Bayesian Analysis of Case-control Genetic Association Studies in the Presence of Population Stratification or Genetic Model Uncertainty

by Linglu Wang

B.S. in Statistics 2007, Peking University, Beijing, China

A Dissertation Submitted to

The Faculty of
Columbian College of Arts and Sciences
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

August 31, 2011

Dissertation Directed by

Zhaohai Li

Professor of Statistics and Biostatistics

Gang Zheng

Mathematical Statistician

National Heart, Lung and Blood Institute

National Institute of Health

The Columbian College of Arts and Sciences of The George Washington University certifies that Linglu Wang has passed the Final Examination for the degree of Doctor of Philosophy as of May 6, 2011. This is the final and approved form of the dissertation.

Bayesian Analysis of Case-control Genetic Association Studies in the Presence of Population Stratification or Genetic Model Uncertainty

Linglu Wang

Dissertation Research Committee:

Zhaohai Li, Professor of Statistics and Biostatistics,
Dissertation Director

Gang Zheng, Mathematical Statistician,
National Heart, Lung and Blood Institute, National Institute of Health,
Dissertation Co-Director

Dante Verme, Professor of Epidemiology and Biostatistics,
Committee Member

Jonathan Stroud, Associate Professor of Statistics,
Committee Member

Acknowledgments

First and foremost, I would like to express my deepest appreciation to my advisor, Professor Zhaohai Li. This dissertation would have been impossible without his strong support through the past four years. He introduced me into the area of genetic statistics, and he provided me with great guidance not only in statistics, but also in many aspects of life and career.

I would like to thank my co-advisor Dr. Gang Zheng for his insight and enthusiasm in genetic statistics, which have had an important impact on the success of my research work. I truly appreciate the enormous amount of time he spent mentoring my work. I would also like to thank Dr. Qizhai Li for his constructive comments, which greatly improved the quality of my work.

I would like to extend my appreciation to all my committee members: Prof. Dante Verme, Prof. Jonathan Stroud, Prof. Heather Hoffman, Prof. Yinglei Lai, and Dr. Aiyi Liu. Their questions and suggestions have been very helpful for me to improve my dissertation. I would also like to thank the faculty of the Department of Statistics and the Department of Epidemiology and Biostatistics, who taught me fundamental knowledge in statistics and biostatistics. Special thanks goes to Prof. Joseph Gastwirth for the helpful discussion about my dissertation.

I am very thankful to all my supportive friends, both in DC area and in many other places. They stood by me in good times and bad times. Because of them I never feel I am thousands of miles away from home.

Finally, I would like to express my gratitude to my father Zhihong Wang, my mother Minghui Dong, and my grandmother Zhengzhuang Yao. Their never-ending love is always the motivation for me to do better in my life.

Abstract

Bayesian Analysis of Case-control Genetic Association Studies in the Presence of Population Stratification or Genetic Model Uncertainty

Genetic association studies, which examine the association between genetic markers and diseases, have become an important area in human health research. Population-based case-control study is one of the major designs in genetic association studies. Traditionally, the strength of genetic association is measured by the frequentist p-value of statistical tests. Bayesian analysis, which has been increasingly used in population-based case-control genetic studies, provides a different perspective of how to interpret the evidence of association in the data. An important measurement of strength of association in Bayesian analysis is the Bayes factor, which unifies both p-value and statistical power.

Bayesian analysis in case-control genetic association studies is susceptible to two pitfalls. The presence of hidden subpopulation, also known as population stratification, could potentially deviate the distribution of Bayes factor and hence lead to invalid conclusion in Bayesian analysis. In addition, when the underlying genetic model (i.e. recessive, additive, dominant, or sometimes multiplicative) is uncertain, how to provide a robust measurement of association in Bayesian analysis has become a challenge. The objective of this dissertation is to develop appropriate approaches to resolve the issues of hidden population stratification and genetic model uncertainty when conducting Bayesian analysis in case-control genetic association studies.

This dissertation is organized as follows:

Chapter 1 provides an introduction to the study topics and a summary of the results from the dissertation.

Chapter 2 provides a detailed review with respect to the case-control genetic association studies and the application of Bayesian analysis in those studies, as well as existing methods dealing with population stratification and model uncertainty in classical hypothesis testing.

Chapter 3 studies how the hidden population stratification would affect the asymptotic distribution of the estimate of odds ratio and its confidence interval, and how to correct for the impact using a principal components analysis method. Then we apply these corrected estimates and the asymptotic variances to correct for the hidden population stratification in Bayesian analysis. Our method is evaluated by both simulation studies and the application to HapMap data.

Chapter 4 develops a Bayes factor that measures the evidence of association contained in a maximum statistic (MAX) which is robust under genetic model uncertainty. The asymptotic distribution of MAX is derived and then confirmed by simulations. Our proposed Bayes factor is evaluated by simulation studies and real data application, and compared to an existing BF based on a standard test with Bayesian model averaging. We also discussed the relationship between our Bayes factor and the p-value, as well as their implications.

Finally, Chapter 5 discusses the strength and limitations of the proposed methods. Future research needs are also presented.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Summary of Results

The development of modern technology in genetic testing has enabled researchers to investigate the association between genetic markers and diseases. Population-based case-control study is one of the most widely used designs in genetic studies. Traditionally, the frequentist p-value has been used to measure the strength of association. It has come to the attention of statisticians that, in genome-wide association (GWA) studies in which 500,000 to a million markers are tested, small p-values are always observed, even if the power to detect true genetic association is low. Bayesian analysis has been used as an alternative to the frequentist approach in genetic association studies. In contrast to the classic hypothesis testing, Bayesian analysis uses Bayes factor as a measurement of evidence in the data supporting the alternative hypothesis of association against the null hypothesis of no association. Other types of Bayes factor measuring the evidence in test statistics rather than in the data have also been studied.

When there are hidden subpopulations and the allele frequencies and disease prevalence vary across those subpopulations, the measurement of association in population-based case-control studies can be biased. The impact and correction of hidden pop-

ulation stratification on classical hypothesis testing has been studied by many researchers. We investigate on how the hidden population stratification would affect the asymptotic distribution of the estimate of odds ratio for the genetic effect and its confidence interval, and how to adjust for the impact. It is notable that Bayesian analysis is also susceptible to hidden population stratification. We study the impact and correction of hidden population stratification in Bayesian analysis.

In genetic association studies, the statistical tests are genetic model specified. In other words, the optimal test statistic is contructed under one of the genetic models such as dominant, recessive, additive or multiplicative. However, in practice the genetic model is often unknown, especially for complex diseases. Misspecifying the genetic model could cause substantial power loss in both classical and Bayesian hypothesis testing. The MAX statistic, a robust measurement of association based on the Cochran-Armitage trend test, has been proposed and studied by researchers. We are going to study its asymptotic distribution, and then propose a Bayes factor type of measurement that measures the evidence contained in the MAX statistic. Our proposed Bayes factor has similar performance to the Bayesian model averaging method except that our method is overall much less sensitive to the prior of the genetic model than the Bayesian model averaging method.

1.1 Bayes factor in the presence of population stratification

Many methods to correct for population stratification have been studied in classical hypothesis testing. It is not clear, however, how the hidden population stratification would affect Bayesian hypothesis testing and how to correct for it in Bayesian analysis. In Chapter 3, we considered using the approximate Bayes factor proposed

by Wakefield [2007, 2009] which is expressed in terms of the odds ratio and its asymptotic variance. By doing this, we transformed the problem from Bayesian hypothesis testing to the estimates of odds ratio and its asymptotic variance. We started with the investigation on how the hidden population stratification would affect the asymptotic distribution of the estimate of odds ratio and its confidence interval. Using a large panel of null markers scanned in a genome-wide association studies, we applied a principal components analysis (PCA) method to identify the hidden subgroups, and then adjusted for the impact of population stratification on the estimate of odds ratio and its asymptotic variance. Then we applied these corrected estimate and the asymptotic variance to Wakefield's Bayes factor. We showed in simulation studies that the hidden population stratification would bias the estimates of odds ratio and distort its asymptotic variance and hence it has a serious impact on Bayesian analysis, and the proposed method yields an appropriate correction for the hidden population stratification. We also applied our approach to HapMap data as an illustration.

1.2 Bayes factor based on a maximum statistic

In case-control genetic association studies, the maximum of three trend tests (MAX) is more robust than any single trend test and Pearson's chi-squared test when the underlying genetic model is unknown. Zang et al. [2010] has derived the asymptotic distribution of MAX under the null hypothesis of no genetic association, but its distribution under alternative hypothesis was unknown. In Chapter 4, we started with the computation of the mean and covariance matrix of the three trend tests, and then derived the asymptotic and approximate distributions of MAX under the alternative hypothesis. Bayes factor based on the MAX test, as a robust Bayesian measurement of association, has not been studied. We proposed a Bayes factor based

on MAX, and discussed its computation. Simulations were conducted for several purposes: (1) to confirm the distribution of MAX we obtained; (2) to compare the performance of the proposed Bayes factor with an existing BF based on a standard test with Bayesian model averaging; (3) to show that our proposed Bayes factor agrees with the p-value of MAX when the statistical power is fixed. Using real data from genome-wide association studies, we further illustrated how this new measure can be used and compared the implications of p-value and our proposed Bayes factor. Based on our studies, we recommended that our proposed Bayes factor should be reported in addition to small p-values of MAX in genome-wide association studies, as p-value alone does not convey any knowledge about power but our Bayes factor unifies both p-value and power.

Chapter 2

Literature Review

2.1 Genetic models

Genetic studies address the association between genotype, which is an organism's hereditary information, and phenotype, which is the actual observed properties of the organism, such as the disease status. Denote the alleles of a diallelic single-nucleotide polymorphism (SNP) as A and B, and let A be the minor allele (the less common allele). Three genotypes are denoted as $(G_0, G_1, G_2) = (BB, AB, AA)$. The relationship between the three genotypes and many genetic diseases can be described by, but not limited to, three major genetic models: the recessive, additive, and dominant models. If the risk of having the disease is the same for individuals with 0 or 1 minor allele A (genotype BB or AB), the model is said to be recessive. If the risk of having the disease is the same for individuals with 2 or 1 minor allele A (genotype AA or AB), the model is said to be dominant. If the risk of having the disease for individuals with genotype AB is the average of the risk for those with AA and the risk for those with BB, the model is said to be additive.

Disease penetrance is often used to describe the risk of having a disease given a certain genotype, denoted as $\eta_i = \Pr(\text{case}|G_i)$ for $i = 0, 1, 2$. Denote genotype

relative risk (GRR) as $\lambda_j = \eta_j/\eta_0$ $(j = 1, 2)$. The genetic model is recessive (REC) when $\lambda_1 = 1$ and $\lambda_2 > 1$, additive (ADD) when $\lambda_1 = (1 + \lambda_2)/2$ and $\lambda_2 > 1$, and dominant (DOM) when $\lambda_1 = \lambda_2$ and $\lambda_2 > 1$. Another common genetic model, called multiplicative (MUL) model, assumes $\lambda_2 = \lambda_1^2$. When the genotype relative risk λ_j is close to 1, which is true for most risk-associated genetic markers, the multiplicative model (MUL) is almost equivalent to the additive model, proved by the first order Taylor expansion $\lambda_2 = \lambda_1^2 \approx 1 + 2(\lambda_1 - 1) = 2\lambda_1 - 1$. Hence we focus our study on the recessive, additive, and dominant models.

2.2 Genetic study designs and statistical tests

Genetic studies can be divided into two categories, i.e., population-based studies and family-based studies. Population-based studies use genetic information from unrelated individuals while family-based studies employ information from nuclear families. Each of these study designs has its own advantages and disadvantages. Generally speaking, population-based study designs provide greater power and are easier to implement when compared with family-based studies; however, the population-based study design is subject to hidden population stratification. Family-based study design is more robust to population stratification. But it requires more efforts in terms of designing the study and collecting samples. It has less power compared with population-based study design.

It has been shown that population-based case-control studies are a powerful strategy for identifying the loci that are associated with complex diseases [Risch and Merikangas, 1996; Risch, 2000]. Those studies consist of independent cases (subjects with the disease) and controls (subjects without the disease). In population-based case-control studies, a common measurement of the strength of association between

an allele and a disease is the odds ratio (OR). Sasieni [1997] compared five different types of ORs and their interpretations. They are:

1) Heterozygous OR: the odds ratio of the heterozygous genotype AB comparing to the genotype BB.

2) Homozygous OR: the odds ratio of the heterozygous genotype AA comparing to the genotype BB.

3) Linear OR: the maximum likelihood estimate (MLE) of the odds ratio for the counts of A alleles in a logistic regression model.

4) Allele-based OR: the odds ratio of the allele A comparing to allele B.

5) Serological OR: the odds ratio of the individuals with at least one copy of allele A comparing to those with no allele A.

Sasieni [1997] has also shown that, when the Hardy-Weinberg Equilibrium (HWE) holds in both case and control populations, the allele-based odds ratio is equal to the heterozygous one, and the homozygous odds ratio is the square of the heterozygous one. The Hardy-Weinberg Equilibrium states that under the condition of random mating in a large population, the two alleles of an individual come independently from both parents, and the genotype frequencies are products of the allele frequencies.

There are two commonly used statistical tests for population-based case-control studies. The first approach, the Pearson's χ^2 test, is based on the 2×2 allelic contingency table that classifies cases and controls according to their alleles. The number of observations is doubled because each individual has two alleles. This approach ignores the genotypic information. Moreover, the Pearson's χ^2 test assumes the Hardy-Weinberg Equilibrium. When the Hardy-Weinberg Equilibrium is violated

due to various reasons such as selection, mutation, migration, inbreeding, or small population size, the Pearson's χ^2 test is biased and invalid [Sasieni, 1997].

The second approach, the Cochran-Armitage trend test (CATT) [Armitage 1955], is based on the 2×3 genotype contingency table that classifies cases and controls according to their genotypes. The Cochran-Armitage trend test is optimal under the additive genetic effect with respect to the number of risk alleles in the genotype. Other trend tests based on the recessive or dominant genetic models may also be used under non-additive models. Under simple random sampling, these tests can also be applied to cross-sectional studies such as surveys where prevalent disease (cases) and non-disease (controls) individuals are sampled together. One advantage of this approach is that it can be used when the sampled population is not in the state of Hardy-Weinberg Equilibrium. One disadvantage of Cochran-Armitage trend test is that mis-specification of genetic model through the scores assigned to different genotypes can result in a substantial loss of power.

It has been shown that using any trend test optimal for one genetic model is not robust when the model is incorrectly specified [Freidlin et al., 2002]. To avoid such situation, Freidlin et al. [2002] proposed a robust test statistic, denoted by MAX, which is the maximum of three trend tests optimal respectively for the recessive, additive and dominant models. Note that under the null hypothesis, MAX no longer follows the asymptotic normal distribution as the individual trend test statistic. Zheng et al. [2010] derived its asymptotic distribution under the null hypothesis.

Other robust tests have also been studied for genetic association studies, including: a maximum of three likelihood ratio tests [Gonzalez et al., 2008], constrained likelihood ratio tests [Wang and Sheffield, 2005], and a two-stage test using deviation from Hardy-Weinberg equilibrium (HWE) [Zheng and Ng, 2008]. All these robust

tests have comparable power performance with MAX.

2.3 Population stratification

In case-control genetic association studies, a main concern is that the population may consist of several hidden subpopulations, also known as population stratification (PS) [e.g., Devlin and Roeder, 1999; Zheng et al., 2010]. When allele frequencies and disease prevalence rates vary across subpopulations, it can cause bias and variance distortion in the commonly used allele-based analysis [Li et al. 2009]. This population stratification, although well defined, is hidden and latent in the sense that neither the subpopulation (stratum) nor its size are known. Because population stratification is latent, cases and control are disproportionally sampled from the subpopulations. Hence, varying disease prevalence across the subpopulations in addition to the varying allele frequency can lead to bias and variance distortion in genotype-based analysis.

There are several prevailing methods dealing with population stratification, summarized as follows:

(1) The genomic control (GC) method, proposed by Devlin and Roeder [1999]. It uses markers at uncorrelated null loci to estimate the variance distortion factor (VIF) in the test statistic, and then rescale the test statistic using the estimated VIF. One disadvantage of genomic contro method is that, using a uniform VIF can be insufficient at markers with greater differentiation in their allele frequencies across ancestral populations than others, or can be superfluous at markers with less differentiation [Price et al., 2006; Epstein et al., 2007].

(2) The structured association (SA) method, proposed by Pritchard et al. [2000]. It uses markers that are informative for the ancestry to estimate the population substructure, while simultaneously estimating subpopulation allele frequencies. One

advantage of structured association method is that it can assign individuals to either one subpopulation (without admixture) or multiple subpopulations (with admixture). However, because it employs Markov chain Monte Carlo (MCMC) algorithm, it can become very computationally intensive when a large number of structure informative markers are needed.

(3) The EIGENSTRAT algorithm based on principal components analysis (PCA), proposed by Price et al. [2006] and further studied by Li and Yu [2008]. The EIGENSTRAT algorithm identifies continuous axes with large genetic variation (i.e. the top eigenvectors of a genetic similarity matrix between sampled subjects), and then explicitly model ancestry differences between cases and controls and cluster the subjects into subgroups based on those continuous axes. Intuitively, the axes of variation reduce the large genetic dataset to a small number of dimensions, describing as much variability as possible. When the differences among subpopulations are subtle and the population structure can only be identified by a large number of markers, the principal components analysis (PCA) method is more computationally efficient compared to the structured association (SA) method [Price et al. 2006].

In addition, some recent work on population stratification and other population substructure and their impact can be seen in Zhu et al. [2008], Miclaus et al. [2009], Astle and Balding [2009], and Zheng et al. [2010].

2.4 Bayesian analysis for genetic association studies

In genetic studies, especially in genome-wide association (GWA) studies in which all or most genetic markers (usually 500,000 to a million) for individuals are tested, we are often interested in the strongest associations based on p-values. However, small

p-values are always observed, even though the power to detect these associations is not high. For example, a p-value of 10^{-8} may seemingly indicate very strong evidence against the null hypothesis (H_0) of no association between genetic marker and disease. However, if the power of the test is very low such that the probability to observe the data is as unlikely under the alternative hypothesis H_1 as under H_0, using p-value alone to conclude the existence of genetic association would be misleading. Sawcer [2010] conducted a simulation with four studies each having the same p-value 0.01, but they had sample sizes ranging from 200 to 20,000 with different odds ratios (2.11 to 1.09).

Bayesian hypothesis testing, on the other hand, measures the evidence in the data against the null hypothesis (H_0) by a Bayes factor (BF) along with the prior odds (PO) of the association [Kass and Raftery, 1995]. This measurement, the posterior probability of association (PPA), is expressed as PPA=Pr(H_1|data)=BF$\times$ PO/(1+BF $\times$ PO), where PO = Pr(H_1)/Pr(H_0) is the prior odds of H_1, and the Bayes factor (BF) is given by BF=Pr(data|H_1)/Pr(data|H_0), the ratio of the probability of the data under H_1 over that under H_0. Specifically, Sawcer [2010] illustrates that the BF is a ratio of the statistical power over the significance of association. If PO is chosen to be constant over SNPs, the BF gives the same ranking of SNPs as the PPA. Hence researchers often use BF as the primary summary of the evidence for association at a SNP, which leaves the choice of PO flexible. Jeffreys [1961] suggests the following interpretations of BF: the association is not worth mentioning if the BF is between 1 and 3.2, is substantial if the BF is between 3.2 and 10, is strong if the BF is between 10 and 100, and is decisive if the BF is greater than 100. However, as is pointed out by Stephens and Balding [2009], in genome-wide association (GWA) studies where only a small number of SNPs are associated with diseases among all

the SNPs tested (i.e. $\Pr(H_1)$ is very small), the BF has to be very large to provide convincing evidence for an association.

In practice, while p-value is still a main tool used to evaluate the significance of association, BF has been increasingly reported to support the small p-values in genome-wide association (GWA) studies. For example, WTCCC [2007] reported BF for 24 SNPs that are strongly associated with seven diseases (p-values less than 5×10^{-7} for either trend test or exact test), and an additional 58 SNPs that are moderately associated with those diseases (p-values between 5×10^{-7} and 10^{-5} for either trend test or exact test). Stephens and Balding [2009] summarized both p-values and BF for selected SNPs associated with diseases, as well as their PPAs under different PO. The chosen range of $\Pr(H_1)$ for genome-wide association (GWA) studies has been 10^{-3} and 10^{-6}, based on the results of linkage studies and HapMap data [WTCCC, 2007; Sawcer, 2010].

To compute the Bayes factor, the parameters γ (including the odds ratio and other parameters) are integrated out as $\Pr(\text{data}|H_1) = \int \Pr(\text{data}|\gamma, H_1)\pi(\gamma|H_1)d\gamma$ where $\pi(\gamma|H_1)$ is the prior distribution of the parameters γ under H_1. $\Pr(\text{data}|H_0)$ can be computed in the similar way. Bayes factor is often calculated using a Laplace approximation [Kass and Raftery, 1995]. Some limitations for computing Bayes factor are: 1) It requires explicit assumptions for the distribution of every parameter in γ, and 2) it could be computational demanding as it involves evaluating multiple integrals same as the dimension of γ. For those reason, Bayes factors are employed less frequently than they otherwise would be.

A simple alternative approach is to approximate the Bayes factor by replacing the data with the maximum likelihood estimate (MLE) of $log(\text{OR})$. Wakefield [2007, 2009] proposed an approximate Bayes factor (denoted by BF_W) for case-control genetic

association studies, including genome-wide association (GWA) studies, which is very simple to compute. It is written as:

$$\mathrm{BF}_W = \frac{\Pr(\hat{\beta}|H_1)}{\Pr(\hat{\beta}|H_0)} = \sqrt{\frac{V}{V+\sigma^2}} \exp\left\{\frac{\hat{\beta}^2\sigma^2}{2V(V+\sigma^2)}\right\}$$

where $\hat{\beta}$ is the MLE of the $log(\mathrm{OR})$ and V is its asymptotic variance, and W is the prespecified variance in the normal prior for $\beta = log(\mathrm{OR})$. In computation, V is usually replaced by its estimate $\hat{V}$. The derivation of BF_W is based on the assumptions that: (1) $\hat{\beta}|\beta \sim N(\beta, V)$ and $\beta \sim N(0, \sigma^2)$, and (2) $\hat{\beta}$ is independent from the MLE of other parameters, and β is independent from other parameters. Those assumptions are reasonble in genetic association studies. Comparing to the original BF, Wakefield's approximate BF does not involve integrals and it does not require the assumptions of distributions for parameters other than β.

Johnson [2005, 2008] has also considered BFs based on a standard test statistics that follows normal, chi-squared, F-, or t- distributions under H_0 and H_1. Because the distributions of test statistics do not depend on unknown model parameters, this approach eliminates much of the subjectivity that is usually associated with the definition of Bayes factors, and drastically simplifies their computation.

Chapter 3

Bayes Factors in the Presence of Population Stratification

3.1 Introduction

In case-control genetic association studies, a main concern is that the population may consist of hidden subpopulations, also known as population stratification (PS) [e.g., Devlin and Roeder, 1999; Astle and Balding, 2009; Zheng et al., 2010]. With the existence of latent PS, cases and control are disproportionally sampled from the subpopulations. Li et al. (2009) has shown that, when the allele frequencies vary across subpopulations, the variances of tests for genetic association can be inflated. In addition, if the disease prevalance vary across subpopulations, the test results can be biased.

Correcting for hidden population stratification has been focused on hypothesis testing, such as the genomic control (GC) method [Devlin and Roeder, 1999], the structured association (SA) method [Pritchard et al., 2000a, b], and the EIGENSTRAT algorithm based on principal components analysis (PCA) [Price et al. 2006], extended by Li and Yu [2008]. The genomic control method uses markers at uncorrelated null loci to estimate the variance inflation factor and rescale the test statistic with this factor. However, using a uniform adjustment can be inadequate at markers

with more differentiation in their allele frequencies across ancestral populations than others, or can be superfluous at markers with less differentiation [Price et al., 2006; Epstein et al., 2007]. The structured association method uses Markov chain Monte Carlo (MCMC) algorithm and attempts to assign individuals to one subpopulation (without admixture) or multiple subpopulations (with admixture) on the basis of their genotypes, while simultaneously estimating subpopulation allele frequencies. It can become very computationally intensive when a large number of structure informative markers are needed. The EIGENSTRAT algorithm based on principal components analysis (PCA) method identifies continuous axes with large genetic variation and then explicitly model ancestry differences between cases and controls. When the differences among subpopulations are subtle and the population structure can only be identified by a large number of markers, the principal components analysis method is more computationally efficient compared to the structured association method [Price et al. 2006].

It is natural to think that the existing methods to correct for hidden population stratification in hypothesis testing can be modified and applied to correct for the estimates of the odds ratio (OR) and its confidence interval (CI). In this chapter, one of our tasks is to apply the extended EIGENSTRAT algorithm of Li and Yu [2008] to the genotype data to identify the hidden subpopulation structure, where the number of subpopulations is decided by the Gap-statistic method [Tibshirani et al., 2001]. Once the subjects are clustered into identified subpopulations, we can obtain the stratification-adjusted estimates for different types of odds ratios and their confidence intervals.

To our best knowledge, no research has been reported in the literature to allow hidden population stratification in Bayesian applications to genetic association stud-

ies, although Stephens and Balding [2009] mentioned it. Bayesian analysis is a useful alternative to frequentist statistics in genetic association studies [WTCCC, 2007]. In Bayesian hypothesis testing, the evidence in the data against the null hypothesis (H_0) is often measured by a Bayes factor (BF) along with the prior odds of the association [Kass and Raftery, 1995]. Instead of considering the regular BF, we use an approximate Bayes factor (BF_W) proposed by Wakefield [2007, 2009] for genetic association studies, which is a function of the estimates of odds ratio and its asymptotic variance with a single tuning parameter. Our ultimate objective in this chapter is to examine the impact of hidden population stratification on the BF_W and to correct for it, using the adjusted estimation of odds ratio and its asymptotic variance obtained from EIGENSTRAT algorithm based on PCA method.

3.2 OR in the presence of population stratification

3.2.1 Notation and model

Consider a genetic marker of interest with two alleles A and B, which forms three genotypes $(G_0, G_1, G_2) = (BB, AB, AA)$. The marker can be regarded as a covariate (each of its genotype regarded as one exposure level), and each individual has one of the three genotypes. Assume that the study population consists of well-defined and known case and control populations. Cases and controls are unmatched here. Further, the study population consists of K hidden subpopulations, which is often formed due to lack of random mating among ancestors. This type of subpopulation is defined by the disease prevalence Pr(case) and allele frequency Pr(allele = A) such that Pr(case) and Pr(allele = A) are constant within a subpopulation but they vary among the subpopulations [Zheng et al., 2010]. This population stratification (PS), although well defined, is hidden and latent in the sense that neither the subpopulation

Table 3.1: The genotype-based table for the kth subpopulation. The counts in the table are not observable.

	BB	AB	AA	Total
Case	$r_{0,k}$	$r_{1,k}$	$r_{2,k}$	r_k
Control	$s_{0,k}$	$s_{1,k}$	$s_{2,k}$	s_k
Total	$n_{0,k}$	$n_{1,k}$	$n_{2,k}$	n_k

Table 3.2: The allele-based table for the kth subpopulation. The counts in the table are not observable.

	B	A	Total
Case	a_k	b_k	$2r_k$
Control	c_k	d_k	$2s_k$
Total	$2n_{0,k} + n_{1,k}$	$2n_{2,k} + n_{1,k}$	$2n_k$

(stratum) nor its size are known. This also draws a clear difference between the population stratification that we consider here and the stratified sampling in most survey sampling, in which strata are known.

Consider a random sample of r cases and a random sample of s controls drawn from the case and control populations, respectively. Although not observable, the counts of (G_0, G_1, G_2) in the cases and controls that are from the kth subpopulation are denoted as $(r_{0,k}, r_{1,k}, r_{2,k})$ and $(s_{0,k}, s_{1,k}, s_{2,k})$, respectively. This forms the genotype-based 2×3 table for the kth subpopulation (illustrated by Table 3.1). Let $r_k = \sum_{i=0}^{2} r_{i,k}$, $s_k = \sum_{i=0}^{2} s_{i,k}$, and $n_k = r_k + s_k$. Let $a_k = 2r_{0,k} + r_{1,k}$, $b_k = 2r_{2,k} + r_{1,k}$, $c_k = 2s_{0,k} + s_{1,k}$, and $d_k = 2s_{2,k} + s_{1,k}$, which form the allele-based 2×2 table for the kth subpopulation (Table 3.2). Denote $a = \sum_{k=1}^{K} a_k$, $b = \sum_{k=1}^{K} b_k$, $c = \sum_{k=1}^{K} c_k$, and $d = \sum_{k=1}^{K} d_k$. We observe allele counts a, b, c and d for r cases and s controls, not directly a_k, b_k, c_k and d_k for each k. Let $n = r + s$. Thus, inference can be only done using (a, b, c, d) unless the PCA method, described later, is used to infer (a_k, b_k, c_k, d_k) for each k.

Let $\{S = k\}$ represent the event that a subject is in the kth hidden subpopulation.

Denote $w_{r,k} = \Pr(S = k|\text{case})$ and $w_{s,k} = \Pr(S = k|\text{control})$. As shown in Li et al. [2009], when the disease prevalence $\Pr(\text{case}|S = k)$ varies across k, $w_{r,k} \neq w_{s,k}$ for some k. Let $p_{r,k} = \Pr(\text{allele} = A|S = k, \text{case})$ and $p_{s,k} = \Pr(\text{allele} = A|S = k, \text{control})$. The null hypothesis of no association is given by H_0: $p_{r,k} = p_{s,k}$ for each k, while the alternative hypothesis is specified as H_1: $p_{r,k} \neq p_{s,k}$ for some k. Note $p_r = \Pr(\text{allele} = A|\text{case}) = \sum_k p_{r,k} w_{r,k}$ and $p_s = \Pr(\text{allele} = A|\text{control}) = \sum_k p_{s,k} w_{s,k}$. An impact of hidden and unmodeled PS is that $p_{r,k} = p_{s,k}$ for each k but $p_r \neq p_s$, which is a spurious association [Zheng et al., 2010]. An artificial example with $K = 2$ is that we sample all cases from one subpopulation and all controls from another due to hidden PS. Then, even though there is no association in each subpopulation, there will be an association in the total population just because we sample cases and controls disproportionally from each subpopulation. Note that if we knew the subpopulations, we would design a matched case-control study or correct it in a stratified analysis to avoid this issue.

Denote estimates of p_r and p_s as $\hat{p}_r = b/(2r)$ and $\hat{p}_s = d/(2s)$, respectively. When there is no hidden PS, the maximum likelihood estimate (MLE) of log(OR) for the risk allele A is given by $\hat{\beta} = \log(\widehat{\text{OR}}) = \log\{\hat{p}_r(1 - \hat{p}_s)\}/\{\hat{p}_s(1 - \hat{p}_r)\}$, which asymptotically follows $N(\beta, V_n)$, where $\beta = \log\{p_r(1 - p_s)\}/\{p_s(1 - p_r)\}$ and $V_n = 1/\{2rp_r(1 - p_r)\} + 1/\{2sp_s(1 - p_s)\}$ [Lachin, 2000]. In the presence of the hidden PS and when we only observe counts (a, b, c, d) and (r, s), if the above MLE $\hat{\beta}$ is still used, one ignores the hidden PS.

3.2.2 OR and its confidence interval in the presence of population stratification

In the presence of hidden population stratification, an impact of ignoring it and using the above $\hat{\beta}$ is given in the following theorem:

Theorem 3.1. *Using the above notation, in the presence of hidden PS,* $\hat{\beta} = \log(\widehat{\text{OR}})$*, computed as in Section 3.2.1, has an asymptotic* $N(\beta^*, V_n^*)$*, where* $\beta^* = E(\hat{\beta})$ *and* $V_n^* = Var(\hat{\beta})$ *are given by*

$$\begin{aligned}
\beta^* &= \log\left\{\frac{(\sum_k p_{r,k}w_{r,k})\,(1-\sum_k p_{s,k}w_{s,k})}{(\sum_k p_{s,k}w_{s,k})\,(1-\sum_k p_{r,k}w_{r,k})}\right\},\\
V_n^* &= \frac{1}{2r}\left\{\frac{1}{p_r(1-p_r)} + \frac{\sum_k w_{r,k}(p_{r,k}-p_r)^2}{p_r^2(1-p_r)^2}\right\}\\
&\quad + \frac{1}{2s}\left\{\frac{1}{p_s(1-p_s)} + \frac{\sum_k w_{s,k}(p_{s,k}-p_s)^2}{p_s^2(1-p_s)^2}\right\}.
\end{aligned}$$

Note that $V_n^* \geq V_n$ *for any* n*, where the equality holds if and only if* $p_{r,k} = p_r$ *and* $p_{s,k} = p_s$ *for all* k*, under which there is no PS and* $\beta^* = \beta$*, where* β *is the true common* log(OR) *for all strata.*

Proof: When PS exists, with the correct stratification identification, denote the sample proportions of $(p_{r,l}, ..., p_{s,l}, ..., w_{r,l}, ..., w_{s,l}, ...)$ by $(\hat{p}_{r,l}, ..., \hat{p}_{s,l}, ..., \hat{w}_{r,l}, ..., \hat{w}_{s,l}, ...)$. From our setting, $(\hat{w}_{r,1}, ..., \hat{w}_{r,K}) \sim Mul(r, w_{r,1}, ..., w_{r,K})$. Hence $E(\hat{w}_{r,k}) = w_{r,k}$, $Var(\hat{w}_{r,k}) = w_{r,k}(1-w_{r,k})/r$, and $Cov_{k_1 \neq k_2}(\hat{w}_{r,k_1}, \hat{w}_{r,k_2}) = -w_{r,k_1}w_{r,k_2}/r$. Similar results hold for the controls. Within the kth subpopulation, $b_k \sim Binomial(r_k, p_{r,k})$. Hence $E(\hat{p}_{r,k}) = p_{r,k}$, and $Var(\hat{p}_{r,k}) = p_{r,k}(1-p_{r,k})/(2rw_{r,k})$. Similar results hold for the controls. Then, for the log odds ratio $\hat{\beta} = \log(\widehat{\text{OR}})$, when sample sizes are large enough,

$$E(\hat{\beta}) = E\left\{\log\frac{(\sum_k \hat{p}_{r,k}\hat{w}_{r,k})(1-\sum_k \hat{p}_{s,k}\hat{w}_{s,k})}{(\sum_k \hat{p}_{s,k}\hat{w}_{s,k})(1-\sum_k \hat{p}_{r,k}\hat{w}_{r,k})}\right\} \approx \log\frac{(\sum_k p_{r,k}w_{r,k})(1-\sum_k p_{s,k}w_{s,k})}{(\sum_k p_{s,k}w_{s,k})(1-\sum_k p_{r,k}w_{r,k})}.$$

Using the variance-covariance matrix of $(\hat{p}_{r,l}, ..., \hat{p}_{s,l}, ..., \hat{w}_{r,l}, ..., \hat{w}_{s,l}, ...)$ and the δ-method, we get $Var(\hat{p}_r) = \{p_r(1-p_r) + \sum_k w_{r,k}(p_{r,k}-p_r)^2\}/(2r)$ and $Var(\hat{p}_s) = \{p_s(1-p_s) + \sum_k w_{s,k}(p_{s,k}-p_s)^2\}/(2s)$, which were also obtained by Li et al. [2009]

using a different method. By further applying the δ-method, we obtain

$$\begin{aligned} Var(\hat{\beta}) &= Var\left(\log\frac{\hat{p}_r}{1-\hat{p}_r}\right) + Var\left(\log\frac{\hat{p}_s}{1-\hat{p}_s}\right) \\ &\approx \frac{1}{2r}\left(\frac{1}{p_r(1-p_r)} + \frac{\sum_k w_{r,k}(p_{r,k}-p_r)^2}{p_r^2(1-p_r)^2}\right) \\ &\quad + \frac{1}{2s}\left(\frac{1}{p_s(1-p_s)} + \frac{\sum_k w_{s,k}(p_{s,k}-p_s)^2}{p_s^2(1-p_s)^2}\right). \square \end{aligned}$$

Numerical results show that β^* is always different from the true β in the presence of PS. Thus, $\hat{\beta}$, ignoring the hidden PS, is asymptotically biased and that its asymptotic variance V_n is underestimated. Therefore, when the PS is present but unmodeled, the confidence interval for log(OR) would not have correct coverage probability.

3.2.3 Identifying and correcting for hidden population stratification

Identifying and correcting for hidden PS has been focused on hypothesis testing problems, e.g., testing whether or not the allele frequency in cases equals that in controls [Li et al., 2009]. Our ultimate goal is Bayesian hypothesis testing. To this end, we have to address the estimation problem which will be related to the approximate Bayes factor (BF_W) later.

First, we apply a common procedure to identify and correct for hidden PS. That is, a panel of m null markers (single-nucleotide polymorphisms - SNPs), obtained from existing genome-wide association (GWA) studies and not associated with common diseases, is used to detect the hidden population structure using the principal components analysis (PCA) method. The steps of the PCA clustering [Price et al., 2006; Li and Yu, 2008] are described as follows.

1) Denote the genotype matrix of m null SNPs and n subjects by $G = \{g_{i,j}\}_{m\times n}$, where $g_{i,j}(= 0, 1, 2)$ is the number of allele A on the ith SNP for the jth subject. Denote the row mean of the ith SNP by $\mu_i = \sum_j g_{i,j}/n$. Let $p_i =$

$(1 + \sum_j g_{i,j})/(2 + 2n)$. Then $X = \left\{ \frac{g_{i,j} - \mu_i}{2\sqrt{p_i(1-p_i)}} \right\}_{m \times n}$ is the standardized matrix and $\psi_{n \times n} = X^T X$ is the similarity matrix with the elements describing the ancestry "distance" between subjects, with a dimension equal to the number of subjects in the dataset. The greater an entry in the matrix ψ, the larger ancestry "distance" between two subjects.

2) Let $\lambda_1 \geq ... \geq \lambda_L > 0$ be the top L largest eigenvalues of ψ, and $\vec{v_1}$,...,$\vec{v_L}$ be the corresponding eigenvectors, where $\vec{v_j} = (v_{1,j}, ..., v_{n,j})^T$ and $\sum_i v_{i,j}^2 = \lambda_j$. Then $V = \{\vec{v_1},...,\vec{v_L}\}_{n \times L}$ is the principal component matrix. The eigenvectors $\vec{v_1}$,...,$\vec{v_L}$ can be seen as continuous axes of genetic variation, which intuitively reduce the large genotype dataset to a small number (L) of dimensions, describing as much genetic variability as possible.

3) Using $\vec{v_1}$,...,$\vec{v_L}$ (in this chapter we used $L = 10$, which is the default value in Price et al. [2006]) to cluster the n subjects into $\tilde{K}$ groups, where the value of $\tilde{K}$ is decided by Gap-statistic [Tibshirani et al., 2001]. The brief idea of the Gap-statistic is: if there exists a threshold $\tilde{K}$ such that the decrease of the within-cluster dispersion flattens when the number of clusters exceeds $\tilde{K}$, then the appropriate value for the number of clusters is $\tilde{K}$, as increasing the number of clusters would not dramatically improve the grouping result. An advantage of using the gap statistic is that the estimated cluster number could be 1, which means that no clustering structure has been identified.

After the n subjects are clustered into $\tilde{K}$ groups using PCA method, stratified analysis can be carried out. In Sasieni [1997], five different ORs were discussed. Here we consider four of them: the linear OR based on the logistic regression model, the allele-based OR comparing the risk of allele A to allele B, the heterozygous OR com-

paring the risk of genotype AB to genotype BB, and the homozygous OR comparing the risk of genotype AA to genotype BB. For each of the four ORs, we compute its estimate and CI adjusted for the identified stratification.

For the linear OR, we use logistic regression model

$$\text{logit}(p_j) = \alpha + \beta_{linear} g_{i,j} + \sum_{k=1}^{\tilde{K}-1} \beta_k I_{k,j} + \epsilon_j,$$

where p_j is the risk for the jth subject, $I_{k,j}$ is an indicator taking 1 if the jth subject belongs to the kth identified subpopulation and 0 otherwise, and the number of risk allele A is $g_{i,j} = 0, 1, 2$ for G_0, G_1, G_2, respectively. The MLE of β_{linear} and its asymptotic variance (and the CI) can be obtained by Newton-Raphson iteration. Denote the OR by $\widehat{\text{OR}}_{linear}$.

For the allele-based OR, we use Mantel-Haenszel estimate of common odds ratio, given by $\widehat{\text{OR}}_{allele} = (\sum_{k=1}^{\tilde{K}} b_k c_k/2n_k)/(\sum_{k=1}^{\tilde{K}} a_k d_k/2n_k)$, where $(a_k, b_k, c_k, d_k, n_k)$ are the counts in the kth identified subgroup where k is determined by the PCA method. There are various approximations of its variance. We use the one given by Robins et al. (1986). Denote statistics:

$$S_1 = \sum_k a_k d_k/2n_k;\ S_2 = \sum_k c_k b_k/2n_k;\ S_3 = \sum_k (a_k + d_k) a_k d_k/4n_k^2;$$

$$S_4 = \sum_k (b_k + c_k) b_k c_k/4n_k^2;\ S_5 = \sum_k \{(a_k + d_k) b_k c_k + (b_k + c_k) a_k d_k\}/4n_k^2.$$

Then the estimate of the variance is given by

$$\widehat{V}_n = \widehat{Var}\left\{\log(\widehat{\text{OR}}_{allele})\right\} = \frac{S_3}{2S_1^2} + \frac{S_5}{2S_1 S_2} + \frac{S_4}{2S_2^2}.$$

The estimate of the heterozygous OR is given by

$$\widehat{\text{OR}}_{hetero} = (\sum_k r_{1,k} s_{0,k}/n_k)/(\sum_k r_{0,k} s_{1,k}/n_k).$$

The variance estimate is similar to that of OR_{allele} except that we replace a_k, b_k, c_k and d_k by $r_{0,k}$, $r_{1,k}$, $s_{0,k}$ and $s_{1,k}$, respectively, and replace $2n_k$ by n_k.

For the homozygous OR, $\widehat{\text{OR}}_{homo} = (\sum_k r_{2,k}s_{0,k}/n_k)/(\sum_k r_{0,k}s_{2,k}/n_k)$. The variance estimate is similar to that of OR_{allele} except that we replace a_k, b_k, c_k and d_k by $r_{0,k}$, $r_{2,k}$, $s_{0,k}$ and $s_{2,k}$, respectively, and replace $2n_k$ by n_k.

When Hardy-Weinberg Equilibrium (HWE) holds in the data, $\text{OR}_{hetero} = \text{OR}_{allele}$ and $\text{OR}_{homo} = \text{OR}^2_{allele}$ asymptotically, as Sasieni [1997] pointed out.

3.3 Impact and correction of hidden population stratification in Bayesian analysis

In Bayesian hypothesis testing, the Bayes factor $\text{BF} = Pr(\text{data}|H_1)/Pr(\text{data}|H_0)$ is used to measure the strength of association, which involves the integrations of the likelihoods of the data under H_0 and H_1 with respect to the prior densities of the parameters. The H_0 and H_1 are specified in Section 3.2.1. BF is often calculated using a Laplace approximation [Kass and Raftery, 1995]. A simple alternative approach is the approximate BF by replacing the data in the BF with a test statistic T: $\text{BF}_J = \Pr(T|H_1)/\Pr(T|H_0)$ [Johnson, 2005]. Wakefield [2007] considered such an BF for case-control genetic studies, where the data in the BF are replaced with the MLE of log(OR), which only depends on a single tuning parameter β=log(OR). By using a normal prior on β, $\beta = \log(\text{OR}) \sim N(0, \sigma^2)$, Wakefield [2007] derived the following BF_W

$$\text{BF}_W = \sqrt{\frac{\hat{V}_n}{\hat{V}_n + \sigma^2}} \exp\left\{\frac{\hat{\beta}^2\sigma^2}{2\hat{V}_n(\hat{V}_n + \sigma^2)}\right\} \tag{3.1}$$

where $\hat{\beta}$ is the MLE of the log(OR) and $\hat{V}_n$ is the estimate of its asymptotic variance V_n, and σ^2 is the prespecified variance in the normal prior. A larger value of the BF_W indicates stronger evidence to support H_1.

The following theorem states the asymptotic behavior of the BF_W in terms of hidden population stratification.

Theorem 3.2. *In the absence of PS, as n is large enough and under H_0,*

$$\sqrt{n}\mathrm{BF}_W \approx \sqrt{\frac{I^{-1}(\beta)}{\sigma^2}} \exp\left(\frac{\chi_1^2}{2}\right), \tag{3.2}$$

where $I^{-1}(\beta)$ is the inverse of the Fisher information about β in a single sample and χ_1^2 is a random variate following a chi-squared distribution with 1 degree of freedom (df). In the presence of hidden PS, as n is large enough and under H_0,

$$\sqrt{n}\mathrm{BF}_W \approx \sqrt{\frac{I^{-1}(\beta^*)}{\sigma^2}} \exp\left\{\frac{\chi_1^2(\delta_n^*)}{2}\right\}, \tag{3.3}$$

where $\delta_n^* = (\beta^*)^2/V_n^*$ is the non-central parameter in the generalized chi-square distribution, and β^* and V^* are given in Theorem 3.1.

The proof is outlined below. Note that (3.2) follows from the fact that, when n is large enough, $Z^2 = \hat{\beta}^2/\hat{V}_n$ asymptotically follows a chi-squared distribution with 1 df under H_0. Since $\hat{V}_n \approx 0$ when n is large enough, $\exp\left[\hat{\beta}^2\sigma^2/\{2\hat{V}_n(\hat{V}_n+\sigma^2)\}\right] \approx \exp(\chi_1^2/2)$. On the other hand, $n\hat{V}_n \approx I^{-1}(\beta)$. For (3.3), as n is large enough, $\hat{\beta} \sim N(\beta^*, V_n^*)$ under H_0 (Theorem 3.1). Thus, $\exp\left\{\frac{\hat{\beta}^2\sigma^2}{2\hat{V}_n(\hat{V}_n+\sigma^2)}\right\} \approx \exp\{\frac{1}{2}\chi_1^2(\delta_n^*)\}$. Also note that $n\hat{V}_n \approx I^{-1}(\beta^*)$ in the presence of PS. This leads to (3.3).

Since the BF_W is a function of the MLEs of log(OR) and its asymptotic variance, we can correct the BF_W in the presence of hidden PS by correcting the OR and its variance as in Section 3.2.3. Then we substitute the corrected log(OR) and its asymptotic variance into (3.1). This approach is simple to apply in practice.

3.4 Simulation studies and results

3.4.1 Impact on estimation of log(OR) without correcting for hidden population stratification

We first conducted simulations to examine the asymptotic mean and variance of the MLE ($\hat{\beta}$) of log(OR) in the presence of PS. Four scenarios are considered,

Table 3.3: Four scenarios (denoted as Sc1 to Sc4), each of which contains four subpopulations. The allele frequencies in cases equal to those of controls multiplied by $\mathrm{OR}/(1 - p_{s,k} + \mathrm{OR} \times p_{s,k})$ in each subpopulation.

Scenario	Cases*	Controls*	Allele freq in controls**	OR
Sc1	(0.35, 0.3, 0.2, 0.15)	(0.15, 0.2, 0.3, 0.35)	(0.1, 0.9, 0.2, 0.8)	1
Sc2	(0.25, 0.25, 0.45, 0.05)	(0.05, 0.45, 0.25, 0.25)	(0.8, 0.2, 0.7, 0.3)	1.1
Sc3	(0.35, 0.3, 0.2, 0.15)	(0.15, 0.2, 0.3, 0.35)	(0.7, 0.3, 0.6, 0.4)	1.2
Sc4	(0.25, 0.25, 0.45, 0.05)	(0.05, 0.45, 0.25, 0.25)	(0.4, 0.6, 0.45, 0.55)	1.5

* The proportions of samples from the four subpopulations.
** Allele freq from the four subpopulations $(p_{s,1}, ..., p_{s,4})$.

denoted by Sc1 to Sc4 (Table 3.3). Each scenario contains four subpupulations. The frequencies of allele A in the controls $(p_{s,k}, k = 1, 2, 3, 4)$ in the four subpopulations are given in Table 3.3. The frequencies of allele A in the cases $(p_{r,k}, k = 1, 2, 3, 4)$ are equal to those in the controls multiplied by $\mathrm{OR}/(1 - p_{s,k} + \mathrm{OR} \times p_{s,k})$ (Table 3.3). When OR=1, H_0 holds. The alternative hypothesis H_1 corresponds to OR>1. For each of the scenarios, 1000 cases and 1000 controls were simulated separately from multinomial distributions with proportions for the four subpopulations given in Table 3.3, and their genotypes were simulated independently under the Hardy-Weinberg equilibrium (i.e. the frequency of the genotype is equal to the product of the frequencies of the two alleles). Then we pooled the 1,000 cases and 1,000 controls in the total population with 10,000 replicates. We estimated the mean and variance of $\hat{\beta}$ from the simulations, referred to as simulated mean and variance. The theoretical mean β^* and variance V_n^* of $\hat{\beta}$ were also calculated using the true values in Table 3.3 (based on Theorem 3.1 and also assuming the PS is known). The true β equal to log(OR) which is constant across subgroups, and the V_n as computed in Section 3.2.1, are given for comparison. These values were calculated assuming the PS is known. The results are reported in Table 3.4.

The results show that the simulated means and variances of $\hat{\beta}$ are close to the theoretical β^* and V_n^* of $\hat{\beta}$. In the presence of PS, the usual estimate $\hat{\beta}$ ignoring

Table 3.4: Simulated and theoretical mean and variance of $\log(\widehat{\text{OR}})$ in the presence of PS. The true values when the PS is known are β and V_n.

	Mean				Variance			
Scenario	Simulated	Theoretical	β	Bias*	Simulated	Theoretical	V_n	VIF**
Sc1	-0.282	-0.280	0	0.282	0.00604	0.00602	0.00402	1.502
Sc2	0.884	0.885	0.095	0.789	0.00500	0.00517	0.00420	1.190
Sc3	0.281	0.281	0.182	0.099	0.00442	0.00443	0.00403	1.097
Sc4	0.157	0.156	0.405	0.249	0.00414	0.00415	0.00406	1.020

* Bias = |(simulated mean) $-\ \beta$|; ** VIF = (simulated variance)/V_n.

the PS is biased as $\beta^* \neq \beta$. The bias ranges from 0.282 under H_0 to 0.789 under H_1. Numerical results also show that the simulated variance is always greater than V_n, which are consistent with our analytical results. The simulated variance (actual variance in the presence of PS) of $\hat{\beta}$ over the true variance V_n is also reported as variance inflation factor (VIF). It has largest value 1.502 under the first scenario, and decreases as the differences in allele frequencies decreases. The results in Table 3.4 show that there is a significant impact on the estimate of log(OR) if the PS is not corrected in the analysis.

3.4.2 Correcting $\log(\text{OR})$ and its CI in the presence of hidden population stratification

We then conducted simulations to assess the performance of the PCA method to correct for hidden PS in estimating the OR and its CI. Four hidden subpopulations with 1,000 cases and 1,000 controls from the general population were assumed, each with 100 replicates. Each replicate includes two steps: 1) generate a panel of null SNPs, and apply the PCA method to them in order to infer the subpopulations, and 2) generate the candidate (target) SNPs, and use the identified subpopulations to correct for the OR estimates and CIs of each candidate SNP. The simulation procedure given below is standard [Price et al., 2006; Li and Yu, 2008].

Methods for generating null SNPs

In each replicate, 10,000 null SNPs were used to identify the hidden subpopulation structure. To generate each of the 10,000 null SNPs, we followed Price et al. [2006] and Li and Yu [2008]. We first generated the ancestry population allele frequency $p = \Pr(A)$ from the uniform distribution $U(0.1, 0.9)$. Then for each subpopulation, the allele frequency was generated independently from $Beta\Big(p(1-F_{ST})/F_{ST}, (1-p)(1-F_{ST})/F_{ST}\Big)$, where $F_{ST} = 0.01$ is Wright's inbreeding coefficient that measures the genetic distance between two populations (the PS vanishes if $F_{ST} = 0$ because all allele frequencies in the subpopulations are identical to p). Genotype frequencies of (G_0, G_1, G_2) were calculated as $((1-p)^2, 2p(1-p), p^2)$ under Hardy-Weinberg equilibrium (HWE). We considered two levels (moderate and extreme) of ancestral differences between cases and controls. When the ancestry difference of subpopulations is *moderate*, the proportions of cases from the four subpopulations are $(0.15, 0.2, 0.3, 0.35)$, and controls are $(0.35, 0.3, 0.2, 0.15)$. When the ancestry difference of subpopulations is *extreme*, the proportions of cases from the four subpopulations are $(0.25, 0.25, 0.45, 0.05)$, and controls are $(0.05, 0.45, 0.25, 0.25)$.

Methods for generating candidate SNPs

In each replicate, 1,000 candidate SNPs were simulated, in which the allele frequencies are differentially distributed in subpopulations. We also considered two levels (moderate and extreme) of differences between allele frequencies of the four subpopulations. For *moderately* differential allele frequencies, we follow the same approach to generate the allele frequencies for each SNP as we did for the null SNPs. For *extremely* differential allele frequencies, we randomly drew four allele frequencies from $U(0.1, 0.9)$ for the four subpopulations. The risk model for the allele A of a candidate SNP was implemented as follows: for an individual from population k with population allele frequency p_k, controls were assigned genotypes G_0, G_1, and G_2 with

Table 3.5: Simulated coverage probabilities of the CIs for OR, based on different combinations of subgroup proportion differences and allele frequency differences, before and after correcting the PS.

Subgroup Proportion Differences	Allele Frequency Differences	Before correcting PS				After correcting PS			
		linear	allele	hetero	homo	linear	allele	hetero	homo
Moderate	Moderate	0.839	0.836	0.917	0.876	0.950	0.949	0.952	0.952
More extreme	Moderate	0.784	0.780	0.889	0.835	0.950	0.950	0.952	0.952
Moderate	More extreme	0.312	0.298	0.533	0.328	0.950	0.950	0.951	0.952
More extreme	More extreme	0.255	0.241	0.433	0.261	0.949	0.949	0.950	0.950

probabilities $(1-p_k)^2$, $2p_k(1-p_k)$, and p_k^2, and cases were assigned with probabilities $(1-p_k)^2$, $2\tilde{R}p_k(1-p_l)$, and $\tilde{R}^2p_k^2$ (scaling those probabilities so that their sum is 1), where $\tilde{R}$ was randomly drawn from $U(1,2)$ under H_1. Notice that when $\tilde{R}=1$, the model is under H_0.

Simulation results

In Table 3.5, the percentages of 1000×100 candidate SNPs from all the replicates with 95% CIs covering the true OR are presented. It shows that the proposed correction method has a good coverage probability for all four kinds of ORs, with the coverage very close to the nominal level 95%. Without controlling the hidden PS, however, the coverages for all the four ORs are lower than 95%. The coverage of CI further goes down as the subgroup proportion differences and the allele frequency differences increase.

In addition, we wanted to evaluate how PCA performs in terms of the accuracy of clustering. We only considered the model when the subgroup proportion difference and the allele frequency difference are both moderate. Among the 100 replicates, there were only 15 replicates in which PCA did not cluster all subjects correctly into the true hidden subgroup. We further looked into the coverage of 95% CI for ORs from those 15 replicates (Table 3.6). It is shown that even when PCA wrongly cluster some subjects, our proposed method still provides an appropriate correction

Table 3.6: Minimum and average coverage probabilities of the CIs for OR from only the 15 replicates where PCA wrongly cluster some subjects.

	linear	allele	hetero	homo
Minimum Coverage	0.940	0.931	0.941	0.941
Average Coverage	0.950	0.949	0.952	0.954

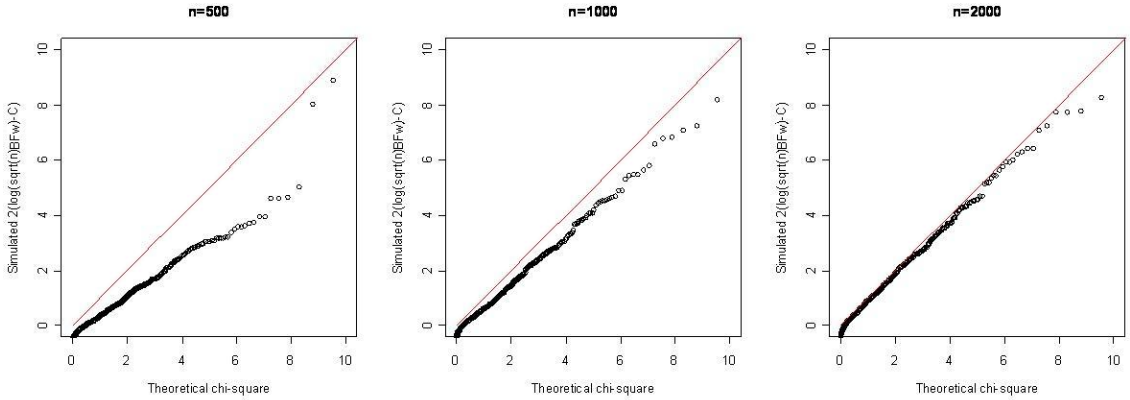
for hidden PS. Intuitively, this is because if the difference among hidden subgroups is subtle, such stratification would not much affect the OR estimate or its CI even when PCA over-clustered or under-clustered the subjects. On the other hand, if the difference among hidden subgroups is non-subtle, PCA method would be able to cluster subjects into their true subgroup, given sufficient number of null SNPs.

3.4.3 Bayesian analysis and population stratification

The BF_W depends on the prior distribution of $\beta = \log(\mathrm{OR}) \sim N(0, \sigma^2)$. The discussion of choosing σ^2 can be found in Wakefield [2007]. Often it depends on the results of prior association and linkage studies and the genetic effect (OR is most likely in the range of 1.1 to 1.5 for genome-wide association studies). In this paper, we chose $\sigma^2 = 0.06$, which was obtained assuming that there is roughly 5% one-sided probability to detect a SNP with OR greater than 1.50.

We studied the asymptotic behavior of BF_W when there is no PS. We conducted simulations when the number of cases and controls was $r = s$ =250, 500, or 1,000. For each sample size, we simulated 1,000 replicates under H_0. The allele frequencies in each replicate were randomly generated from $U(0.1, 0.9)$. Denote $C = \log\sqrt{I(\beta)^{-1}/\sigma^2}$ where $I(\beta)^{-1}$ was estimated by $n\hat{V}_n$ and $\hat{V}_n$ was estimated from logistic regression. The Q-Q plots of the simulated $2(\log(\sqrt{n}\mathrm{BF}_W) - C)$ versus the standard χ^2_1 are presented in Figure 3.1. In the absence of PS, it is expected that $2(\log(\sqrt{n}\mathrm{BF}_W) - C)$ performs asymptotically like χ^2_1, which is mathematically

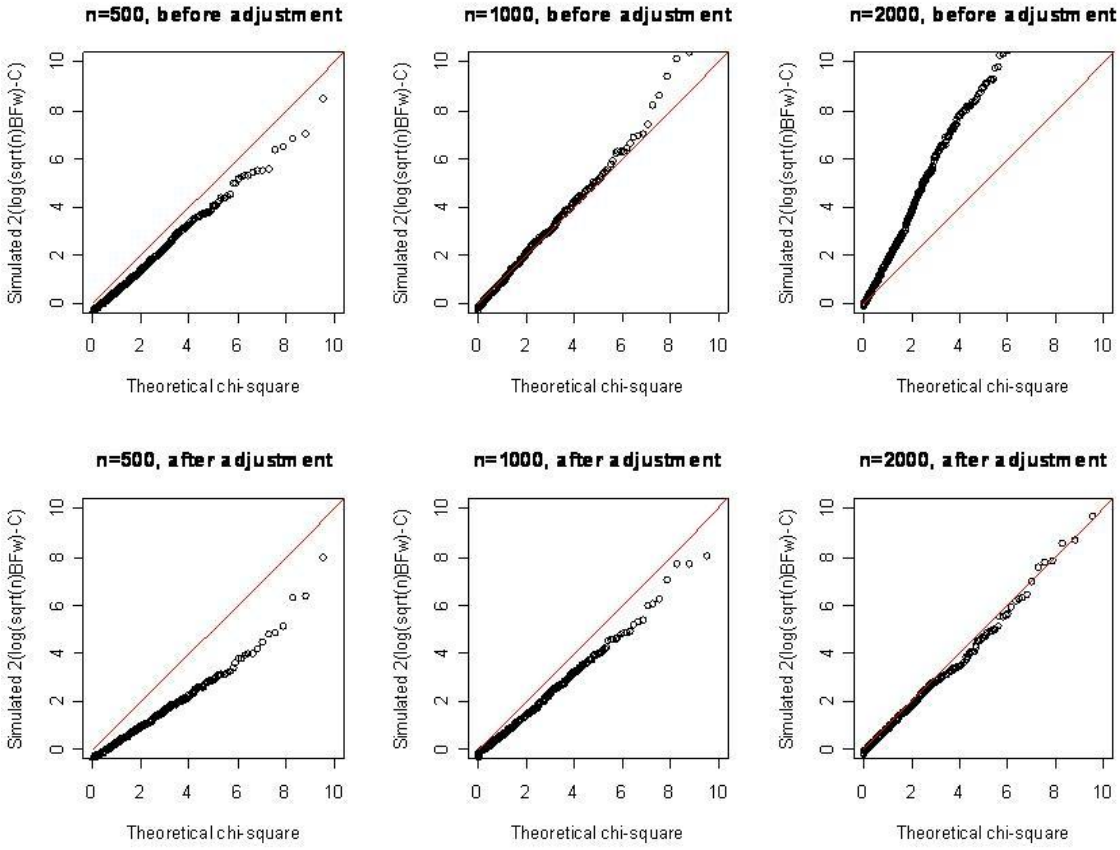
Figure 3.1: Q-Q plots of $2(\log(\sqrt{n}\mathrm{BF}_W) - C)$ without the presence of PS versus the standard χ^2_1 distribution. The diagonal line $x = y$ is indicated.



equivalent to what is implied by Theorem 3.2. The results in Figures 3.1 confirm this.

To study the asymptotic behavior of the BF_W in the presence of hidden PS, we applied the same methods to generate 10,000 null SNPs and 1,000 candidate SNPs as we did in the previous simulation, except that the candidate SNPs only have $\tilde{R} = 1$. We only considered the model when the subgroup proportion difference and the allele frequency difference are both moderate. The number of cases and controls was $r = s =$ 250, 500, or 1,000. Denote $C = \log\sqrt{I(\beta)^{-1}/W}$ where $I(\beta)^{-1}$ was estimated by $n\hat{V}_n$ and $\hat{V}_n$ was estimated from logistic regression. The Q-Q plots of the simulated $2(\log(\sqrt{n}\mathrm{BF}_W) - C)$ before and after corrections for hidden PS versus the standard χ^2_1 are presented in Figure 3.2. It is shown that even when the stratification is moderate, hidden PS still causes distortion to the asymptotic distribution of the BF_W. The proposed method based on the PCA can provide a proper correction for the effect of PS on the BF_W when the sample size is large enough.

Figure 3.2: Q-Q plots of $2(\log(\sqrt{n}\mathrm{BF}_W) - C)$ in the presence of PS (before and after correction) versus the standard chi-squared distribution with 1 df. The diagonal line $x = y$ is indicated.



3.5 Applications to HapMap Data

For illustration, we applied the PCA method and the correction for PS using a subset of data from the HapMap Project, which contains genotype data of 7,415 SNPs in 303 subjects from 3 subgroups: African ancestry in Southwest USA (ASW), Utah residents with Northern and Western European ancestry from the CEPH collection (CEU), and Han Chinese in Beijing, China (CHB). These SNPs are known to be independent and not associated with diseases [Yu et al., 2008].

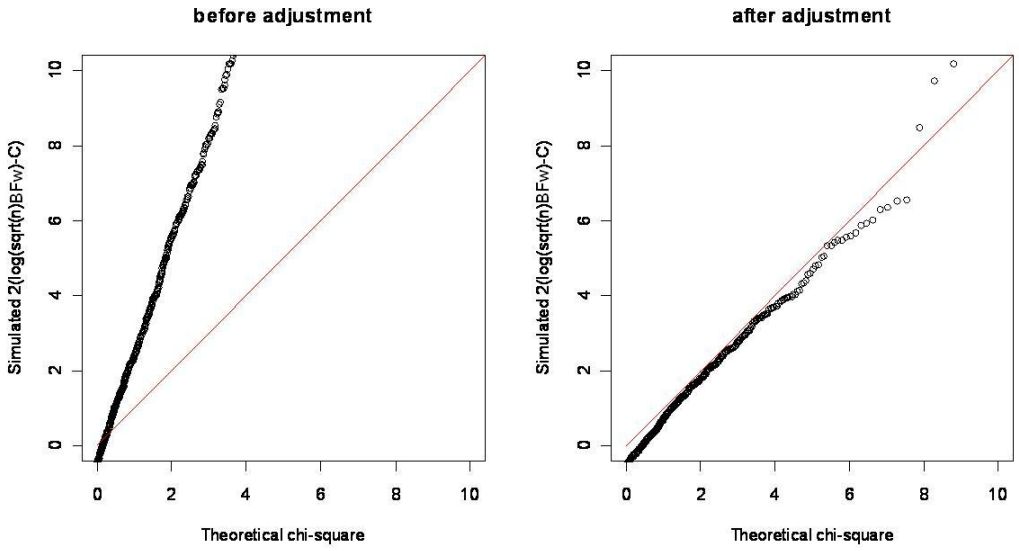
We applied the following simulation procedure with 10 replicates. In each replicate, we randomly generated three different disease rates (f_1, f_2, f_3) for the three subgroups from $U(0.1, 0.9)$. Then the number of cases in each subgroup was generated from $Binomial(n_i, f_i)$ where n_i is the size of subgroup for $i = 1, 2, 3$. The random generation of case-control status is justified by the fact that all the SNPs are known not to be associated with the disease. Since all the SNPs are null SNPs, we randomly selected 100 SNPs as candidate SNPs and the rest 7,315 SNPs as null SNPs for applying the PCA clustering method. The simulation results showed that using Tibshirani's Gap-statistic, the subjects were correctly clustered into $\tilde{K} = 3$ groups for all 10 replicates. After the clustering, we computed the CIs for the four different types of ORs using the proposed method. The overall percentages of 95% CIs covering the true OR = 1 are presented in Table 3.7 along with those without corrections for PS. Like the results that we presented before, all CIs without correcting for PS have lower than 95% coverage, while the CIs with the corrections for PS have coverage probabilities close to 95%.

For Bayesian analysis, the Q-Q plots of $2(\log(\sqrt{n}\mathrm{BF}_W) - C)$ before and after corrections for PS versus the standard χ^2_1 are presented in Figure 3.3. In this illustration, we confirmed that BF_W is affected by the PS and that the proposed clustering

Table 3.7: Coverage probabilities of the CIs for OR, based on the four methods, before and after correcting the PS using the HapMap data.

Correction for PS	linear	allele	hetero	homo
Before correction	0.761	0.711	0.784	0.879
After correction	0.955	0.951	0.965	0.964

Figure 3.3: Q-Q plots of $2(\log(\sqrt{n}\mathrm{BF}_W) - C)$ before and after correction for PS in HapMap data, versus the standard chi-squared distribution with 1 df. The diagonal line $x = y$ is indicated.



method can correct for the impact of PS.

3.6 Discussion

It is known that ignoring hidden population stratification (PS) has serious impact on the analysis of case-control genetic association studies [Devlin and Roeder, 1999; Astle and Balding, 2009; Zheng et al., 2010]. In this chapter, we examined this impact in Bayesian analysis and provided a simple approach to correct it in the presence of hidden PS. To achieve this, we used an approximate BF [Johnson, 2005; Wakefield, 2007], which is expressed in terms of the odds ratio and its asymptotic variance

[Wakefield, 2007], so that the problem is transformed from Bayesian hypothesis testing to the estimates of odds ratio and its asymptotic variance.

We first derived the asymptotic distribution of the odds ratio (OR) when hidden PS exists but is unmodeled, and showed the impact of hidden PS on the OR estimate and its confidence interval (CI). The bias in the OR estimate and its variance distortion calls for an appropriate correction for hidden PS. We then adopted an existing PCA procedure and the EIGENSTRAT algorithm that were developed for testing hypothesis to address the estimation problem. The PCA approach [Price et al., 2006; Li and Yu, 2008] uses a large number of null markers to infer hidden population substructure, so that each individual can be assigned to one identified subpopulation and an appropriate stratified analysis can then be carried out. We presented the asymptotic behavior of BF without and with the presence of hidden PS under H_0, and then proposed to use the previously obtained adjusted OR and its variance estimate to correct for the BF.

Simulation studies confirmed our theorem about the impact of hidden PS on the OR estimate and its asymptotic variance. It was also shown in the simulations that, applying the PCA method to the estimation of four types of odds ratio leads to appropriate correction and good coverage probability of the CI in the presence of hidden PS, while ignoring hidden PS leads to a much smaller converge probability of the CI. Further investigation in simulations showed that PCA method can, in most cases, cluster the subjects into the correct subgroups; even when it does not, our proposed method still provides an appropriate coverage for the CI of OR. The comparison of the asymptotic behavior of BF before and after PCA was also studied in simulations and was illustrated by Q-Q plots. It is shown that the hidden PS will deviate the asymptotic distribution of the BF, and the proposed correction in

estimating odds ratio and its asymptotic variance leads to appropriate correction for the asymptotic behavior of the BF.

In addition to the simulation studies, we also use the HapMap data for illustration of the proposed approach. By using more than 7,000 null SNPs to identify the subpopulation structure, the PCA method can always correctly assign the subjects into one of the three true subgroups. From there we conducted stratified analysis to obtain the appropriate estimates for different types of OR and its variance, and then further correct for the BF. It is illustrated in the Q-Q plot that the distribution of the unadjusted BF deviates from what it is supposed to be under H_0, while the distribution of the adjusted BF agrees with our theorem about the asymptotic behavior of the BF.

Chapter 4

Bayes Factor based on a Maximum Statistic for Case-control Genetic Association Studies

4.1 Introduction

In the analysis of case-control genetic association studies, the data are often summarized in a 2×3 contingency table, which is commonly analyzed by the Cochran-Armitage trend test (CATT) and Pearson's chi-squared test. The former is a one-degree-of-freedom (1-df) test and asymptotically most powerful under the additive model, while the latter is a 2-df test and more powerful than the CATT under the recessive model, or when a large deviation from Hardy-Weinberg equilibrium (HWE) in cases is present. Under the recessive and dominant models, other optimal 1-df trend tests are used, respectively. In the analysis of genome-wide association studies (GWAS), it has been demonstrated that using the CATT alone is not robust, i.e., it may fail to detect markers (SNPs) in linkage disequilibrium (LD) with functional loci under non-additive models, although most SNPs associated with complex diseases have shown an additive effect due to LD [Zheng et al., 2009a]. For example, the

Table 4.1: P-values of the 14 SNPs reported in the WTCCC (2007) with at least moderate association with BD. The trend test p-values in bold are the smallest among the three trend tests for each SNP. The Pearson's p-values in bold are those smaller than the smallest trend tests p-values. The SNP in bold is the only one reported having strong association with BD. The p-values of MAX were calculated using Zang et al. [2010]

	Trend tests				
SNP ID	REC	ADD	DOM	Pearson's	MAX
rs3761218	2.14E-01	4.43E-05	**1.13E-06**	6.71E-06	3.12E-06
rs7570682	9.05E-04	**3.11E-06**	3.37E-05	1.64E-05	8.10E-06
rs10982256	1.21E-04	**8.80E-06**	5.84E-04	4.41E-05	2.39E-05
rs2953145	2.02E-05	**1.12E-05**	6.14E-04	**6.57E-06**	2.86E-05
rs683395	5.44E-01	2.30E-06	**8.55E-07**	5.11E-06	2.07E-06
rs4027132	**3.08E-06**	1.31E-05	9.20E-03	9.69E-06	8.48E-06
rs420259	9.66E-02	2.19E-04	**8.48E-09**	6.29E-08	2.36E-08
rs4276227	2.04E-03	**4.57E-06**	2.64E-05	2.62E-05	1.24E-05
rs10134944	4.79E-01	3.21E-06	**1.15E-06**	6.89E-06	2.81E-06
rs1375144	5.82E-05	**2.43E-06**	2.68E-04	1.31E-05	6.59E-06
rs1344484	2.61E-05	**1.65E-06**	3.96E-04	1.03E-05	4.53E-06
rs6458307	2.11E-01	3.43E-01	**2.81E-05**	**4.35E-06**	7.37E-05
rs2609653	1.41E-02	**6.86E-06**	2.10E-05	2.31E-05	1.54E-05
rs11622475	1.95E-02	**2.10E-06**	2.20E-06	8.14E-06	5.64E-06

WTCCC [2007] reported 14 SNPs with at least moderate association with bipolar disorder (BD). The SNP (rs420259) with strongest association demonstrated a dominant effect. The p-value for the CATT of that SNP is only 2.19E-4, while the p-values for Pearson's test and the 1-df trend test under the dominant model are 6.29E-8 and 8.47E-9, respectively (Table 4.1). On the other hand, for these 14 SNPs reported by the WTCCC [2007], the CATT has smaller p-values than Pearson's test on 9 SNPs, which mostly demonstrated an additive effect, under which Pearson's test is usually less powerful. Eight SNPs have the smallest p-values under the additive model (Table 4.1).

Based on the above discussion, using either the CATT or Pearson's test alone is not a good strategy in the analysis of GWAS. The WTCCC [2007] considered reporting

the minimum of the p-values of the CATT and Pearson's test. Alternatively, the maximum of three 1-df trend tests, most powerful respectively under the recessive, additive and dominant genetic models, can be used. This maximum test, denoted as MAX or MAX3, was used by Sladek et al. [2007] to search SNPs associated with type 2 diabetes in GWAS. Freidlin et al. [2002] showed that MAX is more robust than Pearson's test and the CATT across various genetic models. Yamada and Okada [2009] and Zheng et al. [2009b] further showed that the statistical power of MAX dominates that of Pearson's test among the three common genetic models. MAX, however, does not follow a chi-squared distribution asymptotically under the null hypothesis of no association. Calculations and approximations of the p-value of MAX have been studied [Li et al., 2008; Zang et al., 2010]. Of the 14 SNPs in Table 1, MAX has smaller p-values than Pearson's test for 12 SNPs. MAX has similar performance to the CATT under the additive or dominant models, but MAX is more robust than the CATT, which can be seen from their p-values of SNPs rs420259 and rs6458307 in Table 4.1.

It is known that a p-value alone cannot be used as a measure of evidence of association in GWAS, in which 500,000 to more than a million SNPs are tested. Small p-values are often observed in GWAS, even though the power to detect these associations is not high. On the other hand, Bayes factor (BF), a ratio of the probability of the observed data under the alternative hypothesis H_1 over that under the null hypothesis H_0, provides evidence in favor of H_0 or H_1 and integrates both p-value and power to detect an association. Specifically, Sawcer [2010] illustrated that the BF mimics the ratio of the power over the significance of association. In practice, while the p-value is still the main tool used to evaluate the significance of SNPs in GWAS, the BF has been increasingly reported to support the small p-values [WTCCC, 2007;

Stephens and Balding, 2009]. When MAX and its p-value are used to evaluate association in GWAS [e.g. Sladek et al., 2007], what BF should be reported to accompany MAX? Zheng et al. [2011] considered this problem and provided a simple approach by taking the maximum of three BFs each calculated under one of the three common genetic models. In this way, the maximum of the three BFs and the maximum of the three trend tests are *comparable* and can be reported together. Their maximum BF has high concordance with the p-values of MAX on SNPs showing at least a moderate association in GWAS. However, it cannot be interpreted like a usual BF as a ratio of two probabilities.

In this chapter, we propose a new measure of the strength of association using MAX. It is an extension of the BFs of Johnson [2005, 2008] who considered BFs modeling the distributions of a test statistic under H_1 and H_0. Wakefield [2007, 2009] applied a similar idea of Johnson [2005] and derived a BF modeling the distributions of odds ratio of genetic effect under H_1 and H_0. We consider modeling the distributions of MAX under H_1 and H_0. Thus, we first need to derive an asymptotic distribution of MAX under H_1 and then, for computational purpose, we obtain a simple approximation of this asymptotic distribution to use in the BF. Detailed computations are presented. Simulation studies are used to evaluate the performance of our proposed measure with existing ones, including Bayesian model averaging of Stephens and Balding [2009] and the maximum BFs of Zheng et al. [2011]. Applications to real GWAS illustrate the use of our method and show that p-values cannot be used to compare results across GWAS.

4.2 Distribution of MAX

4.2.1 Mean and covariance of trend tests

Denote the alleles of a single-nucleotide polymorphism (SNP) as A and B with the population frequencies $\Pr(A) = p$ and $\Pr(B) = 1 - p$, where A is the minor allele (the less common one). Three genotypes are denoted as $(G_0, G_1, G_2) = (BB, AB, AA)$ with the population frequency $g_i = \Pr(G_i)$. Among r cases (s controls), the genotype count for G_i is r_i (s_i). Denote $n_i = r_i + s_i$ $(i = 0, 1, 2)$ and $n = r + s$. Let $\kappa = \Pr(\text{case})$ be the disease prevalence in the population. Then $(r_0, r_1, r_2)^T$ and $(s_0, s_1, s_2)^T$ follow multinomial distributions $\text{Mul}(r; \mathbf{p})$ and $\text{Mul}(s; \mathbf{q})$, respectively, where $\mathbf{p} = (p_0, p_1, p_2)^T$ and $\mathbf{q} = (q_0, q_1, q_2)^T$, $p_i = \Pr(G_i|\text{case}) = g_i\eta_i/\kappa$, and $q_i = \Pr(G_i|\text{control}) = g_i(1 - \eta_i)/(1 - \kappa)$, where $\eta_i = \Pr(\text{case}|G_i)$ is the disease penetrance for genotype G_i. Denote genotype relative risk (GRR) as $\lambda_j = \eta_j/\eta_0$ $(j = 1, 2)$. Under H_0, $\lambda_1 = \lambda_2 = 1$ and equivalently $\mathbf{p} = \mathbf{q}$; under H_1, $\lambda_2 \geq \lambda_1 \geq 1$ and $\lambda_2 > 1$ if A is the risk allele. Recall that a genetic model is recessive (REC) when $\lambda_1 = 1$ and $\lambda_2 > 1$, additive (ADD) when $\lambda_1 = (1 + \lambda_2)/2$ and $\lambda_2 > 1$, and dominant (DOM) when $\lambda_1 = \lambda_2$ and $\lambda_2 > 1$. Three genetic models are unified as $\lambda_1 = (1 - x) + x\lambda_2$ with $x = 0$, $1/2$, 1 for the REC, ADD, DOM models, respectively. Although the case-control data are obtained retrospectively, the analysis can be done as if they were obtained prospectively [Prentice and Pyke, 1979; Qin and Zhang, 1997; Seaman and Richardson, 2004].

Denote $\widehat{p}_i = r_i/r$, $\widehat{q}_i = s_i/s$ and $\widehat{h}_i = n_i/n$ $(i = 0, 1, 2)$. The trend tests optimal for different genetic models can be expressed in a single statistic $Z(x)$ indexed by a

parameter x as follows [Freidlin et al., 2002]:

$$Z(x) = \frac{\mathbf{x}^T(\widehat{\mathbf{p}} - \widehat{\mathbf{q}})}{\sqrt{(n/rs)\left\{(\mathbf{x^2})^T\,\widehat{\mathbf{h}} - \left(\mathbf{x}^T\widehat{\mathbf{h}}\right)^2\right\}}}, \tag{4.1}$$

where $\mathbf{x} = (0, x, 1)^T$ and $x = 0, 1/2, 1$ for the REC, ADD, DOM models, respectively. Also $\mathbf{x^2} = (0, x^2, 1)^T$, $\widehat{\mathbf{p}} = (\widehat{p}_0, \widehat{p}_1, \widehat{p}_2)^T$, $\widehat{\mathbf{q}} = (\widehat{q}_0, \widehat{q}_1, \widehat{q}_2)^T$, and $\widehat{\mathbf{h}} = (\widehat{h}_0, \widehat{h}_1, \widehat{h}_2)^T$. Under H_0, given x, $Z(x)$ asymptotically follows $N(0, 1)$.

We first derived the general expressions of the mean and covariance of trend tests regardless of H_0 or H_1, summarized in Theorem 4.1.

Theorem 4.1. Assume $b = r/n \in (0, 1)$ is a constant when $n \to \infty$. Denote $\mathbf{h} = b\mathbf{p} + (1-b)\mathbf{q}$. Let μ_x be the expected value of $Z(x)$ and Cov_{x_1,x_2} be the covariance of $Z(x_1)$ and $Z(x_2)$. Then, when n is large enough,

$$\begin{aligned} \mu_x &= \frac{\sqrt{nb(1-b)}\mathbf{x}^T(\mathbf{p} - \mathbf{q})}{\sqrt{(\mathbf{x}^2)^T\,\mathbf{h} - (\mathbf{x}^T\mathbf{h})^2}}, \\ \mathrm{Cov}_{x_1,x_2} &= \left(\frac{\partial Z_{x_1}}{\partial \mathbf{p}}\right)^T \frac{\Delta_p}{r}\left(\frac{\partial Z_{x_2}}{\partial \mathbf{p}}\right) + \left(\frac{\partial Z_{x_1}}{\partial \mathbf{q}}\right)^T \frac{\Delta_q}{s}\left(\frac{\partial Z_{x_2}}{\partial \mathbf{q}}\right), \end{aligned}$$

where Δ_p is a 3×3 matrix whose (i,i)th and (i,j) th elements are $p_i(1-p_i)$ and $-p_ip_j$ $(i \neq j)$ respectively and Δ_q is defined similarly, $\partial Z_{x_i}/\partial \mathbf{p} = \sqrt{nb(1-b)}\,\{C_i - bD_i\}$, and $\partial Z_{x_i}/\partial \mathbf{q} = -\sqrt{nb(1-b)}\,\{C_i + (1-b)D_i\}$, where

$$C_i = \frac{\mathbf{x_i}}{\sqrt{(\mathbf{x_i}^2)^T\,\mathbf{h} - (\mathbf{x_i}^T\mathbf{h})^2}} \;\; \textit{and} \;\; D_i = \frac{\mathbf{x_i}^T(\mathbf{p} - \mathbf{q})\left(\mathbf{x_i}^2 - 2\mathbf{x_i}^T\mathbf{h}\mathbf{x_i}\right)}{2\left\{(\mathbf{x_i}^2)^T\,\mathbf{h} - (\mathbf{x_i}^T\mathbf{h})^2\right\}^{3/2}}.$$

A proof of Theorem 4.1 can be obtained using the Delta method as follows:

Proof: $\widehat{\mathbf{p}} = (\widehat{p}_0, \widehat{p}_1, \widehat{p}_2)^T$ and $\widehat{\mathbf{q}} = (\widehat{q}_0, \widehat{q}_1, \widehat{q}_2)^T$ are sample statistics with the mean vectors $\mathbf{p} = (p_0, p_1, p_2)$ and $\mathbf{q} = (q_0, q_1, q_2)$ and the covariance matrices Δ_p/r and Δ_q/s, respectively. Denote $\mathbf{Z} = [Z_0(\widehat{\mathbf{p}}, \widehat{\mathbf{q}}), Z_{1/2}(\widehat{\mathbf{p}}, \widehat{\mathbf{q}}), Z_1(\widehat{\mathbf{p}}, \widehat{\mathbf{q}})]^T$. By the Delta method,

$$\mathrm{E}(\mathbf{Z}) = [Z_0(\mathbf{p}, \mathbf{q}), Z_{1/2}(\mathbf{p}, \mathbf{q}), Z_1(\mathbf{p}, \mathbf{q})]^T, \mathrm{Var}(\mathbf{Z}) = \mathbf{Y}(\mathbf{p}, \mathbf{q})^T\mathbf{\Sigma_{p,q}}\mathbf{Y}(\mathbf{p}, \mathbf{q})$$

where $\mathbf{\Sigma_{p,q}} = \begin{pmatrix} \Delta_p/r & 0 \\ 0 & \Delta_q/s \end{pmatrix}$, and

$$\mathbf{Y}(\mathbf{p},\mathbf{q}) = \begin{pmatrix} \partial Z_0(\widehat{\mathbf{p}},\widehat{\mathbf{q}})/\partial\widehat{\mathbf{p}} & \partial Z_0(\widehat{\mathbf{p}},\widehat{\mathbf{q}})/\partial\widehat{\mathbf{q}} \\ \partial Z_{1/2}(\widehat{\mathbf{p}},\widehat{\mathbf{q}})/\partial\widehat{\mathbf{p}} & \partial Z_{1/2}(\widehat{\mathbf{p}},\widehat{\mathbf{q}})/\partial\widehat{\mathbf{q}} \\ \partial Z_1(\widehat{\mathbf{p}},\widehat{\mathbf{q}})/\partial\widehat{\mathbf{p}} & \partial Z_1(\widehat{\mathbf{p}},\widehat{\mathbf{q}})/\partial\widehat{\mathbf{q}} \end{pmatrix}^T \Big|_{\widehat{\mathbf{p}}=\mathbf{p},\,\widehat{\mathbf{q}}=\mathbf{q}}.$$

Denote

$$\frac{\partial Z_x(\widehat{\mathbf{p}},\widehat{\mathbf{q}})}{\partial\widehat{\mathbf{p}}}\Big|_{\widehat{\mathbf{p}}=\mathbf{p},\,\widehat{\mathbf{q}}=\mathbf{q}} = \frac{\partial Z_x}{\partial\mathbf{p}},\ \frac{\partial Z_x(\widehat{\mathbf{p}},\widehat{\mathbf{q}})}{\partial\widehat{\mathbf{q}}}\Big|_{\widehat{\mathbf{p}}=\mathbf{p},\,\widehat{\mathbf{q}}=\mathbf{q}} = \frac{\partial Z_x}{\partial\mathbf{q}}.$$

Then

$$\mu_x = Z_x(\mathbf{p},\mathbf{q}) = \frac{\sqrt{nb(1-b)}\mathbf{x}^T(\mathbf{p}-\mathbf{q})}{\sqrt{(\mathbf{x}^2)^T\mathbf{h} - (\mathbf{x}^T\mathbf{h})^2}},$$

and the (x_1, x_2) entry of the covariance matrix $Var(\mathbf{Z})$ is

$$\mathrm{Cov}_{x_1,x_2} = \left(\frac{\partial Z_{x_1}}{\partial\mathbf{p}}\right)^T \frac{\Delta_p}{r}\left(\frac{\partial Z_{x_2}}{\partial\mathbf{p}}\right) + \left(\frac{\partial Z_{x_1}}{\partial\mathbf{q}}\right)^T \frac{\Delta_q}{s}\left(\frac{\partial Z_{x_2}}{\partial\mathbf{q}}\right).$$

Denote $U = \mathbf{x}^T(\widehat{\mathbf{p}} - \widehat{\mathbf{q}})$, $T = \sqrt{(\mathbf{x^2})^T\widehat{\mathbf{h}} - \left(\mathbf{x}^T\widehat{\mathbf{h}}\right)^2}$. Then $Z_x = \sqrt{nb(1-b)}U/T$. Simple algebra shows that

$$C_i = \frac{\partial U/\partial\mathbf{p}}{T},\quad -C_i = \frac{\partial U/\partial\mathbf{q}}{T},\quad bD_i = \frac{U\partial T/\partial\mathbf{p}}{T^2},\quad (1-b)D_i = \frac{U\partial T/\partial\mathbf{q}}{T^2},$$

then we can readily get

$$\partial Z_{x_i}/\partial\mathbf{p} = \sqrt{nb(1-b)}\left\{C_i - bD_i\right\},\quad \partial Z_{x_i}/\partial\mathbf{q} = -\sqrt{nb(1-b)}\left\{C_i + (1-b)D_i\right\}.[]$$

.

4.2.2 Asymptotic Distributions of MAX under null and alternative hypotheses

MAX is given by

$$\mathrm{MAX} = \max\{|Z(0)|, |Z(1/2)|, |Z(1)|\}.$$

Zang et al. [2010] derived the cumulative distribution function (c.d.f.) of MAX under H_0, summarized here in Lemma 4.1.

Lemma 4.1. The c.d.f. of MAX, denoted as $F_0(t)$, under $H_0 : \mathbf{p} = \mathbf{q}$, can be written as $F_0(t) = 2\int\int_{\Omega}\phi(z_0, z_1, 0, \Sigma_0)dz_0dz_1$, where $\phi(z_0, z_1, \mu, \Sigma)$ is the density of a bivariate normal with mean vector μ and covariance matrix Σ, $\Omega = \Omega_1 \cup \Omega_2$ where $\Omega_1 = \{(z_0, z_1) : z_0 \in (0, t(1-\omega_1)/\omega_0), z_1 \in (-t, t)\}$ and $\Omega_2 = \{(z_0, z_1) : z_0 \in (t(1-\omega_1)/\omega_0, t), z_1 \in (-t, (t - z_0\omega_0)/\omega_1)\}$, $\omega_i = (\rho_{i,1/2} - \rho_{0,1}\rho_{1/2,(1-i)})/(1-\rho_{0,1}^2)$ $(i = 0, 1)$, and $\Sigma_0 = \begin{pmatrix} 1 & \rho_{0,1} \\ \rho_{0,1} & 1 \end{pmatrix}$, where

$$\rho_{x_1,x_2} = \frac{\mathbf{x_1}^T\Delta_p\mathbf{x_2}}{\sqrt{(\mathbf{x_1^2})^T\mathbf{p} - (\mathbf{x_1}^T\mathbf{p})^2}\sqrt{(\mathbf{x_2^2})^T\mathbf{p} - (\mathbf{x_2}^T\mathbf{p})^2}}, \tag{4.2}$$

where $\mathbf{x_i} = (0, x_i, 1)^T$ and Δ_p is a 3×3 matrix whose (i,i)th and (i,j) th elements are $p_i(1-p_i)$ and $-p_ip_j$ $(i \neq j)$, respectively.

It can be shown that, under $H_0 : \mathbf{p} = \mathbf{q}$, $\rho_{x_1,x_2} = \mathrm{Cov}_{x_1,x_2}/\left(\mathrm{Cov}_{x_1,x_1}\mathrm{Cov}_{x_2,x_2}\right)^{1/2}$ obtained from Theorem 4.1 reduces to ρ_{x_1,x_2} given in (4.2). Note that $F_0(t)$ depends on $\mathbf{p}$. Under HWE, $\mathbf{p} = \left((1-p)^2, 2p(1-p), p^2\right)^T$ and $F_0(t)$ depends on the allele frequency p. In Lemma 4.1, the three-fold integration over three trend tests in $F_0(t)$ is reduced to the sum of two double integrations because $Z(1/2) = \omega_0 Z(0) + \omega_1 Z(1)$ is a linear combination of $Z(0)$ and $Z(1)$ under H_0 (Zang et al., 2010).

From Theorem 4.1, when $\Sigma_1 = (\mathrm{Cov}_{i,j})_{3\times 3}$ is positive definite for any p and n is large enough, $(Z(0), Z(1/2), Z(1))^T \sim N(\mu, \Sigma_1)$, where $\mu = (\mu_0, \mu_{1/2}, \mu_1)^T$. When D_i in Theorem 4.1 is zero, $Z(1/2)$ is a linear combination of $Z(0)$ and $Z(1)$, a property holds under H_0 and is used in Lemma 4.1. Under H_1, however, $D_i \neq 0$ and $(Z(0), Z(1/2), Z(1))^T$ are linearly independent unless D_i is dropped. Denote Cov_{x_1,x_2} as $\widetilde{\mathrm{Cov}}_{x_1,x_2}$ when D_i is dropped. Correspondingly, denote Σ_1 as $\widetilde{\Sigma}_1$. The numerical difference between Σ_1 and $\widetilde{\Sigma}_1$ is relatively small. For instance, when the minor allele frequency (MAF) in the population is 0.3, the disease prevalence is $\kappa = 0.01$, $\lambda_2 = 1.5$ and the true model is additive, we have

$$\Sigma_1 = \begin{pmatrix} 0.98746 & 0.68762 & 0.30857 \\ 0.68762 & 0.99092 & 0.89336 \\ 0.30857 & 0.89336 & 0.99572 \end{pmatrix} \text{ and } \widetilde{\Sigma}_1 = \begin{pmatrix} 0.99794 & 0.69417 & 0.30973 \\ 0.69417 & 0.99512 & 0.89431 \\ 0.30973 & 0.89431 & 0.99580 \end{pmatrix}.$$

Table 4.2: The determinant of Σ_1, $|\Sigma_1|$, given $r = s = 1000$, disease prevalence $\kappa = 0.01$, MAF and GRR $\lambda_2 = 1.5$ for a given genetic model. HWE holds in the population.

Model	MAF=0.15	MAF=0.30	MAF=0.45
REC	0.00024	0.00014	0.00028
ADD	0.00018	0.00017	0.00014
DOM	0.00135	0.00105	0.00049

Under H_0, the determinant of Σ_1 is $|\Sigma_1| = 0$. Under H_1, $|\Sigma_1| \approx 0$ even the term D_i is retained. Some numerical results of $|\Sigma_1|$ under H_1 are reported in Table 4.2 with different MAFs and genetic models. The largest observed $|\Sigma_1|$ is 0.00135 under the DOM model with MAF = 0.15. Other numerical results from simulations will be reported in Section 4.4.2.

When Σ_1 in the previous section is used and $|\Sigma_1| > 0$, the c.d.f. of MAX, denoted as $F_1(t)$, is given by

$$\begin{aligned} F_1(t) &= \Pr(\text{MAX} \le t) = \Pr(|Z(0)| \le t, |Z(1/2)| \le t, |Z(1)| \le t) \\ &= \int_{(-t,t)^3} \phi(\mathbf{z}, \mu, \Sigma_1) d\mathbf{z}. \end{aligned} \tag{4.3}$$

The probability density function (p.d.f.) of MAX can be obtained by differentiating $F_1(t)$ with respect to t. However, to simplify the computation, we consider using $\widetilde{\Sigma}_1$, satisfying $|\widetilde{\Sigma}_1| = 0$. Hence the same linear combination of $(Z(0), Z(1/2), Z(1))^T$ in Lemma 4.1 can be used. Thus, we obtain an approximate c.d.f. of MAX under H_1, denoted as $\widetilde{F}_1(t)$, given in the following result.

Theorem 4.2. Under H_1, an approximate c.d.f. of MAX is

$$\widetilde{F}_1(t) = \int\int_{\Omega} \phi(z_0, z_1, \widetilde{\mu}, \widetilde{\Sigma}_0) dz_0 dz_1$$

, where $\Omega = \cup_{j=1}^{3} \Omega_j$, $\widetilde{\mu} = (\mu_0, \mu_1)^T$ and $\widetilde{\Sigma}_0 = \begin{pmatrix} \widetilde{\text{Cov}}_{0,0} & \widetilde{\text{Cov}}_{0,1} \\ \widetilde{\text{Cov}}_{1,0} & \widetilde{\text{Cov}}_{1,1} \end{pmatrix}$, where

$$\omega_i = (\widetilde{\text{Cov}}_{1/2,i}\widetilde{\text{Cov}}_{1-i,1-i} - \widetilde{\text{Cov}}_{0,1}\widetilde{\text{Cov}}_{1/2,1-i})/(\widetilde{\text{Cov}}_{0,0}\widetilde{\text{Cov}}_{1,1} - \widetilde{\text{Cov}}_{0,1}^2).$$

The integration regions Ω_j *are given by* $\Omega_1 = \{(z_0, z_1) : |z_0| < t(1-\omega_1)/\omega_0, |z_1| < t\}$, $\Omega_2 = \{(z_0, z_1) : z_0 \in (t(1-\omega_1)/\omega_0, t), z_1 \in (-t, (t-\omega_0 z_0)/\omega_1)\}$, *and* $\Omega_3 = \{(z_0, z_1) : z_0 \in (-t, -t(1-\omega_1)/\omega_0), z_1 \in (-(t+\omega_0 z_0)/\omega_1, t)\}$.

The main steps for the proof of Theorem 4.2 are described below. It is similar to that of Lemma 4.1 [Zang et al., 2010] except $\widetilde{\mu} \neq 0$.

Proof: Simple algebra shows that the determinant for $\widetilde{\Sigma}_1$, the approximate covariance matrix of $\mathbf{Z} = [Z(0), Z(1/2), Z(1)]^T$, is approximately zero. That is, $Z(0)$, $Z(1/2)$, and $Z(1)$ are approximately linearly correlated. Similar to Lemma 4.1, we can write $Z(1/2) = \omega_0 Z(0) + \omega_1 Z(1)$. Then

$$\widetilde{\mathrm{Cov}}_{0,1/2} = \omega_0 \widetilde{\mathrm{Cov}}_{0,0} + \omega_1 \widetilde{\mathrm{Cov}}_{0,1} \quad \text{and} \quad \widetilde{\mathrm{Cov}}_{1/2,1} = \omega_0 \widetilde{\mathrm{Cov}}_{0,1} + \omega_1 \widetilde{\mathrm{Cov}}_{1,1}.$$

Solving the above two equations, we obtain the expressions of ω_i $(i = 0, 1)$ in Theorem 4.2. Then the c.d.f. of MAX is $\widetilde{F}_1(t) = \Pr(|Z_0| < t, |\omega_0 Z(0) + \omega_1 Z(1)| < t, |z_1| < t)$, which has a bivariate normal distribution with regard to $[Z(0), Z(1)]^T$, on the region Ω given in Theorem 4.2. []

An advantage of using the approximate c.d.f. given in Theorem 4.2, comparing to the asymptotic c.d.f. in (4.3), is that the algorithm to compute the c.d.f. of MAX under H_0 can be readily modified to calculate $\widetilde{F}_1(t)$ under H_1. Denote $\phi(z_0, z_1, \widetilde{\mu}, \widetilde{\Sigma}_0) = \widetilde{\phi}(z_0, z_1)$. Then, the corresponding approximate probability density function (p.d.f.) for MAX, obtained by differentiating the c.d.f., can be written as

$$\begin{aligned}
\widetilde{f}_1(t) = \quad & \int_{-t}^{\frac{t(1-\omega_0)}{\omega_1}} \widetilde{\phi}(t, z)dz + \int_{-\frac{t(1-\omega_0)}{\omega_1}}^{t} \widetilde{\phi}(-t, z)dz \\
& + \int_{-t}^{\frac{t(1-\omega_1)}{\omega_0}} \widetilde{\phi}(z, t)dz + \int_{-\frac{t(1-\omega_1)}{\omega_0}}^{t} \widetilde{\phi}(z, -t)dz \\
& + \frac{1}{\omega_1} \left\{ \int_{\frac{t(1-\omega_1)}{\omega_0}}^{t} \widetilde{\phi}\left(z, \frac{t-\omega_0 z}{\omega_1}\right) dz + \int_{-t}^{-\frac{t(1-\omega_1)}{\omega_0}} \widetilde{\phi}\left(z, \frac{-t-\omega_0 z}{\omega_1}\right) dz \right\}.
\end{aligned}$$

Table 4.3: Means and standard deviations of different notations of Ψ's which are c.d.f.'s of normal distributions.

	Mean	Variance
Ψ_1	$\mu_1 + (t-\mu_0)\widetilde{\text{Cov}}_{0,1}/\widetilde{\text{Cov}}_{0,0}$	$\widetilde{\text{Cov}}_{1,1} - \widetilde{\text{Cov}}_{0,1}^2/\widetilde{\text{Cov}}_{0,0}$
Ψ_2	$\mu_1 + (-t-\mu_0)\widetilde{\text{Cov}}_{0,1}/\widetilde{\text{Cov}}_{0,0}$	$\widetilde{\text{Cov}}_{1,1} - \widetilde{\text{Cov}}_{0,1}^2/\widetilde{\text{Cov}}_{0,0}$
Ψ_3	$\mu_0 + (t-\mu_1)\widetilde{\text{Cov}}_{0,1}/\widetilde{\text{Cov}}_{1,1}$	$\widetilde{\text{Cov}}_{0,0} - \widetilde{\text{Cov}}_{0,1}^2/\widetilde{\text{Cov}}_{1,1}$
Ψ_4	$\mu_0 + (-t-\mu_1)\widetilde{\text{Cov}}_{0,1}/\widetilde{\text{Cov}}_{1,1}$	$\widetilde{\text{Cov}}_{0,0} - \widetilde{\text{Cov}}_{0,1}^2/\widetilde{\text{Cov}}_{1,1}$
Ψ_5	$\mu_0 + (t-\mu_*)(\omega_0\widetilde{\text{Cov}}_{0,0} + \omega_1\widetilde{\text{Cov}}_{0,1})/\sigma_*^2$	$\widetilde{\text{Cov}}_{0,0} - (\omega_0\widetilde{\text{Cov}}_{0,0} + \omega_1\widetilde{\text{Cov}}_{0,1})^2/\sigma_*^2$
Ψ_6	$\mu_0 + (-t-\mu_*)(\omega_0\widetilde{\text{Cov}}_{0,0} + \omega_1\widetilde{\text{Cov}}_{0,1})/\sigma_*^2$	$\widetilde{\text{Cov}}_{0,0} - (\omega_0\widetilde{\text{Cov}}_{0,0} + \omega_1\widetilde{\text{Cov}}_{0,1})^2/\sigma_*^2$

The above density of MAX can also be expressed using the p.d.f. and c.d.f. of univariate normal distributions as follows:

$$\begin{aligned}
\widetilde{f}_1(t) &= \phi(t;\mu_0,\widetilde{\text{Cov}}_{0,0})\left\{\Psi_1(\frac{t(1-\omega_0)}{\omega_1}) - \Psi_1(-t)\right\} \\
&\quad +\phi(-t;\mu_0,\widetilde{\text{Cov}}_{0,0})\left\{\Psi_2(t) - \Psi_2(\frac{t(1-\omega_0)}{\omega_1})\right\} \\
&\quad +\phi(t;\mu_1,\widetilde{\text{Cov}}_{1,1})\left\{\Psi_3(\frac{t(1-\omega_1)}{\omega_0}) - \Psi_3(-t)\right\} \\
&\quad +\phi(-t;\mu_1,\widetilde{\text{Cov}}_{1,1})\left\{\Psi_4(t) - \Psi_4(\frac{t(1-\omega_1)}{\omega_0})\right\} \\
&\quad +\phi(t;\mu_*,\sigma_*^2)\left\{\Psi_5(t) - \Psi_5(\frac{t(1-\omega_1)}{\omega_0})\right\} \\
&\quad +\phi(-t;\mu_*,\sigma_*^2)\left\{\Psi_6(-\frac{t(1-\omega_1)}{\omega_0}) - \Psi_6(-t)\right\},
\end{aligned}$$

where $\mu_* = \omega_0\mu_0 + \omega_1\mu_1$, $\sigma_*^2 = \omega_0^2\widetilde{\text{Cov}}_{0,0} + \omega_1^2\widetilde{\text{Cov}}_{1,1} + 2\omega_0\omega_1\widetilde{\text{Cov}}_{0,1}$, $\phi(;\mu,\sigma^2)$ is the p.d.f. of normal distribution with mean μ and variance σ^2, and Ψ's are the c.d.f.'s of normal distributions with the following means and variances (Table 4.3):

To indicate that $\widetilde{f}_1(t)$ is a function of $\mathbf{p}$ and $\mathbf{q}$, we denote it as $\widetilde{f}_1(t|\mathbf{p},\mathbf{q})$.

4.3 Bayes factor based on MAX

The BF of Kass and Raftery (1995), given the genetic model x, can be written as

$$\mathrm{BF}(x) = \frac{\Pr(\text{data}|H_1)}{\Pr(\text{data}|H_0)} = \frac{\int f(\text{data};\theta,x)\pi_1(\theta)d\theta}{\int f(\text{data};\theta)\pi_0(\theta)d\theta},$$

where $f(\text{data};\theta,x)$ is the likelihood of the data where x vanishes under H_0, θ contains all the parameters, and $\pi_i(\theta)$ is the prior density for θ under H_i, $i = 0, 1$. Note that $\mathrm{BF}(x)$ is used when x is known or when the same x is used to calculate the p-value. For example, if the CATT ($x = 1/2$) is used to calculated all the p-values for SNPs in GWAS, BF(1/2) should be reported accordingly. In Bayesian analysis, if a parameter is unknown, it is usually treated as a random variable and averaged out (like θ) with a prior distribution. If we treat x as a random variable with a prior $\Pi(x)$ for $x = 0, 1/2, 1$, then

$$\mathrm{BF} = \frac{\sum_x \int f(\text{data};\theta,x)\pi_1(\theta)\Pi(x)d\theta}{\int f(\text{data};\theta)\pi_0(\theta)d\theta} = \sum_x \mathrm{BF(x)}\Pi(x),$$

which is referred to as Bayesian model averaging (BMA) (Stephens and Balding, 2009). The above BMA is a typical way to handle model uncertainty in Bayesian analysis. Zheng et al. (2011) considered $\max_x \mathrm{BF}(x)$ when x is unknown. They showed that the BMA overall works concordantly with the p-values of MAX. However, the maximum of BFs often detects more associations than the BMA in GWAS among SNPs whose p-values of MAX are significant at the level of 1E-5. Without conditioning on a small p-value, the maximum of BFs claims more false positive associations than $\mathrm{BF}(x)$ for any $x = 0, 1/2, 1$ and the BMA. This is not surprising as the p-value of MAX, without a correction, tends to claim more false positive than the p-value of any single trend test.

Johnson (2005,2008) considered a BF given below

$$\mathrm{BF}_J = \frac{\Pr(T|H_1)}{\Pr(T|H_0)} = \frac{\int f_1(t;\beta)\pi_1(\beta)d\beta}{\int f_0(t;\beta)\pi_0(\beta)d\beta}, \tag{4.4}$$

where T is a test statistic and $f_i(t;\beta)$ is the asymptotic distribution of T under H_i, $i = 0, 1$. Applying (4.4) with $T = \hat{\beta}_x$ as the estimate of log odds ratio β of genetic effect under model x and using a normal prior for $\beta \sim N(0, \sigma^2)$, Wakefield (2007, 2009) obtained

$$\mathrm{BF}_W(x) = \left(\frac{V_x}{\sigma^2 + V_x}\right)^{1/2} \exp\left(\frac{\hat{\beta}_x^2 \sigma^2}{2V_x(\sigma^2 + V_x)}\right),$$

where V_x is the asymptotic variance of $\hat{\beta}_x$. In this paper we consider BMA $= \sum_x \mathrm{BF}_W(x)\Pi(x)$ and maxBF $= \max_x \mathrm{BF}_W(x)$ as well as the proposed BF measure based on MAX given by

$$\mathrm{BFM} = \frac{\Pr(\mathrm{MAX}|H_1)}{\Pr(\mathrm{MAX}|H_0)}. \tag{4.5}$$

It can be regarded as an application of BF of (4.4) with $T =$ MAX. Here are the motivations of (4.5). First, it is a ratio of two probabilities under H_1 and H_0, respectively. Thus, it is expected that it would perform like a BF and would not claim more false positives like maxBF does. Second, it would be comparable to MAX and its p-value as it directly based on MAX. Third, as the asymptotic distribution of MAX under H_1 depends on the genetic model x, the BFM would be less dependent on the prior of x than BMA because the model averaging using the prior of x is applied after the modeling of MAX.

Denote the densities of MAX under H_0 and H_1 as $f_0(t|\mathbf{p} = \mathbf{q})$ and $\widetilde{f}_1(t|\mathbf{p}, \mathbf{q})$, respectively, because both are functions of genotype probabilities $\mathbf{p}$ in cases and $\mathbf{q}$ in controls. Note that $f_0(t|\mathbf{p} = \mathbf{q}) = \widetilde{f}_1(t|\mathbf{p} = \mathbf{q})$. We reparameterize $p_0 = q_0/\Delta$, $p_1 = \delta_1 q_1/\Delta$ and $p_2 = \delta_2 q_2/\Delta$, where $\Delta = q_0 + \delta_1 q_1 + \delta_2 q_2$. Under H_0, $\delta_1 = \delta_2 = 1$ and under H_1, given $x = 0, 1/2, 1$, we set $\delta_1 = 1$, $\delta_1 = (1 + \delta_2)/2 = (1 + \delta)/2$, $\delta_1 = \delta_2 = \delta$,

respectively. Thus, $\mathbf{p}$ is a function of $\mathbf{q}$, δ and x, denoted as $\mathbf{p} = \xi(\mathbf{q}, \delta, x)$. The parameter $\mathbf{q}$ contains genotype frequencies in controls and is a nuisance parameter. In BF, a prior is determined for $\mathbf{q}$ which is averaged out in BF under H_0 and H_1. We take a different approach by replacing $\mathbf{q}$ by $\widehat{\mathbf{q}}$ using controls. This leads to two parameters left δ for the genetic effect under H_1 and x for the underlying genetic model. We specify prior $\pi(\delta)$ as $\log \delta \sim N(0, \sigma^2)$ or a mixture of normals, the same prior used in the usual BF, and $\Pi(x)$ for x as before. Thus, our BFM given in (4.5) can be written as

$$\text{BFM} = \frac{\sum_x \int \widetilde{f}_1(t|\xi(\widehat{\mathbf{q}}, \delta, x), \widehat{\mathbf{q}}, \delta)\pi(\delta)d\delta\Pi(x)}{\widetilde{f}_1(t|\mathbf{p} = \mathbf{q}, \mathbf{q} = \widehat{\mathbf{q}})} = \sum_x \frac{\int \widetilde{f}_1(t|\xi(\widehat{\mathbf{q}}, \delta, x), \widehat{\mathbf{q}}, \delta)\pi(\delta)d\delta}{\widetilde{f}_1(t|\mathbf{p} = \mathbf{q}, \mathbf{q} = \widehat{\mathbf{q}})}\Pi(x). \tag{4.6}$$

The computation in (4.6) involves evaluations of multiple single integrations in the numerator.

Directly comparing (4.6) to the BMA, we see that the BF based on the data or the MLE of the log odds ratio in the BMA is replaced by the BF based on the MAX statistic. Moreover, in the BMA, the correlations among the estimates of log odds ratios under three different genetic models are not taken into account, which is incorporated in the BFM through MAX. It is known that the correlations play an important role in determining how robust the tests are to detect true associations (Gastwirth, 1985; Joo et al., 2010). Thus, although the model averaging technique is used in both BMA and BFM, it has different roles and it is expected that BFM would be less sensitive to the prior $\Pi(x)$ due to the impact of MAX on BFM. To take into account the above correlations in Bayesian analysis, one can model odds ratios of G_1 to G_0 and G_2 to G_0. That is, one can replace T in (4.4) by the estimates of two odds ratios (using bivariate normal distributions under H_0 and H_1) so that the correlations among them can be incorporated. In this case, priors of bivariate normal

distributions should be specified under H_0 and H_1, This includes specifying priors for the correlations under H_0 and H_1. Specifying these correlations itself implies a genetic model. For example, specifying equal variances for the two odds ratios implies a DOM model *a priori*, and specifying a much larger variance for one odds ratio than the other implies a REC model. In our approach, these correlations are expressed in terms of $\mathbf{p}$ and $\mathbf{q}$. No priors are required for the covariance matrix.

4.4 Simulation studies and results

4.4.1 Priors.

Choosing priors for calculating BFs is important. Our paper is to discuss a new BF measure rather than a discussion of which priors should be used. Thus, we only considered the priors suggested in the literature. For the genetic effect, e.g., the log odds ratio β, we considered a normal prior $N(0, \sigma^2)$ and a mixture of normals. The discussion of choosing σ^2 for a single normal can be found in the WTCCC (2007) and Wakefield (2007, 2009). In this paper, we chose $\sigma^2 = 0.06$, which is in the range of the suggested σ^2. The mixture of normal prior has been considered by Stephens and Balding (2009). We used $0.9N(0, 0.04)+0.05N(0, 0.16)+0.05N(0, 0.64)$ as in Stephens and Balding (2009), in which small weights are placed on larger variances. Usually, a larger variance often leads to a larger BF, a stronger indication of association.

For the genetic models, we considered two types of priors. First, we chose $\Pi(x) = 1/3$ for all three models, which is a non-informative prior. Because it is expected to see more SNPs with associations under the ADD model, an informative prior is $\Pi(0) = 0.1$, $\Pi(0.5) = 0.8$ and $\Pi(1) = 0.1$, which was also used by Stephens and Balding (2009). Therefore, a combination of four different choices of priors was considered in the simulations. The purpose of choosing different priors is to examine

the sensitivity of the performance of different approaches in terms of the choices of priors.

The MAF was 0.15, 0.30 or 0.45, and then the genotype frequencies $g_i = Pr(G_i)$, $i = 0, 1, 2$, were computed under HWE. We specified $\lambda_2 = 1.5$ under H_1 ($\lambda_2 = 1$ under H_0) and computed λ_1 for a given genetic model. Then $\eta_0 = \kappa/(g_0 + \lambda_1 g_1 + \lambda_2 g_2)$ and $\eta_i = \eta_0 \lambda_i$, $i = 1, 2$ were computed, where the prevalence was $\kappa = 0.01$. The genotype counts were randomly generated using $(r_0, r_1, r_2)^T \sim \text{Mul}(r; \mathbf{p})$ and $(s_0, s_1, s_2)^T \sim \text{Mul}(s; \mathbf{q})$, where $\mathbf{p}$ and $\mathbf{q}$ were computed using the formulas given before. The prior of H_1, $\Pr(H_1)$, was 0.01 for testing a single SNP. The sample size was $n = 2000$ ($r = s = 1000$).

4.4.2 The approximate distribution of MAX

The histograms of MAX from 10,000 replicates were presented in Figure 4.1 along with its approximate p.d.f. $\widetilde{f}_1(t)$ (the solid curve) given in Section 4.2.2. The plots show that the distributions of MAX vary across the genetic models, MAFs and GRRs, and that the approximate p.d.f. of MAX given in Theorem 4.2 fits the histograms of simulated MAX very well. Note that, under H_0, the distribution of MAX is skewed, while under H_1, it is quite symmetric because one of the three normally distributed trend tests dominates the others.

Figure 4.2 presents the Q-Q plots comparing the simulated c.d.f. versus the asymptotic c.d.f. in (4.3) (the left column), the simulated versus the approximate c.d.f. in Theorem 4.2 (the middle column), and the asymptotic versus approximate c.d.f. (the right column) of MAX under H_1. MAF=0.3 is used in all plots. The true genetic models are REC, ADD and DOM from top to bottom. The plots show that the approximate c.d.f. computed in Theorem 4.2 provides a very good approximation to both the asymptotic c.d.f. and the simulated c.d.f.

Figure 4.1: Histograms of MAX with $r = s = 1000$. The MAF is 0.15, 0.3 or 0.45 from the left to the right. The first row is under H_0 and the second to the fourth are REC, ADD and DOM models, repsectively, given $\lambda_2 = 1.5$. The solid curve is the approximate p.d.f. of MAX.

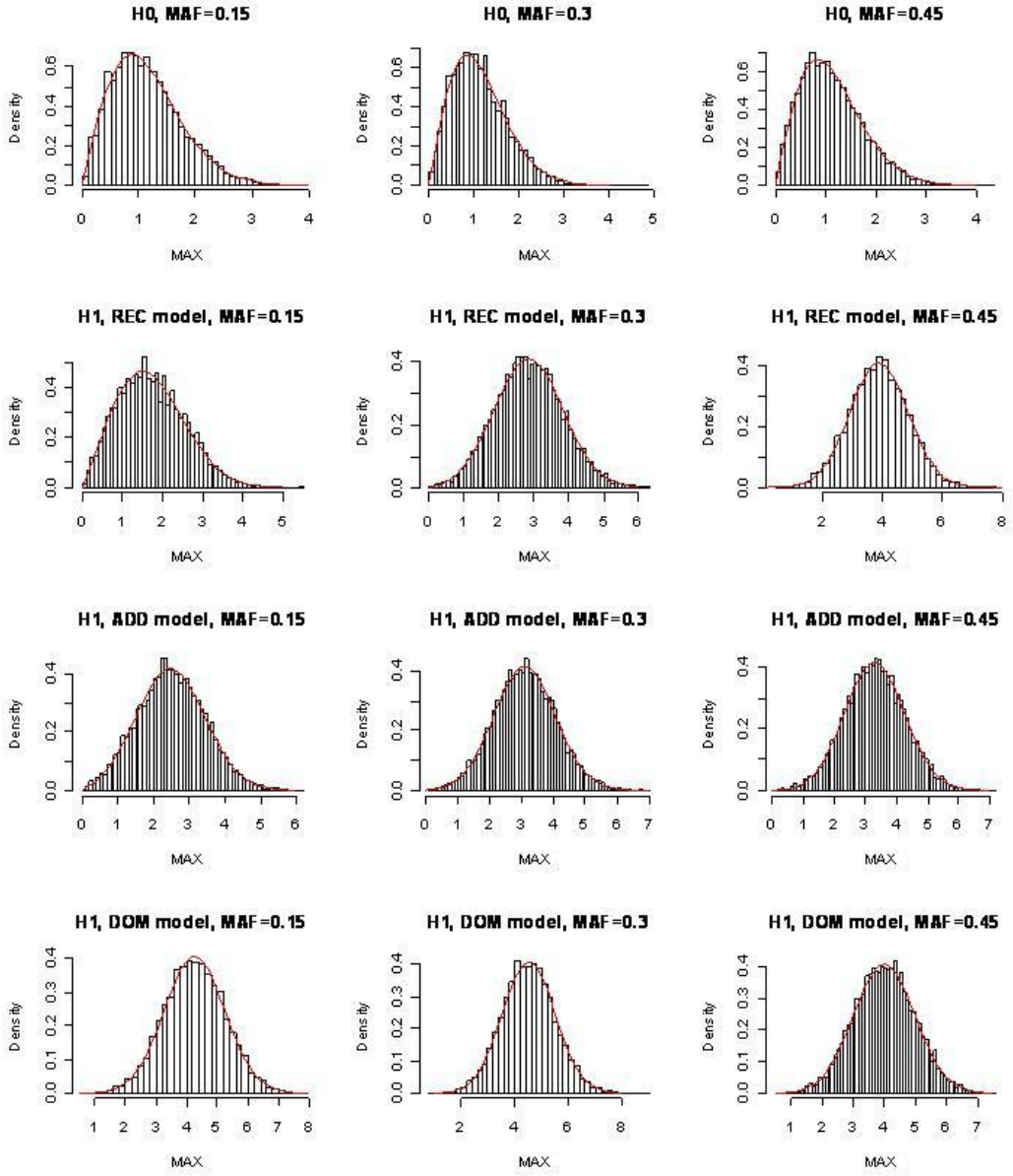


Figure 4.2: Q-Q plots comparing the simulated c.d.f., the asymptotic c.d.f., and the approximate c.d.f. of MAX. The diagonal line is indicated, which almost perfectly overlaps the points on the Q-Q plots.

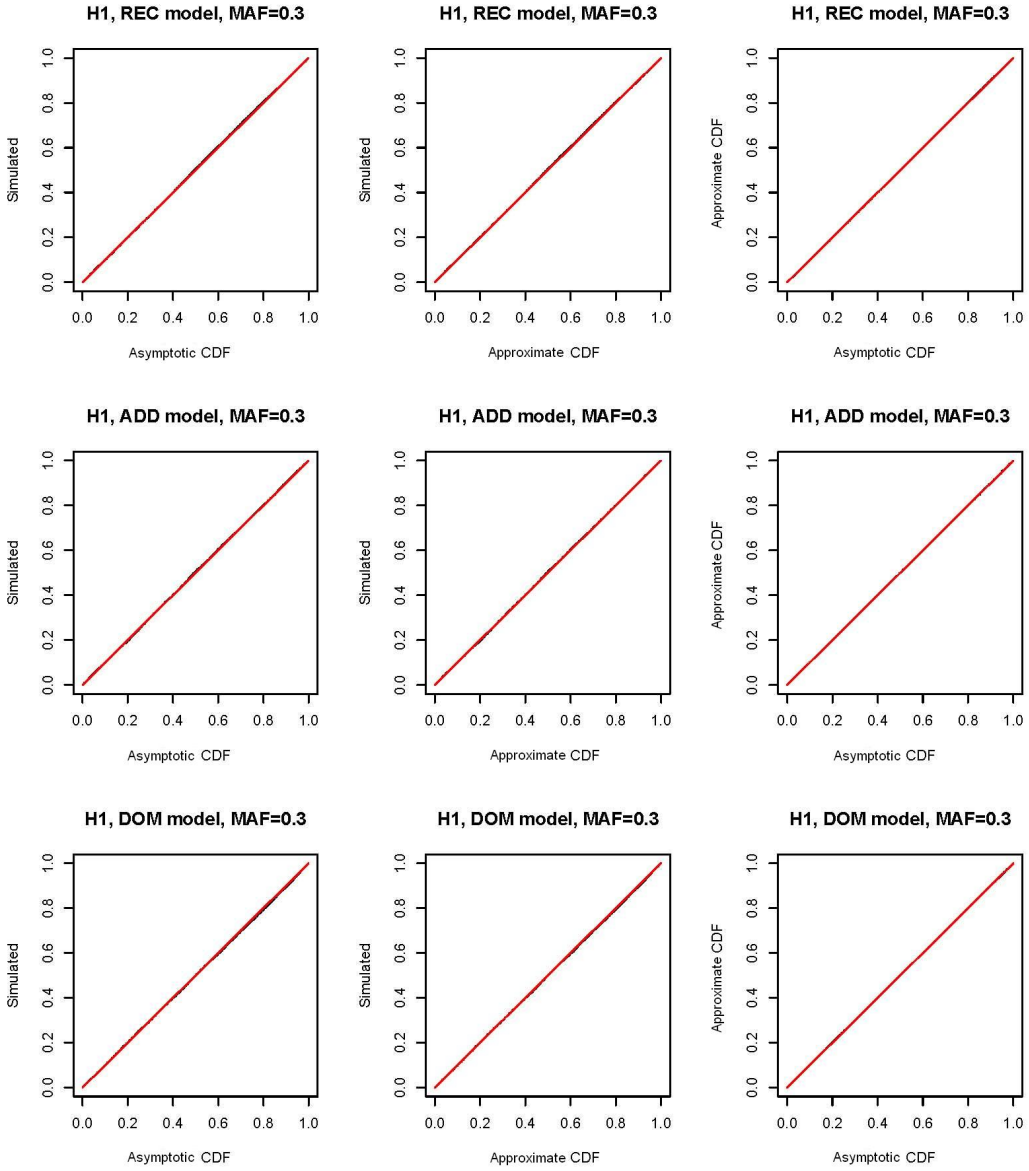
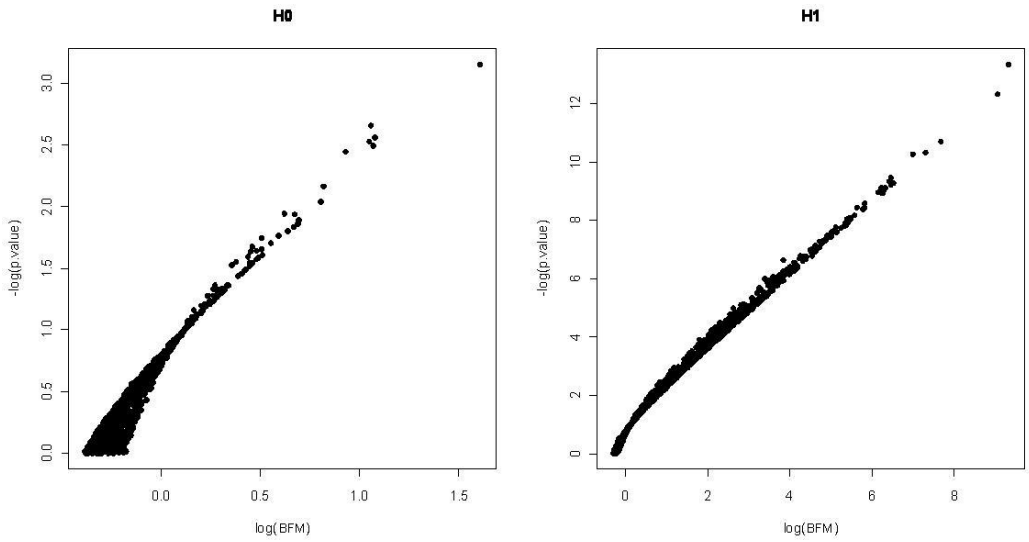


Figure 4.3: Plots of $-\log_{10}$(p-value) versus $\log_{10}$ BFM with $r = s = 1000$ under H_0 and H_1.



4.4.3 Comparison of BFM and the p-value of MAX

We randomly generated 1000 SNPs to study the relationship between BFM and the p-value of MAX. The MAF of each SNP was randomly generated from a uniform distribution $U(0.1, 0.9)$. Under H_1, the genetic model x was generated from the discrete uniform distribution with $\Pi(x) = 1/3$ for $x = 0, 1/2, 1$. Figure 4.3 presents the plots of $-\log_{10}$(p-value) versus $\log_{10}$ BFM. The plots show the consistency between the p-value and BFM across different MAFs and genetic models when the power in the study design is fixed. The smaller the p-value, the greater the BFM indicating stronger evidence of association contained in MAX.

4.4.4 Posterior probability of associations under various genetic models

Given the prior of an association, $\Pr(H_1)$, we can calculate the posterior probability of association (PPA) based on MAX as $\Pr(H_1|\text{MAX})$ given by

$$\text{PPAM} = \Pr(H_1|\text{MAX}) = \frac{\text{BFM} \times \Pr(H_1)}{1 + \text{BFM} \times \Pr(H_1)}.$$

For the BMA, its PPA can be obtained similarly and denoted as PPAA. However, for maxBF, we define maxPPA= $\max_x \text{PPA}(x)$ where, for $x = 0, 1/2, 1$,

$$\text{PPA}(x) = \frac{\text{BF}(x) \times \Pr(H_1)}{1 + \text{BF}(x) \times \Pr(H_1)}.$$

We compare PPAM, PPAA and maxPPA under H_0 and H_1. In all the comparisons, $\Pr(H_1) = 0.01$. To compare PPAs, we first report the means of the three PPAs among 10,000 replicates. Then we examine the percentage the simulated PPAs is greater than 0.20 under H_0 and H_1, respectively for each of the three PPAs. We chosen 0.20 so we can compare how likely the probability of H_1 after observing the data increases to 0.20 from 0.01. The results are reported in Table 4.4 – Table 4.7 for the four combinations of the priors for the genetic model x and genetic effect: a non-informative or an informative prior for x, and a normal or a mixture of normal prior for the genetic effect.

Table 4.4: Means of PPAs and the percentages of PPA > 0.2, given $\kappa = 0.01$ and $\lambda_2 = 1.5$ under H_1. A non-informative prior $\Pi(0) = \Pi(1/2) = \Pi(1) = 1/3$ for x and a normal prior $N(0, 0.06)$ for the genetic effect are used.

Model	PPA	MAF=0.15		MAF=0.30		MAF=0.45	
		mean	% > .2	mean	% > .2	mean	% > .2
Null	PPAM	0.009	0.07	0.009	0.12	0.009	0.15
	maxPPA	0.013	0.13	0.013	0.35	0.013	0.41
	PPAA	0.009	0.05	0.009	0.11	0.009	0.15
REC	PPAM	0.023	1.63	0.181	27.1	0.508	68.1
	maxPPA	0.023	0.75	0.230	35.9	0.604	77.8
	PPAA	0.015	0.33	0.152	23.0	0.504	67.3
ADD	PPAM	0.118	16.6	0.268	39.2	0.319	46.5
	maxPPA	0.192	28.3	0.355	51.6	0.400	57.9
	PPAA	0.142	20.5	0.284	41.7	0.322	47.6
DOM	PPAM	0.582	75.1	0.715	86.5	0.554	72.6
	maxPPA	0.724	87.1	0.811	92.6	0.651	82.1
	PPAA	0.643	80.6	0.743	87.9	0.552	72.3

The results in Table 4.4 – Table 4.7 can be summarized as follows. Under H_0, PPAM and PPAA have similar mean PPAs and similar estimated probabilities that PPAs are greater than the threshold 0.2. maxPPA, by the definition, has the largest PPA among the three PPAs even under H_0. Under H_1, we focus on PPAM and PPAA. When the true model is REC, the proposed PPAM outperforms PPAA, while the latter outperforms PPAM when the ADD model is true. Under the DOM model, however, they have similar performance, depending on the MAF. For a smaller MAF, PPAA outperforms PPAM, while the latter outperforms PPAA for a larger MAF. The choices of priors do not change the above conclusions. However, when comparing Table 4.4 vs. Table 4.5 and Table 4.6 vs. Table 4.7, the BMA-based PPAA is more sensitive to the prior of the genetic model than the proposed PPAM. For example, in Table 4.6 vs. Table 4.7, the % > 0.2 for PPAM under the REC model with

Table 4.5: Continued with an informative prior $(\Pi(0), \Pi(1/2), \Pi(1)) = (0.1, 0.8, 0.1)$ for x and the same normal prior as in (A) for the genetic effect.

Model	PPA	MAF=0.15		MAF=0.30		MAF=0.45	
		mean	% > .2	mean	% > .2	mean	% > .2
Null	PPAM	0.010	0.05	0.009	0.08	0.009	0.17
	maxPPA	0.013	0.20	0.013	0.32	0.013	0.44
	PPAA	0.010	0.08	0.009	0.10	0.009	0.18
REC	PPAM	0.023	1.27	0.175	25.9	0.497	67.9
	maxPPA	0.024	0.66	0.228	35.0	0.606	78.3
	PPAA	0.015	0.46	0.121	17.0	0.442	61.5
ADD	PPAM	0.111	15.0	0.266	39.8	0.310	45.8
	maxPPA	0.190	27.5	0.362	52.5	0.401	58.6
	PPAA	0.145	21.0	0.312	46.6	0.354	52.3
DOM	PPAM	0.556	74.0	0.697	85.7	0.545	72.6
	maxPPA	0.728	87.4	0.806	92.2	0.653	82.0
	PPAA	0.606	79.0	0.695	85.0	0.494	67.0

Table 4.6: Continued with the same non-informative prior as in (A) for x and a mixture of normal prior $0.9N(0, 0.04) + 0.05N(0, 0.16) + 0.05N(0, 0.64)$ for the genetic effect.

Model	PPA	MAF=0.15		MAF=0.30		MAF=0.45	
		mean	% > .2	mean	% > .2	mean	% > .2
Null	PPAM	0.010	0.06	0.009	0.10	0.009	0.11
	maxPPA	0.013	0.21	0.012	0.22	0.013	0.34
	PPAA	0.009	0.09	0.009	0.10	0.009	0.10
REC	PPAM	0.021	1.20	0.165	24.1	0.484	66.0
	maxPPA	0.022	0.68	0.203	31.0	0.582	76.7
	PPAA	0.014	0.31	0.133	19.3	0.478	65.4
ADD	PPAM	0.106	14.3	0.255	37.4	0.291	42.5
	maxPPA	0.175	26.0	0.340	49.9	0.367	53.8
	PPAA	0.128	18.0	0.270	39.8	0.292	43.5
DOM	PPAM	0.561	73.1	0.690	84.3	0.529	70.3
	maxPPA	0.706	86.3	0.794	91.8	0.628	80.5
	PPAA	0.620	79.0	0.720	86.5	0.528	69.9

Table 4.7: Continued with the same informative prior as in (B) for x and the same mixture of normal prior as (C) for the genetic effect.

Model	PPA	MAF=0.15		MAF=0.30		MAF=0.45	
		mean	% > .2	mean	% > .2	mean	% > .2
Null	PPAM	0.010	0.04	0.009	0.09	0.009	0.05
	maxPPA	0.012	0.12	0.012	0.25	0.012	0.24
	PPAA	0.009	0.03	0.009	0.11	0.009	0.09
REC	PPAM	0.022	1.15	0.155	22.4	0.472	64.7
	maxPPA	0.023	0.92	0.201	31.0	0.584	76.3
	PPAA	0.015	0.44	0.104	14.2	0.419	58.5
ADD	PPAM	0.100	13.1	0.238	35.3	0.284	41.8
	maxPPA	0.176	25.6	0.330	48.0	0.372	54.2
	PPAA	0.129	18.0	0.280	41.3	0.325	47.6
DOM	PPAM	0.543	71.6	0.670	83.4	0.509	68.9
	maxPPA	0.712	86.6	0.791	92.0	0.624	80.2
	PPAA	0.584	76.4	0.668	83.0	0.459	63.3

MAF 0.45 is 66.0% with the non-informative prior and 64.7% with the informative prior, but for PPAA they are 65.4% and 58.5% respectively. Under the DOM model, the percentages are 70.3% and 68.9% for PPAM and 69.9% and 63.3% for PPAA. Hence, PPAM is more robust to the choice of prior with respect to the genetic model than PPAA. More comparisons of sensitivity of PPAs based on BFM and BMA are reported in Table 4.8. The results confirm our conclusions, especially for moderate or common MAFs, under which BMA is quite sensitive to the choice of prior for x.

4.5 Applications

Based on the simulation results, we focus on the BMA-based PPAA and our proposed MAX-based PPAM in applications. We first apply PPAM and PPAA to the 14 SNPs in Table 4.1. The results are reported in Table 4.9 with four different

Table 4.8: Sensitivity of the PPAs based on BFM and BMA to the choice of prior for model x. Equal prior denotes the non-informative prior for x and unequal denotes the informative prior for x.

			MAF=0.15		MAF=0.30		MAF=0.45	
Model	PPA	$\Pi(x)$	mean	% > .2	mean	% > .2	mean	% > .2
REC	BFM	equal	0.023	1.63	0.181	27.1	0.508	68.1
		unequal	0.023	1.27	0.175	25.9	0.497	67.9
	BMA	equal	0.015	0.33	0.152	23.0	0.504	67.3
		unequal	0.015	0.46	0.121	17.0	0.442	61.5
ADD	BFM	equal	0.118	16.6	0.268	39.2	0.319	46.5
		unequal	0.111	15.0	0.266	39.8	0.310	45.8
	BMA	equal	0.142	20.5	0.284	41.7	0.322	47.6
		unequal	0.145	21.0	0.312	46.6	0.364	52.3
DOM	BFM	equal	0.582	75.1	0.715	86.5	0.554	72.6
		unequal	0.556	74.0	0.697	85.7	0.545	72.6
	BMA	equal	0.643	80.6	0.743	87.9	0.552	72.3
		unequal	0.606	79.0	0.695	85.0	0.494	67.0

combinations of priors for the genetic model and the genetic effect. The underlying genetic model for each SNP based on the p-values of three trend tests is indicated by R, A and D for the REC, ADD and DOM models after each SNP ID. In Table 4.9, the prior of association is $Pr(H_1) = 10^{-4}$.

Overall, PPAA and PPAM have comparable PPAs under priors I and III, when the non-informative prior is used for the genetic model x. Their performances are somehow different under priors II and IV, when the informative prior is used for x. Like what we have demonstrated in the simulations, PPAA is also more sensitive to the choice of prior for x than PPAM. For example, for the SNP with strongest association with BD (rs420259), PPAM is 0.971 and PPAA is 0.803 using prior I. They are 0.953 and 0.530 using prior II, respectively. For PPAM, both PPAs are greater than 0.95, while a large reduction of PPA from 0.803 to 0.530 is observed

Table 4.9: The PPAM and PPAA of the 14 SNPs associated with BD (WTCCC, 2007). $\Pr(H_1) = 10^{-4}$. Prior I = the non-informative prior for the genetic model x and the normal prior for the genetic effect β; Prior II = the informative prior for x and the mixture normal prior for β; Prior III = the non-informative prior for x and the mixture of normal prior for β; and Prior IV = the informative prior for x and the normal prior for β. SNP rs420259 is the only one that has shown strong association with BD.

	Prior							
	I		II		III		IV	
SNP ID	PPAM	PPAA	PPAM	PPAA	PPAM	PPAA	PPAM	PPAA
rs3761218D	0.347	0.363	0.272	0.152	0.297	0.331	0.330	0.173
rs7570682A	0.144	0.142	0.111	0.199	0.121	0.112	0.143	0.255
rs10982256A	0.087	0.076	0.069	0.122	0.075	0.062	0.084	0.150
rs2953145A	0.056	0.050	0.045	0.073	0.047	0.038	0.058	0.097
rs683395D	0.224	0.389	0.142	0.202	0.189	0.331	0.158	0.245
rs4027132R	0.191	0.192	0.149	0.124	0.162	0.159	0.181	0.151
rs420259D	0.971	0.803	0.953	0.530	0.962	0.789	0.967	0.552
rs4276227A	0.136	0.131	0.106	0.183	0.114	0.107	0.131	0.226
rs10134944D	0.192	0.338	0.123	0.173	0.161	0.287	0.141	0.214
rs1375144A	0.209	0.176	0.162	0.259	0.175	0.139	0.202	0.320
rs1344484A	0.282	0.245	0.221	0.345	0.240	0.199	0.269	0.413
rs6458307D	0.030	0.022	0.024	0.005	0.025	0.017	0.030	0.007
rs2609653A	0.038	0.044	0.022	0.039	0.031	0.037	0.023	0.044
rs11622475A	0.214	0.336	0.166	0.301	0.180	0.294	0.209	0.365

The letter R, A and D at the end of SNP ID denotes the genetic model REC, ADD and DOM, respectively based on the p-values of the trend tests.

Table 4.10: The PPAM and PPAA of the 2 SNPs associated with AMD [Klein et al., 2005]. $\Pr(H_1) = 10^{-4}$. The four combinations of priors are given in Table 4.9. The p-values of MAX are 8.65E-07 (rs380390) and 2.34E-06 (rs1329428), calculated using Zang et al. [2010].

SNP ID	Prior	PPAM	PPAA
rs380390	I	1.36×10^{-3}	0.53×10^{-3}
	II	11.5×10^{-3}	5.37×10^{-3}
	III	18.9×10^{-3}	3.29×10^{-3}
	IV	0.89×10^{-3}	0.56×10^{-3}
rs1329428	I	0.80×10^{-3}	0.77×10^{-3}
	II	4.07×10^{-3}	3.01×10^{-3}
	III	7.03×10^{-3}	6.69×10^{-3}
	IV	0.54×10^{-3}	0.44×10^{-3}

when PPAA is used. As is expected, PPAA is less sensitive to the prior of x when the potential genetic model is ADD. Another example is SNP rs6458307. PPAA reduces from 0.022 (prior I) to 0.005 (prior II), while the corresponding PPAMs are 0.030 and 0.024, respectively.

In the second application, we considered the top two SNPs with age-related macular degeneration (AMD) (Klein et al., 2005). The results are presented in Table 4.10. In this application, PPAM and PPAA have similar performance. However, with the same prior $Pr(H_1) = 10^{-4}$, both PPAM and PPAA in Table 4.10 are much smaller than those reported in Table 4.9. P-values of MAX for the two SNPs are also given in Table 4.10. The p-values of MAX for the 14 SNPs are given in Table 4.1. If we compare the p-values of MAX in Table 4.1 and Table 4.10, we find that the p-values of MAX for AMD are smaller than most of the p-values of MAX in Table 4.1, indicating more significant associations. The reason for a more significant association (p-value) without a stronger PPA is that the sample size of the AMD study was about 150. For example, for the first SNP (rs380390), it contains only 96 cases and 50 controls. But the sample size for BD was about 5000 (2000 cases and 3000 controls). Thus,

the power to detect association in AMD and BD is different, which is not reflected in the p-values of MAX. For the AMD study, with such a small sample size, the power to detect to an association is relatively lower than that of the BD study, although the region covering these top two SNPs is truly associated with AMD because it has been confirmed and replicated by many other subsequent studies.

4.6 Discussion and Conclusion

Robust tests and their p-values have been recently studied for case-control genetic association studies, especially the MAX test. It is more robust than a single trend test and Pearson's chi-squared test under genetic model uncertainty. In this chapter, we firstly derived the means and covariances of the three trend tests optimal for REC, ADD, and DOM models, respectively. Then we studied the asymptotic and approximate distributions for MAX under H_1. By showing that the determinant of the covariance matrix for the three trend tests is approximately 0, we were able obtain an approximte c.d.f. of MAX that reduces the three-fold integration to the sum of several double integrations. Simulation under different genetic models and MAFs showed that the approximated c.d.f. provides a good fit of the distribution of MAX.

In practice, p-values of MAX are used to rank the SNPs and report the significance of associations in GWA studies [e.g. Sladek et al., 2007]. In GWA studies testing half to one million hypotheses, small p-values are always observed regardless of the power to detect associations. Hence p-value cannot be used alone as a measure of strength of association, as understanding the strength of evidence conveyed by a particular p-value also requires knowledge of power. BF and PPA unify the power and the significance of p-value in a single measure to provide strength of association in favor of H_1. Typical BF depends on an underlying genetic model [Johnson, 2005;

Wakefield, 2007, 2009], which is often unknown and cannot be detected under the linkage disequilibrium between the SNP and the function loci [Zheng et al., 2009]. How to calculate a BF corresponding to MAX has not been studied, which also needs the asymptotic distribution of MAX under H_1. On the other hand, the maximum BF of Zheng et al. [2011] has to be applied conditional on SNPs showing at least some moderate association and that it cannot be interpreted as a ratio of two probabilities under H_0 and H_1.

In this chapter, we proposed a BF that measures the strength of association in MAX, denoted as BFM, and used the approximate c.d.f. of MAX to compute it. This BFM is not a traditional BF of Kass and Raftery [1995], which is defined as the probability of data under the alternative hypothesis (H_1) over that under the null hypothesis (H_0). Our BFM is defined as the probability of MAX under H_1 over that under H_0. That is, the data in the usual BF are replaced by MAX. There are some important differences between the measures of Johnson [2005] and Wakefield [2007] from our proposed BFM: MAX does not have standard distributions under H_0 and H_1 and the distribution of MAX under H_0 and H_1 also depends on parameters (e.g., the allele frequency) while in Johnson [2005], for example, the distributions of statistics under H_0 are free of any parameter.

Results from simulations and data applications demonstrate that our proposed method has similar performance to the existing Bayesian model averaging but the latter can be more sensitive to the choice of prior for the genetic model. Numerical results also demonstrate that, when the power to detect an association is fixed, the p-value of MAX and the BFM are comparable. However, when the power is different, BF based on MAX provides more evidence of association than the p-value alone. Sawcer [2010] illustrated this point using simulations and we demonstrated this using

real data from GWA studies, in which small p-value may be caused by the lack of power and our proposed BFM provides a unified measurement of both the p-value and the power.

Chapter 5

Discussion and Future Research in Related Areas

5.1 Discussion

As the genetic testing and genome-wide scanning technologies become widely available, population-based genetic studies examining the association between diseases and genetic markers have gained popularity in recent years. The identification of the loci that are potentially associated with diseases makes it possible for researchers to develop the treatment and prevention of certain diseases from the genetic perspective. However, through our studies we have shown that the existence of hidden subpopulations can cause bias and variance distortion in the estimation of odds ratio and its confidence interval, which would further affect the results of Bayesian analysis. To resolve this problem, we applied a principal components analysis method to assign individuals into subgroups, and then adjust for the effect of hidden population stratification on the estimate of odds ratio and its asymptotic variance. We then applied these corrected estimate and the asymptotic variance to the approximate Bayes factor. The comparison of the asymptotic distribution of the approximate Bayes factor

before and after the proposed correction shows that our method yields an appropriate adjustment for hidden population stratification in Bayesian analysis.

For many complex diseases, the underlying genetic model is unknown, and constructing statistical tests to be optimal to a misspecified model can cause substantial loss of power. To resolve this problem, researchers have proposed robust statistics, among which the MAX statistic has become favorable. The distribution of MAX does not follow any standard distribution such as normal distribution. We started our study from investigating the asymptotic means, variances and covariances of the trend tests, which is necessary for the derivation of the asymptotic distribution of MAX. We then provided an approximate form of the asymtotic distribution of MAX which involves only single integration, and confirmed its good fit by simulation studies. To resolve the issue of model uncertainty in Bayesian analysis, we propose a robust Bayes factor which measures the evidence of association contained in MAX statistic. Simulation studies were conducted to show that our proposed method has similar performance to the Bayesian model averaging method except that our method is overall much less sensitive to the prior of the genetic model than the Bayesian model averaging method. Finally, using both simulation studies and real data application from genome-wide association (GWA) studies, we demonstrated that we can draw comparable conclusions based on either Bayes factor or p-value when the statistical power is consistent across different studies. However, when the power differs across genetic studies, Bayes factor would provide a unified measurement that integrates both p-value and statistical power. For example, the two SNPs (rs380390 and rs1329428) associated with age-related macular degeneration (AMD) [Klein et al., 2005] have very small p-values, but their Bayes factor is not as high as many other differential SNPs due to the small sample size. Therefore it is recommended that, in order to

identify SNPs associated with complex diseases under unknown genetic model, our proposed Bayes factor should be reported in addition to the highly significant p-values of MAX statistic.

5.2 Limitations and research needs

5.2.1 PCA method and the identification of hidden subpopulations

When using the principal components analysis (PCA) method to identify the hidden subgroups based on a large panel of null SNPs, we used the 10 eigenvectors of the genetic similarity matrix corresponding to the top 10 largest eigenvalues, as described in Section 3.2.3. The number of largest eigenvalues ($L = 10$) here are the default value in Price et al. [2006]. A more rigorous approach, as is indicated in Price et al. [2006] and Patterson et al. [2006], is to choose all the eigenvalues that have significance levels below 0.01, as measured by the Tracy–Widom statistic [Tracy and Widom, 1994]. By allowing flexible number of top eigenvalues, the computational burden could be reduced when a smaller number of eigenvectors is enough to capture the genetic variation across hidden subgroups. On the other hand, it may be able to capture more information about hidden substructure if more than a fixed number of eigenvectors are needed.

An alternative approach to correct for the linear (logistic) odds ratio, as suggested by Li and Yu [2009], is to include the matrix (denoted by V in Section 3.2.3) consisting of the top eigenvectors in the logistic regression model, in addition to the dummy variables representing group membership based on the clustering result that we already included in the logistic regression model. By adding the top eigenvectors, which are continuous axes describing the genetic variation, we can adjust adequately

for the potential confounding effect of the hidden population structure. Specifically, in cases when PCA clustering does not assign the correct membership for some subjects, the inclusion of the matrix consisting of the top eigenvectors compensates the incorrect clustering to some extent.

5.2.2 Potential applications in non-human genetic association studies

This dissertation is focused on the Bayesian analysis in case-control genetic association studies in human populations. It is notable that our proposed correction for hidden population stratification can be easily extended to non-human populations, such as the plants in agricultural studies. One important implication of such application is, compared to human populations, it is even more difficult to obtain subpopulation (e.g. races) information for non-humans such as plants. Therefore the genotypic information obtained from genetic scanning become the most easily accessible resources to identify the hidden population stratification using PCA method. From there we could readily adjust for the Bayesian analysis in those studies with our proposed approach.

5.2.3 Bayes factors based on other robust statistics

Besides the maximum of three trend test (MAX), other robust statistics have been proposed under genetic model uncertainty, including the maximum efficient robust statistic (MERT) [Gastwirth, 1966] which is a linear combination of the optimal trend tests. Compared to MAX statistic, MERT has shown less power in situations such as under the dominant or recessive model [Zheng et al., 2006]. However, MERT has its advantage of maintaining asymptotic normal distribution. The same idea of constructing Bayes factor based on MAX can be applied to construct a new robust

Bayes factor measuring the evidence in MERT, which is defined as the probability of observing the MERT statistic under the alternative hypothesis over that under the null hypothesis. Because of the robustness of MERT, we would expect this new type of Bayes factor to be robust across genetic models. Compared to our proposed Bayes factor based on MAX, we would expect the Bayes factor based on MERT to be less powerful in Bayesian hypothesis testing, but simpler in computation because of the normal distribution of MERT.

References

Astle W and Balding DJ (2009) Population structure and cryptic relatedness in genetic association studies. *Stat Sci* **24**: 451–471.

Armitage P (1955) Tests for linear trends in proportions and frequencies. *Biometrics* **11**: 375–386.

Devlin B and Roeder K (1999) Genomic control for association. *Biometrics* **55**: 997–104.

Epstein MP, Allen AS, Satten GA (2007). A simple and improved correction for population stratification in casecontrol studies. *Am J Hum Genet* **80**: 921-930.

Freidlin B, Zheng G, Li Z and Gastwirth JL (2002) Trend tests for case-control studies of genetic markers: Power, sample size and robustness. *Human Heredity.* **53**: 146–152.

Gastwirth, J. L. (1966). On robust procedures. *J. Am. Statist. Assoc.* **61**: 929-948.

Gastwirth, J.L. (1985). The use of maximin efficiency robust tests in combining contingency tables and survival analysis. *J. Am. Stat. Assoc.* **80**: 380–384.

Gonzalez JR, Carrasco JL, Dudbridge F, Armengol L, Estivill X and Moreno V. (2008) Maximizing association statistics over genetic models. *Genet. Epidemiol.* **32**: 246–254.

Hoeting JA, Madigan D, Raftery AE and Volinsky CT (1999) Bayesian Model Averaging: A Tutorial. *Stat. Sci.* **14**: 382-417.

Hunter, D.J., Kraft, P., Jacobs, K.B., Cox, D.G., Yeager, N., Hankinson, S.E., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A., et al. (2007). A genome-wide

association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.* **39**: 870-874.

Jeffreys H (1961) Theory of Probability. *Oxford Univ. Press*

Johnson VE (2005) Bayes factors based on test statistics. *J. R. Statist. Soc. Ser. B* **67**: 689-701.

Johnson VE (2008) Properties of Bayes factors based on test statistics. *Scand. J. Stat.* **35**: 354-368.

Joo, J., Kwak, M., Chen, Z. & Zheng, G. (2010). Efficiency robust statistics for genetic linkage and association studies under genetic model uncertainty. *Stat. Med.* **29**: 158–180.

Kass R and Raftery A (1995) Bayes factors. *J Am Stat Assoc* **90**: 773–795.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S.,Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T. et al. (2005). Complement factor H polymorphism in aged-related macular degeneration. *Science* **308**: 385-389.

Lachin J (2000) Biostatistical methods: the assessment of relative risks. *Wiley Press*

Li CC (1955) Population Genetics. Chicago: University of Chicago Press.

Li Q and Yu K (2008) Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. *Genet Epidemiol* **32**: 215–226.

Li Q, Zheng G, Li Z and Yu K (2008) Efficient approximation of p-value of the maximum of correlated tests, with applications to genome-wide association studies. *Ann. Hum. Genet.* **72**: 397–406.

Li Z, Zhang H, Zheng G, Gastwirth JL, and Gail M (2009) Excess false positive rate caused by population stratification and disease rate heterogeneity in case-control association studies. *Comput Stat Data Anal* **53**: 1767–1781.

Miclaus K, Wolfinger R and Czika W (2009) SNP selection and multidimensional scaling to quantify population structure. *Genet Epidemiol* **33**: 488–496.

Patterson N, Price A, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet* **2**: 2074-2093.

Prentice R and Pyke R (1979) Logistic disease incidence models and case-control studies. *Biometrika* **66**: 403–411.

Price A, Patterson N, Plenge R, Weinblatt M, Shadick N and Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904–909.

Pritchard JK, Stephens M and Donnelly P (2000a) Inference on population structure using multi-locus genotype data. *Genetics* **155**: 945–959.

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000b) Association mapping in structured populations. *Am J Hum Genet* **67**: 170-181.

Qin J and Zhang B (1997) A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika* **84**: 609–618.

Risch N (2000) Searching for genetic determinants in the new millennium. *Nature* **405**: 847–856.

Risch N and Merikangas K (1996) The future of genetic studies of complex human diseases. *Science* **273**: 1516–1517.

Robins J, Breslow N and Greenland S (1986) Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics* **42**: 311–323.

Sasieni PD (1997) From genotypes to genes: Doubling the sample size. *Biometrics.* **53**: 1253–1261.

Sawcer S (2010) Bayes factors in complex genetics. *Euro. J. Hum. Genet.* **18**: 746-750.

Seaman S and Richardson S (2004) Equivalence of prospective and retrospective models in the Bayesian analysis of case-control studies. *Biometrika* **91**: 15–25.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S. et al. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**: 881-885.

Stephens M and Balding DJ (2009) Bayesian statistical methods for genetic association studies. *Nat Rev Genet* **10**: 681–690.

Tibshirani R, Walther G and Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. *J Roy Stat Soc* **B 2**: 411–423.

Tracy CA, Widom H (1994) Level-spacing distributions and the airy kernel. *Commun Math Phys* **159**: 151-174.

The Wellcome Trust Case Control Consortium. 2007. Genome-wide asssociation study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**: 661–683.

Wakefield J (2007) A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet* **81**: 208–227.

Wakefield J (2009) Bayes factors for genome-wide association studies: comparison with p-values. *Genet Epidemiol* **33**: 79–86.

Wang K and Sheffield VC (2005) A constrained-likelihood approach to marker-trait association studies. *Am. J. Hum. Genet.* **77**: 768–780.

Yamada R and Okada Y (2009) An optimal dose-effect mode trend test for SNP genotype tables. *Genet. Epidemiol.* **33**: 114–127.

Yeager, M., Orr, N., Hayes, R. B., Jacobs, K. B., Kraft, P. et al. (2007). Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* **39**: 645-649.

Yu K, Wang Z, Li Q, Wacholder S, Hunter D, Hoover R, Chanock S, and Thomas G (2008) Population substructure and control selection in genome-wide association studies. *PLoS ONE 3(7): e2551. doi:10.1371/journal.pone.0002551.*

Zang, Y., Fung, W.K. and Zheng, G. (2010). Simple algorithms to calculate asymptotic null distributions of robust tests in case-control genetic association studies in R. *J. Stat. Soft.* **33**: issue 8 (http://www.jstatsoft.org/).

Zheng G, Freidlin B, and Gastwirth JL (2006) Comparison of robust tests for genetic association using case-control studies. *IMS Lecture NotesMonograph Series 2nd Lehmann Symposium Optimality* **49**: 253-265

Zheng G and Ng HKT (2008) Genetic model selection in two-phase analysis for case-control association studies. *Biostatistics* **9**: 391–399.

Zheng G, Joo J, Zaykin D, Wu CO and Geller NG (2009a) Robust tests in genome-wide scans under incomplete linkage disequilibrium. *Stat. Sci.* **24**: 503–516.

Zheng, G., Joo, J. & Yang, Y.N. (2009b). Pearsons test, trend test, and MAX are all trend tests with different types of scores. *Ann. Hum. Genet.* **73**: 133–140.

Zheng G, Li Z, Gail MH, Gastwirth JL (2010) Impact of population substructure on trend tests for genetic case-control association studies. *Biometrics* **66**: 196–204.

Zheng, G., Yuan, A. & Jeffries, N. (2011). Hybrid Bayes factors for genome-wide association studies when a robust test is used. *Comput. Stat. Data Analy.* **55**: 2698–2711.

Zhu X, Li S, Cooper RS, Elston RC (2008) A unified association analysis approach for family and unrelated samples correcting for stratification. *Am J Hum Genet* **82**: 352–365.