ABSTRACT

ZHAO, GUOLIN. Assessing Complex Genetic Effects Using Variance Component Based Marker-set Methods. (Under the direction of Dr. Jung-Ying Tzeng and Dr. Daowen Zhang.)

Complex genetic traits have a multi-factorial etiology that involves genetic susceptibility and its interaction with environmental exposures. Because complex traits are polygenic, variants across allelic spectrum may function together to confer disease risk. Effects of individual mutations are heterogeneous, non-specific, and of low individual impact due to modest effect sizes or rarity of the variants. Given the power advantage brought by marker-set methods, we present variance component based marker-set methods to examine either gene-environmental interaction or multi-gene effects. In addition, we show the connections between variance component models and similarity-based regression models. The genetic similarity information of the marker-set within each gene summarized by collapsing the similarity status across multiple markers is incorporated into the variance-covariance matrix for the random effects in the generalized linear mixed models. In this research work, two projects are of interest and presented.

In the first project, we introduce a generalized linear mixed model testing for gene-environment effects on binary traits. In addition, we show the connection between the proposed generalized linear mixed model and a gene-trait similarity regression model. By showing the equivalence, we could take the benefit from the variance-component methods and the similarity collapsing methods. The generalized linear mixed models have a systematic theoretical foundation and flexibility in inference and the similarity collapsing methods could gain power when the genetic variants have nonlinear effects with diverse effect sizes. We construct a score test for examining the gene-environment interactions, and apply an EM algorithm to estimate the nuisance variance components. The challenge of examining GxE interaction is that it may result poor statistical power due to the sample size; and standard regression-based methods using single-SNP technique may cause inflation on the test statistics. To illustrate the utility of the proposed method, we conduct simulations and real data applications on both Genome-wide association

study data and next generation sequencing data.

In the second project, we extend the association analyses from gene level to the level of gene-set. To explore more multi-genic biological structure for marker-sets and investigate undiscovered biological mechanism for complex diseases, we introduce a variance component based method to conduct gene-set association analyses on quantitative traits. A linear mixed model is proposed where genetic information is collapsed at similarity level. This method incorporates a variable selection technique on the gene-specific variance components to assess the overall network effect and to identify significant genes associated with the disease from a gene-set.

Assessing Complex Genetic Effects Using Variance Component Based Marker-set Methods

by

Guolin Zhao

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2013

APPROVED BY:

Dr. Wenbin Lu	Dr. Dahlia Nielsen
Dr. Jung-Ying Tzeng Co-chair of Advisory Committee	Dr. Daowen Zhang Co-chair of Advisory Committee

DEDICATION

To my beloved family

BIOGRAPHY

The author was born in a beautiful city Hangzhou, China, and spent 22 years in her hometown. She graduated from Hangzhou No. 2 High School and attended Zhejiang University of Technology afterwards for her undergraduate study in Information & Computational Science. After obtaining her Bachelor's degree in 2006, she came to United States and studied Applied Statistics at the University of North Carolina at Greensboro. In August 2008, she graduated with a M.A. degree in Mathematics and in the same year she joined the Statistic Department at North Carolina State University to pursue her Ph.D study. Her doctoral dissertation research is under the direction of Dr. Daowen Zhang and Dr. Jung-Ying Tzeng and she will complete her Ph.D. in August 2013.

ACKNOWLEDGEMENTS

I would especially like to express my gratitude to my advisors Dr. Jung-Ying Tzeng and Dr. Daowen Zhang for their guidance, patience and encouragement throughout the years working on the research and writing of this dissertation. I greatly appreciate the time and effort that both of them dedicated to helping me complete the research study smoothly.

I would like to thank my committee members Dr. Wenbin Lu, Dr. Dahlia Nielsen and Dr. Jeffrey Thorne constructive comments and suggestions for my research.

I would also like to thank all the Directors of Graduate Programs (DGP) Dr. Pam Arroway, Dr. Jacqueline M. Hughes-Oliver, Dr. Sujit K Ghosh and Dr. John F Monahan. It is their assistance, effort and encouragement to make my graduate study life meaningful. I am also grateful to Dr. Howard Bondell and Dr. Jason Osborne for their guidance on the course work and help on the research work. I am grateful to all the faculty members, staffs, fellow students and friends who I know during the years that ever helped.

In addition, I am grateful to Dr. Timothy Frayling and Dr. William Rayner and members of the Warren 2 Consortium for providing us the BMI data. This work makes use of data generated by the Wellcome Trust Case Control Consortium (WTCCC). A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk. I am also thankful to Dr. Michael Wu for helping with the simulation design. I am also appreciative to Dr. Mathieu Firmann and Dr. Vladimir Mayor and members in the CoLaus study for providing the CoLaus study data.

Finally, my grateful thanks to my best parents, Weizu Zhao and Zeli Han, who always believe in me and give me the courage to pursue my dreams.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Genetic Associations of Complex Diseases

Complex diseases are multi-factorial disorders for which disease susceptibility, disease development and treatment response are mediated by intricate genetic and environmental factors. Many complex common diseases have drawn public attentions, such as obesity, diabetes, cardiovascular diseases, mental disorders and cancers. Take the type 2 diabetes as an example, more than 347 million people are now suffering from this disease worldwide [1]. On the environmental sides, factors such as family history, daily diet, smoking status, physical activity and body weight interactively affect the disease development. On the genetic sides, at least 18 genes have been discovered to be associated with type 2 diabetes even though only 6% of the heritability is explained by these genes [2]. Another complex disease example is asthma. It is the most common chronic disease that irritates the airways connecting to the lungs, and its onset usually happens early in childhood. Nowadays, around 235 million people worldwide have asthma [3], and over 22 million cases are in the United States [2]. Genetic variants, respiratory infections and living environments all contribute to the disease onset: over 100 candidate genes have been reported to be associated with asthma among which FLG was the first identified gene that leads to asthma [4]. The discovery that FLG mutations have influence on asthma is

innovative because it is the first time that the skin defense system is demonstrated to be as important as the immune system in allergic pathways [5].

Recent genome-wide association studies (GWAS) have revealed a polygenic basis for complex disorders. Both common and rare variants are involved, and individual genetic effects tend to be moderate due to low individual impact or rarity of the mutation. Numerous methods have been proposed to perform association analysis at gene, protein, pathway or network level. These approaches differ in the way on how the genetic information is aggregated to examine the association. Details are reviewed in the subsequent sections.

1.2 Gene-Level Association Analyses

When multiple single-nucleotide polymorphisms (SNPs) in a gene are genotyped to assess the complex genetic effect, both SNP-based and gene-based analyses could be considered to test whether the gene is associated with the disease.

In SNP-based methods, the effect of each SNP is assessed one at a time. The major tasks in the SNP-based methods are how to combine the individual test results at gene level and how to account for multiple testing. The minimum p-value method [6] and the Fisher's combined method [7] are commonly used to combine individual test results. To account for multiple testing, the adjusted threshold, α_i^*, are calculated by certain procedures to ensure the family-wise error rate is under control. For example, Bonferroni correction gives $\alpha_i^* = \alpha/M$, where M is the number of SNPs in the gene and α is the family-wise error rate. Sidak correction gives $\alpha_i^* = 1 - (1-\alpha)^M$. In order to improve the power, efforts have been put into obtaining the effective number number of independent tests k_{eff}. The minimum p-value method applies an improved correction based on the Sidak adjustment in which M is replaced by k_{eff} [6]. Many SNPs have been identified for some common diseases by using single-SNP methods. However, to detect the subtle SNP effect at the extremely small significant level, a large sample size is always needed. In addition, the problems caused by correlation among the SNPs may not be eliminated and there still may be non-negligible power loss by using single-locus approaches.

Alternatively, gene-based methods that analyze a group of SNPs simultaneously are considered, and these approaches are also referred to as "marker-set" or "multi-locus" methods. Even though each variant may have moderate effect, the overall effect of multiple loci from the gene could be clinically significant. Based on how the genetic information is collapsed across loci, marker-set methods can roughly be categorized into two types [8]: genotype collapsing methods and similarity collapsing methods.

Genotype collapsing methods combine the multi-locus information by first obtaining a weighted or unweighted sum of the individual genotypes across markers, and then assess the multi-locus effect based on the genotype sum. For example, PCA-based methods [9, 10] compute the principal components (PCs) of the genotypes and then use the PCs to assess the association between the disease trait and the genetic information. In rare variant analysis, due to the sparseness of mutations, the "newer" generations of genotype collapsing methods incorporate the minor allelic frequencies in the weights [11, 12, 13]. Specifically, these approaches aim to assess the overall genetic burden due to rare variants, which are in essence weighted genotype sums. Methods of this type, such as the counting based burden test (sum test) [14] and weighted sum test [12], implicitly assume all the variants affecting the disease risk in the same direction and thus may lose power when the assumption is violated [15, 16, 17]. The genotype collapsing methods are recommended when additive genetic effects with similar effect size are present among the markers [8].

Similarity collapsing methods combine genetic information by calculating the genetic similarity for each pair of unrelated individuals at each locus, and then obtain the multi-locus similarity scores by taking a weighted or unweighted similarity sum across all the markers. Some representative methods in this category include the kernel machine regression [18, 19, 20], the gene-trait similarity regression [21, 22], and the random-effect model [23, 24]. Though each of the three methods was derived based on rationales and appear differently, all these methods have been shown to be connected with each other [19, 25, 26, 20, 21, 22], which we discussed further in Section 1.2.1.

The kernel machines-based methods [18, 25, 27, 19, 20, 17, 28] incorporate the pairwise genetic similarity via kernel functions. Some commonly used kernel functions include (a) the identity by state (IBS) kernel [29], which is the number of alleles shared by state and is recommended when nonlinearity or interactive effect is present; (b) the linear kernels [19, 30], which offers higher power if only linear additive effect is present among the SNPs [28]; and (c) the polynomial kernel, which is considered when SNPs within a small window have local correlations due to linkage disequilibrium (LD) but no long-range correlations among the SNPs [27].

The idea of the gene-trait similarity regression[21, 22] is inspired by Haseman-Elston regression [31] from linkage analysis and haplotype similarity tests [32] for regional association. It takes pairs of individuals and regresses their trait similarity on genetic similarity of a marker set. If the markers are associated with the trait, the slope should be non-zero.

Finally, the random effect methods [23, 24] incorporate genetic similarity information into the variance-covariance structures of the genetic effects. The gene-trait associations could be studied by testing the variance components in these random effect models.

Compared to the genotype-collapsing methods which detect the association by examining the mean genetic effects, similarity collapsing approaches detect the association by examining the variance of the genetic effects. Similarity collapsing methods have higher power when nonlinear or interactive effects are present among markers with different effect sizes and are more robust under complex genetic architectures [8].

1.2.1 Connections among different similarity-collapsing methods

The methods developed in this dissertation fall in the similarity-collapsing category and detect association by assessing significance of the appropriate variance-components. Below we discuss the variance component models in details and further illustrate how the different similarity collapsing methods can be connected under the generalized linear mixed-effect model (GLMM).

Consider a GLMM which includes fixed covariates effect and genetic random effect:

$$g(\mu) = X\gamma + G, \tag{1.1}$$

where $g(\cdot)$ is the link function connecting the conditional mean vector μ of the complex trait vector Y and the explanatory variables or covariates, which forms the design matrix X with coefficient vector γ, and $G \sim N(0, \Sigma_G)$ is the matrix of genetic effects for N individuals in the sample where Σ_G is a $N \times N$ nonnegative definite variance-covariance matrix which incorporates the genetic information.

If one sets G_i as $h(g_i^1, \ldots, g_i^M)$ for some function, where $g_i^m,\ m = 1, \cdots, M$ are the genotypes of locus m for the individual i and $\Sigma_G = \tau K$ where K is defined using different kernel functions (e.g., linear kernel or weighted IBS kernel [19, 20]), then the test of $H_0 : \tau = 0$ is equivalent to examining $H_0 : h(\mathbf{g}) = 0$ in the kernel machines regression where $\mathbf{g}$ includes all the genotypes for N individuals on M loci [18, 19].

The gene-trait similarity regression model is given as

$$E(T_{ij}|X, S) = c \times S_{ij},\ i \neq j,$$

where T_{ij} is the trait similarity for individuals i and j and S_{ij} is the pairwise genetic similarity. If one sets T_{ij} as the weighted covariance between Y_i and Y_j and lets Σ_G in Equation (1.1) be specified as τS with $S = \{S_{ij}\}$, then it can be shown that $E(T_{ij}|X, S) \approx \tau S_{ij}$ using the optimal weights determined by the distribution of trait vector Y in calculating the conditional trait similarity expectations. Thus, the variance-component test $H_0 : \tau = 0$ is equivalent to the significance test of the regression coefficient in the gene-similarity regression, i.e., $H_0 : c = 0$ [21].

Finally, let $G = H^T\beta$, where H is an $N \times r$ haplotype design matrix with r being the number of distinct haplotypes formed by the M loci, and β is the haplotype random effect. If one set Σ_G in Equation (1.1) as $\tau H^T R_\beta H$ where R_β is a $r \times r$ similarity matrix that quantifies

the similarity between different haplotypes, then the variance component test of $H_0 : \tau = 0$ is equivalent to haplotype-based association test proposed in Tzeng and Zhang [24].

With the connections between variance-component based methods and the similarity collapsing methods, we see the variance component based methods cover a range of similarity-based approaches, and the key lies in how the variance-covariance structure, Σ_G, is specified. The similarity based methods have been shown to have superior power when genetic variants have nonlinear effects with various effect size [8]. In addition, the variance component based methods have well-established theoretical foundation and flexibility in inference. Thus, such connection between these two types of methods makes the variance component based approaches as attractive analytical tools for associations.

1.2.2 Methods for testing Gene-Environment Interaction

For multi-factorial complex disorders, gene-environment interaction (GxE) studies can facilitate the understanding of genetic heterogeneity under different environmental exposures [33, 34]. They help to identify high-susceptible subgroups in the population [35], provide insights into biological mechanisms for complex diseases [35, 36, 37], and improve the ability to discover susceptible genes that interact with other factors but exhibit little marginal effect [37]. In the post genome-wide association studies (GWAS) era, genome-wide GxE studies are undertaken to mine the existing GWAS data, with an aim to uncover the missing inherited risks. In pharmacogenetics, gene-by-treatment analyses are conducted to understand the inter-individual variations in drug responses, with an aim to suggest personalized prevention, detection and treatment regimes.

Just as methods in detecting genetic main effects, the GxE effects can be assessed by either SNP-based or gene-based approaches [38]. Because SNP-based approaches assess the GxE effect by testing the interaction of each SNP and an environmental factor individually under each hypothesis, adjustment in the significance level is required to control the overall type I error rate. These SNP-based methods using various multiple testing corrections are the classical and

standard SNP-based approaches to test the associations. Modifications are developed based on the classical SNP-methods for case-control study to correct the biases caused by the violations for the assumptions of either independence between gene-environment factors or Hardy Weinberg Equilibrium (HWE) [39]. Such modifications include adapting logistic regression models with some quadratic penalization [40], and applying an empirical Bayes method to derive a shrinkage estimator [39] under a semiparametric maximum-likelihood framework [41].

There are limitations with the SNP-based approaches. One major issue, as what encountered in the main-effect tests, is the correlations among SNPs. In the analysis testing for the GxE interaction, the multiple tests for SNP-environment interactions could be more dependent because the same environmental effect is shared in the interaction terms under each hypothesis [42]. Potential efficiency loss is another challenge if a large amount of SNPs are present. Thus, marker-set methods could be considered as a better alternative for GxE tests in presence of multiple SNPs.

The same two major types of collapsing methods, genotype collapsing and similarity collapsing, can be applied to test the GxE interaction. Several genotype collapsing methods testing for the GxE interactions are simply the extensions for the same type of methods testing for the genetic main effects. In addition, new approaches are also developed for assessing the interactions, such as the approach incorporating all the genotypes into a latent variable. Then the interaction is tested under Tukey's 1 degree-of-freedom models where the latent variable interacted with the environmental exposures [43].

Because of the power gains shown in the similarity collapsing methods testing for the genetic main effects under complex genetic structures, more efforts are being contributed under this category for assessing GxE interactions. A gene-trait similarity regression model is introduced to test the interactions between multiple SNPs within a candidate gene and an environmental exposure for quantitative traits [22]. Other random effect methods have also been introduced. For example, the gene-environment set association test (GESAT) [42] incorporates the gene-environment effect into the variance-covariance matrix where interaction is examined on the

corresponding variance component.

In addition, methods under Bayesian frameworks are also developed for examining the GxE effects such as a resampling-based test which identify a latent genetic profile variable to classify the genotypes into different clusters such that individuals sharing the same genetic profile category share the same risk model [38].

1.2.3 Topics Addressed in This Dissertation

In Chapter 2 of this dissertation, we developed variance component based methods for examining the GxE effects on binary traits, commonly encountered traits in genetic applications. While methods exist for testing the genetic main effects on binary traits [21] or for testing GxE effects on quantitative traits [22], the direct extensions to GxE binary traits is not straightforward. There are several challenges with respect to computational and estimation issues especially with large sample data.

First, the GxE effect is assessed by testing a variance component corresponding to the GxE effects. Most commonly used methods for hypothesis testing, such as the Wald test and the likelihood ratio test, are not valid under this situation because the variance components to be tested are on the boundary of the parameter space. To conquer these challenges, we consider score-type of test statistics for assessing GxE interaction since score statistics usually have stable statistical properties when the variance component tested at the boundary points.

Second, when assessing the GxE effect using a score statistic, there exists one nuisance variance component that corresponds to the genetic main effect and needs to be estimated under H_0. For binary traits, the estimation procedure involves a high-dimensional integration that lacks a closed form. To obtain an acceptable variance component estimate, we propose an approximate EM algorithm to estimate the nuisance variance components and other model parameters.

Third, even with the proposed EM algorithm, we still have to invert matrices that are of the same magnitude as the sample size and such inversion may not be feasible for studies with

large samples. To overcome this problem, we considered matrix decompositions in modifying the proposed generalized linear mixed model where the majority of the information is still maintained. Then the proposed approximate EM algorithm proceeds for the reduced GLMM.

1.3 Gene-Set Association Analyses

Because genes function together within biological modules such as pathways and networks, demands in the methodological development of association analyses have shifted from SNP level and gene level to gene-set level in order to facilitate understanding of the biological mechanisms for complex diseases. Gene-set association analyses (GSAA) evaluate the significance of a pre-defined gene set (such as a biological pathway) with the trait values. Gene-set analyses were once primarily applied on the gene expression studies [44, 45, 46, 47, 48, 49, 50] and have now been widely used as a complementary tool to mine GWAS data.

There are two types of statistical hypotheses evaluated under GSAA: the competitive hypotheses and the self-contained hypotheses. A null hypothesis is a competitive one if we examine the association from a gene-set compared to other genes that are not in the pre-defined set. The well-received ALIGATOR (Association LIst Go AnnoTatOR) [51] is a representative competitive test for gene-set analyses. Our focus is on the self-contained tests. The null hypothesis is self-contained if it examines whether genes from the pre-defined gene-set have significant effect on the trait values [52]. Methods such as gene-set ridge regression in association studies (GRASS) [53] and supervised principal component analysis (SPCA) [54] are self-contained tests.

Unlike gene-set expression analysis, in GSAA, the information is measured at the marker level, and multi-marker information has to be aggregated at the gene level before the analysis can be proceeded. Many efforts have been devoted to better account for the within-gene multi-locus information and to assess association at both set (pathway) level and gene levels. For example, in the GRASS, the gene-level information is obtained by the several leading principal components (eigenSNPs) of multi-locus genotypes, and then the significant genes are identified using a

regularized regression [53]. Because of the unnecessary relation between the eigenSNPs and the phenotypes, SPCA is proposed which identifies significant genes using principal components incorporating the correlations between SNPs [54].

1.3.1 Variable Selection Methods for Testing Gene-Set Associations

To identify the genes associated with a disease under the self-contained hypothesis, many variable selection techniques can be applied to identify the significant disease-associated genes. These methods can be roughly classified into two types.

The first type is the grouped variable selection based methods, which treat genetic information within same gene as a group of variables. Regularization methods such as Group-LASSO [55, 56, 57] and group ridge regression [58] could be applied either directly on the SNPs data or on the principal components of multi-locus genotypes within each gene. GRASS is such a method in which a group ridge penalty is imposed to shrink the estimates of the genes and a lasso penalty is used to perform variable selection on the eigenSNPs within each gene [53].

The second type is the individual variable selection based methods such as stepwise selection, LASSO [59] and adaptive LASSO [60] in which genetic information within each gene is either summarized in one variable or corresponded to one variance component. Such gene-specific variables could be obtained by using genotype collapsing methods, such as the counting based burden methods using a single count to represent a single gene. Regularization penalty terms could also be imposed on gene-specific variance components, the proposed variance component based method introduced in this research work is such an example.

1.3.2 Topics Addressed in This Dissertation

Compared to single-gene analyses, genetic information from multiple genes are used to examine associations simultaneously in gene-set association analyses. Because large number of genotypes are known for each gene, ten thousands or even more genotypes are possible to be included in one gene-set association analysis, one challenge in GSAA is to incorporate multi-marker information

within each gene in an efficient way. In addition, due to the modest genetic effect and LD among the SNPs, another challenge in GSAA is to identify the disease-associated genes with subtle genetic effects but significant overall multi-gene effects [53, 54, 61].

In Chapter 3 of the dissertation, we adapt the nonparametric function selection approaches in additive regression models or generalized additive models [62, 63] to the gene-set association analysis. The genotype information of multi-markers within a gene is collapsed at similarity level and incorporated into a variance-covariance structure. The importance of the gene is characterized by a single variance component in a reduced linear mixed model. To identify significant genes associated with the disease trait from a gene-set of interest, adaptive LASSO is implemented on the gene-specific variance components under the self-contained hypothesis. One major challenge in the proposed variance component based method is to handle intensive data where genotypes of tens to hundreds SNPs are known for each gene from thousands of individuals. The pairwise genetic similarity for each gene is calculated which formulate multiple high-dimension genetic similarity matrices, thus matrix decompositions in addition to EM algorithms are used to reduce the dimensions in estimating the nuisance parameters.

Chapter 2

Association Tests for Gene-environment Interaction on Binary Traits using Generalized Linear Mixed Models

2.1 Introduction

Complex diseases are defined as disorders not determined by a sole gene but by multiple genes where those gene factors only partially contribute to the development of diseases. Besides genetic effects, environmental factors are also involved in the determination of the disease risk, mostly are involved interplaying with the genetic susceptibility. To discover the underlying genetic and environmental causes linked to complex forms of human disease, examining gene-environment interaction (denoted by "GxE") is one dynamic area where researchers are taking great effort to explore the mysterious patterns and structures. By studying the GxE interactions, we aim to understand the biological mechanisms of complex diseases [36] and meanwhile to improve both the performance of predicting the disease risk and the ability of detecting the benefit in

changing the modifiable environment [37].

Hundreds of genes have been identified to be significantly associated with multi-factorial disorders by using numerous single-SNP methods developed for genome-wide association studies. Such discoveries encourage researchers to continue their contributions in exploring the underlying genetic and environmental effects associated with complex diseases. Even though it was a great success brought by those single-SNP methods in GWAS, there are even larger part of the trait heritability is remained unexplained and undiscovered [64]. To explore more proportion of trait heritability, approaches assessing the association on multiple markers may be of desire because genetic variants may function together and confer the disease risk at different genetic levels simultaneously. Thus, the overall effect of these variants is more possible to be significant in detecting the association between the complex disease and traits if aggregating at the correct level [65].

With the development of resequencing technology, thousands or millions of rare variants are genotyped. These rare variants may have subtle effects individually or occur sparsely which have a more recent origin [66]. According to this fact, rare variants should not be ignored in association analyses for complex diseases. Due to the sparse mutations, significance is barely detected by using the single-SNP methods in the presence of rare variants. Additionally, these methods mostly ignore the correlation among the markers, which may lead to the consequence that the LD structure for the markers is automatically not taken into account. Therefore, multi-locus approaches are considered as major methods in association studies for genetic variants with low minor allele frequencies.

Because of the multi-factorial etiology involving the interplay of genetic susceptibility and environmental exposures for complex diseases, modeling GxE interaction or modeling gene main effect and GxE interaction jointly becomes a big challenge [36]. One major challenge in GxE studies is the poor statistical power [36] because it typically requires at least four times as large as sample size to detect a GxE effect than a main effect of similar magnitude [67]. As a result, not many major findings have been reported in GWAS for GxE studies and the findings

are often lack of replications by using the single-SNP methods. Abundant marker-set methods have been proposed for genetic main effect analyses to enhance detecting power [8], and we argue that the same strategies should be considered for GxE studies. First, GxE with G being genes, pathways or functions, provides a biologically sensible way to assess jointly whether the gene/pathway effects are modified under different environmental exposures. Second, assessment of the effect G, E and GxE individually often requires large number of parameters, and when the analyses are conducted at gene or pathway level, the dimensionality is even higher. In contrast, marker-set methods are able to perform efficient analysis by summarizing high dimensional information using smaller number of parameters and by aggregating moderate signals across multiple factors. Therefore, analyses using the information based on a marker-set should be advocated and be developed for examining the GxE interaction where multiple markers and environment factors are presented to cause the complex diseases. It is also the reason that analyses based on marker-set have drawn attention in GWAS and next generation sequencing research studies [8]. Attracted by these advantages, several methods have been developed by applying the similarity collapsing technique which included the kernel machine approaches [18, 19, 20] and similarity-based approaches [22, 21].

Specifically in this research study, we focus on examining the GxE interaction effect on binary traits by using the multi-marker analysis. We propose a generalized linear mixed model (GLMM) and conduct score tests on both the GxE interaction and joint effects. Because of the flexibility in GLMM, covariates such as treatment, population subgroup could be included as baseline information to adjust in the model. In the proposed model, the genetic information is collapsed at the similarity level which is incorporated into the variance-covariance matrix for the random effects. Compared to methods available for testing the genetic main effect [21] or GxE interaction effect for quantitative trait [22], it has several challenges with respect to computation and estimation issues involving high dimensional data. First, inverting a high dimensional similarity matrix may be required to estimate the variance components in order to calculate the test statistic when both the genetic main and GxE interaction effects are in

the models. Though the same problem is also present for quantitative traits, such problem is bypassed by the properties of the normal distribution. In models where the genetic main effect is of interest with no GxE interaction, only fixed effects are in the model under the null hypothesis. Thus, there is no need to estimate the nuisance variance components when testing the genetic main effect. In that situation, a generalized linear model such as the logistic linear regression model is sufficient to estimate the coefficients for the fixed effects in order to compute the test statistic. Similarly, when the genetic main effect and gene-environment interaction effect are both of interest, the problem will also simplify because there is no nuisance parameters in the model under the null hypothesis. Second, in order to estimate the variance component, a high dimensional integration may be involved where in cases it is impossible or extremely complicated to derive the integration. To avoid such a trouble, we use the approximate EM algorithm to approximate the integration. Third, most common methods testing the association, such as the Wald test and the likelihood ratio test, are not valid under the situations where variance components are tested at its boundary of the parameter space.

In this project, two systematic simulation studies were conducted. The first simulation is to investigate the test validity and compare the power between the proposed method and a genotype collapsing approach on a COSI based simulation data [20]. The second simulation is to study the test validity and evaluate the power by applying the proposed method and by applying a single-SNP method under a variety of scenarios on a Hapmap based simulation data. Finally, we applied our proposed method to two real data applications. The first is a next generation sequencing data collected from the CoLaus study [68] in which smoking status was treated as the environmental factors and the second is a Type 2 Diabetes GWAS data from the Wellcome Trust Case Control Consortium in which the Body Mass Index (BMI) was used as the environmental modifier.

2.2 Material and Method

2.2.1 Generalized Linear Mixed Model for Gene-Environment Effect

Generalized linear mixed models are not specifically developed for association tests, but more and more studies have shown the power by using the GLMM framework. Even though only few studies have directly proposed the generalized linear mixed models for association studies [24], those variance component based models have been shown to have the connection with various methods [19, 25, 26, 20]. For example, the connection between GLMM and similarity collapsing methods shows that the genetic similarity information among individuals could be incorporated into the variance-covariance matrix for the random effects in the GLMM. By showing the equivalence, we could take the benefit from both types of methods. The similarity collapsing method could gain power when the genetic variants have nonlinear effects with diverse effect sizes, and the generalized linear mixed models have a systematic theoretical foundation and flexibility in inference.

In this work, we propose a generalized linear mixed model incorporating both the genetic main effect and gene-environment effect to assess the gene-environmental interaction. This work is motivated by the previous research by Tzeng and Zhang [24] in which they proposed a generalized linear mixed model using the haplotype information. In the proposed GLMM, the genetic information could be either haplotype information or genotypes from a market-set. It is an extension work with a generalized form of the variance component based model that has been proposed for the quantitative traits [22]. In this work, score tests are conducted on the variance components to examine either the GxE interaction or joint effect.

To illustrate the proposed generalized linear mixed model, we define some notation. Suppose there are a total of N individuals in a study. For individuals i, $i = 1, \ldots, N$, let Y_i be the observed binary trait; $Y_i = 1$ if the individual has the disease and $Y_i = 0$ if the individual does not the disease. Let X_{Ei}^T be a $K_E \times 1$ vector containing the environmental factors for assessing gene and environmental interactions where K_E is the number of such environmental factors. Let

X_{Ci}^T be the $K_C \times 1$ design vector for the confounders and K_C is the number of such covariates. Environmental factors are not used for assessing the gene-environmental interactions are all included in the confounder vector. Let $X_i^T = (1, X_{Ci}^T, X_{Ei}^T)$ be a $(K_C + K_E + 1) \times 1$ covariate vector which includes the intercept. All covariates are standardized to mean 0 and variance 1. Let γ be the coefficient vector for the intercept and covariates. Let D be a $N \times N$ diagonal matrix with covariate of environment effect on its diagonal (i.e. $D_{ii} = X_{Ei}$ for $i = 1, \ldots, N$). Let G be the random effects containing the genetic main effects and η be the random effects for the genetic effect in the GxE interaction. Let $\mu = (\mu_1, \ldots, \mu_N)$ be the vector of the conditional means where each conditional mean $\mu_i = E(Y_i|X, S, G, \eta)$. Let S be the similarity matrix where each element calculates the genetic similarity between two individuals. There are several approaches to characterize the similarity [18, 29]. Suppose either haplotype or genotypes at M loci (plural of "locus") are known, let g_i^m for $i = 1, \ldots, N$ and $m = 1, \ldots, M$ be the genetic information for individual i at locus m. We value $g_i^m = 0$ if the genotype at a single locus is homozygous for minor alleles, $g_i^m = 1$ if the genotype is heterozygous, and $g_i^m = 2$ if the genotype is homozygous for major alleles. To calculate the genetic similarity for individual i and j, we compute the number of alleles they share purely by state at each locus, denoted as $IBS(g_i^m, g_j^m)$ [29]. That is $IBS(g_i^m, g_j^m) = 0$ if both individuals are homozygous at locus m with different SNP alleles (e.g. $g_i^m = AA$ and $g_j^m = aa$), $IBS(g_i^m, g_j^m) = 1$ if they share one allele (e.g. $g_i^m = AA$ and $g_j^m = Aa$), and $IBS(g_i^m, g_j^m) = 2$ if they are both homozygous with the same alleles (e.g., $g_i^m = AA$ and $g_j^m = AA$). In this work, similarity based on identity-by-state (IBS) allele sharing (S_{ij}^{IBS}) and similarity based on weighting by allele frequency (S_{ij}^W) are applied. The former similarity is defined as $S_{ij}^{IBS} = \sum_{m=1}^M IBS(g_i^m, g_j^m)/2M$. The latter similarity is defined as $S_{ij}^W = \sum_{m=1}^M w_m IBS(g_i^m, g_j^m) / \sum_{m=1}^M w_m$ where w_m is a positive number reflecting the weight assigned to locus m. Without prior information, a weight of $w_m = Beta(q_m; 1, 25) \propto (1 - q_m)^{24}$ was proposed in order to improve power in presence of rare variants which assigns much lower weights on non-causal common variants and higher weights on rare variants where q_m is the minor allele frequency at locus m [20]. According to the definitions, it is apparent that S_{ij}^{IBS}

calculates the unweighted similarity and S_{ij}^{W} calculates the weighted similarity incorporating the minor allele frequency information. In cases where almost all variants are common alleles, the unweighted similarity S_{ij}^{IBS} could be considered. In the presence of rare alleles, S_{ij}^{W} is preferred because it is believed that individuals who share rare alleles are more likely to share similar genomes and thus the weighted similarity might improve the power for detecting variants significantly associated with the phenotypes [29].

To illustrate the generalized linear mixed model, we consider $K_E = 1$ but the proposed work can be extended straightforwardly to $K_E > 1$. Thus, the proposed GLMM is expressed as

$$g(\mu) = X\gamma + G + D\eta, \tag{2.1}$$

where $G \sim N(0, \tau S)$ and τ is the variance component corresponding to the genetic main effect; $\eta \sim N(0, \phi S)$ and ϕ is the variance component corresponding to the GxE effect; and $g(.)$ is the link function specified as the logit of μ_i under the binary cases which is expressed as $g(\mu_i) = logit(\mu_i) = \log\{\mu_i/(1-\mu_i)\}$.

However, because of the high dimensional calculation with respect to the $N \times N$ similarity matrix S, we rewrite the Model (2.1) in a different form in order to make it more computationally efficient by reducing the dimension. All the further tests on the variance components will be conducted by applying the following model

$$g(\mu) = X\gamma + Zb + DZ\lambda, \tag{2.2}$$

where Z is a $N \times L$ matrix such that $ZZ^T = S$, $b \sim N(0, \tau I)$, $\lambda \sim N(0, \phi I)$, I is the $L \times L$ identity matrix and L is the rank of similarity matrix S. In some situations, a lot of nuisance parameters may be present if using all information in similarity matrix S. As a consequence, the power of detecting the significance of either the GxE effects or the joint effects will be compromised [69]. For that purpose, we use the principal component technique to capture the majority information in the similarity matrix, denoted by $\tilde{S}$. The eigen-decomposition is

applied on the similarity matrix such that $S = \Sigma_{l=1}^{L} \upsilon_l e_l e_l^T$ where $\upsilon_1 \geq \upsilon_2 \geq \ldots \upsilon_L \geq 0$ are the eigenvalues of S in decreasing order and e_l's $(l = 1, \ldots, L)$ are the corresponding eigenvectors. The matrix Z is then formed as $Z = [\sqrt{\upsilon_1} e_1, \ldots, \sqrt{\upsilon_L} e_L]$ which leads to $ZZ^T = S$. In order to reduce the dimensionality of Z, we identify a $p \in (0, 1)$. For this p, there exists a minimum value of L^* such that $\Sigma_{l=1}^{L^*} \upsilon_l / \Sigma_{l=1}^{L} \upsilon_l \geq p$. With an appropriate choice of p, $\tilde{S} = \Sigma_{l=1}^{L^*} \upsilon_l e_l e_l^T$ could hold the most important information and thus to approximate to the original similarity matrix S. Consequently, we denote $\tilde{Z} = [\sqrt{\upsilon_1} e_1, \ldots, \sqrt{\upsilon_L} e_{L^*}]$ such that $\tilde{Z}\tilde{Z}^T = \tilde{S}$. Thus, we expect the $\tilde{Z}$ replaced in the Model (2.2) to solve the high dimensionality problem. To investigate the patterns for the two types of data, next generation sequencing data and GWAS data, Figure 2.1 shows eigenvalues and cumulative percentages of the information explained by the first several leading eigenvalues. We could observe that for next generation sequencing data where most alleles are rare variants, the first eigenvalue could explain as high as 99.6% of the total information. On the contrary, we need much more eigenvalues to explain up to 99% of the information for GWAS data. Therefore, in this research study, the principal component technique will be considered for next generation sequencing data in the presence of rare variants.

Under Model (2.2), it is straightforward that the GxE interaction can be examined by testing the null hypothesis of $H_0 : \phi = 0$. Although the main focus of this work is to test the interaction effect of gene and environment, a test examining the "joint effects" of genetic main effect and gene-environment interaction effect simultaneously is also covered. We use the term "joint test" to represent the score test examining the joint effects for the null hypothesis of $H_0 : \tau = \phi = 0$. The purpose of latter test is to evaluate the overall genetic association. Because the joint test incorporates information about both genetic main effect and the GxE interaction, it is recommended if either genetic heterogeneity or gene-environment interactive mechanism is unknown or unsure [22].

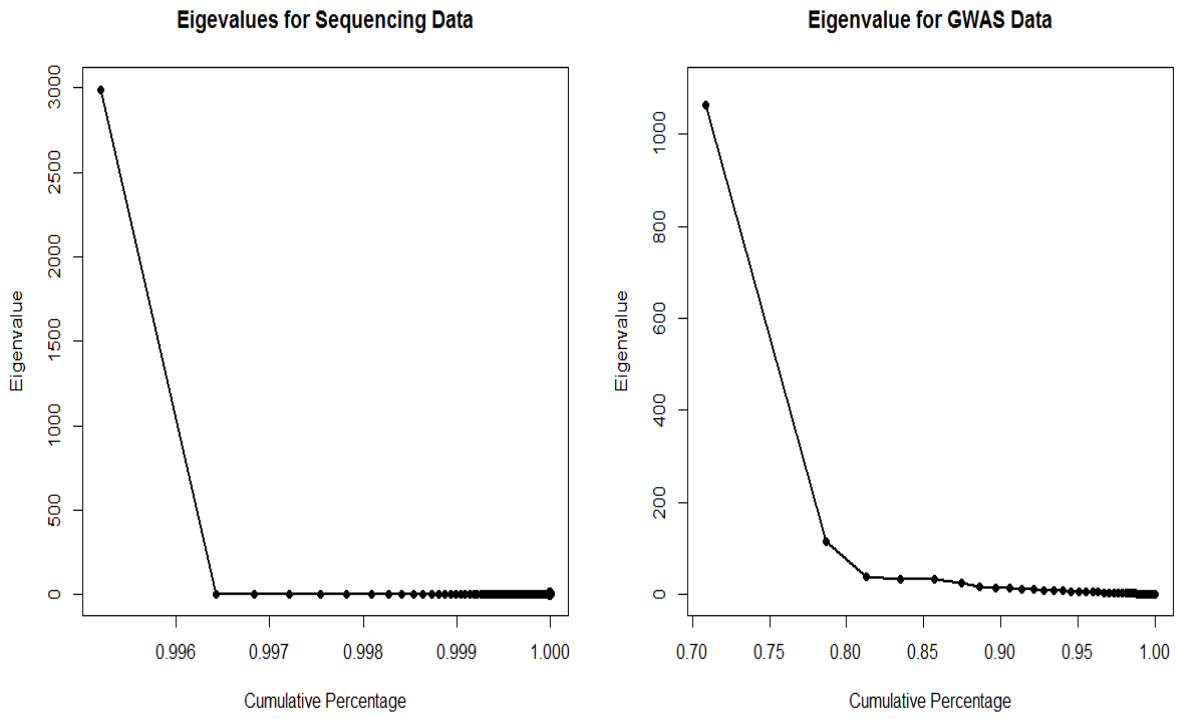

Figure 2.1: The Eigenvalues and the corresponding cumulative proportions of the total information in the similarity matrix S by using a next-generation sequence data set and a genome-wide association study data set.

2.2.2 Connection between the Generalized Linear Mixed Model and a corresponding Gene-Trait Similarity Regression Model

Motivated by the previous work showing the connection between the variance component models and gene-trait similarity regression models [22, 21], we want to show that examining the variance components in Equation (2.1) is equivalent to testing the GxE effect and the joint effect in a gene-trait similarity regression model.

Inspired by the work of Elston et al. [70] and Wang and Elston [71], a gene-trait similarity regression model including the gene-environmental interaction could be expressed as

$$E(T_{ij}|X,S) = c \times S_{ij} + d \times S_{ij} \times X_{Ei}X_{Ej},\ i \neq j, \tag{2.3}$$

where T_{ij} is the trait similarity for individuals i and j using the weighted covariance between Y_i and Y_j. Specifically, $T_{ij} = \{\omega_i(Y_i - \mu_i^0)\}\{\omega_j(Y_j - \mu_j^0)\}$ [21], where $\mu_i^0 = E(Y_i|X_i, S_{i.})$ is the subject-specific trait mean accounting for the covariates but assuming no genetic effects. For binary phenotypes, we consider a logistic model that regresses the binary traits on the covariates such that $\mu_i^0 = e^{X_i\gamma}/(1 + e^{X_i\gamma})$. Quantity ω_i is a pre-specified weight to account for the unequal variance of Y_i. In this work, we use the optimal weight based on the logistic model, i.e., $\omega_i = 1/\mu_i^0(1 - \mu_i^0)$ [21].

As indicated above, we consider one environmental effect interacting with the gene effects in this work, which implies that X_{Ei} and X_{Ej} in the Model (2.3) are scalars. In Model (2.3), S_{ij} could be either unweighted similarity S_{ij}^{IBS} or weighted similarity S_{ij}^{W} defined in Section 2.2.1. Due to the reason of the standardization in T_{ij}, there is no intercept and interaction of the two covariates $X_{Ei}X_{Ej}$ [22]. Under the Model (2.3), the GxE effect is examined under the null hypothesis of $H_0 : d = 0$ and the joint test could be conducted under the null hypothesis of $H_0 : c = d = 0$.

Under the Model (2.1) for binary cases, we assume Y_i to be conditionally independent with the conditional mean $\mu_i = \delta(X_i\gamma + G_i + X_{Ei}\eta_i)$ where $\delta(.) = g^{-1}(.)$, variance $var(Y_i|X, S, G, \eta) = \mu_i(1 - \mu_i)$. Taking the Taylor expansion on mean function $\mu_i = g^{-1}(X_i\gamma + G_i + X_{Ei}\eta_i)$ with respect to G and η around $E(G) = 0$ and $E(\eta) = 0$ respectively, we show that

$$\begin{aligned}
E(Y_i|X,S) &= E_G\{E_\eta(\mu_i|X,S,G,\eta)\} \\
&\approx E_G\{E_\eta(g^{-1}(X_i\gamma + G_i) + [g'(X_i\gamma + G_i)]^{-1}X_{Ei}\eta_i|X,S,G,\eta)\} \\
&= E_G\{g^{-1}(X_i\gamma + G_i)|X,S,G\} \\
&\approx E_G\{g^{-1}(X_i\gamma) + [g'(X_i\gamma)]^{-1}G_i|X,S,G\} \\
&= \mu_i^0.
\end{aligned} \tag{2.4}$$

Because of the approximation shown in Equation (2.4), we thus derive the approximation for the covariance of Y_i and Y_j under the GLMM which is expressed as

$$\begin{aligned} cov(Y_i, Y_j|X, S) &= E[\{Y_i - E(Y_i|X, S)\}\{Y_j - E(Y_j|X, S)\}|X, S] \\ &\approx E[(Y_i - \mu_i^0)(Y_j - \mu_j^0)|X, S]. \end{aligned}$$

As a result, the expectation of the gene-trait similarity T_{ij} conditional on X and S is approximate to the quantity shown as follows

$$\begin{aligned} E(T_{ij}|X, S) &= E[\{\omega_i(Y_i - \mu_i^0)\}\{\omega_j(Y_j - \mu_j^0)\}|X, S] \\ &\approx \omega_i\omega_j \times cov(Y_i, Y_j|X, S) \\ &\approx \omega_i\omega_j \times [g'(X_i\gamma)g'(X_j\gamma)]^{-1}\{\tau S_{ij} + \phi X_{Ei}X_{Ej}S_{ij}\}. \end{aligned} \tag{2.5}$$

The details of the derivation shown in Equation (2.5) are provided in Appendix A. If we choose $\omega_i = g'(X_i\gamma)$ as the weight, the expected gene-trait similarity $E(T_{ij}|X, S)$ could be expressed as

$$E(T_{ij}|X, S) \approx \tau S_{ij} + \phi X_{Ei}X_{Ej}S_{ij}.$$

It is straightforward that the $E(T_{ij}|X, S)$ derived from the generalized linear mixed model shown in Equation (2.2.2) is equivalent to the Model (2.3). Thus, testing the coefficients in the gene-trait similarity regression model given in Equation (2.3) is equivalent to examining the variance components in the generalized linear mix model shown in Equation (2.1).

2.2.3 Score Test

Under the proposed generalized linear mixed model, the score test to examine GxE effect could be derived as [72]

$$U_{\text{GxE}} = \frac{1}{2}\{(y^W - X\hat{\gamma})^T V_0^{-1}\Sigma V_0^{-1}(y^W - X\hat{\gamma}) - tr(P_0\Sigma)\}|_{\phi=0,\tau=\hat{\tau},\gamma=\hat{\gamma}}, \tag{2.6}$$

where y^W is the working vector calculated by $y^W = X\hat{\gamma} + Z\hat{b} + \Delta(y - \hat{\mu}^b)$ under the model assuming no GxE interaction effect in it and $\Sigma = DSD$. Under the same assumption, $g(\mu) = X\gamma + Zb$, $\Delta = diag\{g'(\mu_i)\}$ with $g'(\mu_i) = 1/\{\mu_i(1-\mu_i)\}$, $V_0 = W^{-1} + \tau S$ where $W = diag\{\mu_i(1-\mu_i)\}$ and $P_0 = V_0^{-1} - V_0^{-1}X(X^TV_0^{-1}X)^{-1}X^TV_0^{-1}$.

Similarly, the score test which examines the joint effect of both gene and gene-environment effects is represented as

$$U_{\text{joint}} = \frac{1}{2}\{(y^{*W} - X\hat{\gamma})^T V_0^{*-1}\Sigma^* V_0^{*-1}(y^{*W} - X\hat{\gamma}) - tr(P_0\Sigma^*)\}|_{\phi=0,\tau=0,\gamma=\hat{\gamma}}, \tag{2.7}$$

where $y^{*W} = X\hat{\gamma} + \Delta(y - \hat{\mu}^0)$, $\Sigma^* = DSD + S$ and $V_0^* = W^{-1}$. The details of estimating the fixed effects and variance components are given in Appendix B. The details in computing the test score statistics, their corresponding distributions and p-values calculation are provided in Appendix C.

In cases when variants are common such as in the simulation on Hapmap based data and real data application on GWAS data, the unweighted similarity matrix S^{IBS} is considered. In cases of multiple rare variants are present such as in the simulation on COSI based data and real data application on CoLaus data, the weighted similarity matrix S^W is preferred.

2.2.4 Simulation Studies

To investigate the performance of the proposed method, we conducted simulation studies on two different types of data. The first systematic simulation study was conducted on the COSI model based data where all the markers are rare variants. We considered the two study designs where the disease was either from a random sample study or from a case control study. When analyzing the COSI model based data, the weighted similarity matrix S^W was applied to deal with the rare variants. The second systematic simulation study examines the proposed method on a Hapmap based data where most markers are common variants mixed with one rare variant and one low frequency variant. When the Hapmap based data were analyzed, the unweighted

similarity matrix S^{IBS} was considered.

Simulation Study using the COSI Based Data

In the simulation study for COSI based data, the haplotype information was provided on 10,000 chromosomes, where 1Mb region on each chromosome was generated according to a coalescent model where the LD pattern and population history was mimicked by using COSI [20]. Genetic variants with minor allele frequency less than 5% are defined as "rare variants" and the first 100 rare variants were chosen from the 1Mb region and used in the analysis. The simulation study was conducted based on the haplotype information from the variants in the gene bank.

We used the logistic model shown in Equation (2.8) to generate the COSI based data

$$logit\ P(Y_i = 1|X_i, G) = \gamma_0 + X_{Ei}\gamma_E + G_{1i}\gamma_G^1 + \cdots + G_{Ri}\gamma_G^R + G_{1i}X_{Ei}\gamma_{\text{GxE}}^1 + \cdots + G_{Ri}X_{Ei}\gamma_{\text{GxE}}^R, \tag{2.8}$$

where $G_{1i}, \ldots, G_{Ri}$ are the genotype information of the first R rare variants selected as the group of disease-risk contributing variants (D-variants) from the gene bank. We also used "causal SNPs" to indicate these D-variants in this article exchangeably.

In the Equation (2.8), the coefficients $\gamma_G^1, \ldots, \gamma_G^R$ for these R D-variants were computed as

$$\gamma_G^r = \log\left\{\frac{\alpha_G}{(1-\alpha_G)q_r} + 1\right\}, \quad r = 1, \ldots, R,$$

where q_r is the minor allele frequency for rth variant based on the haplotype information of 10,000 simulated chromosomes and α_G is the marginal population attributable risk (PAR) for gene effect which controls the low marginal effect on each rare SNPs [12]. The coefficients for each gene-environment effect $\gamma_{\text{GxE}}^1, \ldots, \gamma_{\text{GxE}}^R$ in the same Equation were assigned as

$$\gamma_{\text{GxE}}^r = \log\left\{\frac{\alpha_{\text{GxE}}}{(1-\alpha_{\text{GxE}})q_r} + 1\right\}, \quad r = 1, \ldots, R.$$

In the spirit of controlling the PAR for each group of the causal SNPs, we assigned each causal

variant a coefficient using a same marginal PAR (α) to make sure the low effect for each variant. Therefore, we calculated the marginal PAR for gene main effect as $\alpha_G = \text{group-PAR}_G/R$. Without loss of generality, $\text{group-PAR}_G = 0.02$ throughout this simulation study for the COSI based data. Similarly, $\alpha_{\text{GxE}} = \text{group-PAR}_{\text{GxE}}/R$ was defined as the marginal PAR for each GxE effect. In this simulation study, $\text{group-PAR}_{\text{GxE}} = 0.1$ was applied for the COSI based data.

For simplicity, no confounders are included in the models for this stimulation study and only one environmental effect was included in the models. To calculate the similarity based on the genotype information, 2 haplotypes were randomly selected from the gene bank each time which then determined an individual's genotype.

When testing the power of GxE effect and joint effect, Model (2.8) was applied to generate the simulation data. However, before carrying out any further tests and comparisons, we needed to test if the proposed method controlled the type I error rate in a proper manner for the COSI based data. Both the random sample simulation data and the case control simulation data were generated under the two reduced logistic models based on Model (2.8) respectively. Under the null hypothesis $H_0 : \phi = 0$ which assumes no GxE effect but having gene effect, the reduced model is

$$logit\ P(Y_i = 1|X_i, G) = \gamma_0 + X_{Ei}\gamma_E + G_{1i}\gamma_G^1 + \cdots + G_{Ri}\gamma_G^R.$$

Under the null hypothesis $H_0 : \tau = \phi = 0$ assuming there are no gene and GxE effects simultaneously, the reduced model is

$$logit\ P(Y_i = 1|X_i) = \gamma_0 + X_{Ei}\gamma_E.$$

In this simulation study, we considered a sample size of $N = 1500$, and 1000 data sets were generated for each scenario under both null hypotheses using Models (2.2.4) and (2.2.4), respectively. In order to test the power, 500 data sets were generated for each scenario using Equation (2.8). For the random sample simulation data, around 30% of the individuals have the disease. For the case control simulation data, in order to obtain 50% cases in the data

set, the same three models were used to generate the data until we obtained $N/2$ cases in the simulated data. To investigate the type I error controlling issue and conduct the power comparison for both study designs, 5 groups of D-variants were tested. The 5 considered groups were the first 20, 40, 60, 80, and 100 variants selected as the groups of D-variants. In addition, situations where 20, 40, 60, or 80 variants were randomly selected as the groups of D-variants were also tested for validation of type I error and conducted for power comparisons. We only showed the results where the first 20, 40, 60, 80, and 100 variants were selected as the groups of causal SNPs in this article, because the results from the other situations are very similar to the results where the first R variants were selected as D-variants. In this simulation study, the counting based burden test [15, 20, 73] was used as a comparison to the proposed method. The counting based burden test use one "super-variant" which contains all the genotype information from each individual [14]. Then the information for this super-variant was applied to test for either the GxE effect or the joint effect based on the logistic model $logit\ P(Y_i = 1|X_i, G) = \gamma_0 + X_{Ei}\gamma_E + G_{Ci}\gamma_C + X_{Ei}G_{Ci}\gamma_{CE}$ where G_{Ci} is the genotype information of the super-variant for individual i, γ_C is the fixed effect coefficient corresponding to G_{Ci} and γ_{CE} is the fixed effect coefficient for the interaction of $X_{Ei}G_{Ci}$.

Simulation Study using the Hapmap Based Data

In the simulation study for Hapmap based data, samples were selected from the population of CEU. The CEU population represents Utah residents with Northern and Western European ancestry from the CEPH (Centre d'Etude du Polymorphisme Humain) collection, and it is one of the 11 populations in HapMap Phase 3 project. Genotype information with a total of 234 phased haplotypes was provided on gene TCF7L2 from chromosome 10. Although there are 111 SNPs on the gene TCF7L2 from the HapMap project data, 29 of the SNPs were genotyped in the real data application. Thus, only those 29 variants were used as the gene bank to perform all the analysis in this simulation study. To find out the association between genetic variants and disease risk, a variety of scenarios from a case control study were considered. Although all

causal SNPs were assumed to increase the disease risk for practical purpose in the simulation, it had no impact on the direction to affect the risk in real studies.

The following logistic model shown in Equation (2.9) was used to generate the Hapmap based simulation data

$$logit\ P(Y_i = 1|X_i, G) = \gamma_0 + X_{Ei}\gamma_E + G_{1i}\gamma_G^1 + G_{2i}\gamma_G^2 + G_{1i}X_{Ei}\gamma_{\text{GxE}}^1 + G_{2i}X_{Ei}\gamma_{\text{GxE}}^2, \quad (2.9)$$

where G_{1i} and G_{2i} are the two causal SNPs selected from the 29 SNPs. Without loss of generality, we set $X_{Ei} \sim N(0, 6)$, $\gamma_0 = -2.5$, $\gamma_E = log(1.5) = 0.4055$, $\gamma_G^1 = \gamma_G^2 = \log(1.2)/2 = 0.0912$ and $\gamma_{\text{GxE}}^1 = \gamma_{\text{GxE}}^2 = \log(1.055)/2 = 0.0268$.

In this simulation, we considered a same sample size of $N = 1500$ with $N/2$ cases and $N/2$ controls. Using the formula shown in the (2.9), a random sample with around 30% cases was generated. In order to obtain 50% cases for $Y_i = 1$, such model was used to generate the data until we obtained $N/2$ cases in the simulated data. For each scenario, 2 SNPs were chosen to be the causal SNPs. Therefore, there were $\binom{29}{2} = 406$ scenarios considered and 100 data sets were generated for each scenario to test the power. The power rates were grouped into three categories based on the LD values under each scenario where the LD values were categorized into one of the three conditions: low-LD, median-LD and high-LD structures. The three LD levels were defined based on the empirical values of LD where each LD value was defined as the average of all R^2 values under each scenario. Each R^2 number was calculated as the LD between an observed marker and a risk locus. Because the causal SNPs were excluded, there were 27 observed markers. Each of the 27 SNPs was used to calculate the R^2 with either of the 2 causal SNPs, thus 54 R^2 values were computed under each scenario for this study.

Similarly, the simulation to test type I error rates was conducted on the Hapmap based data before any further tests and comparisons were carried out. The same two null hypotheses were examined with 1000 data sets under each scenario. Under the null hypothesis of $H_0 : \phi = 0$,

reduced model to generate the data is

$$logit\ P(Y_i = 1|X_i, G) = \gamma_0 + X_{Ei}\gamma_E + G_{1i}\gamma_G^1 + G_{2i}\gamma_G^2.$$

Under the null hypothesis of $H_0 : \tau = \phi = 0$, the reduced model is

$$logit\ P(Y_i = 1|X_i) = \gamma_0 + X_{Ei}\gamma_E.$$

Because there is one rare variant and one low frequency variant among the 29 SNPs, 4 scenarios were tested to investigate the type I error rates. The 4 scenarios are: (1) The two causal SNPs are the one rare variant and the low-frequency allele, denoted as "RL"; (2) one causal SNP is the rare allele and the other SNP is a common allele, denoted as "CR"; (3) one causal variant is the low frequency variant and the other causal SNP is a common allele, denoted as "CL"; (4) both causal SNPs are common alleles, denoted as "CC". To assess the type I error rates affected by rare or low-frequency alleles with different minor allele frequency (MAF) values, we used the same SNP rs4918796 as the common allele for the scenarios (2) and (3). When testing the association under each scenario, the test statistic and the corresponding p-value were calculated based on the data set excluding the 2 causal SNPs. The motivation of testing GxE effect under 4 scenarios is to verify whether rare or low-frequency alleles deflate the type I error rates for binary traits as which was observed for quantitative traits [22]. Among the 29 SNPs, rs7089262 is the rare allele with MAF of 0.0085, rs10787476 is the low-frequency allele with MAF of 0.0940. The other 27 genotyped SNPs are all common alleles with MAF > 0.1. In scenario (4), two pairs of SNPs were randomly selected from the 27 common alleles. As a result, the SNPs rs4918796 and rs1555485 were one pair, and the SNPs rs7917983 and rs4132670 from another pair.

For comparison, the same 27 SNPs in each scenario have been analyzed by applying the single-SNP minimum p-value methodology. It is known that SNPs are correlated with each other to some extent, the minimum p-value approach estimates the effective number of the

independent tests to control the overall type I error rate and make the p-value adjustments. The adjusted p-value was computed via the formula $1-(1-\text{minimum p-value})^{k_{eff}}$ where k_{eff} the number of independent LD tests [6]. In the simulation study, each of the 27 SNPs has a p-value, the minimum of those p-values was then chosen and compared to the significance threshold after adjustment.

2.3 Results for Simulation Studies

2.3.1 Simulation Study using the COSI Based Data

Under the null hypothesis of no gene and no gene-environment effects, type I errors for testing GxE effect and joint effect are shown in Table 2.1. The results were obtained by applying the proposed variance component based method ("VC-Based") and the counting based burden test ("Burden") for both the random sample study and the case control study. From the results, we observed that the type I error rates by using both methods are within an acceptable region around the significance level of 0.05, i.e. all type I errors are within 95% confidence interval of the significance level at 0.05. It might indicate that in general the type I error rates by using the variance component based method are closer to the significant level for both tests under the two study designs than the rates by applying the counting based burden test.

Table 2.1: Type I error rates for examining the joint effect over 1000 runs using the COSI based simulation data.

Study Design	GxE test		Joint test	
	VC-Based	Burden	VC-Based	Burden
Random Sample	0.043	0.044	0.052	0.059
Case-Control Sample	0.047	0.047	0.052	0.045

To investigate type I error rates under the null hypothesis of no gene-environment effect, the first 20, 40, 60, 80 and 100 variants were selected as the groups of D-variants. The results under

this null hypothesis are listed in Table 2.2 which provides the type I error rates by using the same methods as under the null hypothesis of no gene and gene-environment effects simultaneously. Similarly, all type I errors are still within 95% confidence interval of the significance level at 0.05 which implies that both tests are valid for the COSI based data with rare variants. Thus, power comparisons could be conducted between the proposed method and count based burden test.

Table 2.2: Type I error rates for examining the Gene-Environment effect over 1000 runs using the COSI based simulation data.

Number of causal SNPs	Random Sample		Case-Control Sample	
	VC-Based	Burden	VC-Based	Burden
20	0.046	0.053	0.064	0.060
40	0.046	0.049	0.049	0.055
60	0.043	0.043	0.045	0.049
80	0.042	0.044	0.050	0.050
100	0.045	0.049	0.041	0.039

The power results with group-PAR for gene effect of 0.02 and group-PAR for gene-environment effect of 0.1 are shown in Figure 2.2. We observed that the power increases as the number of causal SNPs increases by using both methods. The power increase by applying the proposed method is more than the increase in power by using the competing method under the same scenario. In addition, we noticed that for most situations, the power for joint test is slightly smaller than the power for GxE test. It might be the reason that several data sets have significant GxE effect but negligible gene effect. However, compare to the degrees of freedom testing for GxE effect, more degrees of freedom are needed for joint test. When the test statistic value for GxE test is very similar to the test statistic value for joint test, such results become possible.

2.3.2 Simulation Study using the Hapmap Based Data

To investigate type I error rate for Hapmap based data, we first tested the type I error rare under null hypothesis of no gene and gene-environment effects. The results listed in Table 2.3 show the type I error rates provided by both the proposed method and the minimum p-value method ("minPval"). From the observation of the results presented in this table, it demonstrates that the type I error rates fall into the 95% confidence interval at significant level of $\alpha = 0.05$.

Table 2.3: Type I error rates for examining the joint effect over 1000 runs using the Hapmap based simulation data

Type of Test	GxE test		Joint test	
Method	VC-Based	minPval	VC-Based	minPval
Type I error rate	0.037	0.054	0.044	0.06

The results listed in Table 2.4 are the type I error rates under the null hypothesis of no gene-environment effect. Similarly, the type I error rates are all around the nominal level of $\alpha = 0.05$ for the scenario that both causal SNPs are common alleles. When SNPs rs4918796 and rs1555485 are the two causal SNPs, the type I error rate by applying the minimum p-value method is slightly deflated but in an acceptable range. The reason we showed the results from 2 pairs of causal SNPs which are both common alleles is that we wanted to use them as a comparison. The two type I error rates for scenario (4) by applying the proposed method are both closer to the nominal level at 0.05. Thus, the type I error rate of 0.044 by applying the minimum p-value method when both rs4918795 and rs1555485 are common alleles could not be due to the MAF of the causal SNPs, but might be due to the LD structure with the other 27 SNPs. Under this scenario, the average LD value is 0.2145 which is relatively low among all the 406 scenarios. Under the scenario when the common SNPs rs7917983 and rs4132670 are the two causal variants, the average LD value is 0.287 which is in the median level among all the LD values. Therefore, it might indicate that the LD value also affects the ability of the minimum P-value method to detect the significance of GxE effect. From the same table, we observed that

when at least one of the causal SNPs is rare or low frequent, the type I error rates are deflated by applying both methods. The type I error rates are around 0.04 when we set our nominal level at $\alpha = 0.05$ even though they are all acceptable. In addition, the low LD structure under the first 3 scenarios may be another reason deflating the type I error rates.

Table 2.4: Type I error rates for examining the Gene-Environment effect over 1000 runs using Hapmap based simulation data. 4 scenarios are considered where 2 causal SNPs are used under each scenario. For scenario (1), a rare variant and a low frequency variant are used as the causal SNPs ("RL"). For scenario (2), a rare variant and a common variant are used as the causal SNPs ("CR"). For scenario (3), a low frequency variant and a common variant are used as the causal SNPs ("CL"). For scenario (4), two common variants are used as the causal SNPs ("CC"). The reference SNP ID numbers are provided corresponding to each causal SNP with its MAF using the Hapmap3 data.

Scenario		Causal SNPs	MAF	VC-Based	minPval
Scenario 1	RL	rs7089262 rs10787476	0.008547 0.094017	0.042	0.04
Scenario 2	CR	rs7089262 rs4918796	0.008547 0.196581	0.04	0.043
Scenario 3	CL	rs10787476 rs4918796	0.094017 0.196581	0.04	0.045
Scenario 4	CC	rs4918796 rs1555485	0.196581 0.222222	0.047	0.044
	CC	rs7917983 rs4132670	0.418803 0.299145	0.049	0.05

A power comparison was performed to compare the power of the proposed method and the minimum p-value method. Each two of the 29 SNPs were taken turns to be the causal SNPs such that there were total 406 scenarios. Based on the empirical LD values, we categorized each LD value into one of the three LD levels. Let LD values be ordered in an increasing manner. The first 33.33% smallest LD values were defined as low-LD values which fall in the interval of $[0, 0.2434)$. The empirical LD values were defined as median-LD values if they fall into the range of $[0.2434, 0.2873)$. Values greater than 0.2873 were defined as high-LD values.

The box-plots in Figure 2.3 showed the minimum, first quantile, median, third quantile,

maximum values and the mean power for each LD-level. When LD is low, the median power by applying the proposed method is slightly lower than the median power by using the minimum p-value method. Other than median power, the proposed method has higher power in terms of other quantities shown in the box-plot for low-LD. It is obvious that the average power for the proposed method is higher than the average power by applying the minimum p-value method. Among the 135 scenarios with low-LD structures, 37 (27.4%) scenarios show that the proposed method has lower power but 92 (68.1%) scenarios present higher power by applying the proposed method. For median-LD values, both the median and mean of the power for the proposed method are higher than the ones by using the minimum p-value method. Among the 136 scenarios with median-LD structures, 32 (23.5%) scenarios indicate that the power by applying the proposed method is not as high as the ones by using the minimum p-value method. But 99 (72.8%) scenarios show that the proposed method is more powerful in detecting the significance. When the LD value is relatively higher, both the median power and the mean power by using the proposed method are higher. Though the proposed method has lower power under 23 (17.0%) scenarios, 111 (82.2%) scenarios show that the proposed method is more powerful. By comparing the power at different LD levels, we observed that when the LD values is lower, the power by applying both methods are lower than the power when the LD value is either in median-level or higher-level in general. This observation is reasonable because low-LD scenarios represent the cases in which the markers contain little information about the two risk loci.

In addition, we also conducted the power comparison for the joint test. Very similar patterns are observed which is also shown in Figure 2.3. When the LD value is low, the proposed method and minimum p-value method show relatively equivalent power. Although the median power is slightly lower by applying the proposed method when testing the gene and gene-environment effects jointly, the mean power is still slightly higher. By comparing the power at low-LD structure, the performance of joint test by applying the minimum p-value method is slightly better than its performance in the GxE test. It is understandable because the minimum p-value

method provides good power in detecting the gene effect which improves its performance in the joint test. For median-LD and high-LD scenarios, both median power and mean power by using the proposed method are higher than the ones by applying the minimum p-value method. The reason is that the more information about the causal SNPs is correlated with the other SNPs, the more powerful the proposed method is to detect its significance.

2.4 Real Data Applications

2.4.1 Real Data Analysis for CoLaus Study Data

We applied the proposed model on a sub-sample collected from the CoLaus study which is a population-based study approved by the Institutional Ethic's Committee of the University of Lausanne, Switzerland [68]. The primary interest of the CoLaus study is to evaluate the prevalence of risk factors in cardiovascular disease. The secondary interest is to find out any genetic determinants which are associated with those risk factors in the Caucasian population. A total sample of 6,188 individuals aged from 35 to 75 were randomly selected, their genotype information was obtained by applying the Affymetrix 500K chip.

To explore the association between the available genes and smoking status in obesity which is one of the cardiovascular disease risk factors, we were interested in testing the interaction effect of the genes available in CoLaus study and smoking status. The smoking status was treated as the environmental factor in a sub-sample of the data set from the CoLaus study. Genotype information for 8 genes is available in the study, which includes the gene GPBAR1.

Because of the missing genotypes in the data set for gene GPBAR1, data from 1937 observations are available for the tests where 252 are cases. Genotype information on 11 SNPs was used in the analysis where all the 11 genetic variants are rare variant with minor allele frequency being less than 1%. As mentioned in Section 2.2.2, when most alleles are rare variants, S^W is suggested.

Because it was of primary interest to examine the interaction between the gene GPBAR1

and smoking status, we performed the GxE test on this data set. In addition to the smoking status, 9 confounders were also included in the model such age, gender, alcohol drinking status, physical activity status. In this data application, these confounders were not treated as the environmental factors interacting with the gene effects. By applying the proposed method, we obtained a p-value of 4.87×10^{-3}. It was a strong evidence to indicate that the gene variants interplay with the smoking status in affecting obesity. By using the counting based burden test, a p-value of 0.127 was obtained which indicated non-significance of the interaction of the gene GPBAR1 and smoking status. We observed that the p-value calculated by using the proposed method is much smaller than the one by using the competing method. This result may support the observation in the simulation on the COSI based data that the proposed method is in general more powerful in detecting the significance of GxE effect.

In addition of testing the gene-environment interaction effect, we also conducted a test to examine the gene and gene-environment effects simultaneously. The proposed method had a p-value of 6.427×10^{-3} and the counting based burden test had a p-value of 0.15. Similar to the result we observed in the GxE test, the proposed method indicated the significance of the joint effect. However, non-significance was concluded by using the competing method. As shown in the simulation study, the results from the real data application also show that the proposed method gains more power for COSI based data in presence of rare variants.

2.4.2 Real Data Analysis for WTCCC Study Data

We utilized the proposed method on a sample collected from the Wellcome Trust Case Control Consortium (WTCCC) studies. WTCCC is a consortium which aims to help better understanding and studying the etiological causes of several global diseases. Some previous studies [74, 75] have already discovered the association between the gene TCF7L2 and BMI in Type 2 Diabetes (T2D) studies by applying different methods and manners in Genome-Wide Association Studies. To use a marker-set analysis exploring this association, we are interested in testing the interaction effect of gene TCF7L2 and BMI by applying the proposed approach, and BMI is

the genetic modifier which is expressed as the "environmental factor" in the GxE interaction for a T2D data set from the WTCCC studies. In this data set, there are total 3368 individuals remained after the quality control procedure, among them 1913 are cases and 1455 are from the control group. The case samples were collected from various sites across the UK in purpose of allowing it to be efficiently compared to the controls. The 1455 controls included in the data set were samples from the 1958 British Birth Cohort. The genotyping was conducted by applying the Affymetrix 500K chip.

We performed the GxE test by applying the proposed method, and we obtained a p-value of 4.05×10^{-5}. There is strong evidence showing that the gene variants interact with the Body Mass Index in affecting the disease of Type 2 Diabetes. By using the minimum p-value method, the adjusted p-value of 2.72×10^{-3} was obtained. We observed that the p-value calculated by using the proposed method is much smaller than the one obtained by applying the minimum p-value method though both p-values are smaller than the nominal level at $\alpha = 0.05$. Even though it is not absolutely that the proposed method always have lower p-values in every scenario through the results of the simulation, it still showed that the proposed method are more powerful in detecting the significance in most cases.

Additionally, the joint test was also conducted for the application on this GWAS data set. The p-value by using the proposed method is 1.81×10^{-10} and the adjusted p-value by applying the minimum p-value method is 1.39×10^{-9}. Compared to the p-values testing for the GxE effect, the p-values for the joint test are smaller. It might be the reason that the gene effect in this real data set has a stronger effect on an individual's chance to have Type 2 Diabetes. However, the gene-environment interaction effect cannot be ignored because of its significance.

2.5 Discussion

The existing methods currently available do not have the capacity to examine GxE interaction effects for studies on binary traits. In this article, we proposed a flexible approach which is able to accommodate covariates into the model. In addition, this proposed method is powerful,

computationally feasible and efficient when conducting marker-set analyses for examining the gene-environment interaction on binary traits. By showing the connection between the generalized linear mixed model and the gene-similarity regression model, the proposed approach has flexibility in inference on the test statistics, while it maintains the power gain brought by aggregating the genetic information at a similarity level. In addition, the proposed method is more robust than model-based regression approaches treating the genotype information as fixed effect, such as the counting based burden test considered as a competing method in the research work. When the mean-model is misspecified or population substructure exists in the model, the counting based burden test may lead to inflation in testing the GxE interaction [76]. Even though this situation did not occur in our simulation studies, we should take great care when applying those model-based regression methods to real data sets. Because the proposed method collapses the genotype information at a similarity level and treats the genetic effects as random effects, it is comparably stable under different assumptions.

In this current work, examining the interaction of gene-environment effect is the primary interest when we only considered one environmental factor in the model. However, if it is believed that more than one environmental factor interacts with genes, a similar model involving more GxE interaction effects could be applied. Another extension to the current work is that one may apply the idea to test gene-gene interactions. When one gene is suspected to be interacted with other genes, a generalized linear mixed model involving multiple genes and gene-gene interactions could be constructed. Especially for the next generation sequencing data, different genes with rare variants are believed to have more effect size than common variants identified in GWAS [77], the examination of gene-gene interaction may therefore discover even more associations between the genes and complex diseases.

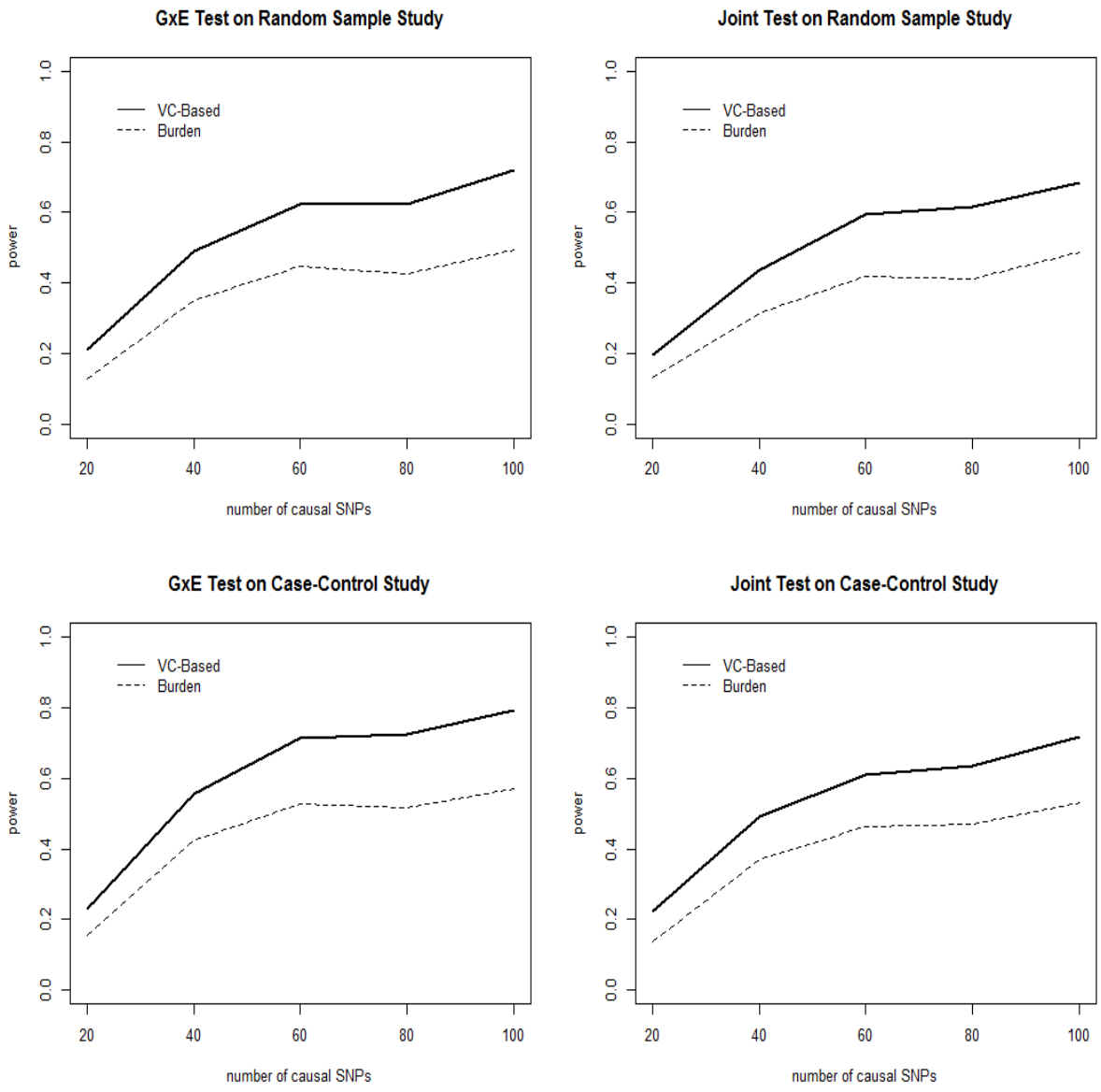

Figure 2.2: The power comparisons for gene-environment test and joint test over 500 simulation runs using the COSI based simulation data. The solid lines are the power by applying the proposed variance component based method. The dashed lines are the power by applying the counting based burden test.

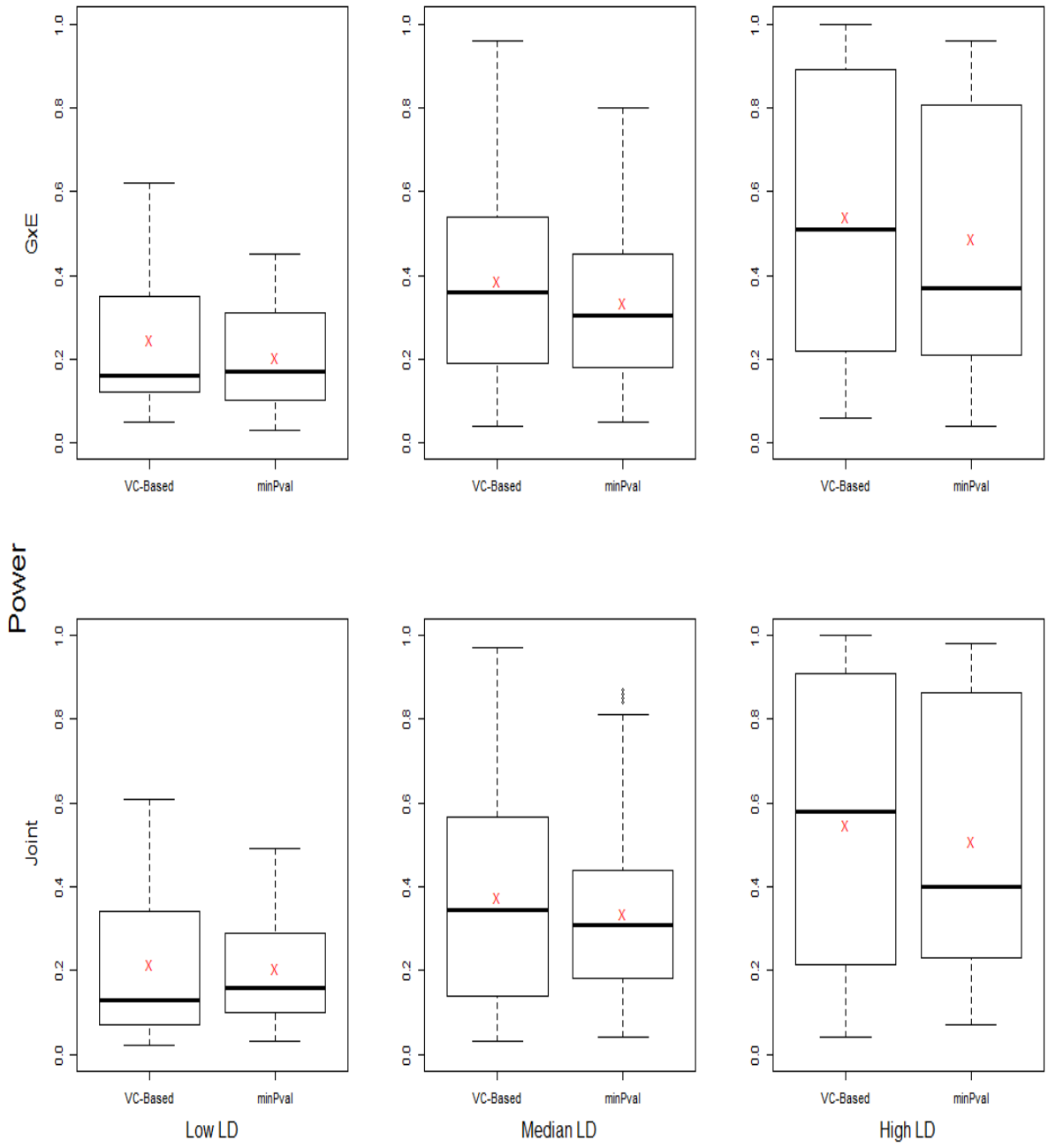

Figure 2.3: The power comparisons for gene-environment test and joint test over 500 runs under the three LD-level conditions using the Hapmap based simulation data. In each box-plot, the summary for the power by using the proposed variance component based method (VC-Based) is shown on the left and the summary for the power by using the minimum P-value (minPval) approach is shown on the right. A red "X" in each box-plot is the power average under each LD condition.

Chapter 3

Gene-Set Association Analyses for Quantitative Traits using Linear Mixed Models

3.1 Introduction

Since the first successful association detected in genome-wide association study, association analysis has become a dynamic area of study that has drawn researchers' great effort in order to discover the genetic variants determining the multi-factorial disorders. There are over one thousand associations that have been identified during the last decade, where analyses examined the associations at either SNP-level or gene-level. However, those associations only make up a small portion of the genetic heritability. There are still myriad of unknown and undiscovered associations that need to be explored. Instead of assessing the genetic effect for multi-marker from a same gene, association analyses have extended the focus range from a single gene to a gene set. It is believed that multi-gene analysis may be more promising to capture the important effects on pathways missed by single-gene analyses. Researchers also indicated that most significant associations have not been identified replicatively by different SNP-level or gene-

level association studies on a same disease [45]. For that purpose, gene set association analysis (GSAA) could be used as a complementary tool to mine either existing or new association study data sets aiming to identify more multi-genic biological structure for multi-factorial disorders.

To uncover the pathway mechanisms that modulate disease risk, gene set association approaches offer great potential but encounter challenges brought by the large amount of information in the existing association study data sets. Different from single-gene analyses, multi-gene analyses examine associations for networks or pathways which include multiple genes with complicated relationship among them. It requires a lot computational capacity in calculation which may involve inverting extraordinary high-dimensional matrices in order to estimate nuisance parameters. Therefore, we need to be careful in dealing with two major challenges: (1) how to better account for the within-gene multi-marker information[53, 54]; (2) how to appropriately assess association at both gene set level and gene level [61, 53]. To address the challenge of incorporating the genetic information within each gene, we collapse such information at similarity level. For each pair of individuals, a similarity score which measures the genetic distance between the two individuals within a gene is calculated. Thus, a large similarity matrix with $N \times N$ dimension is computed for each gene where N is the sample size number. In order to reduce the dimensions of the similarity matrices, we consider the eigen-decomposition technique. To handle the challenge of assessing the associations, a variable selection technique is incorporated to the variable selection on gene-specific variance components to identify significant genes from a gene set.

In this research work, we introduce a gene set association analysis approach via a mixed effects selection framework. A linear mixed model is proposed for quantitative traits in which covariates could be included as baseline information to adjust in the model. This method applies the adaptive LASSO technique on the variance components to select genes significantly associated with disease risk. Such method provides sparse estimates of gene-specific variance components to assess the overall gene set effect. In order to derive valid variance component estimators, an EM algorithm is considered to estimate these parameters. In this study, we as-

sume there is no gene-gene interactions present in the proposed model in order to investigate the underlying impact of gene sets under the additive models.

3.2 Material and Method

3.2.1 Linear Mixed Model for Gene-set Association Analysis

To demonstrate the proposed method, the following notations are defined. Suppose there are N individuals in the study. Let $Y = (Y_1, \ldots, Y_N)^T$ be the response vector of the observed phenotypes. Specifically, quantitative traits are of interest in this project. Let X be the design matrix where each row includes the intercept and covariates for each individual. Let γ be the coefficient vector for the intercept and covariates. We assume there are K genes in the gene set, and genetic information is available for multiple markers within each gene. Let M_k be the number of loci for gene k. Let S_k be the similarity matrix where each element calculates the genetic similarity for kth gene between two individuals. Because most alleles are rare variants, we considered the weighted similarity matrix S_k^W which incorporates minor allele frequencies for markers within the kth gene. Among various choices of the weights, we define the weighted similarity element for kth gene as $S_{kij}^W = \sum_{m=1}^{M_k} w_m IBS(g_i^m, g_j^m) / \sum_{m=1}^{M} w_m$ with weights $w_m = (1 - q_m)^{24}$ for $m = 1, \ldots, M_k$ and q_m is the minor allele frequency for locus m on the kth gene [20].

Therefore, the proposed linear mixed model for quantitative traits is represented as

$$Y = X\gamma + G_1 + \cdots + G_K + \varepsilon, \tag{3.1}$$

where $G_k \sim N(0, \phi_k S_k^W)$ for $k = 1, \ldots, K$ and $\varepsilon \sim N(0, \sigma I_N)$.

However, using Model (3.1) to conduct a gene set association analysis may involve high-dimensional calculation with respect to multiple $N \times N$ similarity matrices in estimating the variance components. To reduce the dimensions of these matrices and make the computation more efficient, we re-write the proposed model to the following one

$$Y = X\gamma + Z_1 b_1 + \cdots + Z_K b_K + \varepsilon, \tag{3.2}$$

where Z_k is a $N \times r_k$ matrix such that $Z_k Z_k^T = S_k^W$ for $k = 1, \ldots, K$ and r_k is the rank of S_k^W, $b_k \sim N(0, \phi_k I_{r_k})$. To identify significant genes from the gene set, a variable selection technique is considered to eliminate genes with nonsignificant associations with the disease risk; or equivalently, genes with ϕ_k's being zero. In this work, the adaptive LASSO is applied to handle the sparseness on ϕ_k's for $k = 1, \ldots, K$ due to its optimal properties.

3.2.2 Connection with a corresponding Gene-Trait Similarity Regression Model

Motivated by the previous work where the variance component model has been shown its connections to a gene-trait similarity regression model when markers from one gene is of interest [21], we prove the variable selection on the variance components in Equation (3.1) is equivalent to select the gene-specific coefficients for the gene-trait similarity regression model in Model (3.3). Let T_{ij} be the trait similarity between individuals i and j which is defined as $T_{ij} = (Y_i - \mu_i^0)(Y_j - \mu_j^0)$ where the conditional mean $\mu_i^0 = X_i\gamma$ [21, 22]. Similar to the association analysis for marker-set on a single gene, the corresponding gene-trait similarity regression model for gene set association analysis is

$$E(T_{ij}|X, S) = a_1 \times S_{1ij} + a_2 \times S_{2ij} + \cdots + a_K \times S_{Kij}, \tag{3.3}$$

where S_{kij} is the similarity matrix for the gene k and the weighted similarity matrix S_{kij}^W is considered in this work.

Let $G = (G_1, \ldots, G_K)$ be the genetic information from the gene set. Thus the covariance

between two individuals under the Model (3.1) could be represented as

$$\begin{aligned} cov(Y_i, Y_j|X, G) &= Z_{1i}cov(b_1)Z_{1j}^T + Z_{2i}cov(b_2)Z_{2j}^T + \cdots + Z_{Ki}cov(b_K)Z_{Kj}^T \\ &= \phi_1 Z_{1i}Z_{1j}^T + \phi_2 Z_{2i}Z_{2j}^T + \cdots + \phi_K Z_{Ki}Z_{Kj}^T \\ &= \phi_1 S_{1ij} + \phi_2 S_{2ij} + \cdots + \phi_K S_{Kij}. \end{aligned}$$

Therefore, it is straightforward that a variable selection on the variance components of $\phi_1, \ldots, \phi_K$ in the proposed linear mixed model is equivalent to a variable selection among $a_1, \ldots, a_K$ in the gene-trait similarity regression model. By showing the connection between the two models, we expect to take benefit from the advantages of each model where linear mixed model has the flexibility in inference and gene-trait similarity regression model could gain power to detect more associations in the presence of nonlinear genetic effects or diverse effect sizes.

3.2.3 Gene Selection using Adaptive LASSO

For the purpose of identifying significant genes associated with complex diseases, a variable selection technique is considered on the gene selection procedure. In this work, the adaptive LASSO [60] which assigns different weights to each variance component is adopted. The variable selection and parameter estimation are processed simultaneously using the proposed method. To estimate the variance components ϕ_k for $k = 1, \ldots, K$, a penalized likelihood based approach is used where unimportant variance components will be shrunk to zero [62].

Inspired by the previous work by Zou [60] and Zhang and Lu [78], a penalized log-likelihood function with adaptive LASSO penalties is applied for the proposed linear mixed model which is expressed as

$$l_p(\gamma, \phi, \sigma; y, \lambda) = l(\gamma, \phi, \sigma; y, \lambda) - n\lambda \sum_{k=1}^{K} \frac{\phi_k}{\tilde{\phi}_k}, \tag{3.4}$$

where $\phi_k \geq 0$ for $k = 1, \ldots, K$, λ is a tuning parameter and $\tilde{\phi}_k$'s are some estimators of ϕ_k's. One option for those $\tilde{\phi}_k$'s are the maximum likelihood estimates derived by maximizing Equation

(3.4) in the case when $\lambda = 0$ [62]. The coefficients for fixed effects γ, variance components $\phi = (\phi_1, \ldots, \phi_k)$ and σ are estimated by maximizing the penalized log-likelihood function shown in the Equation (3.4) with respect to γ, ϕ and σ.

To calculate the estimators for variance components ϕ_k's, we could derive the estimation by directly maximizing the Equation (3.4). However, it may yield invalid estimators of ϕ_k's where negative values may be obtained in some cases. To avoid such a trouble, we use EM algorithm to estimate all parameters including the coefficients for the covariates and variance components by maximizing the same objective function (3.4). For each pre-specified λ value, the algorithm details deriving the corresponding maximum penalized likelihood estimation is provided in Appendix D.

Based on the estimators obtained by adopting the EM algorithm for each known λ value, we eliminate the genes with nonsignificant associations with zero variance component estimators. In situations where $K^* \leq K$ genes with non-zero variance component estimators are selected for a known λ, the linear mixed model shown in Equation (3.2) could be reduced to the following model with less variables

$$Y = X\gamma + Z_1 b_1 + \cdots + Z_{K^*} b_{K^*} + \varepsilon^*. \tag{3.5}$$

Under the Model (3.5), we apply the EM algorithm again to derive the estimations of all the parameters based on the likelihood function without penalty terms. We show the steps of deriving the maximum likelihood estimation adopting EM algorithm in Appendix E

In order to choose the optimal λ, Bayesian information criteria (BIC) is used as the measurement to conduct the model selection which is defined as

$$BIC = -2l(\gamma, \phi, \sigma; y) + K^* ln(n) = -2l(\theta; y) + K^* ln(n), \tag{3.6}$$

where $l(\gamma, \phi, \sigma; y)$ is log-likelihood under each reduced model, N is the number of individuals in the data and K^* is the number of parameters in the reduced model (3.5). The λ value which

provides the minimum BIC score is chosen as the optimal estimator for the tuning parameter. The model with the optimal λ value is selected as the optimal model. The genes in the optimal model are identified to be significantly associated with the disease risk. The details of deriving the likelihood function and its logarithm formula are shown in Appendix F.

3.3 Simulation Studies

To investigate the performance of the proposed method, we conducted a systematic simulation study based on the genotypes from a real data set. It is a sample from the CoLaus study which is conducted at population level in Switzerland [68]. The genetic information for 1962 individuals is available on seven genes in the data set, where most of genes are from different chromosomes. Among the 7 genes, MLNR from Chromosome 13 has the smallest gene size with 24 SNPs in the data. Gene GPBAR1 from Chromosome 2 has genotype information on 44 loci. There are two genes from Chromosome 19. One is gene SIRT6 which has 56 genotyped SNPs in the data; and SIRT2 is another gene from this chromosome where 60 loci have genotype information. Gene SIRT3 is a gene from Chromosome 11 with 60 SNPs available for analysis. SIRT1 from Chromosome 10 has 96 genotyped SNPs. Gene PLA2G7 from Chromosome 6 have the largest gene size in the data where 114 SNPs have genotypes information. Within each gene, around 90% of the SNPs are rare variants with minor allele frequencies less than 0.01.

In this study, a sample size of $N < 1962$ was considered and two simulation settings were used to generate the simulation data. The first setting is that the genetic information is randomly chosen from N different individuals for one gene at a time. Under this setting, the genetic information for sample i may take the multi-marker genotypes within one gene from individual A and take the multi-marker genotype within another gene from individual B. Such setting could guarantee that the genotypes from different genes are mutually uncorrelated. The second setting is that the genetic information for all genes is arbitrarily selected from N different individuals. Under this setting, the genotypes for sample i are used the genetic information from the same individual in the real data where the LD structure is maintained for all available

SNPs among these genes. Therefore, there may be some correlations existed among the genes. To explore the LD structures among the seven genes, we calculate the R^2 values for each pair of SNPs from two different genes. In Figure 3.1, we showed the boxplots for the R_2 values between each pair of genes where the minimums, first quartiles, medians, third quartiles, maximums and the averages were provided. We observed that most R^2 values are distributed near zero according to the different quartile values. However, we also discovered that a few pairs of SNPs are in high LD between each pair of genes. For example, a SNP in MLNR is highly correlated with a SNP in SIRT2 with R^2 of 1. The exact average R^2 values and the medians for each pair of genes were given in Table 3.1 for a better understanding of the LD structures among the genes. All the average R^2 values are below 0.002 where GPBAR1 and SIRT1 have the largest average R^2 value of 0.001654. According to the fact that all the medians are close to zero, it indicates that most pairs of SNPs from different genes are not correlated or in extremely low LD. However, the outliers observed in the Figure 3.1 may have an influence on the results in the association analyses. The possible impact of these extreme high LD values was investigated by comparing the results under the two data settings.

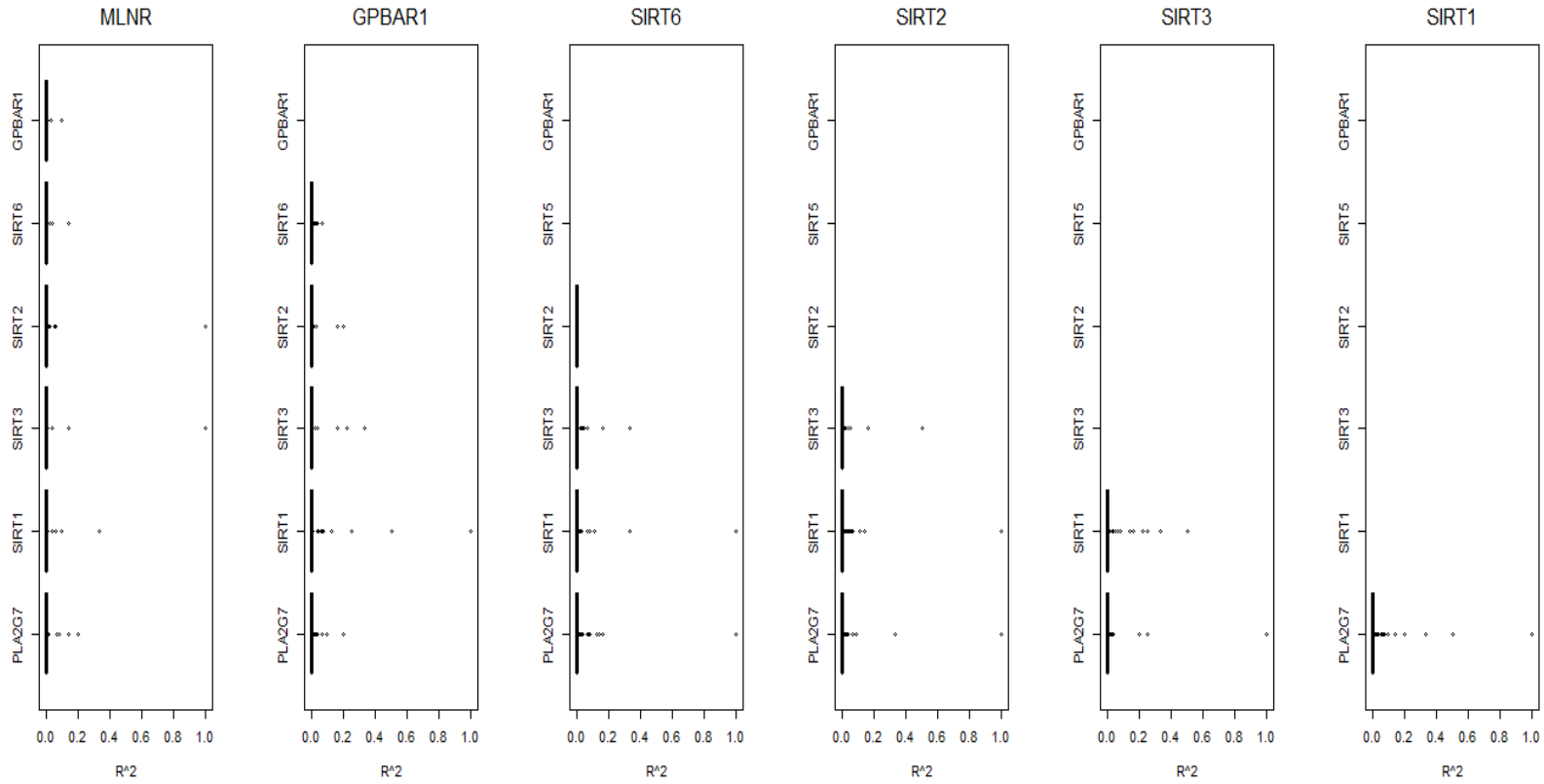

Figure 3.1: The Boxplots for the R_2 values between each pair of genes. Each R_2 value calculates the correlation between two SNPs from different genes. In each boxplot, it present the minimum, first quartile, median, third quartile, maximum and the outliers of the R_2 values between two genes.

Table 3.1: The averages and medians of R^2 values between the SNPs from two different genes. The numbers without parentheses are the averages and the numbers with parentheses are the median values.

	MLNR	GPBAR1	SIRT6	SIRT2	SIRT3	SIRT1
GPBAR1	0.0002094 (5.2×10^{-7})					
SIRT6	0.0002233 (5.2×10^{-7})	0.0002157 (7.8×10^{-7})				
SIRT2	0.001019 (5.2×10^{-7})	0.0003466 (7.8×10^{-7})	8.6×10^{-6} (2.6×10^{-7})			
SIRT3	0.001654 (7.8×10^{-7})	0.0007237 (1.4×10^{-6})	0.0004797 (1.6×10^{-6})	0.0004366 (1.8×10^{-6})		
SIRT1	0.0003204 (5.2×10^{-7})	0.000914 (7.8×10^{-7})	0.0004306 (1.0×10^{-6})	0.0003931 (1.0×10^{-6})	0.0006227 (1.8×10^{-6})	
PLA2G7	0.0004208 (5.2×10^{-7})	0.000325 (7.8×10^{-7})	0.0004535 (1.0×10^{-6})	0.0005154 (1.3×10^{-6})	0.0004045 (1.6×10^{-6})	0.0007492 (1.0×10^{-6})

In the simulation, suppose C genes are considered as the causal genes which determine the disease risk. The simulation data was generated by using the following formula

$$Y_i = X_i\gamma + G_{i11}\gamma_{11}^G + \cdots + G_{i1M_1}\gamma_{1M_1}^G + \cdots + G_{iC1}\gamma_{C1}^G + \cdots + G_{iCM_C}\gamma_{CM_C}^G + \varepsilon.$$

where $G_{ic1}, \ldots, G_{icM_c}$ for $c = 1, \ldots, C$ are the genotype information from M_c SNPs within the cth causal gene for sample i, and γ_{cj}^G for $c = 1, \ldots, C$ and $j = 1, \ldots, M_c$ are the coefficients for each SNPs on the causal genes. The coefficient γ_{cj}^G was calculated using the following equation

$$\gamma_{cj}^G = \log\left\{\frac{\alpha_G^c}{(1-\alpha_G^c)q_j} + 1\right\}, \quad j = 1, \ldots, M_c, \tag{3.7}$$

where q_j is the minor allele frequency for jth variant on the cth causal gene, α_G^c is the marginal attributable risk for cth gene effect. The spirit of using such formula is that we wanted to control the overall population attributable risk (PAR) for each causal gene. By assigning each variant within a causal gene using a same marginal PAR α_G^c, low effect for each variant can be ensured [12]. In addition, $\varepsilon \sim N(0, \sigma I_N)$ and $\sigma = 1$ was used for all scenarios in this simulation study.

In this simulation study, we considered a sample size of $N = 1000$, and 100 data sets were generated under each scenario. Based on the number of SNPs in each gene, we considered genes MLNR (24 SNPs), GPBAR1 (44 SNPs), and SIRT1 (96 SNPs) to consist the combinations of the causal genes. There are 7 scenarios under each data setting where one gene at a time was used as the causal gene (3 scenarios); two genes were taken as the causal genes (3 scenarios); and all the three genes were assigned to be the causal genes (1 scenario). In the simulation, three methods were considered as comparisons to the proposed method. The first competing method is the counting based burden approach [14] where genotype information within a same gene is summed up to constitute a "super-variant" for each gene. Then the variable selection using the counting based burden test is based on the model of $Y = X\gamma + G_1^*\gamma_1 + \cdots + G_7^*\gamma_7$ where G_i^* for $i = 1, \ldots, 7$ are the genetic information for each super-variant. The second competing method uses principal component technique to obtain the first several leading principal components maintaining 90% of the total genetic information for each gene. Considering the principal components for each gene as a group of variables, variable selection is then conducted by applying the group lasso technique [55]. The third competing method applies the group lasso method on all genetic information where genotypes within each gene were considered as a group of variables [55].

We presented the results for each of the 7 scenarios under the two data settings in tables. In each table, we provided the marker numbers for each gene which fall into the range from 24 to 114. In addition, we also showed the rank of similarity r_k for $k = 1, \ldots, 7$ in the same table.

In Table 3.2, we exhibited the results when one gene was used as the causal gene under the assumption that the genetic information is not correlated with each other. There were three genes we considered in the simulation which are MLNR, GPBAR1 and SIRT1. By using the Equation (3.7) to define the coefficient for each SNP, genes with smaller gene sizes have larger marginal PAR values. To adjust such effect due to the different α_G's, we used a smaller group PAR value when genes with fewer SNPs were used as the causal genes. When large genes with more SNPs were used as the causal genes, bigger group PAR values were used. In this simulation, we assigned gene MLNR a group PAR value of 0.05 when it was the causal gene,

which leaded to a marginal PAR of $\alpha_G = 0.002083$. The group PAR for gene GPBAR1 was 0.08 which resulted $\alpha_G = 0.001818$. When SIRT1 was used as the causal gene, the group PAR was 0.12 and thus the marginal PAR $\alpha_G = 0.00125$. The identification rates for the causal gene and the false positives for the non-causal genes were compared by using the proposed variance component based method ("VC-Based"), the counting based burden method ("Burden"), group lasso applying on principal components ("PCA"), and group lasso method on all genotype information ("GRPLASSO"). Among the 4 competing methods, the first two approaches are the type of individual variable selection based methods, and the latter two approaches are in the category of grouped variable selection based methods.

When MLNR was used as the causal gene, the proposed method is the best among all the four approaches in the comparison. The results by applying variance component based method indicated that the causal gene MLNR is the gene significantly associated with the disease risk. The counting based burden approach identified gene MLNR highly associated with the disease with percentage of 42%. Even though group lasso applying on the principal components identified 75% times for the causal gene MLNR, such approach also indicated other genes with higher probabilities as the genes strongly associated with the disease, such as genes SIRT3 (95%), SIRT1 (90%) and PLA2G7 (100%). As a result, if the gene MLNR was identified as the significant gene by using this method, SIRT3, SIRT1 and PLA2G7 should also be included as disease associated genes. However, only the MLNR was used as the causal gene under this scenario. Therefore, the results by using the PCA method cannot identify the true causal gene under this scenario. When group lasso method was applied on all the genotype information, it identified 10% times that the MLNR is associated with the diseases. Meanwhile, it also identified 14% times that the gene SIRT3 is associated with the disease risk. Both the identification rates and false positives by using the GRPLASSO method were lower than the other three methods.

When GPBAR1 was used as the causal gene for the disease, the proposed method indicated the significant association existed between the true causal gene GPBAR1 and disease risk with probability of 97%. There is 3% chance that the non-causal gene PLA2G7 was indicated as

the important gene. But such low positive false rate is in a reasonable range. The counting based burden method identified GPBAR1 highly associated with disease 89% times, which is the second best among the four approaches. The results using the group lasso technique on principal components showed a high percentage in the association between the causal gene GPBAR1 and the phenotypes, but the probabilities of getting false positives for the non-causal genes are also high. When group lasso technique was applied on the whole genetic information, it indicated that GPBAR1 and SIRT3 are associated with the disease with the probability of 10% where the latter probability for SIRT3 is false positive rate.

Very similar results are observed when SIRT1 was the causal gene. The proposed method provides the best results among all the four methods with identification rate of 93% for the causal gene and low false positive rates for the non-causal genes. The counting based burden approach identified the causal gene with the probability of 68%. The group lasso method applying on the principal components showed both high identification rate for the causal gene and high false positive rates for the non-causal genes. When the group lasso method was used on the whole genotype information, it identified the causal gene with the probability of 75% which is higher than the counting based burden approach. But it had 14% and 11% false positive rates for the non-causal genes SIRT3 and PLA2G7, respectively.

Through the results shown in the Table 3.2 when one gene was used as the causal gene, we observed that the proposed method is the best approach among the four methods under these three scenarios. The counting based burden approach provided the second best results under the same scenarios. Even though group lasso method applying on the principal components identified the causal gene with a high probability, it also indicated that other non-causal genes are highly associated with the disease risk. When group lasso technique was used on the whole genotype information, it gave a lower identification rate for the causal gene. Meanwhile, it may indicate other non-causal genes to be significantly associated with the phenotypes by using the PCA method. Though the last method provided good identification rates in some situations, the performance by using this method was not consistent according to the results when each of

the three genes was considered as the causal gene.

Table 3.2: The gene identification rates (percents) and the probabilities (percents) of detecting the false positives over 100 runs when one of the three genes MLNR, GPBAR1, and SIRT1 is used as the causal gene. All the seven genes are assumed to be mutually uncorrelated where we use ✠ to denote the causal gene under each scenario. When MLNR is the causal gene, a group PAR value of 0.05 is assigned to the causal gene. When GPBAR1 is the causal gene, a group PAR value of 0.08 is assigned to the causal gene. When SIRT1 is the causal gene, a group PAR value of 0.12 is assigned to the causal gene.

	MLNR	GPBAR1	SIRT6	SIRT2	SIRT3	SIRT1	PLA2G7
num.snps.SG	24	44	56	60	60	96	114
Rank.S.SG	16	31	42	45	51	69	89
group PAR=0.05							
	✠						
VC-Based	100	0	0	0	0	0	0
Burden	42	1	0	1	1	0	0
PCA	75	24	68	22	95	90	100
GRPLASSO	10	0	0	0	14	4	6
group PAR=0.08							
		✠					
VC-Based	0	97	0	0	0	0	3
Burden	0	89	0	0	0	0	0
PCA	75	90	72	20	96	92	99
GRPLASSO	5	10	0	0	10	6	5
group PAR=0.12							
						✠	
VC-Based	0	0	1	1	0	93	9
Burden	0	1	0	0	1	68	0
PCA	68	33	66	20	94	99	100
GRPLASSO	1	0	0	0	14	75	11

We presented the results under the scenarios when two genes were used as the causal genes simultaneously assuming the genes were not correlated with each other in Table 3.3. Genes MLNR, GPBAR1 and SIRT1 were considered again where two of them were chosen to be the causal genes at a time. We used a same group PAR value for the two causal genes under each scenario though different gene sizes have different marginal PAR values. When MLNR was

used as one of the two causal genes, the probabilities that MLNR was identified by using the proposed method are high above 90%. But under the same scenarios, the other causal genes with more SNPs have lower detection rates. Such phenomenon is due to the different marginal PAR values where the gene with smaller gene size has larger α_G value and the gene with more SNPs has smaller α_G. It is observed that when MLNR (24 SNPs) and GPBAR1 (44 SNPs) were used as the two causal genes, GPBAR1 is identified 34% times by using the proposed method. When MLNR and SIRT1 (96 SNPs) were considered as the two causal genes, SIRT1 is only identified by 11% times. Though similar pattern is observed under the scenario when GPBAR1 and SIRT1 were the two causal genes, we noticed that the difference between the identification rates for the two causal genes is not as big as the difference under the scenario when MLNR and SIRT1 were considered as the two causal genes. It may imply that marginal PAR α_G affects the gene identification rates by using the proposed method. However, the counting based burden approach does not have the similar pattern. Instead, the numbers of SNPs, the marginal PAR values and the minor allele frequencies have effects on the gene identification rates simultaneously when the counting based burden method is applied. When group lasso technique was applied on the principal components, both the causal gene identification rates and non-causal gene false positive rates are high. Thus, this approach may not be able to identify the causal genes correctly. Using the group lasso on the whole genetic information, the chance that the causal gene was identified as significantly disease-associated gene is low. The probabilities that the noncausal genes are falsely identified to be highly associated with the phenotype could be higher than the causal gene identification rates in some cases.

In Table 3.4, we showed the results when different group PAR values were assigned to the two causal genes MLNR and SIRT1, respectively. Compared to the results in the Table 3.7 where the same two genes were used as the causal genes with a same group PAR, the identification rates for the causal gene SIRT1 are increased by using both the proposed method and the counting based burden approach. It may explain the fact that the identification rates for larger genes are lower when a same group PAR value was assigned to the genes with different SNP

Table 3.3: The gene identification rates (percents) and the probabilities (percents) of detecting the false positives over 100 runs when two of the three genes MLNR, GPBAR1, and SIRT1 are used as the causal genes. All the seven genes are assumed to be mutually uncorrelated where we use ✠'s to denote the two causal genes under each scenario. When the pair of MLNR and GPBAR1 and the pair of MLNR and SIRT1 are considered as the causal genes, a group PAR value of 0.05 are assigned to the two pair of the causal genes. When GPBAR1 and SIRT1 are the two causal genes, a group PAR value of 0.10 is assigned to the two causal genes.

	MLNR	GPBAR1	SIRT6	SIRT2	SIRT3	SIRT1	PLA2G7
num.snps.SG	24	44	56	60	60	96	114
Rank.S.SG	16	31	42	45	51	69	89
group PAR=0.05							
	✠	✠					
VC-Based	94	34	0	0	0	0	1
Burden	24	52	0	1	0	0	0
PCA	74	72	73	24	94	93	99
GRPLASSO	2	3	0	0	13	4	11
group PAR=0.05							
	✠					✠	
VC-Based	99	0	0	0	0	11	0
Burden	28	1	0	1	1	11	0
PCA	76	23	67	21	95	98	99
GRPLASSO	5	0	0	0	15	26	10
group PAR=0.10							
		✠				✠	
VC-Based	0	100	0	0	1	73	1
Burden	1	96	0	0	0	43	0
PCA	85	95	74	26	96	100	100
GRPLASSO	11	28	4	4	13	51	17

numbers shown in the Table 3.7. The latter two methods could not provide good results for the gene selection for this scenario.

In Table 3.5, we presented the results when three genes were used as the causal genes assuming the genes are mutually uncorrelated. We assigned the group PAR to be 0.05 for the three causal genes which are MLNR, GPBAR1 and SIRT1. The proposed method identified the most three important genes associated with the disease are the three causal genes. However, the gene SIRT1 with 96 SNPs has been only identified 11% times while the gene MLNR with 24 SNPs has been identified 92% times. This situation is similar with the cases when two causal

Table 3.4: The gene identification rates (percents) and the probabilities (percents) of detecting the false positives over 100 runs when genes MLNR and SIRT1 are used as the causal genes. Different group PAR values of 0.05 and 0.08 are assigned to the two causal genes MLNR and SIRT1, respectively. All the seven genes are assumed to be mutually uncorrelated where we use ✠'s to denote the two causal genes under each scenario.

	MLNR	GPBAR1	SIRT6	SIRT2	SIRT3	SIRT1	PLA2G7
num.snps.SG	24	44	56	60	60	96	114
Rank.S.SG	11	32	40	40	56	63	81
group $\text{PAR}_{\text{MLNR}} = 0.05$ & group $\text{PAR}_{\text{SIRT1}} = 0.08$							
	✠					✠	
VC-Based	90	0	0	0	0	48	0
Burden	26	1	0	1	1	34	0
PCA	73	25	69	22	94	98	99
GRPLASSO	6	1	1	1	15	48	14

genes were considered under each scenario shown in the Table 3.3. The three genes identified most times by applying the counting based burden approach are the three causal genes as well. Instead of the gene with the smallest SNP number, the gene GPBAR1 was identified most of the times as the important gene by this method. It may support our saying that multiple factors have an effect on the results by using this method. The other two competing methods where group lasso technique was applied, the results are not reliable which may indicate the non-causal genes as the most significant genes associated with the disease risk.

The same 7 scenarios were considered for the data setting where LD structures are maintained among the genes. Thus the assumption that genes are mutually uncorrelated is not guaranteed. The results when one causal gene was considered are presented in Table 3.6. When two causal genes were used, the results are exhibited in Table 3.7. Table 3.9 showes the results when three genes were applied as the causal genes. When one gene at a time was considered as the causal gene, we observed very similar findings as for the data setting when genes are not correlated. Under these scenarios, there are some low false positive rates for the non-causal genes by using both the proposed method and the counting based burden approach. Such false positive rates may be due to the underlying correlation among the genes, though these values are all below 5% within an acceptable range. When two or three genes were considered as the

Table 3.5: The gene identification rates (percents) and the probabilities (percents) of detecting the false positives over 100 runs when all the three genes MLNR, GPBAR1, and SIRT1 are used as the causal genes. All the seven genes are assumed to be mutually uncorrelated where we use ✠'s to denote the three causal genes under each scenario. A group PAR value of 0.05 is assigned to the three causal genes.

	MLNR	GPBAR1	SIRT6	SIRT2	SIRT3	SIRT1	PLA2G7
num.snps.SG	24	44	56	60	60	96	114
Rank.S.SG	16	31	42	45	51	69	89
group PAR=0.05							
	✠	✠				✠	
VC-Based	92	29	0	0	0	9	1
Burden	18	50	0	1	1	9	0
PCA	74	71	75	23	96	98	99
GRPLASSO	3	3	1	1	13	24	19

causal genes at a time, we assigned a same group PAR value which resulted in different marginal PAR values. Under these scenarios, the smaller gene with higher α_G is detected more than the gene with larger SNP numbers and lower marginal PAR by using the proposed method. In Table 3.8, we showed the results when the genes MLNR and SIRT1 were considered as the two causal genes where the number of SNPs in SIRT1 is quadruple of the SNP number in MLNR. We assigned the group PAR for MLNR to be 0.05 and the group PAR for SIRT1 to be 0.08. Compared to the results in the Table 3.7 where the same two genes were used as the causal genes, the probability of SIRT1 being identified is increased by using both the proposed method and the counting based burden approach. It may explain the fact that the identification rates for larger genes are lower when the same group PAR value was assigned to genes with different SNP numbers. The latter two methods may not be able to correctly identify the causal genes for all the scenarios.

3.4 Real Data Applications

We applied the proposed method to the real data from the CoLaus study. This study aims to find out if there are any genetic determinants associated with risk factors in cardiovascular

Table 3.6: The gene identification rates (percents) and the probabilities (percents) of detecting the false positives over 100 runs when one of the three genes MLNR, GPBAR1, and SIRT1 is used as the causal gene. The LD structure is maintained among the seven genes where we use ✠ to denote the causal gene under each scenario. When MLNR is the causal gene, a group PAR value of 0.05 is assigned to the causal gene. When GPBAR1 is the causal gene, a group PAR value of 0.08 is assigned to the causal gene. When SIRT1 is the causal gene, a group PAR value of 0.12 is assigned to the causal gene.

	MLNR	GPBAR1	SIRT6	SIRT2	SIRT3	SIRT1	PLA2G7
num.snps.SG	24	44	56	60	60	96	114
Rank.S.SG	11	32	40	40	56	63	81
group PAR=0.05							
	✠						
VC-Based	99	3	3	3	4	4	5
Burden	15	2	1	1	1	0	1
PCA	64	19	68	19	99	96	100
GRPLASSO	2	1	0	0	31	8	12
group PAR=0.08							
		✠					
VC-Based	0	98	0	2	1	2	2
Burden	1	93	1	3	0	3	0
PCA	78	94	75	56	98	91	99
GRPLASSO	1	14	1	0	15	9	22
group PAR=0.12							
						✠	
VC-Based	0	0	1	0	1	99	3
Burden	0	1	1	0	0	83	1
PCA	62	27	67	18	97	100	100
GRPLASSO	0	0	0	0	14	84	11

diseases among the Caucasian population [68]. To explore the association at gene set level, Body Mass Index (BMI) was considered as the disease risk which is used as the obesity indicator. The obesity could increase the chance for many diseases including the heart disease [79]. There are 1961 individuals available with the phenotypes. Thus gene set association analysis was conducted using the information from those individuals. The genetic information on the same 7 genes in the simulation study was considered. The 7 seven genes are MLNR, GPBAR1, SIRT6, SIRT2, SIRT3, SIRT1 and PLA2G7. In addition, there are 10 confounders are included in the model as the baseline information, such as age, gender, smoking stats, alcohol drinking status,

Table 3.7: The gene identification rates (percents) and the probabilities (percents) of detecting the false positives over 100 runs when two of the three genes MLNR, GPBAR1, and SIRT1 are used as the causal genes. The LD structure is maintained among the seven genes where we use ✠ to denote the two causal genes under each scenario. When the pair of MLNR and GPBAR1 and the pair of MLNR and SIRT1 are considered as the causal genes, a group PAR value of 0.05 are assigned to the two pair of the causal genes. When GPBAR1 and SIRT1 are the two causal genes, a group PAR value of 0.10 is assigned to the two causal genes.

	MLNR	GPBAR1	SIRT6	SIRT2	SIRT3	SIRT1	PLA2G7
num.snps.SG	24	44	56	60	60	96	114
Rank.S.SG	11	32	40	40	56	63	81
group PAR=0.05							
	✠	✠					
VC-Based	62	56	0	0	0	0	1
Burden	11	72	1	3	1	2	0
PCA	74	82	73	38	99	95	100
GRPLASSO	0	7	0	0	31	8	21
group PAR=0.05							
	✠					✠	
VC-Based	87	2	2	2	3	21	4
Burden	17	1	1	0	1	27	1
PCA	72	20	65	20	100	96	100
GRPLASSO	4	0	0	0	31	40	14
group PAR=0.10							
		✠				✠	
VC-Based	0	99	0	0	0	86	0
Burden	0	97	1	1	0	52	0
PCA	81	96	83	74	97	100	100
GRPLASSO	12	26	12	12	24	77	37

physical activity status and 5 principal components values. Table 3.10 presented the results on the real data application. The gene MLNR was identified as the gene significantly associated with BMI by using the proposed method. The counting based burden approach identified none of the genes are in high association with the obesity indicator. The two group lasso approaches, PCA and GRPLASSO, indicated all the genes are the obesity-associated genes. Based on the results obtained in the simulation study, the results using these two methods might not be reliable. The gene MLNR identified by using the proposed method in this real data application was considered as the causal gene in the simulation study discussed in Section 3.3. The results

Table 3.8: The gene identification rates (percents) and the probabilities (percents) of detecting the false positives over 100 runs when genes MLNR and SIRT1 are used as the causal genes. Different group PAR values of 0.05 and 0.08 are assigned to the two causal genes MLNR and SIRT1, respectively. The LD structure is maintained among the seven genes. We use ✠ to denote the two causal genes under this scenario.

	MLNR	GPBAR1	SIRT6	SIRT2	SIRT3	SIRT1	PLA2G7
num.snps.SG	24	44	56	60	60	96	114
Rank.S.SG	11	32	40	40	56	63	81
group-PAR$_{\text{MLNR}} = 0.05$ & group-PAR$_{\text{SIRT1}} = 0.08$							
	✠					✠	
VC-Based	54	2	2	2	2	61	2
Burden	17	1	1	0	1	55	1
PCA	75	19	66	19	99	100	100
GRPLASSO	4	0	0	0	31	58	15

showed in Table 3.2 and Table 3.6 provided the identification rates for the gene MLNR when it was used as the causal gene for both data settings. Through the findings from these two tables, we could indicate that the gene MLNR and BMI are in significant association by applying the proposed method.

3.5 Discussion

As a dynamic and ongoing area, researchers are making great effort in the field of gene set association analysis. We proposed a GSAA method which identifies significant genes associated with the disease by using a variance component based method. In this work, we introduced a linear mixed model assuming additive gene effect for quantitative traits. In such a preliminary study, we are interested in investigating whether the correlation among the genes affects the results by using the proposed GSAA method. Through the simulation study and real data application, we observed that the proposed method could successfully identify the significant genes under different scenarios.

In the simulation study, we noticed that it is important to provide an appropriate range aiming to find out the optimal tuning parameter λ. In Table 3.11, we observed that the causal

Table 3.9: The gene identification rates (percents) and the probabilities (percents) of detecting the false positives over 100 runs when all the three genes MLNR, GPBAR1, and SIRT1 are used as the causal genes. The LD structure is maintained among the seven genes where we use ✠ to denote the three causal genes under each scenario. A group PAR value of 0.05 is assigned to the three causal genes.

	MLNR	GPBAR1	SIRT6	SIRT2	SIRT3	SIRT1	PLA2G7
num.snps.SG	24	44	56	60	60	96	114
Rank.S.SG	11	32	40	40	56	63	81
group PAR=0.05							
	✠	✠				✠	
VC-Based	63	51	0	0	0	16	2
Burden	11	64	1	3	1	16	0
PCA	75	75	72	41	99	96	100
GRPLASSO	5	5	2	2	31	34	25

Table 3.10: The results for the gene selection using the CoLaus Study Data.

	MLNR	GPBAR1	SIRT6	SIRT2	SIRT3	SIRT1	PLA2G7
num.snps.SG	24	44	56	60	60	96	114
Rank.S.SG	26	47	60	64	73	102	123
VC-Based	1	0	0	0	0	0	0
Burden	0	0	0	0	0	0	0
PCA	1	1	1	1	1	1	1
GRPLASSO	1	1	1	1	1	1	1

gene MLNR was identified as the significant gene by using the proposed method, but the non-causal genes were also identified as important genes with high probabilities which are false positives. In this simulation, we pre-specified the search of the optimal tuning parameter in the range of $[e^{-10}, e^{5}]$. When the group PAR of 0.05 was assigned to the causal gene shown in the Table 3.6, the proposed method have a good performance in identifying the significant gene. However, under the scenario when the group PAR is 0.08 shown in the Table 3.11, the proposed method gave high probabilities of false positives. Such phenomenon is due to the reason that the optimal tuning parameter minimizing the BIC score does not lie in the pre-specified interval of $[e^{-10}, e^{5}]$. Thus, the λ value corresponding to the local minimum BIC score lies on the boundary of the pre-specified λ's range in the simulation. As a consequence, we observed these high false

positive rates by using the proposed method in the Table 3.11. With respect to the false positive rates by using the group lasso technique on either principal components or the whole information of the genotypes, the choices of the tuning parameters could also be a problem. However, the high similarities among the first several principal components within each gene could be another reason causing the high false positive rates. When the group lasso technique was used on the whole genotype information, both the causal gene identification rate and the false positive rate for the non-causal genes are lower than the other methods may indicate that such method might not be an appropriate method to identify significant disease-associated genes.

Table 3.11: The gene identification rates (percents) and the probabilities (percents) of detecting the false positives over 100 runs when MLNR is used as the causal gene. A group PAR value of 0.08 is assigned to the causal gene. High false positives are present because the optimal λ is outside the range of λ's considered in this simulation study. Under this scenario, the LD structure is maintained among the seven genes where we use ✠ to denote the causal gene.

	MLNR	GPBAR1	SIRT6	SIRT2	SIRT3	SIRT1	PLA2G7
num.snps.SG	24	44	56	60	60	96	114
Rank.S.SG	11	32	40	40	56	63	81
group PAR=0.08							
	✠						
VC-Based	100	20	60	60	61	61	61
Burden	51	1	1	2	1	0	1
PCA	77	20	67	20	97	96	100
GRPLASSO	5	1	0	0	38	8	14

To demonstrate the issue of obtaining the optimal λ, we used the real data application to plot the change of BIC scores corresponding to each λ value in the range of $e^{-10} \leq \lambda \leq e^{6}$, which is shown at the top of the Figure 3.2. Because the BIC score for $\lambda = e^{6}$ is much larger than the other BIC scores, all other BIC values are distributed on a horizontal line shown in the same plot. For this reason, it may be hard to identify the optimal λ through this plot. To present in more details with respect to the point minimizing the BIC score, we present two zoomed-in plots shown at the bottom of the Figure 3.2. The left bottom plot shows all other

BIC scores except the one for $\lambda = e^(6)$. In this plot, we observe the decreasing in the change of BIC scores for $e^{-10} \leq \lambda \leq e^{5.7974}$. In order to identify the point which minimizes the BIC score, we exhibit the BIC scores for $e^{1.9494} \leq \lambda \leq e^{5.7974}$ in the plot at the right bottom of the Figure 3.2. According to the observation in the right bottom plot, the minimum BIC score is obtained when $\lambda = e^{5.7974} = 329.46$. Suppose we used the interval of $[e^{-10}, e^{5}]$ to search the optimal tuning parameter, only the local minimum BIC score could be reached at the boundary value of $\lambda = 148.41$. If we extended the search from $[e^{-10}, e^{5}]$ to $[e^{-10}, e^{6}]$, the global minimum BIC could be obtained. Therefore, the pre-specified λ values may have an influence on the results by using the proposed method according to the simulation study. The problem of providing an appropriate range for pre-specified λ values should be taken into account in future work.

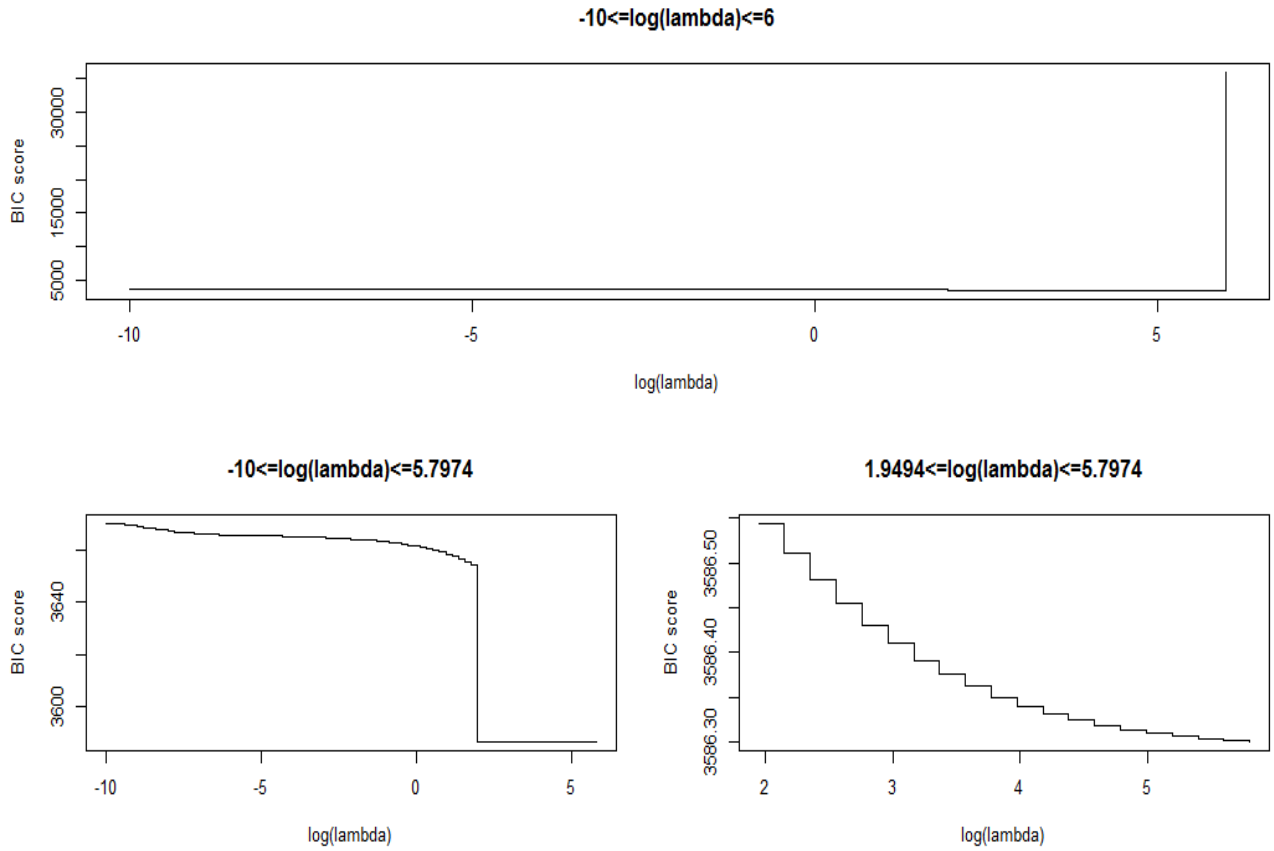

Figure 3.2: Plots of BIC scores corresponding to pre-specified $\log(\lambda)$'s using the CoLaus study data. The first plot at the top shows the BIC scores for $e^{-10} \leq \lambda \leq e^{6}$. The second plot at the left bottom shows the BIC scores for $e^{-10} \leq \lambda \leq e^{5.7974}$. The third plot at the right bottom shows the BIC scores for $e^{1.9494} \leq \lambda \leq e^{5.7974}$.

Through the simulation under two data settings where genes are either mutually uncorrelated or LD structures are maintained among the genes, we observed that there might be few false positive rates due to the correlation among genes. It is advocated that incorporating correlation information in the gene set analysis may help to detect more associations [80, 81, 82]. Thus, the topology information for the gene set could be considered and be incorporated in the similarity matrices for each gene.

As an extension of the current work, gene-gene interactions may be taken into account as well in future. Additionally, variance component based methods for general traits, such as binary traits, could be considered for gene set association analysis using generalized linear mixed models.

REFERENCES

[1] G. Danaei, M. M. Finucane, Y. Lu, G. M. Singh, M. J. Cowan, C. J. Paciorek, J. K. Lin, F. Farzadfar, Y.-H. Khang, G. A. Stevens, M. Rao, M. K. Ali, L. M. Riley, C. A. Robinson, and M. Ezzati, "National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 27 million participants," *The Lancet*, vol. 378, no. 9785, pp. 31–40, 2011.

[2] N. W. Gillham, *Genes, Chromosomes, And Disease: From Simple Traits, To Complex Traits, To Personalized Medicine*. FT Press Science, 2011.

[3] World Health Organization, "World health organization fact sheet fact sheet no 307: Asthma," May 2011. $http://www.who.int/mediacentre/factsheets/fs307/en/$.

[4] J. Zhang, P. D. Paré, and A. J. Sandford, "Recent advances in asthma genetics," *Respiratory Research*, vol. 9, no. 1, pp. 1–8, 2008.

[5] C. Palmer, A. Irvine, A. Terron-Kwiatkowski, Y. Zhao, H. Liao, S. Lee, D. Goudie, A. Sandilands, L. Campbell, F. Smith, G. O'Regan, R. Watson, J. Cecil, S. Bale, J. Compton, J. DiGiovanna, P. Fleckman, S. Lewis-Jones, G. Arseculeratne, A. Sergeant, and C. Munro, "Common loss-of-function variants of the epidermal barrier protein major predisposing factor for atopic dermatitis," *Nature Genetics*, vol. 38, pp. 441–446, March 2006.

[6] V. Moskvina and K. M. Schmidt, "On multiple-testing correction in genome-wide association studies," *Genetic Epidemiology*, vol. 32, p. 567C573, 2008.

[7] R. Fisher, *Statistical methods for research workers*. London: OLiver and Boyd, 4 ed., 1932.

[8] M. L. N. Monnat Pongpanich and J.-Y. Tzeng, "On the aggregation of multimarker information for marker-set and sequencing data analysis: genotype collapsing vs.similarity collapsing," *Frontiers in Genetics:Statistical Genetics and Methodology*, vol. 2, pp. 1–14, 2012.

[9] W. J. Gauderman, C. Murcray, F. Gilliland, and D. V. Conti, "Testing association between disease and multiple snps in a candidate gene," *Genetic Epidemiology*, vol. 31, no. 5, pp. 383–395, 2007.

[10] K. Wang and D. Abbott, "A principal components regression approach to multilocus genetic association studies," *Genetic Epidemiology*, vol. 32, no. 2, pp. 108–118, 2008.

[11] B. Li and S. M. Leal, "Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data," *The American Journal of Human Genetics*, vol. 83, pp. 311 – 321, 2008.

[12] B. E. Madsen and S. R. Browning, "A groupwise association test for rare mutations using a weighted sum statistic," *PLoS Genet*, vol. 5, p. e1000384, 02 2009.

[13] A. L. Price, G. V. Kryukov, P. I. W. de Bakker, S. M. Purcell, J. Staples, L.-J. Wei, and S. R. Sunyaev, "Pooled Association Tests for Rare Variants in Exon-Resequencing Studies," *The American Journal of Human Genetics*, vol. 86, pp. 832–838, 2010.

[14] W. Pan, "Asymptotic tests of association with multiple snps in linkage disequilibrium," *Genetic epidemiology*, vol. 33, no. 6, pp. 497–507, 2009.

[15] F. Han and W. Pan, "A data-adaptive sum test for disease association with multiple common or rare variants," *Human Heredity*, vol. 70, no. 1, pp. 42–54, 2010.

[16] B. M. Neale, M. A. Rivas, B. F. Voight, D. Altshuler, B. Devlin, M. Orho-Melander, S. Kathiresan, S. M. Purcell, K. Roeder, and M. J. Daly, "Testing for an unusual distribution of rare variants," *PLoS Genetics*, vol. 7, p. e1001322, 2011.

[17] S. Lee, M. C. Wu, and X. Lin, "Optimal tests for rare variant effects in sequencing association studies," *Biostatistics*, vol. 13, no. 4, pp. 762–775, 2012.

[18] L. C. Kwee, D. Liu, X. Lin, D. Ghosh, and M. P. Epstein, "A powerful and flexible multilocus association test for quantitative traits," *The American Journal of Human Genetics*, vol. 82, pp. 386–397, February 2008.

[19] M. C. Wu, P. Kraft, M. P. Epstei, D. M. Taylor, S. J. Chanock, D. J. Hunter, and X. Lin, "Powerful snp-set analysis for case-control genome-wide association studies," *The American Journal of Human Genetics*, vol. 86, pp. 929–942, 2010.

[20] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin, "Rare variant association testing for sequencing data using the sequence kernel association test (skat).," *The American Journal of Human Genetics*, vol. 89, pp. 82–93, 2011.

[21] J.-Y. Tzeng, D. Zhang, S.-M. Chang, D. C. Thomas, and M. Davidian, "Gene-trait similarity regression for multimarker-based association analysis.," *Biometrics*, vol. 65, pp. 822–832, 2009.

[22] J.-Y. Tzeng, D. Zhang, M. Pongpanich, C. Smith, M. I. McCarthy, M. M. Sale, B. B. Worrall, F.-C. Hsu, D. C. Thomas, and P. F. Sullivan, "Studying gene and gene-environment effects of uncommon and common variants on continuous traits: A marker-set approach using gene-trait similarity regression," *The American Journal of Human Genetics*, vol. 89(2), pp. 277–288, 2011.

[23] J. J. Goeman, S. A. van de Geer, F. de Kort, and H. C. van Houwelingen, "A global test for groups of genes: testing association with a clinical outcome," *Bioinformatics*, vol. 20, pp. 93–99, 2004.

[24] J.-Y. Tzeng and D. Zhang, "Haplotype-based association analysis via variance component score test.," *The American Journal of Human Genetics*, vol. 81, pp. 927–938, 2007.

[25] D. J. Schaid, "Genomic Similarity and Kernel Methods I: Advancements by Building on Mathematical and Statistical Foundations," *Human Heredity*, vol. 70, pp. 109–131, 2010.

[26] W. Pan, "Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing," *Genetic Epidemiology*, vol. 35, pp. 211–216, 2011.

[27] D. J. Schaid, "Genomic Similarity and Kernel Methods II: Methods for Genomic Information," *Human Heredity*, vol. 70, pp. 132–140, 2010.

[28] M. C. Wu, A. Maity, S. Lee, E. M. Simmons, Q. E. Harmon, X. Lin, S. M. Engel, J. J. Molldrem, and P. M. Armistead, "Kernel machine snp-set testing under multiple candidate kernels," *Genetic Epidemiology*, vol. 37, no. 3, pp. 267–275, 2013.

[29] J. Wessela and N. J. Schork, "Generalized genomic distance-based regression methodology for multilocus association analysis," *The American Journal of Human Genetics*, vol. 79, pp. 792–806, 2006.

[30] X. Lin, T. Cai, M. C. Wu, Q. Zhou, G. Liu, D. C. Christiani, and X. Lin, "Kernel machine snp-set analysis for censored survival outcomes in genome-wide association studies," *Genetic Epidemiology*, vol. 35, no. 7, pp. 620–631, 2011.

[31] J. Haseman and R. Elston, "The investigation of linkage between a quantitative trait and a marker locus," *Behavior genetics*, vol. 2, no. 1, pp. 3–19, 1972.

[32] J.-Y. Tzeng, B. Devlin, L. Wasserman, and K. Roeder, "On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit," *The American Journal of Human Genetics*, vol. 72, no. 4, pp. 891–902, 2003.

[33] P. Kraft, Y.-C. Yen, D. O. Stram, J. Morrison, and W. J. Gauderma, "Exploiting gene-environment interaction to detect genetic associations," *Human Heredity*, vol. 63, no. 2, pp. 111–119, 2007.

[34] J. van Os and B. P. Rutten, "Gene-environment-wide interaction studies in psychiatry," *Am Journal Psychiatry*, vol. 166, pp. 964–966, 2009.

[35] C. E. Murcray, J. P. Lewinger, and W. J. Gauderman, "Gene-environment interaction in genome-wide association studies," *American Journal of Epidemiology*, vol. 169, no. 2, pp. 219–226, 2009.

[36] L. E. Mechanic, H.-S. Chen, C. I. Amos, N. Chatterjee, N. J. Cox, R. L. Divi, R. Fan, E. L. Harris, K. Jacobs, P. Kraft, S. M. Leal, K. McAllister, J. H. Moore, D. N. Paltoo, M. A. Province, E. M. Ramos, M. D. Ritchie, K. Roeder, D. J. Schaid, M. Stephens, D. C. Thomas, C. R. Weinberg, J. S. Witte, S. Zhang, S. Zöllner, E. J. Feuer, and E. M. Gillanders, "Next generation analytic tools for large scale genetic epidemiology studies of complex diseases," *Genetic Epidemiology*, vol. 36, pp. 22–35, 2012.

[37] D. Thomas, "Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies," *Annual Review of Public Health*, vol. 31, pp. 21–36, 2010.

[38] K. Yu, S. Wacholder, W. Wheeler, Z. Wang, N. Caporaso, M. T. Landi, and F. Liang, "A flexible bayesian model for studying gene-environment interaction," *PLoS Genet*, vol. 8, p. e1002482, 01 2012.

[39] B. Mukherjee and N. Chatterjee, "Exploiting gene-environment independence for analysis of casecontrol studies: An empirical bayes-type shrinkage estimator to trade-off between bias and efficiency," *Biometrics*, vol. 64, no. 3, pp. 685–694, 2008.

[40] M. Y. Park and T. Hastie, "Penalized logistic regression for detecting gene interactions," *Biostatistics*, vol. 9, no. 1, pp. 30–50, 2008.

[41] N. Chatterjee and R. J. Carroll, "Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies," *Biometrika*, vol. 92, no. 2, pp. 399–418, 2005.

[42] X. Lin, S. Lee, D. C. Christiani, and X. Lin, "Test for interactions between a genetic marker set and environment in generalized linear models," *Biostatistics*, 2013.

[43] N. Chatterjee, Z. Kalaylioglu, R. Moslehi, U. Peters, and S. Wacholder, "Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions," *American Journal of Human Genetics*, vol. 79, no. 6, pp. 1002–1016, 2006.

[44] V. K. Mootha, C. M. Lindgren, K.-F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, and J. N. Hirschhorn, "Pgc-1a-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.," *Nature Genetics*, vol. 34, no. 3, p. 267, 2003.

[45] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–15550, 2005.

[46] L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, and P. J. Park, "Discovering statistically significant pathways in expression profiling studies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 38, pp. 13544–13549, 2005.

[47] B. Efron and R. Tibshirani, "On testing the significance of sets of genes," *The Annals of Applied Statistics*, vol. 1, June 2007.

[48] Z. Jiang and R. Gentleman, "Extensions to gene set enrichment," *Bioinformatics*, vol. 23, no. 3, pp. 306–313, 2007.

[49] I. Dinu, J. Potter, T. Mueller, Q. Liu, A. Adewale, G. Jhangri, G. Einecke, K. Famulski, P. Halloran, and Y. Yasui, "Improving gene set analysis of microarray data by sam-gs," *BMC Bioinformatics*, vol. 8, no. 1, p. 242, 2007.

[50] J. J. Goeman and P. Buhlmann, "Analyzing gene expression data in terms of gene sets: methodological issues," *Bioinformatics*, vol. 23, no. 8, pp. 980–987, 2007.

[51] P. Holmans, E. K. Green, J. S. Pahwa, M. A. R. Ferreira, S. M. Purcell, P. Sklar, M. J. Owen, M. C. O'Donovan, and N. Craddock, "Gene Ontology Analysis of GWA Study Data Sets Provides Insights into the Biology of Bipolar Disorder," *Am J Hum Genet*, vol. 85, pp. 13–24, July 2009.

[52] L. Wang, P. Jia, R. D. Wolfinger, X. Chen, and Z. Zhao, "Gene set analysis of genome-wide association studies: Methodological issues and perspectives," *Genomics*, vol. 98, no. 1, pp. 1 – 8, 2011.

[53] L. S. Chen, C. M. Hutter, J. D. Potter, Y. Liu, R. L. Prentice, U. Peters, and L. Hsu, "Insights into colon cancer etiology via a regularized approach to gene set analysis of gwas data," *The American Journal of Human Genetics*, vol. 86, no. 6, pp. 860–871, 2010.

[54] X. Chen, L. Wang, B. Hu, M. Guo, J. Barnard, and X. Zhu, "Pathway-based analysis for genome-wide association studies using supervised principal components," *Genetic Epidemiology*, vol. 34, no. 7, pp. 716–724, 2010.

[55] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, pp. 49–67, 2006.

[56] L. Meier, S. Van De Geer, and P. Bühlmann, "The group lasso for logistic regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 1, pp. 53–71, 2008.

[57] J. Friedman, T. Hastie, and R. Tibshirani, "A note on the group lasso and a sparse group lasso," *arXiv preprint arXiv:1001.0736*, 2010.

[58] J. Huang, S. Ma, H. Xie, and C.-H. Zhang, "A group bridge approach for variable selection," *Biometrika*, vol. 96, no. 2, pp. 339–355, 2009.

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

[60] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[61] P. Carbonetto and M. Stephens, "Integrated analysis of variants and pathways in genome-wide association studies using polygenic models of disease," *ArXiv e-prints*, Aug. 2012.

[62] H. Miao, "Model selection and estimation in additive regression models," *Doctoral dissertation, North Carolina State University*, 2009.

[63] D. Wang, "Model selection and estimation in generalized additive models and generalized additive mixed models," *Doctoral dissertation, North Carolina State University*, 2013.

[64] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. C. amd Alan E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, pp. 747–753, 2009.

[65] E. E. Eichler, J. Flint, G. Gibson, A. Kong, S. M. Leal, J. H. Moore, and J. H. Nadeau, "Missing heritability and strategies for finding the underlying causes of complex disease," *Nature Reviews Genetics*, vol. 11, pp. 446–450, June 2010.

[66] P. Sham and S. Cherny, *Analysis of complex disease association studies [electronic resource] : a practical guide*. London ; Burlington, MA: Academic Press/Elsevier, 2011.

[67] D. Thomas, "Response to 'gene-by-environment experiments: a new approach to finding the missing heritability' by van ijzendoorn et al.," *Nature Reviews Genetics*, vol. 12, no. 12, p. 881, 2011.

[68] M. Firmann, V. Mayor, P. Vidal, M. Bochud, A. Pecoud, D. Hayoz, F. Paccaud, M. Preisig, K. Song, X. Yuan, T. Danoff, H. Stirnadel, D. Waterworth, V. Mooser, G. Waeber, and P. Vollenweider, "The colaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome," *BMC Cardiovascular Disorders*, vol. 8, no. 1, 2008.

[69] T. Cai, G. Tonini, and X. Lin, "Kernel machine approach to testing the significance of multiple genetic markers for risk prediction," *Biometrics*, vol. 67, pp. 975–986, September 2011.

[70] R. C. Elston, S. Buxbaum, K. B. Jacobs, and J. M. Olson, "Haseman and elston revisited," *Genetic Epidemiology*, vol. 19, pp. 1–17, 2000.

[71] T. Wang and R. C. Elston, "Two-level haseman-elston regression for general pedigree data analysis," *Genetic Epidemiology*, vol. 29, pp. 12–22, 2005.

[72] D. Zhang and X. Lin, "Hypothesis testing in semiparametric addictive mixed models," *Biostatistics*, vol. 4, pp. 57–74, 2003.

[73] S. Basu and W. Pan, "Comparison of statistical tests for disease association with rare variants," *Genetic Epidemiology*, vol. 35, pp. 606–619, 2011.

[74] A. Helgason, S. Pálsson, G. Thorleifsson, S. F. A. Grant, V. Emilsson, S. Gunnarsdottir, A. Adeyemo, Y. Chen, G. Chen, I. Reynisdottir, R. Benediktsson, A. Hinney, T. Hansen, G. Andersen, K. Borch-Johnsen, T. Jorgensen, H. Schäfer, M. Faruque, A. Doumatey, J. Zhou, R. L. Wilensky, M. P. Reilly, D. J. Rader, Y. Bagger, C. Christiansen, G. Sigurdsson, J. Hebebrand, O. Pedersen, U. Thorsteinsdottir, J. R. Gulcher, A. Kong, C. Rotimi, and K. Stefánsson, "Refining the impact of tcf7l2 gene variants on type 2 diabetes and adaptive evolution," *Nature Genetics*, vol. 39, pp. 218–225, 2007.

[75] L. J. Scott, K. L. Mohlke, L. L. Bonnycastle, C. J. Willer, Y. Li, W. L. Duren, M. R. Erdos, H. M. Stringham, P. S. Chines, A. U. Jackson, L. Prokunina-Olsson, C.-J. Ding, A. J. Swift, N. Narisu, T. Hu, R. Pruim, R. Xiao, X.-Y. Li, K. N. Conneely, N. L. Riebow, A. G. Sprau, M. Tong, P. P. White, K. N. Hetrick, M. W. Barnhart, C. W. Bark, J. L. Goldstein, L. Watkins, F. Xiang, J. Saramies, T. A. Buchanan, R. M. Watanabe, T. T. Valle, L. Kinnunen, G. R. Abecasis, E. W. Pugh, K. F. Doheny, R. N. Bergman, J. Tuomilehto, F. S. Collins, and M. Boehnke, "A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants," *Science*, vol. 316, pp. 1341–1345, 2007.

[76] A. Voorman, T. Lumley, B. McKnight, and K. Rice, "Behavior of qq-plots and genomic control in studies of gene-environment interaction," *PLoS ONE*, vol. 6, p. e19416, 05 2011.

[77] W. Bodmer and C. Bonilla, "Common and rare variants in multifactorial susceptibility to common diseases," *Nature Genetics*, vol. 40, pp. 695–701, June 2008.

[78] H. H. Zhang and W. Lu, "Adaptive-lasso for cox's proportional hazards model," *Biometrika*, vol. 94, pp. 1–13, 2007.

[79] National Heart, Lung, and Blood Institute (NHLBI), *Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults: The Evidence Report*. 1998.

[80] M. Chen, J. Cho, and H. Zhao, "Incorporating biological pathways via a markov random field model in genome-wide association studies.," *PLoS genetics*, vol. 7, p. e1001353, April 2011.

[81] P. F. Sullivan, "Puzzling over schizophrenia: schizophrenia as a pathway disease," *Nature Medicine*, vol. 18, pp. 210–211, 2012.

[82] P. Sullivan, M. Daly, and M. O'Donovan, "The genetic architechtures of psychiatric disorders: Apprehending the outline, glimpsing the details," *Submitted*, 2012.

[83] X. Lin and D. Zhang, "Inference in generalized additive mixed models by using smoothing splines," *Journal of the Royal Statistical Society - Series B: Statistical Methodology*, vol. 61, pp. 381–400, 1999.

[84] J. P. Imhof, "Computing the distribution of quadratic forms in normal variables," *Biometrika*, vol. 48, pp. 419–426, 1961.

[85] N. E. Breslow and D. G. Clayton, "Approximate inference in generalized linear mixed models," *Journal American Statistical Association*, vol. 88, pp. 9–25, 1993.

APPENDICES

Appendix A

Covariance of Y_i and Y_j conditional on X and S

By Law of total covariance, we have

$$\begin{aligned} cov(Y_i, Y_j|X, S) &= E_G\{cov_G(Y_i, Y_j|X, S, G)\} + cov_G\{E_G(Y_i|X, S, G), E_G(Y_j|X, S, G)\} \\ &= E_G\{cov_\eta[E(Y_i|X, S, G, \eta), E(Y_j|X, S, G, \eta)]\}+ \\ &\quad E_G\{E_\eta[cov(Y_i, Y_j|X, S, G, \eta)]\} + cov_G\{E_G(Y_i|X, S, G), E_G(Y_j|X, S, G)\}. \end{aligned} \tag{A.1}$$

With the assumption of the conditional independence, we derive $E_G\{E_\eta[cov_G(Y_i, Y_j|X, S, G, \eta)] = 0$. We take the Taylor expansion on the term $E_G\{cov_\eta[E(Y_i|X, S, G, \eta), E(Y_j|X, S, G, \eta)]\}$ in Equation (A.1) with respect to η around $E(\eta) = 0$, we have

$$\begin{aligned} & E_G\{cov_\eta[E(Y_i|X, S, G, \eta), E(Y_j|X, S, G, \eta)]\} \\ = \; & E_G\{cov_\eta(\mu_i, \mu_j)\} \\ \approx \; & E_G\{cov_\eta[(g^{-1}(X_i\gamma + G_i) + [g'(X_i\gamma + G_i)]^{-1}X_{Ei}\eta_i), \\ & (g^{-1}(X_j\gamma + G_j) + [g'(X_j\gamma + G_j)]^{-1}X_{Ej}\eta_j)]\} \\ = \; & E_G\{[g'(X_i\gamma + G_i)]^{-1}X_{Ei}\phi S_{ij}X_{Ej}[g'(X_j\gamma + G_j)]^{-1}\} \\ = \; & [g'(X_i\gamma)g'(X_j\gamma)]^{-1}\phi X_{Ei}X_{Ej}S_{ij}. \end{aligned}$$

Similarly, we apply the Taylor expansion on the term $cov_G\{E_G(Y_i|X,S,G,\eta), E_G(Y_j|X,S,G,\eta)\}$ in Equation (A.1) with respect to G around $E(G)=0$,

$$\begin{aligned}
& cov_G\{E_G(Y_i|X,S,G,\eta), E_G(Y_j|X,S,G,\eta)\} \\
= \; & cov_G\{g^{-1}(X_i\gamma + G_i), g^{-1}(X_j\gamma + G_j) \\
\approx \; & cov_G\{g^{-1}(X_i\gamma) + [g'(X_i\gamma)]^{-1}G_i, g^{-1}(X_j\gamma) + [g'(X_j\gamma)]^{-1}G_j]\} \\
= \; & [g'(X_i\gamma)]^{-1}\tau S_{ij}[g'(X_j\gamma)]^{-1} \\
= \; & [g'(X_i\gamma)g'(X_j\gamma)]^{-1}\tau S_{ij}.
\end{aligned}$$

Appendix B

EM Algorithm to Estimate Approximate Maximum Likelihood Estimations when Testing for the Gene-Environment Effect

In section 2.2.2, we show the conditional mean and variance of binary value Y_i to be equal to μ_i and $\mu_i(1-\mu_i)$, respectively. Under the assumption of $H_0 : \phi = 0$, we denote the conditional mean of Y_i to be $\mu_i(b) = E(Y_i|b, X, G) = \delta(X_i\gamma + Z_i b)$. Thus, we observe that $Y_i|b \sim \text{bin}(1, \mu_i(b))$, we then apply the technique of ML to estimate the variance component τ based on the marginal distribution $\delta(y, b; \theta) = \delta(y|b; \theta)\delta(b; \theta)$.

An EM algorithm is motivated and used to get the estimation of variance component τ and the fixed effects in the vector γ. Let $Y = (Y_1, \ldots, Y_n)$ be the vector of binary traits. Let the parameter vector to be expressed as $\theta = (\gamma, \tau)$, then the complete-data log likelihood is represented by $\log \delta(Y, b; \theta)$. By given the assumption that $b \sim N(0, \tau I_r)$, we computer $Q(\theta|\theta^{(t)})$

in the expectation step (E-step) which is

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= E\{\log\delta(Y,b;\theta)|Y,\theta^{(t)}\} \\ &= E\{\log\delta(Y|b;\theta)|Y,\theta^{(t)}\} + E\{\log\delta(b;\theta)|Y,\theta^{(t)}\}, \end{aligned} \tag{B.1}$$

where $\delta(Y,b;\theta) = \delta(Y|b;\theta)\delta(b;\theta)$. The reason of writing the complete-data log likelihood into two parts is that the part $\delta(b;\theta)$ is free of the fixed effects in the vector γ. It only contains the variance component τ in which the equation could be written as

$$\begin{aligned} \log\delta(b;\theta) &= \log\delta(b;\tau) \\ &= \log\{(2\pi)^{-r/2}|\tau I_r|^{-1/2}\exp\{-\tfrac{1}{2}b^T(\tau I_r)^{-1}b\}\} \\ &= -\tfrac{r}{2}\log 2\pi - \tfrac{r}{2}\log\tau - \tfrac{b^Tb}{2\tau}, \end{aligned} \tag{B.2}$$

where $|\tau I_r| = \tau^r$. The expectation of $\log\delta(b;\theta)$ is

$$\begin{aligned} E\{\log\delta(b;\theta)|Y,\theta^{(t)}\} &= E\{(-\tfrac{r}{2}\log 2\pi - \tfrac{r}{2}\log\tau - \tfrac{b^Tb}{2\tau})|Y,\theta^{(t)}\} \\ &= -\tfrac{r}{2}\log 2\pi - \tfrac{r}{2}\log\tau - \tfrac{E(b^Tb|Y,\theta^{(t)})}{2\tau}. \end{aligned} \tag{B.3}$$

Because this part is free of the fixed effects, by solving the Equation (B.3) with respect to the variance component τ, we obtain the estimate of τ. This step is the maximization step (M-step) in EM algorithm. We have the equation of

$$\frac{\partial E\log\delta(b;\theta)|Y,\theta^{(t)}}{\partial\tau} = -\frac{r}{2\tau} + \frac{E(b^Tb|Y,\theta^{(t)})}{2\tau^2},$$

where it is set to be 0, then we get

$$\begin{aligned} \hat{\tau}^{(t+1)} &= E(b^Tb|Y,\theta^{(t)})/r \\ &= \tfrac{1}{r}[b^{(t)T}b^{(t)} + \mathrm{trace}(\Sigma^{(t)})]. \end{aligned} \tag{B.4}$$

It follows the assumption that $(b|Y,\theta^{(t)}) \sim N(b^{(t)},\Sigma^{(t)})$ approximately. We now show how to get this approximate distribution for $(b|Y,\theta^{(t)})$.

If we rewrite the equation $\delta(Y, b; \theta)$ as $\delta(Y, b; \theta) = \delta(b|Y; \theta)\delta(Y; \theta)$, we could express $\delta(b|Y^{(t)}; \theta^{(t)}) = \delta(Y, b; \theta^{(t)})/\delta(Y^{(t)}; \theta^{(t)})$. By this expression, we rewrite the complete-data log-likelihood by using the distribution of $\delta(b|Y^{(t)}; \theta^{(t)})$ which is

$$\begin{aligned}
\delta(Y, b; \theta^{(t)}) &= \delta(b|Y^{(t)}; \theta^{(t)}) \times \delta(Y^{(t)}; \theta^{(t)}) \\
&= \textstyle\prod_{i=1}^{n} \{\mu_i^{Y_i}(1-\mu_i)^{1-Y_i}(2\pi)^{-r/2}\tau^{-\frac{r}{2}} \exp(-\frac{b^T b}{2\tau})\} \\
&= \exp\{\textstyle\sum_{i=1}^{n}[Y_i \log \mu_i + (1-Y_i)\log(1-\mu_i)] - \frac{r}{2}\log 2\pi - \frac{r}{2}\log\tau - \frac{b^T b}{2\tau}\} \\
&= \exp\{h(b)\},
\end{aligned} \tag{B.5}$$

where

$$\begin{aligned}
h(b) &= \textstyle\sum_{i=1}^{n}[Y_i \log \mu_i + (1-Y_i)\log(1-\mu_i)] - \frac{r}{2}\log 2\pi - \frac{r}{2}\log\tau - \frac{b^T b}{2\tau} \\
&= \textstyle\sum_{i=1}^{n}[Y_i \log \frac{\exp(X_i\gamma + Z_i b)}{1+\exp(X_i\gamma + Z_i b)} + (1-Y_i)\log \frac{1}{1+\exp(X_i\gamma + Z_i b)}] \\
&\quad -\frac{r}{2}\log 2\pi - \frac{r}{2}\log\tau - \frac{b^T b}{2\tau}.
\end{aligned} \tag{B.6}$$

We calculate the first derivative of the Equation (B.6) with respect to b which is

$$\begin{aligned}
\tfrac{\partial h(b)}{\partial b} &= \textstyle\sum_{i=1}^{n}\{Z_i^T Y_i - Z_i^T Y_i \hat{\mu_i}^b + Z_i^T Y_i \hat{\mu_i}^b - Z_i^T \hat{\mu_i}^b\} - \frac{b}{\tau} \\
&= \textstyle\sum_{i=1}^{n}\{Z_i^T Y_i - Z_i^T \hat{\mu_i}^b\} - \frac{b}{\tau} \\
&= Z^T Y - Z^T \hat{\mu}^b - \frac{b}{\tau} \\
&= Z^T (Y - \hat{\mu}^b) - \frac{b}{\tau},
\end{aligned}$$

where $\hat{\mu}^b = (\hat{\mu_1}^b, \hat{\mu_2}^b, \ldots, \hat{\mu_n}^b)^T$ and $\hat{\mu_i}^b = \frac{\exp(X_i\gamma + Z_i b)}{[1+\exp(X_i\gamma + Z_i b)]}$.

The second derivative of the Equation (B.6) with respect to b is also computed which is

$$\begin{aligned}
\tfrac{\partial h'(b)}{\partial b^T} &= -\textstyle\sum_{i=1}^{n} \frac{\exp(X_i\gamma + Z_i b)}{[1+\exp(X_i\gamma + Z_i b)]^2} Z_i^T Z_i - \frac{1}{\tau} I_r \\
&= -\textstyle\sum_{i=1}^{n}(1-\hat{\mu_i}^b)\hat{\mu_i}^b Z_i^T Z_i - \frac{1}{\tau} I_r \\
&= -Z^T W^{(b)} Z - \frac{1}{\tau} I_r \\
&= -(Z^T W^{(b)} Z + \frac{1}{\tau} I_r),
\end{aligned}$$

where $W^{(b)} = diag\{(1-\hat{\mu_i}^b)\hat{\mu_i}^b\}$.

By applying the Newton-Raphson method, we get the iteration of estimator of b as

$$\begin{aligned} b_t^{(k+1)} &= b_t^{(k)} - [h''(b_t^{(k)})]^{-1} h'(b_t^{(k)}) \\ &= b_t^{(k)} + [Z^T W^{(b)} Z + \tfrac{1}{\tau} I_r]^{-1} [Z^T (Y - \hat{\mu}^b) - \tfrac{b_t^{(k)}}{\tau}]. \end{aligned}$$

Then let $b^{(t)} = b_t^{(\infty)}$ where $Y = Y^{(t)}$ and $\tau = \tau^{(t)}$. By using the Taylor expansion on $h(b)$ with respect of b around $b^{(t)}$, we get

$$\begin{aligned} h(b) &= h(b^{(t)}) + h'(b^{(t)})(b - b^{(t)}) + \tfrac{1}{2}(b - b^{(t)})^T h''(b^{(t)})(b - b^{(t)}) \\ &= h(b^{(t)}) + \tfrac{1}{2}(b - b^{(t)})^T h''(b^{(t)})(b - b^{(t)}). \end{aligned}$$

Because $b^{(t)}$ maximizes $h(b)$ which means that $h'(b^{(t)}) = 0$, we could estimate the complete-data log-likelihood by applying Taylor expansion as

$$\begin{aligned} \delta(Y, b; \theta^{(t)}) &\approx \exp\{h(b^{(t)}) + \tfrac{1}{2}(b - b^{(t)})^T h''(b^{(t)})(b - b^{(t)})\} \\ &= \exp\{h(b^{(t)})\} \exp\{\tfrac{1}{2}(b - b^{(t)})^T h''(b^{(t)})(b - b^{(t)})\}. \end{aligned} \tag{B.7}$$

In the Equation (B.7), $\exp\{-\frac{1}{2}(b - b^{(t)})^T[-h''(b^{(t)})](b - b^{(t)})\}$ is a gaussian kernel where we notice that $-h''(b^{(t)}) = [\Sigma^{(t)}]^{-1}$. Thus, we show the conditional distribution of $(b|Y; \theta^{(t)})$ follows a multivariate normal distribution with mean vector $b^{(t)} = b_t^{(\infty)}$ and variance-covariance matrix $\Sigma^{(t)} = [-h''(b^{(t)})]^{-1}$.

Back to the expectation shown in Equation (B.1), we take the expectation of $\log \delta(Y|b; \theta)|Y, \theta^{(t)}$ to derive the estimate of the fixed effects in the vector γ.

Let $d(\gamma) = E\{\log \delta(Y|b; \theta)|Y; \theta^{(t)}\} = \sum_{i=1}^n E\{Y_i \log \mu_i + (1 - Y_i) log(1 - \mu_i)|Y, \theta^{(t)}\}$. To get

the estimator of γ, we take the first derivative of $d(\gamma)$ with respect to γ which is

$$\begin{aligned}
\frac{\partial d(\gamma)}{\partial \gamma} &= \frac{\partial \sum_{i=1}^{n} E\{Y_i \log \mu_i + (1-Y_i)\log(1-\mu_i)|Y,\theta^{(t)}\}}{\partial \gamma} \\
&= \sum_{i=1}^{n} E\{X_i^T \frac{Y_i-\mu_i}{\mu_i(1-\mu_i)}\mu_i(1-\mu_i)\} \\
&= \sum_{i=1}^{n} X_i^T(Y_i-\mu_i) \\
&= X^T(Y-\hat{\mu}^b).
\end{aligned}$$

In addition, we take the second derivative of $d(\gamma)$ with respect to γ as

$$\begin{aligned}
\frac{\partial d'(\gamma)}{\partial \gamma^T} &= -\sum_{i=1}^{n} \mu_i(1-\mu_i)X_i^T X_i \\
&= -X^T W^{(b)} X.
\end{aligned}$$

By using the first and second derivatives of $d(\gamma)$, the iteration for the estimation of γ can be derived by the M-step in EM algorithm which is

$$\begin{aligned}
\gamma_t^{(k+1)} &= \gamma_t^{(k)} - [d''(\gamma_t^{(k)})]^{-1} d'(\gamma_t^{(k)}) \\
&= \gamma_t^{(k)} + [X^T W^{(b)} X]^{-1} X^T (Y-\hat{\mu}^b).
\end{aligned}$$

Then, let $\gamma^{(t)} = \gamma_t^{(\infty)}$.

To sum up, we have the iterations for each estimate as follows.

- $\hat{\tau}^{(t+1)} = -\frac{1}{r}[b^{(t)T} b^{(t)} + \text{trace}(\Sigma^{(t)})]$.
- $b^{(k+1)} = b^{(k)} + [Z^T W^{(b)} Z + \frac{1}{\tau} I_r]^{-1}[Z^T(Y-\hat{\mu}^b) - \frac{b^{(k)}}{\tau}]$ and $b^{(t)} = b_t^{(\infty)}$.
- $\gamma^{(k+1)} = \gamma^{(k)} + [X^T W^{(b)} X]^{-1} X^T (Y-\hat{\mu}^b)$ and $\gamma^{(t)} = \gamma_t^{(\infty)}$.

Appendix C

Derivation of the Score Test Statistics and Their Corresponding Asymptotic Distributions

Zhang and Lin [72] proposed the REML version to conduct the polynomial test for the non-Gaussian responses which includes the case of binary response. By applying the smoothing technique introduced by Lin and Zhang [83], we could derive the score of ϕ tested when $\phi = 0$. The score test statistic is presented as

$$U_\phi = \frac{1}{2}\{(Y^W - X\hat{\gamma})^T V_0^{-1}\Sigma V_0^{-1}(y^W - X\hat{\gamma}) - tr(P_0\Sigma)\}|\phi = 0,\ \gamma = \hat{\gamma},\ \tau = \hat{\tau},\ \sigma = \hat{\sigma}, \quad \text{(C.1)}$$

where $\hat{\gamma}$ and $\hat{\tau}$ are approximate ML estimates of γ and τ, respectively. Y^W is the working vector which equals to $y^W = X\hat{\gamma} + Z\hat{b} + \Delta(Y - \hat{\mu}^b)$, where $\Delta = diag\{g'(\mu_i)\}$ in Equation (C.1) with $g'(\mu_i) = \frac{1}{\mu_i(1-\mu_i)}$, $V_0 = W^{-1} + \tau S$ and $P_0 = V_0^{-1} - V_0^{-1}X(X^T V_0^{-1} X)^{-1} X^T V_0^{-1}$.

To estimate the asymptotic distribution, we write U_ϕ into two parts such that $U_\phi = T_\phi + \tilde{e}$

where T_ϕ and $\tilde{e}$ are expressed as

$$\begin{aligned} T_\phi &= \tfrac{1}{2}(Y^W - X\hat{\gamma})^T V_0^{-1}\Sigma V_0^{-1}(Y^W - X\hat{\gamma}), \\ \tilde{e} &= tr(P_0\Sigma)/2. \end{aligned}$$

T_ϕ follows a scale chi-square distribution approximately. The mean of T_ϕ could be evaluated by $\tilde{e}$ and the variance is approximated by

$$I_{\tau\tau} = \frac{1}{2}tr(P_0\Sigma P_0\Sigma) - (\frac{1}{2}tr(P_0\Sigma P_0 S))^2/(\frac{1}{2}tr(P_0 S P_0 S)).$$

By applying the Laplace approximation, the working vector Y^W follows a multivariate normal distribution. A Gaussian fourth-moment assumption could be employed to estimate the variance of T_ϕ [72].

Due to the reason that the test statistic U_ϕ under the null hypothesis of $\phi = 0$ does not follow a normal distribution asymptotically, we use the first term T_ϕ in the score statistics as the testing statistics. Given that Y^W follows a normal distribution of $N(X\gamma, V_0)$, we denote $Y^s = (Y^W - X\gamma)^T V_0^{-1/2}$ which standardizes the work vector Y^W into a standard normal distribution. In the equation of obtaining the T_ϕ, we need to derive the estimation of γ in order to calculate this quantity. The estimate could be computed as

$$\hat{\gamma} = (X^T V_0^{-1} X)^{-1} X^T V_0^{-1} Y^W. \tag{C.2}$$

Consequently, we have

$$\begin{aligned} Y^W - X\hat{\gamma} &= [I_n - X(X^T V_0^{-1} X)^{-1} X^T V_0^{-1}](Y^W - X\gamma) \\ &= K_0(Y^W - X\gamma), \end{aligned}$$

where $K_0 = [I_n - X(X^T V_0^{-1} X)^{-1} X^T V_0^{-1}]$.

Therefore, the expression of T_ϕ could be rewritten as

$$\begin{aligned} T_\phi &= \tfrac{1}{2}\{(Y^W - X\gamma)^T K_0^T V_0^{-1} \Sigma V_0^{-1} K_0 (Y^W - X\gamma)\} \\ &= \tfrac{1}{2}\{(Y^W - X\gamma)^T V_0^{-1/2} V_0^{1/2} K_0^T V_0^{-1} \Sigma V_0^{-1} K_0 V_0^{1/2} V_0^{-1/2} (Y^W - X\gamma)\} \\ &= \tfrac{1}{2}\{(Y^s)^T A (Y^s)\}, \end{aligned} \tag{C.3}$$

where $A = V_0^{1/2} K_0^T V_0^{-1} \Sigma V_0^{-1} K_0 V_0^{1/2}$.

Define e_i^* and λ_i^* to be the eigenvector and eigenvalue of matrix A respectively for $i = 1, \ldots, n$. For the concern of the time consuming, define a $L \times L$ matrix B such that the non-zero eigenvalues of A and eigenvalues of B are the same. The matrix B is shown as

$$B = Z^T D V_0^{-1} [V_0 - X(X^T V_0^{-1} X)^{-1} X^T] V_0^{-1} D Z.$$

We define the e_j and λ_j to be the eigenvector and eigenvalue of matrix B for $i = 1, \ldots, L$. Thus $T_\phi = \Sigma_{i=1}^n \lambda_i^* (e_i^{*T} Y^s)^2 = \Sigma_{j=1}^L \lambda_j \tilde{Y}^{s2}$ with $\tilde{Y}^{s2}$ follows a 1 degree of freedom chi-square distribution. Followed by the distribution shown in the Tzeng and Zhang [24], this testing statistic T_ϕ could be approximate by the distribution of $\Sigma_{j=1}^L \hat{\lambda_j} \chi_{j1}^2$. By applying the three-moment approximation [84], the level-α significance threshold could estimated by $\kappa_1 + (\chi_\alpha - h')\sqrt{\kappa_2/h'}$, where $\kappa_k = \Sigma_k \lambda_k^k$, $h' = \kappa_2^3/\kappa_3^2$ and χ_α is the αth quantile of $\chi_{h'}^2$. Alternatively, one can report the p-value of the observed statistic T_ϕ by $P(\chi_{h'}^2 > \chi^*)$ where $\chi^* = (T_\phi - \kappa_1)\sqrt{n'/\kappa_2} + h'$.

Appendix D

EM Algorithm to Penalized Maximum Likelihood Estimations on Quantitative Traits

With each pre-specific λ value, we need to estimate all the parameters by maximizing the penalized log-likelihood function shown in Equation (3.4). Below we show in details of deriving the estimators for both the coefficients and variance components.

Let Y be the quantitative trait vector for all the individuals. Let $b = (b_1^T, \ldots, b_K^T)^T$ to be the random effect coefficients which are missing in the data. Therefore, maximizing the Equation (3.4) is equivalent to maximize the following penalized Q-function

$$Q_p(\theta|\theta^{(t)}) = E\{\log f(Y, b; \theta)|Y, \theta^{(t)}\} - n\lambda \sum_{k=1}^{K} \frac{\phi_k}{\tilde{\phi}_k}, \tag{D.1}$$

where $\theta = (\gamma, \phi, \sigma)$. Let $f(Y, b; \theta) = f(Y|b; \theta)f(b; \theta)$ such that the distribution of $f(b; \theta)$ is free of the fixed effect coefficients γ. Thus, the probability density function $f(b; \theta)$ only contains the

variance components ϕ_k's for $k = 1, \ldots, K$ which is

$$f(b;\theta) = (2\pi)^{-r/2}|\Sigma_b|^{-1/2}\exp\{-\frac{1}{2}b^T\Sigma_b^{-1}b\}.$$

where $r = \sum_{k=1}^{K} r_k$ with $r_k = \text{rank}(S_k)$ and $\Sigma_b = \begin{pmatrix} \phi_1 I_{r_1} & & \\ & \ddots & \\ & & \phi_K I_{r_K} \end{pmatrix}$.

By splitting $f(Y, b; \theta)$ into two parts, the penalized Q-function shown in the Equation (D.1) could also be expressed as

$$\begin{aligned} Q_p(\theta|\theta^{(t)}) &= E\{\log f(Y|b;\theta)|Y,\theta^{(t)}\} + E\{\log f(b;\theta)|Y,\theta^{(t)}\} - n\lambda\textstyle\sum_{k=1}^{K}\frac{\phi_k}{\tilde{\phi}_k} \\ &= -\tfrac{n}{2}\log(2\pi) - \tfrac{n}{2}\log|\sigma| - \tfrac{r}{2}\log(2\pi) - \tfrac{1}{2}\log|\Sigma_b| \\ &\quad + E\left\{\left[-\tfrac{1}{2\sigma}\textstyle\sum_{i=1}^{n}(Y_i - X_i\gamma - Z_i b)^2 - \tfrac{1}{2}b^T\Sigma_b^{-1}b\right]|Y,\theta^{(t)}\right\} - n\lambda\textstyle\sum_{k=1}^{K}\frac{\phi_k}{\tilde{\phi}_k} \\ &= -\tfrac{n}{2}\log(2\pi) - \tfrac{n}{2}\log|\sigma| - \tfrac{r}{2}\log(2\pi) - \tfrac{1}{2}\Sigma_{k=1}^{K} r_k\log\phi_k \\ &\quad -\tfrac{1}{2\sigma}E\{(Y - X\gamma - Zb)^T(Y - X\gamma - Zb)|Y,\theta^{(t)}\} - \tfrac{1}{2}\Sigma_{k=1}^{K}\tfrac{1}{\phi_k}E\{b_k^T b_k|Y,\theta^{(t)}\} \\ &\quad -n\lambda\textstyle\sum_{k=1}^{K}\frac{\phi_k}{\tilde{\phi}_k}. \end{aligned} \tag{D.2}$$

The steps shown above are the E-steps in the EM algorithm, the M-steps in the EM algorithm which derive the iterated estimators for the parameters and variance components are given as follows

$$\frac{\partial Q_p}{\partial \phi_k} = -\frac{1}{2}\frac{r_k}{\phi_k} + \frac{1}{2\phi_k^2}E(b_k^T b_k|Y,\theta^{(t)}) - n\lambda\frac{1}{\tilde{\phi}_k}.$$

$$\frac{\partial Q_p}{\partial \gamma} = -2E\{X^T(Y - X\gamma - Zb)|Y,\theta^{(t)}\}$$

$$\frac{\partial Q_p}{\partial \sigma} = -\frac{n}{2\sigma} + \frac{1}{2(\sigma)^2}E\{(Y - X\gamma - Zb)^T(Y - X\gamma - Zb)|Y,\theta^{(t)}\}.$$

Under the quantitative cases, the mode of $f(b|Y;\theta)$ equals to $E(b|Y,\theta^{(t)})$, thus we use the

mode to replace $E(b|Y,\theta^{(t)})$ in the penalized Q-function and to calculate the estimators. To find out the mode, first we need to obtain the conditional distribution $(b|Y,\theta^{(t)})$. Because in the joint function $f(Y,b;\theta) = f(b|Y;\theta)f(Y;\theta)$, the function of $f(Y;\theta)$ is a constant, we could derive that $f(b|Y;\theta) = \frac{f(Y,b;\theta)}{f(Y;\theta)} \propto f(Y,b;\theta)$. Thus, the conditional distribution of $(b|Y,\theta^{(t)})$ could be obtained based on the function of $f(Y,b;\theta)$.

Define $h(b) = -\frac{1}{2\sigma}(Y - X\gamma - Zb)^T(Y - X\gamma - Zb) - \frac{1}{2}b^T\Sigma_b^{-1}b$, and $\hat{b}^{(t)}$ be the mode of $f(Y,b;\theta^{(t)})$. Using the Taylor expansion, we could show the following

$$\begin{aligned} h(b) &= h(\hat{b}^{(t)}) + h'(\hat{b}^{(t)})(b - \hat{b}^{(t)}) + \tfrac{1}{2}(b - \hat{b}^{(t)})^T h''(\hat{b}^{(t)})(b - \hat{b}^{(t)}) \\ &= h(\hat{b}^{(t)}) + \tfrac{1}{2}(b - \hat{b}^{(t)})^T h''(\hat{b}^{(t)})(b - \hat{b}^{(t)}), \end{aligned}$$

where the first and second derivative of $h(b)$ with respect to b are

$$\frac{\partial h(b)}{\partial b} = \frac{1}{\sigma}Z^T(Y - X\gamma - Zb) - \Sigma_b^{-1}b = h'(b),$$

$$\frac{\partial h'(b)}{\partial b^T} = -\frac{1}{\sigma}Z^TZ - \Sigma_b^{-1} = h''(b).$$

Because $h'(\hat{b}^{(t)}) = 0$, $\hat{b}^{(t)}$ which maximizes the function $h(b)$ is the mode of $f(Y,b;\theta^{(t)})$. Then we could express the function $f(b|Y;\theta^{(t)})$ as

$$
\begin{aligned}
f(b|Y;\theta^{(t)}) &= f(Y,b;\theta^{(t)})/f(Y;\theta^{(t)}) \\
&= (1/f(Y;\theta^{(t)}))f(Y,b;\theta^{(t)}) \\
&= (1/f(Y;\theta^{(t)}))(2\pi)^{-\frac{n}{2}}|\sigma I_n|^{-\frac{1}{2}}(2\pi)^{-\frac{r}{2}}|\Sigma_b|^{-\frac{1}{2}} \\
&\quad \exp\left\{-\tfrac{1}{2\sigma}(Y-X\gamma-Zb)^T(Y-X\gamma-Zb)-\tfrac{1}{2}b^T\Sigma_b^{-1}b\right\} \\
&= (1/f(Y;\theta^{(t)}))(2\pi)^{-\frac{n}{2}}|\sigma I_n|^{-\frac{1}{2}}(2\pi)^{-\frac{r}{2}}|\Sigma_b|^{-\frac{1}{2}} \\
&\quad \exp\{h(\hat{b}^{(t)})+\tfrac{1}{2}(b-\hat{b}^{(t)})^T h''(\hat{b}^{(t)})(b-\hat{b}^{(t)})\} \\
&= (1/f(Y;\theta^{(t)}))(2\pi)^{-\frac{n}{2}}|\sigma I_n|^{-\frac{1}{2}}(2\pi)^{-\frac{r}{2}}|\Sigma_b|^{-\frac{1}{2}}\exp\{h(\hat{b}^{(t)})\} \\
&\quad \exp\{-\tfrac{1}{2}(b-\hat{b}^{(t)})^T[-h''(\hat{b}^{(t)})](b-\hat{b}^{(t)})\} \\
&= (1/f(Y;\theta^{(t)}))(2\pi)^{-\frac{n}{2}}|\sigma I_n|^{-\frac{1}{2}}(2\pi)^{-\frac{r}{2}}|\Sigma_b|^{-\frac{1}{2}}\exp\{h(\hat{b}^{(t)})\} \\
&\quad \exp\{-\tfrac{1}{2}(b-\hat{b}^{(t)})^T[\tfrac{1}{\sigma}Z^TZ+\Sigma_b^{-1}](b-\hat{b}^{(t)})\}.
\end{aligned}
\tag{D.3}
$$

In the Equation (D.3), $\exp\{-\frac{1}{2}(b-\hat{b}^{(t)})^T[\frac{1}{\sigma}Z^TZ+\Sigma_b^{-1}](b-\hat{b}^{(t)})\}$ is a gaussian kernel. Thus, we show the conditional distribution of $(b|Y;\theta^{(t)}) \sim N(\hat{b}^{(t)}, [\frac{1}{\sigma}Z^TZ+\Sigma_b^{-1}]^{-1})$ and $\hat{b}^{(t)}$ can be obtained by the following equation based on the fact that $h'(\hat{b}^{(t)})=0$

$$
\begin{aligned}
\hat{b}^{(t)} &= (Z^TZ+\sigma\Sigma_b^{-1})^{-1}Z^T(Y-X\gamma^{(t)})|_{\theta=\theta^t} \\
&= [\sigma(\tfrac{1}{\sigma}Z^TZ+\Sigma_b^{-1})]^{-1}Z^T(Y-X\gamma^{(t)})|_{\theta=\theta^t} \\
&= \tfrac{1}{\sigma}V_b^{(t)}Z^T(Y-X\gamma^{(t)})|_{\theta=\theta^t}.
\end{aligned}
$$

Based on the mode of the $(b|Y;\theta^{(t)})$, we could derive the iterated estimators $\hat{\phi}_k^{(t+1)}$'s as

$$
\hat{\phi}_k^{(t+1)} = \frac{2[(\hat{b}_k^{(t)})^T\hat{b}_k^{(t)}+tr(\Sigma_{b_k}^{(t)})]}{r_k+\sqrt{r_k^2+8\frac{n\lambda}{\tilde{\phi}_k}[(\hat{b}_k^{(t)})^T\hat{b}_k^{(t)}+tr(\Sigma_{b_k}^{(t)})]}}.
$$

Because one option to calculate the estimators of $\tilde{\phi}_k$'s could be the maximum likelihood estimates by maximizing the Equation (D.1) when $\lambda=0$. These $\tilde{\phi}_k$ values for $k=1,\ldots,K$ could be computed as

$$
\tilde{\phi}_k^{(t+1)} = \frac{(\hat{b}_k^{(t)})^T\hat{b}_k^{(t)}+tr(\Sigma_{b_k}^{(t)})}{r_k}.
$$

Similarly, the estimator for γ at $(t+1)$th iteration could be calculated as

$$\hat{\gamma}^{(t+1)} = (X^T X)^{-1} X^T (Y - Z\hat{b}^{(t)}).$$

In a similar manner, the iterated estimator for σ at $(t+1)$th iteration could be expressed as

$$\hat{\sigma}^{(t+1)} = \frac{(Y - X\gamma^{(t)} - Z\hat{b}^{(t)})^T (Y - X\gamma^{(t)} - Z\hat{b}^{(t)}) + tr(Z^T Z V^{(t)_b})}{n}.$$

Appendix E

EM algorithm to Maximum Likelihood Estimations on Quantitative Traits

Similar to the EM algorithm applied to obtain the penalized maximized likelihood estimations, a Q-function without the penalty term is used to derive the maximize likelihood estimation which is expressed as

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= E\{\log f(Y,b;\theta)|Y,\theta^{(t)}\} \\ &= E\{\log f(Y|b;\theta)|Y,\theta^{(t)}\} + E\{\log f(b;\theta)|Y,\theta^{(t)}\} \\ &= -\tfrac{n}{2}\log(2\pi) - \tfrac{n}{2}\log|\sigma^*| - \tfrac{r^*}{2}\log(2\pi) - \tfrac{1}{2}\Sigma_{k=1}^{K^*} r_k^* \log\phi_k^* \\ &\quad -\tfrac{1}{2\sigma}E\{(Y-X\gamma-Z^*b)^T(Y-X\gamma-Z^*b)|Y,\theta^{(t)}\} - \tfrac{1}{2}\Sigma_{k=1}^{K^*}\tfrac{1}{\phi_k^*}E\{b_k^T b_k|Y,\theta^{(t)}\}. \end{aligned} \tag{E.1}$$

Based on the Q-function shown in the Equation (E.1), we could get the iterated estimators for the parameters and variance components as follows. First, we derive the estimation for ϕ_k^*'s

where $k = 1, \ldots, K^*$ using the following formula

$$\frac{\partial Q}{\partial \phi_k^*} = -\frac{1}{2}\frac{r_k^*}{\phi_k^*} + \frac{1}{2\phi_k^{*2}} E(b_k^T b_k | Y, \theta^{(t)}).$$

By solving the equation for ϕ_k^*'s, we calculate the estimator for ϕ_k^*'s at $(t+1)$th iteration using

$$\hat{\phi}_k^{*(t+1)} = \frac{(b_k^{(t)})^T b_k^{(t)} + tr(\Sigma_{b_k}^{(t)})}{r_k^*}.$$

To estimate the coefficient vector γ, the following partial derivative with respect to γ is used which is

$$\frac{\partial Q}{\partial \gamma} = -2E\{X^T(Y - X\gamma - Z^* b) | Y, \theta^{(t)}\}$$

By solving the above equation for γ, we derive its iterated estimator at iteration $(t+1)$ as

$$\hat{\gamma}^{(t+1)} = (X^T X)^{-1} X^T (Y - Z^* \hat{b}^{(t)}),$$

where $\hat{b}^{(t)} = E\{b | Y, \theta^{(t)}\}$ using the reduce model shown in Equation (3.5).

In addition, we estimate the variance σ by using the following formula

$$\frac{\partial Q}{\partial \sigma^*} = -\frac{n}{2\sigma^*} + \frac{1}{2(\sigma^*)^2} E\{(Y - X\gamma - Z^* b)^T (Y - X\gamma - Z^* b) | Y, \theta^{(t)}\}.$$

By solving the equation with respect to σ, the estimator at iteration $(t+1)$ is calculated as

$$\hat{\sigma}^{*(t+1)} = \frac{(Y - X\gamma^{(t)} - Z^* \hat{b}^{(t)})^T (Y - X\gamma^{(t)} - Z^* \hat{b}^{(t)}) + tr(Z^{*T} Z^* V_b^{(t)})}{n}.$$

Appendix F

Derivation of Likelihood and Log-Likelihood Functions

To calculate the log-likelihood function, we first show that the integrated likelihood function as [85]

$$\begin{aligned}
L(\theta; y) &= \int f(Y|b;\theta) f(b) db \\
&= (2\pi)^{-\frac{n}{2}} |\sigma^* I_n|^{-\frac{1}{2}} (2\pi)^{-\frac{r^*}{2}} |\hat{\Sigma}_b|^{-\frac{1}{2}} \\
&\quad \int \exp\left\{ -\tfrac{1}{2\sigma^*} (Y - X\gamma - Z^* b)^T (Y - X\gamma - Z^* b) - \tfrac{1}{2} b^T \hat{\Sigma}_b^{-1} b \right\} db \\
&= (2\pi)^{-\frac{n}{2}} |\sigma^* I_n|^{-\frac{1}{2}} (2\pi)^{-\frac{r^*}{2}} |\hat{\Sigma}_b|^{-\frac{1}{2}} \int \exp\{h(b)\} db \\
&= (2\pi)^{-\frac{n}{2}} |\sigma^* I_n|^{-\frac{1}{2}} (2\pi)^{-\frac{r^*}{2}} |\hat{\Sigma}_b|^{-\frac{1}{2}} \\
&\quad \int \exp\{h(\hat{b})\} \exp\{\tfrac{1}{2}(b-\hat{b})^T h''(\hat{b})(b-\hat{b})\} db \\
&= (2\pi)^{-\frac{n}{2}} |\sigma^* I_n|^{-\frac{1}{2}} |\hat{\Sigma}_b|^{-\frac{1}{2}} \exp\{h(\hat{b})\} |V|^{\frac{1}{2}} \\
&\quad (2\pi)^{-\frac{r^*}{2}} |V|^{-\frac{1}{2}} \int \exp\{\tfrac{1}{2}(b-\hat{b})^T h''(\hat{b})(b-\hat{b})\} db \\
&= (2\pi)^{-\frac{n}{2}} |\sigma^* I_n|^{-\frac{1}{2}} |\hat{\Sigma}_b|^{-\frac{1}{2}} \exp\{h(\hat{b})\} |V|^{\frac{1}{2}},
\end{aligned}$$

where $V = [-h''(\hat{b})]^{-1}$ and $|V|^{-\frac{1}{2}} \int \exp\{\frac{1}{2}(b-\hat{b})^T h''(\hat{b})(b-\hat{b})\} db = 1$.

Because we have shown in the Appendix D that the conditional distribution of $(b|y;\theta^{(t)})$ follows a multivariate normal distribution with mean vector $\hat{b}^{(t)}$ and variance-covariance matrix $V = [-h''(\hat{b}^{(t)})]^{-1}$ based on the Model (3.5). Then, it is straightforward to obtain the log-

likelihood as follows

$$
\begin{aligned}
l(\theta; y) &= -\tfrac{n}{2}\log(2\pi) - \tfrac{n}{2}\log(\sigma^*) - \tfrac{1}{2}\log|\hat{\Sigma}_b| + h(\hat{b}) - \tfrac{1}{2}\log|\Sigma^{(t)-1}| \\
&= -\tfrac{n}{2}\log(2\pi) - \tfrac{n}{2}\log(\sigma^*) - \tfrac{1}{2}\log|\hat{\Sigma}_b| \\
&\quad -\tfrac{1}{2\hat{\sigma^*}}(Y - X\hat{\gamma} - Z\hat{b})^T(Y - X\gamma - Z\hat{b}) - \tfrac{1}{2}\hat{b}^T\hat{\Sigma}_b^{-1}\hat{b} - \tfrac{1}{2}\log|-h''(\hat{b})| \\
&= -\tfrac{n}{2}\log(2\pi) - \tfrac{n}{2}\log(\sigma^*) - \tfrac{1}{2\hat{\sigma^*}}(Y - X\hat{\gamma} - Z\hat{b})^T(Y - X\gamma - Z\hat{b}) \\
&\quad -\tfrac{1}{2}\hat{b}^T\hat{\Sigma}_b^{-1}\hat{b} - \tfrac{1}{2}\log|\hat{\Sigma}_b[-h''(\hat{b})]| \\
&= -\tfrac{n}{2}\log(2\pi) - \tfrac{n}{2}\log(\sigma^*) - \tfrac{1}{2\hat{\sigma^*}}(Y - X\hat{\gamma} - Z\hat{b})^T(Y - X\gamma - Z\hat{b}) \\
&\quad -\tfrac{1}{2}\hat{b}^T\hat{\Sigma}_b^{-1}\hat{b} - \tfrac{1}{2}\log|\hat{\Sigma}_b[\tfrac{1}{\sigma^*}Z^TZ + \Sigma_b^{-1}]| \\
&= -\tfrac{n}{2}\log(2\pi) - \tfrac{n}{2}\log(\sigma^*) - \tfrac{1}{2\hat{\sigma^*}}(Y - X\hat{\gamma} - Z\hat{b})^T(Y - X\gamma - Z\hat{b}) \\
&\quad -\tfrac{1}{2}\Sigma_{k=1}^{K^*}\tfrac{1}{\phi_k}\hat{b}_k^T\hat{b}_k - \tfrac{1}{2}\log|\tfrac{1}{\sigma}\hat{\Sigma^*}_b Z^TZ + I_{r^*}|.
\end{aligned}
$$