Statistical Methods in Mapping Complex Diseases

Jing He

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2011

Supervisor of Dissertation

*Signature*_____

Mingyao Li

Assistant Professor of Biostatistics

Co-supervisor

*Signature*_____

Hongzhe Li

Professor of Biostatistics

Graduate Group Chairperson

*Signature*_____

Daniel F. Heitjan, Professor of Biostatistics

Dissertation Committee

Mary E. Putt, Associate Professor of Biostatistics

Thomas Cappola, Assistant Professor of School of Medicine

Warren Ewens, Professor of Biology

Acknowledgments

This dissertation would never come into existence without the guidance of my advisors, Professor Mingyao Li and Professor Hongzhe Li. I cannot express my gratitude enough for their support, advice and understanding in the past few years. As Prof. Mingyao Li's first student, she spent plenty of time and energy on supervising me. She has always been trying to train me to be independent, making me exposed to the frontier of the field, encouraging me to attend conferences, needless to say all the challenges and difficulties she guided me through. I also feel lucky to have Prof. Hongzhe Li to mentor me for the last project. His discussion on research and future goals always inspired me. He is also full of brilliant ideas, when I feel frustrated, he always suggested an alternative and made me confident in pursuing our project. At the same time, I would like thank all the members of my committee, Prof. Mary Putt, Prof. Thomas Cappola and Prof. Warren Ewens, for their insightful comments and help to improve the quality of the work.

Special thanks go to Prof. Thomas Cappola and Prof. Muredach Reilly. They not only provide me part of financial support, but also influence me in my way of doing research and collaboration. I had been involved in the analysis of the data generated

from their studies. Their passion for research and keen insights into the data have always motivated me to be a good researcher.

At the same time I would like to express my thanks to my friends: Seunghee Baek, Yimei Li, Jichun Xie and Rongmei Zhang, who arrived at the department in the same year as me. Life will be different without their company. Special thanks go to Baeky, who always understands me in every aspect and Yimei, who is always there ready to answer my questions.

Finally I dedicate this dissertation to my parents who taught me everything that matters in my life. Their unconditional love and support have always kept me going when everything else fails.

ABSTRACT

Statistical Methods in Mapping Complex Diseases

Jing He, Hongzhe Li, Mingyao Li

Genome-wide association studies have become a standard tool for disease gene discovery over the past few years. These studies have successfully identified genetic variants attributed to complex diseases, such as cardiovascular disease, diabetes and cancer. Various statistical methods have been developed with the goal of improving power to find disease causing variants. The major focus of this dissertation is to develop statistical methods related to gene mapping studies with its application in real datasets to identify genetic markers associated with complex human diseases.

In my first project, I developed a method to detect gene-gene interactions by incorporating linkage disequilibrium (LD) information provided by external datasets such as the International HapMap or the 1000 Genomes Projects. The next two projects in my dissertation are related to the analysis of secondary phenotypes in case-control genetic association studies. In these studies, a set of correlated secondary phenotypes that may share common genetic factors with disease status are often collected. However, due to unequal sampling probabilities between cases and controls, the standard regression approach for examination of these secondary phenotype can yield inflated type I error rates when the test SNPs are associated with the disease. To solve this issue, I propose a Gaussian copula approach to jointly model the disease status and the secondary phenotype. In my second project, I consider only one marker in the model and perform a test to access whether the marker is associated with the

secondary phenotype in the Gaussian copula framework. In my third project, I extend the copula-based approach to include a large number of candidate SNPs in the model. I propose a variable selection approach to select markers which are associated with the secondary phenotype by applying a lasso penalty to the log-likelihood function.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

For most common diseases, including heart disease, diabetes, hypertension, and cancer, multiple genetic and environmental factors jointly influence an individual's risk of being affected. Rapid advances in SNP genotyping technology and the deposition of millions of SNPs into public databases have set the stage for genome-wide association studies. Most of these studies have used a single marker based analysis strategy in which each SNP is tested individually for association with a specific phenotype. There is a growing evidence suggesting that complex diseases are the results of marginal gene effect and the gene-gene interactions (Moore and Williams (2009)). Detecting such interactions will allow us to elucidate the biological and biochemical pathways underpinning complex diseases (Moore (2003)). However, detection of gene-gene interactions has long been a challenge due to their complexity. The standard method aiming at detecting SNP-SNP interactions may be inadequate as it does not model linkage disequilibrium (LD) among SNPs in each gene and may lose power due to a large number of comparisons. To improve power, in Chapter 2,

we propose a principal component (PC)-based framework for gene-based interaction analysis. We analytically derive the optimal weight for both quantitative and binary traits based on genotypes and pairwise LD information. We then use PCs to summarize the information in each gene and test for interactions between the PCs. We further extend this gene-based interaction analysis procedure to allow the use of imputation dosage scores obtained from a popular imputation software package, MACH, which incorporates multi-locus LD information. To evaluate the performance of the gene-based interaction tests, we conduct extensive simulations under various settings. We demonstrate that gene-based interaction tests are more powerful than SNP-based tests when more than two variants interact with each other; moreover, tests that incorporate external LD information are generally more powerful than those that use genotyped markers only. We also apply the proposed gene-based interaction tests to a candidate gene association study on high-density lipoprotein cholesterol (HDL-C). As our method operates at the gene level, it can be applied to a genome-wide association setting and used as a screening tool to detect gene-gene interactions.

Another problem in genome-wide association studies is the analysis of secondary phenotypes. Many genome-wide association studies measure a variety of quantitative or qualitative traits other than the disease trait that defines the case control status. Exploring these secondary phenotypes can maximize return considering the massive time and money invested. For example, recent genome-wide association studies for height (Lettre et al. (2008); Sanna et al. (2008); Weedon et al. (2007)) and body mass index (Loos et al. (2008)) were conducted using samples from multiple genome-wide association studies which were originally conducted to find markers associated with

risk of diabetes, breast and prostate cancer, and other traits. In addition, examination of these secondary phenotypes that may share common genetic factors with disease status can yield valuable insights about the disease etiology and supplement the main studies. However, due to unequal sampling probabilities in cases and controls, standard regression analysis that assesses the effect of SNPs on secondary phenotypes using cases only, controls only, or combined samples of cases and controls can be very misleading when the test SNP is associated with the disease status. To solve this issue, in Chapter 3, we propose a Gaussian copula-based approach that efficiently models the dependence between disease status and secondary phenotype. Through simulations, we show that our method yields correct type I error rates for the analysis of secondary phenotype under a wide range of settings. We further show that the type I error rates of our test are under control even when the model is mis-specified. To illustrate the effectiveness of our method in the analysis of real data, we applied our method to a genome-wide association study on HDL-C, where cases are defined as individuals with extremely high HDL-C level and controls are defined as those with low HDL-C level. We treated four quantitative traits with varying degrees of correlation with HDL-C as secondary phenotypes and tested for association with SNPs in *LIPG*, a gene that is well-known to be associated with HDL-C. We show that when the correlation between the primary and secondary phenotypes is >0.2, the P-values from case-control combined unadjusted analysis are much more significant than methods that aim to correct for ascertainment bias. Our results suggest that to avoid false positive associations, it is important to appropriately model secondary phenotypes in case-control genetic association studies.

One disadvantage of single marker based analysis is that the obtained P-value may not be significant when adjusted for multiple testing. An alternative approach is to impose a lasso penalty to the log-likelihood function to select the relevant variants. The lasso penalty is an effective device for model selection, especially in problems where the number of predictors far exceeds the number of observations. In Chapter 4, we extende the copula-based approach to include a large number of candidate SNPs in the analysis of secondary phenotypes. A Lasso penalty is applied to control the sparsity of the solution. We proposed an efficient computational algorithm using the coordinate gradient descent method to solve the likelihood. For a given value of the tuning parameter, the penalized likelihood is maximized by the coordinate gradient descent method. This method is compared with the "Penalized" package which uses the regular lasso assuming the secondary trait is normally distributed. We test this method on both simulated data as well as the LPL gene on HDL-C study. We demonstrate that the copula-based approach is efficient in controlling the false discovery rate.

Chapter 2

Gene-based Interaction Analysis by Incorporating External Linkage Disequilibrium Information

2.1 Introduction

With continued decreasing cost of high throughput genotyping technology, genome-wide association studies (GWAS) are becoming increasingly popular for gene mapping of complex human diseases. Most of the published GWAS papers report results from single-marker-based analysis in which each SNP is analyzed individually. Although this simple approach has led to the discovery of disease susceptibility genes for many diseases, the identified SNPs often only explain a small fraction of the phenotypic variation, suggesting a large number of disease variants are yet to be discovered. There is growing evidence that gene-gene interactions are important contributors to

genetic variation in complex human diseases (Cordell et al. (1995); Cox et al. (1990); Howard et al. (2002); Moore and Williams (2002); Xu et al. (2004); Ochoa et al. (2004)). However, detecting gene-gene interactions remains a challenge due to the lack of powerful statistical methods. The most commonly used statistical approach for studying gene-gene interactions is to use a regression framework in which a pair of markers and their interaction terms are included as predictors. When a large number of markers are available, one might consider doing a stepwise regression (Hoh and Ott (2003)) or a two-stage analysis (Marchini et al. (2005)). Although such methods have been proven useful in simulation studies, they may lose power when multiple interacting variants exist in each gene.

One potential solution to the aforementioned problem is to perform interaction analysis at the gene level. There is increasing recognition for the importance of gene-based analysis (Neale and Sham (2004)). Several methods have been developed to test whether a gene is associated with the trait of interest (Gauderman et al. (2007); Wang and Abbott (2007); Wei et al. (2008); Li et al. (2009)). The central idea of these methods is to summarize marker genotypes into a few components so that the overall degrees of freedom are reduced while most information in the data is retained. Extensive simulations demonstrate that gene-based association analysis can increase the power of detecting genetic association compared to single-marker-based analysis. It is therefore reasonable to expect that gene-based interaction analysis may outperform SNP-based interaction analysis and lead to identification of novel disease susceptibility genes.

Recently, Li et al. (2009) proposed a novel gene-based association test - ATOM,

by combining optimally weighted markers within a gene. For each marker in the gene, either genotyped or untyped, an optimally weighted score is derived based on observed genotypes and linkage disequilibrium (LD) information in a reference dataset such as the HapMap (The International HapMap Consortium (2005), The International HapMap Consortium (2007)). To reduce the dimensionality of the data, ATOM tests for association using selected principal components (PC) of these derived scores. Simulations and analysis of real data showed improved power of ATOM over methods that do not incorporate external LD information, especially when the disease loci are not directly genotyped.

The success of ATOM motivated us to extend it to the analysis of gene-gene interactions. Here we describe a PCs framework for gene-based interaction analysis. We analytically derive the optimal weight for both quantitative and binary traits based on pairwise LD information. We then use PCs to summarize the information in each gene, and test for interactions between the PCs. We further extend this gene-based interaction analysis procedure to allow the use of imputation dosage scores obtained from popular imputation software packages MACH (Li et al. (2009)) or IMPUTE (Marchini et al. (2007)), which incorporates multilocus LD information. We evaluate the performance of the proposed tests by extensive simulations and the analysis of a candidate gene study on high-density lipoprotein cholesterol (HDL-C).

2.2 Methods

We consider the problem of gene-based interaction analysis between two genes with multiple markers in each gene.We first present the analytical solutions for quantitative and binary traits assuming the interacting trait loci are known.We then extend the method to the more realistic situation in which the interacting trait loci are unknown.

2.2.1 Quantitative Trait

Suppose the quantitative trait of interest Y, is influenced by the interaction between two diallelic quantitative trait loci (QTLs) located in two different genes. Let T_j and t_j (with frequencies p_{T_j} and p_{t_j}, respectively) denote the two alleles at QTL $j(j = 1, 2)$. Assume the mean of the trait value Y given genotypes g_{T_1} and g_{T_2} can be written as

$$E(Y|g_{T_1}, g_{T_2}) = \alpha_T + \beta_{T_1} g_{T_1} + \beta_{T_2} g_{T_2} + \beta_{T_1,T_2} g_{T_1} g_{T_2}, \tag{2.2.1}$$

where $g_{T_j} \in \{0, 1, 2\}$ is the number of allele T_j at QTL j. To detect interaction between the two QTLs, we wish to test $H_0 : \beta_{T_1,T_2} = 0$. However, as g_{T_1} and g_{T_2} may not be directly observed, the test of interaction is often accomplished through examination of genetic interactions between genotyped markers. Assume a diallelic marker j in gene j (with alleles A_j and a_j and allele frequencies p_{A_j} and p_{a_j}, respectively) is in LD with QTL j. We will show that the mean of Y given genotypes g_1, g_2, at marker 1 and marker 2 can be written as

$$E(Y|g_1, g_2) = \alpha + \beta_1 g_1 + \beta_2 g_2 + \beta_{1,2} g_1 g_2. \tag{2.2.2}$$

Equation 2.2.2 allows indirect assessment of interaction between the QTLs by testing $H_0 : \beta_{T_1,T_2} = 0$.

The regression coefficients β_{T_1,T_2} and $\beta_{1,2}$ reflect the magnitude of interaction between the QTLs and between the markers, respectively, and their relationship depends on the degree of LD between the QTLs and the markers. To explicitly derive their relationship, we note that $E(Y|g_1, g_2)$ can be written as

$$\begin{aligned} E(Y|g_1, g_2) &= \sum_{g_{T_1}} \sum_{g_{T_2}} E(Y|g_{T_1}, g_{T_2}) P(g_{T_1}|g_1) P(g_{T_2}|g_2) \\ &= \alpha_T + \beta_{T_1} H_1(g_1) + \beta_{T_2} H_2(g_2) + \beta_{T_1,T_2} H_1(g_1) H_2(g_2), \end{aligned} \tag{2.2.3}$$

where $H_j(g_j) = \sum_{g_{T_j}} g_{T_j} P(g_{T_j}|g_j), j = 1, 2$. Li et al. (2009) have shown that when both the QTLs and the markers are in Hardy-Weinberg equilibrium in the population,

$$H_j(g_j) = 2(p_{T_j} - \triangle_j / p_{a_j}) + [\triangle_j / (p_{A_j} p_{a_j})] \times g_j, \tag{2.2.4}$$

where $\triangle_j = p_{A_j T_j} - p_{A_j} p_{T_j}$ is the LD coefficient between QTL j and marker j. Therefore,

$$\begin{aligned} H_1(g_1) H_2(g_2) =& 4[p_{T_1} - \triangle_1 / p_{a_1}][p_{T_2} - \triangle_2 / p_{a_2}] \\ &+ 2[p_{T_2} - \triangle_2 / p_{a_2}][\triangle_1 / (p_{A_1} p_{a_1})] \times g_1 \\ &+ 2[p_{T_1} - \triangle_1 / p_{a_1}][\triangle_2 / (p_{A_2} p_{a_2})] \times g_2 \\ &+ [\triangle_1 / (p_{A_1} p_{a_1})][\triangle_2 / (p_{A_2} p_{a_2})] \times g_1 \times g_2. \end{aligned} \tag{2.2.5}$$

If we replace the items in 2.2.3 accordingly by those in 2.2.4 and 2.2.5, it becomes apparent that 2.2.2 holds, with

$$\beta_{1,2} = [\triangle_1 / (p_{A_1} p_{a_1})][\triangle_2 / (p_{A_2} p_{a_2})] \times \beta_{T_1,T_2}. \tag{2.2.6}$$

Equation 2.2.6 indicates that the two interaction coefficients β_{T_1,T_2}, $\beta_{1,2}$ differ only by a factor $[\triangle_1/(p_{A_1}p_{a_1})][\triangle_2/(p_{A_2}p_{a_2})]$, which is a function of the marker allele frequencies and the LD coefficients between the QTLs and the markers. The above derivation can be readily extended to binary traits such as disease affection status (see Supplementary Material).

From the above derivation, we can see that if we define a weighted genotype score at marker j, $g_j^* = [\triangle_j/(p_{A_j}p_{a_j})]g_j$, then the corresponding mean model for Y given the weighted genotype score becomes

$$\begin{aligned} E(Y|g_1^*, g_2^*) &= \alpha^* + [\beta_{T_1} + 2\beta_{T_1T_2}(p_{T_2} - \triangle_2/p_{a_2})]g_1^* \\ &\quad + [\beta_{T_2} + 2\beta_{T_1T_2}(p_{T_1} - \triangle_1/p_{a_1})]g_2^* + \beta_{T_1T_2}g_1^*g_2^* \\ &= \alpha^* + \beta_{T_1}g_1^* + \beta_{T_2}g_2^* + \beta_{T_1,T_2}g_1^*g_2^*, \end{aligned} \tag{2.2.7}$$

with the interaction coefficient being the same as that in equation 2.2.1. This indicates that for any pair of markers with one from each of the two genes, using weighted genotype scores will result in models that share the same interaction coefficient $\beta_{T_1T_2}$, a fact that we will use below to combine multiple markers within a gene.

Suppose m_j diallelic markers in gene j are genotyped, with alleles $1_{l_j}^{(j)}$ and $0_{l_j}^{(j)}$ for marker $l_j (1 \leq l_j \leq m_j)$ and allele frequencies $p_{l_j}^{(j)}$ and $q_{l_j}^{(j)}$, respectively. The above derivations suggest that for individual $i (1 \leq i \leq n)$ and marker $l_j (1 \leq l_j \leq m_j)$ in gene j, we may consider the weighted genotype score $g_{i,l_j}^{(j)*} = [\triangle_{l_j}^{(j)}/(p_{l_j}^{(j)}q_{l_j}^{(j)})]g_{i,l_j}^{(j)}$, where $\triangle_{l_j}^{(j)}$ is the LD coefficient between QTL j and marker l_j in gene j, and $g_{i,l_j}^{(j)}$ denotes the number of allele $1_{l_j}^{(j)}$ carried by individual i. Then the mean of the trait value Y_i given weighted genotype scores at marker l_1 in gene 1 and marker l_2 in gene 2 will be $E(Y|g_{i,l_1}^{(1)*}, g_{i,l_2}^{(2)*}) = \alpha_{l_1l_2}^* + \beta_{l_1}^* g_{i,l_1}^{(1)*} + \beta_{l_2}^* g_{i,l_2}^{(2)*} + \beta_{T_1T_2}g_{i,l_1}^{(1)*}g_{i,l_2}^{(2)*}$. When all possible

marker combinations in the two genes are considered, then all $m_1 \times m_2$ interaction terms share a common coefficient β_{T_1,T_2}. This suggests that for individual i, we can aggregate the information from m_j markers in gene j by defining a score

$$S_i^{(j)} = \frac{1}{m_j} \sum_{l_j=1}^{m_j} \frac{\triangle_{l_j}^{(j)}}{p_{l_j}^{(j)} q_{l_j}^{(j)}} g_{i,l_j}^{(j)} = \frac{1}{m_j} \sum_{l=1}^{m_j} w_{l_j}^{(j)} g_{i,l_j}^{(j)}, \tag{2.2.8}$$

and then assess interaction between the two QTLs by examining the cross product of their scores $S_i^{(1)} \times S_i^{(2)}$. In situations in which the trait locus is in complete or strong LD with a genotyped marker or is itself genotyped, the score in equation 2.2.8 may not work well as the other markers will simply add noise and dilute the association signal. Given this consideration, an alternative weighted genotype score is

$$S_i^{(j)} = w_{l_{j,max}}^{(j)} g_{i,l_{j,max}}^{(j)}, \tag{2.2.9}$$

where $l_{j,max}$ is the genotyped marker that has the strongest LD with the trait locus as measured by r^2.

2.2.2 Estimation of Weights

In the previous sections, we assumed the trait loci are known. In real data analysis, the locations of the trait loci are unknown. It is reasonable to assume that each of the known polymorphisms in the gene, either genotyped or untyped in the study sample, is equally likely to be the trait locus. For each such locus, we can estimate the weights for all the genotyped markers and calculate a score for the locus. Following Li et al. (2009), we propose to estimate the weight using LD information obtained from a reference database such as that generated by the HapMap, other publicly available

dense SNP data sets or resequencing data from a subset of the study sample. Suppose M_j markers are available for gene j in the reference data set, and they are a superset of the markers genotyped in the study sample. If marker $k_j (1 \leq k_j \leq M_j)$ in the reference data set is the trait locus, then the weight for marker k_j in the study sample is

$$w_{k_j,l_j}^{(j)} = \frac{\triangle_{k_j,l_j}^{(j)}}{p_{l_j}^{(j)} q_{l_j}^{(j)}}, \tag{2.2.10}$$

where $\triangle_{k_j,l_j}^{(j)}$ is the LD coefficient between markers k_j and l_j, and $p_{l_j}^{(j)}$ and $q_{l_j}^{(j)}$ are allele frequencies at marker l_j in gene j. These quantities can be estimated from the reference data set. For individual i and each marker k_j in the reference data set, we can calculate a weighted genotype score

$$S_{i,k_j}^{(j)} = \frac{1}{m} \sum_{l_j=1}^{m_j} w_{k_j,l_j}^{(j)} g_{i,l_j}^{(j)}, \tag{2.2.11}$$

or an alternative weighted genotype score

$$S_{i,k_j}^{(j)} = w_{k_j,l_{j,max}}^{(j)} g_{i,l_{j,max}}^{(j)}, \tag{2.2.12}$$

where $l_{j,max}$ is the genotyped marker that has the strongest LD with marker k_j.

We note that the weighted genotype scores in equations 2.2.11 and 2.2.12 share similarity with imputation dosage scores, and they can be considered as a simple version of the multilocus LD-based imputation dosage scores obtained from software packages such as MACH and IMPUTE. Although using pairwise-LD information only, the weighted genotype scores in equations 2.2.11 and 2.2.12 provide an intuitive justification of why incorporating external LD information may provide power gain for

association testing. In the following sections, we will consider both the pairwise LD-based weighted genotype scores and multilocus LD-based imputation dosage scores in the testing procedure.

2.2.3 Gene-based Interaction Analysis

Once we have calculated the scores, either weighted genotype scores in equation 2.2.11 and 2.2.12 or imputation dosage scores in MACH or IMPUTE, for each marker in the reference data set, we can then test for gene-gene interaction based on the scores $(S_1^{(j)}, \ldots, S_{M_j}^{(j)})$ for gene j, where $S_{k_j}^{(j)} = (S_{1,k_j}^{(j)}, \ldots, S_{n,k_j}^{(j)})$ and n is the total number of individuals in the study. As the trait loci are unknown, a simple test of interaction could be to include all pairwise interactions of the imputation dosage scores in a regression framework and then test for their overall significance. However, this approach may suffer from low power due to the large number of degrees of freedom involved. To efficiently aggregate all information while reducing the degrees of freedom, we propose to test for gene-gene interaction using PCs obtained from the scores. Without loss of generality, for gene j, we order the PCs such that $PC_1^{(j)}$ has the largest variance and $PC_2^{(j)}$ has the second largest variance and so on.

Once the PCs are computed, we can then test for gene-gene interaction by conducting a regression analysis with a set of selected PCs and their interactions as covariates. As the PCs are ordered by the magnitude of explained variance, for each gene, we select the first several PCs that explain a prespecified fraction of the total variance. Suppose L_j PCs are selected for gene j. For a binary trait, we can fit the

data with the following logistic regression model (see Supplementary Material)

$$\begin{aligned}\text{logit}[P(Y=1|\text{genotypes})] = \alpha &+ \sum_{l_1=1}^{L_1} \beta_{1,l_1} PC_{1,l_1} \\ &+ \sum_{l_2=1}^{L_2} \beta_{2,l_2} PC_{2,l_2} \\ &+ \sum_{l_1=1}^{L_1} \sum_{l_2=1}^{L_2} \beta_{l_1,l_2} PC_{1,l_1} PC_{2,l_2}.\end{aligned} \tag{2.2.13}$$

For a quantitative trait, we can fit the data with a linear regression model.

Under the null hypothesis of no interaction between the two genes,$H_0 : \beta_{l_1,l_2} = 0$ for $l_1 = 1, \ldots, L_1$ and $l_2 = 1, \ldots, L_2$. We can test this null hypothesis by a likelihood ratio test, and the corresponding test statistic is approximately distributed as a χ^2 distribution with $L_1 \times L_2$ degrees of freedom. We call this test as a *global* test. Alternatively, we can conduct *pairwise* interaction analysis between all selected PCs and choose the statistic for the most significant pair as the test statistic and evaluate its significance by Bonferroni correction. We call this test as a pairwise test. In our analyses, we used 90% threshold for the fraction of variance as it generally provides better power than other variance thresholds in scenarios we considered. We note that for binary traits, the null hypothesis tested by logistic regression is only an approximation to the null hypothesis that $(f_{2,2} - f_{2,0}) - (f_{0,2} - f_{0,0}) = 0$, and thus the weighted genotype scores we derived earlier may not be 'optimal'. However, as shown in the Supplementary Material, this approximation is probably valid as long as the interaction effects are not too strong and the disease is not common.

2.3 Results

In this section, we evaluate the performance of the gene-based interaction tests for binary traits and compare with SNP-based interaction test. We considered four gene-based interaction tests: (1) ATOM-AVG, which uses weighted genotype scores from equation 2.2.11; (2) ATOM-MAX, which uses weighted genotype scores from equation 2.2.12; (3) MACH, which uses imputation dosage scores from MACH; and (4) PCA, which uses genotyped markers only. For each test T, we considered two versions: (1) the global version, which tests for the joint interaction effect of all selected PCs; and (2) the pairwise version, which tests for the pairwise interaction among all selected PCs. Significance for the pairwise version is adjusted by Bonferroni correction. For the SNP-based interaction analysis, we only considered the pairwise version as the power of the global version is extremely low due to the large number of degrees of freedom.

2.3.1 Comparison of Type I Error and Power

We simulated data based on the LD structures of two genes *CHI3L2* (Figure 2.1) and *PTPN22* (Figure 2.2), both are located on chromosome 1 but are in linkage equilibrium with each other. For each gene, we considered common SNPs with minor allele frequency ≥ 0.05 and selected tagSNPs using the program Tagger (De et al. (2005)) with pairwise tagging at $r^2 \geq 0.8$. We identified 25 common SNPs for *CHI3L2* and selected seven tagSNPs; for *PTPN22*, 29 common SNPs and 9 tagSNPs. We assumed that only the tagSNPs were genotyped and available for analysis, a common scenario

in both candidate gene and GWAS studies. To simulate case-control data with LD, we first estimated the haplotype frequencies of the tagSNPs for each gene, and then simulated the genotype data according to the estimated haplotype frequencies. We considered two situations: (1) each gene has only one disease locus; and (2) each gene has two disease loci.

For the first situation, we designated one locus in each gene as the disease locus, and the case-control status for individual i was simulated according to the following model

$$\text{logit}[P(Y = 1|g_{i,D}^{(1)}, g_{i,D}^{(2)})] = \alpha + 0.2(g_{i,D}^{(1)} + g_{i,D}^{(2)}) + 0.3g_{i,D}^{(1)}g_{i,D}^{(2)},$$

where α is determined in a way such that the overall disease prevalence is 5%. Power was estimated based on 1000 replicate data sets each consisting of 2000 cases and 2000 controls and significance was assessed at the 1% level. The type I error rate was evaluated based on 10000 data sets by setting the interaction effect in the above logit model to 0. Since gene-based interaction tests based on ATOM and MACH require external LD information, we simulated 60 individuals (mimicking the HapMap CEU samples) as a reference data set and then calculated the weighted genotype scores or the imputation dosage scores using the LD information estimated from these 60 individuals.

As the performance of different tests may vary depending on whether the disease loci are genotyped or not, we considered three scenarios: 1) both disease loci are genotyped; 2) only one of the disease loci is genotyped; and 3) both disease loci are untyped. A thorough evaluation of all tests would require the consideration

of $25 \times 29 = 725$ combinations. To avoid extensive simulations from all marker combinations, we classified the markers in each gene into three categories according to LD levels. Specifically, a marker is classified into the 'strong LD' category if five or more markers in the gene have $r^2 > 0.8$ with it; a marker is in the 'moderate LD' category if three to five markers in the gene have $r^2 > 0.8$ with it; the rest are in the 'weak LD' category. On the basis of this classification, markers in *CHI3L* fall into either strong or weak LD categories. By classifying markers in this manner, we were able to investigate the performance of various tests under a wide range of settings, yet avoided simulations of all marker combinations.

Table 2.1 displays the estimated type I error rates under two-locus interaction model. The type I error rates of all tests are under control. Not surprisingly, for each test, the pairwise version of the test is more conservative than the global version due to the correction of a large number of pairwise comparisons. Table 2.2 shows the estimated power. As expected, when there is a single disease locus in each gene, $T_{SNP-pairwise}$ consistently outperforms the other tests. Among the other tests we considered, $T_{ATOM-AVG-global}$, $T_{ATOM-MAX-global}$ and $T_{MACH-global}$, which incorporate external LD information, offer better power, followed by $T_{PCA-global}$.We note that the powers of ATOM- and MACH-based tests are similar, despite that MACH is much more computationally intensive. For example, it requires 210s to finish one simulation for MACH-based tests with 2000 cases and 2000 controls; however, the required computing time for ATOM-based tests is only 5s, 40 times faster.

For complex diseases, it might be an oversimplification to consider only one disease locus per gene. To evaluate the performance of different tests under a more

complicated setting, we considered a model in which two loci in *CHI3L2* interact with two loci in *PTPN22*. Specifically, we simulated case-control status according to the model:

$$\text{logit}[P(Y_i = 1|g^{(1)}_{i,D_1}, g^{(1)}_{i,D_2}, g^{(2)}_{i,D_1}, g^{(2)}_{i,D_2})] = \alpha + 0.3(g^{(1)}_{i,D_1} + g^{(1)}_{i,D_2} + g^{(2)}_{i,D_1} + g^{(2)}_{i,D_2}) +$$
$$0.4(g^{(1)}_{i,D_1}g^{(2)}_{i,D_1} + g^{(1)}_{i,D_1}g^{(2)}_{i,D_2} + g^{(1)}_{i,D_2}g^{(2)}_{i,D_1} + g^{(1)}_{i,D_2}g^{(2)}_{i,D_2}).$$

Again, the overall disease prevalence was set at 5% by adjusting the value of α. For type I error estimation, we set the coefficient for the interaction effect at 0.

As shown in Table 2.3, the type I error rates of all interaction tests are under control. Table 2.4 shows the power comparison results. These results indicate that all gene-based interaction tests outperform the SNP-based test. For example, the power advantage of $T_{ATOM-MAX-global}$ over $T_{SNP-pairwise}$ as measured by mean power difference ranged from 12.7 to 27.3%. This is much higher than the mean power difference $(4.1 - 9.7\%)$ between the two tests under the simpler disease models in Table 2.2. This indicates that SNP-based interaction analysis is not sufficient when multiple loci in a gene interact with multiple loci in another gene. Among all gene-based tests we considered, $T_{MACH-global}$ is generally the most powerful test, followed by $T_{ATOM-MAX-global}$, $T_{ATOM-AVG-global}$ and $T_{PCA-global}$. It is worth noting that the power of $T_{ATOM-MAX-global}$ is only slightly lower than $T_{MACH-global}$ despite that MACH is much more computationally intensive. The pairwise versions of these three tests are typically less powerful than the global versions of the tests. Moreover, our results clearly indicate the advantage of incorporating external LD information in the analysis. The power gain of $T_{ATOM-MAX-global}$ over $T_{PCA-global}$ as measured by the

mean power difference ranged from 3.3 to 7.9%, and the power gain of $T_{MACH-global}$ over $T_{PCA-global}$ ranged from 4.6 to 10.7%.

2.3.2 Application to IBC HDL Data Set

We applied the three gene-based interaction tests to an ongoing candidate gene study on subjects with extreme levels of HDL-C. In this study, 625 subjects of European ancestry with HDL > 90th percentile were considered as cases and 606 subjects with HDL < 25th percentile were considered as controls. All study subjects were genotyped using the IBC 50K SNP array (Keating et al. (2008)). Our previous SNP pairwise interaction analysis on this data set reveals that a number of SNPs in *CETP* significantly interact with several SNPs in *BCAT1*. It is well known that *CETP* promotes the transfer of cholesteryl esters from HDL to low-density lipoprotein, and individuals that are genetically deficient for CETP often have extremely high HDL levels (Brown et al. (1989), Inazu et al. (1990)). In a recent GWAS on biochemical traits, *BCAT1* is shown to be significantly associated with serum albumin concentration (Zemunik et al. (2009)). As albumin is correlated with HDL (Gillum (1993)), it is possible that *CETP* and *BCAT1* interact in modulating the level of HDL-C.

Figure 2.3 and 2.4 display the LD structures of *CETP* and *BCAT1* estimated using the HDL data set. We downloaded genotype data at these two genes for the CEU samples from the HapMap website. For *CETP*, there are 31 common SNPs in the HapMap, whereas the HDL data set has 57, with 27 common SNPs in both data sets. As the HapMap data set does not provide much additional LD information,

for ATOM-based tests, we calculated the weighted genotype scores using the LD information provided by the 57 SNPs in the HDL controls. For *BCAT1*, 164 common SNPs are in the HapMap and 79 are in the HDL data set, with 56 in both. For the 164 SNPs in the HapMap, we calculated their weighted genotype scores using the LD information provided by the HapMap; for the 23 SNPs in the HDL data set but not in the HapMap, we used their observed genotypes in the HDL data set.

The *BCAT1* SNPs are in several LD blocks with weak LD between some of the blocks, requiring 23 PCs to explain 90% of the variance. Testing interaction using all SNPs in *BCAT1* may have low power due to the large number of degrees of freedom. To reduce the dimensionality, we divided the SNPs in *BCAT1* into four blocks (Figure 2.4) and tested interaction between *CETP* and each of the four blocks. We found significant interaction between *CETP* and the third block of *BCAT1*. The P-value of $T_{ATOM-AVG-global}$ is 0.0034. In comparison, the P-values of $T_{ATOM-MAX-global}$, $T_{MACH-global}$ and $T_{PCA-global}$ are 0.22, 0.25 and 0.072, respectively. The P-values of the pairwise versions of the four tests are 0.035, 0.062, 0.38 and 0.029, respectively. The P-value of $T_{SNP-pairwise}$ is 0.078. Compared with other gene-based interaction tests, $T_{ATOM-AVG-global}$ clearly revealed stronger evidence of association.

2.4 Discussion

We have proposed a PC framework for gene-based interaction analysis. Our tests are based on the aggregation of information from weighted genotype scores using pairwise LD information or imputation dosage scores using multilocus LD informa-

tion in a gene. To reduce dimensionality, the scores within a gene are further summarized into PCs and then used in a regression framework for interaction analysis. By extensive simulations under various settings and the analysis of a real data set, we demonstrated that gene-based interaction tests are a powerful alternative to the conventional SNP-based interaction test and to approaches that do not incorporate external LD information.

The gene-based interaction tests consider each gene as a testing unit and tests for interaction at the gene level. Compared with methods that operate at the marker level, a key advantage of gene-based interaction tests lies in their ability to capture all potential risk conferring variants in a gene. This makes gene-based interaction tests particularly attractive when multiple disease loci in a gene interact with multiple disease loci in another gene. We note that when a single locus in a gene interacts with a single locus in another gene, or when some of the interaction effects are weak when more than two loci interact, the SNP-based interaction test may perform well, as such a simple test can capture the interaction effect more effectively than the gene-based interaction tests.

Another advantage of gene-based interaction analysis over the conventional SNP-based interaction analysis is that it requires much less number of tests. For example, for the IBC data with 50000 SNPs genotyped in 2000 candidate genes, the conventional SNP pairwise interaction analysis will involve 1.25 billion tests. In contrast, using gene-based interaction analysis, the number of tests is reduced to 2 million. For large-scale candidate gene and GWAS data sets, gene-based interaction tests can be used as a screening tool. After a pair of significant interacting genes is identified,

one can then conduct further investigation to evaluate which SNPs within the genes significantly interact.

Our method concerns with gene-based tests of interaction effect. We note that there exist gene-based methods that jointly test for the main effect and the interaction effect (Chatterjee et al. (2006); Chapman and Clayton (2007)). Although the goals of these tests are slightly different from ours, they all aim to incorporate information contributed by multiple markers in a gene. How to extend the proposed PC framework to jointly test for the main and interaction effects would merit further research.

Figure 2.1: LD structure of the *CHI3L2* gene on chromosome 1 in the HapMap CEU samples. Displayed is estimated r^2 for 25 SNPs with minor allele frequency (MAF) ≥ 0.05. SNPs within the black boxes are tagSNPs selected using the Tagger program at r^2 threshold of 0.8.

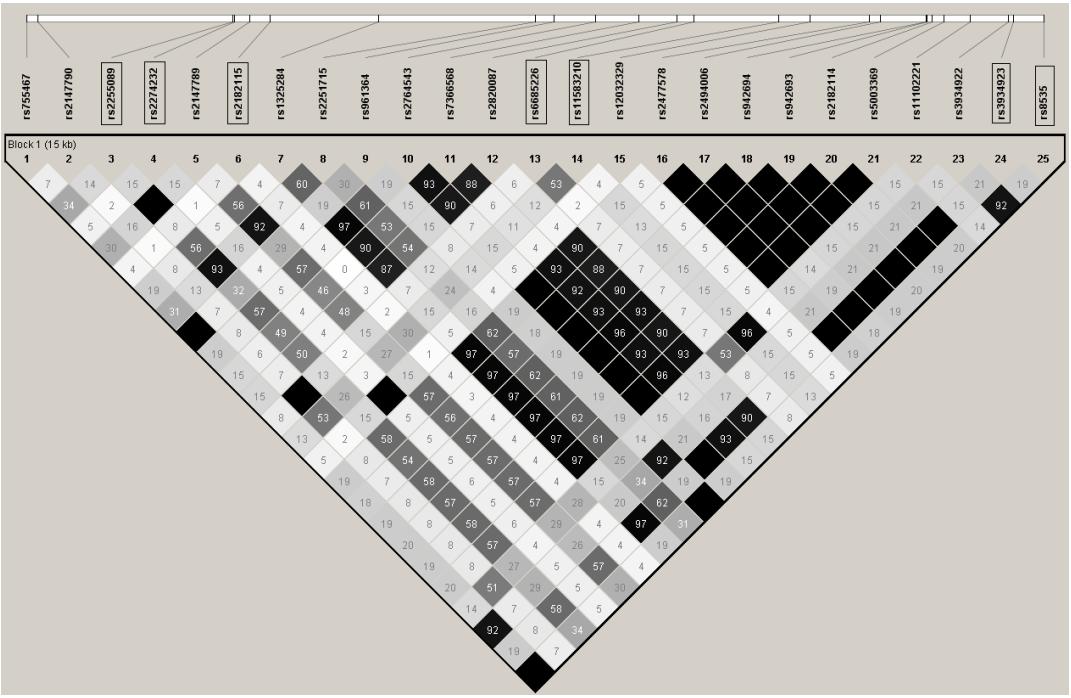


Figure 2.2: LD structure of the *PTPN22* gene on chromosome 1 in the HapMap CEU samples. Displayed is estimated r^2 for 29 SNPs with MAF ≥ 0.05. SNPs within the black boxes are tagSNPs selected using the Tagger program at r^2 threshold of 0.8.

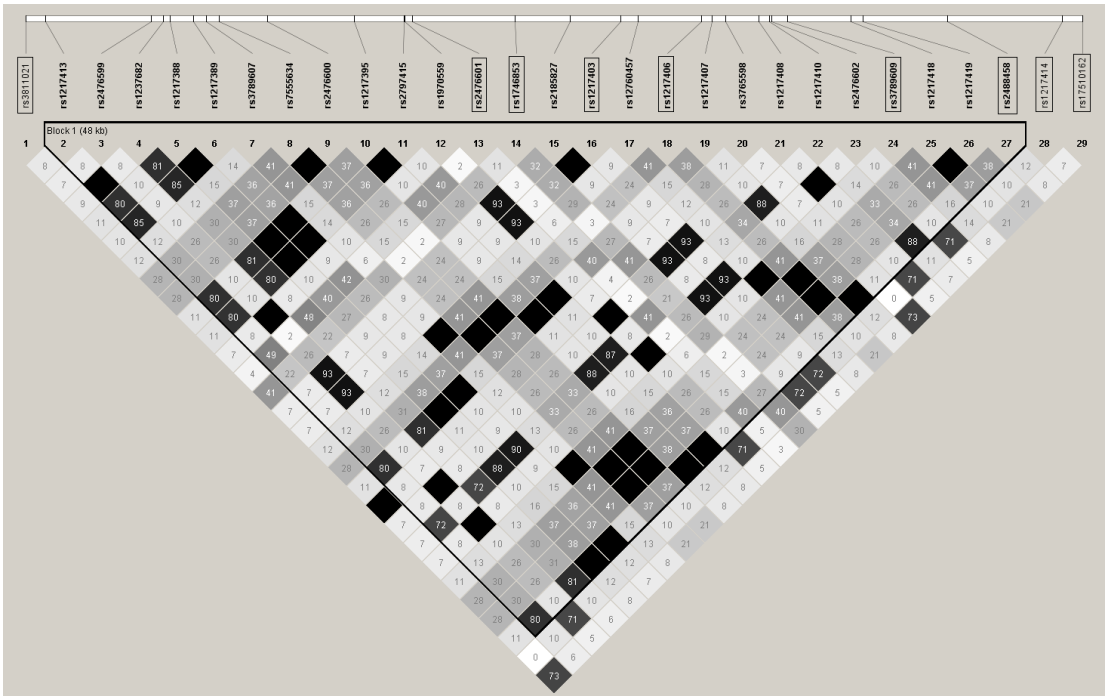


Figure 2.3: LD structure of the *CETP* gene on chromosome 16 in the IBC samples.

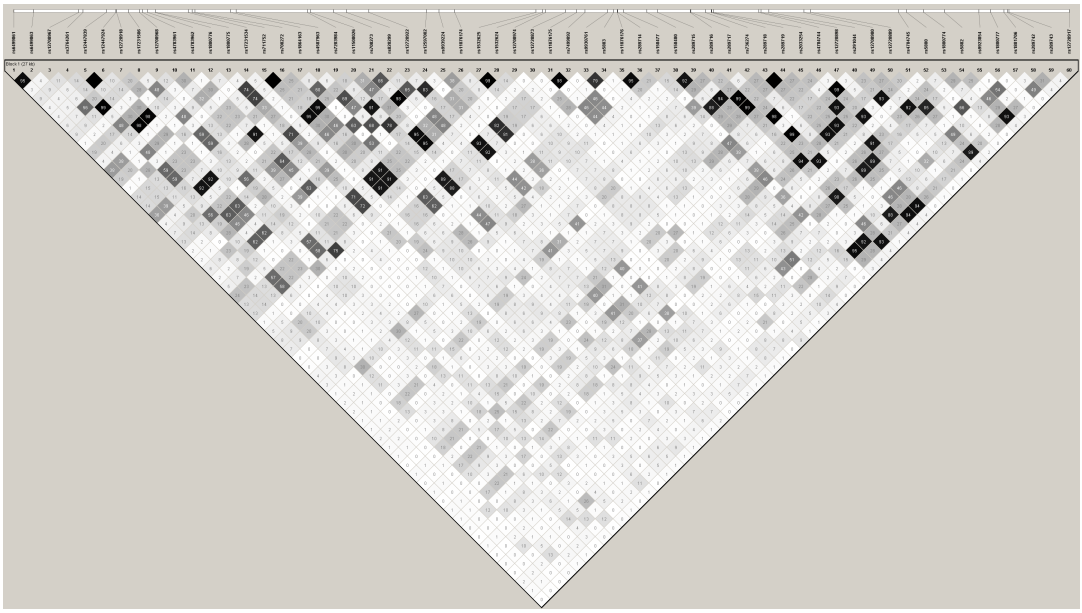


Figure 2.4: LD structure of the *BCAT1* gene on chromosome 12 in the IBC samples.

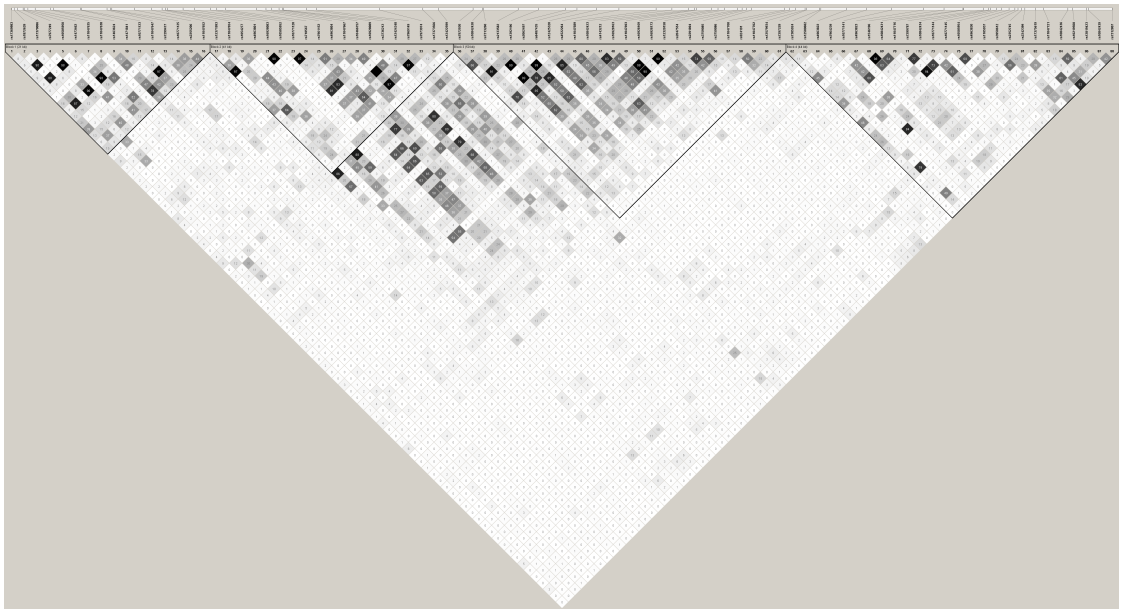


Table 2.1: Type I error rates (%) under a two-locus interaction model in which one locus in *CHI3L2* interacts with one locus in *PTPN22*

Disease locus in CHI3L2		*Disease locus in* PTPN22		*Interaction tests*								
				SNP	*PCA*		*ATOM-AVG*		*ATOM-MAX*		*MACH*	
LD category	*SNP*	*LD category*	*SNP*	*Pairwise*	*Global*	*Pairwise*	*Global*	*Pairwise*	*Global*	*Pairwise*	*Global*	*Pairwise*
G		G										
W	3	W	13	0.82	1.04	0.79	1.15	0.70	1.10	1.00	1.08	0.80
S	24	S	27	0.72	1.10	0.94	0.93	0.68	1.03	0.91	1.13	1.03
G		U										
W	3	S	5	0.65	1.09	0.81	1.10	0.66	1.00	0.98	1.17	0.77
S	24	M	10	0.73	1.16	0.95	0.99	0.51	1.20	0.82	0.98	0.88
U		G										
W	5	S	27	0.71	1.18	0.82	1.15	0.57	1.04	0.90	1.09	0.96
S	7	W	24	0.74	1.06	0.97	1.12	0.84	1.08	1.13	0.91	0.75
U		U										
W	5	M	12	0.60	0.85	0.86	1.00	0.48	0.86	0.91	0.93	0.77
S	7	W	7	0.72	0.98	1.05	1.14	0.61	1.00	1.06	0.92	0.91

Abbreviations: G, genotyped; M, moderate LD category; S, strong LD category; U, untyped; W, weak LD category.

Table 2.2: Comparison of power (%) under a two-locus interaction model in which one locus in *CHI3L2* interacts with one locus in *PTPN22*

Disease locus in CHI3L2		*Disease locus in* PTPN22		*Interaction tests*								
				SNP	*PCA*		*ATOM-AVG*		*ATOM-MAX*		*MACH*	
LD category	*SNP*	*LD category*	*SNP*	*Pairwise*	*Global*	*Pairwise*	*Global*	*Pairwise*	*Global*	*Pairwise*	*Global*	*Pairwise*
G		G										
W	3	W	13	40.6	23.9	21.0	29.4	24.2	18.5	6.1	20.1	17.7
W	3	M	16	60.3	51.9	52.1	56.1	48.7	54.2	29.2	56.6	40.3
W	3	S	27	71.6	65.6	66.0	67.9	65.6	68.3	40.2	66.3	58.1
S	24	W	13	36.4	21.6	14.6	24.9	18.7	17.6	5.6	16.3	17.8
S	24	M	16	55.9	44.5	38.9	47.4	39.3	51.3	26.3	50.0	46.5
S	24	S	27	68.0	61.5	52.1	58.7	55.0	64.7	38.7	66.7	72.1
			Mean	55.5	44.8	40.8	47.4	41.9	45.8	24.4	46.0	42.1
U		G										
W	5	W	13	40.1	20.3	18.1	27.6	24.0	18.6	14.5	19.7	14.8
W	5	M	16	60.1	51.4	54.7	58.6	50.5	54.3	37.1	56.5	41.2
W	5	S	27	72.7	66.7	62.2	69.1	63.2	67.5	54.9	66.3	58.5
S	7	W	14	80.1	73.7	65.5	72.6	60.1	75.0	76.9	70.0	56.1
S	7	M	16	55.3	46.3	41.2	50.8	39.2	52.4	47.7	50.1	46.0
S	7	S	27	68.6	60.2	50.8	59.5	52.8	62.7	68.4	66.8	72.8
			Mean	62.8	53.1	48.8	56.4	48.3	55.1	49.9	54.9	48.2
G		U										
W	3	W	17	73.6	69.3	58.7	71.7	59.3	70.4	52.1	67.3	56.8
W	3	M	3	60.1	52.8	52.0	57.4	50.1	56.3	38.6	57.7	42.5
W	3	S	5	74.3	67.5	62.8	69.7	63.5	69.0	55.2	65.9	57.8
S	24	W	17	67.4	59.4	42.2	59.8	48.2	63.4	63.4	65.8	67.5
S	24	M	15	58.5	50.3	42.9	52.4	40.4	54.2	49.1	52.8	46.5
S	24	S	5	66.5	59.8	50.9	58.5	51.8	63.2	67.9	63.6	68.4
			Mean	66.9	59.8	51.6	61.6	52.2	62.8	54.4	62.2	56.6
U		U										
W	5	W	17	74.1	66.5	60.7	72.5	60.1	69.3	53.7	68.9	59.1
W	5	M	12	61.9	55.5	53.8	60.7	50.5	56.3	36.9	54.8	39.9
W	5	S	5	72.9	65.0	62.6	69.8	63.6	65.8	53.6	67.5	61.3
S	7	W	7	66.4	59.0	45.7	61.0	51.2	62.4	66.6	63.0	66.1
S	7	M	3	66.6	47.1	40.7	50.5	40.5	53.3	47.2	54.3	50.7
S	7	S	5	66.1	61.5	46.6	57.6	51.7	63.6	68.7	62.0	69.0
			Mean	66.2	59.1	51.7	62.0	52.9	61.8	54.5	61.8	57.7

Abbreviations: G, genotyped; M, moderate LD category; S, strong LD category; U, untyped; W, weak LD category.

Table 2.3: Type I error rates (%) under a four-locus interaction model in which two loci in *CHI3L2* interact with two loci in *PTPN22*

Disease loci in CHI3L2		*Disease loci in* PTPN22		*Interaction tests*								
				SNP	*PCA*		*ATOM-AVG*		*ATOM-MAX*		*MACH*	
LD category	*SNP*	*LD category*	*SNP*	*Pairwise*	*Global*	*Pairwise*	*Global*	*Pairwise*	*Global*	*Pairwise*	*Global*	*Pairwise*
G,G		G,G										
W,W	4,14	M,S	16,27	0.83	1.10	0.91	1.12	1.01	1.18	1.08	1.05	1.00
W,S	4,24	W,M	1,16	0.76	0.99	0.92	1.06	0.98	0.98	0.94	0.97	1.03
G,U		G,U										
W,W	2,4	W,M	1,3	0.73	1.05	1.00	1.03	0.92	1.04	0.77	1.17	1.03
W,S	4,7	M,M	3,18	0.62	1.28	0.92	1.21	1.05	1.13	0.94	1.08	1.09
U,U		U,U										
W,S	15,1	W,W	7,20	0.82	1.15	0.98	1.18	1.05	1.13	0.97	1.19	1.06
W,S	15,11	M,S	23,19	0.73	0.87	0.97	0.96	0.99	0.88	0.95	1.07	1.01

Abbreviations: G, genotyped; M, moderate LD category; S, strong LD category; U, untyped; W, weak LD category.

Table 2.4: Comparison of power (%) under a four-locus interaction model in which two loci in *CHI3L2* interact with two loci in *PTPN22*

Disease loci in CHI3L2		*Disease loci in* PTPN22		*Interaction tests*								
				SNP	*PCA*		*ATOM-AVG*		*ATOM-MAX*		*MACH*	
LD category	*SNP*	*LD category*	*SNP*	*Pairwise*	*Global*	*Pairwise*	*Global*	*Pairwise*	*Global*	*Pairwise*	*Global*	*Pairwise*
G,G		G,G										
W,W	4,14	W,W	1,13	25.5	44.8	30.6	51.9	22.7	47.8	22.2	53.2	25.7
W,W	4,14	W,M	1,16	44.8	68.5	39.5	74.1	33.7	72.5	52.3	73.3	41.2
W,W	4,14	M,M	16,18	59.7	71.7	50.0	75.0	56.6	74.8	61.3	75.7	65.7
W,W	4,14	M,S	16,27	18.8	58.0	39.7	67.3	52.3	69.5	57.7	73.3	56.7
W,W	4,14	W,S	24,27	39.9	60.7	37.0	67.3	36.0	66.6	32.8	68.7	39.8
W,S	4,24	W,W	13,28	42.8	55.5	62.1	60.5	71.8	47.7	46.7	45.4	37.5
W,S	4,24	W,M	1,16	77.3	66.2	62.1	72.9	60.8	72.5	56.9	75.2	46.9
W,S	4,24	M,M	16,18	73.8	69.7	64.3	74.4	66.7	75.4	62.8	75.1	67.8
W,S	4,24	M,S	16,27	42.3	60.6	68.7	68.3	80.5	73.8	66.5	73.3	64.9
W,S	4,24	W,S	24,27	71.2	60.6	62.1	67.7	58.1	67.7	39.4	71.7	49.9
			Mean	49.6	61.6	51.6	67.9	53.9	66.8	49.9	68.5	49.6
G,U		G,U										
W,W	3,15	W,W	1,7	84.6	94.9	93.8	96.9	95.4	98.1	88.8	98.4	86.7
W,W	2,4	W,M	1,3	31.6	58.0	34.3	50.1	23.0	51.3	28.6	53.2	24.3
W,W	2,4	M,M	3,18	48.6	67.8	48.4	63.6	44.7	63.5	48.2	64.5	50.6
W,W	3,15	M,S	16,19	82.2	96.1	92.7	97.6	95.3	98.2	94.6	98.4	94.2
W,W	2,4	W,S	1,22	30.1	60.2	37.8	49.5	22.8	50.1	25.6	55.3	25.3
W,S	4,7	W,W	1,7	40.9	58.2	67.9	67.8	77.3	71.9	55.1	73.4	57.0
W,S	4,7	W,M	1,3	74.7	66.0	61.4	72.5	62.7	73.0	58.3	74.7	48.5
W,S	4,7	M,M	3,18	75.9	69.8	63.9	74.4	60.8	76.4	62.1	77.1	70.7
W,S	4,7	M,S	16,19	40.3	59.8	67.8	68.1	77.3	71.8	62.4	72.5	66.2
W,S	4,7	W,S	1,19	78.5	66.5	64.6	73.8	61.6	75.4	56.8	75.2	56.1
			Mean	58.7	69.7	63.3	71.4	62.1	73.0	58.1	74.3	58.0
U,U		U,U										
W,W	1,15	W,W	7,20	20.8	56.8	54.1	67.0	57.9	70.0	41.5	70.8	39.3
W,W	1,15	W,M	20,3	46.4	66.3	49.7	72.9	45.0	71.6	43.3	71.9	31.6
W,W	1,15	M,M	3,8	29.0	46.1	25.5	52.2	24.9	51.8	21.2	74.3	59.2
W,W	1,15	M,S	3,22	18.0	54.5	51.5	61.9	54.7	63.3	45.6	63.2	41.7
W,W	1,15	W,S	7,19	41.8	62.4	53.9	69.1	39.3	67.6	28.4	70.8	33.2
W,S	15,11	W,W	7,20	41.9	59.1	68.2	65.7	75.2	70.4	53.7	74.2	56.0
W,S	15,11	W,M	7,15	74.7	62.7	56.0	71.1	63.8	71.6	60.3	72.0	59.6
W,S	15,11	M,M	3,8	76.0	68.9	63.5	75.1	64.9	75.6	62.8	76.6	71.3
W,S	15,11	M,S	12,19	17.7	63.0	55.4	56.3	45.9	64.9	37.6	62.2	37.6
W,S	15,11	M,S	23,19	45.4	65.5	71.9	74.0	80.6	76.7	66.2	76.3	67.3
			Mean	41.2	60.5	55.0	66.5	55.2	68.4	65.1	71.2	49.7

Abbreviations: G, genotyped; M, moderate LD category; S, strong LD category; U, untyped; W, weak LD category

2.5 Supplementary Material

Derivation of ATOM Weighted Genotype Scores for Binary Trait

We now consider gene-based interaction analysis for binary traits, such as disease status. Assuming there is no interaction on the penetrance scale, i.e., the two-locus penetrance $f_{g_{D_1},g_{D_2}} = P(Y = 1|g_{D_1}, g_{D_2})$ can be written as the sum of the disease risks due to each disease locus. Here $g_{D_j} \in \{0, 1, 2\}$ is the number of disease allele D_j at disease locus $j(= 1, 2)$. It is easy to show that this definition of no interaction is equivalent to $(f_{2,2} - f_{2,0}) - (f_{0,2} - f_{0,0}) = 0$. Suppose marker j is in LD with disease locus j. Let $\varphi_{g_1,g_2} = P(Y = 1|g_1, g_2)$ denote the penetrance for genotypes (g_1, g_2) at markers 1 and 2. Then

$$\begin{aligned}\varphi_{g_1,g_2} = P(Y = 1|g_1, g_2) &= \sum_{g_{D_1}} \sum_{g_{D_2}} P(Y = 1, g_{D_1}, g_{D_2}|g_1, g_2) \\ &= \sum_{g_{D_1}} \sum_{g_{D_2}} f_{g_{D_1},g_{D_2}} P(g_{D_1}|g_1)P(g_{D_2}|g_2).\end{aligned}$$

For gene j, the conditional probability $P(g_{D_j}|g_j)$ depends on the LD between the marker j and disease locus j. Let $\triangle_j = p_{D_jA_j} - p_{D_j}p_{A_j} = -p_{d_jA_j} + p_{d_j}p_{A_j}$ be the LD coefficient between them. Assume both disease loci follow an additive model. Let $K_0 = f_{0,0}p_{d_1}^2 + 2f_{1,0}p_{d_1}p_{D_1} + f_{2,0}p_{D_1}^2$, $K_1 = f_{0,1}p_{d_1}^2 + 2f_{1,1}p_{d_1}p_{D_1} + f_{2,1}p_{D_1}^2$ and $K_2 = f_{0,2}p_{d_1}^2 + 2f_{1,2}p_{d_1}p_{D_1} + f_{2,2}p_{D_1}^2$. It can be shown that

$$\begin{aligned}\varphi_{0,0} = &\frac{p_{a_2d_2}^2}{p_{a_2}^2}\{K_0 - \frac{1}{p_{a_1}}[2\triangle_1(f_{1,0} - f_{0,0})]\} + \frac{2p_{a_2d_2}p_{a_2D_2}}{p_{a_2}^2}\{K_1 - \frac{1}{p_{a_1}}[2\triangle_1(f_{1,1} - f_{0,1})]\} \\ &+ \frac{p_{a_2D_2}^2}{p_{a_2}^2}\{K_2 - \frac{1}{p_{a_1}}[2\triangle_1(f_{1,2} - f_{0,2})]\}.\end{aligned}$$

Similarly,

$$\varphi_{0,2} = \frac{p^2_{A_2d_2}}{p^2_{A_2}}\{K_0 - \frac{1}{p_{a_1}}[2\triangle_1(f_{1,0} - f_{0,0})]\} + \frac{2p_{d_2}p_{A_2D_2}}{p^2_{A_2}}\{K_1 - \frac{1}{p_{a_1}}[2\triangle_1(f_{1,1} - f_{0,1})]\}$$
$$+ \frac{p^2_{A_2D_2}}{p^2_{A_2}}\{K_2 - \frac{1}{p_{a_1}}[2\triangle_1(f_{1,2} - f_{0,2})]\}$$

$$\varphi_{2,0} = \frac{p^2_{a_2d_2}}{p^2_{a_2}}\{K_0 + \frac{1}{p_{A_1}}[2\triangle_1(f_{1,0} - f_{0,0})]\} + \frac{2p_{a_2d_2}p_{a_2D_2}}{p^2_{a_2}}\{K_1 + \frac{1}{p_{A_1}}[2\triangle_1(f_{1,1} - f_{0,1})]\}$$
$$+ \frac{p^2_{a_2D_2}}{p^2_{a_2}}\{K_2 + \frac{1}{p_{A_1}}[2\triangle_1(f_{1,2} - f_{0,2})]\}$$

$$\varphi_{2,2} = \frac{p^2_{A_2d_2}}{p^2_{A_2}}\{K_0 + \frac{1}{p_{A_1}}[2\triangle_1(f_{1,0} - f_{0,0})]\} + \frac{2p_{A_2d_2}p_{A_2D_2}}{p^2_{A_2}}\{K_1 + \frac{1}{p_{A_1}}[2\triangle_1(f_{1,1} - f_{0,1})]\}$$
$$+ \frac{p^2_{A_2D_2}}{p^2_{A_2}}\{K_2 + \frac{1}{p_{A_1}}[2\triangle_1(f_{1,2} - f_{0,2})]\}.$$

Under additive model, we have $2(f_{1,i} - f_{0,i}) = f_{2,i} - f_{0,i}$ for $j = 0, 1$, and 2. Since $\triangle_2 = p_{D_2A_2} - p_{D_2}p_{A_2} = p_{d_2a_2} - p_{d_2}p_{a_2} = -p_{d_2A_2} + p_{d_2}p_{A_2} = -p_{D_2a_2} + p_{D_2}p_{a_2}$, thus the relationship between the penetrance of genotyped markers can be simplified as

$$\begin{aligned}
(\varphi_{2,2} - \varphi_{2,0}) - (\varphi_{0,2} - \varphi_{0,0}) &= \frac{2\triangle_1(f_{1,0} - f_{0,0})}{p_{A_1}p_{a_1}}(\frac{p^2_{A_2d_2}}{p^2_{A_2}} - \frac{p^2_{a_2d_2}}{p^2_{a_2}}) \\
&+ \frac{2\triangle_1(f_{1,1} - f_{0,1})}{p_{A_1}p_{a_1}}(\frac{2p_{A_2d_2}p_{A_2D_2}}{p^2_{A_2}} - \frac{2p_{a_2d_2}p_{a_2D_2}}{p^2_{a_2}}) \\
&+ \frac{2\triangle_1(f_{1,2} - f_{0,2})}{p_{A_1}p_{a_1}}(\frac{p^2_{A_2D_2}}{p^2_{A_2}} - \frac{p^2_{a_2D_2}}{p^2_{a_2}}) \\
&= \frac{\triangle_1(f_{2,0} - f_{0,0})}{p_{A_1}p_{a_1}}\frac{\triangle_2}{p_{A_2}p_{a_2}}[\frac{\triangle_2}{p_{A_2}p_{a_2}}(p_{a_2} - p_{A_2}) - 2p_{d_2}] \\
&+ \frac{\triangle_1(f_{2,1} - f_{0,1})}{p_{A_1}p_{a_1}}\frac{\triangle_2}{p_{A_2}p_{a_2}}[\frac{\triangle_2}{p_{A_2}p_{a_2}}(p_{A_2} - p_{a_2}) + p_{d_2} - p_{D_2}] \\
&+ \frac{\triangle_1(f_{2,2} - f_{0,2})}{p_{A_1}p_{a_1}}\frac{\triangle_2}{p_{A_2}p_{a_2}}[\frac{\triangle_2}{p_{A_2}p_{a_2}}(p_{a_2} - p_{A_2}) + 2p_{D_2}] \\
&= \frac{\triangle_1}{p_{A_1}p_{a_1}}\frac{\triangle_2}{p_{A_2}p_{a_2}}[(f_{2,2} - f_{2,0}) - (f_{0,2} - f_{0,0})].
\end{aligned}$$

Similarly to what we have shown for quantitative traits, the interaction effect between the markers also relates to the interaction effect between the disease loci, with the same multiplying factor $[\frac{\triangle_1}{p_{A_1}p_{a_1}}][\frac{\triangle_2}{p_{A_2}p_{a_2}}]$. This suggests that the weighting scheme as defined previously for quantitative traits can also be used for binary traits to appropriately aggregate information from all markers within each gene.

Test of Interaction for Binary Trait

Our goal is to test the null hypothesis $(f_{2,2}-f_{2,0})-(f_{0,2}-f_{0,0})=0$, i.e.,$f_{2,2}-f_{0,0}=(f_{2,0}-f_{0,0})-(f_{0,2}-f_{0,0})$. However, this null hypothesis cannot be directly tested in case-control studies because the risks (i.e., penetrances) are not directly estimable. Hence, approximations are required to test for interaction. Let r denote the genotype relative risk, then the null hypothesis of no interaction can be rewritten as $r_{2,2}-1=(r_{2,0}-1)-(r_{0,2}-1)$. If all the relative risks are near one, then $r_{2,2}-1\approx 0$. Since $x\approx \log(1+x)$ for $x\approx 0$, thus the null hypothesis of no interaction can be approximated by $\log(r_{2,2})=\log(r_{2,0})+\log(r_{0,2})$. If, in addition, all the penetrances are near zero (e.g. rare diseases), then $1-f_{i,j}\approx 1$, and the genotype relative risks can be approximated by odds ratios, $r_{i,j}=f_{i,j}/f_{i,j}\approx[f_{i,j}/(1-f_{0,0})]/[f_{0,0}/(1-f_{0,0})]=OR_{i,j}$. In this situation, the null hypothesis of no interaction can be further approximated by $\log(OR_{2,2})=\log(OR_{2,0})+\log(OR_{0,2})$ (i.e. additivity on the log odds scale), which suggests that interaction can be tested by logistic regression. The above derivations indicate that the weighted genotype scores we derived earlier may not be "optimal" when used in logistic regression since it is based on a two-step approximation. However, when the genotype relative risks are near one and the penetrances are near zero, these scores may still perform well.

Chapter 3

A Gaussian Copula Approach for the Analysis of Secondary Phenotypes in Case-Control Genetic Association Studies

3.1 Introduction

Genome-wide association studies (GWAS) offer a powerful tool to identify genes that confer moderate disease risks. In these studies, the main outcome of interest is often disease status. However, in many of these studies, a set of correlated secondary phenotypes that may share the same genetic factors with disease status are also collected. Examination of these secondary phenotypes may provide important clues about the disease etiology and supplement the main studies. Various secondary phenotypes have been suggested as useful for gene mapping of complex diseases. For example, low-density lipoprotein cholesterol (LDL-C) and high-density lipoprotein cholesterol (HDL-C) levels for coronary artery disease (Grundy et al. (2004)), and angiotensin-converting enzyme activity for hypertension (Kammerer et al. (2004)) are clear examples of useful secondary phenotypes. In some situations, the analysis of secondary phenotypes may become the primary focus of subsequent studies. Recently, there have been several GWAS on secondary phenotypes, such as BMI and lipid levels (Kathiresan et al. (2008); Loos et al. (2008); Teslovich et al. (2010); Willer et al. (2008)), where most of the data came from case-control studies of complex diseases, such as diabetes, hypertension, and heart disease.

Commonly used approaches for the analysis of secondary phenotypes rely on standard regression that assesses the effect of SNPs using controls only, cases only, combined data of cases and controls, or joint analysis of cases and controls adjusting for disease status. However, none of these methods are statistically correct and may lead to false positive associations (Lin and Zeng (2009)). The reason is that in a study

where the samples are ascertained according to disease status, cases and controls no longer constitute a random sample of the general population. As a result, the population association between a marker and a secondary phenotype can be distorted in the case-control data.

Several methods have been developed to correct for the sampling bias in the analysis of secondary phenotypes. Monsees et al. (2009) proposed an inverse-probability-of-sampling weighted regression approach to incorporate selection probability in likelihood calculation. Lin and Zeng (2009) proposed a retrospective likelihood approach that conditions on disease status. Although these methods are useful, they make restrictive assumptions on the distribution of the secondary phenotypes. It is thus desirable to develop methods that allow the modeling of secondary phenotypes that do not fit the above-mentioned methods distributional assumptions. Since the primary and secondary phenotypes are often correlated, a critical first step in devising such methods is on modeling the joint distribution of the primary and secondary phenotypes.

In statistics, copulas are used as a general way of formulating multivariate distributions (Nelsen (1999)). The rationale is that a simple transformation can be made for each marginal variable in a way such that each transformed marginal variable has a uniform distribution. Once the transformation is done, the dependence structure can then be expressed as a multivariate distribution on the obtained uniforms, and a copula is precisely a multivariate distribution on marginally uniform random variables. A commonly used copula is the Gaussian copula, which is constructed from multivariate normal distribution via Sklar's theorem. Gaussian copulas share many

similarities with the multivariate normal distribution, and are useful for constructing joint distributions of continuous, discrete, or mixed types of outcomes (Song et al. (2009); De Leon and Wu (2011)).

Gaussian copulas have been previously employed in linkage studies for mapping quantitative trait loci (Li et al. (2006)). Here we extend the analysis to association mapping in which we use Gaussian copulas to model the joint distribution of the disease status variable and secondary phenotypes. An advantage of our method is that it can handle a variety of secondary phenotypes as long as the distribution comes from an exponential family, making it much more flexible than existing methods. We show through extensive simulations that our method yields correct type I error rates even when the model is mis-specified. We also demonstrate the effectiveness of our method in the analysis of a genome-wide association study on high HDL-C in which LDL-C and three apolipoprotein levels including ApoA1, ApoB and ApoC3 are treated as secondary phenotypes.

3.2 Methods

3.2.1 Gaussian Copula and Joint Analysis of Correlated Mixed Outcomes

The central idea of our method is to jointly model the distribution of disease status and secondary phenotypes using Gaussian copulas. This is based on the likelihood framework that we previously developed for joint regression analysis of

correlated mixed outcomes (Li et al. (2006); Song et al. (2009)). Below we briefly review results from our previous work. Consider m dependent random variables $y_1, \ldots, y_m$. Let $F_j(y_j)$ denote the cumulative distribution function (CDF) of y_j, which is assumed to be from an exponential family. The density function of y_j is $f(y_j; \eta_j, \varphi_j) = \exp\{(y_j\eta_j - b(\eta_j))/a(\varphi_j) + c(y_j, \varphi_j)\}$, where a, b and c are known functions, φ_j is the dispersion parameter, and η_j is the canonical parameter. The mean and variance of y_j are given by $E(y_j) = \mu_j = b'(\eta_j)$ and $\text{var}(y_j) = b''(\eta_j)a(\varphi_j)$, respectively. Since the CDF is uniformly distributed on the $[0, 1]$ interval, the joint distribution function of $y_1, \ldots, y_m$ can be modeled through a Gaussian copula, defined as $\Phi_m(\Phi^{-1}(F_1(y_1)), \ldots, \Phi^{-1}(F_m(y_m))|\Gamma)$, where Φ and Φ_m are the standard univariate and multivariate normal distribution functions, and Γ is an $m \times m$ correlation matrix. With Gaussian copulas, the handling of a multivariate distribution can be separated into a marginal model for the inverse normal score $\Phi^{-1}(F_j(y_j))$ and a model for the joint distribution of the inverse normal scores.

In our previous work (Li et al. (2006)), we showed that when the first m_1 of the outcome variables are discrete and the remaining outcome variables are continuous, the joint density of $y_1, \ldots, y_m$ is

$$P(y_1, \ldots, y_m) = \prod_{j=m_1+1}^{m} f_j(y_j; \eta_j, \varphi_j) \sum_{j_1}^{2} \cdots \sum_{j_{m_1}=1}^{2} (-1)^{j_1+\ldots+j_m} (2\pi)^{-\frac{m_1}{2}} |\Gamma|^{-\frac{1}{2}} \int_{-\infty}^{\Phi^{-1}(\mu_{1,j_1}))} \cdots \int_{-\infty}^{\Phi^{-1}(\mu_{m_1,j_{m_1}}))} \exp[-\frac{1}{2}(y, q)\Gamma^{-1}(y^T, q^T) + \frac{1}{2}qq^T]dy, \tag{3.2.1}$$

where $y = (y_1, \ldots, y_m)$, $q = (\Phi^{-1}(F_{m_1+1}(y_{m_1+1})), \ldots, \Phi^{-1}(F_m(y_m)))$, $\mu_{j,1} = F(y_j-; \eta_j, \varphi_j)$ and $\mu_{j,2} = F_j(y_j; \eta_j, \varphi_j)$. Here $F(y_j-; \eta_j, \varphi_j)$ is the left-hand limit of F_j at y_j, which is equal to $F(y_j - 1; \eta_j, \varphi_j)$ when y_j takes integer values as for the Poisson and Binomial

distributions.

The above likelihood allows us to model the joint distribution between disease status and those continuous secondary phenotypes. Let y_1(1=affected; 0=unaffected) denote the disease status, and $y_2, \ldots, y_m$ denote those secondary phenotypes. Based on the likelihood given in equation 3.2.1, we can show that the joint density function of $y_1, \ldots, y_m$ is

$$\begin{aligned} P(y_1, \ldots, y_m) = & \prod_{j=2}^{m} f(y_j; \eta_j, \varphi_j) \\ & [1 - \int_{-\infty}^{\Phi^{-1}(1-\mu_1)} \frac{1}{\sqrt{2\pi|\Gamma|}} \exp\{-\frac{1}{2}(z,q)\Gamma^{-1}(z,q)^T + \frac{1}{2}qq^T\}dz]^{I(y_1=0)} \\ & [\int_{-\infty}^{\Phi^{-1}(1-\mu_1)} \frac{1}{\sqrt{2\pi|\Gamma|}} \exp\{-\frac{1}{2}(z,q)\Gamma^{-1}(z,q)^T + \frac{1}{2}qq^T\}dz]^{I(y_1=1)}, \end{aligned} \tag{3.2.2}$$

where $I\{\cdot\}$ is an indicator function. In particular, if there is only a single secondary phenotype and it is normally distributed, then the joint density function of y_1 and y_2 can be simplified as

$$\begin{aligned} P(y_1, y_2) = \phi(y_2; \mu_2, \sigma_2)[1 - \Phi(\frac{\Phi^{-1}(\mu_1) + \gamma(y_2 - \mu_2)/\sigma_2}{\sqrt{1-\gamma^2}})]^{I(y_1=0)} \\ \Phi(\frac{\Phi^{-1}(\mu_1) + \gamma(y_2 - \mu_2)/\sigma_2}{\sqrt{1-\gamma^2}})^{I(y_1=1)}, \end{aligned} \tag{3.2.3}$$

where ϕ is the density of a normal random variable, σ_2 is the standard deviation of y_2, and γ is the correlation parameter that characterizes the degree of correlation between y_1 and y_2 with $|\gamma| \neq 1$. Obviously, when $\gamma = 0$, the cases and controls share the same density for the secondary phenotype; when $\gamma \neq 0$, the distributions are different with both mean shift and rescaling, and this makes the copula approach flexible in capturing the shape of real data. From equation 3.2.3, we can easily derive

the conditional probability of y_1 given y_2

$$P(y_1 = 1|y_2) = \Phi(\frac{\Phi^{-1}(\mu_1) + \gamma(y_2 - \mu_2)/\sigma_2}{\sqrt{1-\gamma^2}}), \tag{3.2.4}$$

which suggests that when $\gamma > 0$, the probability of being affected increases with y_2.

3.2.2 Retrospective Likelihood for Secondary Phenotype

In a case-control study, since the data are ascertained based on disease status, y_1, to reflect the sampling scheme and to make valid inference on the secondary phenotypes, a retrospective likelihood would be appropriate (Kraft and Thomas (2000)). Let $g(= 0, 1, 2)$ denote the genotype (counting the number of minor alleles) at the test SNP for an individual. Then the retrospective likelihood of the individual is

$$P(y_2, \ldots, y_m, g|y_1) = \frac{P(y_1, \ldots, y_m|g)P(g)}{P(y_1)} = \frac{P(y_1, \ldots, y_m|g)P(g)}{\sum_{g=0}^{2} P(y_1|g)P(g)}, \tag{3.2.5}$$

where $P(g)$ is the genotype frequency at the test SNP. Assume Hardy-Weinberg equilibrium, then the genotype frequencies can be calculated as $P(g = 0) = (1 - p)^2, P(g = 1) = 2p(1 - p)$, and $P(g = 2) = p^2$, where p is the minor allele frequency (MAF) of the test SNP. We can relate g with the marginal mean model for each of the phenotypes through the use of a link function $h(\mu_j) = \beta_{0,j} + \beta_{1,j} \times g$. The specification of the link function depends on the distribution of the phenotype. For the disease status variable, we can model its marginal mean by the logit link function $\log[\mu_1/(1 - \mu_1)] = \beta_{1,0} + \beta_{1,1} \times g$, where $\mu_1 = P(y_1 = 1|g)$. For the j-th secondary phenotype, $h(\mu_j) = \beta_{j,0} + \beta_{j,1} \times g$, the marginal mean model will be determined by its distribution; for example, for a normally distributed random variable, $h(\mu_j) = \mu_j$; for

a Poisson distributed random variable, $h(\mu_j) = \log(\mu_j)$; for a binary random variable, $h(\mu_j) = \log[\mu_j/(1-\mu_j)]$; for a gamma distributed random variable, one can use either a reciprocal link function $h(\mu_j) = 1/\mu_j$ or a log link function $h(\mu_j) = \log(\mu_j)$. For a case-control study with a total of n subjects, the overall likelihood is simply the product of the likelihoods across all individuals.

After the likelihood is specified, the next step is to establish simultaneous maximum likelihood inference for all parameters $\theta = (\{\beta_{j,0}, \beta_{j,1}, \gamma_j\}_{j=1}^m, p)$, where γ_j is the parameter that characterizes the correlation between y_1 and y_j (note that $\gamma_1 = 1$ by definition). When the data only contain unrelated cases and controls, these parameters are not all identifiable. In our analysis, we fixed the disease prevalence and updated the intercept parameter for the primary phenotype by $\beta_{1,0} = \beta_{1,0}^* + \log[K/(1-K)]$, where $\beta_{1,0}^*$ is the intercept estimate obtained from a logistic regression on the disease status variable with the SNP genotype included as a covariate. By using this strategy, we avoided estimating the intercept parameter $\beta_{1,0}$.

Let $l(\theta)$ denote the log-likelihood function. To find the maximum likelihood estimate (MLE) $\hat{\theta} = \text{argmax}_\theta l(\theta)$ numerically, we implement a Gauss-Newton type algorithm (Ruppert (2005)), which only requires the first derivatives of the log-likelihood function. We search for the MLE by taking step-halving to guarantee that the likelihood increases progressively over iterations. Specifically, the $(k+1)$-th iteration updates the parameter θ by

$$\theta^{k+1} = \theta^k + \varepsilon\{B_n(\theta^k)\}^{-1}\frac{\partial l(\theta^k)}{\partial \theta}, \tag{3.2.6}$$

where $B_n = \sum_{i=1}^n (\frac{\partial l(\theta)}{\partial \theta})(\frac{\partial l(\theta)}{\partial \theta})^T$ and ε is the step-halving term that starts at 1 and

halves until $l(\theta^{k+1}) > l(\theta^k)$ at iteration k. The algorithm stops when the increase in the likelihood is no longer possible or the difference between the two consecutive update is smaller than a prespecified precision level. To determine the initial values in the optimization, we conduct univariate analysis for the primary and secondary phenotypes and use parameter estimates obtained from these analyses as initial values. For the correlation parameter, we estimate the Pearson correlation between the primary and secondary phenotypes and use that as the initial value. The variances of the parameters are approximated by numerical Fisher information. This Gauss-Newton type algorithm works well for Gaussian copulas and it generally converges in less than 10 iterations.

With the previously developed likelihood framework, we can evaluate whether the test SNP is associated with the j-th secondary phenotype using a Wald test $W_S = \hat{\beta_{j,1}}^2/\text{var}(\hat{\beta}_{j,1})$. Under the null hypothesis of no association, W_S is asymptotically distributed as a chi-squared distribution with one degree of freedom.

3.3 Simulation Studies

3.3.1 Simulation Setup

We conducted extensive simulations to examine the performance of the copula-based approach and to compare it with several existing methods. We considered a SNP with MAF of 0.3 and a single secondary phenotype that is normally distributed. To generate correlated disease status and secondary phenotype, we first simulated y_2

from a normal distribution $N(\mu_2, 1)$, where $\mu_2 = \beta_{2,0} + \beta_{2,1} \times g$. We then simulated y_1 from the conditional distribution of $y_1|y_2$ based on equation 3.2.4. An individual's disease status was determined by comparing a randomly generated number with the pre-specified penetrance function $P(y_1 = 1|g) = \mu_1 = \exp(\beta_{1,0} + \beta_{1,1} \times g)/[1 + \exp(\beta_{1,0} + \beta_{1,1} \times g)]$. The value of $\beta_{1,0}$ was determined such that the overall disease prevalence is 5%, and the value of $\beta_{1,1}$ was determined by the odds ratio OR $= \exp(\beta_{1,1})$. For the secondary phenotype, we set $\beta_{2,0} = 0$, and the value of $\beta_{2,1}$ was determined by heritability $h^2 = 2\beta_{2,1}^2 p(1-p)/[1+2\beta_{2,1}^2 p(1-p)]$, i.e., the proportion of phenotypic variation explained by the test SNP. After a large pool of individuals was simulated, we then sampled 1,000 cases and 1,000 controls from the pool. We analyzed the secondary phenotype in each simulated dataset using the following methods: 1) the copula-based approach, 2) linear regression with cases only, 3) linear regression with controls only, 4) linear regression with cases and controls combined without adjustment of disease status, 5) linear regression with cases and controls combined with adjustment of disease status, and 6) Lin and Zeng's method (Lin and Zeng (2009)). Type I error rates were estimated based on 100,000 simulations, and power was estimated based on 10,000 simulations.

3.3.2 Analysis of Secondary Phenotype

Empirical Type I Error Rates

The upper part of Table 3.1 shows the empirical type I error rates of different methods for the analysis of the secondary phenotype when the test SNP is not asso-

ciated with the disease. As expected, the type I error rates of all methods are close to the 1% nominal level. We next evaluated the performance of different methods when the test SNP is associated with the disease at OR = 1.2. As shown in the lower part of Table 3.1, only the copula-based approach and the control-only analysis have controlled type I error rates. In contrast, the type I error rates of the other methods are inflated when the disease status and secondary phenotype are correlated, and the degree of inflation increases with the degree of correlation. Of note, Lin and Zeng's method is a parametric based method. It is not surprising that its type I error rate is slightly inflated as the simulated data do not fit their distributional assumption.

Empirical Power

As shown in the upper part of Figure 3.1, when the test SNP is not associated with the disease, the copula-based approach, Lin and Zeng's approach, and the case-control combined analysis adjusting disease status have comparable power and outperform the other methods. The analysis using both cases and controls without adjustment of disease status has comparable power with the above-mentioned three methods at low correlation but becomes less powerful when the degree of correlation increases. Not surprisingly, case-only and control-only analyses have less power than analysis methods that include the entire sample.

When the test SNP is associated with the disease, we excluded the case-only, and case-control combined unadjusted analyses from power comparison due to their serious inflation of type I error rates. Not surprisingly, methods that use the entire sample, such as the copula-based approach, Lin and Zeng's approach, and the

combined analysis adjusting disease status, are more powerful than the control-only analysis (lower part of Figure 3.1). The power of Lin and Zeng's analysis and the combined-adjusted analysis are similar and both decrease as the correlation increases in the positive direction. We note that the type I error rates of these two tests are slightly inflated when $|\gamma| > 0.3$, and thus their power should be interpreted with caution.

Evaluation of Robustness of the Copula-based Approach

Results in the previous sections were obtained based on data simulated from the copula model. To evaluate the robustness of our method, we simulated data from a model employed by Lin and Zeng (2009). In this model, the secondary phenotype was assumed to be normally distributed with mean $\mu_2 = \alpha_{2,0} + \alpha_{2,1} \times g$ and standard deviation 1, and the disease status was simulated according to a logit model $\log[\mu_1/(1-\mu_1)] = \alpha_{1,0} + \alpha_{1,1} \times g + \alpha_{1,2} \times y_2$. In this model, the correlation between the disease status and secondary phenotype is introduced through $\alpha_{1,2}$ in the logit function.

Table 3.2 shows the estimated type I error rates of the copula-based approach and Lin and Zeng's approach. As indicated by its controlled type I error rates, the copula-based approach is robust to misspecification of the models. We also compared the power of the copula-based approach and Lin and Zeng's approach. As shown in Figure 3.2, these two methods have comparable power when the test SNP is not associated with the disease; when the test SNP is associated with the disease, the copula-based approach is more powerful than Lin and Zeng when the disease status

variable and the secondary phenotype are positively correlated, but Lin and Zeng's approach is more powerful when the correlation is negative.

3.4 Application to a Genetic Association Study on High HDL

The University of Pennsylvania High HDL Cholesterol Study is a cross-sectional study of genetic factors contributing to elevated HDL-C levels and is composed of subjects with extreme levels of HDL-C. In this study, subjects of European ancestry with HDL $>$ 90th percentile for age and gender were considered as cases and subjects with HDL $<$ 30th percentile for age and gender were considered as controls. 605 cases and 724 controls were genotyped using Affymetrix 6.0 SNP array. The primary phenotype is the HDL-C case-control status. In addition, all study subjects have measurements on LDL-C and chemical measures of apolipoprotein A1 (ApoA1), apolipoprotein B (ApoB) and apolipoprotein C3 (ApoC3) protein concentration.

Our GWAS analysis revealed strong association between HDL case-control status and a number of SNPs in *LIPG*, a gene located on chromosome 18. It is well-known that LIPG plays an important role in the regulation of HDL-C level. In a recent meta-analysis of 46 GWAS of lipids in >100,000 individuals (Teslovich et al. (2010)), *LIPG* was found to be strongly associated with HDL-C (rs7241918: P-value $= 3 \times 10^{-49}$). However, this study did not find evidence of association between *LIPG* and LDL-C. In another GWAS of plasma lipoprotein size, concentration, and cholesterol content

in >17,000 individuals (Chaseman et al. (2009)), *LIPG* was found to be significantly associated with ApoA1 (rs4939883: P-value $= 5.60 \times 10^{-14}$) but not with ApoB. Since ApoC3 was not included in this study, it is unclear what the role is for *LIPG* in the regulation of ApoC3 levels.

In our analysis, we treated LDL-C, ApoA1, ApoB and ApoC3 levels as secondary phenotypes. We analyzed each of them separately with the HDL case-control status. The purpose is to investigate whether SNPs in *LIPG* are associated with these secondary phenotypes. For each of the 85 SNPs in *LIPG*, we tested association with the secondary phenotype using 1) the copula-based approach, 2) Lin and Zeng's approach, 3) case-control combined analysis adjusting case-control status, and 4) case-control combined unadjusted analysis. For the copula-based approach, we standardized each secondary phenotype by subtracting its mean estimated from controls and then divided by its standard deviation estimated using all samples. Figure 3.3 displays the P-values for the analysis of secondary phenotypes together with the P-values for the analysis of the HDL-C case-control status.

For ApoB, whose level is weakly correlated with HDL-C case-control status ($\gamma = -0.06$), the four methods have comparable P-values for all tested SNPs. For LDL-C, which shows moderate correlation with HDL-C ($\gamma = 0.16$), the P-values from the case-control combined unadjusted analysis are more significant than the other methods for most SNPs in *LIPG*. Since *LIPG* is not associated with LDL-C (Chaseman et al. (2009); Teslovich et al. (2010)), the significant results from the case-control combined unadjusted analysis are likely due to false positive associations.

For ApoC3, which is more strongly correlated with HDL-C ($\gamma = 0.20$), the results

from the case-control combined unadjusted analysis become even more significant than approaches that aim to correct for ascertainment bias. For example, for the three SNPs (rs10438978, rs2156552, and rs4939883) that are significantly associated with HDL-C case-control status (P-value < 0.001), they all show significant association with ApoC3 (P-value < 0.005) using the case-control combined unadjusted analysis; however, the corresponding P-values by the other three approaches are all greater than 0.01. In particular, the copula-based approach yielded the least significant results with the P-values greater than 0.05 for two of the three SNPs (rs2156552: P-value $= 0.0521$; rs4939883: P-value $= 0.0502$). Although the role of *LIPG* on ApoC3 level is still unclear, based on our simulations (Table 3.1), we found that when the correlation between the primary and secondary phenotypes is 0.2 and the OR for the primary phenotype is 1.2, the type I error rate for the case-control combined unadjusted analysis can be as high as 2% (at the 1% significance level). Since the ORs for the above-mentioned three *LIPG* SNPs are >1.4, the degree of type I error inflation for the case-control combined unadjusted analysis might be even higher than 2%. This suggests that the small P-values from the case-control combined unadjusted analysis for ApoC3 are possibly due to false positive signals.

For ApoA1, which is strongly correlated with HDL-C ($\gamma = 0.80$), the copula-based method revealed evidence of association at rs4939883 (P-value $= 0.0073$), the same SNP reported by (Chaseman et al. (2009)). The corresponding P-value from the case-control combined unadjusted analysis is slightly more significant (P-value $= 0.0059$). This reduced P-value is possibly due to the ignorance of ascertainment bias in the analysis. On the other hand, the case-control combined adjusted analysis yielded a

P-value of 0.099 at rs4939883. Of note, Lin and Zeng's method failed to converge for all SNPs due to high correlation between ApoA1 and HDL-C. Our results for the analysis of ApoA1 suggest that the copula-based method not only prevents false positive associations, but also does not remove true associations.

3.5 Discussion

We have developed a Gaussian copula-based approach for the analysis of secondary phenotypes collected from case-control genetic association studies. Through extensive simulations, we showed that our method has controlled type I error rates even when the model was mis-specified. We also demonstrated the effectiveness of our method in the analysis of a real dataset and showed that inappropriate analysis of secondary phenotypes may lead to false positive association signals and produce misleading results.

Compared with existing methods, our method has several advantages. First, it is flexible and can handle a variety of secondary phenotypes as long as the trait distribution comes from an exponential family. Second, it is able to incorporate multiple correlated secondary phenotypes, whereas the existing methods such as Monsees et al. (2009) and Lin and Zeng (2009) can only handle a single secondary phenotype. Although we only showed results for the analysis of a single secondary phenotype, we also conducted simulations with two secondary phenotypes, and found that the copula-based approach still maintains controlled type I error rates, and moreover, it offers even more power gain over univariate analysis (data not shown). Third, our

method can analyze secondary phenotypes that have different marginal distributions. This feature is particularly appealing for practical analysis. For example, we are currently working on a GWAS on coronary artery disease (CAD). In this study, the primary outcome of interest is CAD status. However, in addition to CAD status, we also have information on many secondary phenotypes that are closely related to CAD, including lipid levels such as HDL-C, LDL-C, triglyceride, and coronary artery calcium score. These traits are all correlated with CAD, but have different marginal distributions. Flexible methods such as the copula-based approach are critical for proper analysis of such type of data.

In this paper, we focused on the analysis of the secondary phenotypes. However, our likelihood framework is general and also allows us to test for association with the primary phenotype, i.e., the disease status. Our preliminary results indicate that by incorporating secondary phenotypes that are moderately correlated with the disease status (e.g. correlation $\gamma > 0.3$), there can be noticeable power improvement over the standard logistic regression that only considers the disease status.

Table 3.1: Comparison of type I error rates (%) for the analysis of secondary phenotype. Significance was assessed at 1% significance level based on 100,000 simulations. γ is the correlation parameter between the primary and secondary phenotypes.

γ	Copula	Lin-Zeng	Case-only	Control-only	Case-control combined unadjusted	Case-control combined adjusted
$h^2 = 0, OR = 1.0$						
-0.4	1.00	1.01	1.03	1.02	0.90	0.98
-0.3	1.04	1.01	0.99	1.00	1.01	1.00
-0.2	0.93	0.99	1.03	1.06	0.99	0.97
-0.1	0.97	1.01	1.02	0.95	1.00	1.00
0	1.00	1.04	1.04	1.01	1.02	1.03
0.1	1.00	1.00	0.96	1.02	0.98	0.98
0.2	1.00	1.09	1.06	0.97	1.05	1.08
0.3	0.96	1.03	0.95	1.04	1.03	1.00
0.4	0.98	1.08	0.98	1.02	1.02	1.05
$h^2 = 0, OR = 1.2$						
-0.4	1.05	1.38	2.31	0.97	5.05	1.68
-0.3	1.06	1.20	1.57	0.98	3.44	1.31
-0.2	1.07	1.14	1.23	0.93	2.00	1.16
-0.1	0.95	1.02	1.09	0.97	1.26	1.01
0	1.00	0.97	1.00	1.01	0.97	0.96
0.1	0.99	1.09	1.06	0.99	1.24	1.08
0.2	0.98	1.09	1.21	1.06	1.99	1.12
0.3	1.02	1.23	1.65	0.97	3.13	1.35
0.4	1.09	1.29	2.35	0.98	5.69	1.73

Table 3.2: Comparison of type I error rates (%) for the analysis of secondary phenotype when data were simulated from Lin and Zeng's model. Significance was assessed at 1% significance level based on 100,000 simulations. $\alpha_{1,2}$ is a parameter that determines the correlation between the primary and secondary phenotypes.

	$OR = 1.0$		$OR = 1.2$	
$\alpha_{1,2}$	Copula	Lin-Zeng	Copula	Lin-Zeng
-0.4	0.97	1.05	1.04	1.04
-0.3	1.01	1.05	1.05	1.01
-0.2	1.01	1.00	1.03	1.02
-0.1	0.99	0.96	1.00	1.05
0	0.94	0.96	0.99	1.00
0.1	0.96	1.02	0.98	1.00
0.2	0.99	1.05	0.97	1.05
0.3	0.96	1.00	1.00	1.00
0.4	0.98	1.02	1.03	0.94

Figure 3.1: Comparison of power for the analysis of secondary phenotype. Significance was assessed at 1% significance level based on 10,000 simulations.

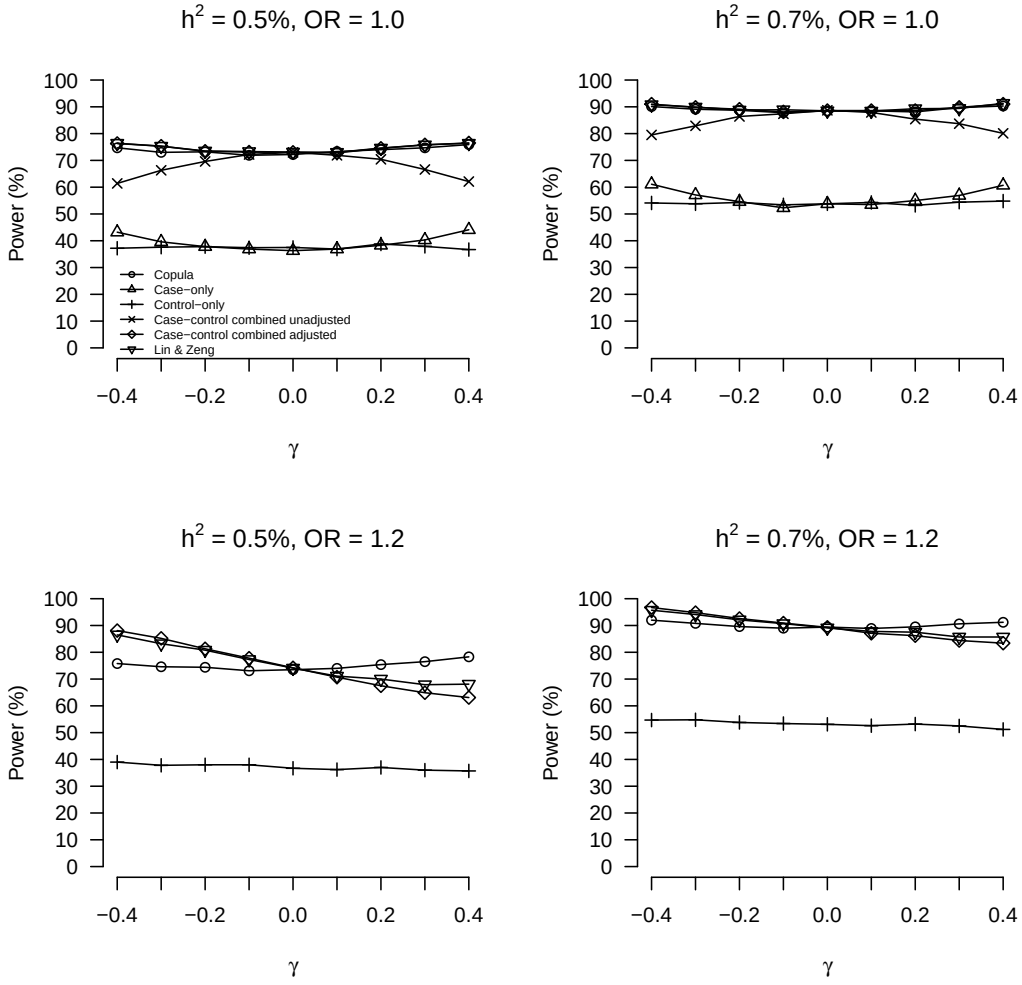


Figure 3.2: Comparison of power when data were simulated from Lin and Zeng's model. Significance was assessed at 1% significance level based on 10,000 simulations.

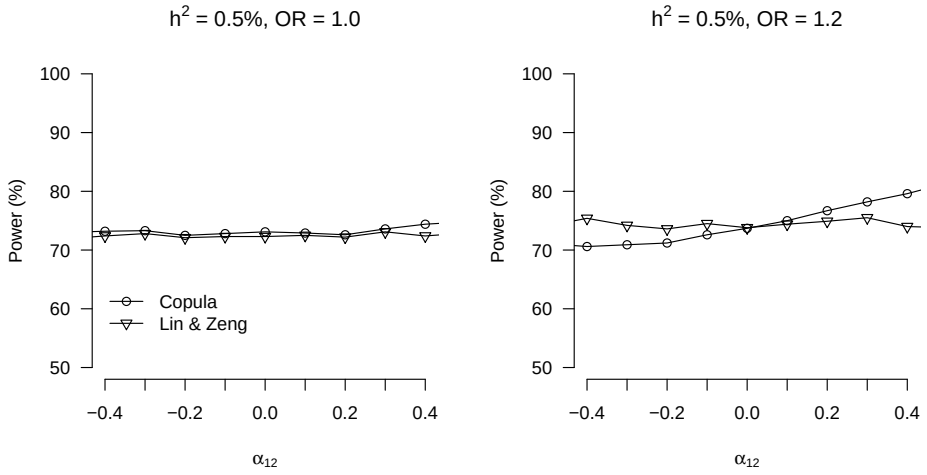
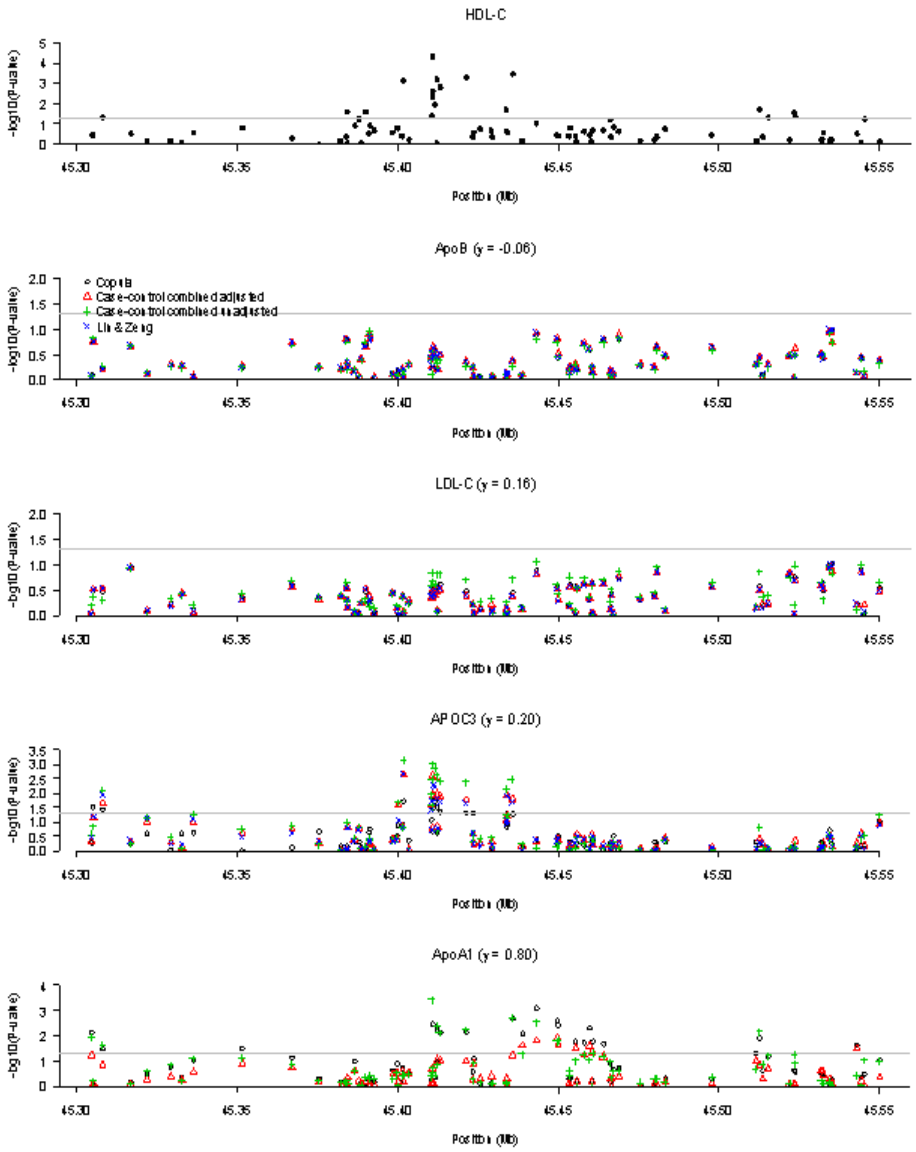


Figure 3.3: Association results from four different procedures for the analysis of LDL-C, ApoB, ApoC3, and ApoA1 levels in the high HDL study. HDL-C association was evaluated by logistic regression. The grey line corresponds to P-value = 0.05.



Chapter 4

Penalized Estimation of Gaussian Copula Model for Secondary Phenotype Analysis

4.1 Introduction

In genome-wide association studies, a set of correlated phenotypes other than the phenotype that defines the disease status are often collected. These phenotypes are often called the secondary phenotypes. Given the already genotyped data and available phenotypes, it is of interest to identify markers associated with the secondary phenotypes. Examination of the secondary phenotypes can yield valuable insights about disease etiology and supplement the main study. Given the large number of SNPs, the signals from single marker association may no longer be significant when adjusted for multiple testing. Instead of single marker tests, an alternative is to perform variable selection in order to identify the disease-associated genetic variants. However, direct application of the commonly used variable selection methods developed for linear regression models such as Lasso (Tibshirani (1996)) for analysis of the secondary phenotypes are statistically incorrect because cases and controls are collected at different rates from their subpopulations, thus case control sample does not constitute a random sample of the general population. To correct for the sampling bias we propose to develop variable selection method in a Gaussian copula modeling framework that effectively utilizes the information provided by the disease status. The Gaussian copula model can effectively model the joint distribution of two or more random variable where only the marginal distributions of the variables need to be specified. Such models can effectively model the between-outcomes association and provide unbiased marginal parameter estimates when the association parameter is correctly modeled.

In this paper, we propose to develop a penalized likelihood approach based on the Gaussian copula model for selecting the relevant markers that are associated with the secondary phenotypes. Our procedure involves two steps: first we employ variable selection methods such as Lasso or sure independence screening (Fan and Song (2010)) to select the markers that are associated with the primary outcome. This can be done unbiased by fitting logistic or penalized logistic regression models. We then treat these markers as known and in the second step, we propose to develop a penalized likelihood approach to select the markers that are associated with the secondary outcomes. We develop an efficient gradient descent algorithm to perform the optimization. We choose the tuning parameters using the BIC or extended BIC (eBIC) that was developed for high dimensional regressions (Chen (2008)).

This paper is organized as follows. We first introduce the Gaussian copula model for joint modeling of primary and secondary outcomes and for analysis of multiple markers. We then present a penalized likelihood approach for variable selection. An efficient computational algorithm based on the coordinate gradient descent is then presented to solve the optimization problem. We perform simulation studies to evaluate our method and compare the results with the Lasso and adjusted Lasso that includes the case control status in the regression model. Finally, we apply the method to a genetic association study on HDL.

4.2 Statistical Methods

4.2.1 Gaussian Copula Model for Correlated Phenotypes

For simplicity, we consider the setup where we have a binary primary outcome such as the case-control status and a continuous secondary phenotype. Let Y_1 denote the disease status that takes value of 0 and 1, and let Y_2 denote the secondary continuous phenotype. Let $\mathbf{g} = (g_1, ..., g_p)$ be the set of p SNP markers that we consider. Suppose there are n *i.i.d.* samples and let $(y_{1i}, y_{2i})_{i=1,...,n}$ denote the observed phenotypes for the ith individual. We assume that following marginal models for the phenotype Y_1 and Y_2: given the genotypes of the p markers $\mathbf{g_i}$, we assume that $y_{1i} \sim Bernoulli(\mu_1)$ and $y_{2i} \sim N(\mu_2, \sigma^2)$, where

$$\begin{cases} \text{logit}(\mu_1) = \beta_0 + \sum_{i=1}^{p} g_i \beta_i \\ \\ \mu_2 = \alpha_0 + \sum_{i=1}^{p} g_i \alpha_i, \end{cases} \tag{4.2.1}$$

where $\alpha = (\alpha_0, \alpha_1, \cdots, \alpha_p)$ and $\beta = (\beta_0, \beta_1, \cdots, \beta_p)$ denote the parameter vectors in the marginal models, and σ^2 is the error variance. Our goal is to select the SNPs that are associated with the secondary phenotype from the p candidate SNPs. In other words, we aim to select the markers such that the corresponding regression coefficients α_i are non-zero.

Assuming that the joint distribution of (y_1, y_2) is determined by a Gaussian copula, i.e.,

$$F(y_1, y_2) = \Phi_2(\Phi^{-1}(F_1(y_1)), \Phi^{-1}(F_2(y_2)); \Gamma_2), \tag{4.2.2}$$

where Φ is the standard normal distribution, Φ_2 is the standard bivariate normal

distribution with correlation Γ_2 and $F_i(y_i)$ is the cumulative distribution function (CDF) of Y_i. In our previous work (He (in revision)), we show that when the correlation $\text{corr}[\Phi^{-1}(F_1(y_1)), \Phi^{-1}(F_2(y_2))]$ between the two traits is $\gamma(|\gamma| \neq 1)$, the density function in cases and controls can be simplified as

$$f(y_1, y_2) = \begin{cases} \phi(y_2; \mu_2)(1 - \Phi(\frac{\Phi^{-1}(\mu_1)+\gamma(y_2-\mu_2)}{\sqrt{1-\gamma^2}})), \text{if } y_1 = 0; \\ \phi(y_2; \mu_2)\Phi(\frac{\Phi^{-1}(\mu_1)+\gamma(y_2-\mu_2)}{\sqrt{1-\gamma^2}}), \text{if } y_1 = 1, \end{cases} \tag{4.2.3}$$

where ϕ is the density function of a normal random variable. The conditional probability of having disease given y_2 can be easily derived as

$$P(y_1 = 1|y_2, \mathbf{g}) = \Phi(\frac{\Phi^{-1}(\mu_1) + \gamma(y_2 - \mu_2)}{\sqrt{1-\gamma^2}}). \tag{4.2.4}$$

Since the data are ascertained based on disease status, the joint retrospective likelihood function for the case control samples can be written as

$$L = \prod_{i=1}^{n} P(y_{2i}, \mathbf{g_i}|y_{1i}) = \prod_{i=1}^{n} \frac{P(y_{1i}, y_{2i}|\mathbf{g_i})P(\mathbf{g_i})}{P(y_{1i})}, \tag{4.2.5}$$

where $P(y_{1i} = 1)$ is the disease prevalence, which is assumed to be known.

4.2.2 Penalized Likelihood Estimation

When the dimension of the genotype vector is high ($p > 100$), direct optimization of the log-likelihood function (4.2.5) becomes infeasible. In addition, we expect that there are only a few markers that are associated with the secondary phenotype Y_2 and therefore the model (4.2.1) should be sparse. Instead of attempting to select the markers that are associated with both the primary and the secondary phenotypes

jointly, we propose a two step procedure, where in Step 1, we use the sure independence screening (Fan and Song (2010)) or the penalized logistic regression model to first select the markers that are associated with the primary phenotype and obtain the parameter estimates of β. We then treat these parameters as known and focus on variable selection for the model of the secondary outcomes.

We propose the following penalized likelihood procedure for variable selection and for estimating the parameter α:

$$\hat{\alpha} = \text{argmin}_{\alpha}\{Q_{\lambda}(\alpha, \gamma) := -l(\alpha, \gamma) + \lambda(\sum_{j=0}^{p} |\alpha_j|)\}, \tag{4.2.6}$$

where $l(\alpha, \gamma)$ is the logarithm of the retrospective likelihood function, λ is a tuning parameter that controls the degree of sparsity, and $\|\alpha\| = \sum_{j=0}^{p} |\alpha_j|$ is the L_1 norm. We include α_0 in this optimization procedure to simplify the computation. In practice, we can re-scale this parameter to make it almost unpenalized.

4.3 A Coordinate Gradient Descent Method

To minimize the objective function (4.2.6), we propose to develop a profile minimization approach to obtain the estimate of the association parameter γ. Specifically, for each tuning parameter λ, we vary γ in the range $[-0.5, 0.5]$ with an interval of 0.1 and estimate $\hat{\alpha}(\gamma)$ by minimizing $Q_{\lambda}(\alpha, \gamma)$. We than estimate γ to be the value that maximizes the likelihood $l(\hat{\alpha}(\gamma), \gamma)$.

For a fixed γ, the block coordinate gradient descent (CGD) method is employed to solve the optimization problem (4.2.6) and obtain the estimate α, denoted by $\hat{\alpha}$. This approach is useful when the objective function consists of a smooth function and

a block separable penalty function. The key idea of the method is to approximate the log-likelihood by a convex quadratic and then apply block coordinate descent to generate a descent direction (Tseng (2009)). We use the second-order Taylor expansion $l(\alpha, \gamma)$ at $\hat{\alpha}$ and approximate $Q_\lambda(\alpha, \gamma)$ by $M_\lambda(\mathbf{d})$ defined as

$$M_\lambda(\mathbf{d}) = -\{l(\hat{\alpha}) + \mathbf{d}^T \nabla l(\hat{\alpha}) + \frac{1}{2}\mathbf{d}^T H \mathbf{d}\} + \lambda \|\hat{\alpha} + \mathbf{d}\|_1, \tag{4.3.1}$$

where $\mathbf{d} \in R^{p+1}$ and H is a diagonal matrix approximating the Hessian matrix. Next we choose a nonempty index subset $J \subseteq N = \{0, 1, 2, ..., p\}$ and move along the direction $\mathbf{d_J}$, where $\mathbf{d_J}$ is the estimated descent direction at $\widehat{\alpha}$ and is defined as $\mathbf{d_J} = \text{argmin}(M_\lambda(\mathbf{d}))$. Specifically, the jth component of $\mathbf{d_J}$ is

$$d_j(\hat{\alpha}) = -\text{mid}(\frac{\nabla l(\hat{\alpha})_j - \lambda}{H_{jj}}, \hat{\alpha}_j, \frac{\nabla l(\hat{\alpha})_j + \lambda}{H_{jj}}), j \in J, \tag{4.3.2}$$

where $\text{mid}(a, b, c)$ denotes the median or mid-point of (a, b, c).

The choice of index subject J is important for the convergence of the CGD method. We use the Gauss-Southwell-q rule (Fletcher (1982)) to select the subset J. We define

$$q_J(\hat{\alpha}, H) = \{-\mathbf{d^T} \nabla l(\hat{\alpha}) - \frac{1}{2}\mathbf{d^T} H \mathbf{d} + \lambda \|\hat{\alpha} + \mathbf{d}\|_1\} - \lambda \|\hat{\alpha}\|_1, \tag{4.3.3}$$

which estimates the descent in Q_λ from $\hat{\alpha}$ to $\hat{\alpha} + \mathbf{d}$. And J is selected to satisfy

$$\{j \in J | q_j(\hat{\alpha}, H) \leq v \text{min} q_j(\hat{\alpha}, H)\}, \tag{4.3.4}$$

where $0 < v \leq 1$ and smaller v results in more coordinates being updated.

An inexact line search using Armijo rule (Conn (2000)) is performed for the approximate step size of the descent direction. Given $\hat{\alpha}$ and $\mathbf{d}$, the step size s is the largest value in $(s_0 \delta^l)_{l \geq 0}$ such that

$$Q_\lambda(\hat{\alpha} + s\mathbf{d}) - Q_\lambda(\hat{\alpha}) \leq c_0 s\Delta, \tag{4.3.5}$$

where $0 < \delta < 1, 0 < c_0 < 1, s_0 > 0$ and Δ is the improvement in the objective function $Q_\lambda(\alpha, \gamma)$ when using a linear approximation for the log-likelihood and

$$\Delta = -\mathbf{d}^T \nabla l(\hat{\alpha}) + \lambda \|\hat{\alpha} + \mathbf{d}\|_1 - \lambda \|\hat{\alpha}\|_1. \tag{4.3.6}$$

In the implementation of the CGD method, the tuning parameters for the Gauss-Southwell-q rule (4.3.4) are set as $v^{[0]} = 0.5$,

$$v^{[t+1]} = \begin{cases} \max(10^{-4}, v^{[t]}/10) \text{ if } s^{[t]} > 10^{-3}; \\ \min(0.5, 50v^{[t]}) \text{ else,} \end{cases} \tag{4.3.7}$$

and the stepsize $s^{[t]}$ is chosen by the Armijo rule (4.3.5) with $c_0 = 0.1, \delta = 0.5, s_0^{[0]} = 1, s_0^{[t+1]} = \min(s^{[t]}/0.5, 1)$.

We summarize the block coordinate gradient descent optimization procedure as following:

Given $\hat{\alpha}^{[t]}$,

1. Compute $\nabla l_\gamma(\hat{\alpha}^{[t]})$ and $\nabla^2 l_\gamma(\hat{\alpha}^{[t]})$, and let $H^{[t]} = -\text{diag}(\max(\nabla^2 l_\gamma(\hat{\alpha}^{[t]})_{jj}, c^*))$, where $c^* > 0$ is a lower bound to ensure convergence.

2. Obtain $\mathbf{d_J}(\hat{\alpha}^{[\mathbf{t}]})$ for $J = N = \{0, 1, 2, ..., p\}$.

3. Choose a nonempty index set $J^{[t]} \subseteq N$ using Gauss-Southwell-q rule.

4. Set $d_j(\hat{\alpha}^{[t]}) = 0$ for all $j \notin J^{[t]}$, and let $\mathbf{d}^{[\mathbf{t}]} = \mathbf{d_J}(\hat{\alpha}^{[\mathbf{t}]})$.

5. Choose a stepsize $s^{[t]}$ according to Armijo rule.

6. Update $\hat{\alpha}^{[t+1]} = \hat{\alpha}^{[t]} + s^{[t]}\mathbf{d}^{[\mathbf{t}]}$.

The iteration is repeated until some convergence criterion is met. We stop the interactions when the change of the likelihood functions is less than 10^{-5}. We show in Appendix that the objective function is convex with probability converging to 1 when the sample size goes to ∞, and therefore the algorithm converges to the global minimum with high probability.

4.4 Choosing the Tuning Parameter λ

The Bayes information criteria (BIC) can be employed to select the tuning parameter λ that minimizes the following quantity

$$BIC(\lambda) = -2\log\{L(\lambda; \hat{\gamma}, \hat{\alpha})\} + k\log(n), \tag{4.4.1}$$

where $L(\lambda; \hat{\gamma}, \hat{\alpha})$ is the value of the likelihood function with tuning parameter λ and the estimate of $\hat{\gamma}$ and $\hat{\alpha}$, k is the number of nonzero elements of $\hat{\alpha}$, and n is the total number of observations.

The BIC works well for variable selection in relatively low or moderate dimensional settings. When p is very large, alternatively, the choice of tuning parameter λ can be selected based on an extended Bayes information criteria (eBIC) (Chen (2008)) that was recently developed for high dimensional settings. Specifically, we can select λ by minimizing the following criteria,

$$eBIC(\lambda) = -2\log(L(\lambda; \hat{\gamma}, \hat{\alpha})) + k\log(n) + 2\gamma^{'}\log\binom{p}{k},\ 0 \leq \gamma^{'} \leq 1, \tag{4.4.2}$$

where the last term is added to the usual BIC to account for the complexity of the model space defined by $\binom{p}{k}$, where p is the number of variables under consideration

and k the size of the model. It works well when the number of covariates p is much larger than the sample size n. Three values of $\gamma^{'}$ are of special interest, that is $\gamma^{'} = 0, 0.5, 1$. The value of $\gamma^{'} = 0$ corresponds to the original BIC, value of $\gamma^{'} = 1$ ensures consistency of eBIC when $p = p_n = O(n^{k_0})$ for any $k_0 \geq 0$ not depending on n. The value of 0.5 ensures consistency when $k_0 < 1$.

4.5 Simulation Studies

4.5.1 Simulation

In this section, we conduct simulations to evaluate the variable selection performance of the proposed penalized Gaussian copula model approach in secondary trait analysis. We simulated $p = 100$ SNPs with the first 5 elements of the coefficients α being nonzero and the rest being zero. We set the minor allele frequency (MAF) of these SNPs to be 0.3 and the regression coefficients of the secondary trait α by controlling the heritability $h^2 = \frac{\alpha^2 2p(1-p)}{1+\alpha^2 2p(1-p)}$, which is the proportion of phenotypic variation explained by the test SNP. We consider two sets of heritability $h^2 = 0.007$, which corresponds to strong association and $h^2 = 0.005$, which corresponds to moderate association. We consider three different degrees of correlation $\gamma = 0, 0.2, 0.5$. For each scenario the 5 SNPs that are correlated with the secondary phenotype are either associated with disease status ($\beta_i = \log(1.2), i = 1..5$) or not associated with the disease status ($\beta_i = 0, i = 1..5$).

To generate the binary-normal variable, we first simulate y_{2i} from a normal distri-

bution $N(\mu_2, 1)$, where $\mu_2 = \sum_1^5 g_i \alpha_i$, and then simulate y_{1i} by conditional probability $y_{1i}|y_{2i}$ (Eqn 4.2.4). The case control status is determined by comparing a randomly generated number with the conditional probability. We sample 500 cases and 500 controls for each dataset. We repeat the simulations 50 times.

4.5.2 Variable Selection

To evaluate the variable selection performance of our method, we compare results of the proposed penalized likelihood approach with those from Lasso and Lasso adjusting the case-control status. For each method, we select the tuning parameter using BIC, eBIC with $\gamma' = 0.5$ and eBIC with $\gamma' = 1$. For each setting, we repeat the simulation 50 times and calculate the average number of nonzero coefficients correctly estimated to be nonzero (denoted by C) and the average number of zero coefficients incorrectly estimated to be nonzero (denoted by IC) over 50 simulations. The results were summarized in Table 4.1. The false positive rate (IC) for penalized likelihood approach and Lasso adjusting for case-control status is stable and relatively low across all settings. Lasso tends to select more irrelevant variables especially when the correlation between the primary and secondary traits is strong. In terms of correctly selected variables (C), when $OR = 1.2$, Lasso performs slightly better than the penalized approach while the Lasso adjusting for the case control status performs the worst. When $OR = 1.0$, the penalized approach is more efficient than the Lasso and Lasso adjusting for case-control status. In addition, the original BIC works well in this simulation setup in which $p(100) << n(1000)$.

The proposed penalized likelihood approach also out-performs the single marker analysis when Bonferonni correction or the FDR control is used to adjust for multiple comparisons. The single marker analysis is often conservative and leads to loss of power for detecting the true secondary phenotype associated genetic variants.

4.5.3 Parameter Estimation

Table 4.2 presents the predictive risk $R(\alpha, \hat{\alpha}) = E_\alpha|\mathbf{g}\hat{\alpha} - \mathbf{g}\alpha|^2 = \sum_{i=1}^{n} \frac{|\mathbf{g}\hat{\alpha}-\mathbf{g}\alpha|_i^2}{n}$ (Foster and George (1994)) using Lasso, Lasso adjusting for the case-control status and penalized likelihood approaches. We observe that in general the penalized likelihood and Lasso adjusting for case-control status consistently have smaller predictive risks under all the settings considered. The predictive risk using Lasso is close to the true value when the correlation is zero. However, its predictive risk deviates from the true value as the correlation increases, caused by the biased estimate of the intercept.

4.6 Real Data Analysis

To illustrate the effectiveness of our method in the analysis of real data, we applied the penalized likelihood approach to a genetic association study on high HDL Cholesterol (HDL-C). HDL-C is a cross-sectional study of genetic factors contributing to elevated HDL-C levels and is composed of subjects with extreme levels of HDL-C. In this study, subjects of European ancestry with HDL $> 90^{th}$ percentile for age and gender were considered as cases and subjects with HDL $< 30^{th}$ percentile for age and gender were considered as controls. A total of 605 cases and 624 controls were

genotyped using Affymetrix 6.0 SNP array. The primary phenotype is the HDL-C case-control status. In addition, all study subjects have measurements on LDL-C and separate measures on apolipoprotein concentration such as ApoB.

Our GWAS analysis revealed strong association between HDL case-control status and a number of SNPs in gene *LPL*, a gene located on chromosome 8. Fig 4.1 displays the linkage disequilibrium(LD) structure of *LPL*. This gene has 95 SNPs with most SNPs not in strong LD with each other. It is well-known that *LPL* plays an important role in the regulation of HDL-C level (Lettre (2011)). In our analysis, we treated LDL-C and ApoB level as the secondary phenotypes and aimed to identify the possible variants in this genes that are associated with these secondary phenotypes. The analysis with single marker association did not find any association between *LPL* and LDL. However, we did see strong association between *LPL* and ApoB.

We applied the proposed penalized likelihood approach to select the SNPs that are associated with the secondary phenotypes. We also applied the Lasso and Lasso adjusting for case-control status for comparison. For the penalized likelihood approach, we standardized each secondary phenotype by subtracting its mean estimated from the controls and then divided by the standard deviation estimated using all samples.

Table 4.3 listed the SNPs identified to be associated with the two secondary phenotypes using different approaches along with the p-values from single marker association analysis using the copula approach (He (in revision)). For ApoB, whose level is weakly correlated with HDL-C case control status ($\gamma = -0.1$), SNPs rs1018078 (p-value=0.0213) and rs11994862 (p-value=0.0362) are associated with ApoB in single marker association. However, the p-values for the two SNPs are no longer significant

when adjusted by multiple testing. Using the approach of variable selection, Lasso and Lasso adjusting for case-control status identified one single SNP rs1018078, Lasso with eBIC ($\gamma^{'} = 0.5, 1$) identified the same two SNPs as in single marker association and one addition SNP rs10503670. Lasso adjusting for case-control status with eBIC ($\gamma^{'} = 0.5, 1$) identified two SNPs rs11994862 and rs10503670. The penalized likelihood approach ($\gamma^{'} = 0, 0.5, 1$) identified the same set of SNPs as the Lasso with eBIC ($\gamma^{'} = 0.5, 1$).

LDL-C shows a moderate correlation with HDL-C ($\gamma = 0.15$), but none of *LPL* SNPs are associated with this trait in single marker association. SNPs identified in variable selection are probably due to false discovery. From Table 4.3, one SNP was selected by penalized likelihood approach ($\gamma^{'} = 0, 0.5, 1$), Lasso and Lasso adjusting for the case-control status. Lasso and Lasso when adjusting for the case-control status selected two SNPs for LDL-C when eBIC ($\gamma^{'} = 0.5, 1$) was used. These results indicate that the proposed penalized likelihood approach can potentially lead to better power of identifying the genetic variants that are associated with the secondary phenotypes.

4.7 Discussion

In this paper we have proposed a Gaussian copula model and a penalized likelihood approach for selecting variables that are associated with the secondary phenotype in case control genetic association study. This work is an extension of the previous work of single marker association but allows for multiple genetic effects on the secondary phenotypes. A tuning parameter in the penalized likelihood function is employed to

control the sparsity of the solution and serve the purpose of variable selection. We performed simulations to demonstrate that our method is more efficient in selecting variables that are associated with the secondary phenotypes than direct application of Lasso or Lasso with adjustment of case-control status. We also demonstrated that the penalized likelihood approach has well controlled false discovery rate when comparing with Lasso. We have further demonstrated that penalized likelihood approach has best sensitivity and specificity in selecting variables associated with LDL-C and ApoB for gene *LPL* in a real HDL-C study.

In analysis of real data sets, we performed a two-step approach of the secondary phenotype analysis where in step 1, we propose to apply sure independence screening or penalized logistic regression to first identify the variants that are associated with the primary phenotype. We then treat these variants as known in our penalized estimation. We observed through simulation our method is very robust to the number and specification of the loci selected in step 1. An alternative approach is to develop methods that can identify the variants that are associated with the primary and secondary phenotypes simultaneously.

4.8 Appendix

We verify that the Fisher's information associated with the logarithm of the retrospective likelihood function (4.2.5) is positive, which implies that the negative of the log-likelihood function is convex with high probability when $n \to \infty$.

let $x = \frac{\Phi^{-1}(\mu_1)+\gamma(y_2-\mu_2)}{\sqrt{1-\gamma^2}}$ let $t = (y_2 - \mu_2) + \gamma\Phi^{-1}(\mu_1)$

(A) The copula model is identifiable for α_0, α_1.

$$\begin{aligned}
E[\frac{\partial l}{\partial \alpha_i}] &= E[\frac{\partial l}{\partial \alpha_i}|_{D=1}] + E[\frac{\partial l}{\partial \alpha_i}|_{D=0}] \\
&= \int \frac{\partial l}{\partial \alpha_i}|_{D=1} \times P(y=1)dy + \int \frac{\partial l}{\partial \alpha_i}|_{D=0} \times P(y=0)dy \\
&= \frac{\partial \mu_2}{\partial \alpha_i}[\int (y-\mu_2)\phi(y_2;\mu_2)\Phi(x)dy - \int \frac{\gamma}{\sqrt{1-\gamma^2}}\phi(x)\phi(y_2;\mu_2)dy \\
&+ \int (y-\mu_2)\phi(y_2;\mu_2)(1-\Phi(x))dy + \int \frac{\gamma}{\sqrt{1-\gamma^2}}\phi(x)\phi(y_2;\mu_2)dy] \\
&= \frac{\partial \mu_2}{\partial \alpha_i}\int (y-\mu_2)\phi(y_2;\mu_2)dy \\
&= 0
\end{aligned} \tag{4.8.1}$$

$$\begin{aligned}
E[\frac{\partial l}{\partial \alpha_j}\frac{\partial l}{\partial \alpha_k}] &= E[\frac{\partial l}{\partial \alpha_j}\frac{\partial l}{\partial \alpha_k}|_{D=1}] + E[\frac{\partial l}{\partial \alpha_j}\frac{\partial l}{\partial \alpha_k}|_{D=0}] \\
&= \frac{\partial \mu_2}{\partial \alpha_j}\frac{\partial \mu_2}{\partial \alpha_k}[\int (y-\mu_2)^2\phi(y_2;\mu_2)\Phi(x)dy - 2\int (y-\mu_2)\frac{\gamma}{\sqrt{1-\gamma^2}}\phi(x)\phi(y_2;\mu_2)dy \\
&+ \int \frac{\gamma^2}{1-\gamma^2}\frac{\phi(x)^2\phi(y_2;\mu_2)}{\Phi(x)}dy] + \frac{\partial \mu_2}{\partial \alpha_j}\frac{\partial \mu_2}{\partial \alpha_k}[\int (y-\mu_2)^2\phi(y_2;\mu_2)(1-\Phi(x))dy \\
&+ 2\int (y-\mu_2)\frac{\gamma}{\sqrt{1-\gamma^2}}\phi(x)\phi(y_2;\mu_2)dy + \int \frac{\gamma^2}{1-\gamma^2}\frac{\phi(x)^2\phi(y_2;\mu_2)}{(1-\Phi(x))}dy] \\
&= \frac{\partial \mu_2}{\partial \alpha_j}\frac{\partial \mu_2}{\partial \alpha_k}[1 + \int \frac{\gamma^2}{1-\gamma^2}\frac{\phi(x)^2\phi(y_2;\mu_2)}{\Phi(x)(1-\Phi(x))}dy]
\end{aligned} \tag{4.8.2}$$

As

$$\frac{\partial^2 l}{\partial \alpha_j \partial \alpha_k}|_{D=1} = -\frac{\partial \mu_2}{\partial \alpha_j}\frac{\partial \mu_2}{\partial \alpha_k}(1 + \frac{\gamma^2}{1-\gamma^2}(\frac{x\phi(x)}{\Phi(x)} + \frac{\phi(x)^2}{\Phi(x)^2})) \tag{4.8.3}$$

$$\frac{\partial^2 l}{\partial \alpha_j \partial \alpha_k}|_{D=0} = -\frac{\partial \mu_2}{\partial \alpha_j}\frac{\partial \mu_2}{\partial \alpha_k}(1 + \frac{\gamma^2}{1-\gamma^2}(-\frac{x\phi(x)}{1-\Phi(x)} + \frac{\phi(x)^2}{(1-\Phi(x))^2})) \tag{4.8.4}$$

So

$$-E[\frac{\partial^2 l}{\partial\alpha_j\partial\alpha_k}] = \frac{\partial\mu_2}{\partial\alpha_j}\frac{\partial\mu_2}{\partial\alpha_k}[1+\int\frac{\gamma^2}{1-\gamma^2}\frac{\phi(x)^2\phi(y_2;\mu_2)}{\Phi(x)}dy+\int\frac{\gamma^2}{1-\gamma^2}\frac{\phi(x)^2\phi(y_2;\mu_2)}{1-\Phi(x)}dy]$$
$$= \frac{\partial\mu_2}{\partial\alpha_j}\frac{\partial\mu_2}{\partial\alpha_k}[1+\int\frac{\gamma^2}{1-\gamma^2}\frac{\phi(x)^2\phi(y_2;\mu_2)}{\Phi(x)(1-\Phi(x))}dy] \tag{4.8.5}$$

Thus we have

$$E[\frac{\partial l}{\partial\alpha_j}\frac{\partial l}{\partial\alpha_k}] = -E[\frac{\partial^2 l}{\partial\alpha_j\partial\alpha_k}] \tag{4.8.6}$$

(B) The Fisher information matrix M is positive definite. The element of M

$$M_{j,k} = E[-\frac{\partial^2 l}{\partial\alpha_j\partial\alpha_k}] = \frac{\partial\mu_2}{\partial\alpha_j}\frac{\partial\mu_2}{\partial\alpha_k}[1+\int\frac{\gamma^2}{1-\gamma^2}\frac{\phi(x)^2\phi(y_2;\mu_2)}{\Phi(x)(1-\Phi(x))}dy] > \frac{\partial\mu_2}{\partial\alpha_j}\frac{\partial\mu_2}{\partial\alpha_k}$$
$$= \sum_{i=1}^{n} g_{i,j}\times\sum_{i=1}^{n} g_{i,k} = \mathbf{g_j^{'}}\times\mathbf{g_k} \tag{4.8.7}$$

The information matrix M can be written as

$$M = \begin{pmatrix}\mathbf{g_1^{'}}\\ \mathbf{g_2^{'}}\\ \cdots\\ \mathbf{g_p^{'}}\end{pmatrix}\times\begin{pmatrix}\mathbf{g_1} & \mathbf{g_2} & \cdots & \mathbf{g_p}\end{pmatrix} = G^{'}\times G \tag{4.8.8}$$

For all nonzero vector z,

$$z^{'}\times M\times z = z^{'}\times G^{'}\times G\times z = (Gz)^{'}\times(Gz) > 0 \tag{4.8.9}$$

so M is positive definite.

Figure 4.1: LD structure of *LPL* gene on chromosome 8 in the Affymetrix samples.

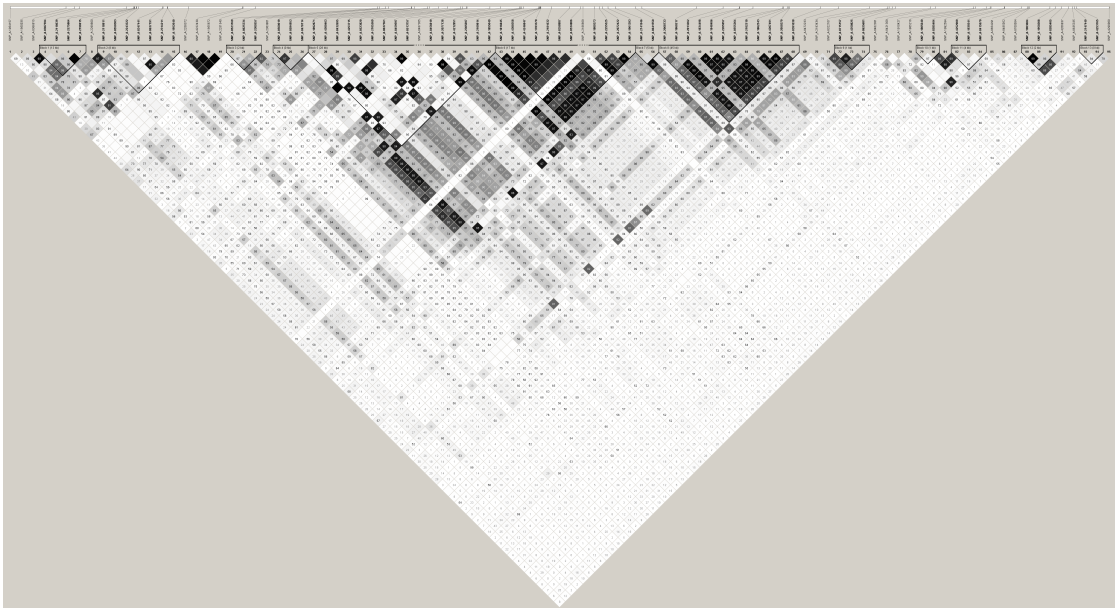


Table 4.1: Simulation results - variable selection. The column labeled with C (or IC) represents the average number of correctly (or incorrectly) identified variables and their SEs.

	C	IC	C	IC	C	IC
	$h^2 = 0.007, OR = 1.2, \gamma = 0.5$		$h^2 = 0.007, OR = 1.2, \gamma = 0.2$		$h^2 = 0.007, OR = 1.2, \gamma = 0$	
Lasso	3.84(0.89)	1.88(1.56)	3.12(0.90)	0.84(0.84)	2.72(0.93)	0.92(0.97)
Lasso-E.5	3.92(0.85)	2.10(1.62)	3.24(0.77)	0.88(0.85)	2.78(0.93)	1.02(0.91)
Lasso-E	**3.94(0.84)**	2.16(1.62)	**3.30(0.81)**	0.92(0.85)	2.82(0.90)	1.10(0.97)
Adjusted Lasso	1.90(0.99)	0.26(0.60)	2.12(1.12)	0.36(0.69)	2.64(0.98)	0.56(0.84)
Adjusted Lasso-E.5	2.34(0.75)	0.48(0.79)	2.48(0.93)	0.50(0.76)	2.80(0.95)	0.68(0.87)
Adjusted Lasso-E	2.38(0.75)	0.50(0.79)	2.64(0.94)	0.64(0.80)	2.86(0.95)	0.68(0.87)
Penalized copula	3.20(0.88)	0.28(0.67)	2.68(0.89)	0.30(0.58)	2.76(0.89)	0.46(0.79)
Penalized copula-E.5	3.34(0.80)	0.32(0.68)	2.92(0.72)	0.34(0.59)	2.88(0.87)	0.56(0.79)
Penalized copula-E	3.38(0.78)	0.38(0.75)	3.04(0.64)	0.42(0.64)	**2.96(0.83)**	0.58(0.78)
Bon-0.05	1.68(1.10)	**0.04(0.20)**	1.08(0.85)	**0.14(0.35)**	0.78 (0.79)	**0.02(0.14)**
FDR-0.05	2.10(1.41)	0.22(0.46)	1.42(1.05)	0.28(0.54)	1.10(1.03)	0.06(0.24)
FDR-0.10	2.64(1.41)	0.60(0.99)	1.92(1.10)	0.48(0.74)	1.80(1.41)	0.24(0.48)
	$h^2 = 0.007, OR = 1.0, \gamma = 0.5$		$h^2 = 0.007, OR = 1.0, \gamma = 0.2$		$h^2 = 0.007, OR = 1.0, \gamma = 0$	
Lasso	2.38(1.03)	1.94(1.36)	2.58(1.13)	1.08(0.99)	2.50(1.07)	1.06(1.24)
Lasso-E.5	2.48(0.95)	2.06(1.24)	2.64(1.10)	1.26(0.94)	2.74(1.01)	1.22(1.23)
Lasso-E	2.52(0.97)	2.22(1.23)	2.64(1.10)	1.38(0.95)	2.78(0.97)	1.28(1.25)
Adjusted Lasso	2.50(0.99)	0.30(0.51)	2.46(1.27)	0.38(0.60)	2.50(1.25)	0.50(0.76)
Adjusted Lasso-E.5	2.72(0.83)	0.38(0.57)	2.60(1.20)	0.54(0.65)	2.76(1.13)	0.70(0.84)
Adjusted Lasso-E	2.82(0.80)	0.46(0.58)	2.66(1.19)	0.64(0.69)	**2.88(1.04)**	0.82(0.90)
Penalized copula	2.76(1.12)	0.60(0.81)	2.56(1.15)	0.40(0.64)	2.58(1.16)	0.58(0.93)
Penalized copula-E.5	2.98(1.00)	0.68(0.79)	2.66(1.08)	0.58(0.67)	2.72(1.09)	0.68(0.91)
Penalized copula-E	**3.06(0.96)**	0.76(0.77)	**2.82(1.06)**	0.68(0.71)	2.80(1.05)	0.84(0.98)
Bon-0.05	1.60(1.18)	**0.18(0.44)**	1.14(1.03)	**0(0)**	0.94(0.71)	**0.02(0.14)**
FDR-0.05	1.98(1.36)	0.44(0.73)	1.42(1.30)	0.10(0.30)	1.18(0.96)	0.14(0.53)
FDR-0.10	2.36(1.35)	0.94(1.38)	2.04(1.38)	0.28(0.57)	1.50(1.05)	0.34(0.72)
	$h^2 = 0.005, OR = 1.2, \gamma = 0.5$		$h^2 = 0.005, OR = 1.2, \gamma = 0.2$		$h^2 = 0.005, OR = 1.2, \gamma = 0$	
Lasso	3.28(0.99)	1.66(1.44)	2.14(1.11)	1.10(0.95)	1.84(1.08)	0.84(0.77)
Lasso-E.5	3.34(0.96)	1.76(1.45)	2.30(1.11)	1.20(0.93)	2.10(0.91)	1.02(0.87)
Lasso-E	**3.40(0.93)**	1.88(1.38)	**2.38(1.09)**	1.32(0.89)	2.14(0.90)	1.20(0.86)
Adjusted Lasso	1.30(0.89)	0.32(0.55)	1.52(1.01)	0.54(0.68)	1.72(0.99)	0.46(0.54)
Adjusted Lasso-E.5	1.74(0.94)	0.64(0.88)	1.80(0.99)	0.84(0.87)	1.94(0.89)	0.68(0.62)
Adjusted Lasso-E	1.86(0.95)	0.70(0.93)	1.86(0.99)	0.98(0.84)	2.04(0.88)	0.72(0.64)
Penalized copula	2.44(1.11)	0.32(0.65)	1.82(0.98)	0.40(0.67)	1.98(0.96)	0.36(0.48)
Penalized copula-E.5	2.66(0.94)	0.46(0.73)	2.14(0.95)	0.62(0.75)	2.22(0.82)	0.42(0.50)
Penalized copula-E	2.68(0.96)	0.48(0.76)	2.20(0.99)	0.76(0.80)	**2.38(0.78)**	0.50(0.58)
Bon-0.05	0.96(0.97)	**0.16(0.51)**	0.50(0.54)	**0.10(0.36)**	0.62(0.75)	**0.08(0.27)**
FDR-0.05	1.12(1.13)	0.32(0.62)	0.70(0.79)	0.16(0.47)	0.72(0.93)	0.14(0.35)
FDR-0.05	1.56(1.21)	0.62(0.90)	1.02(1.10)	0.30(0.58)	1.12(1.08)	0.20(0.45)
	$h^2 = 0.005, OR = 1.0, \gamma = 0.5$		$h^2 = 0.005, OR = 1.0, \gamma = 0.2$		$h^2 = 0.005, OR = 1.0, \gamma = 0$	
Lasso	1.60(1.03)	2.18(1.30)	1.72(0.90)	1.18(1.14)	1.72(1.11)	0.96(0.95)
Lasso-E.5	1.68(0.96)	2.34(1.24)	1.96(0.88)	1.42(1.07)	1.98(0.94)	1.32(1.00)
Lasso-E	1.76(0.89)	2.50(1.22)	2.04(0.83)	1.52(1.03)	2.16(1.00)	1.40(0.97)
Adjusted Lasso	1.80(0.88)	0.26(0.56)	1.62(0.81)	0.32(0.71)	1.72(1.14)	0.52(0.65)
Adjusted Lasso-E.5	2.08(0.80)	0.46(0.65)	2.10(0.84)	0.60(0.76)	2.14(1.01)	0.84(0.82)
Adjusted Lasso-E	2.24(0.87)	0.64(0.80)	2.24(0.80)	0.76(0.85)	**2.28(1.01)**	0.88(0.80)
Penalized copula	2.08(0.97)	0.46(0.73)	1.66(0.92)	0.44(0.73)	1.76(1.08)	0.54(0.68)
Penalized copula-E.5	2.42(0.91)	0.70(0.76)	2.20(0.78)	0.68(0.84)	2.16(0.96)	0.78(0.74)
Penalized copula-E	**2.48(0.91)**	0.72(0.78)	**2.36(0.72)**	0.80(0.86)	2.26(1.01)	0.92(0.72)
Bon-0.05	0.82(0.80)	**0.08(0.27)**	0.58(0.67)	**0.08(0.27)**	0.36(0.53)	**0.04(0.20)**
FDR-0.05	1.04(0.97)	0.24(0.48)	0.72(0.93)	0.16(0.37)	0.46(0.68)	0.08(0.27)
FDR-0.10	1.58(1.18)	0.62(1.00)	1.00(0.99)	0.34(0.56)	0.68(0.78)	0.20(0.40)

Lasso-E is eBIC with $\gamma = 1$ using Lasso. Adjusted Lasso-E.5 is eBIC with $\gamma = 0.5$ using Lasso adjusting for case control status; Adjusted Lasso-E is eBIC with $\gamma = 1$ using Lasso adjusting for case control status; Penalized copula-E.5 is eBIC with $\gamma = 0.5$ using the penalized likelihood approach; Penalized copula-E is eBIC with $\gamma = 1$ using the penalized likelihood approach; Bon-0.05 is the single SNP analysis adjusted for Bonferroni correction; FDR-0.05 is FDR of 0.05 based on single SNP analysis p-value; FDR-0.10 is FDR of 0.10 based on single SNP analysis p-value.

Table 4.2: Predictive risk.

	$h^2 = 0.007$	$h^2 = 0.007$	$h^2 = 0.007$
	$OR = 1.2, \gamma = 0.5$	$OR = 1.2, \gamma = 0.2$	$OR = 1.2, \gamma = 0$
Lasso	0.392	0.082	0.028
Lasso-E.5	0.393	0.081	0.028
Lasso-E	0.394	0.081	0.028
Adjusted Lasso	0.028	0.030	0.026
Adjusted Lasso-E.5	0.026	0.028	0.025
Adjusted Lasso-E	0.026	0.027	0.025
Penalized copula	0.018	0.024	**0.023**
Penalized copula-E.5	**0.017**	**0.022**	**0.023**
Penalized copula-E	**0.017**	**0.022**	**0.023**
	$h^2 = 0.007$	$h^2 = 0.007$	$h^2 = 0.007$
	$OR = 1.0, \gamma = 0.5$	$OR = 1.0, \gamma = 0.2$	$OR = 1.0, \gamma = 0$
Lasso	0.474	0.098	0.028
Lasso-E.5	0.474	0.099	0.028
Lasso-E	0.474	0.100	0.029
Adjusted Lasso	0.023	0.026	0.027
Adjusted Lasso-E.5	0.022	0.025	0.026
Adjusted Lasso-E	0.022	0.026	0.026
Penalized copula	0.022	**0.024**	0.025
Penalized copula-E.5	**0.021**	**0.024**	**0.024**
Penalized copula-E	**0.021**	**0.024**	**0.024**
	$h^2 = 0.005$	$h^2 = 0.005$	$h^2 = 0.005$
	$OR = 1.2, \gamma = 0.5$	$OR = 1.2, \gamma = 0.2$	$OR = 1.2, \gamma = 0$
Lasso	0.402	0.091	0.026
Lasso-E.5	0.402	0.091	0.026
Lasso-E	0.403	0.091	0.027
Adjusted Lasso	0.025	0.027	0.024
Adjusted Lasso-E.5	0.024	0.027	0.024
Adjusted Lasso-E	0.024	0.027	0.023
Penalized copula	**0.019**	**0.024**	0.022
Penalized copula-E.5	**0.019**	**0.024**	0.021
Penalized copula-E	**0.019**	**0.024**	**0.020**
	$h^2 = 0.005$	$h^2 = 0.005$	$h^2 = 0.005$
	$OR = 1.0, \gamma = 0.5$	$OR = 1.0, \gamma = 0.2$	$OR = 1.0, \gamma = 0$
Lasso	0.468	0.099	0.029
Lasso-E.5	0.469	0.100	0.029
Lasso-E	0.469	0.100	0.029
Adjusted Lasso	0.022	0.024	0.026
Adjusted Lasso-E.5	0.022	0.023	0.026
Adjusted Lasso-E	0.022	0.024	0.025
Penalized copula	0.021	0.023	0.025
Penalized copula-E.5	**0.020**	**0.022**	**0.024**
Penalized copula-E	**0.020**	**0.022**	**0.024**

Lasso-E.5 is eBIC with $\gamma = 0.5$ using Lasso. Lasso-E is eBIC with $\gamma = 1$ using Lasso. Adjusted Lasso-E.5 is eBIC with $\gamma = 0.5$ using Lasso adjusting for case control status; Adjusted Lasso-E is eBIC with $\gamma = 1$ using Lasso adjusting for case control status; Penalized copula-E.5 is eBIC with $\gamma = 0.5$ using the penalized likelihood approach; Penalized copula-E is eBIC with $\gamma = 1$ using the penalized likelihood approach.

Table 4.3: Real data analysis results - Analysis of secondary phenotypes ApoB and LDL-C.

	ApoB	coefficient	P-value	LDL-C	coefficient	P-value
Lasso	rs1018078	0.110	0.0213	rs13263508	-0.0987	0.0567
Lasso-E.5	rs1018078	0.0695	0.0213	rs326	0.0793	0.0984
	rs11994862	0.0523	0.0362	rs13263508	-0.0791	0.0567
	rs10503670	0.0611	0.242			
Lasso-E	rs1018078	0.0695	0.0213	rs326	0.0793	0.0984
	rs11994862	0.0523	0.0362	rs13263508	-0.0791	0.0567
	rs10503670	0.0611	0.242			
Adjusted Lasso	rs1018078	0.109	0.0213	rs13263508	-0.0782	0.0567
Adjusted Lasso-E.5	rs1018078	0.0897	0.0213	rs13263508	-0.0811	0.0567
	rs10503670	0.0677	0.242	rs10503670	0.0770	0.0835
Adjusted Lasso-E	rs1018078	0.0897	0.0213	rs13263508	-0.0811	0.0567
	rs10503670	0.0677	0.242	rs10503670	0.0770	0.0835
Penalized copula	rs1018078	0.0682	0.0213	rs10503670	0.0749	0.0835
	rs11994862	0.0492	0.0362			
	rs10503670	0.0675	0.242			
Penalized copula-E.5	rs1018078	0.0682	0.0213	rs10503670	0.0749	0.0835
	rs11994862	0.0492	0.0362			
	rs10503670	0.0675	0.242			
Penalized copula-E	rs1018078	0.0682	0.0213	rs10503670	0.0749	0.0835
	rs11994862	0.0492	0.0362			
	rs10503670	0.0675	0.242			

Lasso-E is eBIC with $\gamma = 1$ using Lasso. Adjusted Lasso-E.5 is eBIC with $\gamma = 0.5$ using Lasso adjusting for case control status; Adjusted Lasso-E is eBIC with $\gamma = 1$ using Lasso adjusting for case control status; Penalized copula-E.5 is eBIC with $\gamma = 0.5$ using the penalized likelihood approach; Penalized copula-E is eBIC with $\gamma = 1$ using the penalized likelihood approach.

Chapter 5

Conclusion

In this dissertation, we have developed new statistical methods to map genetic variants that are associated with complex human diseases, motivated by analysis of the high HDL candidate gene association study conducted at Penn cardiovascular institute. In Chapter 2, we have developed a method to detect gene-gene interactions by incorporating the external LD information obtained from the HapMap project. This method has better power that the SNP-based tests when more than two variants interact with each other. We conducted simulations to demonstrate that tests that incorporate external LD information are generally more powerful than those that use genotyped markers only. This method can be applied to genome-wide association setting and can be used as a screening tool to detect gene-gene interactions. We expect more power improvement when incorporating external information such as data released from 1000 Genomes Project. This idea can be extended to other statistical methods for rare variants analysis.

In Chapters 3 and 4, we have developed a Gaussian copula approach to analyze

secondary phenotypes in case control genetic association studies. The Gaussian copula model provides a natural way of jointly modeling the secondary phenotype with the disease status via an association parameter. When only one marker is considered in the model, we estimate the parameters using a Gauss-Newton algorithm and perform a Wald test to access whether the marker considered is associated with the secondary phenotype. Our model improves over the existing methods in several aspects. First, it allows us to correct the sampling bias of the secondary phenotype by modeling its dependence with the primary phenotype. Second, the secondary phenotype does not need to have a marginal normal distribution, any phenotype from the exponential family can be easily incorporated into the analysis. Third, this method can be naturally extended to the analysis of multiple secondary phenotypes. Through simulations, we have demonstrated that our method yields correct type I error rates under a wide range of settings. Although our method is a full parametric model, we also showed that our method is robust to model specification when the data is simulated from Lin and Zeng's method (Lin and Zeng (2009)).

In Chapter 4, we further consider the Gaussian copula model to include a large number candidate SNP markers. Instead of performing a statistical test, we take a variable selection approach to identify the genetic variants that are associated with the secondary phenotype. We have developed a penalized likelihood approach using the L_1 penalty to select the genetic variants. We have developed an efficient coordinate gradient descent algorithm to solve the optimization problem. In contrast to single SNP analysis, this method avoids the multiple testing problem and can lead to better power of identifying the genetic variants that are associated with the sec-

ondary phenotypes. By using the retrospective likelihood function, our method can also adjust for the sampling bias and result in more precise estimates of the genetic effects on the secondary phenotypes.

In summary, we have develop several novel statistical methods for identifying the genetic variants that are associated with complex phenotypes. The methods presented in this dissertation provide a set of valuable tools for different statistical analysis issues emerged in genome wide association study. Finally, we have developed software package that is freely available.

Bibliography

Brown ML, Inazu A, Hesler CB et al. (1989). Molecular basis of lipid transfer protein deficiency in a family with increased high-density lipoproteins. *Nature*; **342**: 448–451.

Chapman J and Clayton D. (2007). Detecting association using epistatic information. *Genet Epidemiol*; **31**: 894–909.

Chaseman DI, Pare G, Mora S, Hopewell JC, Peloso G, Clarke R, Cupples A et al. (2009). Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis. *PLoS Genet*; **5**: e1000730.

Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S. (2006). Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. *Am J Hum Genet*; **79**: 1002–1016.

Chen J, Chen Z. (2008). Extended Baysian information criteria for model selection with large model spaces. *Biometrika*; **95**: 759–771.

Conn AR, Gould NIM, Toint PL. (2000). Trust-region methods. *Philadelphia: SIAM*.

Cordell HJ, Todd JA, Bennett ST et al. (1995). So many correlated tests, so little time! Rapid adjustment of p-values for multiple correlated tests. *Am J Hum Genet*; **57**: 920–934.

Cox NJ, Frigge M, Nicolae DL et al.(1990). Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to type 2 diabetes. *Nat Genet*; **21**: 213–215.

De Bakker P, Yelensky R, Pe'er Itsik, Gabriel SB, Daly MJ, Altshuler D. (2005). Efficiency and power in genetic association studies. *Nat Genet*; **37**: 1217–1223.

De Leon AR, Wu B. (2004). Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Stat Med*; **30**: 175–185.

Fan J, Song R. (2010). Sure independence screening in generalized linear models whith np-denemsionality. *Ann Stat*; **38(6)**: 3567-3604.

Fletcher R. (1982). A model algorithm for composite nondifferentiable optimazation problems. *Math Program Stud*; **17**: 67–76.

Foster D. and George E. (1994) The risk inflation criterion for multiple regression. *Ann Statist*; **22(4)**: 1947–1975.

Gauderman WJ, Murcray C, Gilliland F, Conti DV.(2007). Testing association between disease and multiple SNPs in a candidate gene. *Genet Epidemiol*; **31**: 383–395.

Gillum RF. (1993). The association between serum albumin and HDL and total cholesterol. *J Nat Med Assoc*; **85**: 290–292.

Grundy SM, Brewer HB, Cleeman JI, Smith SC and Lenfant C. (2004). Definition of metabolic syndrome: report of the National Heart, Lung, and Blood Institute/American Heart Association conference on scientific issues related to definition. *Circulation*; **109**: 433–438.

He J, Li H, Edmondson A, Rader D, Li M. A gaussina copula approach for the analysis of secondary phenotypes in case-control genetic association studies. *Biostatistics*; in revision.

Hoh J, Ott J. (2003). Mathematical multi-locus approaches to localizing complex human trait genes. *Nat Rev*; **4**: 701–709.

Howard TD, Koppelman GH, Xu J, Zheng SL, Postma DS, Meyers DA, Bleecker ER. (1990). Gene-gene interaction in asthma: *IL4RA* and *IL13* in a dutch Population with asthma. *Am J Hum Genet*; **70**: 230–236.

Inazu A, Brown ML, Hesler CB et al. (1990). Increased high-density lipoprotein levels caused by a common cholesteryl-ester transfer protein gene mutation. *N Engl J Med*; **323**: 1234–1238.

Kammerer CM, Gouin N, Samollow PB, VandeBerg JF, Hixson JE, Cole SA, MacCluer JW and Atwood LD. (2004). Two quantitative trait loci affect ACE activities in Mexican-Americans. *Hypertension*; **43**: 466–470.

Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, Rieder M J, Cooper GM et al. (2008).Six new loci associated with blood low-density lipoprotein cholesterol,

high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet*; **40**: 189–197.

Keating BJ, Tischfield S, Murray SS, Bhangale T, Price TS, Glessner JT, Galver L et al. (2008). Concept, design and implementation of a cardiovascular gene-centric 50 K SNP array for large-scale genomic association studies. *PLoS ONE*; **3(10)**: e3583.

Kraft P, Thomas DC. (2000). Bias and efficiency in family-based gene-characterization studies: Conditional, prospective, retrospective, and joint likelihoods. *Am J Hum Genet*; **66**: 1119–1131.

Lettre G, Jackson AU, Gieger C, Schumacher FR, Bernde SI, Sanna S et al. (2008). Idetification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet*; **40**: 584–591.

Lettre G, Palmer CD et al. (2011). Genome-Wide Association Study of Coronary Heart Disease and Its Risk Factors in 8,090 African Americans: The NHLBI CARe Project. *PLoS Genet*; **Feb 10;7(2)**: e1001300.

Li M, Boehnke M, Abecasis GR and Song P X-K. (2006). Quantitative trait linkage analysis using Gaussian copulas. *Genetics*; **173**: 2317–2327.

Li M, Wang K, Grant SFA, Hakonarson H, Li C. (2009). A powerful gene-based association test by combining optimally weighted markers. *Bioinformatics*; **25**: 497–503.

Li Y, Willer CJ, Sanna S and Abecasis GR. (2009). Genotype imputation. *Annu Rev Genomics Hum Genet*; **10**: 387–406.

Lin DY and Zeng D. (2009). Proper analysis of secondary phenotype data in case-control association studies. *Genet Epidemiol*; **33**: 256–265.

Loos RJ, Lindgren CM, Li S, Wheeler E, Zhao JH, Prokopenko I, Inouye M et al. (2008). Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat Genet*; **40**: 768–775.

Marchini J, Donnelly P, Cardon LR.(2005). Genome-wide strategies for detecting multiple loci that influence complex disease. *Nat Genet*; **37**: 413–417.

Marchini J, Howie B, Myers S, McVean G and Donnelly P. (2007). A new multi-point method for genome-wide association studies via imputation of genotypes. *Nat Genet*; **39**: 906–913.

Monsees GM, Tamimi RM and Kraft P. (2009). Genome-wide association scans for secondary traits using case-control samples. *Genet Epidemiol*; **33**: 717–728.

Moore JH (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human Hered*; **56**: 73–82.

Moore JH, Williams SM. (2002). New strategies for identifying gene-gene interactions in hypertension. *Ann Med*; **34**: 88–95.

Moore JH, Williams SM. (2009). Epistasis and its implications for personal genetics. *Am J Hum Genet*; **85**: 309–320.

Neale BM, Sham PC.(2004). The future of association studies: gene-based analysis and replication. *Am J Hum Genet*; **75**: 353–362.

Nelsen RB. (1999). An introduction to copulas. *New York: Springer*.

Ochoa MC, Marti M, Azcona C, Chueca M, Oyarzábal M, Pelach R, Patiño A,Moreno-Aliaga MJ, Martínez-González MA, Martínez JA. (2004). Gene-gene interaction between PPARγ2 and ADRβ3 increases obesity risk in children and adolescents. *Int J Obes*; **28**: S37–S41.

Ruppert D. (2005). Ruppert D. Discussion of "Maximization by parts in likelihood inference". *J Am Stat Assoc*; **100**: 1161–1163.

Sanna S, Jackson AU, Nagaraja R, Willer CJ, Chen WM, Bonnycastle LL et al. (2008). Common variants in the GDF5-UQCC region are associatted with variation in human height. *Nat Genet*; **40**: 198–203.

Song P X-K, Li M and Yuan Y. (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics*; **65**: 60–68.

Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, Pirruccello JP et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*; **466**: 707–713.

The International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*; **437**: 1299–1320.

The International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*; **449**: 851–861.

Tibshirani, RJ (1996). Regression shrinkage and selection via the LASSO. *J. Roy. Statistical Society, Ser. B*; **58**: 267–288.

Tseng P, Yun S. (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Math. Program., Ser. B*; **117**: 387–423.

Wang K, Abbott D. (2007). A principal components regression approach to multilocus genetic association studies. *Genet Epidemiol*; **32**: 108–118.

Weedon MN, Lettre G, Freathy RM, Lindgren CM, Vioght BF, Perry JR et al. (2007). A common variant of HMGA2 is associated with adult and childhood height in the general population. *Nat Genet*; **39**: 1245–1250.

Wei Z, Li M, Rebbeck T, Li H. (2008). U-statistics-based tests for multiple genes in genetic association studies. *Ann Hum Genet*; **72**: 821–833.

Willer CJ, Speliotes EK, Loos RJ, Li S, Lindgren CM, Heid IM, Berndt SI et al. (2008). Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet*; **41**: 125–134.

Xu J, Langefeld CD, Zheng SL, Gillanders EM, Chang B, Isaacs SD, Williams AH, Wiley KE, Dimitrov L, Meyers DA, Walsh PC, Trent JM, Isaacs WB. (2004). Interaction effect of PTEM and CDKN1B chromosomal regions on prostate cancer linkage. *Hum Genet*; **115**: 255–262.

Zemunik T, Boban M, Lauc G et al. (2009). Genome-wide association study of biochemical traits in Korcula Island, Croatia. *Croat Med J*; **50**: 23–33.