The application of a hidden Markov random field model in genome-wide association studies

Mengyuan Chen

A Thesis submitted for the degree of

Doctor of Philosophy

of

University College London.

Department of Statistical Science
University College London

I, Mengyuan Chen, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Genome-wide association studies (GWASs) are widely used to detect single nucleotide polymorphisms (SNPs) associated with diseases. Commonly, we use hypothesis testing to identify associations. When analysing multiple SNPs, people usually use multivariate analysis methods which are usually based on individual genotype data or meta analysis methods which integrate summary statistics from single SNP analysis. However, individual genotype data are difficult to obtain due to privacy and most meta analysis methods do not consider and utilize correlations between SNPs. All these multiple SNPs analysis methods can only test whether multiple SNPs are associated with one disease and cannot identify specific SNPs associated with disease within multiple SNPs. In this thesis, we study how to leverage linkage disequilibrium (LD) information, which summarises the degree of association between different SNPs, and use summary statistics to discover SNPs associated with disease. We propose to use a hidden Markov random field model (HMRF) to model the correlation structure between SNPs and FDR control procedure to identify the association. Simulation experiments show that our method is better than other methods in terms of controlling false discovery rate and the power of discovering true associated SNPs. Then the proposed method is extended into gene association analysis. Simulation studies demonstrate that our approach outperforms other methodologies concerning the control of false discovery rate and the efficacy in detecting associated genes.

Impact Statement

A genome-wide association study is a powerful and widely method used in genetics and genomics research to identify single nucleotide polymorphisms associated with particular traits or diseases in human populations. Generally, multiple testing methods are utilized to identify associations between specific genetic variants and the trait of interest. However, when identifying associated genetic variants, few researchers consider the linkage disequilibrium, which exists commonly when two or more genetic variants are located close to each other on the same chromosome. In this thesis, we propose to use a hidden Markov random field model to leverage the linkage disequilibrium between genetic markers, which is rarely used in multiple testing problems. In addition, summary statistics are applied in our proposed method, which may be better than using individual genotype data since genotype data are usually unavailable due to the privacy. We illustrate how our proposed method can be applied to real datasets to identify genetic variants associated with the trait Bipolar disorder. We also extend the proposed method to gene association analysis, which can identify multiple associated genes with one disease. This can find more associated genetic variants or genes, which is helpful to explain a higher percentages of trait variance. By identifying genetic variants associated disease from GWAS, people can aggregate the effects of multiple genetic variants to estimate an individual's polygenic risk scores to a particular trait or disease, which can be used to predict individual disease risk, help doctors to provide clinical guidance for disease prevention and facilitate the improvement of personalized medicine.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In genetics, people usually use ”genome” to represent the complete set of genetic material present in an organism, which consists of deoxyribonucleic acid (DNA). DNA is made up four nucleotides: adenine (A), thymine (T), cytosine (C), and guanine (G), the sequence of which determines the characteristics and traits of an individual. The completion of the human genome sequence and the initiation of the International HapMap Project [1] has resulted in the development of genome-wide association studies (GWAS), which are used to identity the associated genetic markers or genes with different traits. For mapping the genes which are associated with common disease and quantitative traits, there are two categories of approaches: candidate-gene studies and genome-wide studies, which contain, respectively, linkage mapping and genome-wide association studies [2]. Genes that are close to each other on a chromosome are inclined to be passed down together during the process of genetic recombination. Based on this, linkage mapping is usually used to study the relationship between the transmission of a genetic marker and the disease or trait within families, while GWAS mainly focuses on identifying the association between genetic markers and traits within a large population. In GWAS, people detect genetic variants associated with diseases, by genotyping single nucleotide polymorphisms (SNPs) in diseased and normal individuals. Compared with traditional candidate-gene studies, GWAS does not need prior information about a gene’s biological functional impact on the disease. In addition, for identifying genetic variants associated with disease with modest effect sizes, GWASs are more powerful than linkage mapping [3].

For complex diseases, such as prostate and breast cancers [4] and type 2 diabetes [5], GWASs have been shown to be an effective method to detect associated genetic variants. In 2007, the Wellcome Trust Case Control Consortium (WTCCC) [6] carried out a GWAS of 7 diseases using 14000 cases and 3000 controls, which started the era of large-scale GWAS [7]. Because of new genotyping technologies, the cost of large GWAS has become lower and more and more large-scale GWASs have been carried out.

Although recently GWAS has discovered many genetic variants associated with complex diseases, these findings can only explain a small fraction of phenotype variance. For example, it is known that height is a phenotype with high heritability, which could explain about 80% of phenotype variance. However, 40 variants identified by GWAS by 2010 can only explain approximately 5% of height variance [8]. In general, there exists a wide gap between the estimates of heritability and the percentage of phenotype variance explained by GWAS, which is usually called "missing heritability". There are many possible reasons for missing heritability. For example, there are many genetic variants with small effects, which are hard to detect. It is common for current GWAS to involve sample sizes of 2000 to 5000, which can detect common variants (minor allele frequency (MAF)$>$5%) with odds ratios larger than 1.5 with 80% power [9], where MAF is the frequency of the second most common allele in a given population. See Section 2.1 for more information. However, for variants with odds ratios of 1.1, a sample size of 60000 is needed to detect them with sufficient power, which will lead to large cost. What is more, for variants with low MAF (1%$<$MAF$<$5%) or rare variants (MAF$<$1%), current GWA genotyping technologies struggle to capture them. However, sometimes these variants may have large effects on complex disease, which is the common disease/rare variant hypothesis [10]. For instance, 20 variants with an allele frequency of 1% and an odds ratio of three may explain most familial aggregation of type 2 diabetes. To solve the problem of missing heritability, it is necessary to develop new statistical methods to detect more associated genetic variants with moderate sample sizes.

1.2 Contributions

The first contribution in this thesis is developing a new method to deal with the dependence among multiple testing, which is usually ignored in classical multiple hypothesis testing problem. Our study is motivated by a genome-wide association study, which aims to identify the SNPs which are associated with disease. A hidden Markov random field model, which is usually used on image segmentation problems, is proposed to leverage the dependence between different SNPs. The proposed method outperforms the other methods when identifying the SNPs which are associated with disease in GWAS in several aspects, such as controlling false discovery rate at a predefined level and improving the power of multiple testing by leveraging correlation between SNPs.

The second contribution is in the area of GWAS application. Our proposed method can be used in identifying specific SNPs or genes which are associated with disease. Only summary statistics are required in our model, which are easy to be obtained. Furthermore, by detecting more SNPs or genes associated with disease, we can better estimate individual risk scores associated with a particular trait or disease, which is helpful for the development of disease prevention and personalized medicine.

1.3 Outline of this thesis

The focus of this thesis is to investigate new statistical methods to improve power in genome-wide association studies. The rest of this thesis is organised as follows:

- Chapter 2 introduces some concepts and terminology used in GWAS and conducts a literature review in GWAS.

- Chapter 3 introduces a hidden Markov random field model and describes how to utilize a hidden Markov random field model to leverage dependence information between SNPs.

- Chapter 4 presents a simulation study to compare the performance of the proposed method with other methods and discusses the effect of different choices of threshold values and how the proposed method perform when hypothesis is violated.

- Chapter 5 shows the application of proposed method on real dataset.

- Chapter 6 introduces an extension of the proposed method to gene association and studies its performance using a simulation study and real datasets.

- Chapter 7 discusses limitations of the proposed method and future work.

Chapter 2

Genome-wide association study

This chapter is organised as follows. In Section 2.1, we introduce some concepts, and terminology used in GWAS and describe what the data look like. In Section 2.2, related methods of association analysis in GWAS are provided.

2.1 Preliminaries

2.1.1 Single nucleotide polymorphism (SNP)

Normally, human cells have 23 pairs of chromosomes, which contains 22 pairs of autosomal chromosomes. Each chromosome is a deoxyribonucleic acid (DNA) molecule, which is a long string of paired nucleotides that have four different types: A, T, C, and G. Most of the genetic information stored in DNA is the same for all human beings. The remaining small percentage of genetic variants make each individual unique.

A single nucleotide polymorphism (SNP) represents a single nucleotide difference among individuals. For example, if a specific base position is nucleotide C for many genomes in the population, but for some genomes, this position is occupied by an T, then this position is a SNP with alleles C and T. SNPs are the most prevalent form of genetic variation (there are at least 11 million common SNPs [11]), which makes them good genetic markers for genetic mapping. The main aim of GWAS is to determine whether a SNP is associated with a disease or trait.

Figure 2.1 illustrates SNPs from the human genome. For example, alleles G and C appear at SNP1. The allele with higher frequency among the population is called "the major allele", while the other one is called "the minor allele". For instance, the major allele and the minor allele for SNP1 are G and C, respectively. In genetics, a set of SNP alleles at different locations of a gene or DNA sequence which are inherited together is termed as a haplotype. For example, under the assumption that the five SNPs in Figure 2.1 are inherited together, we have two haplotypes for the 1st individual: GCGTT and GAAGC.

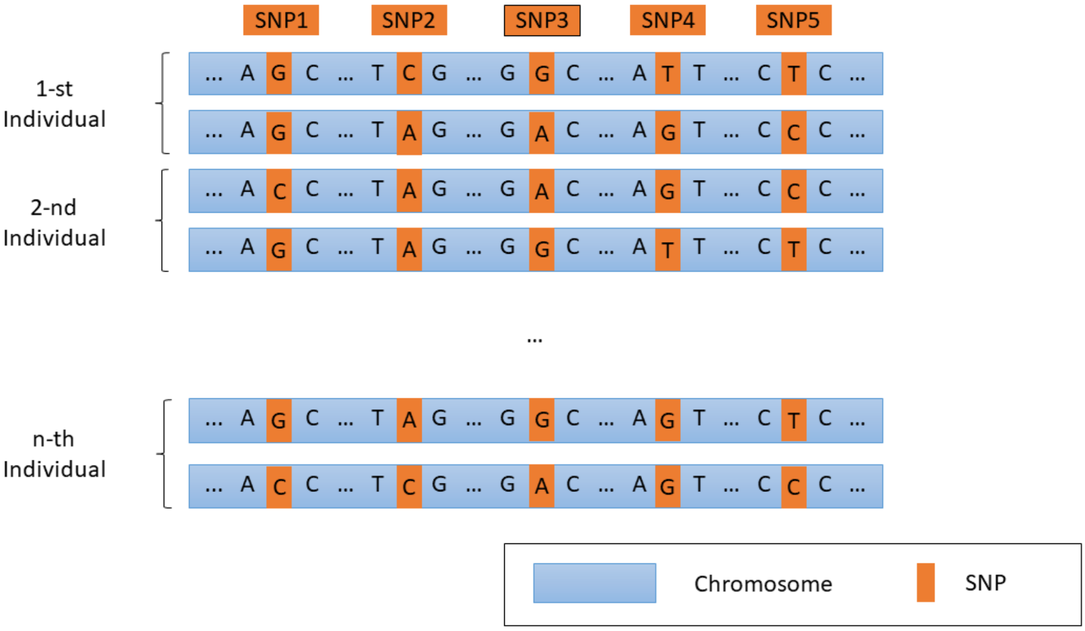

SNP1
SNP2
SNP3
SNP4
SNP5
1-st Individual
... A G C ... T C G ... G G C ... A T T ... C T C ...
... A G C ... T A G ... G A C ... A G T ... C C C ...
2-nd Individual
... A C C ... T A G ... G A C ... A G T ... C C C ...
... A G C ... T A G ... G G C ... A T T ... C T C ...
...
n-th Individual
... A G C ... T A G ... G G C ... A G T ... C T C ...
... A C C ... T C G ... G A C ... A G T ... C C C ...
Chromosome
SNP

Figure 2.1: An illustration of SNPs and their alleles.

For current GWAS, we can only observe the genotype data of each SNP instead of haplotypes of each individual. Let us take the 1st individual in Figure 2.1 as an example. The SNP data set in Figure 2.1 contains genotype information of five SNPs as follows:

$$\begin{array}{cccccc} \text{Sample ID} & \text{SNP1} & \text{SNP2} & \text{SNP3} & \text{SNP4} & \text{SNP5} \\ 1 & GG & CA & GA & TG & TC \end{array} \tag{2.1}$$

In a real application, the above SNP genotypes are usually encoded as 0, 1 or 2, dependent on the number of copies of variant alleles. For instance, suppose a SNP has alleles G (reference) and C (variant), then

- genotype GG is coded as 0 (homozygous reference);
- genotype GC is coded as 1 (heterozygous);
- genotype CC is coded as 2 (homozygous variant).

Usually, the reference allele is the major one while the variant is the minor one. The frequency of this minor allele in a population is called the minor allele frequency (MAF).

2.1.2 Hardy-Weinberg equilibrium

The Hardy-Weinberg equilibrium (HWE) describes a probabilistic relationship between allele frequencies and genotype frequencies, which states that the genetic variation in a population will remain constant from one generation to the next when a set of assumptions are satisfied. These assumptions include random mating, large population size, no immigration or emigration, no mutations and no natural selection, which means all genotypes have an equal chance of surviving and reproducing.

Under these assumptions, a population is not evolving and in Hardy-Weinberg equilibrium, which means that the frequencies of alleles and genotypes in the population remain constant. Let p be the frequency of allele A and q be the frequency of the allele a, and AA, Aa, and aa be the frequencies of three possible genotypes in the population. Under HWE, the following relationship holds:

$$\begin{aligned} &p+q=1, \\ &p^2+2pq+q^2=1. \end{aligned} \tag{2.2}$$

In practice, Pearson's chi-squared test is usually applied to check if HWE is satisfied for each allele, which compares the observed genotype frequencies obtained from the data and the expected genotype frequencies obtained using the equation (2.2). Since many genetic association studies are based on the assumption of HWE, it is necessary to conduct HWE test for every SNP before analysing data.

2.1.3 Linkage disequilibrium

In genetics, linkage disequilibrium (LD) is used to describe the dependency relationship between different SNPs. For a combination of alleles of two SNPs, LD is defined as the difference between observed frequency and expected frequency.

Now we illustrate how to measure LD. Consider two SNPs and their alleles (A,a) and (B,b), respectively. Denote the allele frequency of A as p_A and the allele frequency of B as p_B. Then the level of linkage disequilibrium between two SNPs can be quantified by D. Its definition is as follows:

$$D = P(AB) - p_A p_B. \tag{2.3}$$

When $D = 0$, it means that these two SNPs are independent of each other. Otherwise, there is correlation between them. Since the value of D depends on the frequencies of the alleles, it is difficult to use D to compare the level of linkage disequilibrium between different pairs of alleles. So LD is usually measured by normalizing D as follows:

$$r = \frac{D}{\sqrt{p_A(1-p_A)p_B(1-p_B)}}. \tag{2.4}$$

When the major and minor alleles are encoded using 0 and 1 as we state in Section 2.1.1, r is equivalent to the Pearson correlation coefficient between two SNPs [12].

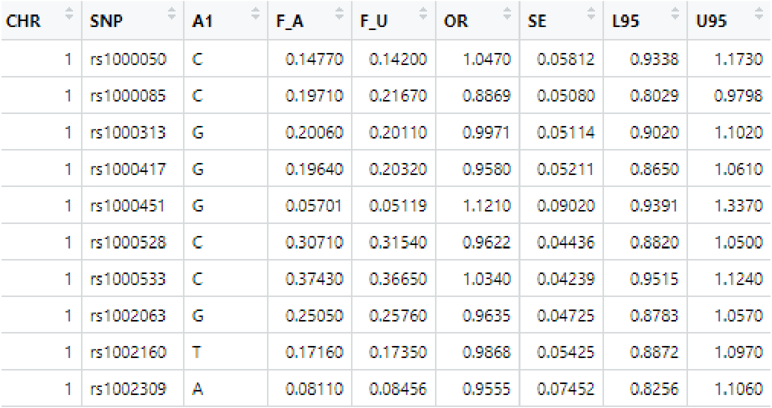
2.1.4 LD matrix calculation

In general, the LD matrix is calculated from a reference panel with individual genotype data such as the 1000 Genomes Project [13]. In this study, we use the European population genotype data from 1000 Genomes Project phase3 as the reference panel, which contains 503 individuals. Usually, the PLINK1.9 software [14] is used to select the genotype data of corresponding SNPs in our study and VCFtools [15] is used to recode the genotype data as 0,1, or 2, which is dependent on the number of copies of variant alleles. Then the LD matrix can be calculated based on the genotype data of 503 individuals.

2.1.5 Data Description

A genome-wide association study is an example of a large-scale hypothesis testing problem, which considers hundreds or thousands of test statistics at once. Before describing multiple testing problems, we describe the data format firstly in this part and use a plot to show the effect of correlation on the null distribution of Z values.

In this section, a GWAS dataset that will be analysed in Chapter 5 is introduced . We give an overview of the structure of the data in this section and discuss issues concerned with multiple testing in Section 2.1.6. The data come from the Wellcome Trust Case Control Consortium (WTCCC) [16]. They conducted the genome-wide association studies of 2000 cases and 3000 shared controls for several complex human diseases: bipolar disorder (BD), coronary artery disease (CAD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D) [16]. The format of summary statistics are illustrated in Figure 2.2.



CHR	SNP	A1	F_A	F_U	OR	SE	L95	U95
1	rs1000050	C	0.14770	0.14200	1.0470	0.05812	0.9338	1.1730
1	rs1000085	C	0.19710	0.21670	0.8869	0.05080	0.8029	0.9798
1	rs1000313	G	0.20060	0.20110	0.9971	0.05114	0.9020	1.1020
1	rs1000417	G	0.19640	0.20320	0.9580	0.05211	0.8650	1.0610
1	rs1000451	G	0.05701	0.05119	1.1210	0.09020	0.9391	1.3370
1	rs1000528	C	0.30710	0.31540	0.9622	0.04436	0.8820	1.0500
1	rs1000533	C	0.37430	0.36650	1.0340	0.04239	0.9515	1.1240
1	rs1002063	G	0.25050	0.25760	0.9635	0.04725	0.8783	1.0570
1	rs1002160	T	0.17160	0.17350	0.9868	0.05425	0.8872	1.0970
1	rs1002309	A	0.08110	0.08456	0.9555	0.07452	0.8256	1.1060

Figure 2.2: This is a small subset of the sample data from whole data of disease bipolar disorder. The CHR means Chromosome number, SNP is the SNP ID. A1 represents the minor allele. F_A is the frequency of this allele in cases, while F_U is the frequency of this allele in controls. OR represents the estimated odds ratio for A1 when the other allele is the reference allele. SE is the standard error for $\log(\text{OR})$. L95 and U95 are the lower bound and upper bound of a 95% confidence interval for this odds ratio, respectively. The Z values we used in our study can be calculated using $\log(\text{OR})/\text{SE}$ (see Section 2.2.1 for details).

The aim of a genome-wide association study is to identify those SNPs which are associated with disease. In the context of a null distribution of Z values based on the assumption of no association, SNPs with odds ratios that are unusual, that is, significantly different from 1, will be regarded as significant discoveries. The Z

values we used in our study can be calculated using $\log(\mathrm{OR})/\mathrm{SE}$ (see Section 2.2.1 for details).

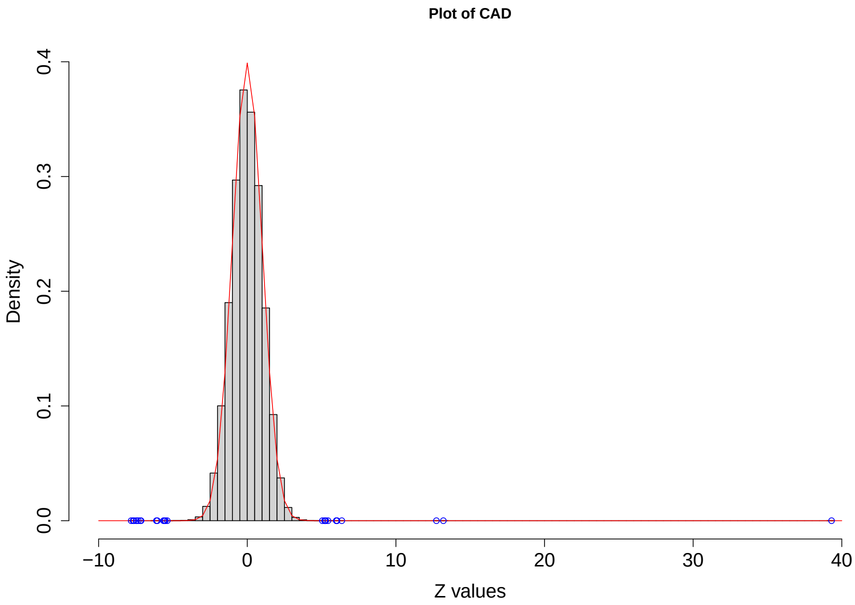


Figure 2.3: The histogram of Z values from diseases CAD. The red curve represents the theoretical null distribution, which is N(0,1). The blue circles on heavy tail have large or small Z values, which may be non null cases and associated with disease. These are what we want to find.

Figure 2.3 shows the histogram of Z values. The blue circles, which have unusual Z values, are the SNPs we are interested in. In Figure 2.3, when blue circles are identified as significant discoveries, the standard normal distribution N(0,1) is assumed as the null distribution of Z values. In genome-wide association studies, the different SNPs are not independent and there exists linkage disequilibrium, which makes the Z values not independent. The correlation between Z values will have twofold effect for multiple testing [17, 18]:

1. Correlation will have effect on the null distribution of Z values. In general, the Z values are assumed to follow standard normal distribution under the null hypothesis, which is called the theoretical null distribution. However, correlation may make the null distribution widen or narrow.

2. Correlation may affect the number of SNPs that are reported as non-null when conducting simultaneous testing.

To illustrate how different SNPs are correlated with each other, the correlation r (see equation (2.4)) between a small sample of 1000 SNPs from disease bipolar disorder is plotted in Figure 2.4.

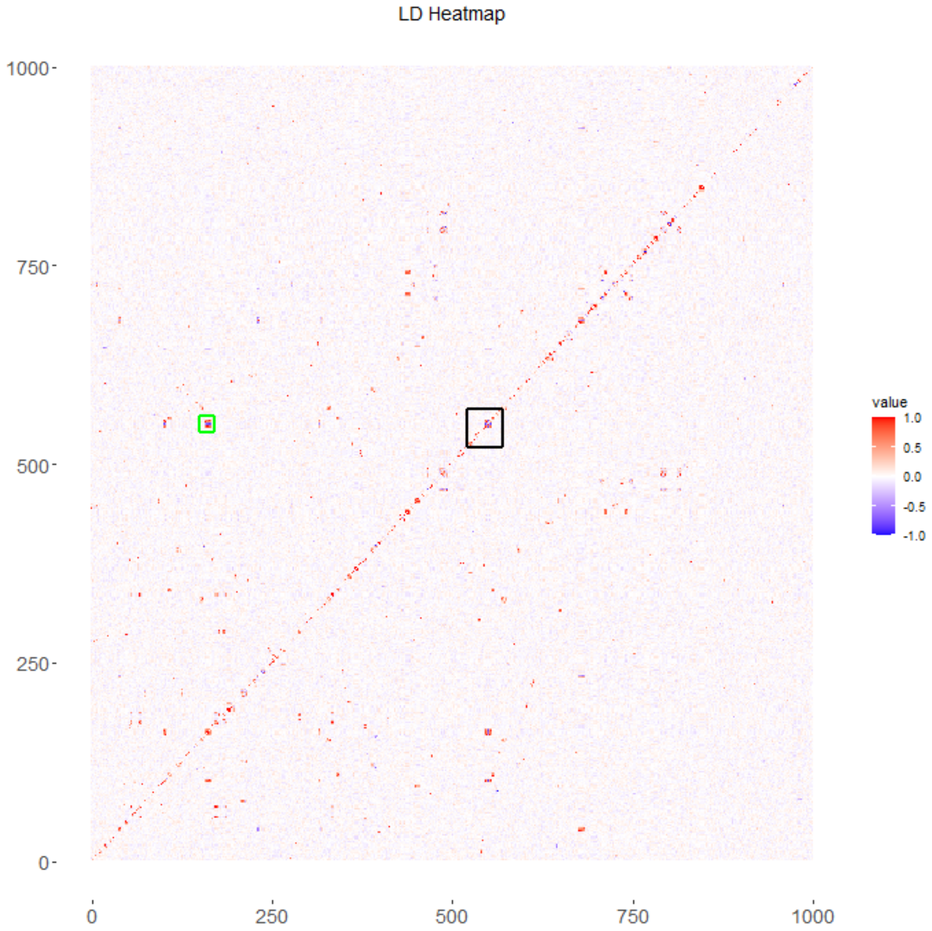


Figure 2.4: The correlation between 1000 SNPs. The blue points represent that the correlation between SNPs is close to -1, while the red points represent that the correlation is close to 1. The black square indicates 50 SNPs with numbers 521 to 570, which will be analysed in Figure 2.5. The green square indicates correlation structure at (150:170,540:560), which will be analysed in Figure 2.6.

Figure 2.4 shows the correlation between 1000 SNPs, which is from rs2554622 to rs2920090 on chromosome 8. It can be seen that for most of the pairs of SNPs, the correlation is close to 0, because most of the plot has colors close to the white, which represents a value of 0. When a correlation is not close to 0, more correlation of the pairs of SNPs are positive since more colors are close to red rather than blue. Also, when points close to diagonal line from the lower left to upper right, they have red or blue colors, which means SNPs tend to have stronger correlation when they are close to each other. Figure 2.4 suggests that there is some local structure in the correlations for groups of neighbouring SNPs. To make the correlation structure clearer, one black square area is selected and zoomed in. The black square indicates 50 SNPs with numbers 521 to 570, and the correlation between these 50 SNPs is displayed in Figure 2.5.

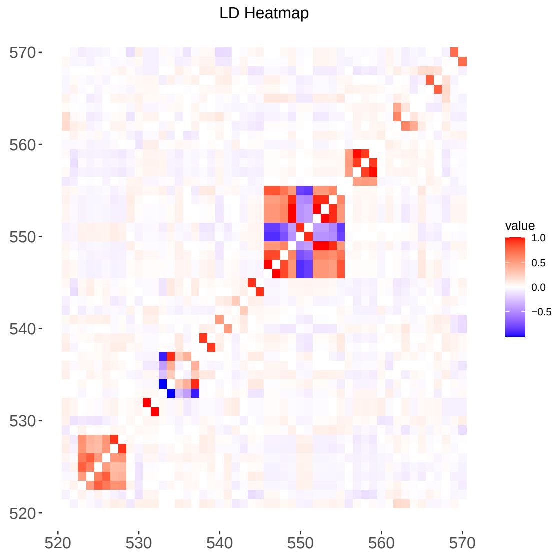


Figure 2.5: The correlation between 50 SNPs.

It can be seen from Figure 2.5 that the diagonal line is white, since the diagonal values are set as 0. For some SNPs such as SNPs between 545 and 555, they have both positive and negative correlation with other nearby SNPs. Especially, SNPs 550 and 551 are strongly positively correlated while these SNPs are negatively correlated with the SNPs that are near them. The group of SNPs from 546 to 549 and another group of SNPs between 552 and 555 are positively correlated with each other, to varying degrees. If we cluster these different groups of correlated SNPs as haplotypes, it can be seen that 2 haplotypes are correlated. Here we focus on observing the correlations between SNPs since our analysis in this thesis is based on SNPs. For SNPs between 523 and 528, they are positively correlated with nearby SNPs, while there are groups of SNPs such as those SNPs 528 to 532 with correlations close to 0.

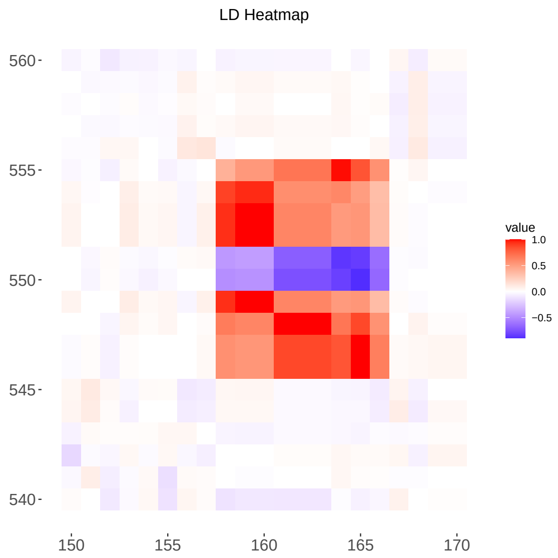
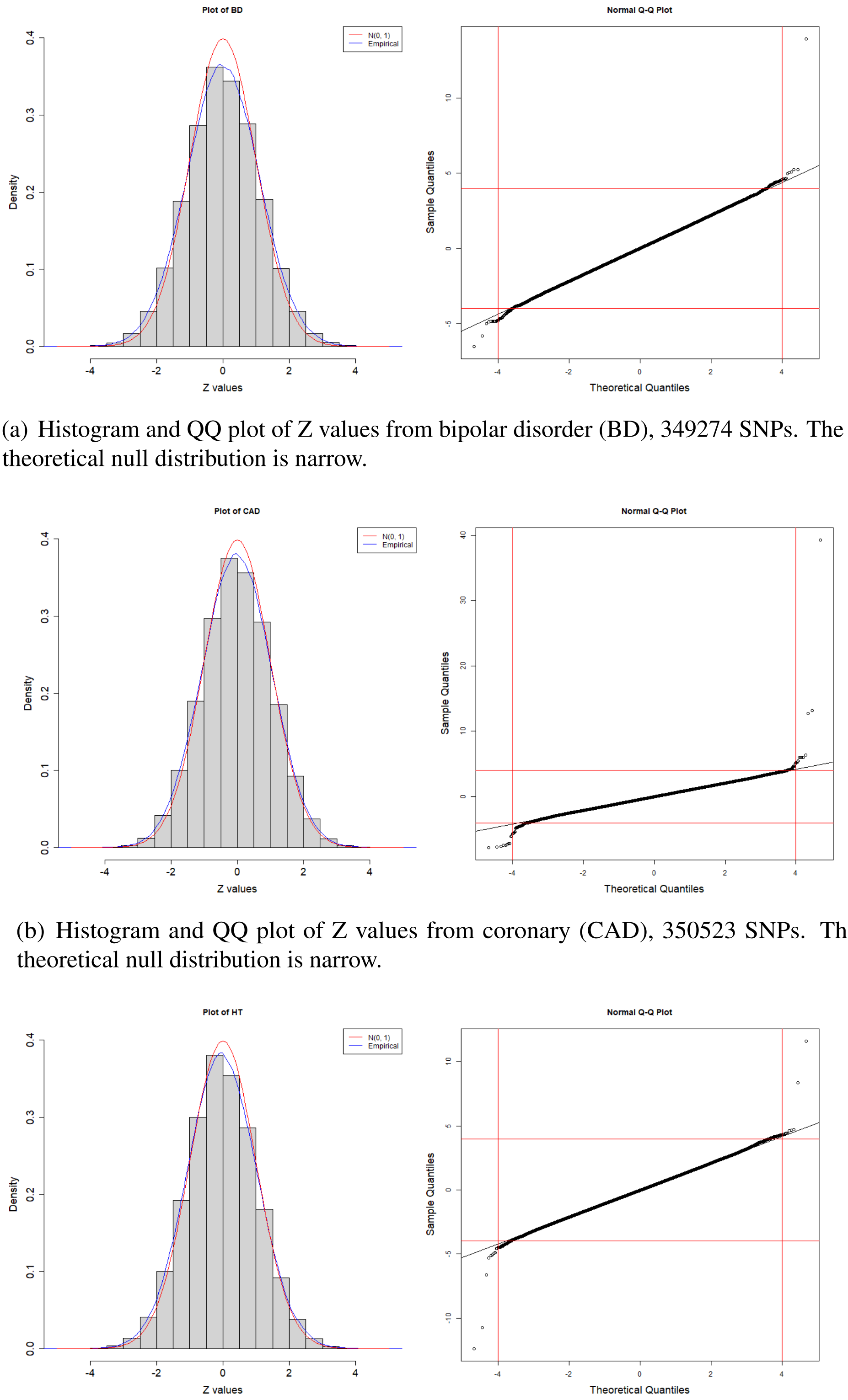


Figure 2.6: The correlation structure between SNPs at (150:170,540:560), which is the off-diagonal green square in Figure 2.4.

Figure 2.6 shows that except SNPs close to each other in Figure 2.5, there are some strong correlations between sets of SNPs that are not located closely together. For example, SNPs with numbers 158 to 166 have positive correlations with SNPs from 546 to 549 and groups of SNPs from 552 to 556, while they are negatively correlated with SNPs 550 and 551. Figures 2.4 to 2.6 show that there are some strong positive and negative associations between SNPs and that while many of these strong associations are between SNPs that are located close to each other on the chromosome, there are also some strong associations between distant pairs of SNPs.

To explore the distribution of Z values for these data, plots of Z values for several diseases are presented in Figures 2.7 and 2.8. The plots on the left are histograms of Z values with the $N(0,1)$ density function superimposed in red and a kernel density estimate of the empirical density function superimposed in blue. The horizontal axis of these plots is constrained to $(-4,4)$ to focus on a comparison of the $N(0,1)$ probability density. The plots on the right are normal QQ plots of the Z values, with horizontal and vertical red lines superimposed at -4 and 4 to make comparison with the plots on the left, but the QQ plots also show instances where Z have magnitudes that are very much larger than expected under a $N(0,1)$ null distribution.

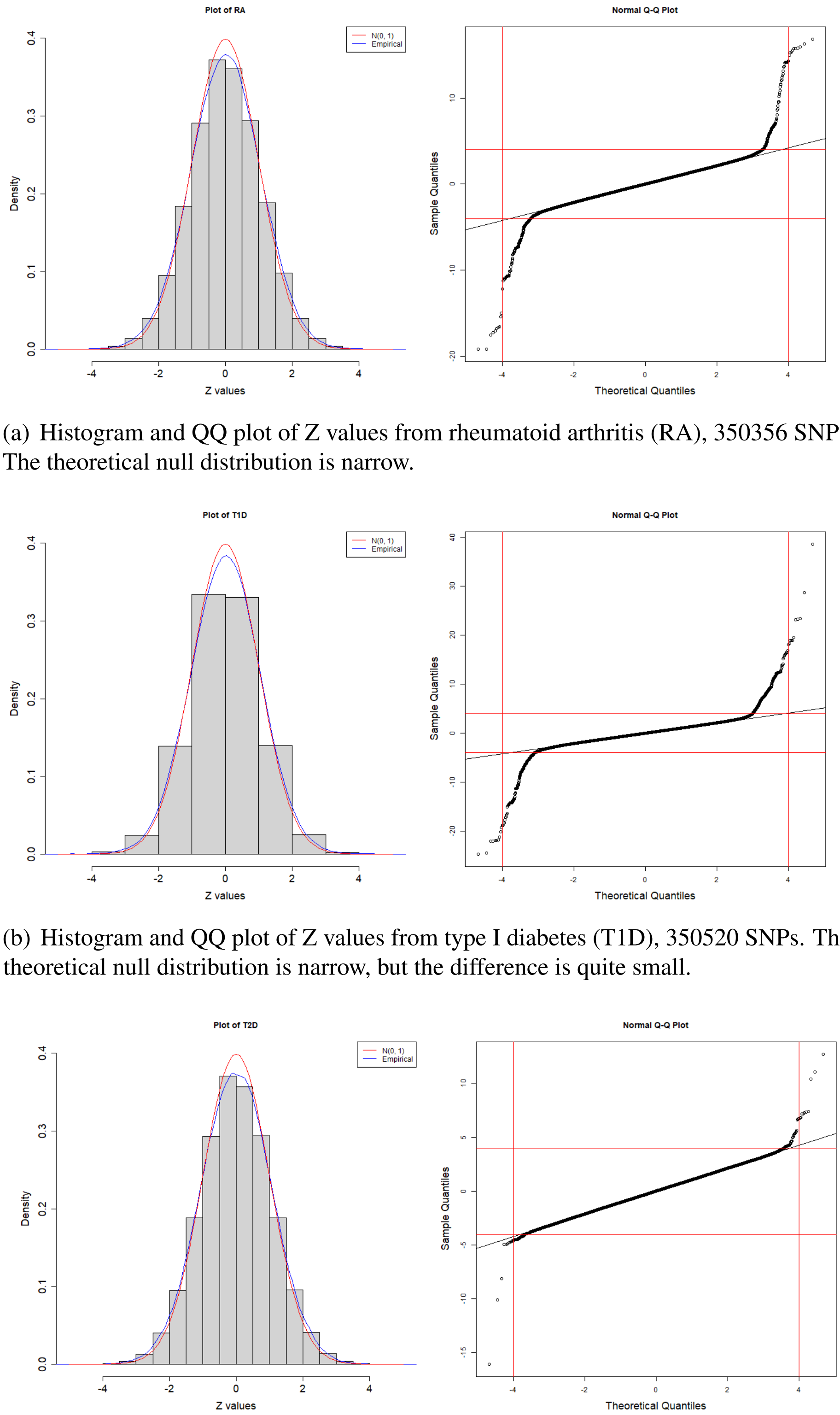


(a) Histogram and QQ plot of Z values from bipolar disorder (BD), 349274 SNPs. The theoretical null distribution is narrow.

(b) Histogram and QQ plot of Z values from coronary (CAD), 350523 SNPs. The theoretical null distribution is narrow.

(c) Histogram and QQ plot of Z values from hypertension (HT), 350271 SNPs. The theoretical null distribution is narrow.

Figure 2.7: The histogram and QQ plots of Z values from diseases BD, CAD, HT. The red line in histogram represents the theoretical distribution, while the blue line in histogram is the empirical distribution.The red lines in QQ plot represent the lines -4 and 4, which are consistent with the $(-4,4)$ scale in histogram,



(a) Histogram and QQ plot of Z values from rheumatoid arthritis (RA), 350356 SNPs. The theoretical null distribution is narrow.

(b) Histogram and QQ plot of Z values from type I diabetes (T1D), 350520 SNPs. The theoretical null distribution is narrow, but the difference is quite small.

(c) Histogram and QQ plot of Z values from type 2 diabetes (T2D), 349767 SNPs. The theoretical null distribution is narrow.

Figure 2.8: The histogram of Z values from diseases RA, T1D and T2D. The red line in histogram represents the theoretical distribution, while the blue line in histogram is the empirical distribution.The red lines in QQ plot represent the lines -4 and 4, which are consistent with the $(-4,4)$ scale in histogram,

It can be seen from the histograms in Figure 2.7 and Figure 2.8 that the empirical distributions have slightly heavier tails than the theoretical null distribution. This may be caused by the correlation between Z values or just reflect the influence of the Z values with unusually large magnitudes seen in the QQ plots. However, the difference is small, especially for disease type I diabetes in Figure 2.8 (b). When the empirical distribution appears to be very different from the theoretical null distribution, it may be not appropriate to use the theoretical null distribution in the testing procedure. For the QQ plots, it can be seen from Figures 2.7 and 2.8 that most of the points lie on the straight lines, especially for middle bulk of the Z values, which suggests that many Z values are sampled from a $N(0,1)$ distribution, while some Z values have unusually large magnitudes. These unusual large Z values are what we are interested in.

For the second effect of correlation, Efron [18] used simulation experiments to show that correlation effect may cause misleading estimates of false discovery rate (FDR), which is usually used to decide the non-null cases in multiple testing procedure (see details in Section 2.1.6.2). In their experiments, correlation may make the estimates of FDR decline as the actual false discovery proportion increases, which will cause false discoveries. So considering correlation in a multiple testing procedure can be important.

2.1.6 Multiple testing

In this part, some concepts in multiple testing problems and several correction methods are introduced, which will be useful in comparing performance during experiments.

2.1.6.1 Error Criteria

Let us consider testing m hypotheses simultaneously, which include m_0 true hypotheses and $m_1 = m - m_0$ false hypotheses. The possible outcomes are summarised in Table 2.1.

	H_0 is true	H_0 is false	Total
H_0 is rejected	V	S	R
H_0 is not rejected	$m_0 - V$	$m_1 - S$	$m - R$
Total	m_0	m_1	m

Table 2.1: The number of hypotheses in each category when testing m hypotheses.

We call V the number of false positives (Type I error), and S the number of true positives. $m_1 - S$ is the number of false negatives (Type II error) and $m_0 - V$ is the number of true negatives.

The following describes a multiple testing problem. The more hypotheses we consider, the higher the probability of getting at least one false positive result (Type I errors) [19]. For example, if assuming that the $m = 100$ hypotheses are independent, and the probability of a false positive for each test is 0.05, then the probability of getting at least 1 false positive for 100 hypotheses is $1-(1-0.05)^{100} \approx 0.994$. For a single-marker test, the hypothesis test is conducted for each SNP. Usually, there are a large number of SNPs in a GWAS. So to avoid a multiple testing problem, it is necessary to correct the significance level by considering the multiplicity of simultaneously testing all SNPs.

The classic method for multiple testing correction is to use the family-wise error rate (FWER) [20] to control the type I error of multiple hypotheses. FWER is the probability of making at least one false discovery: $\text{FWER} = P(V \geq 1)$. The most conservative correction to control FWER is the Bonferroni correction [21], which assumes that all tests are independent and simply uses $\alpha^* = \alpha/m$ as the significance level for a single test. Here α is the desired overall significance level for m hypotheses, and m is the number of hypotheses. Another more powerful method is called Holm-Bonferroni method [22], which is a modification of the Bonferroni correction. The procedure is as follows: (1) Order the p-values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$, and the corresponding hypotheses are $H_{(1)}, \ldots, H_{(m)}$; (2) For a given significance level α, let k be the minimal index such that $p_{(k)} \geq \frac{\alpha}{m+1-k}$; (3) Reject the null hypotheses $H_{(1)}, \ldots, H_{(k-1)}$.

FWER control is a stringent statistical method used to minimize the risk of making any false positives in a family of tests. However, in large-scale multiple testing problem such as GWAS, it is quite common to have false rejections due to the large number of statistical tests performed and FWER is too conservative to find interesting results. Under these situations, a more proper approach is to control the false discovery rate (FDR) [23]. The FDR is defined as the expected proportion of false positives among the set of rejected hypotheses: FDR $= E(V/(R \vee 1))$, where $R \vee 1$ means $\max(R, 1)$, and FDR $= 0$ when $R = 0$. Compared with controlling the FWER, controlling the FDR provides a less stringent significance level, thus achieving a higher power. The most common method to control the FDR is the Benjamini and Hochberg (BH) procedure [23]. This procedure is as follows: (1) Order the p-values: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$; (2) Given the desired FDR level q, the index k is computed as follows:

$$k = \max\left\{i : p_{(i)} \leq \frac{iq}{m}\right\}. \tag{2.5}$$

If there is no such k, none of the hypotheses is rejected. Otherwise H_i $(i = 1, \ldots, k)$ are rejected.

2.1.6.2 Local False Discovery Rate Methods

Efron and others [24] first proposed the local false discovery rate methods. Consider that m hypothesis tests are conducted simultaneously and the corresponding test statistics are $Z_1, \ldots, Z_m$. Denote the probability of the null hypothesis as $p_0 = P(H_0)$, while the probability of the alternative hypothesis is set as $p_1 = P(H_1)$. Given a test statistic $Z = z$, the local false discovery rate (lfdr) is defined as the probability that a hypothesis is null:

$$lfdr(z) = P(H_0 \mid z) = \frac{p_0 f_0(z)}{p_0 f_0(z) + p_1 f_1(z)}, \tag{2.6}$$

where f_0 denotes the probability density function (PDF) of test statistics when the null hypothesis is true, while f_1 is the PDF of test statistics when alternative hypothesis is true. If F_0 and F_1 are used to represent the corresponding cumulative distribution functions (CDFs), then FDR can be defined as the posterior probability of a null hypothesis given that the test statistic Z is less or equal to some value z:

$$FDR(z) = P(H_0 \mid Z \leq z) = \frac{p_0 F_0(z)}{p_0 F_0(z) + p_1 F_1(z)}. \tag{2.7}$$

Efron et al. [25] showed the connections between FDR and lfdr. It can be seen from equations (2.6) and (2.7) that the local false discovery rate depends on densities, while false discovery rate is based on CDFs. $FDR(z)$ can be seen as the average of $lfdr(Z)$ for all $Z \leq z$ [25]. lfdr can be also used to make decisions like P values, since lfdr provides the probability of false discovery for each hypothesis. The smaller lfdr one hypothesis has, the smaller probability of making a type I error when rejecting this hypothesis.

2.1.6.3 Local Significance Index Methods

Local false discovery methods only consider individual test statistic when making decisions, which may lose information especially when test statistics are not independent. The local significance index method is a generalization of the local false discovery rate method that considers all test statistics to make decisions, which can incorporate dependence information when test statistics are correlated. The local index of significance (LIS) [26] is defined as:

$$LIS_i = P_\Theta(H_i \text{ is null} \mid \text{all the observations at } m \text{ hypotheses}), \tag{2.8}$$

where Θ are the parameters which are used to specify the dependence structure of the m hypotheses. Sun and Cai [26] assumed the dependence between test statistics are from a chain structure across the gene and used hidden Markov models to leverage the dependence information. Then they made inference for parameters using a forward-backward procedure and proposed the decision rule as $\delta = I(LIS_i < \lambda),\ i = 1, \ldots, m$, where λ is a threshold. If R_λ, V_λ and $Q(\lambda)$ are used to denote the number of rejections, the number of false rejections and the marginal false discovery rate, respectively, these values can be calculated as follows:

$$\begin{aligned} R_\lambda &= \sum_{i=1}^{m} I(LIS_i < \lambda) \\ V_\lambda &= \sum_{i=1}^{m} I(LIS_i < \lambda, H_i \text{ is null}) \\ Q(\lambda) &= \frac{E(V_\lambda)}{E(R_\lambda)}. \end{aligned} \tag{2.9}$$

Sun and Cai [26] showed that LIS could be applied to approximate the marginal false discovery rate:

$$\hat{Q}(k) = \frac{1}{k}\sum_{i=1}^{k} I(LIS_{(i)}), \tag{2.10}$$

where k is the number of rejected hypotheses.

2.1.7 Multiple testing for grouped hypotheses

For grouped hypotheses, it is more difficult to control the rate of false discoveries for multiple testing. There are several people who have developed current multiple testing methods for grouped hypotheses. For reference, Sun and Cai [27] proposed two strategies. One is called pooled FDR analysis [28], which ignores the group labels and calculates the pooled lfdr statistic for all hypotheses. If there are m hypotheses, which can be divided in to K groups, then the pooled lfdr statistic (PLfdr) is defined by

$$\text{PLfdr}(Z_i) = P(H_0|Z) = \frac{p_0 f_0(Z_i)}{p_0 f_0(Z_i) + p_1 f_1(Z_i)}, \quad i = 1, \ldots, m, \tag{2.11}$$

where Z_i denotes the test statistic and $p_0 = P(H_0)$. f_0 denotes the PDF of test statistics when the null hypothesis is true, while f_1 is the PDF of test statistics when alternative hypothesis is true. Then let $\text{PLfdr}_{(1)}, \ldots, \text{PLfdr}_{(m)}$ be the ranked PLfdr values and $H_{(1)}, \ldots, H_{(m)}$ be the corresponding hypotheses. The pooled FDR analysis procedure is to reject all $H_{(i)}$ for $i = 1, \ldots, l$ where

$$l = \max\left\{ i : (1/i)\sum_{j=1}^{i} \text{PLfdr}_{(j)} \leq \alpha \right\}, \quad i = 1, \ldots, l. \tag{2.12}$$

The pooled FDR analysis is similar as the FDR analysis for single group hypotheses, since it ignores the group labels and the equation (2.12) is the same as equation (2.5).

Another method is separate FDR analysis, which analyses each group separately at the same FDR level and then combines the decision results together. Sun and Cai [27] defined the conditional lfdr for group k as

$$\mathrm{CLfdr}^k(Z_{ki}) = P(H_{k0}|Z) = \frac{p_{k0}f_{k0}(Z_{ki})}{p_{k0}f_{k0}(Z_{ki}) + p_{k1}f_{k1}(Z_{ki})}, \quad i = 1,\ldots,m_k;\ k = 1,\ldots,K, \tag{2.13}$$

where m_k represents the number of hypotheses in group k. Then for the separated FDR analysis, let $\mathrm{CLfdr}^k_{(1)},\ldots,\mathrm{CLfdr}^k_{(m)}$ be the ranked CLfdr values in group k and $H^k_{(1)},\ldots,H^k_{(m_k)}$ be the corresponding hypotheses. The separate FDR procedure for group k is to reject all $H^k_{(i)}$ for $i = 1,\ldots,l_k$ where

$$l_k = \max\left\{ i : (1/i)\sum_{j=1}^{i} \mathrm{CLfdr}^k_{(j)} \leq \alpha \right\}. \tag{2.14}$$

Then the final set of rejection hypotheses for separate FDR analysis is equal to $\bigcup_{k=1}^{K}\{H^k_{(i)} : i = 1,\ldots,l_k\}$, which combined the K rejection sets together.

In 2009, Cai and Sun [28] combined the above two methods and proposed to calculate the lfdr values for separate groups based on equation (2.13). Then they combined these lfdr values from all groups together to make decisions as in equation (2.12). Benjamini and Cohen [29] developed a hierarchical weighted false discovery rate (FDR) method to control FDR with equal weights. Basu et al. [30] also developed a weighted FDR method, but they used some predefined weights. For multiple groups, they estimated the local false discovery rate (lfdr) for separate group and used the weighted FDR method for multiple groups. Zhao and Zhang [31] used weighted p-value procedures to control FDR, where the weights were estimated by maximizing a power-related objective function. Hu et al. [32] proposed a group BH procedure, which weighted p-values for each group and pooled all weighted

p-values together. Then a BH procedure was used to make a decision. The weights for each group were estimated from the proportion of true null hypotheses in each group.

Liu et al. [33] proposed a framework to control within-group false discoveries while controlling false discoveries from all hypotheses from a Bayesian viewpoint, which is under the assumption of one-way classified hypotheses. They expressed the hypotheses as follows: $\theta_{ki} = \theta_k \times \theta_{i|k}$, where $k = 1, \ldots, K$ represent the group and $i = 1, \ldots, m_k$ represent the index within the group. So it can be seen that if $\theta_k = 0$, then all $\theta_{ki} = 0$. If $\theta_k = 1$, then $\theta_{ki} = 0$ or 1. Then they defined

$$\mathrm{fdr}_k(Z) = P(\theta_k = 0|Z), \tag{2.15}$$

and

$$\mathrm{fdr}_{i|k}(Z) = P(\theta_{i|k} = 0|\theta_k = 1, Z), \tag{2.16}$$

where Z represents the observed data. Their proposed procedures are as follows:

1. For each group k, let $\mathrm{fdr}_{(1)|k}, \ldots, \mathrm{fdr}_{(m_k)|k}$ be the ranked $\mathrm{fdr}_{i|k}$ values and $H_{k(1)}, \ldots, H_{k(m_k)}$ be the corresponding hypotheses. Then they rejected all $H_{k(i)}$, $i = 1, \ldots, R_k$, where

$$R_k = \max\left\{ l_k : \frac{1}{l_k} \sum_{i=1}^{l_k} \mathrm{fdr}_{(i)|k} \leq \eta \right\}, \tag{2.17}$$

   where $0 < \eta \leq \alpha$, and α is the significance level. They chose $\eta = \alpha$.

2. Calculate $\eta_k = \frac{1}{R_k} \sum\limits_{i=1}^{R_k} \mathrm{fdr}_{(i)|k}$, and define $\mathrm{fdr}_k^* = 1 - (1 - \eta_k)(1 - \mathrm{fdr}_k)$, for each group k. Then let $\mathrm{fdr}_{(1)}^*, \ldots, \mathrm{fdr}_{(K)}^*$ be the ranked fdr_k^* values and $H_{(1)}, \ldots, H_{(K)}$ be the corresponding hypotheses. Then the testing procedure was to reject all $H_{(k)}$, $k = 1, \ldots, l$, where

$$l = \max\left\{ j : \frac{\sum\limits_{k=1}^{j} R_{(k)} \mathrm{fdr}_{(k)}^*}{\sum\limits_{k=1}^{j} R_{(k)}} \leq \alpha \right\}, \tag{2.18}$$

where $R_{(k)}$ is the value of R in equation (2.17) for the group that corresponds to $\text{fdr}^*_{(k)}$.

In 2021, Sarker and Nandi [34] extended Liu's framework to two-way classified hypotheses. Beside controlling FDR, Zhao [35] proposed weighted p-value procedures to control the family-wise error rate. Wang [36] used a weighted testing procedure to control the generalized family-wise error rate, which assumed p-values within each group had weak dependence.

2.2 Related work

2.2.1 Single-marker methods

In GWAS, hypothesis testing is a broadly used approach to decide if one SNP is associated with a disease. The most common method is single-marker method, which means that one SNP is considered at a time. For case-control studies, the frequently used testing methods are derived from logistic regression, which contains allele-based tests and genotype-based tests according to their independent variables. For each SNP, a contingency table can be created for case-control studies as in Table 2.2 (allele-based), Table 2.3 (genotype-based), Table 2.4(haplotype-based) or as in Table 2.5(groups of genotype distribution). Here we use A and a to denote the two alleles. There are connections between Table 2.2 and Table 2.3. Let g_{00}, g_{01}, g_{02} be the frequencies of genotypes AA, Aa, and aa in control group, respectively. Since genotype AA contains allele A twice, and genotype Aa includes one allele A, we can get $n_{00} = 2g_{00} + g_{01}$, where n_{00} is the frequency of allele A in control group. Similarly, we can have $n_{01} = g_{01} + 2g_{02}$, while n_{01} is the frequency of allele a in control group.

	A	a	Total
Control	n_{00}	n_{01}	$2n_0$
Case	n_{10}	n_{11}	$2n_1$
Total	$n_{00}+n_{10}$	$n_{01}+n_{11}$	$2n$

Table 2.2: Allele distribution for case-control studies. Here $n_{0i},\ i=0,1$ are the frequencies of alleles in control group, while $n_{1i},\ i=0,1$ are the frequencies of alleles in case group. $2n_0 = n_{00}+n_{01}$, $2n_1 = n_{10}+n_{11}$.

	AA	Aa	aa	Total
Control	g_{00}	g_{01}	g_{02}	n_0
Case	g_{10}	g_{11}	g_{12}	n_1
Total	$g_{00}+g_{10}$	$g_{01}+g_{11}$	$g_{02}+g_{12}$	n

Table 2.3: Genotype distribution for case-control studies. Here $g_{0i},\ i=0,1,2$ are the frequencies of genotypes in control group, while $g_{1i},\ i=0,1,2$ are the frequencies of genotypes in case group. $n_0 = g_{00}+g_{01}+g_{02}$, $n_1 = g_{10}+g_{11}+g_{12}$.

	AB	Ab	aB	ab	Total
Control	h_{00}	h_{01}	h_{02}	h_{03}	n_0
Case	h_{10}	h_{11}	h_{12}	h_{13}	n_1
Total	$h_{00}+h_{10}$	$h_{01}+h_{11}$	$h_{02}+h_{12}$	$h_{03}+h_{13}$	n

Table 2.4: Haplotype distribution for case-control studies. There are two loci, each with alleles: A, a at locus 1 and B, b at locus 2, which gives us four possible haplotypes: AB, Ab, aB, ab. Here $h_{0i},\ i=0,1,2,3$ are the frequencies of haplotypes in control group, while $h_{1i},\ i=0,1,2,3$ are the frequencies of haplotypes in case group. $n_0 = h_{00}+h_{01}+h_{02}+h_{03}$, $n_1 = h_{10}+h_{11}+h_{12}+h_{13}$.

	Control	Case	Total
AABB	g_{000}	g_{100}	$g_{000}+g_{100}$
AABb	g_{001}	g_{101}	$g_{001}+g_{101}$
AAbb	g_{002}	g_{102}	$g_{002}+g_{102}$
AaBB	g_{010}	g_{110}	$g_{010}+g_{110}$
AaBb	g_{011}	g_{111}	$g_{011}+g_{111}$
Aabb	g_{012}	g_{112}	$g_{012}+g_{112}$
aaBB	g_{020}	g_{120}	$g_{020}+g_{120}$
aaBb	g_{021}	g_{121}	$g_{021}+g_{121}$
aabb	g_{022}	g_{122}	$g_{022}+g_{122}$
Total	n_0	n_1	n

Table 2.5: Groups of genotype distribution for case-control studies for two genotypes. Here $g_{0ij},\ i=0,1,2,\ j=0,1,2$ are the frequencies of genotypes in control group, while $g_{1ij},\ i=0,1,2,\ 0,1,2$ are the frequencies of genotypes in case group, where i represents the number of variants in the first genotype with alleles A, a, and j represents the number of variants in the second genotype with alleles B, b.

The log-odds ratio (OR) test is a typical allele-based test in which allele frequencies are compared between control group and case group. The log-OR μ can be estimated as follows:

$$\hat{\mu} = \log \frac{n_{00}n_{11}}{n_{01}n_{10}}. \tag{2.19}$$

Using Woolf's method [37], the approximate asymptotic standard error of $\hat{\mu}$can be calculated:

$$\sigma \approx \sqrt{\frac{1}{n_{00}} + \frac{1}{n_{01}} + \frac{1}{n_{10}} + \frac{1}{n_{11}}}. \tag{2.20}$$

The test statistic is $z = \hat{\mu}/\sigma$, which follows a standard normal distribution under the null hypothesis. The log-OR test can be equivalent to the Wald test of the following logistic regression model:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \alpha + \beta x, \tag{2.21}$$

where $Y=1$ denotes the case group and $Y=0$ denotes the control group, and $x=0$ is encoded for allele A while $x=1$ represents allele a.

For genotype-based tests, people compare whether the observed genotype frequencies in cases differs from that in controls. The Pearson's χ^2 test is the most common one. Based on counts in Table 2.3, the test statistics is:

$$\chi^2 = \sum_{t=0}^{2} \left[\frac{(g_{0t} - n_0(g_{0t} + g_{1t})/n)^2}{n_0(g_{0t} + g_{1t})/n} + \frac{(g_{1t} - n_1(g_{0t} + g_{1t})/n)^2}{n_1(g_{0t} + g_{1t})/n} \right]. \tag{2.22}$$

Under the null hypothesis, χ^2 follows a χ^2 distribution with 2 degrees of freedom asymptotically. Pearson's χ^2 test can be equivalent to the score test of the following logistic regression:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \alpha + \beta_1 \mathbb{1}_{x=1} + \beta_2 \mathbb{1}_{x=2}, \tag{2.23}$$

where $\mathbb{1}$ is the indicator function, and we denote $x = 0, 1, 2$ are denoted for genotypes AA, Aa, aa, respectively.

Another popular genotype-based method is the Cochran-Armitage trend test (CATT) [38]. Its test statistic is:

$$Z_{CATT} = \frac{n_0(2g_{12} + g_{11}) - n_1(2g_{02} + g_{01})}{\sqrt{\frac{n_0 n_1}{n}(n(g_{01} + g_{11} + 4g_{02} + 4g_{12}) - (g_{01} + g_{11} + 2g_{02} + 2g_{12})^2)}}. \tag{2.24}$$

Z_{CATT} follows a standard normal distribution asymptotically under null hypothesis. It is equivalent to the score test of the following logistic regression model:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \alpha + \beta x. \tag{2.25}$$

For the test statistics above, it is easy to calculate a p-value according to their asymptotic distributions under the null hypothesis. Also exact p-values can be calculated using Fisher's exact test or permutation test, but the computation cost is high.

For studies with quantitative phenotypes, people usually use a test derived from linear regression. For example, consider the following additive linear regression model:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{2.26}$$

where X is the number of copies of allele a and $\varepsilon \sim N(0,\sigma^2)$, for which σ is a constant. The test statistic is:

$$T = \frac{\hat{\beta}_1}{\sqrt{var(\hat{\beta}_1)}}, \tag{2.27}$$

where $\hat{\beta}_1$ is the maximum likelihood estimate. Under the null hypothesis, T follows a t distribution with degrees of freedom equal to $N-2$, and N is the sample size.

In addition to these frequentist approaches, Bayesian methods have also been developed in GWAS. WTCCC has used Bayes factors [6] to discover the associated SNPs, which can provide similar ranking with p-values for common variants [39]. A Bayes factor (BF) describes the ratio of the likelihood of one particular hypothesis to the likelihood of another:

$$\text{BF} = \frac{P(y|H_1)}{P(y|H_0)}, \tag{2.28}$$

where y is the observed phenotype vector. BF > 1 favors the alternative hypothesis, while BF < 1 favors the null hypothesis. Since people need to specify a prior distribution for all unknown parameters to calculate Bayes factor, the cost of computation is high. To make it simpler, Wakefield [39] described an alternative asymptotic Bayes factor. If $\hat{\beta}$ and $\sqrt{V}$ are used to denote the maximum likelihood estimate and standard error from the above logistic regression model (2.25), and the prior distribution of β is assumed as $N(0,W)$, the Wakefield approximate Bayes factor (WABF) is:

$$\text{WABF} = \sqrt{\frac{V}{V+W}}\exp\left(\frac{Z^2W}{2(V+W)}\right), \tag{2.29}$$

where $Z = \hat{\beta}/\sqrt{V}$ is the usual Wald statistic. This is easy to calculate because $\hat{\beta}$ and V are usually available from the results of a standard frequentist analysis.

After obtaining a Bayes factor, the posterior odds (PO) can be calculated on H_1 according to the Bayes theorem:

$$\text{PO} = \text{BF} \times \frac{1-\pi_0}{\pi_0}, \tag{2.30}$$

where $\pi_0 = P(H_0)$. Then the posterior probability of association (PPA) can be calculated as:

$$\text{PPA} = \frac{\text{PO}}{1+\text{PO}}. \tag{2.31}$$

For GWAS, the probability of association $1-\pi_0$ is quite small, so BF has to be large enough (for instance, $> 10^4 - 10^6$) to provide strong evidence that the SNP is associated with the disease, which will provide PPA close to 1 [40].

2.2.2 Multiple-SNP based methods using genotypes

The single-marker methods neglect the association effects jointly expressed by multiple SNPs. Based on this, people develop multi-marker methods, which test multiple SNPs (e.g., all the SNPs in a gene or a pathway) simultaneously. The underlying null hypothesis tested is that none of the SNPs in the set are associated with the disease, while the alternative hypothesis is that at least one SNP in the set is associated with the disease.

One class of multiple-marker methods is based on multivariate analysis, such as methods based on multivariate regression [41, 42, 43]. Multivariate Hotelling's T^2 test [44] compares the means of genotype scores between different groups. Suppose there are m SNPs to be considered. Let X_{ij} denote the genotype score of the jth SNP for the ith individual from cases, which are encoded as 0, 1, 2 according to different genotypes. Similarly, let Y_{ij} denote the genotype score from controls. The Hotelling's T^2 statistic can be written as:

$$T^2 = \frac{n_0 n_1}{n}(\bar{X}-\bar{Y})^T S^{-1}(\bar{X}-\bar{Y}), \tag{2.32}$$

where $\bar{X} = (\bar{X}_1, \cdots, \bar{X}_m)^T$ and $\bar{Y} = (\bar{Y}_1, \cdots, \bar{Y}_m)^T$ represent the mean vectors of genotype scores of each SNP for case group and control group, respectively, and $S = \frac{1}{n-1}\left[\sum_{i=1}^{n_1}(X_i-\bar{X})(X_i-\bar{X})^T + \sum_{i=1}^{n_1}(Y_i-\bar{Y})(Y_i-\bar{Y})^T\right]$. Under the null hypothesis, $(n-m-1)T^2/m(n-2)$ follows a central F distribution with m and $n-m-1$ degrees of freedom asymptotically.

Multivariate regression provides a flexible framework to accommodate additional covariates and interaction forms. For GWAS, the number of independent variables (SNPs) is usually larger than the sample size, which will cause over-fitting problem in standard linear regression models. Based on this, people add regularization and shrinkage parameters in regression models to overcome this difficulty, such as ridge regression [45], Lasso penalized logistic regression [46], and Bayesian shrinkage methods [47]. For example, Let X be an $n \times m$ matrix representing genotype scores, and Y be an n-dimensional vector containing phenotype values for each individual. Consider the standard linear regression model: $Y = X\beta + \varepsilon$, with regression coefficients estimated by:

$$\hat{\beta} = \arg\min(Y - X\beta)'(Y - X\beta) = (X'X)^{-1}X'Y. \tag{2.33}$$

For ridge regression, the estimates of regression coefficients become:

$$\hat{\beta}^{Ridge} = (X'X + kI)^{-1}X'Y, \tag{2.34}$$

where k is the ridge parameter to control the degree of shrinkage. Then a Wald-test can be used to test the significance of the coefficient of each SNP. Based on a multivariate regression framework, people can choose the subset of significant SNPs at one time.

Another class of multi-marker method is based on an individual marker test, which can be divided into three classes: linear test statistics, quadratic statistics, and combined statistics. Using the above notation, denote $S_j = \sum_{i=1}^{n} (Y_i - \bar{Y})X_{ij}$, and the linear test statistics have the following form:

$$W_L = \sum_{j=1}^{m} w_j S_j. \tag{2.35}$$

When $w_j = 1$, this is the cohort allele sums test (CAST) [48] and if w_j is a function of the estimated MAF, W_L is the weighted sum method [49]. The form of quadratic statistics is as follows:

$$W_Q = S'AS, \tag{2.36}$$

where A is a positive definite symmetric matrix. When $A = \text{diag}\{a_1, \ldots, a_m\}$, this is the SKAT statistic [50], where a_j depends on the MAF.

There are some tests to combine two types of tests such as SKAT-O [51], Fisher's method, and minimum-p method [52]. Their test statistics are follows

$$T_{SKAT-O} = \max_{\rho \in [0,1]} (\rho W_L + (1-\rho) W_{SKAT}). \tag{2.37}$$

$$T_{Fisher} = -2\log(p_L) - 2\log(p_Q). \tag{2.38}$$

$$T_{min} = \min(p_L, p_Q), \tag{2.39}$$

where p_L represents the p-value for linear test statistic W_L, while p_Q denotes the p-value for W_Q. Yoo et.al [53] developed a linear combination test which combines clustering method. If m SNPs were partitioned into l clusters, they used a $m \times l$ matrix J to represent SNP assignments, where $J_{ij} = 1$ if ith SNP was assigned into the jth cluster, otherwise $J_{ij} = 0$. The test statistic is:

$$T = (W^T \hat{\beta})(W^T \Sigma W)^{-1}(\hat{\beta}^T W), \tag{2.40}$$

where $W = (\Sigma^{-1} J)(J^T \Sigma^{-1} J)^{-1}$. Yoo et.al [53] separated SNPs into clusters based on the LD information between SNPs. A threshold is needed to decide the neighbours in a cluster. However, different choices of threshold values will affect the power.

For the above test statistics, no test statistic can be powerful for all situations. When the directions of individual effects are different, or when the directions are the same, but the proportion of SNPs associated with phenotype is small, quadratic test statistics are more powerful than linear test statistics. When the proportion of associated SNPs is high and the directions are the same, linear test statistics are more powerful [54]. For combined test statistics, Fisher's method is better than minimum-p value when the direction of effect is the same [52].

Beside these statistics, there is another popular test statistic: C-alpha test [55]. C-alpha detects unusual numbers of counts of alleles. If the target region has no alleles associated with the phenotype, the distribution of counts should follow a binomial distribution. The binomial (n, p) distribution evaluates the probability of observing a particular variant y times in the cases out of n total, assuming the rare variants are distributed at random across the subjects. For the ith variant, if the total observation is n_i, y_i is assumed to follow binomial (n_i, p_i). Under the null hypothesis, $p_i = p_0$, where $p_0 = 1/2$ if cases and controls are equal in number and rare variants fall in either sample at random. The alternative hypothesis is that p_i follows a mixture distribution across the m variants.

The C-alpha test statistic is as follows:

$$T = \sum_{i=1}^{m} [(y_i - n_i p_0)^2 - n_i p_0 (1 - p_0)], \tag{2.41}$$

where T contrasts the variance of each observed count with the expected variance, assuming the binomial distribution. The resulting test statistic is defined as $Z = T/\sqrt{c}$, where c is the variance of T. Then the null hypothesis is rejected when Z is larger than expected using a one-tailed standard normal distribution.

Besides above model, to leverage LD information to improve the power, Li [56] proposed to use hidden Markov random field model. They set a random indicator variable for a given SNP s as:

$$X_s = \begin{cases} 1 & \text{if SNP } s \text{ is associated with the disease} \\ 0 & \text{if SNP } s \text{ is not associated with the disease .} \end{cases} \tag{2.42}$$

Then they used a Markov random field (MRF) model to model the dependency as follows:

$$p(X;\Phi) \propto \exp\left(\gamma \sum_{s=1}^{p} X_s + \beta \sum_{s \sim s'} w_{ss'} I(X_s = X_s')\right), \tag{2.43}$$

where γ and β are model parameters, and $w_{ss'} = I(r_{ss'}^2 > \tau) r_{ss'}^2$, where $r_{ss'}^2$ is the LD between SNP s and s', and τ is a predetermined cutoff value.

By assuming the genotype frequencies had a Dirichlet prior, they estimated $f(Y_s \mid X_s)$, where Y_s represents the observed genotype data, and used the ICM algorithm to estimate parameters. Simulation experiments showed that their methods could find more true positives than the Empirical Bayes method and Cochran-Armitage trend test. However, their method involves the genotype frequency information of individuals, which is usually difficult to obtain due to the privacy. Also when using their methods on a real data set, they grouped SNPs into groups of 1000 SNPs, which may lose some information. In our study, a hidden Markov random field model based on summary statistics is developed.

2.2.3 Multiple-SNP based methods using haplotypes

Besides the multiple-SNP methods based on genotypes discussed above, another popular strategy to capture the correlation structure between SNPs is to use haplotypes, which consider a set of alleles at different locations of a DNA sequence which are inherited together. The analysis based on haplotypes can have fewer degrees of freedom, since they consider a block of SNPs together. However, in practice, haplotypes of SNPs are not observed from the data directly. To solve this, people often use the phase procedure to infer haplotypes from the genotype data. To implement this procedure, many methods have been proposed, such as parsimony approaches [57, 58], maximum-likelihood methods [59, 60] or Bayesian methods [61, 62]. However, most of those methods cannot deal with large data sets. Therefore, as in the development of large GWAS, more computationally efficient methods are developed [62, 63, 64, 65]. For example, Stephens and Donnelly [66] introduced a new algorithm named as PHASE version 2.0 and showed that the new method outperformed 3 existing Bayesian methods using simulation and real datasets. Halperin and Eskin [67] proposed to partition the SNPs into blocks, and predicted the common haplotypes and each individual's haplotype for each block. This method is called HAP, which is more efficient than PHASE and can deal with large datasets. Marchini et al. [68] compared the performance of several leading algorithms, and showed that PHASE is the most accurate algorithm, but it is the slowest.

For association tests based on haplotype, the simplest method is to test the independence in a contingency table [69]. However, this method does not account for the uncertainty in inferring the haplotypes when testing the association of haplotypes. Then people develop methods to integrate the procedure of phasing and testing [70, 71, 72, 73, 74]. For example, Zhu et al. [75] developed a two-stage procedure, which can identify and classify rare risk haplotypes using a relatively small sample.

Ali and Zhang [76, 77] improved Zhu's method by grouping genotypes before the association testing. They select risk genotypes in stage 1 and infer risk haplotypes in stage 2 based on results from stage 1. They used a simulation study to show that their methods are better.

Chapter 3

Hidden Markov random field model

In this chapter, the hidden Markov random field model and parameter estimation methods will be introduced.

3.1 Model Description

A hidden Markov random field (HMRF) model is a generalization of a hidden Markov model (HMM). For a HMM, the hidden variable is an underlying one-dimensional Markov chain, which can only have two neighbours for one variable and can not be directly used in two-dimensional or three-dimensional problems. Therefore, HMRF is developed, which has an underlying Markov random field for hidden variables and defines neighbours within a network. So a HMRF model is usually used in 2D or 3D problems such as image segmentation or modelling spatial dependence [78, 79, 80].

Consider a random variable $Y = \{Y_i,\ i = 1, \ldots, n\}$. A HMRF assumes that Y is determined by the unobservable Markov random field $X = \{X_i,\ i = 1, \ldots, n\}$. The values of X have the following distribution:

$$P(X|\boldsymbol{\theta}) = \frac{1}{\psi(\boldsymbol{\theta})} \exp\{-U_{\boldsymbol{\theta}}(X)\}, \tag{3.1}$$

where $\boldsymbol{\theta}$ represent the parameters, $\psi(\boldsymbol{\theta})$ denotes the normalizing constant which has the form $\psi(\boldsymbol{\theta}) = \sum_X \exp\{-U_{\boldsymbol{\theta}}(X)\}$. $U_{\boldsymbol{\theta}}(X)$ denotes the energy function of the form:

$$U_{\boldsymbol{\theta}}(X) = \sum_{c \in C} V_{\boldsymbol{\theta},c}(X), \tag{3.2}$$

which is a sum of potential functions $V_{\boldsymbol{\theta},c}(X)$ over all possible cliques C. A clique in a graphical model is defined as a subset of nodes in which every pair of distinct nodes are neighbours, except for single-site cliques, which means the set of nodes in a clique are fully connected. The clique potential functions represent the local relationships between variables within cliques of the graphical model.

The most common form of energy function for binary spatial process is the Ising model:

$$U_{\boldsymbol{\theta}}(X) = -\alpha \sum_{i=1}^{n} X_i - \beta \sum_{i \sim j} X_i X_j, \tag{3.3}$$

where $\{X_i, i = 1\ldots, n\}$ take values in $\{-1, 1\}$. The notation $i \sim j$ represents that X_j is a neighbour of X_i and each neighbouring pair is calculated once in the summation. The parameter α controls the relative abundance of -1's and $+1$'s. The parameter β represents the interaction strength between X_i and X_j. When $\beta > 0$, neighbouring nodes tend to encourage the same sign, while adjacent notes are more likely to have opposite signs if $\beta < 0$[81, 82].

For each particular configuration X, every Y_i follows a known conditional probability distribution $p(Y_i \mid X_i)$ of the same functional form such as a normal distribution. For any X, the random variables Y_i are conditionally independent:

$$P(Y|X) = \prod_{i=1}^{n} P(Y_i \mid X_i). \tag{3.4}$$

This is called the conditional independence assumption.

3.2 Inference

To make it simple to show the parameter estimation procedure, consider the simplest Ising model as follows:

$$P_\beta(X) = \frac{1}{\psi(\beta)} \exp\left(\beta \sum_{i=1}^{n} X_i \sum_{j \in N(i)} X_j\right), \tag{3.5}$$

where $N(i)$ denotes the neighbours of X_i. Given X, each component in Y has a distribution $P_\phi(Y \mid X)$, where ϕ represents parameters in the conditional distribution

$P(Y \mid X)$. So the joint distribution of hidden state X and observed Y is as follows:

$$P_\theta(X_1 \ldots, X_n, Y_1, \ldots, Y_n) = P_\beta(X_1, \ldots, X_n) \prod_{i=1}^{n} P_\phi(Y_i \mid X_i), \tag{3.6}$$

where $\theta = (\beta, \phi)$. Since the model involves a latent variable X, it is common to use the EM algorithm to obtain the parameter estimates based on the likelihood function [83, 84]. Except EM algorithm, MCMC algorithm [85] or iterative conditional mode (ICM) algorithm [56] are applied in estimating parameters in some papers. The EM algorithm is an effective method for performing maximum likelihood estimation when there are latent variables. It evaluates the expectation of the complete data log likelihood using current parameter in E step. Then it estimates new parameters by maximizing the expectation of complete data log likelihood in M step. These two steps are repeated until convergence [86]. The conditional expectation of the complete data log-likelihood $Q(\theta|\theta^{(old)})$ is as follows:

$$\begin{aligned} Q(\theta|\theta^{(old)}) = E_{\theta^{(old)}} &\Big[\beta \sum_{i=1}^{n} X_i \sum_{j \in N(i)} X_j - \log \psi(\beta) \mid Y\Big] \\ &+ E_{\theta^{(old)}} \Big[\sum_{i=1}^{n} \mathbb{1}(X_i = 1) \log P_\phi(Y_1 \mid X_i = 1) \\ &+ \sum_{i=1}^{n} \mathbb{1}(X_i = -1) \log P_\phi(Y_1 \mid X_i = -1) \mid Y\Big] \\ &= l(\beta) + l(\phi). \end{aligned} \tag{3.7}$$

The estimation of $\theta = (\beta, \phi)$ can be separable. We consider in turn the estimation of ϕ and β.

(1) Estimation of ϕ.

For the first indicator function in $l(\phi)$ of equation (3.7), the conditional expectation can be calculated:

$$E_{\theta^{(old)}}[\mathbb{1}(X_i = 1) \mid Y] = P_{\theta^{(old)}}(X_i = 1 \mid Y). \tag{3.8}$$

Then $l(\phi)$ can be written as:

$$\sum_{i=1}^{n}\left\{P_{\phi}(X_i=1\mid Y)\log P_{\phi}(Y_i\mid X_i=1)+P_{\phi}(X_i=-1\mid Y)\log P_{\phi}(Y_i\mid X_i=-1)\right\}. \tag{3.9}$$

Since $P_{\phi}(X_i=1\mid Y)$ and $P_{\phi}(X_i=-1\mid Y)$ do not involve the parameter ϕ, the estimation of ϕ is the solution to:

$$\begin{aligned}&\sum_{i=1}^{n}\{P_{\phi}(X_i=1\mid Y)\frac{\partial\log P_{\phi}(Y_i\mid X_i=1)}{\partial\phi}\\&+P_{\phi}(X_i=-1\mid Y)\frac{\partial\log P_{\phi}(Y_i\mid X_i=-1)}{\partial\phi}\}=0.\end{aligned} \tag{3.10}$$

Since the distribution form $P(Y_i\mid X_i)$ is known, we can get the closed form solution of ϕ for equation (3.10).

(2) Estimation of β.

The estimate of parameter β is the solution of following equation:

$$\frac{\partial}{\partial\beta}l(\beta)=\frac{\partial}{\partial\beta}\left\{\beta E_{\theta^{(old)}}\left[\sum_{i=1}^{n}X_i\sum_{j\in N(i)}X_j\mid Y\right]-\log\psi(\beta)\right\}=0, \tag{3.11}$$

where $\psi(\beta)=\sum_{X}\exp\{\beta\sum_{i=1}^{n}X_i\sum_{j\in N(i)}X_j\}$. The above equation can be rewritten as follows:

$$E_{\theta^{(old)}}\left[\sum_{i=1}^{n}X_i\sum_{j\in N(i)}X_j\mid Y\right]=\frac{\frac{\partial}{\partial\beta}\psi(\beta)}{\psi(\beta)}=E_{\beta}\left[\sum_{i=1}^{n}X_i\sum_{j\in N(i)}X_j\right]. \tag{3.12}$$

Since the above equation does not have a closed form solution, a recursive algorithm like Newton-Raphson method is usually used. The parameter updating rule is:

$$\beta^{(t)}=\beta^{(t-1)}-\frac{l'(\beta^{(t-1)})}{l''(\beta^{(t-1)})}, \tag{3.13}$$

where

$$l''(\beta)=\frac{\partial^2}{\partial^2\beta}(-\log\psi(\beta))=-\left[\frac{\psi''(\beta)}{\psi(\beta)}-\left(\frac{\psi'(\beta)}{\psi(\beta)}\right)^2\right], \tag{3.14}$$

$$\frac{\psi''(\beta)}{\psi(\beta)} = E_\beta \left[\left(\sum_{i=1}^{n} X_i \sum_{j \in N(i)} X_j \right)^2 \right]. \tag{3.15}$$

Then the double derivative of likelihood function with respect to β is equal to

$$l''(\beta) = -var_\beta \left[\sum_{i=1}^{n} X_i \sum_{j \in N(i)} X_j \right]. \tag{3.16}$$

Then the updating equation of β is:

$$\beta^{(t)} = \beta^{(t-1)} + \frac{E_{\theta^{(old)}}\left[\sum_{i=1}^{n} X_i \sum_{j \in N(i)} X_j \mid Y\right] - E_{\beta^{(t-1)}}\left[\sum_{i=1}^{n} X_i \sum_{j \in N(i)} X_j\right]}{var_{\beta^{(t-1)}}\left[\sum_{i=1}^{n} X_i \sum_{j \in N(i)} X_j\right]}, \quad (3.17)$$

where $E_{\theta^{(old)}}\left[\sum_{i=1}^{n} X_i \sum_{j \in N(i)} X_j \mid Y\right]$ can be approximated by Monte-Carlo sampling of the posterior distribution $P_{\theta^{(old)}}(X \mid Y)$. $E_{\beta^{(t-1)}}\left[\sum_{i=1}^{n} X_i \sum_{j \in N(i)} X_j\right]$ and $var_{\beta^{(t-1)}}\left[\sum_{i=1}^{n} X_i \sum_{j \in N(i)} X_j\right]$ can also be approximated by Monte-Carto sampling of $P(X \mid \beta^{(t-1)})$.

To calculate the conditional expectation of the log-likelihood function in the E step for the EM algorithm, the mostly likely state of X needs to be estimated [85].

3.3 A hidden Markov random field model for GWAS using summary statistics

3.3.1 Introduction

Genome-wide association studies have been increasingly used to detect genetic variants associated with diseases, by genotyping single nucleotide polymorphisms (SNPs) in diseased and normal individuals. For complex diseases including prostate and breast cancers [4] and type 2 diabetes [5], GWAS have been shown to be an effective method to detect associated genetic variants.

In traditional GWAS, it is common to use hypothesis testing to identify SNPs associated with a disease. The most frequently used method is single-marker method (see Chapter 2). However, the large number of hypotheses in GWAS will cause a multiple testing problem. To solve this problem, several statistical techniques such as Bonferroni correction or FDR control have been used to make correction for multiple testing.

However, the single-SNP analysis neglects the association effects jointly expressed by multiple SNPs. Based on this, multi-marker methods have been developed such as multivariate analysis based on multivariate regression such as ridge regression [45], Lasso penalized logistic regression [46], and Bayesian shrinkage method [47], which usually need the individual genotype information. To avoid this problem, some meta analysis methods have been developed such as cohort allele sums test (CAST) [48], SKAT [50] , SKAT-O [51], Fisher's method, and minimum-p method [52] (see Chapter 2 for details).

However, since nearby SNPs are usually in linkage disequilibrium (LD), these methods do not utilize LD information between SNPs when jointly analysing multiple SNPs. If there are multiple SNPs which are in strong LD, utilizing LD information between SNPs effectively may increase the power of identifying SNPs associated with disease, especially for those SNPs having weak effect on disease. Furthermore, previous multi-marker methods can only identify whether a set of SNPs are associated with disease or not, but they can not discover which SNPs are associated with disease. Li [56] proposed to use a hidden Markov random field model to leverage LD information from multiple SNPs. However, individual level genotype data were needed in their study. Using only summary statistics, Sun and Cai [26] proposed to consider LD information in a Hidden Markov Model and proposed a local index of significance, which could be used to select associated SNPs when SNPs are correlated.

In this study, we propose to leverage LD information using a hidden Markov random field model (HMRF) and use summary statistics to detect association between SNPs and disease. We regard the true associated statuses as hidden variables, and build a weighted LD graph based on LD information between them. The dependence of hidden variables are assumed to follow a Markov random field model. Then we choose the two-component mixture prior for all SNPs and estimate model parameters using the EM algorithm. Finally, we propose to use the Gibbs sampling method to estimate posterior probability of true association status and select associated SNPs using a false discovery rate (FDR) procedure.

3.3.2 Method

3.3.2.1 Weighted graph and hidden Markov random field model

Suppose there are m SNPs. Let $S = \{1, \ldots, m\}$ denote the SNP index. For any SNP i, the hypotheses in which we are interested are:

H_{i0} SNP i is not associated with the disease

and

H_{i1} SNP i is associated with the disease.

Firstly, for a given SNP, we define a random indicator variable as

$$\theta_i = \begin{cases} 1 & \text{if SNP } i \text{ is associated with the disease} \\ 0 & \text{if SNP } i \text{ is not associated with the disease.} \end{cases} \tag{3.18}$$

Usually, nearby SNPs are highly correlated, which means that the dependence between SNP i and SNP j is stronger when i and j are close. Therefore, we model the dependency between SNPs using a discrete Markov random field model with the following joint probability function for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$:

$$p(\boldsymbol{\theta}; \Phi) \propto \exp\left(\gamma \sum_{i=1}^{m} \theta_i + \beta \sum_{i \sim j} w_{ij} I(\theta_i = \theta_j)\right), \tag{3.19}$$

where $\Phi = (\gamma, \beta)^T$ represent the model parameters. $\beta > 0$ will encourage SNPs with LD to have similar states. w_{ij} represents the LD information between SNP i and SNP j. We define w_{ij} using the following method:

$$w_{ij} = I(r_{ij}^2 > \tau) r_{ij}^2 \tag{3.20}$$

where r_{ij} is the $r2$ measurement (see equation (2.4) in Section 2.1.2) of LD between SNP i and j, and τ is a predetermined cutoff value. For a larger τ, it will generate a more sparse weight matrix since $w_{ij} = 0$ if $r_{ij}^2 < \tau$, so this will reduce the computation cost. For a smaller τ, it can preserve more dependence information, but the computation cost will increase. Li [56] used $\tau = 0.4$. Here to utilize more dependence information, we use $\tau = 0.1$ in our study. Given that the value of τ may matter, a useful avenue of future research could be to devise a way to estimate the value of τ empirically from the data.

The conditional association state for SNP i, given the states of all neighbouring SNPs, is

$$p(\theta_i \mid \theta_{N_i};\Phi) \propto \exp\left(\gamma\theta_i + \beta \sum_{j \in N_i} w_{ij} I(\theta_i = \theta_j)\right), \tag{3.21}$$

where N_i represents the neighbours of the SNP i on the LD graph.

3.3.2.2 Gaussian mixture model

Suppose we observe summary statistics Z values, $Z = (Z_1, \ldots, Z_m)$, which represent the individual test statistics for m SNPs. We assume that Z values are conditionally independent given the hidden indicators, so that:

$$P(Z \mid \theta) = \prod_{i=1}^{m} P(Z_i \mid \theta_i). \tag{3.22}$$

For an arbitrary SNP, we do not know whether it is associated or not. So the distribution of Z values are assumed as following mixture distribution:

$$Z \mid \theta \sim (1-\theta)N(0,1) + \theta N(\mu, \sigma^2). \tag{3.23}$$

The Z values are assumed to follow a Gaussian mixture distribution according to their hidden states. For SNP i, When hidden state $\theta_i = 0$, the corresponding Z value Z_i has the standard normal distribution $N(0,1)$, while under alternative hypothesis $\theta_i = 1$, the Z value Z_i follows a shifted normal distribution with mean μ and variance σ^2.

3.3.3 Parameter estimation and FDR control procedure

In order to estimate parameters γ, β, μ and σ^2, EM algorithm is applied.

Firstly, the conditional expectation of complete data log-likelihood $Q\left(\phi \mid \phi^{\{old\}}\right)$ can be written as follows:

$$\begin{aligned} Q\left(\phi \mid \phi^{\{old\}}\right) &= E_{\phi^{\{old\}}}\left(\gamma \sum_{i=1}^{m} \theta_i + \beta \sum_{i\sim j} w_{ij} I(\theta_i = \theta_j) - \log \psi(\theta) \mid Z\right) \\ &+ E_{\phi^{\{old\}}}\left(\sum_{i=1}^{m} \log P(Z_i \mid \theta_i)\right) = l_1(\boldsymbol{\phi}_1) + l_2(\boldsymbol{\phi}_2). \end{aligned} \tag{3.24}$$

Where $\psi(\theta) = \sum\limits_{\theta \in \{0,1\}^m} \exp(\gamma\theta_i + \beta \sum\limits_{j \in N_i} w_{ij} I(\theta_i = \theta_j))$, $\boldsymbol{\phi}_1 = (\gamma, \beta)$, $\boldsymbol{\phi}_2 = (\mu, \sigma^2)$. Taking the first and second derivatives with respect to $\boldsymbol{\phi}_1$, we obtain:

$$U(\boldsymbol{\phi}_1) = \frac{\partial}{\partial \boldsymbol{\phi}_1} l_1(\boldsymbol{\phi}_1) = E_{\boldsymbol{\phi}^{(old)}}[\boldsymbol{H}(\boldsymbol{\theta}) \mid Z] - E_{\boldsymbol{\phi}_1}[\boldsymbol{H}(\boldsymbol{\theta})]. \tag{3.25}$$

$$I(\boldsymbol{\phi}_1) = -\frac{\partial^2}{\partial \boldsymbol{\phi}_1 \partial \boldsymbol{\phi}_1^T} l_1(\boldsymbol{\phi}_1) = Var_{\boldsymbol{\phi}_1}[\boldsymbol{H}(\boldsymbol{\theta})]. \tag{3.26}$$

where $\boldsymbol{H}(\boldsymbol{\theta}) = (H_1, H_2)^T = (\sum\limits_{i=1}^{m} \theta_i, \sum\limits_{i\sim j} w_{ij} I(\theta_i = \theta_j))^T$. Since the distribution of $P(\boldsymbol{\theta}|Z)$ and $P(\boldsymbol{\theta})$ both involve a normalizing term $\psi(\boldsymbol{\theta})$, which needs to be calculated by considering all configuration of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$. This is quite difficult. So we use Gibbs sampling to estimate the expectation and variance in equation (3.25) and (3.26). The steps to estimate parameters using EM algorithm are as follows:

HMRF parameter estimation

Input: Z: observed Z values and weight matrix w;

Output: optimal γ^*, β^*, μ^*,σ^*

1: Initialize the states of $\hat{\theta}_i$. Calculate the corresponding p-values according to Z values. According to the Bonferroni correction, for a large number of hypothesis tests, to control the type I error at α, when p-value is smaller than α/m, the hypothesis tests are rejected, where m is the number of hypothesis tests. Here we choose a small value 0.0001 as the threshold to generate the initial state of $\hat{\theta}_i$. Since this is just a threshold to generate initial value, we do not follow the Bonferroni correction criterion strictly and the choice of 0.0001 is the same as Li's setting [56]. That means if $p_i < 0.0001$, $\hat{\theta}_i$ is set as 1 and 0 otherwise.

2: Generate 5000 Gibbs samplers, where 1500 of Gibbs samplers are regarded as burn-in period. Use Gibbs samplers from $P(\theta_i \mid Z, \hat{\theta}_{S\backslash i})$ to estimate

$E_{\boldsymbol{\phi}^{(old)}}[\boldsymbol{H(\theta)} \mid Z]$ and $P(\theta_i = 1 \mid Z_i)$. Use Gibbs sampler from $p(\theta_i \mid \theta_{N_i})$ to estimate $E_{\boldsymbol{\phi}_1}[\boldsymbol{H(\theta)}]$ and $Var_{\boldsymbol{\phi}_1}[\boldsymbol{H(\theta)}]$; $\boldsymbol{\phi}_1 = (\gamma, \beta)$. The distributions of $P(\theta_i \mid Z, \hat{\theta}_{S \setminus i})$ and $p(\theta_i \mid \theta_{N_i})$ are as follows:

$$P(\theta_i \mid Z, \hat{\theta}_{S \setminus i}) \propto P(Z_i \mid \theta_i; \mu, \sigma) P(\theta_i \mid \hat{\theta}_{N_i}; \gamma, \beta) \tag{3.27}$$

$$p(\theta_i \mid \theta_{N_i}) \propto \exp\left(\gamma\theta_i + \beta \sum_{j \in N_i} w_{ij} I(\theta_i = \theta_j)\right). \tag{3.28}$$

3: update μ and σ^2 using the following equations:

$$\mu = \frac{\sum_{i=1}^{m} P(\theta_i = 1 \mid Z_i) Z_i}{\sum_{i=1}^{m} P(\theta_i = 1 \mid Z_i)} \tag{3.29}$$

$$\sigma^2 = \frac{\sum_{i=1}^{m} P(\theta_i = 1 \mid Z_i)(Z_i - \mu)^2}{\sum_{i=1}^{m} P(\theta_i = 1 \mid Z_i)}. \tag{3.30}$$

4: update the value of $\boldsymbol{\phi}_1 = (\gamma, \beta)$:

4.1 Maximizing $l(\boldsymbol{\phi}_1)$ is equivalent to solving the equation: $U(\boldsymbol{\phi}_1) = 0$. To avoid to search for the solution over all $\boldsymbol{\phi}_1$, we find a new $\boldsymbol{\phi}_1$ to increase $l_1(\boldsymbol{\phi}_1)$ [83]. A set of decreasing positive values λ_m are introduced:

$$\boldsymbol{\phi}_1^{(t+1,m)} = \boldsymbol{\phi}_1^{(t)} + \lambda_h I(\boldsymbol{\phi}_1^{(t)})^{-1} U(\boldsymbol{\phi}_1^{(t)}), \tag{3.31}$$

where $\lambda_h = 2^{-h}$, $U(\boldsymbol{\phi}_1) = E_{\boldsymbol{\phi}^{(old)}}[\boldsymbol{H(\theta)} \mid Z] - E_{\boldsymbol{\phi}_1}[\boldsymbol{H(\theta)}]$ and $I(\boldsymbol{\phi}_1) = Var_{\boldsymbol{\phi}_1}[\boldsymbol{H(\theta)}]$. Then the new $\boldsymbol{\phi}_1$ is equal to $\boldsymbol{\phi}_1^{(t+1,m)}$, which is the first one satisfying the following Armijo condition:

$$l_1(\boldsymbol{\phi}_1^{(t+1,m)}) - l_1(\boldsymbol{\phi}_1^{(t)}) \geq \alpha \lambda_h U(\boldsymbol{\phi}_1^{(t)})^T I(\boldsymbol{\phi}_1^{(t)})^{-1} U(\boldsymbol{\phi}_1^{(t)}). \tag{3.32}$$

4.2 For $l_1(\boldsymbol{\phi}_1^{(t+1,m)}) - l_1(\boldsymbol{\phi}_1^{(t)})$,

$$\begin{aligned} l_1(\boldsymbol{\phi}_1^{(t+1,m)}) - l_1(\boldsymbol{\phi}_1^{(t)}) \approx {} & \frac{1}{n}(\boldsymbol{\phi}_1^{(t+1,m)} - \boldsymbol{\phi}_1^{(t)})^T \sum_{i=1}^{n} \boldsymbol{H}(\boldsymbol{\theta}^{(t,i)}) \\ & + \log\left(\frac{\sum_{i=1}^{n} \exp\{-\boldsymbol{\phi}_1^{(t+1,m)^T} \boldsymbol{H}(\boldsymbol{\theta}^{(i,\phi_1^{(t+1,m)})})\}}{\sum_{i=1}^{n} \exp\{-\boldsymbol{\phi}_1^{(t)^T} \boldsymbol{H}(\boldsymbol{\theta}^{(i,\phi_1^{(t)})})\}} \right). \end{aligned} \tag{3.33}$$

where $\theta^{(t,i)}$ are Gibbs samplers from $P(\theta_i \mid Z, \hat{\theta}_{S\setminus i})$. $\theta^{(i,\phi_1^{(t+1,m)})}$ and $\theta^{(i,\phi_1^{(t)})}$ are Gibbs sampler from $p(\theta_i \mid \theta_{N_i})$. See the Appendix for the derivation of equation (3.33).

5: Repeat step 2,3,4 until convergence;

After estimating parameters, we can estimate the posterior probability of $LIS_i = P(\theta_i = 0 \mid Z)$ using Gibbs sampling. Then to select SNPs associted with the disease, we use the FDR control procedure. Let $LIS_{(1)}, \ldots, LIS_{(m)}$ be the sorted values of LIS_i in descending order and $H_{(1)}, \ldots, H_{(m)}$ be the corresponding hypothesis for m SNPs. Then define $k = \max\{t : \frac{1}{t} \sum_{i=1}^{t} LIS_{(i)} \leq \alpha\}$ and reject all $H_{(i)},\ i = 1, \ldots, k$.

Chapter 4

Simulation study

4.1 Simulation study 1

We conduct simulation experiments to evaluate the performance of the proposed model and demonstrate that the proposed model is more powerful than an lfdr-based procedure which does not consider LD information between SNPs.

To simulate summary statistics Z values, the hidden states θ are simulated firstly from model (3.21). We randomly set initial values of half of θ as 1, while the remaining half of θ are set as 0. Then Gibbs sampling is applied to update values of θ according to the conditional distribution $p(\theta_i \mid \theta_{N_i}; \Phi)$ in equation (3.21) to generate the states of θ. In model (3.21), we choose different values of γ and β to assess the performance. A negative and smaller γ means that we encourage to generate more 0s and fewer 1s, which is reasonable since a small number of SNPs are associated with disease in reality. When β is positive and large, the nearby SNPs are encouraged to have the same hidden state. For the LD matrix $W = \{w_{ij}, i = 1, \ldots, m, j = 1, \ldots, m\}$, where m is the number of SNPs, the following AR model is used.

$$W = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{m-1} \\ \rho & 1 & \rho & \cdots & \rho^{m-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{m-1} & \rho^{m-2} & \rho^{m-3} & \cdots & 1 \end{pmatrix}, \tag{4.1}$$

where $\rho = 0.7$ in our simulation. After simulating hidden states θ, summary statistics Z values are generated according to equation (3.23), and different values for μ are set to compare the performance.

To compare the performance with different parameters, three scenarios are considered in Table 4.1.

	The number of SNPs	ρ	γ	β	μ	σ^2
Setting 1	1000	0.7	-0.3	0.6	2	1
Setting 2	1000	0.7	-0.3	0.6	3	1
Setting 3	1000	0.7	-0.2	0.6	3	1

Table 4.1: The parameters for different settings

Table 4.1 describes different parameter settings in simulation study 1. For settings 2 and 3, the value of μ increases, which means disease related SNPs will have larger effect sizes. This will make these SNPs identified more easily. For setting 3, the value of γ increases, which is expected to generate more disease related SNPs.

For each simulation, 3500 Gibbs samplers are generated from 3500 iterations after a burn-in period of 1500 iterations. The maximum iteration number for HMRF parameter estimation is set as 500. If the iteration number is reached or the convergence condition is satisfied, the parameter estimation procedure will end. For simulation data, firstly we generate true states of θ by Gibbs sampling based on equation (3.21) and the iteration number is 10000. After having true θ, Z values are generated according to a different normal distribution for different states of θ. the proposed model runs on UCL's high performance research computing platform using Myriad. The code will take about 48 hours when running 10 tasks in parallel. So to study the effect of random sampling and consider the time cost the code takes, we run 10 simulations for each setting, which are based on 10 different simulated datasets. The initial values of parameters are set as: $\gamma = \beta = \mu = 0$ and $\sigma = 0.1$. The description of the simulation data and parameter estimation results are summarised in Table 4.2.

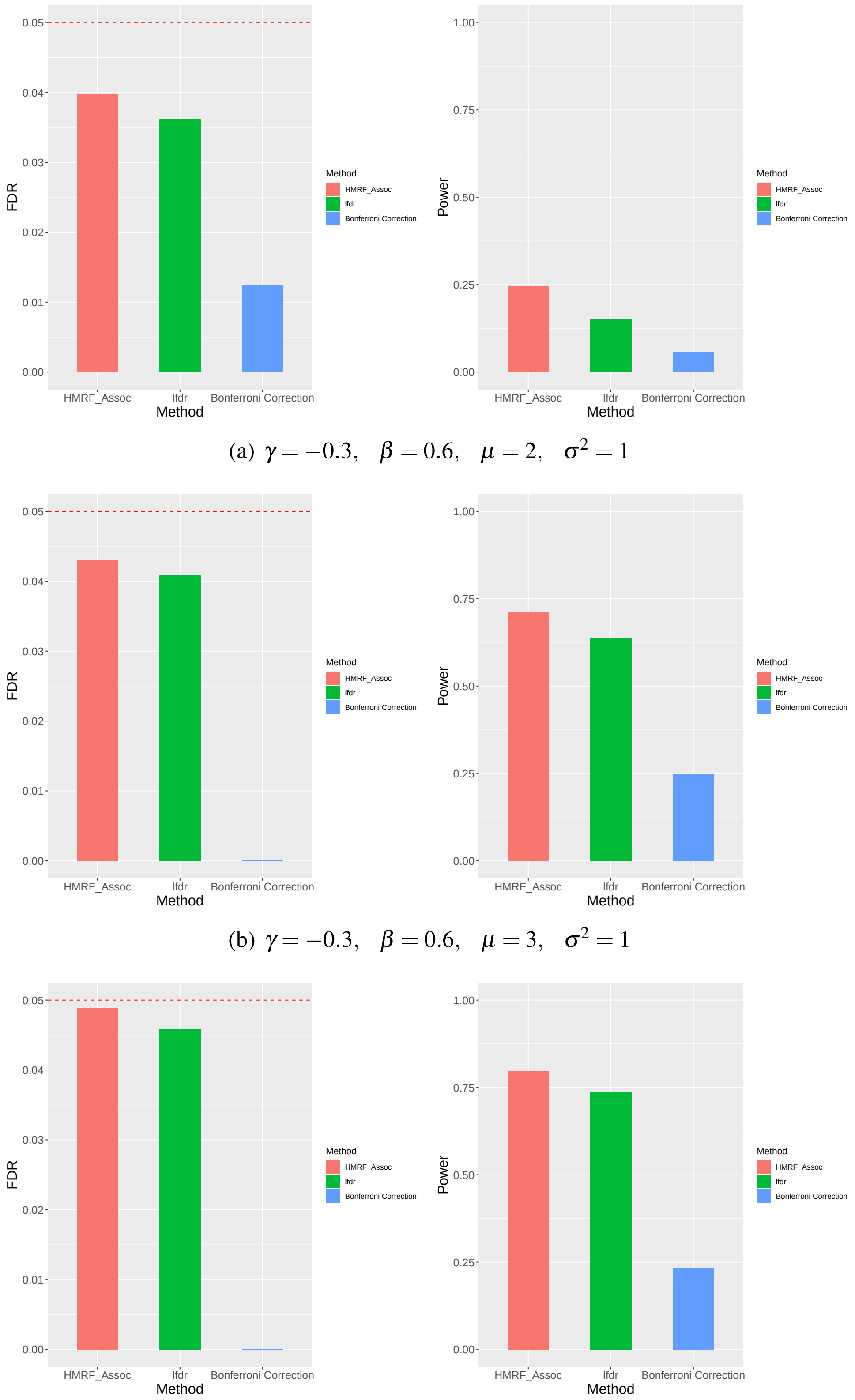
	The mean of $\sum_{i=1}^{m} \theta_i$		γ	β	μ	σ^2
Setting 1	135.8	Mean estimate	-0.28	0.59	1.80	1.29
		True value	-0.3	0.6	2	1
Setting 2	137.8	Mean estimate	-0.32	0.59	2.93	1.09
		True value	-0.3	0.6	3	1
Setting 3	185.8	Mean estimate	-0.24	0.57	2.97	1.05
		True value	-0.2	0.6	3	1

Table 4.2: The parameters for different settings. The $\sum_{i=1}^{m} \theta_i$ is the number of SNPs which are associated with disease. For setting 1 and 2, we have the same γ and β, but different mean values of $\sum_{i=1}^{m} \theta_i$. This is because Gibbs sampling for generating simulation data is random. The mean values are based on 10 simulation data for each parameter setting. The mean estimated values are also based on 10 different simulation data.

Table 4.2 contains the mean estimates of 10 simulation datasets for different parameter settings. It can be seen that except σ^2 in setting 1, other estimated parameters are quite close to true values. The estimated results for setting 3 are most accurate compared with other two settings. When few SNPs are associated with disease, the data are more sparse, which make parameter estimation more difficult.

After estimating parameters and having LIS_s for each SNP, we select SNPs associated with disease using the FDR control procedure (see Section 3.3). The significance level α is set as 0.05. After selecting associated SNPs, we compare them with the true states of SNPs and compute the false discovery rates (FDR) and the power. Then we compare them with the Bonferroni Correction method and lfdr. For the Bonferroni Correction method, the p values which are smaller than $\alpha/1000$ are considered as significant. For lfdr, lfdr for each SNP is computed using R package ”locfdr” [87] and then associated SNPs are selected using FDR control procedure, which does not consider dependence information between SNPs. The results of FDR and power are shown in Figure 4.1. Although the aim of all methods is to control the FDR at approximately 5%, for some methods, notably the Bonferroni correction, the realised FDR is far below 5%. As may be expected, in these cases the power is also relatively low.

For all settings, it can be seen from the plots on the left of Figure 4.1 that the average FDR of our proposed method, which will call HMRF_Assoc is around the predefined level ($\alpha = 0.05$), which means the type I error is well controlled. Comparatively, the FDR for Bonferroni Correction methods is too conservative for all settings, while the FDR for lfdr is a little conservative in setting1. For the comparison of power on the right of Figure 4.1, the average empirical power of the proposed method are higher than that of other two methods for all three settings, which means proposed method perform better for identifying true associated SNPs. It can be seen that the power increases dramatically when μ increases from 2 to 3. Since the difference between null distribution and alternative distribution becomes larger, it is easier to discover disease related SNPs. For different choices of γ, the performance of proposed method does not change much.



(a) $\gamma = -0.3, \quad \beta = 0.6, \quad \mu = 2, \quad \sigma^2 = 1$

(b) $\gamma = -0.3, \quad \beta = 0.6, \quad \mu = 3, \quad \sigma^2 = 1$

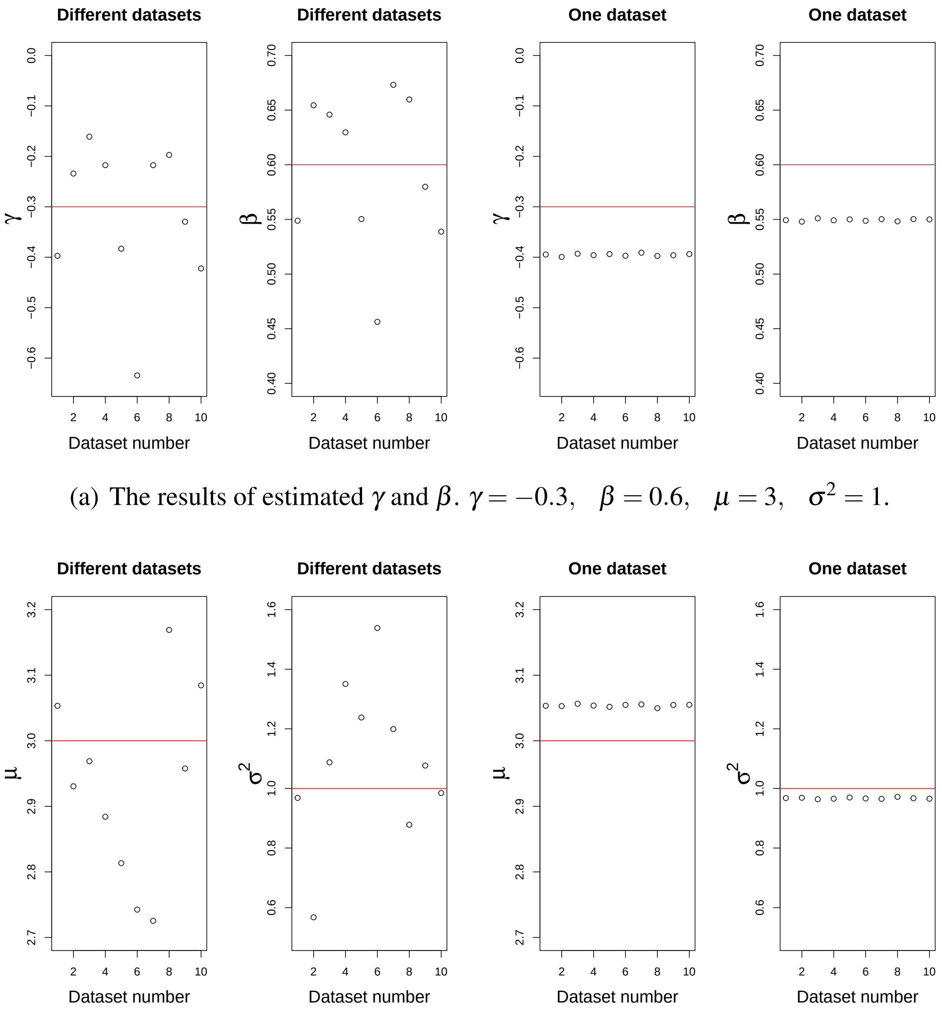
(c) $\gamma = -0.2, \quad \beta = 0.6, \quad \mu = 3, \quad \sigma^2 = 1$

Figure 4.1: The average empirical power and FDR in simulation experiments. The left three figures are the FDR for each setting. The red dotted lines represent the significance level $\alpha = 0.05$. If the FDR is just below 0.05, it means the type I error is well controlled. The right three figures are the power for each setting. The methods with larger power means they can select more true associated SNPs.

From above results, it can be seen that when considering dependence information, the proposed method can control type I error better and have larger power than other two methods, which can show that the proposed method is effective and better. However, when generating simulation data and estimating parameters, it involves a large number of random sampling, which will generate two sources of variability. One is from different fits to the same simulated dataset, which is just from parameter estimation procedure. Another is from fits to different simulated dataset, which are from generating simulation data. We will discuss this in detail in next section.

4.2 Discussion about the bias and variability

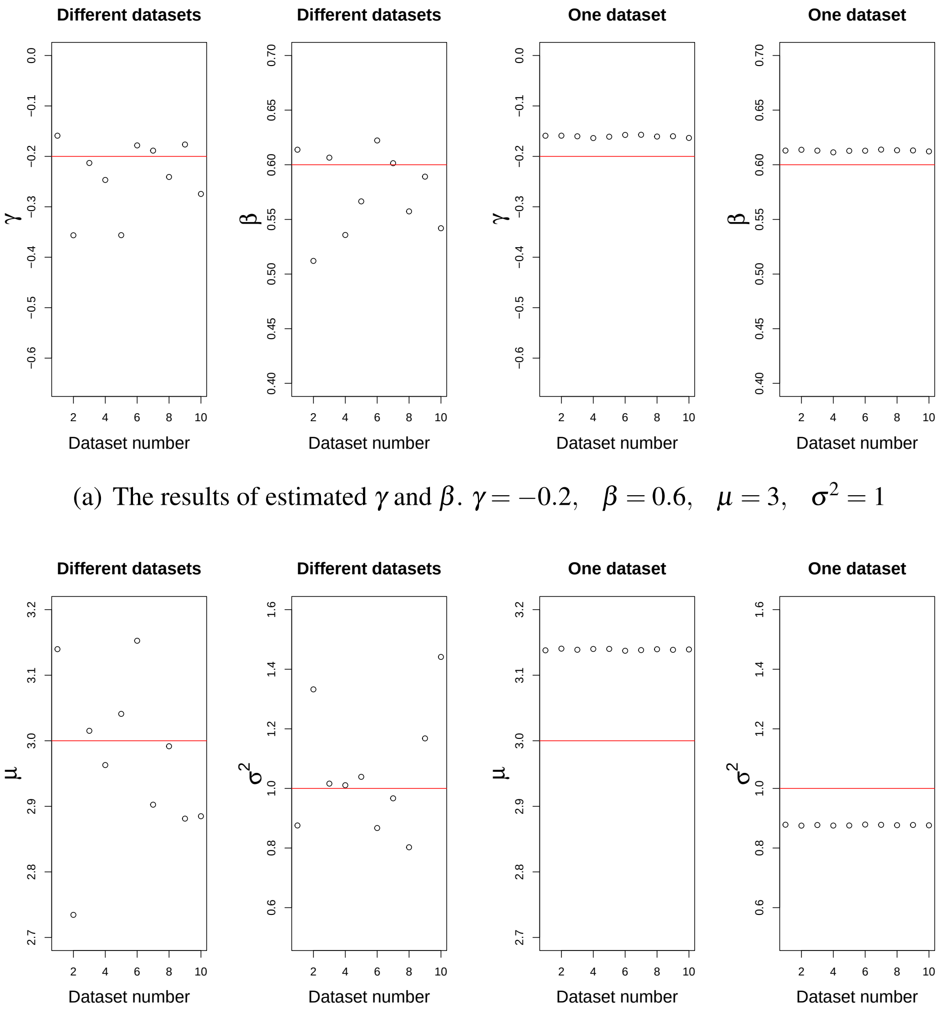
As discussed before, the estimated parameters are the different for 10 simulated datasets. There are two sources of variability. When generating simulation dataset, half of SNPs are randomly assigned with 0 and another half with 1. Then Gibbs sampling are used to update the state of SNPs until it converges to a stationary distribution. So this will lead to different configurations of SNPs for 10 simulations, which will cause different estimates of parameters. Besides this, when using EM algorithm, Gibbs sampling needs to be used to approximate the required quantities. So even if we use the same simulated dataset and run proposed method for many times, we will still get different estimates of parameters. So the 10 fits for different simulated datasets in Section 4.2 combines these two variability. To study the variability from repeated sampling of the data, 10 simulations for the same dataset are conducted and the estimated parameters are plotted, which are expected to have lower variability than 10 fits of different datasets. The results are shown in Figures 4.2 (Setting 2 in Table 4.1) and 4.3 (Setting 3).



(a) The results of estimated γ and β. $\gamma = -0.3, \quad \beta = 0.6, \quad \mu = 3, \quad \sigma^2 = 1.$

(b) The results of estimated of μ and σ^2. $\gamma = -0.3, \quad \beta = 0.6, \quad \mu = 3, \quad \sigma^2 = 1$

Figure 4.2: The estimated parameters for 10 different simulation datasets and 10 identical simulation datasets for Setting 2. For figure (a) and (b), the left two figures are the parameter results for 10 different simulation datasets, while the right two are results from the same dataset.The red line represent the true value.



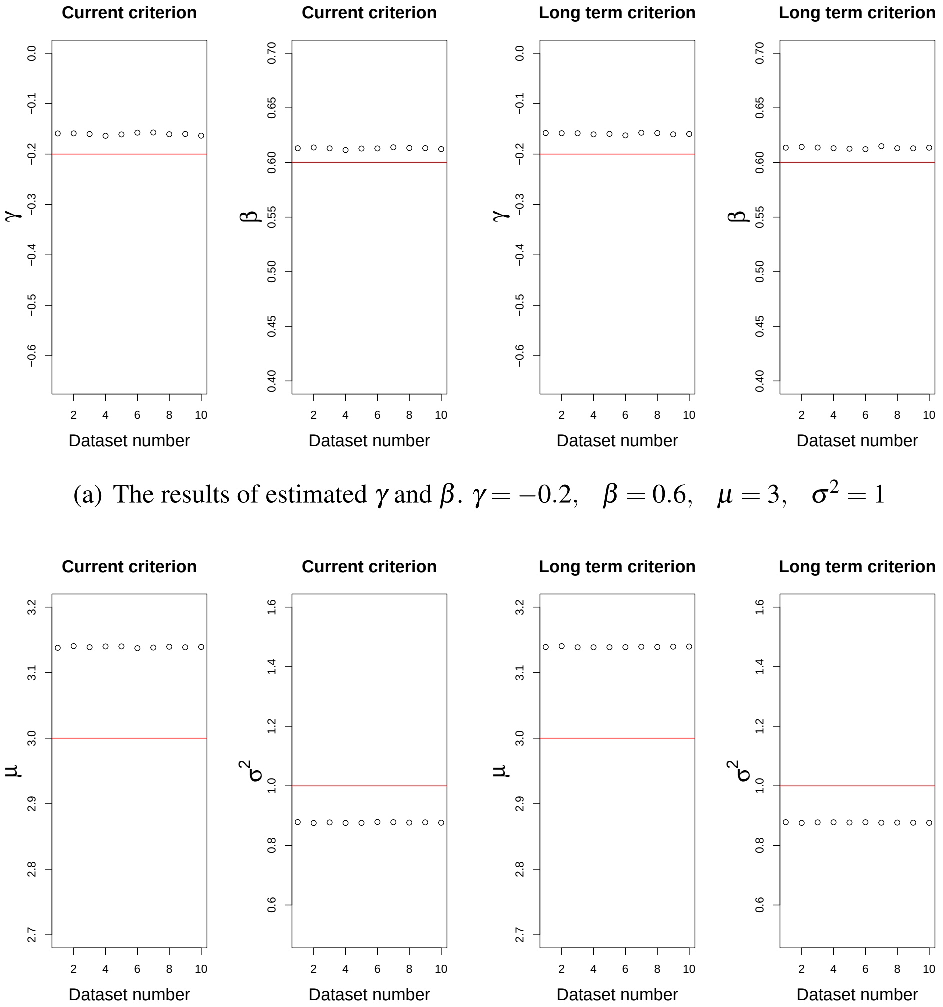
(a) The results of estimated γ and β. $\gamma = -0.2, \quad \beta = 0.6, \quad \mu = 3, \quad \sigma^2 = 1$

(b) The results of estimated of μ and σ^2. $\gamma = -0.2, \quad \beta = 0.6, \quad \mu = 3, \quad \sigma^2 = 1$

Figure 4.3: The estimated parameters for 10 different simulation datasets and 10 identical simulation datasets for Setting 3. For figure (a) and (b), the left two figures are the parameter results for 10 different simulation datasets, while the right two are results for the same dataset.The red line represent the true value.

The main message from Figures 4.2 and 4.3 is clear. Although there is some variability in the estimates resulting from different fits to a single dataset, this variability is negligible compared to the variability in estimates from fits to 10 different datasets. Therefore, we should not be overly concerned about the impact of randomness resulting from the use of Gibbs sampling within the EM algorithm.

A related source of variability is from the convergence criterion used to stop the EM algorithm, on which the random nature of the Gibbs sampler may have an impact. Wu and Ma [88] mentioned one method to reduce this variability, which is changing the convergence criterion. In general, the convergence criterion is set by $(Q_t - Q_{t-1})/|Q_{t-1}| < \varepsilon$, where t is the iteration number. Due to the problem of random simulated samplers, they adopted a relative long-term convergence criterion as $[(Q_t + Q_{t-1}) - (Q_{t-2} + Q_{t-3})]/|Q_{t-2} + Q_{t-3}| < \varepsilon$. To evaluate if the long-term convergence criterion can reduce the variability, 10 fits for a single simulation dataset are used to compare the difference of two convergence criterion. The results are shown in Figure 4.4.



(a) The results of estimated γ and β. $\gamma = -0.2, \quad \beta = 0.6, \quad \mu = 3, \quad \sigma^2 = 1$

(b) The results of estimated of μ and σ^2. $\gamma = -0.2, \quad \beta = 0.6, \quad \mu = 3, \quad \sigma^2 = 1$

Figure 4.4: The estimated parameters for same simulation datasets using two different convergence criterion for Setting 3. The left four figures are the parameter results using $(Q_t - Q_{t-1})/|Q_{t-1}| < \varepsilon$, while the right four are results using $[(Q_t + Q_{t-1}) - (Q_{t-2} + Q_{t-3})]/|Q_{t-2} + Q_{t-3}| < \varepsilon$. $\varepsilon = 0.0001$. The red line represent the true value.

It seems that using long term convergence criterion does not reduce the variability very much, and it will increase the computation cost since this convergence criterion is more difficult to reach. In our experiment, the common convergence criterion $(Q_t - Q_{t-1})/|Q_{t-1}| < \varepsilon$ is applied.

4.3 Simulation study 2

In simulation study 1, the structure of the weight matrix is idealised and may be unlikely observed in real data. To make the simulation data closer to real data, we use the weight matrix from real data to generate simulation data in this study, and then compare the performance of the proposed model with other methods.

In this study, the weight matrix represent the correlation between the 1000 SNPs on Chromsome 8 from the disease Bipolar disorder data set. To describe the structure of LD matrix, we use a heatmap (Figure 4.5) and histogram (Figure 4.6).

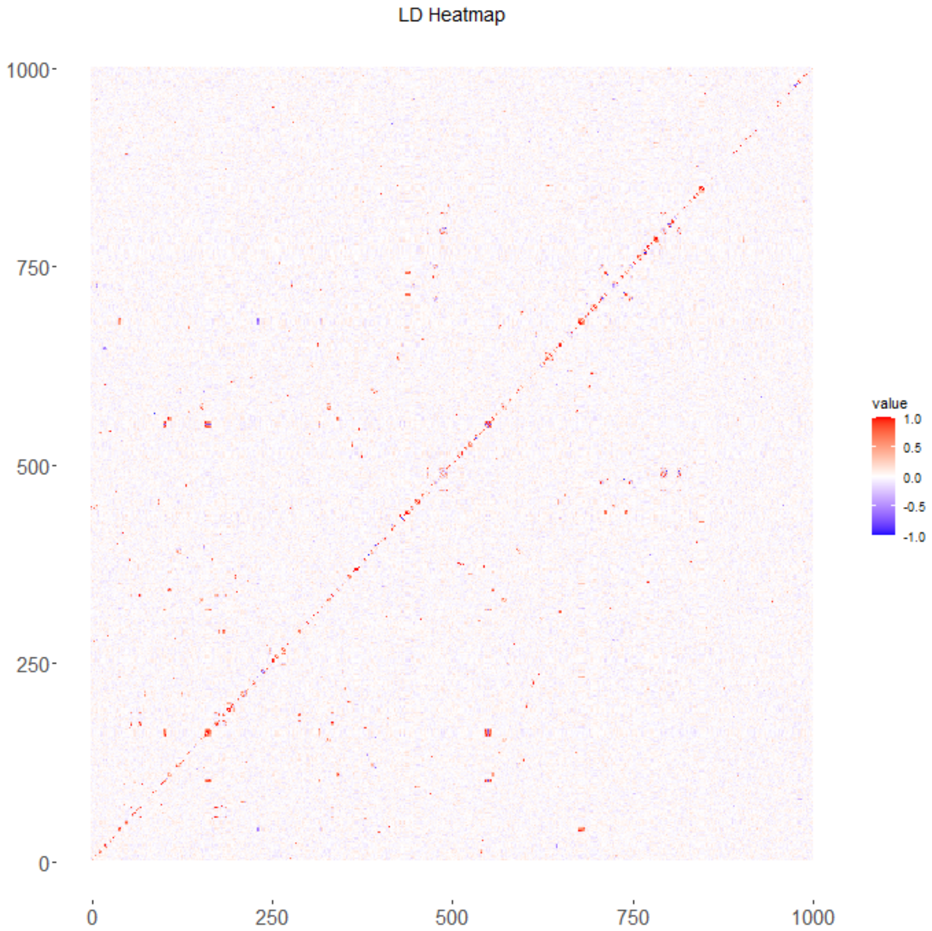


Figure 4.5: The correlation between 1000 SNPs. The blue points represent that the correlation between SNPs is close to -1, while the red points represent that the correlation is close to 1.

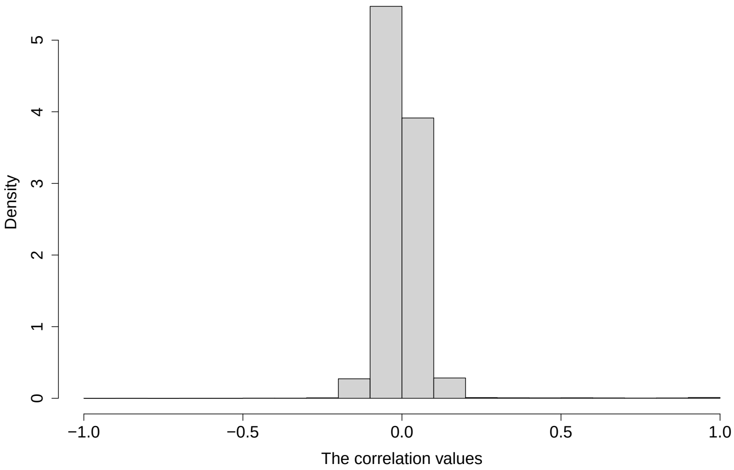


Figure 4.6: The histogram of values in the LD matrix, which is generated from the upper triangle values of the weight matrix.

Figure 4.5 is a repeat of Figure 2.4. As we noted in Section 2.1.5, many of the correlations between pairs of SNPs are neighbouring SNPs, and the correlations are mostly positive. In Figure 4.6, it can be seen that most LD values are between [-0.3,0.3]. To make a comparison between the weight matrix in this study and that in simulation study 1, we generate the summary values in Table 4.3.

Weight matrix	Min	1st quantile	Median	Mean	3rd quantile	Max
simulation study 1	0	0	0	0.004	0	0.7
simulation study 2	-1	-0.03	0	0.002	0.036	1

Table 4.3: The comparison of weight matrix from two studies

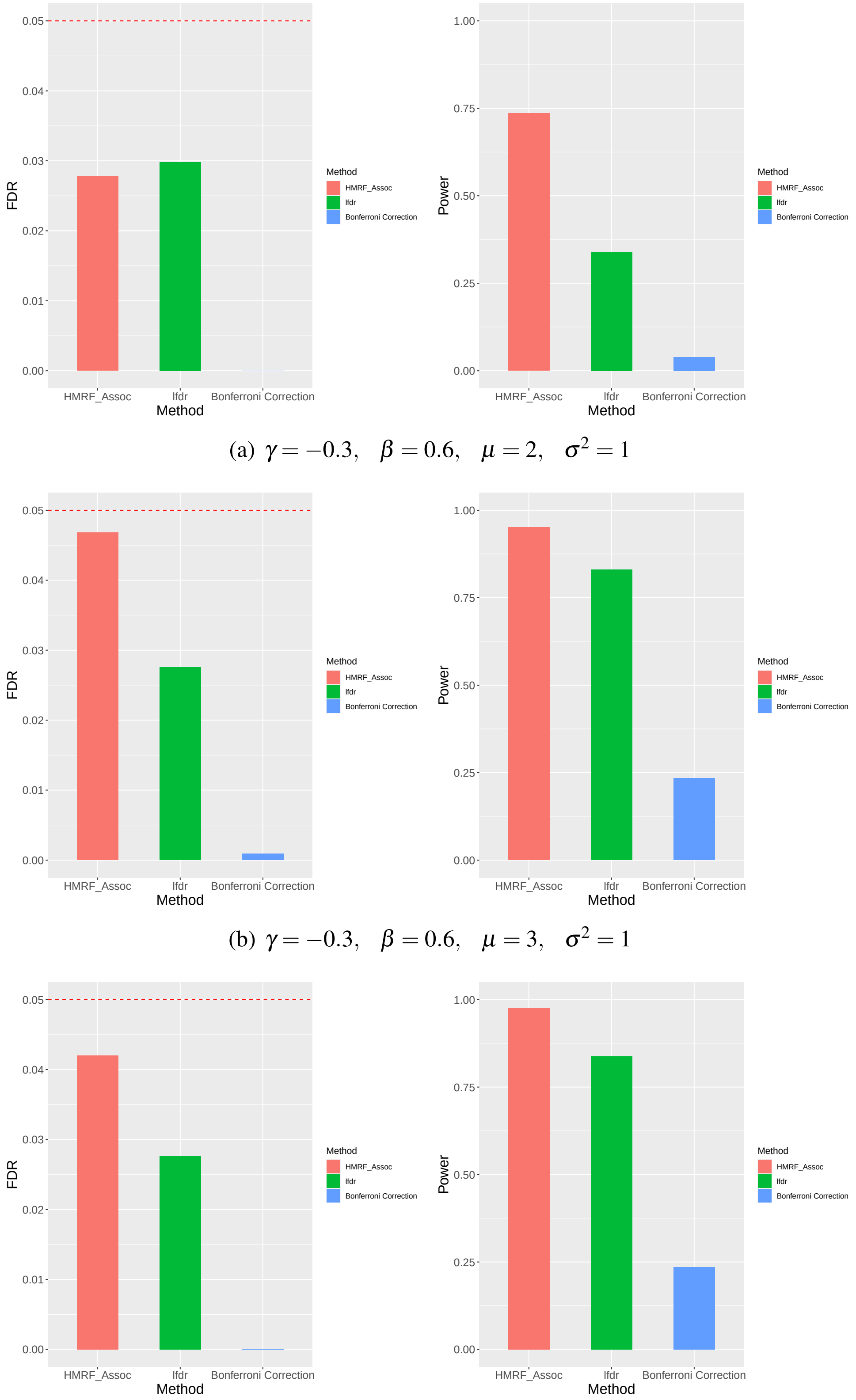
The simulation data generating process and parameters for setting 1 to setting 3 are the same as in simulation study 1 (see Table 4.1). Since the mean values of $\sum_{i=1}^{m} \theta_i$ for these 3 settings are quite larger than that in simulation 1, 3 additional settings are also considered, so that the mean values of $\sum_{i=1}^{m} \theta_i$ are similar to the values in simulation study 1. The description of simulation data and parameter estimation results are summarised in Table 4.4.

Model setting	The mean of $\sum_{i=1}^{m} \theta_i$		γ	β	μ	σ^2
Setting 1	461.5	Mean estimate	0.03	0.80	1.96	1.00
		True value	-0.3	0.6	2	1
Setting 2	460.4	Mean estimate	0.08	1.05	3.01	0.996
		True value	-0.3	0.6	3	1
Setting 3	475.5	Mean estimate	0.07	1.28	3.00	1.03
		True value	-0.2	0.6	3	1
Setting 4	122.6	Mean estimate	0.49	0.69	1.84	1.11
		True value	-3	0.6	2	1
Setting 5	118.6	Mean estimate	0.58	0.73	2.83	1.10
		True value	-3	0.6	3	1
Setting 6	256	Mean estimate	0.23	0.77	2.96	1.09
		True value	-2.5	0.6	3	1

Table 4.4: The parameters for different settings. The $\sum_{i=1}^{m} \theta_i$ is the number of SNPs which are associated with disease. For setting 1 and 2, we have the same γ and β, but different mean values of $\sum_{i=1}^{m} \theta_i$. This is because Gibbs sampling for generating simulation data is random. The mean values are based on 10 simulation data for each parameter setting. The mean estimated values are also based on 10 different simulation datasets.

It can be seen from the above table, except the estimates of γ and β, the estimated values of μ and σ^2 are very close to the true values. In particular, there is evidence of large bias in the estimation of γ in several cases. However, this seems not to have an obviously detrimental effect on the empirical performance of this method, with the combination of estimates for parameters γ and β producing relatively good values of FDR and power. Compared with simulation study 1, the values of $\sum_{i=1}^{m} \theta_i$ are quite larger in this study when parameter settings are the same, which are the results of different weight matrix. For model settings 4, 5 and 6, the $\sum_{i=1}^{m} \theta_i$ is much smaller due to the smaller negative γ.

After estimating parameters, $LIS_i = P(\theta_i = 0 \mid Z)$ are estimated by Gibbs sampling, and associated SNPs with disease are selected based on FDR control procedure. The significance level α is still 0.05. The comparisons of the FDR and power with other two methods are shown in Figure 4.7 and 4.8.

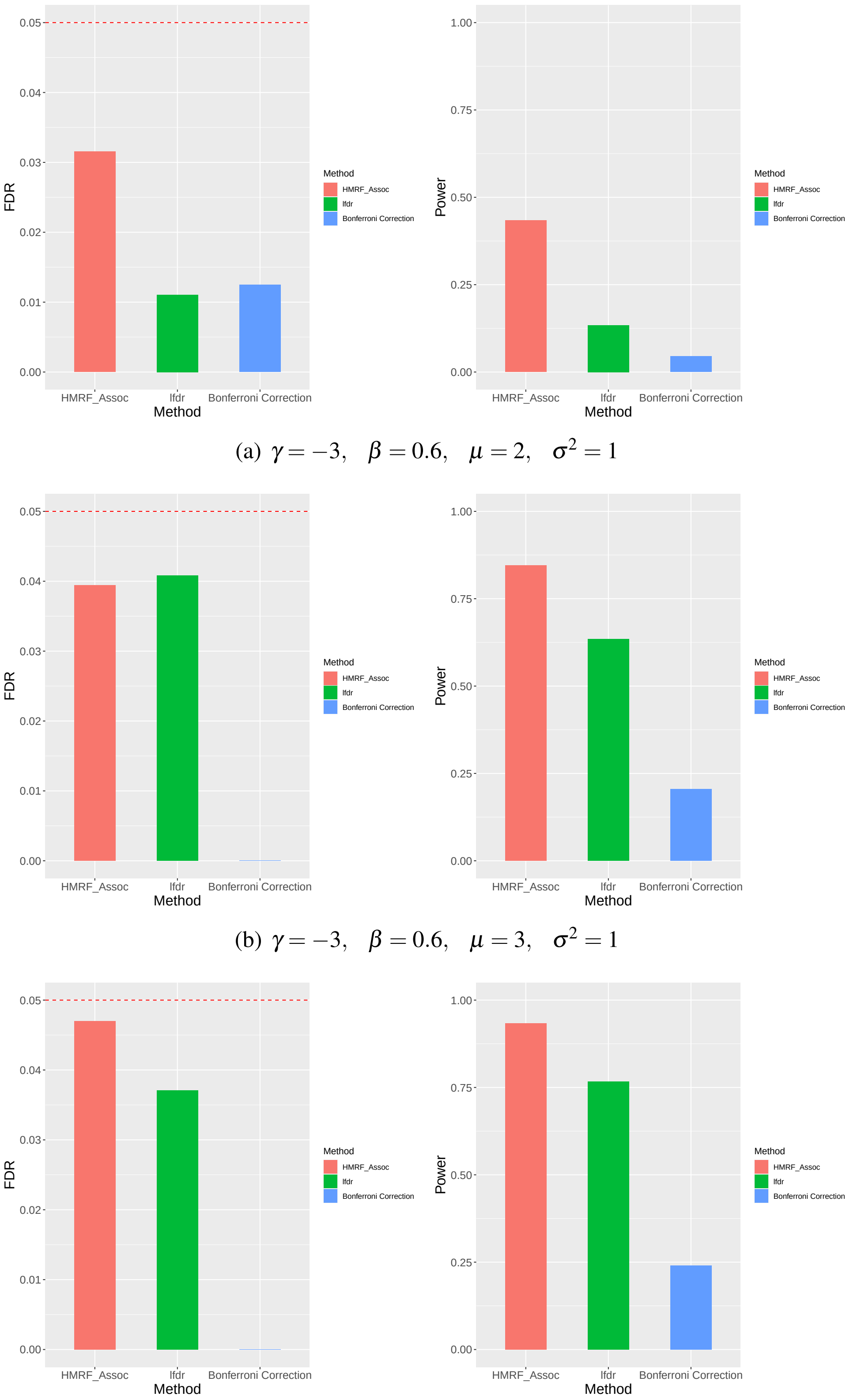


(a) $\gamma = -0.3, \quad \beta = 0.6, \quad \mu = 2, \quad \sigma^2 = 1$

(b) $\gamma = -0.3, \quad \beta = 0.6, \quad \mu = 3, \quad \sigma^2 = 1$

(c) $\gamma = -0.2, \quad \beta = 0.6, \quad \mu = 3, \quad \sigma^2 = 1$

Figure 4.7: The average empirical power and FDR in simulation experiments for Setting 1-3. The left three figures are the FDR for each setting. The red dotted lines represent the significance level $\alpha = 0.05$. If the FDR is just below 0.05, it means FDR is well controlled. The right three figures are the power for each setting. The methods with larger power means they can select more true associated SNPs.



(a) $\gamma = -3, \quad \beta = 0.6, \quad \mu = 2, \quad \sigma^2 = 1$

(b) $\gamma = -3, \quad \beta = 0.6, \quad \mu = 3, \quad \sigma^2 = 1$

(c) $\gamma = -2.5, \quad \beta = 0.6, \quad \mu = 3, \quad \sigma^2 = 1$

Figure 4.8: The average empirical power and FDR in simulation experiments for Setting 4-6, which causes smaller number of SNPs associated with disease. The left three figures are the FDR for each setting. The red dotted lines represent the significance level $\alpha = 0.05$. If the FDR is just below 0.05, it means the FDR is well controlled. The right three figures are the power for each setting. The methods with larger power means they can select more true associated SNPs.

It can be seen from the plots on the left of Figures 4.7 and 4.8 that the average FDR of the proposed method HMRF_Assoc is just below the significance level for all 6 settings, which means FDR are well controlled. The FDR of lfdr methods are conservative compared with the proposed method for most of settings, while the FDR of Bonferroni Correction is conservative for all settings. For the power based on the plots on the right of Figures 4.7 and 4.8, it can be seen that the average power of proposed method is higher than that of other two methods for all settings. Especially when $\mu = 2$, the power of proposed method is much higher than other two methods. From $\mu = 2$ to $\mu = 3$, the power increased a lot, which is consistent with simulation study 1.

4.4 Discussion of the model fitting

There are various aspects of the proposed method that might have an impact on its performance. In the following sections we examine these aspects.

4.4.1 The effect of initial estimates

To speed up the computation time, it may be useful to set a good initial values for parameters. Before initialising the parameters, we need to initialise the states of θ firstly. Since the states of θ are corresponding to whether the SNPs are associated with disease or not, the states are relate to their individual test statistics. Therefore, the initial states of θ are set based on their p-values as follows:

- **Initialize θ:** Based on observed Z values, calculate the corresponding p values, then initialize the configuration of θ:

$$\theta_i = \begin{cases} 1 \text{ If } p_i < P_{\text{thres}} \\ 0 \text{ Otherwise,} \end{cases} \tag{4.2}$$

where P_{thres} represents the threshold values. We will discuss how different values of P_{thres} will affect the results later. For initialising γ and β, recall from chapter 3 that the proposed model is as follows:

$$p(\boldsymbol{\theta};\Phi) \propto \exp\left(\gamma\sum_{i=1}^{m}\theta_i + \beta\sum_{i\sim j} w_{ij}I(\theta_i=\theta_j)\right), \tag{4.3}$$

and the conditional association state for SNP i, given the states of all neighbouring SNPs, is

$$p(\theta_i|\theta_{N_i};\Phi) \propto \exp\left(\gamma\theta_i + \beta\sum_{j\in N_i} w_{ij}I(\theta_i=\theta_j)\right), \tag{4.4}$$

where N_i represents the neighbours of the SNP i on the LD graph.

It can be seen from above model that if we take log on both sides, the relationship between $\log p(\boldsymbol{\theta};\Phi)$ and γ, β are linear, which is like the logistic regression. So we can initialise the γ and β as follows:

- **Initialize γ and β:** Use the following logistic regression model to estimate the values of γ_0 and β_0, which are regarded as the starting values of γ and β.

$$\log\frac{p_i}{1-p_i} = \gamma + \beta\sum_{j\in N_i} w_{ij}(I(\theta_j=1)-I(\theta_j=0)),$$

where $p_i = p(\theta_i = 1|\theta_{N_i};\Phi)$, and θ is regarded as dependent variable Y in the logistic regression model, while $\sum_{j\in N_i} w_{ij}(I(\theta_j=1)-I(\theta_j=0))$ is set as an independent variable X. So the initial value of γ is the estimated intercept, while initial value of β is the estimated coefficient.

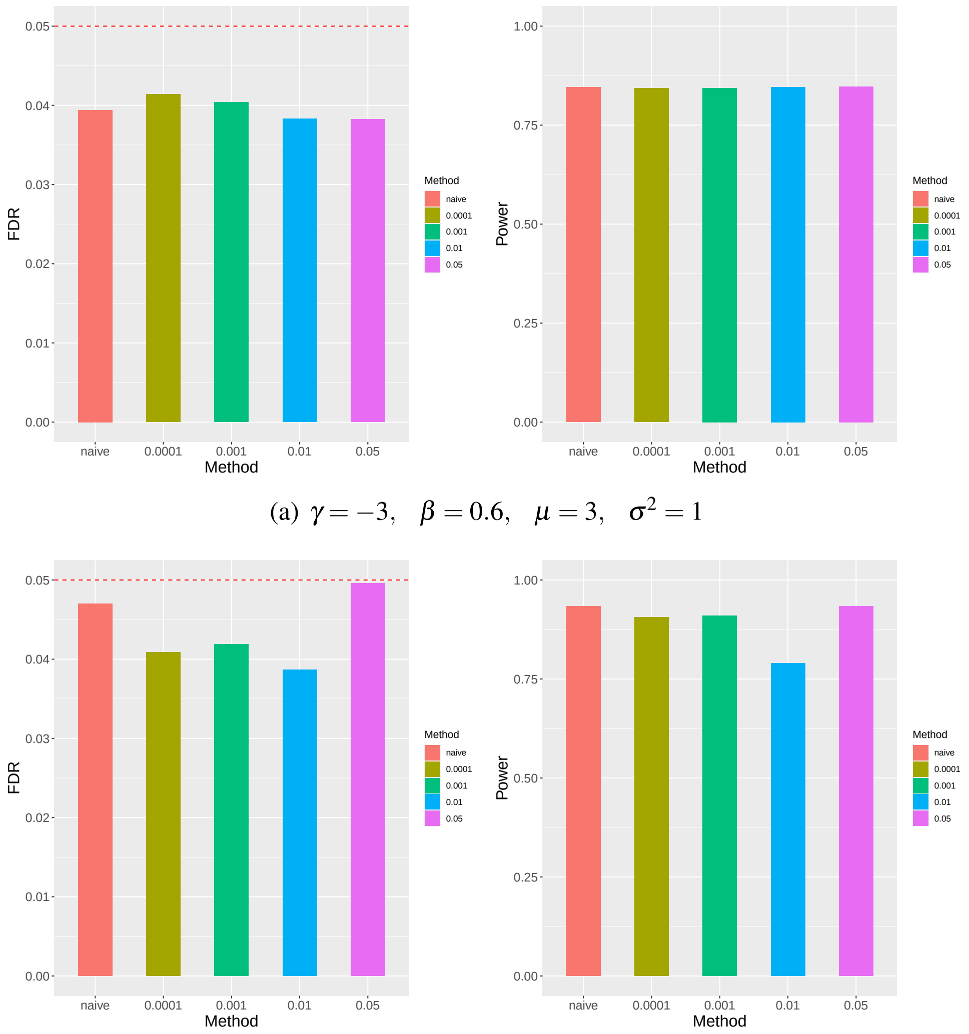
Lastly, we need to initialise the μ and σ^2. Since they are the parameters from Gaussian mixture model, we use the sample mean and sample variance as their initial values, respectively.

- **Initialize μ and σ^2:** The starting values of μ and σ^2 are the sample mean and variance of Z values whose $\theta = 1$, that is:

$$\mu_0 = \frac{1}{n_s}\sum_{i\in S_1} Z_i, \quad S_1 = \{j:\ \theta_j = 1,\ j = 1,\ldots,m\},$$

$$\sigma_0^2 = \frac{1}{n_s-1}\sum_{i\in S_1}(Z_i-\mu_0)^2, \quad S_1 = \{j:\ \theta_j = 1,\ j = 1,\ldots,m\},$$

where n_s is the size of S_1. Different values of p_{thres} will cause different starting values. A larger threshold will cause more values of 1 in the initial configuration of θ_i, while a smaller threshold will generate less values of 1. The current choice is threshold $= 0.0001$. To study the effect of different choices of threshold, we use different values of threshold to run the same simulation model on the 10 simulation datasets. For naive choice, the initial values are set as $\gamma = \beta = \mu = 0$, $\sigma = 0.1$ and $P_{\text{thres}} = 0.0001$. Figure 4.9 shows a comparison of power and FDR.



(a) $\gamma = -3, \quad \beta = 0.6, \quad \mu = 3, \quad \sigma^2 = 1$

(b) $\gamma = -2.5, \quad \beta = 0.6, \quad \mu = 3, \quad \sigma^2 = 1$

Figure 4.9: The comparison of FDR and power for different threshold values for different model settings. The left represents the FDR, while the right is the comparison of power. For naive choice, the initial values are set as $\gamma = \beta = \mu = 0$, $\sigma = 0.1$ and $P_{\text{thres}} = 0.0001$, which represent the results in Figure 4.8 and does not use the methods in this section to estimate initial values.

It can be seen that the all FDR can be well controlled at the level of 0.05.The differences of power are quite small for different threshold values, which may mean that the threshold does not have large effect on the results. The estimated parameters and computation time are in Table 4.5.

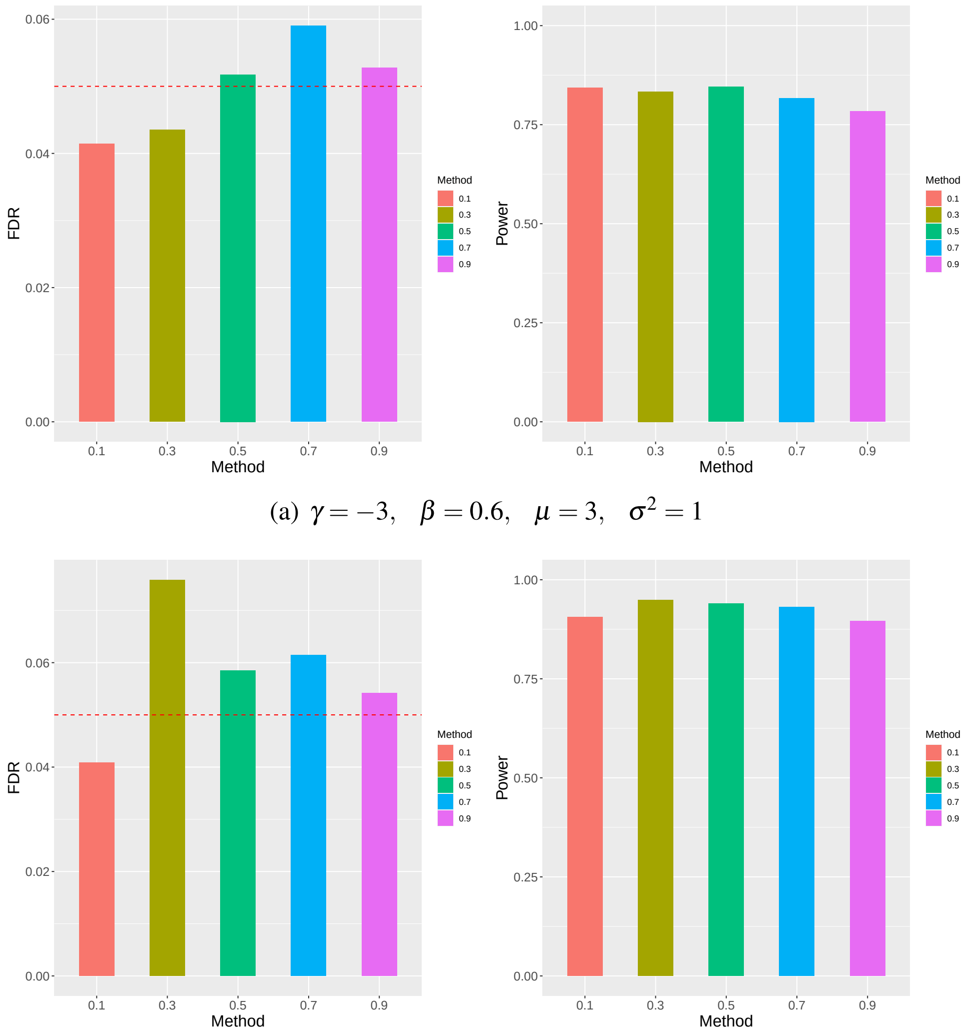
Model	P_{thres}	γ	β	μ	σ^2	time(h))
Setting5	naive	0.58	0.73	2.83	1.10	10.49
	0.0001	0.51	0.72	2.92	0.97	15.07
	0.001	0.54	0.73	2.91	0.98	14.00
	0.01	0.61	0.75	2.92	0.97	10.75
	0.05	0.60	0.75	2.92	0.97	10.90
	true value	-3	0.6	3	1	
Setting6	naive	0.23	0.77	2.96	1.09	24.78
	0.0001	0.07	0.80	2.95	1.09	24.73
	0.001	0.18	0.72	3.02	1.07	34.20
	0.01	-0.04	0.93	3.02	1.05	34.30
	0.05	0.20	0.53	2.96	1.06	33.64
	true value	-2.5	0.6	3	1	

Table 4.5: The comparison for estimate parameters and computation time for different threshold values. For naive choice, the initial values are set as $\gamma = \beta = \mu = 0$, $\sigma = 0.1$ and $P_{\text{thres}} = 0.0001$, which represent the results in Figure 4.8 and does not use the methods in this section to estimate initial values.

It can be seen from Table 4.5 that the differences between estimated parameters are not large. For computation time, when we use the above methods to estimate the initial values, the computation time does not decrease in comparison to the naive choice. Also for different choices of P_{thres}, the computation time does not change much. In the following experiment, we use the method in this section to estimate the initial values and set $P_{thres} = 0.0001$.

4.4.2 The effect of choice of τ in weight matrix

Different τ will generate a different weight matrix in the model (Equation 3.20). A large τ will generate a more sparse weight matrix, which will include less dependence information and may decrease the computation burden, while a weight matrix with small τ will contain more dependence information. To study if the choice of τ for weight matrix has the effect on final results, we compare the FDR and power for different choices of τ in Figure 4.10.



(a) $\gamma = -3, \quad \beta = 0.6, \quad \mu = 3, \quad \sigma^2 = 1$

(b) $\gamma = -2.5, \quad \beta = 0.6, \quad \mu = 3, \quad \sigma^2 = 1$

Figure 4.10: The comparison of FDR and power for different threshold choices of τ. The plots on the left represent the FDR, while the plots on the right give the comparison of power. The threshold value for initial estimates of θ is $P_{thres} = 0.0001$.

Similarly, it can be seen that the difference of power between different choices pf τ is quite small. However, for FDR, except when $\tau = 0.1$, the estimated FDR is larger than 0.05 when τ increases. Then we see the comparison of estimated parameters and computation time in Table 4.6.

Model	τ	γ	β	μ	σ^2	time(h)
Setting5	0.1	0.51	0.72	2.92	0.97	15.07
	0.3	-1.14	1.04	2.95	0.98	4.99
	0.5	-1.16	1.51	2.87	1.04	7.78
	0.7	-1.25	1.68	2.84	1.09	9.12
	0.9	-1.43	1.62	2.84	1.08	8.93
	true value	-3	0.6	3	1	
Setting6	0.1	0.07	0.80	2.95	1.09	24.73
	0.3	-0.39	1.37	2.80	1.37	7.20
	0.5	-0.45	1.50	2.87	1.23	7.62
	0.7	-0.57	2.39	2.89	1.22	6.95
	0.9	-0.70	2.37	2.92	1.17	7.97
	true value	-2.5	0.6	3	1	

Table 4.6: The comparison of estimated parameters and computation time for different choices of τ.

It can be seen that the estimates of μ and σ^2 are quite accurate for all choices of τ. For estimates of γ and β, the estimates are not so accurate for all choices of τ. About the computation time, it can be seen that the time decrease when $\tau > 0.1$. This makes sense since a larger τ will decrease the dependent information and make the weight matrix more sparse. During the computation, zero values are not considered, so a more spare matrix will reduce the computation. However, since larger τ keeps less dependence information, it increased the proportion of false discoveries as in Figure 4.10. So in the following section, τ is set as 0.1. Given that the value of τ may matter, a useful avenue of future research could be to devise a way to estimate the value of τ empirically from the data.

4.4.3 Monitoring convergence and stopping criterion

It can be seen from above results that even though the estimated parameters are not so accurate, it still has a high estimated power. To speed up the computation, it may be useful to consider a more relaxed stopping criterion. Before changing the stopping criterion, we need to study what happens for each iteration during EM algorithm. Since the aim is to recover the configuration of true θ and find more associated SNPs, we will study how quantities relate to final power change during the EM algorithm. Here we consider two quantities: power and $\frac{1}{m}\sum_{i=1}^{m}|P_i^{(k+1)} - P_i^{(k)}|$ for each iteration, where k is the iteration number during the EM algorithm and $P_i^{(k)} = P(\theta_i = 0|Z)$ at iteration k. For each iteration, the configuration of θ will be updated, which could be used to calculate the power for each iteration. In our study, the associated SNPs are identified by estimated posterior probability $LIS_i = P(\theta_i = 0|Z)$, so if the quantity $\frac{1}{m}\sum_{i=1}^{m}|P_i^{(k=1)} - P_i^{(k)}|$ is small, it means that for iteration $k-1$ and k, it will cause similar power and FDR. I illustrate 1 dataset in Setting 6 as an example and the results are in Figure 4.11.

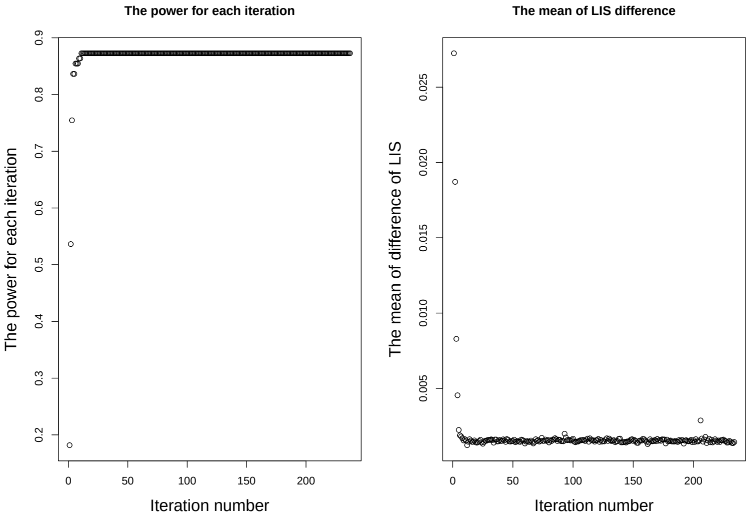


Figure 4.11: The power and $\frac{1}{m}\sum_{i=1}^{m}|P_i^{(k=1)} - P_i^{(k)}|$ for each iteration during EM algorithm.

It can be seen the power and $\frac{1}{m}\sum_{i=1}^{m}|P_i^{(k=1)} - P_i^{(k)}|$ becomes relatively stable after a few iterations. The aim of the study is to find more true associated SNPs, that is to find the true configuration of θ. During the EM algorithm, we will update the states of θ for each iteration. Therefore, we could consider using the difference between the configuration of θ for iteration k and that for iteration $k-1$ as the stopping criterion. If the configuration of θ does not change, then the algorithm could stop. Since it involves random sampling, to reduce the random error, if the configuration of θ does not change for successive l times, the algorithm stops. To check if the new algorithm works, we calculate the FDR and power using the new stopping criterion for different choices of l and compare them with the results using the old stopping criterion in Figure 4.12.

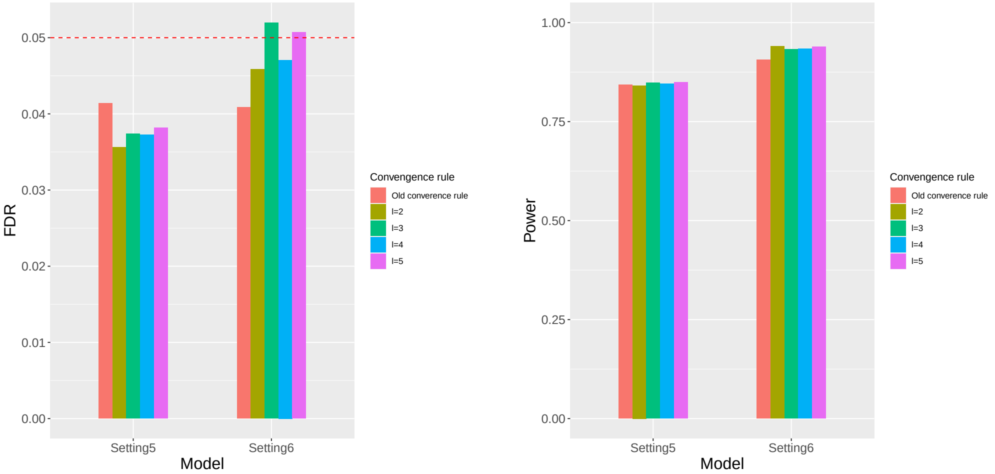


Figure 4.12: The comparison of FDR and power for two convergence criteria.

It can be seen from Figure 4.12 that the difference between power for different criterion is not large. The time comparison is illustrated as follows:

Model	Old convergence rule(h)	New convergence rule(h)			
		$l = 2$	$l = 3$	$l = 4$	$l = 5$
Setting5	15.07	0.27	0.37	0.37	0.48
Setting6	24.73	0.43	1.54	3.36	4.35

Table 4.7: The time comparison for different stopping rule.

It can be seen that as l increase the computation time required increases, but compared with old convergence rule, the computation time of new convergence rule decreases a lot.

4.4.4 The form of Gaussian mixture model

In previous sections, the non-null distribution is assume to be Gaussian distribution with one μ and σ. However, this assumption is hard to be satisfied in real data. To make the non-null distribution approximate the real distribution of Z values well, we consider two components of Gaussian mixture model, which means the distribution of Z values are assumed as following distributions:

$$Z|\theta \sim (1-\theta)N(0,1) + \theta(p_1 N(\mu_1, \sigma_1^2) + p_2 N(\mu_2, \sigma_2^2)), \tag{4.5}$$

where $p_1 + p_2 = 1$. Then the estimates of parameters $(p_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ will satisfy the following updating rules in EM algorithm:

$$p_k^{(t+1)} = \frac{\sum\limits_{i=1}^{m} P(\theta_i = 1|Z) r_i^{(t)}(k)}{\sum\limits_{i=1}^{m} P(\theta_i = 1|Z)}, \tag{4.6}$$

$$\mu_k^{(t+1)} = \frac{\sum\limits_{i=1}^{m} P(\theta_i = 1|Z) r_i^{(t)}(k) Z_i}{\sum\limits_{i=1}^{m} P(\theta_i = 1|Z) r_i^{(t)}(k)}, \tag{4.7}$$

$$(\sigma_k^2)^{(t+1)} = \frac{\sum\limits_{i=1}^{m} P(\theta_i = 1|Z) r_i^{(t)}(k) (Z_i - \mu_k^{(t+1)})^2}{\sum\limits_{i=1}^{m} P(\theta_i = 1|Z) r_i^{(t)}(k)}. \tag{4.8}$$

where $r_i(k) = \frac{p_k N(\mu_k, \sigma_k^2)}{p_1 N(\mu_1, \sigma_1^2 + p_2 N(\mu_2, \sigma_2^2)}$.

The parameter estimation for two components of Gaussian mixture model

Input: Z: observed Z values and weight matrix w;

Output: optimal γ^*, β^*, p_1^*, μ_1^*, σ_1^*, μ_2^*, σ_2^*.

1: **Initialize θ:** Based on observed Z values, calculate the corresponding p values, then initialize the configuration of θ:

$$\theta_i = \begin{cases} 1 \text{ If } p_i < p_{\text{thres}} \\ 0 \text{ Otherwise.} \end{cases} \tag{4.9}$$

Where p_{thres} represents the threshold values.

2: **Initialize γ and β:** Use the following logistic regression model to estimate the values of γ_0 and β_0, which are regarded as the starting values of γ and β.

$$\log \frac{p_i}{1 - p_i} = \gamma + \beta \sum_{j \in N_i} w_{ij} (I(\theta_j = 1) - I(\theta_j = 0)),$$

where $p_i = p(\theta_i = 1|\theta_{N_i}; \Phi)$, and θ is regarded as dependent variable Y in the logistic regression model, while $\sum\limits_{j \in N_i} w_{ij}(I(\theta_j = 1) - I(\theta_j = 0))$ is set as independent variable X. So the initial value of γ is the estimated intercept, while initial value of β is the estimated coefficient.

3: **Initialize $(p_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$:** Firstly, we use K-means clustering to group SNPs with $\theta_i = 1$ into two clusters C_1 and C_2, where $C_1 \cup C_2 = \{i,\ \theta_i = 1\}$. Then we initialize $(p_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$ according to the following equations:

$$p_1^{(0)} = \frac{|C_1|}{|C_1| + |C_2|}, \quad |C_1| \text{ is the size of } C_1,$$

$$\mu_1^{(0)} = \frac{1}{|C_1|} \sum_{i \in C_1} Z_i, \quad \sigma_1^{2(0)} = \frac{1}{|C_1| - 1} \sum_{i \in C_1} (Z_i - \mu_1^{(0)})^2,$$

$$\mu_2^{(0)} = \frac{1}{|C_2|} \sum_{i \in C_2} Z_i, \quad \sigma_2^{2(0)} = \frac{1}{|C_2| - 1} \sum_{i \in C_2} (Z_i - \mu_2^{(0)})^2.$$

4: Generate 5000 Gibbs samplers, where 1500 of Gibbs samplers are regarded as burn-in period. Use Gibbs samplers from $P(\theta_i|Z, \hat{\theta}_{S\backslash i})$ to estimate $E_{\boldsymbol{\phi}^{(old)}}[\boldsymbol{H(\theta)}|Z]$ and $P(\theta_i = 1|Z_i)$. Use Gibbs sampler from $p(\theta_i|\theta_{N_i})$ to estimate $E_{\boldsymbol{\phi}_1}[\boldsymbol{H(\theta)}]$ and $Var_{\boldsymbol{\phi}_1}[\boldsymbol{H(\theta)}]$; $(\boldsymbol{\phi}_1 = (\gamma, \beta))$. The distributions of $P(\theta_i|Z, \hat{\theta}_{S\backslash i})$ and $p(\theta_i|\theta_{N_i})$ are as follows:

$$P(\theta_i|Z, \hat{\theta}_{S\backslash i}) \propto P(Z_i|\theta_i; p_1, \mu_1, \sigma_1, \mu_2, \sigma_2) P(\theta_i|\hat{\theta}_{N_i}; \gamma, \beta) \tag{4.10}$$

$$p(\theta_i|\theta_{N_i}) \propto \exp(\gamma\theta_i + \beta \sum_{j \in N_i} w_{ij} I(\theta_i = \theta_j)). \tag{4.11}$$

5: **update $(p_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$:** The parameters are updated using the following equations:

$$p_k^{(t+1)} = \frac{\sum_{i=1}^{m} P(\theta_i = 1|Z) r_i^{(t)}(k)}{\sum_{i=1}^{m} P(\theta_i = 1|Z)}, \tag{4.12}$$

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{m} P(\theta_i = 1|Z) r_i^{(t)}(k) Z_i}{\sum_{i=1}^{m} P(\theta_i = 1|Z) r_i^{(t)}(k)}, \tag{4.13}$$

$$(\sigma_k^2)^{(t+1)} = \frac{\sum_{i=1}^{m} P(\theta_i = 1|Z) r_i^{(t)}(k) (Z_i - \mu_k^{(t+1)})^2}{\sum_{i=1}^{m} P(\theta_i = 1|Z) r_i^{(t)}(k)}. \tag{4.14}$$

where $r_i(k) = \frac{p_k N(\mu_k,\sigma_k^2)}{p_1 N(\mu_1,\sigma_1^2 + p_2 N(\mu_2,\sigma_2^2)}$.

6: **update the value of $\boldsymbol{\phi}_1 = (\gamma, \beta)$:**

4.1 Maximizing $l(\boldsymbol{\phi}_1)$ is equivalent to solving the equation: $U(\boldsymbol{\phi}_1) = 0$. To reduce the effect of Newton-Raphson method, we find a new ϕ_1 to increase $l_1(\boldsymbol{\phi}_1)$ as follows [83]. A set of decreasing positive values λ_m are introduced:

$$\boldsymbol{\phi}_1^{(t+1,m)} = \boldsymbol{\phi}_1^{(t)} + \lambda_h I(\boldsymbol{\phi}_1^{(t)})^{-1} U(\boldsymbol{\phi}_1^{(t)}), \tag{4.15}$$

where $\lambda_h = 2^{-h}$, $U(\boldsymbol{\phi}_1) = E_{\boldsymbol{\phi}^{(old)}}[\boldsymbol{H}(\boldsymbol{\theta})|Z] - E_{\boldsymbol{\phi}_1}[\boldsymbol{H}(\boldsymbol{\theta})]$ and $I(\boldsymbol{\phi}_1) = Var_{\boldsymbol{\phi}_1}[\boldsymbol{H}(\boldsymbol{\theta})]$. Then the new $\boldsymbol{\phi}_{(t+1)}$ is equal to $\boldsymbol{\phi}_1^{(t+1,m)}$, which is the first one satisfying the following Armijo condition:

$$l_1(\boldsymbol{\phi}_1^{(t+1,m)}) - l_1(\boldsymbol{\phi}_1^{(t)}) \geq \alpha \lambda_h U(\boldsymbol{\phi}_1^{(t)})^T I(\boldsymbol{\phi}_1^{(t)})^{-1} U(\boldsymbol{\phi}_1^{(t)}). \tag{4.16}$$

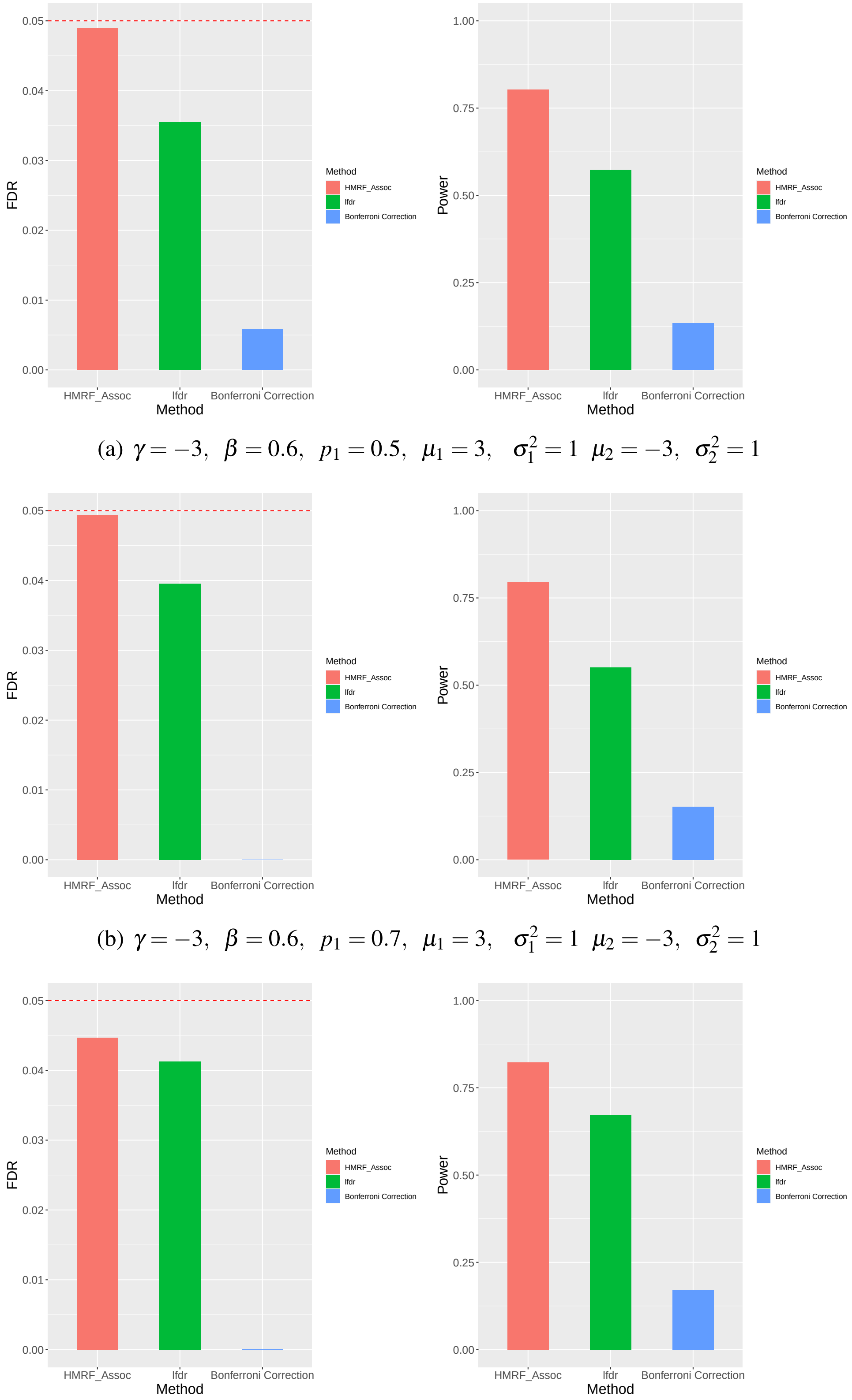
4.2 For $l_1(\boldsymbol{\phi}_1^{(t+1,m)}) - l_1(\boldsymbol{\phi}_1^{(t)})$,

$$\begin{aligned} l_1(\boldsymbol{\phi}_1^{(t+1,m)}) - l_1(\boldsymbol{\phi}_1^{(t)}) \approx{} & \frac{1}{n}(\boldsymbol{\phi}_1^{(t+1,m)} - \boldsymbol{\phi}_1^{(t)})^T \sum_{i=1}^{n} \boldsymbol{H}(\boldsymbol{\theta}^{(t,i)}) \\ & + \log\left(\frac{\sum_{i=1}^{n} \exp\{-\boldsymbol{\phi}_1^{(t+1,m)^T} \boldsymbol{H}(\boldsymbol{\theta}^{(i,\phi_1^{(t+1,m)})})\}}{\sum_{i=1}^{n} \exp\{-\boldsymbol{\phi}_1^{(t)^T} \boldsymbol{H}(\boldsymbol{\theta}^{(i,\phi_1^{(t)})})\}} \right). \end{aligned} \tag{4.17}$$

where $\theta^{(t,i)}$ are Gibbs samplers from $P(\theta_i|Z, \hat{\theta}_{S\backslash i})$, $\theta^{(i,\phi_1^{(t+1,m)})}$ and $\theta^{(i,\phi_1^{(t)})}$ are Gibbs sampler from $p(\theta_i|\theta_{N_i})$.

7: Repeat step 2,3,4 until convergence;

We compare the FDR and power for different parameter settings in Figure 4.13.



(a) $\gamma = -3, \quad \beta = 0.6, \quad p_1 = 0.5, \quad \mu_1 = 3, \quad \sigma_1^2 = 1 \quad \mu_2 = -3, \quad \sigma_2^2 = 1$

(b) $\gamma = -3, \quad \beta = 0.6, \quad p_1 = 0.7, \quad \mu_1 = 3, \quad \sigma_1^2 = 1 \quad \mu_2 = -3, \quad \sigma_2^2 = 1$

(c) $\gamma = -3, \quad \beta = 0.6, \quad p_1 = 0.7, \quad \mu_1 = 3, \quad \sigma_1^2 = 1 \quad \mu_2 = -4, \quad \sigma_2^2 = 1$

Figure 4.13: The average empirical power and FDR in simulation experiments for different parameter settings. The left three figures are the FDR for each setting. The red dotted lines represent the significance level $\alpha = 0.05$. If the FDR is just below 0.05, it means the FDR is well controlled. The right three figures are the power for each setting. The methods with larger power means they can select more true associated SNPs.

It can be seem from above figure that the FDR can be well controlled at significance level for all settings, and for all settings, the proposed methods gain a higher power than other two methods.

4.4.5 The form of null distribution

The differences between the theoretical null distribution and empirical null distribution are shown in Figure 4.14.

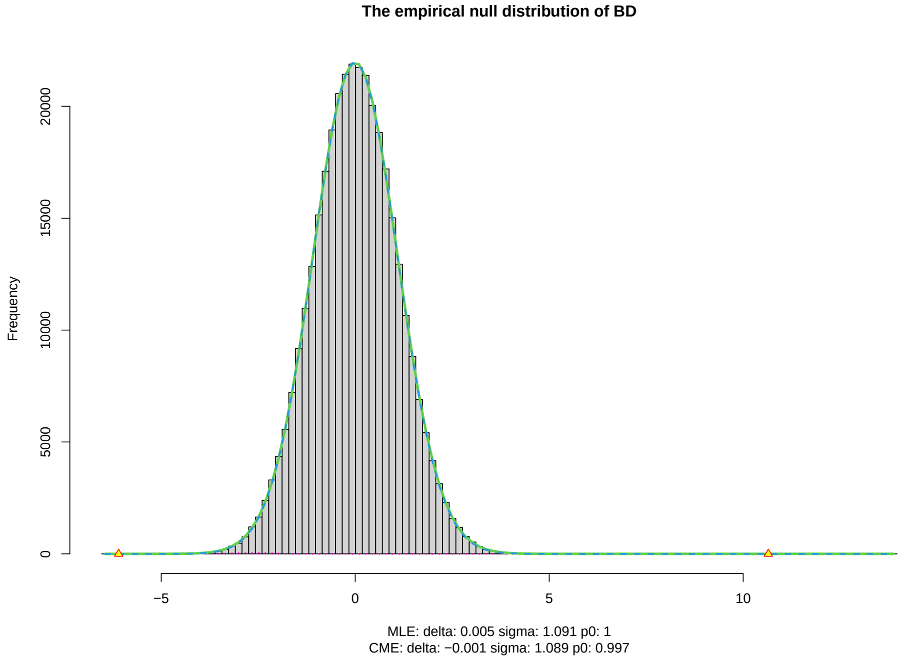
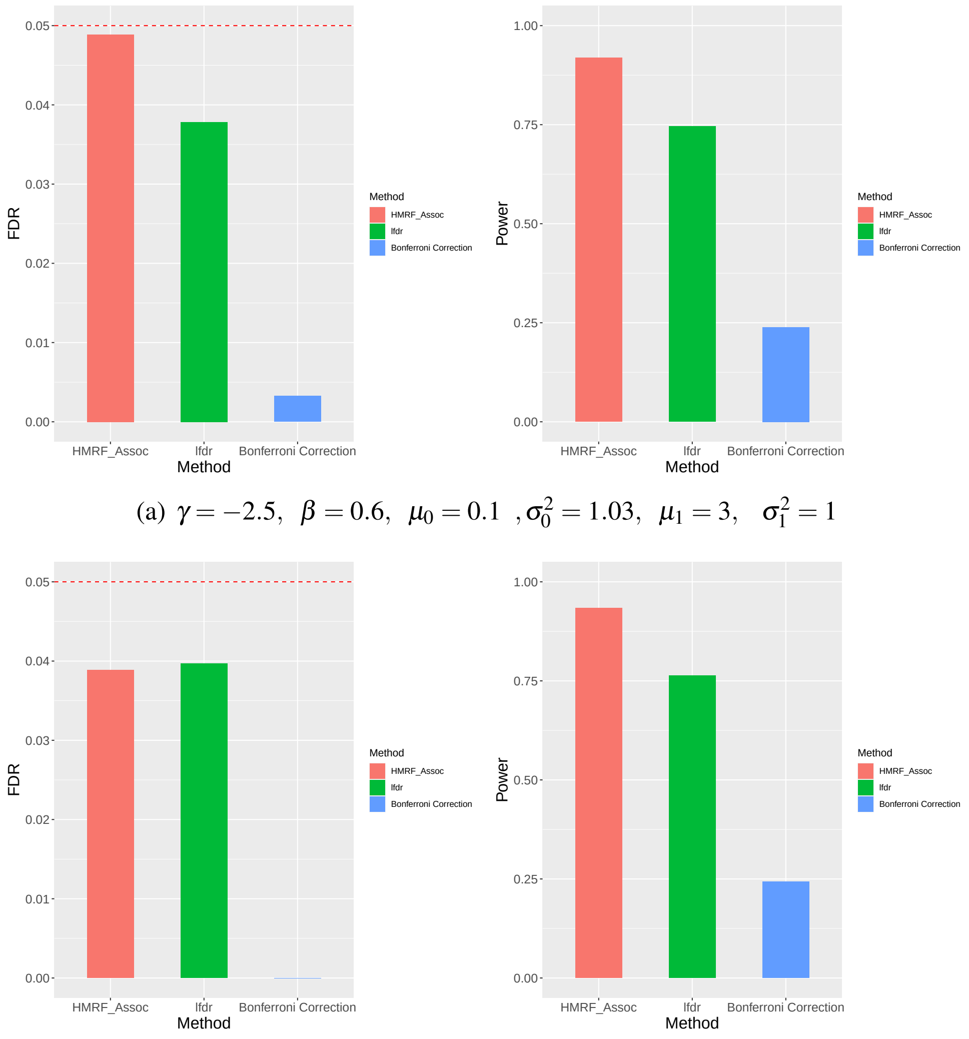


Figure 4.14: The empirical null distribution estimated from R package "locfdr". The "MLE" means maximum likelihood estimation results, while "CME" represents the central matching estimation results. "delta" and "sigma" represent the mean and standard deviation of estimated empirical null distribution.p_0 denotes the proportion of null hypothesis.

The main aim of R package "locfdr" [87] is to calculate local false discovery rates given Z values. It also provides the estimation of empirical null distribution. The basic empirical null idea is that $f_0(z)$ is assumed to be normal but not necessarily standard normal distribution, which means $f_0(Z) \sim N(\delta_0, \sigma_0^2)$. Then the estimates δ_0, σ_0 and the null proportion p_0 are estimated using the histogram data around $z = 0$. "locfdr" provides two estimation methods: central matching method and maximum likelihood estimation methods [17]. If the estimated δ_0 and σ_0 are far away from 0 and 1, respectively, using the empirical null distribution may be better than using standard normal distribution, since the null density is important to the calculation of false discovery rate [17]. It can be seen from Figure 4.14 that the empirical null distribution is close to $N(0,1)$, then it may be sensible to use $N(0,1)$ in our model.

To conduct the sensitive analysis about the form of null distribution, we use the null distribution different from $N(0,1)$. When generating simulation data, the null distributions are set as $N(0.1,1.03)$ and $N(-0.1,1.03)$, respectively. Then during the EM algorithm, we still use $N(0,1)$ to estimate parameters and discover the results. To check if misspecified null distribution have the effect on the results, the FDR and power are shown in Figure 4.15.



(a) $\gamma = -2.5, \;\; \beta = 0.6, \;\; \mu_0 = 0.1 \;\; , \sigma_0^2 = 1.03, \;\; \mu_1 = 3, \;\;\; \sigma_1^2 = 1$

(b) $\gamma = -2.5, \;\; \beta = 0.6, \;\; \mu_0 = -0.1 \;\; , \sigma_0^2 = 1.03, \;\; \mu_1 = 3, \;\;\; \sigma_1^2 = 1$

Figure 4.15: The average empirical power and FDR when the null distribution does not follow N(0,1). The left three figures are the FDR for each setting. The red dotted lines represent the significance level $\alpha = 0.05$. If the FDR is just below 0.05, it means the FDR is well controlled. The right three figures are the power for each setting.

It can be seen that when the null distribution is a little away from the assumption of standard normal distribution, the proposed method could still control the FDR at the significance level. The power of proposed method is still the largest.

4.5 Summary

For Section 4.1 and 4.2, we conducts the simulation study using a weight matrix with AR structure and discusses the sources of bias and variability in the model parameter estimates. Experimental results shows that our proposed method performs better than other methods in identifying more SNPs associated with disease. For Section 4.3, the simulation studies are conducted with a weight matrix based on real data, which demonstrates the superior performance of our proposed method compared to others for all parameter settings. Then in Section 4.4, different aspects which may have an effect on results are discussed. Firstly we discuss the effect of different initial values during EM algorithm and different choices of τ in weight matrix in Sections 4.4.1 and 4.4.2, respectively. Simulation results shows that different initial values have little effect on estimated parameters and computation time, while a larger τ can reduce the computation time effectively, but it will also cause a large FDR. To speed up the computation, a new stopping criterion is considered in Section 4.4.3 and results shows that an improvement in computation time can be achieved. In the meantime, Figure 4.12 shows that the new stopping criterion does not reduce the performance when improving the computation time. In Section 4.4.4, we extend our proposed method from one component non-null distribution to two components of Gaussian mixture model. The parameter estimation procedure is described and simulation studies shows that our proposed method still behave better than others. Finally, we conduct a sensitive analysis that considers a misspecified null distribution. Results demonstrate that our proposed method still shows a superior performance when the null distribution is a little away from $N(0,1)$.

Chapter 5

Application to real data set

5.1 Application to a Bipolar disorder data set

Bipolar disorder is a common disease related to mood abnormalities, which will recur and be accompanied by thinking and behavior disorders. Although the pathogenesis of bipolar disorder is complex, there is conclusive evidence that genetics contribute a lot [89]. To demonstrate our proposed method, we try to identify SNPs associated with several common human diseases using data set from the Wellcome Trust Case Control Consortium (WTCCC). For a large scale experiment, errors may be generated from various procedures such as sample selection bias, sample labelling error and genotyping error, which will cause inaccurate results if we do not remove those data which are likely to be errors. Therefore, before analysing the data set, we use the following quality control procedure in the experiment [90, 91].

- Missing data control: Individuals or SNPs with a high level of missingness can potentially lead to bias or technical problems. So the individuals with more than 10% missing genotypes are removed. The SNPs with more than 10% missing entries are removed.

- Minor allele frequency control: SNPs with a low MAF are difficult to be detect owing to a lack of statistical power. So we just consider the common variants and remove the SNPs with MAF smaller than 0.05.

- Hardy–Weinberg (dis)equilibrium (HWE): The allele and genotype frequencies are assumed to satisfy the HWE law. So the SNPs that do not satisfy Hardy-Weinberg equilibrium are removed. In practice, SNPs for which the p-values in a HWE test are smaller than 0.001 are removed [92].

After applying this quality control procedure, we analysed 27108 SNPs on chromosome 1. It is hard to fit the model for whole data set with 27108 SNPs since the EM algorithm for estimating parameters will take quite a long time, which will exceed the time limitation on UCL's high performance computing platform Myriad. To accelerate the computation, these SNPs are separated into groups of 1000 SNPs according to their locations since nearby SNPs tend to have high correlation. That means only the orders of the SNPs decides their groups. For example, SNP rs1000050 to rs10733059 are in the dataset 1 7and SNP rs10733078 to rs10864698 are in the dataset 2. This will lead to 26 datasets of size 1000 SNPs and 1 dataset of size 1108 SNPs. Then the model is fitted to each of these datasets using proposed method and we get 27 groups of estimated parameters such as $\gamma_1,\ldots,\gamma_{27}$. After getting estimated parameters for each dataset, to leverage LD information within all SNPs when selecting associated SNPs, we use the mean of estimated parameters such as $\bar{\gamma} = \frac{1}{27}\sum_{i=1}^{27}\gamma_i$, $\bar{\beta} = \frac{1}{27}\sum_{i=1}^{27}\beta_i$, $\bar{\mu} = \frac{1}{27}\sum_{i=1}^{27}\mu_i$ and $\bar{\sigma}^2 = \frac{1}{27}\sum_{i=1}^{27}\sigma_i^2$ to estimate the posterior probability $LIS_i = P(\theta_i = 0|Z)$ based on equations (5.1) and (5.2). Unlike the EM algorithm, the calculation for estimating LIS is not time-consuming, so we could consider all LD information within a whole dataset with 27108 SNPs. To complete this step, we use $\bar{\gamma}$, $\bar{\beta}$, $\bar{\mu}$, $\bar{\sigma}^2$ in equations (5.1) and (5.2) and run 12000 Gibbs sampler with first 2000 as the burn-in period.

$$P(\theta_i|Z,\hat{\theta}_{S\backslash i};\bar{\gamma},\bar{\beta},\bar{\mu},\bar{\sigma}^2) \propto P(Z_i|\theta_i;\mu,\sigma)P(\theta_i|\hat{\theta}_{N_i};\gamma,\beta) \tag{5.1}$$

$$p(\theta_i|\theta_{N_i};\bar{\gamma},\bar{\beta}) \propto \exp(\gamma\theta_i + \beta\sum_{j\in N_i} w_{ij}I(\theta_i = \theta_j)). \tag{5.2}$$

Having LIS, we select associated SNPs using the FDR control procedure, and the SNPs with FDR smaller than 0.05 are considered as being associated with Bipolar disorder (see Section 3.3 for details). To compare the proposed method with other selection criterion such as a Bonferroni correction, which identifies a

SNP as associated with the disease if its p-value is smaller than α/m, where m is the number of SNPs. However, when we use Bonferroni correction criterion, which is 1.8×10^{-5} in our study, it does not select any associated SNPs, since the smallest $p-$value is 1.83×10^{-5} in our dataset. So to select some SNPs to make a comparison with proposed method, we choose 10^{-4} as the criterion and select 11 SNPs with p-values smaller than 10^{-4} from all SNPs which are identified using the FDR control procedure of the proposed method, since small p-values are highly likely to be associated with Bipolar disorder. Table 5.1 shows these SNPs and their corresponding posterior probabilities of $P(\theta_i = 1|Z)$, which is $1 - LIS_i$. The FDR of these SNPs are smaller than 0.05, so they are all selected by the proposed method.

SNP	Posterior probability $P(\theta_i = 1\|Z)$	p-values	dataset
rs2989476	1.00	1.83e-05	16
rs396954	1.00	2.12e-05	18
rs1461356	1.00	3.02e-05	9
rs11207909	1.00	4.14e-05	5
rs387176	1.00	4.22e-05	18
rs6691577	1.00	5.96e-05	23
rs1776905	1.00	6.13e-05	12
rs10779279	1.00	7.31e-05	2
rs1187995	1.00	7.68e-05	6
rs2209307	1.00	8.12e-05	14
rs10889189	1.00	8.44e-05	3

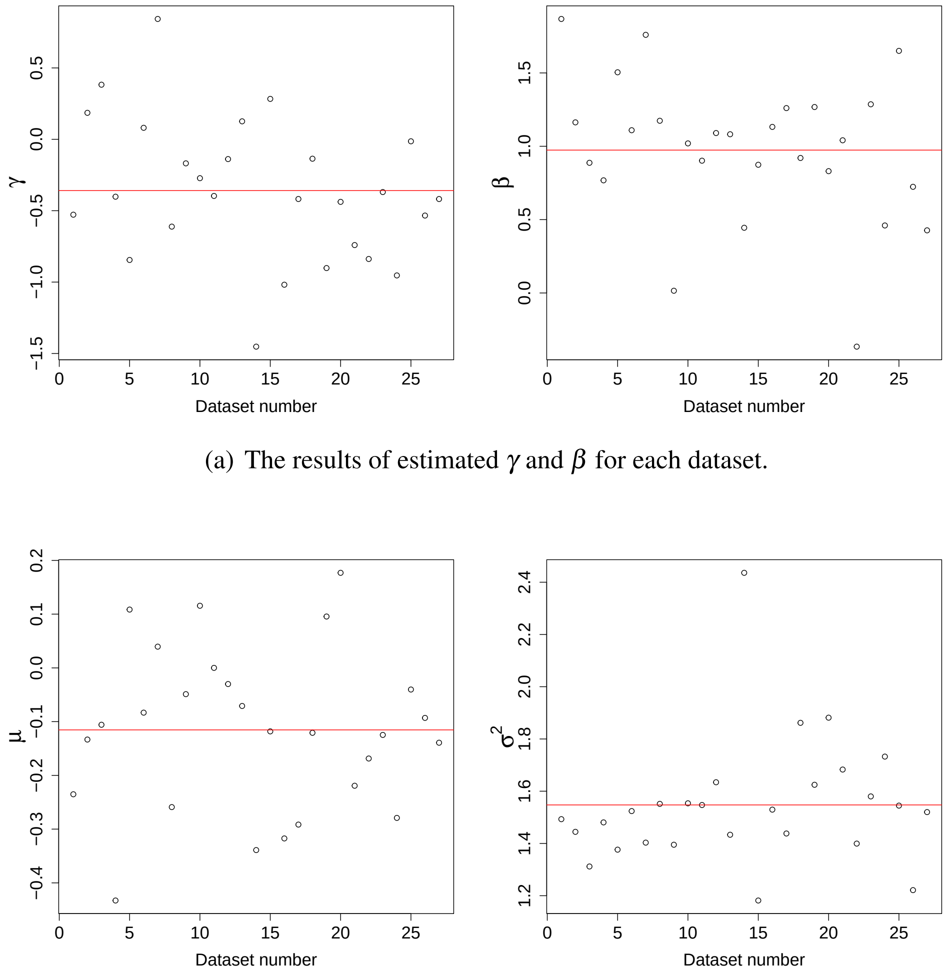
Table 5.1: Results of analysis of Bipolar disorder data set with p-values smaller than 10^{-4}. Posterior probability $P(\theta_i = 1|Z) = 1 - LIS_i$ are calculated using proposed method. The SNPs have been ordered in ascending order of p-values. Column 'dataset' is the dataset number which the SNP belongs when fitting EM algorithm.

SNP rs1187995 shows the association with Bipolar disorder with posterior probability close to 1. No other SNPs have high LD with it, so the posterior probability is mostly determined by itself. In contrast, SNPs rs2989476, rs1461356 and rs10889189 all show large posterior probability of being associated with Bipolar disorder, and they are in high LD with the r^2 between each other ranging from 0.88 to 0.98. Considering that SNP rs2989476 has been shown to be associated with Bipolar disorder by previous studies [93, 89], other two SNPs may work together with SNP rs2989476 to be associated with Bipolar disorder since their high LD and posterior probability.

For SNPs rs2989476 and rs1461356, another SNP, rs555070, also has a high posterior probability of 0.95, which indicates that it may be associated with Bipolar disorder. However, if we use Bonferroni correction, it is hard to be detected since its p-value is 1.5×10^{-4}, which is not small enough for Bonferroni correction criterion since the number of SNPs analysed is large. SNP rs555070 are in high LD with rs2989476 and rs1461356 with r^2 values equal to 0.81 and 0.785, respectively. This shows that the proposed method is better in detecting SNPs associated with Bipolar disorder than single SNP method without considering LD information.

5.2 Discussion

When analysing the above dataset, to accelerate the computation, we separate the whole dataset into 27 smaller datasets and then fit the EM algorithm to each dataset to estimate parameters. When calculating the LIS and selecting associated SNPs, we combine these smaller datasets together and use the mean of estimated parameters so that we can leverage LD information among all SNPs. However, if the estimated parameters have large variance, using the mean of estimated parameter may lead to inaccurate results. To study the effect of using the mean of estimated parameters, we generate a plot of the estimated parameters from the different datasets, which are shown in Figure 5.1.



(a) The results of estimated γ and β for each dataset.

(b) The results of estimated of μ and σ^2 for each dataset.

Figure 5.1: The estimated parameters for each dataset. The red line represent the mean values. The y-axis represent the estimated parameter values.

In Figure 5.1, we can see that the estimated γ for each dataset has both positive and negative values. Most of γ are around -0.35, which is meaningful since the proportion of associated SNPs is usually small. About the effect of different γ on results, consider the equation (5.3), which is derived from equation (3.21) in Section 3.3.2.1.

$$p(\theta_i = 1|\theta_{N_i}) = \frac{\exp(\gamma + \beta \sum_{j \in N_i} w_{ij} I(\theta_j = 1))}{\exp(\beta \sum_{j \in N_i} w_{ij} I(\theta_j = 0)) + \exp(\gamma + \beta \sum_{j \in N_i} w_{ij} I(\theta_j = 1))}. \quad (5.3)$$

If we consider the simplest setting as $\beta = 0$, we can calculate the difference of $p(\theta_i = 1|\theta_{N_i})$ between the largest and smallest γ in Figure 5.1 as follows:

$$\frac{p(\theta_i = 1|\theta_{N_i}, \gamma_{\max})}{p(\theta_i = 1|\theta_{N_i}, \gamma_{\min})} = \frac{\exp(0.85)}{\exp(0) + \exp(0.85)} \Big/ \frac{\exp(-1.45)}{\exp(0 + \exp(-1.45))} \approx 3.69 \quad (5.4)$$

So for positive γ, it is expected to have more associated SNPs than negative γ.

For β, most of β are positive and around 1, which means the dependence between SNPs will affect their states. Suppose that β is positive. When the correlation w_{ij} between SNPs is positive, the tendency for states i and j to have the same state is increased. Similarly, when the correlation between SNPs is negative, the tendency for states i and j to have the same state is decreased. This is what we expect, and therefore it is appropriate that the estimates of β are mostly positive. The estimated μ in Figure 5.1 also have both positive and negative values and most values are around -0.1, which represents the distribution of Z values under alternative hypothesis. For σ^2 in Figure 5.1, all values are larger than 1 and most values are between 1.3 and 1.8. Hence, the distribution of the values of Z under the alternative hypothesis tends to be more widely spread than that under the null hypothesis. The mean value of the estimated value of σ^2 over the datasets (approximately 1.55) corresponds to approximately a 24.5% increase in the standard deviation of Z compared to the null hypothesis, which is 1.

In Figure 5.1, estimated parameters for some datasets are far from the means values. For these datasets, when we use mean values to replace raw parameter values, we may get inaccurate FDR and identify wrong associated SNPs. To quantify how much difference it makes to a given subset of the data if we use the sample mean of the estimated parameter values across all datasets, compared to using the estimated parameter values that are specific to that subset, we use the two groups of parameters to compute the FDR and compare the FDR difference. For example, we set the estimated parameter values for specific subset as $\hat{\gamma} = (\gamma_1, \ldots, \gamma_{27})$, $\hat{\beta} = (\beta_1, \ldots, \beta_{27})$, $\hat{\mu} = (\mu_1, \ldots, \mu_{27})$ and $\hat{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_{27}^2)$. Then the corresponding sample mean values are $\hat{\hat{\gamma}} = \frac{1}{27}\sum_{i=1}^{27}\gamma_i$, $\hat{\hat{\beta}} = \frac{1}{27}\sum_{i=1}^{27}\beta_i$, $\hat{\hat{\mu}} = \frac{1}{27}\sum_{i=1}^{27}\mu_i$ and $\hat{\hat{\sigma}}^2 = \frac{1}{27}\sum_{i=1}^{27}\sigma_i^2$. Then we calculate two set of FDR values. Firstly, we use $(\gamma_i, \beta_i, \mu_i, \sigma_I^2)$, $i = 1,\ldots,27$ on dataset i to calculate FDR values. Hence, we will get 27 sets of FDR values and denote them as $\text{FDR}_1,\ldots,\text{FDR}_{27}$. Then, the mean values $(\hat{\hat{\gamma}}, \hat{\hat{\beta}}, \hat{\hat{\mu}}, \hat{\hat{\sigma}}^2)$ are also applied on 27 datasets to get the FDR values, which are denoted as $\text{FDR}'_1,\ldots,\text{FDR}'_{27}$.Then the FDR differences are $\text{FDR}_i - \text{FDR}'_i, i = 1,\ldots,27$. A plot of FDR differences is shown in Figure 5.2.

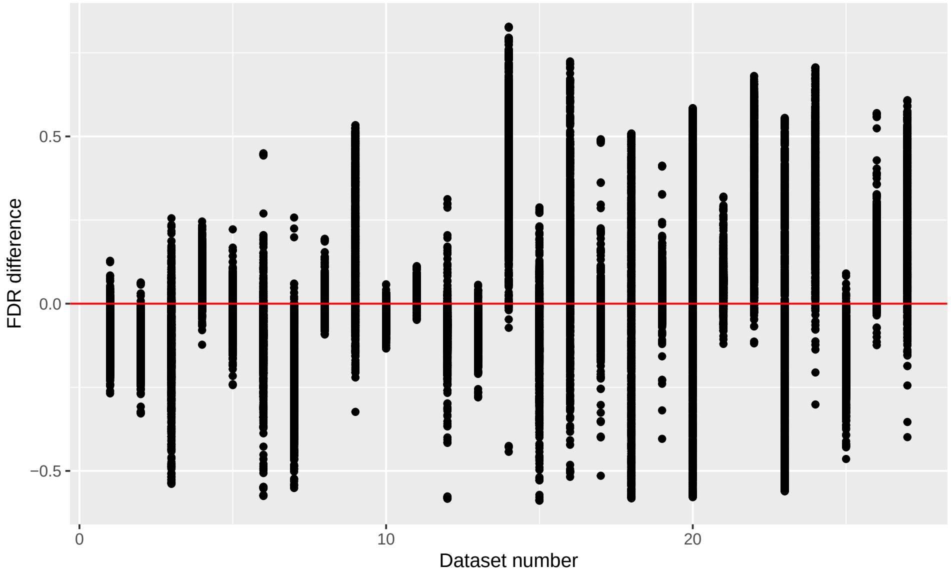


Figure 5.2: The plot of FDR difference between using estimated parameters in individual group and using the mean parameters in the whole group.

In the Figure 5.2, we see that there are datasets such as datasets 10 and 11, for which the FDR differences are close to zero on average and do not vary much. For these datasets, it can be seen from Figure 5.1 that their parameters are quite close to the mean values. For dataset 14, whose estimated γ, β and σ^2 are further from the mean in Figure 5.1, it can be seen that the average FDR difference is also far from the zero and there is a greater variability. Therefore, when estimated parameters have great variability, using the mean values to calculate LIS may make results inaccurate. If we can find method to accelerate the computation or find a better method to combine results of different datasets rather than using the mean, this method may be improved.

Chapter 6

The extension to Gene association

6.1 Introduction

Although numerous common genetic variants associated with complex traits have been identified by genome-wide association studies, these variants can only explain a small proportion of estimated trait heritabilities. This motivated researchers to study the role of rare variants (MAF smaller than 1-5%) and Cohen et al. [94] have found that rare variants are important for complex traits. Because of the low frequency of rare variants, testing for the association between rare variants and complex traits is challenging. Therefore, methods that combine information of multiple rare variants in a genome region like a gene have been considered. The region-based test collects the relevant variants in a region and tests their association with traits jointly [95]. The regions can be a gene or a moving window across the genome. This chapter regards the gene as a testing unit, so we use a gene-based method instead of a region-based method in the following context.

The gene-based methods can be divided into three categories: linear test statistics, quadratic statistics and combined statistics. Linear tests combine the individual SNP effects linearly with certain weights. For example, Morgenthaler and Thilly [48] proposed the "cohort allelic sums test" (CAST) with all weights equal to 1. Madsen and Browning [49] used weights that are inversely proportional to the MAF, which means SNPs with a smaller MAF will have larger weights. Price et al. [96] proposed a threshold so that the weights are larger than 0 only when the MAF is below a specified threshold. Li and Leal [95] proposed a combined multivariate and collapsing method (CMC), which divided SNPs into subgroups based on predefined criteria. Within each group, marker data are collapsed and then a Hotelling's T^2 test is used to analyse the groups of marker data. Quadratic tests combine the individual test statistics using a quadratic form with a weight matrix. For example, the C-alpha method assumes that the distribution of counts follows a binomial distribution under the null hypothesis, and constructs a test statistic to contrast the variance of each observed count with the expected variance [55]. Wu et al. [50] proposed the sequence kernel association test (SKAT), which combines rare variants using a kernel function and then uses a variance component test. SKAT can be regarded as a generalization of the C-alpha test [50]. Yoo et al. [53] proposed a multi-bin linear combination test (MLC), which utilizes the correlation among SNPs to divide SNPs into clusters of highly correlated SNPs using a clique-based algorithm. Then they constructed the test statistic by combining linear combination statistics quadratically.

The linear test statistics and quadratic test statistics are powerful in some situations, but no test statistic can be powerful under all situations. For example, when the directions of effects are different, or when the directions are the same, but the proportion of SNPs associated with trait is small, quadratic test statistics are more powerful than linear test statistics. In contrast, when the proportion of associated SNPs is high and their effect directions are the same, linear test statistics are more powerful [54]. However, real data are complex and people do not have prior information for proportions of causal SNPs and their effect direction. Therefore, some people proposed to combine linear test statistics and quadratic test statistics. The common methods to combine two types of tests are Fisher's method, and the minimum-p value method [52]. Derkach et al. [52] also found that Fisher's method is better than the minimum-p value method when the directions the same, while the minimum-p value method is better when the directions of the effects are different. Lee et al. [51] proposed a data-adaptive optimal test within a class of tests called SKAT-O, which combines linear test statistics and SKAT using a correlation parameter. Pan et al. [97] proposed a class of sum or powered score

(SPU) tests and a data adaptive SPU (aSPU) test so that it can maintain high power for different scenarios. Most of above methods are based on Z values. Some people combined individual tests based on p-values [98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109]. For example, Liu and Xie [108] used a weighted sum of Cauchy transformation to combine individual p values, while Vsevolozhskaya [106] combined top-ranking test statistics using the augmented rank truncation method.

For most of above methods, when they combine the different test statistics into one final test statistics, they actually use part of information contained in different test statistics such as minimum-p value methods. In this chapter, we extend the HMRF methods in Chapter 3 to perform a gene association test. We propose two possible ways. One is to analyse the SNPs in all genes in one HMRF model, and then calculate the posterior probability that a gene is associated with the disease. Finally, we use an FDR control procedure for grouped hypotheses to select associated genes. The second way is to combine summary test statistics in one gene into one test statistic, and then apply the HMRF model on the aggregated test statistics, which uses a gene as one unit in the HMRF model.

6.2 Method

6.2.1 A brief review of some existing methods

We will illustrate some test statistics for the three types of methods. Most of methods are based on a regression model. Here we use the logistic regression model as an example to describe different types of methods. Let $Y_i = 0$ or 1 be an indicator of disease state, where $i = 1, \ldots, n$ and n is the number of subjects. In a case-control study, $Y_i = 1$ represents a case while $Y_i = 0$ represents a control. Let $\boldsymbol{G_i} = (G_{i1}, \ldots, G_{ik})'$ represent the genotypes for the k variants within one gene, where $G_{ij} = 0$,1 or 2 denotes the number of copies of the minor allele at SNP j for subject i and k is the number of SNPs within the gene. Consider the following logistic regression model:

$$\text{Logit}\, P(Y_i = 1) = \beta_0 + \sum_{j=1}^{k} G_{ij}\beta_j, \tag{6.1}$$

where β_0 is an intercept term and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ denotes the vector of regression coefficients for the k SNPs in the gene. Testing whether the gene is associated with the disease corresponds to testing the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$, that is, $\beta_1 = \cdots = \beta_k = 0$.

Many of the existing tests are based on the score vector $S = (S_1, \ldots, S_k)'$ for β in the logistic regression model [97]:

$$\boldsymbol{S} = \sum_{i=1}^{n} (Y_i - p_i)\boldsymbol{G_i}, \tag{6.2}$$

where $p_i = P(Y_i = 1)$ represents the probability of Y_i equal to 1. It can be found that:

$$S_j = \sum_{i=1}^{n} (Y_i - p_i) G_{ij}, \quad j = 1, \ldots, k. \tag{6.3}$$

$Z_j = S_j/sd(S_j)$ denotes the Z values for test statistics and $sd(S_j)$ represents the standard deviation of S_j. The Z values are the summary statistics used in our study. Then the linear test statistics have the following form:

$$T_L = \sum_{j=1}^{k} w_j S_j. \tag{6.4}$$

When $w_j = 1$, this is the cohort allele sums test (CAST) [48] and if w_j is a function of the estimated MAF, then this is the weighted sum method [49]. The form of quadratic statistics is as follows:

$$T_Q = S'AS, \tag{6.5}$$

where A is a $k \times k$ positive definite symmetric matrix. When $A = diag\{a_1, \ldots, a_k\}$, this is the SKAT statistic [50], where a_j depends on MAF.

There are some tests to combine two types of tests such as SKAT-O [51], Fisher's method, and minimum-p method [52]. Their test statistics are as follows:

$$T_{SKAT-O} = \max_{\rho \in [0,1]} (\rho T_L + (1-\rho) T_{SKAT}), \tag{6.6}$$

$$T_{Fisher} = -2\log(p_L) - 2\log(p_Q), \tag{6.7}$$

$$T_{min} = \min(p_L, p_Q), \tag{6.8}$$

where p_L is the p-value for test statistic W_L, while p_Q is the p-value for W_Q. After having p_L and p_Q, combined test statistics can be calculated. For combined test statistics for whom asymptotic approximations are unreliable, p-values are calculated using permutation methods.

Pan [97] proposed another data-adaptive test as follows:

$$T_{\mathrm{SPU}_{(\gamma)}} = \sum_{j=1}^{k} S_j^{\gamma}. \tag{6.9}$$

With different values of $\gamma \geq 1$, they obtained a class of tests. When $\gamma \to \infty$, T_{SPU_∞} is defined as:

$$T_{\mathrm{SPU}_\infty} = \max_{j=1}^{k} |S_j|. \tag{6.10}$$

Then the combining procedure is to calculate minimum-p values:

$$T_{\mathrm{aSPU}} = \min_{\gamma} P_{\mathrm{SPU}_{(\gamma)}}, \tag{6.11}$$

where $P_{\mathrm{SPU}_{(\gamma)}}$ is the p-value for test statistics $T_{\mathrm{SPU}_{(\gamma)}}$ with a specified γ. In practice, they usually choose a series of γ values and calculate different values of $T_{\mathrm{SPU}_{(\gamma)}}$. Finally, the p-value for T_{aSPU}, P_{aSPU} is calculated using permutation methods.

Barnett et. al [110] proposed to use the Generalized Higher Criticism (GHC) test to test the gene association. They defined $S(t)$ as

$$S(t) = \sum_{j=1}^{k} \mathbb{1}(|Z_j| \geq t), \tag{6.12}$$

then the generalized higher criticism test statistics is defined as:

$$\mathrm{GHC} = \sup_{t \geq t_0} \left\{ \frac{S(t) - p * 2\Phi(t)}{\sqrt{\widehat{\mathrm{var}}(S(t))}} \right\}. \tag{6.13}$$

$t_o = 0$ is assumed for simplicity. For the independent case without correlation between SNPs, the GHC statistic reduces to the original higher criticism statistics [111].

Sum et al. [112] proposed a generalized Berk-Jones statistic, which is called as GBJ in the following text. The GBJ statistic is:

$$\begin{aligned} GBJ = \max_{1\leq j\leq k/2} \log & \left[\frac{P\left\{S(|Z|_{(k-j+1)} = j | E(\boldsymbol{Z}) = \hat{\mu}_{j,k} \cdot \boldsymbol{J}_k, \mathrm{Cov}(\boldsymbol{Z}) = \Sigma)\right\}}{P\left\{S(|Z|_{(k-j+1)} = j | E(\boldsymbol{Z}) = 0 \cdot \boldsymbol{J}_k, \mathrm{Cov}(\boldsymbol{Z}) = \Sigma)\right\}} \right] \\ & \times \mathbb{1}\left\{ 2\Phi\left(|Z|_{(k-j+1)}\right) < \frac{j}{k} \right\}, \end{aligned} \tag{6.14}$$

where $S(t)$ is defined in equation (6.12). $\boldsymbol{J}_k^T = (1,\ldots,1)_{1\times k}$, and $\hat{\mu}_{j,k} > 0$ solves the equation:

$$j/k = 1 - \left\{ \Phi(|Z|_{(k-j+1)} - \hat{\mu}_{j,k}) - \Phi(-|Z|_{(k-j+1)} - \hat{\mu}_{j,k}) \right\}. \tag{6.15}$$

GBJ is the maximum of a set of likelihood ratio type tests, whose p-values can be calculated analytically.

Liu and Xie [108] proposed a Cauchy combination test (CCT), which is a weighted sum of Cauchy transformation of individual p-values.

$$T = \sum_{i=1}^{k} w_i \tan\left\{(0.5 - p_i)\pi\right\}, \tag{6.16}$$

where the weights w_i are nonnegative and $\sum_{i=1}^{k} w_i = 1$. In our simulation study, the equal weights for w_i are used, which is the same as their paper.

Different from considering LD structure when combining individual test statistics, Vsevolozhskaya [113] proposed to decorrelate associated statistics and combined them,that is,

$$DOT = \sum_{i=1}^{k} X_i^2, \tag{6.17}$$

where $\boldsymbol{X} = \boldsymbol{HZ}$, and $\boldsymbol{H} = \boldsymbol{EDE'}$. The columns of the matrix $\boldsymbol{E}$ are orthogonalized and normalized eigenvectors of correlation matrix $\boldsymbol{R}$ of $\boldsymbol{Z} = (Z_1,\ldots,Z_k)'$. $\boldsymbol{D} = \frac{1}{\sqrt{\lambda}}\boldsymbol{I}$ and λ is the eigenvalues of $\boldsymbol{R}$.

Wilson [105] proposed a harmonic mean p-value method, which is based on the likelihood ratio test:

$$P_R = \frac{\sum_{i \in R} w_i}{\sum_{i \in R} w_i / p_i}, \tag{6.18}$$

where p_i is the p-value for likelihood ratio test: $p_i = P(r_i \geq R_i | \theta \in \Theta_{M_i})$ and $R_i = \frac{\sup\{P(X|\theta):\theta \in \Theta_{M_i}\}}{\sup\{P(X|\theta):\theta \in \Theta_{M_0}\}}$ measures the evidence for the alternative hypothesis M_i against the null M_0 given the data X. w_i denotes the weights and satisfy $\sum_{i=1}^{k} w_I = 1$, which are set as equal in their paper.

Vsevolozhskaya et al. [106] proposed the augmented rank truncation method (ARTA), which combines top-ranking association statistics.

$$T_l = -\ln W_{l-1} + (l-1)\ln P_{(l)} + G_{\lambda}^{-1}(1 - B_l(P_{(l)})), \tag{6.19}$$

where $W_l = -\sum_{i=1}^{l} \ln P_{(i)}$ and $l < k$. $B_l(\cdot)$ is the CDF of a Beta(l, $k-l+1$) random variable, G_l^{-1} is inverse CDF of Gamma(l,1) and $\lambda = (l-1)E[-\ln P_{(l)}]$. For the choice of l, they calculated the minimum-p value based on various candidate truncation points.

6.2.2 The proposed method

6.2.2.1 Method 1

In the results presented in Sections 6.3 and 6.4, we refer to method 1 as HMRF. Let $G = \{1, \ldots, m\}$ denote the gene index, where m is the number of genes which are analysed one time. For any gene i, suppose that this gene includes k_i SNPs. The hypotheses can be described as follows:

H_{i0} Gene i is not associated with the disease, that is, all k_i SNPs in this gene are not associated with the disease.

and

H_{i1} Gene i is associated with the disease, that is, at least one SNP in these k_i SNPs is associated with the disease.

Then for a given gene, the random variable $\boldsymbol{\theta} = (\theta_{11}, \ldots, \theta_{1k_1}, \ldots, \theta_{m1}, \ldots, \theta_{mk_m})'$ is defined as

$$\theta_{ij} = \begin{cases} 1 & \text{if SNP } j \text{ within gene } i \text{ is associated with the disease} \\ 0 & \text{if SNP } j \text{ within gene } i \text{ is not associated with the disease.} \end{cases} \tag{6.20}$$

Therefore, the hypotheses can be written as:

$$H_{i0}\colon \quad \boldsymbol{\theta_i} = (\theta_{i1}, \ldots, \theta_{ik_i})' = 0,$$

and

$$H_{i1}\colon \quad \prod_{j=1}^{k_i}(1 - \theta_{ij}) = 0.$$

To identify genes that are associated with disease, we apply the HMRF model (equation (6.21))) to all SNPs. This means that we model the dependency between all SNPs using a discrete Markov random field model with the following joint probability function for $\boldsymbol{\theta} = (\theta_{11}, \ldots, \theta_{1k_1}, \ldots, \theta_{m1}, \ldots, \theta_{mk_m})$:

$$p(\boldsymbol{\theta}; \Phi) \propto \exp\left(\gamma \sum_{i=1}^{N} \theta_i + \beta \sum_{i \sim j} w_{i_j} I(\theta_i = \theta_j)\right), \tag{6.21}$$

where θ_i and θ_j are two variables in the vector $\boldsymbol{\theta}$, $N = \sum_{i=1}^{m} k_i$ is the total number of SNPs across all genes. γ and β are two model parameters. w_{ij} represents the LD information between SNP i and SNP j. $\beta > 0$ will encourage SNPs with positive LD values to have similar states.

After estimating parameters using the EM algorithm, we calculate the posterior probability $\text{LIS}_i = P(H_{i0} \text{ is true}|Z) = P(\theta_i = 0|Z)$ and $\text{LIS}_{j|i} = P(\theta_{j|i} = 0|\theta_i = 1, Z)$ using Gibbs sampling, where $\text{SNP}_{i1}, \ldots, \text{SNP}_{ik_i}$ belong to gene i, $i = 1, \ldots, m$ represent the gene index, and Z represents the Z values vector for test statistics of coefficients in the logistic regression model. After having these posterior probabilities, we follow Liu's FDR control procedure for grouped hypotheses [33] to select associated genes (See Section 2.1.7). The steps are as follows:

1. For each gene i, let $\mathrm{LIS}_{(j)|i},\ldots,\mathrm{LIS}_{(k_i)|i}$ be the ranked $\mathrm{LIS}_{j|i}$ values and $H_{i(1)},\ldots,H_{i(k_i)}$ be the corresponding hypotheses. Then we rejected all $H_{i(j)}$, $j=1,\ldots,R_i$, where

$$R_i = \max\left\{ l_i : \frac{1}{l_i}\sum_{j=1}^{l_i} \mathrm{LIS}_{(j)|i} \leq \eta \right\}, \tag{6.22}$$

where $0<\eta\leq\alpha$, and α is the significance level. We chose $\eta=\alpha$, which is the same as Liu's choice.

2. Calculate $\eta_i = \frac{1}{R_i}\sum_{j=1}^{R_i}\mathrm{LIS}_{(j)|i}$, and define $\mathrm{LIS}_i^* = 1-(1-\eta_i)(1-\mathrm{LIS}_i)$, for each gene i. Then let $\mathrm{LIS}_{(1)}^*,\ldots,\mathrm{LIS}_{(m)}^*$ be the ranked LIS_i^* values and $H_{(1)},\ldots,H_{(m)}$ be the corresponding hypotheses. Then the testing procedure was to reject all $H_{(i)}$, $i=1,\ldots,l$, where

$$l = \max\left\{ h : \frac{\sum_{i=1}^{h} R_{(i)}\mathrm{LIS}_{(i)}^*}{\sum_{i=1}^{h} R_{(i)}} \leq \alpha \right\}, \tag{6.23}$$

where $R_{(i)}$ is the value of R in equation (6.22) for the gene that corresponds to $\mathrm{LIS}_{(i)}^*$.

6.2.2.2 Method 2

In the results presented in Sections 6.3 and 6.4, we refer to method 2 as HMRF_gene. If we partition N SNPs into m clusters, we use an $N\times m$ matrix J to represent SNP assignments, where $J_{ij}=1$ if ith SNP is assigned into the jth cluster, otherwise $J_{ij}=0$. For Z values of all SNPs across all genes $Z=(Z_{11},\ldots,Z_{1k_1},\ldots,Z_{m1},\ldots,Z_{mk_m})'$, where Z_{ij} represent the Z values for coefficients β_{ij} in the logistic regression model for SNP j within gene i. The aggregated test statistic for one gene is:

$$T = J^T Z, \tag{6.24}$$

where $T=(T_1,\ldots,T_m)$ represents the aggregated test statistics for each gene. Then we need to calculate the correlation matrix for the new test statistics vector T. For example, suppose that we have 2 genes and one gene includes 2 SNPs while another includes 3 SNPs. Then we have the Z values $Z = (Z_{11}, Z_{12}, Z_{21}, Z_{22}, Z_{23})'$, and we will have aggregated test statistic for two genes:

$$T_1 = Z_{11} + Z_{12}, \qquad T_2 = Z_{21} + Z_{22} + Z_{23}, \tag{6.25}$$

The correlation between T_1 and T_2 can be estimated:

$$\begin{aligned} cor(T_1, T_2) &= cor(Z_{11} + Z_{12}, Z_{21} + Z_{22} + Z_{23}) \\ &= \frac{cov(Z_{11} + Z_{12}, Z_{21} + Z_{22} + Z_{23})}{\sqrt{(var(Z_{11} + Z_{12}))}\sqrt{var(Z_{21} + Z_{22} + Z_{23}))}}. \end{aligned} \tag{6.26}$$

In practice, the LD values are estimated from sample genotype data for each gene. We sum genotype data within one gene, then calculates the correlation between genes. The hypothesis can be written as:

H_{i0} Gene i is not associated with the disease

and

H_{i1} Gene i is associated with the disease.

Then for a given gene, the random indicator variable θ is defined as

$$\theta_i = \begin{cases} 1 & \text{if gene } i \text{ is associated with the disease} \\ 0 & \text{if gene } i \text{ is not associated with the disease.} \end{cases} \tag{6.27}$$

Since we assume that the Z-scores are conditionally independent given the hidden indicators, that is:

$$P(Z|\theta) = \prod_{i=1}^{m} P(Z_i|\theta_i), \tag{6.28}$$

we will have $P(T|\theta) = \prod\limits_{i=1}^{n} P(T_i|\theta_i)$. For an arbitrary SNP, the distribution of Z values are assumed to follow mixture distribution:

$$Z|\theta \sim (1-\theta)N(0,1) + \theta N(\mu, \sigma^2). \tag{6.29}$$

Then for one gene j, the distribution of T values are assumed as follows:

$$T_j|\theta \sim (1-\theta)N(0,k_j)+\theta N(\mu_1,\sigma_1^2). \tag{6.30}$$

Where μ_1 and σ_1^2 are parameters of the distributions for each gene under alternative hypothesis. Since we assume that Z-scores are conditional independent given hidden indicators, when hidden state $\theta_i = 0$, the corresponding Z values in gene i all have the standard normal distribution $N(0,1)$, then T_i also follows normal distribution with its mean being the sum of means (equal to 0), and its variance being the sum of variances (equal to k_j). While under the alternative hypothesis $\theta_i = 1$, at least one Z value within gene i follows a normal distribution with mean μ and variance σ^2. Then similarly T_i also follows a normal distribution with unknown mean μ_1 and variance σ_1^2, since the number of SNPs associated with disease within gene i is unknown. The mean μ_1 and variance σ_1^2 can be estimated from data.

Then we can apply HMRF models for all genes. After estimating parameters, we can estimate the posterior probability of $LIS_i = P(\theta_i = 0|T)$ using Gibbs sampling. Then to select associated SNPs with the disease, we use the FDR control procedure to select associated genes.

6.2.3 Data presentation

As mentioned before, when the directions of effects are different, or when the directions are the same, but the proportion of SNPs associated with trait is small, quadratic test statistics are more powerful than linear test statistics. In contrast, when the proportion of associated SNPs is high and their effect directions are the same, linear test statistics are more powerful [54]. To study the gene structure in real data, we analysed 17340 genes for low-density-lipoprotein cholesterol (LDL) [114]. In particular, we compare the strategies of summarising Z values within a gene using their sum or by their largest absolute value. The results are shown in Table 6.1.

Gene	nsnp	≥ 5.3	$\leq(-5.3)$	max	min	sum	abs_max	sum_sign
A1BG	10	0	0	1.0	-1.8	-3.0	-1.8	TRUE
A1CF	5	0	0	3.6	-2.2	3.1	3.6	TRUE
A2M	33	0	0	3.4	-1.6	4.1	3.4	TRUE
A2ML1	65	0	0	2.5	-2.3	3.3	2.5	TRUE
A4GALT	6	0	0	1.2	-1.3	1.9	-1.3	FALSE
A4GNT	10	0	0	1.5	-1.4	-1.5	1.5	FALSE
AAAS	21	0	0	2.5	-2.5	-3.7	2.5	FALSE
AACS	15	0	0	2.0	-1.2	0.2	2.0	TRUE
AADAC	13	0	0	1.5	-0.8	1.6	1.5	TRUE
AADACL2	10	0	0	2.8	-0.6	8.2	2.8	TRUE

Table 6.1: The summary of Genes for disease LDL. The "nsnp" represents the number of SNPs within one gene. The columns "≥ 5.3" and "$\leq(-5.3)$" indicate the number of SNPs whose Z values are larger than 5.3 or smaller than -5.3 within one gene. The colums "max" and "min" show the maximum and minimum Z values, respectively. The "sum" is the sum of Z values. The column "abs_max" represents the maximum Z values of all SNPs within one gene, but keeping its sign. The last column "sum_sign" indicates whether the sum of Z values have the same sign with column "abs_max".

In Table 6.1, the "nsnp" represents the number of SNPs within one gene. The columns "≥ 5.3" and "$\leq(-5.3)$" indicate the number of SNPs whose Z values are larger than 5.3 or smaller than -5.3 within one gene. Here the value 5.3 is chosen because the p-value for quantile -5.3 is 5.8×10^{-8}, which is close to the general GWAS significance level 5×10^{-8}. The columns "max" and "min" show the maximum and minimum Z values, respectively. The "sum" is the sum of Z values. The column "abs_max" represents the maximum absolute Z values of all SNPs within one gene, but keeping its sign. That means, if $|\text{max}| > |\text{min}|$, abs_max $=$ max, while abs_max $=$ min if $|\text{max}| < |\text{min}|$. For example, for gene "A1BG", its absolute maximum value is 1.8, which is equal to $|\text{min}|$. Then abs_max $= -1.8$, keeping its negative sign. The last column "sum_sign" indicates whether the sum of Z values have the same sign with column "abs_max". If the sign is the same, the entry is "TRUE".

Although none of the genes in Table 6.1 has Z values that are greater than 5.3 in absolute value, among the 17340 genes, 200 genes have the maximum absolute values larger than 5.3, such as some of those genes in Table 6.2, which are possibly associated with LDL. Within these 200 genes, there are 72 genes, whose absolute sum of Z values are smaller than absolute Z values. Part of these genes are shown in Table 6.2.

Gene	nsnp	>=5.3	<=(-5.3)	max	min	sum	abs_max	sum_sign
ACOX1	17	1	0	5.5	-1.5	-0.4	5.5	FALSE
AMIGO1	6	0	1	3.2	-7.2	-2.6	-7.2	TRUE
APOA5	12	1	1	8.4	-9.3	3.0	-9.3	FALSE
APOB	138	7	8	35.6	-18.6	-10.7	35.6	FALSE
APOC1	6	1	1	40.2	-74.2	-33.0	-74.2	TRUE
APOE	7	2	2	44.3	-84.3	-41.4	-84.3	TRUE
APOH	14	1	0	12.8	-2.2	7.4	12.8	TRUE
BBS1	17	0	1	1.4	-5.6	-5.3	-5.6	TRUE
BUD13	23	1	1	6.7	-6.8	-0.7	-6.8	TRUE
CBLC	7	0	1	2.1	-5.5	1.0	-5.5	FALSE

Table 6.2: Part of genes whose absolute sum of Z values are smaller than absolute Z values.

For gene "ACOX1" in Table 6.2, the sum of Z values is -0.4, while the maximum values is 5.52, which means that the sum of Z values mitigates the effect of association. In this case, using sum of test statistics as the summary may be also misleading, since the sum of test statistics make it more difficult to identify the association.

Within these 200 genes which have the maximum absolute values larger than 5.3, the sum of test statistics do not have the same sign with maximum absolute Z values for some genes, which is shown in Figure 6.1.

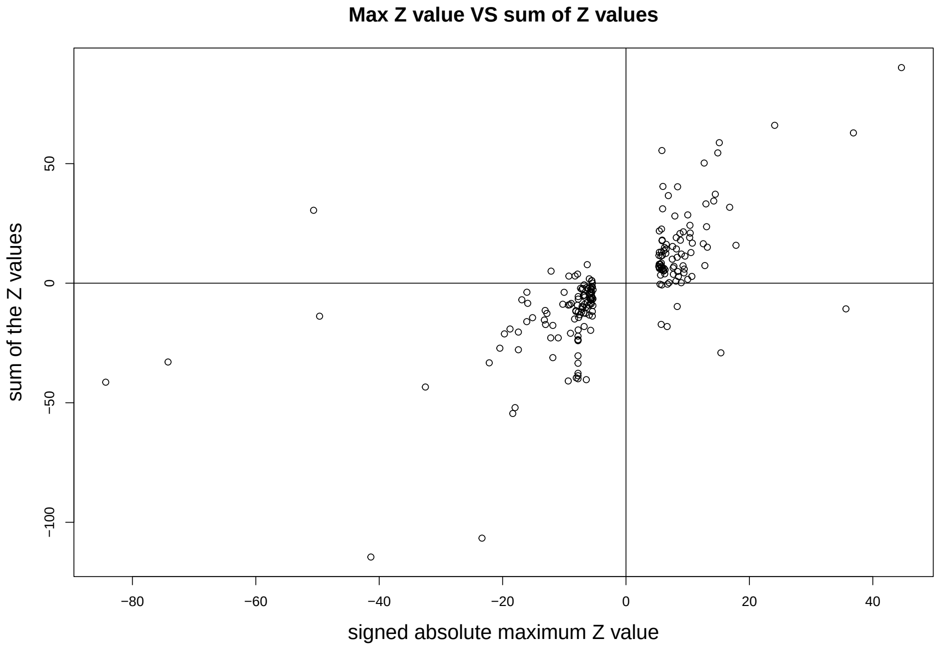


Figure 6.1: The sum of the *Z* values is plotted against the signed absolute maximum *Z* value and only genes that have maximum absolute values larger than 5.3 are included in the plot.

There are 18 genes with sum_sign equal to FALSE, which means that the effect direction of sum of *Z* values is different from raw *Z* values. In this case, using sum of test statistics as the summary may be misleading.

Futhurmore, in Table 6.2, gene "APOB" has a large maximum *Z* values (35.6) and minimum values (-18.6), which may indicate that the gene "APOB" is associated with LDL in different directions. A plot of the the *Z* values for each of the SNPs in gene "APOB" is shown in Figure 6.2.

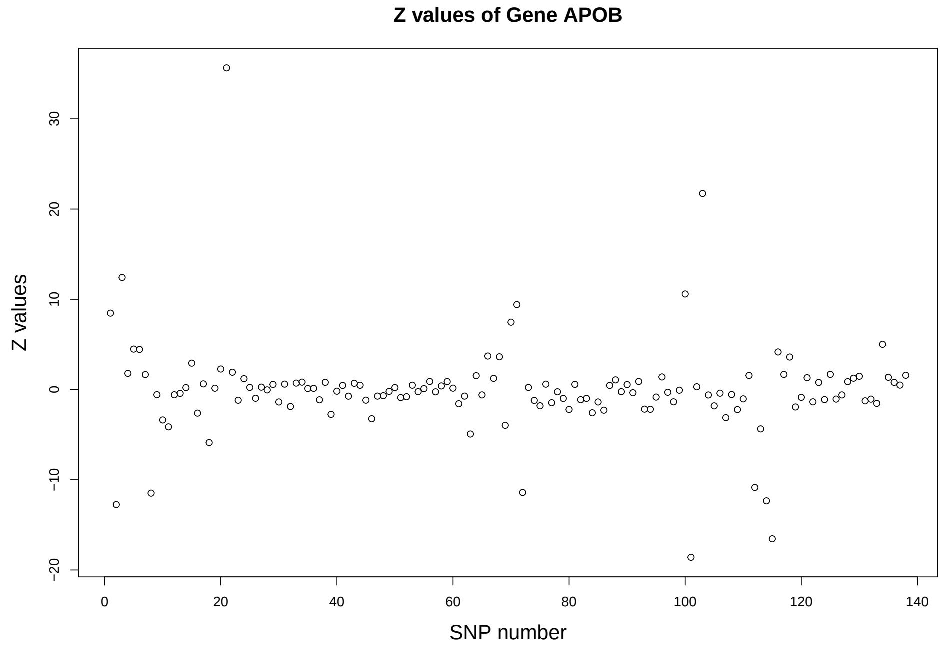


Figure 6.2: The plot of Z values for gene “APOB”.

It can be seen from Figure 6.2 that there are two SNPs whose Z values are within (10, 20), while Z values of 7 SNPs are within (-20, -10). It shows that different SNPs affect LDL in different directions. Several studies have shown that gene “APOB” is associated two genetic disorders about LDL, which are e familial hypobetalipoproteinemia and d familial ligand-defective apoB-100 [115, 116, 117].

6.3 Simulation study

To show that the proposed methods can control the Type I error well and compare the power with other methods, we conduct the simulation study and estimate the power. To generate the simulation data, I selected at random 300 genes with at least one variant within one gene from the 1000 Genomes project [13]. Then, 2000 individuals available from the 1000 Genomes project are selected randomly to generate disease state and the LD values are calculated based on corresponding genotypes of 503 individuals from the 1000 Genomes project European population. After obtaining the genotype data, we randomly select 10 or 40 variants for each gene and generate traits according to the following equation.

$$\text{Logit}\, P(Y_i = 1) = \beta_0 + \sum_{j=1}^{k} G_{ij}\beta_j, \tag{6.31}$$

where Y_i denotes the disease states and $Y_i = 1$ means that this person has the disease. G_{ij} represent the genotype data for SNP j of individual i. Each row represents one sample while each column represents one SNP. The entry $G_{ij} = 0,\ 1$ or 2 is based on the numbers of copies of the minor allele. Define $\boldsymbol{G_i} = (G_{i1}, \ldots, G_{ik})$, which is the genotype data for individual i, where k denotes the number of SNPs in one gene. Then the distribution of Y_i given $\boldsymbol{G_i}$ is a Bernouli distribution with probability:

$$P(Y_i = 1|\boldsymbol{G_i}) = \frac{\exp(\beta_0 + \sum_{j=1}^{k} G_{ij}\beta_j)}{1 + \exp(\beta_0 + \sum_{j=1}^{k} G_{ij}\beta_j)}. \tag{6.32}$$

The parameter settings for this simulation study are given in Table 6.3:

Parameter	Description	Values
n	Sample size	2000
k	Total number of SNPs within one gene	10 or 40
p_c	proportion of the causal SNPs ($\beta \neq 0$)	from 0.1 to 0.7
p_D	Proportion of the deleterious SNPs among the causal ones ($\beta > 0$)	1 or 0.7
OR_j	Odds ratio of SNP j for neutral SNPs ($\beta = 0$)	1
	Odds ratio of SNP j for causal SNPs	2 or 0.05
p_0	background disease prevalence	0.05

Table 6.3: The parameter values in simulation study 1. The background disease prevalence $p_0 = P(Y_i = 1|\boldsymbol{G_i} = 0) = 0.05$. Then the intercept in equation (6.27) is $\beta_0 = \log(0.05/0.95)$.

I compare methods HMRF and HMRF_gene with several popular methods, which are summarized in Table 6.4.

Year	Methods	Reference	Equation number
2013	SKATO	[51]	(6.6)
2014	aSPU	[97]	(6.11)
2017	GHC	[110]	(6.13)
2019	GBJ	[112]	(6.14)
2019	ARTA	[106]	(6.19)
2019	HMP	[105]	(6.18)
2020	DOT	[113]	(6.17)
2020	CCT	[108]	(6.16)

Table 6.4: SKATO: Optimal sequence kernel association test; aSPU: Adaptive sum of powered score tests; GBJ: Generalized Berk-Jones test; ARTA: Adaptive augmented rank truncation method; HMP: The harmonic mean *P* value test; DOT: Decorrelation by orthogonal transformation; CCT: Aggregated Cauchy association test.

The results are summarised in Figure 6.3, which contains the FDR of different methods in simulation study. In Figures 6.3 and 6.4, $p_D = 1$, which means all causal SNPs have positive effect. It can be seen that the proposed method HMRF can control the FDR well regardless of the proportion of causal SNPs. However, methods DOT and HMRF_gene always maintains a higher FDR than significance level of 5% no matter what the proportions of causal SNPs are, which means DOT and HMRF_gene will identify many genes which are not associated with the disease. Besides, the FDR of other methods are very close to 0, which means these methods are conservative for identifying the associated genes. In terms of FDR, the proposed method HMRF perform better than other methods since the FDR can be well controlled around the significance level of 5% compared with other methods, regardless of the proportion of causal SNPs in one gene.

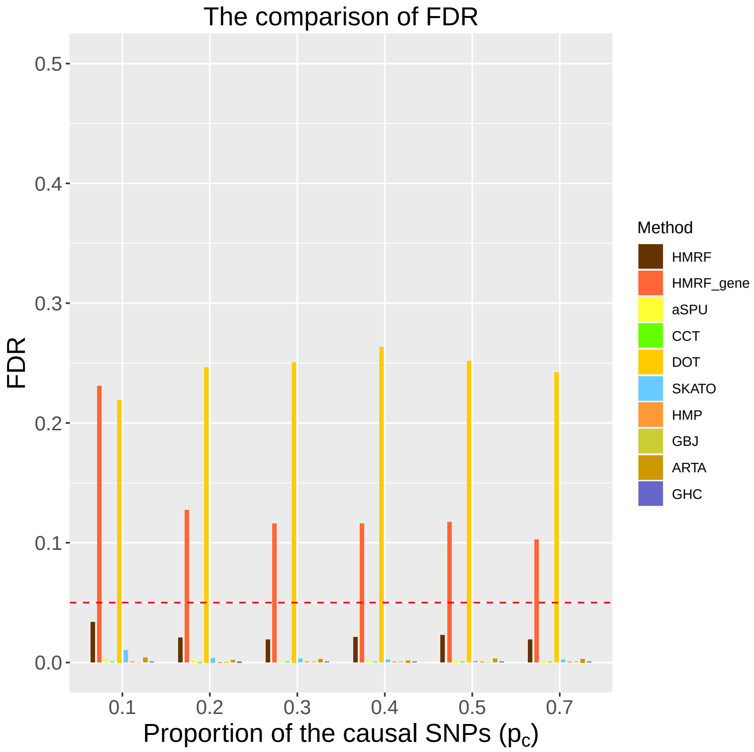


Figure 6.3: The FDR for simulation study. The X-axis represents different proportion of causal SNPs. The red dashed line represents the significance level of 5%. $p_c = 0.1$ means that there is one causal SNP among 10 SNPs. The results are based on the mean values of 10 repeats.Method HMRF means method 1.

Similarly, the results of power for different methods are summarised in Figure 6.4.

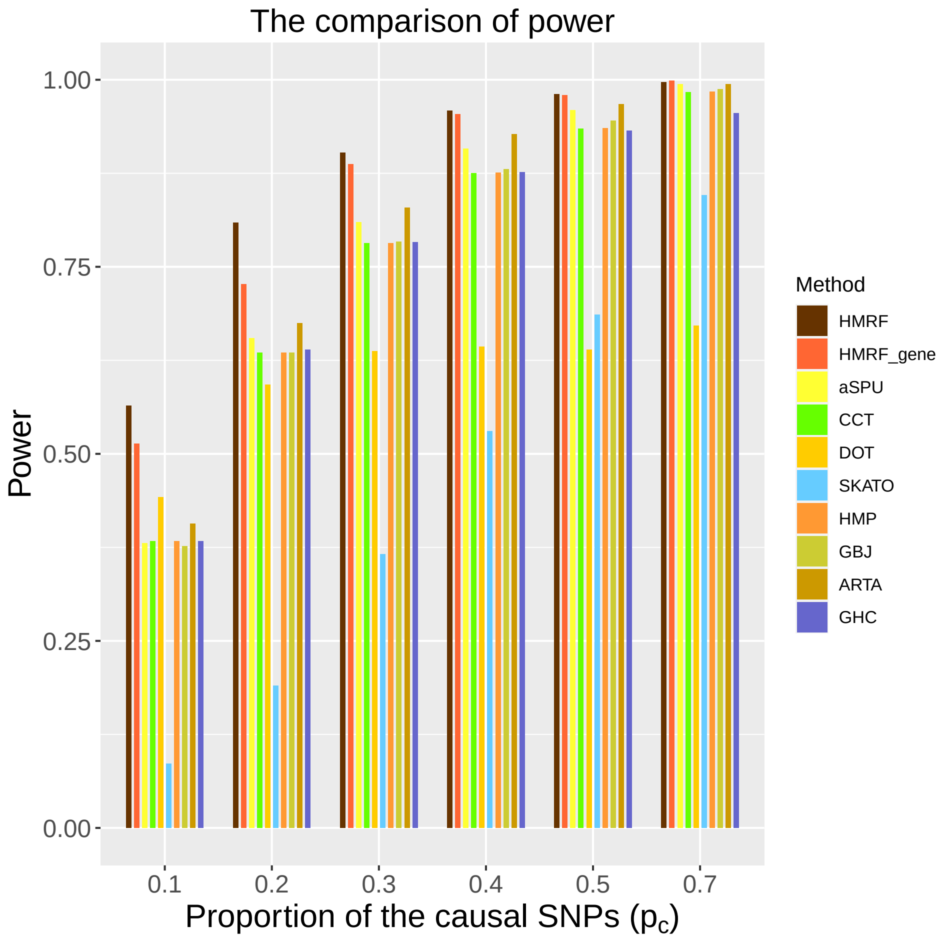


Figure 6.4: The Power for simulation study. The X-axis represents different proportion of causal SNPs. $p_c = 0.1$ means that there is one causal SNP among 10 SNPs. The odds ratio is 2. All causal SNPs have the same effect direction. The results are based on the mean values of 10 repeats.Method HMRF represents method 1. HMRF_gene represents the gene level HMRF model, which is method 2.

It can be seen from above figures that for all methods, as the proportion of causal SNPs (p_c) increases, the power increases, since it is easier to identify the associated genes when more causal SNPs are included in the gene. Method HMRF has a higher power than other methods. Especially when the p_c is small, the difference of power between HMRF and other methods is larger, while this difference became smaller when $p_c = 0.7$. Though methods HMRF_gene and DOT show a higher FDR than significance level in Figure 6.3, they do not have a higher power. In particular, when $p_c = 0.5$ and $p_c = 0.7$, the power of DOT is the lowest. Except for these three methods, the performance of SKATO is the worst since it has a much lower power than the other 6 methods. From Figure 6.3 and 6.4, it can be observed, when $p_c = 0.7$, that the difference in performance between HMRF and methods aSPU, CCT, HMP, GBJ and ARTA is very small. When p_c is small, method HMRF performs much better than other methods.

Since the proposed method performs much better when the causal SNPs are sparse, to further demonstrate this point, another simulation study with lower proportions of causal SNPs is conducted. In this study, 100 genes are randomly selected and 40 variants are randomly chosen for each gene. Also in this study, $p_D = 1$. The comparison of FDR and power are shown in Figure 6.5 and Figure 6.6, respectively.

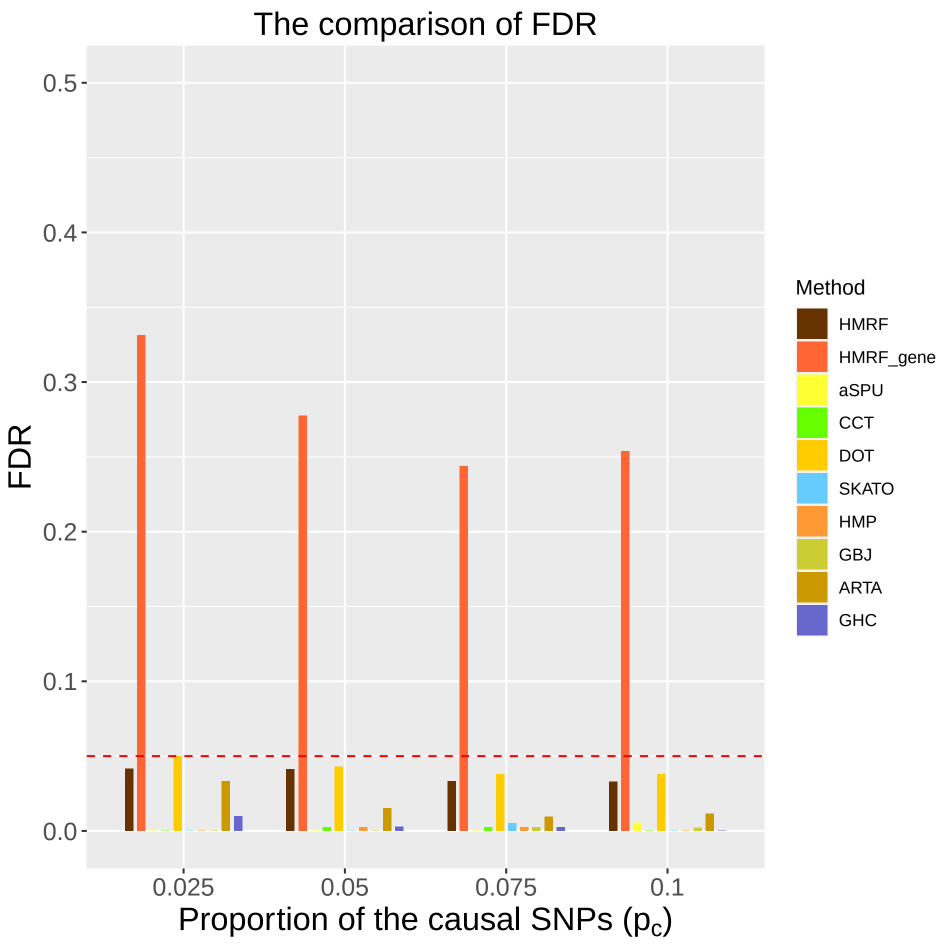


Figure 6.5: The FDR for simulation study. The X-axis represents different proportion of causal SNPs. The red dotted line represents the significance level. $p_c = 0.025$ means that there is one causal SNP among 40 SNPs. The gene number is 100. The results are based on the mean values of 10 repeats.Method HMRF means method 1.

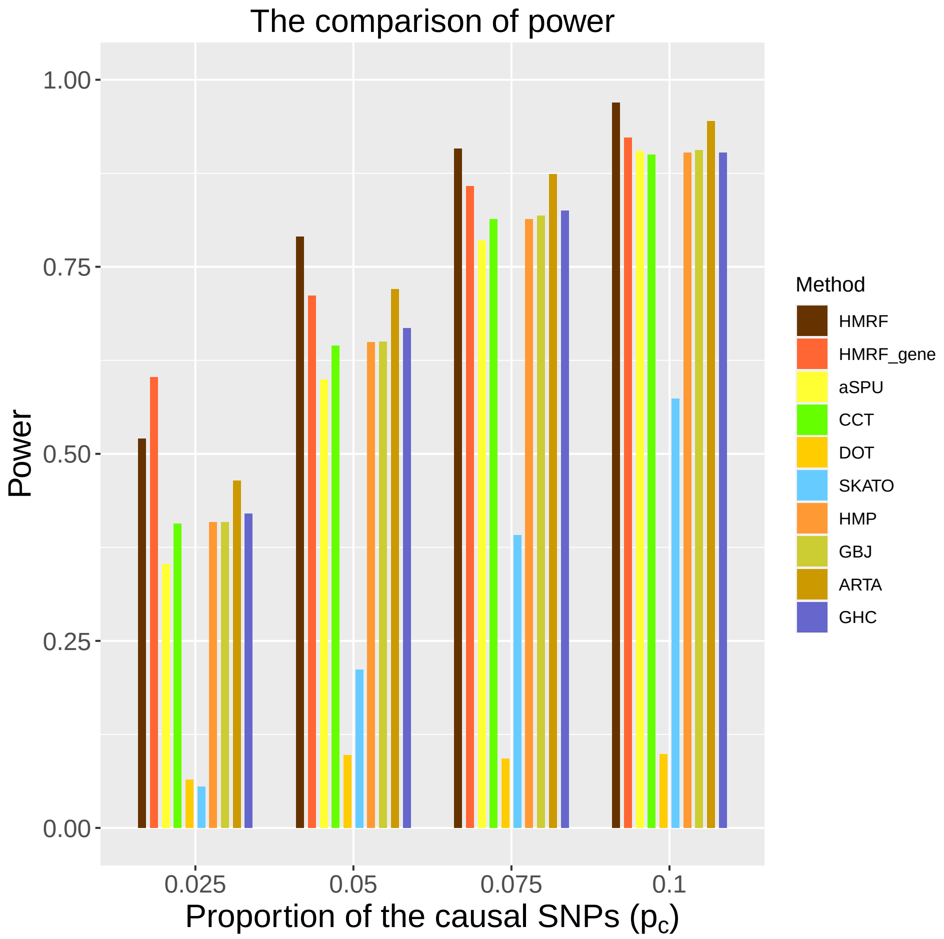


Figure 6.6: The Power for simulation study. The X-axis represents different proportion of causal SNPs. $p_c = 0.025$ means that there is one causal SNP among 40 SNPs. The gene number is 100. The odds ratio of causal SNPs are 2. The results are based on the mean values of 10 repeats.Method HMRF means method 1. HMRF_gene represents the gene level HMRF model, which is method 2.

It can be seen from above figures that method HMRF_gene still shows a much higher FDR than significance level, which means it can not control the FDR very well. The FDR of methods HMRF and DOT can be well controlled, since the FDR are very close to the significance level, while other methods are still too conservative. It can be seen from the Figure 6.6 that the power of HMRF is significantly higher than other methods except for methods HMRF_gene when $p_c = 0.025$. This further demonstrates that HMRF performs much better than other methods when the causal SNPs within one gene is sparse. This means when p_c is small, HMRF is better calibrated, with an FDR close to the target significance level, and has relatively high power to identify associated genes. For other methods, DOT has the lowest power except when $p_c = 0.025$.

For the above simulation study, all associated SNPs affect the disease in the same direction since all associated SNPs have a positive effect size. However, in practice, the association could be positive or negative. To compare the performance of different methods when there exists positive and negative effect size together within one gene, a simulation study is implemented. In this study, similar to before, 100 genes are selected to calculate the power and FDR. For each gene, 40 SNPs are randomly selected. The proportions of causal SNPs change from 0.1 to 0.4. The proportion of deleterious SNPs among causal ones is set as 0.7, which means 70% of causal SNPs have positive association. The odds ratio is 2 for deleterious SNPs while 0.05 for other causal SNPs. The results are summarized in Figure 6.7 and 6.8.

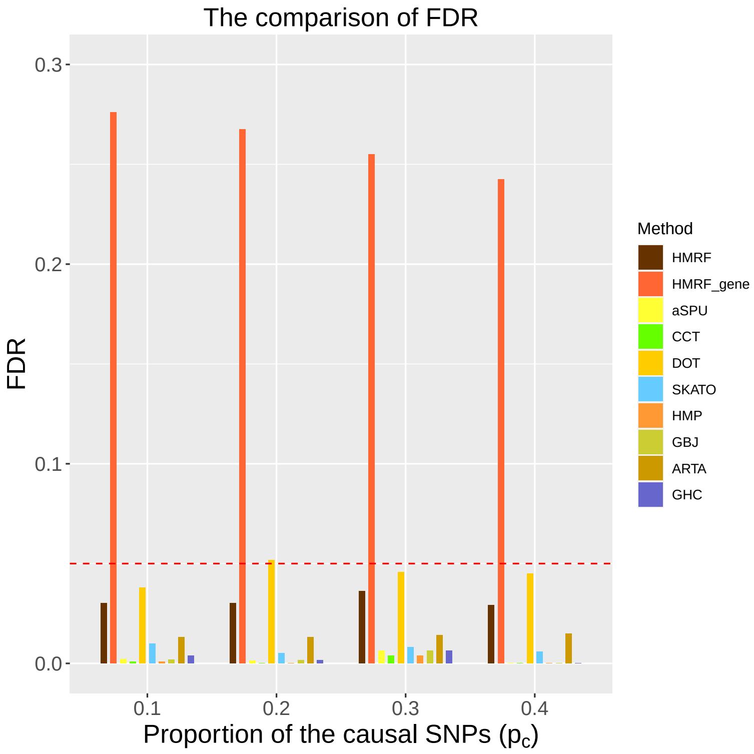


Figure 6.7: The FDR for simulation study. The X-axis represents different proportion of causal SNPs. The red dotted line represents the significance level. $p_c = 0.1$ means that there are 4 causal SNPs among 40 SNPs. The gene number is 100. The results are based on the mean values of 10 repeats. Method HMRF means method 1.

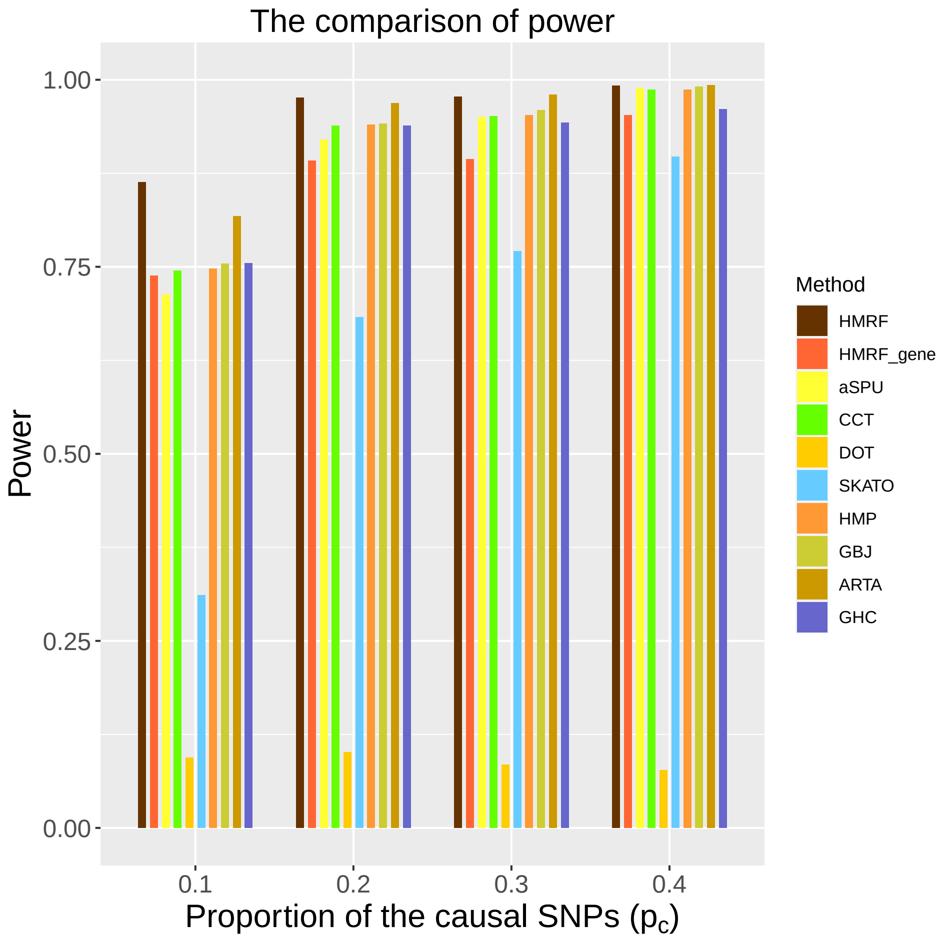
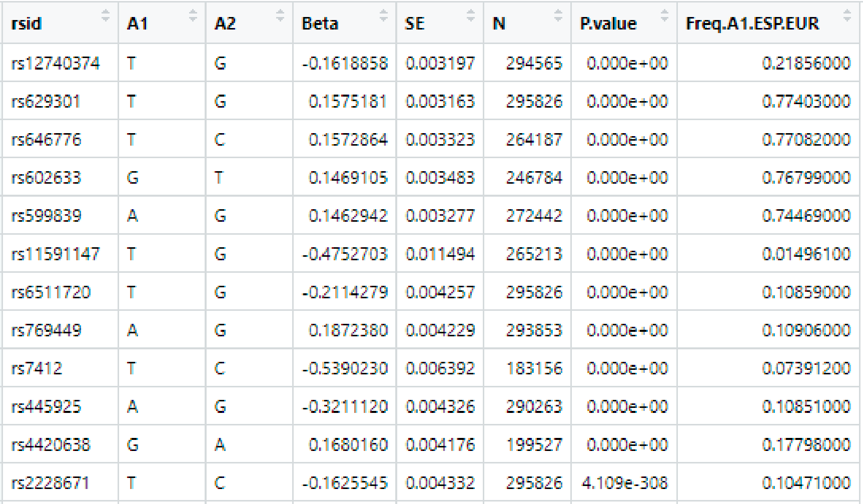


Figure 6.8: The Power for simulation study. The X-axis represents different proportion of causal SNPs. $p_c = 0.1$ means that there are 4 causal SNPs among 40 SNPs. The gene number is 100. The results are based on the mean values of 10 repeats. Method HMRF means method 1. HMRF_gene represents the gene level HMRF model, which is method 2.

It can be seen that the conclusion is similar as before. In Figure 6.7, the FDR of HMRF_gene is still much higher than significance level, since it sums the Z values within one gene and the sum will decrease the effect size when there are both positive effect and negative effect. Methods HMRF and DOT can still control the FDR well, while the FDR of other methods are quite small as before. For the power in Figure 6.8, it can be seen that the power of HMRF is still the highest no matter what p_c is, but the difference of power between HMRF and other methods becomes small when $p_c \geq 0.3$. DOT still has the lowest power among 10 methods. The power of HMRF_gene is lower than HMRF though it has a higher FDR, while the power of other methods except for SKATO are very close to each other.

6.4 Real data application

To see the performance on real data, we perform gene association analysis for four plasma lipid traits, which includes HDL, LDL, TC and TG. The summary statistics are downloaded from GLGC [114], which is shown in Figure 6.9.



rsid	A1	A2	Beta	SE	N	P.value	Freq.A1.ESP.EUR
rs12740374	T	G	-0.1618858	0.003197	294565	0.000e+00	0.21856000
rs629301	T	G	0.1575181	0.003163	295826	0.000e+00	0.77403000
rs646776	T	C	0.1572864	0.003323	264187	0.000e+00	0.77082000
rs602633	G	T	0.1469105	0.003483	246784	0.000e+00	0.76799000
rs599839	A	G	0.1462942	0.003277	272442	0.000e+00	0.74469000
rs11591147	T	G	-0.4752703	0.011494	265213	0.000e+00	0.01496100
rs6511720	T	G	-0.2114279	0.004257	295826	0.000e+00	0.10859000
rs769449	A	G	0.1872380	0.004229	293853	0.000e+00	0.10906000
rs7412	T	C	-0.5390230	0.006392	183156	0.000e+00	0.07391200
rs445925	A	G	-0.3211120	0.004326	290263	0.000e+00	0.10851000
rs4420638	G	A	0.1680160	0.004176	199527	0.000e+00	0.17798000
rs2228671	T	C	-0.1625545	0.004332	295826	4.109e-308	0.10471000

Figure 6.9: This is a small subset of the sample data from whole data of disease LDL. The rsid means the SNP ID. A1 represents the minor allele, while A2 denotes the major allele. Beta represents the estimated effect size. SE is the standard error for Beta. N represents the sample size. P.value denotes the p-values for Beta. The last column Freq.A1.ESP.EUR represents the frequency for A1 in European population. The Z values we used in our study can be calculated using Beta/SE.

We consider 5636 genes which includes 53843 variants and define the set of SNPs located within either 2kb extension upstream of the transcription start site or 2kb downstream of the transcription end site of a given gene. MAGMA software is used to map our SNPs to genes [118]. We only analyse the genes which contain at least five variants and the 1000 Genome European genotype data are used as the reference panel to calculate LD matrix. The results are shown in Table 6.5, and the similarity of the results across the different methods are shown in Table 6.6.

Method	HDL	LDL	TC	TG
HMRF	**492**	**343**	**504**	**362**
aSPU	271	221	313	237
CCT	241	189	281	202
HMP	238	184	276	197
GBJ	252	189	272	198
ARTA	284	208	304	225
GHC	233	180	271	185

Table 6.5: The number of genes identified as being associated with each of four plasma lipid traits for different methods. The real data application just considers those methods which can control FDR well. Since the real data just includes summary statistics, SKATO is not applied since it needs individual genotype data.

Method	HDL	LDL	TC	TG
aSPU	259 (95.6%)	202 (91.4%)	294 (93.9%)	216 (91.1%)
CCT	241 (100%)	189 (100%)	281 (100%)	201 (99.5%)
HMP	238 (100%)	184 (100%)	276 (100%)	196 (99.5%)
GBJ	246 (97.6%)	183 (96.8%)	264 (97.1%)	193 (97.5%)
ARTA	264 (93.0%)	191 (91.8%)	287 (94.4%)	201 (89.3%)
GHC	233 (100%)	180 (100%)	270 (99.6%)	184 (99.5%)

Table 6.6: The similarity of the results across the different methods. The entry in the table represents the number identified by method HMRF meanwhile, among the genes identified by other methods. The numbers in the bracket means the proportion. For example, the first number in the second column 259 denotes that for 271 genes identified by method aSPU, there are 259 genes identified by proposed method HMRF meanwhile. The similarity proportion is 259/271=95.6%

It can be seen from Table 6.5 that the proposed method HMRF has identified more genes which may be associated with disease than other methods, which is consistent with the simulations study results. Also, the number identified by other methods are quite close. Among the other methods, GHC has detected a smaller number of associated genes than other methods. For trait HDL, HMRF identifies 492 associated genes, while CCT detects 241 genes. All these 241 genes are also identified by HMRF. It can be seen from Table 6.6 that the similarity of the results across different methods are high. For example, for traits HDL and LDL, all genes identified by method CCT, HMP and GHC are all identified by proposed method HMRF. Except for trait TG used by method ARTA, for other traits, the proportion of same genes identified by proposed method HMRF and other methods are higher than

90%. Among all traits, 481 genes are identified to be associated with at lease two traits by HMRF, while 134 genes are shared in at least three traits. HMRF identifies 60 genes which are associated with four traits, including genes ABCA1 and CETP, which have been confirmed by precious study [119].

6.5 Conclusion

This chapter develops the proposed method for gene association analysis. Two possible extensions are developed, but the simulation studies demonstrate that only HMRF shows a good performance. HMRF_gene produces an FDR that is too high, which may cause more false discoveries. The reason may be because the sum is applied on the Z values of SNPs within one gene, which may be not a good way to represent gene association effect, particularly when these Z values contain both positive and negative effects. In the future, it may be worth investigating how best to summarise the Z values of multiple SNPs within one gene.

Chapter 7

Conclusion and future research

7.1 Summary of work

In this thesis, we have illustrated how GWAS identifies SNPs that may be associated with disease, from which we understand that the linkage disequilibrium is common between different genetic variants. We have introduced a hidden Markov random field model, which is usually used in graphical analysis, to leverage the dependence between SNPs. In our proposed model, Z values are assumed to follow a mixture normal distribution and the EM algorithm is applied to estimate the parameters. Finally, the FDR control procedure is used to identify associated SNPs. We carried out simulation studies to compare the performances of the proposed method, the lfdr method and the Bonferroni Correction method, showing that the proposed method achieves a higher power than the other two methods in the context of controlling FDR.

Although some assumptions are illustrated in our model, we also conducted simulation studies to show the performance of proposed methods when the model assumptions are violated. When the form of null distribution is not a standard normal distribution, we have found that the proposed method still behave better. Furthermore, we also discussed the effect of initial estimates and choice of the parameter τ in the weight matrix. Simulation studies have showed that initial estimates do not have much effect on performance, while a large τ will cause a FDR larger than the significance level. Given that the value of τ may matter, a useful avenue of future research could be to devise a way to estimate the value of τ empirically from the data. The situation when Z values do not follow a mixture normal distribution has been considered. The distribution of Z values has been extended to two components of Gaussian mixture model, for which proposed method still outperforms other methods.

Besides SNP association analysis, we have extended our method into gene association. Two forms of extensions are introduced. One is putting all gene data together and regarding Z values for each SNP as input data, but using a different grouped FDR control procedure to identify associated genes. Another is summing Z values for each SNP within one gene into one value, then using the proposed HMRF model in Chapter 3 on these values. We have conducted simulation studies to compare the performance with other popular gene association methods. The simulation study results shows that the method HMRF performs better than other methods, especially when the proportion of causal SNPs within one gene is small. Then the method HMRF are applied on four traits and it can identify more associated genes than other methods.

7.2 Future research

7.2.1 Gibbs sampling

In this thesis, as implementation of the EM algorithm involves Gibbs sampling, the computation time for estimating parameters will increase substantially as the number of SNP we analyse increases. Because of the time limitation of UCL's high computation platform, which is not larger than 48 hours, we limit the size of SNPs in the simulation study. However, in Chapter 4 we discussed using a different convergence criterion and have shown that this can decrease the computation time since it requires less iterations to end the algorithm. When the number of SNPs is large, it becomes more difficult to converge even if the new convergence criterion is applied and the Gibbs sampling is still computationally intensive. Therefore, to make the proposed model applicable to a large dataset, considering how to accelerate the parameter estimation procedure is an important task. In Chapter 5, we proposed to split the large datasets into several small datasets and implemented the EM algorithm on each small dataset. This makes it possible to analyse the data within the time limit of computation platform. However, this will lose some LD information between SNPs in different datasets. For instance, for the last SNP in dataset 1 and first SNP in dataset 2, they are located close to each other and may be highly correlated. But when we analyse datasets separately, their correlation is neglected. To overcome this problem, one possibility is to create small datasets with overlapping data between different datasets. For example, suppose that 20 SNPs are split into two datasets. One dataset contains SNP from 1 to 8, second dataset contains SNP 7 to SNP 15 and last dataset contains SNP 14 to 20, by which way, the correlation between every pair of SNPs can be included. However, how to decide the number of overlapping data needs to be investigated further.

7.2.2 Possible improvement

In this thesis, numerical analyses have shown the superiority of the proposed procedure. However, the asymptotic properties of the proposed LIS-based testing procedures are unknown. In 2009, Sun and Cai [26] assumed that the hidden states of multiple testing hypothesis followed a hidden Markov model and proposed an oracle testing procedure in an ideal setting. They showed that under mild conditions, the LIS-based oracle procedure was optimal in the sense that it minimized FNR subject to a constraint on FDR. Mimicking the oracle procedure, they proposed a data-driven procedure, which had been shown to be asymptotically optimal. Therefore, in the future, we may try to follow Sun and Cai's research and study the asymptotic properties of proposed HMRF-LIS based procedures.

In this thesis, it is assumed that Z values follow a one-component of normal mixture distribution. In Chapter 4, this is extended into two components of Gaussian mixture distributions. However, in practice, the actual number of mixture components is usually unknown. Therefore, in the future, it is worth to study how to introduce a Dirichlet process to represent the components of mixture models [120].

7.2.3 Practical application

In this thesis, a testing procedure is proposed to identify association between multiple genetic variants and one trait. In practice, it may be beneficial to identify association based on multiple GWAS data, that is integrating data relating to multiple traits in order to improve power. Then the proposed method can be extended to identifying association between one genetic variant and multiple traits, since some traits are correlated. Then the weight matrix will represent the correlation between different traits rather than different SNPs. Furthermore, it is possible to integrate Z values between multiple genetic variants and multiple traits together. In this case, the weight matrix will not only represent the correlation between different SNPs, but also denote the dependence between different traits [121].

Chapter 8

Appendix

8.1 Derivation of equation (3.33)

The conditional expectation of complete data log-likelihood $Q(\phi|\phi^{\{old\}})$ can be written as follows:

$$\begin{aligned} Q(\phi|\phi^{\{old\}}) &= E_{\phi^{\{old\}}}(\gamma\sum_{i=1}^{m}\theta_i + \beta\sum_{i\sim j} w_{ij}I(\theta_i=\theta_j) - \log\psi(\theta)|Z) \\ &+ E_{\phi^{\{old\}}}(\sum_{i=1}^{m}\log P(Z_i|\theta_i)) = l_1(\phi_1) + l_2(\phi_2). \end{aligned} \tag{8.1}$$

where $\psi(\theta) = \sum_{\theta\in\{0,1\}^m}\exp(\gamma\theta_i + \beta\sum_{j\in N_i} w_{ij}I(\theta_i=\theta_j))$, $\boldsymbol{\phi}_1 = (\gamma,\beta)^T$, $\boldsymbol{\phi}_2 = (\mu,\sigma^2)^T$.

$$\begin{aligned} p(\boldsymbol{\theta};\Phi) &\propto \exp(\gamma\sum_{i=1}^{m}\theta_i + \beta\sum_{i\sim j} w_{ij}I(\theta_i=\theta_j)) \\ &\propto \exp(\boldsymbol{\phi}_1^T\boldsymbol{H}(\boldsymbol{\theta})) = \frac{\exp(\boldsymbol{\phi}_1^T\boldsymbol{H}(\boldsymbol{\theta}))}{\psi(\theta)} \end{aligned} \tag{8.2}$$

where $\psi(\theta) = \sum_{\theta\in\{0,1\}^m}\exp(\gamma\theta_i + \beta\sum_{j\in N_i} w_{ij}I(\theta_i=\theta_j))$, γ and β are two model parameters. $\boldsymbol{\phi}_1 = (\gamma, \beta)^T$, $\boldsymbol{H}(\boldsymbol{\theta}) = (H_1, H_2)^T = (\sum_{i=1}^{m} \theta_i, \sum_{i\sim j} w_{ij} I(\theta_i = \theta_j))^T$.

$$\begin{aligned} E[\exp(-\boldsymbol{\phi}_1^T \boldsymbol{H}(\boldsymbol{\theta}))] &= \sum_{\theta \in \{0,1\}^m} \exp(-\boldsymbol{\phi}_1^T \boldsymbol{H}(\boldsymbol{\theta})) \frac{\exp(\boldsymbol{\phi}_1^T \boldsymbol{H}(\boldsymbol{\theta}))}{\psi(\theta)} \\ &= \sum_{\theta \in \{0,1\}^m} \frac{1}{\psi(\theta)} = \frac{C}{\psi(\theta)}, \end{aligned} \tag{8.3}$$

where C is the number of all possible configurations of θ, which is a constant. $E[\exp(-\boldsymbol{\phi}_1^T \boldsymbol{H}(\boldsymbol{\theta}))]$ can be approximated by Gibbs samplers:

$$E[\exp(-\boldsymbol{\phi}_1^T \boldsymbol{H}(\boldsymbol{\theta}))] \approx \frac{1}{n} \sum_{i=1}^{n} \exp(-\boldsymbol{\phi}_1^T \boldsymbol{H}(\boldsymbol{\theta}^{(i,\phi_1)})), \tag{8.4}$$

where $\{\theta^{(1,\phi_1)}, \theta^{(2,\phi_1)}, \ldots, \theta^{(n,\phi_1)}\}$ are Gibbs samplers generated from $p(\boldsymbol{\theta}; \Phi)$.

The conditional association state for SNP i, given the states of all neighboring SNPs, is

$$p(\theta_i | \theta_{N_i}; \Phi) \propto \exp(\gamma \theta_i + \beta \sum_{j \in N_i} w_{ij} I(\theta_i = \theta_j)), \tag{8.5}$$

where N_i represents the neighbors of the SNP i on the LD graph.

For $l_1(\boldsymbol{\phi}_{\mathbf{1}}^{(t+1,m)}) - l_1(\boldsymbol{\phi}_{\mathbf{1}}^{(t)})$,

$$\begin{aligned} &l_1(\boldsymbol{\phi}_{\mathbf{1}}^{(t+1,m)}) - l_1(\boldsymbol{\phi}_{\mathbf{1}}^{(t)}) \\ &= E_{\boldsymbol{\phi}_1^{(t+1,m)}}[\boldsymbol{\phi}_1^{(t+1,m)^T} \boldsymbol{H}(\boldsymbol{\theta})|Z] - E_{\boldsymbol{\phi}_1^{(t)}}[\boldsymbol{\phi}_1^{(t)^T} \boldsymbol{H}(\boldsymbol{\theta})|Z] - \log \frac{\psi(\theta; \boldsymbol{\phi}_{\mathbf{1}}^{(t+1,m)})}{\psi(\theta; \boldsymbol{\phi}_{\mathbf{1}}^{(t)})} \\ &\approx \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{\phi}_1^{(t+1,m)^T} \boldsymbol{H}(\boldsymbol{\theta}^{(t,i)}) - \boldsymbol{\phi}_1^{(t)^T} \boldsymbol{H}(\boldsymbol{\theta}^{(t,i)}) \right) - \log \left(\frac{\sum_{i=1}^{n} \exp(-\boldsymbol{\phi}_1^{(t)^T} \boldsymbol{H}(\boldsymbol{\theta}^{(i,\phi_1^{(t)})}))}{\sum_{i=1}^{n} \exp(-\boldsymbol{\phi}_1^{(t+1,m)^T} \boldsymbol{H}(\boldsymbol{\theta}^{(i,\phi_1^{(t+1,m)})}))} \right) \\ &= \frac{1}{n} (\boldsymbol{\phi}_1^{(t+1,m)} - \boldsymbol{\phi}_1^{(t)})^T \sum_{i=1}^{n} \boldsymbol{H}(\boldsymbol{\theta}^{(t,i)}) + \log \left(\frac{\sum_{i=1}^{n} \exp\{-\boldsymbol{\phi}_1^{(t+1,m)^T} \boldsymbol{H}(\boldsymbol{\theta}^{(i,\phi_1^{(t+1,m)})})\}}{\sum_{i=1}^{n} \exp\{-\boldsymbol{\phi}_1^{(t)^T} \boldsymbol{H}(\boldsymbol{\theta}^{(i,\phi_1^{(t)})})\}} \right), \end{aligned} \tag{8.6}$$

where $\theta^{(t,i)}$ are Gibbs samplers from $P(\theta_i | Z, \hat{\theta}_{S \setminus i})$, $\theta^{(i,\phi_1^{(t+1,m)})}$ and $\theta^{(i,\phi_1^{(t)})}$ are Gibbs sampler from $p(\theta_i | \theta_{N_i})$.

Bibliography

[1] Gabor Marth & Steve Sherry Ravi Sachidanandam, David Weissman, Steven C. Schmidt, Jerzy M. Kakol & Lincoln D. Stein. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822):928–933, feb 2001.

[2] Joel N. Hirschhorn and Mark J. Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108, feb 2005.

[3] E S Lander and N J Schork. Genetic dissection of complex traits. *Science (New York, N.Y.)*, 265(5181):2037–48, sep 1994.

[4] David J Hunter, Peter Kraft, Kevin B Jacobs, David G Cox, Meredith Yeager, Susan E Hankinson, Sholom Wacholder, Zhaoming Wang, Robert Welch, and Hutchinson et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nature Genetics*, 39(7):870–874, jul 2007.

[5] Laura J Scott, Karen L Mohlke, Lori L Bonnycastle, Cristen J Willer, Yun Li, William L Duren, Michael R Erdos, Heather M Stringham, Peter S Chines, and Jackson et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science (New York, N.Y.)*, 316(5829):1341–5, jun 2007.

[6] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, jun 2007.

[7] Peter M. Visscher, Matthew A. Brown, Mark I. McCarthy, and Jian Yang. Five Years of GWAS Discovery. *The American Journal of Human Genetics*, 90(1):7–24, jan 2012.

[8] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael E Goddard, and Peter M Visscher. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569, jul 2010.

[9] Eun Pyo Hong and Ji Wan Park. Sample size and statistical power calculation in genetic association studies. *Genomics & Informatics*, 10(2):117, 2012.

[10] Paul El-Fishawy. *Common Disease-Rare Variant Hypothesis*, pages 720–722. Springer New York, New York, NY, 2013.

[11] Kelly A. Frazer, Sarah S. Murray, Nicholas J. Schork, and Eric J. Topol. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, 10(4):241–251, apr 2009.

[12] Jonathan K. Pritchard and Molly Przeworski. Linkage Disequilibrium in Humans: Models and Data. *The American Journal of Human Genetics*, 69(1):1–14, jul 2001.

[13] The 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature 2015 526:7571*, 526(7571):68–74, sep 2015.

[14] Christopher C. Chang, Carson C. Chow, Laurent C.A.M. Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, 4(1):7, feb 2015.

[15] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, and Richard Durbin. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, aug 2011.

[16] Paul R. Burton, David G. Clayton, Lon R. Cardon, Nick Craddock, Panos Deloukas, Audrey Duncanson, Dominic P. Kwiatkowski, Mark I. McCarthy, and Ouwehand et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, jun 2007.

[17] Bradley Efron. Size, power and false discovery rates. *Annals of Statistics*, 35(4):1351–1377, aug 2007.

[18] Bradley Efron. Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102(477):93–103, mar 2007.

[19] William S Noble. How does multiple testing correction work? *Nature biotechnology*, 27(12):1135–7, dec 2009.

[20] Karen Young, John W. Tukey, and H. I. Braun. The Collected Works of John W. Tukey: Vol. VIII, Multiple Comparisons: 1948-1983. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158(3):632, 1995.

[21] J M Bland and D G Altman. Multiple significance tests: the Bonferroni method. *BMJ (Clinical research ed.)*, 310(6973):170, jan 1995.

[22] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

[23] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57:289–300, 1995.

[24] Bradley Efron, Robert Tibshirani, John D. Storey, and Virginia Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160, dec 2001.

[25] Bradley Efron and Robert Tibshirani. Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, 23(1):70–86, 2002.

[26] Wenguang Sun and T. Tony Cai. Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 71(2):393–424, apr 2009.

[27] Wenguang Sun and T. Tony Cai. Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901–912, sep 2007.

[28] T. Tony Cai and Wenguang Sun. Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *Journal of the American Statistical Association*, 104(488):1467–1481, dec 2009.

[29] Yoav Benjamini and Rami Cohen. Weighted false discovery rate controlling procedures for clinical trials. *Biostatistics*, 18(1):91–104, jan 2017.

[30] Pallavi Basu, T. Tony Cai, Kiranmoy Das, and Wenguang Sun. Weighted False Discovery Rate Control in Large-Scale Multiple Testing. *Journal of the American Statistical Association*, 113(523):1172–1183, jul 2018.

[31] Haibing Zhao and Jiajia Zhang. Weighted p-value procedures for controlling FDR of grouped hypotheses. *Journal of Statistical Planning and Inference*, 151-152:90–106, aug 2014.

[32] James X. Hu, Hongyu Zhao, and Harrison H. Zhou. False discovery rate control with groups. *Journal of the American Statistical Association*, 105(491):1215–1227, sep 2010.

[33] Yanping Liu, Sanat K. Sarkar, and Zhigen Zhao. A new approach to multiple testing of grouped hypotheses. *Journal of Statistical Planning and Inference*, 179:1–14, dec 2016.

[34] Sanat K. Sarkar and Shinjini Nandi. On the Development of a Local FDR-Based Approach to Testing Two-Way Classified Hypotheses. *Sankhya B*, 83(1):1–11, may 2021.

[35] Haibing Zhao. Adaptive FWER control procedure for grouped hypotheses. *Statistics & Probability Letters*, 95:63–70, dec 2014.

[36] Li Wang. Weighted multiple testing procedure for grouped hypotheses with k-FWER control. *Computational Statistics*, 34(2):885–909, jun 2019.

[37] BARNET WOOLF. On estimating the relation between blood group and disease. *Annals of Human Genetics*, 19(4):251–253, may 1955.

[38] William G. Cochran. Some methods for strengthening the common χ^2 Tests. *Biometrics*, 10(4):417, dec 1954.

[39] Jon Wakefield. Bayes factors for genome-wide association studies: comparison with *P*-values. *Genetic Epidemiology*, 33(1):79–86, jan 2009.

[40] Matthew Stephens and David J. Balding. Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10):681–690, oct 2009.

[41] Gregory C. Chow. Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28(3):591–605, 1960.

[42] Kai Wang and Diana Abbott. A principal components regression approach to multilocus genetic association studies. *Genetic Epidemiology*, 32(2):108–118, feb 2008.

[43] Gulnara R. Svishcheva, Nadezhda M. Belonogova, Irina V. Zorkoltseva, Anatoly V. Kirichenko, and Tatiana I. Axenovich. Gene-based association tests

using GWAS summary statistics. *Bioinformatics*, 35(19):3701–3708, oct 2019.

[44] Momiao Xiong, Jinying Zhao, and Eric Boerwinkle. Generalized T^2 test for genome association studies. *American journal of human genetics*, 70(5):1257–68, may 2002.

[45] Nathalie Malo, Ondrej Libiger, and Nicholas J. Schork. Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *The American Journal of Human Genetics*, 82(2):375–385, feb 2008.

[46] H. Zhou, M. E. Sehl, J. S. Sinsheimer, and K. Lange. Association screening of common and rare genetic variants by penalized regression. *Bioinformatics*, 26(19):2375–2382, oct 2010.

[47] Hui Wang, Yuan-Ming Zhang, Xinmin Li, Godfred L Masinde, Subburaman Mohan, David J Baylink, and Shizhong Xu. Bayesian shrinkage estimation of quantitative trait loci parameters. *Genetics*, 170(1):465–80, may 2005.

[48] Stephan Morgenthaler and William G. Thilly. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 615(1-2):28–56, feb 2007.

[49] Bo Eskerod Madsen and Sharon R. Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, 5(2):e1000384, feb 2009.

[50] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *American journal of human genetics*, 89(1):82–93, jul 2011.

[51] S. Lee, M. C. Wu, and X. Lin. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775, sep 2012.

[52] Andriy Derkach, Jerry F. Lawless, and Lei Sun. Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. *Genetic Epidemiology*, 37(1):110–121, jan 2013.

[53] Yun Joo Yoo, Sun Ah Kim, and Shelley B. Bull. Clique-based clustering of correlated SNPs in a gene can improve performance of gene-based multi-bin linear combination test. *BioMed Research International*, 2015:1–11, aug 2015.

[54] Andriy Derkach, Jerry F. Lawless, and Lei Sun. Pooled association tests for rare genetic variants: a review and some new results. *Statistical Science*, 29(2):302–321, may 2014.

[55] Benjamin M. Neale, Manuel A. Rivas, Benjamin F. Voight, David Altshuler, Bernie Devlin, Marju Orho-Melander, Sekar Kathiresan, Shaun M. Purcell, Kathryn Roeder, and Mark J. Daly. Testing for an unusual distribution of rare variants. *PLoS Genetics*, 7(3):e1001322, mar 2011.

[56] H. Li, Z. Wei, and J. Maris. A hidden Markov random field model for genome-wide association studies. *Biostatistics*, 11(1):139–150, jan 2010.

[57] Yao-Ting Huang, Kun-Mao Chao, and Ting Chen. An approximation algorithm for haplotype inference by maximum parsimony. In *Proceedings of the 2005 ACM symposium on Applied computing*, pages 146–150, 2005.

[58] Lusheng Wang and Ying Xu. Haplotype inference by maximum parsimony. *Bioinformatics*, 19(14):1773–1780, 2003.

[59] Laurent Excoffier and Montgomery Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular biology and evolution*, 12(5):921–927, 1995.

[60] Daniele Fallin and Nicholas J Schork. Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for

unphased diploid genotype data. *The American Journal of Human Genetics*, 67(4):947–959, 2000.

[61] Shin Lin, David J Cutler, Michael E Zwick, and Aravinda Chakravarti. Haplotype inference in random population samples. *The American Journal of Human Genetics*, 71(5):1129–1137, 2002.

[62] Shin Lin, Aravinda Chakravarti, and David J Cutler. Haplotype and missing data inference in nuclear families. *Genome Research*, 14(8):1624–1632, 2004.

[63] Paul Fearnhead and Peter Donnelly. Estimating recombination rates from population genetic data. *Genetics*, 159(3):1299–1318, 2001.

[64] Matthew Stephens, Nicholas J Smith, and Peter Donnelly. A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, 68(4):978–989, 2001.

[65] Paul Scheet and Matthew Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78(4):629–644, 2006.

[66] Matthew Stephens and Peter Donnelly. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *The American Journal of Human Genetics*, 73(5):1162–1169, 2003.

[67] Eran Halperin and Eleazar Eskin. Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics*, 20(12):1842–1849, 2004.

[68] Jonathan Marchini, David Cutler, Nick Patterson, Matthew Stephens, Eleazar Eskin, Eran Halperin, Shin Lin, Zhaohui S Qin, Heather M Munro, Gonçalo R Abecasis, et al. A comparison of phasing algorithms for trios and unrelated individuals. *The American Journal of Human Genetics*, 78(3):437–450, 2006.

[69] Pak Chung Sham. Statistics in human genetics. 1997.

[70] Daniel J Schaid, Charles M Rowland, David E Tines, Robert M Jacobson, and Gregory A Poland. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *The American Journal of Human Genetics*, 70(2):425–434, 2002.

[71] Fadhaa Ali and Jian Zhang. Mixture model-based association analysis with case-control data in genome wide association studies. *Statistical Applications in Genetics and Molecular Biology*, 16(3):173–187, 2017.

[72] John Molitor, Paul Marjoram, and Duncan Thomas. Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. *The American Journal of Human Genetics*, 73(6):1368–1384, 2003.

[73] Jung-Ying Tzeng, Chih-Hao Wang, Jau-Tsuen Kao, and Chuhsing Kate Hsiao. Regression-based association analysis with clustered haplotypes through use of genotypes. *The American Journal of Human Genetics*, 78(2):231–242, 2006.

[74] Sharon R Browning and Brian L Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5):1084–1097, 2007.

[75] Xiaofeng Zhu, Tao Feng, Yali Li, Qing Lu, and Robert C Elston. Detecting rare variants for complex traits using family and unrelated data. *Genetic epidemiology*, 34(2):171–187, 2010.

[76] Fadhaa Ali and Jian Zhang. Screening tests for disease risk haplotype segments in genome by use of permutation. *Journal of Systems Science and Mathematical Sciences*, 35(12):1402–1417, 2015.

[77] Fadhaa Ali and Jian Zhang. Search for risk haplotype segments with gwas data by use of finite mixture models. *Statistics and its interface*, 9(3):267–280, 2015.

[78] Yongyue Zhang, Michael Brady, and Stephen Smith. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–57, jan 2001.

[79] Dipti Patra and Smita Pradhan. Development of fuzzy clustering based unsupervised scheme for medical image segmentation using HMRF model. In *2010 International Conference on Industrial Electronics, Control and Robotics, IECR 2010*, pages 225–229, 2010.

[80] Kristi Kuljus, Fekadu L. Bayisa, David Bolin, Jüri Lember, and Jun Yu. Comparison of hidden Markov chain models and hidden Markov random field models in estimation of computed tomography images. *Communications in Statistics Case Studies Data Analysis and Applications*, 4(1):46–55, jan 2018.

[81] N. Friel, A. N. Pettitt, R. Reeves, and E. Wit. Bayesian inference in hidden Markov random fields for binary data defined on large lattices. *Journal of Computational and Graphical Statistics*, 18(2):243–261, 2009.

[82] Julien Stoehr. A review on statistical inference methods for discrete Markov random fields. *arXiv*, pages 1–30, 2017.

[83] Hai Shu, Bin Nan, and Robert Koeppe. Multiple testing for neuroimaging via hidden Markov random field. *Biometrics*, 71(3):741–750, sep 2015.

[84] Yongyue Zhang, Michael Brady, and Stephen Smith. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57, 2001.

[85] Namjoon Suh. Review on Parameter Estimation in HMRF. *arXiv*, nov 2017.

[86] Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006.

[87] Bradley Efron, Brit Turnbull, and Balasubramanian Narasimhan. *locfdr: Computes Local False Discovery Rates*, 2015. R package version 1.1-8.

[88] Di Wu and Jinwen Ma. An effective EM algorithm for mixtures of Gaussian processes via the MCMC sampling and approximation. *Neurocomputing*, 331:366–374, feb 2019.

[89] Nick Craddock and Pamela Sklar. Genetics of bipolar disorder: successful start to a long journey. *Trends in Genetics*, 25(2):99–105, feb 2009.

[90] Geraldine M. Clarke, Carl A. Anderson, Fredrik H. Pettersson, Lon R. Cardon, Andrew P. Morris, and Krina T. Zondervan. Basic statistical analysis in genetic case-control studies. *Nature Protocols*, 6(2):121–133, feb 2011.

[91] Carl A. Anderson, Fredrik H. Pettersson, Geraldine M. Clarke, Lon R. Cardon, Andrew P. Morris, and Krina T. Zondervan. Data quality control in genetic case-control association studies. *Nature Protocols*, 5(9):1564–1573, aug 2010.

[92] Andries T. Marees, Hilde de Kluiver, Sven Stringer, Florence Vorspan, Emmanuel Curis, Cynthia Marie-Claire, and Eske M. Derks. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, 27(2), jun 2018.

[93] Yuan Jiang and Heping Zhang. Propensity score-based nonparametric test revealing genetic variants underlying bipolar disorder. *Genetic Epidemiology*, 35(2):125–132, feb 2011.

[94] Jonathan C. Cohen, Robert S. Kiss, Alexander Pertsemlidis, Yves L. Marcel, Ruth McPherson, and Helen H. Hobbs. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, 305(5685):869–872, aug 2004.

[95] Bingshan Li and Suzanne M. Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American journal of human genetics*, 83(3):311–321, sep 2008.

[96] Alkes L. Price, Gregory V. Kryukov, Paul I.W. de Bakker, Shaun M. Purcell, Jeff Staples, Lee-Jen Wei, and Shamil R. Sunyaev. Pooled association tests for rare variants in exon-resequencing studies. *The American Journal of Human Genetics*, 86(6):832–838, jun 2010.

[97] Wei Pan, Junghi Kim, Yiwei Zhang, Xiaotong Shen, and Peng Wei. A powerful and adaptive association test for rare variants. *Genetics*, 197(4):1081, 2014.

[98] R. J. Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

[99] R. A. Fisher. *Statistical methods for research workers*, pages 66–70. Springer New York, NY, 1992.

[100] D. V. Zaykin, Lev A. Zhivotovsky, P. H. Westfall, and B. S. Weir. Truncated product method for combining P-values. *Genetic Epidemiology*, 22(2):170–185, feb 2002.

[101] Frank Dudbridge and Bobby P.C. Koeleman. Rank truncated product of P-values, with application to genomewide association scans. *Genetic Epidemiology*, 25(4):360–366, dec 2003.

[102] Karen N Conneely and Michael Boehnke. So many correlated tests, so little time! rapid adjustment of *P* Values for multiple correlated tests. *The American Journal of Human Genetics*, 81:1158–1168, 2007.

[103] Dmitri V. Zaykin, Lev A. Zhivotovsky, Wendy Czika, Susan Shao, and Russell D. Wolfinger. Combining p-values in large scale genomics experiments. *Pharmaceutical statistics*, 6(3):217, 2007.

[104] Xiaoyi Gao, Joshua Starmer, and Eden R. Martin. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology*, 32(4):361–369, may 2008.

[105] Daniel J. Wilson. The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences of the United States of America*, 116(4):1195–1200, jan 2019.

[106] Olga A. Vsevolozhskaya, Fengjiao Hu, and Dmitri V. Zaykin. Detecting Weak Signals by Combining Small P-Values in Genetic Association Studies. *Frontiers in Genetics*, 10:1051, nov 2019.

[107] Miao-Xin Li, Hong-Sheng Gui, Johnny S H Kwan, and Pak C Sham. GATES: A Rapid and Powerful Gene-Based Association Test Using Extended Simes Procedure. *The American Journal of Human Genetics*, 88:283–293, 2011.

[108] Yaowu Liu and Jun Xie. Cauchy Combination Test: A powerful test With analytic p-Value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529):393–402, jan 2020.

[109] Yeonil Kim, Yueh Yun Chi, Judong Shen, and Fei Zou. Robust genetic model-based SNP-set association test using CauchyGM. *Bioinformatics*, 39(1), jan 2023.

[110] Ian Barnett, Rajarshi Mukherjee, and Xihong Lin. The Generalized Higher Criticism for Testing SNP-Set Effects in Genetic Association Studies. *Journal of the American Statistical Association*, 112(517):64–76, 2017.

[111] David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *https://doi.org/10.1214/009053604000000265*, 32(3):962–994, jun 2004.

[112] Ryan Sun, Shirley Hui, Gary D. Bader, Xihong Lin, and Peter Kraft. Powerful gene set analysis in GWAS with the Generalized Berk-Jones statistic. *PLoS Genetics*, 15(3), mar 2019.

[113] Olga A. Vsevolozhskaya, Min Shi, Fengjiao Hu, and Dmitri V. Zaykin. DOT: Gene-set analysis by combining decorrelated association statistics. *PLOS Computational Biology*, 16(4):e1007819, apr 2020.

[114] Dajiang J. Liu, Gina M. Peloso, and et al. Exome-wide association study of plasma lipids in >300,000 individuals. *Nature Genetics 2017 49:12*, 49(12):1758–1766, oct 2017.

[115] Tanya M. Teslovich, Kiran. Musunuru, and et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature 2010 466:7307*, 466(7307):707–713, aug 2010.

[116] Gina M. Peloso, Akihiro Nomura, and et al. Rare Protein-Truncating Variants in APOB, Lower Low-Density Lipoprotein Cholesterol, and Protection Against Coronary Heart Disease. *Circulation. Genomic and precision medicine*, 12(5):e002376, may 2019.

[117] Amanda J. Whitfield, P. Hugh R. Barrett, Frank M. Van Bockxmeer, and John R. Burnett. Lipid Disorders and Mutations in the APOB Gene. *Clinical Chemistry*, 50(10):1725–1732, oct 2004.

[118] Christiaan A. de Leeuw, Joris M. Mooij, Tom Heskes, and Danielle Posthuma. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Computational Biology*, 11(4):e1004219, apr 2015.

[119] Eva Boes, Stefan Coassin, Barbara Kollerits, Iris M Heid, and Florian Kronenberg. Genetic-epidemiological evidence on genes associated with HDL cholesterol levels: A systematic in-depth review. 2008.

[120] Hongliang Lü, Julyan Arbel, and Florence Forbes. Bayesian nonparametric priors for hidden Markov random fields. *Statistics and Computing*, 30(4):1015–1035, jul 2020.

[121] Il-Youp Kwak and Wei Pan. Gene- and pathway-based association tests for multiple traits with GWAS summary statistics. *Bioinformatics*, 33(1):64–71, jan 2017.