8-2013

BAYESIAN STATISTICAL METHODS IN GENE-ENVIRONMENT AND GENE-GENE INTERACTION STUDIES

Changlu Liu

BAYESIAN STATISTICAL METHODS IN GENE-ENVIRONMENT AND GENE-GENE INTERACTION STUDIES

by

Changlu Liu

APPROVED:

_____
Supervisory Professor: Christopher I. Amos, Ph.D.

_____
Ralf Krahe (On-site advisor), Ph.D.

_____
Jianzhong Ma, Ph.D.

_____
Yunxin Fu, Ph.D.

_____
Richard E. Davis, M.D.

APPROVED:

_____
Dean, The University of Texas
Graduate School of Biomedical Sciences at Houston

BAYESIAN STATISTICAL METHODS IN GENE-ENVIRONMENT AND GENE-GENE INTERACTION STUDIES

A DISSERTATION

Presented to the Faculty of

The University of Texas

Health Science Center at Houston

and

The University of Texas

M.D. Anderson Cancer Center

Graduate School of Biomedical Sciences

in Partial Fulfillment of Requirements

for the Degree of

Doctor of Philosophy

by

Changlu Liu

Houston, Texas, USA

August 2013

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincerest gratitude to my supervisor, Dr. Christopher I. Amos for inspiring my research work and guiding me with endless patience in the past few years. There have been several tough times from which I would not be able to reach this point without his encouragement and support. I owe many thanks to Dr. David Ma as my study mentor and friend who generously offered me many invaluable points of advices in both research and soft skills. I am particularly thankful to my on-site advisor, Dr. Ralf Krahe for his continuous support, care and openness to me whenever I needed helps. I also want to acknowledge and thank Dr. Yunxin Fu and Dr. Richard E. Davis for their insightful suggestions that substantially improved my research and presentation. All the committee members deserve my highest respects. I am especially grateful to Dr. Victoria P. Knutson for providing phenomenal guidances and oversights during my whole study at GSBS.

Many thanks also belong to my family and friends during these years. I will never forget the time you gave: you are the ones I could rely on to overcome those insuperable difficulties.

In the next journey as being a statistician at Novartis right after school, I will endeavor in my career to make all of you I love be proud of me.

ABSTRACT

BAYESIAN STATISTICAL METHODS IN GENE-ENVIRONMENT AND GENE-GENE INTERACTION STUDIES

by

Changlu Liu

Biomathematics and Biostatistics

Graduate School of Biomedical Sciences

UTHealth and MDACC

August, 2013

Complex diseases such as cancer result from multiple genetic changes and environmental exposures. Due to the rapid development of genotyping and sequencing technologies, we are now able to more accurately assess causal effects of many genetic and environmental factors. Genome-wide association studies have been able to localize many causal genetic variants predisposing to certain diseases. However, these studies only explain a small portion of variations in the heritability of diseases. More advanced statistical models are urgently needed to identify and characterize some additional genetic and environmental factors and their interactions, which will enable us to better understand the causes of complex diseases. In the past decade, thanks to the increasing computational capabilities and novel statistical developments, Bayesian methods have been widely applied in the genetics/genomics researches and demonstrating superiority over some regular approaches in certain research areas. Gene-environment and gene-gene interaction studies are among the areas where Bayesian

methods may fully exert its functionalities and advantages.

This dissertation focuses on developing new Bayesian statistical methods for data analysis with complex gene-environment and gene-gene interactions, as well as extending some existing methods for gene-environment interactions to other related areas. It includes three sections: (1) Deriving the Bayesian variable selection framework for the hierarchical gene-environment and gene-gene interactions; (2) Developing the Bayesian Natural and Orthogonal Interaction (NOIA) models for gene-environment interactions; and (3) extending the applications of two Bayesian statistical methods which were developed for gene-environment interaction studies, to other related types of studies such as adaptive borrowing historical data.

We propose a Bayesian hierarchical mixture model framework that allows us to investigate the genetic and environmental effects, gene by gene interactions (epistasis) and gene by environment interactions in the same model. It is well known that, in many practical situations, there exists a natural hierarchical structure between the main effects and interactions in the linear model. Here we propose a model that incorporates this hierarchical structure into the Bayesian mixture model, such that the irrelevant interaction effects can be removed more efficiently, resulting in more robust, parsimonious and powerful models. We evaluate both of the 'strong hierarchical' and 'weak hierarchical' models, which specify that both or one of the main effects between interacting factors must be present for the interactions to be included in the model. The extensive simulation results show that the proposed strong and weak hierarchical mixture models control the proportion of false positive discoveries and yield a powerful approach to identify the predisposing main effects and interactions in the studies with complex gene-environment and gene-gene interactions. We also compare these two models with the 'independent' model that does not impose this hierarchical constraint

and observe their superior performances in most of the considered situations. The proposed models are implemented in the real data analysis of gene and environment interactions in the cases of lung cancer and cutaneous melanoma case-control studies. The Bayesian statistical models enjoy the properties of being allowed to incorporate useful prior information in the modeling process. Moreover, the Bayesian mixture model outperforms the multivariate logistic model in terms of the performances on the parameter estimation and variable selection in most cases. Our proposed models hold the hierarchical constraints, that further improve the Bayesian mixture model by reducing the proportion of false positive findings among the identified interactions and successfully identifying the reported associations. This is practically appealing for the study of investigating the causal factors from a moderate number of candidate genetic and environmental factors along with a relatively large number of interactions.

The natural and orthogonal interaction (NOIA) models of genetic effects have previously been developed to provide an analysis framework, by which the estimates of effects for a quantitative trait are statistically orthogonal regardless of the existence of Hardy-Weinberg Equilibrium (HWE) within loci. Ma et al. (2012) recently developed a NOIA model for the gene-environment interaction studies and have shown the advantages of using the model for detecting the true main effects and interactions, compared with the usual functional model. In this project, we propose a novel Bayesian statistical model that combines the Bayesian hierarchical mixture model with the NOIA statistical model and the usual functional model. The proposed Bayesian NOIA model demonstrates more power at detecting the non-null effects with higher marginal posterior probabilities. Also, we review two Bayesian statistical models (Bayesian empirical shrinkage-type estimator and Bayesian model averaging), which were developed for the gene-environment interaction studies. Inspired by these Bayesian models, we develop two novel statistical methods that are able to handle the

related problems such as borrowing data from historical studies. The proposed methods are analogous to the methods for the gene-environment interactions on behalf of the success on balancing the statistical efficiency and bias in a unified model. By extensive simulation studies, we compare the operating characteristics of the proposed models with the existing models including the hierarchical meta-analysis model. The results show that the proposed approaches adaptively borrow the historical data in a data-driven way. These novel models may have a broad range of statistical applications in both of genetic/genomic and clinical studies.

TABLE OF CONTENTS

Page

Page

Page

List of Figures

List of Tables

1. Introduction

1.1 Background

Complex diseases have been recognized as resulting from the effects of multiple genetic changes and environmental exposures, which usually interact to cause the diseases and therefore are difficult to investigate by traditional approaches. The advances in genome-wide association studies (GWAS) have provided powerful tools to study the genetic contribution to complex diseases. In the past few years, a large number of robust associations between chromosomal loci and diseases have been identified by GWAS [1] [2]. However, the genetic variants identified by GWAS only explain a small proportion of the heritability of diseases [3] and it is speculated that gene-environment and gene-gene interactions may explain some of the missing heritability [5]. Moreover, typical applications of GWAS analysis have not considered the gene-environment and gene-gene interactions [4]. Accurately completing association studies with interactions remains a challenging task for researchers. We are therefore trying to develop advanced statistical methods to investigate this problem to better understand the disease heritability and causality.

Unlike the Mendelian diseases for which the single genetic variants affect the outcomes, complex diseases and disorders have been known to be affected by multiple genetic and environmental factors [4] [5]. Approaches for modeling gene-gene (GxG) and gene-environment (GxE) interactions among risk factors have been moderately investigated in research areas such as genetics, genomics, evolution and epidemiology in recent years [6] [8]. The term 'interaction' has several meanings in different study

backgrounds. Here we focus our interests on the definition of interaction as an effect departure from the additivity of the main effects. The genetic expression level of certain genes can be inhibited or induced by certain environmental factors. Also, the environmental factors' effects could be adjusted by the effects from genetic factors. In this dissertation, we will explore the models of interactions based on the statistical definitions.

1.2 Definition of Gene-Environment and Gene-Gene Interactions

In order to correctly analyze and interpret gene-environment and gene-gene interactions, it is first essential to describe how the interactions are defined. The challenge in statistical analysis of interaction studies is that the gene-environment and gene-gene interactions are not uniquely defined in the literatures. To avoid ambiguity of concepts, in this section I will first discuss the 'interaction' in statistical views and then introduce the concrete concepts commonly applied in genetic studies. Also, the biological interpretation of statistical interactions will be discussed.

1.2.1 Statistical Interaction

The term 'interaction' usually refers to a phenomenon in which several variables jointly influence the outcome. The target in Quantitative Trait Locus (QTL) and association studies is to investigate the relationship between the outcome (trait) Z and the genetic factors $X = (X_1, ..., X_J)$ and environmental factors $Y = (Y_1, ..., Y_K)$, where J and K are the total number of factors considered respectively. This can be expressed in a statistical model as

$$f(E(Z)) = \Phi(X, Y) = \Phi(X_1, ..., X_J, Y_1, ..., Y_K) \tag{1.1}$$

where $f(.)$ is the link function, which connect the observed outcome and the factors; $E(.)$ is the expectation of response trait and $\Phi(.)$ is an unknown function that connects

the factors to the transformation of the expectation. For a quantitative (continuous) trait, a normally distributed response may be assumed. For a qualitative (discrete) trait, we may assume it follows Bernoulli distribution. Were the parameters in $\Phi(.)$ presented in a linear way, the model would be generally termed as 'Linear Model'. For example, in the QTL analysis, the trait is usually assumed normally distributed and a linear regression model would be built to identify the predisposing loci from a candidate locus set.

Usually multiple genetic and environmental factors are believed to jointly affect the outcome. In this dissertation, we first restrict our attention to the simplest case of two-way (factor) interactions. There are three possible types of interactions for the genetic and environmental factors: gene-environment(GxE), gene-gene(GxG) and environment-environment(ExE) interactions. The ExE interactions could be included as covariates in the model and the environmental measures usually include considerable measurement errors, so researchers mainly focus on GxE and GxG interactions. It is worthwhile to mention that, although the formulation of GxE and GxG interactions are similar, the interpretations are rather different. With only one genetic factor X_1 and one environmental factor Y_1, if the two factors do not interact, then by the definition of additivity, $\Phi(X_1, Y_1) = \Phi_x(X_1) + \Phi_y(Y_1)$, where Φ is the function of the variable [9]. This implies that the environmental effect of Y_1 will not affect the genetic effect of X_1. Here, the environmental effect is often not of main interest, but could be an potential confounding factor. In GxG interactions, we first consider two genetic factors X_1 and X_2. If the functions of two factors could be placed by two separate functions $\Phi(X_1, X_2) = \Phi_{x_1}(X_1) + \Phi_{x_2}(X_2)$, then it would be claimed that the genetic effect of X_1 does not depend on the genetic effect of X_2 and vice versa. If the condition is not satisfied, the response due to to a change in one factor will be dependent on the level of another factor. In the model with multiple genetic

and environmental factors, we regard the interactions as a product term of the main effect variables. In the most common cases, the genetic factors have 3 levels (0,1,2) and environmental factor has 2 levels (0,1). We naturally start to build up a model for a two-way GxE interaction as

$$\begin{aligned}
\Phi(X_1, Y_1) &= a + \mu_j + \tau_k + \sigma_{jk} \\
\Phi(0,0) &= a \\
\Phi(1,0) &= a + \mu_1 \\
\Phi(2,0) &= a + \mu_2 \\
\Phi(0,1) &= a + \tau_1 \\
\Phi(1,1) &= a + \mu_1 + \tau_1 + \sigma_{11} \\
\Phi(2,1) &= a + \mu_2 + \tau_1 + \sigma_{21} \qquad (1.2)
\end{aligned}$$

where α is the baseline (intercept), μ_1 and μ_2 are the main effects for genetic levels 1 and 2, τ_1 is the main effect for the environment factor and σ_{11} and σ_{21} are the two interactions between the genetic and environmental factors. In practice, we usually place the above terms under the linear model framework. Since the genetic factor has 3 levels, two dummy variable X_{11} and X_{12} should be created to represent the levels of the genetic factor, and then the model will become

$$\Phi(X_1, Y_1) = \alpha + \beta_1 X_{11} + \beta_2 X_{12} + \gamma_1 Y_1 + \theta_1 X_{11} Y_1 + \theta_2 X_{12} Y_1 \qquad (1.3)$$

where for the dummy variables, $X_{11} = 1$ and $X_{12} = 0$ if $X_1 = 1$; $X_{11} = 0$ and $X_{12} = 1$ if $X_1 = 1$; $X_{11} = 0$ and $X_{12} = 0$, if $X_1 = 2$.

1.2.2 Interaction in Genetic Studies

The genetic factors $X_1, ..., X_J$ usually take on three values (0,1,2) for example, to indicate the number of minor alleles at the locus, so one level will be excluded as the baseline or reference effect. When we consider the GxG interactions, in the simplest

case of 2 genetic factors, there will be 4 main effects and 4 interactions. In human genetic association studies, this model is named co-dominant model [10].

$$\begin{aligned}\Phi(X_1, X_2) &= \alpha + (X_{1a}a_1 + X_{1d}d_1) + (X_{2a}a_2 + X_{2d}d_2) \\ &\quad + (X_{1a}X_{2a}aa_{12} + X_{1a}X_{2d}ad_{12} + X_{1d}X_{2a}da_{12} + X_{1d}X_{2d}dd_{12}) \quad (1.4)\end{aligned}$$

Here X_{1a} and X_{1d} are the variables for X_1 with $X_{1a} = 1$ if $X_1 = 1$; $X_{1a} = 0$ otherwise, and $X_{1d} = 1$ if $X_1 = 2$; $X_{1d} = 0$ otherwise. Here a_1, a_2, d_1 and d_2 are the main effects and aa_{12}, ad_{12}, da_{12} and dd_{12} represent the interactions. There are alternative ways to code the variables X_{1a}, X_{1d}, X_{2a}, and X_{2d}. An alternative widely used method is the Cockerham model [7], which codes the main effect variables as

$$\begin{aligned}&X_{1a} = X_1 - 1,\ X_{2a} = X_2 - 1 \\ &X_{1d} = (X_1 - 1)(3 - X_1) - 0.5,\ X_{2d} = (X_2 - 1)(3 - X_2) - 0.5 \quad (1.5)\end{aligned}$$

where a_1 and a_2 are the additive effects, d_1 and d_2 are the dominance effects and aa_{12}, ad_{12}, da_{12} and dd_{12} correspond to the interactions of additive by additive, additive by dominance, dominance by additive and dominance by dominance, respectively.

1.3 Generalized Linear Model for Interaction Studies

For studying various types of phenotypes (disease status or complex trait), the generalized linear models have been implemented for detecting the interactions [10]. A generalized linear model is composite of three parts: link function, linear predictor and the distribution of phenotypes [11]. The linear predictor is a linear combination of the independent variables. The link function relates the mean of the outcome to the linear predictor. So if the link function $f(.)$ is an identify function, the model will degenerate to be a linear regression model. The phenotype Z can be assumed to follow different types of distributions such as Normal, Logistic, Binomial, Log-

normal and Poisson, depending on the response outcomes. Usually for a count type of response, a Poisson distribution is assumed and the link function will be

$$f(E(Z)) = log(E(Z)) \tag{1.6}$$

For a binary outcome that is Bernoulli distributed, the link function $f(E(Z))$ could be any of the followings:

$$\begin{aligned} &\text{logit: } log(\frac{E(Z)}{1-E(Z)}) \\ &\text{probit: } \Psi(E(Z)) \\ &\text{cloglog: } log(-log(1-E(Z))) \end{aligned} \tag{1.7}$$

Modeling the GxE and GxG interactions will be more complicated in the generalized linear model than in the linear regression model due to the link function connecting the response and the linear predictor. In the generalized linear model, the genetic and environmental effects correspond to a transformation of the expectation of the response. For example, in a logistic regression, the estimates of effects will be based on the scale of the log odds of success $log\frac{Pr(Z=1)}{1-Pr(Z=1)}$. However, some generalized linear models may be viewed as a linear regression model with a latent variable [12]. For example, in the logistic model with a latent variable T, the model will be

$$\begin{aligned} &T \sim N(\Phi(X,Y), \sigma^2) \\ &Z = 1 \text{ if } T > 0 \\ &Z = 0 \text{ if } T < 0 \end{aligned} \tag{1.8}$$

This type of formulation by latent variables provides a computational efficient approach to model the effects, that it is widely applied in Bayesian modeling. It also renders a convenient way to explicitly interpret the study results from the generalized linear model.

The genetic and environmental effects and interactions estimates depend on the link function. It is possible that some interaction results will change when the selected link functions change. This is analogous to the scenario of log-normal transformation in a simplified linear regression model [13], for example,

$$\begin{aligned} Z &= X * Y \\ log(Z) &= log(X) + log(Y) \end{aligned} \tag{1.9}$$

An interaction is called 'removable', when a transformation of the outcome scale exists to induce additivity [14]. It is of great importance to decide if the detected interaction is removable or not. If some interactions could be removed by changing the link function or making a simple transformation, the resulting model fitting would be improved and the interpretation would be more straightforward. For a continuous outcome, the Box-Cox transformation could be attempted to transform the outcome or the linear predictor to fit the data. For a binary outcome, the proposed different generalized linear model could be implemented separately for the same dataset to select the most efficient model. When there are no interaction terms in a generalized linear model, it is still possible that the effects of one factor depends on the effects of other factors in the model due to the link function $f(.)$. However, this dependence will not affect the transformed value of the expected value of the response $f(E(Z))$. As long as the interaction terms included improve the model fit, we will claim that the model captures some existing interactions between variables by adding the interaction terms. Bayesian hierarchical methods actually provides effective tools to analyze the data using different models in which the GxE and GxG interactions can be investigated simultaneously.

1.4 Existing Models for Detecting Interactions

To detect interactions involves the specification of statistical models and the findings will depend on the selected models. There are many different methods for detecting GxE and GxG interactions such as the two-stage design [15], multifactor dimensionality reduction (MDR) [16], and tree based model (random forest) [17], among many others. Here, we will only focus on introducing penalized regression models and Bayesian hierarchical models used in the GxE and GxG interactions, based on which our proposed models in chapters 2 and 3 are built.

1.4.1 Penalized Regression Model

Statistical variable selection techniques have been deeply explored in statistics. When we seek to evaluate the effects of many potential variables with a limited number of observations, finding the predictors that parsimoniously explain the variations in the dependent variables is rather challenging. There are also many statistical models that have been proposed for variable selection with considerations of interactions.

In the classical statistical models, the estimation of effects are achieved by obtaining the Maximum likelihood estimation (MLE). But a generalized linear model with large number of variables or correlated variables often fails to identify the true affecting variables. A standard remedy is to add a penalty function to the likelihood function, enabling the selection of variables by certain criteria:

$$J(\Theta) = log(L(\Theta|data)) - \pi(\Theta) \tag{1.10}$$

where $J(\Theta)$ is the penalized likelihood function, $L(\Theta|data)$ represents the likelihood and $\pi(\Theta)$ is the penalty function. Adding parameters will increase the value of $L(\Theta|data)$ but also increase the penalty $\pi(\Theta)$. Therefore, $J(\Theta)$ is a balancing combination of the model fit and complexity. By applying the penalty function $\pi(\Theta)$,

we may constrain the parameters under the traditional frameworks or place prior information under Bayesian framework. While the penalized likelihood model can stabilize the estimation of the model, it also provides a criterion for model selection and comparison. The selection of the penalty function $\pi(\Theta)$ controls the general performance of the penalized likelihood approach. Small penalties will correspond to the selection of models with large number of parameters that are less biased but will bring large variances due to the correlation within parameters. In contrary, large penalties will lead to reduced models that will generate relatively large bias and small variance.

Many statistical methods for variable selection utilize this penalized likelihood framework, such as the Akaike Information Criteria (AIC) [18] and Bayesian Information criteria (BIC) [19]. In AIC, the penalty $\pi(\Theta)$ is related to the total number of parameters $dim(\Theta)$ included in the model. The BIC criteria is related to the number of parameters and the total sample size N of the data.

$$\begin{aligned} \text{AIC:} \quad & -2log(L(\Theta|data)) + 2dim(\Theta) \\ \text{BIC:} \quad & -2log(L(\Theta|data)) + dim(\Theta) * ln(N) \end{aligned} \tag{1.11}$$

These variable selection criteria have been applied in multiple QTL mapping studies [20]. However, due to the large number of potential variables, these criteria tend to incorrectly include many spurious loci. Therefore it may not be appropriate for the QTL mapping [22]. When interactions are also included in the model, the total number of parameters will increase substantially. Some modifications of these criteria have been applied for the interaction studies, such as assigning different penalties to the main effects and interactions [23] as:

$$log(L(\Theta, \Gamma|data)) - \lambda_\Theta dim(\Theta) - \lambda_\Gamma dim(\Gamma) \tag{1.12}$$

where λ_Θ and λ_Γ represent the different tuning parameters for the penalties on the main effects and pairwise interactions. Cross-validation could be applied to find the

optimal values of these tuning parameters. Forward or backward stepwise selection procedures might also be attempted based on these penalized likelihood models. Usually, the above penalized likelihood approach is called L_0 penalty, which involves the total number of parameters in the modeling process while neglecting the scale of the parameters.

Another popular penalized likelihood approach is the Least absolute shrinkage and selection operator (Lasso) proposed by [24], which is referred to as the L_1 penalty. The estimates of parameters in Lasso are obtained by maximizing the likelihood function subject to a constraint on the sum of the absolute values of parameters in the regression models $||\Theta||_1 < L$, which is equivalent to maximizing the penalized likelihood function:

$$log(L(\Theta|data)) - \lambda||\Theta||_1 \tag{1.13}$$

There is also another group of approaches that are based on the L_2 penalty on the regression coefficients:

$$log(L(\Theta|data)) - \lambda||\Theta||_2^2 \tag{1.14}$$

where $||\Theta||_2 = \sqrt{\sum_J \theta_j^2}$. This regression is called ridge regression and it can effectively handle the problem of collinearity and highly correlated variables but can not select variables. A logistic model with L_2 penalty has been proposed to handle the gene-gene interactions in case-control studies [21]. The model with quadratic penalties can simultaneously fit a large number of factors and their interactions in a stable way. However, it can not shrink any parameters directly to zero as Lasso dose and thus does not perform the variable selection to resolve important signals versus noises. People also applied a forward selection method based on this penalty function to select the variables [25].

Lasso is an effective tool for the analysis of genetic interactions by shrinking the parameters to zero to perform variable selection [28]. A variety of optimization approaches for Lasso have been proposed, which make it possible to be applied in large scale studies [26] [27]. Lasso logistic regression was also applied for genome-wide association analysis in case-control analysis [28]. In the studies, the number of predictors is fixed such that the tuning parameter λ can be pre-determined by running cross validation. For a given λ, the coordinate descent algorithm was applied to select the important genes and interactions. Tanck (2006) [29] also applied the Lasso regression model to detect the gene-gene interaction in association studies with L_2 penalty for the main effects and L_1 penalty for the interaction effects. Therefore, in the modeling process the effective interaction terms can be selected, while all the main effects are retained.

1.4.2 Bayesian Hierarchical Model

Hierarchical models have been applied in the high-dimensional data analysis including QTL mapping and association studies [22]. The models are parameterized in a structured way so that some dependences among the variables are incorporated, thereby fitting a model with a large number of predictors. In non-hierarchical models, model fitting is usually not stable due to the correlations of the estimates, so non-hierarchical methods are not able to handle many variables simultaneously. The hierarchical models can be interpreted and handled more readily under the Bayesian framework [30].

In the Bayesian framework, the prior distributions of parameters reflect the knowledge before observing the data and the statistical inferences are based on the posterior

distributions of parameters, which are proportional to the products of the prior distribution $\pi(\Theta)$ and likelihood function $L(data|\Theta)$.

$$f(\Theta|data) \propto \pi(\Theta) \times L(data|\Theta) \tag{1.15}$$

The posterior distributions will include all the current knowledge/information about the parameters and can be updated as new information accumulates. The primary objective of Bayesian inference is to explore the full posterior distributions of all the parameters, which can be achieved by Markov chain Monte Carlo (MCMC) algorithms. In certain circumstances, we may be interested in certain statistics that could be obtained by more efficient computational algorithms [31]. For example, we may want to find out the posterior modes of the parameters by maximizing the logarithm of the products of the prior and likelihood:

$$log(f(\Theta|data)) = log(L(\Theta|data)) + log(\pi(\Theta)) + C \tag{1.16}$$

where C is a constant irrelative to Θ. This is analogous to the penalized likelihood model solution (1.10) with the logarithm of the prior $log(\pi(\Theta))$ as the penalty function. With certain priors, Bayesian hierarchical models can lead to the penalized likelihood methods mentioned earlier. In Bayesian inferences, it is more comprehensive to investigate the full posterior distributions rather than merely seek the posterior modes as in the penalized likelihood methods.

The specification of priors in the hierarchical models is essentially important. Different types of priors could be proposed for dealing with high-dimensional data. Some of the methods have been employed in the QTL mapping and association studies [22]. In the studies with a large number of predictors, it is reasonable to assume that many of the predictors have weak or no effects on the traits. Therefore, for the regression coefficients, we may set up the priors that assign low probabilities of being significant

to the majority of the predictors.

Shrinkage Priors

One effective way for achieving the target of selecting the significant factors is to assign shrinkage priors for the parameters. There are two commonly used shrinkage priors for this Bayesian hierarchical modeling: double exponential (DE) and Student's t distribution [32]. Suppose that $\beta_j,\ j = 1\ to\ J$ are the regression coefficients in the generalized linear model, in which we want to obtain the posterior distributions by (1.15). Their priors are

$$\begin{aligned} &\text{DE:}\ \ \pi(\beta_j) = (\lambda/2)e^{-\lambda|\beta|} \\ &\text{t:}\ \ \pi(\beta_j) = t_{v_j}(0, \sigma_j^2) \end{aligned} \tag{1.17}$$

where λ and v_j and σ_j^2 are the tuning parameters to control the amount of shrinkage in the hierarchical modeling. Both of the shrinkage priors could be translated to a two-level hierarchical models, where the shrinkage priors are set up by controlling the variances of the priors for the coefficients. The first level of the hierarchical model assumes that the coefficients β_j are normally distributed with mean 0 and variance σ_j^2 and the second level assumes that the variances follow some specific prior distributions:

$$\begin{aligned} \beta_j|\sigma_j^2 &\sim N(0, \sigma_j^2) \\ \sigma_j^2|\tau_j &\sim \eta(\sigma_j^2|\tau_j) \end{aligned} \tag{1.18}$$

where η is the specific prior distribution and τ_j are the hyper-parameters. Then for a DE prior:

$$\eta(\sigma_j^2|\tau_j) = Gamma(1, \tau_j^2/2) \tag{1.19}$$

And for a t prior:

$$\eta(\sigma_j^2|\tau_j) = Inverse - \chi^2(v_j, \tau_j^2) \tag{1.20}$$

There are several advantages for these two-level hierarchical models. Due to the conjugacy of certain priors π_j^2, the Bayesian inferences could be easier to implement, since the optimization is more readily to accomplish. By controlling the different amount of shrinkage variances, some of the parameters ultimately shrink to zero. Also, there is some flexibility in specifying the priors for the hyper-parameters in a hierarchical structure.

Mixture Priors

The second popular class of priors assume two components in a single model: one represents the parameters with null effects and the other with non-null effects. Here we will describe two typical Bayesian mixture priors structure: 'Spike and Slab' and Stochastic Search Variable Selection (SSVS).

A Spike and Slab [33] [34] prior mixture structure assumes that the prior of each parameter is a mixture of a diffuse distribution and a point mass at 0 as

$$\beta_j|I_j, \sigma_j^2 \sim I_j N(0, \sigma_j^2) + (1 - I_j)\mathbf{1}_\emptyset \tag{1.21}$$

where $\mathbf{1}_\emptyset$ is a point mass at 0 and I_j is the binary variable indicating if the effect is present or not in the model. In the second level, a Bernoulli distribution is assumed for the indicators and a hyper prior is assumed for the variance components.

$$\begin{aligned} I_j &\sim Bernoulli(p_j) \\ \sigma_j^2 &\sim Inverse - \chi^2(v_j, \tau_j^2) \end{aligned} \tag{1.22}$$

So the variable selection performance such as sparseness can be controlled by the specifications of hyper-priors by (1.22). Stochastic Search Variable Selection structures were proposed by George and McCullogh [35] [36]. Rather than assuming a

point mass at zero as the 'Spike and Slab' priors, SSVS applies a prior component condensed at the 'Null' effect part $N(0, \sigma^2_{\epsilon j})$ as

$$\beta_j | I_j, \sigma^2_j, \sigma^2_{\epsilon j} \sim I_j N(0, \sigma^2_j) + (1 - I_j) N(0, \sigma^2_{\epsilon j}) \tag{1.23}$$

These mixture prior approaches have been applied in QTL mapping studies [37] and [38]. In the following chapters, we choose to focus on the development of models with SSVS, because the variances within the 'Null' effect part may capture some effect bias arising from genotyping errors or confounding due to study designs.

2. Bayesian Variable Selection for Hierarchical Gene-Environment and Gene-Gene Interactions

2.1 Motivation

2.1.1 Biological Background

Complex diseases are influenced by multiple genetic and environmental factors. The factors may interact and are hard to directly discern. The advances in genome-wide association studies (GWAS) have provided powerful tools to study the genetic contributions to complex diseases [1]. During the past few years, a large number of robust associations between chromosomal loci and complex diseases have been identified by GWAS. However, the genetic variants identified by GWAS only explain a small proportion of the heritability of most diseases and it is speculated that gene-gene and gene-environment interactions could explain some of the missing heritability [5]. Moreover, typical applications of GWAS analysis do not study gene-gene and gene-environment interactions. We therefore developed advanced statistical methods jointly modeling interactions and main effects to address the problem to explaining better disease heritability and causality.

Gene-gene and gene-environment interactions among risky factors have already been widely investigated in genetics, genomics, evolution and epidemiology. The term 'interaction' has many meanings. Here we focus on the statistical definition of interaction as a departure from additivity of the main effects. The genetic expression level of certain genes can be inhibited or induced by certain environmental factors, which causes a biological interaction that may result in a statistical interaction. Also,

the environmental factors' effects could be modified according to the effects from genetic factors.

2.1.2 Statistical Background

Bayesian Variable Selection

Variable selection has been extensively studies in statistics. When we seek to evaluate the effects of many potential variables and have a limited number of observations, finding the predictors that parsimoniously explain the variations in the dependent variables is challenging. There are many statistical models that have been proposed for variable selection. Here we focus on the Bayesian hierarchical mixture model. In the Bayesian framework, we can view variable selection as identifying the non-zero regression parameters based on the posterior distributions of the parameters. Different priors have been considered for this variable selection purpose. A spike and slab [33] prior mixture structure was proposed by assuming that the prior of each parameter is a mixture of a diffuse distribution and a point mass at 0. These two components in the priors represent the prior belief about an effect's existence or not. George and MuCulloch (1993, 1997) [35] [36] proposed the Stochastic Search Variable Selection (SSVS) model that assumes the prior for each parameter as a mixture of two distributions, both of which are typically centered at 0 but with different magnitudes of variances for the corresponding normal density functions.

Hierarchical Interaction

For statistical modeling of interactions, there may be a hierarchical structure among the predictors. For example, if we consider setting up a regression model with

dependent variable Z and three independent variables X_1, X_2, X_3, then the model with all the two-way interactions will be

$$f(E(Y)) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \gamma_{12} X_1 X_2 + \gamma_{13} X_1 X_3 + \gamma_{23} X_2 X_3$$

where $f(.)$ is the link function. In a usual statistical modeling approach, when inferring each effect or variable selection, the main effect parameters are not treated differently from interaction parameters, which means there is no constraint on the parameter space for these two kinds of parameters. However, the resulting model can be difficult to interpret. Moreover, usually if an interaction is present, the main effects will be nonzero. In practice, a statistical researcher commonly requires that the interaction effects should be included with their corresponding main effects, otherwise the resulting model tends to be rather unstable. In the statistical literature, Hamada and Wu (1992) [39] first introduced the concept of hierarchical interaction that was named 'Heredity principle'. There are two versions of hierarchical constraints on effects [13] [44]. Under the strong hierarchical constraint, for any two-way interaction term to be included in the model, both of the main effects must also be included; whereas under the weak hierarchical constraint, for any two-way interaction term to be included in the model, one of the main effects must be included. In other words,

Strong Hierarchical Interaction: $\gamma_{12} \neq 0 \Rightarrow \beta_1 \neq 0 \cap \beta_2 \neq 0$,

Weak Hierarchical Interaction: $\gamma_{12} \neq 0 \Rightarrow \beta_1 \neq 0 \cup \beta_2 \neq 0$.

These effect constraints have been studies extensively in the variable selection literature. Choi (2009) [40] and Jacob (2013) [41] imposed this constraint on the popular variable selection approach Least Absolute Shrinkage and Selection Operator (Lasso). Yuan (2007, 2009) [42] [43] improved other original variable selection approaches (LARS and Nonnegative Garrote) with this constraint. The Bayesian model with this constraint was first introduced by Chipman (1996) [13]. In his paper,

they proposed a hierarchical prior structure for modeling the interaction effects with the constraint. But their approach does not explain explicitly the rationale of the prior specifications. In this project, we specify the priors of the main effects and interactions and incorporate them into the Bayesian mixture model we proposed.

2.2 Methodology

2.2.1 Classic Bayesian Hierarchical Model

Suppose we study n observations. We denote z_i as the disease status (1 for positive, 0 for negative) with $i = 1, ..., n$. Let x_{ij} denote the number of minor alleles for the jth SNP of ith observation, y_{ik} as the kth environmental exposure for ith observations. In logistic regression, we will include the main effects of both gene and environment as well as gene-environment and gene-gene interaction analysis as

$$z_i \sim Bernoulli(p_i)$$

$$log\frac{p_i}{1-p_i} = \alpha + \sum_j \beta_j x_{ij} + \sum_k \gamma_k y_{ik} + \sum_j \sum_k \theta_{jk} x_{ij} y_{ik} + \sum_{j<l} \eta_{jl} x_{ij} x_{il}$$

where p_i indicates the probability of disease. α is the general intercept and $exp(\alpha)$ is the baseline odds for the disease if we do not include any predictors in the model. β_j is the genetic effect and $exp(\beta_j)$ denotes the increase of odds with one minor allele count increase. γ_k is the environmental effect for the kth exposure and $exp(\gamma_k)$ denotes the increase of odds with the environmental exposure present in the model. θ_{jk} denotes the gene-environment interaction effects between the jth SNP and the kth environmental factor and η_{jl} denotes the interaction effect between jth and lth SNP. Since the gene-gene products $x_{ij}x_{il}, j \neq l$ are internally symmetric, it is reasonable to assume that $\eta_{jl} = \eta_{lj}$, therefore we index the parameters as $j < l$. For example, when we consider 6 candidate SNPs with 1 candidate EXPs, 6 gene-environment in-

teractions and 15 gene-gene interaction parameters will also be included in the model.

Then for each parameter, we shall assume normal priors for the parameters in the regression:

$$\begin{aligned}
\beta_j &\sim N(0, \sigma_j^2) \\
\gamma_k &\sim N(0, \sigma_k^2) \\
\theta_{jk} &\sim N(0, \sigma_{jk}^2) \\
\eta_{jl} &\sim N(0, \sigma_{jl}^2)
\end{aligned} \tag{2.1}$$

On the next level, the hyperpriors are assumed for the variance components for each parameter:

$$\begin{aligned}
\sigma_j^{-2} &\sim Gamma(a_j, b_j) \\
\sigma_k^{-2} &\sim Gamma(a_k, b_k) \\
\sigma_{jk}^{-2} &\sim Gamma(a_{jk}, b_{jk}) \\
\sigma_{jl}^{-2} &\sim Gamma(a_{jl}, b_{jl})
\end{aligned} \tag{2.2}$$

These structures are from the conventional Bayesian hierarchical model for proposing the priors for the parameters [12]. Recently, researchers have started to modify this framework for the purpose of variable selection. Park and Casella (2008) [32] and Yi (2009) [22] proposed the Bayesian Lasso model by proposing some common priors for the parameters, by which the trivial parameters would be flattened out the model. In this paper, we work on proposing different structures for the variances by assuming the mixture components.

2.2.2 Bayesian Mixture Model

Under the null hypothesis of no effects, we impose a prior distribution for each parameter, while tiny variances imply a condensed mass distribution centered at 0:

$$\begin{aligned}
\beta_j &\sim N(0, \sigma^2_{\epsilon j}) \\
\gamma_k &\sim N(0, \sigma^2_{\epsilon k}) \\
\theta_{jk} &\sim N(0, \sigma^2_{\epsilon jk}) \\
\eta_{jl} &\sim N(0, \sigma^2_{\epsilon jl})
\end{aligned} \tag{2.3}$$

Under the alternative hypothesis of nonzero effects, similar priors with larger variances are given as:

$$\begin{aligned}
\beta_j &\sim N(0, \sigma^2_{j}) \\
\gamma_k &\sim N(0, \sigma^2_{k}) \\
\theta_{jk} &\sim N(0, \sigma^2_{jk}) \\
\eta_{jl} &\sim N(0, \sigma^2_{jl})
\end{aligned} \tag{2.4}$$

Motivated by the Stochastic Search Variable Selection framework [35] [36], a mixture model is proposed here for the modeling of the effects in the logistic regression model. Therefore, the priors with the indicators can be written as:

$$\begin{aligned}
\beta_j &\sim N(0, I_j\sigma^2_{j} + (1 - I_j)\sigma^2_{\epsilon j}) \\
\gamma_k &\sim N(0, I_k\sigma^2_{k} + (1 - I_k)\sigma^2_{\epsilon k}) \\
\theta_{jk} &\sim N(0, I_{jk}\sigma^2_{jk} + (1 - I_{jk})\sigma^2_{\epsilon jk}) \\
\eta_{jl} &\sim N(0, I_{jl}\sigma^2_{jl} + (1 - I_{jl})\sigma^2_{\epsilon jl})
\end{aligned} \tag{2.5}$$

Each of the indicators in the model follows a Bernoulli distribution as:

$$
\begin{aligned}
I_j &\sim Bernoulli(\pi_j) \\
I_k &\sim Bernoulli(\pi_k) \\
I_{jk} &\sim Bernoulli(\pi_{jk}) \\
I_{jl} &\sim Bernoulli(\pi_{jl})
\end{aligned} \tag{2.6}
$$

2.2.3 Bayesian Mixture Model for Hierarchical Interactions

Under the strong hierarchical interaction model, we propose a scheme to describe the relationships among the indicators of these parameters as

$$
\begin{aligned}
\pi_{jk} &= I_j \times I_k \times \pi^s_{jk} \\
\pi_{jl} &= I_j \times I_l \times \pi^s_{jl}
\end{aligned} \tag{2.7}
$$

where π^s_{jk} and π^s_{jl} are the conditional prior probability of the indicator for interaction effects being non-null given both the main effects being non-null under the weak hierarchical interaction model. For each conditional prior probabilities, we assume that

$$
\begin{aligned}
\pi^s_{jk} &= min(\pi_j, \pi_k) \\
\pi^s_{jl} &= min(\pi_j, \pi_l)
\end{aligned} \tag{2.8}
$$

Of course, other frameworks for the priors for these conditional priors can be proposed. By assuming the strong hierarchical effects model, the prior probability should be more consistent with each of the main effects.

Under the weak hierarchical interaction model, we propose a scheme for describing the relationship between the interaction effects and the main effects as

$$
\begin{aligned}
\pi_{jk} &= (I_j + I_k - I_j * I_k) \times \pi^w_{jk} \\
\pi_{jl} &= (I_j + I_l - I_j * I_l) \times \pi^w_{jl}
\end{aligned} \tag{2.9}
$$

where π_{jk}^{w} and π_{jl}^{w} are the conditional prior probability of the indicator for interaction effects being non-null given both the main effects being non-null under the strong hierarchical interaction model. For each of the conditional prior probability, we assume that

$$\pi_{jk}^{w} = min(\pi_j, \pi_k) \times \frac{I_j\pi_j + I_k\pi_k}{\pi_j + \pi_k}$$
$$\pi_{jl}^{w} = min(\pi_j, \pi_l) \times \frac{I_j\pi_j + I_l\pi_l}{\pi_j + \pi_l} \tag{2.10}$$

So when both of the corresponding main effects for the gene-environment interactions are present in the model, we denote the conditional probability as $\pi_{jk}^{w11} = min(\pi_j, \pi_k)$;
When both of the main effects are missing in the model, $\pi_{jk}^{w00} = 0$;
When $I_j = 0$ and $I_k = 1$, we denote the conditional prior probability as

$$\pi_{jk}^{w01} = min(\pi_j, \pi_k) \times \frac{\pi_k}{\pi_j + \pi_k} \tag{2.11}$$

When $I_j = 1$ and $I_k = 0$, we denote the conditional prior probability as

$$\pi_{jk}^{w10} = min(\pi_j, \pi_k) \times \frac{\pi_j}{\pi_j + \pi_k}$$

Therefore, $\pi_{jk}^{w00} \leq (\pi_{jk}^{w01}, \pi_{jk}^{w10}) \leq \pi_{jk}^{w11}$ and $\pi_{jk}^{w01}/\pi_{jk}^{w10} = \pi_k/\pi_j$. These structures reflect our belief that the interaction terms will be more likely to be non-null in the model when both main effects are present than one of the main effect is missing. Also, larger main effects are more likely to bring appreciable interactions than small main effects. As mentioned in Jacob (2013) [41], Cox(1984) [9] first brought up this model constraint in the statistical literature. Our Bayesian model will naturally take into account these properties in the prior setups. It would be rather difficult to consider all these constraints in the frequentist model framework.

In our models, we fix the two mixture variances components for each of the parameters. Usually the total number of the factors considered is large, so it may be

possible to estimate the variances from the data. In logistic regression, for the main effect parameters, $exp(\beta_j)$ and $exp(\gamma_k)$ correspond to the relative odds for the disease when only including these individual parameters are included in the model. Therefore, we fix the variance components by restricting their confidence interval within a prespecified range of disease odds. So under the null hypothesis, we assume the 95% confidence interval will be $e^{1.96\sigma_{\epsilon j}} = e^{\pm 0.05}$ and $e^{1.96\sigma_{\epsilon k}} = e^{\pm 0.05}$. These specification will provide that the odds will be within interval $(0.951, 1.051)$. Similarly, we assume the same variances for the null-component of the interaction effects. Under the alternative hypothesis, we set up the variances components by restricting the odds within a certain range. We assume the 90% confidence interval of odds would be at $(\frac{1}{3}, 3)$ for genetic main and gene-gene Interaction effects and $(\frac{1}{4}, 4)$ for environmental exposure main and gene-environment interaction effects. Therefore, we set the values for the hyper priors for the variances as:

$$\begin{aligned}
&\sigma^2_{\epsilon j} = \sigma^2_{\epsilon k} = \sigma^2_{\epsilon jk} = \sigma^2_{\epsilon jl} = 0.00065 \\
&\sigma^2_j = \sigma^2_{jl} = 0.446 \\
&\sigma^2_k = \sigma^2_{jk} = 0.710
\end{aligned} \tag{2.12}$$

2.2.4 Posterior Inference

We have provided the prior structures for the Bayesian mixture model for the Hierarchical Interaction Model. The full posterior distribution of the parameters given the data would be

$$\begin{aligned}
&f(\alpha, \beta, \gamma, \theta, \eta, I^\beta, I^\gamma, I^\theta, I^\eta | data) \propto f(data|\alpha, \beta, \gamma, \theta, \eta, I^\beta, I^\gamma, I^\theta, I^\eta) \times \\
&f(\alpha) f(\beta|I^\beta) f(\gamma|I^\gamma) f(\theta|I^\theta) f(\eta|I^\eta) \times f(I^\beta) f(I^\gamma) f(I^\theta|I^\beta, I^\gamma) f(I^\eta|I^\beta)
\end{aligned} \tag{2.13}$$

where β is the parameter vector for genetic main effects, γ for environmental exposure main effects, θ for gene-environment interaction effects and η for gene-gene Interaction

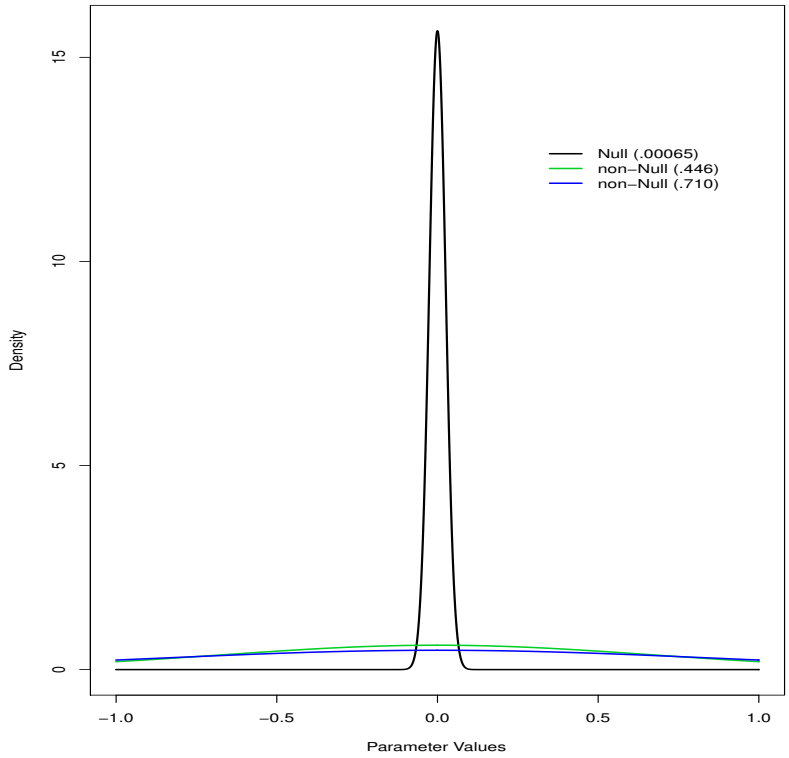

Figure 2.1.: Two components for the mixture priors of parameters. 'Null (.00065)' corresponds to the 95% C.I. of Odds (.951,1.051), 'Non-Null(.446)' corresponds to 90% C.I. of Odds (1/3,3) and 'Non-null(.771)' corresponds to 90% C.I. of Odds (1/4,4).

effects. Also, I^{β}, I^{γ}, I^{θ} and I^{η} are the corresponding indicator vectors. We apply WinBUGS to implement the Markov chain Monte Carlo algorithms for the posterior inferences of the parameters, such that the posterior samples of parameters can be drawn [45].

2.3 Empirical Results

In this section, we use simulation studies to present the efficacy of the proposed approach and to compare the results with the model that does not consider the hier-

archical interactions in the models. Certainly, the efficacy of the models will depend on the true model generating the data. To provide broad evaluations, we conducted three simulation studies to cover a range of possible scenarios in practice. We propose three different studies:

Study I: 6 genetic factors (SNP), 1 environmental exposure factor (EXP) and 6 gene-environment Interaction factors (GEI)

Study II: 50 SNPs, 1 EXP and 50 INTs

Study III: 6 SNPs, 1 EXP, 6 GEI and 15 pairwise gene-gene Interaction factors (GGI)

We compared the performance of four models based on Bayesian mixture models. Wakefield (2010) [46] proposed a model structure for the interaction that includes the interaction term when both of the main effects are present in the model as: $I_{jk} = I_j \times I_k$. We will denote this as 'effect enforce' model as in Chipman et al. (2006) [47]. Also we considered the independent model which does not impose any constraint on the relationships between the interaction parameters and the main effect parameters. Then we will consider the proposed hierarchical interaction models: Strong hierarchical model and weak hierarchical model.

In the simulation studies, we compared the four models in terms of the prediction accuracy and variable selection performance. In the 3 studies, we generated 100 replicates with 1000 cases and 1000 controls for each scenario under the same parameters. For measuring the prediction accuracy, we compare the prediction errors (PE) on a test set with 20000 cases and 20000 controls by

$$PE = \frac{1}{N} \sum_{i=1}^{N} |y_i^{test} - \hat{p}_i| \tag{2.14}$$

where $N = 40000$ here, y_i^{test} is the disease status of ith patient and $\hat{p}_i$ is the estimated probability of having the disease by

$$\hat{p}_i = \frac{1}{1 + exp(-(L_i)} \tag{2.15}$$

where L_i is the linear predictor for the ith observation. We will also add the results from the traditional logistic regression as the benchmark for comparison on the prediction performance.

We also compare the variable selection performance among the four models. Since our main interest in the project is to improve the modeling of the interaction effects, we focused on evaluating the capacity of the models for recovering the non-null gene-environment interaction and gene-gene interaction while controlling the false discovery of the non-null effects. In each scenario, we measured the sensitivity which is the proportion of the non-null effects being selected, and the specificity, which is the proportion of the null effects not being selected.

There is also a criteria in variable selection studies to control the total number of parameters being selected. For example, in Lasso studies by changing the penalty parameter λ, the number of the parameters included in the model will change accordingly. Also, in forward/backward stepwise selection, we need to directly specify the total number of parameters we want to include in a model. In Bayesian variable selection study areas, the median model decision rule is a typical approach. It will select all the covariates in the model with $P(I = 1|data) \geq 0.5$.

2.3.1 Study I

We include 6 additive genetic factors (SNP) and 1 environmental factor (EXP) to simulate the independent variables for generating the data. We focused on the study

of case-control qualitative datasets with 1000 cases and 1000 controls in all settings. For simulating the dataset, we fixed the prevalence of the EXP and the minor allele frequency (MAF) of the SNPs and set the effect parameters corresponding to the odds of disease. The MAF of the non-null SNP is fixed as 0.1 and non-null SNP as 0.3. The environmental exposure factor is fixed at 0.1. As shown in Table 1, we are considering 8 different scenarios for the interaction studies. In all scenarios, we assumed the odds for the non-null genetic factor was 1.25 and for the non-null exposure factor was 1.5, which corresponds to the parameter values in Table 1.

The 8 scenarios reflect the conditions that are frequently encountered in practical gene-environment interaction studies. Scenario 1 is the 'Null' model that does not include any significant effect factors. Scenario 2 includes a non-null environmental factor without any genetic main or interaction effects. In scenario 3, the environmental factor is significant and it also has a significant gene environment interaction SNP 6. In scenario 4 and 5, the environmental factor is absent while an effect from SNP 6 is present. And a non-null gene-environment interaction between SNP 6 and the exposure is absent in scenario 4 and present in scenario 5. In scenario 6, there is one genetic main effect and one environmental effect present without any significant interaction effects. In scenario 7, the interaction between SNP 6 and the environmental factor exists as well as a corresponding main effect. In scenario 8, one additional gene environment interaction effect was added to the model compared with scenario 7, including a main effect that is not significant.

Figure 2.2 to Figure 2.11 show the variable selection performance of the four Bayesian models. Compared with the independent model, the other three models tend to select main effects more often than interaction effects. In scenario 1, the three hierarchical models control the probability of selecting the interaction effects to

be under 0.5. In scenario 2, the strong hierarchical model has the largest probability of selecting the non-null environmental factor. In scenario 3, the strong hierarchical model has higher probability of selecting the non-null environmental factor while failing to select the interaction effects. The weak hierarchical model has a similar performance to the independent model, inferior to the strong hierarchical model. In scenario 4, all the models perform similarly in selecting the non-null genetic effect. The effect enforce, strong hierarchical and weak hierarchical model perform very well in controlling the false positive discovery of the null effect for the environmental factor and interaction effects. In scenario 5, we observe a similar phenomenon as scenario 3 for the hierarchical model's failure to detect the interaction effects. In this case, the weak hierarchical model performs better than the strong hierarchical model. In scenario 6, all the three hierarchical models perform better than the independent model on the controlling the false positive discovery of the non-null interactions. In scenario 7, the strong hierarchical model has a larger probability of selecting the non-null environmental factor but did not control well the false discovery of interactions. The effect enforce model and the weak hierarchical model controlled the number of false positive discoveries better for this scenario. In scenario 8, we observed a similar pattern to scenario 7 for detecting the interactions. Scenario 9's setup violates the hierarchical assumption for the interaction effect. In this scenario, we observe that the independent model identifies the non-null interaction effect while the others do not work very well. However, the independent model could not identify well the environmental factor effect. The weak hierarchical model performs similarly as the independent model. In scenario 10, there is one non-null interaction effect and the other factors are null. The independent model outperforms the other models. This is mainly because the truth violates the hierarchical assumption and the interaction effect is non trivial.

Figure 2.12 and Figure 2.13 show the prediction performance of the five models in each scenario. The frequentist logistic regression generates the largest prediction error when compared with the other Bayesian models. In general, the strong hierarchical model has the lowest level of prediction error. Scenario 3, 5 and 8 violate the strong hierarchical assumption. In these 3 scenarios, the strong hierarchical model surprisingly still outperforms the others. In scenario 5, there is one non-null genetic factor and the environmental factor is also non-null, while the interaction effect is null. In this scenario, the effect enforce model, strong hierarchical model and weak hierarchical model outperform the independent model, because these three models favor the main effect. In scenario 9, the independent model outperforms the other models because of the capability of finding out interactions without constraints on main effects.

	SNP						Exposure	GE Interaction						Pattern
	β_1	β_2	β_3	β_4	β_5	β_6	γ_1	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	
Scenario 1	0	0	0	0	0	0	0	0	0	0	0	0	0	null
Scenario 2	0	0	0	0	0	0	.405	0	0	0	0	0	0	exp only
Scenario 3	0	0	0	0	0	0	.405	0	0	0	0	0	.405	exp & int
Scenario 4	0	0	0	0	0	.223	0	0	0	0	0	0	0	SNP only
Scenario 5	0	0	0	0	0	.223	0	0	0	0	0	0	.405	SNP & int
Scenario 6	0	0	0	0	0	.223	.405	0	0	0	0	0	0	SNP & exp
Scenario 7	0	0	0	0	0	.223	.405	0	0	0	0	0	.405	SNP & exp & int
Scenario 8	0	0	0	0	0	.223	.405	0	0	0	0	.405	.405	SNP & exp & int*
Scenario 9	0	0	0	0	0	.223	0	0	0	0	0	.405	0	SNP & int**
Scenario 10	0	0	0	0	0	0	0	0	0	0	0	0	.405	int only

Table 2.1: Parameter setups for Study I

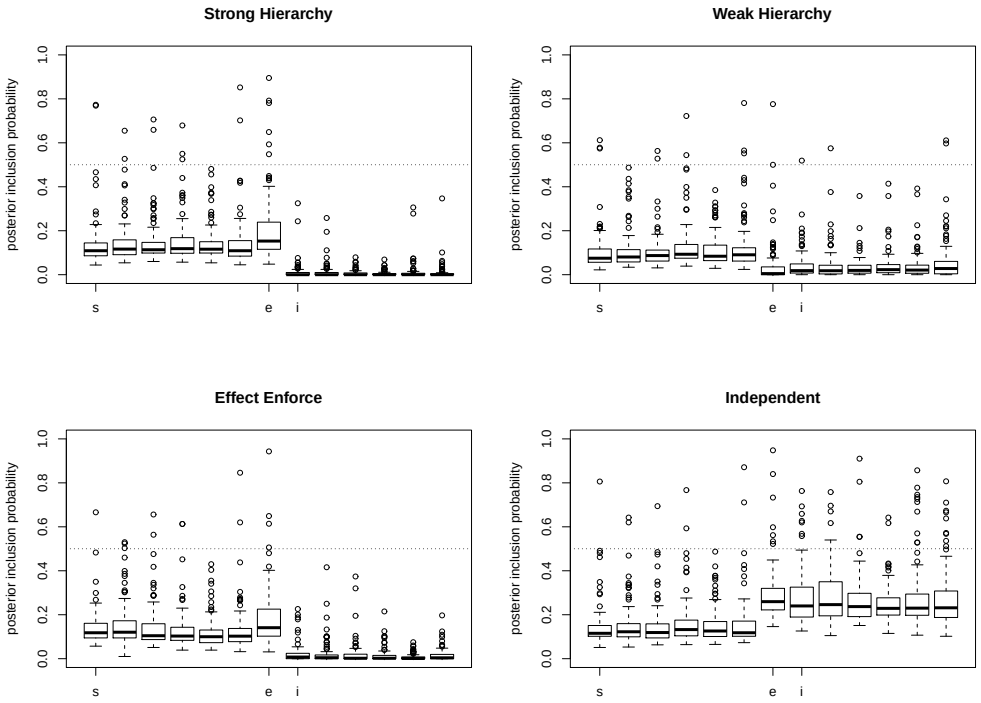

Figure 2.2.: Scenario 1 in Study I. The boxplots of each factor being selected in 100 replicates. 's' refers to the 6 SNPs. 'e' refers to EXP and 'i' refers to 6 GxE interaction effects

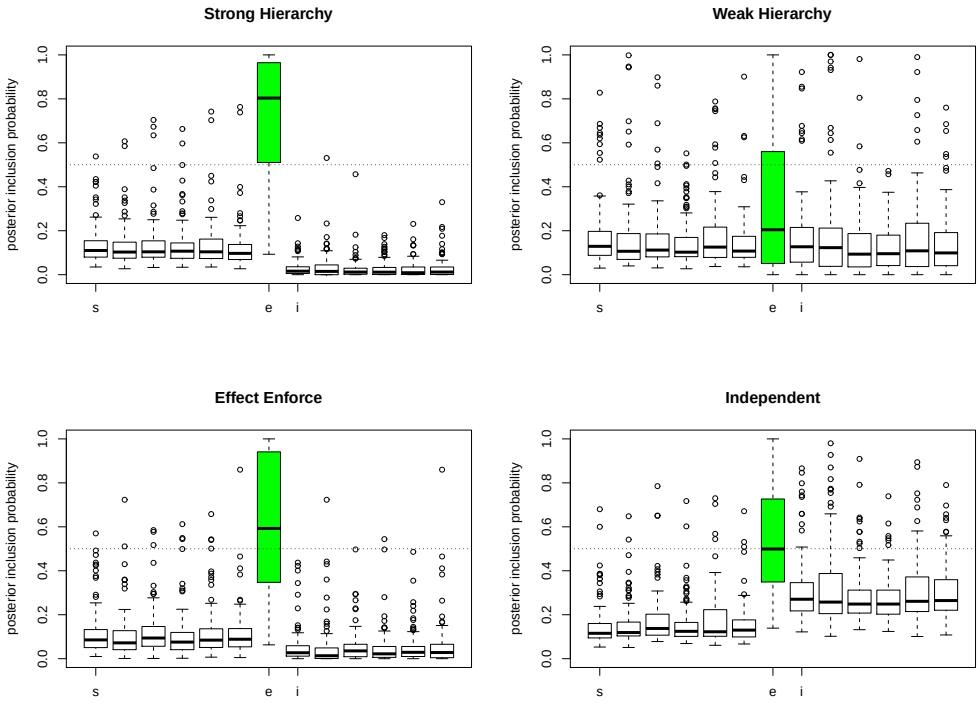

Figure 2.3.: Scenario 2 in Study I. The boxplots of each factor being selected in 100 replicates. 's' refers to the 6 SNPs. 'e' refers to EXP and 'i' refers to 6 GxE interaction effects

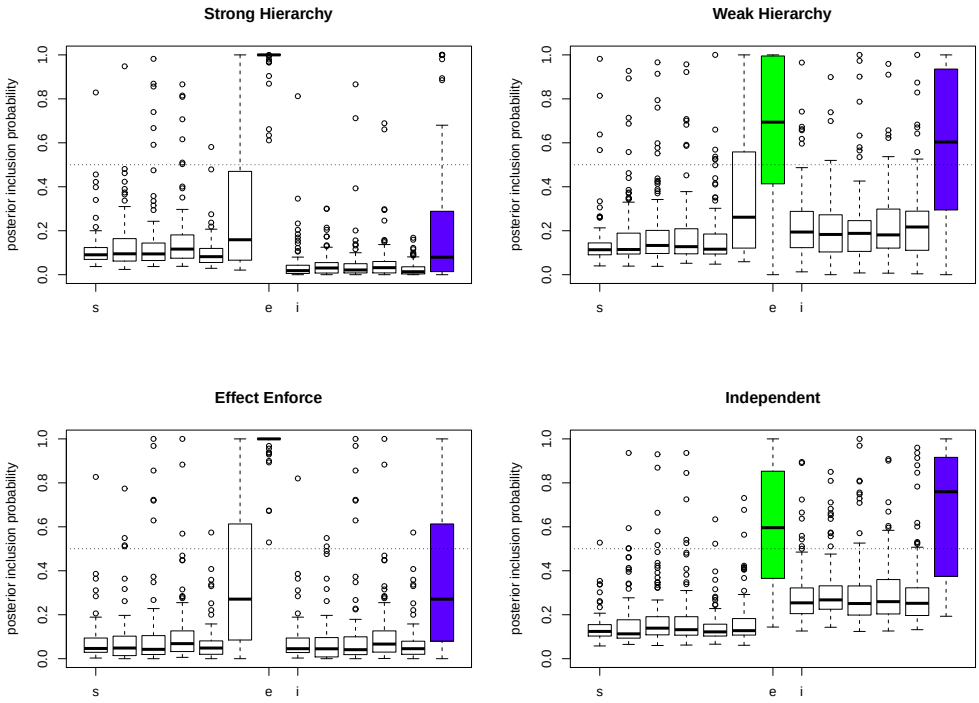

Figure 2.4.: Scenario 3 in Study I. The boxplots of each factor being selected in 100 replicates. 's' refers to the 6 SNPs. 'e' refers to EXP and 'i' refers to 6 GxE interaction effects

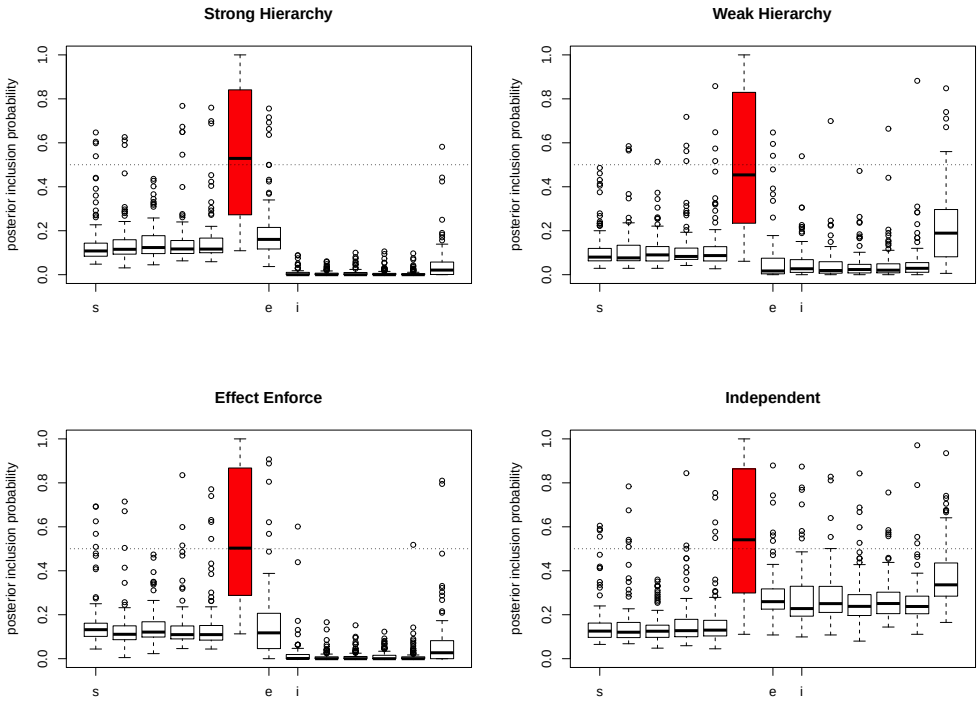

Figure 2.5.: Scenario 4 in Study I. The boxplots of each factor being selected in 100 replicates. 's' refers to the 6 SNPs. 'e' refers to EXP and 'i' refers to 6 GxE interaction effects

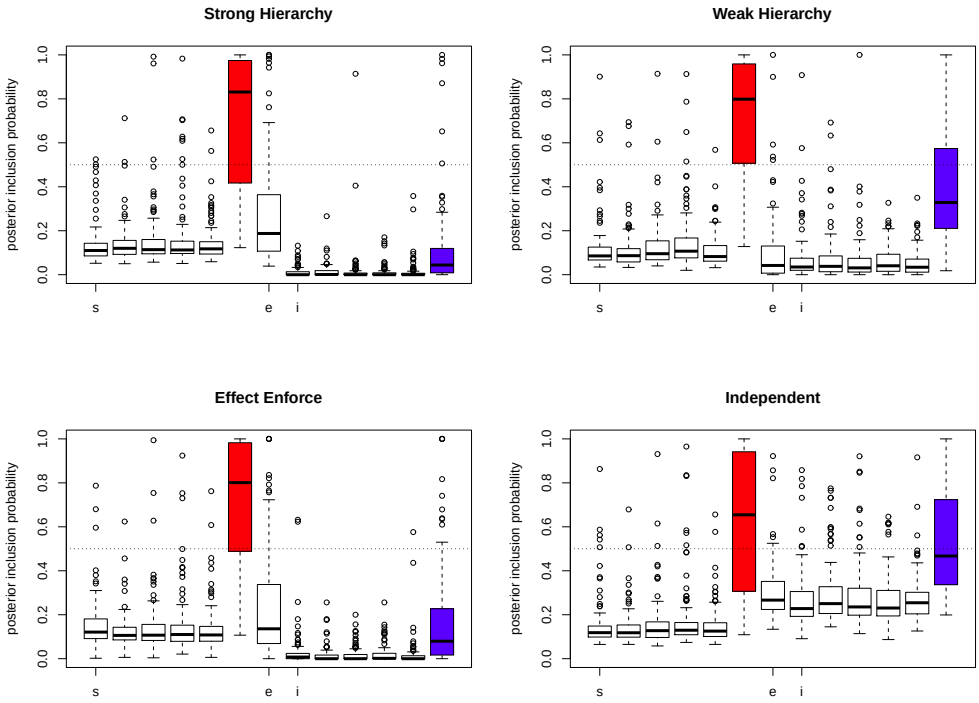

Figure 2.6.: Scenario 5 in Study I. The boxplots of each factor being selected in 100 replicates. 's' refers to the 6 SNPs. 'e' refers to EXP and 'i' refers to 6 GxE interaction effects

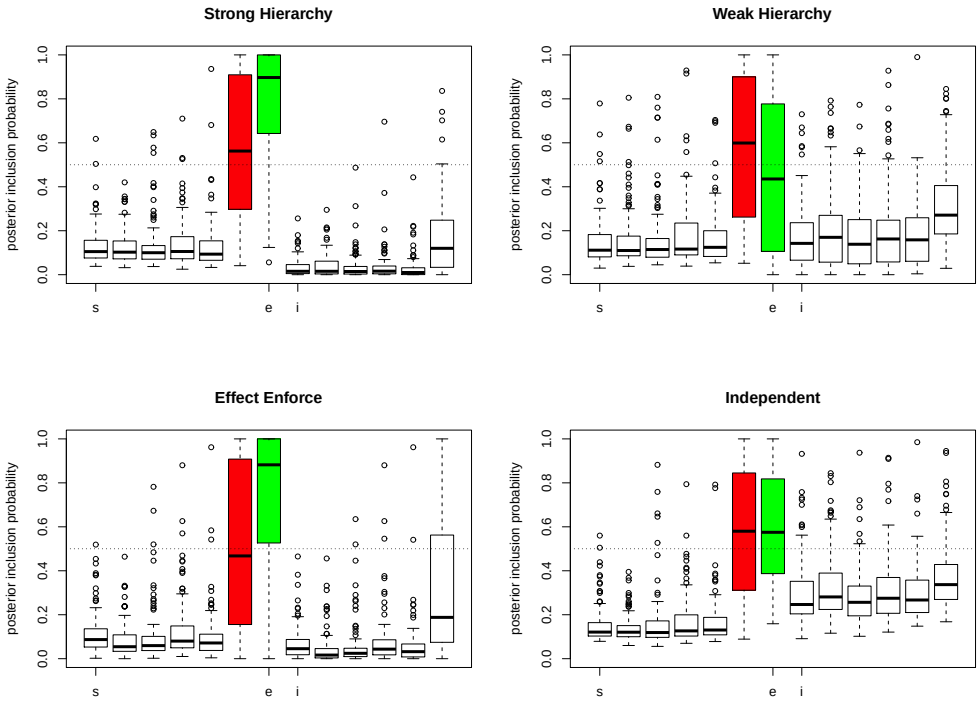

Figure 2.7.: Scenario 6 in Study I. The boxplots of each factor being selected in 100 replicates. 's' refers to the 6 SNPs. 'e' refers to EXP and 'i' refers to 6 GxE interaction effects

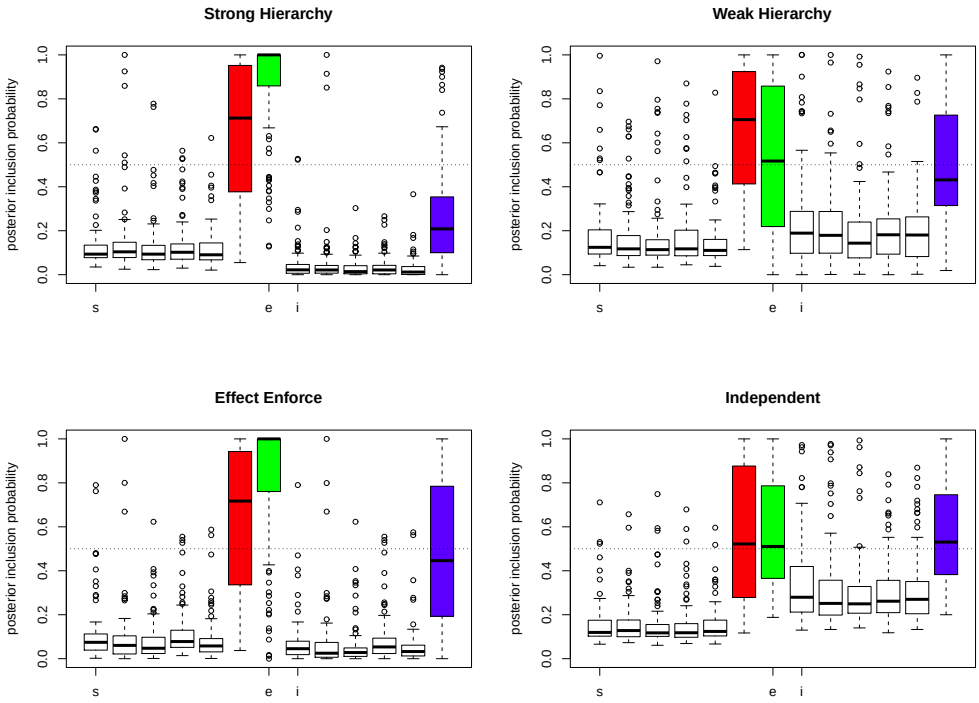

Figure 2.8.: Scenario 7 in Study I. The boxplots of each factor being selected in 100 replicates. 's' refers to the 6 SNPs. 'e' refers to EXP and 'i' refers to 6 GxE interaction effects

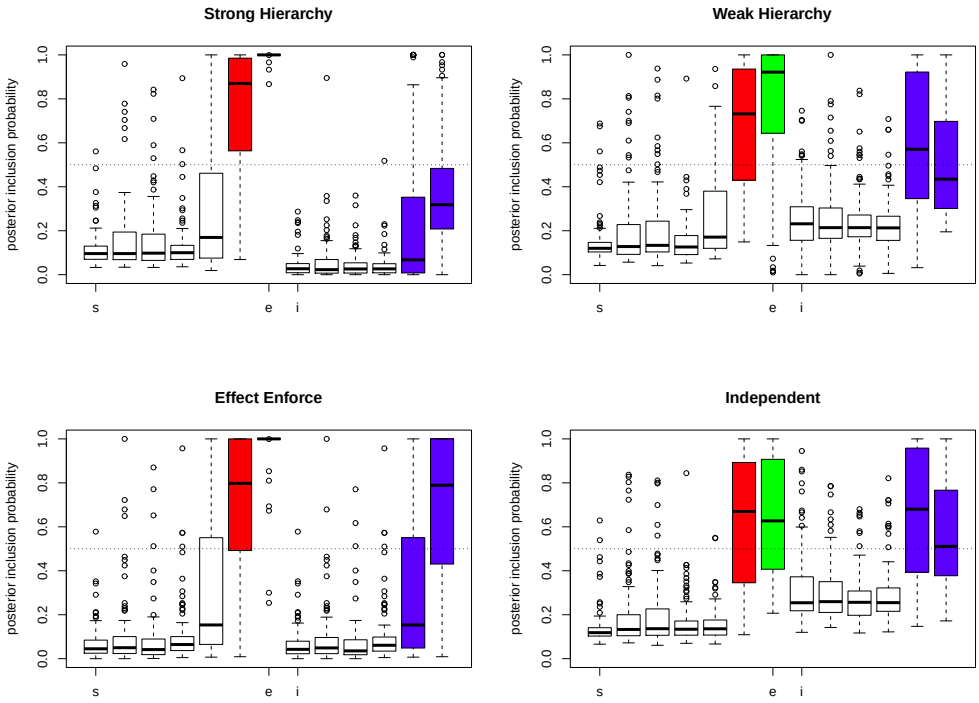

Figure 2.9.: Scenario 8 in Study I. The boxplots of each factor being selected in 100 replicates. 's' refers to the 6 SNPs. 'e' refers to EXP and 'i' refers to 6 GxE interaction effects

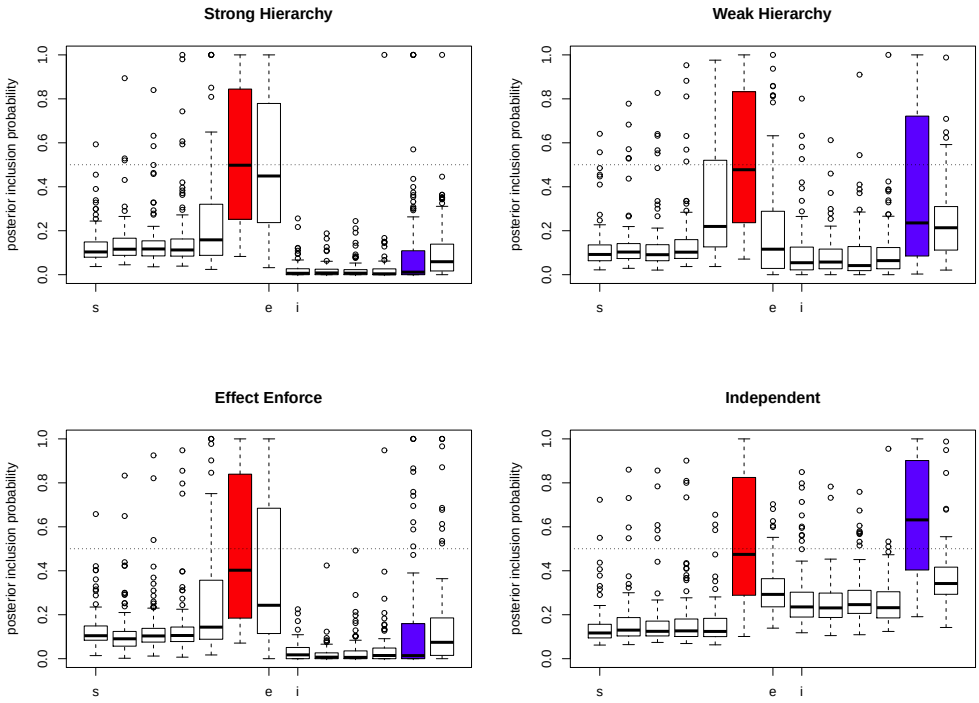

Figure 2.10.: Scenario 9 in Study I. The boxplots of each factor being selected in 100 replicates. 's' refers to the 6 SNPs. 'e' refers to EXP and 'i' refers to 6 GxE interaction effects

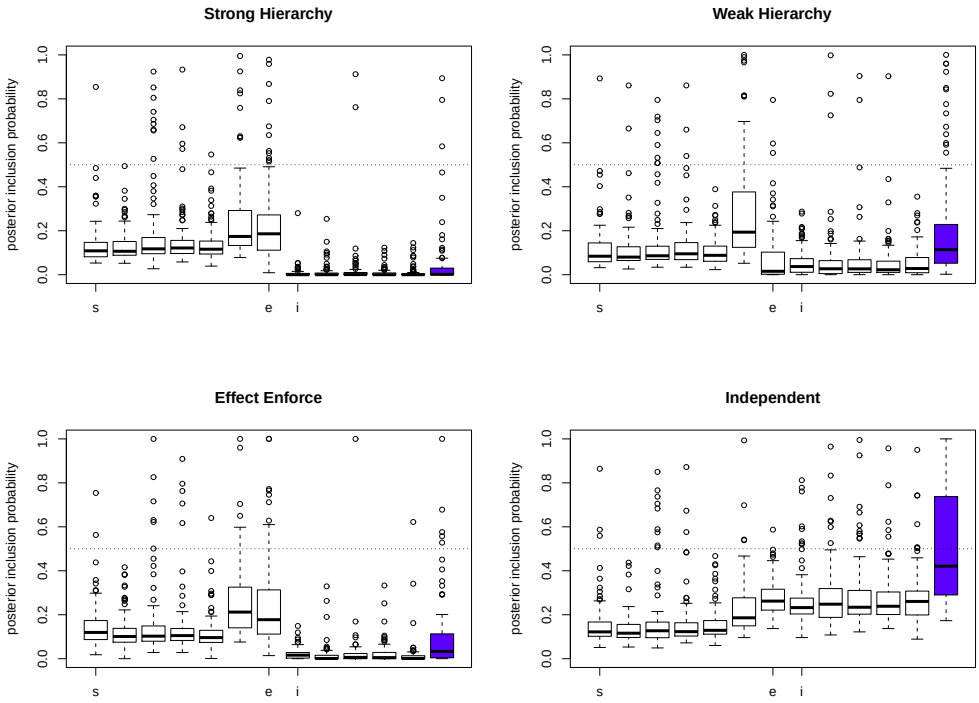

Figure 2.11.: Scenario 10 in Study I. The boxplots of each factor being selected in 100 replicates. 's' refers to the 6 SNPs. 'e' refers to EXP and 'i' refers to 6 GxE interaction effects

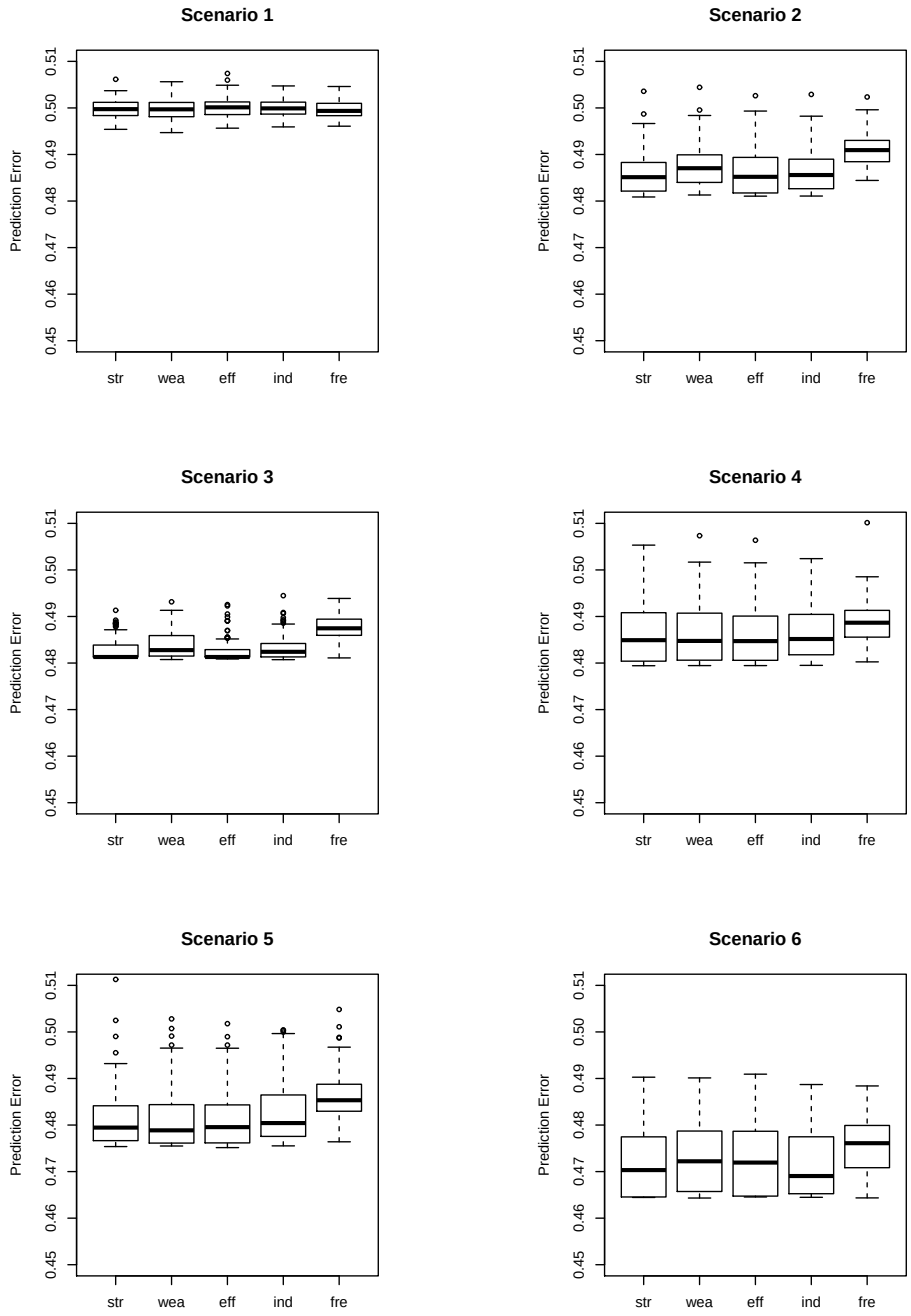

Figure 2.12.: Prediction performance for each model in each scenario of study I. 'ind':independent, 'enf': effect enforce, 'str': strong hierarchical, 'wea': weak hierarchical, 'fre' frequentist logistic regression.

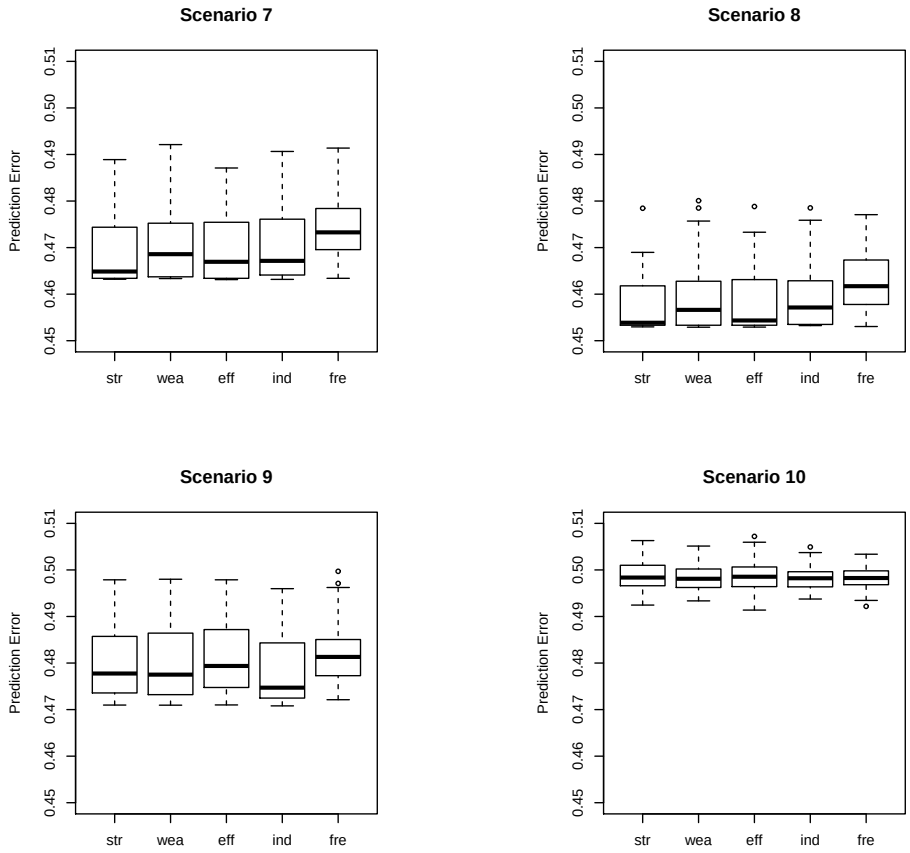

Figure 2.13.: Prediction performance for each model in each scenario of study I. 'ind':independent, 'enf': effect enforce, 'str': strong hierarchical, 'wea': weak hierarchical, 'fre' frequentist logistic regression.

	SNP			Exposure	GE Interaction			
	β_{48}	β_{49}	β_{50}	γ_1	θ_{47}	θ_{48}	θ_{49}	θ_{50}
Scenario 1	.223	.223	.223	.405	0	0	0	0
Scenario 2	.223	.223	.223	.405	0	0	0	.405
Scenario 3	.223	.223	.223	.405	0	.223	.223	0
Scenario 4	.223	.223	.223	.405	0	.223	.223	.405
Scenario 5	.223	.223	.223	.405	.223	0	0	0

Table 2.2: Parameter setups for Study II

2.3.2 Study II

We simulate 50 SNPs with only the additive effects and 1 EXP to simulate the data. Here, we only consider the genetic and environmental main effects as well as gene environment interaction effects. This study imitate the real practice in which we wish to find out the predisposing SNPs and also the positive interactions among the SNP and environmental factor. In this type of study, the environmental factor is usually already confirmed as an non-null factor, so here in all scenarios we simulate the environmental factor having non-null effects. The MAF for SNP and frequency for EXP are set the same as in Study I.

In all scenarios, we assume 3 non-null genetic effects among 50 SNPs and one non-null effect. In scenario 1, there is no gene-environment interaction. In scenario 2, there is one non-null larger interaction effect. In scenario 3, two relatively smaller interaction effects exist. In scenario 4, we assume there are non-null interaction effects with with different values. In scenario 5, there is one small interaction effect which corresponds to the null SNP. Here we still want to examine the prediction and variable selection performance of the proposed models. Due to the same setting for the

main effects, here we will only show the variable selection performance on selecting the interaction effects. The results also have shown the results on the interaction part are very similar to each other. The results are also based on 50 replicates in each scenario. Since we assume that the environmental factors are non-null in the model, the weak hierarchical model is the same as the independent model. The existence of the non-null interaction effect will not depend on the corresponding genetic factors.

So as shown in Figure 2.14, the strong hierarchical model does best among the four models on the prediction performance. This is because the strong hierarchical model favors the main effects more than the interaction effects. In Figure 2.15, in all scenarios, the strong hierarchical model has a superior result for controlling the false positive. Although the power of detecting the interaction is limited by the feature of model, the capacity of controlling false positive make up for its inferior ability to detect non-null effects. In scenarios 2, 3 and 4, the effect enforce model outperforms the other models in selecting the true positive interaction effects, while controlling the false positive comparably to the strong hierarchical model. This is because the model partially or fully matched the assumption in the truth of the simulated data. In this scenario, the non-null interaction is present without the corresponding genetic main effect. Then we observe that the effect enforce model performs poorly as the strong hierarchical model on detecting the non-null interaction.

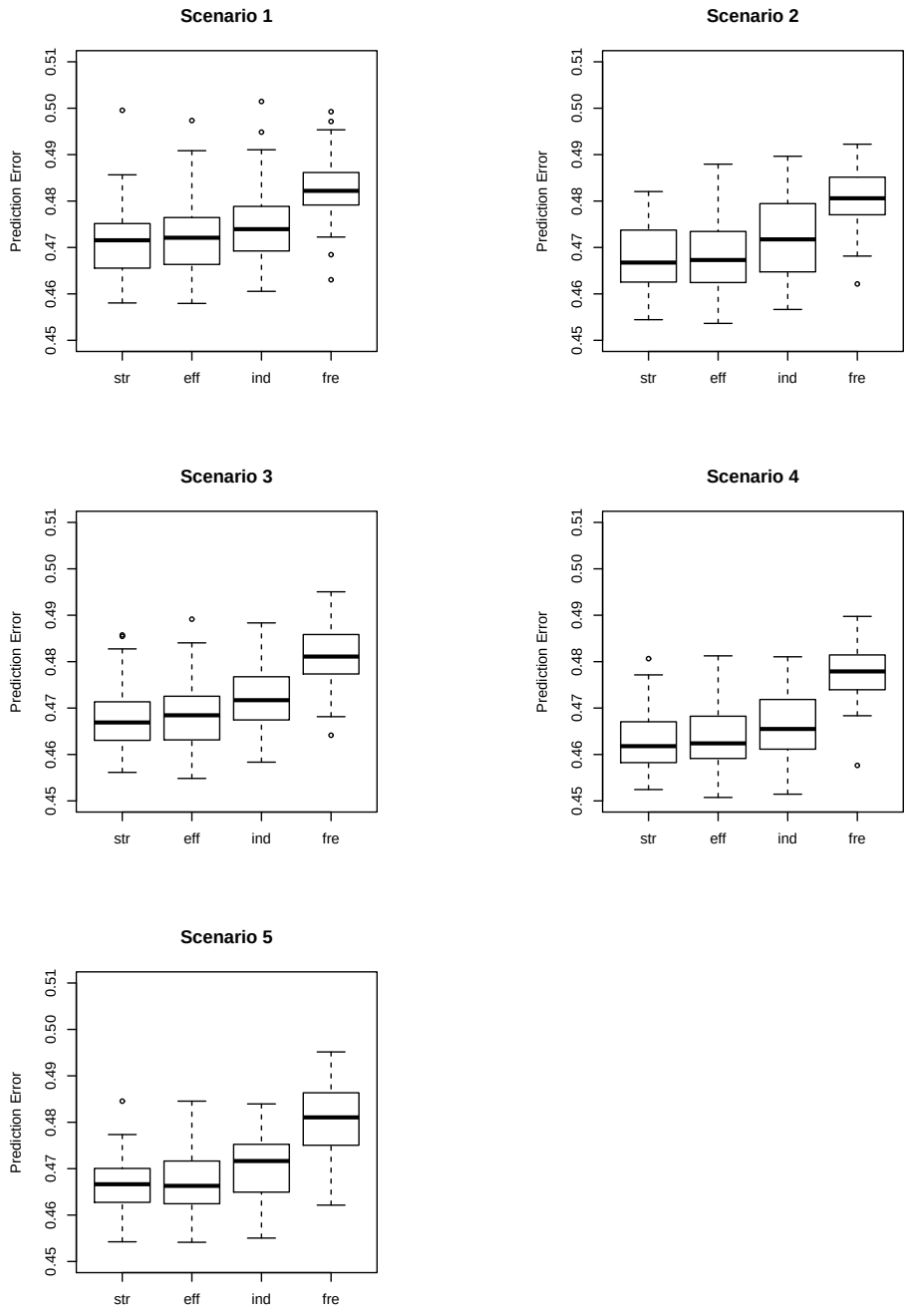

Figure 2.14.: Prediction performance for each model in each scenario of study II. 'ind':independent, 'enf': effect enforce, 'str': strong hierarchical, 'wea': weak hierarchical, 'fre' frequentist logistic regression.

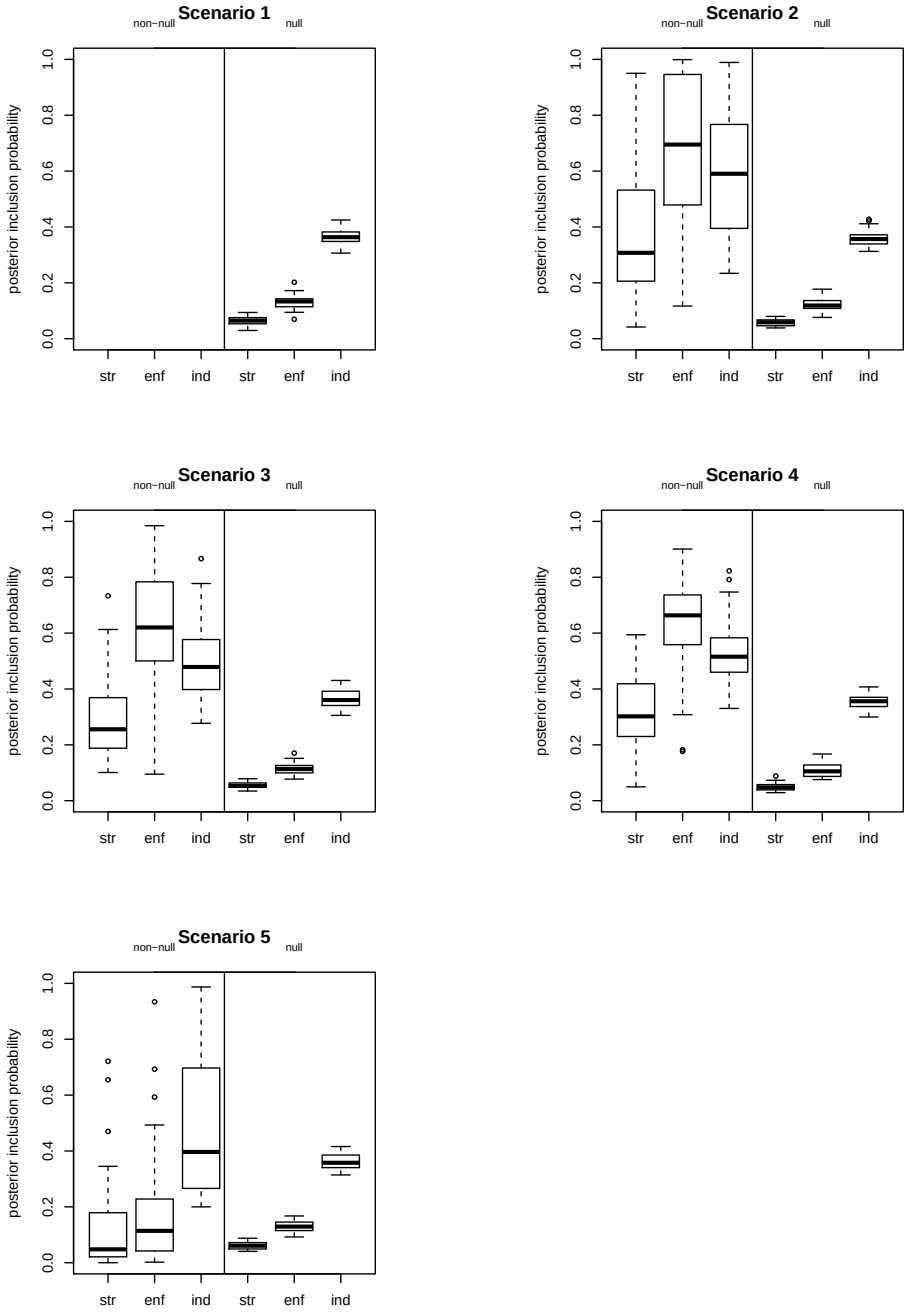

Figure 2.15.: Variable selection performance for each model in each scenario of study II. 'ind':independent, 'enf': effect enforce, 'str': strong hierarchical, 'wea': weak hierarchical, 'fre' frequentist logistic regression.

	SNP		Exposure	GE Interaction	GG Interaction		
	β_5	β_6	γ_1	θ_6	τ_1	τ_{14}	τ_{15}
Scenario1	.223	.223	.405	.405	0	0	0
Scenario2	.223	.223	.405	.405	0	0	.223
Scenario3	.223	.223	.405	.405	0	.223	0
Scenario4	.223	.223	.405	.405	.223	0	0

Table 2.3: Parameter setups for Study III

2.3.3 Study III

In Study III, we further incorporate the gene-gene interaction effects into the complete model based on Study I. In total we have 6 genetic main effects, 1 environmental factor, 6 gene-environment interactions and 15 pairwise gene-gene interactions. In this study, we consider 4 different scenarios by changing the gene-gene interaction parameter values while fixing the other effects. The true parameter values are presented as in Table 2.3.

In scenario 1, the strong hierarchical model assigned all the null GxG interaction effects to have a lower posterior probability and yielded a better selection of the non-null main effects than the other models. The independent model performed best at identifying the GxE interaction. In scenario 2, the GxG interaction between SNP 5 and SNP 6 is set as non-null. The independent model performs well on identifying the GxG interaction effect. However, it performed worse on finding the non-null genetic effects for SNP 5 and SNP 6. In scenario 3, which partially violates the hierarchical structure, shows the similar phenomenon as scenario 2. Compared with the independent model, the strong hierarchical model is good at identifying the main effect. The weak hierarchical model performs similarly as the independent model and

the effect enforce model similarly as the strong hierarchical model. In scenario 4, the non-null GxG interaction corresponds to null genetic main effects at SNP 1 and SNP 2. So here we observed that the independent model could identify the significant interaction while the other models failed. This is because that the independent model dose not put any constraint on the hierarchical structure. Also, due to the hierarchical structure, the strong hierarchical, weak hierarchical and effect enforce model tend to have a larger probability of including the corresponding SNPs. When we compared the prediction performances in these four scenarios, the strong hierarchical mode performed best in overall. In scenario 4, the independent model perform similarly to the strong hierarchical model. The strong hierarchical model detected the non-null main effects to make up for its inferior ability to detect the interactions and the tendencies to include the null main effects.

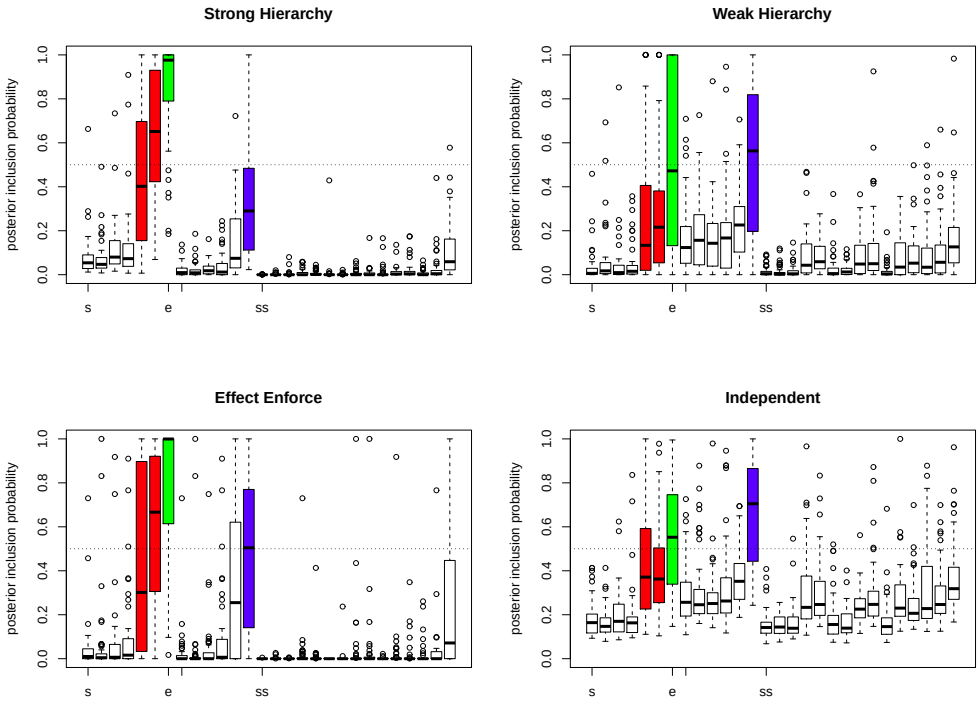

Figure 2.16.: Scenario 1 in Study III. The boxplots of each factor being selected in 100 replicates. 's' refers to the 6 SNPs. 'e' refers to EXP and 'ss' refers to 15 GG interaction effects

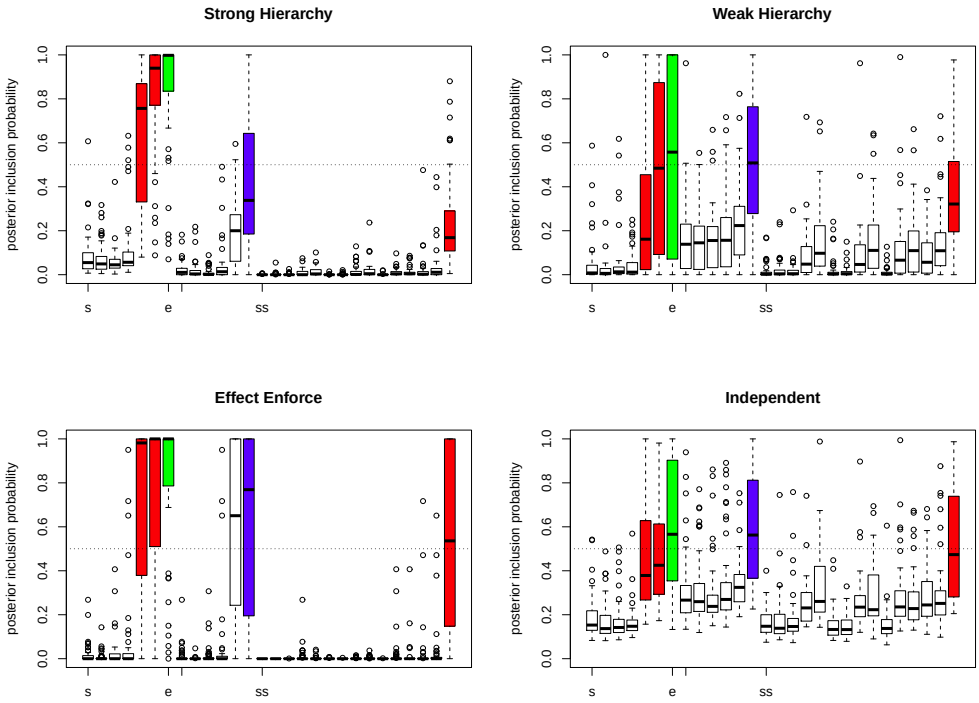

Figure 2.17.: Scenario 2 in Study III. The boxplots of each factor being selected in 100 replicates. 's' refers to the 6 SNPs. 'e' refers to EXP and 'ss' refers to 15 GG interaction effects

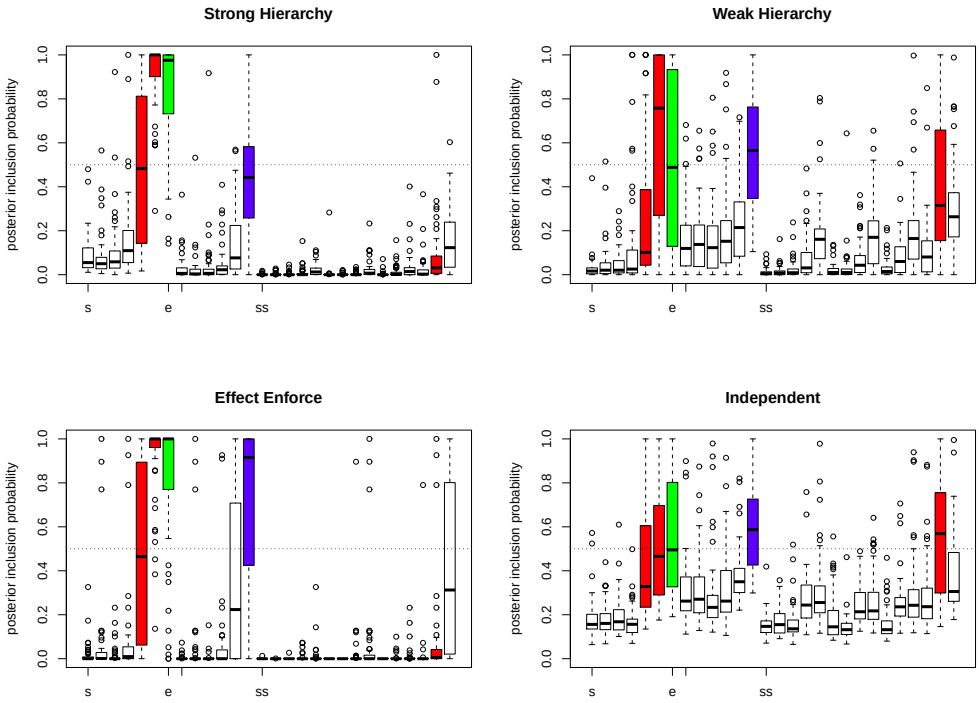

Figure 2.18.: Scenario 3 in Study III. The boxplots of each factor being selected in 100 replicates. 's' refers to the 6 SNPs. 'e' refers to EXP and 'ss' refers to 15 GG interaction effects

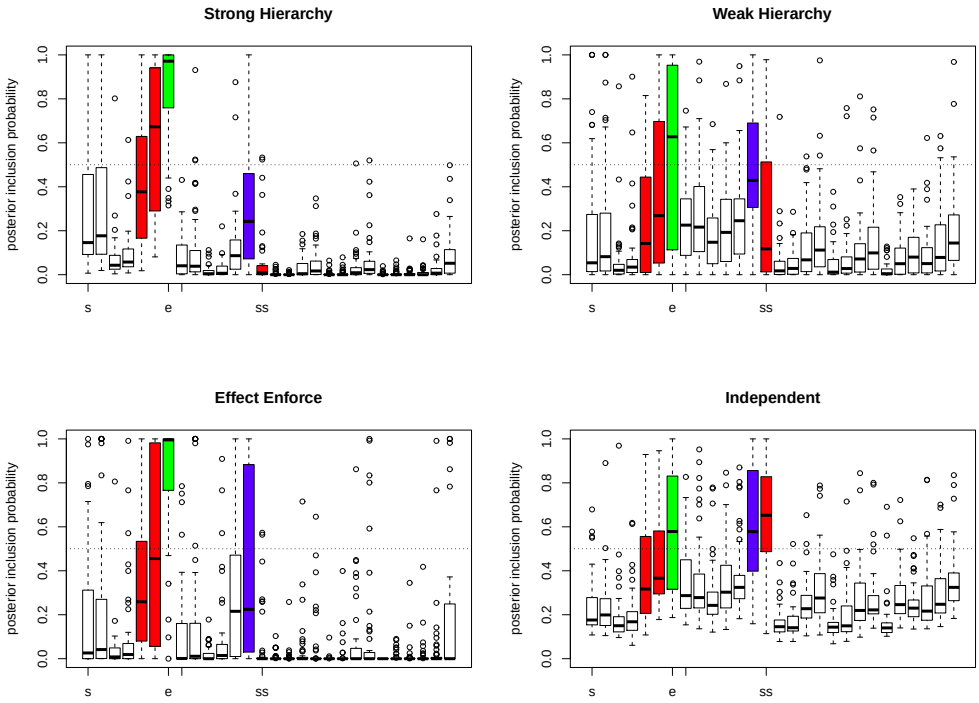

Figure 2.19.: Scenario 4 in Study III. The boxplots of each factor being selected in 100 replicates. 's' refers to the 6 SNPs. 'e' refers to EXP and 'ss' refers to 15 GG interaction effects

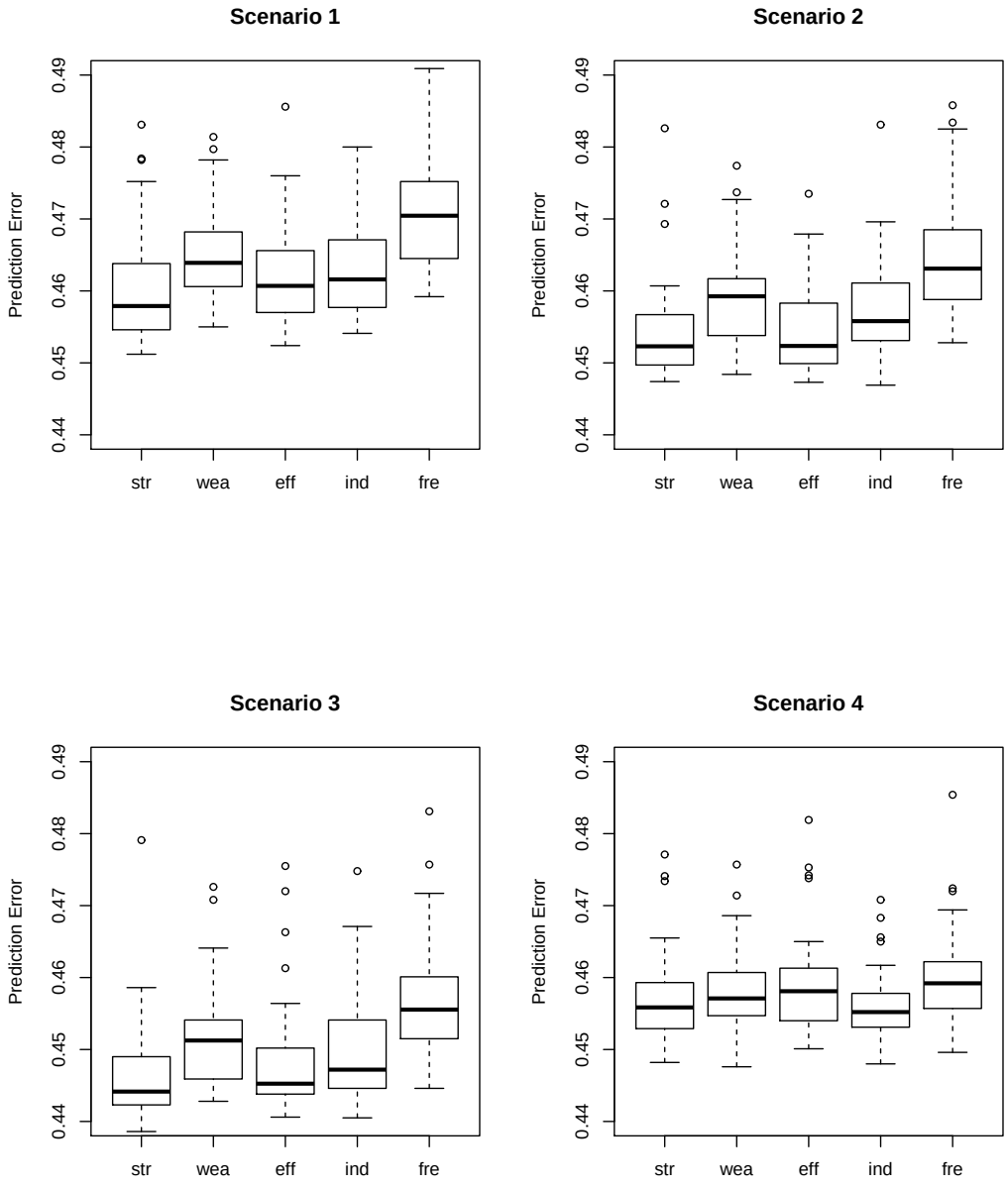

Figure 2.20.: Prediction performance for each model in each scenario of study III. 'ind':independent, 'enf': effect enforce, 'str': strong hierarchical, 'wea': weak hierarchical, 'fre' frequentist logistic regression.

2.4 Real Data Examples

2.4.1 Lung Cancer

We applied our proposed Bayesian methods and the independent model to the data from the International Lung Cancer Consortium. The data include 17 different studies in 13 countries. For illustration of the model, we focused on 6 SNPs in 3 regions: $rs2736100$ and $rs402710$ at 5p15, $rs2256543$ and $rs4324798$ at 6p21, and $rs16969968$ and $rs8034191$ at 15q25. The data we applied include 8867 participants with complete data. Among all the participants, there were 5,217 controls and 3,650 cases. In the case group, there were 2,434 males and 1,216 females and 3,378 were smokers. In the control group, there were 3,642 males and 1,575 females and 3,703 were smokers. We include sex as a covariate and the smoking indicator as the environmental factor. In the model, we wanted to detect the genetic and environmental main effects and the gene-gene interaction and gene-environment interactions simultaneously.

Figure 2.21 shows the result for running the three different models on the data set. Since the six SNPs have been tested to cause the cancer in prior studies [48], we include them with the prior of each SNP being significant with probability 0.9. Also, for the gene-environment and gene-gene interactions we assume the prior for the effects being significant with probability 0.5. There were 6 SNPs, 1 EXP, 6 gene-environment interaction and 15 pairwise gene-gene interaction. In the graph, we found that there were no gene-gene interactions that can be regarded as significant. In the main effects part, all the models identified the smoking factor with probability 1 and the independent models identified all the SNPs with probability larger than 0.5. The strong hierarchical model identify $rs402710$ as the most significant factor with the other two non-null SNPs. The two SNPs at 6p21 are substantially non-

significant compared with the SNPs found at the other regions. The three models also do not identify obvious gene environment interactions. The weak hierarchical model only identified the interaction between $rs16969968$ and the smoking. Also, the hierarchical model lower the probability of interaction between $rs2256543$ and smoking. These two interaction worth mentioning.

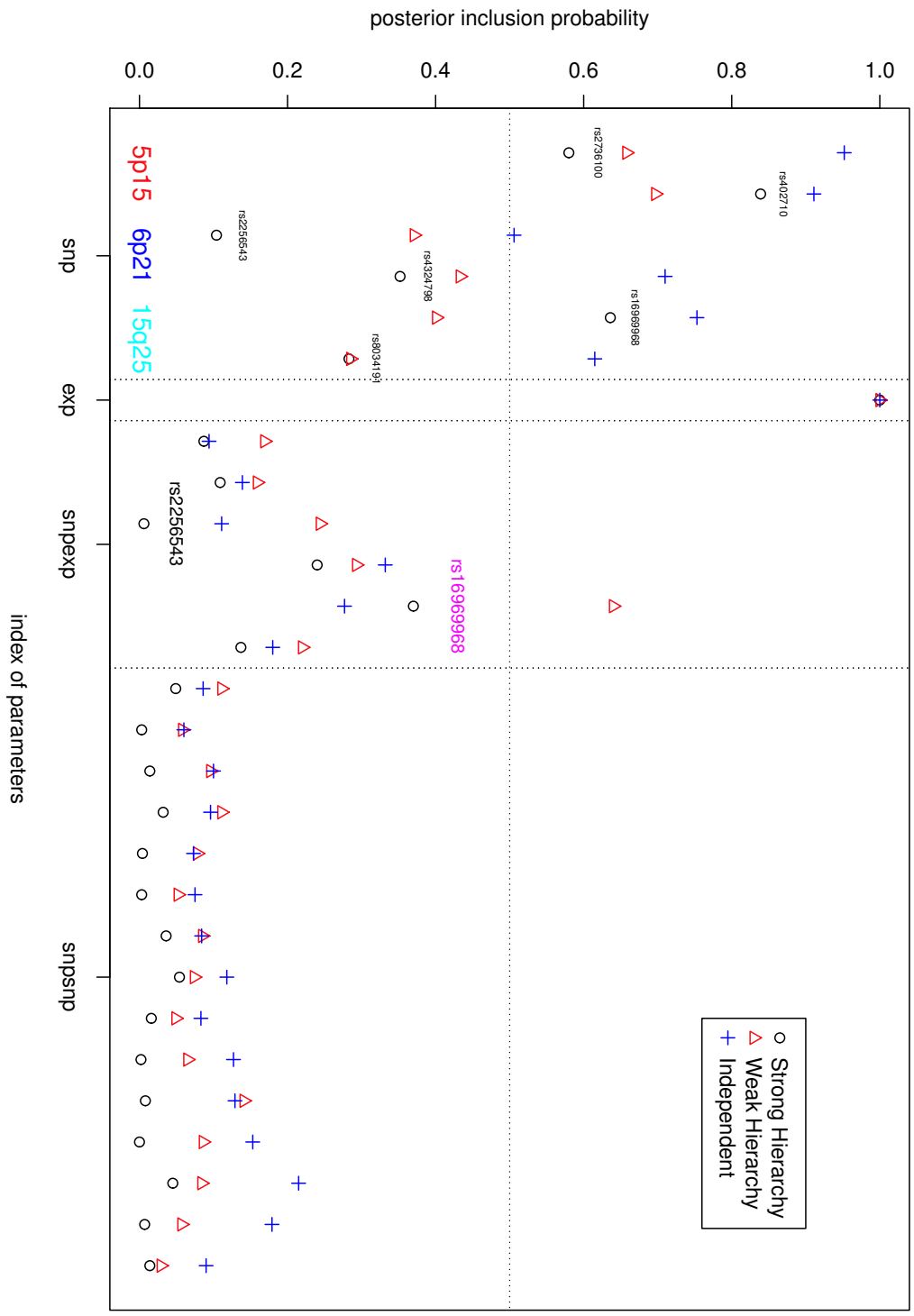

Figure 2.21.: Real data results for the lung cancer study.

2.4.2 Cutaneous Melanoma

We also applied our models in the cutaneous melanoma studies. We first selected 24 SNPs that have been found previously in GWAS. It is also well known that eye color correlates with certain genetic factors. So in this study, we want to find out the interaction between genetic factors and eye color that can relate to the occurrence of cutaneous melanoma. After removing the data with missing values, the data were composed with 929 cases and 1,024 controls. There are five different colors for eyes as in Table 2.4 . We coded the color factor into 2 dummy variables which represent the comparisons between 3 level of eye colors (blue/grey, brown and green/hazel). Green and heel colors were combined because they have similar hues and similarly blue and grey were combined because grey eyes are are and of similar grade in hue to blue. We included 24 SNPs as the candidate genetic factors: $rs1015362$, $rs1042602$, $rs10757257$, $rs10830253$, $rs12896399$, $rs12913832$, $rs1335510$, $rs1393350$, $rs1408799$, $rs16891982$, $rs17305573$, $rs1805007$, $rs1806319$, $rs1847142$, $rs1885120$, $rs2218220$, $rs2284063$, $rs28777$, $rs4911414$, $rs4911442$, $rs6001027$, $rs7023329$, $rs910873$ and $rs935053$.

Figure 2.22 shows the variable selection results by the three Bayesian models. Totally we are considering 24 genetic factors, 2 environmental factors and 48 gene environment interactions. The interaction between $rs12913832$ and the eye color has been found significant in the study of Amos et al. (2011) [49]. The hierarchical model also finds the interaction between $rs17305573$ and the eye color. For the other interaction terms, the posterior probability does not exceed 50%, indicating no evidence of an interactions. The Independent model also tends to generate higher probability for the interactions. The hierarchical model successfully controlled the probability of false positive discovery that enable us to focus on the significant interactions.

Eye color	Case	Control	Total
Blue	396	331	727
Grey	12	18	30
Brown	189	312	501
Green	150	164	314
Hazel	182	199	381
Total	929	1,024	1,953

Table 2.4: The distribution of study samples with different eye colors in the case-control study

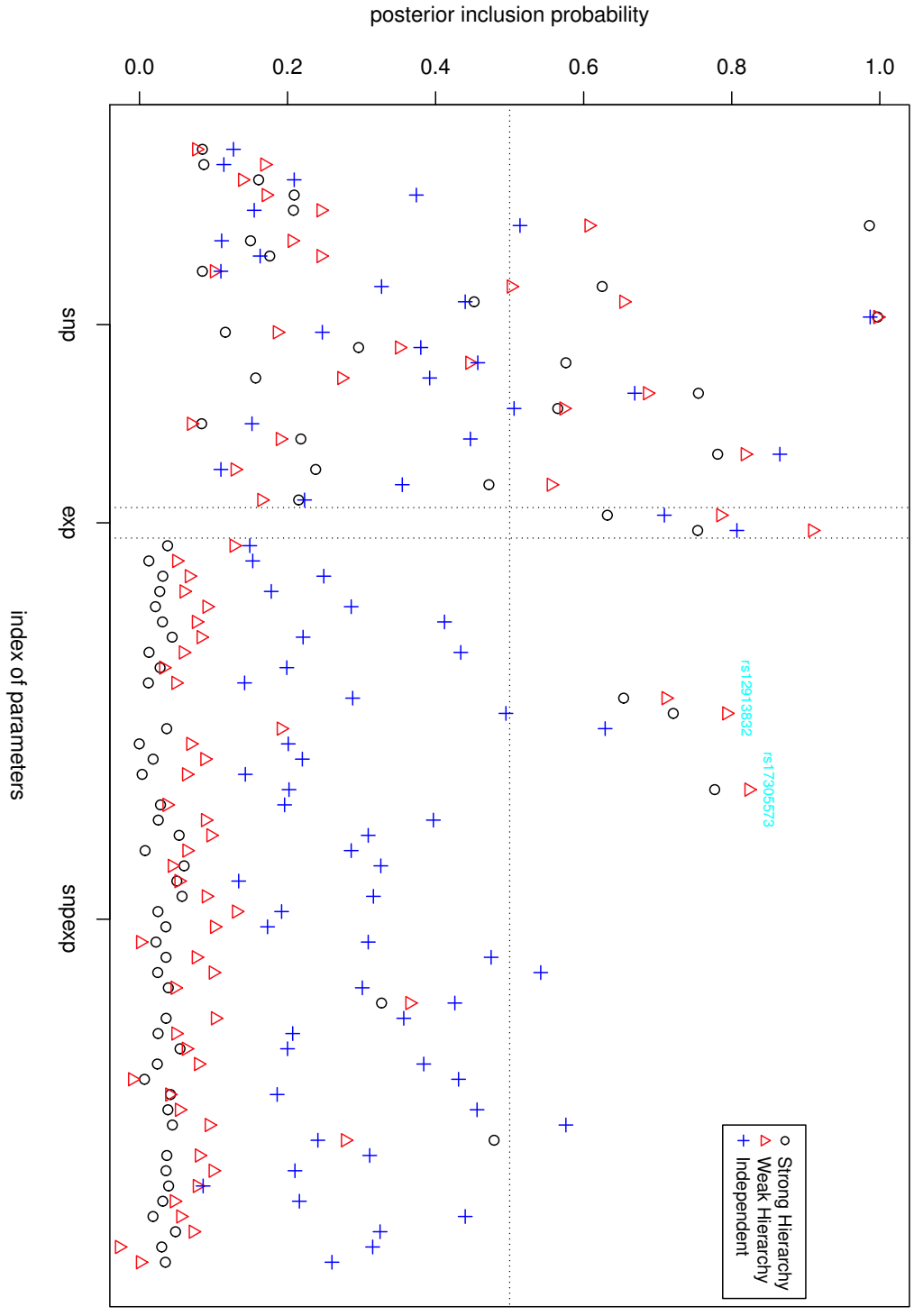

Figure 2.22.: Real data results for the cutaneous melanoma study. 2 EXP variables: gray/blue vs. brown and hazel/green vs. brown,

2.5 Discussion

In this project, we introduced the Bayesian mixture model for modeling the gene-gene and gene-environment interactions simultaneously. Compared with the traditional logistic regression model, the Bayesian model showed good performance in simulation studies on parameter estimation and variable selection. Traditionally, there is no constraint on the modeling of the main effects and higher order interactions in the same model. When several genetic and environmental factors are jointly modeled, there will be a possibly huge number of potential interactions, which will make model fitting and variable selection difficult. Although the hierarchical approaches we studied have been developed in the statistical machine learning literatures, there is a clear need for extending the model application into the genetic studies. Several articles Cheongeun (2003), Wakefield (2010) [46] [50] mentioned applying a hierarchical constraint. Our research provided a systematic approach to setting up the priors among the variables to impose a hierarchical constraint.

Here we only considered case-control analysis, but the model is readily applied to continuous outcome studies. Due to the computational burden for the Bayesian analysis, we limit the total number of genetic and environmental factors in the model. So the proposed model could be recognized as the second step analysis after the GWAS studies, from which we can filter the interesting genetic factors for further explorations. Another important aspect for the Bayesian modeling is the specification of the priors. In our approach, we provided fixed values for the two variance components for the prior distributions of the main effects and interactions. For the factor indicators, we assigned non-informative prior with 0.5 probability for each of the indicators. As shown in the lung cancer real data analysis, when we already know that some environmental factors are directly related with the outcome, then we can specify a very high prior probability for that indicator. In practice, without further

information we may set the prior probability as the proportion of the factors that we have prior belief will be associated with disease risk.

2.6 Appendix: Prior Elicitations by Bayesian Model Averaging

Assumption and Notation

In the previous sections, we introduced the Bayesian mixture methods. The priors for the indicators were fixed a priori. In practice, we may have prior information on the probability for each factor being non-null among all the considered factors. In this Appendix, we tried to apply a Bayesian model averaging approach to elicit the priors when evaluating gene and environment main effects and interactions in the epidemiological case-control studies. This approach allows some uncertainties on the specification of the priors, which will result in a more robust model.

We denote $x_i (i = 1, 2, ..., N)$ as the disease status (1 positive, 0 negative) for the ith patient. Further, we denote $y_i^s (s = 1, 2, ..., S)$ as the minor allele count (0,1,2) for SNP s of patient i; $z_i^e (e = 1, 2, ..., E)$ as the environmental exposure e of patient i (1 exposed, 0 non-exposed).

Basic Modeling

We use a regression model to include both main and interaction effects for the ith patient by

$$logit(p_i) = \alpha + BY_i + \Gamma Z_i + \Theta(Y_i \otimes Z_i)$$

where p_i indicates the probability of disease, α is the general intercept, $Y_i = [y_i^1, ..., y_i^S]^T$, $Z_i = [z_i^1, ..., z_i^E]^T$, and $B = [\beta_1, ..., \beta_S]$, $\Gamma = [\gamma_1, ..., \gamma_E]$, $\Theta = [\theta_{11}, \theta_{12}, ..., \theta_{SE}]$. $\otimes$ is a Kronector operator. Also, we denote β_s, γ_e as the main effects of gene and envi-

ronment and θ_{se} as the interaction effect. Note that here we only consider the one by one ($\mathbf{G} \times \mathbf{E}$) interactions, whereas consideration of higher order interactions(e.g. $\mathbf{G} \times \mathbf{G} \times \mathbf{E}$) is possible by the extension of our models.

Hierarchical Mixture

We put the posterior inference of parameters into two stages:

1st stage. Under the null hypothesis of no effects, we impose a prior distribution for each parameter, where small variances $\sigma^2_{s_\epsilon}, \sigma^2_{e_\epsilon}, \sigma^2_{se_\epsilon}$ imply a condensed mass distribution centred at 0:

$$\beta_s \sim N(0, \sigma^2_{s_\epsilon})$$

$$\gamma_e \sim N(0, \sigma^2_{e_\epsilon})$$

$$\theta_{se} \sim N(0, \sigma^2_{se_\epsilon})$$

Under the alternative hypothesis of effects existing, similar priors with larger variances (AKA: widespread distributed)are given as:

$$\beta_s \sim N(0, \sigma^2_s)$$

$$\gamma_e \sim N(0, \sigma^2_e)$$

$$\theta_{se} \sim N(0, \sigma^2_{se})$$

In Conti (2003) [51] and Wakefield (2010) [46], a mixture model indicator approach is introduced for the modeling of testing of effects in their models. Similarly the priors with indicators can be written as:

$$\beta_s \sim N(0, I^S_s \sigma^2_s + (1 - I^S_s)\sigma^2_{s_\epsilon})$$

$$\gamma_e \sim N(0, I^E_e \sigma^2_e + (1 - I^E_e)\sigma^2_{e_\epsilon})$$

$$\theta_{se} \sim N(0, I^{SE}_{se} \sigma^2_{se} + (1 - I^{SE}_{se})\sigma^2_{se_\epsilon})$$

where, if $I_s^S, I_e^E, I_{se}^{SE} = 1$, we will reject the null hypothesis and confess the existence of effects and *Vice Versa*. We propose that if at least one of $I_s^S = 1$ and $I_e^E = 1$ suffices, there could exist $I_{se}^{SE} = 1$. This reflects our belief that higher order effects(interactions) exist if at least one of the one lower order effects exists. Then, we can simplify the interaction effect parameters as:

$$\theta_{se} \sim N(0, [1 - (1 - I_s^S)(1 - I_e^E)]\sigma_{se}^2 + (1 - I_s^S)(1 - I_e^E)\sigma_{se_\epsilon}^2)$$

This corresponds to

$$P(I_{se}^{SE} = 1 | I_s^S, I_e^E) = \begin{cases} p_{00} = 0 & I_s^S = I_e^E = 0 \\ p_{01} = 1 & I_s^S = 0, I_e^E = 1 \\ p_{10} = 1 & I_s^S = 1, I_e^E = 0 \\ p_{11} = 1 & I_s^S = I_e^E = 1 \end{cases}$$

As noted in Chipman (1996) [13], the higher-order interaction effects are always not independent from the lower order effects. It satisfies $p_{00} \leq min\{p_{01}, p_{10}\} \leq p_{11}$, As example shown in their model, given $P(I_s^S = 1) = 0.25$ and $P(I_e^E = 1) = 0.25$,

$$P(I_{se}^{SE} = 1 | I_s^S, I_e^E) = \begin{cases} 0.01 & I_s^S = I_e^E = 0 \\ 0.1 & I_s^S = 0, I_e^E = 1 \\ 0.1 & I_s^S = 1, I_e^E = 0 \\ 0.25 & I_s^S = I_e^E = 1 \end{cases}$$

However, the authors did not explicitly explain the rationale for selecting the parameters. In Wakefield (2010) [46], it is modeled as

$$P(I_{se}^{SE}=1|I_s^S,I_e^E)=\begin{cases} p_{00}=0 & I_s^S=I_e^E=0 \\ p_{01}=0 & I_s^S=0,I_e^E=1 \\ p_{10}=0 & I_s^S=1,I_e^E=0 \\ p_{11}=1 & I_s^S=I_e^E=1 \end{cases}$$

Or we may release the dependency between main effects and interactions, by setting

$$P(I_{se}^{SE}=1|I_s^S,I_e^E)=\begin{cases} p_{00}=\pi_s^S\pi_e^E & I_s^S=I_e^E=0 \\ p_{01}=\pi_s^S\pi_e^E & I_s^S=0,I_e^E=1 \\ p_{10}=\pi_s^S\pi_e^E & I_s^S=1,I_e^E=0 \\ p_{11}=\pi_s^S\pi_e^E & I_s^S=I_e^E=1 \end{cases}$$

2nd stage. Priors are given to the mixture indicators respectively:

$$I_s^S \sim Bin(\pi_s^S)$$

$$I_e^E \sim Bin(\pi_e^E)$$

Even when both main effects of each SNP $\times$ Exposure pair exist, the interaction effect still can be missing. To solve this, the variances of interaction effects can be proposed randomly inverse gamma distributed as:

$$\sigma^2_{se_\epsilon} \sim IG(a_{se_\epsilon}, b_{se_\epsilon})$$

$$\sigma^2_{se} \sim IG(a_{se}, b_{se})$$

Prior Structure

Hinted by Chipman (2006) [47], we propose to use the priors structure:

$$P(I_{se}^{SE} = 1 | I_s^S, I_e^E) = \frac{I_s^S * \pi_s^S + I_e^E * \pi_e^E}{2} +$$

This relates to:

$$P(I_{se}^{SE} = 1 | I_s^S, I_e^E) \begin{cases} p_{00} = 0 & I_s^S = I_e^E = 0 \\ p_{01} = \frac{\pi_e^E}{2} & I_s^S = 0, I_e^E = 1 \\ p_{10} = \frac{\pi_s^S}{2} & I_s^S = 1, I_e^E = 0 \\ p_{11} = \frac{\pi_s^S + \pi_e^E}{2} & I_s^S = I_e^E = 1 \end{cases}$$

Bayesian Model Averaging

The Bayesian Model Averaging has been applied in several statistical models [52] [53]. Here we modify the model by incorporating several competing candidate priors we called **Module**'s for π_s^S, π_e^E. The K **Module**s are:

$$(\pi_1^{S_{(1)}}, ..., \pi_S^{S_{(1)}}, \pi_1^{E_{(1)}}, ..., \pi_E^{E_{(1)}})$$

$$...$$

$$(\pi_1^{S_{(k)}}, ..., \pi_S^{S_{(k)}}, \pi_1^{E_{(k)}}, ..., \pi_E^{E_{(k)}})$$

$$...$$

$$(\pi_1^{S_{(K)}}, ..., \pi_S^{S_{(K)}}, \pi_1^{E_{(K)}}, ..., \pi_E^{E_{(K)}})$$

We denote each Module k as M_k. Then the likelihood function under M_k and parameters $A_{(k)}$ will be:

$$L(D|A_{(k)}, M_k) = \prod_{i=1}^{N} \{p_{i(k)}\}^{y_i} \{1 - p_{i(k)}\}^{1-y_i}$$

where

$$p_{i(k)} = \frac{e^{\alpha_{(k)}+B_{(k)}Y_i+\Gamma_{(k)}Z_i+\Theta_{(k)}(Y_i\otimes Z_i)}}{1+e^{\alpha_{(k)}+B_{(k)}Y_i+\Gamma_{(k)}Z_i+\Theta_{(k)}(Y_i\otimes Z_i)}}$$

The posterior model probability of M_k is given by

$$pr(M_k|D) = \frac{L(D|M_k)pr(M_k)}{\sum_{k=1}^{K} L(D|M_k)pr(M_k)}$$

We assume the Modules are equally possible *a priori* with $pr(M_k) = 1/K$ and $L(D|M_k)$ is the marginal likelihood of model M_k,

$$L(D|M_k) = \int L(D|A_{(k)}, M_k) f(A_{(k)}|M_k) dA_{(k)}$$

in which, $f(A_{(k)}|M_k)$ corresponds to prior Modules with:

$$I_s^{S(k)} \sim Bin(\pi_s^{S(k)})$$

$$I_e^{E(k)} \sim Bin(\pi_e^{E(k)})$$

And the full expression of all parameters included in $f(A_{(k)}|M_k)$ is:

$$\prod_{s=1;e=1}^{s=S,e=E} f(\beta_{s(k)}, \gamma_{e(k)}, \theta_{se(k)}|I_s^{S(k)}, I_e^{E(k)}, \sigma^2_{se_\epsilon(k)}, \sigma^2_{se(k)})$$

$$\times f(I_s^{S(k)}) \times f(I_e^{E(k)}) \times f(\sigma^2_{se_\epsilon(k)}) \times f(\sigma^2_{se(k)}) \times f(\alpha_{(k)})$$

where we impose a normal prior for $\alpha_{(k)}$. Therefore, the posterior density of parameters will be

$$f(A|D) = \sum_{k=1}^{K} f(A|M_k, D) pr(M_k|D)$$

The BMA estimates $\hat{A}$ for the parameters of main effects and interactions will be based on the model averaging,

$$E(A|D) = \sum_{k=1}^{K} E(A_{(k)}|M_k, D) pr(M_k|D) \approx \sum_{k=1}^{K} \hat{A_{(k)}} pr(M_k|D) := \hat{A}$$

by which, $\hat{A_{(k)}}$ is the unbiased estimator of $E(A_{(k)}|M_k, D)$ (Newton, 1997) [55] and

$$Var(A|D) = \{\sum_{k=1}^{K} (Var(A_{(k)}|M_k, D) + (E(A_{(k)}|M_k, D))^2) pr(M_k|D)\} - (E(A|D))^2$$

$$\approx \{\sum_{k=1}^{K}(Var(A_{(k)}|M_k, D) + (\hat{A_{(k)}})^2)pr(M_k|D)\} - \hat{A}^2$$

$$\approx \{\sum_{k=1}^{K}(\hat{\sigma^2_{A(k)}} + (\hat{A_{(k)}})^2)pr(M_k|D)\} - \hat{A}^2 := \hat{\sigma^2_A}$$

where $\hat{\sigma^2_{A(k)}}$ is the unbiased estimator of $Var(A_{(k)}|M_k, D)$ and $A_{(k)}$ follows

$$f(A_{(k)}|M_k, D) = \frac{f(D, A_{(k)}|M_k)}{f(D|M_k)} = \frac{L(D|A_{(k)}, M_k)f(A_{(k)}|M_k)}{\int L(D|A_{(k)}, M_k)f(A_{(k)}|M_k)dA_{(k)}}$$

which we will sample from by Markov chain Monte Carlo methods via R compatible free software WinBUGS [45].

Numerical Solution

The computation process is as follows:

Step 1 Under each M_k, use MCMC to generate a whole sample set denoted by $A^j_{(k)}$for all parameters $A_{(k)}$ in jth iteration of the simulation.

Step 2 After obtaining all samples $A^j_{(k)}$ $(j = 1, ..., J)$, calculate the values of $L(D|A^j_{(k)}, M_k)$. Approximate the marginal likelihood of $L(D|M_k)$ by

$$\{\frac{1}{J}\sum_{j=1}^{J}(\frac{1}{L(D|A^j_{(k)}, M_k)})\}^{-1}$$

which is called the harmonic estimator (Rafetery, et al. 1997) [54]. Or we can use the Laplace-Metropolis approximation

$$\log \widehat{L(D|M_k)} = \frac{1}{2}P\log(2\pi) + \frac{1}{2}\log|\mathbf{S}| + \sum_{p=1}^{P} logs_p + \sum_{i=1}^{N} logf(y_i|\bar{A_{(k)}}, M_k) + logf(\bar{A_{(k)}}|M_k)$$

in which $|\mathbf{S}|$ is the determinant of variance-covariance matrix for all parameters, s_p is the standard deviation of parameter p and $\bar{A_{(k)}}$ is mean of the posterior distribution

sampled by MCMC.

Step 3 Repeat *Step* 1 and *Step* 2 for Module 1 to K. Calculate the posterior distribution of $Pr(M_k|D)$ for each Module. By Occam's Window criterion, we can exclude some Modules which do not fulfill

$$\frac{P(M_k|D)}{max_{k\in 1,\ldots,K}P(M_k|D)} > \eta$$

by which η is the criterion set up based on the operating characteristics. Then the adjusted $\overline{Pr}(M_k|D)$ of $Pr(M_k|D)$ will be used in the following calculations.

Step 4 By the samples generated and the formula in Section 1.4 with $\overline{Pr}(M_k|D)$, we compute the estimates of interested parameters.

To escape the identifiability problem due to over-parameterization, we fix the priors for the variances of main effect parameters under both null and alternative hypothesis. When we set up the distribution for variance of β_{s_ϵ} and γ_{e_ϵ}, we want the $e^{\pm 1.96\sigma_{s_\epsilon}} = e^{\pm 0.05}$ and $e^{\pm 1.96\sigma_{e_\epsilon}} = e^{\pm 0.05}$, corresponding to the 95% C.I. of parameters at $(0.951, 1.051)$ for odds ratio. Under the alternative hypothesis, we set up the distribution of β_s and γ_e as $e^{\pm 1.645\sigma_s} = e^{\pm log3}$ and $e^{\pm 1.645\gamma_e} = e^{\pm log4}$ for 90% C.I., corresponding to odds ratio intervals $(\frac{1}{3}, 3)$ for gene effects and $(\frac{1}{4}, 4)$ for environmental effects. So $\sigma^2_{s_\epsilon} = \sigma^2_{e_\epsilon} = 0.00065$, and $\sigma^2_s = .446$ and $\sigma^2_e = .710$. As shown in the simulation studies, the prior settings for the variance do not affect the inferences dramatically. We will omit those parameters by fixing the variances in the mixture for the computation reasons. Similarly as main effects, we set $\sigma^2_{se_\epsilon} = 0.00065$ and $\sigma^2_{se} = .710$.

Data Generation

Suppose we want to generate a big data set with totally 1000 cases and 1000 controls. Firstly, we consider a fixed number of SNPs and exposure factors, say 6 SNPs with 2 non-null as bolded (SNP1, SNP2, SNP3, SNP4, **SNP5**, **SNP6**) and 2 exposure factors (exp1, **exp2**) with 1 non-null. And 5 interactions are non-null among 12 (int11, int12, int21, **int22**, int31, **int32**, **int41**, int42, int51, **int52**, **int61**, int62). The Minor Allele Frequency is 0.3 for null allele and 0.1 for non-null allele, and the exposure prevalence is 0.1 for both the null and non-null factors. Therefore, **int41** reflects 'null null' pair interaction, **int22 int32** 'null non-null', **int61** 'non-null null' and **int52** 'non-null non-null' interactions.

We assume the prevalence of disease for all factors other than those within the model is 0.2, which corresponds to intercept $\alpha = -1.386$ by $\alpha = log(\frac{p}{1-p})$. The odds ratio for SNP minor allele is supposed to be 1 for null allele and 1.25 for non-null, which corresponds to 0 and **0.223** by $\beta = log(OR)$. Remark that we only use the complete recessive genotype model with no dominant effects, in that the odds radio between AA and aa for disease will be $e^{2\beta}$. The odds radio for the exposure factors are 1 for null and 2 for non-null, which corresponds to 0 and **0.693** of the coefficient.

One important assumption of the simulation is that the interaction terms are generated independently away from each factor. The odds radio for the interactions is 1 for null and 2 for non-null, which corresponds to 0 and **0.693** for the regression coefficients of interactions.

	SNP						Exposure		$Pr(M\|Data)$
	$s1$	$s2$	$s3$	$s4$	$s5^*$	$s6^*$	$e1$	$e2^*$	
Module1	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.197
Module2	0.2	0.2	0.2	0.2	0.8	0.8	0.2	0.8	0.788
Module3	0.8	0.8	0.8	0.8	0.2	0.2	0.8	0.2	4.6×10^{-8}
Module4	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.011
Module5	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.003

Table 2.5: Posterior probability of each proposed prior in the elicitation

Model Evaluation

Table 2.5 shows that the proposed BMA model correctly identified the best prior module thru the Bayesian Model Averaging process. Then we can use the BMA way or the best module way or Occam's Window [52] way to get the estimates and compare the power of detecting true positive and negative effects among them as well as with other methods like the single and full logistic regression models.

Discussion

Here we proposed an approach with several candidate priors for the mixture indicators. By the Bayesian models averaging, the best fit prior Modules will play important roles and worse fit Modules be removed automatically from the model thru the posterior interpretations of models, in other words it is like the Bayes Factor way of model selection. The information about some SNPs or Exposure factors being significant or trivial may exist from the previous studies or biological pathway information. Then we can readily set some indictor priors to reflect this background information thru the Modules. Also, the competing priors structure for the variances

of the effects enable us to assign certain priors to reflect the proposals based on pre-existing knowledge about the effects and researcher's preference. Wakefield (2010) [46] gives an idea of hierarchical models but fix all parameters about indicators. By this modeling approach, we hope to gain more knowledge and model selection mechanisms can greatly help increase the power and control the Type-I error.

3. Bayesian Natural and Orthogonal Interaction Model for Gene-Gene and Gene-Environment Interactions

3.1 Motivation

In the past few years, genome-wide association studies (GWAS) have substantially contributed to the success of searching for causal genetic variants for complex diseases and traits. In GWAS, hundreds of thousands of single nucleotide polymorphism (SNP) are assayed from the participants in a relatively large study group. Usually each genetic factor (locus) is analyzed separately. It has been proved a powerful tool for investigating the underlying genetic reasons for the complex diseases and traits [3] [56]. GWAS have found a large number of causal regions for the complex diseases, which provided deep insight about the genetic mechanisms of diseases. However, a large portion of the heritability in those diseases is still missing by GWAS studies. One limitation of GWAS is that the interactions are ignored in those studies. So in recent years more efforts are being made to investigate the interactions among the genetic factors and between genetic and environmental factors that can explain the missing heritability.

In this project, we first reviewed the Natural Orthogonal Interaction (NOIA) Model that was proposed for studying the gene-gene and gene-environment interactions. Then a Bayesian mixture model was proposed to be combined with the NOIA model and the usual functional model. The proposed approaches were applied in the lung cancer case-control studies.

3.2 Natural and Orthogonal Interaction (NOIA) Model

3.2.1 Functional Model for Genotype-Phenotype Mapping

We first introduced the usual functional model for the genotype-phenotype mapping. Suppose that we have a study sample with n individuals. In the usual genotype-phenotype quantitative trait locus (QTL) studies, the trait is usually assumed normally distributed. For the single locus study, the trait is influenced by a diallelic locus with major allele A and minor allele a. For the ith individual, let y_i be denoted as the observed quantitative trait and G_i as the genotypic value for the diallelic locus. There is a linear relationship between the quantitative trait and the genotypic value as

$$y_i = G_i + \epsilon_i \tag{3.1}$$

where G_i can take three values G^{11}, G^{12} and G^{22} which are the genotypic values for genotypes AA, Aa and aa. So for the n individuals in the sample, the vector of genotypic values can be modeled with the individual design matrix Z

$$\begin{pmatrix} G_1 \\ G_2 \\ ... \\ ... \\ G_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ ... & ... & ... \\ 0 & 1 & 0 \\ ... & ... & ... \\ 0 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} G^{11} \\ G^{12} \\ G^{22} \end{pmatrix} \tag{3.2}$$

The genotypic values G_{11}, G_{12} and G_{22} can be modeled as product of the genetic design matrix D and the genetic effect values E

$$G = D \times E \tag{3.3}$$

So let X denote the design matrix for the whole sample of study $X = Z \times D$. Then we can model the relation between the quantitative trait and genotypic value for all the individuals by $Y = X * E + \epsilon$. We then have

$$Y = Z \times D \times E + \epsilon = Z \times G + \epsilon \tag{3.4}$$

The question is about how to code the design matrix of the genotypic values D. Different mapping methods make different assumptions. One of the widely applied ways, which is referred as the usual functional model is coded as

$$\begin{pmatrix} G^{11} \\ G^{12} \\ G^{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 0 \end{pmatrix} \times \begin{pmatrix} R \\ a \\ d \end{pmatrix} \tag{3.5}$$

where R denotes the intercept, a denotes the additive displacement due to each increment in the number of variant alleles, and d denotes the deviation of the heterozygous value from the additive model. When we inverted the design matrix D, we can get the genetic effect value from the genotypic values

$$\begin{pmatrix} R \\ a \\ d \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -1/2 & 0 & 1/2 \\ -1/2 & 1 & -1/2 \end{pmatrix} \begin{pmatrix} G^{11} \\ G^{12} \\ G^{22} \end{pmatrix} \tag{3.6}$$

Here, R is also called the reference point which corresponds to the genotypic value G^{11} of the homozygotes AA. The additive effect a is half of the difference of the genotypic values G^{11} and G^{22} for the homozygotes AA and aa. The dominance effect d is the difference between the genotypic value G^{12} for heterozygote Aa and the average of the genotypic values of G^{11} and G^{22}. This is the model set up for quantitative trait study. For case-control studies, we can propose the same design matrix with different generalized linear models. As mentioned in Zeng (2002, 2005) [7] [57], there are other coding approaches for functional models and the main difference depends on the modeling for the reference point. For example, we may want to use the average of

the two homozygotes AA and aa as the reference point, then the design matrix will be

$$\begin{pmatrix} G^{11} \\ G^{12} \\ G^{22} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix} \times \begin{pmatrix} R \\ a \\ d \end{pmatrix} \tag{3.7}$$

Here we will only consider the first model (3.6). These models are called functional models because they model the genetic effects as allele substitutions on genotypes which provide the biological functions.

3.2.2 NOIA Model for Genotype-Phenotype Mapping

Another coding scheme was proposed by Alvarez-Castro and Carlborg (2007) [58] for estimating the genetic effects for a quantitative traits and gene-gene interactions. As shown in Ma et al (2012) [60], the design matrix G could be modeled as

$$\begin{pmatrix} G^{11} \\ G^{12} \\ G^{22} \end{pmatrix} = \begin{pmatrix} 1 & -\bar{N} & -2p_{12}p_{22}/V \\ 1 & 1-\bar{N} & 4p_{11}p_{22}/V \\ 1 & 2-\bar{N} & -2p_{11}p_{12}/V \end{pmatrix} \times \begin{pmatrix} \mu \\ \alpha \\ \delta \end{pmatrix} \tag{3.8}$$

where p_{11}, p_{12} and p_{22} denote the frequencies for the genotype of AA, Aa and AA. And $\bar{N} = p_{12} + 2p_{22}$ and $V = p_{12} + 4p22 - (p_{12} + 2p_{22})^2 = p_{11} + p_{22} - (p_{11} - p_{22})^2$. N is regarded as the frequency of the minor allele a in the population, $\bar{N}$ is the estimate of N in the sample and V is the variance of $\bar{N}$. Similarly, we can invert the design matrix to get the estimate of the genetic effects from the genotypic values

$$\begin{pmatrix} \mu \\ \alpha \\ \delta \end{pmatrix} = \begin{pmatrix} p_{11} & p_{12} & P_{22} \\ p_{11}^{'} & p_{12}^{'} & p_{22}^{'} \\ -1/2 & 1 & -1/2 \end{pmatrix} \times \begin{pmatrix} G^{11} \\ G^{12} \\ G^{22} \end{pmatrix} \tag{3.9}$$

where

$$p_{ij}^{'} = p_{ij}\frac{N_{ij} - \bar{N}}{V}$$

Here the design matrix in (3.8) is dependent on the genotype frequency for the locus in the sample. Then the proposed statistical model is an orthogonal model which produce the uncorrelated estimates of the parameters and the variance component decomposition is also orthogonal [58]. Also, in the statistical model, the reference point is $\mu = p_{11}G^{11} + p_{12}G^{12} + p_{22}G^{22}$.

There are two advantages for this statistical models as pointed out by Alvarez (2007, 2008) [58] [59]. First, the estimate of each genetic effect is not influenced by the estimate of other effects. So when we do variable selections, the results will be consistent no matter of the number of loci and genotype frequencies. Second, the variance components can be explained by being separated into different components according to different effects. As shown in Alvarez (2007) [58], the design matrix of the statistical model fulfill the orthogonality conditions specified by Zeng (2005) [57]:

$$X^T X = D^T \times Z^T \times Z \times D = nD^T \times Q \times D \tag{3.10}$$

where

$$Q = \begin{pmatrix} p_{11} & 0 & 0 \\ 0 & p_{12} & 0 \\ 0 & 1 & p_{22} \end{pmatrix} \tag{3.11}$$

Therefore, to attain the orthogonality feature, the design matrix D should fulfill the conditions:

$$\begin{aligned} d_{11}p_{11} + d_{22}p_{12} + d_{12}p_{22} &= 0 \\ d_{13}p_{11} + d_{23}p_{12} + d_{33}p_{22} &= 0 \\ d_{12}d_{13}p_{11} + d_{22}d_{23}p_{12} + d_{32}d_{33}p_{22} &= 0 \end{aligned}$$

The NOIA statistical model fulfills the above conditions as shown in Alvarez (2007) [58] and the functional model does not hold this feature. This feature for the statistical model is held even if the Hardy-Weinberg Equilibrium (HWE) is violated, because the

coding scheme is based on the genotype frequencies not allele frequencies. This model is appealing because it provides a consistent approach for the hypothesis testing as well as the variable/model selection in the genetic studies, which is important for the studies with multiple genetic and environmental factors with their interactions.

3.3 NOIA Model for Gene-Gene and Gene-Environment Interactions

We have already introduced the NOIA statistical model for the single locus study. As mentioned earlier, under the NOIA model the estimate of the effect parameters are uncorrelated. So when considering multiple loci for the quantitative trait, under the condition of orthogonality, the estimates for one locus will not be affected by the others. Here we first introduced the NOIA statistical model for two loci and the model can be readily extended to more loci. Then we will also introduced the NOIA statistical model for the gene-environment interaction as in Ma et al (2012) [60]. Suppose that we consider two diallelic loci $A_{ij} = AA, Aa, aa$ and $B_{ij} = BB, Bb, bb$. p_{ij} and q_{ij} are the genotype frequencies for A_{ij} and B_{ij}. Let N_A denote the allele frequency for allele a in the population and N_B for allele b. We let $\bar{N_A}$ and $\bar{N_B}$ denote the estimate of the allele frequency in the sample and V_A and V_B are denoted as the variance of $\bar{N_A}$ and $\bar{N_B}$, respectively. Therefore, we can get

$$\begin{aligned} \bar{N_A} &= p_{12} + 2p_{22} \\ V_A &= p_{12} + 4p_{22} - (p_{12} + 2p_{22})^2 \\ \bar{N_B} &= q_{12} + 2q_{22} \\ V_B &= q_{12} + 4q_{22} - (q_{12} + 2q_{22})^2 \end{aligned} \tag{3.12}$$

By Alvarez-Castro and Carborg (2007) [58], the genotypic value G_{AB} for the two loci gene-gene interaction model are

$$G_{AB} = D_{AB} \times E_{AB} = (D_B \otimes D_A) \times E_{AB}$$

where D_B and D_A are the NOIA statistical model design matrix for loci A and B. The genotypic value G_{AB}, the design matrix D_{AB} and the genetic effect vector E_{AB}

$$G_{AB} = G_B \otimes G_A = \begin{pmatrix} G_B^{11} \\ G_B^{12} \\ G_B^{22} \end{pmatrix} \otimes \begin{pmatrix} G_A^{11} \\ G_A^{12} \\ G_A^{22} \end{pmatrix} = \begin{pmatrix} G^{1111} \\ G^{1211} \\ G^{2211} \\ G^{1112} \\ G^{1212} \\ G^{2212} \\ G^{1122} \\ G^{1222} \\ G^{2222} \end{pmatrix} \tag{3.13}$$

$$D_{AB} = D_B \otimes D_A = \begin{pmatrix} 1 & -\bar{N}_B & -2p_{12}p_{22}/V_B \\ 1 & 1-\bar{N}_B & 4p_{11}p_{22}/V_B \\ 1 & 2-\bar{N}_B & -2p_{11}p_{12}/V_B \end{pmatrix} \otimes \begin{pmatrix} 1 & -\bar{N}_A & -2p_{12}p_{22}/V_A \\ 1 & 1-\bar{N}_A & 4p_{11}p_{22}/V_A \\ 1 & 2-\bar{N}_A & -2p_{11}p_{12}/V_A \end{pmatrix} \tag{3.14}$$

and

$$E_{AB} = E_B \otimes E_A = \begin{pmatrix} \mu \\ \alpha_A \\ \delta_A \\ \alpha_B \\ \alpha\alpha \\ \delta\alpha \\ \delta_B \\ \alpha\delta \\ \delta\delta \end{pmatrix} \tag{3.15}$$

Therefore we can get the statistical model for the two-loci study as

$$\begin{pmatrix} G^{1111} \\ G^{1211} \\ G^{2211} \\ G^{1112} \\ G^{1212} \\ G^{2212} \\ G^{1122} \\ G^{1222} \\ G^{2222} \end{pmatrix} = \begin{pmatrix} 1 & -\bar{N}_B & -2q_{12}q_{22}/V_B \\ 1 & 1-\bar{N}_B & 4q_{11}q_{22}/V_B \\ 1 & 2-\bar{N}_B & -2q_{11}q_{12}/V_B \end{pmatrix} \otimes \begin{pmatrix} 1 & -\bar{N}_A & -2p_{12}p_{22}/V_A \\ 1 & 1-\bar{N}_A & 4p_{11}p_{22}/V_A \\ 1 & 2-\bar{N}_A & -2p_{11}p_{12}/V_A \end{pmatrix} \times \begin{pmatrix} \mu \\ \alpha_A \\ \delta_A \\ \alpha_B \\ \alpha\alpha \\ \delta\alpha \\ \delta_B \\ \alpha\delta \\ \delta\delta \end{pmatrix} \tag{3.16}$$

So for the two loci studies, there are 9 parameters including one baseline(intercept) μ, 4 main effects α_A, δ_A, α_B and δ_B and 4 pairwise interactions $\alpha\alpha$, $\alpha\delta$, $\delta\alpha$ and $\delta\delta$.

Similarly, for the functional 2 loci model, the design matrix the Kronecker product of the design matrix for each locus. And we can get

$$\begin{pmatrix} G^{1111} \\ G^{1211} \\ G^{2211} \\ G^{1112} \\ G^{1212} \\ G^{2212} \\ G^{1122} \\ G^{1222} \\ G^{2222} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 0 & 1 & 2 & 0 & 1 & 2 & 0 \\ 1 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 2 & 2 & 2 & 0 & 0 & 0 \\ 1 & 2 & 0 & 2 & 4 & 0 & 0 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} R \\ a_A \\ d_A \\ a_B \\ aa \\ da \\ d_B \\ ad \\ dd \end{pmatrix} \tag{3.17}$$

For the usual functional model, there are one intercept R, 4 main effects a_A, d_A, a_B and d_B and 4 pairwise interactions aa, ad, da and dd. In the functional model, the

design matrix is not based on the frequency of the genotype so the effect parameters have different meanings from the statistical model (3.16). Actually the effect parameter values of both models can be converted to each other by

$$E_{AB}^{stat} = (D_{AB}^{stat})^{-1} D_{AB}^{func} E_{AB}^{func}$$

So far we have introduced the NOIA statistical model for the gene-gene interaction models, but the NOIA model was also developed by Ma et al (2012) [60] for the gene-environment interactions. Suppose we consider a binary environmental factor with a genetic factor in the quantitative trait analysis. Let M denote the phenotypic values for the environmental factor. In the usual functional model, we assume that

$$M = \begin{pmatrix} M_1 \\ M_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \times \begin{pmatrix} R_m \\ a_m \end{pmatrix} \tag{3.18}$$

where M_1 and M_2 are the two levels of the phenotypic values of the environmental factors, R_m is the baseline and a_m is the additive effect.Then we can get the effect parameters as

$$E_m^{func} = \begin{pmatrix} R \\ a_m \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \times \begin{pmatrix} M_1 \\ M_2 \end{pmatrix}$$

Suppose that $m_1 = m$ and $m_2 = 1 - m$ are the frequency of the environmental factors in the population. By using the same technique as in the NOIA model for GxG studies, Ma el al (2012) [60] derived the orthogonal interaction model for the environmental factor as

$$M = \begin{pmatrix} M_1 \\ M_2 \end{pmatrix} = \begin{pmatrix} 1 & m-1 \\ 1 & m \end{pmatrix} \times \begin{pmatrix} \mu \\ \alpha_m \end{pmatrix} \tag{3.19}$$

and the effects can be estimated by

$$E_m^{stat} = \begin{pmatrix} \mu \\ \alpha_m \end{pmatrix} = \begin{pmatrix} m & 1-m \\ -1 & 1 \end{pmatrix} \times \begin{pmatrix} M_1 \\ M_2 \end{pmatrix}$$

Therefore, by applying the Kronecker derivations, we get the following usual functional model for the gene-environment interaction for one locus and one exposure as

$$G_{GM} = \begin{pmatrix} 1 & -\bar{N} & -2p_{12}p_{22}/V & m-1 & -(m-1)\bar{N} & -2p_{12}p_{22}/V \\ a_G & 1-\bar{N} & 4p_{11}p_{22}/V & m-1 & (m-1)(1-\bar{N}) & 4p_{11}p_{22}/V \\ d_G & 2-\bar{N} & -2p_{11}p_{12}/V & m-1 & (m-1)(2-\bar{N}) & -2p_{11}p_{12}/V \\ a_M & -\bar{N} & -2p_{12}p_{22}/V & m & -m & -2p_{12}p_{22}/V \\ aa & 1-\bar{N} & 4p_{11}p_{22}/V & m & m(1-\bar{N}) & 4p_{11}p_{22}/V \\ da & 2-\bar{N} & -2p_{11}p_{12}/V & m & m(2-\bar{N}) & -2p_{11}p_{12}/V \end{pmatrix} \times \begin{pmatrix} \mu \\ \alpha_G \\ \delta_G \\ \alpha_M \\ \alpha\alpha \\ \delta\alpha \end{pmatrix} \tag{3.20}$$

and the usual functional model as

$$G_{GM} = \begin{pmatrix} G^{11}M_1 \\ G^{12}M_1 \\ G^{22}M_1 \\ G^{11}M_2 \\ G^{12}M_2 \\ G^{22}M_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 0 & 1 & 2 & 0 \end{pmatrix} \times \begin{pmatrix} R \\ a_G \\ d_G \\ a_M \\ aa \\ da \end{pmatrix} \tag{3.21}$$

The parameters of effects in (3.20) and (3.21) can be converted to each other as

$$\begin{pmatrix} \mu \\ \alpha_G \\ \delta_G \\ \alpha_M \\ \alpha\alpha \\ \delta\alpha \end{pmatrix} = \begin{pmatrix} 1 & \bar{N} & p_{12} & 1-m & (1-m)\bar{N} & (1-m)p_{12} \\ 0 & 1 & p_{12}^{'} & 0 & 1-m & (1-m)p_{12}^{'} \\ 0 & 0 & 1 & 0 & 0 & 1-m \\ 0 & 0 & 0 & 1 & \bar{N} & p_{12} \\ 0 & 0 & 0 & 0 & 1 & p_{12}^{'} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} R \\ a_G \\ d_G \\ a_M \\ aa \\ da \end{pmatrix} \tag{3.22}$$

where $p_{12}^{'} = p_{12}\frac{N_{12}-\bar{N}}{V}$

3.4 Bayesian NOIA Model for Interactions

In the previous sections, we reviewed the NOIA statistical model for the gene-gene and gene-environment interaction models. In reality, we may consider multiple genetic and environmental factor simultaneously in the same model. Here we propose a Bayesian model that combines the Bayesian mixture model and the NOIA statistical model and we contrast model fitting with the Bayesian approach for the usual functional model.

We consider gene-environment interaction model (3.16) and the other models for the gene-gene interaction model can be similarly set up. In quantitative trait studies, we assume that the trait will follow a normal distribution with the mean as the genotypic value

$$y_i = G_i + \epsilon_i$$

where G_i is the genotypic value for the genotypes of ith observation.

$$y_i = Z_i \times D \times E \tag{3.23}$$

where Z_i is the design matrix for the genotype of ith observation and D is the design matrix as in 3.16 and E is the effect parameters. Therefore, the parameter μ is the intercept in the linear regression, and α_A, δ_A, α_B, δ_B are the parameters for the main effects and $\alpha\alpha$, $\delta\alpha$, $\alpha\delta$ and $\delta\delta$ are the parameters for the pairwise interactions. Therefore, we propose the Bayesian mixture model for each of the parameters as

$$\begin{aligned}
\alpha_A &\sim N(0, I_{1\alpha}\sigma^2_{1\alpha} + (1 - I_{1\alpha})\sigma^2_{1\alpha\epsilon}) \\
\delta_A &\sim N(0, I_{1\delta}\delta^2_{1\delta} + (1 - I_{1\delta})\sigma^2_{1\delta\epsilon}) \\
\alpha_B &\sim N(0, I_{2\alpha}\sigma^2_{2\alpha} + (1 - I_{2\alpha})\sigma^2_{2\alpha\epsilon}) \\
\delta_B &\sim N(0, I_{2\delta}\delta^2_{2\delta} + (1 - I_{2\delta})\sigma^2_{2\delta\epsilon})
\end{aligned}$$

for the interaction parameters

$$\begin{aligned}
\alpha\alpha &\sim N(0, I_{\alpha\alpha}\sigma^2_{\alpha\alpha} + (1 - I_{\alpha\alpha})\sigma^2_{\alpha\alpha\epsilon}) \\
\alpha\delta &\sim N(0, I_{\alpha\delta}\sigma^2_{\alpha\delta} + (1 - I_{\alpha\delta})\sigma^2_{\alpha\delta\epsilon}) \\
\delta\alpha &\sim N(0, I_{\delta\alpha}\sigma^2_{\delta\alpha} + (1 - I_{\delta\alpha})\sigma^2_{\delta\alpha\epsilon}) \\
\delta\delta &\sim N(0, I_{\delta\delta}\sigma^2_{\delta\delta} + (1 - I_{\delta\delta})\sigma^2_{\delta\delta\epsilon})
\end{aligned}$$

for the intercept $\mu \sim N(0, \sigma^2_\mu)$

In the second level we assume the hyper-priors for each of the indicators, for SNP A:

$$\begin{aligned}
I_{1\alpha} &\sim bernoulli(p_{1\alpha}) \\
I_{1\delta} &\sim bernoulli(p_{1\delta})
\end{aligned}$$

for SNP B:

$$\begin{aligned}
I_{2\alpha} &\sim bernoulli(p_{2\alpha}) \\
I_{2\delta} &\sim bernoulli(p_{2\delta})
\end{aligned}$$

for the interactions:

$$\begin{aligned}
I_{\alpha\alpha} &\sim bernoulli(p_{\alpha\alpha}) \\
I_{\alpha\delta} &\sim bernoulli(p_{\alpha\delta}) \\
I_{\delta\alpha} &\sim bernoulli(p_{\delta\alpha}) \\
I_{\delta\delta} &\sim bernoulli(p_{\delta\delta})
\end{aligned}$$

Therefore, we can conduct the Bayesian inference for the parameters and select the significant genetic effects from the model with multiple genetic factors. In the original NOIA statistical model, all the results are based on the parameter estimation. Ma et al (2012) [60] compared the hypothesis test for the gene-environment interactions. By our proposed model, we can also do the variable selection by the mixture model.

This framework can be extended to the gene-environment interaction, usual functional models. For the case-control binary outcome, we may propose a similar framework for a generalized linear model. Suppose that Y is the vector indicating the disease status of all observations

$$\begin{aligned} Y_i &\sim bernoulli(p_i) \\ logit(p_i) &= G_i = Z_i \times D \times E \end{aligned}$$

Therefore, the similar prior structures can be proposed for logistic regression models.

3.5 Real Data Example

We conduct the real data analysis using the Bayesian NOIA model and the Bayesian Functional model on the lung cancer dataset Totally we consider 6 SNPs and 1 environmental factor. For each SNP, we model both the additive effect and dominance effect. For the smoking, there is only one additive effect. For interactions, there are 12 pairwise gene-environment interactions. To illustrate the methods, we do not take into account the gene-gene interactions in the model. The results are shown in Figure 3.1. Both models identifies the smoking effects. The NOIA statistical model find the interaction between the additive effect and the smoking, while the usual functional model does not identify it.

3.6 Discussion

In this chapter, we introduced the NOIA statistical model and the usual functional model for the Genotype-Phenotype mapping. Then the NOIA statistical model for the gene-gene interactions and gene-environment interactions was introduced. In practice, we may be interested in making the variable selection on the SNP level. Then a group selection approach could be attempted for achieving this goal. To find out the predisposing SNPs from a candidate sets, we may put extra constraint on

the priors of parameters. One approach to prior structure set up for the gene-gene interactions can be

for SNP A:

$$I_{1\alpha} = I_{1\delta} \sim bernoulli(p_1)$$

for SNP B:

$$I_{2\alpha} = I_{2\delta} \sim bernoulli(p_2)$$

and for interactions:

$$I_{\alpha\alpha} = I_{\alpha\delta} = I_{\delta\alpha} = I_{\delta\delta} \sim bernoulli(p_3)$$

By this modeling scheme, each SNP and the interaction can be selected as a group comprising several genetic effects.

Another extension could be using the Lasso-type hyper-priors for the variance components, so the parameter will shrink to zero as well as being selected in the model. For example, we assume that

for SNP A:

$$\begin{aligned} \alpha_A &\sim I_{1\alpha} \times dexp(\lambda_1) + (1 - I_{1\alpha}) \times \mathbf{1}_{\emptyset} \\ \delta_A &\sim I_{1\delta} \times dexp(\lambda_1) + (1 - I_{1\delta}) \times \mathbf{1}_{\emptyset} \end{aligned}$$

for SNP B:

$$\begin{aligned} \alpha_B &\sim I_{2\alpha} \times dexp(\lambda_2) + (1 - I_{2\alpha}) \times \mathbf{1}_{\emptyset} \\ \delta_B &\sim I_{2\delta} \times dexp(\lambda_2) + (1 - I_{2\delta}) \times \mathbf{1}_{\emptyset} \end{aligned}$$

for the interactions:

$$\begin{aligned} \alpha\alpha &\sim I_{\alpha\alpha} \times dexp(\lambda_3) + (1 - I_{\alpha\alpha}) \times \mathbf{1}_{\emptyset} \\ \alpha\delta &\sim I_{\alpha\delta} \times dexp(\lambda_3) + (1 - I_{\alpha\delta}) \times \mathbf{1}_{\emptyset} \\ \delta\alpha &\sim I_{\delta\alpha} \times dexp(\lambda_3) + (1 - I_{\delta\alpha}) \times \mathbf{1}_{\emptyset} \\ \delta\delta &\sim I_{\delta\delta} \times dexp(\lambda_3) + (1 - I_{\delta\delta}) \times \mathbf{1}_{\emptyset} \end{aligned}$$

Here σ_1^2, σ_2^2 and σ_3^2 are the common tuning parameters for SNP A, SNP B and the interactions.

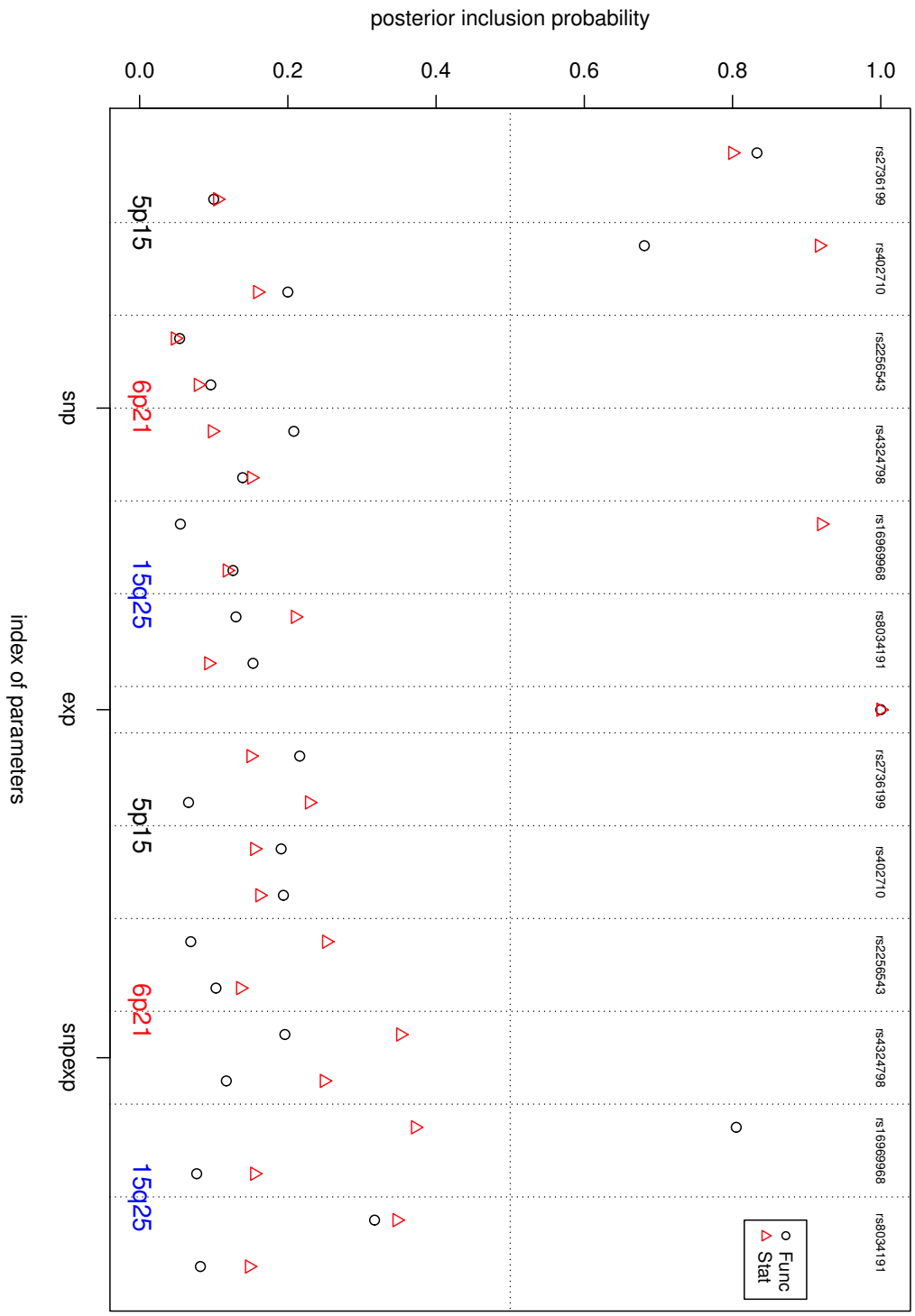

Figure 3.1.: Bayesian NOIA coding analysis results for the lung cancer study.

4. Bayesian Empirical Shrinkage-Type Estimator and Bayesian Model Averaging Approaches for Gene-Environment Interactions and Extensions to Adaptive Borrowing Historical Data in Clinical Studies

4.1 Motivation

In the previous sections, we introduced the proposed Bayesian statistical models for gene-gene and gene-environment interactions based on the framework of variable selection in generalized linear models. Alternative Bayesian statistical models exist for exploring the gene-environment interactions under the other statistical framework. In this section, we will first review two Bayesian approaches that were recently developed to adaptively model the gene-environment interactions depending on the independence between the genetic and environmental factors. Then, we extend the idea to the problem of adaptive borrowing the data from historical studies. We will focus on the evaluations of the event rate, such as in clinical or epidemiological studies. Simulation studies are conducted to demonstrate the advantages of the novel approaches.

4.2 Bayesian Hierarchical Meta-Analysis Model

Suppose that we are interested in borrowing the information from the previous studies such that the sample size can be enriched. Let p_1 and p_0 denote the current and historical event rates. y_1 and n_1 are denoted as the total number of events and total sample size in current study. y_{0i} and n_{0i} with $i = 1, 2, ..., n$ are denoted as the

total number of events and total sample size in each of the previous ith study. For illustration purpose, we first consider the simple case with one current study and one historical study. The maximum likelihood estimator for the event rate of the historical study is $\frac{y_0}{n_0}$ and for the current study is $\frac{y_1}{n_1}$. When we assume the existence of exchangeability between the two studies, they can be pooled together to provide the estimator of the event rate as $\frac{y_1+y_0}{n_1+n_0}$.

Here we use a Bayesian hierarchical model to model the relationship between the two studies. For the current study, we assume

$$\begin{aligned} y_1 &\sim binomial(n_1, p_1) \\ logit(p_1) &= \alpha_1 \\ \alpha_1 &\sim N(\mu_a, 1/\tau_a) \\ \mu_a &\sim N(a, b) \\ \tau_a &\sim Gamma(c, d) \end{aligned} \tag{4.1}$$

For the historical study, we assume

$$\begin{aligned} y_0 &\sim binomial(n_0, p_0) \\ logit(p_0) &= \alpha_0 \\ \alpha_0 &\sim N(\mu_a, 1/\tau_a) \end{aligned} \tag{4.2}$$

Now we want to compare this with the frequentist approach. Since the endpoint of the study is event rate, we used the Binomial exact test to compare with the Bayesian model's performance. For the Binomial exact test, the p-value is equal to the probability of obtaining the event number as equal or less than the observed one. We performed the exact test for both of the current study and the pooled study which blindly combined the two data sets together. For the Bayesian approach, we implemented two different priors for the variance components of the model to compare it

with the exact test. $c = 10, d = 10$ is the prior that is relatively informative due to its smaller variance of 0.1. $c = 0.001, d = 0.001$ is a more diffuse prior with a larger variance of 1000. This reflects our belief that we expect to see similarities between the two studies due to the exchangeability.

4.2.1 Example

Suppose that in one historical study we collect 314 observations. In the current study, we want to evaluate the study power and Type-I error given the sample size 100. We want to test the hypothesis about the rate:

$$H_0 : p_0 \geq 0.08$$
$$H_a : p_1 < 0.08 \tag{4.3}$$

To calculate the Type-I error rate, the simulation steps are:

1. Fix $n_1 = 100$, $n_0 = 314$, $p_1 = \mathbf{0.08}$ and $p_0 = 0.04$

2. Randomly generate $simN$ numbers of event numbers y_1^j and y_0^j, $j = 1, 2, ..., simN$ by

$$y_1 \sim binomial(p_1, n_1)$$
$$y_0 \sim binomial(p_0, n_0)$$

3. Under each y_1^j, y_0^j with the fixed number n_1 and n_0, we implemented 4 approaches mentioned above and use $\alpha = 0.05$ as the significance level for exact test. For the Bayesian approach, we calculate $\pi = Pr(p_1 < 0.08)$. When $\pi > 0.95$ we reject H_0, otherwise we accept it.

4. Then we count how many rejections happened for each of the approaches. $\frac{no\ of\ rejection}{simN}$ is the estimated Type-I error rate.

enumerate To calculate the Power for Scenario 2, the simulation steps are:

1. Fix $n_1 = 100$, $n_0 = 314$, $p_1 = \mathbf{0.04}$ and $p_0 = 0.04$

2. Randomly generate $simN$ numbers of event numbers y_1^j and y_0^j, $j = 1, 2, ..., simN$ by

$$y_1 \sim binomial(p_1, n_1)$$

$$y_0 \sim binomial(p_0, n_0)$$

3. Under each y_1^j, y_0^j with the fixed number n_1 and n_0, we perform the 4 approaches mentioned above and use $\alpha = 0.05$ as the significance level for exact test. For the Bayesian approach, we calculate $\pi = Pr(p_1 < 0.08)$. When $\pi > 0.95$ we reject H_0, otherwise we accept it.

4. Then we count how many rejections happened for each of the approaches. $\frac{no\ of\ rejection}{simN}$ is the estimated power.

4.2.2 Result

	Scenario 1		Scenario 2		Scenario 3		Scenario 4		Scenario 5	
p_1, p_0	0.04, 0.04		0.04, 0.03		0.04, 0.03		0.03, 0.03		0.03, 0.04	
sample size n_1	100		100		50		100		100	
	Type-I	Power	Type-I	Power	Type-I	Power	Type-I	Power	Type-I	Power
No borrowing	0.03	0.36	0.03	0.36	0.02	0.12	0.03	0.58	0.03	0.58
Full borrowing	0.73	0.91	0.89	1.00	0.93	1.00	0.89	1.00	0.73	0.95
$Gamma(0.001, 0.001)$	0.34	0.83	0.37	0.90	0.53	0.89	0.37	0.98	0.34	0.93
$Gamma(10, 10)$	0.10	0.58	0.10	0.60	0.09	0.42	0.10	0.74	0.10	0.74

Table 4.1: Frequentist operating characteristics for the 5 scenarios

We performed 100 times of replicates for comparing the performances of the four approaches in Table 4.1. In Scenario 1, the 'no borrowing' exact test controls the

Type-I error well with a lower power. On the contrary, the 'full borrowing' exact test has a large power while the type-I error is poorly inflated. The Bayesian hierarchical model with prior $Gamma(0.001, 0.001)$ has a larger power than the single study while the type-I error is not that inflated. The Bayesian model with $Gamma(10, 10)$ has a better power than the 'no borrowing' approach while the type-I error is still controlled under 0.1. In Scenario 2, the historical rate is a little smaller as 0.03. When compared with Scenario 1, the 'No borrowing' approach has no improvement. The type-I error of 'Full borrowing' is more inflated than in Scenario 1. Both of the Bayesian model has some improvement on power. The 'Gamma(10,10)' also control the false positive rate as in Scenario 1. When we decrease the sample size of the current study from 100 to 50, we observed that all methods perform worse than in Scenario 2. Still, the Bayesian approach performed better than the frequentist exact test. In Scenario 4 and Scenario 5, the current rate was changed from 0.04 to 0.03 in the simulation setup. We find that the Bayesian approaches have a better power than the single study without borrowing any information. They also control the Type-I error rate better than the fully borrowing study. Especially 'Gamma(10,10)' can control the Type-I error under the level of 0.1 and have comparably larger power as the full borrowing approach.

4.2.3 Comment

In this study, the rate estimate without borrowing any information is the unbiased estimate. If we pool the historical study with the current study, the rate estimate will be the biased estimate. However, the power of the study can be improved if the historical study is confirmed to be similar as the current study. In this study, the Bayesian hierarchical model helped to borrow information from the previous trial to improve the study. The power was put higher than the single study and the type-I error rate is controlled relatively well. Here for the Bayesian hierarchical model, we

use two priors which reflect two kinds of belief about the connection between the current study and historical study. Overall, the Bayesian approach performed more adaptively to the information we have. This means that our study can borrow partial information from the historical study thus help improve our current trial study.

4.3 Bayesian Empirical Shrinkage-Type Estimator

4.3.1 Motivating Background

Population based case control studies are commonly used to investigate the effects of gene-environment interactions on complex diseases. It is well known that case-control studies have poor power for detecting the multiplicative interactions [5]. On the other hand, under the assumption of independence for the underlying population, we can conduct the test with more power for the multiplicative interactions based on the odds ratio of dichotomous genotypes and environmental factors in the case group of observations. However, when the independence assumption is violated, the method can result in inflated Type-I error. To solve this uncertainty on the bias-efficiency trade-off, people have proposed approaches such as two-step testing procedure [61] that test the assumption of the independence assumption first and then decide whether to use the case-only approach for reaching larger power or case-control complete approach for controlling lower bias. However, this two-step testing procedure requires relatively large sample size to reach the power for testing the independence. Also, the two-step procedure is a discrete process in which the variance of the test statistics is complicated to measure. Mukherjee [62] [63] proposed an empirical Bayesian shrinkage estimator to solve the bias and efficiency dilemma by taking a composite estimator by a weighted average of the case-only and case-control estimators. The weights are dependent on the dependence between the genotypes

and environmental factors but irrespective of the test of independence.

Suppose that we are considering the case-control study with one dichotomous genetic factor and one dichotomous environmental factor. SNP=1 denotes the susceptible genotype carrier and EXP=1 denotes exposed to the environmental factor. So we will have eight numbers according to each category as in table 4.3. Then $x_1 = (x_{100}, x_{101}, x_{110}, x_{111})$ is the vector of the observed frequency according to each genotype and exposure combination in the cases. And $x_0 = (x_{000}, x_{001}, x_{010}, x_{011})$ denotes the data vector for the controls. Let the corresponding probability for each category will be $p_1 = (p_{100}, p_{101}, p_{110}, p_{111})$ and $p_0 = (p_{000}, p_{001}, p_{010}, p_{111})$ Therefore, the observed number for each category can be viewed as two multinomial distribution as

$$\begin{aligned} x_{1jk} &\sim Multinomial(n_1, p_{1jk}) \\ x_{0jk} &\sim Multinomial(n_0, p_{0jk}) \end{aligned} \tag{4.4}$$

where

$$\begin{aligned} \sum_{j,k} p_{1jk} &= 1 \\ \sum_{j,k} p_{0jk} &= 1 \end{aligned} \tag{4.5}$$

As illustrated in Muckerjee (2008) [62], Let $OR_{10} = \frac{p_{000}p_{110}}{p_{010}p_{100}}$ denotes the odds ratio associated with the genotype for the non-exposed observations. Similarly, we let $OR_{01} = \frac{p_{000}p_{101}}{p_{001}p_{100}}$ denotes the odds ratio for exposed observations. Also, we let $OR_{11} = \frac{p_{000}p_{111}}{p_{011}p_{100}}$ denotes the odds ratio for the observation with the exposure and susceptible genotype to the observation without either risk factor. Therefore, the multiplicative interaction parameter as in Chatterjee (2008) [62] is

$$\phi = \frac{OR_{11}}{OR_{10}OR_{01}} = \frac{p_{001}p_{010}p_{100}p_{111}}{p_{000}} \tag{4.6}$$

	SNP=0		SNP=1		
	EXP=0	EXP=1	EXP=0	EXP=1	Total
Case	x_{100}	x_{101}	x_{110}	x_{111}	n_1
Control	x_{000}	x_{001}	x_{010}	x_{011}	n_0

Table 4.2: Data structure for the case-control study with a dichotomous genotype (SNP) and a dichotomous environmental factor (EXP).

Let β denotes the multiplicative interaction parameter as $\beta = log(\phi)$. Then the unbiased estimator of this parameter in the case-control studies will be

$$\hat{\beta}_{CC} = log(\frac{x_{001}x_{010}x_{100}x_{111}}{x_{000}x_{011}x_{101}x_{110}}) \tag{4.7}$$

Also, for the multiplicative interaction parameter β, there exists

$$\beta_{CC} = log(\frac{p_{001}p_{010}p_{100}p_{111}}{p_{000}p_{011}p_{101}p_{110}}) = log(\frac{p_{100}p_{111}}{p_{101}p_{110}}/\frac{p_{000}p_{011}}{p_{001}p_{010}}) \tag{4.8}$$

Therefore, if in the control group the genotype and exposure do not interact,

$$\beta_{CO} = log(\frac{p_{100}p_{111}}{p_{101}p_{110}}) \tag{4.9}$$

So in the case, the estimator for β under the conditional of being independent in the control group will be $\hat{\beta}_{co}$:

$$\hat{\beta}_{CO} = log(\frac{x_{100}x_{111}}{x_{101}x_{110}}) \tag{4.10}$$

Then the difference between β_{CC} and β_{CO} is the criteria to evaluate the independency assumption in the general population

$$\begin{aligned} \theta_{GE} &= \beta_{CO} - \beta_{CC} \\ \hat{\theta}_{GE} &= \hat{\beta}_{CO} - \hat{\beta}_{CC} \\ &= log(\frac{x_{000}x_{011}}{x_{001}x_{010}}) \end{aligned} \tag{4.11}$$

Then they proposed a new Bayesian empirical shrinkage-type estimator $\hat{\beta}_{EB}$ which is the weighted average of the estimator $\hat{\beta}_{CC}$ under the case-control studies and the biased estimator $\hat{\beta}_{CO}$ [62]:

$$\hat{\beta}_{EB} = \frac{\hat{\sigma}^2_{CC}}{\theta^2_{GE} + \hat{\sigma}^2_{CC}}\hat{\beta}_{CC} + \frac{\hat{\theta}^2_{GE}}{\theta^2_{GE} + \hat{\sigma}^2_{CC}}\hat{\beta}_{CO} \tag{4.12}$$

where all the parameter estimates are based from table 4.3. $\theta_{GE} = \frac{p_{000}p_{011}}{p_{001}p_{010}}$ and $\hat{\sigma}^2_{CC}$ is the standard error of the estimator of β_{CC}. By taylor's approximation, they generate the variance of the Bayesian empirical shrinkage-type estimator $\hat{\beta}_{EB}$. Intuitively, as the data provide evidence about the independence between the genetic and environmental factor in the controls $\hat{\theta}_{GE} \to 0$, $\hat{\beta}_{EB} \to \hat{\beta}_{CO}$. When the independence is violated and as the sample size increase $\hat{\sigma^2_{CC}} \to 0$, then $\hat{\beta}_{EB} \to \hat{\beta}_{CC}$.

4.3.2 Methodology

In the previous study, suppose we only have two study. $\hat{p}_1 = \frac{x_1}{n_1}$ is the MLE of the true rate p_1. If we assume the exchangeability of the historical study and current study we can derive an estimate of p_1 as $\hat{p}_{pool} = \frac{x_1+x_0}{n_1+n_0}$. So we can regard p_{pool} as the restricted MLE.

	Event	Total	Rate
Current	x_1	n_1	$\frac{x_1}{n_1}$
Historical	x_0	n_0	$\frac{x_0}{n_0}$
Pool	$x_1 + x_0$	$n_1 + n_0$	$\frac{x_0+x_1}{n_0+n_1}$

Table 4.3: Data structure for borrowing data from single historical study

We define the difference between $\hat{p}_1$ and $\hat{p}_{pool}$ as $\hat{\gamma} = \hat{p}_{pool} - \hat{p}_1$, which is an estimator of true difference $p_{pool} - p_1$. We propose the following Bayes empirical shrinkage-type estimator for p_1 as:

$$\hat{p}_{EB} = \frac{var(\hat{p}_1)}{var(\hat{p}_1) + \hat{\gamma}^2} \times \hat{p}_{pool} + \frac{\hat{\gamma}^2}{var(\hat{p}_1) + \hat{\gamma}^2} \times \hat{p}_1$$

4.3.3 Asymptotic Property

This estimator is a kind of estimator that weighs adaptively on the MLE and restricted MLE. The weigh depends on the estimation property of p_1 and the difference between the two estimators.

- As $\hat{\gamma}^2 \to 0$, the event rate of the two studies getting closer, $\hat{p}_{EB} \to \hat{p}_{pool}$
- As $\hat{\gamma}^2 \to \infty$, the event rate of the two studies getting larger, $\hat{p}_{EB} \to \hat{p}_1$
- As $n_1 \to 0$, then $var(\hat{p}_0) \to \infty$, then $\hat{p}_{EB} \to \hat{p}_{pool}$
- As $n_1 \to \infty$, then $var(\hat{p}_0) \to 0$, then $\hat{p}_{EB} \to \hat{p}_1$

We set up a simulation to study the property of the estimator. In Figure 4.1, we plot the results for 4 different scenarios based on the different sample size. We fixed the true value of p_1 at 0.2. We use $\Delta = p_{pool} - p_1$ and change the value of Δ to get the true value of p_0 as

$$\frac{(p_1 + \Delta) \times (n_1 + n_0) - p_1 n_1}{n_0}$$

Therefore, we run an 1000 times simulation and under each simulated data, we calculated mean absolute error (MAE) by $\frac{1}{1000} \sum |\hat{p}_{EB} - 0.2|$ for the empirical Bayesian estimator $\frac{1}{1000} \sum |\hat{p}_1 - 0.2|$ for the current study estimate and $\frac{1}{1000} \sum |\hat{p}_{pool} - 0.2|$ for the pooled study.

In the first study with sample size $n_1 = 50$ for current study and $n_0 = 50$ for historical study. We change the setup in the difference between the current study

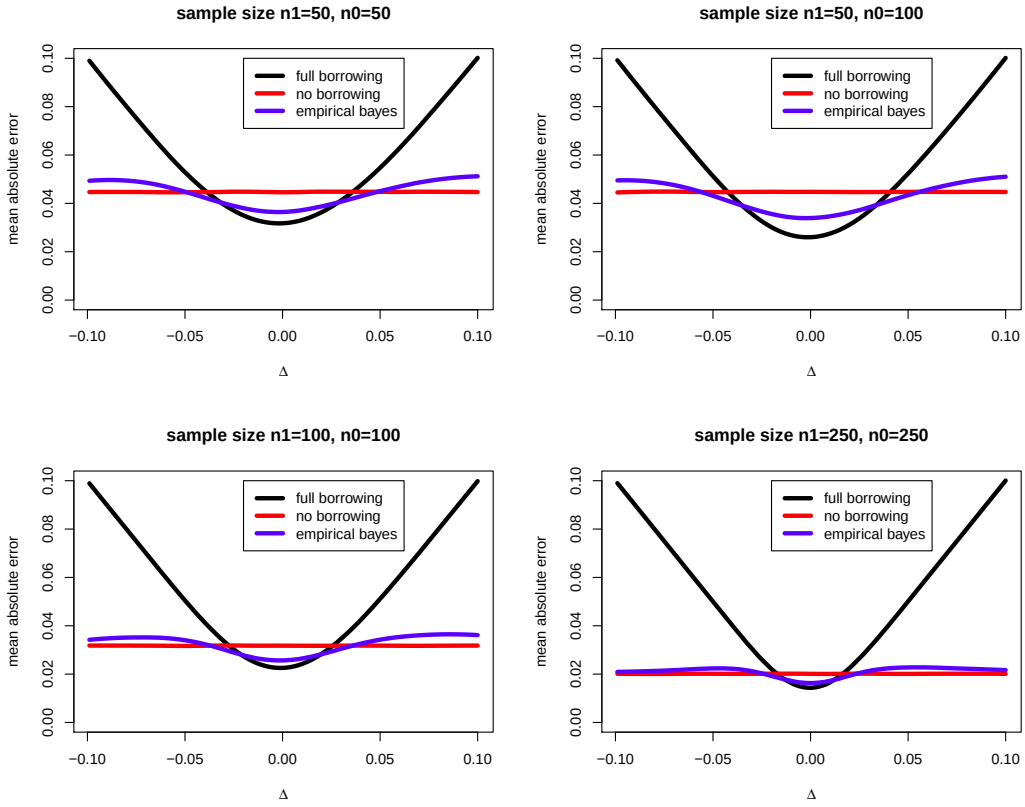

Figure 4.1.: Simulation results for comparing the borrowing performances

rate and pool study rate while controlling the event rate of current study at fixed value 0.2. We want to examine the estimator shrinkage property when changing the difference between these two estimators.

We can observe that as the difference Δ is around 0, meaning that the two studies are very alike, the empirical Bayesian estimator $\hat{p}_{EB}$ shrink to the pool estimator $\hat{p}_{pool}$. While the Δ getter larger, the shrinkage estimator will be more close to the single study estimator $\hat{p}_1$. As we increase the sample size in the historical study to 100, $\hat{p}_{EB}$ shrink more to the pool estimator. When we increase the sample size of the current study, all estimators get improved while the empirical estimator shrink more to the current study when the difference get larger. And as we increase both sample sizes

of the two studies, the empirical estimator perform adaptively better. It means that the empirical Bayes estimator only shrink to the pool study estimator p_{pool} when the two study rates are very close. It is true for the practical meaning since we already have large sample size in the current study.

4.3.4 Inference

As derived similarly as in Mukerjee (2008) [62] [63], we derive the approximation of the estimate variance of p_{EB}.

$$Var(\hat{p}_{EB}) = Var(\hat{p}_{pool}) + (\frac{\hat{\gamma}^2(\hat{\gamma}^2 + 3Var(\hat{p}_1))}{(Var(\hat{p}_1) + \hat{\gamma}^2)^2})^2 \times Var(\hat{\gamma})$$

where,

$$Var(\hat{p_1}) = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1}$$

$$Var(\hat{p}_{pool}) = \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_1+n_0}$$

$$Var(\hat{\gamma}) = \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_1+n_0} + \hat{p}_1(1-\hat{p}_1) \times (\frac{1}{n_1} - \frac{2}{n_1+n_0})$$

We can construct the 1-sided Wald confidence interval $(-\infty, \hat{p}_{EB} + Z(1-\alpha) \times se(\hat{p}_{EB}))$ to test the null hypothesis $p_1 > 0.2$.

4.4 Bayesian Model Averaging and Bayesian Model Selection

4.4.1 Motivating Background

In the previous section, we have derived the Bayesian empirical shrinkage-type estimator for borrowing the historical data in device trials. The current study data provides the unbiased estimate, while the combined study provides larger power depending on the homogeneity. This is analogous to the bias-variance dilemma in gene-environment interaction studies. The criteria for weighting the estimate in the

gene-environment interaction is the independency assumption between genetic and environmental factor. In the borrowing historical data study, the criteria is the homogeneity assumption. This borrowing process is a continuous process by which the empirical estimate will change smoothly. The Bayesian model averaging is another approach that has been developed for taking the uncertainty in the model comparison/selection for the estimation. There is a posterior probability assigned to each model, which will play parts in the final model averaging process.

The Bayesian Model Averaging approach was initially proposed by Raftery (1997) [52] for the linear regression model, in which the uncertainty when selecting the model among all possible is considered. The BMA approach was also applied in the case-control studies by Raftery (2001) [64] to investigate the generalized linear model with interactions. Li and Conti (2009) [65] extend the BMA approach to the gene-environment interactions studies by balancing the power and bias among the case-only and case-control studies. They proposed the Bayesian model averaging approach to combine case-control and case-only studies.

Suppose that we have a case-control dataset with a disease status variable Z, a binary genetic factor X and an environmental factor Y. The interaction between the genetic factor and environmental factor could be tested using the logistic regression model:

$$\begin{aligned} Z &\sim bernoulli(P) \\ logit(P) &= \alpha + \beta X + \gamma Y + \theta XY \end{aligned} \tag{4.13}$$

Then by 4.13,

$$log(N * Pr(Z = 1)) - log(N * Pr(Z = 0)) = \alpha + \beta X + \gamma Y + \theta XY \tag{4.14}$$

where N is the total number. Therefore, the model can be viewed as a log-linear model, in which the logarithm of the expected number of observations in each cell can also be modeled in [66] as

$$log(\mu|Z,X,Y) = \alpha_0 + \beta_0 X + \gamma_0 Y + \theta_0 XY + \alpha Z + \beta ZX + \gamma ZY + \theta ZXY \quad (4.15)$$

And 4.13 and 4.15 have the same parameters α, β, γ and θ.

From 4.15, we can have the following equations:

$$\begin{aligned} log(\mu|Z=0,X=1,Y) &= \alpha_0 + \gamma_0 Y + \theta_0 Y \\ log(\mu|Z=0,X=0,Y) &= \alpha_0 + \gamma_0 Y \end{aligned} \quad (4.16)$$

So under the assumption of independence, in the control date $log(\mu|Z=0,X=1,Y) = log(\mu|Z=0,X=0,Y)$, therefore $\theta_0 = 0$. Under this assumption,

$$\begin{aligned} log(\mu|Z=1,X=1,Y) &= \alpha_0 + \beta_0 + \gamma_0 Y + \alpha + \beta + \gamma Y + \theta Y \\ log(\mu|Z=1,X=0,Y) &= \alpha_0 + \gamma_0 Y + \alpha + \beta + \gamma Y \end{aligned} \quad (4.17)$$

And the interaction can be evaluated based on the model in the case-only study as

$$\begin{aligned} X &\sim bernoulli(\pi) \\ logit(\pi) &= \beta_0 + \theta Y \end{aligned} \quad (4.18)$$

So here the case-control estimator of the interaction parameter is θ_{CC} is under the model 4.13 and the case-only estimator of the interaction parameter is θ_{CO} under the model 4.18. And all these estimations are under the framework of 4.15. (Li and Conti) [65] then applied the Bayesian model averaging (BMA) approach to combine these two parameters according to different models. Model averaging has been proposed to account for uncertainty brought in by different models. Unlike the Model Selection, the model averaging explicitly presents the uncertainty of models and make inferences

based on all competing models rather the best model. Here, θ is the parameter of interest for evaluating the gene-environment interactions. Let M_1 denote the model 4.13 as M_1 and 4.18 as M_2 and the BMA estimator θ_{BMA} would be

$$\begin{aligned} \theta_{BMA} &= Pr(M_1|data) \times \theta_{CC} + Pr(M_2|data) \times \theta_{CO} \\ Pr(M_k|data) &= \frac{Pr(data|M_k)Pr(M_k)}{\sum Pr(data|M_k)Pr(M_k)} \\ Pr(data|M_k) &= \int Pr(data|\theta_k) \times f(\theta_k|M_k)d\theta_k \end{aligned} \tag{4.19}$$

Also, the variance of the parameters could be measured and then the Wald test could be applied for the hypothesis test.

4.4.2 Methodology

First of all, we will consider the case of only two studies and then generate the approach to multiple studies. Our hypothesis is that

$$H_0 : p_1 = p_0$$

$$H_a : p_1 \neq p_0$$

So if we believe H_0 is right, we will pool the two study together. Otherwise we will only use the current study to estimate p_1. We have to specify our prior belief about H_0 and H_a according to $Pr(H_0) + Pr(H_a) = 1$. Then we can calculate the posterior of each model as:

$$p(H_0|data) = \frac{p(H_0) \times p(data|H_0)}{p(H_0) \times p(data|H_0) + p(H_a) \times p(data|H_a)}$$

where we can assume $Pr(H_0) = Pr(H_a) = \frac{1}{2}$ that reflects an non-informative belief about the parameters. Here, 'data' refers to the patients data in both studies. After some integration, we can get the closed form of the posterior of each model as

$$p(H_0|data) = \frac{Beta_{10}}{Beta_1 \times Beta_0}$$

where,

$$Beta_{10} = Beta(x_1 + x_0 + 1, n_1 + n_0 - x_1 - x_0 + 1)$$

$$Beta_1 = Beta(x_1 + 1, n_1 - x_1 + 1)$$

$$Beta_0 = Beta(x_0 + 1, n_0 - x_0 + 1)$$

And

$$p(H_a|data) = 1 - p(H_0|data)$$

Therefore, we can get the posterior of the $f(p_1|data)$ as

$$f(p_1|H_0, data) \times p(H_0|data) + f(p_1|H_a, data) \times p(H_a|data)$$

where $f(p_1|H_0, data)$ is the posterior of p_1 under the model $H_0 : p_0 = p_1$ and $f(p_1|H_a, data)$ is the posterior of p_1 under the model $H_a : p_0 \neq p_1$. So the posterior of $p_1|data$ is a weighted mixture of two parts. As illustrated in Hoeting (1999) [53], if we want to get the point estimate of p_1, the BMA estimate denoted as $\hat{p}_1^{BMA}$ should be

$$\hat{p}_1^{BMA} = E(p_1|x_1, x_0)p(H_0|data) + E(p_1|x_1)p(H_a|data)$$

In Figure 4.2, we plot the distribution of $p_1|data$ under different situations. If we observed 3 events among 20 patients in the current study and 14 events among 30 patients in the historical study, the posterior distribution will be approximately the same as the distribution of the single study. If we observed fewer events in the historical data, the posterior is approaching the historical study and stay between the single study and the pool study. Finally suppose we observe 5 events among 30 patients in the historical study, the posterior shrinks to the pooled distribution. So the distribution of the proposed estimate of p_1 has the property of adaptively deciding the borrowing information based on the similarities between the current study and historical study. Since we can draw the posterior samples for $p_1|data$ by Monte Carlo approach, we can do hypothesis test based on the sampled posterior distribution. For example, if $Pr(p_1 < 0.2|data) > P_{cut}$ we reject H_0.

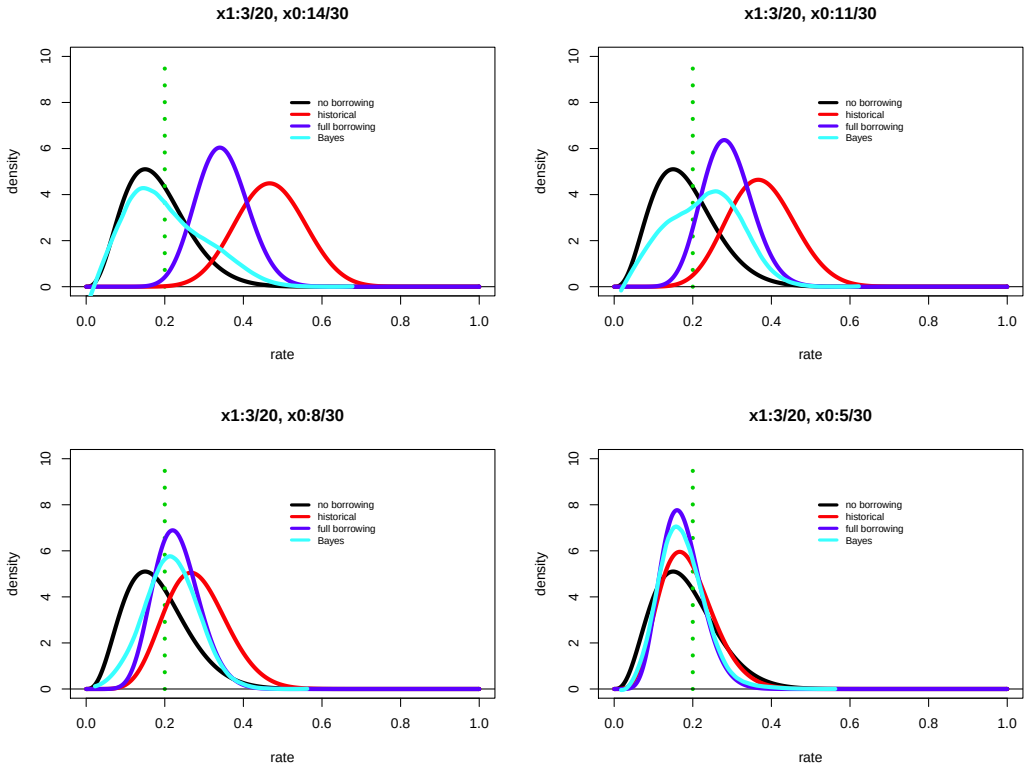

Figure 4.2.: Posteiror distributions of the estimator according to degress of homogeneity

4.4.3 Simulation

Here we run 1000 simulations to compare the 'Empirical Bayes', 'Bayesian Model Averaging' (BMA) and frequentist Binomial exact test with fully borrowing and no borrowing approach for the event rate test. So we define

$$H_0 : p_1 \geq 0.2$$

$$H_a : p_1 < 0.2$$

We compare these approaches based on the Frequentist Operating Characteristics. We change the true current event rate and historical event rate and calculate the power or Type-I error based on the setup.

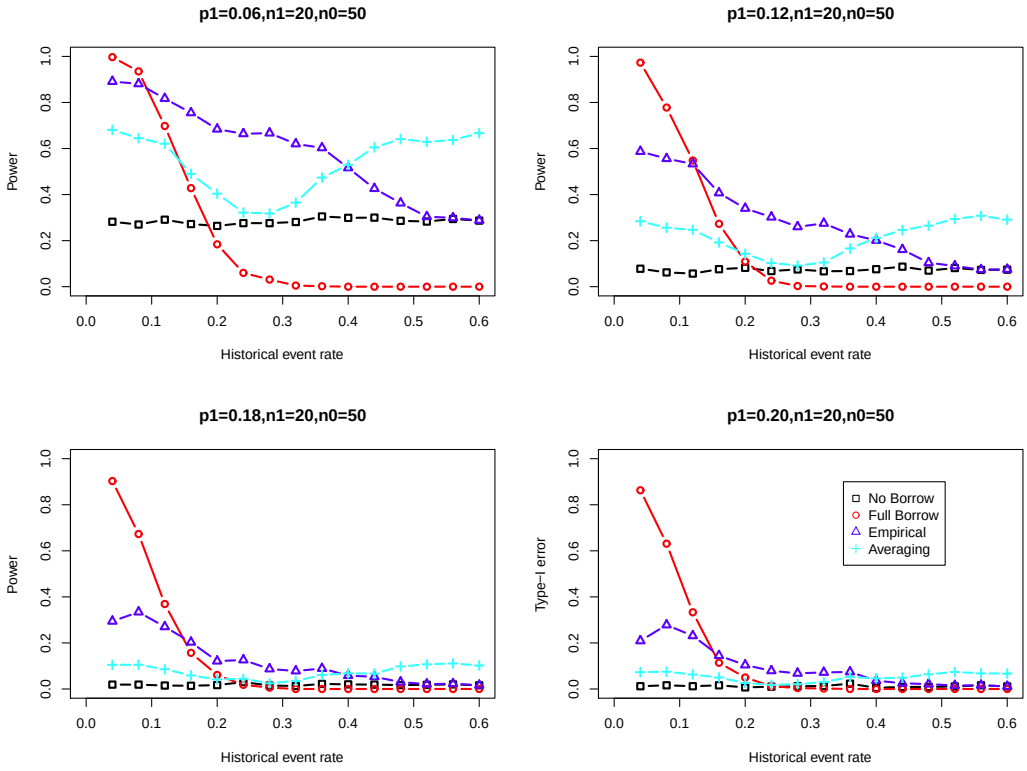

Figure 4.3.: Frequentist operating characteristics for Scenario 1 with 20 observations in the current study and 50 observations in the historical study

4.4.4 Results

In Figure 4.3, when the current study sample size is 20 and historical is 50, both of the empirical Bayes approach and BMA outperform the full borrowing and no borrowing approaches. The empirical Bayes approach has a larger power than the BMA approach, but produce a higher level of Type-I error. When we increase the sample size from 20,50 to 50,100 as in Figure 4.4, the performance of all models get improved. The BMA shrink more obviously to the current study as the similarity between the current and pool get weak. When we have a relative larger sample size as in Figure 4.5, the graph shows that the two approaches proposed outperform the no-borrowing approach. When the true rate of the current study is in the same area

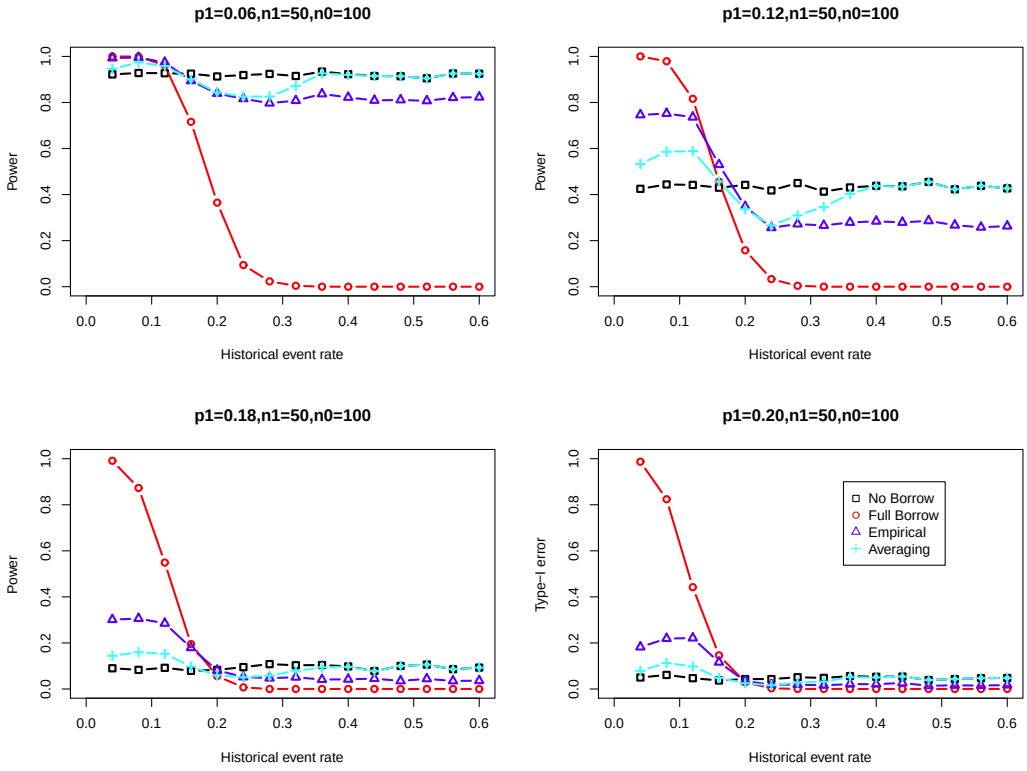

Figure 4.4.: Frequentist operating characteristics for Scenario 1 with 50 observations in the current study and 100 observations in the historical study

of the alternative hypothesis, the methods effectively borrow information to increase the power. In Figure 4.6, since the sample size is already large enough to detect the difference, the two approaches proposed perform similar as the no borrowing approach.

Overall, the empirical Bayesian approach and Bayesian model averaging approach has the adaptive property that can borrow the information from the historical study information to improve the frequentist operating characteristics.

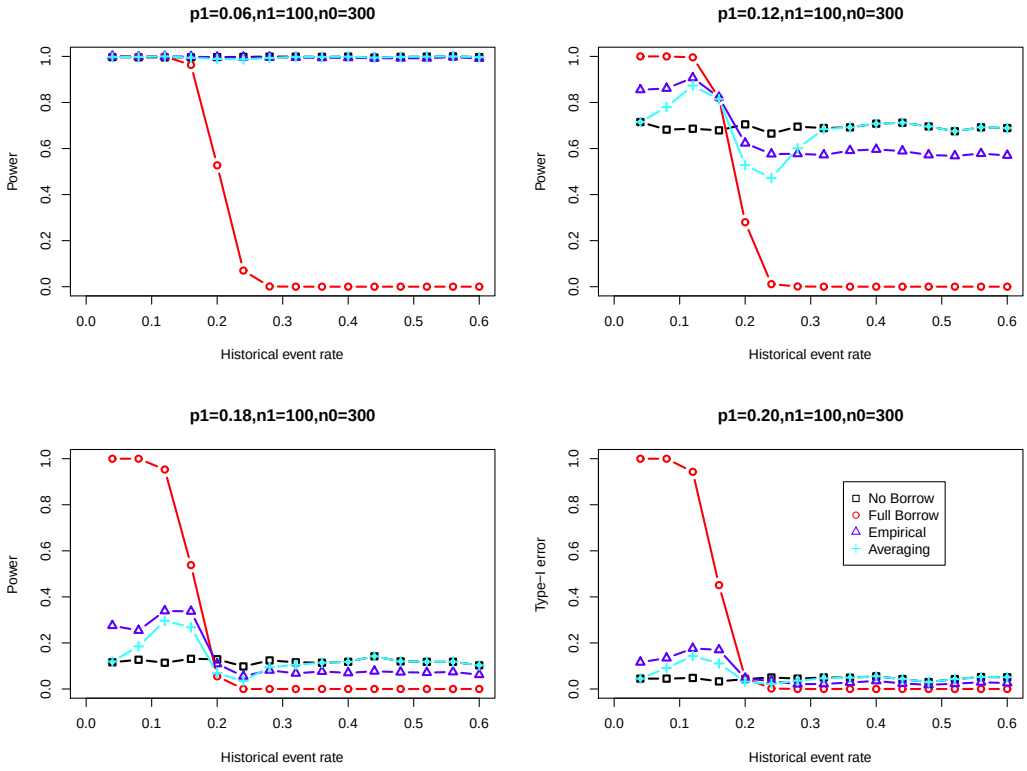

Figure 4.5.: Frequentist operating characteristics for Scenario 1 with 100 observations in the current study and 300 observations in the historical study

4.4.5 Alternative Extensions

Multiple Study/Arm BMA

The BMA approach has the property to be extended to multiple studies. Now suppose we have 3 studies with their event rate as p_0, p_1 and p_2, in which p_2 is the current rate for the current study. Then we can make a group of hypothesis,

$$H_1 : p_0 = p_1, p_0 = p_2, p_1 = p_2$$

$$H_2 : p_0 = p_1, p_0 \neq p_2, p_1 \neq p_2$$

$$H_3 : p_0 \neq p_1, p_0 = p_2, p_1 \neq p_2$$

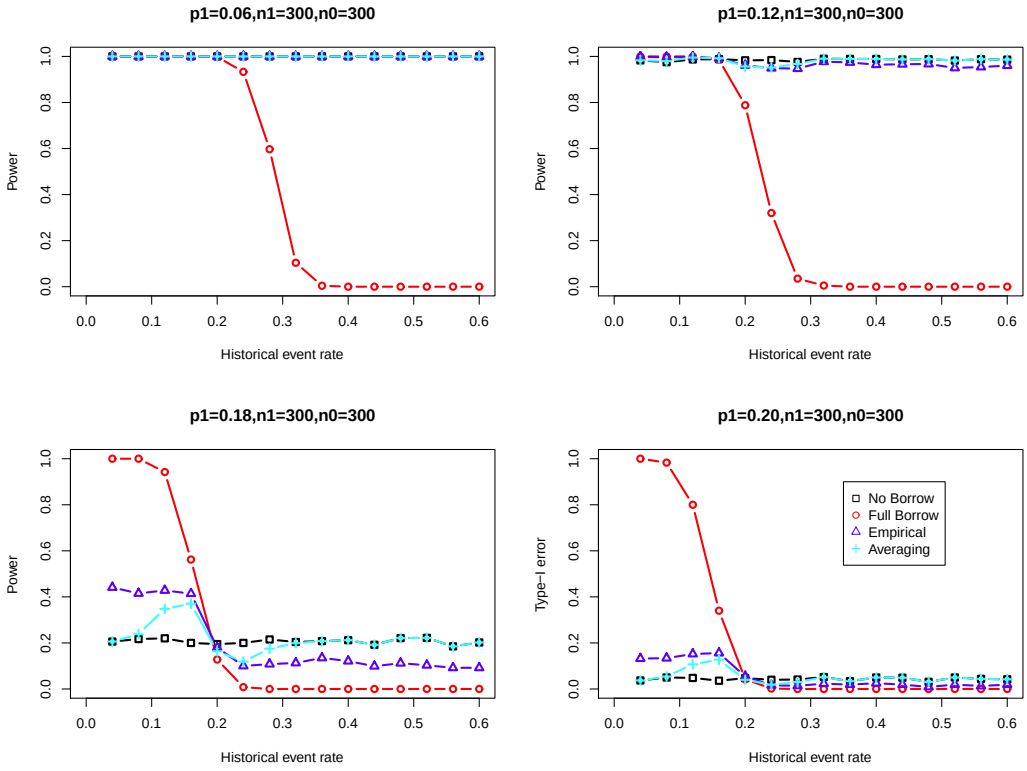

Figure 4.6.: Frequentist operating characteristics for Scenario 1 with 300 observations in the current study and 300 observations in the historical study

$$H_4 : p_0 \neq p_1, p_0 \neq p_2, p_1 = p_2$$

$$H_5 : p_0 \neq p_1, p_0 \neq p_2, p_1 \neq p_2$$

Therefore, we can simulate get the posterior $p_2|data$, which is a finite mixture with all $H_i|data,\ i = 1, 2, 3, 4, 5$. One unique feature of this approach is that it can fully consider all the possibility of the rations between the different studies. As illustrated in below, the 'commensurate' will assume the homogeneity among the historical groups. But in here, we do not need that assumption.

Equivalence Margin

Also, based the on original hypothesis, we may set up the hypothesis as:

$$H_0 : |p_1 - p_0| > \delta$$

$$H_a : |p_1 - p_0| \leq \delta$$

So compared the original BMA model, this model's focus on a region of hypothesis rather than a diagonal line for the null hypothesis. This set-up provides us some flexibility to measure the difference between the event rate. Also, it has clinical meaning that the physicians can assign clinical meaningful margin into the study. The calculation will be more complex. We have tried the simplest two studies cases, it shows that the inference can be improved than the original BMA approach.

Bayesian Decision-Theoretic Optimal Boundary

In the last section, we introduced the concept of equivalence margin. For the hypothesis in the Bayesian model Averaging

$$\begin{aligned} T_0 &: p_1 - p_0 = 0 \\ T_a &: p_1 - p_0 \neq 0 \end{aligned} \tag{4.20}$$

We can borrow the equivalence trial test idea to evaluate the plausibility to 'pool' the data together or not by introducing

$$\begin{aligned} T_0 &: |p_1 - p_0| \geq \delta \\ T_a &: |p_1 - p_0| < \delta \end{aligned} \tag{4.21}$$

where δ is the equivalence margin, by which we want to pool the data together if T_a is true and vice versa. This is similar as equivalence hypothesis test.

As for the most simple case, p_0 is assumed known without uncertainty(large historical study) and the sample size of our current study is n_1. We use composite single point hypothesis as

$$\begin{aligned} T_0 &: p_1 = p_0 \\ T_{a1} &: p_1 = p_0 - \delta \\ T_{a2} &: p_1 = p_0 + \delta \end{aligned} \tag{4.22}$$

Same as before we assume that the three hypothesis are equally likely apriori

$$p(T_0) = p(T_{a1}) = p(T_{a2}) = 1/3 \tag{4.23}$$

Here we propose to use the Bayesian decision-theoretic framework to find the 'optimal' solution about deciding if we should pool the data(BMS). We want to find the lower bound ϕ_L and upper bound ϕ_U of the event rate, such that when the observed event number x_1 fulfills $n_1\phi_L < x_1 < n_1\phi_U$ we pool the data otherwise we will not pool. We define a loss function as the probability of making the wrong decision (pool or not)

$$\begin{aligned} L(\phi_L, \phi_U) &= p(x_1 \geq n_1\phi_U \bigcup x_1 \leq n_1\phi_L | T_0) \times p(T_0) \\ &\quad + p(n_1\phi_L < x_1 < n_1\phi_U | T_{a_1}) \times p(T_{a_1}) \\ &\quad + p(n_1\phi_L < x_1 < n_1\phi_U | T_{a_2}) \times p(T_{a_2}) \end{aligned} \tag{4.24}$$

So for each given p_0, n_1 and δ, we can solve ϕ_L and ϕ_U by minimizing $L(\phi_L, \phi_U)$. Then the optimal bound solution is:

$$\begin{aligned} \hat{\phi_L} &= \underset{\phi_L}{argmin} \; p(T_0)F_{bin}(x_i \leq n1\phi_L; n_1, p_0) - p(T_{a_1})F_{bin}(x_i \leq n1\phi_L; n_1, p_0 - \delta) \\ &\quad - p(T_{a_2})F_{bin}(x_i \leq n1\phi_L; n_1, p_0 + \delta) \\ \hat{\phi_U} &= \underset{\phi_U}{argmin} \; p(T_0)(1 - F_{bin}(x_i < n1\phi_U; n_1, p_0)) + p(T_{a_1})F_{bin}(x_i < n1\phi_U; n_1, p_0 - \delta) \\ &\quad + p(T_{a_2})F_{bin}(x_i < n1\phi_U; n_1, p_0 + \delta) \end{aligned} \tag{4.25}$$

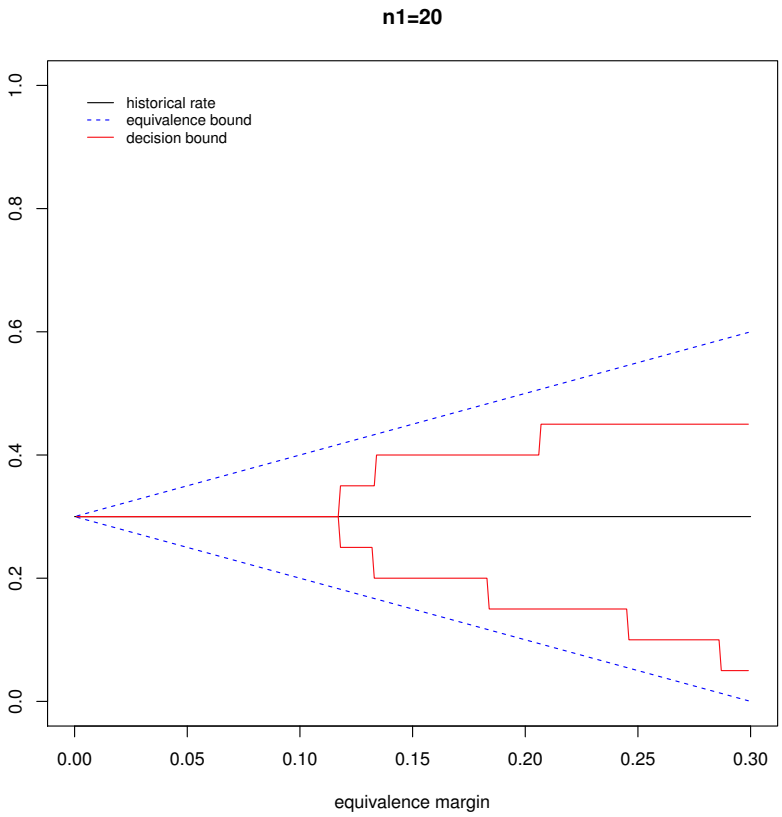

Figure 4.7.: Decision boundary for different equivalence margins with 20 observations in the current study

δ	0.05	0.10	0.15	0.20	0.25
No.of events	6	6	$[5, 7]$	$[4, 7]$	$[3, 8]$

Table 4.4: Decision boundary for different equivalence margins with 20 observations in the current study

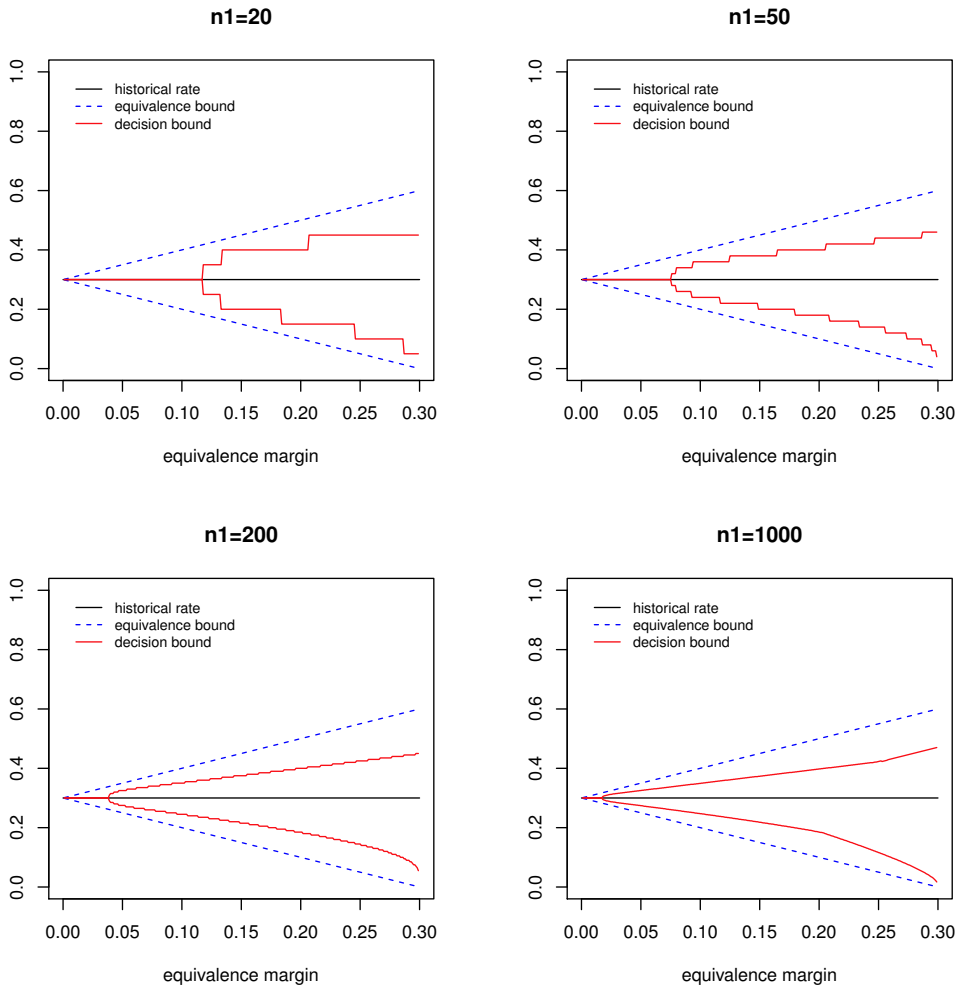

Figure 4.8.: Decision boundary for different equivalence margins with different numbers of observations in the current study

Predictive Probability

All our previous Bayesian hypothesis testing is based on the calculation of the posterior probability of certain claim. If the probability is larger than some cut-off point, we will reject or accept the corresponding hypothesis. Next we will introduce the general idea of predictive probability approach and use a simple example to illustrate the difference from posterior probability approach.

Suppose that for single arm trial, we have observed y_1 events out of n_1 patients. In **posterior probability** we will reject $H_0 : p_1 > 0.2$ if

$$Pr(p_1 < 0.2|y_1) > 0.95$$

Now suppose we have total another n_2 patients who will be treated by the current arm. So similarly the posterior probability we will reject H_0 based on all the data is

$$Pr(p_1 < 0.2|y_1, y_2) > 0.95$$

However, we actually have not observed y_2, we can regard the above as a random variable taking value 0 or 1 depending on the value of y_2 .Then we need to average out y_2 from the above conditional probability with the **predictive posterior probability**, which follows a beta-binomial models.

$$f(y_2|y_1) \sim Beta - Binomial(n_2, \alpha + y_1, \beta + n_1 - y_1)$$

Then we want to get our **predictive probability** as

$$\sum_{y_2=1}^{n_2} I(Pr(p_1 < 0.2|y_1, y_2) > 0.95) \times f(y_2|y_1)$$

So we can regard the posterior probability model as the direct evidence to make our conclusion. The predictive probability model is a finite mixture of each posterior model under different potential future outcomes within the total sample size.

There are comprehensive introductions about these ideas and practical examples for working on. The idea of predictive probability model can fulfill the purpose of continuous/group monitoring the study observations, since the results depend on the future sample size.

4.5 Bayesian Commensurate Model

4.5.1 Introduction

In all the approaches mentioned before, we mainly consider the borrowing information between two studies based on their end point only. For example, we focus on evaluating the event rate of the studies. In practice, the trial is always complicated and have factors influencing the results. So we need some tool to systematically incorporate all those kinds of factors into the analysis. In Bayesian hierarchical model section, we have introduced a model based on the binomial likelihood function. The model is actually can be equally written as:
For the current study, we assume

$$y_{1i} \sim bernoulli(p_{1i}), \quad i = 1, ..., n_1$$

$$logit(p_{1i}) = \alpha_1$$

$$\alpha_1 \sim N(\mu_a, 1/\tau_a)$$

$$\mu_a \sim N(a, b)$$

$$\tau_a \sim Gamma(c, d)$$

For the historical study, we assume

$$y_{0j} \sim bernoulli(p_{0j}), \quad j = 1, ..., n_0$$

$$logit(p_{0i}) = \alpha_0$$

$$\alpha_0 \sim N(\mu_a, 1/\tau_a)$$

So here we start to model based on each individual level's data. Since each individual data has their specific covariate level or treatment level, then we can incorporate those effects in the model. First, we will consider to incorporate one covariate effect into the current study. Then the likelihood of the models will become:
For the current study, we assume

$$y_{1i} \sim bernoulli(p_{1i}), \quad i = 1, ..., n_1$$

$$logit(p_{1i}) = \alpha_1 + \beta_1 \times x_{1i}$$

For the historical study, we assume

$$y_{0j} \sim bernoulli(p_{0j}), \quad j = 1, ..., n_0$$

$$logit(p_{0i}) = \alpha_0 + \beta_0 \times x_{0j}$$

where, β_1 and β_0 are the covarate effect parameters.

4.5.2 Methodology

The random effect meta-analysis model is the most commonly used approach for modeling the results from multiple studies. It generally consider all the corresponding effect parameters are coming a common distribution. By controlling the variance of this distribution, we can control how the information can be borrowed between different studies. So for the above example, the meta-analysis model will regard the study effects α_1 and α_0 are coming from the same distribution and β_1 and β_0 from another common distribution. For example,

$$\alpha_1, \alpha_0 \sim N(\mu_a, \tau_a)$$

$$\beta_1, \beta_0 \sim N(\mu_b, \tau_b)$$

So in the above models, we can control the effect parameters between the studies by tuning τ_a and τ_b. If $\tau_a \to 0$ meaning small variances, the study effect parameter of α_1 will borrow from the data that influence the parameter α_0.

In Hobbs (2011, 2012) [67] [68], a novel approach is proposed to more directly describe the relations between the different studies by 'commensurate' model. In the above model, the meta-analysis assumes the parameters are coming from a common distribution, while 'commensurate' model assumes that one parameter is a non-systematically biased representation of another parameter. The model is like this

$$\alpha_0 \sim N(\mu_{a_0}, \tau_{a_0})$$

$$\alpha_1 \sim N(\alpha_0, \gamma_a)$$

$$\beta_0 \sim N(\mu_{b_0}, \tau_{b_0})$$

$$\beta_1 \sim N(\beta_0, \gamma_b)$$

As $\tau_a \to 0$, $\alpha_1 \to \alpha_0$ which means that the two parameters are very close. $\tau_a \to \infty$, α_1 and α_0 are independent. The 'commensurate' means that the two parameters are not that different, so the idea sounds like the BMA approach which construct the model based on the closeness between the studies' endpoint values. Now if we have more than 1 historical study to borrow, the model for the two study will be extended. Then it apply the homogeneity model to model all the historical data set and model the parameter of the current based on the parameter from the historical parameters. Next we will first take a look at an example of two studies with one covariate.

4.5.3 Case 1: Two Studies with One Covariate in Each Study

Methodology

Here, suppose that we have two studies with responses y_{0i} and y_{1j} that are binary data for each patient. We assume that we will have diabetic patient in the current study, such that one covariate is considered in the current study. The full model of the **commensurate** model is as follows:
For the current study, we assume

$$y_{1i} \sim bernoulli(p_{1i}), \quad i = 1, ..., n_1$$

$$logit(p_{1i}) = \alpha_1 + \beta_1 \times x_{1i}$$

$$\alpha_1 \sim N(\alpha_0, \gamma_a)$$

$$\beta_1 \sim N(\beta_0, \gamma_b)$$

For the historical study, we assume

$$y_{0j} \sim bernoulli(p_{0j}), \quad j = 1, ..., n_0$$

$$logit(p_{0i}) = \alpha_0$$

$$\alpha_0 \sim N(\mu_{a_0}, \tau_{a_0})$$

We want to evaluate the covariate effect parameter β_1 with the study effect α_1. As mentioned earlier, the random-effect **meta-analysis** model is:
For the current study, we assume

$$y_{1i} \sim bernoulli(p_{1i}), \quad i = 1, ..., n_1$$

$$logit(p_{1i}) = \alpha_1 + \beta_1 \times x_{1i}$$

For the historical study, we assume

$$y_{0j} \sim bernoulli(p_{0j}), \quad j = 1, ..., n_0$$

$$logit(p_{0i}) = \alpha_0 + \beta_0 \times x_{0j}$$

$$\alpha_1, \alpha_0 \sim N(\mu_a, \tau_a)$$

$$\beta_1, \beta_0 \sim N(\mu_b, \tau_b)$$

For comparison reason, we also consider the **single study** model with the current study group only,

$$y_{1i} \sim bernoulli(p_{1i}), \quad i = 1, ..., n_1$$

$$logit(p_{1i}) = \alpha_1 + \beta_1 \times x_{1i}$$

$$\alpha_1 \sim N(\mu_a, \tau_a)$$

$$\beta_1 \sim N(\mu_b, \tau_b)$$

In the last, a **homogeneity** model is considered for comparison:

$$y_{1i} \sim bernoulli(p_{1i}), \quad i = 1, ..., n_1$$

$$logit(p_{1i}) = \alpha + \beta \times x_{1i}$$

$$y_{0j} \sim bernoulli(p_{0j}), \quad j = 1, ..., n_0$$

$$logit(p_{0i}) = \alpha$$

$$\alpha \sim N(\mu_a, \tau_a)$$

$$\beta \sim N(\mu_b, \tau_b)$$

Simulation

For comparison reasons, We assume the prevalence of the covariate is 0.5 in the current study. The null hypothesis for the event rate of patient without diabetes is

$$H_0 : p_{10} \geq 0.2$$

The null hypothesis for the event rate of patient with diabetes is

$$H_0^{'} : p_1 \geq 0.3$$

This is not a multiple test case. In Bayesian frame work, if $pr(p_1 < 0.2|data) > 0.95$ we will reject null hypothesis for $pr(p_1^{'} < 0.3|data) > 0.95$. For the parameter setup we will use logit transformation to get the true value for the parameters. For example, for the current study, if we set $p_0 = \mathbf{0.2}$, $p_{11} = 0.3$ and $p_{10} = \mathbf{0.2}$, then

$$\alpha_0 = \alpha_1 = logit(0.2) = -1.386$$

$$\beta = logit(0.3) - logit(0.2) = 0.539$$

if we set $p_0 = \mathbf{0.1}$

$$\alpha_0 = -2.197$$

Results

Table 4.5 and 4.6 summarize the results for analyzing the data from 2 studies with the covariate in the current study. When the sample size is 80 for historical study and 40 for the current study, we find that overall the performance of the power is not appealing. When commensurate model performs slightly better than the meta-analysis approach. If the historical rate p_0 is same as the current rate for the non-diabetic patient, the commensurate and meta-analysis are performing similarly as the best model 'homogeneity' in this situation. However, when the true rate for the historical study is 0.1, the Type-I error rate is unacceptably inflated, although the Type-I error for the commensurate and meta-analysis is also inflated. So overall the commensurate is the best among all the models considered.

When we increase the sample size of the study to $n_0 = 200$ and $n_1 = 100$, we observed that commensurate model perform also the best among all the four models.

When we even increase more sample size on both of the studies, we can see that the Type-I error is better controlled than the single study, while they all have acceptable power. This study is useful because we always want to evaluate the effects on the subgroup of the current arm.

4.5.4 Case 2: Two Studies with One Covariate and One Treatment

Methods

In practical clinical trial studies, we often meet the case in which the current trial is a randomized trial. This is a kind of typical trial scenario in device clinical trial study, because we often has some historical information on one arm and we want to evaluate the treatment effect from the other arm. So we have the model for the data: For the current study, we assume

$$y_{1i} \sim bernoulli(p_{1i}), \quad i = 1, ..., n_1$$

$$logit(p_{1i}) = \alpha_1 + \beta_1 \times x_{1i} + \beta_{trt} \times x_{trt}$$

For the historical study, we assume

$$y_{0j} \sim bernoulli(p_{0j}), \quad j = 1, ..., n_0$$

$$logit(p_{0i}) = \alpha_0 + \beta_2 \times x_{2j}$$

Simulations

By doing simulations, we want to see how the different Bayesian approaches can estimate the parameter of interest. In the setup we assume that there are 100 patients in the current study and 200 in the historical study. In both of the studies we assume that the prevalence of the covariate (diabetic factor) is 0.5 in both. In the current study of randomized trial, we assume that the treatment allocation is 0.5. We

$n_0 = \mathbf{80}$	p_0	0.2		0.2		0.2		0.2	
$n_1 = \mathbf{40}$	p_{10} p_{11}	0.2	0.3	0.2	0.25	0.2	0.2	0.2	0.15
		Type-I	Type-I	Type-I	Power	Type-I	Power	Type-I	Power
Commensurate		0.03	0.05	0.03	0.13	0.02	0.25	0.01	0.49
Meta-analysis		0.04	0.06	0.04	0.14	0.04	0.28	0.04	0.54
Single study		0.18	0.06	0.19	0.14	0.19	0.27	0.17	0.52
Homogeneity		0.04	0.06	0.04	0.14	0.04	0.27	0.04	0.52
	p_0	0.1		0.1		0.1		0.1	
	p_{10} p_{11}	0.2	0.3	0.2	0.25	0.2	0.2	0.2	0.15
		Type-I	Type-I	Type-I	Power	Type-I	Power	Type-I	Power
Commensurate		0.46	0.05	0.47	0.13	0.48	0.25	0.47	0.49
Meta-analysis		0.55	0.06	0.55	0.14	0.54	0.27	0.52	0.51
Single study		0.18	0.06	0.19	0.14	0.19	0.27	0.17	0.52
Homogeneity		0.72	0.06	0.76	0.14	0.76	0.27	0.77	0.52
$n_0 = \mathbf{200}$	p_0	0.2		0.2		0.2		0.2	
$n_1 = \mathbf{100}$	p_{10} p_{11}	0.2	0.3	0.2	0.25	0.2	0.2	0.2	0.15
		Type-I	Type-I	Type-I	Power	Type-I	Power	Type-I	Power
Commensurate		0.01	0.05	0.01	0.17	0.01	0.47	0.01	0.81
Meta-analysis		0.03	0.06	0.03	0.18	0.03	0.49	0.03	0.82
Single study		0.08	0.07	0.07	0.20	0.09	0.54	0.09	0.85
Homogeneity		0.05	0.06	0.05	0.20	0.04	0.53	0.04	0.84
	p_0	0.1		0.1		0.1		0.1	
	p_{10} p_{11}	0.2	0.3	0.2	0.25	0.2	0.2	0.2	0.15
		Type-I	Type-I	Type-I	Power	Type-I	Power	Type-I	Power
Commensurate		0.55	0.05	0.53	0.17	0.51	0.47	0.52	0.79
Meta-analysis		0.56	0.05	0.54	0.16	0.53	0.48	0.49	0.79
Single study		0.08	0.07	0.07	0.20	0.09	0.54	0.09	0.85
Homogeneity		0.97	0.06	0.97	0.20	0.97	0.54	0.97	0.85

Table 4.5: Frequentist operating characteristics for different models with different sample sizes in the historical and current studies

$n_0 = \mathbf{500}$	p_0	0.2		0.2		0.2		0.2	
$n_1 = \mathbf{300}$	p_{10} p_{11}	0.2	0.3	0.2	0.25	0.2	0.2	0.2	0.15
		Type-I	Type-I	Type-I	Power	Type-I	Power	Type-I	Power
Commensurate		0.02	0.03	0.02	0.39	0.02	0.92	0.02	1.00
Meta-analysis		0.03	0.05	0.03	0.41	0.03	0.92	0.02	1.00
Single study		0.08	0.05	0.08	0.41	0.08	0.94	0.07	1.00
Homogeneity		0.07	0.05	0.07	0.41	0.07	0.92	0.07	1.00
	p_0	0.1		0.1		0.1		0.1	
	p_{10} p_{11}	0.2	0.3	0.2	0.25	0.2	0.2	0.2	0.15
		Type-I	Type-I	Type-I	Power	Type-I	Power	Type-I	Power
Commensurate		0.27	0.03	0.28	0.39	0.26	0.92	0.26	1.00
Meta-analysis		0.28	0.05	0.28	0.41	0.30	0.92	0.28	1.00
Single study		0.08	0.05	0.08	0.41	0.08	0.94	0.07	1.00
Homogeneity		1.00	0.05	1.00	0.41	1.00	0.92	1.00	1.00
$n_0 = \mathbf{1000}$	p_0	0.2		0.2		0.2		0.2	
$n_1 = \mathbf{1000}$	p_{10} p_{11}	0.2	0.3	0.2	0.25	0.2	0.2	0.2	0.15
		Type-I	Type-I	Type-I	Power	Type-I	Power	Type-I	Power
Commensurate		0.01	0.03	0.01	0.75	0.01	1.00	0.01	1.00
Meta-analysis		0.01	0.03	0.02	0.77	0.01	1.00	0.01	1.00
Single study		0.03	0.03	0.03	0.74	0.03	1.00	0.04	1.00
Homogeneity		0.07	0.03	0.07	0.74	0.07	1.00	0.07	1.00
	p_0	0.1		0.1		0.1		0.1	
	p_{10} p_{11}	0.2	0.3	0.2	0.25	0.2	0.2	0.2	0.15
		Type-I	Type-I	Type-I	Power	Type-I	Power	Type-I	Power
Commensurate		0.06	0.03	0.06	0.69	0.06	1.00	0.06	1.00
Meta-analysis		0.05	0.03	0.05	0.76	0.05	1.00	0.05	1.00
Single study		0.03	0.03	0.03	0.74	0.03	1.00	0.04	1.00
Homogeneity		1.00	0.03	1.00	0.74	1.00	1.00	1.00	1.00

Table 4.6: (cont.)Frequentist operating characteristics for different models with different sample sizes in the historical and current studies

totally run two scenarios, in the first scenario we assume that the treatment effect is $\beta_{trt} = -0.887$ in the current study, which corresponds to decrease the event rate from 0.3 to 0.15. In the second scenario we just assume no effect of the treatments. In both of the studies, we assume that the study effects are same and also for the covariate effect.

Results

In figure 7 and 8, we showed the results of mean estimates of the treatment effect parameters. Under these two scenarios, the homogeneity should be the 'true' model which means that its model assumption best fit the simulation setup of parameters. So in both of scenarios, the commensurate model perform better than the meta-analysis approach and single study. In Hobbs (2012) [68], it mentioned that the random effects meta-analysis will not effectively borrow information when the similarity between the studies is obvious. So this simulation results show the aspect that the commensurate model can improve the inference of treatment effects.

4.5.5 Summary

In this section, we compare the 'commensurate', 'meta-analysis', 'single study' and 'homogeneity' models. First we setup a two groups study with one covariate considered. This model is useful in practice when we want to evaluate the device improvements or confirm the device's efficacy in specific subgroups. The results show that the commensurate and meta-analysis approaches effectively increase the power and comparably control the Type-I error. There exist certain situation in which the two study we investigate are similar but they are conflicting with our target, such as rejecting null hypothesis to claim the power. In such cases, the more powerful borrowing approach definitely will decrease the study effects. So in this situation, the

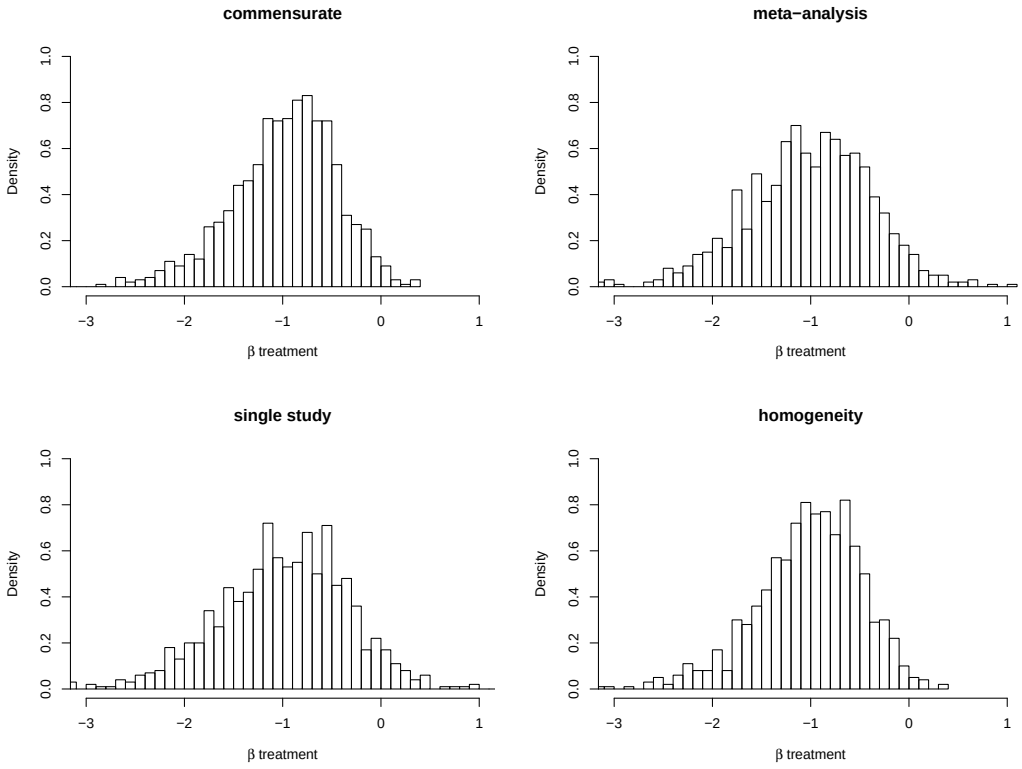

Figure 4.9.: Parameter estimate performance in 1000 simulations for Scenario 1

study is subjective that we can not improve the results. However, we often have some useful historical studies which means they will make efficient our current study. That is the case in which we should borrow the information. And both commensurate and meta-analysis models have been shown to reach this goal.

We also compare the commensurate and meta-analysis approaches. In the first study, in most situations the commensurate outperform the meta-analysis. In certain situations, as we mentioned earlier, when both of the studies conflict with our target, the more powerful approach will make worse results.

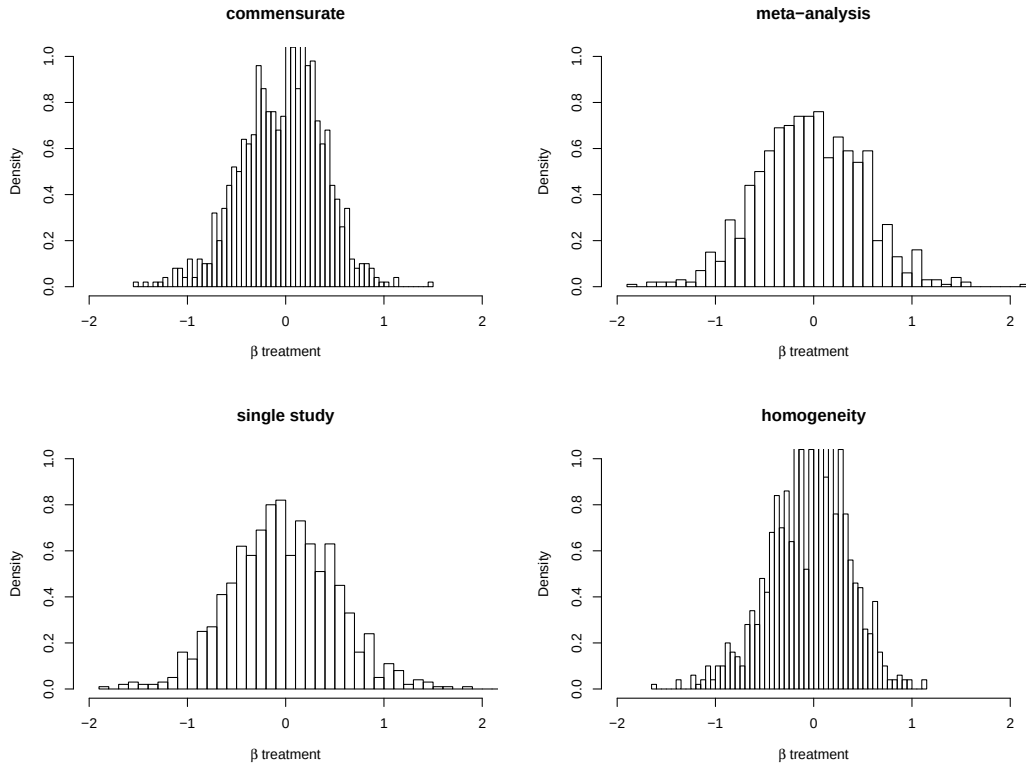

Figure 4.10.: Parameter estimate performance in 1000 simulations for Scenario 2

One of the most important feature of commensurate model is the 'commensurate prior' that describe the closeness between the parameters from different studies. There can be different hyper-priors models for the commensurate variance parameters. We have also tried to use 'spike and slab' prior as in Hobbs (2012) [68]. Just as the paper points out, all prior models are sort of subjective. The performance of the priors proposed will highly depend on the pattern of data from different groups and also also depends on the question we want to answer(for example, different hypothesis target). So this work surely need to be improved and calibrated well to fit the objective of study. Also, in all studies we tried, we only assume two studies. One of assumptions for the 'commensurate' model is that it assumes the homogeneity between the different historical groups. This is also a strong assumption that need to be evaluated before

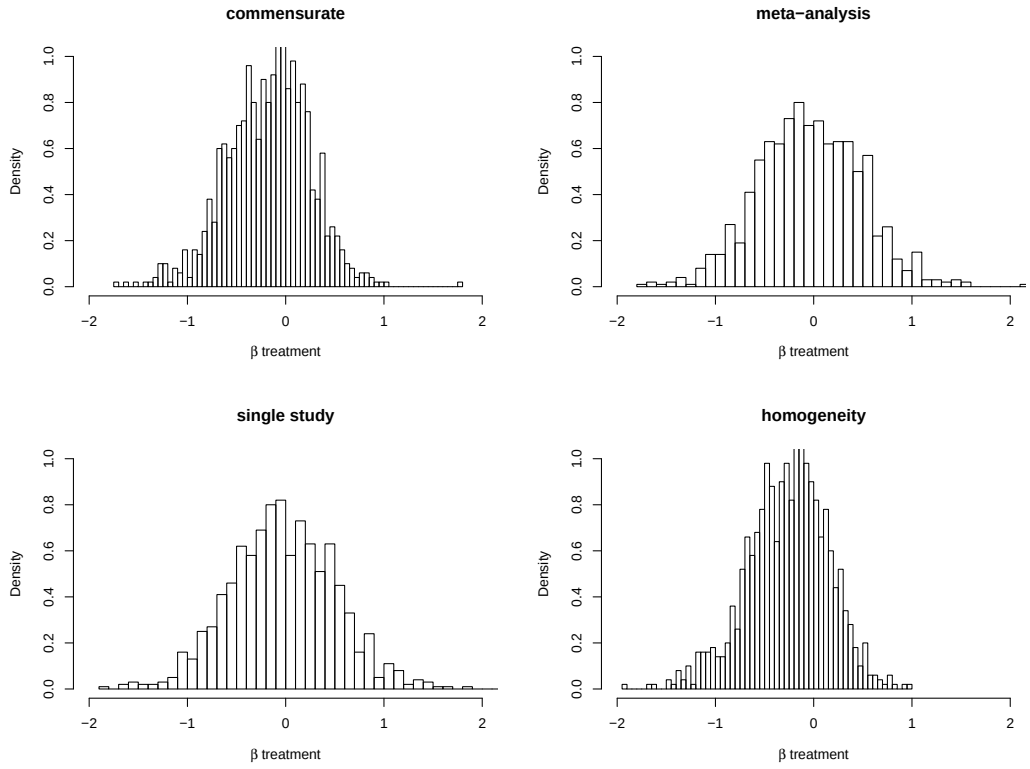

Figure 4.11.: Parameter estimate performance in 1000 simulations for Scenario 3

any practical studies. One possible solution for this problem is that we may condition each historical study on the current study parameters. This could be an interesting extension for the future work.

Discussion

In this project, we mainly worked on four kinds of models which are designed for different type of questions about adaptive borrowing information from historical studies. In each study, we have shown the advantages of the models compared with frequentist approaches and existing Bayesian approach. We novelly develop the empirical Bayesian shrinkage type estimate and the Bayesian model averaging approaches. These approaches have shown the property of borrowing information based

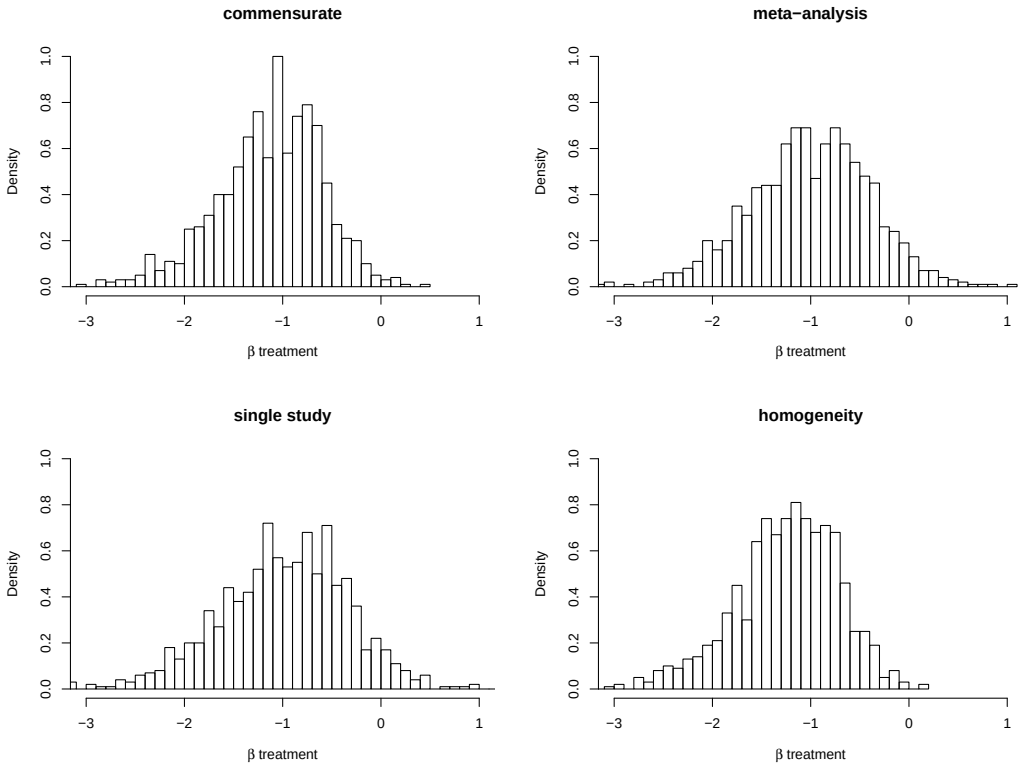

Figure 4.12.: Parameter estimate performance in 1000 simulations for Scenario 4

on the similarity between studies. The hierarchical model is a good way to decide the sample size when we want to use Bayesian approach to improve the performance of the study based on the current data as shown in Pennello and Thompson (2008) [69]. The commensurate is the most flexible approach among the four approaches we considered. We can incorporate different covariate factors in the study and also incorporate the treatment effect as in a randomized trial study. These flexibility properties will facilitate the subgroup analysis and the combining study of randomized clinical trials and single arm historical results. In conclusion, we show the non-comparable capability of Bayesian approach to handle the problem of borrowing information for medical device clinical trial studies.

5. Conclusion and Future Research

In this dissertation, we focused on investigating and developing Bayesian statistical methods for the modeling of genetic interactions. First, we introduced the definition of interactions in statistics. There are several statistical approaches available for detecting the gene-gene (epistasis) and gene-environment interactions. To provide the background about our proposed methods, we discussed two models: the penalized regression model and the shrinkage prior model that are widely applied in interaction studies.

In Chapter 2, we introduced a logistic regression based variable selection approach for gene-environment and gene-gene interaction studies. When building up the models with multiple factors, statisticians have a natural preference to include the interaction terms with the corresponding main effect terms. This can make the model more parsimonious and interpretable. In frequentist machine learning literature, this statistical constraint has been incorporated into different methods, such as Linear models, Lars and Lasso. In the Bayesian areas, there are relatively fewer papers exploring this topic. Therefore, we merged this constraint with the Bayesian variable selection framework. Under the logistic regression model, we proposed a Bayesian mixture model which represented the two hypotheses about the effect of each factor. We designed the 'Strong hierarchical' and 'Weak hierarchical' models by proposing different prior structures for the indicators of the priors. The extensive simulation and real data studies demonstrated the superiority of the proposed models over the other models considered. In the future research, we may consider comparing the performances with the other statistical models such as Random Forest or Multi-factor

Dimension Reductions for detecting the interactions. Due to the computational burdens of the Bayesian methods, the methods proposed in this project can be considered as a second step analysis after finishing the Genome-wide interaction studies by the frequentist approaches, such as the Bayesian empirical shrinkage-type estimator. Selecting subjective priors is another challenge in Bayesian variable selection. We also tried to develop a method using Bayesian model averaging technique to account for the uncertainty while eliciting the priors among the competing models. In the future research, we may strive to explore a systematic way to specify the priors, on which the Bayesian model averaging/selection can be applied more efficiently.

In Chapter 3, we investigated the Bayesian complementary version of the Natural and Orthogonal Interaction (NOIA) model for gene-environment interaction studies. The NOIA model allows us to estimate the genetic and environment orthogonally and decompose the total variances according to each effect. This feature is especially useful when dealing with multiple genetic and environmental factors that are usually confounded with each other. So we developed the Bayesian NOIA model, in which the estimates of the parameters lead to higher power and lower Type-I error. Also, we can assign proper priors based on the existing biological information or previous study results for yielding more satisfactory operating characteristics. We compared the proposed Bayesian NOIA model with the Bayesian usual functional model by analyzing the lung cancer data set. The results have shown that the Bayesian model correctly detect more positive effects which have been reported before. In the next level of research, we can try to combine this with the hierarchical interaction models from Chapter 2 which will effectively take advantages of both approaches.

In chapter 4, we first reviewed two Bayesian approaches that have been successfully applied in the gene-environment interaction studies. The case-control study yielded

the unbiased result for detecting the interactions but it tended to result in under-expected powers. The case-only study has a larger power to detect the interaction. However, the departure from the independency assumption may dramatically bias the study result. So the Bayesian empirical shrinkage-type estimator and Bayesian model averaging approaches were proposed to let the data decide on borrowing based on the uncertainty within the selection of types of the studies. These approaches would solve the dilemma of bias and variance trade-off in the gene-environment case-control studies. In the real data analysis, we always encounter the situation where we need to consider borrowing historical data into the current study to enrich the sample size. There are many characteristics that determine the eligibility of borrowing from the historical studies. In this project, we focused on making the decisions about borrowing based on the homogeneity among the studies. When the data from historical studies are pooled together with the current ones, the sample size will be enlarged. The departure from the homogeneity assumption will significantly risk the result of borrowing data. Inspired by the two Bayesian approaches in gene-environment interactions, we proposed the analogous Bayesian empirical shrinkage-type estimator and the Bayesian model averaging approaches for borrowing the historical data in an adaptive way. The simulation results have shown that the two Bayesian approaches have the property to automatically decide the strength borrowed from the historical data. The asymptotic property of the empirical estimator has also been discussed, so the hypothesis test could be conducted. In the Bayesian model averaging approach, the uncertainty among the model selection is considered in finding the posterior distribution of the model averaging estimator. Further more, we developed the Bayesian model averaging approach for the scenarios where some degree of departure from the homogeneity is allowed. The results from borrowing will depend on the equivalence margin selected. In the last, we discussed some other advanced Bayesian methods that could be applied on these studies. Overall, these could be one of the future

research directions that have a broad range of meanings for various studies, such as the meta-analysis problem in the genetic and clinical studies, the sequencing data analysis with the previous results and the adaptive clinical trial design with the effective historical information, which I will continue to explore in my future work.

Bibliography

[1] Teri A. Manolio, Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, Aravinda Chakravarti, Judy H. Cho, Alan E. Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N. Rotimi, Montgomery Slatkin, David Valle, Alice S. Whittemore, Michael Boehnke, Andrew G. Clark, Evan E. Eichler, Greg Gibson, Jonathan L. Haines, Trudy F. C. Mackay, Steven A. McCarroll, and Peter M. Visscher (2009) Finding the missing heritability of complex diseases. *Nature 461:747-753*

[2] Goldstein, D. B. (2009). Common genetic variation and human traits. *New England Journal of Medicine 360, 1696-1698*

[3] Teri A. Manolio (2010). Genome-wide association studies and disease risk assessment. *New England Journal of Medicine 363, 166-176.*

[4] Teri A. Manolio (2013). Bringing genome-wide association findings into clinical use. *Nature Reviews Genetics 14, 549-558*

[5] Aschard H, Chen J, Cornelis MC, Chibnik LB, Karlson EW, Kraft P. (2012). Inclusion of gene-gene and gene-environment interactions unlikely to dramatically improve risk prediction for complex diseases. *Am. J. Hum. Genet. 90, 962-972*

[6] Duncan Thomas (2010). Gene-environment-wide association studies: emerging approaches *Nature Reviews Genetics 11, 259-272*

[7] Kao, C., Zeng, Z. (2002). Modeling epistasis of quantitative trait loci using Cockerhams model. *Genetics 160, 1243-1261.*

[8] Duncan Thomas (2010). Methods for Investigating Gene-Environment Interactions in Candidate Pathway and Genome-Wide Association Studies.*Annu Rev Public Health 31:21-36*

[9] Cox, D. R. (1984). Interaction. *International Statistical Review 52, 1-31*

[10] Duncan Thomas (2004). *Statistical Methods in Genetic Epidemiology.* Oxford: Oxford University Press.

[11] McCullagh, P., Nelder, J. A. (1989).*Generalized Linear Models.* CHAPMAN and HALL.

[12] Andrew Gelman, John B. Carlin, Hal S. Stern, Donald B. Rubin (2003) *Bayesian Data Analysis.* CHIPMAN & HALL/CRC

[13] H Chipman (1996) Bayesian variable selection with related predictors. *The Canadian Journal of Statistics, Vol 24, No. 1, 1996, Pages 17-36*

[14] Berrington, A., Cox, D. R. (2007). Interpretation of interaction: a review. *Annals of Applied Statistics 1, 371-385.*

[15] Kooperberg, C., Leblanc, M., Dai, J., Rajapakse, I. (2009). Structures and assumptions: strategies to harness gene gene and gene environment interactions in GWAS. *Statistical Science 24, 472-488.*

[16] Ritchie M, Hahn L, Moore J. (2003). Power of multi factor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genetic Epidemiology 24:150-157*

[17] Goldstein BA, Polley EC, Briggs FB (2011). Random forests for genetic association studies. *Stat Appl Genet Mol Biol. 2011;10(1):32.*

[18] Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics 21, 243-247.*

[19] Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6, 461-464.*

[20] Zeng, Z., Kao, C., Basten, C. (1999). Estimating the genetic architecture of quantitative traits. *Genetic Research 74, 279-289.*

[21] Sun, W., Ibrahim, J., Zou, F. (2010). Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalized regression. *Genetics 185, 349-359.*

[22] Yi, N., Banerjee, S. (2009). Hierarchical generalized linear models for multiple quantitative trait locus mapping. *Genetics 181, 1101-1113.*

[23] Manichaikul, A., Moon, J., Sen, S., Yandell, B., Broman, K. (2009). A model selection approach for the identification of quantitative trait loci in experimental crosses, allowing epistasis. *Genetics 181, 1077-1086.*

[24] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society B 58, 267-288.*

[25] Park, M and Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics 9, 30-50.*

[26] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004). Least angle regression. *Annals of Statistics 32, 407-451.*

[27] Friedman, J., Hastie, T., Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33, 1-22.*

[28] Wu, T., Chen, Y., Hastie, T., Sobel, E., Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics 25, 714-721.*

[29] Tanck, M., Jukema, J., Zwinderman, A. (2006). Simultaneous estimation of gene-gene and gene environment interactions for numerous loci using double penalized log-likelihood. *Genetic Epidemiology 30, 645-651.*

[30] R.B.O'Hara and M.J.Sillanapaa. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis, 4, 85-117*

[31] Peter Hoff (2010). *A First Course in Bayesian Statistical Methods* Springer Texts in Statistics.

[32] Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association 103, 681-686.*

[33] T. J. Mitchell; J. J. Beauchamp (1988). Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association, Vol. 83, No. 404*

[34] Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhya Series B 60, 65-81.*

[35] Edward I. Georgea and Robert E. McCulloch (1993) Variable selection via Gibbs Sampling. *Journal of American Statistical Association 88:881-889*

[36] Edward I. Georgea and Robert E. McCulloch (1997) Approaches for bayesian variable selection. *Statistica Sinica 7, 339-373*

[37] Yi, N. (2004). A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci. *Genetics 167, 967-975.*

[38] Meuwissen, T. H. E., and M. E. Goddard, (2004). Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genet. Sel. Evol. 36: 261-279.*

[39] M. Hamada and C. F. J. Wu (1992). Analysis of designed experiments with complex aliasing. *Journal of Quality Technology, 24, 130-137*

[40] Nam Hee Choi, William Li, Ji Zhu (2010). Variable Selection with the Strong Heredity Constraint and its Oracle Property. *Journal of the American Statistical Association. Vol 105, 489*

[41] Jacob Bien, Jonathan Taylor, Robert Tibshirani (2013). A Lasso for Hierarchical Interactions. *The Annals of Statistics. In press*

[42] M. Yuan, R. Joseph and Y. Lin (2007). An Efficient Variable Selection Approach for Analyzing Designed Experiments. *Technometrics, 49(4), 430-439*

[43] M.Yuan, R. Joseph and H. Zou (2009). Structured Variable Selection and Estimation. *The Annals of Applied Statistics, 3(4), 1738-1757*

[44] H Chipman, M. Hamada, C.F.J.Wu (1997) A Bayesian Variable Selection Approach for Analyzing Designed Experiments with Complex Aliasing. *Technometrics , 39, 372–381.*

[45] Ioannis Ntzoufras (2009) *Bayesian Modeling using WinBUGS.* WILEY

[46] Jon Wakefield, Frank De Vocht and Rayjean J.Hung. (2010) Bayesian Mixture Modeling of Gene-Environment and Gene-Gene Interactions. *Genetic Epidemiology 34:16-25*

[47] Chipman, H. (2006). Prior Distributions for Bayesian Analysis of Screening Experiments, Chapter in Screening: *Methods for Experimentation in Industry, Drug Discovery, and Genetics*, A. Dean and S.M.Lewis, Editors, 235-267.

[48] Truong T, Hung RJ, Amos CI, Wu X, Bickebller H, Rosenberger A, Sauter W, Illig T, Wichmann HE, Risch A, Dienemann H, Kaaks R, Yang P, Jiang R, Wiencke JK, Wrensch M, Hansen H, Kelsey KT, Matsuo K, Tajima K, Schwartz AG, Wenzlaff A, Seow A, Ying C, Staratschek-Jox A, Nrnberg P, Stoelben E, Wolf J, Lazarus P, Muscat JE, Gallagher CJ, Zienolddiny S, Haugen A, van der Heijden HF, Kiemeney LA, Isla D, Mayordomo JI, Rafnar T, Stefansson K, Zhang ZF, Chang SC, Kim JH, Hong YC, Duell EJ, Andrew AS, Lejbkowicz F, Rennert G, Mller H, Brenner H, Le Marchand L, Benhamou S, Bouchardy C, Teare MD, Xue X, McLaughlin J, Liu G, McKay JD, Brennan P, Spitz MR. (2010) Replication of lung cancer susceptibility loci at chromosomes 15q25, 5p15, and 6p21: A pooled analysis from the international lung cancer consortium. *J Natl Cancer Inst 102: 959-971.*

[49] Amos CI, Wang LE, Lee JE, Gershenwald JE, Chen WV, Fang S, Kosoy R, Zhang M, Qureshi AA, Vattathil S, Schacherer CW, Gardner JM, Wang Y, Bishop DT, Barrett JH; GenoMEL Investigators, MacGregor S, Hayward NK, Martin NG, Duffy DL; Q-Mega Investigators, Mann GJ, Cust A, Hopper J; AMFS Investigators, Brown KM, Grimm EA, Xu Y, Han Y, Jing K, McHugh C, Laurie CC, Doheny KF, Pugh EW, Seldin MF, Han J, Wei Q. (2011). Genome-wide association study identifies novel loci predisposing to cutaneous melanoma. *Human Molecular Genetics doi:10.1093/hmg/ddr415*

[50] Cheongeun Oh, Kenny Q Ye, Qimei He and Nancy R Mendell (2003). Locating disease genes using Bayesian variable selection with the Haseman-Elston method. *BMC Genetics, 4: S69*

[51] Conti DV, Cortessis V, Molitor J, Thomas DC (2003) Bayesian modeling of complex metabolic pathways. *Human Heredity 56, 83-93*

[52] Adrian E. Raftery, David Madigan and Jennifer A. Hoeting (1997) Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association, 92:179-191.*

[53] Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), Bayesian Model Averaging: A Tutorial. *Statistical Science, 14, 382-401.*

[54] Michael A. Newton; Adrian E. Raftery (1994). Approximate Bayesian inference by the weighted likelihood bootstrap (with Discussion) *Journal of the Royal Statistical Society. Series B (Methodological), Vol. 56, No. 1. (1994), 3-48.*

[55] Newton MA, Raftery AE (1994) Approximate Bayesian inference by the Weighted Likelihood Bootstrap (with discussion). *Journal of Royal Statistical Society ,Series b, 56, 3-48*

[56] Hardy J, Singleton A. (2009). Genomewide association studies and human disease. *New England Journal of Medicine. 360(17): 1759-68.*

[57] Zeng, Z., Wang, T., Zou, W. (2005). Modeling quantitative trait Loci and interpretation of models. *Genetics 169, 1711-1725.*

[58] Alvarez Castro J, Carlborg O. (2007). A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis. *Genetics 176: 1151-1167.*

[59] Alvarez Castro J, Le Rouzic A, Carlborg O. (2008). How to perform meaningful estimates of genetic effects. *PLoS Genet 4: e1000062.*

[60] Ma J, Xiao F, Xiong M, Andrew AS, Brenner H, Duell EJ, Haugen A, Hoggart C, Hung RJ, Lazarus P, Liu C, Matsuo K, Mayordomo JI, Schwartz AG, Staratschek-Jox A, Wichmann E, Yang P, Amos CI. (2012).Natural and Orthog-

onal Interaction framework for modeling gene-environment interactions with application to lung cancer. *Human Heredity 73(4):185-194, 8/2012.*

[61] Albert PS, Ratnasinghe D, Tangrea J, Wacholder S. 2001. Limitations of the case-only design for identifying gene-environment interactions. *American Journal of Epidemiology 154:687-693.*

[62] Mukherjee, Bhramar and Ahn, Jaeil and Gruber, Stephen B. and Rennert, Gad and Moreno, Victor and Chatterjee, Nilanjan (2008). Tests for gene-environment interaction from case-control data: a novel study of type I error, power and designs. *Genetic Epidemiology 32:615-626*

[63] Mukherjee, Bhramar and Chatterjee, Nilanjan (2008). Exploiting gene-environment independence for analysis of case-control studies: An empirical-Bayes type shrinkage estimator to trade off between bias and efficiency. *Biometrics, 64, 685-694.*

[64] Viallefont, V., Raftery, A. E. and Richardson, S. (2001). Variable selection and Bayesian model averaging in case-control studies. Statist. Med., 20: 3215-3230

[65] Li, Dalin, Conti, David V (2009). Detecting gene-environment interactions using a combined case-only and case-control approach *American Journal of Epidemiology 169:497-504*

[66] Bishop YMM, Fienberg SE, Holland PW. *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, MA: The MIT Press; 1975.

[67] Brian Hobbs, Bradley Carlin, Sumithra Mandrekar, Daniel Sargent (2011). Hierarchical Commensurate and Power Prior Models for Adaptive Incorporation of Historical Information in Clinical Trials. *Biometrics 67, 1047-1056*

[68] Brian Hobbs, Daniel Sargent, Bradley Carlin (2012). Commensurate Priors for Incorporating Historical Information in Clinical Trials Using General and Generalized Linear Models. *Bayesian Analysis 7, no. 2, pp. 1-36*

[69] Gene Pennello, Laura Thompson (2008). Experience with Reviewing Bayesian Medical Device Trials. *Journal of Biopharmaceutical Statistics 18:1,81-115*

VITA

February, 1984 ……...……...… Born - Tianjin, China

July, 2006 ……...……... Bachelor, Information and Computing Science
Department of Mathematics
Beijing Institute of Technology

May, 2008 ……...…….. Master, Biostatistics
Georgia Southern University

August, 2013 ……...…….. Ph.D., Biostatistics
GSBS, UTHealth and MD Anderson