Biometrics 63, 834–844
September 2007
DOI: 10.1111/j.1541-0420.2007.00750.x

Semiparametric Bayesian Analysis of Case–Control Data under Conditional Gene-Environment Independence

Bhramar Mukherjee,[1,*] Li Zhang,[2] Malay Ghosh,[3] and Samiran Sinha[4]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
[2]Department of Quantitative Health Sciences, The Cleveland Clinic Foundation, Cleveland, Ohio 44195, U.S.A.
[3]Department of Statistics, University of Florida, P.O. Box 118545, Gainesville, Florida 32611, U.S.A.
[4]Department of Statistics, Texas A&M University, TAMU 3143, College Station, Texas 77843, U.S.A.
[*]*email:* bhramar@umich.edu

Summary. In case–control studies of gene-environment association with disease, when genetic and environmental exposures can be assumed to be independent in the underlying population, one may exploit the independence in order to derive more efficient estimation techniques than the traditional logistic regression analysis (Chatterjee and Carroll, 2005, *Biometrika* **92,** 399–418). However, covariates that stratify the population, such as age, ethnicity and alike, could potentially lead to nonindependence. In this article, we provide a novel semiparametric Bayesian approach to model stratification effects under the assumption of gene-environment independence in the control population. We illustrate the methods by applying them to data from a population-based case–control study on ovarian cancer conducted in Israel. A simulation study is conducted to compare our method with other popular choices. The results reflect that the semiparametric Bayesian model allows incorporation of key scientific evidence in the form of a prior and offers a flexible, robust alternative when standard parametric model assumptions do not hold.

Key words: Dirichlet process prior; Exponential family; Gene-environment interaction; Logistic regression; Ovarian cancer; Stratification factors; Zero inflated.

1. Introduction

Except for some rare diseases, such as Huntington or Tay Sachs disease that may be the result of a deficiency of a single gene product, most common human diseases have a multifactorial etiology involving complex interplay of many genetic and environmental factors. By identifying and characterizing such complicated gene-environment interactions, one has more opportunities to study etiology, diagnosis, prognosis, and treatment of complex diseases.

The case–control study design, where sampling is conditional on the presence or absence of disease, is a powerful epidemiologic tool for studying potential risk factors of rare diseases. It has been established that prospective logistic regression analysis of case–control data is "efficient" in the modern semiparametric sense with respect to a completely unrestricted covariate density model (Breslow, Robins, and Wellner, 2000). A special aspect of the gene-environment association problem is that it may often be reasonable to assume that a subject's genetic susceptibility (G) is independent of the environmental exposure (E). Consequently, one may be able to obtain more efficient estimation techniques than the traditional logistic regression, by exploiting the additional G–E independence restriction.

However, methods that use G–E independence produce severely biased estimates if the assumption is violated (Schmidt and Schaid, 1999; Albert et al., 2001). Nonindependence is less likely to occur when the environmental exposure is external (pollution, pesticide, or radioactive substance) or a randomized treatment in a clinical trial. One has to be much more cautious with the independence assumption when considering behavioral risk factors and metabolic polymorphisms that could alter an individual's behavior. Gatto et al. (2004) discuss several such potential sources of nonindependence. In fact, genetic susceptibility factors and environmental exposures, though unlikely to be causally related at an individual level may be correlated at a population level due to their dependence on other variables that stratify the population, such as age, ethnicity, family history, and alike. For example, a woman with a strong family history of breast cancer is more likely to carry BRCA1/2 (two major genes identified for breast and ovarian cancer) mutation and knowing her family history, less likely to use postmenopausal hormones. This may result in a negative association between BRCA1/2 mutation and hormone use. In such instances, G–E independence does not hold marginally, but may hold when conditioned on the stratification variables (for instance, family history). Modeling stratification effects can thus be viewed as a possible remedy to guard against resultant bias due to violation of the G–E independence assumption. One of the major goals of this article is to develop techniques to model stratification effects in a flexible, data-adaptive way in an estimation framework that exploits *conditional* G–E independence.

Piegorsch, Weinberg, and Taylor (1994) first observed that one can estimate the multiplicative *G–E* interaction parameter with data from cases alone, provided that G and E are independent in the population and the disease is rare. They also noted that the estimate of the *G–E* interaction parameter from case-only data is more efficient than its counterpart obtained from case–control data. Umbach and Weinberg (1997) showed that using data on both cases and controls, one can estimate the main effects and interaction by fitting a suitably constrained log-linear model under a rare disease assumption. In a population-based case–control study of ovarian cancer in Jewish women in Israel, Modan et al. (2001) argued that under *G–E* independence and rare disease assumption, the disease odds ratio associated with E among subjects with genotype $G = g$ can be estimated by a logistic regression analysis that compares $P(E \mid D = 0)$ with $P(E \mid D = 1, G = g)$. However, the method proposed in Modan et al. (2001) also does not allow for the estimation of all main effects of interest. The extension of these methods in the presence of stratification factors is not immediate as well.

Chatterjee and Carroll (2005) hereafter referred as CC propose a semiparametric maximum likelihood method of estimation of *all* the logistic regression parameters. They exploit the *G–E* independence assumption and use data from both cases and controls. Their method addresses many of the limitations of the existing methods as discussed above. CC derive a robust profile-likelihood-based estimation technique that does not require the rare disease assumption. They also consider the issue of stratification effects and propose a method when the *G–E* independence assumption only holds conditional on the set of stratification variables (S). CC consider a logistic disease probability model for $P(D \mid G, E, S)$. They proceed to work with the joint retrospective likelihood of the form $P(G, E, S \mid D)$, factorized as,

$$P(G, E, S \mid D) = \frac{P(D \mid G, E, S)P(G \mid E, S)P(E, S)}{\sum_{\text{G,E, and S}} P(D \mid G, E, S)P(G \mid E, S)P(E, S)}.$$

Due to the *G–E* independence assumption conditional on S, $P(G \mid E, S)$ reduces to $P(G \mid S)$, and thus it remains to model $P(E, S)$ and $P(G \mid S)$. CC leave the joint distribution of E and S, $P(E, S)$, to be fully nonparametric, but model $P(G \mid S)$ by assuming a logistic regression with S as covariate. As we will note, the parametric logistic model for the $P(G \mid S)$ is often inadequate, especially for a genetic mutation that is rarely detected in healthy controls but commonly prevalent in the case population. To overcome this problem, we use a factorization of the partially retrospective likelihood $P(G, E \mid D, S)$ that allows us to model the genotype frequencies separately in the case and the control population. Moreover, a flexible Bayesian model can use information obtained from genetic models (Parmigiani, Berry, and Aguilar, 1998) as well as empirical data (Couch et al., 1997), which predict population frequencies for BRCA1/2 mutation after adjusting for covariates. This information could be used in the form of an informative prior distribution assigned to $P(G \mid S)$ and lead to more accurate estimation. To elicit this advantage of the Bayesian paradigm while estimating all the parameters in the *G–E* logistic regression model, and *not just G–E* interaction, remains another goal of this article.

The data set we use is a replica of the one that CC use, based on a case–control study on ovarian cancer patients in Israel (Modan et al., 2001). We consider the presence of mutation of BRCA1/2 as the genetic risk factor and the number of years of oral contraceptive (OC) use and parity as the environmental exposures. The stratification variables we consider are age group, ethnicity, personal history of breast cancer (PHB), and family history of breast and ovarian cancer (FHBO). We model the control distribution of the continuous environmental exposures conditional on S as a Dirichlet process mixture of normals (DPM), which provides a natural measure of the degree of stratification and leads to model-robust inference. An extensive simulation study providing an in-depth comparison of the proposed Bayesian methods with the powerful estimation techniques provided by CC, the case-only method and ordinary logistic regression is a very important feature of this article. Our simulation explores several scenarios, with changing distributions for G and E as well as under violation of the *G–E* independence assumption even when conditioned on observable confounders.

It appears that under *G–E* independence, the proposed semiparametric Bayesian method has a real advantage over the competing methods under any of the following situations (i) the individual genotype frequencies in each stratum do not follow the logistic multiplicative odds model in terms of stratification variables, and (ii) the genetic mutation is rare in the control population and is commonly prevalent in the case population. The gain is significant when the number of strata defined by S is relatively large. If the *G–E* independence assumption even when conditional on S fails, all the methods that use this assumption perform poorly, least so for the Bayesian semiparametric method, which is more robust to model misspecification.

The rest of the article is organized as follows. In Section 2, we present the model, likelihood, priors, and posteriors. Section 3 contains analysis of the Israeli ovarian cancer data. Section 4 presents our simulation study and the results. Section 5 contains concluding discussion, while proofs and computational details are relegated to the Appendix available online.

2. Model, Likelihood, Priors, and Posteriors

Consider a case–control study with n subjects, of which n_1 are cases and n_0 are controls. Let D be the binary disease variable, i.e., $D_j = 1$ if the jth subject is a case, and $D_j = 0$ if the subject is a control. The genetic risk factor G is essentially the genotype at a single locus within a candidate gene. We will consider G as a categorical variable with $M + 1$ levels, namely $g_0, \ldots, g_M$. In addition, the data are assumed to be stratified based on some other covariates, say $\boldsymbol{S}$. We consider the following logistic regression function to model the disease probability in terms of G, E, and $\boldsymbol{S}$,

$$\begin{aligned} &P(D = 1 \mid G, E, \boldsymbol{S}) \\ &\quad = H\left\{\beta_0(\boldsymbol{S}) + \sum_{m=0}^{M} I(G = g_m)\beta_{1m} + \beta_2 E \right. \\ &\qquad \left. + E\sum_{m=0}^{M} \beta_{3m} I(G = g_m)\right\}, \end{aligned} \tag{1}$$

where $H(u) = \{1 + \exp(-u)\}^{-1}$. The intercept $\beta_0(\boldsymbol{S})$ captures stratification effects due to the covariates $\boldsymbol{S}$ on the risk of disease. Let $\boldsymbol{\beta}_1 = (\beta_{10}, \ldots, \beta_{1M})$, β_2, and $\boldsymbol{\beta}_3 = (\beta_{30}, \ldots, \beta_{3M})$ represent the main effect of the genetic factor, the main effect of the environmental factor, and their interaction effect, respectively. For parameter identifiability, we set $\beta_{10} = 0$ and $\beta_{30} = 0$. For simplicity, we present our model with only one continuous environmental exposure. Extension to multiple continuous exposures E is straightforward (see Remark 3). One could adapt this method when E is a mixed set of continuous and categorical exposures by using other nonparametric methods like the Bayesian bootstrap.

As we continue to compare and contrast our methods with CC and traditional logistic regression, we would first like to point out that each method is based on a different likelihood; the CC method uses a fully retrospective likelihood, $P(G, E, S \mid D)$, the traditional logistic model uses a fully prospective likelihood, $P(D \mid G, E, S)$, whereas our method uses the following partially retrospective likelihood $P(G, E \mid D, S)$ factorized as

$$L_R = \prod_{j=1}^{n} P(G_j, E_j \mid \boldsymbol{S}_j, D_j) = \prod_{j=1}^{n} \{P(G_j \mid E_j, \boldsymbol{S}_j, D_j) P(E_j \mid \boldsymbol{S}_j, D_j)\}. \quad (2)$$

As illustrated in Prentice and Pyke (1979) and discussed in Roeder, Carroll, and Lindsay (1996), the form of the retrospective likelihood considered here is compatible with the logistic form of the prospective likelihood. Evaluation of the likelihood function (2) requires the conditional distributions of $[G \mid E, \boldsymbol{S}, D]$ and $[E \mid \boldsymbol{S}, D]$. We make the following assumption:

Assumption 1: *Conditional on* $\boldsymbol{S}$, *G and E are independent in the control population, i.e.,* $P(G \mid D = 0, E, \boldsymbol{S}) = P(G \mid D = 0, \boldsymbol{S})$.

When the disease is rare in each stratum, the control population mimics the entire population, thus the usual *G–E* independence assumption in source population, i.e., $P(G \mid E, \boldsymbol{S}) = P(G \mid \boldsymbol{S})$ is approximately equivalent to Assumption 1. The two assumptions of *G–E* independence in source population and rare disease are made by Piergorsch et al. (1994), Umbach and Weinberg (1997), and Modan et al. (2001), while CC do not need the rare disease assumption. Our analysis is exact under Assumption 1. As pointed out in Schmidt and Schaid (1999), the rare disease assumption is quite subtle and may not hold, for example, in situations where the disease risk is much higher for the carriers of a particular gene mutation or for certain strata of the population. In the dataset we consider, the risk of ovarian cancer is known to be higher for BRCA1/2 carriers and for subjects with family history of breast or ovarian cancer. Fortunately, the bias due to the rare disease assumption has less impact when the overall disease prevalence $P(D = 1)$ is small, even with highly penetrant genes (Schmidt and Schaid, 1999).

We do recognize that directly verifying Assumption 1 empirically could be quite difficult based on the given study at hand, as tests of independence will have little power. Many researchers have considered this issue of verifying *G–E* independence in the control population in the context of using this as a screening tool to validate the use of case-only analysis (Albert et al., 2001). The *G–E* association pattern in controls reflect *G–E* association in source population if baseline disease risk is less than 0.1% (Gatto et al., 2004). To address this issue, in our simulations, we do consider performance of all the methods under various departures from Assumption 1. We advocate that when substantial uncertainty remains on the validity of the independence assumption, statistically significant results based on the proposed methods should be treated as precursors for high-priority investigations for future epidemiologic studies.

Assuming that the first n_0 observations are controls and the next $n - n_0$ observations are cases, under Assumption 1, the retrospective likelihood in (2) reduces to

$$L_R = \prod_{j=1}^{n_0} \{P(G_j \mid \boldsymbol{S}_j, D_j = 0) P(E_j \mid S_j, D_j = 0)\} \times \prod_{j=n_0+1}^{n} \{P(G_j \mid E_j, \boldsymbol{S}_j, D_j = 1) P(E_j \mid \boldsymbol{S}_j, D_j = 1)\}.$$

Consequently, to evaluate the likelihood contributed from control data we will need to specify probability models for $P(G \mid \boldsymbol{S}, D = 0)$ and $P(E \mid \boldsymbol{S}, D = 0)$. Following the technique first suggested by Satten and Kupper (1993), we present the following Lemmas which will then furnish expressions for $P(G \mid \boldsymbol{S}, E, D = 1)$ and $P(E \mid \boldsymbol{S}, D = 1)$, once we have the control distributions and the prospective model as in (1)

Lemma 1:

$$\frac{P(G = g_m \mid E, \boldsymbol{S}, D = 1)}{P(G = g_m \mid E, \boldsymbol{S}, D = 0)} = \frac{P(D = 1 \mid G = g_m, E, \boldsymbol{S}) / P(D = 0 \mid G = g_m, E, \boldsymbol{S})}{P(D = 1 \mid E, \boldsymbol{S}) / P(D = 0 \mid E, \boldsymbol{S})}.$$

Lemma 2:

$$\frac{P(D = 1 \mid E, \boldsymbol{S})}{P(D = 0 \mid E, \boldsymbol{S})} = \sum_{m=0}^{M} \frac{P(D = 1 \mid G = g_m, E, \boldsymbol{S})}{P(D = 0 \mid G = g_m, E, \boldsymbol{S})} P(G = g_m \mid D = 0, E, \boldsymbol{S}).$$

Lemma 3:

$$\frac{P(E \mid \boldsymbol{S}, D = 1)}{P(E \mid \boldsymbol{S}, D = 0)} = \frac{P(D = 1 \mid E, \boldsymbol{S}) / P(D = 0 \mid E, \boldsymbol{S})}{\int \frac{P(D = 1 \mid E, \boldsymbol{S})}{P(D = 0 \mid E, \boldsymbol{S})} P(E \mid \boldsymbol{S}, D = 0)\, dE}.$$

The proofs of the Lemmas are collected in Web Appendix A.

Remark 1. With our likelihood conditional on $\boldsymbol{S}$, we do not intend to estimate the relative risks due to $\boldsymbol{S}$ and focus only on our parameter of interest $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \beta_2, \boldsymbol{\beta}_3)$. As we proceed,

we note that under our formulation, we would tacitly avoid direct estimation of $\beta_0(\boldsymbol{S})$.

Before describing the estimation theory, we first would like to address the identifiability of the parameters in the prospective model (1) and the retrospective likelihood L_R. As stated in Prentice and Pyke (1979), if there are no assumptions made on the covariate distribution $\mathcal{H}(g, e \mid \boldsymbol{s}) = P(G = g, E = e \mid \boldsymbol{S} = \boldsymbol{s})$, neither $\mathcal{H}(\cdot, \cdot \mid \boldsymbol{s})$, nor $\beta_0(\boldsymbol{S})$ is identifiable. But $\boldsymbol{\beta}$ is always identifiable under any choice of $\mathcal{H}$. Following lemma 1 of Roeder et al. (1996) it can be easily shown that under Assumption 1 on the covariate density, $\boldsymbol{\beta}$ remains identifiable in our likelihood L_R.

We consider $\boldsymbol{S}$ as a vector of $q \geq 1$ categorical covariates, with the kth variable having r_k categories or levels. Therefore, the level combinations of $\boldsymbol{S}$ define $I = \prod_{k=1}^{q} r_k$ possible strata. For instance, in the Israeli ovarian cancer data we consider $q =$ 4 stratification variables: (Age group, ethnicity, PHB, FHBO), the first three each having two categories and FHBO having three categories. Therefore, $\boldsymbol{S}$ defines $I = 2 \times 2 \times 2 \times 3 =$ 24 possible strata. Define a categorical variable Z which can assume I possible values, with each value corresponding to a distinct level combination of the set of stratification variables. We can now rewrite L_R after replacing $\boldsymbol{S}_j$ by the stratum membership indicator of subject j, namely Z_j.

$$L_R = \prod_{j=1}^{n_0} \{P(G_j \mid Z_j, D_j = 0) P(E_j \mid Z_j, D_j = 0)\} \times \prod_{j=n_0+1}^{n} \{P(G_j \mid E_j, Z_j, D_j = 1) P(E_j \mid Z_j, D_j = 1)\}. \tag{3}$$

We consider the following model for the control distribution of the genetic factor in stratum i,

$$\log \frac{P(G = g_m \mid Z = i, D = 0)}{P(G = g_0 \mid Z = i, D = 0)} = \gamma_{im}, \quad m = 1, \ldots, M. \tag{4}$$

Note that $\gamma_{i0} = 0$. The above model does not assume any stringent parametric form for $P(G \mid D = 0, \boldsymbol{S})$ in terms of $\boldsymbol{S}$ but simply treats the probabilities in each stratum to be the model parameters, allowing complete distributional flexibility.

Result 1. Using (1), (4), and Lemma 1, we obtain the case distribution of G as:

$$P(G = g_m \mid E, Z = i, D = 1) = \frac{\exp(\beta_{1m} + \beta_{3m} E + \gamma_{im})}{1 + \sum_{k=1}^{M} \exp(\beta_{1k} + \beta_{3k} E + \gamma_{ik})}, \quad m = 1, \ldots, M. \tag{5}$$

Proof of Result 1 is presented in Web Appendix A. Note that although in the control population by virtue of the independence assumption, $P(G \mid E, D = 0, Z = i) = P(G \mid D = 0, Z = i)$, in the case population, $P(G \mid E, D = 1, Z = i)$ does depend on E.

Due to the high-dimensional nature of the stratification variables $\boldsymbol{S}$, it is often hard to model the effect of $\boldsymbol{S}$ on the distribution of E explicitly. We consider a flexible semiparametric Bayesian approach to model the distribution $[E \mid D = 0, Z = i]$, which allows for possible stratification effects on the distribution of E and does so in a data-adaptive way. We consider the case when E is continuous, as in our data example. Our DPM with a normal kernel can be expressed in the following hierarchical structure

$$[E \mid D = 0, Z = i] \sim N\big(\mu_i, \sigma_i^2\big) \text{ with } \boldsymbol{\theta}_i = \big(\mu_i, \sigma_i^2\big) \,\big|\, \mathcal{P} \sim \mathcal{P} \text{ and } \mathcal{P} \sim \mathrm{DP}(\alpha \mathcal{P}_0), \tag{6}$$

where $\mathcal{P}$, serving as a prior on the $\boldsymbol{\theta}_i$, $i = 1, \ldots, I$, is itself a *random* probability measure. We assume that $\mathcal{P}$ is realization of a Dirichlet process (DP) with a scalar precision parameter $\alpha \geq 0$ and base measure (or base prior) $E[\mathcal{P}] = \mathcal{P}_0$ which is a bivariate cumulative distribution function on $\mathcal{R} \times \mathcal{R}^+$. A property of the DP prior is that the random probability measure $\mathcal{P}$ is almost surely discrete, leading to the following properties which reinterpret the DPM model structure (see Antoniak, 1974): (i) Any realization of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_I$ generated from $\mathcal{P}$ lies in a set of $K(\leq I)$ distinct values, denoted by $\boldsymbol{\omega} = \{\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_K\}$; (ii) $\boldsymbol{\omega}_l$, $(l = 1, \ldots, K)$ are a random sample from the base prior $\mathcal{P}_0$; (iii) $K(\leq I)$ is drawn from an implicitly determined prior distribution depending on the precision parameter α and I; and (iv) Given $K \leq I$, the I values are selected from the set $\boldsymbol{\omega}$ according to a uniform multinomial distribution.

The above discussion is conditional on α and the hyperparameters which determine $\mathcal{P}_0$.

With this hierarchical mixture prior structure for the control distribution of E and the prospective logistic model (1), it now remains to investigate the nature of the case distribution of E. The following result provides an answer.

Result 2. Assume that the $\boldsymbol{\theta}_i$ take values $\boldsymbol{\omega}_l$ from the set $\boldsymbol{\omega}$ as described in (i). Then

$$[E \mid Z = i,\ D = 1, \boldsymbol{\theta}_i = \boldsymbol{\omega}_l] = \sum_{m=0}^{M} p_{ilm} \phi\big(E; \boldsymbol{\omega}_{lm}^*\big), \tag{7}$$

where $\phi(\cdot; \boldsymbol{\theta})$ denotes the normal density with a parameter vector $\boldsymbol{\theta}$, $\boldsymbol{\omega}_l = (\mu_l, \sigma_l^2)$, say, and $\boldsymbol{\omega}_{lm}^* = (\mu_l + \beta_2 \sigma_l^2 + \beta_{3m} \sigma_l^2, \sigma_l^2)$ and $p_{ilm} = \exp\{\beta_{1m} + (\mu_l + \beta_2 \sigma_l^2 + Z\beta_{3m}\sigma_l^2)^2/(2\sigma_l^2) + \gamma_{im}\} / \sum_{k=0}^{M} \exp\{\beta_{1k} + (\mu_l + \beta_2 \sigma_l^2 + \beta_{3k}\sigma_l^2)^2/(2\sigma_l^2) + \gamma_{ik}\}$. Hence, the distribution of E in the case population, conditional on all other parameters is again a DPM but not with a normal kernel but with a mixture kernel given by equation (7).

The exact expression of the likelihood (3) and proof of Result 2 are deferred to Web Appendices B and A, respectively. We will refer to this model as EDPM for future references.

Prior structure: The likelihood in equation (3) involves the association parameters $\boldsymbol{\beta}_1$, β_2, $\boldsymbol{\beta}_3$, and $\gamma_{i1}, \ldots, \gamma_{iM}$, and $\theta_i = (\mu_i, \sigma_i^2)$, $i = 1, \ldots, I$. We use independent normal priors for all the association parameters and also on γ_{im}'s, $m = 1, \ldots, M$. We will note in our real data example (with only two possible values of G, so that $m = 0, 1$) that if we a priori know that the mutation is rare in the control population, and have an established prediction model for mutation frequencies $P(G \mid \boldsymbol{S})$, we should select an informative prior on γ_{i1}, so that the effective range of the carrier probabilities in the control/case population for each stratum reflects the scientific guesses for these values.

It now remains to describe the hierarchical prior structure involved in the DPM model. Note that the mean of the random probability measure $\mathcal{P}$ is $\mathcal{P}_0$ is a bivariate distribution, and we consider the following standard normal-inverse gamma (IG) structure, namely, under $\mathcal{P}_0, \mu_i \,|\, \sigma_i^2 \sim N(m_0, \tau\sigma_i^2), (\sigma_i^2) \sim \text{IG}(s/2, S/2)$. We add an extra layer of uncertainty in $\mathcal{P}_0$ by using Normal$(\mu_{m_0}, \sigma_{m_0}^2)$ and IG $(a_\tau/2, b_\tau/2)$ prior on m_0 and τ, respectively. Lastly, following Escobar and West (1995), we assume a Gamma (a_α, b_α) prior on α, and the hyperparameters a_α and b_α are so chosen that a priori mean of K is reasonably large (compared to I) and the variance is modest. Choosing such a "further" prior is suggested in West, Müller, and Escobar (1994).

None of the full conditional distributions follows a standard distributional form and posterior inference is made by using the Markov chain Monte Carlo (MCMC) numerical integration technique. Conditional on $\boldsymbol{\theta}_i$, drawing random numbers from the respective conditional distributions are straightforward applications of the Metropolis–Hastings algorithm. To update the $\boldsymbol{\theta}_i$ we use the no gaps algorithm prescribed by MacEachern and Müller (1998). We describe the computational details of our algorithm in Web Appendix C.

Remark 2. An interesting feature of the EDPM model is that it selects K, the number of distinct values in I realizations from $\mathcal{P}$ or the cardinality of the set $\boldsymbol{\omega}$ in a data-adaptive way depending on the extent of stratification effects on the control distribution of E. In the presence of strong stratification effects, all of the $\boldsymbol{\omega}_j$ could be distinct, i.e., $K = I$; in the complete absence of stratification effects, $K = 1$. Typically K will lie somewhere in between. The posterior mode of K thus serves as an indicator of the degree of stratification effects on the control distribution of E.

In the above discussion we have assumed $\boldsymbol{S}$ to be a set of categorical stratification variables which is most often the case. If any of the stratification variables is continuous, we recommend categorizing them for implementing the EDPM model.

Remark 3. The proposed method can be applied to a multivariate continuous environmental exposure variable by considering multivariate normal kernel in the DP structure, which has been illustrated in our additional real data analysis available online.

Remark 4. Note that, as indicated in Remark 1, via the above formulation, the nuisance parameters $\beta_0(\boldsymbol{S})$ do not present themselves in the case distributions of G and E as presented through Lemmas 1 and 3, respectively. Because the analysis is conditional on $\boldsymbol{S}$, $\beta_0(\mathbf{S})$ appears as a common term in both the numerator and denominator of equations (5) and (7) and thus gets canceled in the ratio. Hence, the retrospective likelihood does not involve $\beta_0(\mathbf{S})$.

Remark 5. One could naturally think of modeling the distribution of G as $\log\{P(G_j = g_m \,|\, D_j = 0, \boldsymbol{S}_j)/P(G_j = g_0 \,|\, D_j = 0, \boldsymbol{S}_j)\} = \nu_0 + \boldsymbol{\nu}_m^T \boldsymbol{S}_j$, $m = 1, \ldots, M$, where $\boldsymbol{\nu}_m$ is a vector of regression parameters capturing the effect of stratification variables on the incidence of the genetic susceptibility factor in the control population. Because BRCA1/2 is very rare in the control population, it is hard to predict BRCA1/2 carrier probabilities using this logistic structure. Thus we only consider model (4).

3. The Israeli Ovarian Cancer Data

In this section, we apply the proposed methodology to the data from a population-based case–control study on all ovarian cancer patients identified in Israel between March 1, 1994 and June 30, 1999 (Modan et al., 2001). Blood samples were collected from the cases and the controls in order to test for the presence of mutation in the two major breast and ovarian cancer susceptibility genes BRCA1 and BRCA2. In addition, the subjects were interviewed to collect data on reproductive/gynecological history such as parity, number of years of OC use, and gynecological surgery. The main goal of the study was to examine the interplay of the BRCA1/2 genes and known reproductive/gynecological risk factors of ovarian cancer. Because the actual data had confidentiality issues, a replica was generated by replacing only the original genetic susceptibility factor by a simulated binary genetic risk factor, retaining all the features as in the original dataset. The dataset we used contained 832 cases and 747 controls.

This is a real example where OC use and BRCA1/2 mutation may appear to be correlated simply because both could be related to the stratification variables $\boldsymbol{S}$ like age, ethnicity, and family history, and it is more realistic to assume independence between these two genetic and gynecological risk factors conditional on $\boldsymbol{S}$. However, it is hard to verify Assumption 1 based on this single dataset as only 7 out of the 747 controls were BRCA1/2 carriers. We ran a logistic regression of G on the exposures of interest E in the controls in each stratum, and though the tests of association were insignificant, the sparsity of the data makes the results of these tests for association unstable and less reliable. However, Modan et al. (2001, p. 236) and CC both indicate that it is reasonable to assume that carrier status is independent of the exposures under consideration, namely parity and number of years of OC use, and we also employ this assumption in our analysis.

It is known that the risk of ovarian cancer is higher for certain strata (e.g., for the subgroup with family history of both breast and ovarian cancer) as well as for BRCA1/2 carriers. So the rare disease assumption may not hold for all levels of the genetic factor or for certain subgroups. However, Modan et al. (2001) reported only 1326 cases of epithelial ovarian cancer during the 5-year study period with a baseline population of approximately 1.5 million, suggesting an empirical estimate of $P(D = 1) = 8.7 \times 10^{-4}$. Thus the odds-ratio estimates obtained through our analysis under Assumption 1 will provide adequate approximations to the ones obtained via exact analysis using *G–E* independence in source population.

All analyses are carried out conditional on four stratification variables $\boldsymbol{S}$ = [age group (=0 if age <50 years and = 1 if age $\geq$ 50 years), ethnicity (=1 for Ashkenazi Jews and 0 otherwise), PHB (=1 if present and 0 if absent), FHBO (= 0 if no history, 1 if one breast cancer case in family and 2 if ovarian cancer or two or more breast cancer cases in the family)]. So the total number of strata defined by the level combinations of $\boldsymbol{S}$ is $I = 24$.

We analyze the data using the EDPM method as described in the previous section. For modeling the distribution of the genetic factor, we use equation (4). The genetic factor G is binary with $G = 0$ for absence of any BRCA1/2 mutation

and $G = 1$ for carrying at least one BRCA1/2 mutation. Due to rare BRCA1/2 mutation among the controls, traditional logistic regression analysis would yield imprecise estimates of the parameters of interest (Modan et al., 2001). Compounding to the sparsity is the fact that we do have a relatively large number of strata defined by $\boldsymbol{S}$ and as a result, estimation of genotype frequencies individually in each stratum would be imprecise in a classical setup. CC adopt a parametric logistic model for $P(G \mid \boldsymbol{S})$ to circumvent this problem which is also not satisfactory. In a Bayesian paradigm, we effectively use the prior knowledge on BRCA1/2 carrier probabilities with varying levels of family history, age, and ethnicity based on genetic algorithms (BRCAPRO: Parmigiani et al., 1998) and empirical data (Couch et al., 1997). We allow uncertainty in these predictions by allowing the informative prior on γ_{i1} to vary around the scientific guesses and in this process relax the stringent logistic assumption of CC. The effective range of prior probabilities for $P(G = 1 \mid \boldsymbol{S}, D = 0)$ typically varied from 10^{-1} to 10^{-4} across different strata.

We consider OC use as the only environmental exposure (E) as a direct illustration of the methods formulated in Section 2. With a binary G, the disease-risk model (1) becomes $\text{logit}\, P(D = 1 \mid G,\ E,\ \boldsymbol{S}) = \beta_0(\boldsymbol{S}) + \beta_{\text{BRCA}} I[\text{BRCA1/2} = 1] + \beta_{\text{OC}}\text{OC} + \beta_{\text{BRCA}*\text{OC}} I[\text{BRCA1/2} = 1] *\text{OC}$. For each of β_{BRCA}, β_{OC}, $\beta_{\text{BRCA}*\text{OC}}$ we use $N(0, 16)$ prior. Because scientific theory suggests high positive value of β_{BRCA}, and a protective effect of OC for the noncarriers, one could also select a sharper prior for β_{BRCA} and β_{OC}, resulting in slightly more precise parameter estimates (web Table 1). For the EDPM model as described in equation (6), under the base-measure $\mathcal{P}_0$, we assume that the variance component $\sigma^2 \sim IG(2, 1)$ and $\mu \mid \sigma^2 \sim N(m_0, \tau\sigma^2)$. The exposure variable, number of years of OC use typically ranges from 0 to 20 years. We chose the $N(3, 9)$ prior for m_0, and the IG(3,1) prior on τ. Choosing priors for α is a challenging task as α has the dual role of capturing the degree of faith in the base measure, as well as determining the number of distinct values of $\boldsymbol{\theta}$. As prescribed by Escobar and West (1995), we choose a Gamma prior on α which allows for prior probabilities for larger values of $K \leq I = 24$. We experimented with various choices of the shape and scale parameters of the Gamma prior, and the results are presented for Gamma (4, 1).

For comparison purposes, we also analyzed these data with a parametric model, largely targeted towards this dataset. As 81% of the cases and 86% of the controls in the data did not use OCs at all, we used a zero-inflated model (EZIM) for the control distribution of OC use. For individual j, we consider p_j as the probability of nonexposure ($E_j = 0$), and with probability $(1 - p_j)$, the exposure values follow $N(\mu_j, \sigma^2)$, where $\mu_j = \delta_0 + \boldsymbol{\delta}_1^T \boldsymbol{S}_j$. The mixing probabilities are also modeled through the four observed stratification factors, $\text{logit}(p_j) = \eta_0 + \boldsymbol{\eta}_1^T \boldsymbol{S}_j$. The case distributions can be obtained as mixture distributions via Lemmas 1–3. For the EZIM model, we consider mutually independent N(0, 16) prior for the regression parameters, β_{BRCA}, β_{OC}, and $\beta_{\text{BRCA}*\text{OC}}$, as well as on δ_0, η_0's and each component of $\boldsymbol{\delta}_1$ and $\boldsymbol{\eta}_1$. For the scale parameter σ^2 we use IG(2,1) prior. Posterior inference is again based on MCMC samples from the full conditional distribution of the parameters.

We analyzed these data through the method proposed by CC and the case-only method after adjusting for the covariates $\boldsymbol{S}$. The case-only method only furnishes estimate of the BRCA*OC interaction parameter. The results are presented in Table 1. There is little in the way of differences for estimation of β_{OC} and $\beta_{\text{BRCA}*\text{OC}}$ by all the four methods that use *G–E* independence. But for estimating the main effect of BRCA 1/2 carrier status as measured by β_{BRCA}, the Bayesian methods have much smaller posterior standard deviations and narrower HPD intervals compared to the standard error and the CI for the estimate of β_{BRCA} in the CC method. The results indicate that standard logistic assumption is less likely to hold for $P(G \mid \boldsymbol{S})$ in this dataset, and the more flexible model

Table 1

*Analysis of Israeli ovarian cancer data by all five methods, with presence/absence of BRCA*1/2 *mutation as the genetic factor and OC use as the environmental exposure. Estimates of the log odds ratios corresponding to the main effects and the interaction parameter are presented with* 95% *HPD and confidence intervals. The analysis is adjusted for age, ethnicity, family history of breast or ovarian cancer, and PHB cancer.*

Model		β_{BRCA}	β_{OC}	$\beta_{\text{BRCA}*\text{OC}}$	α	K
EZIM[a]	Estimate	3.78	−0.05	0.09		
	post. st.dev.	0.13	0.02	0.03		
	HPD	(3.46, 3.98)	(−0.13, −0.01)	(0.03, 0.15)		
EDPM[b]	Estimate	3.75	−0.07	0.11	14.75	5.76
	post. st.dev.	0.13	0.03	0.04	5.84	1.88
	HPD	(3.44, 3.93)	(−0.14, −0.02)	(0.04, 0.18)	(5.89, 28.57)	(2, 10)
CC[c]	Estimate	3.63	−0.06	0.11		
	std. error	0.40	0.03	0.03		
	CI	(2.85, 4.42)	(−0.11, −0.01)	(0.04, 0.18)		
Ordinary	Estimate	3.77	−0.06	0.05		
Logistic	std. error	0.44	0.03	0.10		
	CI	(2.91, 4.63)	(−0.12, −0.01)	(−0.15, 0.24)		
Case only	Estimate			0.09		
	std. error			0.03		
	CI			(0.03, 0.16)		

[a]The parametric Bayesian approach with an EZIM for the control distribution of OC use.
[b]The semiparametric Bayesian approach with a DPM model for the control distribution of OC use.
[c]The profile likelihood method proposed by Chatterjee and Carroll (2005).

for G as given in equation (4), boosted with the scientifically validated priors, adapts itself more naturally to the features of the data. Interestingly, the semiparametric EDPM model performs quite comparably to the parametric EZIM that is designed specifically to capture the distribution of OC use.

We also analyzed the data by ordinary logistic regression analysis that does not exploit *G–E* independence in any manner. The wider confidence intervals, especially for the interaction parameter indicates that any method using *G–E* independence is able to estimate the interaction parameter more precisely. Whereas all the other four methods declare *G–E* interaction to be statistically significant, the ordinary logistic model cannot detect significance.

In summarizing the results, we first observe that for women who never used OC (E = 0), there is an almost astronomic increase in risk of ovarian cancer for a BRCA1/2 mutation carrier. The estimated odds ratio by the EDPM method is $\exp(3.75) = 42.52$. On the other hand, among noncarriers, longer use of OC is related to decrease in disease risk with associated odds ratio $\exp(-0.0748) = 0.92$. However, the estimate of the interaction parameter $\beta_{\text{BRCA}*\text{OC}}$ suggests that among BRCA1/2 carriers, the risk of ovarian cancer increases slightly with OC use, with an odds ratio $\exp(-0.0748) \times \exp(0.1091) = 1.03$, but this effect is not statistically significant (95% credible interval=[0.97, 1.11]). The precision estimates and the credible intervals all indicate that the main effect of BRCA1/2 and the BRCA–OC interaction are statistically significant whereas the main effect of OC use is only marginally significant.

Figure 1 presents plots of posterior distributions for the log odds-ratio parameters. To explore the degree of stratification, we also present a plot of posterior distribution of α, K, a boxplot of $\text{var}(\mu_i)$ and $\text{var}(\sigma_i)$ in the EDPM model ($i = 1, \ldots, 24$) in Figure 2. We notice that the μ_i and σ_i values do reflect variation, the variability in σ_i being greater. The posterior mode of K is at 5, suggesting that though there are 24 possible strata, not all of them have distinct effects on the distribution of number of years of OC use. The analysis with OC and parity both considered as environmental exposures is collected in web Table 2.

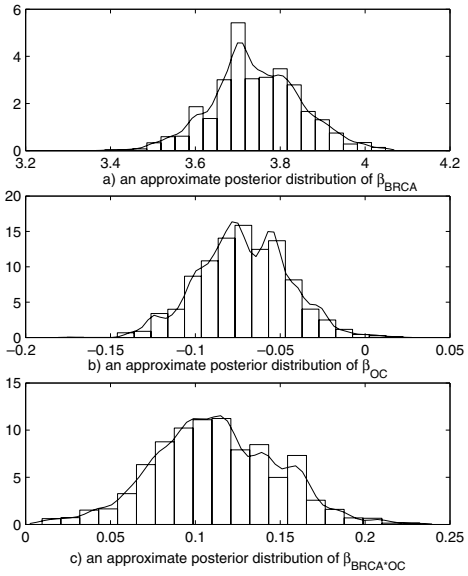

Figure 1. Israeli ovarian cancer data analyzed by the EDPM[a] model with presence/absence of BRCA1/2 mutation as the genetic factor and OC use as an environmental exposure: Histogram of last 5000 MCMC values for the main effects and the interaction parameter with overlayed smoothed kernel density estimate.
[a]The proposed semiparametric Bayesian approach using a DPM model for the control distribution of OC use.

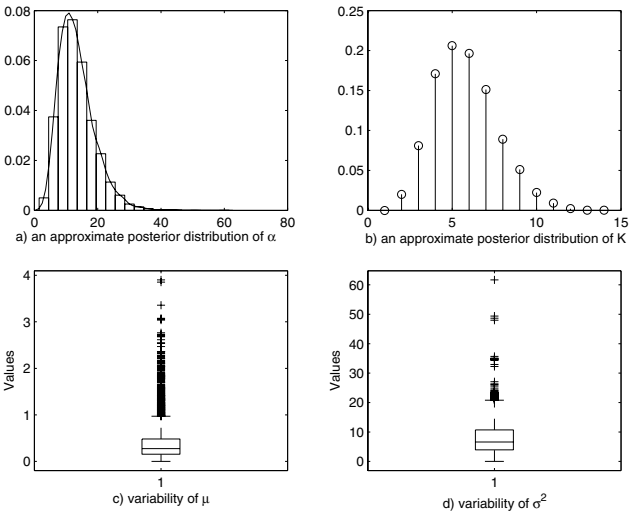

Figure 2. Secondary parameters related to the EDPM[a] model for analyzing the Israeli ovarian cancer data: Histograms corresponding to the approximate posterior distributions of α and K in the DPM model. Also presented are boxplots of variances of the μ_i's and σ_i's $i = 1, \ldots, 24$. Each plot is based on the last 5000 MCMC runs.
[a]The proposed semiparametric Bayesian approach using a DPM model for the control distribution of OC use.

In the following section, we conduct an extensive simulation study to assess the performances of the methods under different scenarios and provide recommendations for the practitioner.

4. Simulation

In order to simulate a dataset for comparing the Bayesian methods with the method proposed by CC, case-only analysis and ordinary logistic regression, we used the ovarian cancer data as a prototype. We set the true values close to the results we obtained in our analysis of real data by EDPM method in Table 1, $\beta_1 = 3$, $\beta_2 = -0.07$, and $\beta_3 = 0.12$. We generated 1500 observations following the scheme as below:

(1) We first generated the stratification factors $\boldsymbol{S}$ = (Age group, Ethnicity, PHB, FHBO) from a multinomial distribution, the stratum probabilities being consistent with the real study.
(2) Given $\boldsymbol{S}$, we generated a binary variable D representing the disease status, with probabilities $P(D = 1 \mid \boldsymbol{S})$ in agreement with the ovarian cancer study, the marginal disease probability in the generated population being around 0.1%. Results for a more common disease are included in the online supplementary material.

Table 2

Simulation scenarios: distribution of E is zero inflated; G: rare or common; G−E independence assumption holds ($\gamma_E = 0$) or does not hold ($\gamma_E = 0.25$). Mean denotes the mean estimate based on 100 replications, whereas MSE is the estimated mean squared error based on 100 replications.

G	γ_E	Model	True	β_1 3.00	β_2 −0.07	β_3 0.12
Rare	0	EZIM[a]	Mean	2.98	−0.06	0.11
			MSE	0.020	0.001	0.002
		EDPM[b]	Mean	2.93	−0.06	0.11
			MSE	0.024	0.001	0.002
		CC[c]	Mean	2.90	−0.06	0.12
			MSE	0.263	0.001	0.002
		Ordinary	Mean	2.91	−0.06	0.19
		Logistic	MSE	0.374	0.001	0.062
		Case-Only	Mean			0.12
			MSE			0.002
Rare	0.25	EZIM[a]	Mean	2.87	−0.19	0.31
			MSE	0.045	0.018	0.043
		EDPM[b]	Mean	2.85	−0.13	0.25
			MSE	0.056	0.003	0.027
		CC[c]	Mean	2.38	−0.20	0.33
			MSE	0.453	0.248	0.041
		Ordinary	Mean	3.16	−0.19	0.06
		Logistic	MSE	0.163	0.271	0.012
		Case-Only	Mean			0.37
			MSE			0.069
Common	0	EZIM[a]	Mean	2.98	−0.08	0.13
			MSE	0.010	0.002	0.001
		EDPM[b]	Mean	2.97	−0.08	0.13
			MSE	0.011	0.002	0.002
		CC[c]	Mean	2.77	−0.08	0.12
			MSE	0.081	0.002	0.001
		Ordinary	Mean	2.82	−0.08	0.12
		Logistic	MSE	0.066	0.002	0.003
		Case-Only	Mean			0.13
			MSE			0.001
Common	0.25	EZIM[a]	Mean	2.86	−0.28	0.34
			MSE	0.031	0.049	0.054
		EDPM[b]	Mean	2.90	−0.15	0.21
			MSE	0.022	0.006	0.009
		CC[c]	Mean	2.42	−0.31	0.37
			MSE	0.358	0.065	0.070
		Ordinary	Mean	2.80	−0.26	0.18
		Logistic	MSE	0.064	0.043	0.011
		Case only	Mean			0.40
			MSE			0.082

[a]The parametric Bayesian approach with an EZIM for the control distribution of E.
[b]The semiparametric Bayesian approach with a DPM model for the control distribution of E.
[c]The profile likelihood method proposed by Chatterjee and Carroll (2005).

(3) We generated E from two distributions: (i) An EZIM model, exactly mimicking the distribution of OC use as in the real dataset. The true values of all associated parameters were chosen as the estimates obtained from our real data when analyzed by the EZIM model. (ii) Mixture of two normal distributions: To deviate from the exact pattern of real data and to put our semiparametric and parametric methods to test, we considered the case when $[E \mid D = 0, Z = i]$ comes from the following mixture: $0.5 \times N(2, 1) + 0.5 \times N(5, 1)$.

(4) Finally, we generated a binary variable G standing for BRCA1/2 mutation status using the probability structure $P(G \mid D, E, Z)$ as given in equations (4) and (5). We select the true values for γ_{i1} in such a way that $\Pr(G = 1 \mid D = 0) \approx 3.3\%$ and $\Pr(G = 1 \mid D = 0) \approx 46.9\%$ to represent the two situations with a moderately rare and a common genetic mutation, respectively. Simulation results are also presented when G was generated from a logistic model in terms of $\boldsymbol{S}$.

Table 3

Simulation scenarios: distribution of E: Mixture of two normals; G: with parametric logistic in terms of $\boldsymbol{S}$ *(as in Remark* 5*) or commonly prevalent as in* (4)*; G−E independence holds* ($\gamma_E = 0$) *or does not hold* ($\gamma_E = 0.25$)*. Mean denotes the mean estimate based on* 100 *replications, whereas MSE is the estimated mean squared error based on* 100 *replications.*

G model	γ_E	Model	True	β_1 3.00	β_2 −0.07	β_3 0.12
Logistic in terms of $\boldsymbol{S}$	0	EZIM[a]	Mean	3.00	−0.04	0.13
			MSE	0.069	0.080	0.009
		EDPM[b]	Mean	2.99	−0.07	0.13
			MSE	0.059	0.002	0.002
		Ordinary	Mean	3.02	−0.08	0.13
		Logistic	MSE	0.128	0.002	0.008
		Case-Only	Mean			0.13
			MSE			0.002
Generated by (4)	0	EZIM[a]	Mean	2.82	−0.12	0.14
			MSE	0.036	0.089	0.008
		EDPM[b]	Mean	2.99	−0.07	0.12
			MSE	0.030	0.002	0.002
		CC[c]	Mean	2.75	−0.07	0.12
			MSE	0.110	0.002	0.002
		Ordinary	Mean	2.77	−0.07	0.13
		Logistic	MSE	0.186	0.002	0.009
		Case-Only	Mean			0.13
			MSE			0.002
Generated by (4)	0.25	EZIM[a]	Mean	2.86	−0.31	0.38
			MSE	0.157	0.213	0.189
		EDPM[b]	Mean	2.89	−0.27	0.35
			MSE	0.045	0.044	0.055
		CC[c]	Mean	1.93	−0.30	0.37
			MSE	0.890	0.056	0.065
		Ordinary	Mean	2.77	−0.25	0.15
		Logistic	MSE	0.137	0.037	0.008
		Case only	Mean			0.32
			MSE			0.089

[a]The parametric Bayesian approach with an EZIM for the control distribution of E.
[b]The semiparametric Bayesian approach with a DPM model for the control distribution of E.
[c]The profile likelihood method proposed by Chatterjee and Carroll (2005).

In order to study robustness under violation of *G–E* independence assumption, we simulate G using the model

$$\log\left\{\frac{\Pr(G=1\,|\,Z=i,E,D=0)}{\Pr(G=0|Z=i,E,D=0)}\right\}=\gamma_{i1}+\gamma_E E. \qquad (8)$$

To introduce moderate dependence between G and E, we consider $\gamma_E = 0.25$, that is, the odds of having $G = 1$ with one unit increase in E increases by a factor of 1.284 (Tables 2–4). The strategies followed for choosing priors for the Bayesian methods in the simulation study are essentially the same as discussed in the real data analysis. We simulated 100 datasets for each scenario, and the results are presented in Tables 2–4.

The simulation results are fairly clear. If interest lies in estimating the main effect of the genetic factor β_1, the Bayesian EDPM model performs the best for any choice of distributions of G and E. The fully parametric Bayesian EZIM model suffers when E is originated from any other model, for example the mixture of two normal distributions (Tables 3 and 4). When the parametric logistic assumption for $P(G\,|\,\boldsymbol{S})$ does not hold, there is a clear dominance of the Bayesian methods over the CC method for estimating β_1. Even when the data are generated from an exactly logistic model for $P(G\,|\,\boldsymbol{S})$ (Table 3), the Bayesian methods perform quite comparably with the CC method. The efficiency gain (for estimating β_1) in Bayesian methods is larger when the genetic mutation rarely occurs in the control population (Tables 2 and 4), which could be due to the flexibility of the likelihood in modeling the control distributions separately in the Bayesian methods, whereas CC model the marginal distribution of $G\,|\,\boldsymbol{S}$. If interest lies in estimating the main effect of E, both the CC method and the EDPM method are comparable, with CC method having a slight edge in some cases. One may note that the mean square error (MSE) corresponding to β_2 for the EDPM model is often larger than the other methods as with the DPM structure we are adding another level of model uncertainty. Indeed, the advantage of the DPM is not in terms of gain in efficiency for estimating β_2 across all scenarios, but because of its robustness. One may note that instead of modeling $P(E\,|\,\boldsymbol{S})$, CC model $P(E, \boldsymbol{S})$ nonparametrically. Their profile likelihood technique works extremely well across many different data

Table 4

Simulation scenarios: distribution of E: Mixture of two normals; G: rarely prevalent; G−E independence holds ($\gamma_E = 0$) or does not hold ($\gamma_E = 0.25$). Mean denotes the mean estimate based on 100 replications, whereas MSE is the estimated mean squared error based on 100 replications.

G	γ_E	$E \mid D = 0, \mathbf{Z}$	True value	β_1 3.00	β_2 −0.07	β_3 0.12
Rare	0	EZIM[a]	Mean	2.92	−0.04	0.16
			MSE	0.082	0.031	0.021
		EDPM[b]	Mean	2.93	−0.07	0.13
			MSE	0.067	0.002	0.004
		CC[c]	Mean	2.85	−0.07	0.13
			MSE	0.222	0.001	0.005
		Ordinary	Mean	2.87	−0.07	0.15
		Logistic	MSE	0.828	0.001	0.061
		Case-Only	Mean			0.13
			MSE			0.005
Rare	0.25	EZIM[a]	Mean	3.07	−0.20	0.31
			MSE	0.169	0.193	0.048
		EDPM[b]	Mean	3.15	−0.13	0.30
			MSE	0.097	0.006	0.034
		CC[c]	Mean	2.07	−0.14	0.31
			MSE	0.848	0.007	0.040
		Ordinary	Mean	3.15	−0.14	0.05
		Logistic	MSE	0.711	0.006	0.043
		Case only	Mean			0.35
			MSE			0.056

[a]The parametric Bayesian approach with an EZIM for the control distribution of E.
[b]The semiparametric Bayesian approach with a DPM model for the control distribution of E.
[c]The profile likelihood method proposed by Chatterjee and Carroll (2005).

generating mechanisms for *E*. For estimating the *G–E* interaction β_3, one could choose either case-only, EDPM, or the CC method. When simultaneous estimation of all three parameters is considered, and Assumption 1 is fairly reasonable, the EDPM model appears to be a superior choice. Under violation of the independence assumption, performance of all the methods worsen (Tables 2–4), least so for EDPM model; this could be attributed to robustness of the DP prior as well as the informative prior on *G*. The ordinary logistic regression model that is less efficient under *G–E* independence, especially for the interaction parameter, does not suffer under violation of *G–E* independence as it does not use any restrictions on the *G–E* distribution.

5. Discussion

The term "interaction" has a diverse meaning to the scientific community. "No statistical interaction" in our model means constant multiplicative effect of genotype on the disease odds across all levels of the environmental exposure. A biologist might define "interaction" in a broader mechanistic sense that interaction exists if the genetic factor and environmental exposure work on the same pathway (Clayton and McKeigue, 2001). Assessing the joint effects of genetic and environmental factors within strata defined by other variables may provide useful insight into disease etiology and help to determine effective public health intervention strategies. The article by CC is thus a major breakthrough, which emphasizes that retrospective analysis of case–control studies of gene-environment "interaction" go well beyond estimating the statistical interaction parameter β_3. However, as emphasized throughout the text, scientific and empirical validation of the *G–E* independence assumption is of utmost importance before using the proposed methods.

To conclude, we would like to highlight some of the new features of the article. In this article, we proposed a fully flexible, robust Bayesian semiparametric model for estimating not only the interaction parameter, but the main effects under conditional gene-environment independence. The method outperforms the existing methods in some instances and performs comparably in others. With genetic mutation that has unequal frequencies in case and control population, the ability to model them separately through the proposed likelihood has a natural justification. When the *G–E* independence assumption does not hold, the method performs better when compared to other contenders. The article introduces some interesting statistical techniques especially for handling the high-dimensional stratum effects on the genetic and environmental exposure distribution in a data-adaptive way. The use of the DPM model as illustrated in Result 2 in conjunction with transition from control to case distribution is a nice application of the theory on DP. Using prior biological information on the frequencies of the genetic mutation reiterates the fundamental advantage of a Bayesian paradigm. The exhaustive simulation study comparing the Bayesian methods with other frequentist methods, including the one recently proposed by CC remains an additional asset of this article. The Bayesian framework appears to be a promising route for relaxing the *G–E* independence assumption in a data-adaptive way where further work is necessary.

6. Supplementary Materials

Web Appendices and more numerical results are available under the Paper Information Link at the Biometrics website `http://www.tibs.org/biometrics`.

Acknowledgements

The first two authors contributed equally to this work. The authors thank Nilanjan Chatterjee, the editor, the associate editor, and the reviewers for many helpful comments leading to substantial improvements in the article. The research of BM was partially supported by NSA young investigator grant H98230-06-1-0033.

References

Albert, P. S., Ratnastingle, D., Tangrea, J., and Wacholder, S. (2001). Limitations of the case-only design for identifying gene-environment interaction. *American Journal of Epidemiology* **154,** 687–693.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to non-parametric problems. *The Annals of Statistics* **2,** 1152–1174.

Breslow, N. E., Robins, J. M., and Wellner, J. A. (2000). On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli* **6,** 447–455.

Chatterjee, N. and Carroll, R. (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* **92,** 399–418.

Clayton, D. and McKeigue, P. M. (2001). Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet* **358,** 1356–1360.

Couch, F. J. DeShano, M. L., et al. (1997). BRCA1 mutations in women attending clinics that evaluate the risk of breast cancer. *The New England Journal of Medicine* **336,** 1409–1415.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90,** 577–588.

Gatto, N. M., Campbell, U. B., Rundle, A. G., and Ahsan, H. (2004). Further development of the case-only design for assessing gene-environment interaction: Evaluation of and adjustment for bias. *International Journal Epidemiology* **33**(5), 1014–1024.

MacEachern, S. N. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics* **7,** 223–238.

Modan, M. D. Hartge, P., et al. (2001). Parity, oral contraceptives and the risk of ovarian cancer among carriers and noncarriers of a BRCA1 or BRCA2 mutation. *The New England Journal of Medicine* **345,** 235–240.

Parmigiani, G., Berry, D. A., and Aguilar, O. (1998). Determining carrier probabilities for breast cancer susceptibility genes BRCA1 and BRCA2. *American Journal of Human Genetics* **62,** 145–158.

Piegorsch, W. W., Weinberg, C. R., and Taylor, J. A. (1994). Non-hierarchical logistic models and case-only designs for assessing suspectibility in population based case-control studies. *Statistics in Medicine* **13,** 153–162.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **13,** 403–411.

Roeder, K., Carroll, R. J., and Lindsay, B. G. (1996). A nonparametric mixture approach to case-control studies with errors in variables. *Journal of the American Statistical Association* **91,** 722–732.

Satten, G. and Kupper, L. (1993). Inference about exposure-disease associations using probability-of-exposure information. *Journal of the American Statistical Association* **88,** 200–208.

Schmidt, S. and Schaid, D. J. (1999). Potential misinterpretation of the case-only study to assess gene-environment interaction. *American Journal of Epidemiology* **150,** 878–885.

Umbach, D. M. and Weinberg, C. R. (1997). Designing and analyzing case-control studies to exploit independence of genotype and exposure. *Statistics in Medicine* **16,** 1731–1743.

West, M., Müller, P., and Escobar, M. D. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. In *Aspects of Uncertainty. A Tribute to D. V. Lindley*, A. F. M. Smith and P. Freeman (eds), 363–386. New York: Wiley.

Received January 2006. Revised October 2006. Accepted August 2006.