Published in final edited form as:

Hum Genet. 2015 January ; 134(1): 23–36. doi:10.1007/s00439-014-1478-5.

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

Changlu Liu,

Biomathematics and Biostatistics program, Graduate School of Biomedical Sciences, The University of Texas Health Science Center at Houston and The University of Texas MD Anderson Cancer Center, Houston, TX, USA, and Novartis Pharmaceuticals, East Hanover, NJ 07936, USA

Jianzhong Ma, and

Division of Clinical and Translational Sciences, Department of Internal Medicine, Medical School, and Biostatistics/Epidemiology/Research Design (BERD) component, Center for Clinical and Translational Sciences (CCTS), University of Texas Health Science Center at Houston, Houston, TX 77030, USA

Christopher I. Amos

Department of Community and Family Medicine, Geisel School of Medicine, Dartmouth College, Lebanon, NH 03766, USA

Christopher I. Amos: Christopher.I.Amos@Dartmouth.edu

Abstract

We propose a Bayesian hierarchical mixture model framework that allows us to investigate the genetic and environmental effects, gene by gene interactions and gene by environment interactions in the same model. Our approach incorporates the natural hierarchical structure between the main effects and interaction effects into a mixture model, such that our methods tend to remove the irrelevant interaction effects more effectively, resulting in more robust and parsimonious models. We consider both strong and weak hierarchical models. For a strong hierarchical model, both of the main effects between interacting factors must be present for the interactions to be considered in the model development, while for a weak hierarchical model, only one of the two main effects is required to be present for the interaction to be evaluated. Our simulation results show that the proposed strong and weak hierarchical mixture models work well in controlling false positive rates and provide a powerful approach for identifying the predisposing effects and interactions in gene-environment interaction studies, in comparison with the naive model that does not impose this hierarchical constraint in most of the scenarios simulated. We illustrated our approach using data for lung cancer and cutaneous melanoma.

Keywords

Bayesian Variable Selection; Bayesian Mixture Model; Hierarchical Interaction

1 Introduction

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 (Manolio, 2010). 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 (Aschard et al, 2012). 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 factor's effects could be modified according to the effects from genetic factors.

1.2 Statistical Background

1.2.1 Bayesian Variable Selection—Variable selection has been extensively studied 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 (Mitchell and Beauchamp, 1988) 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 McCulloch (1993, 1997) 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.

1.2.2 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 $f(\cdot)$ is the link function. In a usual statistical modeling approach, when inferring each effect of 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) first introduced the concept of hierarchical interaction and named it the *Heredity principle*. There are two versions of hierarchical constraints on effects (Chipman, 1996; Chipman and Wu, 1997). 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 studied extensively in the variable selection literature. Choi et al (2010) and Bien et al (2013) imposed this constraint on the popular variable selection approach Least Absolute Shrinkage and Selection Operator (Lasso). Yuani et al (2007); Yuan et al (2009) 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). Chipman (1996) proposed a hierarchical prior structure for modeling the interaction effects with the constraint. However, the approach does not explicitly explain the rationale of the prior specifications. In this paper, we specify the priors of the main effects and interactions and incorporate them into a Bayesian mixture model.

2 Methodology

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 $i = 1, \dots, n$. Let x_{ij} denote the number of minor alleles for the j th single nucleotide polymorphism (SNP) of i th observation, y_{ik} as the k th environmental exposure (EXP) for i th observation. In logistic regression, we will model the main effects of both gene and environment as well as gene-environment and gene-gene interaction analysis as

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

where p_i indicates the probability of disease. σ is the general intercept and $\exp(\alpha)$ is the baseline odds for the disease. 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 k th 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 j th SNP and the k th environmental factor and η_{jl} denotes the interaction effect between j th and l th SNP. Since the gene-gene products $x_{ij}x_{il}$, $j < l$ 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 EXP, 6 gene-environment interactions and 15 gene-gene interaction parameters could be included in the model.

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

$$\beta_j \sim N(0, \sigma_j^2) \quad \gamma_k \sim N(0, \sigma_k^2) \quad \theta_{jk} \sim N(0, \sigma_{jk}^2) \quad \eta_{jl} \sim N(0, \sigma_{jl}^2)$$

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

$$\sigma_j^{-2} \sim \text{Gamma}(a_j, b_j) \quad \sigma_k^{-2} \sim \text{Gamma}(a_k, b_k) \quad \sigma_{jk}^{-2} \sim \text{Gamma}(a_{jk}, b_{jk}) \quad \sigma_{jl}^{-2} \sim \text{Gamma}(a_{jl}, b_{jl})$$

These structures are from the conventional Bayesian hierarchical model for proposing the priors for the parameters. Recently, researchers have started to modify this framework for the purpose of variable selection. Park and Casella (2008) and Yi and Banerjee (2009) proposed the Bayesian Lasso model by proposing some common priors for the parameters, by which the trivial parameters with estimated values near 0 would be flattened out the model, but this model requires specifying a tuning parameter, whose value is complex to estimate. In this paper, we evaluate different structures for the variances by proposing and studying different approaches to modeling the mixture components.

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:

$$\beta_j \sim N(0, \sigma_{\varepsilon j}^2) \quad \gamma_k \sim N(0, \sigma_{\varepsilon k}^2) \quad \theta_{jk} \sim N(0, \sigma_{\varepsilon jk}^2) \quad \eta_{jl} \sim N(0, \sigma_{\varepsilon jl}^2)$$

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

$$\beta_j \sim N(0, \sigma_j^2) \quad \gamma_k \sim N(0, \sigma_k^2) \quad \theta_{jk} \sim N(0, \sigma_{jk}^2) \quad \eta_{jl} \sim N(0, \sigma_{jl}^2)$$

Motivated by the Stochastic Search Variable Selection framework (George and McCulloch, 1993, 1997), 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:

$$\beta_j \sim N(0, I_j \sigma_j^2 + (1 - I_j) \sigma_{\varepsilon j}^2) \quad \gamma_k \sim N(0, I_k \sigma_k^2 + (1 - I_k) \sigma_{\varepsilon k}^2) \quad \theta_{jk} \sim N(0, I_{jk} \sigma_{jk}^2 + (1 - I_{jk}) \sigma_{\varepsilon jk}^2) \quad \eta_{jl} \sim N(0, I_{jl} \sigma_{jl}^2 + (1 - I_{jl}) \sigma_{\varepsilon jl}^2)$$

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

$$I_j \sim \text{Bernoulli}(\pi_j) \quad I_k \sim \text{Bernoulli}(\pi_k) \quad I_{jk} \sim \text{Bernoulli}(\pi_{jk}) \quad I_{jl} \sim \text{Bernoulli}(\pi_{jl})$$

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

$$\pi_{jk} = I_j \times I_k \times \pi_{jk}^s \quad \pi_{jl} = I_j \times I_l \times \pi_{jl}^s$$

where π_{jk}^s and π_{jl}^s 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 conditional prior probabilities, we assume that

$$\pi_{jk}^s = \min(\pi_j, \pi_k) \quad \pi_{jl}^s = \min(\pi_j, \pi_l)$$

Of course, other frameworks for the priors for these conditional priors could be proposed. By assuming strong hierarchical effects model, the prior probabilities 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

$$\pi_{jk} = (I_j + I_k - I_j * I_k) \times \pi_{jk}^w \quad \pi_{jl} = (I_j + I_l - I_j * I_l) \times \pi_{jl}^w$$

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

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

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

$$\pi_{jk}^{w01} = \min(\pi_j, \pi_k) \times \frac{\pi_k}{\pi_j + \pi_k} \quad (1)$$

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 either or 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 smaller main effects. Cox (1984) 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. Therefore, we fix the variance components by restricting their credible interval within a prespecified range of disease odds. So under the null hypothesis, we assume the 95% credible interval will be $e^{1.96\sigma_{\varepsilon j}} = e^{\pm 0.05}$ and $e^{1.96\sigma_{\varepsilon k}} = e^{\pm 0.05}$. These specifications will provide that under the null model the odds will be sampled from within the 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 variance components by restricting the odds within a certain range. We assume the 90% credible interval of odds would be $(\frac{1}{3}, 3)$ for genetic main and gene-gene Interaction effects and $(\frac{1}{4}, 4)$ for environmental exposure main and gene-environment interaction effects (Figure 1). In most genetic studies, the proportion of positive effects is expected to be low and therefore estimating the priors from the data is unlikely to be a very effective approach. Therefore, we set the values for the hyper priors for the variances as:

$$\sigma_{\varepsilon j}^2 = \sigma_{\varepsilon k}^2 = \sigma_{\varepsilon jk}^2 = \sigma_{\varepsilon jl}^2 = 0.00065 \quad \sigma_j^2 = \sigma_{jl}^2 = 0.446 \quad \sigma_k^2 = \sigma_{jk}^2 = 0.710$$

2.4 Bayesian 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

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

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 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 (Ntzoufras, 2009).

3 Empirical Study

In this section, we use simulation studies to evaluate the efficacy of the proposed approach and to compare the results with the model that does not consider the hierarchical 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 exposure and 50 INTs

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

We compared the performance of four models based on Bayesian mixture models. Wakefield et al (2010) 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, 2006). 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

$$PE = \frac{1}{N} \sum_{i=1}^N |y_i^{test} - \hat{p}_i| \quad (2)$$

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

$$\hat{p}_i = \frac{1}{1 + \exp(-(L_i))} \quad (3)$$

where L_i is the linear predictor for the i th 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 $P(I = 1 | \text{data}) > 0.5$.

3.1 Study I

We include 6 additive genetic factors (SNP) and 1 environmental factor (exposure) 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 exposure 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 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.

Figures S1 to S10 (Online Resource) 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 and Figure 3 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. Overall, 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.

3.2 Study II

We simulate 50 SNPs with only the additive effects and 1 exposure 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 exposure 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 environmental 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 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.

As shown in Figure 4, 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 5, 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 scenario 5, 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.

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 3.

As shown in Figure S11 (Online Resource), 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, as shown in Figure S12 (Online Resource), 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. The results for scenario 3 is shown in Figure S13 (Online Resource). This scenario 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 to the independent model and the effect enforce model similarly to the strong hierarchical model. In scenario 4, the non-null GxG interaction corresponds to null genetic main effects at SNP 1 and SNP 2

(Figure S14, Online Resource). Here we observed that the independent model could identify the significant interaction while the other models failed. This is because the independent model does 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 higher probability of including the corresponding SNPs. When we compared the prediction performances in these four scenarios, the strong hierarchical mode performed best overall (Figure S15, Online Resource). 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.

4 Analysis of gene by exposure studies

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 from 13 countries. For illustration of the model, we focused on 6 SNPs in 3 regions: rs2736100 and rs402710 at 5p15, rs2256543 and rs4324798 and rs16969968 at 6p21, and rs16969968 and rs8034191 at 15q25, which have previously been associated with lung cancer risk, but the chromosome 6p region is weakly associated with risk for squamous carcinoma while the chromosome 5p SNPs associate with risk of adenocarcinoma and the chromosome 15q SNPs with both subtypes. The data we applied include 8867 participants with complete data. Among all the participants, there were 5217 controls and 3650 cases. In the case group, there were 2434 males and 1216 females and 3378 were smokers. In the control group, there were 3642 males and 1575 females and 3703 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 6 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 (Truong et al, 2010), 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 exposure, 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 less associated 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 smoking. Also, the hierarchical model provided weak evidence for an interaction between rs2256543 and smoking.

Supplemental figures S16 to S21 depict the impact that variations in priors have upon the posterior probabilities for the main effects and interactions under independence, strong and weak hierarchy models. Figures S16 to S18 depict results when the effects were systematically varied between 0.1 and 0.9 by 0.2 and figures S19 to S21 depict results when the main effects were varied but the interaction prior was set at 0.5. Results show that the independence prior identifies many main effects and interactions under most scenarios and particularly when priors are set high. The strong hierarchy constraint identifies few interactions when priors were systematically varied and also fails to detect the known interaction with rs16969968 for all prior models when the interaction prior was set to 0.5. The weak hierarchy model provides the best performance, correctly identifying a known interaction with rs16969968 under either the systematic or interaction fixed prior when the prior probabilities of effects were 0.5 or higher and with low posterior probabilities of detecting gene-gene interactions, which seems a likely explanation for these data, given that ongoing meta-analyses including additional data sets have yet to identify significant gene-gene interactions using standard logistic regression.

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. In this study we wanted to investigate 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 1024 controls. There are five different colors for eyes as in Table 4. We coded the color factor into 2 dummy variables which represent 3 level of eye colors (blue/grey, brown and green/hazel). Green and hazel colors were combined because they have similar hues and similarly blue and grey were combined because grey eyes are of similar grade in hue to blue.

Figure 7 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, 2012). 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.

5 Discussion

In this study, we introduced a Bayesian mixture model for modeling the gene-gene and gene-environment interactions simultaneously. Compared with the traditional logistic regression model, the Bayesian model has shown good performance on the 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. Even with a limited number of main effect terms, as we have modeled here, a large number of interaction terms need to be

considered, which will make the model fit and variable selection difficult for traditional approaches. A computational limitation of the method we developed is that it could not be applied in its current form to a full genome-wide association analysis because of the reliance on MCMC methods that would be excessively computationally intensive for modeling millions of main effects and interactions. A first step of modeling is to reduce the space of main effects to a few hundred SNPs. The selection process could use usual univariate logistic regression with a low significance criterion or some version of multivariable forward selection to reduce the space of main effects and interactions to be considered. The next step would be applying the proposed Bayesian approaches here for the candidate gene mapping. The selection process could use usual univariate logistic regression with a low significance criterion or a version of multivariable forward selection (Gu et al, 2009) to reduce the space of main effects and interactions to be considered. After a subset of variables are chosen further reduction of the search space can be accomplished by hyperlasso Hoggart et al (2008). The next step would be applying the proposed Bayesian approaches described here for the candidate gene characterization, which will give an accurate assessment of the joint effects of variables influencing disease susceptibility.

Although the hierarchical relation we studied has been described in the statistical machine learning literature, there is a clear need for extending the model application into the genetic studies. Several articles (Wakefield et al, 2010) mentioned hierarchical constraints, but did not compare their performances. This article provided a systematic approach to set up the priors among the variables to impose the hierarchical constraint. The approach to that we have taken assumes a specific form for the strength of the prior when fitting hierarchical constraints that is based on mild assumptions about the strength of the interaction. It may be possible to further relax this assumption as has been suggested by George and McCulloch (1997), but that would require further evaluation and may be hard to implement in SNP studies if relatively few positive associations are expected.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

CIA has been supported by NIH grant U19CA148127 and P30CA023108. JM has been supported by NIH grant R01CA134682. JM also acknowledge the support provided by the Biostatistics/ Epidemiology/ Research Design (BERD) component of the Center for Clinical and Translational Sciences (CCTS) for this project. CCTS is mainly funded by the NIH Centers for Translational Science Award (NIH CTSA) grant (UL1 RR024148), awarded to University of Texas Health Science Center at Houston in 2006 by the National Center for Research Resources (NCRR) and its renewal (UL1 TR000371) by the National Center for Advancing Translational Sciences (NCATS).

References

Amos C, Wang L, Lee J, Gershenwald J, Chen W, Fang S, Kosoy R, Zhang M, Qureshi A, Vattathil S, Schacherer C, Gardner J, Wang Y, Bishop D, Barrett J, Investigators G, Macgregor S, Hayward N, Martin N, Duffy D, Investigators QM, Mann G, Cust A, Hopper J, Brown K, Grimm E, Xu Y, Han Y, Jing K, McHugh C, Laurie C, Doheny K, Pugh E, Seldin M, Han J, Wei Q. AMFS-Investigators. Genome-wide association study identifies novel loci predisposing to cutaneous melanoma. *Human Molecular Genetics*. 2012 2011.

- Aschard H, Chen J, Cornelis M, Chibnik L, Karlson E, Kraft P. Inclusion of gene-gene and gene-environment interactions unlikely to dramatically improve risk prediction for complex diseases. *Am J Hum Genet.* 2012; 90:962–972. [PubMed: 22633398]
- Bien J, Taylor J, Tibshirani R. A lasso for hierarchical interactions. *The Annals of Statistics.* 2013; 41:1111–1141.
- Chipman H. Bayesian variable selection with related predictors. *The Canadian Journal of Statistics.* 1996; 24:17–36.
- Chipman H. Prior Distributions for Bayesian Analysis of Screening Experiments, Chapter in *Screening. Methods for Experimentation in Industry Drug Discovery, and Genetics, A.* 2006
- Chipman MH, Hamada, Wu C. A bayesian variable selection approach for analyzing designed experiments with complex aliasing. *Technometrics.* 1997; 39:372–381.
- Choi N, Li W, Zhu J. Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association.* 2010; 105:489.
- Cox D. Interaction. *International Statistical Review.* 1984; 52:1–31.
- George E, McCulloch R. Variable selection via gibbs sampling. *Journal of American Statistical Association.* 1993; 88:881–889.
- George E, McCulloch R. Approaches for bayesian variable selection. *Statistica Sinica.* 1997; 7:339–373.
- Gu X, RF F, Rosner G, Relling M, Peng B, Amos C. A modified forward multiple regression in high-density genome-wide association studies for complex traits. *Genet Epidemiol.* 2009; 33:518–525.
- Hamada M, Wu C. Analysis of designed experiments with complex aliasing. *Journal of Quality Technology.* 1992; 24:130–137.
- Hoggart C, Whittaker J, de Iorio M, Balding D. Simultaneous analysis of snps in genome-wide and resequencing association studies. *PLoS Genet.* 2008; 4:e1000130. [PubMed: 18654633]
- Manolio T. Genome-wide association studies and disease risk assessment. *New England Journal of Medicine.* 2010; 363:166–176. [PubMed: 20647212]
- Mitchell Y, Beauchamp J. Bayesian variable selection in linear regression. *Journal of the American Statistical Association.* 1988; 83
- Ntzoufras I. Bayesian modeling using winbugs. 2009
- Park T, Casella G. The bayesian lasso. *Journal of the American Statistical Association.* 2008; 103:681–686.
- Truong T, Hung R, Amos C, Wu X, Bickeboller H, Rosenberger A, Sauter W, Illig T, Wichmann H, Risch A, Dienemann H, Kaaks R. 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.* 2010; 102:959–971. [PubMed: 20548021]
- Wakefield J, De Vocht F, Hung R. Bayesian mixture modeling of gene-environment and gene-gene interactions. *Genetic Epidemiology.* 2010; 34:16–25. [PubMed: 19492346]
- Yi N, Banerjee S. Hierarchical generalized linear models for multiple quantitative trait locus mapping. *Genetics.* 2009; 181:1101–1113. [PubMed: 19139143]
- Yuan M, Joseph R, Zou H. Structured variable selection and estimation. *The Annals of Applied Statistics.* 2009; 3:1738–1757.
- Yuani M, R J, Lin Y. An efficient variable selection approach for analyzing designed experiments. *Technometrics.* 2007; 49:430–439.

Fig. 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).

Fig. 2. Prediction performance for each model in each scenario of study I. 'ind': independent, 'eff': effect enforce, 'str': strong hierarchical, 'wea': weak hierarchical, 'fre' frequentist logistic regression.

Fig. 3. Prediction performance for each model in each scenario of study I. 'ind': independent, 'eff': effect enforce, 'str': strong hierarchical, 'wea': weak hierarchical, 'fre' frequentist logistic regression.

Fig. 4. Prediction performance for each model in each scenario of study II. 'ind':independent, 'eff': effect enforce, 'str': strong hierarchical, 'wea': weak hierarchical, 'fre' frequentist logistic regression.

Fig. 5. Variable selection performance for each model in each scenario of study II. 'str': strong hierarchical, 'eff': effect enforce, 'ind': independent,

Fig. 6.
Real data results for the lung cancer study.

Fig. 7.

Real data results for the cutaneous melanoma study. 24 SNPs included in the study: rs1015362, rs1042602, rs10757257, rs10830253, rs12896399, rs12913832, rs1335510, rs1393350, rs1408799, rs16891982, rs17305573, rs1805007, rs1806319, rs1847142, rs1885120, rs2218220, rs2284063, rs28777, rs4911414, rs4911442, rs6001027, rs7023329, rs910873, rs935053; 2 exposure: gray/blue vs. brown and hazel/green vs. brown

Table 1

Parameter setups for Study I

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

Table 2

Parameter setups for Study II

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

Table 3

Parameter setups for Study III

Scenario	SNP		Exposure		GE Interaction		GG Interaction		
	β_5	β_6	γ_1	γ_2	θ_6	θ_7	γ_1	γ_4	γ_{15}
1	.223	.223	.405	.405	.405	.405	0	0	0
2	.223	.223	.405	.405	.405	.405	0	0	.223
3	.223	.223	.405	.405	.405	.405	0	.223	0
4	.223	.223	.405	.405	.405	.405	.223	0	0

Table 4

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

Eye color	Case	Control	Total
Blue	396	331	727
Green	150	164	314
Grey	12	18	30
Brown	189	312	501
Hazel	182	199	381
Total	929	1024	1953