Semiparametric Bayesian Analysis of Nutritional Epidemiology Data in the Presence of Measurement Error

Samiran Sinha,1,* Bani K. Mallick,1,*** Victor Kipnis,2,**** and Raymond J. Carroll1,*****

1Department of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.
2Biometry Research Group, Division of Cancer Prevention and Control, National Cancer Institute, Bethesda, Maryland 20892, U.S.A.
*email: sinha@stat.tamu.edu
**email: bmallick@stat.tamu.edu
***email: kipnisv@mail.nih.gov
****email: carroll@stat.tamu.edu

SUMMARY: We propose a semiparametric Bayesian method for handling measurement error in nutritional epidemiological data. Our goal is to estimate nonparametrically the form of association between a disease and exposure variable while the true values of the exposure are never observed. Motivated by nutritional epidemiological data, we consider the setting where a surrogate covariate is recorded in the primary data, and a calibration data set contains information on the surrogate variable and repeated measurements of an unbiased instrumental variable of the true exposure. We develop a flexible Bayesian method where not only is the relationship between the disease and exposure variable treated semiparametrically, but also the relationship between the surrogate and the true exposure is modeled semiparametrically. The two nonparametric functions are modeled simultaneously via B-splines. In addition, we model the distribution of the exposure variable as a Dirichlet process mixture of normal distributions, thus making its modeling essentially nonparametric and placing this work into the context of functional measurement error modeling. We apply our method to the NIH-AARP Diet and Health Study and examine its performance in a simulation study.

KEY WORDS: B-splines; Dirichlet process prior; Gibbs sampling; Measurement error; Metropolis–Hastings algorithm; Partly linear model.

1. Introduction

We consider the logistic regression of disease status $Y$ on covariates $(X, Z)$, where $X$ is unobservable and can only be measured with an error, and the goal is to understand its effect on $Y$. In doing so, we assume that the logit of the disease probability is a linear function of $Z$ and possibly a nonlinear function of $X$ but its exact form is unknown, thus yielding a partially linear logistic model. Although we work with a logistic model, the method is applicable to any distributional model of partially linear form.

The commonly used classical measurement error model assumes that instead of observing $X$, for every subject one observes, possibly after data transformation, a surrogate $W$ that is unbiased for $X$ involving purely random error with a constant variance (Carroll et al., 2006). However, in large nutritional epidemiological studies, dietary intake for all individuals is usually measured by a food frequency questionnaire (FFQ), which we denote here by $Q$. It has become well appreciated in the literature that FFQs have substantial measurement error, both random and systematic, therefore violating the classical measurement error assumptions. Our motivating example is the NIH-AARP Diet and Health Study (http://dietandhealth.cancer.gov/), the details of which can be found in Schatzkin et al. (2001). It is important to note that any case–control study for finding diet-cancer association is subject to differential recall bias between cases and controls. Secondly, a homogeneous population with a narrow range of fat intake usually fails to find any association between fat intake and breast cancer. To circumvent these issues, in the NIH-AARP Diet and Health Study a large and diverse population was targeted where diet was assessed prior to diagnosis. In this study, the initial cohort of 617,119 men and women who responded to an FFQ in 1995–1996 has been followed for the evaluation of possible diet-cancer associations. The latest database till the year 2003 contained information on 27 types of cancer, and breast cancer is one of them, which is considered as the response variable in this article. To adjust for FFQ measurement error in estimated relationships, the NIH-AARP study includes a so-called calibration substudy with approximately 2000 men and women who, in addition to the FFQ, were administered two nonconsecutive 24-hour dietary recalls (24hr) denoted by $W$. Following the study design, we assume $W$ follows the classical measurement error model. Thus the study design consists of the binary response $Y$ (occurrence or nonoccurrence of invasive breast cancer during follow up), covariates without error $Z$, true unobserved main exposure $X$, observed exposure $Q$ that measures $X$ with both bias and random error. In addition, a calibration substudy includes both $Q$ and $W$ along with $Z$. 
To handle the substantial random and systematic measurement error in the FFQ in a semiparametric fashion, we assume that conditional on \((X, Z)\), the surrogate variable \(Q\) has a partially linear model, where the effect of \(Z\) is linear and the effect of \(X\) on \(Q\) is still unknown, but assumed to be a smooth and monotone function. In fact for our data example, we found that a linear regression between \(Q\) and the average of \(W\) is not sufficient to explain the nature of association between these two variables, which is an indication of a possible nonlinear association between \(Q\) and \(X\). Although a simple logistic regression of the occurrence of invasive breast cancer \((Y)\) on the percentage of nonalcohol energy from total fat measured via FFQ \((Q)\) revealed a significant association, we want to measure how the risk changes with the true value of the percentage of nonalcohol energy from total fat \((X)\) taking into account that \(X\) is unobserved and \(Q\) contains substantial measurement error. Therefore, we will model the effect of \(X\) on the logit of the success probability of \(Y\) via splines, and the effect of \(X\) on \(Q\) via monotone splines. Since we will develop a likelihood-based inference for our models, we need to specify the distribution of the latent variable \(X\). Since the histogram of the average of \(W\) obtained from the calibration data and that of \(Q\) from the cohort study (see Web Figure 1) do not show strong evidence for the normal distribution assumption for \(X\), we model the distribution of \(X\) nonparametrically via an infinite mixture of normal distributions. In fact the pattern of food intake is likely to vary as the cohort members have a diverse background and are from six different U.S. cities and two metropolitan areas with large minority populations.

Before describing the novelty of our disease model and the calibration model and our approach to handling them, we first point out that the regression calibration \((RC)\) approach is not applicable in our context. In parametric linear logistic model for the disease probability, i.e.,

\[
\Pr(Y = 1 | X, Z) = H\{X \xi + Z^T \zeta_{\text{ink}}\},
\]

where \(H(u) = 1/(1 + \exp(-u))\) is the logistic distribution function, it has become common to apply RC adjustment for measurement error, where one regresses \(W\) on \(Q\) in the calibration substudy, then replaces \(X\) in (1) by the predictions from this regression. However, this method is well known to be undesirable in semiparametric models such as the partially linear logistic model, and indeed in our simulations it performs quite poorly. The reason is that RC assumes that in the induced observed data model \(\Pr(Y = 1 | W, Z)\) the effect of \(W\) is conferred only through \(E(X | W, Z)\). However, if the effect of \(X\) in the logit of \(\Pr(Y = 1 | X, Z)\) is nonlinear, the actual observed data model may be quite far from the assumed induced observed data model. An example of this is known in ordinary nonparametric regression, where if \(E(Y | X) = \sin(2X), X, U \sim \text{Normal}(0, 1)\), and \(W = X + U\), then the approximation \(E(Y | W) \approx \sin(2E(X | W))\) is systematically biased and out of phase with the true regression function.

Carroll and Hall (1988), Fan and Truong (1993), and Delaigle and Hall (2008) considered deconvolution method to deal with the classical measurement error in \(X\) for a nonparametric regression of \(Y\) on \(X\). However, such methods did not consider systematically biased surrogate such as FFQ. Also, the methods were not designed to handle the partially linear logistic model.

There are some related Bayesian methodologies, but none of them handle the generality of the problem we confront. Berry, Carroll, and Ruppert (2002) used smoothing splines and regression splines in the classical measurement error problem to a linear model set up, but not to the important case of binary data. Carroll et al. (2004) used Bayesian spline-based regression when an instrument is available for all study participants. In addition, both papers assumed that the unknown \(X\) is normally distributed. Mallick and Gelfand (1996) considered covariate measurement error in the generalized linear model with an unknown link function, where the distribution of \(X\) was modeled via a multivariate normal distribution. Müller and Roeder (1997) considered the multivariate normal mixture of the Dirichlet process (DP) prior for handling covariate measurement error in case–control studies. Bayesian nonparametric regression approaches without measurement error in the covariate for binary data have been considered by Wood and Kohn (1998), Wood et al. (2002), and Holmes and Mallick (2003), among others. In summary, all the above-mentioned papers considered either (a) a nonparametric regression without any measurement error or (b) measurement error in covariates while the regression model is parametrically specified. None of these papers allowed for the partially linear model for the response variable \(Y\), nor did they even begin to address partially linear calibration model to handle systematic bias in FFQ. Johnson et al. (2007) considered a problem similar to ours, although the data structures are slightly different, since in their example the FFQ is replicated. The important differences are that in place of the semiparametric risk model for \(Y\) given \((X, Z)\), they used the parametric model (1), and in place of the partially linear calibration model for \(Q\) given \((X, Z)\), they used a linear model. The important similarity is that they, like us, used a mixture of the DP model for the distribution of the latent variable.

In summary, the three novel features of our approach are the following. First, we consider a semiparametric logistic model with a nonparametric component subject to measurement error. Second, we allow for the fact that in actual epidemiological practice, the vast majority of the data only have a systematically biased measure of the true risk covariate, and we handle this feature via a semiparametric model with a monotone nonparametric component, the monotonicity being natural in the scientific context. Although the idea of systematic and random bias in \(Q\) has appeared in many papers in nutritional epidemiology (Kipnis et al., 2001, 2003), our consideration of a semiparametric model for the systematic component of the bias is new. Third, we model the distribution of the unobserved covariate nonparametrically via the DP mixture of normal distributions. While the use of such models in general is not new (Johnson et al., 2007), using the idea in a semiparametric context has not previously been investigated.
estimation of two nonparametric functions, where one is dependent on the other, is not an easy task, especially for a binary regression when the covariate distribution is unknown. The simulation study and data analysis show the importance of such flexible models.

An outline of the article is as follows. The model and assumptions are described in Section 2, while the method of estimation is described in Section 3. Section 4 contains the data analysis of the NIH-AARP Study. A simulation study is described in Section 5. Section 6 contains concluding remarks.

2. Model and Assumptions

2.1 Outline

In this section, we give the model structure for the observed data \((Y, Q, Z, W)\) given the latent long-term diet \(X\) (Section 2.2), the model structure for \(X\) (Section 2.3), and the models for the semiparametric functions that are in the model (Section 2.4).

2.2 Model Structure for Observed Data

Let \(Y\) be the binary response variable, \(X\) the covariate of interest, \(Q\) the surrogate variable for \(X\), and \(Z\) a vector of error-free covariates. The primary data consist of \((Y_i, Q_i, Z_i, W_i)\) for \(i = 1, \ldots, n\). In order to obtain information about the relationship of \(Q_i\) and \(X_i\), we assume that there is an external calibration data where we observe \(Q, Z\), and measured data of some unbiased surrogate variable \(W\). Therefore, the calibration data are \((Q_i, Z_i, W_{ij})\) for \(j = 1, \ldots, J\), and \(i = n + 1, \ldots, N\), where \(N = n + m\). The basic risk model is

\[
\Pr(Y = 1|X, Z) = H\{\theta_{\text{risk}}(X) + Z^T \zeta_{\text{risk}}\},
\]

where \(\theta_{\text{risk}}(\cdot)\) is an unknown function, and \(Q\) is related to \((X, Z)\) via

\[
Q = \theta_{\text{cal}}(X) + Z^T \zeta_{\text{cal}} + U_Q,
\]

where \(\theta_{\text{cal}}(\cdot)\) is an unknown smooth monotone function. We further assume that the within-person random measurement error \(U_Q\) is nondifferential, and we make the standard assumption that \(U_Q\) follows a normal distribution with mean zero and variance \(\sigma_Q^2\). Let \(\Delta_i\) be an indicator variable corresponding to subject \(i\), which takes on value 1 if the \(i\)th subject belongs to the calibration study and 0 otherwise. Therefore, when \(\Delta_i = 1\) we observe \((Q_i, Z_i, W_{ij}, j = 1, \ldots, J)\), and when \(\Delta_i = 0\) we observe \((Y_i, Q_i, Z_i)\) for the \(i\)th subject. For identifiability, we require that the distribution of \(Q\) given \((X, Z)\) is the same in both primary and calibration data, for example the latter is a randomly chosen subset of the primary data.

For the unbiased surrogate variable, there are \(j = 1, \ldots, J\) replicates and we assume \(W_{ij} = X_i + U_{W_{ij}}\). Furthermore, we assume that conditional on \(X, U_{W_{ij}}\) are independent and identically distributed Normal(0, \(\sigma_W^2\)), and conditional on \((X, Z), U_Q\) is assumed to be independent of \(U_{W_{ij}}\). Note that the assumptions of normality and independence for the errors \(U_Q\) and \(U_{W_{ij}}\) can be relaxed to any bivariate parametric model.

2.3 Distribution of the Unobserved Covariate \(X\)

A likelihood-based approach requires a distribution for the unobserved covariate \(X\). Since the misspecification of this distribution may result in biased estimates of the quantities of interest, we model the distribution in a flexible semiparametric fashion. We assume that conditional on \(Z, \mu, \sigma^2\), and \(\tau\), \(X \sim \text{Normal}(\mu + Z^T \zeta, \sigma^2)\), and assume that a priori the parameters \((\mu, \sigma^2)\) come from a distribution \(G\), that means, \(\phi_i = (\mu, \sigma^2) \sim G\), and a priori \(G \sim DP(\alpha_0 G_0)\), where \(G_0\) is the base probability measure defined on \((\chi, \mathcal{B})\). Here \(\chi = (-\infty, \infty) \times (0, \infty)\) and \(\mathcal{B}\) is the \(\sigma\)-algebra generated by \(\chi\). Under \(G_0, [\mu, \sigma^2, \tau] \sim \text{Normal}(0, \tau^2), \sigma^2 \sim IG(a_2, b_2)\), and the hyperparameter \(\tau \sim IG(a_1, b_1)\), where \(IG(a_1, b_1)\) denotes the inverse-gamma distribution with mean \(1/(b_1(a_1 - 1))\). Since for any set \(\lambda(A)\) has mean \(G_0(A)(1 - G_0(A))/(\alpha_0 + 1)\), the parameter \(\alpha_0\) plays an important role in determining the concentration of the random process \(G\) around the base probability measure \(G_0\). The stick-breaking representation of the DP indicates that \(G\) is almost surely a discrete probability measure with infinite many mass points, which in turn implies that the distribution of \(X\) is an infinite mixture of normal distributions, and the parameters of the component distribution come from the base probability measure \(G_0\). In addition, the mixing probabilities \(\pi_i\)’s are obtained via \(\pi_k = \gamma_k \prod_{j=1}^{k-1} (1 - \gamma_j)\), where \(\gamma_k = \text{Beta}(1, \alpha_0)\), and \(\sum_{k=1}^{\infty} \pi_k = 1\).

2.4 Semiparametric Models for \(\theta_{\text{risk}}(\cdot)\) and \(\theta_{\text{cal}}(\cdot)\)

We wish to estimate \(\theta_{\text{risk}}(X)\) in the interval \([a, b]\) and for this purpose we use B-splines. Let \(B^T_{\text{risk}}(X) = \{B_{\text{risk}}(X), \ldots, B_{\text{risk}}(X)\}\) be a set of cubic spline basis functions for \(k_{\text{risk}}\) fixed knot points. With \(p = k_{\text{risk}} + 4\), our model is

\[
\theta_{\text{risk}}(X) = \sum_{j=1}^{p} B_j \theta_{\text{risk}}(X) \beta_j.
\]

Also based on \(k_{\text{cal}}\) fixed knot points, with \(q = k_{\text{cal}} + 4\), we express \(\theta_{\text{cal}}(\cdot)\) as

\[
\theta_{\text{cal}}(X) = \sum_{j=1}^{q} B_{\text{cal}}(X) \alpha_j,
\]

where \(B^T_{\text{cal}}(X) = \{B_{\text{cal}}(X), \ldots, B_{\text{cal}}(X)\}\) is also a set of cubic spline basis. The use of cubic splines implies that the nonparametric functions have continuous second derivatives. The chosen knot points for the two nonparametric functions could be the same, and consequently the spline basis functions would be the same. In order to estimate \(\theta_{\text{risk}}(\cdot)\) properly, \(\theta_{\text{cal}}(\cdot)\) should be a monotonic nondecreasing function of its argument, which is a natural assumption regarding the conditional mean of \(Q\) given \(X\). Because the B-spline basis functions are nonnegative, under the monotonicity constraint for \(\theta_{\text{cal}}(\cdot)\), the first derivative \(\theta'_{\text{cal}}(X)\) is nonnegative if the coefficients \(\alpha_i\)’s are nondecreasing, i.e., \(\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_q\). This property of B-splines has been previously used by Leit- enstoffer and Tutz (2007) among many others. Also we chose a moderately large number of knot points that will make the inference insensitive to the location of knots (Ruppert, 2002). Next, we describe the method of estimation within the Bayesian paradigm.

3. Method of Estimation

Let \(\beta = (\beta_1, \ldots, \beta_J), \alpha^T = (\alpha_1, \ldots, \alpha_q), \phi = (\phi_1, \ldots, \phi_{n+m})\), and define \(\theta = (\alpha^T, \beta^T, \phi_{\text{risk}}^T, \zeta_{\text{cal}}^T, \zeta_{\text{risk}}^T, \sigma^2, \sigma^2_0)\). The observed
data likelihood function is
\[ L_0 = \int \prod_{i=1}^{n+m} f^{i-\Delta} (Y_i, Q_i | Z_i, \theta, \phi) \times f^{\Delta} (Q_i, W_{ij}, j = 1, \ldots, J | Z_i, \theta, \phi) dG(\phi), \]
where
\[ f(Y_i, Q_i | Z_i, \theta, \phi) = \int f(Y_i, Q_i, X_i | Z_i, \theta, \phi) dX_i \]
\[ = \int p^{Y_i} (Y_i = 1 | X_i, Z_i, \theta) p^{Y_i - Y_i} \times (Y_i = 0 | X_i, Z_i, \theta) f(Q_i | X_i, Z_i, \theta) \times f(X_i | Z_i, \theta, \phi) dX_i \]
\[ = \int \frac{\exp[Y_i (\theta_{\text{ink}}(X_i) + Z_i^T \zeta_{\text{ink}}))]}{1 + \exp(\theta_{\text{ink}}(X_i) + Z_i^T \zeta_{\text{ink}})) (\sigma^2_\alpha)^{1/2}} \times \exp \left\{ -\frac{1}{2\sigma^2_\alpha} (Q_i - \theta_{\text{cal}}(X_i) - Z_i^T \zeta_{\text{cal}}) \right\} \]
\[ \times \frac{\exp \left\{ -\frac{1}{2\sigma^2_\zeta} (X_i - \mu - Z_i^T \zeta_{\text{ink}}) \right\}}{dX_i}, \]
\[ f(Q_i, W_{ij}, \ldots, W_{ij} | Z_i, \theta, \phi) \]
\[ = \int f(Q_i, W_{ij}, \ldots, W_{ij} | X_i, Z_i, \theta, \phi) dX_i \]
\[ = \int f(Q_i, X_i, Z_i, \theta) f(W_{ij}, \ldots, W_{ij} | X_i, Z_i, \theta) \times f(X_i | Z_i, \theta, \phi) dX_i \]
\[ = \frac{1}{(\sigma^2_\alpha)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2_\alpha} (Q_i - \theta_{\text{cal}}(X_i) - Z_i^T \zeta_{\text{cal}}) \right\} \]
\[ \times \frac{1}{(\sigma^2_\zeta)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2_\zeta} (X_i - \mu - Z_i^T \zeta_{\text{ink}}) \right\} \]
\[ \times \frac{1}{(\sigma^2_\zeta)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2_\zeta} (X_i - \mu - Z_i^T \zeta_{\text{ink}}) \right\} dX_i. \]

Before describing prior specification and posterior inference, we would like to mention that one could adopt a complete nonparametric method by modeling the joint distribution of \( X, Q, \) and \( Z \) for \( Y = 0 \) and \( Y = 1 \) using a dependent DP prior. However, that method would involve multidimensional DP integrals, which is quite challenging in terms of computation. Therefore, to reduce the dimensionality and computational complexity of the problem we model the distribution of \( X \) conditional on \( Z \) where \( Z \) is always observed without any error.

3.1 Specification of Priors
In order to avoid oversmoothing due to large number of knot points in the spline models, following Ruppert, Wand, and Carroll (2003), we use the following prior distribution for \( \alpha \) and \( \beta \), \( \pi(\alpha | \delta_\alpha) \propto \exp(-\delta^2/2) \sum_{j=1}^{m} (\Delta^2 \alpha_j)^2 - (\delta_j/2V) \alpha_j^2 \alpha_j \), and \( \pi(\beta | \delta_\beta) \propto \exp(-\delta^2/2) \sum_{j=1}^{m} (\Delta^2 \beta_j)^2 - (\delta_j/2V) \beta_j^2 \beta_j \), where \( \delta_\alpha \) and \( \delta_\beta \) are the two penalty parameters corresponding to \( \alpha \) and \( \beta \). Note that the priors are proper for any choice of \( V > 0 \), and we set \( V = 10^6 \).

Also, \( \Delta_{ij} = \alpha_j - \alpha_{j-1} \), and \( \Delta^2 \alpha_j = \alpha_j - 2\alpha_{j-1} + \alpha_{j-2} \). We assume \( \delta_\alpha \sim \text{Gamma}(a_{\alpha}, b_{\alpha}) \) and \( \delta_\beta \sim \text{Gamma}(a_{\beta}, b_{\beta}) \). For \( \sigma^2_\alpha, \sigma^2_\beta \), we use \( IG(\sigma^2_\alpha, b_{\sigma^2_\alpha}) \) and \( IG(\sigma^2_\beta, b_{\sigma^2_\beta}) \) priors respectively. For each component of \( \zeta_{\text{ink}}, \zeta_{\text{cal}}, \zeta_{\text{kal}} \), we use a Normal(0, \( \sigma^2_\zeta \)) prior.

As mentioned earlier, for \( \mu, \sigma^2 \) \( \sim \text{G} \) and a priori \( G \sim DP(\alpha_0 G_0) \). The DP prior involves two parameters \( G_0 \) and \( \alpha_0 \). Note that the tuning parameter of the DP prior \( \alpha_0 \) has a dual role. On the one hand it determines how concentrated \( G \) is around \( G_0 \), and on the other hand it determines the number of clusters, that is, the number of distinct mass points of \( G \). Hence, the choice of \( \alpha_0 \) is an important one. Following Escobar and West (1995) one may put a prior on \( \alpha_0 \). However, we will estimate \( \alpha_0 \) empirically following an existing characterization of its maximum-likelihood estimator (McAuliffe, Blei, and Jordan, 2006). In summary, for our proposed method one needs to specify the following quantities: \( V, a_{\alpha}, b_{\alpha}, a_{\beta}, b_{\beta}, a_V, b_V, \sigma^2_\alpha, \) and \( \alpha_0 \).

3.2 Posterior Inference
Our primary goal is to estimate \( \zeta_{\text{ink}}, \zeta_{\text{cal}}, \beta, \alpha, \sigma^2_\alpha, \) and \( \sigma^2_\beta \). Posterior means are used as the parameter estimates. As we can see the posterior means cannot be calculated analytically, we propose an efficient posterior sampling algorithm for simulating random numbers from the posterior distributions, and the details are given in the Web Appendix. At each Markov chain Monte Carlo (MCMC) iteration, we estimate \( \alpha_0 \) empirically by solving the following equation:
\[ \sum_{i=1}^{N} \frac{\alpha_0}{\alpha_0 + r - 1} = E(k | X_i, i = 1, \ldots, N, \alpha_0), \]
where the right-hand side of (4) is obtained by taking average of \( k \) the number of distinct \( \phi_i \)'s, of the last 30 Gibbs sampling iterations, i.e., we take \( E(k | X_i, i = 1, \ldots, N, \alpha_0) \approx \sum_{i=1}^{N} k_i/30 \).

Note, as \( \phi_1, \ldots, \phi_N \) are independent and identically distributed observations from \( G \), and \( G \) has a \( DP(\alpha_0 G_0) \) prior, the posterior distribution of \( G \) given \( \phi_1, \ldots, \phi_N \) is again a \( DP(\alpha^*_0 G^*_0) \), where \( \alpha^*_0 = \alpha_0 + N \) and \( G^*_0 = (\alpha_0 + N)^{-1} (\alpha_0 G_0 + \sum_{i=1}^{N} \delta_\phi) \). Thus, for future realization \( \phi_{N+1} \), \( \text{pr}(\phi_{N+1} \in A | \phi_1, \ldots, \phi_N) = E(G(A) | G \sim DP(\alpha^*_0 G^*_0)) \) is obtained. Therefore, the predictive distribution of \( \phi_{N+1} | \phi_1, \ldots, \phi_N \) is \( G^*_0 \).

4. Application to the NIH-AARP Diet and Health Data
4.1 Background
In this illustration of our methodology, we consider the association between percentage of nonalcohol energy from total fat and the risk of invasive breast cancer (\( Y \)) in the NIH-AARP cohort. Here are some details about this study.

During 1995–1996, 3.5 million baseline questionnaires were mailed to current members of the AARP (formally the American Association of Retired Persons), aged 50–71 years, and 567,169 members completed them satisfactorily (Schatzkin et al., 2001). Of these, the investigators excluded an additional withdrawal (\( n = 1 \)), duplicate records (\( n = 179 \)), subjects who moved out of the eight states included in the study before returning the baseline questionnaire (\( n = 321 \)), or were
found to have died before study entry \((n = 261)\). After these exclusions, the NIH-AARP baseline cohort included 566,407 persons that we considered in our analysis. Thiebaut et al. (2007) analyzed this cohort with nonalcohol energy from total fat as well as fatty acid as the main exposure variables, in addition to several potential confounders. In our illustration, we use nonalcohol energy from total fat as the main exposure and body mass index (BMI) measured in the metric unit kg/m² as the error-free covariate \(Z\). We considered only postmenopausal women who did not report a personal history of any cancer, did not develop breast cancer in situ during the follow-up, and whose BMI values were not missing. Of the remaining 144,366 subjects, we excluded those women who reported extreme values (i.e., more than two interquartile range above the 75th percentile or below the 25th percentile on the logarithmic scale) for total fat, energy, or percentage of nonalcohol energy from total fat, thus leaving us with 142,364 subjects, of whom 2,724 women developed invasive breast cancer during the follow-up period. The average follow-up time for these subjects was 4.52 years. Based on the daily alcohol consumption in grams, total fat intake in grams, and total energy intake in Kcals, the percentage of nonalcohol energy from total fat was calculated by using the following formula:

\[
\text{Percentage of nonalcohol energy from total fat} = \frac{\text{energy from total fat}}{\text{nonalcohol energy}} \times 100,
\]

where energy from total fat was obtained by multiplying total daily consumption of fat by 9 Kcals, and nonalcohol energy was calculated by subtracting energy from alcohol from total energy. Energy from alcohol was computed by multiplying alcohol consumption measured in grams by 7 Kcals. The percentage of nonalcohol energy from total fat measured from the FFQ was treated as the surrogate measurement \(Q\) of \(X\).

The NIH-AARP calibration substudy was designed to calibrate the FFQ used in the main study with two nonconsecutive 24-hour recall telephone interviews as the unbiased surrogate measure. While there is a concern that 24-hour recalls may not be actually unbiased for true intake, these are the data we have and their use can be justified as a so-called auxiliary data. The result showed that the estimated \(\theta_{\text{cal}}(\cdot)\) function along with a 95% credible interval. This figure also shows the scatter plot of \((Q - \hat{\phi}_0 Z, W)\) from the calibration data. Both figures clearly indicate that there is some sort of nonlinearity in both \(\theta_{\text{cal}}(\cdot)\) and \(\theta_{\text{cal}}(\cdot)\). Table 1 presents the posterior mean, posterior standard deviation, and 95% credible intervals for \(\zeta_{\text{cal}}, \zeta_{\text{cal}}, \sigma^2, \text{and} \sigma^2\).

Using the DP prior, we found that the average number of mixing components is 2.11 with a posterior standard deviation of 0.33, indicating that the true distribution of \(X\) given \(Z\) is not a normal distribution but a mixture of normals that is also evident from the histogram of an MCMC sample of \(X\) values (Figure 3). Figure 4 shows the histogram of \(\alpha_0\), the tuning parameter of the DP prior. Web Figure 2 shows a plot of the number of clusters of the DP process prior in the MCMC samples. Following a referee’s comment we also have added a summary of cluster sizes using boxplot based on the MCMC sample (Web Figure 3) when the number of clusters are 2, 3, or 4. Web Figure 4 shows an estimated \(G^*_M = E(G|X)\) data based on the MCMC sample of \(\phi_{N+1}\).

The null hypothesis of interest is that there is no risk of fat on breast cancer: under this null hypothesis \(\theta_{\text{cal}}\) should be constant and thus it is equivalent to test \(\beta_i = \cdots = \beta_p\) as \(\sum_{j=1}^p B_i \beta_j\) given \(X\).

Suppose we consider \(M\) MCMC samples taking every 300th observation after the burn-in period. Then the test statistic for testing the null hypothesis is \(F = \left[\{(M - p + 1)/{(M - 1)(p - 1)}\}T^2\right]^{1/2}\), where \(T^2 = M D^T S_p^{-1} D, D^T = (D_{-1}, \ldots, D_{p-1})\), \(D_k = \sum_{j=1}^M D_{kj}/M, k = 1, \ldots, (p - 1)\). Here \(D_{kj} = \beta_{kj} - \beta_{\text{cal}, kj}\) is the \(j\) th value of \(\beta_k\) of \(M\) observations, \(S_0 = \sum_{j=1}^M (D_k - \bar{D}) (D_k - \bar{D})^T/\{(M - 1)\) with \(\bar{D}^T = (D_{-1}, \ldots, D_{p-1})\). Under the null hypothesis, for sufficiently large \(M\), the statistic follows an \(F_{(p - 1), (M ^{p + 1})}\) distribution. Therefore, we use our semiparametric Bayesian (SPB) approach. We experimented with 15, 11, 9, and 6 knot points, and the results were fairly insensitive to the number of knot points. Therefore, here we discuss the results for 6 knot points: (2.9, 3.0, 3.2, 3.4, 3.6, 3.80), and we used cubic B-splines, so there were \(p = 10\) spline basis functions. We used the same knot points for \(\theta_{\text{cal}}(\cdot)\) and \(\theta_{\text{cal}}(\cdot)\). For \(\delta_0\), and \(\delta_3\) we used Gamma(0.001, 1000) priors. We used \(IG(25, 0.25)\) prior for \(\sigma_0^2\) and \(\sigma_0^2\), and Normal(0,10²) prior for \(\zeta_{\text{cal}}, \zeta_{\text{cal}}, \zeta_{\text{cal}}, \sigma^2, \text{and} \sigma^2\). For the distribution \(G_0\) of the DP prior, the priors chosen were \(\tau \sim IG(2.1, 2.0)\) and \(\sigma^2 \sim IG(2.1, 0.91)\). The prior parameters were chosen in such a way that the priors cover a wide range of values for the parameter of interest. We ran the Gibbs sampling for 20,000 iterations and we discarded the first 5000 iterations as burn-in samples. The choice of the initial parameters is an important step for proper and timely convergence of the MCMC method, and in the next section we discuss this issue in detail. The solid lines of Figure 1 show the estimated \(\theta_{\text{cal}}(\cdot)\) along with a 95% credible interval. Overall, the method shows that the risk of having breast cancer increases with the logarithm of the percentage of nonalcohol energy from total fat. Figure 2 shows the estimated \(\theta_{\text{cal}}(\cdot)\) function along with 95% credible interval. This figure also shows the scatter plot of \((Q - \tilde{\phi}_0 Z, W)\) from the calibration data. Both figures clearly indicate that there is some sort of nonlinearity in both \(\theta_{\text{cal}}(\cdot)\) and \(\theta_{\text{cal}}(\cdot)\).
Figure 1. Plot of the estimated $\theta_{\text{val}}(X)$ along with a pointwise 95% credible interval using the semiparametric Bayesian approach (SPB, solid lines) and the parametric Bayesian approach (PRB, dashed lines) approach.

Figure 2. This figure shows the estimated $\theta_{\text{cal}}(X)$ using the SPB method, along with a 95% pointwise credible interval. This figure also contain the scatter plot of $(Q - \hat{C}_{\text{cal}}Z, W)$ of the calibration data, and a line with zero intercept and slope one.
Table 1
Results of the NIH-AARP data analysis by the semiparametric Bayesian (SPB) and parametric Bayesian (PRB) methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameter</th>
<th>Posterior mean</th>
<th>Posterior standard deviation</th>
<th>95% credible interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ζ_risk</td>
<td>0.0516</td>
<td>0.0173</td>
<td>(0.0205, 0.0830)</td>
</tr>
<tr>
<td></td>
<td>ζ_cal</td>
<td>0.0008</td>
<td>0.0039</td>
<td>(−0.0046, 0.0096)</td>
</tr>
<tr>
<td></td>
<td>σ²Q</td>
<td>0.0260</td>
<td>0.0025</td>
<td>(0.0229, 0.0295)</td>
</tr>
<tr>
<td></td>
<td>ζ_0</td>
<td>0.0697</td>
<td>0.0038</td>
<td>(0.0598, 0.0767)</td>
</tr>
<tr>
<td>SPB</td>
<td>σ²W</td>
<td>0.1104</td>
<td>0.0145</td>
<td>(0.0889, 0.1368)</td>
</tr>
<tr>
<td></td>
<td>ζ_risk</td>
<td>0.0397</td>
<td>0.0234</td>
<td>(−0.0134, 0.0754)</td>
</tr>
<tr>
<td></td>
<td>ψ_0</td>
<td>1.3960</td>
<td>0.1365</td>
<td>(1.0577, 1.6327)</td>
</tr>
<tr>
<td></td>
<td>ψ_1</td>
<td>0.5915</td>
<td>0.0463</td>
<td>(0.5137, 0.7036)</td>
</tr>
<tr>
<td></td>
<td>ζ_cal</td>
<td>0.0268</td>
<td>0.0052</td>
<td>(0.0134, 0.0327)</td>
</tr>
<tr>
<td></td>
<td>σ²Q</td>
<td>0.0496</td>
<td>0.0019</td>
<td>(0.0458, 0.0533)</td>
</tr>
<tr>
<td></td>
<td>γ_0</td>
<td>3.1550</td>
<td>0.0562</td>
<td>(3.0576, 3.2253)</td>
</tr>
<tr>
<td></td>
<td>σ²X</td>
<td>0.0839</td>
<td>0.0199</td>
<td>(0.0563, 0.1187)</td>
</tr>
<tr>
<td></td>
<td>σ²W</td>
<td>0.0528</td>
<td>0.0065</td>
<td>(0.0370, 0.0637)</td>
</tr>
<tr>
<td>PRB</td>
<td>σ²W</td>
<td>0.0916</td>
<td>0.0043</td>
<td>(0.0836, 0.1065)</td>
</tr>
</tbody>
</table>

distribution. The observed value of the test statistic was 14.78 with a p-value 2.01 × 10⁻¹⁰. Thus, we rejected the null hypothesis at 1% level of significance, and concluded that θ_risk was not a constant function. Before conducting this test we made sure that there is no significant correlation between D_kj and D_kj+1 for any k. Note that if X changes from 3.0 to 3.1, according to our SPB method, θ_risk(3.0) = −4.226 and θ_risk(2.9) = −4.223, and thus the risk increases by 0.35%. In contrast, θ_risk(3.4) = −4.113, and θ_risk(3.3) = −4.162 thus the risk increases by 5.03% for changing X from 3.3 to 3.4, which signifies that the rate of change of the relative risk is not constant in our method. In the SPB method, the estimated odds ratio for BMI is 1.057 with a 95% credible interval (1.021, 1.087) for an increase of 10 kg/m² BMI. In addition, following a referee’s comment we fitted model (1) by replacing X by ˆX = ˆη_0 + ˆη_1Q + ˆη_3Z, where ˆη_0, ˆη_1, ˆη_3 are the MLE of the simple linear regression of the average of W on Q and Z based on the substudy. The estimated odds ratio is 1.36 for one unit change in X, with a p-value = 0.03, or more specifically the risk increases by 3.04% for 0.1 unit increase in X for all values of X.

4.3 Parametric Calibration Model
One of the unique features of our approach is the semiparametric model for the calibration function given in (3), where θ_cal(·) is an unknown smooth monotone function. In addition, we model the distribution of X in a semiparametric manner. Here, we compare the results to those that assume this calibration function is linear in the unmeasured X, and that the unmeasured X is normally distributed given Z. With a slight abuse of terminology, we will refer to this method as the Parametric Bayes (PRB) approach. In the

Figure 3. Histogram of X obtained from the MCMC sample using our SPB method along with the density plot of a normal distribution with the same mean and variance as of X.
Figure 4. An empirical distribution of $\alpha_0$ of the Dirichlet process prior that is determined according to equation (4) in each MCMC sample of the SPB method.

PRB method we modeled the surrogate variable as $Q_i = \psi_0 + \psi_1 X_i + Z_i \zeta_{cal} + U_{Q_i}, U_{Q_i} \sim \text{Normal}(0, \sigma^2_Q)$ and assumed that $X_i \sim \text{Normal}(\mu_0 + Z_i \zeta_x, \sigma^2_X)$. All other components and assumptions remained the same as the SPB approach. The parameters were estimated by MCMC. For this method, we used the same prior that we had used in the SPB method. For $\psi_0, \psi_1$, and $\mu_0$, we used $\text{Normal}(0, 10^2)$ priors, and used $IG(2, 1)$ prior for $\sigma^2_X$. The posterior mean, standard deviation, and 95% credible intervals for the parameters are given in Table 1. It is obvious that the estimate of $\sigma^2_Q$ is significantly larger in the PRB approach compared to the SPB method, which is an indicator that possibly $E(Q | X, Z)$ is better explained by the partly linear function $\theta_{cal}(X) + Z \zeta_{cal}$ than a linear function of $X$ and $Z$. The dashed lines in Figure 1 show the estimated $\theta_{lin}$ for the PRB approach along with a 95% credible interval. For this method, using the above F-test, we again concluded that $\theta_{lin}$ is not a constant function in the interval of interest.

4.4 Model Checking

It is possible to compare the SPB and PRB models, i.e., to compare whether either $\theta_{cal}(\cdot)$ is nonlinear or the distribution of $X$ is nonnormal. In order to compare these two models we calculated the marginal likelihood under each model. Suppose $\Theta^T = (\Theta^T, \phi)$ is the set of all parameters, then the marginal likelihood under the SPB model is $m_{SPB} = \prod_{i=1}^{n} \int f^{1-\Delta_i}(Y_i, Q_i | Z_i; \Theta) f^{\Delta_i}(Q_i, W_{ij} | Z_i; \Theta) \pi(\Theta) d\Theta$, where the expressions for $f(Y_i, Q_i | Z_i; \Theta)$ and $f(Q_i, W_{ij} | Z_i; \Theta)$ are given in Section 3. Following Newton and Raftery (1994), we estimate the marginal likelihood via

$$\hat{m}_{SPB} = \left\{ \frac{1}{M} \sum_{m=1}^{M} f^{1-\Delta_i}(Y_i, Q_i | Z_i; \Theta^{(m)}) f^{\Delta_i}(Q_i, W_{ij} | Z_i; \Theta^{(m)}) \right\}^{-1},$$

where $\Theta^{(1)}, \ldots, \Theta^{(M)}$ are $M$ draws from the posterior distribution of $\Theta$. Note that $f(Y_i, Q_i | Z_i; \Theta^{(m)}) = \int f(Y_i, Q_i | X_i, Z_i; \Theta^{(m)}) f(X_i | Z_i, \Theta^{(m)}) dX_i$ and $f(Q_i, W_{ij} | Z_i; \Theta^{(m)}) = \int f(Q_i, W_{ij} | Z_i, \Theta^{(m)}) f(X_i | Z_i, \Theta^{(m)}) dX_i$; these integrals were calculated by a numerical quadrature. In a similar manner we estimated $m_{PRB}$ via $\hat{m}_{PRB}$. Finally, we obtain $\hat{m}_{BSP}/\hat{m}_{PRB} = 111.21$, which suggests that the SPB fits the data far better than does the PRB model.

5. Simulation

To test our methodology, we performed a simulation study that has some of the complexities of the NIH-AARP study, in terms of design and nonlinearity. First we created a
cohort of size \( n = 10,000 \) with \( Y, X, Q, W_1, \) and \( W_2 \). We considered two different cases. In case I, we simulated \( X \) from the Normal(\( \gamma_0 = 3.5, \sigma_X = 0.5 \)) distribution and in case II, \( X \) was simulated from (1/2)\( \text{Gamma}(65, 0.0455) + (1/2)\( \text{Normal}(3.8, 0.1^2) \). The second case clearly indicates that the distribution of \( X \) is not normal. For each case, after simulating \( X \), we generated the unbiased variable \( W \) as \( W_j = X_j + U_{W,j} \), where \( U_{W,j} = \text{Normal}(0, \sigma_W^j) \). We took two different values for \( \sigma_W \), 0.2 and 0.5, where the value 0.5 mimics the large measurement error situation common in nutrition epidemiology. The surrogate variable \( Q \) was generated as \( Q_i = \theta_{al}(X_i) + U_{Q,i}, U_{Q,i} = \text{Normal}(0, \sigma_Q^i) \) and we used two different values for \( \sigma_Q \), 0.1 and 0.2. We considered a situation where \( \theta_{al}(X) = 1.25 + 0.6X \), a linear function of \( X \) and another situation where \( \theta_{al}(X) = 2 + 3 \exp(4(X - 3.5))/[1 + \exp(4(X - 3.5))] \). Conditional on \( X \), the binary disease variable \( Y \) was generated with success probability \( H(\theta_{risk}(X)) \), where \( \theta_{risk}(X) = -2.8 + 1.5 \exp(10(X - 3.5))/[1 + \exp(10(X - 3.5))] \). On average the above probability resulted in 10–12% subjects with \( Y = 1 \). To create a calibration data from the same cohort, we randomly drew \( m = 1000 \) subjects and obtained the variables \( Q \) and the unbiased surrogate variables \( W_1 \) and \( W_2 \). Note that for the primary data we only considered \((Y_i, Q_i), i = 1, \ldots, 10,000 \) whereas for the calibration data we considered \((Q_j, W_{j1}, W_{j2}), j = 1, \ldots, 1000 \).

Under each scenario we generated \( R = 200 \) data sets, and each simulated data set was analyzed by the SPB approach and by the PRB approach. In both the SPB and PRB approaches, the nonparametric effect of \( X \) on the disease risk \( \theta_{risk} \) was modeled via B-splines with 9 knot points, and the knots were placed at every 10th percentile of the distribution of the average values of \( W \) starting with the 10th percentile. One should note the two main differences between the SPB and PRB approach. In the PRB approach \( X \) was always modeled as a normal random variable whereas in the SPB method \( X \) was modeled as a DP mixture of normal random variable. In addition, in the PRB approach, \( \theta_{al}(X) \) was assumed to be linear in \( X \), whereas in the SPB method we do not assume any specific form for \( \theta_{al} \), rather we simply impose the restriction that \( \theta_{al}(X) \) is a smooth and a monotone nondecreasing function of \( X \). One can of course construct many other methods to be compared with the SPB method. We have chosen to use the PRB approach to investigate the extent to which less-flexible models in terms of the distribution of the latent covariate \( X \) and the calibration model affect results.

For both the SPB and PRB methods, we used a Gamma(0.001, 10000) distribution for \( \delta \). In the analysis, we used \( IG(25, 0.25) \) prior for \( \sigma_Q^2 \) and \( \sigma_W^2 \). The PRB approach involves \( \psi_Q \) and \( \psi_W \), and we put a Normal(0, 10^2) prior on each of them. Also, in the PRB approach we used Normal(0, 10^2) prior for \( \mu_Q \) and \( IG(2.1, 0.91) \) prior for \( \sigma_Q^2 \). In the SPB method, we need to estimate \( \theta_{al}(X) \) using B-splines, which involves \( \alpha_{j1}, j = 1, \ldots, q \), and we used a Gamma(0.001, 10000) prior for \( \delta_{al} \). The previously mentioned knot points are also used for estimating \( \theta_{al}(X) \). In the DP prior, we used \( IG(2.1, 2) \) and \( IG(2.1, 0.91) \) priors for \( \tau \) and \( \sigma_Q^2 \), respectively.

For the initial values of \( X \) and \( \alpha \), we first regress \( (W_i + W_{i2})/2 \) on \( Q \) in the calibration study. Let \( \eta_{Q}, \eta_{W}, \) and \( \sigma^2_{Q} \) be the estimated intercept, slope parameter, and the residual variance of the regression obtained using the calibration study. For the true exposure variable, we set the initial value of \( X_i = \tilde{\eta}_Q + \eta_{W}Q_i, i = 1, \ldots, n \) and set \( X_i = (W_{i1} + W_{i2})/2 \) for \( i = (n + 1), \ldots, (n + m) \). The initial value of \( \alpha \) was obtained by fitting penalized nonparametric regression through \( \{Q_i, (W_{i1} + W_{i2})/2\} \) using the calibration data. We set the initial penalty parameters \( \delta = 0.005 \) and \( \delta_{al} = 0.005 \).

The performance of the methods was assessed via the following two statistics. The major concern of the article is how well we estimate the risk function \( \theta_{risk}(X) \). We considered a set of grid points from 2.8 to 4.2 with 0.01 increment. Letting \( g_j \) be the \( j \)th grid point, for the \( s \)th data set, we estimated \( \theta_{risk}(g_j) \) by \( \hat{\theta}_{risk}(g_j) \) and let \( \hat{\theta}_{risk}(g_j) = \sum_{i=1}^{200} \hat{\theta}_{risk}(g_j)/200 \). We computed integrated square bias as \( \sum_{j=1}^{141} \{(\hat{\theta}_{risk}(g_j) - \theta_{risk}(g_j))^2/141 \} \) and integrated mean square error as \( \sum_{j=1}^{141} \sum_{i=1}^{200} \{(\hat{\theta}_{risk}(g_j) - \theta_{risk}(g_j))^2/(141 \times 200) \} \). We compared the performance of SPB and PRB methods by these statistics. The major concern of the article is how well we estimate the risk function \( \theta_{risk}(X) \).

The results of the simulation study are presented in Table 2. From the results one can quickly make the following important observations. For both the SPB and PRB methods, bias increases with \( \sigma_Q \) and \( \sigma_W \). The value of integrated squared bias due to PRB method is significantly higher than the SPB method when the underlying model assumptions of the PRB approach are violated. When \( E(Q \mid X) \) is a linear function of \( X \) and \( \theta_{al}(X) \) is a smooth and a monotone nondecreasing function of \( X \). One can of course construct many other methods to be compared with the SPB method. We have chosen to use the PRB approach to investigate the extent to which less-flexible models in terms of the distribution of the latent covariate \( X \) and the calibration model affect results.

### 6. Discussion

In this article, we considered the logistic regression when an error-prone covariate \( X \) is modeled nonparametrically, while exactly measured covariates \( Z \) are modeled parametrically, the result being a partially linear model. The logistic framework could be readily generalized to any distributional model.

There are two additional important and novel components to our approach. First, we recognize that in current practice, the surrogate \( Q \) for the unobserved \( X \) is not unbiased for \( X \), and hence we have a situation where the classical measurement error model does not hold. We model the relationship of \( Q \) to \((X, Z)\) also in a partially linear fashion, while forcing the nonparametric function \( \theta_{al}(X) \) for \( X \) to be monotone.
Finally, in a small subset of individuals, a so-called calibration study, we observed replicated unbiased versions $W$ of $X$. We then model the distribution of the unobserved $X$ nonparametrically through the DP. As stated in the Introduction, this appears to be the first paper in semiparametric measurement error models where the two functions, one of them monotone, are estimated jointly nonparametrically, while the distribution of $X$ is modeled nonparametrically.

The simulation study shows that our method generally vastly outperforms the RC method. In terms of integrated mean square error our SPB method shows much better performance than the less-flexible PRB approach. The simulation study also indicates that the prices of no model assumption (naive method) and wrong model assumption (PRB method) could be significant. Instead of a single covariate with nonlinear effect, our method could also be used for the generalized additive model with multiple covariates.

One of the challenging parts of the method is the computation that requires appropriate choice of the initial values of the parameters and proposal distributions of the Metropolis–Hastings algorithm for the Gibbs sampling that are discussed in the appendix. The computation were done using Fortran and R, and the computer code is available at http://www.stat.tamu.edu/~sinha/research.html.

7. Supplementary Materials

The Web Appendix, and Web Figures 1–4 referenced in Sections 1, 3, and 4 are available under the Paper Information link at the Biometrics website http://www.biometrics.tibs.org.

Acknowledgements

The research of BKM and RJC was supported by grants from the National Cancer Institute (CA57030, CA104620) and in part by Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST).

References


Received December 2008. Revised May 2009.

Accepted May 2009.