Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 151259, 15 pages
doi:10.1155/2012/151259

Research Article

Genotype-Based Bayesian Analysis of Gene-Environment Interactions with Multiple Genetic Markers and Misclassification in Environmental Factors

Iryna Lobach[1] and Ruzong Fan[2]

[1] *Department of Population Health, Division of Biostatistics, School of Medicine, New York University, New York, NY 10016, USA*
[2] *Biostatistics and Bioinformatics Branch, Division of Epidemiology, Statistics and Prevention Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Rockville, MD 20852, USA*

Correspondence should be addressed to Iryna Lobach, iryna.lobach@nyumc.org

Received 1 March 2012; Revised 23 May 2012; Accepted 25 May 2012

Academic Editor: Wei T. Pan

A key component to understanding etiology of complex diseases, such as cancer, diabetes, alcohol dependence, is to investigate gene-environment interactions. This work is motivated by the following two concerns in the analysis of gene-environment interactions. First, multiple genetic markers in moderate linkage disequilibrium may be involved in susceptibility to a complex disease. Second, environmental factors may be subject to misclassification. We develop a genotype based Bayesian pseudolikelihood approach that accommodates linkage disequilibrium in genetic markers and misclassification in environmental factors. Since our approach is genotype based, it allows the observed genetic information to enter the model directly thus eliminating the need to infer haplotype phase and simplifying computations. Bayesian approach allows shrinking parameter estimates towards prior distribution to improve estimation and inference when environmental factors are subject to misclassification. Simulation experiments demonstrated that our method produced parameter estimates that are nearly unbiased even for small sample sizes. An application of our method is illustrated using a case-control study of interaction between early onset of drinking and genes involved in dopamine pathway.

1. Introduction

A key component to prevention and control of complex diseases, such as cancer, hypertension, diabetes, and alcoholism, is to study the independent, cumulative, and interactive effects of genetic and environmental factors. This analysis has the potential to impact the

understanding of the role of genetic influences under various environmental exposures, thus providing valuable information to (1) better understand the biological pathways involved in the disease and its progression, thus providing major clues to the underlying causes of alcohol dependence; (2) design personalized interventions targeted to individuals with enhanced vulnerability to the disease (the risk genes may help identify patients at higher risk long before any symptoms occur); (3) gain critical understanding for drug discovery.

This work is motivated by the following two concerns in the analysis of gene-environment interactions. **First**, complex diseases are caused by multiple variants with small-to-moderate effect sizes working in concert [1]. Most of the results of published genome-wide association studies are based on single nucleotide polymorphism (SNP) analysis [2]. This approach may suffer from low power due to a large number of tests and small effect sizes of individual SNPs. Furthermore, the true causal genetic marker is often not genotyped, rather is captured through linkage disequilibrium (LD) with the typed markers. Since each SNP has only partial linkage disequilibrium with the causal SNP, the observed effect size of the typed SNP is lower than the effect size of the causal SNP. In light of this concern, we propose to use a risk function that allows the genetic markers in linkage disequilibrium to enter the model directly [3]. This model eliminates the need to estimate haplotype phase and hence protects against bias due to the uncertainty that may arise due to the haplotype phase ambiguity [4–8]. In addition, the computation burden can be significantly reduced since the proposed approach uses genotype data directly. **Second**, many variables that are of interest to biomedical researchers are subject to misclassification, for example, due to uncertainty associated with a recall or a measurement at an individual level. Misclassification may result in bias and loss of power to detect gene-environment interactions [9]. Oftentimes uncertainty associated with these variables may not be avoided in practice. The loss of power prevents the ability to discover gene-environment interactions in small studies or studies involving analysis of subtypes of complex diseases.

An example of biomedical problem of gene-environment interactions is the analysis of role of age when first got drunk in the etiology of alcohol dependence. The age at which a person gets drunk for the first time may influence genes linked to alcoholism, making the youngest drinkers most susceptible to severe problems [10]. Twin study found that when twins started drinking early (age < 13 years old), genetic factors contributed greatly to risk for alcohol dependence, at rates as high as 90 percent in the youngest drinkers [10]. Some early-onset drinkers do not develop alcohol problems and some late-onset drinkers do, hence it is important to investigate genetic and environmental influences that predispose for or protect against alcohol dependence in these two groups. However, the definition of early age of getting drunk is subject to misclassification due to uncertainty associated with the recall.

In light of these concerns, we develop a Bayesian methodology for analysis of gene-environment interactions in case-controls studies. Estimation and inference are based on a pseudolikelihood function [3, 11, 12]. This pseudolikelihood function offers the following advantages. One is that environmental variables measured exactly are modeled completely nonparametrically. Furthermore, *a priori* information about the probability of disease can be incorporated directly. The pseudolikelihood function exploits gene-environment independence assumption which is a reasonable assumption in many practical applications. If the gene-environment interaction is not significantly present in the population, then the distribution of genotype can be specified within strata defined by an environmental covariate. The proposed analysis is based on a pseudolikelihood function hence conventional Bayesian techniques may not be applied directly. Validity of Bayesian techniques need to be examined when the likelihood function is not a proper likelihood [13]. We followed Monahan and Boos

[13] and Lobach et al. [3] to validate our Bayesian approach under this pseudolikelihood function. Our Bayesian approach has the ability to shrink the parameter estimates towards prior and hence reduce variability in parameter estimates. This property is essential when environmental exposure is subject to misclassification, especially in studies with smaller sample sizes, for example, of subtypes of complex disease. On the other hand, if sample size is large enough, estimation and inference can be based on the asymptotic posterior distribution that we derived which will ease the computational burden.

An outline of this paper is as follows. In Section 2 we introduce notation and formally state the problem. In Section 2 we present the Bayesian model under various scenarios. Section 3 describes asymptotic posterior distribution. Section 4 describes simulation experiment. Section 5 describes application of the Bayesian model to the analysis of alcoholism study. Section 6 gives concluding remarks.

2. Bayesian Model Based on Pseudolikelihood

2.1. Notation and Risk Function

Consider a sample consisting of n_0 controls and n_d cases at disease stage or type $d = 1, \ldots, K$. Define D as the disease status. Following Lobach et al. [11], we pretend that this case-control sample is collected using a simple Bernoulli scheme, where the selection probability of a subject given disease status is proportional to $n_d/\mathrm{pr}(D = d)$, $d = 0, 1, \ldots, K$. Let $R = 1$ denote the indicator of whether or not a subject is selected into the case-control sample. All participants of the study will have this selection status $R = 1$. The observed genetic data consist of unphased genotypes $G = (G_1, \ldots, G_I)$ at I loci. Let $Q(G; \theta)$ be a model describing Hardy-Weinberg equilibrium (HWE).

Let (T, Z) denote all nongenetic variables of interest. Suppose T is the set of factors subject to misclassification, and Z is the set of variables observed exactly. We assume that the observed genetic data does not contain any additional information on disease status and the true environmental covariate given the genetic variable of interest. Let X denote the error-prone version of T. Suppose the misclassification process is defined by the following parametric structure $p_{\mathrm{miss}}(x \mid T, G, Z, D, \xi)$. This model is general enough to capture differential misclassification. The joint distribution of the environmental factors in the underlying population can be specified in the following form $p_{T|Z}(t \mid z, \xi) f_Z(z)$. While T may be a vector of factors, for simplicity of presentation in what follows we suppose that T is a factor.

Given the environmental covariates T and Z, genotype data G, the risk of disease in the underlying population is given by the following polytomous logistic model:

$$\mathrm{pr}(D = k \geq 1 \mid \mathbf{G}, T, Z) = \frac{\exp\{\beta_{k0} + m_k(\mathbf{G}, T, Z; \beta)\}}{1 + \sum_{j=1}^{K} \exp\{\beta_{j0} + m_j(\mathbf{G}, T, Z; \beta)\}}, \tag{2.1}$$

where $m(\bullet)$ is a function of known form parameterizing the risk of disease in terms of parameters β. For the ith marker, denote the two alleles by M_i and m_i, with frequencies P_{M_i} and P_{m_i}, respectively. Following Lobach et al. [3], we define the following dummy variables and two risk models: genotype effect model and additive effect model.

Define the following dummy variables:

$$A_i = \begin{cases} 1, & \text{if } G_i = M_iM_i, \\ 0, & \text{if } G_i = M_im_i, \\ -1, & \text{if } G_i = m_im_i, \end{cases} \qquad B_i = \begin{cases} -P_{m_i}^2, & \text{if } G_i = M_iM_i, \\ P_{M_i}P_{m_i}, & \text{if } G_i = M_im_i, \\ -P_{M_i}^2, & \text{if } G_i = m_im_i. \end{cases} \tag{2.2}$$

Notice that $A_i + 1$ is the number of allele M_i at the ith marker, and hence A_i can be used to model the allele or additive effect of M_i. Let $\mathrm{pr}(g;\theta)$ be a parametric form of the joint distribution of the observed genetic markers. In the following, we provide two examples of function $m_k(\cdot)$ using the genotype information $\mathbf{G} = (G_1, G_2, \ldots, G_I)$.

2.1.1. Genotype Effect Model (GEM)

The following specification of the risk function incorporates both additive and dominance effects of genotype, as well as the multiplicative gene-environment interactions

$$\begin{aligned} m_k(\mathbf{G}, T, Z; \beta) = m_k(\mathcal{A}, \mathcal{B}, T, Z; \beta) &= T\beta_{kT} + Z\beta_{kZ} + \sum_{i=1}^{I} A_i\beta_{kAi} \\ &+ \sum_{i=1}^{I} TA_i\beta_{kATi} + \sum_{i=1}^{I} ZA_i\beta_{kAZi} + \sum_{i=1}^{I} B_i\beta_{kDi} + \sum_{i=1}^{I} TB_i\beta_{kDTi} + \sum_{i=1}^{I} ZB_i\beta_{kDZi}. \end{aligned} \tag{2.3}$$

In this formulation, the regression coefficients β_{kAi} and β_{kDi} model risk due to the additive and dominance effect, respectively [14, 15]. The remaining terms capture the multiplicative gene-environmental interaction.

2.1.2. Additive Effect Model (AEM)

Suppose that the dominance effect is not significantly present in the model (2.3). In this situation, the risk function takes the following form:

$$m_k(\mathbf{G}, T, Z; \beta) = m_k(\mathcal{A}, T, Z; \beta) = T\beta_{kT} + Z\beta_{kZ} + \sum_{i=1}^{I} A_i\beta_{kAi} + \sum_{i=1}^{I} TA_i\beta_{kATi} + \sum_{i=1}^{I} ZA_i\beta_{kAZi}. \tag{2.4}$$

2.2. Pseudolikelihood

Let us denote $\kappa_k = \beta_{k0} + \log(n_k/n_0) - \log(\pi_k/\pi_0)$, $k = 1, 2, \ldots, K$, and $\tilde{\kappa} = (\kappa_1, \ldots, \kappa_K)^T$. In addition, let $\tilde{\beta}_0 = (\beta_{10}, \ldots, \beta_{K0})^T$, $\Omega = (\tilde{\beta}_0^T, \beta^T, \Theta^T, \tilde{\kappa}^T)^T$, $\mathcal{B} = (\Omega^T, \eta^T)^T$, and $\upsilon = (\eta^T, \xi^T)^T$. Define

$$S(k, \mathbf{g}, t, z; \Omega) = \frac{\exp\left[1_{(k\geq 1)}(k)\{\kappa_k + m_k(\mathbf{g}, t, z; \beta)\}\right]}{1 + \sum_{j=1}^{K} \exp\{\beta_{j0} + m_j(\mathbf{g}, t, z; \beta)\}} \mathrm{pr}(\mathbf{g}; \Theta). \tag{2.5}$$

We assume that G and (X, Z) are independently distributed in the underlying population. Only changes in notation are needed to model genotype and environment within strata thus relaxing gene-environment independence assumption. An example of gene-environment dependence is polymorphisms in nicotine metabolism pathway that may regulate the degree of addiction to nicotine, thus creating gene-environment interaction. Furthermore, these polymorphisms may interact with smoking status while being involved in lung cancer [16]. We suppose that the type of genetic covariate measured does not depend on the individual's true genetic covariate, given disease status, environmental covariates and the measured genetic information. Furthermore, we suppose that the observed genetic variable does not contain any additional information on disease status and true environmental covariate given the genetic variable of interest.

Similarly to Lobach et al. [11], we propose to use the following pseudolikelihood function in place of the likelihood function to estimate the parameters:

$$\begin{aligned} L_{\text{Pseudo}}(k, \mathbf{g}, x, z; \Omega, \eta, \xi) &\equiv \text{pr}(D = k,\ \mathbf{G} = \mathbf{g},\ X = x \mid Z = z,\ R = 1) \\ &= \frac{\sum_{t^*} S(k, \mathbf{g}, t^*, z; \Omega) p_{\text{miss}}(x \mid k, \mathbf{g}, t^*, z; \xi) f_T(t \mid z; \eta)}{\sum_{k^*=0}^{K} \sum_{t^*} \sum_{\mathbf{g} \in \mathcal{G}} \int S(k^*, \mathbf{g}, t^*, z; \Omega) f_T(t^* \mid z; \eta)}, \end{aligned} \tag{2.6}$$

where $\mathcal{G}$ is the set of all possible genotypes in the population. Lobach et al. [12] proved that maximization of L_{Pseudo}, although not the actual retrospective likelihood for case-control data, leads to consistent and asymptotically normal parameter estimates. Observe that conditioning on Z in L_{Pseudo} allows it to be free of the nonparametric density function $f_Z(z)$, thus avoiding the difficulty of estimating potentially high-dimensional nuisance parameters.

2.3. Bayesian Analysis Based on Pseudolikelihood

Since in our setting the retrospectively collected data is analyzed as if they were coming from a random sample, function (2.6) is not a real likelihood function and hence the traditional Bayesian analysis is not technically correct. Conventional approaches to validity of posterior probability statements follow from the definition of the likelihood as the joint density of observations.

For simplicity of presentation we introduce new notation for this section only.

Monahan and Boos [13] introduced a definition based on coverage of posterior sets that are constructed to contain the correct probability of including a parameter τ, if the underlying distribution of τ is the prior $p(\tau)$ and the model $f(Y \mid \tau)$ of data Y is correct. This approach has been used in gene-environment interaction setting [3]. For example, in the one-dimensional case, the natural posterior coverage set functions are the one-sided intervals $I_\alpha^* = R_\alpha(Y) = (-\infty, \tau_\alpha^*)$, where τ_α^* is α-percentile of the posterior $f(Y \mid \tau)$. Validity for such a posterior means that all these intervals I_α^* have the correct coverage α. In practice, it is often challenging to verify the required probability analytically. Monahan and Boos [13] proposed a convenient numerical method. Briefly, define τ_k, $k = 1, \ldots, m$ to be a sample generated independently from a continuous prior $p(\tau)$. For each τ_k, let Y^k denote a value generated from $f(Y \mid \tau_k)$. In addition, for each k, define H_k to be a variable in the following form:

$$H_k = \int_{-\infty}^{\tau_k} f\left(\tau \mid Y^k\right) d\tau. \tag{2.7}$$

This corresponds to posterior coverage set functions of the form $(-\infty, \tau_\alpha^k)$, where τ_α^k is the αth percentile point of posterior density $f(\tau \mid Y^k)$. Monahan and Boos [13] argued that if the distribution of H_k fails to follow the uniform distribution for any prior, then the likelihood function cannot be a coverage proper Bayesian likelihood.

We propose to use the methodology described above to validate the likelihood function and apply conventional MCMC techniques to estimate parameters. We note that the method developed by Monahan and Boos is devised to invalidate a pseudolikelihood. Therefore to validate a pseudolikelihood, we propose to consider a comprehensive set of scenarios to examine coverage probabilities of posterior sets, and if these scenarios fail to invalidate a pseudolikelihood, we suppose that it is valid.

2.4. Fully Bayesian Model

We consider the case when the environmental covariates (T, X), genetic variant G, and disease status D are binary. Let $\text{pr}(G = 1) = \theta$, $\text{pr}(T = 1) = \eta$. For simplicity of presentation, consider an additive model. Define the vector of risk parameters $\mathcal{B} = (\beta_t, \beta_A, \beta_B, \beta_{tA}, \beta_{tB})^T$. Consider a multiplicative interaction and let $m(t, g, \mathcal{B}) = \beta_t t + \beta_A A + \beta_{tA} tA + \beta_B B + \beta_{tB} tB$. Make the following definition:

$$S(d, g, t, \mathcal{B}, \theta) = \frac{\exp\left[I_{(d\geq 1)}(d)\left\{\kappa_d + m(t, g, \mathcal{B})\right\}\right]}{1 + \exp\left\{\beta_0 + m(t, g, \mathcal{B})\right\}} \theta^g (1-\theta)^{1-g}. \tag{2.8}$$

If X is an observed environmental covariate prone to misclassification, denote the misclassification probabilities as $\text{pr}(X = 1 \mid T = 0) = \xi_1$ and $\text{pr}(X = 0 \mid T = 1) = \xi_0$. Hence, the distribution of misclassification process is $f_{\text{mem}}(x \mid t, \xi_0, \xi_1) = \{x\xi_1 + (1-x)(1-\xi_1)\}(1-t) + \{x(1-\xi_0) + (1-x)\xi_0\}t$.

On the risk parameters, we impose a normal prior with mean $\mu_{\mathcal{B}}$ and covariance matrix $\Sigma_{\mathcal{B}}$.

Similarly to the appendix in Fan and Xiong [14] and Lobach et al. [3], the following expectations, variances, and covariances can be derived. $E(A_i) = P_{M_i} - P_{m_i}$, $E(B_i) = 0$, $\text{Var}(A_i) = 2P_{M_i}P_{m_i}$, $\text{Var}(B_i) = P_{M_i}^2 P_{m_i}^2$, $\text{Cov}(A_i, A_j) = 2\Delta_{M_iM_j}$, $\text{Var}(B_i, B_j) = \Delta_{M_iM_j}^2$, $i \neq j$. And $\text{Cov}(A_i, B_i) = 0$ for all i and j;

$$\mathbf{V_A} = 2\begin{pmatrix} P_{M_1}P_{m_1} & \Delta_{M_1M_2} & \cdots & \Delta_{M_1M_I} \\ \Delta_{M_1M_2} & P_{M_2}P_{m_2} & \cdots & \Delta_{M_2M_I} \\ \vdots & \vdots & \cdots & \vdots \\ \Delta_{M_1M_I} & \Delta_{M_2M_I} & \cdots & P_{M_I}P_{m_I} \end{pmatrix}, \qquad \mathbf{V_D} = \begin{pmatrix} P_{M_1}^2 P_{m_1}^2 & \Delta_{M_1M_2}^2 & \cdots & \Delta_{M_1M_I}^2 \\ \Delta_{M_1M_2}^2 & P_{M_2}^2 P_{m_2}^2 & \cdots & \Delta_{M_2M_I}^2 \\ \vdots & \vdots & \cdots & \vdots \\ \Delta_{M_1M_I}^2 & \Delta_{M_2M_I}^2 & \cdots & P_{M_I}^2 P_{m_I}^2 \end{pmatrix}. \tag{2.9}$$

Define $\mathcal{A} = (A_1, \ldots, A_I)$ and $\mathcal{B} = (B_1, \ldots, B_I)$. Let $\mathbf{O_I}$ be a $I \times I$ matrix with zero elements. Based on the expectations and covariances described above, we have $\text{Cov}(\mathcal{A}, \mathcal{B}) = \begin{pmatrix} \mathbf{V_A} & \mathbf{O_I} \\ \mathbf{O_I} & \mathbf{V_D} \end{pmatrix}$.

In the case when misclassification is large, the sampling distribution of risk parameter estimates is likely to be skewed [11, 17]. However, because the shape of the normal distribution is symmetric, this prior is likely to bring the sampling distribution of the risk parameter estimates closer to normal. For the frequency parameters η and θ, we use

noninformative uniform (0,1) priors. In this setting, the prior information imposed on θ is noninformative. If *a priori* information is available about the genotype frequencies, it can be specified using a corresponding distribution or HWE.

Then, the joint posterior distribution for the model unknowns is proportional to

$$\prod_{i=1}^{n} \frac{\sum_{t^*=0}^{1} S(d_i, g_i, t^*, \mathcal{B}, \theta) p_{\text{miss}}(x_i \mid t^*, \xi_0, \xi_1) \eta^{t^*} (1-\eta)^{1-t^*}}{\sum_{t^*=0}^{1} \sum_{d=0}^{1} \sum_{g=0}^{1} S(d_i, g_i, t^*, \mathcal{B}, \theta) \eta^{t^*} (1-\eta)^{1-t^*}} \times |\Sigma_{\mathcal{B}}|^{-1/2} \exp\left\{ -\frac{1}{2} (\mathcal{B} - \mu_{\mathcal{B}})^T \Sigma_{\mathcal{B}}^{-1} (\mathcal{B} - \mu_{\mathcal{B}}) \right\} \times I_{(0,1)}(\eta) \times I_{(0,1)}(\theta). \tag{2.10}$$

Note that in this formulation, we specify a known misclassification process. We recommend performing sensitivity analysis to see whether parameter estimates change when misclassification probabilities are specified slightly differently. Furthermore, we recommend conservative setting when LD is set to be zero as *a priori*.

3. Asymptotic Posterior Distribution

We now consider properties of an asymptotic posterior distribution based on the pseudolikelihood (2.6). MCMC techniques can be computationally challenging. Knowing the form of an asymptotic posterior distribution would ease the computational burden.

For simplicity, we suppose that the parameter ξ that controls misclassification error distribution is known, although this is not required. Denote Θ_0 and $\widehat{\Theta}_n$ to be values that maximize prior and pseudolikelihood, respectively. Let $\Psi(d, g, x, z, \Theta, \xi)$ be the derivative of $\log\{L_i(d, g, x, z, \Theta, \xi)\}$ with respect to Θ and

$$\Lambda = \sum_{d} \frac{n_d}{n} E\{\Psi(D, G, X, Z, \Omega, \eta, \xi) \mid D = d\} \times E\{\Psi(D, G, X, Z, \Omega, \eta, \xi) \mid D = d\}^T. \tag{3.1}$$

Furthermore, if $p(\Theta)$ is the prior distribution of the vector of parameters, define $l(\Theta)$ to be the derivative of $\log\{p(\Theta)\}$ with respect to Θ. Then define

$$\mathcal{L}_n(\Theta, \xi) = \sum_{i=1}^{n} \Psi(D_i, G_i, X_i, Z_i, \Theta, \xi) \tag{3.2}$$

and matrices

$$\mathcal{I}(\Theta) = -E\left[\frac{\partial\{\mathcal{L}_n(\Theta, \xi)\}}{\partial(\Theta)}\right]; \qquad \mathcal{J}(\Theta) = -E\left[\frac{\partial\{l(\Theta)\}}{\partial(\Theta)}\right]. \tag{3.3}$$

Bernardo and Smith [18] showed that under suitable regularity conditions the posterior distribution of vector of parameters $\widehat{\Theta}_n$ converges to normal $\mathcal{N}(\mathcal{M}, \Sigma)$ distribution. Mean vector and covariance matrix can be consistently estimated as follows:

$$\begin{aligned} \widehat{\mathcal{M}_n} &= \widehat{\Sigma}_n^{-1}\left\{\mathcal{J}\left(\widehat{\Theta}_n\right)\widehat{\Theta}_n + \mathcal{J}(\Theta_0)\Theta_0\right\}, \\ \widehat{\Sigma}_n &= \left\{\mathcal{J}(\widehat{\Theta}_n) + J(\Theta_0)\right\}^{-1}. \end{aligned} \tag{3.4}$$

It can be easily seen that $n^{-1}\partial\{\mathcal{L}_n(\widehat{\mathcal{B}}, \xi)\}/\partial\mathcal{B}^T$ is a consistent estimate of $\mathcal{J}(\Theta)$. Alternatively, if $\widehat{\Sigma}$ is the sample covariance matrix of the terms $\Psi(D_i, G_i, X_i, Z_i, \widehat{\mathcal{B}}, \xi)$, then $\widehat{\Sigma} + \widehat{\Lambda}$ consistently estimates $\mathcal{J}(\Theta)$.

Note that the posterior distribution has precision equal to the sum of precision provided by the observed data and the prior precision matrix. This formulation suggests an approximation, namely, that for large n, prior is small compared to the one provided by the observed data. Hence, with a large sample size, one can reduce computational burden by using the asymptotic distribution and using precision provided by the observed data while specifying the posterior distribution.

4. Simulation Experiments

We investigated the case of small $n_0 = n_1 = 350$ and large ($n_0 = n_1 = 1{,}500$) sample sizes.

We validated the pseudolikelihood function using methodology described by Monahan and Boss [13] in a few scenarios by varying sample size, effect size, and misclassification probabilities. In 96% of cases that we considered, the Kolmagorov-Smirnov test failed to reject the null hypothesis that the sample of H_k (2.7) comes from the uniform (0,1) distribution at the 0.05 significance level. Hence, we concluded that the pseudolikelihood is valid for subsequent analysis. Hence, we proceeded to estimating parameters.

We implemented Metropolis-Hastings algorithm in the following setting. On the risk parameters $\mathcal{B}$, we imposed a normal $\mathcal{N}(\mathcal{B}^{\text{mean}}, \Sigma_{\mathcal{B}})$ prior, where $\mathcal{B}^{\text{mean}}$ is equal to the pseudo-MLE estimates. To examine sensitivity of the estimates to this specification, we considered a case when $\mathcal{B}^{\text{mean}}$ is a vector of zero values. Covariance matrix was specified as a diagonal matrix with diagonal elements equal to 3^2. Alternatively, we specified the corresponding matrix according to the known structure that is a function of LD. In all of these scenarios, the results we obtained were comparable. Table 1 presents results based on $\mathcal{B}^{\text{mean}} = (0, 0, 0)$ and covariance matrix with diagonal elements equal to 3^2.

To examine performance of our approach, we performed two simulation experiments. In the first experiment, we investigated performance of Bayesian method compared to pseudo-MLE. The goal of this experiment was to examine the ability of Bayesian approach to shrink the parameter estimates towards prior when misclassification causes the estimates to have skewed distribution. In the second experiment, we examined performance of the asymptotic posterior distribution.

Experiment 1. We generated the true environmental variables T from a binomial distribution with $\text{pr}(T = 1) = 0.5$. The misclassification probabilities are $\text{pr}(X = 0 \mid T = 1) = 0.20$ and $\text{pr}(X = 1 \mid T = 0) = 0.25$. We simulated three genetic markers in LD corresponding to $\Delta = 0.03$

Table 1: Biases and root mean squared errors (RMSEs) of risk parameters for the naive approach that ignores existence of misclassification and the proposed method in the case when $\mathrm{pr}(D = 1)$ is known and when it is estimated. The results are based on 500 samples of 1,500 cases and 1,500 controls. Genotype is simulated at the three marker loci with $P_{M_i} = 0.25$, $i = 1, 2, 3$, with linkage disequilibrium corresponding to $\Delta_{M_iM_j} = 0.03$. The environmental covariate (X) is binary and measured with error with misclassification probabilities being 0.20 for exposed and 0.25 for nonexposed subjects. The data is simulated and analyzed under the genotype effect model.

Parameter	True value	Naive analysis		Pseudo-MLE		MCMC	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
κ	0.484	0.481	0.231	−0.054	0.020	−0.032	0.013
β_X	0.693	−0.351	0.132	0.014	0.039	0.008	0.021
β_{A1}	0.406	0.257	0.073	−0.011	0.016	−0.003	0.009
β_{A2}	0.789	0.194	0.046	−0.003	0.015	−0.002	0.006
β_{A3}	0.693	0.283	0.089	−0.005	0.016	−0.003	0.008
β_{AX1}	0.916	−0.425	0.193	0.039	0.046	0.017	0.025
β_{AX2}	0.693	−0.317	0.113	0.038	0.041	0.023	0.021
β_{AX3}	1.099	−0.515	0.282	0.039	0.058	0.019	0.032
β_{D1}	0.262	0.299	0.133	0.026	0.152	0.009	0.368
β_{D2}	0.095	0.258	0.105	0.005	0.099	0.003	0.039
β_{D3}	0.693	0.231	0.092	0.018	0.128	0.008	0.087
β_{DX1}	1.099	−0.495	0.326	0.018	0.301	0.006	0.093
β_{DX2}	0.916	−0.413	0.235	0.006	0.208	0.005	0.121
β_{DX3}	1.099	−0.486	0.313	0.023	0.286	0.017	0.138
P_{M_i}	0.250	<0.001	<0.001	−0.001	<0.001	<0.001	<0.001
$\mathrm{pr}(X = 1)$	0.500			0.003	0.001	<0.001	<0.001
$\mathrm{pr}(D = 1)$	0.005			0.003	<0.001	<0.001	<0.001

and $P_{M_i} = 0.25$. In the study with $1,500$ cases and $1,500$ controls, we generated a binary disease status according to the following logistic model:

$$\mathrm{logit}\{\mathrm{pr}(D = 1 \mid G, X)\} = \beta_0 + \beta_t t + \sum_{j=1}^{3} \beta_{A_j} A_j + \sum_{j=1}^{3} \beta_{B_j} B_j + \sum_{j=1}^{3} \beta_{ATj} A_j T + \sum_{j=1}^{3} \beta_{TBj} B_j T. \quad (4.1)$$

To examine the case when genetic data is missing, we simulated a similar set of 1,500 cases and 1,500 controls with 50% of genetic information missing completely at random. To investigate a smaller study, we simulated 350 cases and 350 controls with the disease status defined by the risk model with all β_{B_j} and β_{BTj} set to 0. Results presented in Tables 1 and 2 illustrate that the proposed Bayesian approach produced parameter estimates that are less variable and less biased. We examine the empirical distribution of parameter estimates based on a small sample and found that it is skewed, which may be due to small sample size and presence of misclassification. We observed this phenomena in our previous work [3, 11]. The Bayesian solution brings the advantage, that is, a symmetric prior can shrink parameter estimates towards normal distribution. Furthermore, we presented performance of the naive approach that ignores existence of misclassification.

Table 2: Biases and root mean squared errors (RMSEs) of risk parameters obtained based on pseudo-MLE and the proposed MCMC. The results are based on 500 samples of 350 cases and 350 controls. Genotype is simulated at the two marker loci with $P_{M_i} = 0.25$, $i = 1, 2$. The environmental covariate (X) is binary and measured with error misclassification probabilities being 0.20 for exposed and 0.25 for nonexposed subjects. The data is simulated and analyzed under the additive effect model and the LD measure $\Delta_{M1M2} = 0.03$.

Parameter	True value	Pseudo-MLE		MCMC	
		Bias	RMSE	Bias	RMSE
β_X	1.099	0.035	0.392	0.013	0.236
β_{A1}	0.406	−0.268	1.035	−0.079	0.397
β_{A2}	0.789	−0.319	1.062	−0.085	0.372
β_{A3}	0.693	−0.293	1.043	−0.092	0.365
β_{AX1}	0.916	0.432	1.135	0.103	0.432
β_{AX2}	0.693	0.391	1.047	0.085	0.481
β_{AX3}	1.099	0.293	1.113	0.097	0.427

Table 3: Biases and root mean squared errors (RMSEs) of risk parameters obtained based on asymptotic posterior distribution. The results are based on 500 samples of 1,500 cases and 1,500 controls. Genotype is simulated at the two marker loci with $P_{M_i} = 0.25$, $i = 1, 2$. The environmental covariate (X) is binary and measured with error with misclassification probabilities being 0.20 for exposed and 0.25 for nonexposed subjects. The data is simulated and analyzed under the additive effect model and the LD measure $\Delta_{M1M2} = 0.03$.

Parameter	True value	Bias	Estimated SE	SE
β_X	0.693	0.010	0.032	0.039
β_{A1}	0.406	−0.005	0.012	0.015
β_{A2}	0.789	−0.004	0.011	0.014
β_{A3}	0.693	−0.004	0.016	0.016
β_{AX1}	0.916	0.023	0.045	0.044
β_{AX2}	0.693	0.019	0.061	0.058
β_{AX3}	1.099	0.020	0.052	0.054
β_{D1}	0.262	0.016	0.431	0.410
β_{D2}	0.095	0.009	0.052	0.063
β_{D3}	0.693	0.013	0.099	0.100
β_{DX1}	1.099	0.011	0.013	0.015
β_{DX2}	0.916	0.013	0.025	0.027
β_{DX3}	1.099	0.016	0.027	0.030

Experiment 2. We examined performance of estimation based on the derived asymptotic posterior in the simulation setup described in Experiment 1 corresponding to $n_1 = n_2 = 1,500$. Results presented in Table 3 illustrate that the parameter estimates are nearly unbiased. Moreover, estimated variances of parameter estimates are very close to the observed variability with one exception, namely, β_x. Variability of β_x may be inflated due to the misclassification in environmental exposure.

5. Analysis of Alcohol Dependence

The Collaborative Studies on the Genetics of Alcoholism (COGA) is a nine-center nationwide study that was initiated in 1989 and has had as its primary aim the identification of genes that contribute to alcoholism susceptibility and related characteristics [19–21]. COGA is funded through the National Institute on Alcohol Abuse and Alcoholism (NIAAA). The focus of this study is a case-control design of unrelated individuals for a genetic association analysis of addiction. Analyses that include incorporation of important demographic and environmental factors such as age when first got drunk, sex, income, and education into association studies are pursued. Our project involves analysis of 40 SNPs residing in genes involved in dopamine pathways. Specifically, we consider D2 dopamine receptor gene (DRD2) encoding a protein which plays a central role in reward-mediating mesocorticolimbic pathways; a member of the immunoglobulin gene superfamily NCAM1 encoding protein involved in various neural functions; tetratricopeptide repeat domain 12 gene (TTC12); CHRNA3 gene shown to be involved in higher craving after quitting and increased withdrawal symptoms over time. Cases are defined as individuals with DSM-IV alcohol dependence (lifetime). Controls are defined as individuals who have been exposed to alcohol, but have never met lifetime diagnosis for alcohol dependence or dependence on other illicit substances. The sample consists of 50.7% of male and 49.3% female participants; 60% report their race as Caucasian and 40% are non-Caucasian. We categorized age when first got drunk as "Early" if it is less or equal to 13 (EAD = 1, 45.2% of all participants) and people with low income are the ones who make less than 30 K per year (LI = 1, 45% of all participants).

Define T to be the true unobserved indicator of early drinking, that is, $T = 1$ corresponds to the early onset of drinking, $T = 0$ to the late onset. Let X be the observed value of the early onset of drinking. Because we do not have external data or internal replicates to estimate misclassification probability, we performed sensitivity analysis for various values of misclassification.

We used the following risk model:

$$\operatorname{logit}\{\operatorname{pr}(D = 1 \mid G = (A, B), T)\} = \beta_0 + \beta_t T + \beta_A A + \beta_B B + \beta_{AT} AT + \beta_{BT} BT. \tag{5.1}$$

The results of sensitivity analysis (not shown) suggest that when $\operatorname{pr}(X = 0 \mid T = 1)$ is ignored or underestimated, the interaction effect is not significant. The setting corresponds to the case when exposed subjects are defined as nonexposed, thus reducing the association signal. However, the estimation procedure appears to be robust to underestimation of $\operatorname{pr}(X = 1 \mid T = 0)$. This scenario corresponds to the case when a nonexposed subject is considered to be exposed.

Parameter estimates obtained using our method corresponding to $\operatorname{pr}(X = 0 \mid T = t) = 0.25$ and $\operatorname{pr}(X = 1 \mid T = 0) = 0.25$ are presented in Table 4 demonstrating significant interaction between various genetic markers and early onset of drinking.

6. Discussion

Motivated by concerns in the analysis of gene-environment interactions, we proposed a genotype-based Bayesian approach for the analysis of case-control studies when environmental exposure cannot be observed directly and is subject to misclassification. The formulation of risk functions and the estimation procedure are along the lines of our previous work:

Table 4: Risk parameter estimates and standard errors in the alcohol dependence data.

Gene, SNP	Estimate of log(OR)	Standard error
NCAM1, rs586903	1.78	0.06
NCAM1, rs2303377	2.58	0.11
NCAM1, rs2156485	1.87	0.07
TTC12, rs7103866	2.21	0.03
TTC12, rs723077	1.92	0.03
TTC12, rs2288159	2.21	0.01
CHRNA3, rs1051730	1.77	0.03
CHRNA3, rs8192475	1.62	0.02

genotype and additive effect models [14, 15] and pseudolikelihood approach [3, 11, 12]. The risk function of genotype effect model involves both the additive and dominance effect while taking into account possible interactions between genes expressed in terms of interaction between their additive and dominance components, while the additive effect model only involves the additive effect and possible interactions. The additive effect model contains less parameters than the genotype effect model. In applications, the additive effect models should be used in analyzing data as the first step. If the dominance effect is strong enough to compensate the increase of the number of the parameters in the genotype effect models, one may use the genotype effect models.

The proposed method has several unique advantages. First, the observed genetic information enters the model directly and the LD structure is captured in the regression coefficients. This aspect offers advantages from the practical point of view, the computational burden is less demanding because haplotype phase need not to be estimated. In the cases when LD is moderate, which is the focus of our work, the computational demands can be substantial even with the current state of technology. Furthermore, the risk due to uncertainty associated with the haplotype phase estimation can be avoided. Second, the estimating procedure is based on a pseudolikelihood model, similarly to the method investigated previously, that allows efficient estimation of parameters, models environmental covariates completely nonparametrically, and incorporates information about the probability of disease [3, 11, 12]. In epidemiologic studies, the vector of environmental covariates measured exactly is often, high dimensional, and a good estimate about probability of disease in a population is known. Additionally, the Bayesian formulation of the proposed method allows shrinking parameter estimates towards prior which offers advantage in cases when misclassification is present.

Because of the Bayesian formulation and the need to validate posterior sets obtained using a pseudolikelihood, the proposed method can be highly computationally intensive. Moreover, the validation of pseudolikelihood requires evaluation of ratio of two likelihood functions. For example, in our simulation experiments and data analysis, this part required us to obtain a precise value of ratios similar to $\exp(3000)/\exp(2908)$. Hence, we employed GNU Multiple Precision Arithmetic Library (http://gmplib.org/).

The form of our pseudolikelihood function is complex and it does not seem feasible to validate a pseudolikelihood function algebraically. Instead, we propose to apply Monahan and Boos method to examine coverage probabilities of posterior sets. If a comprehensive set of scenarios fails to invalidate a pseudolikelihood function, we suppose that the pseudolikelihood is valid. This reasoning may be similar to the conventional hypothesis testing where the null hypothesis is assumed to be true (pseudolikelihood is valid), and the observed

data is used to quantify evidence in favor of the alternative hypothesis (pseudolikelihood is not valid). Of course, a strong basis for validity of a pseudolikelihood is needed. We employ the following arguments. Our previous research approach [3, 11, 12] demonstrated validity of this pseudolikelihood in frequentist sense, that is, we have shown that estimation and inferences are correct when this pseudolikelihood is used in place of a real likelihood function. Hence, posterior distribution based on a pseudolikelihood may be invalid only for certain prior distributions. Therefore, to invalidate a pseudolikelihood, one should find a prior distribution for which the posterior is not valid. However, in our setting, the number of possible prior settings is narrow, because what we advocate is the use of symmetry of prior distribution as a way to improve precision of estimation and inference. We are restricting the prior of regression coefficients to be Gaussian and advocate mean zero and large variance. While one can try other priors for other parameters, the number of possible prior settings is still reasonable and it is practically feasible to look at their performance in terms of probability of coverage sets.

While the major motivation of the proposed work is dictated by the need of a symmetric prior on risk coefficients, other types of *a priori* information can enter our model. For example, if *a priori* information about the LD structure is available, it can be modeled in the *a priori* distribution. Furthermore, if misclassification probabilities are not known precisely, one can specify uncertainty associated with values of misclassification.

A major practical advantage of this proposed work is that it allows the model to exploit recent advances in genotyping technology. Specifically, with the recent advances genetic markers become more and more densely typed and multiple markers are likely to be observed in a functional unit of interest. These units of interest may be defined in terms of LD blocks using information available in linkage maps. While in situations when linkage disequilibrium is strong, the haplotype-based analysis is advantageous; in more common scenarios when linkage disequilibrium is moderate, our approach provides advantages.

However, in the context when the number of genetic markers in a functional unit of interest is large our methodology may require model averaging and model selection component. Hence, behavior of this pseudolikelihood needs to be examined in this setting. A practical strategy can be that one starts with screening analysis first to get interesting genetic variants and SNPs using traditional methods which is computationally less demanding. Then, one may apply the proposed approaches for possible gene-environment interactions and further investigations by focusing on these important and interesting genetic variants and SNPs.

Acknowledgments

R. Fan was supported by the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Maryland, USA. Genetics and Environment (SAGE) was provided through the NIH Genes, Environment and Health Initiative (GEI) (U01 HG004422). SAGE is one of the genomewide association studies funded as part of the Gene Environment Association Studies (GENEVA) under GEI. Assistance with phenotype harmonization and genotype cleaning as well as with general study coordination was provided by the GENEVA Coordinating Center (U01 HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Support for collection of datasets and samples was provided by the Collaborative Study on the Genetics of Alcoholism (COGA; U10 AA008401), the Collaborative Genetic Study of Nicotine Dependence (COGEND; P01 CA089392), and

the Family Study of Cocaine Dependence (FSCD; R01 DA013423). Funding support for genotyping, which was performed at the Johns Hopkins University Center for Inherited Disease Research, was provided by the NIH GEI (U01HG004438), the National Institute on Alcohol Abuse and Alcoholism, the National Institute on Drug Abuse, and the NIH contract "High throughput genotyping for studying the genetic contributions to human disease" (HHSN268200782096C). The datasets used for the analyses described in this paper were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000092.v1.p1. This work has utilized computing resources at the High Performance Computing Facility of the Center for Health Informatics and Bioinformatics at New York University Langone Medical Center.

References

[1] D. Thomas, "Gene-environment-wide association studies: emerging approaches," *Nature Reviews Genetics*, vol. 11, no. 4, pp. 259–272, 2010.

[2] J. N. Hirschhorn, K. Lohmueller, E. Byrne, and K. Hirschhorn, "A comprehensive review of genetic association studies," *Genetics in Medicine*, vol. 4, no. 2, pp. 45–61, 2002.

[3] I. Lobach, B. Mallick, and R. J. Carroll, "Semiparametric Bayesian analysis of gene-environment interactions with error in measurement of environmental covariates and missing genetic data," vol. 4, no. 3, pp. 305–316, 2011.

[4] S. Lin, D. J. Cutler, M. E. Zwick, and A. Chakravarti, "Haplotype inference in random population samples," *American Journal of Human Genetics*, vol. 71, no. 5, pp. 1129–1137, 2002.

[5] J. Marchini, D. Cutler, N. Patterson et al., "A comparison of phasing algorithms for trios and unrelated individuals," *American Journal of Human Genetics*, vol. 78, no. 3, pp. 437–450, 2006.

[6] Z. S. Qin, T. Niu, and J. S. Liu, "Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms," *American Journal of Human Genetics*, vol. 71, no. 5, pp. 1242–1247, 2002.

[7] M. Stephens and P. Donnelly, "A comparison of bayesian methods for haplotype reconstruction from population genotype data," *American Journal of Human Genetics*, vol. 73, no. 5, pp. 1162–1169, 2003.

[8] M. Stephens, N. J. Smith, and P. Donnelly, "A new statistical method for haplotype reconstruction from population data," *American Journal of Human Genetics*, vol. 68, no. 4, pp. 978–989, 2001.

[9] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu, *Measurement Error in Nonlinear Models Edition*, Chapman & Hall CRC Press, 2nd edition, 2006.

[10] A. Agrawal, C. E. Sartor, M. T. Lynskey et al., "Evidence for an interaction between age at first drink and genetic influences on DSM-IV alcohol dependence symptoms," *Alcoholism*, vol. 33, no. 12, pp. 2047–2056, 2009.

[11] I. Lobach, R. J. Carroll, C. Spinka, M. H. Gail, and N. Chatterjee, "Haplotype-based regression analysis and inference of case-control studies with unphased genotypes and measurement errors in environmental exposures," *Biometrics*, vol. 64, no. 3, pp. 673–684, 2008.

[12] I. Lobach, R. Fan, and R. J. Carroll, "Genotype-based association mapping of complex diseases: gene-environment interactions with multiple genetic markers and measurement error in environmental exposures," *Genetic Epidemiology*, vol. 34, no. 8, pp. 792–802, 2010.

[13] J. F. Monahan and D. D. Boos, "Proper likelihoods for Bayesian analysis," *Biometrika*, vol. 79, no. 2, pp. 271–278, 1992.

[14] R. Fan and M. Xiong, "High resolution mapping of quantitative trait loci by linkage disequilibrium analysis," *European Journal of Human Genetics*, vol. 10, no. 10, pp. 607–615, 2002.

[15] R. Fan, J. Jung, and L. Jin, "High-resolution association mapping of quantitative trait loci: a population-based approach," *Genetics*, vol. 172, no. 1, pp. 663–686, 2006.

[16] M. K. Ho and R. F. Tyndale, "Overview of the pharmacogenomics of cigarette smoking," *Pharmacogenomics Journal*, vol. 7, no. 2, pp. 81–98, 2007.

[17] D. W. Schafer and K. G. Purdy, "Likelihood analysis for errors-in-variables regression with replicate measurements," *Biometrika*, vol. 83, no. 4, pp. 813–824, 1996.

[18] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*, John Wiley & Sons, Chichester, UK, 1994.

[19] H. J. Edenberg, "The collaborative study on the genetics of alcoholism: an update," *Alcohol Research and Health*, vol. 26, no. 3, pp. 214–218, 2002.

[20] L. J. Bierut, N. L. Saccone, J. P. Rice et al., "Defining alcohol-related phenotypes in humans: the collaborative study on the genetics of alcoholism," *Alcohol Research and Health*, vol. 26, no. 3, pp. 208–213, 2002.

[21] H. J. Edenberg and T. Foroud, "The genetics of alcoholism: identifying specific genes through family studies," *Addiction Biology*, vol. 11, no. 3-4, pp. 386–396, 2006.