DOI: 10.1111/j.1541-0420.2009.01309.x

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. Carroll[1,****]**

[1]Department of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.
[2]Biometry 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 $(X, \boldsymbol{Z})$, 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 $\boldsymbol{Z}$ 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 $\boldsymbol{Z}$, 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 $\boldsymbol{Z}$.

To handle the substantial random and systematic measurement error in the FFQ in a semiparametric fashion, we assume that conditional on $(X, \boldsymbol{Z})$, the surrogate variable Q has a partially linear model, where the effect of $\boldsymbol{Z}$ 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.,

$$\mathrm{pr}(Y = 1 \mid X, \boldsymbol{Z}) = H\{X\xi + \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\zeta}_{\mathrm{risk}}\}, \tag{1}$$

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 $\mathrm{pr}(Y = 1|W, \boldsymbol{Z})$ the effect of W is conferred only through $E(X|W, Z)$. However, if the effect of X in the logit of $\mathrm{pr}(Y = 1|X, \boldsymbol{Z})$ 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 $E(Y \mid X) = \sin(2X)$, $X, U \sim \mathrm{Normal}(0, 1)$, and $W = X + U$, then the approximation $E(Y \mid W) \approx \sin\{2E(X \mid W)\}$ 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 abovementioned 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 $(X, \boldsymbol{Z})$, they used the parametric model (1), and in place of the partially linear calibration model for Q given $(X, \boldsymbol{Z})$, 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. To the best of our knowledge then, this is the first work in the semiparametric measurement error field where two smooth nonparametric functions are estimated simultaneously, one in the exposure–response association, and the other in the association between the surrogate variable and the true exposure, while at the same time treating the distribution of the true exposure variable essentially nonparametrically. Moreover, the 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 $(Y, Q, \boldsymbol{Z}, W)$ 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 $\boldsymbol{Z}$ a vector of error-free covariates. The primary data consist of $(Y_i, Q_i, \boldsymbol{Z}_i)$ 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 $Q, \boldsymbol{Z}$, and repeated measurements of some unbiased surrogate variable W. Therefore, the calibration data are $(Q_i, \boldsymbol{Z}_i, W_{ij})$ for $j = 1, \ldots, J$, and $i = n+1, \ldots, N$, where $N = n + m$. The basic risk model is

$$\text{pr}(Y = 1 | X, \boldsymbol{Z}) = H\{\theta_{\text{risk}}(X) + \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\zeta}_{\text{risk}}\}, \tag{2}$$

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

$$Q = \theta_{\text{cal}}(X) + \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\zeta}_{\text{cal}} + U_Q\,, \tag{3}$$

where $\theta_{\text{cal}}(\cdot)$ is an unknown smooth monotone function. We further assume that the within-person random measurement error $U_{Q\,i}$ is nondifferential, and we make the standard assumption that $U_{Q\,i}$ follows a normal distribution with mean zero and variance σ_Q^2. Let Δ_i be an indicator variable corresponding to subject i, which takes on value 1 if the ith subject belongs to the calibration study and 0 otherwise. Therefore, when $\Delta_i = 1$ we observe $(Q_i, \boldsymbol{Z}_i, W_{ij}, j = 1, \ldots, J)$, and when $\Delta_i = 0$ we observe $(Y_i, Q_i, \boldsymbol{Z}_i)$ for the ith subject. For identifiability, we require that the distribution of Q given $(X, \boldsymbol{Z})$ 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 $(X, \boldsymbol{Z}), U_{Q\,i}$ is assumed to be independent of $U_{W\,ij}$. Note that the assumptions of normality and independence for the errors $U_{Q\,i}$ 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 $\boldsymbol{Z}_i, \mu_i$, and $\sigma_i^2, X_i \sim \text{Normal}(\mu_i + \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\zeta}_x, \sigma_i^2)$, and assume that a priori the parameters (μ_i, σ_i^2) come from a distribution G, that means, $\boldsymbol{\phi}_i^{\mathrm{T}} = (\mu_i, \sigma_i^2) \sim G$, 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 σ-algebra generated by χ. Under $G_0, [\mu \mid \sigma^2, \tau] \sim \text{Normal}(0, \tau\sigma^2), \sigma^2 \sim IG(a_\sigma, b_\sigma)$, and the hyperparameter $\tau \sim IG(a_\tau, b_\tau)$, where $IG(a_\tau, b_\tau)$ denotes the inverse-gamma distribution with mean $1/\{b_\tau(a_\tau - 1)\}$. Since for any set $A, G(A)$ has mean $G_0(A)$ and variance $G_0(A)\{1 - G_0(A)\}/(\alpha_0 + 1)$, the parameter α_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 π_k'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 $\theta_{\text{risk}}(\bullet)$ and $\theta_{\text{cal}}(\bullet)$*

We wish to estimate $\theta_{\text{risk}}(X)$ in the interval $[a, b]$ and for this purpose we use B-splines. Let $\boldsymbol{B}_{\text{risk}}^{\mathrm{T}}(X) = \{B_{1\text{risk}}(X), \ldots, B_{p\,\text{risk}}(X)\}$ be a set of cubic spline basis functions for k_{risk} fixed knot points. With $p = k_{\text{risk}} + 4$, our model is

$$\theta_{\text{risk}}(X) = \sum_{j=1}^{p} B_{j\,\text{risk}}(X)\beta_j\,.$$

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

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

where $\boldsymbol{B}_{\text{cal}}^{\mathrm{T}}(X) = \{B_{1\text{cal}}(X), \ldots, B_{q\,\text{cal}}(X)\}$ 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 $\theta_{\text{risk}}(\bullet)$ properly, $\theta_{\text{cal}}(\bullet)$ 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 $\theta_{\text{cal}}(X)$, the first derivative $\theta_{\text{cal}}^{(1)}(X)$ is nonnegative if the coefficients α_j's are nondecreasing, i.e., $\alpha_1 \leqslant \alpha_2 \leqslant \cdots \leqslant \alpha_q$. This property of B-splines has been previously used by Leitenstorfer 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 $\boldsymbol{\beta}^{\mathrm{T}} = (\beta_1, \ldots, \beta_p), \boldsymbol{\alpha}^{\mathrm{T}} = (\alpha_1, \ldots, \alpha_q), \boldsymbol{\phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{n+m})$, and define $\boldsymbol{\theta} = (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\zeta}_{\text{risk}}^{\mathrm{T}}, \boldsymbol{\zeta}_{\text{cal}}^{\mathrm{T}}, \boldsymbol{\zeta}_x^{\mathrm{T}}, \sigma_Q^2, \sigma_W^2)$. The observed data likelihood function is

$$L_{\mathrm{o}} = \int \prod_{i=1}^{n+m} \Bigg\{ f^{1-\Delta_i}(Y_i, Q_i \mid \boldsymbol{Z}_i, \boldsymbol{\theta}, \boldsymbol{\phi}) \times f^{\Delta_i}(Q_i, W_{ij}, j=1,\ldots,J \mid \boldsymbol{Z}_i, \boldsymbol{\theta}, \boldsymbol{\phi}) \Bigg\} dG(\boldsymbol{\phi}), \text{where}$$

$$\begin{aligned} f(Y_i, Q_i \mid \boldsymbol{Z}_i, \boldsymbol{\theta}, \boldsymbol{\phi}) &= \int f(Y_i, Q_i, X_i \mid \boldsymbol{Z}_i, \boldsymbol{\theta}, \boldsymbol{\phi}) dX_i \\ &= \int pr^{Y_i}(Y=1 \mid X_i, \boldsymbol{Z}_i, \boldsymbol{\theta}) pr^{1-Y_i} \\ &\quad \times (Y=0 \mid X_i, \boldsymbol{Z}_i, \boldsymbol{\theta}) f(Q_i \mid X_i, \boldsymbol{Z}_i, \boldsymbol{\theta}) \\ &\quad \times f(X_i \mid \boldsymbol{Z}_i, \boldsymbol{\theta}, \boldsymbol{\phi}) dX_i \\ &= \int \frac{\exp[Y_i\{\theta_{\text{risk}}(X_i) + \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\zeta}_{\text{risk}}\}]}{1 + \exp\{\theta_{\text{risk}}(X_i) + \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\zeta}_{\text{risk}}\}} \frac{1}{(\sigma_Q^2)^{1/2}} \\ &\quad \times \exp\left[-\frac{1}{2\sigma_Q^2} \left\{ Q_i - \theta_{\text{cal}}(X_i) - \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\zeta}_{\text{cal}} \right\}^2 \right] \\ &\quad \times \frac{1}{(\sigma_i^2)^{1/2}} \exp\left\{ -\frac{1}{2\sigma_i^2} \left(X_i - \mu_i - \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\zeta}_x \right)^2 \right\} dX_i, \end{aligned}$$

$$\begin{aligned} & f(Q_i, W_{i1}, \ldots, W_{iJ} \mid \boldsymbol{Z}_i, \boldsymbol{\theta}, \boldsymbol{\phi}) \\ &= \int f(Q_i, W_{i1}, \ldots, W_{iJ}, X_i \mid \boldsymbol{Z}_i, \boldsymbol{\theta}, \boldsymbol{\phi}) dX_i \\ &= \int f(Q_i \mid X_i, \boldsymbol{Z}_i, \boldsymbol{\theta}) f(W_{i1}, \ldots, W_{iJ} \mid X_i, \boldsymbol{Z}_i, \boldsymbol{\theta}) \\ &\quad \times f(X_i \mid \boldsymbol{Z}_i, \boldsymbol{\theta}, \boldsymbol{\phi}) dX_i \\ &= \int \frac{1}{(\sigma_Q^2)^{1/2}} \exp\left[-\frac{1}{2\sigma_Q^2} \left\{ Q_i - \theta_{\text{cal}}(X_i) - \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\zeta}_{\text{cal}} \right\}^2 \right] \\ &\quad \times \frac{1}{(\sigma_W^2)^{J/2}} \exp\left\{ -\frac{1}{2\sigma_W^2} \sum_{j=1}^{J} (W_{ij} - X_i)^2 \right\} \\ &\quad \times \frac{1}{(\sigma_i^2)^{1/2}} \exp\left\{ -\frac{1}{2\sigma_i^2} \left(X_i - \mu_i - \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\zeta}_x \right)^2 \right\} dX_i. \end{aligned}$$

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 $\boldsymbol{Z}$ 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 $\boldsymbol{Z}$ where $\boldsymbol{Z}$ 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 $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, $\pi(\boldsymbol{\alpha} \mid \delta_\alpha) \propto \exp\{-(\delta_\alpha/2) \sum_{j=3}^{q} (\Delta^2 \alpha_j)^2 - (\delta_\alpha/2V) \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{\alpha}\}$, and $\pi(\boldsymbol{\beta} \mid \delta_\beta) \propto \exp\{-(\delta_\beta/2) \sum_{j=3}^{p} (\Delta^2 \beta_j)^2 - (\delta_\beta/2V) \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\beta}\}$, where δ_α and δ_β are the two penalty parameters corresponding to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Note that the priors are proper for any choice of $V > 0$, and we set $V = 10^8$. Also, Δ represents the first-order difference operator, i.e., $\Delta\alpha_j = \alpha_j - \alpha_{j-1}$, 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 σ_Q^2, σ_W^2, we use $IG(a_Q, b_Q)$ and $IG(a_W, b_W)$ priors respectively. For each component of $\boldsymbol{\zeta}_{\text{cal}}, \boldsymbol{\zeta}_{\text{risk}}, \boldsymbol{\zeta}_x$ we use a Normal$(0, \sigma_\zeta^2)$ prior.

As mentioned earlier, for $(\mu_i, \sigma_i^2) \sim G$ and a priori $G \sim DP(\alpha_0 G_0)$. The DP prior involves two parameters G_0 and α_0. Note that the tuning parameter of the DP prior α_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 α_0 is an important one. Following Escobar and West (1995) one may put a prior on α_0. However, we will estimate α_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: $V, a_\alpha, b_\alpha, a_\beta, b_\beta, a_Q, b_Q, a_W, b_W, \sigma_\zeta^2$, and G_0.

3.2 *Posterior Inference*

Our primary goal is to estimate $\boldsymbol{\zeta}_{\text{risk}}, \boldsymbol{\zeta}_{\text{cal}}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_Q^2$, and σ_W^2. 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 α_0 empirically by solving the following equation:

$$\sum_{r=1}^{N} \frac{\alpha_0}{\alpha_0 + r - 1} = E(k \mid X_i, i = 1, \ldots, N, \alpha_0), \qquad (4)$$

where the right-hand side of (4) is obtained by taking average of k, the number of distinct $\boldsymbol{\phi}_i$'s, of the last 30 Gibbs sampling iterations, i.e., we take $E(k \mid X_i, i = 1, \ldots, N, \alpha_0) \approx \sum_{l=1}^{30} k_l/30$.

Note, as $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N$ are independent and identically distributed observations from G, and G has a $DP(\alpha_0 G_0)$ prior, the posterior distribution of G given $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N$ is again a $DP(\alpha_0^* G_0^*)$, where $\alpha^* = \alpha_0 + N$ and $G_0^* = (\alpha_0 + N)^{-1}(\alpha_0 G_0 + \sum_{i=1}^{N} \delta_{\boldsymbol{\phi}_i})$. Thus, for future realization $\boldsymbol{\phi}_{N+1}$, $\text{pr}(\boldsymbol{\phi}_{N+1} \in A \mid \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N) = E\{G(A) \mid G \sim DP(\alpha_0^* G_0^*)\} = G_0^*(A)$. Therefore, the predictive distribution of $\boldsymbol{\phi}_{N+1} \mid \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N$ 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. Thiébaut 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/m^2 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 75$^{\text{th}}$ percentile or below the 25$^{\text{th}}$ 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 alloyed gold standard. Of the 1953 subjects of both genders, we considered only the 919 women who completed two 24-hour recall interviews and whose BMI values were not missing. Note that these women were all postmenopausal and cancer free at the start of the study and they were a random sample from the main study population. The percentage of nonalcohol energy from total fat obtained from the two 24-hour recalls is considered as the unbiased surrogate, W. We defined the response variable Y as one if a person develops invasive breast cancer during the follow-up period and zero otherwise. Before the analysis we transformed BMI by dividing by 10, and X was the logarithm of the percentage of nonalcohol energy from total fat. In the preliminary analysis, we fit model (1) without any adjustment for measurement error using the primary data. The result showed that X was significantly positively associated with the risk of invasive breast cancer with an odds ratio of 1.176 with a 95% confidence interval $(1.016, 1.363)$ and p-value $= 0.030$. Also, we found that BMI was positively associated with the disease with an odds ratio 1.069 with a 95% confidence interval (1.008, 1.135) and p-value $= 0.026$.

4.2 *Semiparametric Results*

First, we used 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 $\theta_{\text{risk}}(\cdot)$ and $\theta_{\text{cal}}(\cdot)$. For δ_α and δ_β we used $\text{Gamma}(0.001, 1000)$ priors. We used $IG(25, 0.25)$ prior for σ^2_Q and σ^2_W, and $\text{Normal}(0, 10^2)$ prior for $\zeta_{\text{risk}}, \zeta_{\text{cal}}$, and ζ_x. 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 $\theta_{\text{risk}}(\cdot)$ 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 $(Q - \widehat{\zeta}_{\text{cal}} Z, \overline{W})$ from the calibration data. Both figures clearly indicate that there is some sort of nonlinearity in both $\theta_{\text{risk}}(\cdot)$ and $\theta_{\text{cal}}(\cdot)$. Table 1 presents the posterior mean, posterior standard deviation, and 95% credible intervals for $\zeta_{\text{risk}}, \zeta_{\text{cal}}, \zeta_x, \sigma^2_Q$, and σ^2_W.

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 α_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 $G_0^* = E(G \,|\, \text{data})$ based on the MCMC sample of $\boldsymbol{\phi}_{N+1}$.

The null hypothesis of interest is that there is no risk of fat on breast cancer: under this null hypothesis θ_{risk} should be constant and thus it is equivalent to test $\beta_1 = \cdots = \beta_p$ as $\sum_{j=1}^{p} B_{j\,\text{risk}}(X) = 1$.

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 $F = [(M - p + 1)/\{(M-1)(p-1)\}]T^2$, where $T^2 = M\bar{\boldsymbol{D}}^{\text{T}} S_{\text{D}}^{-1} \bar{\boldsymbol{D}}, \bar{\boldsymbol{D}}^{\text{T}} = (\bar{D}_1, \ldots, \bar{D}_{p-1}), \bar{D}_k = \sum_{j=1}^{M} D_{kj}/M, k = 1, \ldots, (p-1)$.

Here $D_{kj} = \beta_{kj} - \beta_{pj}, \beta_{kj}$ is the jth value of β_k of M observations, $S_{\text{D}} = \sum_{j=1}^{M} (\boldsymbol{D}_j - \bar{\boldsymbol{D}})(\boldsymbol{D}_j - \bar{\boldsymbol{D}})^{\text{T}}/(M-1)$ with $\boldsymbol{D}_j^{\text{T}} = (D_{1j}, \ldots, D_{p-1j})$. Under the null hypothesis, for sufficiently large M, the statistic follows an $F_{(p-1),(M-p+1)}$

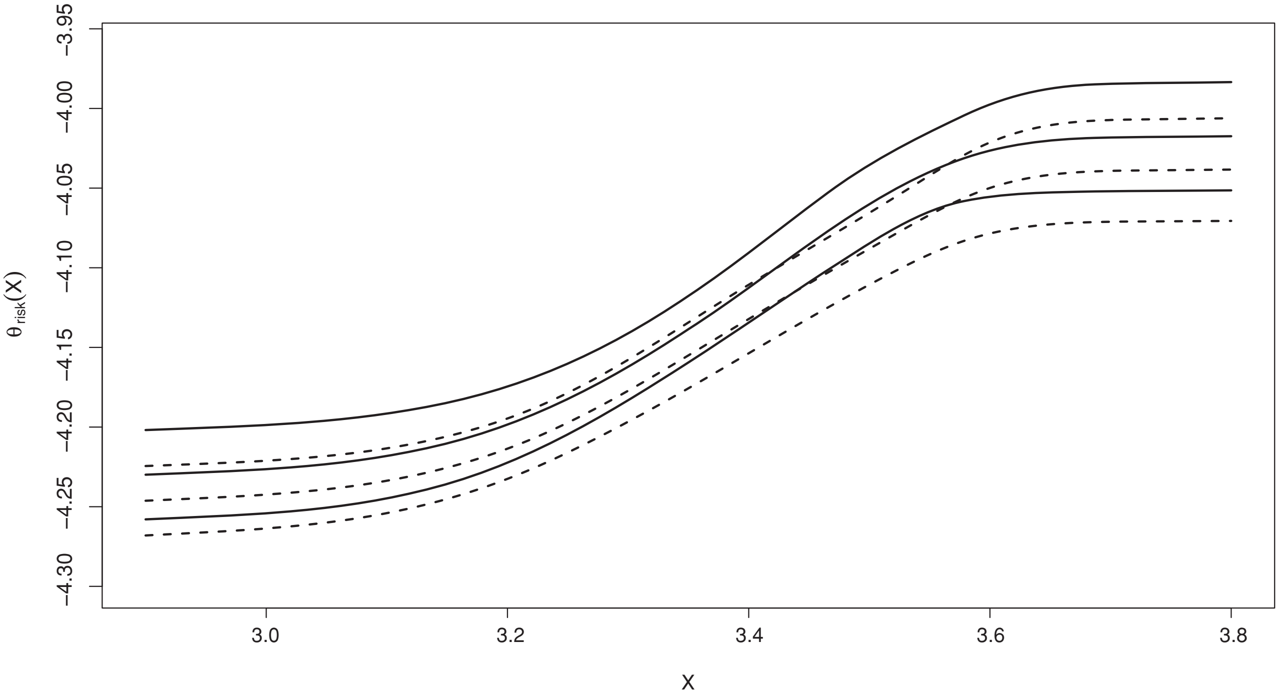

Figure 1. Plot of the estimated $\theta_{\text{risk}}(X)$ along with a pointwise 95% credible interval using the semiparametric Bayesian approach (SPB, solid lines) and the parametric Bayesian approach (PRB, dashed lines) approach.

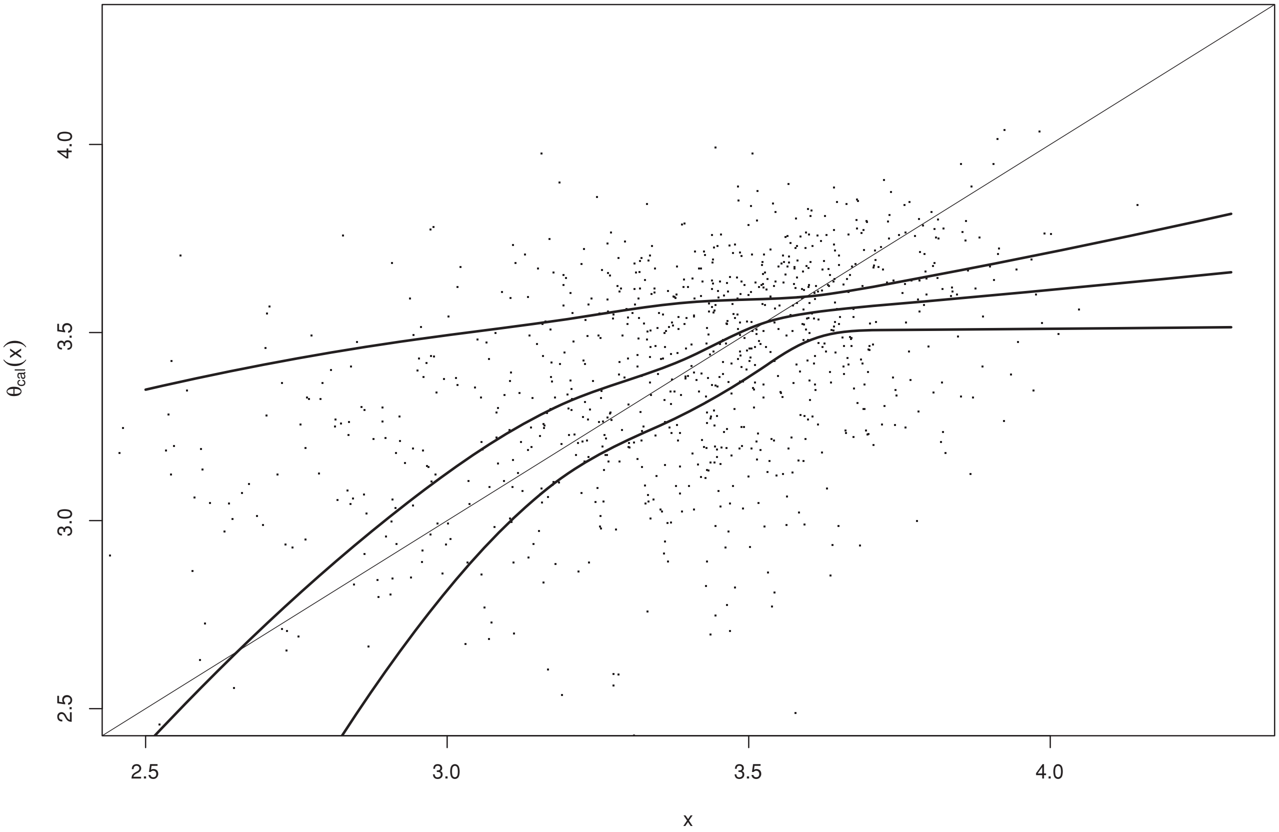

Figure 2. This figure shows the estimated $\theta_{\text{cal}}(X)$ using the SPB method, along with a 95% pointwise credible interval. This figure also contain the scatter plot of $(Q - \widehat{\zeta}_{\text{cal}} Z, \overline{W})$ of the calibration data, and a line with zero intercept and slope one.

Table 1

Results of the NIH-AARP data analysis by the semiparametric Bayesian (SPB) and parametric Bayesian (PRB) methods

Method	Parameter	Posterior mean	Posterior standard deviation	95% credible interval
SPB	ζ_{risk}	0.0516	0.0173	(0.0205, 0.0830)
	ζ_{cal}	0.0008	0.0039	($-$0.0046, 0.0096)
	σ_Q^2	0.0260	0.0025	(0.0229, 0.0295)
	ζ_x	0.0697	0.0038	(0.0598, 0.07677)
	σ_W^2	0.1104	0.0145	(0.0889, 0.1368)
PRB	ζ_{risk}	0.0397	0.0234	($-$0.0134, 0.0754)
	ψ_0	1.3960	0.1365	(1.0577, 1.6327)
	ψ_1	0.5915	0.0463	(0.5137, 0.7036)
	ζ_{cal}	0.0268	0.0052	(0.0134, 0.0327)
	σ_Q^2	0.0496	0.0019	(0.0458, 0.0533)
	γ_0	3.1550	0.0562	(3.0576, 3.2253)
	ζ_x	0.0839	0.0199	(0.0563, 0.1187)
	σ_X^2	0.0528	0.0065	(0.0370, 0.0637)
	σ_W^2	0.0916	0.0043	(0.0836, 0.1065)

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 $D_{k\overline{j+1}}$ for any k. Note that if X changes from 3.0 to 3.1, according to our SPB method, $\theta_{\text{risk}}(3.0) = -4.226$ and $\theta_{\text{risk}}(2.9) = -4.223$, and thus the risk increases by 0.35%. In contrast, $\theta_{\text{risk}}(3.4) = -4.113$, and $\theta_{\text{risk}}(3.3) = -4.162$ 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/m^2 BMI. In addition, following a referee's comment we fitted model (1) by replacing X by $\hat{X} = \widehat{\eta}_0 + \widehat{\eta}_1 Q + \widehat{\eta}_3 Z$, where $\widehat{\eta}_0, \widehat{\eta}_1, \widehat{\eta}_3$ 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 $\theta_{\text{cal}}(\cdot)$ 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

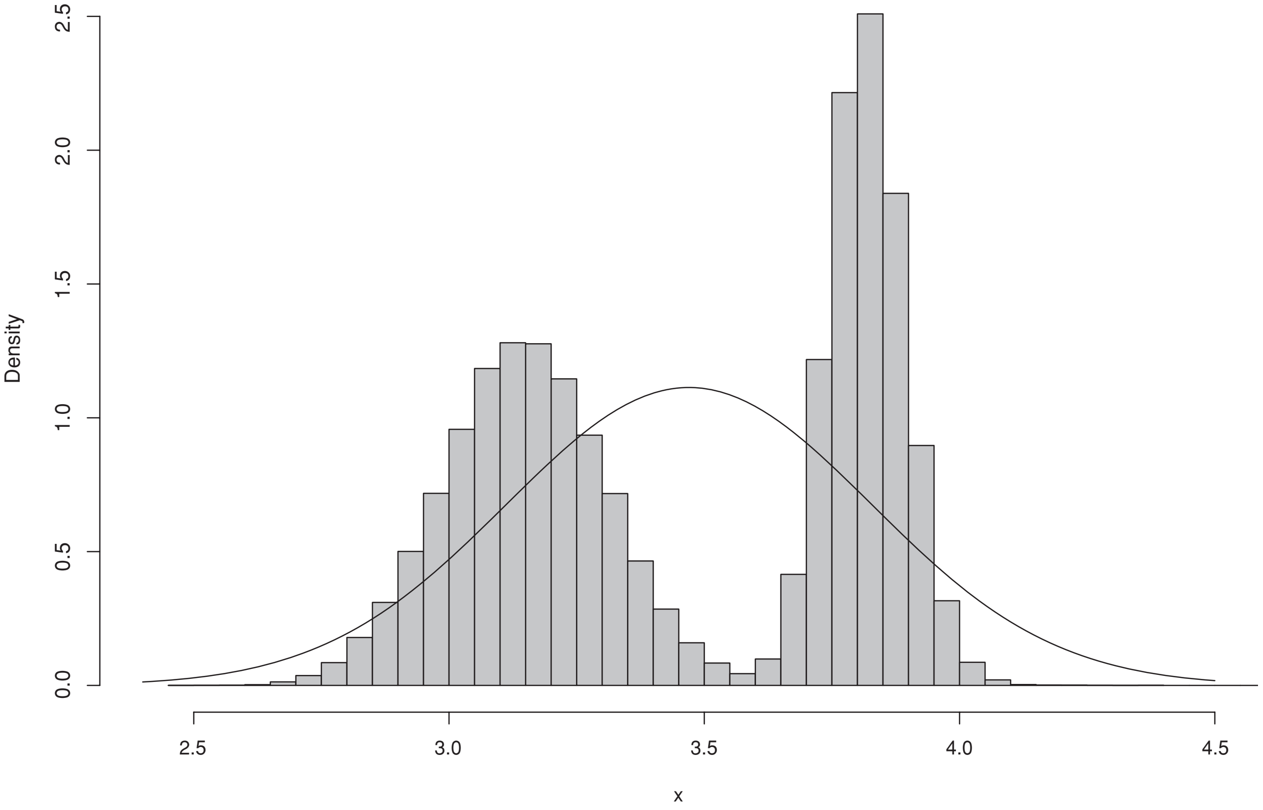

Figure 3. Histogram of X obtained from the MCMC sample using our SPB method along with the density plot of a normal distribution with the same mean and variance as of X.

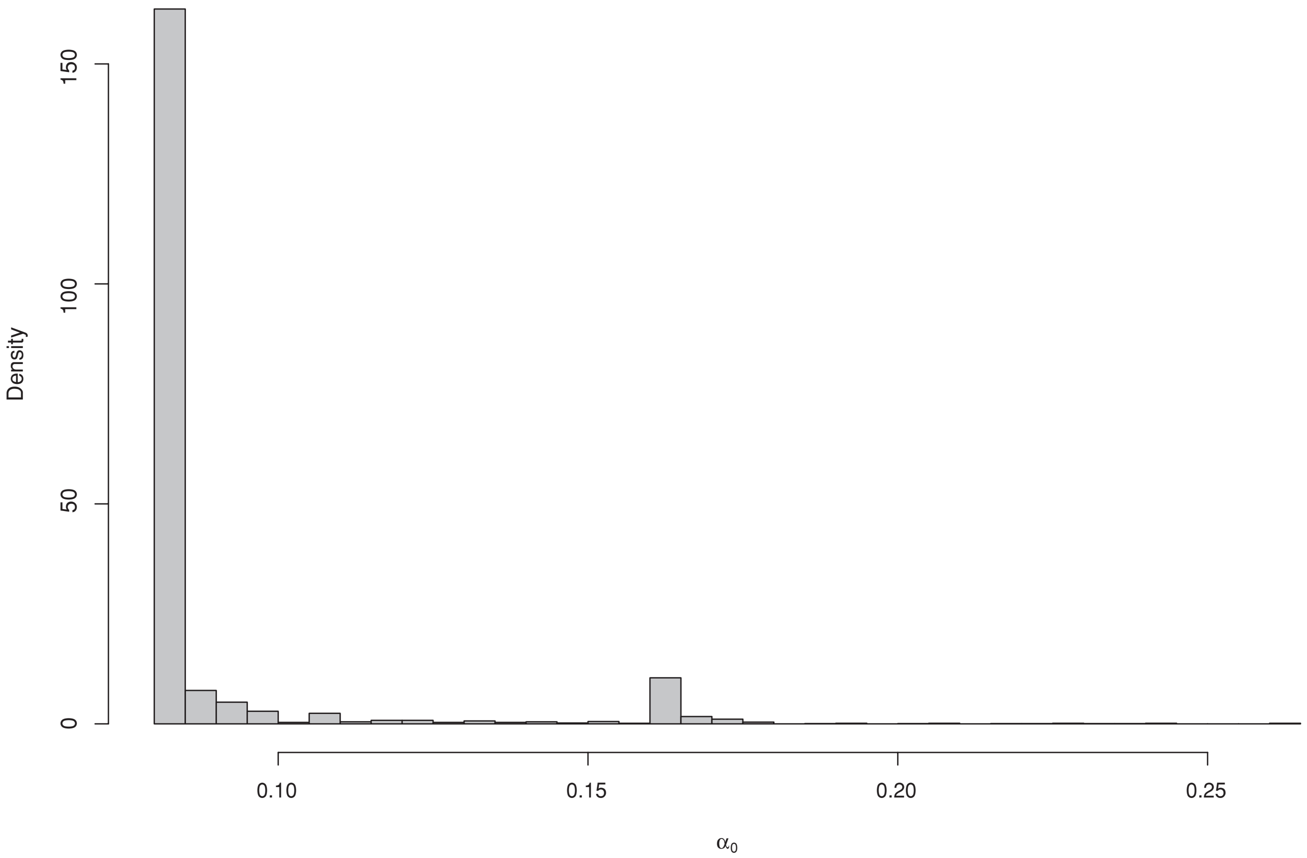

Figure 4. An empirical distribution of α_0 of the Dirichlet process prior that is determined according to equation (4) in each MCMC sample of the SPB method.

PRB method we modeled the surrogate variable as $Q_i = \psi_0 + \psi_1 X_i + Z_i\zeta_{\mathrm{cal}} + U_{Qi}, U_{Qi} \sim \mathrm{Normal}(0, \sigma_Q^2)$ and assumed that $X_i \sim \mathrm{Normal}(\mu_0 + Z_i\zeta_x, \sigma_X^2)$. 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 ψ_0, ψ_1, and μ_0, we used $\mathrm{Normal}(0, 10^2)$ priors, and used $IG(2.1, 0.91)$ prior for σ_X^2. The posterior mean, standard deviation, and 95% credible intervals for the parameters are given in Table 1. It is obvious that the estimate of σ_Q^2 is significantly larger in the PRB approach compared to the SPB method, which is an indicator that possibly $E(Q \mid X, Z)$ is better explained by the partly linear function $\theta_{\mathrm{cal}}(X) + Z\zeta_{\mathrm{cal}}$ than a linear function of X and Z. The dashed lines in Figure 1 show the estimated θ_{risk} for the PRB approach along with a 95% credible interval. For this method, using the above F-test, we again concluded that θ_{risk} 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 $\theta_{\mathrm{cal}}(\cdot)$ 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 $\Theta^{\mathrm{T}} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\phi})$ is the set of all parameters, then the marginal likelihood under the SPB model is $m_{\mathrm{SPB}} = \prod_{i=1}^{n} \int f^{1-\Delta_i}(Y_i, Q_i \mid \boldsymbol{Z}_i; \Theta) f^{\Delta_i}(Q_i, W_{ij} \mid \boldsymbol{Z}_i; \Theta)\, \pi(\Theta) d\Theta$, where the expressions for $f(Y_i, Q_i \mid \boldsymbol{Z}_i; \Theta)$ and $f(Q_i, W_{ij} \mid \boldsymbol{Z}_i; \Theta)$ are given in Section 3. Following Newton and Raftery (1994), we estimate the marginal likelihood via

$$\widehat{m}_{\mathrm{SPB}} = \left\{ \frac{1}{M} \sum_{m=1}^{M} \frac{1}{\prod_{i=1}^{n} f^{1-\Delta_i}(Y_i, Q_i \mid \boldsymbol{Z}_i; \Theta^{(m)}) f^{\Delta_i}(Q_i, W_{ij} \mid \boldsymbol{Z}_i; \Theta^{(m)})} \right\}^{-1},$$

where $\Theta^{(1)}, \ldots, \Theta^{(M)}$ are M draws from the posterior distribution of Θ. Note that $f(Y_i, Q_i | \boldsymbol{Z}_i; \Theta^{(m)}) = \int f(Y_i, Q_i | X_i, \boldsymbol{Z}_i; \Theta^{(m)}) f(X_i | \boldsymbol{Z}_i, \Theta^{(m)}) dX_i$ and $f(Q_i, W_{ij} | \boldsymbol{Z}_i; \Theta^{(m)}) = \int f(Q_i, W_{ij} | X_i, \boldsymbol{Z}_i; \Theta^{(m)}) f(X_i | \boldsymbol{Z}_i, \Theta^{(m)}) dX_i$: these integrals were calculated by a numerical quadrature. In a similar manner we estimated m_{PRB} via $\widehat{m}_{\mathrm{PRB}}$. Finally, we obtain $\widehat{m}_{\mathrm{BSP}} / \widehat{m}_{\mathrm{PRB}} = 111.21$, 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 $(1/2)$Gamma$(65, 0.0455) +$ $(1/2)$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 $W_{ij} = X_i + U_{W\,ij}$, where $U_{W\,ij} = \text{Normal}(0, \sigma_W^2), j = 1, 2$. We took two different values for σ_W, 0.2 and 0.5, where the value 0.5 mimics the large measurement error situation common in nutritional epidemiology. The surrogate variable Q was generated as $Q_i = \theta_{\text{cal}}(X_i) + U_{Q\,i}, U_{Q\,i} = \text{Normal}(0, \sigma_Q^2)$ and we used two different values for σ_Q, 0.1 and 0.2. We considered a situation where $\theta_{\text{cal}}(X) = 1.25 + 0.6X$, a linear function of X and another situation where $\theta_{\text{cal}}(X) = 2 + 3\exp\{4(X - 3.5)\}/[1 + \exp\{4(X - 3.5)\}]$. Conditional on X, the binary disease variable Y was generated with success probability $H\{\theta_{\text{risk}}(X)\}$, where $\theta_{\text{risk}}(X) = -2.8 + 1.5\exp\{10(X - 3.5)\}/[1 + \exp\{10(X - 3.5)\}]$. 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 θ_{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, $\theta_{\text{cal}}(X)$ was assumed to be linear in X, whereas in the SPB method we do not assume any specific form for θ_{cal}, rather we simply impose the restriction that $\theta_{\text{cal}}(X)$ 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$(0.001, 1000)$ distribution for δ_β. In the analysis, we used $IG(25, 0.25)$ prior for σ_Q^2 and σ_W^2. The PRB approach involves ψ_0 and ψ_1, 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 μ_0 and $IG(2.1, 0.91)$ prior for σ_X^2. In the SPB method, we need to estimate $\theta_{\text{cal}}(X)$ using B-splines, which involves $\alpha_j, j = 1, \ldots, q$, and we used a Gamma$(0.001, 1000)$ prior for δ_α. The previously mentioned knot points are also used for estimating $\theta_{\text{cal}}(X)$. In the DP prior, we used $IG(2.1, 2)$ and $IG(2.1, 0.91)$ priors for τ and σ^2, respectively.

For the initial values of X and α, we first regress $(W_1 + W_2)/2$ on Q in the calibration study. Let $\widehat{\eta}_0, \widehat{\eta}_1$, and $\widehat{\sigma}_\eta^2$ 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 $X_i = \widehat{\eta}_0 + \widehat{\eta}_1 Q_i$, for $i = 1, \ldots, n$ and set $X_i = (W_{i1} + W_{i2})/2$ for $i = (n+1), \ldots, (n+m)$. The initial value of $\boldsymbol{\alpha}$ 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 $\delta_\alpha = 0.005$ and $\delta_\beta = 0.005$.

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 $\theta_{\text{risk}}(X)$. We considered a set of grid points from 2.8 to 4.2 with 0.01 increment. Letting g_j be the jth grid point, for the sth data set, we estimated $\theta_{\text{risk}}(g_j)$ by $\widehat{\theta}_{s,\text{risk}}(g_j)$ and let $\widehat{\theta}_{\text{risk}}(g_j) = \sum_{s=1}^{200} \widehat{\theta}_{s,\text{risk}}(g_j)/200$. We computed integrated square bias as $\sum_{j=1}^{141} \{\widehat{\theta}_{\text{risk}}(g_j) - \theta_{\text{risk}}(g_j)\}^2/141$ and integrated mean square error as $\sum_{j=1}^{141} \sum_{s=1}^{200} \{\widehat{\theta}_{s,\text{risk}}(g_j) - \theta_{\text{risk}}(g_j)\}^2/(141 \times 200)$. In addition to the above methods, each data set was analyzed by the *naive* method where we estimated $\theta_{\text{risk}}(X)$ using the data on (Y_i, Q_i) and assuming $X_i = Q_i$. For this method, we modeled $\theta_{\text{risk}}(X)$ via B-splines and estimated $\boldsymbol{\beta}$ and δ_β using the logistic likelihood in a Bayesian framework.

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 σ_Q and σ_W. The value of integrated squared bias due to PRB method is significantly higher than the SPB approach when the underlying model assumptions of the PRB approach are violated. When $E(Q \mid X)$ is a linear function of X, and X follows a normal distribution, the performance of the SPB and PRB methods is equivalent for moderate values of σ_W. Note that the integrated squared bias of the naive method is significantly larger than that for the SPB method. Also we found that except for the situation when $\theta_{\text{cal}}(X)$ is a nonlinear function of its argument and X is generated from a mixture distribution, the bias of the PRB approach is smaller than that for the naive method. Generally the variance of the SPB method is higher than the PRB approach, and the variance of the naive approach is smaller among all three approaches.

Along with the proposed method each data set was analyzed by the RC technique assuming X was a linear function of Q. As expected from Carroll et al. (2006), the method behaved very badly in comparison to the SPB and PRB methods, and we do not present the results here.

6. Discussion

In this article, we considered the logistic regression when an error-prone covariate X is modeled nonparametrically, while exactly measured covariates $\boldsymbol{Z}$ 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 $(X, \boldsymbol{Z})$ also in a partially linear fashion, while forcing the nonparametric function $\theta_{\text{cal}}(X)$ for X to be monotone.

Table 2

Results of the simulation study based on 200 simulations. Here, ISB and IMSE represent integrated square bias and the integrated mean square error. Here $n = 10,000$ and $m = 1,000$. In case I, X is generated from $\text{Normal}(3.5, 0.5^2)$, *and in case II, X is generated from* $(1/2)\text{Normal}(3.8, 0.1^2) + (1/2)\text{Gamma}(65, 0.0455)$. *Note that* $U_{Qi} \sim \text{Normal}(0, \sigma_Q^2)$, *and* $W = X + U_W$, *where* $U_W \sim \text{Normal}(0, \sigma_W^2)$. *SPB and PRB stand for the proposed semiparametric Bayesian approach and the parametric Bayesian approach to the analysis of the simulated data, respectively.*

Case	σ_Q	Method	ISB	IMSE	ISB	IMSE	ISB	IMSE
I					$\theta_{\text{cal}} = 1.25 + 0.6X_i$			
			$\sigma_W = 0.2$		$\sigma_W = 0.5$		Naive	
	0.1	SPB	0.0062	0.0149	0.0080	0.0155	0.0905	0.1004
		PRB	0.0058	0.0140	0.0067	0.0147		
	0.2	SPB	0.0065	0.0152	0.0082	0.0164	0.0915	0.1011
		PRB	0.0060	0.0145	0.0069	0.0150		
I			$\theta_{\text{cal}}(X) = 2 + 3\exp\{4(X_i - 3.5)\}/[1 + \exp\{4(X_i - 3.5)\}]$					
			$\sigma_W = 0.2$		$\sigma_W = 0.5$		Naive	
	0.1	SPB	0.0087	0.0161	0.0089	0.0164	0.0462	0.0551
		PRB	0.0114	0.0194	0.0144	0.0236		
	0.2	SPB	0.0088	0.0169	0.0092	0.0189	0.0470	0.0575
		PRB	0.0118	0.0198	0.0150	0.0249		
II					$\theta_{\text{cal}}(X) = 1.25 + 0.6X_i$			
			$\sigma_W = 0.2$		$\sigma_W = 0.5$		Naive	
	0.1	SPB	0.0149	0.0354	0.0221	0.0429	0.1550	0.1635
		PRB	0.0283	0.0423	0.0560	0.0801		
	0.2	SPB	0.0168	0.0368	0.0245	0.0450	0.2004	0.2208
		PRB	0.0289	0.0439	0.0805	0.1030		
II			$\theta_{\text{cal}}(X) = 2 + 3\exp\{4(X_i - 3.5)\}/[1 + \exp\{4(X_i - 3.5)\}]$					
			$\sigma_W = 0.2$		$\sigma_W = 0.5$		Naive	
	0.1	SPB	0.0184	0.0394	0.0335	0.0552	0.0610	0.0668
		PRB	0.0606	0.0749	0.0684	0.0831		
	0.2	SPB	0.0232	0.0450	0.0347	0.0560	0.0650	0.0710
		PRB	0.0799	0.0965	0.0874	0.1054		

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

Berry, S. M., Carroll, R. J., and Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association* **97,** 160–169.

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association* **83,** 1184–1186.

Carroll, R. J., Ruppert, D., Tosteson, T. D., Crainiceanu, C., and Karagas, M. R. (2004). Nonparametric regression and instrumental variables. *Journal of the American Statistical Association* **99,** 736–750.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd edition. New York: Chapman and Hall.

Delaigle, A. and Hall, P. (2008). Using SIMEX for smoothing parameter choices in errors in variables problems. *Journal of the American Statistical Association* **103,** 280–287.

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

Fan, J. and Truong, Y. K. (1993). Nonparametric regression with errors in variables. *Annals of Statistics* **21,** 1900–1925.

Holmes, C. C. and Mallick, B. K. (2003). Generalized nonlinear modeling with multivariate free-knot regression splines. *Journal of the American Statistical Association* **98,** 352–368.

Johnson, B. A., Herring, A. H., Ibrahim, J. G., and Siega-Riz, A. M. (2007). Structured measurement error in nutritional epidemiology: Applications in the pregnancy, infection, and nutrition (PIN) study. *Journal of the American Statistical Association* **102,** 856–866.

Kipnis, V., Midthune, D., Freedman, L. S., Bingham, S., Schatzkin, A., Subar, A., and Carroll, R. J. (2001). Empirical evidence of correlated biases in dietary assessment instruments and its implications. *American Journal of Epidemiology* **153,** 394–403.

Kipnis, V., Subar, A. F., Midthune, D., Freedman, L. S., Ballard-Barbash, R., Troiano, R., Bingham, S., Schoeller, D. A., Schatzkin, A., and Carroll, R. J. (2003). The structure of dietary measurement error: Results of the OPEN biomarker study. *American Journal of Epidemiology* **158,** 14–21.

Leitenstorfer, F. and Tutz, G. (2007). Generalized monotonic regression based on B-splines with an application to air pollution data. *Biostatistics* **8,** 654–673.

Mallick, B. K. and Gelfand, A. E. (1996). Semiparametric errors-in-variables models: A Bayesian approach. *Journal of Statistical Planning and Inference* **52,** 307–321.

McAuliffe, J. D., Blei, D. M., and Jordan, M. I. (2006). Nonparametric empirical Bayes for the Dirichlet process mixture model. *Statistics and Computing* **16,** 5–14.

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

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Series B* **56,** 3–26.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* **11,** 735–757.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

Schatzkin, A., Subar, A. F., Thompson, F. E., Harlan, L. C., Tangrea, J., Hollenbeck, A. R., Hurwitz, P. E., Coyle, L., Schussler, N., Michaud, D. S., Freedman, L. S., Brown, C. C., Midthune, D., and Kipnis, V. (2001). Design and serendipity in establishing a large cohort with wide dietary intake distributions. *American Journal of Epidemiology* **154,** 1119–1125.

Thiébaut, A. C. M., Kipnis, V., Chang, S.-C., Subar, A. F., Thompson, F. E., Rosenberg, P. S., Hollenbeck, A. R., Leitzmann, M., and Schatzkin, A. (2007). Dietary fat and postmenopausal invasive breast cancer in the National Institutes of Health-AARP diet and health study cohort. *Journal of the National Cancer Institute* **99,** 451–462.

Wood, S. and Kohn, R. (1998). A Bayesian approach to robust binary nonparametric regression. *Journal of the American Statistical Association* **93,** 203–213.

Wood, S., Kohn, R., Shivley, T., and Jiang, W. (2002). Model selection in spline nonparametric regression. *Journal of the Royal Statistical Society, Series B* **64,** 119–139.

Received December 2008. *Revised May* 2009.
Accepted May 2009.