ABSTRACT

OUYANG, HAOJUN. Bayesian Approach for Nonlinear Dynamic System and Genome-Wide Association Study. (Under the direction of Dr. Sujit K. Ghosh and Dr. Jung-Ying Tzeng).

Genome-wide association studies (GWAS) have been widely used to identify single-nucleotide polymorphisms (SNPs) that are responsible for diseases. A challenging aspect of this study is to resolve the multiple testing issue. We propose a new Bayesian method to measure statistical significance in these genome-wide studies based on the concept of false discovery rate (FDR). Our proposed method provides a convenient way to integrate prior knowledge obtained from external resources into current study. By controlling Bayesian FDR at a given level, the realized FDR is controlled. Our simulations show that the power can be substantially improved with precise prior information while the FDR is controlled at the desired level. When prior information is imprecise, our method can still improve the power of detecting signals and while keeping the FDR under control. The proposed Bayesian method is applied to a GWAS for schizophrenia.

Meta-analysis is another approach to utilize information from multiple sources by combining results from multiple independent studies. A major concern in carrying out meta-analysis involves the proper characterization of heterogeneity (refers to the variation in results among studies) among populations. To account for heterogeneity, the most commonly used approach is to implement a random-effects model, where the random-effects are assumed to be normally distributed with an unknown population mean and an unknown variance. We relax the normality assumption and show that a broad class of distributions can be approximated by a class of mixture distributions. The population mean and variance estimates based on the mixture density are then obtained by the EM algorithm. Our results show that the proposed method greatly improves the accuracy in estimating the overall mean effect and heterogeneity variance in various realistic cases. We illustrate our method to combine results from six association studies on C957T polymorphism in DRD2 gene with schizophrenia.

Dynamic systems defined by ordinary differential equations are important tools for modeling complicated biology system. To estimate parameters in a dynamic system for which the analytic, closed form solution is not available and involves missing or censored data, we extend the Bayesian Euler's Approximation method based on data augmentation algorithm. Our simulation study shows the method is robust in both cases. The proposed method is applied to analyze a HIV viral load dataset, which enables us to retrieve information from the censored data.

Bayesian Approach for Nonlinear Dynamic System and Genome-Wide Association Study

by
Haojun Ouyang

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

Bioinformatics

Raleigh, North Carolina

2009

APPROVED BY:

Dr. Jose M. Alonso	Dr. Jason Osborne
Dr. Sujit K. Ghosh Co-Chair of Advisory Committee	Dr. Jung-Ying Tzeng Co-Chair of Advisory Committee

DEDICATION

To my parents

BIOGRAPHY

Haojun Ouyang was born in Shanghai, the People's Republic of China. He received his B.A. in Computational Mathematics from Nanjing Normal University in 1998 and his M.S. in Industrial Economics from Nanjing University in 2001. In 2005, he earned his Master degree in Biomathematics from North Carolina State University and in the same year, he enrolled in the Ph.D program in bioinformatics at North Carolina State University. During his study at NCSU, he worked under the direction of Dr. Sujit Ghosh and Dr. Jung-Ying Tzeng. After receiving his doctoral degree, he will join Eli Lilly and Company Inc. as a statistician in the genomic science department.

ACKNOWLEDGMENTS

After four years of studying in Bioinformatics program at North Carolina State University, I am just about to finish my graduate life. There are some people whom I want to thank for at NC State for helping me to achieve my dream. First of all, I want to express my deepest thanks and appreciation to my two advisors, Dr. Sujit K. Ghosh and Dr. Jung-Ying Tzeng, who provided me wonderful guidance throughout my research and my dissertation. Professor Ghosh has always been so encouraging, understanding and showed great patience whenever I needed. I learned a lot from Professor Tzeng, not only on some challenging questions, but also on how to do research work. Without their help and guidance, I could not have accomplished it.

I want to thank Dr. Jason Osborne for serving in my committee and for being a great teacher. Professor Osborne taught me the first applied statistics course. From his class, I learned a lot of useful statistical skill in analyzing data. I would also like to thank Jose M. Alonso for serving in my committee, who taught me lots of basic knowledge in modern genetic research field. I am highly appreciated that they took the time to serve in my committee.

I also want to thank Dr. Zhao-Bang Zeng for great support and faith in me during these years. I would like to thank to my friends, Sherry Gu, Alexander Griffing, Ming Zhang, Sihui Zhao, Youfang Liu, Xiaohua Gong, Na Cai, Josh Sailsbery and Yufeng Wu for the great help and wonderful friendship in my life. Also, I want to thank all the professors, staffs and graduate students in Bioinformatics Research Center.

Finally and most importantly, I want to thank my parents, Jianhong Wang and Zixiang Ouyang, and my sister. Their great love is the source of my strength.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Genetic Association Studies

Many human diseases are considered as the results of complex effects of several or many genes interacting with the environment. The purpose of a genetic association study is to identify association between one or multiple genetic polymorphism and disease. Discovering those causal genes may provide some clues about pathophysiology and possibly lead to treatments and prevention diseases.

Most recently, with the development in sequencing technology, genetic association studies have become an important approach for locating the genes, which may affect complex traits (Risch, 2000). Generally, there are two approaches to genetic association, family-based studies and population-based studies.

1.1.1 Family-based association study

In family-based association studies, individuals from the same family in a control group can reduce the effect of population structure. The Transmission Disequilibrium Test (TDT), which was first proposed by Spielman and Ewens (1993), is commonly used to detect association in family-based design, which uses both parents and an affected child. The disadvantage of the TDT test is that it requires complete parent-child trios and only a heterozygous parent is informative with regard to allelic transmission, and is applicable only in one allele case. Some extensions to the TDT test have also been developed. For example, TDT for multiple alleles (Sham and Curtis, 1995), Pedigree Disequilibrium Test (PDT) (Martin *et al.* 2000, 2001), and TDT for quantitative traits. (Abecasis *et al.*, 2000a, b).

1.1.2 Population-based association study

Population-based association studies do not focus on families with parents and an affected child; instead, they use affected and unaffected individuals that are not related. The frequency of particular alleles presented in cases and control groups is tested for association with a disease (Risch, 2000). Unlike family-based studies,

samples in population-based studies are collected independently. There are several steps involved in carrying out population-based association studies. First, candidate genes are selected, based on the knowledge of the clinical and biological aspects of the disease of interest. Second, SNPs in candidate genes are selected and genotyped for the people in cases and control groups. Finally, single or multiple SNP analysis is performed. The simplest association test is single SNP based test. For binary traits, a number of different methods have been proposed for testing association with a disease, e.g. chi-squared. Likelihood ratio test, score test and Wald test are three commonly used tests. In large samples, all three approaches lead to chi-squared tests with one degree of freedom. With the number of markers increasing, multiple markers association tests are developed, which can be used to test association between a gene and the phenotype given those SNPs are subject to a linkage disequilibrium (LD) block within a gene.

1.1.3 GWAS and multiple testing problem

In the recent years, with the development of commercial sequencing technology, genome-wide association studies (GWAS) have become feasible. These studies genotype individuals in case and control groups for 10^5 or more SNPs, for example, Affymetrix 500K chips. Several candidate genes are identified successfully, like type I diabetes mellitus and type II diabetes mellitus (Saxena *et al.*, 2007, Scott *et al.*, 2007). Besides concerns in single SNP tests (population stratification and so on), another challenge is multiple-testing.

It becomes difficult to control overall (type I) error when many hypothesis tests are performed simultaneously. The point-wise error rate is not applicable in this situation. Failure to adjust the outcome may lead to a huge false positive rate. However, traditional Bonferroni correction, which controls the family-wise error rate, is too conservative for the purpose of GWAS. Alternatively, Benjamini and Hochberg (1995) proposed false discovery rate (FDR) to control for the overall type I error. There are many extensions to FDR. For example, pFDR (Storey, 2001) and local FDR (Efron,

2004). Please refer to page 32-33 for more a detailed review.

1.1.4 Bayesian method for multiple testing problem

Many experiments are performed for common diseases from various aspects, and integrating information from other sources into a study may improve the power of the study. Roeder *et al.*, (2006) combine linkage information to improve power of GWAS. Quantitative and qualitative information can arise from other studies. The Bayesian approach provides a formal way to incorporate the available knowledge as a prior distribution into current analysis. The idea is to use observation data to update knowledge. Scott and Berger (2006) explored a Bayesian approach to a multiple testing problem by assigning different priors.

1.2 Meta-analysis for association study

The other approach to combine results from multiple studies is based on the so-called meta-analysis. Meta-analysis is a statistical tool to combine results from multiple independent studies, which has become an important method for resolving inconsistency in genetic association studies. During the recent decades, meta-analysis has become a powerful tool in medical research and clinical studies. In health-related research, meta-analysis has increased to 400 publications per year since the year 2000 (Lee *et al.*, 2001).

For complex diseases, where multiple loci may contribute to the common disease, each variant is likely to have small effect, e.g. odds ratio in the median level (around 1.3) is probably a typical value. Under this scenario, Zondervan and Cardon (2004) show that in the best cases, where the marker and disease allele frequencies match, sample sizes of 2000-10000 cases and controls, respectively are required to obtain 80% power. Meta-analysis provides a way to aggregate samples from individual studies to improve power.

Inconsistency is the major issue encountered in performing a meta-analysis. In 2002, Hirschhorn pointed out that in 600 positive associations between genetic variants and disease, 166 associations were repeated three or more times, and only six associations were consistent. The inconsistency is due to the so called "heterogeneity", which refers to the variation in results among studies. Heterogeneity among studies is common. Ioannidis *et al.* (2001) conducted a meta-analysis of 370 studies addressing 36 genetic associations. They found that significant between-study heterogeneity is very common. There are many reasons why association studies between genes and disease are not consistent.

1.2.1 Investigating heterogeneity

A forest plot is often used to visualize the heterogeneity between studies, such as in Figures 3.13 and 3.14. The left column lists all the names of studies. The right column is a plot of the odds ratio for each study. Each horizontal line is the confidence interval for the estimate of the odds ratio from each study. The size of each square is proportional to the study's weight in the meta-analysis. A vertical line representing no effect is also plotted. If the confidence intervals for individual studies overlap with this line, it means that at the given level of confidence their effect sizes do not differ significantly from common effect. Lohmueller *et al.* (2003) pointed out that heterogeneity should be carefully examined by either clinical or statistical methods, and that the magnitude of the heterogeneity should be quantified, wherever possible. Several statistical methods have been developed to test heterogeneity. Cochran (1954) proposed a Q statistic, which has a chi-squared distribution with degree of freedom k-1, where k is the number of studies. The chi-squared test is not powerful when the number of studies is small. In addition, the Q statistic may be significant when k is large, even when the studies are not heterogeneous. Higgins and Thompson (2002) proposed the index of heterogeneity test statistic, I^2, which measures the proportion of deviation from its expectation. The value of I^2 indicates the degree of heterogeneity.

1.2.2 Statistical methods of meta-analysis

To analyze results from multiple studies, two models are often used in meta-analysis, one is the fixed-effects model, and the other is the random-effects model. The fixed-effects model assumes all studies are estimating the same underlying overall effect. The differences across studies are due to within study variation. The methods for fixed-effects model are relatively standard, for example the Mantel-Haenszel method and Peto method for multiple 2 by 2 tables and standard weighted regression. The random-effects model is considered more reasonable, since it incorporates between-study variability into the overall estimate of the treatment effect. It assumes that the underlying effect that each study attempts to estimate is from a population distribution. The popular DerSimonian and Laird (1986) approach to random-effects meta-analysis uses a simple moment-based estimate of the among-study variance. The simple moment estimator is a special case of a more general moment-based approach (Higgins and Thompson, 2002). Several other statistical methods are developed under the random-effects model from frequentist and Bayesian aspects. Sidik *et al.* (2007) compared seven estimators including DerSimonian and Laird approach. One limitation of DerSimonian and Laird approach is that the estimator of heterogeneity is biased. We can observe from our simulation studies that the bias of the estimator of heterogeneity increases greatly when the true heterogeneity is large. Our proposed method relaxed the normality assumption. The estimates for overall effect and heterogeneity are obtained by the EM algorithm. In our simulation studies, the bias of estimate of heterogeneity by our proposed method is close to the true value in each case.

1.3 Dynamic system involving missing and censored data

Computational biology has become an important research area in which various complicated biological phenomena are studied through well-defined mathematical models. These models enable scientists to understand the underlying mechanism of biological phenomena. Dynamic system provides a flexible class of mathematical models, which is defined by a set of ordinary differential equations with some parameters. Recently, scientists use ODEs in modeling genetic data. For example, genetic regulatory network consists of set of genes, proteins, small molecules, and their mutual regulatory interactions. Dynamics of large and complex genetic regulatory processes are hard to understand by intuitive approaches alone. Mathematical methods for modeling and simulation are needed. ODEs have been used to model variety genetic regulatory networks, including lambda phage infection of E. coli (McAdams and Shapiro, 1998) and the developmental cycle of bacteriophage T7 (Endy *et al.*, 2000). More recently, Lin *et al.* (2005) modeled cis-regulatory circuits and gene expression prediction via dynamic system and determined which factors may play the most important role. Elowitz and Leibler (2000) proposed a new way to model genetic regulatory networks.

In my project, we use a classic ODE system to study HIV viral load data. Ho *et al.* (1995) used ODEs to analyze the dynamics of HIV viral load measurements on AIDS patients. Perelson (1996) proposed a mathematical model to analyze a detailed set of human immunodeficiency viral load data, which has been widely used. The model adopts three ODEs to reveal the relationship of rate of change among productively infected cells (T^*) and concentrations of viral particles in plasma (V_I and V_{NI}). The major interest for scientists is to estimate these parameters, which may classify different patients and choose the best treatment strategy for different patients. Also through those parameters, scientists may better understand the pathophysiology of the disease. However, it is difficult to estimate these parameters without solving the ODEs analytically. With complicated dynamic systems, the analytic closed form so-

lution is often not available. In general, there are two global approaches to solve the problem, maximum likelihood method, and Bayesian method. In this dissertation, we take the Bayesian perspective to solve this problem.

Missing data and censored data are very common in practice. For example, in genetic association study, phenotype information at one or more loci may not be available, or haplotype frequencies for multiple locus systems are unknown due to incomplete information (Chiano and Clayton, 1998). Not properly accounting for missing data may lead to biased estimates or reduce accuracy in prediction. By some statistical methods, for example EM algorithm, Gibbs sampling, data augmentation, we may be able to recover the information from the missing part. In biological experiments, limitation of measuring equipment, which provides limited measurement is one of the reasons for censored data. For example, in the HIV study, the RNA array used in the experiment can only provide measurement of as few as 100 copies of RNA. In analyzing the censored data, scientists either delete the censored observations or replace the censored data by the limit value that the equipment can provide. Obviously, such ad-hoc methods will lead to biased estimates.

Chapter 2

Bayesian approach for multiple testing on partitioned data for genome-wide association study

2.1 Abstract

Genome-wide association studies (GWAS) have become possible with the development of modern technologies. With GWAS data, one challenge is the multiple testing problem, as hundreds of thousands of tests are performed simultaneously for a null hypothesis. We propose a new method to measure statistical significance in GWAS based on the concept of false discovery rate (FDR). One advantage of the proposed method is that prior knowledge obtained from other resources, such as linkage studies, microarray analysis, biological pathways and others, can be readily integrated into the testing framework to up-weight or down-weight genomic regions to improve the power. The data are partitioned based on prior knowledge, and a Bayesian rule is applied to compute the posterior probabilities of null hypothesis and estimate the Bayesian positive FDR. By controlling Bayesian FDR at level α, the realized FDR is controlled. Our simulations show that the power can be substantially improved with precise prior information while the FDR is controlled at the desired level. When prior information is mis-specified, our method can still improve the power of detecting signals and keeps the FDR under control. The proposed procedure is directly applicable to other high-throughput studies such as microarray studies. We applied our proposed Bayesian method to a GWAS for schizophrenia.

2.2 Introduction

With the development of inexpensive genotyping technology and the completion of the International Haplotype Mapping (HapMap) Project (Altshuler *et al.* 2005), genome-wide association studies (GWAS) are now widely used to study the genetic basis of complex traits. GWAS investigates the genetic variation across the whole genome, with an aim to identify possible genes associations with some specific traits. Several results have been published (Ozaki *et al.*, 2002, Maraganore *et al.*, 2005, Yeager em et al., 2007) and more studies are ongoing. In GWAS, hundreds of thousands of markers are evaluated and one immediate issue is the multiple testing problem. That is, scientists have to control the number of incorrect significant results obtained while retaining the optimal power for testing a large number of hypothesis at the same time.

To balance between power and type I errors with multiple tests, we need an appropriate quantity to measure the overall type I error rate. One is the family-wise error rate (FWER), which is the probability of at least one hypothesis being falsely rejected. Several methods have been developed to control FWER at level α, such as the Bonferroni correction. However, as controlling FWER aims to prevent any type I error, the FWER procedures tend to results in stringent conservative error rate and little discoveries. To overcome this issue, Lehmann and Romano (2005) proposed *k-FWER*, which controls the probability of k or more rejections at some designated level.

Alternatively, Benjamini and Hochberg (1995) proposed the false discovery rate (FDR), which is the expected value of the ratio of false rejections over total rejections. They also developed a step-up procedure, which compares the ordered p-values with a sequence of critical values to control FDR. Controlling for FDR provides a better balance between power and false discovery, especially when scientists are more interested in looking for a subset of potentially true positive tests for further study. In addition, when all null hypotheses are true, controlling FDR is equivalent to controlling FWER. Other methods have also been developed to control FDR or its alternative

forms. Storey (2002, 2003) proposed the positive FDR (pFDR), which is FDR given the number of rejections is positive. To controlling pFDR, they proposed the q-value algorithm, which is to estimate pFDR for each test and compare with the confidence level α. Assume that the true null hypotheses are from a Bernoulli trial with probability π. Storey showed that the B-H procedure is a special case of the q-value procedure where π is estimated by 1. Some other methods consider to reorder the p-values by introducing other information. Genovese *et al.* (2006) proposed a weighted FDR procedure to incorporate prior information. Roeder *et al.* (2006) showed how linkage information could be incorporated into the weighted FDR framework. A weight is calculated based on the log odds ratio (LOD score) and the weighted "P" value , which is the original p-value divided by the weights, is used to rank the statistic evidence. They show that the procedure can improve the power if the prior information is truly helpful to current study, and suffers only little power loss if the prior information is uninformative or misleading. Roeder *et al.* (2007) used external information to group data and to assign weight for each group, then applied an extended weighted FDR method to the grouped data.

Bayesian approaches are also widely considered to deal with multiple testing problem. Bayesian inference has some advantages in dealing with complex models and incorporates prior knowledge into model. Efron *et al.* (2001) and Efron and Tibshirani (2002) proposed local FDR and applied an empirical Bayes approach for FDR analysis. Muller *et al.* (2006) devised various kinds of loss functions and solved for the corresponding optimal rules. Newton *et al.* (2004) used prior information within the FDR framework to compute a "Bayesian FDR", which is the sum of posterior probabilities of null of the rejections divided by the number of the rejections. Do *et al.* (2005) proposed a fully Bayesian framework with a Dirichlet process mixture model for differential gene expressions. Later, Lewinger *et al.* (2007) used a Bayesian hierarchical model to analyze genome-wide association studies.

Scott and Berger (2003) explored a general Bayesian approach to multiple testing problem. They pointed out that the effect of prior distribution had little effect on the posterior probability when the number of null hypotheses (i.e., noise) is huge. If one

can reduce the ratio of noise to signal, the prior information may have more influence on the results (i.e., the posterior probabilities of being a null hypothesis).

This observation motivated our method. In Section 2.3, we briefly introduce the multiple testing problem and review several error rates, including FWER, FDR, pFDR and Bayesian FDR . In Section 2.4, we introduce a Bayesian approach to the multiple testing problem, and describe the proposed Bayesian method for grouped data, which using external information. In Section 2.5, we evaluate the proposed procedure, using simulations and compare the results with the q-value algorithm. In Section 2.6, we show a real data example by applying the proposed method to the CATIE GWAS study for schizophrenia. In Section 2.7, we discuss possible extensions and limitations of our methods.

2.3 Multiple testing and error rates

Consider the m hypotheses, $H_1, \ldots, H_m$, to be tested simultaneously. Suppose m_0 of those hypotheses are true and R of the m hypotheses are rejected. Also let $Y_1, \ldots, Y_m$ be the m testing data of the m hypotheses, such as the p-values or the test statistics. We can use the following table to describe different error rates.

[Table 2.1 is here.]

The FWER is defined as the probability of at least one hypothesis being falsely rejected, i.e., $FWER = P(V \geq 1)$. The Bonferroni procedure is to set the confidence level at α/m to control the overall error rate of α. When m is very large, this procedure is very conservative and often nothing can be discovered. Several adjustment methods are proposed, Simes (1986) introduced a particular sequence of critical values $\alpha_i = i\alpha/m$, to compare with each p-value.

The false discovery rate (FDR), proposed by Benjamini and Hochberg (1995) defined as the expected ratio of false rejections to total rejections, $FDR = E(V/R)$. If no rejections, define V/R to be zero. If all hypotheses are true, then controlling FDR is equivalent to controlling the FWER. Many desirable properties are discussed

in Benjamini's paper. They also proposed a step-up procedure to control FDR. The procedure can be described as follows. Let $k = \max\{i : Y_{(i)} \leq \alpha_i\}$, where $\alpha_i = i\alpha/m$ and $Y_{(i)}$ are ordered p-values with $Y_{(1)} \leq \ldots \leq Y_{(m)}$. Reject all hypothesis if $j \leq k$. Benjamini and Yekutieli and others have shown that in the independent case, with the B-H procedure, the FDR is exactly equal to $m_0\alpha/m \leq \alpha$. Recently, some modified sequential FDR controlling procedures are developed. One of the idea is to use new critical values $\alpha_i = i\alpha/(m\hat{\pi})$, where $\hat{\pi}$ is the estimate of m_0/m.

Nowadays, scientists are more interested in looking for potential true signals for the next step of study. Storey argued it is meaningful to talk about false discoveries only when there are some discoveries. Therefore, we should look at the expected proportion of false discoveries conditional on the fact that there have been some discoveries. Storey (2002) proposed the positive false discovery rate (pFDR) as $pFDR = E(V/R|R > 0)$. Storey (2002) expresses the pFDR based on p-value rejection. All rejections are made when p-values are in the interval [0,λ], for some $\lambda > 0$. The expression for pFDR can be rewritten as

$$pFDR(\lambda) = \frac{\pi_0 Pr(P \leq \lambda | H = 0)}{Pr(P \leq \lambda)}$$

where π_0 is the prior probability that a null hypothesis is true and P is the p-value obtained from any test. They proposed the q-value algorithm to control pFDR, which rejects all null hypotheses with p-value less than or equal to the critical value $\hat{\gamma}$, where $\hat{\gamma} = max\{\gamma : \hat{pFDR}(\gamma) \leq \alpha\}$.

The Bayesian FDR (Newton, 2004) uses the posterior probability of hypothesis being true (w_i) and the rejection set J. It can be defined as $\hat{FDR}_{Bayes}(J) = \sum_{i \in J} w_i/|J|$. We control the Bayesian FDR to obtain the goal of controlling FDR.

2.4 Bayesian approach to the multiple testing problem

In this section, we first briefly explain the Bayesian hierarchical model to the multiple testing problem. Then we introduce the modified Bayesian approach.

2.4.1 A general Bayesian approach

Suppose there are m independent tests. Let $\boldsymbol{\gamma}$ be a vector denoting the true status of the hypotheses, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_m)$. If the null hypothesis is true for the ith test, $\gamma_i = 0$. If the alternative hypothesis is true for the ith test, $\gamma_i = 1$. For test i, we observe the data Y_i (e.g. the test statistic or the p-value), and we assume the density of Y_i under the null hypothesis is f_0 and under alternative hypothesis is f_1. We model γ_i, $i = 1, \ldots, m$, as m independent Bernoulli trials with success probability $Pr(\gamma_i = 1) = 1 - p$. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$ and $\boldsymbol{\theta}_i = (\theta_{i0}, \theta_{i1})$ to be the vector of parameters. Define density function $f_j(Y_i|\theta_i) = f(Y_i|\gamma_i = j, \theta_{ij})$, for $j = 0$ or 1. Then Y_i can be regarded as from the following two-component mixture model:

$$f(Y_i|\theta_i) = pf_0(Y_i|\theta_{i0}) + (1-p)f_1(Y_i|\theta_{i1}) \tag{2.1}$$

For multiple testing problem, one major interest is the posterior probability of the null hypothesis being true, $w_i = Pr(\gamma_i = 0|\boldsymbol{Y})$ (Scott and Berger, 2003). To estimate this quantity, a hierarchical mixture model with parameter $\boldsymbol{\theta}$ is shown below:

$$Y_i|\theta_i \overset{i.i.d}{\sim} pf_0(Y_i|\theta_{i0}) + (1-p)f_1(Y_i|\theta_{i1})$$

$$p = Pr(\gamma_i = 0), \gamma_i|p \sim Ber(1-p)$$

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}), p \sim \pi(p)$$

The marginal likelihood function of Y, with parameters $\boldsymbol{\theta}$, can be written as

$$L(\boldsymbol{Y}|\boldsymbol{\theta}) = \prod_{i=1}^{m} \{pf_0(Y_i|\theta_{i0}) + (1-p)f_1(Y_i|\theta_{i1})\} \tag{2.2}$$

For simplicity, we use f_j to represent $f_j(Y_i|\theta_{ij})$ and omit the parameters θ_{ij}, for $j = 0, 1$. Without loss of generality, here we assume that f_0 and f_1 are normal densities. Specifically, we assume the null distribution, f_0, is $N(0, 1)$ and the alternative distribution, f_1, is $N(\mu, \sigma^2)$. In this case, $\theta_{i0} = (0, 1)$ and $\theta_{i1} = (\mu_i, \sigma_i^2)$ for $i = 1, \ldots, m$. Without loss of generality, we assume $\sigma_i^2 = \sigma^2$ for all i. So Let $\boldsymbol{\Theta} = (\boldsymbol{\mu}, \sigma^2, p) = (\mu_1, \ldots, \mu_m, \sigma^2, p)$. In other cases, f_0 can be a central χ^2 and f_1 a noncentral χ^2 distribution (Lewinger, 2007). The corresponding likelihood function of all Y_i's can be written as (Scott and Berger, 2003)

$$L(\boldsymbol{Y}|\boldsymbol{\Theta}) = \prod_{i=1}^{m} \left\{ \frac{p}{\sqrt{2\pi}} \exp \frac{-Y_i^2}{2} + \frac{1-p}{\sqrt{2\pi\sigma^2}} \exp \frac{-(Y_i - \mu_i)^2}{2\sigma^2} \right\} \tag{2.3}$$

We can choose the following priors for $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$, σ^2 and p:

$\mu_i \sim N(\beta_0, V)$ for $i = 1, \ldots, m$ and $(V, \sigma^2) \sim \pi(V, \sigma^2)$

$p \sim Beta(\alpha, 1)$

For simplicity, we set $\beta_0 = 0$. For the choice of $\pi(\sigma^2, V)$, we use $\pi(\sigma^2, V) = (V + \sigma^2)^{-2}$ (Scott and Berger, 2003). To specify α in the Beta distribution for the prior of p, we use the prior median to make an "initial guess" for p (Berger, 2003), $\alpha = \frac{\log(0.5)}{\log \hat{p}} - 1$. The corresponding posterior density of $(\boldsymbol{\mu}, \sigma^2, \boldsymbol{\gamma}, V, p)$ is

$$\pi(\boldsymbol{\mu}, \sigma^2, \boldsymbol{\gamma}, V, p|\boldsymbol{y}) \propto L(\boldsymbol{y}|\boldsymbol{\mu}, \sigma^2)[\prod_{i=1}^{m} N(\mu_i|\mu_0, \tau^2)]\pi(\boldsymbol{\gamma}|p)\pi(V, \sigma^2)\pi(p)$$

The marginal posterior distribution of (V, σ^2, p) is given by

$$\begin{aligned} \pi(V, \sigma^2, p|\boldsymbol{y}) &= m(\boldsymbol{y})^{-1} \int f(\boldsymbol{y}|\boldsymbol{\mu}, \sigma^2, p)\pi(\boldsymbol{\mu}, \sigma^2, p)d\boldsymbol{\mu} \\ &= m(\boldsymbol{y})^{-1} \prod_{j=1}^{m} \left[\frac{p}{\sqrt{2\pi}} \exp\left(\frac{-y_j^2}{2}\right) \right. \\ &+ \left. \frac{1-p}{\sqrt{2\pi(\sigma^2+V)}} \exp\left(\frac{-y_j^2}{2(\sigma^2+V)}\right) \right] \pi(V, \sigma^2)\pi(p) \end{aligned} \tag{2.4}$$

where $m(\boldsymbol{y})$ is the marginal distribution of $\boldsymbol{y}$. Berger (2003) has shown that $m(\boldsymbol{y})$ is finite and the posterior is proper. The posterior probability of $\gamma_i = 0$ can be

calculated in the following way.

$$\begin{aligned} w_i &= Pr(\gamma_i = 0|\boldsymbol{y}) \\ &= \frac{f(y_i|\gamma_i = 0)p}{f(y_i|\gamma_i = 0)p + (1-p)m_1(y_i)} \\ &= \int_0^1 \int_0^\infty \int_0^\infty \left[1 + \frac{1-p}{p}\sqrt{\frac{\sigma^2}{\sigma^2+V}}\exp\left(\frac{y_i V}{2\sigma^2(\sigma^2+V)}\right)\right]^{-1} \\ &\quad \pi(V, \sigma^2, p|\boldsymbol{y})dV d\sigma^2 dp \end{aligned}$$

$m_1(y_i)$ is the marginal distribution on the alternative. Importance sampling is used to sample $(V, \sigma^2, p\ )$ from (2.4) and compute w_i, which is the posterior probability of $\gamma_i = 0$. Scott and Berger(2003) explored the effect of different prior distributions on w_i and shows that when the ratio of true null hypotheses (noises) versus true alternative hypotheses (signals) increases, the posterior probability of null is true goes to 1 for different prior distributions. For example, consider a relative strong prior on $1 - p$, with $\pi(p) \sim Beta(1, 1)$, that says the test is less likely to be null. When there are 10 noise observations, the posterior probability of $\gamma = 0$ given observation $y = -2.62$ is 0.29. When the number of noise observations increases to 5000, this posterior probability becomes 0.89. In other words, the prior information has little effect on the posterior probability of being null when the number of noise observations increases. The prior information may effect the posterior probability of null being true, when the ratio of noises to the signals is small.

2.4.2 Partition data by external information

Many experiments have been done to study same interest area from different aspects. If we can incorporate these available data sources as much as possible, we might improve power to verify our hypothesis or get better estimates for the quantities we are interested in. Roeder *et al.* (2006) incorporated linkage information to up-weight or down-weight different regions and improved power in simulated GWAS. We will utilize external information in a more flexible way, so that it can be used to facilitate current study.

The Bayesian approach is a natural way to incorporate information of other sources as prior knowledge, which can be used in the data analysis. In the previous section, we show that when the ratio of noises to signals is large, the prior information has little effect on the posterior probability of being null. If we want to borrow the strength of prior information, we need to constrain the ratio of noise to signal to a certain range so that the effect of the prior will not vanish by noise.

One way to enhance the strength of the prior information is to partition data into different groups based on external information. In those "promising" groups, the ratios of noises to signals are in a moderate range. For example, in disease study, scientists may have pathway information about the disease studied. This prior knowledge can be used to group those single nucleotide polymorphisms (SNPs) within the genes, which are involved in the pathway into one set and the remaining SNPs into another group. By assign different prior distributions, we can up-weight or down-weight markers in each group. The partition can also be implemented with quantitative external information, such as the LOD scores from linkage studies.

[Figure 2.1 here.]

To illustrate the idea, we consider to partition data into two subgroups A and B as shown in figure 2.1. Given $y_i \in A(or B)$, the probability of null hypothesis is true is $p_A(p_B)$. Then for $Y_i \in A$, it follows density function $p_A f_0 + (1 - p_A) f_1$. For $Y_i \in B$, it follows density function $p_B f_0 + (1 - p_B) f_1$. We assume both sets share the same f_0 and f_1.

The likelihood function of the partitioned data can be written as:

$$L(\boldsymbol{Y}|p_A, p_B) = \prod_{Y_i \in A} (p_A f_0 + (1 - p_A) f_1) \prod_{Y_i \in B} (p_B f_0 + (1 - p_B) f_1)$$

A hierarchical model for the partitioned data can be written as:

$Y_i \sim p_A f_0 + (1 - p_A) f_1$, if $Y_i \in A$

$Y_i \sim p_B f_0 + (1 - p_B) f_1$, if $Y_i \in B$

$p_A \sim Beta(\alpha_A, 1)$ and $p_B \sim Beta(\alpha_B, 1)$

Suppose group A is the "promising" set, which means it more likely contains signal. So p_A is probably less than p_B. Similarly, to specify α_A (or α_B), an "initial guess" of $\hat{p}_A$ (or $\hat{p}_B$) is needed, with $\hat{p}_A \leq \hat{p}_B$, we can choose $\alpha_A = \frac{\log(0.5)}{\log \hat{p}_A} - 1$ and $\alpha_B = \frac{\log(0.5)}{\log \hat{p}_B} - 1$. To be general, let $\Gamma = A$ (or B). Then the posterior probability of $\gamma_i = 0$ can be obtained by

$$\begin{aligned} w_i &= Pr(\gamma_i = 0|\boldsymbol{y}) \\ &= \int_0^1 \int_0^1 \int_0^\infty \int_0^\infty \left[1 + \frac{1-p_\Gamma}{p_\Gamma}\sqrt{\frac{\sigma^2}{\sigma^2+V}}\exp\left(\frac{y_i^2 V}{2\sigma^2(\sigma^2+V)}\right)\right]^{-1} \\ &\quad \pi(V, \sigma^2, p_A, p_B|\boldsymbol{y}) dV d\sigma^2 dp \end{aligned} \tag{2.5}$$

The posterior distribution of (V, σ^2, p_A, p_B) is

$$\begin{aligned} \pi(V, \sigma^2, p_A, p_B|\boldsymbol{Y}) \quad \propto \quad & f(\boldsymbol{y}|\boldsymbol{\mu}, \sigma^2, p_A, p_B)[\prod_{j=1}^{m} N(\mu_j|\mu_0, \tau^2)]\pi(\boldsymbol{\gamma}|p_A, p_B)\pi(V, \sigma^2) \\ & \pi(p_A)\pi(p_B) \end{aligned}$$

After computing $\{w_i\}$, we order them from smallest to largest, $w_{(1)} \leq \ldots \leq w_{(m)}$. To estimate the expected number of false discoveries given rejection of first n tests , we use the posterior probabilities of nulls being true (Newton *et al.*, 2004; Broët *et al.*, 2004):

$$C(n) = \sum_{i=1}^{n} w_i \tag{2.6}$$

Let $K = \max\{n : C(n)/n \leq \alpha\}$. All hypotheses will be rejected if $i \leq K$. In summary, we proposed the following Bayesian approach for partitioned data:

1. Partition data based on external information.

2. Sample parameters from their corresponding posterior density to compute w_{iA} and w_{iB} as in (2.5), the posterior probability of each test in group A or B being null, and order them from the smallest to the largest.

3. Use (2.6) to estimate the expected number of false discoveries, and find the largest index j such that $C(j)/j \leq \alpha$. The first j tests will be rejected and claimed to be signals.

2.5 Simulation Study

We conduct simulation studies to evaluate the validity and performance of our proposed method. First, we verify FDR is controlled. We simulate a null dataset with $m = 1000$ and 5000 (i.e., all test data are generated from the null distribution). Specifically, we generate $\{Y_i\}$s are from $N(0,1)$, for $i = 1, \ldots, m$. For the partition, we randomly choose 100 or 500 tests as subgroup A and rest of tests as subgroup B. We set $\alpha_A = 1$ and $\alpha_B = 10$. So $p_A \sim Beta(1,1)$ and $p_B \sim Beta(10,1)$. We control the Bayesian FDR at level 0.05. The results are shown as following:

[Table 2.2 is here.]

In the above table, the realized FDR in each case is under controlled. It seems that the Bayesian FDR is more stringent than FDR.

Next, we generate $m = 5000$ independent test statistics $Y = (Y_1, \cdots, Y_{5000})$. There are 10 signals and 4990 noises in the dataset. The signals are generated from $N(3.5,1)$ and the noises are from $N(0,1)$. The data is partitioned into two groups $\{A,B\}$ with $|A| < |B|$ (where $|A|$ is the number of tests in group A). Generally, we require $|A| > 20$ to avoid extreme case in estimating parameters. In order to investigate the effect of prior information, we consider the combination of different number of tests in A (30, 60, 110 or 510) and different number of signals in A (10, 8, 4 or 0). We choose the prior distributions of p_A and p_B as $p_A \sim Unif(0,1)$ and $p_B \sim Beta(10,1)$, which means subgroup A is more likely to be from the alternative than subgroup B. The results are shown in the following table.

[Table 2.3 is here.]

In each study case, the realized FDR is controlled (less than 0.05). When the ratio of signals to noises in subgroup A increases, we gain the power to detect signals. In the cases where subgroup A contains many signals (for example the number of signals is 10 or 8), we are more powerful than the q-value procedure, which the corresponding power is 0.33 with FDR at level 0.044. Applying the q-value algorithm on both group A and B separately and adding all significant discoveries together is not acceptable,

since the FDR is above the desired level (0.14). We also plot the FDR and the power of our method compared with q-value algorithm (figure 2.2).

[Figure 2.2 here.]

The left panel of figure 2.2 shows that our proposed method and q-value procedure successfully control FDR below 0.05. This suggests that controlling Bayesian FDR is more conservative than FDR. The right panel of figure 2.2 shows that the power of our proposed method decreases as the size of group A increases when the number of signals in group A fixed. In addition, the power of our proposed method decreases as the number of signals in group A decreasing when the size of group A fixed. The above simulation studies explain several properties of our proposed method. First, If the external information is precise and the ratio of noises to signals is small, we can improve the power substantially. Second, If the external information is imprecise, we prefer to cover most of signals by increasing the number of tests in the "promising" group. For example, the performance of covering only 4 out of 10 signals is worse than the cases of covering 8 out of 10 signals under different ratios of signals and noises. Third, FDR is under controlled at level 0.05 in all cases. Finally, if the external information does not contain any signals , q-value algorithm is better than our proposed method.

Based on the simulation results, we suggest to apply the q-value algorithm to data first. If we do not satisfy the result and external information is available, we may partition the data based on external information and apply our proposed method.

2.6 Application in GWAS for schizophrenia

In this section, we apply our proposed method to a genome-wide association study for schizophrenia, in the CATIE study. Schizophrenia is a common disorder disease (0.4% - 0.6% population are affected) caused by the interaction of multiple genetic and environmental factors. The heritability of schizophrenia is relatively high, about 80%. Non genetic factors also affect the disease risk. Currently, about 150 genetic

association studies are published each year. No significant causal gene(s) or genetic variant has been reported yet, but not with as much confidence as for other complex disease, such as APOE in Alzheimer's disease or CFH in macular degeneration.

The data we used, is deposited with the controlled-access repository of the NIMH in June 2007 (http://www.nimhgenetics.org (accessed 28 June 2007)). Affymetix 500K chipset were used for individual genotyping. There are 492,900 SNPs that were genotyped in 738 participants with the disease and 733 group-matched controls from a US population-based sample (Sullivan *et al.* 2008). No high-priority genomic region has been identified by genome-wide linkage studies. Logistic regression was applied for association with SCZ under a log-additive mode of inheritance (one degree of freedom; each SNP coded as the number of minor alleles). All regression models contain the first seven principal components as covariates to adjust for stratification and any artifacts detectable in the GWAS data (details are in Sullivan 2008 supplement file). The logistic model can be written as:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 g + \sum_{i=1}^{C} \eta_i c_i$$

where π denotes the probability of diagnostic status (coded as 0 for control and 1 for case). g stands for the numerically coded genotype of the test SNP as the number of minor alleles (0, 1 or 2). c_i is the ith covariate among a total of C covariates, and η_i denotes the effect coefficient for the covariate.

To test the hypothesis $H_0 : \beta_1 = 0$ for a given SNP, a Wald statistic was used. Each regression analysis was carried out at each locus in the GWAS data. All logistic regression analyses were conducted in PLINK.

In our procedure, we use the value of the Wald test for each locus as observation Y_i, for $i = 1, \ldots, 492,900$. The histograms of the SCZ data is shown in figure 2.3.

[Figure 2.3 here.]

For the prior information, we search the published literature. 26 candidate genes are chosen as the prior information. They are AKT1, CSF2RA, IL3RA, PRODH, RGS4, ZDHHC8, COMT, DAOA, DISC1, DRD3, DTNBP1, HTR2A, NRG1, PLXNA2,

SLC5A4, APOE, DRD1, DRD2, DRD4, GABRB2, GRIN2B, HP, IL1B, MTHFR, TP53, TPH1. We use Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway) to look for set of SNPs located within those genes and classify them in the "promising" set. We use *Genome Browser* to identify SNPs that are located in those candidate genes, and 799 SNPs are identified in the GWAS data. We group those "promising" SNPs into group A (coded as 1) and the rest into group B (coded as 2). The range of P-value at these loci is between $9.1e-4$ and 0.9979.

We apply our proposed method for the partitioned data and compute the posterior probability of being null for each SNP by *WinBUGS* and the *R* package *R2WinBUGS* with an initial burn-in of 3000 iterations followed by 2000 post-burn-in samples. By controlling Bayesian FDR at level 0.1, no statistically significant SNPs are found. We list top five SNPs, which have largest posterior probabilities of being alternative, from group A and group B.

[Table 2.4 is here.]

In the list of top five SNPs from group B, rs16977195 is located in AGBL1 and rs151222 is located in ACSM1 and others are not located within any genes we know. The top five SNPs with largest posterior probabilities of being alternative ($\text{Prob}(\gamma_i = 1|\boldsymbol{Y})$) in group A, are located within gene NRG1 and gene GRIN2B. We plot the posterior probability of being alternative at each SNP as following:

[Figure 2.4 is here.]

Several reasons may explain why our proposed method did not find any statistically significant SNPs. First, the sample size of GWAS is relative small. As Zondervan and Cardon (2004) show that in the best cases, sample sizes of 2000-10000 cases and controls are required to obtain 80% power in a single study. However, in this study, the number of individuals in case and control groups together is less than 1500. Second, several genes may contribute to the disease with small effect for each gene, which means single SNP association test may not be powerful to detect "weak" signals. Third, it may also be possible that prior knowledge about those 26 "promising" genes

is imprecise, which means the true causal genes are not located within those candidate genes, or the ratio of noises to signals is too large. In these cases, our proposed method can not improve the power.

2.7 Discussion

The Bayesian approach for partitioned data provides a formal way to incorporate external information into the study. When the external information is precise (the ratio of noise to signal is in a relative small range), the method may improve the power to detect signals with FDR under control. If the information is imprecise, the performance is adequate but not better than the q-value method. Therefore, we recommend to use the q-value algorithm first. If the result is not satisfied, we may use the proposed method by incorporating the external information into the prior. In our method, the external information is relatively flexible. It can be a set of genes in some pathway related to the disease, or results from linkage studies. If those information are precise and the ratio of noises to signals is small, the gain in power is obvious.

To better use external quantitative information, like the LOD score, we may link probability of null for each test (p_i, $i = 1, \ldots, m$) as a function of the external information, which based on the assumption that regions with higher LOD scores are more likely associated with higher probabilities of being signals. We will explore this idea further in the future.

Appendix

To sample from posterior density of $\pi(V, \sigma^2, p_A, p_B|\boldsymbol{y})$, we first transform the parameters by

$$\begin{aligned}
\xi &= \log(V) \rightarrow dV = e^{\xi} d\xi \\
\eta &= \log(\sigma^2) \rightarrow d\sigma^2 = e^{\eta} d\eta \\
\lambda_A &= \log \frac{p_A}{1-p_A} \rightarrow dp_A = e^{\lambda_A}(1+e^{\lambda_A})^{-2} d\lambda_A \\
\lambda_B &= \log \frac{p_B}{1-p_B} \rightarrow dp_B = e^{\lambda_B}(1+e^{\lambda_B})^{-2} d\lambda_B
\end{aligned}$$

Then the integration domains become $(-\infty, \infty)$.
The transformation of $\pi(V, \sigma^2, p_A, p_B|\boldsymbol{y})$ is give by

$$\pi^*(\xi, \eta, \lambda_A, \lambda_B) = \pi(e^{\xi}, e^{\eta}, (1+e^{-\lambda_A})^{-1}, (1+e^{-\lambda_B})^{-1}|\boldsymbol{y}) e^{\xi+\eta+\lambda_A+\lambda_B}(1+e^{\lambda_A})^{-2}(1+e^{\lambda_B})^{-2}$$

the posterior probability of null is true can be computed as

$$\begin{aligned}
Pr(\gamma_i = 0|\boldsymbol{y}) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(e^{\xi}, e^{\eta}, (1+e^{\lambda_A})^{-1}, (1+e^{\lambda_B})^{-1}) \pi^*(\xi, \eta, \lambda_A, \lambda_B) \\
&\quad d\xi d\eta d\lambda_A d\lambda_B
\end{aligned}$$

To avoid computing the Hessian matrix, Metropolis-Hasting algorithm is used to obtain samples from the posterior density of $(\xi, \eta, \lambda_A, \lambda_B)$. Let $\boldsymbol{\theta} = (\xi, \eta, \lambda_A, \lambda_B)$. Given the current $\boldsymbol{\theta}^{(t)}$, the *random-walk* Metropolis algorithm iterates the following step:

1. Draw $\boldsymbol{\epsilon} \sim Multinormal(0, \Sigma)$ and set $\boldsymbol{\theta'} = \boldsymbol{\theta}^{(t)} + \boldsymbol{\epsilon}$, where Σ is the covariance matrix. Generally, we set $\Sigma = \tau^2 I$.

2. Simulate $\mu \sim DU(0, 1)$ and update

$$\boldsymbol{\theta}^{(t+1)} = \begin{cases} \boldsymbol{\theta'} & \text{if } \mu \leq \frac{\pi^*(\boldsymbol{\theta'})}{\pi(\boldsymbol{\theta}^{(t)})} \\ \boldsymbol{\theta}^{(t)} & \text{o.w.} \end{cases}$$

Roberts and Gilks (1997) suggested a rule in choosing τ^2, such that the acceptance rate is maintained in $25\% \sim 35\%$. Finally, take a sample $\{(\xi^i, \eta^i, \lambda_A^i, \lambda_B^i)\}_{i=1}^N$ from $\pi^*(\xi, \eta, \lambda_A, \lambda_B|\boldsymbol{y})$, $Pr(\gamma_i = 0|\boldsymbol{y})$, which is approximately

$$Pr(\gamma_i = 0|\boldsymbol{y}) = \frac{1}{N}\sum_{i=1}^{N} h(e^{\xi^i}, e^{\eta^i}, (1+e^{\lambda_A^i})^{-1}, (1+e^{\lambda_B^i})^{-1})$$

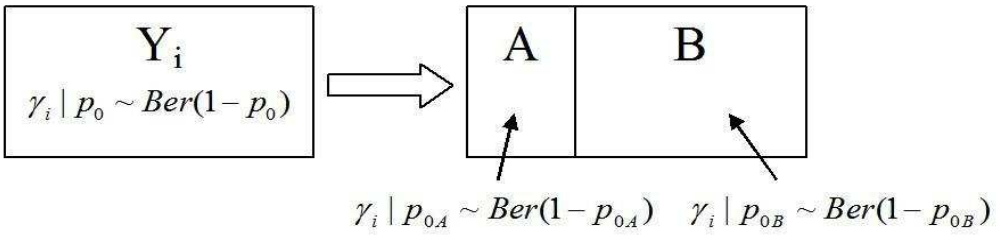

Figure 2.1: The Bayesian approach for partitioned data based on external information

Table 2.1: Number of true status of m hypotheses and decisions made for those hypotheses.

Hypothesis	Accept H_0	Reject H_0	Total
True	U	V	m_0
False	T	S	m_1
Total	W	R	m

Table 2.2: Results for our proposed method in the case of no signals.

Total number of tests (m)	Size of group A	False discovery rate (FDR)
1000	100	0.040
	500	0.040
5000	100	0.036
	500	0.036

Table 2.3: Results of our proposed method for different combinations of sizes of group A and numbers of signals in group A.

Size of group A	# of signals in A signal: noise	Power	FDR	FNR
30	10 (1:2)	0.91	0.040	0.186
60	10 (1:5)	0.83	0.039	0.349
110	10 (1:10)	0.75	0.039	0.509
510	10 (1:50)	0.50	0.045	0.993
30	8 (1:2.75)	0.73	0.035	0.540
60	8 (1:6.5)	0.65	0.040	0.693
110	8 (1:14)	0.59	0.038	0.817
510	8 (1:62.75)	0.40	0.038	1.193
30	4 (1:6.5)	0.42	0.033	1.153
60	4 (1:14)	0.38	0.037	1.243
110	4 (1:26.5)	0.33	0.026	1.333
510	4 (1:126.5)	0.24	0.027	1.521
30	0	0.20	0.031	1.611
60	0	0.20	0.031	1.615
110	0	0.20	0.025	1.599
510	0	0.19	0.037	1.621

Table 2.4: Top five results from group A and B with p-value, and posterior probability that the alternative hypothesis is true

SNPs in group A				
SNP ID	Chromosome	Position	P-value	Prob($\gamma = 1$| Data)
rs327405	8	31937783	6.5e-03	0.179
rs1487154	8	32310979	3.0e-03	0.195
rs16879298	8	32359504	1.6e-03	0.235
rs16879809	8	32692415	9.1e-04	0.239
rs17760889	12	13711681	7.8e-03	0.161
SNPs in group B				
SNP ID	Chromosome	Position	P-value	Prob($\gamma = 1$| Data)
rs4846033	1	11722830	4.36e-06	0.315
rs10911902	1	183363974	1.85e-06	0.327
rs9512730	13	26975144	4.52e-06	0.315
rs16977195	15	84785244	1.71e-06	0.327
rs151222	16	20581993	6.08e-06	0.312

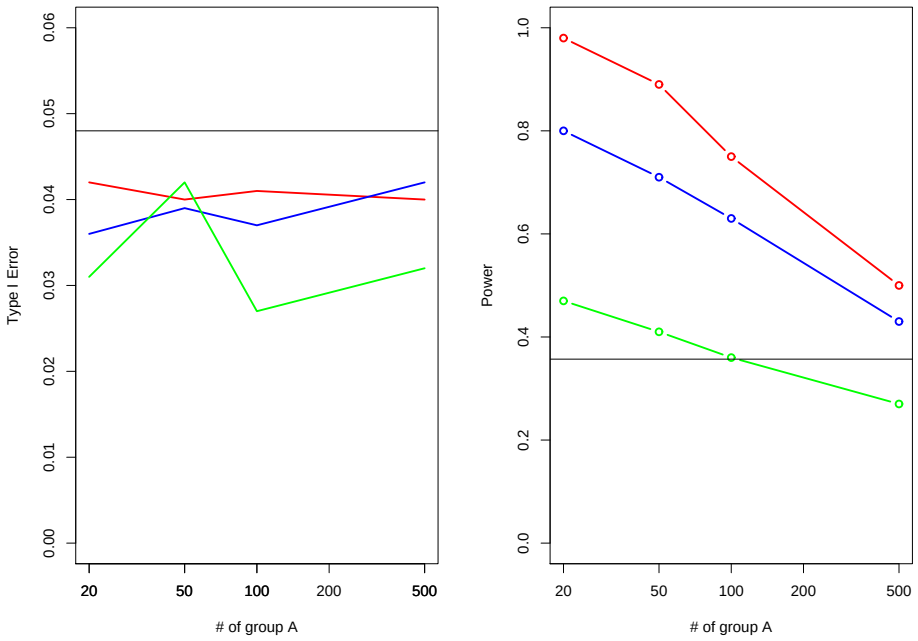

Figure 2.2: Plot of realized FDR and power for different cases of partition compared to q-value results. The red line is the case, which group A contains 10 signals. The blue line is the case, which group A contains 8 signals. The green line is the case, which A contains 4 signals. The horizontal line is the q-value result.

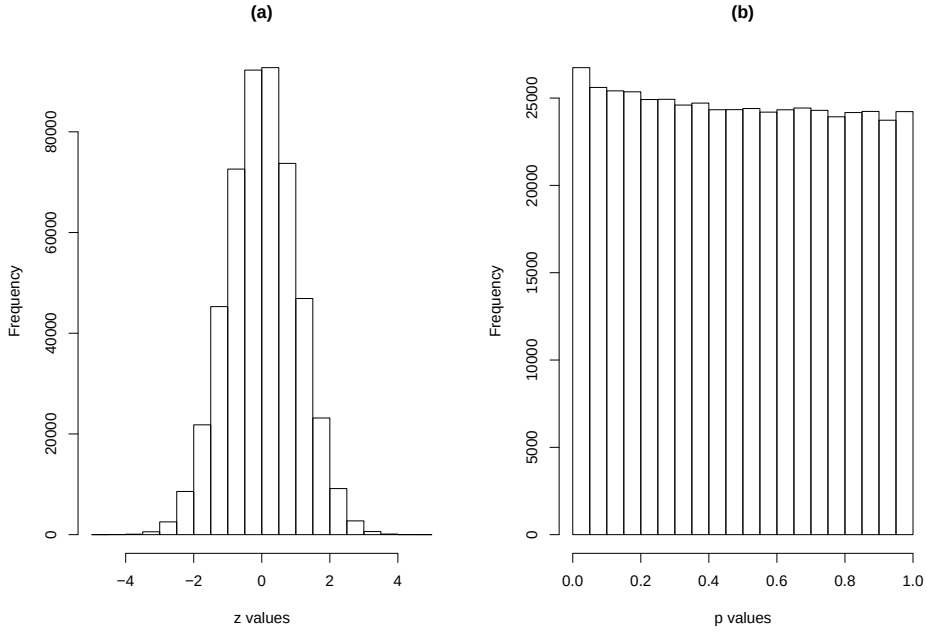

Figure 2.3: Histograms the SCZ data. (a) z values; (b) p-values

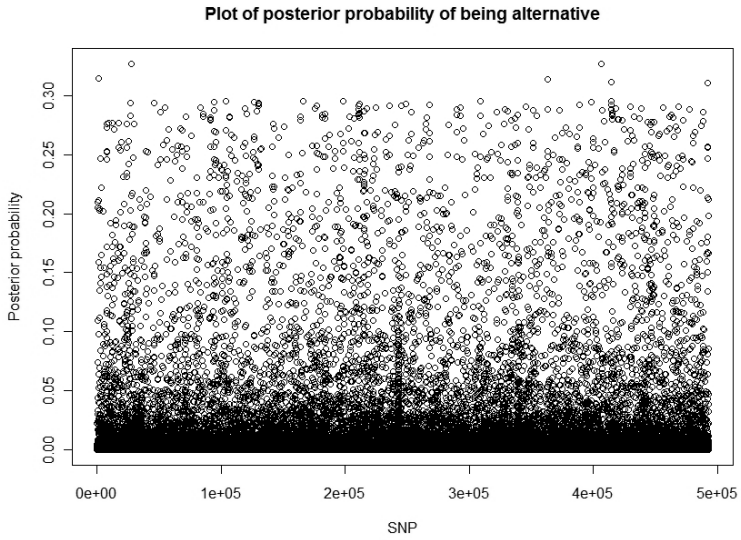

Figure 2.4: The posterior probability of the alternative hypothesis being true at each locus for GWAS with schizophrenia

Chapter 3

Nonparametric approach for meta-analysis with an application to association study

3.1 Abstract

Meta-analysis is useful in combining results (i.e. estimators) from multiple studies when the original raw data from each study is not available. A major concern in performing meta-analysis is to address the heterogeneity issue. Summary statistics or estimates obtained from various studies usually arise from a heterogeneous population. One of the main goals of a meta-analysis is to produce a combined estimate to properly account for the inherent heterogeneity present in various studies. Most of the popular meta-analysis methods are based on a random-effects model where the random effects are assumed to be normally distributed with an unknown population mean and an unknown variance. In this paper, we relax the normality assumption and show that a broad class of distributions can be approximated by a mixture distribution. The population mean and variance estimates are then obtained by using EM algorithm. To assess the performance of our method, we compare our results with DerSimonian and Laird method based on a simulation study. Our results show that the proposed method greatly improves the accuracy in estimating overall effect and heterogeneous variance in various realistic cases. We apply our method to combine results from multiple association studies on C957T polymorphism (located in the DRD2 gene) multiple association studies with schizophrenia.

3.2 Introduction

In the study of complex disease, many linkage and association studies are not able to provide solid evidence of linkage, even with large sample sizes. The relative risks of disease susceptibility loci are from low to moderate due to small sample size. For example, the DRD2 gene, identified as a susceptibility locus for schizophrenia (Allen *et al.*, 2008) has an odds ratio of 1.52, which can be considered in a follow-up study for further verification. For those loci, which have small effect on common disease, Zondervan and Cardon (2004) show that in the best cases, sample sizes of 2000-10000 cases and controls are required to obtain 80% power. However, in most single studies, it is almost impossible to have such large sample sizes. In addition, due to some privacy policies, complete data may not be available, and usually only few summary statistics (e.g., log odds ratio, standard error) are usually reported. In such situations, meta-analysis provides an effective way to combine results from multiple independent studies to produce a more reliable estimate of log odds ratios or other measurement of association.

Meta-analysis is defined as *"the statistical analysis of a large collection of analysis results for the purpose of integrating the findings"* (Glass, 1976). The study of meta-analysis can be traced back to almost a century ago. The earliest example appears in 1904 by Karl Pearson who combined several studies of typhoid vaccine effectiveness. During the recent decades, meta-analysis has become a powerful tool in medical research and clinical studies. The number of published articles on health-related, meta-analyses has increased to 400 per year since the year 2000 (Sutton and Higgins, 2008). One of the issues in combining different studies is that the data collection protocols are not usually consistent with each other, which results into different levels of heterogeneity among studies. Hirschhorn (2002) pointed out in 600 positive associations between genetic variants and disease, 166 associations were repeated for three or more times and only six associations were consistent.

In order to account for possible inconsistencies among study protocols, two different models are generally used in combining multiple studies. One of them is known

as the fixed-effects model; which assumes all study samples are from one population with a common fixed but unknown effect size, and the differences between estimates across multiple studies are due to the sampling error. The other one is known as the random-effects model, which assumes that the study samples are from a distribution of populations. The differences between estimates across studies arise from sampling error and heterogeneity among populations. The causes of heterogeneity are complicated, it may be due to sampling scheme, or deviation from equilibrium (e.g. Hardy-Weinberg equilibrium) in some studies.

The estimate of the population mean within a meta-analysis method is obtained by using weighted average of point estimates obtained from multiple studies. The weights are usually chosen based on the standard errors of these point estimates (Whitehead, 1991). For example, the weights can be chosen as the reciprocal of standard error. Accounting for heterogeneity is one of the major issues in meta-analysis. To test the presence of heterogeneity, a chi-squared Q test (Cochran, 1954) is generally used. Other methods, such as Wald, likelihood ratio, and score tests for heterogeneity have also been developed (Hardy and Thompson, 1996, Viechtbauer, 2007) and some improvements to the Q test have been suggested (Lipsitz *et al.*, 1998, Takkouche *et al.*, 1999). Due to the simplicity of the chi-squared test, Higgins and Thompson (2002) proposed a statistic, I^2, which measures the extent of heterogeneity as the proportion of total variation and I^2, and is now widely used. However, it may be more important to quantify the extent of the heterogeneity than to rely on an overall statistical test to detect its presence (Thompson, 1994). Due to the limited information, poor estimation of heterogeneous variance is always a problem in a random-effects model. Sidik and Jonkman (2007) compared performance of seven estimators (including the DSL method, the maximum likelihood method, an empirical Bayes method and various other methods) of the heterogeneous variance by conducting a simulation study. Except DSL's moment based approach, other chosen methods were based on the normality assumption of underlying random effect. Often it is difficult to check the suitability of the normality assumption. If such an assumption is found to be false, those methods with normality assumption are not reliable.

Hardy and Thompson (1998) proposed to use Anderson-Darling A^2 test statistic to check the normality assumption, but such a test would require large sample sizes with each study.

In this paper, we propose to use a suitable mixture model to approximate the underlying distribution of the random effects and hence avoid making the normality assumption. EM algorithm is then used to obtain optimal weights of the mixture distribution. The estimates of overall population mean effect and heterogeneous variance are then computed based on the estimated weights. In Section 3.3, we introduce the random-effects model and a test statistic for heterogeneous variance and a brief review of the DerSimonia-Laird (DSL) moment based estimator for heterogeneous variance. In Section 3.4, we present the proposed mixture model for random effects and derive an EM algorithm to obtain weights and corresponding estimates for overall mean effect and heterogeneous variance. In Section 3.5, results of several simulation studies are reported to compare the performance of our method to that of the DSL method. Finally, in Section 3.6, we apply our method to combine results from six association studies with schizophrenia on C957T polymorphism and compare our results to several other methods.

3.3 Methods for meta-analysis

Suppose $\hat{Y}_1, \ldots, \hat{Y}_k$ are the summary estimates from k independent studies with standard errors $\hat{\sigma}_1, \ldots, \hat{\sigma}_k$ respectively. For example, $\hat{Y}_i$ could be the observed log odds ratio in trial i, and $\hat{\sigma}_i$ could be the estimated standard error of $\hat{Y}_i$. One of the major interests is to estimate overall treatment effect (e.g., population level log odds ratio) from these k studies and estimate the population level heterogeneity. Two models are often used to obtain those estimates; one is fixed-effects model, and the other is random-effects model.

3.3.1 Fixed-effects model

In a fixed-effects model, all the studies are assumed to arise from a common population with a fixed but unknown mean θ but possibly with different variance. The model can be expressed as:

$$\hat{Y}_i = \theta + \epsilon_i \qquad \epsilon_i \overset{indep.}{\sim} N(0, \hat{\sigma}_i^2) \qquad i = 1, \ldots, k.$$

The overall treatment effect θ may be estimated as a weighted average,

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i Y_i}{\sum_{i=1}^{k} w_i}$$

where w_i is the weight assigned to ith study. The weight for study i is usually taken to be the reciprocal of the estimate of variance of study i (i.e. $w_i = 1/\hat{\sigma}_i^2$) which results in minimizing the overall variability of the estimate $\hat{\theta}$ across all weights. The fixed-effects model may have large power, but it may not be reasonable.

To test the null hypothesis of no between-study variance, a chi-squared test, called Cochran's Q statistic, (Cochran 1954) is used

$$Q = \sum_{i=1}^{k} w_i (\hat{Y}_i - \hat{\theta})^2$$

However, the chi-squared test has several disadvantages. First, it is not powerful, when the number of studies is small. Second, the Q statistic may be significant, when the number of studies is large, while the individual effect size estimates do not really differ much. And finally the distribution of Q is based on the normality assumption, which may not be proper when k is small.

Recently, Higgins and Thompson (2002) proposed the index of heterogeneity test statistic, I^2, which is defined as

$$\begin{aligned} I^2 &= \left(\frac{Q - (k-1)}{Q} \right)_+ \times 100 \\ &= \max \left\{ 0, \frac{Q - (k-1)}{Q} \right\} \times 100 \end{aligned}$$

The value of I^2 is between 0 and 1. Smaller values of I^2 indicates less heterogeneity and larger value of I^2 indicates greater heterogeneity among groups. However, the

exact quantification of "large" and "small" can be tricky and in general may depend on the value of k itself.

3.3.2 Random-effects model

A random-effects model is usually considered more reasonable, as it incorporates between-study variability into the overall estimate of the treatment effect. The model can be written as :

$$\begin{aligned} \hat{Y}_i &= \theta_i + \epsilon_i & \theta_i &= \theta + \delta_i \\ \epsilon_i &\sim N(0, \hat{\sigma}_i^2) & \delta_i &\sim N(0, \tau^2) \qquad i = 1, \ldots, k. \end{aligned}$$

here θ_i is the study-specific effect for the i^{th} study. $\hat{\sigma}_i^2$ is the within-study variance in the i^{th} study. θ is the overall treatment effect and τ^2 is the heterogeneous variance. Notice that the random effects model reduces to fixed effects model when $\tau^2 = 0$. Unlike in fixed-effects model, increasing the number of studies does not necessarily result in an increase in power, because the added studies may increase the total heterogeneity variance (Cohn, 2003).

3.3.3 The DerSimonian and Laird method (DSL)

DerSimonian and Laird (1986) proposed a non-iterative moment-based procedure to estimate the heterogeneous variance. The DSL approach is a simple and straight forward method, which has been widely used. Their proposed moment estimator is a special case of a more general moment-based approach (DerSimonian and Kacker, 2006). Several suggestions for adapting the random-effects method have been discussed, by Hartung and Knapp (1999), Sidik and Jonkman (2005) and others. The DSL approach is based on taking the expectation of Cochran's statistic (1954),

$$E(Q) = (k-1) + \tau^2 \left(\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i} \right)$$

where $w_i = 1/\sigma_i^2$, $Q = \sum_{i=1}^{k} w_i (Y_i - \hat{\theta})^2$ and $\hat{\theta} = \frac{\sum_{i=1}^{k} w_i Y_i}{\sum_{i=1}^{k} w_i}$. The proof can be found in Böhning (1999). Replacing the expected value of Q by observed value gives us the

DSL estimator of τ^2:

$$\hat{\tau^2} = \frac{Q-(k-1)}{\sum_{i=1}^{k} w_i - \sum_{i=1}^{k} w_i^2 / \sum_{i=1}^{k} w_i}$$

In the software *R*, there is a standard package called *rmeta* to implement the algorithm. However, using the conditional study-specific variance instead of the population-averaged versions in situations other than the normal leads to considerable bias which will not disappear even with large study sizes (Böhning *et al.* 2002). In our simulation study, large biases in several cases were observed with large study sizes.

3.4 A nonparametric mixture model

As there is usually a limited knowledge available about the mechanism of the underlying effect in each study, it would be preferable to use a model for θ_i's that are not necessarily normally distributed. Li and Barron (1999) have shown that any bounded continuous density f can be approximated by finite mixture of density function in class $G = \{\phi(\frac{x-\mu}{\sigma}), \mu \in R, \sigma > 0, x \in R\}$, where $\phi(x)$ denotes a density of a standard normal distribution; such that the Kullback-Leibler divergence (defined as $D(f||g) = \int f(x) log[f(x)/g(x)]dx$) between f and a suitable mixture with K components from G can be bounded by a quantity which converges to zero at the rate of $1/k$. We use this result to propose a mixture model as follows:

$$\theta_i \overset{indep.}{\sim} F_N(\cdot), \qquad dF_N(x) = \sum_{l=1}^{N} w_{l,N} I(x = \xi_l^*)$$

where N and ξ_l^*'s are chosen properly. The weights $w_{l,N}$'s are estimated using EM algorithm. Notice that the log-likelihood of the weights w_i's given the $\hat{Y}_i$'s and $\hat{\sigma}_i$'s is given by

$$\begin{aligned} \log L(\boldsymbol{w}) &= \sum_{i=1}^{k} \log \int \phi\left(\frac{\hat{y}_i - \theta_i}{\hat{\sigma}_i}\right) dF_N(\theta_i) \\ &= \sum_{i=1}^{k} \log \left(\sum_{l=1}^{N} \phi(\frac{\hat{y}_i - \xi_l^*}{\hat{\sigma}_i}) w_l \right) \end{aligned}$$

By introducing missing indicators for the mixture components, the above log-likelihood can be maximized using EM algorithm. The details of the algorithm is presented in Appendix A. The iterative solution to obtain the optimal weights can be expressed as

$$\hat{w}_l^{(new)} = \frac{1}{k}\sum_{i=1}^{k} \frac{\hat{w}_l^{(old)}\phi(\frac{\hat{y}_i-\xi_l^*}{\hat{\sigma}_i})}{\sum_{l=1}^{N}\hat{w}_l^{(old)}\phi(\frac{\hat{y}_i-\xi_l^*}{\hat{\sigma}_i})} \qquad l = 1,\ldots,N \tag{3.1}$$

It can be shown that the above algorithm converges to a local maximum under some mild regularity condition (Wu, 1983). We have used a starting value of weights, $\hat{\boldsymbol{w}}^{(old)} = (\frac{1}{k},\ldots,\frac{1}{k})$ for all of our computations. The stopping rule that we have used is based on the maximum absolute value of difference between $\hat{\boldsymbol{w}}^{(new)}$ and $\hat{\boldsymbol{w}}^{(old)}$ not exceeding a small number; i.e., $\max_{1\le l\le N}|\hat{w}_l^{(new)} - \hat{w}_l^{(old)}| \le \epsilon$, where say $\epsilon \approx 10^{-5}$. Once we obtain the estimate of weights $\hat{w}_l$'s, we are able to estimate the treatment effect θ and heterogeneous variance τ^2 by the following expressions:

$$\hat{\theta} = \sum_{l=1}^{N} w_l\xi_l^* \tag{3.2}$$

$$\hat{\tau}^2 = \sum_{l=1}^{N} w_l(\xi_l^* - \hat{\theta})^2 \tag{3.3}$$

In our implementation, we have chosen ξ_l^* to be the equi-spaced partition of the range of $\{\hat{Y}_{i=1}^k\}$. And N is selected using Akaike's information criterion (AIC).

3.5 Simulation Study

To assess the performance of our method, we conducted several simulation studies and compared our results with the DSL method. We have chosen to use the number of studies, $k = 10, 30, 50, 80$ and heterogeneous variances were chosen as $\tau^2 = 0, 0.25, 0.5, 1, 1.25, 1.5, 1.75, 2$ and the true overall effect was set at $\theta = 0.5$. The choice of the characteristics of our simulation studies cover a lot of practical situations that we have found in the literature. Next, we generated 2×2 contingency tables by using a method similar to Platt *et al.* (1999). The procedure is described

as follow. First, generate θ_i from $N(\theta, \tau^2)$ for $i = 1, \ldots, k$. Second, for a given k, the sample sizes n_{iC} of the ith $(i = 1, \ldots, k)$ control groups were randomly chosen from a discrete uniform distribution $DU(20, 200)$, and the sample sizes n_{iT} for the ith $(i = 1, \ldots, k)$ treatment groups were randomly chosen from a discrete uniform distribution $DU(30, 300)$. Third, we generated the responses x_{iC} for the control groups from a binomial distribution, $Bin(n_{iC},p_{iC})$, for $i = 1, \ldots, k$, where the true binomial probability p_{iC} was randomly selected from a uniform distribution, $U(0.05, 0.65)$. The responses x_{iT} for the treatment groups were generated from $Bin(n_{iT},p_{iT})$ distribution with $p_{iT} = \frac{p_{iC} \exp(\theta_i)}{1-p_{iC}+p_{iC} \exp(\theta_i)}$, for $i = 1, \ldots, k$. Finally, the θ_i's were generated from various distributions to be specified below.

In general, the data generation can be summarized as

$$\begin{array}{ll} \theta = 0.5 & \theta_i \sim F(\cdot) \\ n_{iC} \sim DU(20, 200) & n_{iT} \sim DU(30, 300) \\ p_{iC} \sim DU(0.05, 0.65) & p_{iT} = \frac{p_{iC} \exp(\theta_i)}{1-p_{iC}+p_{iC} \exp(\theta_i)} \\ x_{iC} \sim Bin(n_{iC}, p_{iC}) & x_{iT} \sim Bin(n_{iT}, p_{iT}) \end{array}$$

where $\int x dF(x) = \theta$ and $\int (x - \theta)^2 dF(x) = \tau^2$.

Notice that $\theta_i = logit(p_{iT}) - logit(p_{iC})$, where $logit(x) = \log(x/(1 - x))$ is our parameter of interest. We computed estimated log odds ratio and standard error using the following formula:

$$\begin{aligned} \hat{y}_i &= \log \left\{ \frac{(x_{iC} + 0.5)(n_{iT} - x_{iT} + 0.5)}{(x_{iT} + 0.5)(n_{iC} - x_{iC} + 0.5)} \right\} \\ \hat{\sigma}_i^2 &= \frac{1}{x_{iC} + 0.5} + \frac{1}{x_{iT} + 0.5} + \frac{1}{n_{iT} - x_{iT} + 0.5} + \frac{1}{n_{iC} - x_{iC} + 0.5} \end{aligned}$$

We repeated the above data generation procedure $B = 1000$ times and obtained 1000 estimates of θ and τ^2 using the DSL method by using the R package *rmeta* and by using our proposed method. The biases for θ and τ^2 were calculated as $B(\theta) = \frac{1}{B} \sum_{j=1}^{B} \hat{\theta}_j - \theta$ and $B(\tau^2) = \frac{1}{B} \sum_{j=1}^{B} \hat{\tau}^2 - \tau^2$. Three distributions for $F(\cdot)$ were considered: (i) $F(x) = \Phi(\frac{x-\theta}{\tau})$, where $\Phi(\cdot)$ is the CDF of a standard normal distribution; (ii) $F(x) = 0.5[1 + sgn(x - \theta)(1 - \exp(-|x - \theta|/b))]$, where $F(\cdot)$ is the CDF of so called Laplace distribution with mean θ, and variance $\tau^2 = 2b^2$; (iii) $F(x) =$

$[1 + \exp(-(x-\theta)/s)]^{-1}$, where $F(\cdot)$ is the CDF of so called logistic distribution with mean θ, and variance $\tau^2 = \pi^2/3s^2$. We selected $\{\xi_l^*\}_{l=1}^N$ to be the end points which equi-spaced partition of the range of $\{\hat{y}_i\}_{i=1}^k$, *i.e.*, $\xi_l^* = \hat{y}_{(1)} + \frac{l-1}{N-1}(\hat{y}_{(k)} - \hat{y}_{(1)})$, for $l = 1, \ldots, N$. N was chosen based on AIC. We also explored the cases with $N = 3, 4, 6, 7, 10$ and 15. The choice of $N = 5$ usually provided the best estimates for all cases. When N increases ($N > 5$), the bias of τ^2 begins to increase. It means when N is larger, overfitting might be an issue. On the other hand, smaller values of N may lead to underfitting. One can use AIC to choose the best N.

[Table 3.7, 3.7 and 3.7 are here.]

From table 3.7, 3.7 and 3.7, it can be seen that the DSL method underestimates the heterogeneous variance for all values of k and τ^2. The magnitude of biases of the heterogeneous variance estimates increases, as the true value of τ^2 increases. The bias issue of DSL method has also been mentioned by Böhning *et al.* (2002). When the density of underlying effects is not normal, the biases of the estimates of overall mean effect and heterogeneous variance increase dramatically. The proposed method, which estimates the weights nonparametrically appears much better in estimating the overall effect and heterogeneous variance for most distributions (including the non-normal cases). The biases of θ and τ^2 decrease as the number of studies (k) increases. In case of Laplace and logistic distribution, our method provides a much better estimate for overall treatment effect and heterogeneous variance.

Figure 3.1-3.6 are here.

We also present the box plots for all possible true values of k and τ^2 in figure 3.1-3.6. In the box plots of overall treatment effect, the estimates obtained by the proposed method are as good as the DSL method with a little bit larger variance. In the box plots of heterogeneous variance, the estimates obtained by the proposed method are much closer to the true value than the DSL estimates, especially when the number of studies (k) is large. The biases of the estimates by DSL method do not decrease when the number of studies (k) increases, which was mentioned by Böhning *et al.*

(2002). The estimate of the proposed method has a larger variance compared to DSL method, however with the number of studies (k) increasing, the bias and variance decrease.

[Figure 3.7-3.12 are here.]

To compare mean square error, which is defined as $MSE = Bias^2 + s.e.^2$, we also present plots of MSE for the DSL method and our proposed method in figure 3.7-3.12. When the number of studies is small, say $k < 30$, the DSL method has a smaller MSE than our method for estimates of θ and τ^2. In other cases, our method has a smaller MSE for θ and τ^2. As mentioned by Sidik (2007), the DSL estimator tends to concentrate its sampling distribution and its variance is reduced to offset the squared bias, which leads to a decrease in MSE. This leads us to believe that MSE may not be a good criterion in choosing an estimator. In summary, we can clearly see the benefit of using the new estimator.

3.6 Meta-analysis for C957T polymorphism

Schizophrenia is considered to be a complex disease involving genetic and environmental factors with prevalence of about 0.5-1%. The disease is highly heritable, with heritability estimates around 80% (Cardno *et al.*, 1999). Although many genes and regions have been investigated to look for association with schizophrenia, no single causal gene has been identified.

Recently, systematic meta-analysis of genetic association studies in schizophrenia (Allen *et al.*, 2008) has been performed. 24 genetic variants in 16 different genes have been reported as potential causal genes. The DRD2 gene is one of the 16 reported genes. A number of clinical studies have shown that schizophrenia involves a dysregulation of the dopaminergic system. The type 2 dopamine receptor (D2) is a major target of anti psychotic drugs; and the genes involved in dopaminergic transmission, especially DRD2 (chromosome 11q22-q23) have been largely analyzed in case-control

association studies in schizophrenic patients (Hoenicka *et al.*, 2006). *In vitro* studies have shown that the DRD2 C957T polymorphism has marked functional consequences for DRD2 mRNA stability and dopamine-regulated DRD2 expression: the T allele of this SNP is associated with decreased mRNA translation and stability, while the C allele is not (Duan *et al.*, 2003). We apply our proposed method to combine results from six association studies on the C957T polymorphism located in DRD2 gene, published between 2005 and 2008. The summary of the six studies is given in table 3.4.

[Table 3.4 is here]

Since the complete data is accessible, a full nonparametric Bayesian model can be fitted to the complete data. The model is summarized as following.

$$\begin{aligned} x_{iC} &\sim Bin(p_{iC}, n_{iC}) & x_{it} &\sim Bin(p_{iT}, n_{iT}) \\ p_{iC} &= DU(0,1) & p_{iT} &= \frac{p_{iC}e^{\theta_i}}{1-p_{iC}+p_{iC}e^{\theta_i}} \\ \theta_i &\sim F(\cdot) & \boldsymbol{w} &\sim ddirch(\alpha) \end{aligned}$$

where θ_i follows categorical distribution, which means $Prob(\theta_i = \xi_l^*) = w_l$, for $l = 1, \ldots, N$, and w_l follows Dirichlet distribution with parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$. $\xi_1^*, \ldots, \xi_N^*$ can be chosen by the following way. First, calculate $\hat{y}_i = logit\frac{x_{iT}}{n_{iT}-x_{iT}} - logit\frac{x_{iC}}{n_{iC}-x_{iC}}$ and order them by $\hat{y}_{(1)} \leq \ldots \leq \hat{y}_{(k)}$. Next, let $\xi_1^* = \hat{y}_{(1)}$, and $\xi_l^* = \hat{y}_{(1)} + \frac{l-1}{N-1}[\hat{y}_{(k)} - \hat{y}_{(1)}]$. So $\{\xi_l^*\}_{l=1}^N$ equally partition the range of $\hat{y}_i$'s. After obtaining the posterior median of w_l using (3.1), for $l = 1, \ldots, N$, we can estimate overall mean effect and heterogeneous variance as

$$\begin{aligned} \hat{\theta} &= \sum_{l=1}^{N} \xi_l^* w_l \\ \hat{\tau}^2 &= \sum_{l=1}^{N} (\xi_l^* - \hat{\theta})^2 w_l \end{aligned}$$

To choose the optimal N, we compute the deviance for $N = 1, \ldots, k$. A model with smaller deviance is desirable.

To investigate potential association between C/C and the disease, we tested the genotype frequency (C/C vs. C/T, T/T) between the schizophrenia patients and control individuals across all studies. To visually check the heterogeneity, a forest plot is helpful.

[Figure 3.13 is here]

The left column lists all the names of studies. The right column is a plot of odds ratio for each study. Each horizontal line is the confidence interval for estimate of OR from each study. The size of each square is proportional to the study's weight in a meta-analysis. A vertical line representing no effect is also plotted. If the confidence intervals for individual studies overlap with this line, it means that at the given level of confidence their effect sizes do not differ from common effect. For DSL method, the overall mean effect ($\hat{\theta}$) is 0.37, with 95% confidence interval (0.10, 0.65). The estimated heterogeneous variance is 0.08. For our proposed method, first, we choose $N = 2$ based on AIC, which is defined as $2N - 2\log(L(\boldsymbol{w}))$. A model with smaller values of AIC is desirable. For this data, $N = 2$ gives the smallest AIC. The overall mean effect ($\hat{\theta}$) is 0.41, with 95% confidence interval (0.10, 0.75). The corresponding estimated heterogeneous variance is 0.12. The overall mean effect by the Bayesian nonparametric model is 0.37, with 95% equal-tailed posterior interval (0.19, 0.54). The heterogeneous variance is 0.06 using the posterior median of $\hat{\tau}^2$.

Next, we tested the allele frequency (C vs. T) between the schizophrenia patient and control group. The forest plot is shown in figure 3.14.

[Figure 3.14 is here]

The overall mean effect ($\hat{\theta}$) by the DSL method is 0.28, with 95% confidence interval (0.08, 0.48). The estimated heterogeneous variance is 0.04. For our proposed method, we choose $N = 3$ in this case. The overall mean effect is 0.24, with 95% confidence interval (0.05, 0.43). The estimated heterogeneous variance is 0.02. The overall mean effect by the Bayesian nonparametric model is 0.28, with confidence interval (0.15, 0.53). The heterogeneous variance is 0.03.

3.7 Discussion

Meta-analysis is a powerful tool for estimating population-level effects of candidate genes on some complex phenotypes. Without accessing the original data, meta-analysis provides a way to combine results (some reported statistics) from multiple studies. A random-effects model is commonly used in meta-analysis, which is considered more realistic. The normality assumption of the underlying treatment effects is required for most methods that estimate the overall effect and heterogeneity. We show that a broad class of distributions can be approximated by a mixture distribution. Our simulation studies suggest that the nonparametric approach for meta-analysis reduces the biases in overall treatment effect and heterogeneous variance. The performance of our proposed method is much better in various distribution assumptions of underlying effect. We illustrate our proposed method by applying the method to six association studies on C957T polymorphism with schizophrenia. However, meta-analysis is not a replacement for adequately powered genetic association studies. The greatest value of a meta-analysis suggests what we need to aim for. For example, the small odds ratios, a mean of 1.33 in 55 meta-analyses (Ioannidis, *et al.*, 2003) indicates that studies will need to include many thousands of subjects, if they are to provide unequivocal evidence of an association between a genetic variant and a phenotype (Zondervan, 2004).

Appendix

Appendix A

Let $\boldsymbol{Z} = \{z_i\}_{i=1}^k$ be a vector of indicator variables. $z_i \in \{1, \ldots, k\}$ for each i, and $z_i = l$ if the y_i is from l^{th} study with $P(Z_i = l) = w_l$, with $i = 1, \ldots, k$ and $l = 1, \ldots, N$. If we know the values of $\boldsymbol{Z}$, the log likelihood becomes

$$\begin{aligned} \log\left(L(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{Z})\right) &= \log(f(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{\theta})) \\ &= \sum_{i=1}^{k} log(f(y_i|z_i)f(z_i)) \\ &= \sum_{i=1}^{k} log(w_{z_i} f(y_i|\theta_{z_i})) \end{aligned}$$

Let $\boldsymbol{w}^{(\nu)} = (\boldsymbol{w}_1^{(\nu)}, \ldots, \boldsymbol{w}_N^{(\nu)})$ be the true parameter values. Using Bayes's rule, we have

$$\begin{aligned} w^*(z_i|y_i, \boldsymbol{w}^{(\nu)}) &= \frac{w_{z_i} f(y_i|\boldsymbol{w}_{z_i}^{(\nu)})}{f(y_i|\boldsymbol{w}^{(\nu)})} \\ &= \frac{w_{z_i} f(y_i|\boldsymbol{w}_{z_i}^{(\nu)})}{\sum_{l=1}^{N} w_l f(y_i|\boldsymbol{w}^{(\nu)})} \end{aligned}$$

In the "E" step: we define

$$\begin{aligned} Q(\boldsymbol{w}|\boldsymbol{w}^{(old)}) &= \sum_{i=1}^{k} E[\log w_{z_i}|y_i, w^{(old)}] \\ &= \sum_{i=1}^{k}\sum_{l=1}^{N} \log w_l \cdot w_l^*(y_i; \xi_l^*, \sigma_i) \end{aligned}$$

In the "M" step: we need to solve the following optimization problem:

$$\begin{aligned} &\max_{w} Q(\boldsymbol{w}|\boldsymbol{w}^{(old)}) \\ &\text{s.t } \sum_{l=1}^{N} w_l = 1 \end{aligned}$$

We introduce the Lagrange multiplier λ, and solve the following equation:

$$\frac{\partial}{\partial w_l} Q(\boldsymbol{w}|\boldsymbol{w}^{(old)}) - \lambda = 0$$

or

$$\begin{aligned} \frac{1}{w_l}\sum_{i=1}^{k} w_l^*(y_i, \xi_l^*, \sigma_i) - \lambda &= 0 \\ \sum_{i=1}^{k} w_l^*(y_i, \xi_l^*, \sigma_i) &= w_l\lambda \end{aligned}$$

Summing both sides over l, since $\sum_{l=1}^{N} w_l = 1$, we have $\lambda = 1/k$. So the iterative solution is given by

$$\hat{w}_l^{(new)} = \frac{1}{k}\sum_{i=1}^{k} \frac{\hat{w}_l^{(old)}\phi(\frac{\hat{y}_i - \xi_l^*}{\hat{\sigma}_i})}{\sum_{l=1}^{N}\hat{w}_l^{(old)}\phi(\frac{\hat{y}_i - \xi_l^*}{\hat{\sigma}_i})} \qquad l = 1, \ldots, N$$

Appendix B

The overall effect can be estimated as following:

$$\begin{aligned}\hat{\theta} &= g(\hat{\boldsymbol{w}}) \\ &= \sum_{l=1}^{N} \hat{w}_l \hat{\xi}_l \\ &= \sum_{l=1}^{N-1} \hat{w}_l (\hat{\xi}_l - \hat{\xi}_N) + \hat{\xi}_N\end{aligned}$$

The variance of $\hat{\theta}$ is computed as

$$var(\hat{\theta}) \simeq \frac{\partial g(\boldsymbol{w})}{\partial \boldsymbol{w}^T} \Sigma_{\boldsymbol{w}} \frac{\partial g(\boldsymbol{w})}{\partial \boldsymbol{w}}$$

Here $\Sigma_{\boldsymbol{w}}$ is the covariance matrix of $\boldsymbol{w}$. Sinc $\Sigma_{\boldsymbol{w}} = I^{-1}(\boldsymbol{w})$, we can compute the information matrix from $Q(\boldsymbol{w}'|\boldsymbol{w})$.

A simple expression for the information matrix of the general EM algorithm was obtained by Oakes (1999). Suppose observations $\{y\}$s with likelihood $L(\phi, y)$ depending on the parameter vector ϕ. In each step of EM algorithm, we choose $\phi^{'}$ to maximize $Q(\phi^{'}|\phi)$. Then by Oakes,

$$\frac{\partial^2 L(\phi, y)}{\partial \phi^2} = \left\{ \frac{\partial^2 Q(\phi'|\phi)}{\partial \phi'^2} + \frac{\partial^2 Q(\phi'|\phi)}{\partial \phi' \partial \phi} \right\}_{\phi'=\phi}$$

Let $f(y_i|\xi_l^*, \sigma_i) = a_{il}$,

$$\Sigma_{\boldsymbol{w}}^{-1} = E \left\{ -\frac{\partial^2}{\partial \boldsymbol{w} \partial \boldsymbol{w}^T} L(y; \boldsymbol{w}) \right\}_{N \times N}$$

We can rewrite

$$\begin{aligned}\sum_{l=1}^{N} w_l a_{il} &= \sum_{l=1}^{N-1} w_l a_{il} + w_N a_{iN} \\ &= a_{iN} + \sum_{l=1}^{N-1} w_l (a_{il} - a_{iN})\end{aligned}$$

Let $b_i = \sum_l w_l a_{il}$. Then

$$\begin{aligned}
\frac{\partial b_i}{\partial w_m} &= a_{im} - a_{iN} \text{ for m=1, } \ldots \text{, N-1} \\
Q(\boldsymbol{w}^*|\boldsymbol{w}) &= \sum_{l=1}^{N} w_l \log w_l^* \sum_{i=1}^{k} \frac{a_{il}}{b_i} \\
&= \sum_{l=1}^{N-1} w_l \log w_l^* \sum_{i=1}^{k} \frac{a_{il}}{b_i} + (1 - \sum_{l=1}^{N-1} w_l) \log (1 - \sum_{l=1}^{N-1} w_l^*) \sum_{i=1}^{k} \frac{a_{iN}}{b_i}
\end{aligned}$$

By simple calculus, we can get each element in the covariance matrix.

$$\frac{\partial Q(\boldsymbol{w}^*|\boldsymbol{w})}{\partial w_s^*} = \frac{1}{w_s^*} \sum_{i=1}^{k} \frac{w_s a_{is}}{b_i} - \frac{w_N}{w_N^*} \sum_{i=1}^{k} \frac{a_{iN}}{b_i}$$

For $s, t = 1, \ldots, N-1$, the second derivative is given by

$$\begin{aligned}
\left[\frac{\partial^2 Q(\boldsymbol{w}^*|\boldsymbol{w})}{\partial w_s^{*2}}\right]_{\boldsymbol{w}^*=\boldsymbol{w}} &= -\frac{1}{w_s} \sum_i \frac{a_{is}}{b_i} - \frac{1}{w_N} \sum_i \frac{a_{iN}}{b_i} \\
\left[\frac{\partial^2 Q(\boldsymbol{w}^*|\boldsymbol{w})}{\partial w_s^* \partial w_t^*}\right]_{\boldsymbol{w}^*=\boldsymbol{w}} &= -\frac{1}{w_N} \sum_i \frac{a_{iN}}{b_i}
\end{aligned}$$

$$\begin{aligned}
\left[\frac{\partial^2 Q(\boldsymbol{w}^*|\boldsymbol{w})}{\partial w_s^* \partial w_s}\right]_{\boldsymbol{w}^*=\boldsymbol{w}} &= \left[\frac{1}{w_s^*}\sum_i \frac{a_{is}}{b_i} - \frac{w_s}{w_s^*}\sum_i \frac{a_{is}(a_{is}-a_{iN})}{(b_i)^2}\right. \\
&+ \left.\frac{1}{w_N^*}\sum_i \frac{a_{iN}}{b_i} + \frac{w_N}{w_N^*}\sum_i \frac{a_{iN}(a_{is}-a_{iN})}{b_i^2}\right]_{\boldsymbol{w}^*=\boldsymbol{w}} \\
&= \frac{1}{w_s}\sum_i \frac{a_{is}}{b_i} - \sum_i \frac{a_{is}(a_{is}-a_{iN})}{b_i^2} \\
&+ \frac{1}{w_N}\sum_i \frac{a_{iN}}{b_i} + \sum_i \frac{a_{iN}(a_{is}-a_{iN})}{b_i^2} \\
\left[\frac{\partial^2 Q(\boldsymbol{w}^*|\boldsymbol{w})}{\partial w_s^* \partial w_t}\right]_{\boldsymbol{w}^*=\boldsymbol{w}} &= \left[-\sum_i \frac{w_s a_{is}(a_{it}-a_{iN})}{w_s^* b_i^2} + \sum_i \frac{a_{iN}}{w_N^* b_i}\right. \\
&+ \left.\sum_i \frac{w_N a_{iN}(a_{is}-a_{iN})}{w_N^* b_i^2}\right]_{\boldsymbol{w}^*=\boldsymbol{w}} \\
&= \frac{1}{w_N}\sum_i \frac{a_{iN}}{b_i} + -\sum_i \frac{(a_{is}-a_{iN})(a_{it}-a_{iN})}{b_i^2}
\end{aligned}$$

So

$$\begin{aligned}
\left[\frac{\partial^2 Q(\boldsymbol{w}^*|\boldsymbol{w})}{\partial w_s^{*2}} + \frac{\partial^2 Q(\boldsymbol{w}^*|\boldsymbol{w})}{\partial w_s^* \partial w_s}\right]_{\boldsymbol{w}^*=\boldsymbol{w}} &= -\sum_i \frac{(a_{is}-a_{iN})^2}{b_i^2} \\
\left[\frac{\partial^2 Q(\boldsymbol{w}^*|\boldsymbol{w})}{\partial w_s^* \partial w_t^*} + \frac{\partial^2 Q(\boldsymbol{w}^*|\boldsymbol{w})}{\partial w_s^* \partial w_t}\right]_{\boldsymbol{w}^*=\boldsymbol{w}} &= -\sum_i \frac{(a_{is}-a_{iN})(a_{it}-a_{iN})}{b_i^2}
\end{aligned}$$

The covariance matrix can be computed as the inverse the information matrix of $\boldsymbol{w}$.

$$\hat{\Sigma}_{\boldsymbol{w}} = I^{-1}(\boldsymbol{w})$$

As we have

$$\begin{aligned}
E(\hat{\tau}^2) &= E\left[\sum_{l=1}^{N} w_l(\xi_l^* - \hat{\theta})^2\right] \\
&= \sum_{l=1}^{N} E(w_l)\xi_l^{*2} + E[\hat{\theta}^2] - 2E[\hat{\theta})^2] \\
&= \tau^2 + var(\hat{\theta})
\end{aligned}$$

We can adjust the estimate of τ^2 by

$$\begin{aligned} \tau^2 &= \hat{\tau}^2 - var(\hat{\theta}) \\ &= \hat{\tau}^2 - \frac{\partial g(\boldsymbol{w})}{\partial \boldsymbol{w}^T}|_{\boldsymbol{w}=\hat{\boldsymbol{w}}} \hat{\Sigma}_{\hat{\boldsymbol{w}}} \frac{\partial g(\boldsymbol{w})}{\partial \boldsymbol{w}}|_{\boldsymbol{w}=\hat{\boldsymbol{w}}} \end{aligned}$$

The 95% confidence interval of θ can be obtained as follows:

$$\left[\hat{\theta} - 1.96\sqrt{var(\hat{\theta})}, \hat{\theta} + 1.96\sqrt{var(\hat{\theta})}\right]$$

Table 3.1: Bias of θ and τ^2. We assume $\theta_i \sim N(\theta, \tau^2)$, with $\theta = 0.5$. The results are based on 1000 replicates.

		Inference of θ				Inference of τ^2			
		EM		DL		EM		DL	
K	τ^2	$Bias_{\hat{\theta}}$	$se_{\hat{\theta}}$	$Bias_{\hat{\theta}}$	$se_{\hat{\theta}}$	$Bias_{\hat{\tau}^2}$	$se_{\hat{\tau}^2}$	$Bias_{\hat{\tau}^2}$	$se_{\hat{\tau}^2}$
10	0.25	-0.006	0.191	0.004	0.189	-0.009	0.161	-0.019	0.154
10	0.50	0.007	0.249	0.018	0.249	-0.010	0.280	-0.035	0.266
10	0.75	0.002	0.293	0.013	0.294	-0.022	0.404	-0.091	0.344
10	1.00	0.015	0.334	0.023	0.334	-0.032	0.520	-0.142	0.451
10	1.25	0.013	0.358	0.020	0.354	-0.065	0.617	-0.228	0.525
10	1.50	0.003	0.401	0.015	0.392	-0.099	0.711	-0.325	0.577
10	1.75	0.007	0.427	0.025	0.423	-0.058	0.919	-0.386	0.654
10	2.00	0.015	0.449	0.035	0.447	-0.128	0.940	-0.498	0.715
30	0.25	-0.003	0.105	0.006	0.105	-0.004	0.100	-0.018	0.091
30	0.50	0.003	0.143	0.014	0.142	-0.004	0.176	-0.047	0.147
30	0.75	-0.006	0.171	0.006	0.170	0.009	0.244	-0.093	0.193
30	1.00	0.002	0.187	0.010	0.188	0.001	0.325	-0.163	0.244
30	1.25	0.000	0.208	0.014	0.205	-0.004	0.387	-0.253	0.275
30	1.50	0.016	0.233	0.030	0.228	-0.040	0.449	-0.352	0.316
30	1.75	0.018	0.253	0.028	0.250	-0.029	0.503	-0.441	0.360
30	2.00	0.015	0.262	0.032	0.256	-0.085	0.563	-0.556	0.397
50	0.25	-0.002	0.082	0.007	0.081	0.004	0.078	-0.011	0.066
50	0.50	-0.001	0.111	0.008	0.109	0.000	0.138	-0.050	0.113
50	0.75	0.013	0.137	0.025	0.136	0.015	0.202	-0.092	0.154
50	1.00	0.010	0.150	0.022	0.146	0.016	0.260	-0.165	0.192
50	1.25	0.010	0.164	0.020	0.163	0.009	0.306	-0.254	0.217
50	1.50	0.001	0.182	0.016	0.177	0.018	0.358	-0.332	0.248
50	1.75	0.007	0.195	0.020	0.193	-0.015	0.415	-0.460	0.270
50	2.00	0.008	0.206	0.023	0.203	-0.015	0.457	-0.576	0.293
80	0.25	-0.005	0.068	0.004	0.068	0.006	0.063	-0.014	0.053
80	0.50	-0.002	0.090	0.008	0.090	0.011	0.115	-0.049	0.088
80	0.75	0.004	0.109	0.013	0.107	0.019	0.158	-0.101	0.118
80	1.00	0.007	0.122	0.019	0.120	0.023	0.208	-0.167	0.148
80	1.25	0.008	0.125	0.021	0.122	0.021	0.249	-0.265	0.165
80	1.50	0.020	0.150	0.032	0.143	0.029	0.295	-0.349	0.190
80	1.75	0.012	0.148	0.025	0.146	0.010	0.331	-0.459	0.217
80	2.00	0.009	0.163	0.026	0.157	-0.004	0.367	-0.565	0.233

Table 3.2: Bias of θ and τ^2. We assume $\theta_i \sim Laplace(\theta, b)$ with $\theta = 0.5$. The results are based on 1000 replicates.

		Inference of θ				Inference of τ^2			
		EM		DL		EM		DL	
K	τ^2	$Bias_{\hat{\theta}}$	$se_{\hat{\theta}}$	$Bias_{\hat{\theta}}$	$se_{\hat{\theta}}$	$Bias_{\hat{\tau}^2}$	$se_{\hat{\tau}^2}$	$Bias_{\hat{\tau}^2}$	$se_{\hat{\tau}^2}$
10	0.25	-0.008	0.188	0.002	0.184	-0.001	0.223	-0.023	0.192
10	0.50	0.005	0.240	0.014	0.240	0.004	0.421	-0.082	0.306
10	0.75	0.008	0.291	0.016	0.266	0.004	0.602	-0.156	0.418
10	1.00	0.005	0.330	0.014	0.311	-0.043	0.752	-0.246	0.493
10	1.25	0.007	0.366	0.023	0.352	-0.109	0.807	-0.384	0.506
10	1.50	0.006	0.394	0.032	0.356	-0.155	0.966	-0.525	0.596
10	1.75	0.013	0.415	0.031	0.375	-0.224	1.088	-0.636	0.637
10	2.00	0.029	0.441	0.001	0.416	-0.239	1.173	-0.797	0.667
30	0.25	-0.003	0.108	0.006	0.109	0.015	0.139	-0.023	0.106
30	0.50	-0.006	0.138	0.009	0.134	-0.004	0.233	-0.095	0.165
30	0.75	0.010	0.175	0.011	0.165	0.014	0.363	-0.185	0.208
30	1.00	0.008	0.194	0.028	0.180	-0.012	0.428	-0.291	0.251
30	1.25	0.010	0.209	0.021	0.193	-0.003	0.551	-0.424	0.286
30	1.50	0.013	0.229	0.033	0.218	-0.013	0.641	-0.572	0.320
30	1.75	0.014	0.239	0.027	0.223	-0.050	0.690	-0.706	0.350
30	2.00	0.010	0.252	0.027	0.233	-0.092	0.761	-0.870	0.359
50	0.25	-0.002	0.086	0.004	0.084	0.017	0.121	-0.025	0.082
50	0.50	0.004	0.117	0.009	0.106	0.035	0.210	-0.098	0.123
50	0.75	-0.002	0.136	0.017	0.126	0.042	0.291	-0.193	0.171
50	1.00	0.006	0.152	0.015	0.138	0.023	0.372	-0.307	0.188
50	1.25	0.003	0.163	0.023	0.157	0.026	0.440	-0.443	0.211
50	1.50	0.010	0.183	0.016	0.164	0.011	0.488	-0.586	0.239
50	1.75	0.020	0.193	0.025	0.182	-0.024	0.569	-0.719	0.253
50	2.00	0.014	0.214	0.029	0.179	-0.061	0.592	-0.887	0.277
80	0.25	-0.006	0.071	0.007	0.065	0.024	0.104	-0.029	0.065
80	0.50	-0.001	0.094	0.013	0.084	0.043	0.183	-0.095	0.099
80	0.75	-0.003	0.110	0.011	0.099	0.054	0.251	-0.190	0.131
80	1.00	0.000	0.127	0.014	0.116	0.083	0.317	-0.303	0.155
80	1.25	0.006	0.142	0.020	0.115	0.062	0.358	-0.432	0.170
80	1.50	0.008	0.148	0.025	0.134	0.027	0.400	-0.570	0.189
80	1.75	0.010	0.157	0.030	0.137	-0.015	0.452	-0.727	0.199
80	2.00	0.012	0.167	0.031	0.151	-0.060	0.488	-0.892	0.222

Table 3.3: Bias of θ and τ^2.We assume $\theta_i \sim Logistic(\theta, s)$ with $\theta = 0.5$. The results are based on 1000 replicates.

		Inference of θ				Inference of τ^2			
		EM		DL		EM		DL	
K	τ^2	$Bias_{\hat{\theta}}$	$se_{\hat{\theta}}$	$Bias_{\hat{\theta}}$	$se_{\hat{\theta}}$	$Bias_{\hat{\tau}^2}$	$se_{\hat{\tau}^2}$	$Bias_{\hat{\tau}^2}$	$se_{\hat{\tau}^2}$
10	0.25	-0.006	0.193	0.003	0.192	0.001	0.191	-0.017	0.170
10	0.50	0.009	0.255	0.019	0.241	-0.020	0.321	-0.058	0.280
10	0.75	0.005	0.282	0.012	0.289	-0.047	0.460	-0.105	0.387
10	1.00	0.006	0.334	0.014	0.322	-0.073	0.595	-0.178	0.454
10	1.25	0.004	0.373	0.021	0.343	-0.028	0.745	-0.300	0.523
10	1.50	-0.007	0.403	0.026	0.390	-0.048	0.857	-0.422	0.598
10	1.75	0.032	0.422	0.041	0.401	-0.123	0.946	-0.529	0.641
10	2.00	0.019	0.466	0.033	0.416	-0.158	1.047	-0.678	0.689
30	0.25	-0.008	0.109	0.014	0.109	0.001	0.111	-0.021	0.096
30	0.50	0.006	0.149	0.010	0.140	-0.001	0.212	-0.077	0.145
30	0.75	0.009	0.169	0.023	0.164	0.003	0.286	-0.145	0.208
30	1.00	0.011	0.191	0.021	0.188	0.009	0.375	-0.229	0.235
30	1.25	0.005	0.215	0.017	0.207	-0.017	0.441	-0.322	0.278
30	1.50	0.015	0.227	0.025	0.225	-0.018	0.533	-0.438	0.307
30	1.75	0.001	0.245	0.026	0.240	-0.025	0.597	-0.561	0.351
30	2.00	0.013	0.251	0.027	0.244	-0.056	0.667	-0.699	0.376
50	0.25	-0.002	0.085	0.006	0.082	0.009	0.098	-0.022	0.072
50	0.50	-0.004	0.112	0.006	0.110	0.011	0.166	-0.069	0.120
50	0.75	0.009	0.137	0.014	0.132	0.036	0.237	-0.139	0.156
50	1.00	0.008	0.150	0.017	0.139	0.020	0.304	-0.226	0.187
50	1.25	0.011	0.163	0.021	0.154	0.024	0.350	-0.327	0.213
50	1.50	0.010	0.176	0.022	0.168	0.010	0.436	-0.436	0.251
50	1.75	0.025	0.192	0.029	0.181	-0.009	0.493	-0.572	0.268
50	2.00	0.015	0.196	0.030	0.191	-0.038	0.528	-0.702	0.288
80	0.25	-0.001	0.066	0.007	0.067	0.013	0.081	-0.021	0.055
80	0.50	-0.005	0.087	0.007	0.085	0.033	0.148	-0.072	0.094
80	0.75	0.000	0.105	0.018	0.106	0.029	0.196	-0.141	0.123
80	1.00	0.007	0.124	0.018	0.115	0.045	0.276	-0.226	0.149
80	1.25	0.011	0.132	0.027	0.125	0.041	0.292	-0.332	0.169
80	1.50	-0.004	0.148	0.025	0.130	0.045	0.349	-0.444	0.188
80	1.75	0.004	0.153	0.027	0.149	0.022	0.394	-0.578	0.207
80	2.00	0.009	0.162	0.035	0.156	-0.005	0.423	-0.701	0.219

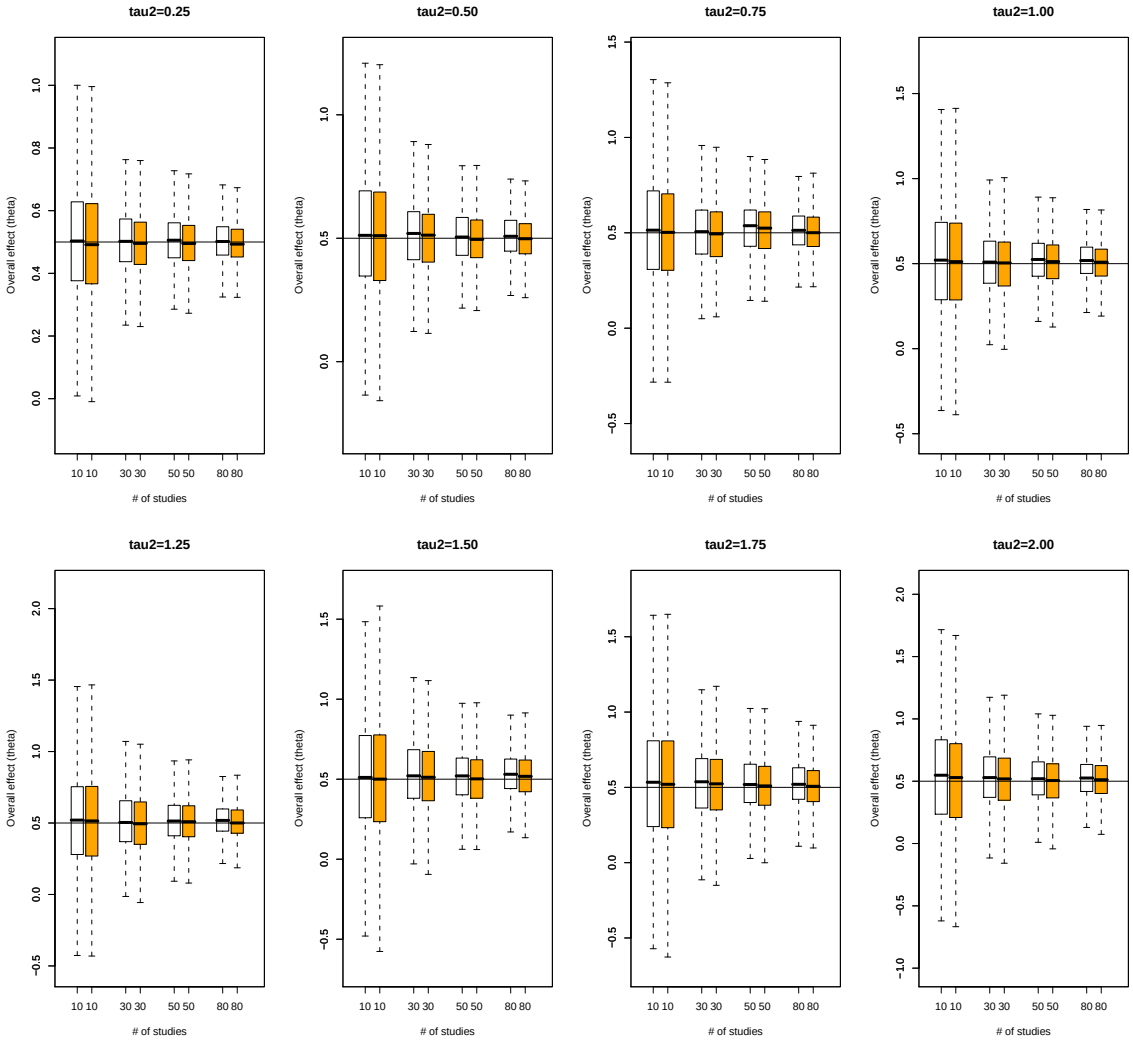

Figure 3.1: We assume $\theta_i \sim N(\theta, \tau^2)$. The white box is based on the estimates by DSL method and the orange box based on the proposed method. The horizontal line is the true overall effect, which is 0.5.

Table 3.4: Genotype and allele frequencies of C957T polymorphism in the DRD2 gene in the patient and control groups of six studies.

Study	Group	Genotypes			Alleles	
		C/C	C/T	T/T	C	T
Lawford, 2005	patient (n=153)	48	75	31	171	137
	control (n=148)	27	70	51	124	172
Hoenicka, 2006	patient (n=131)	30	61	40	121	141
	control (n=364)	46	174	144	266	462
Kukreti, 2006	patient (n=101)	41	38	22	120	82
	control (n=145)	48	64	33	160	130
Hanninen, 2006	patient (n=188)	59	92	37	210	166
	control (n=384)	104	176	104	384	384
Sanders, 2008	patient (n=1870)	411	931	527	1753	1985
	control (n=2002)	435	996	571	1866	2138
Monakhov, 2008	patient (n=311)	99	152	60	350	272
	control (n=364)	78	183	103	339	389

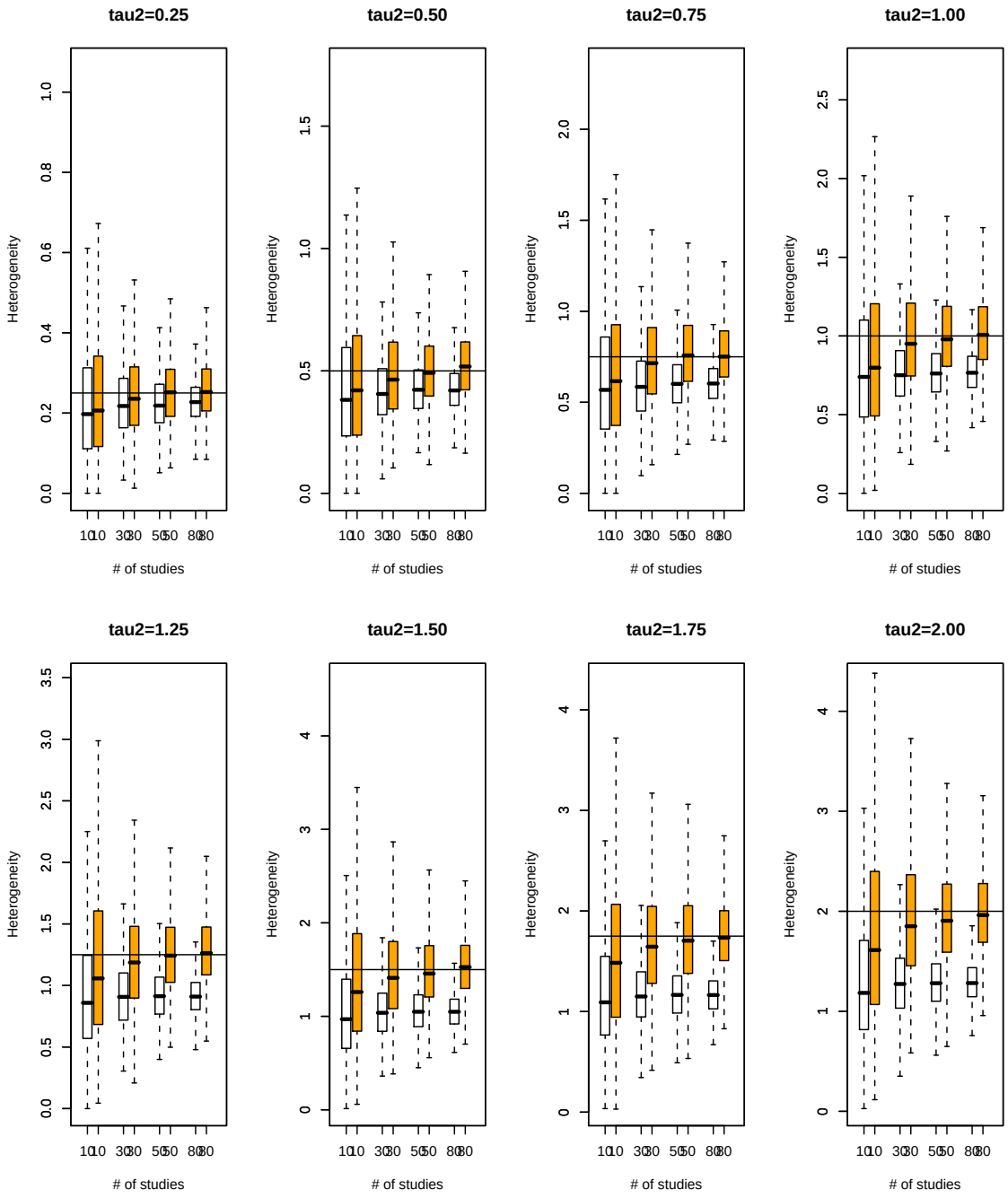

Figure 3.2: We assume $\theta_i \sim N(\theta, \tau^2)$. The white box is based on the estimates by DSL method and the orange box based on the proposed method. The horizontal line is the true heterogeneous variance.

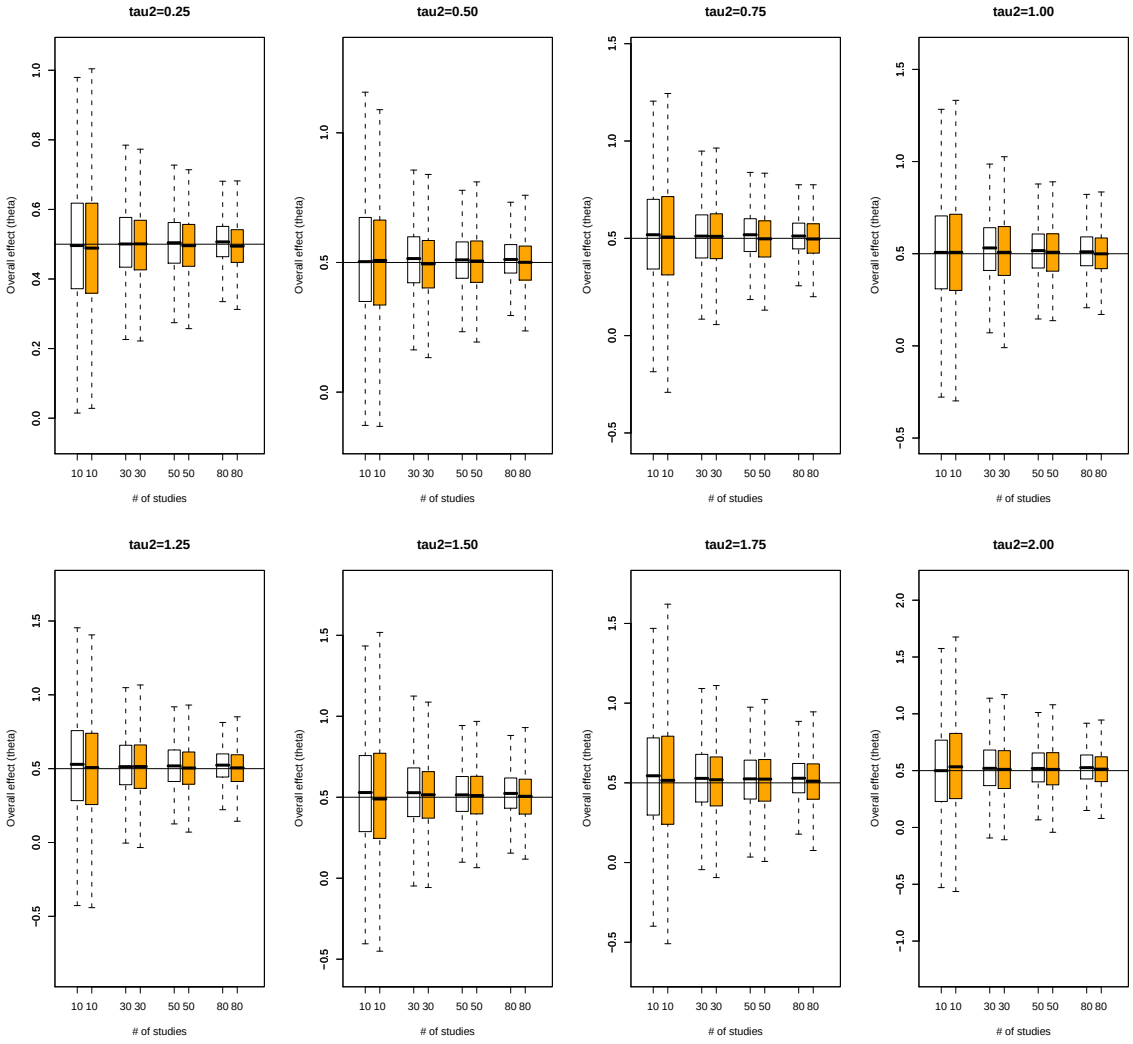

Figure 3.3: We assume $\theta_i \sim Laplace(\theta, b)$. The white box is based on the estimates by DSL method and the orange box based on the proposed method. The horizontal line is the true overall effect, which is 0.5.

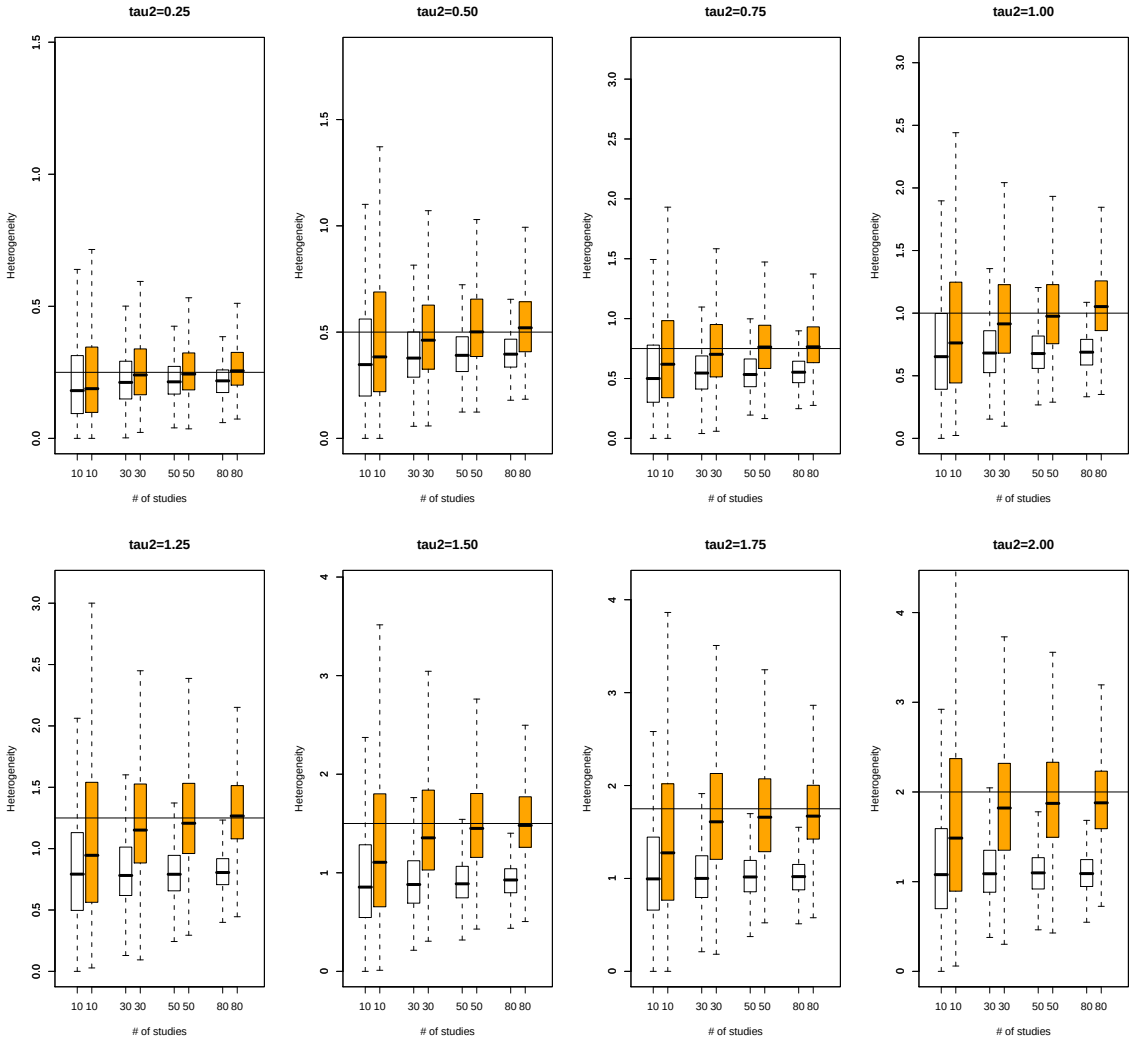

Figure 3.4: We assume $\theta_i \sim Laplace(\theta, b)$. The white box is based on the estimates by DSL method and the orange box based on the proposed method. The horizontal line is the true heterogeneous variance.

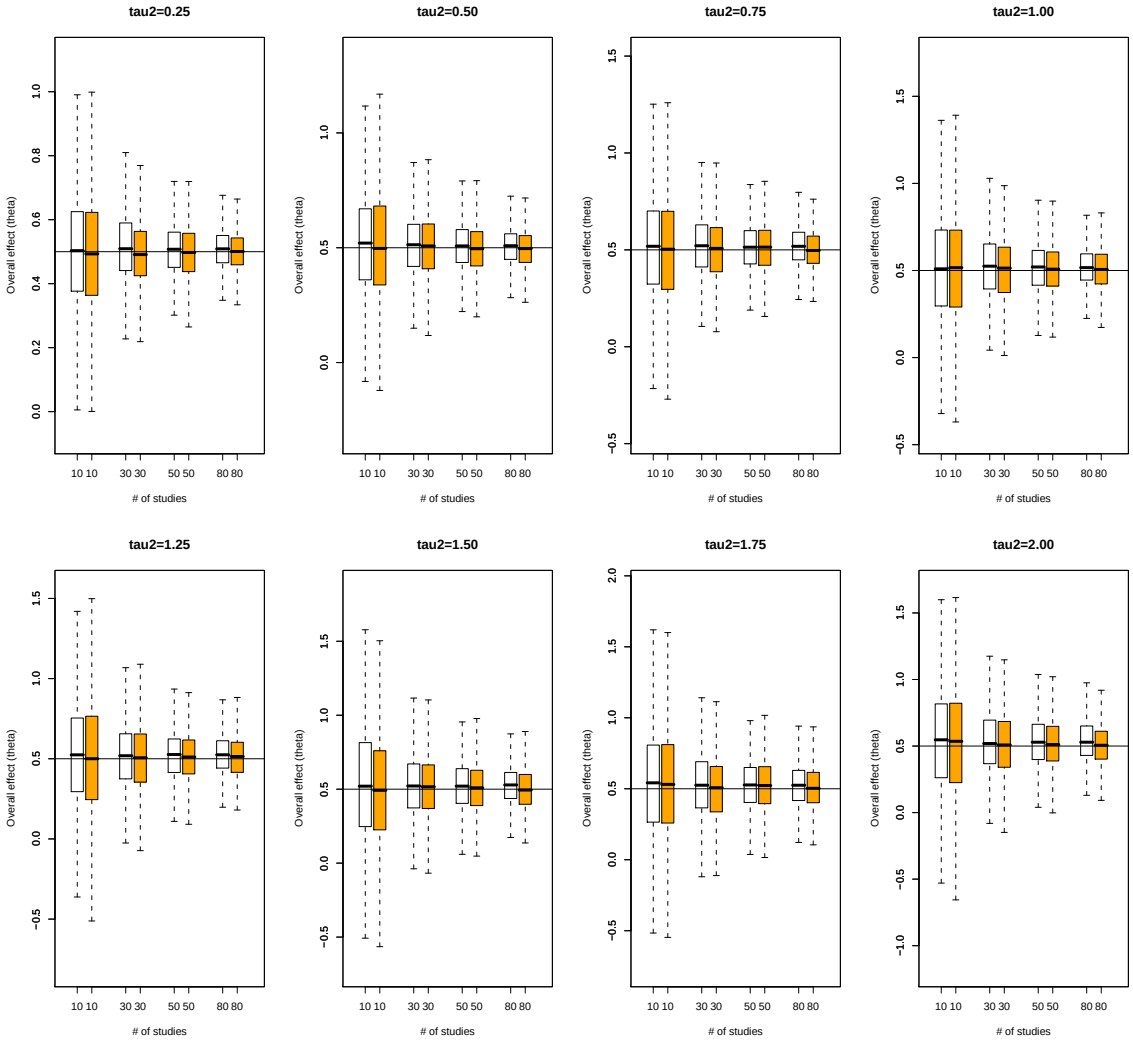

Figure 3.5: We assume $\theta_i \sim Logistic(\theta, s)$. The white box is based on the estimates by DSL method and the orange box based on the proposed method. The horizontal line is the true overall effect, which is 0.5.

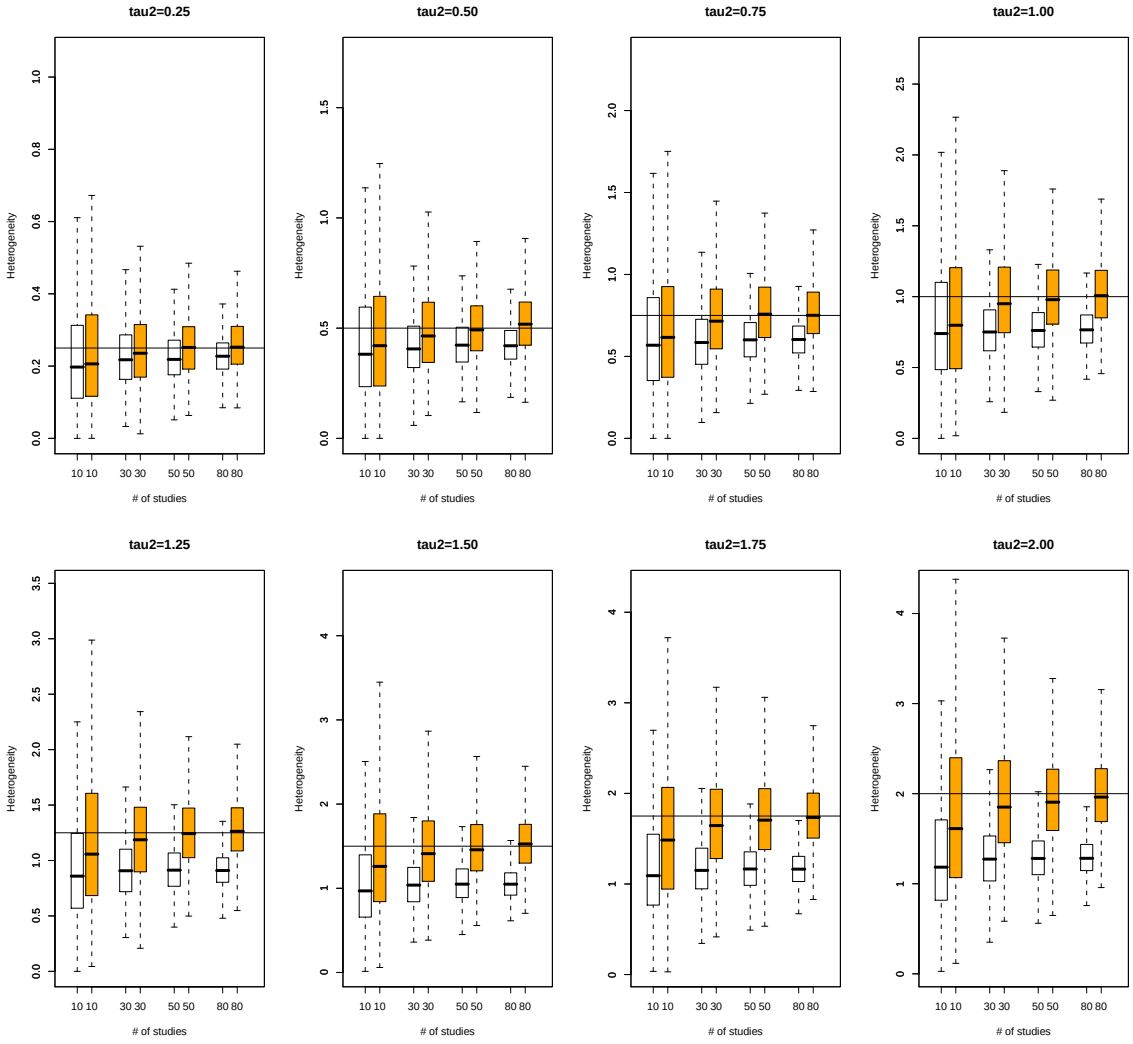

Figure 3.6: We assume $\theta_i \sim Logistic(\theta, s)$. The white box is based on the estimates by DSL method and the orange box based on the proposed method. The horizontal line is the true heterogeneous variance.

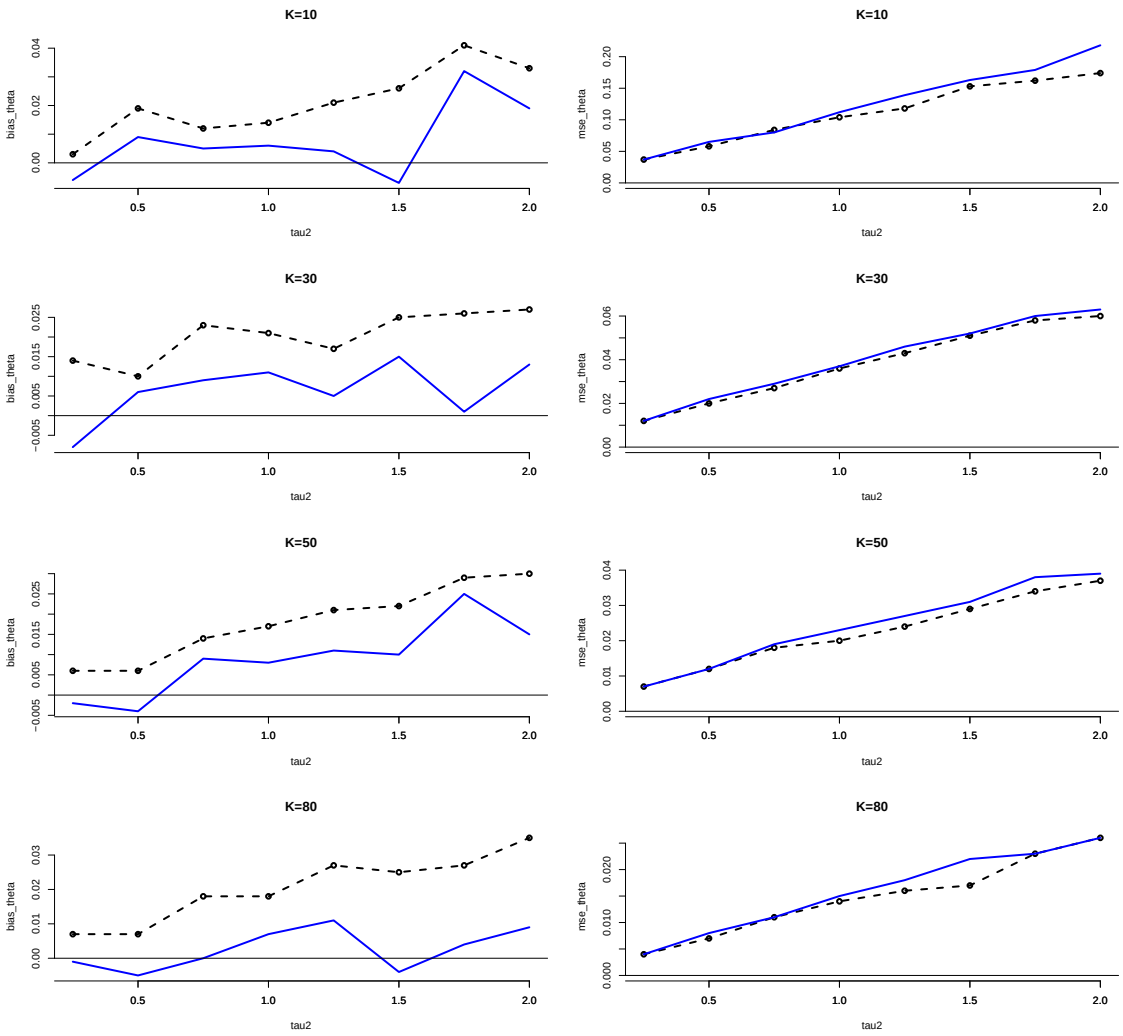

Figure 3.7: Plots of bias and MSE of θ with normality assumption. The blue solid line is by the proposed method and the dash line is DSL method.

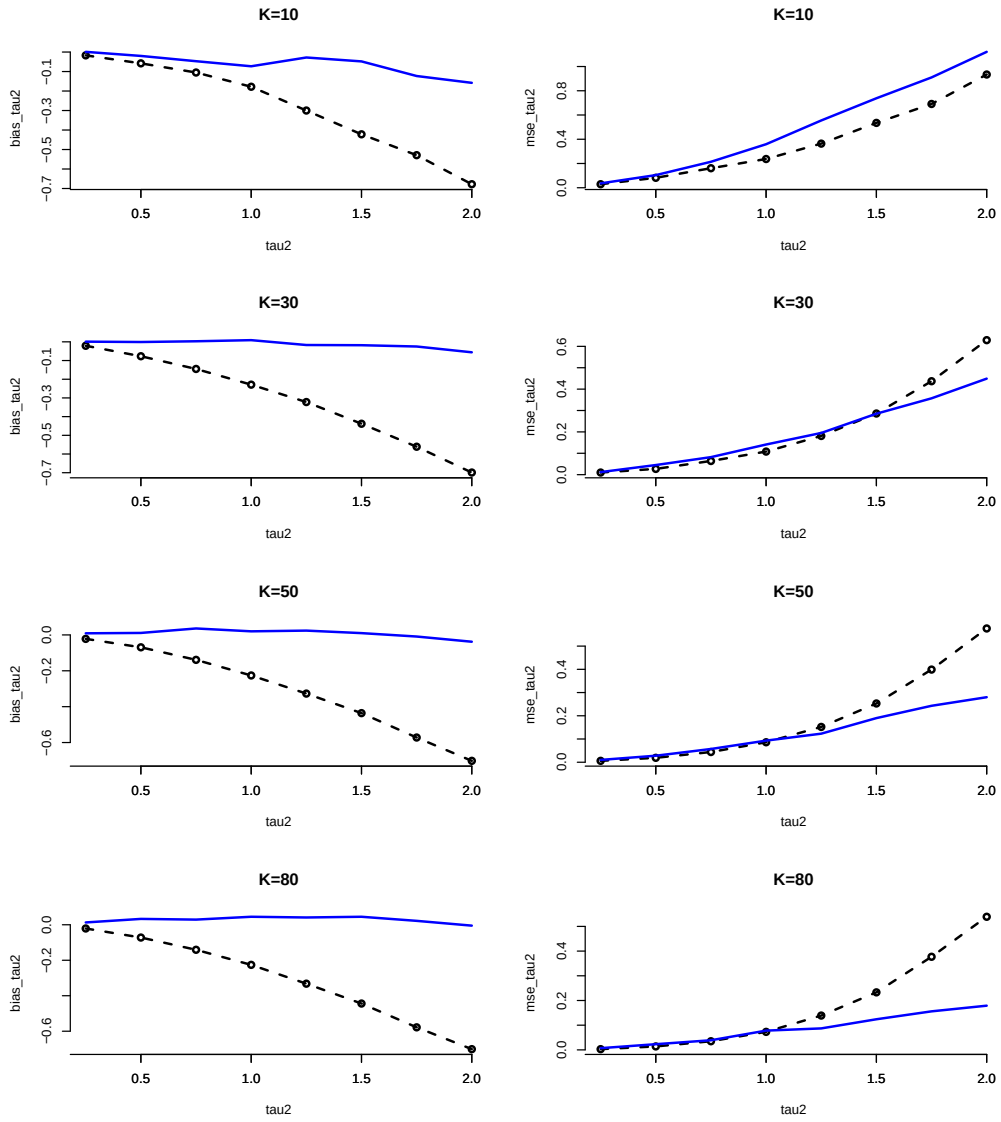

Figure 3.8: Plots of bias and MSE of τ^2 with normality assumption. The blue solid line is by the proposed method and the dash line is DSL method.

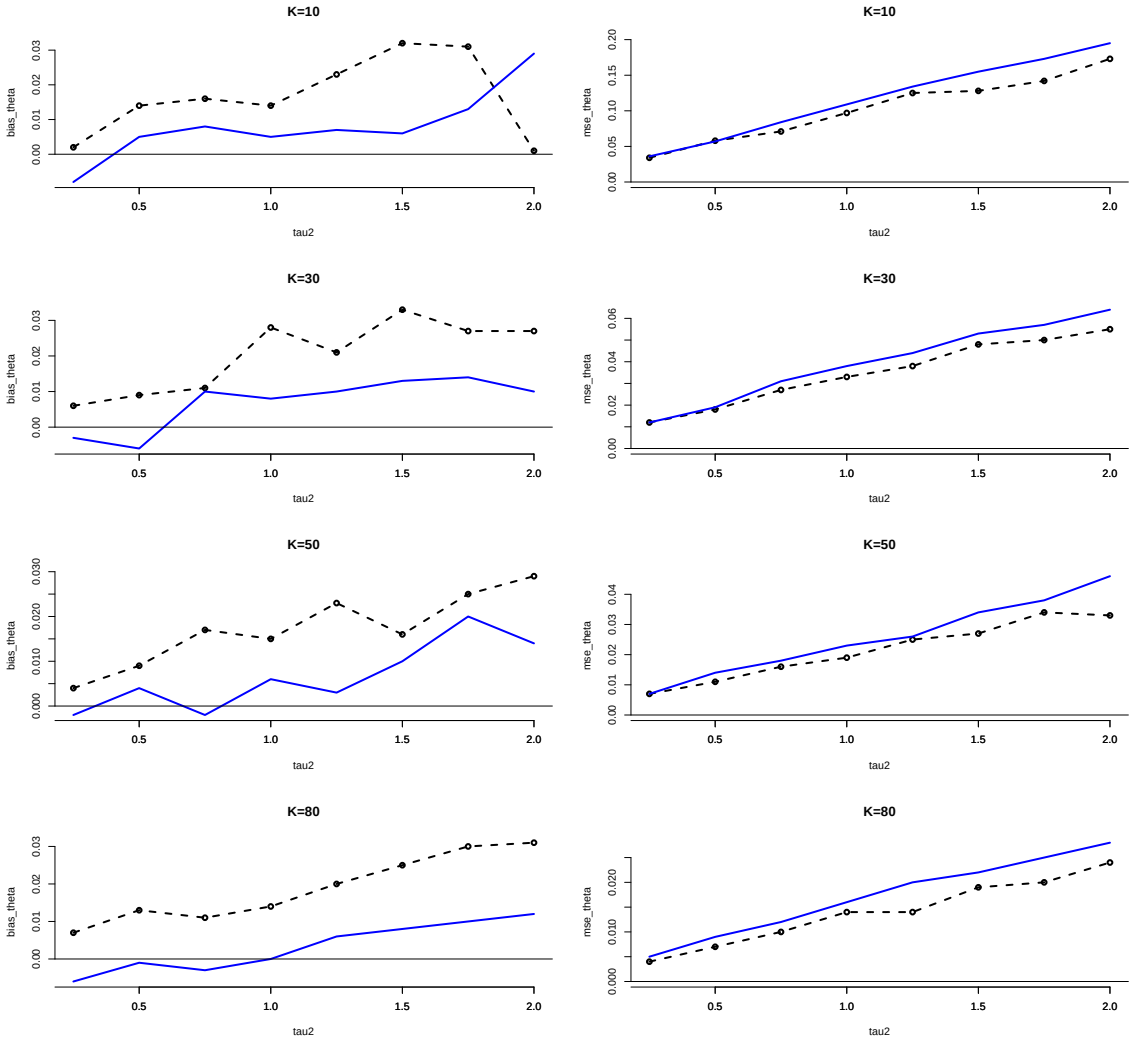

Figure 3.9: Plots of bias and MSE of θ with Laplace assumption. The blue solid line is by the proposed method and the dash line is DSL method.

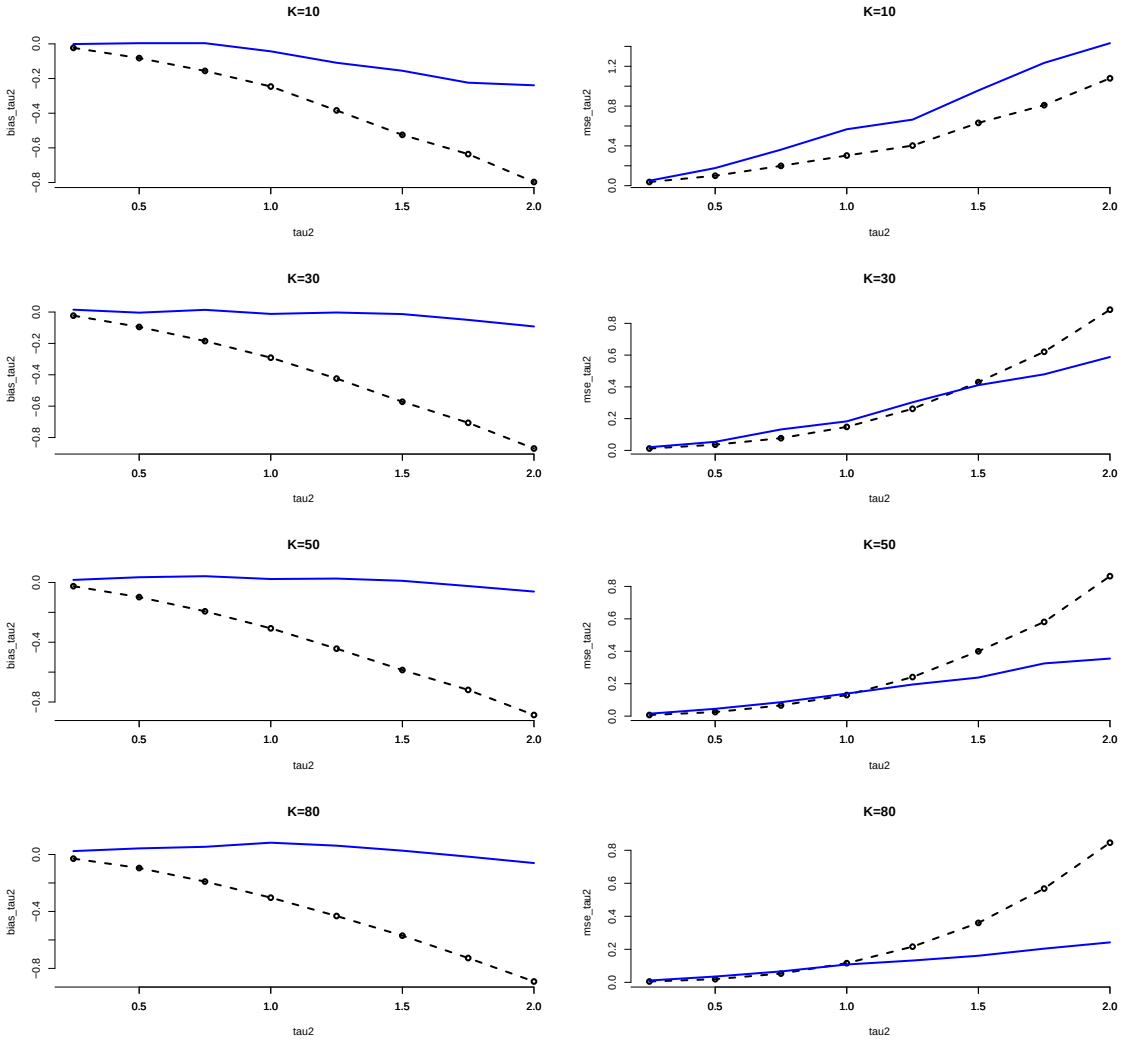

Figure 3.10: Plots of bias and MSE of τ^2 with Laplace assumption. The blue solid line is by the proposed method and the dash line is DSL method.

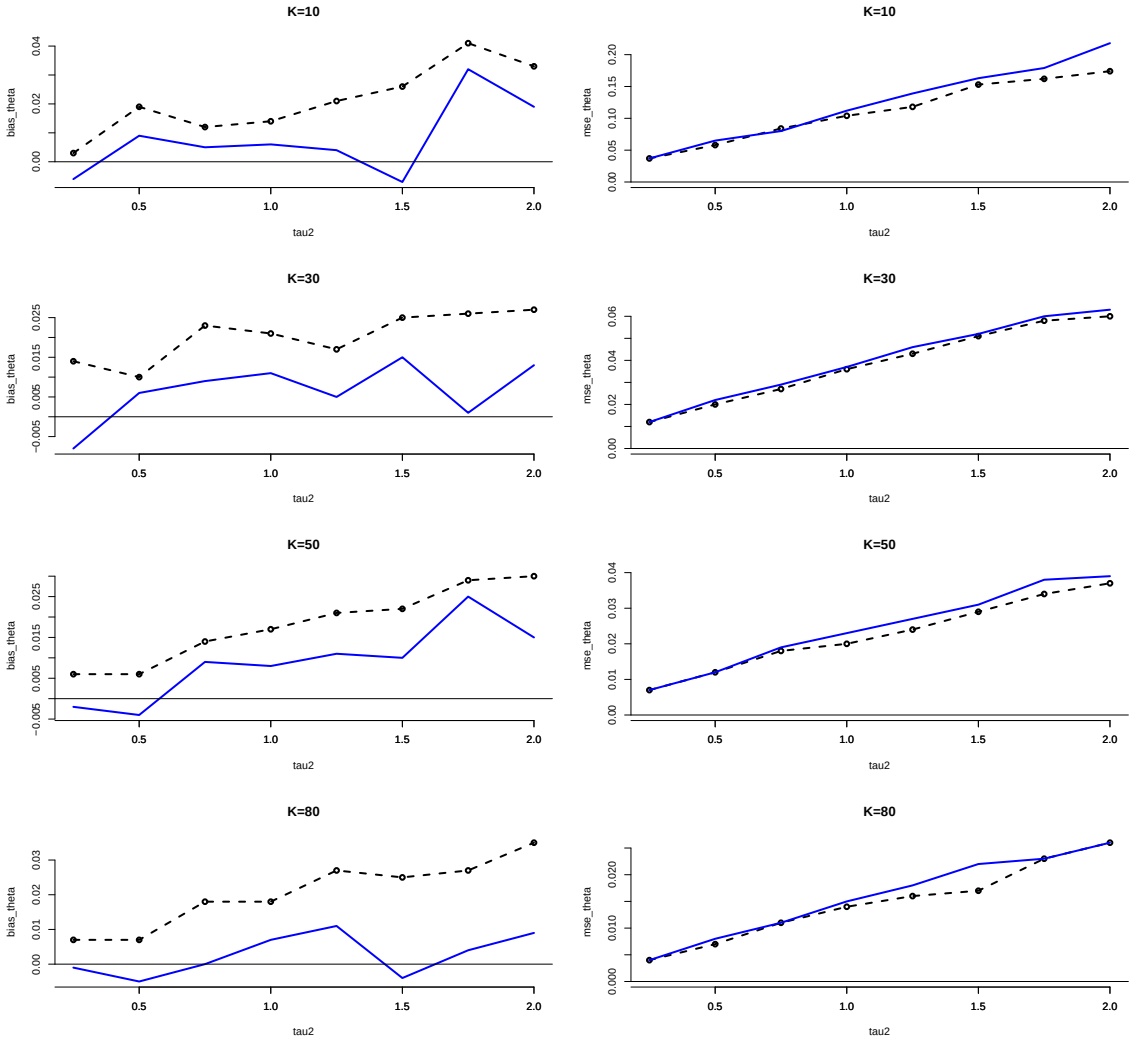

Figure 3.11: Plots of bias and MSE of θ with logistic assumption. The blue solid line is by the proposed method and the dash line is DSL method.

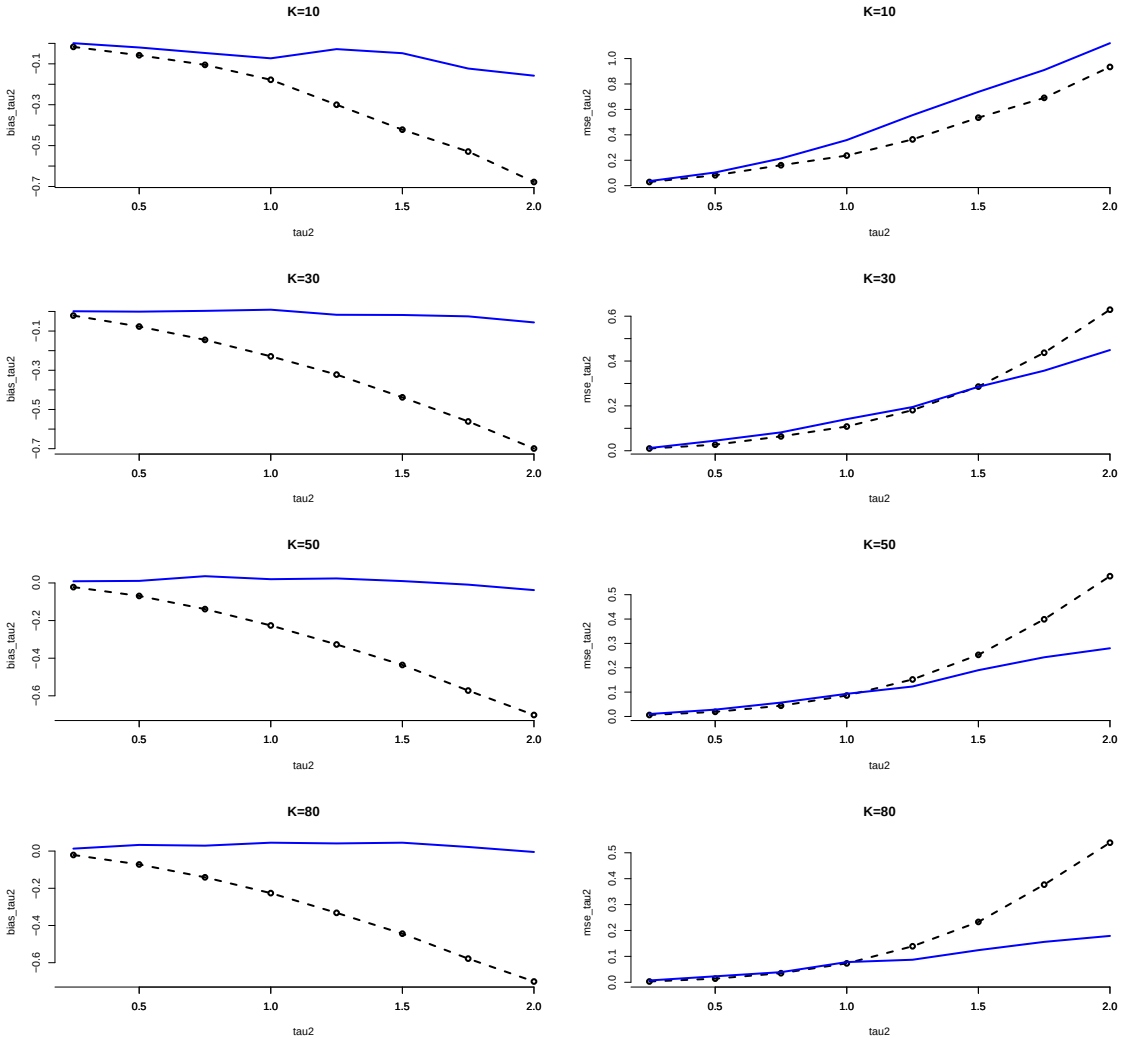

Figure 3.12: Plots of bias and MSE of τ^2 with logistic assumption. The blue solid line is by the proposed method and the dash line is DSL method.

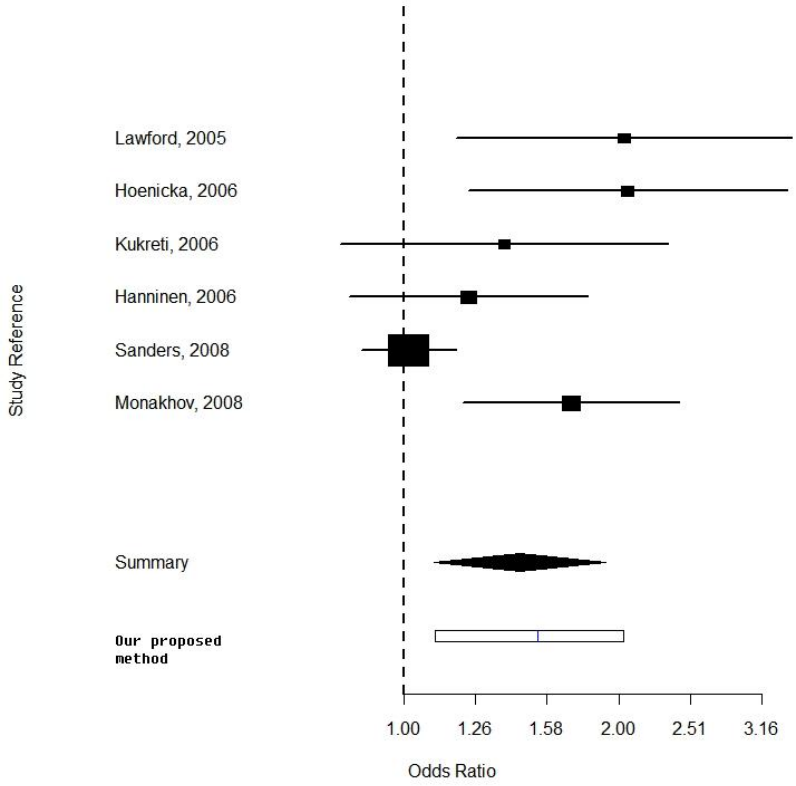

Figure 3.13: Analysis genotype frequency (C/C vs. C/T and T/T) of studies for C957T polymorphism in DRD2 gene associated with schizophrenia. The diamond indicates summary odds ratio (OR) and 95% confidence interval (CI). Studies are in chronological order. The size of the squares is inversely proportional to the variance of the studies. The bar is the 95% confidence interval by our method; the short solid line in the middle is the mean of OR estimated by our method.

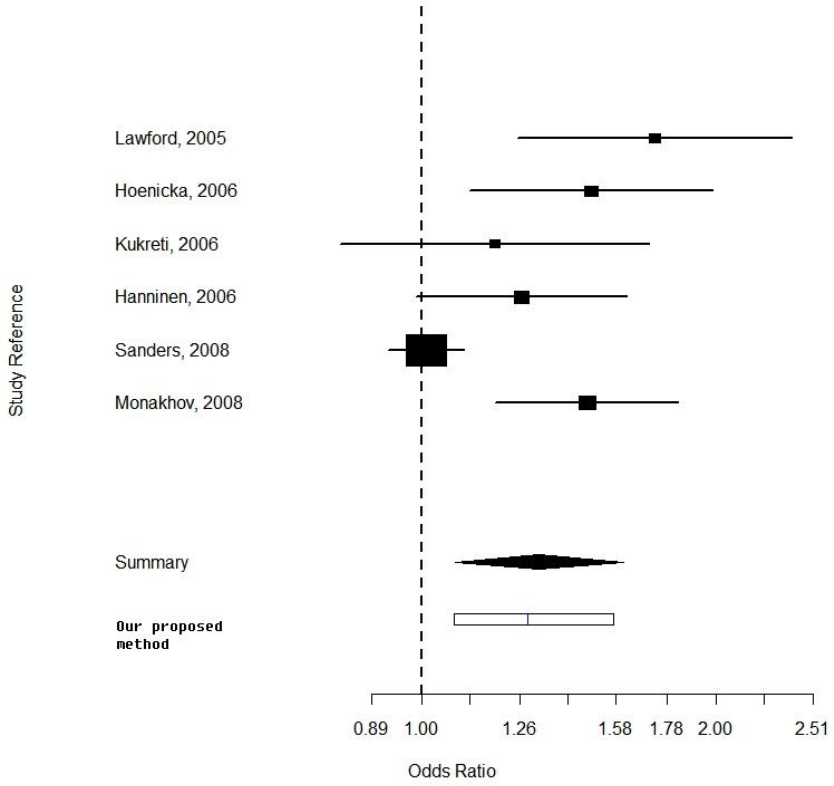

Figure 3.14: Analysis allele frequency (C vs. T) of studies on C957T DRD2 polymorphism associated with schizophrenia. The diamond indicates summary odds ratio (OR) and 95% confidence interval (CI). Studies are in chronological order. The size of the squares is inversely proportional to the variance of the studies. The bar is the 95% confidence interval by our method; the short solid line in the middle is the mean of OR estimated by our method.

Chapter 4

Effects of missing and censored data for Non Linear models involving ODEs

4.1 Abstract

The *Bayesian Euler's Approximation Method* (BEAM) has recently been proposed to estimate the parameters in a non-linear model involving ODEs, especially when analytical closed form solutions to the ODEs are not available. In this article,the *BEAM* is extended to handle datasets with missing or censored observations. The proposed method is based on data augmentation algorithm. A simulation study based on growth colonies of paramecium aurelium is presented to compare the performances of the proposed method for various percentages of missing and censored cases to the results based on completed data cases. Finally, the method is illustrated with a real data obtained from AIDS Clinical Trials Group Protocol 315.

4.2 Introduction

In biomedical sciences, computational biology and many other research fields, scientists often use ordinary differential equations (ODEs) to model various kinds of dynamic systems in order to better understand the underlying mechanism. For example, in HIV virus study, Ho et al. (1995) used ODEs to analyze the dynamics of HIV viral load measurements on AIDS patients. In most cases, the system of ODE is complicated, as the ODEs are nonlinear functions of parameters. In such cases, it is difficult to obtain the analytical closed form solution to the ODEs system.

In the literature, many statistical methods have been proposed to solve those kinds of nonlinear models with ODEs. These methods can be classified into two groups: (i) analytic solution of ODEs is available and hence, methods to estimate parameters of such nonlinear regression are available; (ii) analytical solution is not available and the numerical solution is needed for a given set of parameters. In both cases Bayesian methods (Gelman et al., 1996; Walkefield, 1996; Lunn et al., 2002; Putter et al., 2002) or maximum likelihood methods (Davidian and Giltinan, 1995; Racine-Poon and Wakefield 1998) are used to obtain parameter estimates.

Unfortunately, those methods turn out to be unstable in cases of censored or missing data (Putter et al., 2002). For missing data, Rubin (1976) classified them to three cases: *missing completely at random* (MCAR), *missing at random* (MAR) and *informatively missing or nonignorable missing* (IM). Generally, missing data poses many challenges for data analysis. Several methods were proposed by Gelman, Carlin, Stern and Rubin (1995), Little and Rubin (1987) and Schafer (1997). MCMC methods, such as Gibbs sampling (Geman and Geman 1984), data augmentation (Tanner 1996), exact Monte Carlo (Forster, McDonald, and Smith 1996) and some other methods are developed to impute missing observations and make inference about the parameters based on imputed data.

Censored data is another problem that is typically encountered in medical data. Two general approaches are used to handle censored data. One way is to discard the censored part as missing data and use the rest of the data to analyze the study.

The other way is to treat the censored values as observed data. But both of these naive approaches lead to biased and inefficient estimators. The corresponding results may not be reliable. Robinson (1980) proposed an imputation method for the censored data with its conditional expectation given the completely observed part. Pseudo-likelihood method (Sun and Kalbfleisch, 1995) was proposed when the censoring rate is very high. Hopke, Liu and Rubin (2001) used multiple imputation within a Bayesian framework. However, all of the previous methods require fully specified likelihood function, which is not usually available for an ODE based model.

In this paper, we extend *BEAM* (Goyal and Ghosh, 2006) to missing data and censored data cases. In Section 4.3, we briefly describe *BEAM* for nonlinear models with ODEs. In Section 4.4, we incorporate a data augmentation method (Tanner, 1996) into a MCMC method to deal with missing and censored observations. In Section 4.5, a simulation study is presented to illustrate the performance of the method for missing and censored observations and the results are compared to those obtained from complete data. In Section 4.6, the proposed method is illustrated by using a real data on AIDS clinical trial and the estimated parameters are reported.

4.3 Nonlinear models involving ODEs

Let y_j denote the response variable measured at time t_j for $j = 1, \cdots, n$. A regression model to study the dynamics of response can be written as follows:

$$y_j = \mu(t_j, \boldsymbol{\theta}) + \epsilon_j \tag{4.1}$$

where $j = 1, 2, \ldots, n$ and $\epsilon_j \overset{\text{i.i.d}}{\sim} N(0, \sigma_j^2)$ are independent errors. In (4.1) $\mu(\cdot)$ is the mean function, which describes the population dynamics of the response, and is a function of unknown parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$.

In many biomedical applications (e.g. PK/PD models, HIV dynamics etc.), A system of ODEs are used to describe the dynamics of the system. Assume that $\mu(\cdot) = H(\nu(\cdot))$, where H is a known function, $H : \mathbf{R}^q \to \mathbf{R}$ and the system is described by a set of q compartments $\boldsymbol{\nu}(t) = (\nu_1(t), \cdots, \nu_q(t))^T$, which can be expressed as a

solution of the following ODEs.

$$\begin{aligned} \frac{d\boldsymbol{\nu}_l}{dt} &= \boldsymbol{g}_l(\boldsymbol{\nu}(t), \boldsymbol{\theta}) \text{ and} \\ \boldsymbol{\nu}_l(t_0, \boldsymbol{\theta}) &= \nu_{0l}(\boldsymbol{\theta}), \qquad \text{for } l = 1, \cdots, q \end{aligned} \tag{4.2}$$

where $\boldsymbol{\nu}_0(\cdot)$ is the initial value, which may or may not be completely known. The q-vector valued function $\boldsymbol{g}(\cdot) = (\boldsymbol{g}_1(\cdot), \ldots, \boldsymbol{g}_q(\cdot))^T$ describes the dynamics of the system and are assumed to known but for the unknown parameter $\boldsymbol{\theta}$.

The random error ϵ_j's corresponds to the measurement uncertainties associated with the observed data at time t_j. We assume all these errors are independently distributed with mean 0 and variance that may depend on unknown parameter $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_k)^T$. In other words,

$$E(\epsilon_j) = 0 \text{ and } Var(\epsilon_j) = \sigma^2(t_j, \boldsymbol{\eta}) \tag{4.3}$$

where $\sigma^2(t_j, \boldsymbol{\eta})$ is the variance function and $\boldsymbol{\eta}$ is an unknown parameter. The goal is to obtain the estimate of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ based on the data $\{(y_j, t_j) : j = 1, \cdots, n\}$. In this article, we assume that the variance function is analytically known but for the parameter $\boldsymbol{\eta}$. However, in many applications, we often choose $\sigma^2(t_j, \boldsymbol{\theta}) = \mu^2(t_j, \boldsymbol{\theta})$ or $\sigma^2(t_j, \boldsymbol{\theta}, \boldsymbol{\eta}) = \exp\{\eta_1 + \eta_2 \mu(t_j, \boldsymbol{\theta})\}$. In such case, $\sigma^2(t_j, \boldsymbol{\eta})$ is also not available in closed form. If we have the analytic closed form for the $\boldsymbol{\nu}(\cdot)$, then we can use existing software package like *nlin* or *nlmixed* in SAS to estimate parameters. But in most cases, the analytic closed form for the system of ODEs is not available. Mathematicians have developed many numerical algorithms to solve the system of ODEs (Iserles, 1996). But almost all of those algorithms depend on a known value of $\boldsymbol{\theta}$, which is unknown in the present analysis. In Statistics, two main methods have been developed that use the numerical solution to estimate the parameters in the model based on some initial parameter values. The Bayesian methods (Gelman, Bois and Jing, 1996; Wakefield, 1996; Lunn, et al., 2002; Putter, et al., 2002), and the maximum likelihood methods (Davidian and Giltinan, 1995; Racine-Poon and Wakefield, 1998) both have been developed to obtain parameter estimators

Our work is based on the " *Bayesian Euler's Approximation Method* (BEAM)"

(Goyal and Ghosh, 2008). *BEAM* is based on the existing Bayesian framework (Gelman, Bois and Jing, 1996; Wakefield, 1996; Lunn, et al., 2002. Putter, et al., 2002) and use the naive *Euler*'s approximation method to approximate the likelihood function. The *BEAM* does not require an analytic closed form of the ODEs system.

4.3.1 Likelihood approximation by the Euler's method

Euler's method is a popular numerical approximation method to solve a system of ODEs. The basic approach is to partition the time interval into N small pieces, each with the same length h. Let $t_k^0 = t_0^0 + hk$, for $k = 0, 1, \ldots, N$. So the numerical solution to (4.2) can be expressed as

$$\boldsymbol{\nu}(t, \boldsymbol{\theta}) = \int_{t_0}^{t} \boldsymbol{g}(\boldsymbol{\nu}(s, \boldsymbol{\theta}))ds + \boldsymbol{\nu}_0(\boldsymbol{\theta})$$

It then follows that as $h \to 0$, we have

$$\boldsymbol{\nu}(t + h, \boldsymbol{\theta}) - \boldsymbol{\nu}(t, \boldsymbol{\theta}) = \int_{t}^{t+h} \boldsymbol{g}(\boldsymbol{\nu}(s, \boldsymbol{\theta}))ds \approx h\boldsymbol{g}(\boldsymbol{\nu}(t, \boldsymbol{\theta})) \tag{4.4}$$

The *BEAM* is based on (4.4) to approximate the likelihood function of a regression model with ODEs. We partition the observed time interval $[t_0, t_n]$ into N parts $(t_0 = t_1^0 < t_2^0 < \ldots < t_N^0)$ with lengths $h = t_{k+1}^0 - t_k^0$, $k = 1, \ldots, (N-1)$. In order to cover the whole time interval, we choose $t_N^0 \geq t_n$. Let $\tilde{\boldsymbol{\nu}}_k \equiv \tilde{\boldsymbol{\nu}}(t_k^0, \boldsymbol{\theta})$ and $\tilde{\mu}_k \equiv \tilde{\mu}(t_k^0, \boldsymbol{\theta}) = H(\tilde{\boldsymbol{\nu}}_k)$ for $k = 1, 2, \ldots, N-1$. Then

$$\tilde{\boldsymbol{\nu}}_{k+1} = \tilde{\boldsymbol{\nu}}_k + h\boldsymbol{g}(\boldsymbol{\nu}_k)$$

with initial condition $\tilde{\boldsymbol{\nu}}_1 = \boldsymbol{\nu}_0(\boldsymbol{\theta})$. We use linear interpolation to define $\tilde{\mu}(t, \theta)$, for any $t \in [t_1^0, t_N^0]$. Given a time point $t \in [t_1^0, t_N^0]$, we define the label function $L : [t_1^0, t_N^0] \to \{1, \ldots, N\}$ as:

$$L(t) = \sum_{k=1}^{N} I(t_k^0 \leq t)$$

For any $t \in [t_1^0, t_N^0]$, we can always find $L(t)$, so that $t_{L(t)}^0 < t \leq t_{L(t)+1}^0$. Then, $\tilde{\mu}(t, \boldsymbol{\theta})$ defined as

$$\tilde{\mu}(t, \boldsymbol{\theta}) = \tilde{\mu}_{L(t)} + \frac{t - t_{L(t)}^0}{t_{L(t)+1}^0 - t_{L(t)}^0}(\tilde{\mu}_{L(t)+1} - \tilde{\mu}_{L(t)}) \tag{4.5}$$

We can now approximate the true likelihood function of $(\boldsymbol{\theta}, \boldsymbol{\eta})$ using the $\tilde{\mu}(t, \boldsymbol{\theta})$.

$$\tilde{L}_h(\boldsymbol{\theta}; \boldsymbol{Y}) = \prod_{j=1}^{n} \{ \frac{1}{\sqrt{2\pi}\sigma} \exp(\frac{1}{2\sigma^2}(Y_j - \tilde{\mu}(t_j, \boldsymbol{\theta}))^2) \}$$

It can be shown that $\tilde{\mu}(t, \boldsymbol{\theta}) = \mu(t, \boldsymbol{\theta}) + o(h)$, as $h \to 0$.

Given the observed dataset $\boldsymbol{D} = \{(y_j, t_j), j = 1, 2, \ldots, n\}$, we use a hierarchical model to describe it.

$$y_j | (\boldsymbol{\theta}, \boldsymbol{\eta}) \overset{\text{indep}}{\sim} N(\tilde{\mu}(t_j, \boldsymbol{\theta}), \sigma^2(t_j, \boldsymbol{\eta})) \tag{4.6}$$

where $\tilde{\mu}(t_j, \boldsymbol{\theta})$ is as defined in (4.5) . The priors for parameters are as follows:

$$\boldsymbol{\theta} | \boldsymbol{\eta} \sim MVN_p(\boldsymbol{\theta}_0, \Sigma_0^{\boldsymbol{\theta}}) \text{ and } \boldsymbol{\eta} \sim MVN_k(\boldsymbol{\eta}_0, \Sigma_0^{\boldsymbol{\eta}})$$

The value of $\boldsymbol{\theta}_0$, $\boldsymbol{\eta}_0$, $\Sigma_0^{\boldsymbol{\theta}}$ and $\Sigma_0^{\boldsymbol{\eta}}$ are assumed to be known. For the detailed discussion of the choice of prior distribution, see Natarajan and Kass (2000). The joint posterior distribution for parameters $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ based on above model can be written as:

$$p(\boldsymbol{\theta}, \boldsymbol{\eta} | \boldsymbol{Y}) \propto p(\boldsymbol{Y} | \boldsymbol{\theta}, \boldsymbol{\eta}) p(\boldsymbol{\theta} | \boldsymbol{\eta}) p(\boldsymbol{\eta})$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. The full conditionals of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are attained by

$$p(\boldsymbol{\theta} | \boldsymbol{\eta}, \boldsymbol{Y}) \propto p(\boldsymbol{Y} | \boldsymbol{\theta}, \boldsymbol{\eta}) p(\boldsymbol{\theta} | \boldsymbol{\eta}) \tag{4.7}$$

$$p(\boldsymbol{\eta} | \boldsymbol{\theta}, \boldsymbol{Y}) \propto p(\boldsymbol{Y} | \boldsymbol{\theta}, \boldsymbol{\eta}) p(\boldsymbol{\eta}) \tag{4.8}$$

the Metropolis-Hasting algorithm (Hastings, 1970) can be applied to sample $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ using (4.7) and (4.8).

4.4 Extension to handle missing and censored data

The methodology described in the previous section is based on a completely observed data vector $\boldsymbol{Y}$. In a clinical trial, some observations could be missing due to no response from patients, or censored due to the limitation of equipment that makes the measurements or due to the fact that a clinical trial is stopped. In this section, we will extend the *BEAM* method to obtain parameter estimates in the presence of missing and censored data.

4.4.1 Methods to account for missing data

Consider the complete dataset $\{(Y_j, t_j), j = 1, \ldots, n\}$. Suppose the Y_j's with $j_1 < \ldots < j_k$ are missing, where $j_1, \ldots, j_k \in \{1, 2, \ldots, n\}$. Let $\boldsymbol{Y}^{mis} = (Y_{j_1}, \ldots, Y_{j_k})$. So we can rewrite $\boldsymbol{Y} = (\boldsymbol{Y}^{obs}, \boldsymbol{Y}^{mis})$, where $Y^{obs} = \{Y_j : j \notin \{j_1, \ldots, j_k\}\}$. Let R_j denote the indicator of missing. In other words,

$$R_j = \begin{cases} 1 & \text{if } Y_j \text{ is observed} \\ 0 & \text{if } Y_j \text{ is missing} \end{cases}$$

According to Rubin (1976), we can classify the missing data into three cases based on the $Pr(\boldsymbol{R}|\boldsymbol{Y}, \boldsymbol{\theta})$: (i) Missing completely at random (MCAR), if $Pr(\boldsymbol{R}|\boldsymbol{Y}, \boldsymbol{\theta}) = Pr(\boldsymbol{R}|\boldsymbol{\theta})$, (ii) Missing at random (MAR), if $Pr(\boldsymbol{R}|\boldsymbol{Y}, \boldsymbol{\theta}) = Pr(\boldsymbol{R}|\boldsymbol{\theta}, \boldsymbol{Y}^{obs})$, and (iii) Nonignorable missing, if $Pr(\boldsymbol{R}|\boldsymbol{Y}, \boldsymbol{\theta}) = Pr(\boldsymbol{R}|\boldsymbol{\theta}, \boldsymbol{Y}^{mis})$. In case of MAR, from (4.6) it follows that the density of the missing is given by

$$f(\boldsymbol{Y}^{mis}|\boldsymbol{\theta}, \boldsymbol{\eta}) \propto \prod \frac{1}{\sigma(t_j, \boldsymbol{\eta})} \exp\{-\frac{(y_j^{mis} - \tilde{\mu}(t_j, \boldsymbol{\theta}))^2}{2\sigma^2(t_j, \boldsymbol{\eta})}\} \tag{4.9}$$

We impute the missing observations using (4.9).

MCMC algorithm for missing data

We use Metropolis-Hastings algorithm to generate approximate random samples from the posterior distribution of ($\boldsymbol{\theta}$, $\boldsymbol{\eta}$) using the following steps:

1. Initialize the iteration of the chain at l=0 and start with initial values $\boldsymbol{S}^{(0)} = (\boldsymbol{\theta}^{(0)}, \boldsymbol{\eta}^{(0)})$.

2. Sample the $\boldsymbol{Y}^{mis(l)}$ from (4.9) using $(\boldsymbol{\theta}^{(l-1)}, \boldsymbol{\eta}^{(l-1)})$ and compute complete data $\boldsymbol{Y}^{(l)} = (\boldsymbol{Y}^{obs}, \boldsymbol{Y}^{mis(l)})$.

3. For $\boldsymbol{S}^{(l)}$, generate a new value $\boldsymbol{\phi}$ ($\boldsymbol{\phi} = (\boldsymbol{\theta}, \boldsymbol{\eta})^T$)from a symmetric proposal density $q(\boldsymbol{\phi}|\boldsymbol{S}^{(l-1)})$. Evaluate the acceptance probability of the move given by

$$\alpha(\boldsymbol{\phi}|\boldsymbol{S}^{(l-1)}) = \min\{1, \frac{\pi(\boldsymbol{\phi}|\boldsymbol{S}^{(l-1)}, \boldsymbol{Y}^{(l)})}{\pi(\boldsymbol{S}^{(l-1)}|\boldsymbol{Y}^{(l)})}\}$$

   Generate a sample $u^{(l)} \sim U(0, 1)$. If $u^{(l)} \leq \alpha(\boldsymbol{\phi}|\boldsymbol{S}^{(l-1)})$, then let $\boldsymbol{S}^{(l)} = \boldsymbol{\phi}$, otherwise $\boldsymbol{S}^{(l)} = \boldsymbol{S}^{(l-1)}$.

4. Repeat steps (2) and (3), until the samples arise from the stationary density $p(\boldsymbol{\theta}, \boldsymbol{\eta}|\boldsymbol{Y}^{obs})$.

4.4.2 Methods to account for censored data

Censored data is an common problem with clinical data. Consider observed data $D = \{(Y_j, t_j) : j = 1, \ldots, n\}$. Suppose we may observe Y_j only when Y_j is larger or smaller than a constant value, the minimum or maximum value that the instrument can measure.

Let X_j be the observed value corresponding to the true value Y_j, which may have been censored. We consider the following two cases:

$$X_j = \begin{cases} \min\{Y_j, c_j\} & \text{Right censoring;} \\ \max\{Y_j, c_j\} & \text{Left censoring.} \end{cases}$$

where $c_j \in R$ are the threshold values.

Again, consider the model described in (4.6) for the true data Y_j's. We obtain the density of the censored observations, which will be used to impute the censored observations.

In case of left censoring, the density of X_j is given by

$$f(X_j|\boldsymbol{\theta}, \boldsymbol{\eta}) = \frac{\frac{1}{\sigma(t_j, \boldsymbol{\eta})}\phi(\frac{X_j - \mu(t_j, \boldsymbol{\theta})}{\sigma(t_j, \boldsymbol{\eta})})}{1 - \Phi(\frac{c_j - \mu(t_j, \boldsymbol{\theta})}{\sigma(t_j, \boldsymbol{\eta})})} I(Y_j > c_j) \tag{4.10}$$

In case of right censoring, the density of X_j is given by

$$f(X_j|\boldsymbol{\theta}, \boldsymbol{\eta}) = \frac{\frac{1}{\sigma(t, \boldsymbol{\eta})}\phi(\frac{X_j - \mu(t_j, \boldsymbol{\theta})}{\sigma(t_j, \boldsymbol{\eta})})}{1 - \Phi(\frac{c_j - \mu(t_j, \boldsymbol{\theta})}{\sigma(t_j, \boldsymbol{\eta})})} I(Y_j < c_j) \tag{4.11}$$

When there is no censoring,

$$f(X_j|\boldsymbol{\theta}, \boldsymbol{\eta}) = \frac{1}{\sigma(t_j, \boldsymbol{\eta})}\phi(\frac{X_j - \mu(t_j, \boldsymbol{\theta})}{\sigma(t_j, \boldsymbol{\eta})}) \tag{4.12}$$

Thus, it follows that the likelihood of $(\boldsymbol{\theta}, \boldsymbol{\eta})$ given $\{\boldsymbol{X}_j,\ j = 1, \cdots, n\ \}$ is given by:

$$L(\boldsymbol{\theta}, \boldsymbol{\eta}|\boldsymbol{X}) = \prod_{j=1}^{n} f(X_j|\boldsymbol{\theta}, \boldsymbol{\eta})$$

where $f(X_j|\boldsymbol{\theta},\boldsymbol{\eta})$ is computed using (4.10), (4.11) or (4.12) according as the value X_j is left censored, right censored or not censored.

MCMC algorithm for censored data

The Metropolis-Hastings algorithm to obtain samples from the posterior distribution of $(\boldsymbol{\theta},\boldsymbol{\eta})$ can be obtained as follows:

1. Initialize the iteration of the chain at l=0 and start with initial values $\boldsymbol{S}^{(0)} = (\boldsymbol{\theta}^{(0)},\boldsymbol{\eta}^{(0)})$.

2. Sample the $\boldsymbol{Y}^{cen(l)}$ from (4.10) or (4.11) using $(\boldsymbol{\theta}^{(l)},\boldsymbol{\eta}^{(l)})$ and compute $\boldsymbol{Y}^{(l)}$, which is the set of not censored data and imputed data.

3. For $\boldsymbol{S}^{(l)}$, generate a new value $\boldsymbol{\phi}$ $(\boldsymbol{\phi} = (\boldsymbol{\theta},\boldsymbol{\eta})^T)$ from a symmetric proposal density $q(\boldsymbol{\phi}|\boldsymbol{S}^{(l-1)})$. Evaluate the acceptance probability of the move given by,

$$\alpha(\boldsymbol{\phi}|\boldsymbol{S}^{(l-1)}) = \min\{1, \frac{\pi(\boldsymbol{\phi}|\boldsymbol{S}^{(l-1)},Y^{(l)})}{\pi(\boldsymbol{S}^{(l-1)}|Y^{(l)})}\}$$

   Generate a sample $u^{(l)} \sim U(0,1)$. If $u^{(l)} \leq \alpha(\boldsymbol{\phi}|\boldsymbol{S}^{(l-1)})$, then let $\boldsymbol{S}^{(l)} = \boldsymbol{\phi}$, otherwise $\boldsymbol{S}^{(l)} = \boldsymbol{S}^{(l-1)}$.

4. Repeat steps (2) and (3), until the samples arise from the stationary distribution of density of $\boldsymbol{\theta},\boldsymbol{\eta}$ given true data $\boldsymbol{Y}$.

4.5 Simulation study

A simulation study has be done to evaluate the bias and mean squared error (MSE) for the parameter estimates when observations are subject to missing and censored. Diggle (1990) presents a data set that describes the growth of three closed colonies of paramecium aurelium in a nutritive medium of a period for 19 days . The data is assumed to follow the growth model:

$$\begin{aligned} y_j &= \mu(t_j,\boldsymbol{\theta}) + \epsilon_j \\ \frac{d\mu}{dt} &= g(\mu(t,\boldsymbol{\theta})) \\ &= \theta_1 - \theta_2 e^{\mu} \end{aligned} \tag{4.13}$$

Here $\mu(0) = \log(2)$, $\epsilon_j \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$. By solving above ODEs system, we can get the following closed form solution:

$$\mu(t, \boldsymbol{\theta}) = \log(\theta_1) + \mu(0) + t\theta_1 - \log\{\theta_2 e^{\mu(0)}(e^{t\theta_1} - 1) + \theta_1\} \tag{4.14}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$. For the simulation study, we generate data by using (4.14). The true values of the parameters are set at $\theta_1 = 0.8, \theta_2 = 0.0015$ and $\sigma = 0.25$. The sample size is $n = 19$. We generated samples based on a Markov chain with an initial burn-in of 4000 iterations followed by 2000 post-burn-in samples. To fit the model by *BEAM*, we chose $N = 10$ and the prior hyper parameters were set at: $\boldsymbol{\theta}_0 = (0, 0)^T$, $H_0 = 0.1I_2$, $a_0 = 0.01$ and $b_0 = 0.01$.

4.5.1 Effect of missing data

We randomly choose 5 and 8 data points to be missed, which amounts to about 26% and 42% missing. We report (i)the bias, which is the difference between MC median of the point estimates and the true value of a parameter; (ii)the estimated standard error (ESE), which is the MC mean of the poster standard deviations of the parameters. The estimated parameters are shown in Table 4.1:

[Table 4.1 about here]

Table 4.1 suggests that although the missing rate is high, the *BEAM* method still attained a good estimate of θ_1, θ_2 and σ. The variances of estimates increase as the missing ratio varies from small to large. In order to evaluate the performance of the method in handling censored data, we compute the 90% and 95% posterior interval (PI) of missing data points. The more the PI covers the corresponding missing data point, the better we think the method is. From Table 4.1, the 90% and 95% posterior interval covers more than 90% of missing data, which are very high. The *BEAM* appears to be robust in analyzing missing data case.

[Figure 4.1 about here]

4.5.2 Effect of censoring

In the censoring study, we consider two censoring cases, left censoring and right censoring. For left censoring, observations less than 2 (or 3) will be reported 2 (or 3). For right censoring, observation greater than 6 will be reported 6. Again, let $\theta_1 = 0.8$, $\theta_2 = 0.0015$, $\sigma = 0.25$, data is generated from (4.14). We report (i)the bias, (ii) ESE. The estimated parameters are shown in Table 1. Although the estimates of θ_1 in the censoring case are not in the range of Q1 (25% quartile) and Q3 (75% quartile), 95% posterior intervals still cover the true value. From the boxplot, the variances are increasing as the censoring rate climbs. The method is robust even at high censoring rate case. The 90% and 95% posterior interval cover more than 85% and 90% censoring data.

4.6 Analysis of Virological Data from AIDS Clinical Trials

With the development of assay techniques, viral load (HIV-1 RNA copies) has been used as a surrogate marker to accelerate AIDS clinical trials. HIV dynamic models provide a framework to describe the virus elimination and production process during antiviral treatment for each individual patient. Biomathematicians and theoretical biologists have worked on mathematical dynamic models of HIV since the end of the 1980s (Merrill, 1987; Anderson and May, 1989; Tan and Wu 1998). In evaluating the efficacy of anti-HIV treatments and understanding HIV infection pathogenesis, it is very important to estimate viral dynamic parameters for the whole population and for each individual patient in an AIDS clinical trial.

In AIDS Clinical Trial Group (ACTG) Protocol 315, 53 HIV infected patients were treated with potent antiviral drugs (ritonavir, 3TC, and AZT). Plasma HIV-1 RNA was repeatedly measured from day 0 to at most day 196. Of the 53 patients, 7 either dropped out of the study or showed profiles that did not correspond to the pattern of the rest of the patients, exhibiting no response to the antiviral drugs due

to poor absorption of the drugs, non compliance, drug resistance or other unknown reason. The remaining patients had between four and ten viral load measurements. The HIV-1 RNA assay has the limitation of detection of 100 copies. Measurements below the detectable level have been imputed by the values to be 50 (=1.6990 in log_{10} scale) by early researches. However, we will treat such observation as left censored. We observed the viral load of 16 patients to be below the limit of detection at the last measurement and in 30 patients to be above the limit. There might be two underlying populations for this study. In the analysis, we will fit the model separately for two groups. Since we do not have further information for the study, presently we do not have any systematic way to identify the underlying predictor that may give rise to two groups that we observed by our empirical analysis. We will test whether the differences of parameters are significant for the newly created groups.

The observations are assumed to follow a log-normal distribution with

$$\begin{aligned} y_{ij} &= \log V(t_{ij}, \theta_i) + \epsilon_{ij} \\ &= \mu(t_{ij}, \theta_i) + \epsilon_{ij} \end{aligned} \tag{4.15}$$

The Bayesian nonlinear mixed-effects model can be written in the following three stages (Davidian & Giltinan, 1995):

Stage 1. Within-subject variation model:

$$\boldsymbol{Y}_i = \mu(\boldsymbol{\theta}_i) + \boldsymbol{\epsilon}_i \ , \ \boldsymbol{\epsilon}_i | \sigma^2, \boldsymbol{\theta}_i \sim N(0, \sigma^2 I_{n_i})$$

Where $\boldsymbol{Y}_i = (y_{i1}(t_1), \cdots, y_{in_i}(t_{n_i}))$.

Stage 2. Between subject variation:

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_0 + \boldsymbol{b}_i \ , \ \boldsymbol{b}_i | \Sigma_{\boldsymbol{b}} \sim N(0, \Sigma_{\boldsymbol{b}})$$

Stage 3. Hyperprior distribution:

$$\sigma^2 \sim IG(a_0, b_0) \ , \ \boldsymbol{\theta}_0 \sim N(\eta, \Sigma_{\boldsymbol{\theta}_0}) \ , \ \Sigma_b^{-1} \sim Wi(\Omega, \nu)$$

Perelson (1996) proposed a mathematical model to analyze a detailed set of human immunodeficiency viral load data.$V(t)$ in (4.15) is the solution of the following HIV

viral load dynamic system:

$$\begin{aligned} \frac{dT^*}{dt} &= \alpha V_I - \delta T^* \\ \frac{dV_I}{dt} &= -cV_I \\ \frac{dV_{NI}}{dt} &= N\delta T^* - cV_{NI} \end{aligned} \tag{4.16}$$

Here T^* is the density of the productively infected cells. $V = V_I + V_{NI}$ is the concentration of viral particles in plasma, δ is the rate of loss of virus-producing cells, N is the number of new virions produced per infected cell during its lifetime, and c is the rate constant for virion clearance. The initial settings are $V_I(t=0) = V_0$ and $V_{NI}(t=0) = 0$. Let $\mu_1(t,\theta) = \log(T^*)$, $\mu_2(t,\theta) = \log(V_I)$ and $\mu_3(t,\theta) = \log(V_{NI})$. Equivalently, we can express the above equations in term of $\mu(\cdot) = (\mu_1, \mu_2, \mu_3)^T$:

$$\begin{aligned} \frac{d\mu_1}{dt} &= -e^{\theta_2} + \alpha e^{\mu_2 - \mu_1} \\ \frac{d\mu_2}{dt} &= -e^{\theta_2} \\ \frac{d\mu_3}{dt} &= Ne^{\theta_2 + \mu_1 - \mu_3} - e^{\theta_1} \end{aligned}$$

where

$$\begin{aligned} V = V_I + V_{NI} &= e^{\mu_2} + e^{\mu_3} \\ \log(y_{ij}) &= \log(e^{\mu_2} + e^{\mu_3}) + \epsilon_{ij} \end{aligned} \tag{4.17}$$

We assumed that the random effects vector $\boldsymbol{\theta_i} = (\log(c_i), \log(\delta_i))^T$ also follows a log-normal distribution. The Bayesian framework for the parameter estimation can be expressed as:

$$\begin{aligned} \boldsymbol{\theta}_i &= (\log(c_i), \log(\delta_i))^T \sim MVN_2(\boldsymbol{\theta}, \Sigma_\theta) \\ \boldsymbol{\theta} &= (\log(c), \log(\delta))^T \sim MVN_2(\boldsymbol{\theta}_0, H_0) \\ \Sigma_\theta^{-1} &\sim W_2(R_0, \rho_0) \text{ and } \sigma^2 \sim IG(a_0, b_0) \end{aligned}$$

where $W_2(\cdot,\cdot)$ denotes the Wishart distribution in dimension two and $IG(\cdot,\cdot)$ denotes the Inverse Gamma distribution. We use the following informative priors and the

values for hyper-parameters:

$$a_0 = 4.5, b_0 = 9, \rho_0 = 3, \boldsymbol{\theta}_0 = (1.1, -1.0)^T$$
$$H_0 = diag(0.1, 0.01), R_0 = diag(2.5, 2.5)$$

These values were chosen based on the priors used in Han et al.(2002) and Holte et al.(2003).

[Table 4.2 about here]

The Bayesian Euler's approximation method was used with 10000 MCMC sampler followed by a burn-in of 5000 iterations using a single chain.From the above results, we can see the differences of parameters of "Group A" and "Group B" are significant. A possible reason may be due to different effects of the drugs . In group A, the efficacy is close to 1, which means the drug has good and long lasting effect on the virus. While in group B, the efficacy is decreasing to 0, which means the effect decays to zero.

[Figure 4.2 about here]

Figure 4.3 displays the difference between a naive imputation method (1.6990 in log10-scale) versus the predicted value based on *BEAM*. This figure depicts the actual value would have been observed if the device that would not have had a detection limit.

[Figure 4.3 about here]

From the simulation study and real data application, we illustrate that our method provides good estimates of parameters for missing and censored data without analytically solving the ODEs system. The analysis of censored data by means of regression models is straightforward. The missing and censored data in the case of non-constant variance case will be the topic of future study.

4.7 Discussion

In this paper, we have extended *BEAM* missing or censored data cases for nonlinear models involving ODEs. We impute missing and censored data from the likelihood

function and treat them as observed data in MCMC iterations. The method maximally explores the information in the observed data. Our simulation results suggest that *BEAM* reduces the possible biases and has similar standard errors as those from complete data.

The real HIV data example suggests that imputation of censored data by 50 may not be a safe way. Some reasons are that measurement much lower than 100, which should be explored further may be neglected by imputing a value of 50.

Finally, the method may be extended into more complicated cases, for example, observed data may be highly correlated, or nonignorable missing data.

Table 4.1: Biases and standard error based on the logistic growth model for colonies of the bacteria paramecium aurelium using BEAM with 16% and 42% missing data.

Type	%	θ_1	θ_2	σ	90% PI	95% PI
Complete	0	-0.0124 (0.028)	7.7e-05 (0.00014)	0.027 (0.049)		
Missing	26%	-0.0213 (0.035)	5.0e-05 (1.7E-04)	0.270 (0.055)	89.2%	93.4%
	42%	-0.027 (0.037)	1e-04 (1.9e-04)	0.27 (0.064)	88.9%	93.2%
Left censored	2 (5.8%)	-0.026 (0.03)	3e-05 (1.46e-04)	0.020 (0.05)	89.0%	96.1%
	3 (12.4%)	-0.026 (0.027)	3e-05 (1.3e-04)	0.020 (0.05)	89.4%	96.2%
Right censored	6 (50.5%)	-0.03 (0.033)	-1E-04 (3e-04)	0.046 (0.08)	87.3%	93%

Table 4.2: Parameter estimates for the NLME model with censored data in ACTG 315. Estimates of group A and B are reported separately. 95% posterior intervals are also reported.

	Group A (16 patients)			Group B (30 patients)		
Parameters	Mean	SE	95% PI	Mean	SE	95% PI
θ_1	0.244	0.163	[-0.071,0.566]	-0.991	0.122	[-1.263,-0.759]
θ_2	1.895	1.645	[-1.309,5.107]	-7.371	1.017	[-8.992,-5.752]
σ^2	1.770	0.004	[1.536,2.047]	3.103	0.156	[2.821,3.425]

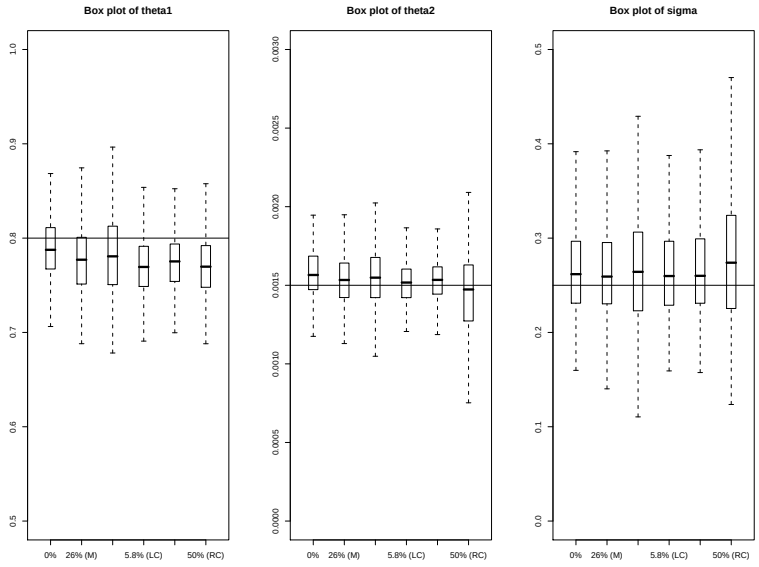

Figure 4.1: Box plot of point estimates based on 500 simulated data sets for complete, missing and censored data cases (The horizontal solid line in each case represents the true value of the parameters). "0%" means complete data, "$x\%(M)$" means "$x\%$" of data missing, "$x\%(LC)$" means average "$x\%$" of data left censored (less than 2 or 3), "$x\%$ (RC)" means average "$x\%$" of data right censored (greater than 6).

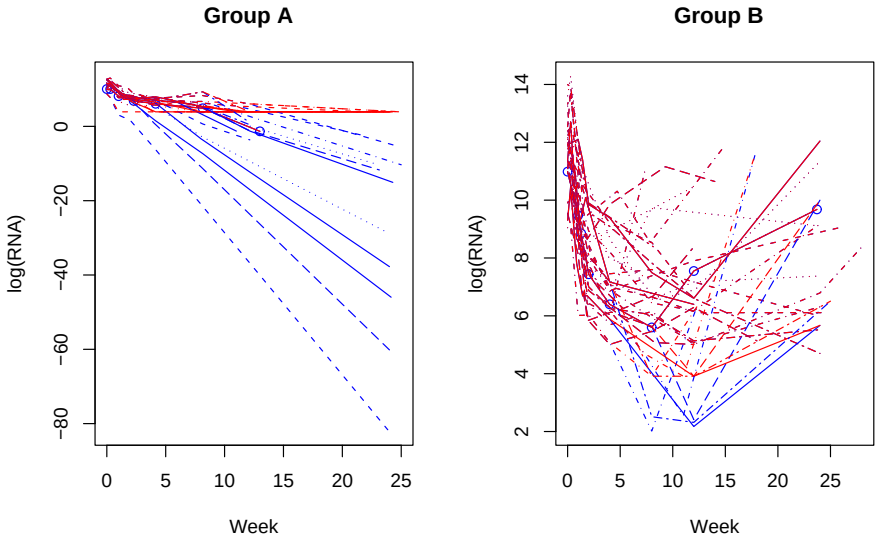

Figure 4.2: The plots of log RNA levels for group A and B containing censored observations and the augmented data. The red lines are the observed data and the blue ones are augmented data

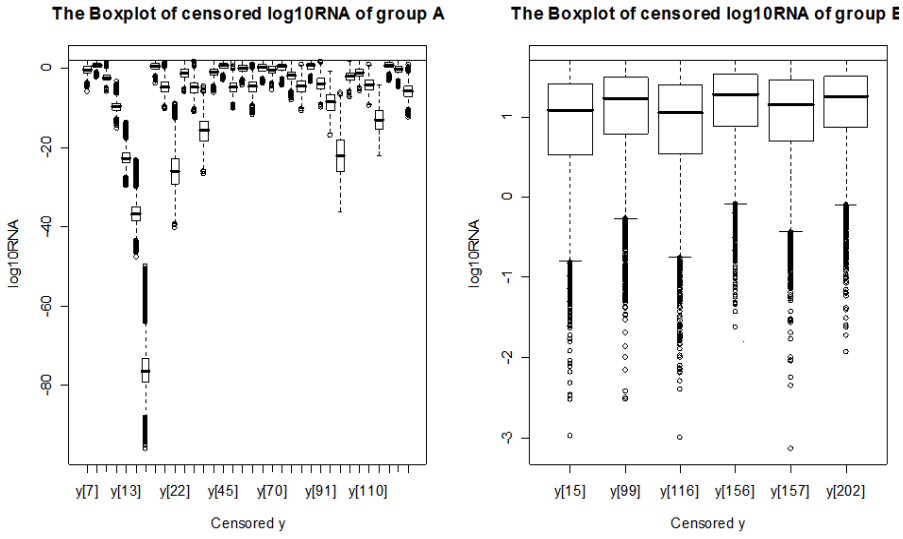

Figure 4.3: The boxplots of posterior distributions of censored log RNA levels for group A and B. The horizontal line is the threshold value for censored data.

Chapter 5

Conclusion and future work

The objective of Bayesian approach for genome-wide association study was to propose a general framework to integrate available information from multiple sources into a study to improve the power to identify causal genes associated with some disease. We proposed the modified Bayesian method for partitioned data, which is based on the external information. In Section 2.4, we briefly review the general Bayesian method for the multiple testing problem. Based on the paper by Scott and Berger (2003), the characteristic of signals can be washed away when noise increases. By controlling the ratio of noises vs. signals, the prior effect may still affect the posterior probability that the null hypothesis is true. In Section 2.4.2, we explain the Bayesian method for partitioned data. The data is partitioned into group A and B based on the information from other sources. The purpose of partitioning the data is to maintain a reasonable range of ratio between noises and signals, so that the effect of the prior may not be washed away. Our proposed method provides a convenient way to incorporate various kinds of information into our prior knowledge. This information can be in qualitative or quantitative form, like pathway information, or linkage study information. After obtaining the posterior probability that the null hypothesis is true, we controlled Bayesian FDR at certain level to achieve the realized FDR under control. In Section 3.5, we compared the results of our proposed method with the q-value algorithm (Storey , 2002) based on simulation studies. The results show that when external information is correct and the ratio of noises and signals is within a certain range, our proposed method can improve the power substantially with FDR under control.

To illustrate our proposed method, in Section 2.6, we applied the method to schizophrenia GWAS data. No statistically significant SNPs have been reported yet. 799 SNPs located in 26 candidate genes, which are mentioned in the literature are chosen as the "promising" group A and rest as group B. Although no significant SNPs were found, we list the top five SNP markers for follow-up study from each group. No statistically significant markers are found. The result was expected because the sample size was small even in a single marker test. However, there are several aspects we may work on in the future. One extension is to perform gene-based GWAS in the

Bayesian framework. As in GWAS for schizophrenia, about 500k tests are carried out, for which the multiple testing issue becomes a major concern. The gene-based GWAS will dramatically reduce the number of tests. The main obstacle is how to combine test statistics (e.g. p-values) in the same gene. Several tests have been developed to combine information from multiple markers. For example, by simultaneously testing the main effect and interaction effect among markers, which are in the same gene in a logistic regression model. Alternatively, one could use haplotype-based methods, which are attractive since genomic variations in humans are structured into haplotypes (Clark, 2004; Schaid, 2004). Recently, several studies have been proposed, such as using principal components of marker genotypes as covariates in multiple regression analysis (Gauderman et al., 2007; Wang and Abbott, 2007), and a weighted score test using a Fourier transformation (Wang and Elston, 2007). The other extension is to model quantitative external information in a more delicate way. Roeder *et al.* (2006) used linkage information to compute the weights for different regions, which are up-weighted and down-weighted. Our goal is to provide a general way to utilize all kinds of external information, not only quantitative information.

Meta-analysis is another way to utilize information from different resources by combining results from multiple studies on some common interest. In Chapter 3, we provide a novel method to summarize those results. Without a normality assumption for the underlying effect, we show that a broad class of distribution can be approximated by a mixture distribution in the random-effects model. The population mean and variance estimates are then obtained by using EM algorithm in Section 3.4. We compare the performance of our proposed method with the commonly used DerSimonian and Laird method based on various simulation studies. Three different distributions (normal, Laplace and logistic distributions) are considered for the underlying effect. The simulation studies show that the estimates using our proposed method greatly improved the accuracy in estimating overall population mean and heterogeneity variance in all three cases in Section 3.5. We applied our proposed method to DRD2 multiple association studies with schizophrenia. In the future, we want to do several extensions to our method. One is to provide the theoretical op-

timal N. From our simulation study, we can see that choosing a correct N is very important. In our paper, we used AIC as a criterion to choose N. Second, based on the estimate of heterogeneity, we want to find the threshold to help us to choose between fixed-effects model and random-effects model. Comparing with a chi-squared test and the index of heterogeneity measurement, estimating heterogeneity variance makes more sense.

ODEs dynamic system has been an important tool to analyze complicated biological systems. Estimating the meaningful parameters in the ODEs system, which involves missing or censored data, is challenging. The objective of this project was to extend "Bayesian Euler's approximation method (BEAM)" to handle the cases of missing and censored data. Our proposed method is based on a data augmentation algorithm. In the case of missing data, we always consider the data to be missing completely at random. In the censored data case, we consider left censored and right censored data. We compared the estimates for complete data, missing data and censored data (Section 4.5). Our results showed that our proposed method provide a way to recover information from missing and censored data.

For the purpose of illustration, we applied our method to the HIV model described in Section 4.6, for which no closed form expression for the mean function is available. The observed data is the number of observed of HIV-1 RNA copies. The dataset contains censored data, which are due to the limitation of RNA array. By imputing the censored data, our method successfully recovers the censored data part, which can not be handled well by traditional methods. We also found a potential partition for the RNA copies data based the status of the last measurement, which has never been mentioned in any literatures related to this HIV study. The partition may arise from two different populations or different effect of the drugs.

At this stage, future research may include in the following aspects. In modeling missing data, we only consider a simple case. In many experimental studies, scientists are more interested in missing at random and non-ignorable missing data cases, which involves a more complicated model. Recovering information from these two type of missing data is meaningful. Also in genetic regulatory network, ODEs systems have

been widely used to model interactions among many factors. Missing and censored data cases are very common. For the censored data, we can extend our proposed method to interval censoring and random censoring cases.

Our proposed method provides estimates for the parameters involved in the system of ODEs, without imposing any restrictive assumptions regarding the dynamics of the data set. Also, we imputed the missing and censored data and incorporated it into a Metropolis-Hastings algorithm to make use of the whole data set.

With more and more genetic studies performed in the recent years, we are continuing to answer the question of how to best use of all kinds of information. Bayesian inference, meta-analysis, ODEs dynamic system and other techniques give us the power to reveal the truth behind phenomena.

Bibliography

Abecasis G, Cardon L, Cookson W (2000a). A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* **66**:279-292.

Abecasis G, Cookson W, Cardon L (2000b). Pedigree tests of transmission disequilibrium. *European journal of human genetics* **8**:545-551.

Allison, D. B. et al. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational statistics & data analysis* **39**: 1-20.

Benjamini, Y. and Hochberg, Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc.* B **57**: 289-300.

Böhning, D. (1999). Computer-assisted Analysis of Mixtures and Applications: Meta-analysis, *Disease Mapping and Others*. London: Chapman and Hall.
DerSimonian, R., Laird, N. Meta-analysis in clinical trials. (1986) *Controlled Clinical Trials*. **7**(3):177-188.

Broët, P., Lewin, A., Richardson,S., Dalmasso, C., and Magdelenat, H. (2004). A mixture model based strategy for selecting sets of genes in multiclass response microarray experiments. *Bioinformatics*. **20**(16), p 2562.

Chiano, M. N. and Clayton, D. G.(1998). Fine genetic mapping using haplotype analysis and the missing data problem *Annals of Human Genetics*. **62**: 55-60.

Clark, A.G. (2004). The role of haplotypes in candidate gene studies. *Genet Epidemiol*. **27**:321-333.

Cochran, W. G. (1954). The combination of estimates from different experiment. *biometrics*. **10**: 101-129.

Cohn, L.D. (2003). How Meta-Analysis Increases Statistical Power. *Psychological Methods*. Vol. **8**, No. 3: 243-253.

DerSimonian, R.,and Kacker, R. (2007). Random-effects model for meta-analysis of clinical trials: an update. *Contemporary Clinical Trials.* **28**:105-114.

Davidian, M. and Giltinan, D. M.(1995). *Nonlinear Models for Repeated Measurement Data.* London: Chapman & Hall.

Duan, J., Wainwright, M.S., Comeron, J.M., Saitou, N., Sanders, A.R., Gelernter, J., Gejman, P.V. (2003). Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor, *Hum. Mol. Genet.* **12**: 205-216.

Efron, B., Storey, J., Tusher, VG and Tibshirani, R (2001). Empirical Bayes analysis of a microarray experiment. *J Am Stat Assoc* **96**: 1151-1160.

Efron, B. and Tibshirani, R (2001). Empirical Bayes methods and false discovery rates for microarrays. *Genet Epidemiol* **23**: 70-86.

Elowitz, M.B., Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. *Nature.* **403**(6767):335-338.

Endy, D. and *et al.* (2000). Computation, prediction, and experimental tests of fitness for bacteriophage T7 mutants with permuted genomes. *Proc. Natl. Acad. Sci.*. Vol. **97**: 5375-5380.

Forster, J. J., McDonald, J. W., and Smith, P. W. F. (1996). Monte Carlo Exact Conditional Tests for Log-Linear and Logistic Models. *Journal of the Royal Statistical Society*, **58**: 445-453.

Gauderman, W. J., Murcray, C., Gilliland, F., Conti, D. V. (2007). "Testing association between disease and multiple SNPs in a candidate gene". *Genetic Epidemiology.* **31**(5):383-395.

Gelman, A., Carlin, J. B., Stem, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*, London: Chapman and Hall.

Gelman, A., Bois, F. and Jing, J.(1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association* **85**: 398-409.

Geman, S., and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,**6**: 721-741.

Goyal, L. and Ghosh, S. K. (2006). Statistical Inference for Nonlinear Models Involving Ordinary Differential Equations. *Joint Statistical Meetings Proceedings.*

Hardy, RJ, Thompson, SG. (1996). A likelihood approach to meta-analysis with random effects. *Statistics in Medicine*. **15**(6):619-629.

Hardy, R. J. and Thompson, S.G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine*. *17*(8): 841-856.

Hartung, J. (1999). An alternative method for meta-analysis. *Biometrical Journal*. **8**: 901-916.

Higgins, J. P. and Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statist. Med.* **21**: 1539-1558.

Hirschhorn, J.N. *et al.* (2002). A comprehensive review of genetic association studies. *Genet. Med.* **4**: 45-61

Ho, D. D., Neumann, A. U., Perelson, A. S., Chen, W., Leonard, J. M. and Markowitz, M. (1995). Rapid turnover of plasma virions and cd4 lymphocytes in HIV-1 infection. *Nature* **373**: 123-126.

Hopke, P. K. and Rubin, D. B. (2001). Multiple imputation for multivariate data with missing and below-threshold measurements: time-series concentrations of pollutions in the Arctic. *Biometrics*, **57**: 22-33.

Ioannidis, J.P. et al. (2003). Genetic associations in large versus small studies: an empirical assessment. *Lancet* **361**: 567-571.

Iserles, Arieh (1996). A First Course in the Numerical Analysis of Differential Equations. *Cambridge University Press*.

Lee, W. L., Bausell, R. B., Berman, B. M. (2001). The growth of health-related meta-analyses published from 1980 to 2000. *Evaluation and the Health Professions*. **24**: 327-335.

Lin, L.H., *et al.* (2005). Dynamic modeling of cis-regulatory circuits and gene expression prediction via cross-gene identification. *BMC Bioinformatics*. **6**:258.

Little, R. J. A,, and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, New York: Wiley.

Lohmueller K. E., Pearce C. L., Pike M., Lander E. S., Hirschhorn J. N. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet.* **33**: 177-82.

Lunn, D. J., Best, N., Thomas, A., Wakefield, J. and Spiegelhalter, D. (2002). Bayesian analysis of population PK/PD models: General concepts and software. *Journal of Pharmacokinetics and Pharmacodynamics* **29**: 271-307.

Maraganore DM, de Andrade M, Lesnick TG, Strain KJ, Farrer MJ, Rocca W, Pant PVK, Frazer KA, Cox DR, Ballinger DG. (2005). High-resolution whole-genome association study of Parkinson disease. *Am J Hum Genet* 77:685C693.

Martin E, Bass M, Kaplan N (2001). Correcting for a potential bias in the pedigree disequilibrium test. *Am. J. Hum. Genet*. **68**: 1065-1067.

Martin E, Monks S, Warren L, Kaplan N (2000). A test for linkage and association in general pedigrees: The pedigree disequilibrium test. *Am. J. Hum. Genet*. **67**: 146-154.

Muller, P., Parmigiani, G. and Rice, K. (2006). FDR and Bayesian Multiple Comparisons Rules. *Proc. Valencia / ISBA 8th World meeting on Bayesian Statistics*

Natarajan, R. and Kass, R. (2000) Reference bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association* **95**: 227-237.

Newton, M. A. *et al.* (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*. **5**,2, 155-176.

Newton, Noueiry, Sarkar and Ahlquist (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5**:, p155.

Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Hori M, Nakamura Y, Tanaka T. (2002). Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat Genet* **32**: 650-654.

Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M. and Ho, D. D. (1996). HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time. **Science**. **271**: 1582-1586.

Putter, H., heisterkamp, S. H., Lange, J. M. A. and Wolf, F. D.(2002). A bayesian approach to parameter estimation in hiv dynamical models. *Statistics in Medicine*. **21**: 2199-2214.

Racine-Poon, A. and Wakefield, J.(1998). Statistical methods for population pharmacokinetic modeling. *Statistical Methods in Medical Research* **7**: 63-84.

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

Roberts,G.O. , Gelman, A. and Gilks, W.R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms, *Ann. Appl. Probab.* **7**: 110-120.

Robinson, R. M. (1980). Estimation and forecasting for time series containing censored or missing observations. *Time Series*, North-Holland Publishing Company, 167-182.

Rubin, D. B. (1976). Inference and Missing Data, *Biometrika*,**63**: 581-592.

Satagopan JM, Elston RC. (2003). Optimal two-stage genotyping in population-based association studies. *Genet Epidemiol* 25: 149-157.

Saxena R., Voight B. F., Lyssenko V., et al. (2007). Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**:1331C1336.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.

Schaid, D. J. (2004). Evaluating associations of haplotypes with traits. *Genetic Epidemiology*. **27**: 348-364.

Scott, G. and Berger, J.O. (2006). An exploration of Bayesian multiple testing. *Journal of Statistical Planning and Inference* .

Sham P, Curtis D (1995). An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Annals of human genetics* **59**:323-336.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73** 751-754.

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

Storey, JD (2002) A direct approach to false discovery rate. *J R Stat Soc B* **64**: 479-498.

Storey,J.D. and Tibshirani,R. (2001). Estimating false discovery rates under dependence, with applications to DNA microarrays. *Technical Report* 2001-28, Department of Statistics, Stanford University, CA.

Storey, JD. and Tibshirani, R (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 100: 9440-9445.

Sullivan PF, Lin D, Tzeng JY, van den Oord E, Perkins D, Stroup TS, Wagner M, Lee S, Wright FA, Zou F, Liu W, Downing AM, Lieberman J, Close SL. (2008). Genomewide association for schizophrenia in the CATIE study: results of stage 1. *Mol Psychiatry* . **13** (6): 570-584.

Sun, J. and Kalbfleisch, J. D. (1995). Estimation of the mean function of point processes based on panel count data. *Statist. Sinica*, **5**: 279-290.

Sullivan PF (2005). The Genetics of Schizophrenia. *PLoS Medicine* . Vol. **2**: 614-618.

Tanner, M. A. (1996). *Tools for Statistical Inference (3rd ed.)*, New York: Springer Verlag.

Wakefield, J. C.(1996). The bayesian analysis to population pharmacokinetic models. *Journal of the American Statistical Association* **91**: 62-75.

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

Wang,T. and Elston,R.C. (2007) Improved power by use of a weighted score test for linkage disequilibrium mapping. *Am. J. Hum. Genet.* **80**: 353-360.

Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, Chatterjee N, Wang Z, Welch R, Staats BJ, Calle

EE, Feigelson HS, Thun MJ, Rodriguez C, Albanes D, Virtamo J, Weinstein S, Schumacher FR, Giovannucci E, Willett WC, Cancel-Tassin G, Cussenot O, Valeri A, Andriole GL, Gelmann EP, Tucker M, Gerhard DS, Fraumeni JF Jr, Hoover R, Hunter DJ, Chanock SJ, Thomas G. (2007). Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet.* **39**(5): 645-649

Zondervan, K.T. and Cardon, L.R. (2004). The complex interplay among factors that influence allelic association. *Nat. Rev. Genet.* **5**:89-100.