 2004.

[81] B. Mukherjee, J. Ahn, S.B. Gruber, and N. Chatterjee. Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. *Am. J. Epidemiol.* 175: 177–190, 2012.

[82] B.M. Neale and P.C.Sham. The future of association studies: Gene-based analysis and replication. *Am. J. Hum. Genet.* 75: 353-362, 2004.

[83] H.S. Wang and Y.C. Xia. Shrinkage Estimation of the Varying Coefficient Model. *J. Amer. Stat. Assoc.* 104: 747–757, 2009.

[84] C.L. Leng. A simple approach for varying-coefficient model selection. *J. Stat. Plan. Inf.*, 139: 2138–2146, 2138

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

[86] F.R. Wei, J. Huang and H.Z. Li. Variable selection and estimation in high-dimensional varying coefficient models. *Stat. Sinica* 21: 1515–1540, 2011.

[87] P. Breheny and J. Huang. Group descent algorithms for nonconvex penalized linear and losistic regression models with grouped predictors. (Under review). 2013

[88] J. Friedman, T. Hastie, H. Höfling and R. Tibshirani. Pathwise Coordinate Optimization. *Ann. Appl. Stat.*, 1: 302–332, 2007.

[89] T.T. Wu and K. Lange. Coordinate Descent Algorithms for Lasso Penalized Regression. *Ann. Appl. Stat.* 1: 224–244. 2008.

[90] J. Friedman, T. Hastie and R. Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 1–22. 2010

[91] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.* 58(1): 267–288, 1996.

[92] H. Zou. The adaptive lasso and its oracle properties. *J. Amer. Stat. Assoc.* 101(476): 1418–1429, 2006.

[93] C. Wu and Y.H. Cui. High dimensional variable selection in gene-environment interactions. (Manuscript) 2013.

[94] H. Zou and R.Z. Li. One-step Sparse Estimates in Nonconcave Penalized Likelihood Models.*Ann. Stat.* 36(4): 1509–1533, 2008.

[95] P. Breheny and J. Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.* 5(1): 232–253, 2011.

[96] D. L. Donoho and I. M. Johnstone. Ideal denoising in an orthonormal basis chosen from a library of bases. *Biometrika* 81: 425–455, 1994.

[97] X. T. Shen, W. Pan and Y.Z. Zhu. Likelihood-based selection and sharp parameter estimation. *J. Amer. Stat. Assoc.* 107: 223–232, 2012.

[98] L. Xue and A. Qu. Variable selection in high-dimensional varying-coefficient models with global optimality. *J. Mach. Learn. Res.* 13: 1973–1998, 2012.

[99] J.Q. Fan and J.C. Lv. Sure independence screening for ultra-high dimensional feature space. (with discussion) *J. R. Statist. Soc. B*, 70: 849–911, 2008.