Statistics and Its Interface Volume 4 (2011) 305–315

Semiparametric Bayesian analysis of gene-environment interactions with error in measurement of environmental covariates and missing genetic data

Iryna Lobach*, Bani Mallick and Raymond J. Carroll

Case-control studies are widely used to detect gene-environment interactions in the etiology of complex diseases. Many variables that are of interest to biomedical researchers are difficult to measure on an individual level, e.g. nutrient intake, cigarette smoking exposure, long-term toxic exposure. Measurement error causes bias in parameter estimates, thus masking key features of data and leading to loss of power and spurious/masked associations. We develop a Bayesian methodology for analysis of case-control studies for the case when measurement error is present in an environmental covariate and the genetic variable has missing data. This approach offers several advantages. It allows prior information to enter the model to make estimation and inference more precise. The environmental covariates measured exactly are modeled completely nonparametrically. Further, information about the probability of disease can be incorporated in the estimation procedure to improve quality of parameter estimates, what cannot be done in conventional case-control studies. A unique feature of the procedure under investigation is that the analysis is based on a pseudo-likelihood function therefore conventional Bayesian techniques may not be technically correct. We propose an approach using Markov Chain Monte Carlo sampling as well as a computationally simple method based on an asymptotic posterior distribution. 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 population-based case-control study of the association between calcium intake with the risk of colorectal adenoma development.

Keywords and phrases: Bayesian inference, Errors in variables, Gene-environment interactions, Markov Chain Monte Carlo sampling, Missing data, Pseudo-likelihood, Semiparametric methods.

*Corresponding author.

1. INTRODUCTION

A key component to prevention and control of complex diseases, such as cancer, diabetes, hypertension, is to analyze genetic and environmental factors that lead to the development of these complex diseases. The analysis is complicated by the fact that the genetic and environmental factors interplay while causing complex diseases (Hunter, 2005). Such gene-environment interactions have the potential to (1) yield insight into the mechanism of action of the environment under various settings of the genetic background; (2) suggest disease prevention strategies; (3) obtain a better estimate of the population-attributable risk for genetic and environmental risk factors by accounting for their joint interaction; and (4) result in improved analysis of the association between environmental factors and complex disease by examining factors in genetically susceptible individuals. A challenge in statistical analyses is that a weak overall association may mask important genetic susceptibility to the effects of the environmental exposure in the population subgroups. Separate estimation of the contributions of genes and environment and ignoring their interaction will lead to an incorrect estimate of the proportion of the disease (the population attributable risk) that is explained by genes, environment, and their joint effect (Hunter, 2005). Restricting analysis of environmental exposure to individuals who are genetically susceptible to the exposure is likely to increase the magnitude of relative risk, thus improving the ability to detect association signal.

Further, many variables that are of interest to biomedical researchers, such as dietary intake and cigarette smoking exposure are very difficult to measure on individuals. Measurement error causes bias in gene-environment parameter estimates, thus masking key features of data and leading to loss of power and spurious/masked associations (Lobach, et al., 2008). Loss of power prevents the ability to detect important relationships among variables (Carroll, et al. 2006). For example, nutrition — defined broadly to indicate diet, body size, physical activity — is likely to be causally related to cancer (Schatzkin, et al., 2009). Nevertheless, nutritional epidemiology of cancer remains problematic, largely because of persistent concerns that standard instruments measure

diet and physical activity with too much error. While it is recognized that information collected about dietary level contains error, considerable uncertainty remains about their qualitative and quantitative characteristics (Subar, et al. 2003). Understanding this error is critical to interpreting findings and surveillance research efforts.

In this paper, we develop a Bayesian methodology for analysis of case-control data in the situation when measurement error is present in an environmental covariate as well as the genetic variable contains missing data (unobserved genotype or haplotype-phase ambiguity). Conventionally, case-control data are analyzed using prospective logistic regression ignoring the fact that under this design subjects are sampled into the study conditionally on their disease status. The validity of this approach relies on the classic results by Cornfield (1956) who showed the equivalence of prospective- and retrospective odds-ratios. The efficiency of the approach was established in two other classic papers by Andersen (1970) and Prentice and Pyke (1979). Recently, Chatterjee and Carroll (2005), Spinka, et al. (2005), and Lobach, et al. (2008, 2010) developed an efficient approach for analysis of case-control studies, the key idea of which is to treat retrospectively collected data as if they were coming from a random sample. Because the retrospectively collected data are analyzed as if they were coming from a random sample, the conventional Bayesian techniques are not valid. The pseudo-likelihood function employed in our analysis is not the same as the conventional prospective likelihood. Validity of the Bayesian analysis needs to be examined when the proposed likelihood function is not a proper likelihood (Monahan and Boos, 1992). Lazar (2003) has examined the validity of Bayesian empirical likelihood based methods. We followed Monahan and Boos (2003) to validate our Bayesian approach under this pseudo-likelihood function and exploit it to obtain posterior inferences about the unknown parameters. Due to the complexity of the pseudo-likelihood function, the posterior distribution of the parameters is not in explicit form, therefore Markov Chain Monte Carlo (MCMC) algorithms are required to sample from this posterior distribution to make necessary inference.

Our work is motivated by a case-control study of colorectal adenoma (Peters et al., 2004). Briefly, the available data consist of measures of dietary calcium intake obtained by a food frequency questionnaire (FFQ), genotype data for three SNPs in the calcium receptor gene CaSR, and various individual-level data such as age, sex and race. The main interest is in studying the interaction between CaSR haplotypes and dietary calcium intake.

The Colorectal Adenoma study thus has unique features, specifically the following.

- First, genetic information is missing. We wish to model the effect of CaSR haplotypes, but these are not observed, and instead we have unphased haplotype information in the form of the three SNPs.
- Second, one of the environmental variables (calcium intake) is subject to substantial measurement error because of the use of a FFQ. It is well known that the FFQ as a measure of long-term diet is subject both to biases and random errors, as illustrated in the OPEN study (Subar, et al., 2003).

FFQs as a measure of long-term diet result in massive amounts of measurement error. It is well known (Schafer and Purdy, 1996) that huge measurement error often results in skewed sampling distribution of parameter estimates and the skewness is more pronounced for small sample sizes. Hence possibly both estimation and inference are not precise (Carroll, et al., 2006). In our motivating example the situation is further complicated by the fact that not only massive amount of measurement error is present in the environmental covariate, furthermore the genotype contains missing values.

We develop a Bayesian approach utilizing the pseudo-likelihood function to quantify the uncertainty of the model parameters exactly. Our approach has the ability to shrink the parameter estimates towards prior using a proper prior distribution and hence reduce variability of these estimates. Moreover, Bayesian methods can incorporate available prior historical or biological information to make inferences more precise. For example, the proposed pseudo-likelihood function allows to incorporate prior information about the probability of disease, which cannot be done in a standard analysis. Typically a good estimate of a probability of disease is available a priori. This information can be used to improve estimation of parameters, especially the intercept.

Our approach is general enough to accommodate any complicated pseudo-likelihood function and use MCMC techniques to obtain the parameter estimates with uncertainty bounds. The method will be particularly useful when this pseudo-likelihood function is multimodal (since MCMC can search the modes) or when the solution lies on the boundary (since prior can constrain the solution space). When the sample size is small or measurement error is massive, the non-Gaussian behavior of an estimate is very common. In terms of the Bayesian model and MCMC based computation, we can perform exact analysis and capture these non-Gaussian behaviors. On the other hand, when the sample size is large enough, we can derive the asymptotic posterior distribution which will ease the computation burden.

Alternative semiparametric Bayesian approach will be to assign Dirichlet process or some other nonparametric prior processes to model the unknown joint distribution of the covariates without measurement error and perform full Bayesian analysis using MCMC (Müller and Roeder, 1997; Sinha et al., 2005). In this process we need to estimate potentially high dimensional nuisance parameters and the MCMC algorithms are computationally demanding. In addition, the analysis could be sensitive towards the specification of the hyper-parameters of these nonparametric processes.

Our approach avoids the complete specification of this distribution, hence reduces the computational burden significantly. Furthermore, we can obtain the asymptotic posterior distribution of the parameters and can avoid MCMC in those situations. However it requires a validation step which could be computationally intensive.

An outline of this paper is as follows. In Section 2 we introduce notation and formally state the problem. Section 3 presents the proposed methodology for parameter estimation based on a pseudo-likelihood function. In Section 4 we describe the full Bayesian model under various scenarios. In Section 5 we derive an asymptotic posterior distribution. Section 6 gives the results of simulation studies, where we show that our methodology results in parameter estimates that are nearly unbiased and error rates close to the nominal. Section 7 analyzes the Colorectal Adenoma Data discussed above. Section 8 gives concluding remarks.

2. NOTATION AND PSEUDO-LIKELIHOOD FUNCTION

Suppose a sample consists of n_0 controls and n_d cases with disease stage $d = 1, 2, \ldots, K$ to accommodate different subtypes of disease. Let $H = (H_1, H_2)$ denote the diplotype status, that is, the two haplotypes that a subject carries at the loci of interest on the pair of homologous chromosomes. Note that typically multilocus genotype data $\mathbf{G} = (G_1, \ldots, G_M)$ are available. Due to lack of haplotype-phase information, multiple configurations of haplotypes can be consistent with the same genotype data. For example, if A/a and B/b denote the major/minor alleles in two bi-allelic loci (e.g. single nucleotide polymorphisms), then subjects with genotypes (Aa) and (Bb) at the first and the second locus, respectively, are considered phase ambiguous: their genotypes could arise from either the haplotype-pair (A-B, a-b), or the haplotype pair (A-b, a-B). Humans are diploid individuals and a pair of haplotypes that a person carries is called diplotype. Let $\mathcal{H}$ denote the set of all possible diplotypes in the underlying population and $\mathcal{H}_{\mathbf{G}}$ denote the set of all possible diplotypes that are consistent with a particular genotype $\mathbf{G}$. We impose a parametric structure on the genetic covariate of interest in the form $\text{pr}(H) = Q(H, \theta)$. In our example we used Hardy-Weinberg Equilibrium (HWE) of the following form.

$$Q(H,\theta) = \text{pr}\{H = (h_j, h_k)|\theta\} = \theta_k^2, \quad \text{if } h_j = h_k;$$
$$= 2\theta_k\theta_j, \quad \text{if } h_j \neq h_k.$$

However, the methodology is general enough to allow various parametric forms of $Q(H, \theta)$. For instance, it is possible to introduce a parameter that models departure from the HWE. Alternatively, one can specify a parametric distribution of H given (X, Z) that could account for gene-environment association (Chatterjee, et al. 2006).

Let (X, Z) denote all of the environmental (non-genetic) covariates of interest with X denoting the factors susceptible to measurement errors. We assume that H 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. 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. Further, 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.

Recall that the environmental covariate X is measured with error. Let W denote the error-prone version of X. We assume a parametric model of the form $f_{\text{mem}}(w|X, H, Z, D; \xi)$ for the conditional distribution of W given the true exposure X, additional environmental factors Z and disease-status D. This model is general enough to capture a differential on the disease status, genetic and other environmental variables $f_{\text{mem}}(w|X, H, Z, D; \xi)$ can be estimated using replications or an external study. We assume that the joint distribution of the environmental factors in the underlying population can be specified according to a semiparametric model of the form $f_{X,Z}(x, z) = f_X(x|z, \eta)f_Z(z)$, where $f_Z(z)$ is left completely unspecified, thus avoiding the need to estimate potentially high-dimensional nuisance parameters.

Given the environmental covariates X and Z and diplotype data H, the risk of the disease in the underlying population is given by the polytomous logistic regression model

$$\text{pr}(D = d|H, X, Z) = \frac{\exp\{\beta_{0d} + m(H, X, Z, \beta)\}}{1 + \sum_{j=1}^{K} \exp\{\beta_{0j} + m(H, X, Z, \beta)\}}, \quad d \geq 1.$$

Here $m(\cdot)$ is a known function parameterizing the joint risk of the disease from H^{dip}, X and Z in terms of the odds-ratio parameters β. Define n_d be the number of subjects with disease status d. Let $\pi_d = \text{pr}(D = d)$, $\kappa_d = \beta_{0d} + \log(n_d/n_0) - \log(\pi_d/\pi_0)$, and $\widetilde{\kappa} = (\kappa_1, \ldots, \kappa_K)^{\text{T}}$. Define $\kappa_0 = \beta_{00}$. Let $\widetilde{\beta}_0 = (\beta_{01}, \ldots, \beta_{0K})^{\text{T}}$. Let $\Omega = (\widetilde{\beta}_0^{\text{T}}, \beta^{\text{T}}, \Theta^{\text{T}}, \widetilde{\kappa}^{\text{T}})^{\text{T}}$, $\mathcal{B} = (\Omega^{\text{T}}, \eta^{\text{T}})^{\text{T}}$. Define $I_{(d\geq 1)}(d)$ be the indicator function. Make the definition

$$S(d, h, x, z, \Omega) = \frac{\exp\left[I_{(d\geq 1)}(d)\{\kappa_d + m(h, x, z, \beta)\}\right]}{1 + \sum_{j=1}^{K} \exp\{\beta_{0j} + m(h, x, z, \beta)\}} Q(h, \theta).$$

Consider a sampling scenario where each subject from the underlying population is selected into the case-control study using a Bernoulli sampling scheme, where the selection probability for a subject given his/her disease status $D = d$ is proportional to $n_d/\text{pr}(D = d)$. Let $R = 1$ denote

the indicator of whether a subject is selected in the case-control sample under the above Bernoulli sampling scheme.

Lobach, et al. (2008) proposed to use the following function in place of the likelihood function, that is to ignore the retrospective design and analyze the data as if it were coming from a random sample. The outlined above Bernoulli sampling connects the retrospective design employed to collect data and the pretend random sample scheme.

$$L_n(d, g, w, z, \mathcal{B}, \xi) = \prod_{i=1}^{n} L_i(d, g, w, z, \mathcal{B}, \xi); \tag{1}$$

where

$$L_i(d, g, w, z, \mathcal{B}, \xi) = \mathrm{pr}(D = d_i, W = w_i, \mathbf{G} = g_i | Z = z_i, R = 1)$$
$$= \frac{\int \sum_{h^* \in \mathcal{H}_G} S(d_i, h^*, x, z_i, \Omega) f_{\mathrm{mem}}(w_i | d_i, h^*, x, z_i, \xi) f_X(x | z_i, \eta) dx}{\int \sum_{d_*=0}^{K+1} \sum_{h_* \in \mathcal{H}} S(d_*, h_*, x, z_i, \Omega) f_X(x | z_i, \eta) dx}.$$

Recall that $S(d, h, x, z, \Omega)$ is a product of the disease risk function and the density of a genetic variable; $f_{\mathrm{mem}}(w|d, h, x, z, \xi)$ defines the measurement error process; and $f_X(x|z, \eta)$ is the density of environmental variables measured with error. Further, recall that $\mathcal{H}$ is the set of all possible haplotypes, $\mathcal{H}_G$ - the set of all haplotypes consistent with the observed genotype G.

It was shown (Lobach, et al. 2008) that maximization of L_n, although not the actual retrospective-likelihood for case-control data, leads to consistent and asymptotically normal parameter estimates. Note that conditioning on Z in L_n allows it to be free of the nonparametric density function $f_Z(z)$. In epidemiologic studies the vector of observations Z is likely to be multidimensional (e.g., age, bmi, race) hence this formulations allows gains of efficiency by not having to model parameters associated with these variables.

3. SEMIPARAMETRIC BAYESIAN ESTIMATION BASED ON PSEUDO-LIKELIHOOD

Since in our setting the retrospectively collected data is analyzed as if they were coming from a random sample, the function (1) 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.

Monahan and Boos (1992) introduced a definition based on coverage of posterior sets that are constructed to contain the correct probability of including a parameter θ, if the underlying distribution of θ is the prior $p(\theta)$, and the model of data X $f(X|\theta)$ are correct. For example, in the one-dimensional case, the natural posterior coverage set functions are the one-sided intervals $I^*_\alpha = R_\alpha(X) = (-\infty, \theta^*_\alpha)$, where θ^*_α is α-percentile of the posterior $f(X|\theta)$. Validity for such a posterior then 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 (1992) proposed a convenient numerical method. Briefly, define θ_k, $k = 1, \ldots, m$ to be a sample generated independently from a continuous prior $p(\theta)$ and for each θ_k let X^k denote a value generated from $f(X|\theta_k)$. Further, for each k define H_k to be a variable in the following form

$$H_k = \int_{-\infty}^{\theta_k} f(\theta | X^k) d\theta. \tag{2}$$

This corresponds to posterior coverage set functions of the form $(-\infty, \theta^k_\alpha)$, where θ^k_α is the αth percentile point of posterior density $f(\theta|X^k)$. Monahan and Boos (1996) 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.

4. SEMIPARAMETRIC BAYESIAN ANALYSIS OF CASE-CONTROL DATA

The Bayesian modeling framework described above provides a conceptually elegant and general method to model gene-environment interactions. Practical implementation requires specification of a prior distribution and computations based on the corresponding posterior distribution. In this section we describe a Bayesian model including likelihood and prior distribution for two cases. The first scenario is based on a setting where all variables are binary. In the second case we model a continuous environmental covariate, e.g., calcium intake. Moreover, the genetic covariate is in the form of a haplotype. In both scenarios, we validate the likelihoods using ideas of Monahan and Boos (1992) as explained in the Section 3.

4.1 Genotype-based case-control studies

Within this setting we consider the case when the environmental covariates (X, W), genetic variant (G) and disease status (D) are binary. Let $\mathrm{pr}(G = 1) = \theta$, $\mathrm{pr}(X = 1) = \eta$. This setting arises in the case when the genetic effect is recessive or dominant. Define the vector of risk parameters $\mathcal{B} = (\beta_x, \beta_g, \beta_{xg})^{\mathrm{T}}$. Suppose that the genotype and environment are independent in the population but they do work together while causing a disease thus creating an interaction. Consider a multiplicative interaction and let $m(x, g, \mathcal{B}) = \beta_g g + \beta_x x + \beta_{xg} xg$. Make the following definition.

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

If W is an observed environmental covariate, denote the misclassification probabilities as $\mathrm{pr}(W=1|X=0)=\xi_1$ and $\mathrm{pr}(W=0|X=1)=\xi_0$, hence the distribution of measurement error process $f_{\mathrm{mem}}(w|x,\xi_0,\xi_1)=\{w\xi_1+(1-w)(1-\xi_1)\}(1-x)+\{w(1-\xi_0)+(1-w)\xi_0\}x$. In this situation W is, e.g. a smoking status reported by the study participant and X is the true long-term smoking exposure of interest. Note that, e.g., lung cancer patients who have the suspected risk factor (e.g., smoking) can blame this risk factor for causing the disease and therefore they are likely to over-report smoking, hence the misclassification probabilities can be differential in the disease status. In this case the measurement error process depends on disease status and misclassification probabilities need to be specified for cases and controls separately.

On the risk parameters we impose a Normal prior with mean $\mu_{\mathcal{B}}$ and covariance matrix $\Sigma_{\mathcal{B}}$. In the case when a massive amount of measurement error is present, the sampling distribution of risk parameter estimates is likely to be skewed (Shafer and Purdy (1996), Lobach, et al. (2008)).

But 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 non-informative. 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

$$\prod_{i=1}^{n}\frac{\sum_{x=0}^{1}S(d_i,g_i,x,\mathcal{B},\theta)f_{\mathrm{mem}}(w_i|x,\xi_0,\xi_1)\eta^x(1-\eta)^{1-x}}{\sum_{x=0}^{1}\sum_{d=0}^{1}\sum_{g=0}^{1}S(d_i,g,x,\mathcal{B},\theta)\eta^x(1-\eta)^{1-x}}\times|\Sigma_{\mathcal{B}}|^{-1/2}\exp\Big\{-\frac{1}{2}(\mathcal{B}-\mu_{\mathcal{B}})^{\mathrm{T}}\Sigma_{\mathcal{B}}^{-1}(\mathcal{B}-\mu_{\mathcal{B}})\Big\}\times I_{(0,1)}(\eta)I_{(0,1)}(\theta).$$

4.2 Haplotype-based case-control studies

Within this setting we consider continuous environmental variables and assumed that the genetic risk depends on the number of copies of a putative haplotype. This setting is particularly useful in the situations when the available genetic information consists of a set of markers that are located closely to each other. The linkage disequilibrium (LD) is generally used to measure the degree of dependence between the genetic markers. When LD is high, the generic markers can be organized in the haplotype blocks according to the LD pattern. The continuous environmental variable can model dietary exposure, such as calcium intake, and X defines the true unobservable calcium intake, W - calcium intake measured using FFQ.

Suppose the true environmental exposure is distributed as Normal with mean μ_x and variance σ_x^2. On mean and variance of the environmental covariate we impose Normal(η_1,η_2) and IG(A,B) prior, respectively. Let θ_j be the frequency of haplotype $j=1,\ldots,T$, then the distribution of diplotypes in the population under consideration is specified using HWE. On all haplotype frequencies we impose a Uniform distribution. The true environmental covariate is not observable, instead W is subject to classic additive measurement error. The distribution of observed environmental covariate $f_{\mathrm{mem}}(w|x,\xi)$ is Normal with mean x and variance ξ. Note, however, that the methodology is general enough to model various types of measurement error including differential errors. Suppose h_1 is a reference haplotype, define $\mathcal{B}=(\beta_x,\beta_{h2},\ldots,\beta_{hk},\beta_{xh2},\ldots,\beta_{xhk})$ to be vector of risk parameters. We use a Normal distribution with mean $\mu_{\mathcal{B}}$ and covariance matrix $\Sigma_{\mathcal{B}}$ as a prior distribution for $\mathcal{B}$. Denote $N_j(H), j=1,\ldots,T$ to be the number of haplotypes h_j observed in a diplotype H. The function $m(x,h,\mathcal{B})$ allows modeling various types of disease, such as additive, multiplicative, recessive, dominant, etc. Additionally, the risk of genotype, environment as well as their interaction are parameterized within this function. Consider a model of an additive disease status and multiplicative interaction defined as $m(x,h,\mathcal{B})=\beta_x x+\beta_{h2}N_2(H)+\cdots+\beta_{hT}N_T(H)+\beta_{xh2}xN_2(H)+\cdots+\beta_{xhT}xN_T(H)$. Finally, define

$$S(d,h,x,\mathcal{B},\theta)=\frac{\exp[I_{(d\geq 1)}(d)\{\kappa_d+m(x,h,\mathcal{B})\}]}{1+\exp\{\beta_0+m(x,h,\mathcal{B})\}}Q(h,\theta).$$

The joint posterior distribution becomes

$$\propto\prod_{i=1}^{n}\frac{\int\sum_{h^*\in\mathcal{H}_G}S(d_i,h^*,x,\mathcal{B},\theta)\exp\Big\{-\frac{(w_i-x)^2}{2\xi}-\frac{(x-\mu_x)^2}{2\sigma_x^2}\Big\}\mathrm{d}x}{\int\sum_{d^*=0}^{1}\sum_{h^*\in\mathcal{H}}S(d,h^*,x,\mathcal{B},\theta)\exp\Big\{\frac{-(x-\mu_x)^2}{2\sigma_x^2}\Big\}\mathrm{d}x}\times|\Sigma_{\mathcal{B}}|^{-1/2}\exp\Big\{-\frac{1}{2}(\mathcal{B}-\mu_{\mathcal{B}})^{\mathrm{T}}\Sigma_{\mathcal{B}}^{-1}(\mathcal{B}-\mu_{\mathcal{B}})\Big\}\times\eta_2^{-1/2}\exp\big\{-(\mu_x-\eta_1)^2/(2\eta_2)\big\}(\sigma_x^2)^{-A-1}\exp(-B/\sigma_x^2)\times\prod_{t=1}^{T}I_{(0,1)}(\theta_t).$$

We propose to validate the likelihood using ideas of Monahan and Boos (1992) and then apply conventional MCMC sampling techniques, such as Metropolis-Hastings algorithm to obtain the samples from the posterior for Bayesian inference.

5. ASYMPTOTIC POSTERIOR DISTRIBUTION

We now consider properties of an asymptotic posterior distribution based on the pseudo likelihood (1). MCMC

techniques can be computationally challenging and knowing the form of an asymptotic posterior distribution would ease the computational burden.

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

$$\Lambda = \sum_d \frac{n_d}{n} \mathrm{E}\{\Psi(D, G, W, Z, \Omega, \eta, \xi) | D = d\} \times \mathrm{E}\{\Psi(D, G, W, Z, \Omega, \eta, \xi) | D = d\}^{\mathrm{T}}.$$

Further, 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 $\mathcal{L}_n(\Theta, \xi) = \sum_{i=1}^n \Psi(D_i, G_i, W_i, Z_i, \Theta, \xi)$ and matrices $\mathcal{I}(\Theta) = -\mathrm{E}[\frac{\partial\{\mathcal{L}_n(\Theta,\xi)\}}{\partial(\Theta)}]$ and $\mathcal{J}(\Theta) = -\mathrm{E}[\frac{\partial\{l(\Theta)\}}{\partial(\Theta)}]$. The following theorem and its heuristic proof motivated by Bernardo and Smith (1994) concerns limiting properties of the posterior distribution.

Theorem 1. *Under suitable regularity conditions the posterior distribution of vector of parameters $\widehat{\Theta}_n$ converges to a Normal distribution with covariance matrix consistently estimated by $\widehat{\Sigma}_n = \{\mathcal{I}(\widehat{\Theta}_n) + J(\Theta_0)\}^{-1}$ and mean vector $\widehat{\mathcal{M}}_n = \widehat{\Sigma}_n^{-1}\{\mathcal{I}(\widehat{\Theta}_n)\widehat{\Theta}_n + \mathcal{J}(\Theta_0)\Theta_0\}$.*

Proof. Note that the posterior distribution of the vector of parameters Θ given data X can be written as

$$f(\Theta|X) \propto p(\Theta)L_n(\Theta) = \exp[\log\{p(\Theta)\} + \log\{L_n(\Theta)\}].$$

Let Θ_0 and $\widehat{\Theta}_n$ be maxima of the prior $p(\Theta)$ and pseudo-likelihood function $L_n(\Theta)$, respectively. They can be obtained by solving $l(\Theta) = 0$ and $\mathcal{L}_n(\Theta) = 0$. Under suitable regularity conditions which ensure that the remainder terms of the following expansion are small for large n, the logarithm of the prior and pseudo-likelihood function can be expanded around their maxima in the following manner.

$$\log\{p(\Theta)\} = \log\{p(\Theta_0)\} - 1/2(\Theta - \Theta_0)^{\mathrm{T}}\mathcal{J}(\Theta_0)(\Theta - \Theta_0);$$
$$\log\{L_n(\Theta)\} = \log\{f(X|\widehat{\Theta}_n)\} - 1/2(\Theta - \widehat{\Theta}_n)^{\mathrm{T}}\mathcal{I}(\widehat{\Theta}_n)(\Theta - \widehat{\Theta}_n).$$

Hence

$$f(\Theta|X) \propto \exp\{-1/2(\Theta - \Theta_0)^{\mathrm{T}}\mathcal{J}(\Theta_0)(\Theta - \Theta_0)\} \times \exp\{-1/2(\Theta - \widehat{\Theta}_n)^{\mathrm{T}}\mathcal{I}(\widehat{\Theta}_n)(\Theta - \widehat{\Theta}_n)\}.$$

Further, it can be easily seen that for large sample sizes

$$f(\Theta|X) \propto \exp\{-1/2(\Theta - \mathcal{M})^{\mathrm{T}}\Sigma^{-1}(\Theta - \mathcal{M})\}. \quad \square$$

Remark 1. The development of Theorem 1 suggests that the posterior based on a pseudo-likelihood function has asymptotic distribution that is the same as Normal with mean that is equal to the weighted average of a maximum pseudo-likelihood estimate and a value that maximizes prior. The precision of this distribution is the sum of the observed information matrix and the prior precision matrix. These considerations suggest one approximation, namely if for large n the prior precision tends to be small compared to the precision provided by the data, it can be ignored.

Remark 2. It can be easily seen that $n^{-1}\partial\{\mathcal{L}_n(\widehat{\mathcal{B}}, \xi)\}/\partial\mathcal{B}^{\mathrm{T}}$ is a consistent estimate of $\mathcal{I}(\Theta)$. Alternatively, if $\widehat{\Sigma}$ is the sample covariance matrix of the terms $\Psi(D_i, G_i, W_i, Z_i, \widehat{\mathcal{B}}, \xi)$, then $\widehat{\Sigma} + \widehat{\Lambda}$ consistently estimates $\mathcal{I}(\Theta)$.

Remark 3. When the sample size is large, we can use this asymptotic posterior distribution for validation purpose rather than using the MCMC based approach. That way, we can reduce the computation burden significantly.

Remark 4. If the parameter ξ controlling the measurement error distribution is unknown, additional data are necessary to estimate it. Consider the case of additive mean-zero measurement error with replications of W. Suppose that there are at most M replications of the W for any individual. Let W_i denote this ensemble of the M replicates, and let m_i be the number of replicates we actually observe. Let $f_{\mathrm{mem}}(w|d, h^{\mathrm{dip}}, x, z, m, \xi)$ be the joint density of the first m replicates for $m = 1, \ldots, M$; $\Psi(D, G, W, Z, \Omega, \eta, \xi, j)$, $\mathcal{I}_j$, and Λ_j be matrices defined earlier for the case with exactly $m = j$ replicates for each individual. Assume that m_i is independent of $(D_i, W_i, Z_i, G_i, X_i, H_i^{\mathrm{dip}})$ and that $\mathrm{pr}(m_i = j) = p(j)$. Further, define $\mathcal{I} = \sum_{j=1}^M p(j)\mathcal{I}_j$. Then Lobach, et al. (2008) showed that the estimating function for $\mathcal{B} = (\Omega^{\mathrm{T}}, \eta^{\mathrm{T}}, \xi^{\mathrm{T}})^{\mathrm{T}}$ can be written in the form

$$0 = \sum_{i=1}^n \sum_{j=1}^M I_{(m_i=j)}(m_i)\Psi(D_i, G_i, W_i, Z_i, \Omega, \eta, \xi, j).$$

and the corresponding consistent sequence of solutions is

$$n^{1/2}(\widehat{\mathcal{B}} - \mathcal{B}_0) \Rightarrow \mathrm{Normal}\left[0, \mathcal{I}^{-1}\left\{\mathcal{I} - \sum_{j=1}^M p(j)\Lambda_j\right\}\mathcal{I}^{-1}\right].$$

The result of Theorem 1 can be readily applied to this situation when measurement error distribution is estimated using replications. Consistent estimates of $\mathcal{I}$ and Λ_j can be obtained by applying formulas that are analogous to those outlined in the Remark 2.

6. SIMULATION EXPERIMENTS

To illustrate performance of the proposed methodology, we performed two simulation studies. First, we compared

Table 1. Biases and Root Mean Squared Errors (RMSEs) for the proposed Bayesian and Pseudo-MLE approaches in a genotype-based case-control study. Genetic (G) and environmental (X) factors are binary with pr$(G=1)=0.5$ and pr$(X=1)=0.5$. pr$(D=1)=0.016$ is assumed to be known in the underlying population. Misclassification probabilities are pr$(W=1|X=0)=0.15$ and pr$(W=0|X=1)=0.10$. The results is based on 500 replications of n_1 cases and n_0 controls

			Proposed Bayesian Analysis		Pseudo-MLE	
	Parameter	True Value	Bias	RMSE	Bias	RMSE
$n_0=200$	β_x	1.099	−0.007	0.229	0.023	0.339
$n_1=200$	β_g	0.693	−0.075	0.336	−0.195	1.023
	β_{xg}	0.693	0.103	0.461	0.217	1.064
	θ	0.500	−0.002	0.002	−0.001	0.021
	η	0.500	−0.005	0.003	0.001	0.048
$n_0=1,000$	β_x	1.099	−0.003	0.095	0.005	0.155
$n_1=1,000$	β_g	0.693	−0.021	0.162	−0.004	0.305
	β_{xg}	0.693	0.029	0.206	0.001	0.327
	θ	0.500	−0.001	0.001	0.000	0.008
	η	0.500	−0.001	0.001	0.002	0.022

performance of the proposed Bayesian approach to the pseudo-MLE using genotype-based setting. In this setting all variables are categorical. Further, we investigated properties of estimation and inference based on the derived form of the Asymptotic Posterior distribution (Theorem 1) and its approximation discussed in Remark 1. This analysis does not require MCMC computations, instead it is uses a derived form of the Asymptotic Posterior Distribution.

6.1 Genotype-based case-control studies: Proposed Bayesian analysis vs. pseudo-MLE approach

We performed a series of simulation experiments to illustrate our approach in the setting of genotype-based case-control study.

We assumed that environmental variables (X, W), genetic variant (G), and disease status (D) are binary. Given the values of (G, X) we generated a binary disease outcome D from the logistic model $\text{logit}\{\text{pr}(D|G,X)\} = \beta_0 + \beta_x X + \beta_g G + \beta_{xg} X * G$, with parameters $(\beta_x, \beta_g, \beta_{xg}) = (1.099, 0.693, 0.693)$. This setting arises in the presence of recessive or dominant effect. The misclassification probabilities were $\text{pr}(W=0|X=1)=0.10$ and $\text{pr}(W=1|X=0)=0.15$. The probability of disease in this setting is 0.016 and we assumed it is known in the population. We investigated the case of small $(n_0 = n_1 = 200)$ and large $(n_0 = n_1 = 1,000)$ sample sizes.

First, it is necessary to validate the likelihood function. We validated coverage of the pseudo-likelihood function using ideas described in Monahan and Boos (1992) for numerous scenarios by setting different values of prior parameters, as well as varying sample size and misclassification probabilities. It was challenging to compute H_k using (2) since it requires calculations with multiple precision. We addressed this challenge by using the General Multiple Precision library in C. Further, the vector of parameters is 5-dimensional and since integration in (2) requires computations with high precision and high-dimensional integration is not feasible, we verified coverage probabilities of each parameter when all others are fixed at their posterior mean. For all cases we considered the Kolmogorov-Smirnov test failed to reject the null hypothesis that the sample H_k comes from the Uniform$(0,1)$ distribution at 0.05 significance level.

Since the likelihood function was validated, we proceeded to parameter estimation using the Metropolis-Hastings algorithm with the following settings. On the risk parameters $\mathcal{B}$ we imposed a Normal$(\mathcal{B}^{\text{mean}}, \Sigma_{\mathcal{B}})$ prior, where $\mathcal{B}^{\text{mean}} = (0,0,0)$ and covariance matrix $\Sigma_{\mathcal{B}} = \text{diag}(3^2, 3^2, 3^2)$. Note that in this setting reflects no a priori knowledge about the risk and mean of the prior distribution is conservatively set to be zero. The only prior information that we are imposing is that the shape of the distribution is symmetric to bring the sampling distribution of the parameter estimates closer to Normal. On both η and θ we imposed a Uniform$(0,1)$ prior. The a priori information specified on the frequency parameter θ is non-informative in this setting. If an estimate about genotype frequencies is available, it can be used while specifying the distribution. The proposal distribution of the new value $\mathcal{B}^t$ given the current $\mathcal{B}^{t-1}$ was set to be Normal$(\mathcal{B}^{t-1}, \Sigma^{\text{prop}})$, where $\Sigma^{\text{prop}} = \text{diag}(0.05^2, 0.05^2, 0.05^2)$. Proposal distribution of a new value of θ^t given θ^{t-1} was chosen to be Uniform$(\theta^{t-1} - 0.05, \theta^{t-1} + 0.05)$. The proposal distribution for η has the same form as that for θ.

Proposed Bayesian approach and Pseudo-MLE The simulation results presented in Table 1 illustrate that for a small sample size the proposed Bayesian approach produced parameter estimates that are less biased and less variable than the estimates obtained using pseudo-MLE approach. Moreover, distribution of the parameter estimates obtained using pseudo-MLE is skewed, while our simulations illustrated that the distribution of parameter estimates produced using our methodology is close to symmetric, when illustrating the ability of Bayesian methodology to shrink toward prior. In

Table 2. Biases and Root Mean Squared Errors (RMSEs) of the proposed Bayesian and Pseudo-MLE approaches in a genotype-based case-control study when genotype is rare. Genetic (G) and environmental (X) factors are binary with pr$(G = 1) = 0.05, 0.025$ and pr$(X = 1) = 0.5$. pr$(D = 1) = 0.0148, 0.0140$ are assumed to be known. Misclassification probabilities are pr$(W = 1|X = 0) = 0.15$ and pr$(W = 0|X = 1) = 0.10$. The results is based on 500 replications of $1,000$ cases and $1,000$ controls

		Proposed Bayesian Analysis		Pseudo-MLE	
Parameter	True Value	Bias	RMSE	Bias	RMSE
β_x	1.099	0.011	0.257	0.003	0.311
β_g	0.693	−0.010	0.318	−0.269	1.379
β_{xg}	0.693	0.004	0.336	0.434	1.473
θ	0.050	-1.6×10^{-4}	0.039	0.001	0.044
η	0.500	0.001	0.005	−0.007	0.011
β_x	1.099	0.010	0.343	0.006	0.303
β_g	0.693	−0.003	0.353	−0.470	1.853
β_{xg}	0.693	0.010	0.039	0.607	1.907
θ	0.025	-1.9×10^{-4}	0.026	0.001	0.044
η	0.500	−0.030	0.002	−0.025	0.028

case of a large sample size, the proposed methodology resulted in parameter estimates that are nearly unbiased with RMSEs that are considerably smaller than the RMSEs of the pseudo-MLE approach.

In the case of massive measurement error, which is the case in our motivating example and simulation experiments, the finite sample distribution of parameter estimates can be skewed (Schafer and Purdy, 1996). We observed this phenomena in Lobach, et al. (2008) and our simulation studies. Hence one of the major advantages of the proposed Bayesian solution is that a symmetric prior can help to bring the finite sample distribution of the parameter estimates closer to Normal.

Rare genotype To investigate performance of the proposed method in the rare genotype case, we performed the following simulation experiment. Genetic and environmental variables, disease status and measurement error were simulated using setup described above. However, the genotype frequency was set up to be $\theta = 5\%, 2.5\%$. On the genotype frequencies we imposed $Beta(A, B)$ distribution with parameters $A = 5$, $B = 95$ and $A = 3$, $B = 97$ for the case of genotype frequency 5% and 2.5%, respectively. These distributions have means that are equal to the true values and support indicating that the genotypes are rare. Table 2 presents simulation results. Pseudo-MLE estimation resulted in genotype frequency estimates that have elevated bias and larger variability. As a result, interaction parameter estimates and main effects of genotype are largely biased. The proposed Bayesian approach produced nearly unbiased estimates and have smaller variability. The sampling distribution of risk parameter estimates obtained using the pseudo-MLE method was heavily skewed, however that of our Bayesian estimates was closer to Normal. This demonstrates the ability of Bayesian approach with symmetric prior to bring posterior estimates closer to Normal.

6.2 Haplotype-based case-control studies: Analysis based on asymptotic posterior distribution

Following the simulation setup of Lobach, et al. (2008), we considered a continuous environmental variables and assumed that the genetic risk depends on the number of copies of a putative haplotype. We simulated the true environmental covariate (X) from a Normal distribution with zero mean and variance 0.1. To simulate observed environmental variables, we used an additive model of the form $W = X + U$, where U is generated from the Normal distribution with zero mean and variance $\xi = 0.25$. Note that we are simulating a case of large measurement error, such as would occur for dietary measurements. This gives a stern test for our methodology.

Given the haplotype frequencies $(h_1, h_2, h_3, h_4, h_5, h_6) = (0.25, 0.15, 0.25, 0.1, 0.1, 0.15)$ we generated diplotypes for each subject under the assumption of Hardy-Weinberg Equilibrium. Then we coded haplotype h_3 as 1 and all the rest as 0. Given the diplotype information H^{dip} and environmental covariate X we generated binary disease status according to the following model

$$\begin{aligned}&\text{pr}(D = d|H^{\text{dip}}, X)\\ &\quad = \frac{\exp\left[d\left\{\beta_0 + \beta_x X + \beta_g N_3(H^{\text{dip}}) + \beta_{xg} X N_3(H^{\text{dip}})\right\}\right]}{1 + \exp\left\{\beta_0 + \beta_x X + \beta_g N_3(H^{\text{dip}}) + \beta_{xg} X N_3(H^{\text{dip}})\right\}},\end{aligned}$$

where $N_3(H^{\text{dip}})$ is the number of copies of h_3 in H^{dip}. In this setting we are interested in estimating the relative risk parameters and the frequency of haplotype h_3. For the sake of computational time we assumed that the probability of disease is known. Moreover, we assessed the effect of missing data by assuming that 50% of subjects were not genotyped and for those who were genotyped, the phase is unknown.

Table 3. Proposed Bayesian Analysis of a haplotype-based case-control study. Biases, Standard Errors (SE) of the estimates, and estimated SEs based on derived asymptotic posterior distribution (Theorem 1). The analysis is performed on the observed data combined with the prior information and the observed data only (see Remark 1). The results are based on a simulation study with 300 replications for 1,000 cases and 1,000 controls, where disease status (D) is binary, environmental variable (X) is Normal with variance 0.1 and the genetic variant h_3 is in the form of diplotype with a multiplicative interaction. The environmental variable is measured with error and the error variance is 0.25. Genotype is missing for 50% of the subjects and when it is observed, haplotype-phase ambiguity is present. The *ed value indicates 10%-trimmed estimate, †ed $-$ 2%-trimmed estimates

		Observed Data and Prior Information			Observed Data Only		
Parameter	True Value	Bias	SE	Estimated SE	Bias	SE	Estimated SE
β_x	1.099	0.0215	$0.0942^\star$	0.0726	0.0201	0.0856	$0.1108^\dagger$
β_g	0.693	-0.0058	0.0023	0.0025	-0.0003	0.0064	$0.0026^\dagger$
β_{xg}	0.693	-0.0201	0.0528	0.0544	-0.0186	0.0972	$0.0742^\dagger$
θ	0.500	0.0006	0.0001	0.0001	0.0005	0.0004	$0.0000^\dagger$
η_1	0.000	-0.0027	0.0010	0.0009	0.0016	0.0005	$0.0003^\dagger$
η_2	0.100	0.0000	0.0001	0.0002	-0.0009	0.0002	$0.0003^\dagger$

The pseudo-likelihood function validated in a similar way as has been described in the discrete situation. Results presented in the Table 3 are based on the estimates obtained using an approximation derived in the Theorem 1. Analysis of the simulation results presented in the Table 3 suggests that the proposed methodology resulted in parameter estimates that are nearly unbiased. Moreover, estimated variances of parameter estimates are very close to observed values, with one exception, namely β_x. This is due to the fact that when a large amount of measurement error is present in the data, parameter estimates can have skewed distributions even for large sample sizes.

Additionally we investigated an approximation discussed in the Remark 1. To recap, the Theorem 1 illustrates that the asymptotic precision is the sum of a precision provided by the observed data and prior precision matrix. Similarly, asymptotic mean is the weighted average of a maximum pseudo-likelihood estimate and a value that maximizes prior. The results presented in Table 3 are based on an approximation that ignores the prior precision and the covariance matrix is constructed using precision provided by the observed data only. Inspection of the results suggests that parameter estimates are unbiased and estimated standard errors are close to the observed standard errors. However, the SE of estimates are generally larger than the SE of estimates obtained with the use of prior information. Recall that in this case the only prior information induced in the model is on the shape of the parameter estimates distribution. And this information helped to bring sampling distribution of the parameter estimates closer to Normal thus reducing the variability and making the inferences more precise. Note, however, that absolute values of biases of parameter estimates in the case when prior information is used are generally slightly larger. The reason is that the the prior mean of the risk parameter estimates is zero, and hence it forces underestimation of risk parameters. In summary, we demonstrated that approximation derived in Theorem 1 can work well in practice and that a symmetric prior can improve inferences.

7. COLORECTAL ADENOMA STUDY

7.1 Modeling

Here we analyze the colorectal adenoma study data described above. To recap, there were 772 cases and 778 controls, the response D was colorectal adenoma status, the genetic data observed were three SNPs in the calcium receptor gene CaSR, the environmental variable X measured with error was log(1+calcium intake), which was measured by W, the result of a food frequency questionnaire. The variables Z measured without error were age, sex and race. The possible haplotypes in the data were ACG, ACT, AGG, GCG, AGT, GGG, and GCT. Since haplotypes AGT, GGG, GCT are rare, we pooled them with the next most common haplotype AGG. The distribution of haplotype frequencies is not significantly deviating from the HWE. A few subjects do not have measurements of calcium intake and we eliminated them from the analysis.

Given calcium intake (X) and diplotype information (H^{dip}) we considered the following risk model

$$\begin{aligned}&\text{logit}\{\text{pr}(D=1|H^{\text{dip}},X)\}\\&\quad=\beta_0+\beta_x*X+\beta_{h2}*N_2(H^{\text{dip}})+\beta_{h4}*N_4(H^{\text{dip}})\\&\qquad+\beta_{h5}*N_5(H^{\text{dip}})+\beta_{xh2}*X*N_2(H^{\text{dip}})\\&\qquad+\beta_{xh4}*X*N_4(H^{\text{dip}})+\beta_{xh5}*X*N_5(H^{\text{dip}}),\end{aligned}$$

where $N_2(H^{\text{dip}})$ is the number of haplotypes ACT observed in a diplotype, $N_4(H^{\text{dip}})$ is the number of haplotypes GCG observed in a diplotype and $N_5(H^{\text{dip}})$ is number of haplotypes AGG, AGT, GGG, or GCT observed in a diplotype.

Using an external data set, Lobach et al. (2008) estimated the measurement error distribution and found that $W = 0.22+0.75X+u$, where $\widehat{\sigma}_u^2=\widehat{\xi}=0.65$. To assess sensitivity

Table 4. Proposed 95% Credible Intervals of the risk estimates in the Colorectal Adenoma Data. Results are based on the last 5,000 of 100,000 iterations of the Metropolis-Hastings algorithm. The estimated error variance is $\widehat{\xi} = 0.65$

	$\xi = 0.10$	$\xi = 0.60$	$\xi = 0.65$	$\xi = 0.70$
κ	$(-0.173, 0.280)$	$(-0.217, 0.215)$	$(-0.172, 0.244)$	$(-0.144, 0.233)$
β_x	$(-0.269, 0.125)$	$(-0.324, 0.017)$	$(-0.367, 0.080)$	$(-0.360, 0.058)$
β_{h2}	$(-0.410, 0.015)$	$(-0.365, -0.010)$	$(-0.365, 0.032)$	$(-0.337, 0.040)$
β_{h4}	$(-0.451, -0.038)$	$(-0.612, -0.252)$	$(-0.622, -0.244)$	$(-0.642, -0.263)$
β_{h5}	$(-0.544, -0.157)$	$(-0.938, -0.528)$	$(-0.946, -0.583)$	$(-0.933, -0.592)$
β_{xh2}	$(-0.163, 0.211)$	$(-0.079, 0.294)$	$(-0.094, 0.290)$	$(-0.101, 0.274)$
β_{xh4}	$(-0.334, 0.029)$	$(-0.731, -0.357)$	$(-0.806, -0.380)$	$(-0.789, -0.411)$
β_{xh5}	$(-0.431, -0.019)$	$(-1.135, -0.692)$	$(-1.088, -0.662)$	$(-1.051, -0.711)$

to the measurement error model specification, we considered several scenarios by imposing measurement error structure estimated using an external data and varying it through σ_u^2.

7.2 Estimation

To estimate parameters we employed Metropolis-Hastings algorithm with the following setting. Denote $\mathcal{B}$ to be the set of risk parameters, Θ to be the vector of haplotype frequencies and η to be parameters of the environmental covariate. Define $\widehat{\mathcal{B}}_{\text{MLE}}$, $\widehat{\Theta}_{\text{MLE}}$, and $\widehat{\eta}_{\text{MLE}}$ to be the set of estimates obtained using pseudo-MLE. We performed the analysis based on zero-mean priors for the risk parameters and obtained almost identical results. On the risk parameters we imposed Normal$(0, \Sigma_{\mathcal{B}})$ prior where $\Sigma_{\mathcal{B}}$ is 8×8 diagonal matrix with elements 3^2. For the haplotype frequencies we used Uniform$(\widehat{\Theta} - 0.5, \widehat{\Theta} + 0.5)$. Mean of the environmental covariate was chosen to follow Uniform$(\widehat{\eta}_{1\text{MLE}}, \sigma^2_{\eta_1})$ distribution, where $\sigma^2_{\eta_1} = 1$. On the variance of the environmental covariate η_2 we imposed Inverse Gamma (IG) prior. Since we considered several scenarios by assuming various measurement error variances, we set the values of the IG distribution such that the mean of the IG distribution is equal to the pseudo-MLE estimate of the variance η_2. The proposal density of the new values $\mathcal{B}^t$ given the current value $\mathcal{B}^{t-1}$ is Normal$(\mathcal{B}^{t-1}, \Sigma^p_{\mathcal{B}})$, where $\Sigma^p_{\mathcal{B}}$ is a 8×8 diagonal matrix with elements 0.5^2. The proposal distribution of a new value η_1^t given the current η_2^{t-1} is Normal$(\eta_1^{t-1}, 1)$. The proposal value of the haplotype frequencies is simulated from the Uniform$(\Theta^{t-1} - D_\Theta, \Theta^{t-1} + D_\Theta)$ distribution, where D_Θ is 0.01 for common and 0.001 for rare haplotypes. The proposal density for a new value η_2^t given the current η_2^{t-1} is IG distribution with parameters $5/\eta_2^{t-1}$ and 5 chosen so that the mean of the IG distribution is equal to the current value η_2^{t-1}.

7.3 Results

The four sets of parameter estimates presented in the Table 5 correspond to different values of measurement error variance. These results illustrate the importance of assessing the measurement error, since its incorrect specification results in substantial bias. Table 4 resents 95% posterior credible intervals obtained based on MCMC sampling. We also performed the analysis based on an asymptotic posterior distribution (not shown here). Both parameter estimates and credible intervals based on the asymptotic posterior are very close to those obtained using MCMC sampling.

Table 5. Bayesian estimates of the Colorectal Adenoma Data risk parameters for various values of the measurement error variance (ξ). Results are based on the last 5,000 of 100,000 iterations of the Metropolis-Hastings algorithm. The estimated error variance is $\widehat{\xi} = 0.65$

Parameter	$\xi = 0.10$	$\xi = 0.60$	$\xi = 0.65$	$\xi = 0.70$
κ	0.054	0.024	0.025	0.018
β_x	−0.067	−0.141	−0.140	−0.179
β_{h2}	−0.198	−0.182	−0.175	−0.144
β_{h4}	−0.229	−0.361	−0.416	−0.522
β_{h5}	−0.366	−0.652	−0.752	−0.932
β_{xh2}	0.028	0.085	0.102	0.123
β_{xh4}	−0.157	−0.467	−0.590	−0.756
β_{xh5}	−0.239	−0.724	−0.887	−0.818

We examined the posterior distribution of risk parameter estimates, including the gene-environment interaction parameters. The distribution of the estimates was roughly Normal (data not shown), which illustrated the ability of prior information to bringing the sampling distribution of parameter estimates to symmetric.

Inspection of the credible intervals reveals that for all measurement error specifications presented in the Table 4 parameters β_{xh4} and β_{xh5} are significantly different from 0 at the 0.05 significance level. This indicates that there is sufficient evidence to conclude that among carriers of h_4 and h_5 increased calcium intake is associated with decreased risk of colorectal tumor development.

Comparison of results for small ($\xi = 0.10$) and large ($\xi = 0.60, 0.65, 0.70$) amounts of measurement error illustrates that ignoring measurement error leads to biased estimates and possibly incorrect inferences. For example, the interaction parameter β_{xh4} is announced to be not significantly different from zero when error variance is set to be small. However when the measurement error is properly accounted for and the error variance is set to be the value

that was estimated from an external dataset, the interaction parameter β_{xh4} is announced to be significant. Further, sensitivity analysis illustrated that when the measurement error variance is close to what was estimated, the conclusion about the risk defined by β_{xh4} did not change.

8. DISCUSSION

We proposed a Bayesian methodology for analysis of gene-environment interactions using interaction and using population based case-control data. A key aspect of our method is that retrospectively collected data is analyzed as a random sample allowing gains of efficiency in parameter estimates (Lobach, et al., 2008). Because the analysis is based on a pseudo-likelihood function, the conventional Bayesian machinery may not be applied directly.

The Bayesian approach allows prior information about risk parameter estimates to enter the estimation and inference procedures, which is particularly useful in the case of massive measurement error. In this case even for large samples the sampling distribution of risk parameter estimates can be skewed and hence inferences that use Normality assumption are not precise. A symmetric distribution helps shrink towards the prior and hence make the sampling distribution of the estimates be closer to Normal, thus improving inferences.

ACKNOWLEDGMENTS

Our research was supported by grants from the National Cancer Institute (CA10462 and CA57030) and the National Science Foundation (DMS 0914951).

Received 19 April 2010

REFERENCES

[1] Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society, Series B*, 32: 283-301. MR0273723

[2] Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: John Wiley & Sons. MR1274699

[3] Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models, Second Edition*. Chapman & Hall CRC Press. MR2243417

[4] Chatterjee, N. and Carroll, R. J. (2005). Semiparametric maximum likelihood estimation in case-control studies of gene-environmental interactions. *Biometrika*, 92: 399–418. MR2201367

[5] Chatterjee, N., Chen, J., Spinka, C. and Carroll, R. J. (2006.) Comment on the paper "Likelihood based inference on haplotype effects in genetic association studies" by D. J. Lin and D. Zhang. *Journal of the American Statistical Association*, 102: 108–110.

[6] Cornfield, J. (1956). A statistical problem arising from retrospective studies. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*. MR0084935

[7] Hunter, D. J. (2005). Gene-environment interactions in human diseases. *Nature Review Genetics*, 6: 287–298.

[8] Lazar, N. A. (2003). Bayesian empirical likelihood. *Biometrika*, 90(2): 319–326. MR1986649

[9] Lobach, I., Carroll, R. J., Spinka, C., Gail, M. and Chatterjee, N. (2008) Haplotype-based regression analysis and inference of case-control studies with unphased genotypes and measurement errors in environmental exposures. *Biometrics*, 60(3): 673–684 MR2526616

[10] Lobach, I., Fan, R. and Carroll, R. J. (2010) Genotype-Based Association Mapping of Complex Diseases: Gene-Environment Interactions with Multiple Genetic Markers and Measurement Errors in Environmental Exposures. *Genetic Epidemiology*, 32: 1–11

[11] Monahan, J. F. and Boos, D. D. (1992). Proper likelihood for Bayesian analysis. *Biometrika*, 79(2): 271–278. MR1185129

[12] Müller, P. and Roeder, K. (1997). A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika*, 84: 523–537. MR1603977

[13] Peters, U., Chatterjee, N., Yeager, M., Chanock, S. J., Schoen, R. E., McGlynn, K. A., Church, T. R., Weissfeld, J. L., Schatzkin, A. and Hayes, R. B. (2004). Association of genetic variants in the calcium-sensing receptor with risk of colorectal adenoma. *Cancer Epidemiology Biomarkers Prevention*, 13(12): 2181–2186.

[14] Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66: 403–412. MR0556730

[15] Sinha, S., Mukherjee, B., Ghosh, M., Mallick, B. K. and Carroll, R. J. (2005). Semiparameteric Bayesian analysis of case-control studies with missing exposure. *Journal of the American Statistical Association*, 100: 591–601. MR2160562

[16] Schafer, D. W. and Purdy, K. G. (1996). Likelihood analysis for errors-in-variables regression with replicate measurements. *Biometrika*, 83: 813–824. MR1440046

[17] Schatzkin, A., Subar, A. F., More, S., Park, Y., Potischman, N., Thompson, F. E., Leitzmann, M., Hollenbeck, A., Morrissey, K. G. and Kipnis, V. (2009) Observational Epidemiologic Studies of Nutrition and Cancer: The Next Generation (with Better Observation). *Cancer Epidemiology, Biomarkers & Prevention*, 18: 1026

[18] Spinka, C., Carroll, R. J. and Chatterjee, N. (2005). Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. *Genetic Epidemiology*, 29: 108–127.

[19] Subar, A.F., Kipnis, V., Troiano, R.P., et al. (2003). Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: The Observing Protein and Energy Nutrition (OPEN) study. *American Journal of Epidemiology*, 158: 1–13.

Iryna Lobach
Division of Biostatistics
New York University School of Medicine
NY 10016, USA
E-mail address: iryna.lobach@nyumc.org

Bani Mallick
Department of Statistics
Texas A&M University, College Station
TX 77840, USA
E-mail address: bmallick@stat.tamu.edu

Raymond J. Carroll
Department of Statistics
Texas A&M University, College Station
TX 77840, USA
E-mail address: carroll@stat.tamu.edu