

Semiparametric Regression Modeling with Mixtures of Berkson and Classical Error, with Application to Fallout from the Nevada Test Site

Bani Mallick,^{1,*} F. Owen Hoffman,^{2,**} and Raymond J. Carroll^{1,***}

¹Department of Statistics, Texas A&M University,
College Station, Texas 77843-3143, U.S.A.

²SENES Oak Ridge, Center for Risk Analysis,
102 Donner Drive, Oak Ridge, Tennessee 37830, U.S.A.

**email:* bmallick@stat.tamu.edu

***email:* fohoff3084@aol.com

****email:* carroll@stat.tamu.edu

SUMMARY. We construct Bayesian methods for semiparametric modeling of a monotonic regression function when the predictors are measured with classical error, Berkson error, or a mixture of the two. Such methods require a distribution for the unobserved (latent) predictor, a distribution we also model semiparametrically. Such combinations of semiparametric methods for the dose-response as well as the latent variable distribution have not been considered in the measurement error literature for any form of measurement error. In addition, our methods represent a new approach to those problems where the measurement error combines Berkson and classical components. While the methods are general, we develop them around a specific application, namely, the study of thyroid disease in relation to radiation fallout from the Nevada test site. We use this data to illustrate our methods, which suggest a point estimate (posterior mean) of relative risk at high doses nearly double that of previous analyses but that also suggest much greater uncertainty in the relative risk.

KEY WORDS: Bayes; Berkson error; Classical error; Dose-response; Latent variables; Likelihood; Measurement error; Pólya trees; Radiation epidemiology; Semiparametric; Thyroid cancer.

1. Introduction

This article develops semiparametric Bayesian methods for regression problems where a predictor is measured with either classical error, Berkson error, or a combination of classical and Berkson measurement error. We allow the regression function and the distribution of the unobservable (latent) covariate to be modeled either parametrically or nonparametrically. Our methods are applied to a study of thyroid cancer induced by fallout from nuclear testing (Stevens et al., 1992).

There is of course an enormous literature on regression problems where the latent covariate is measured either entirely with classical error or entirely with Berkson error (Carroll, Ruppert, and Stefanski, 1995). There have been numerous articles that model the latent variable semiparametrically (Roeder, Carroll, and Lindsay, 1996; Müller and Roeder, 1997; Carroll, Roeder, and Wasserman, 1999; Schafer, 2001; Richardson et al., unpublished manuscript). There are also articles that model the regression function semiparametrically (Carroll, Maca, and Ruppert, 1999). However, to date, no one has exhibited methods that are semiparametric both in the model and in the latent variable distribution. This article exhibits such methods. We focus for specificity on radiation

epidemiology, where the latent variable is the dose to an individual, typically measured with a combination of classical and Berkson errors. The methods developed to date for these models (cf., Reeves et al., 1998; Schafer et al., 2002) rely on approximations to the regression function given the observed data and typically use as the predictor an estimate of its conditional expectation given the observed dose, the so-called regression calibration approach.

This article takes a Bayesian approach. In the problem of interest, the regression function is reasonably thought to be monotone in the latent variable, so we allow either a parametric form or a semiparametric monotone form. In addition, the likelihood of the mixed Berkson-classical model depends on the distribution of the latent variable; this distribution we model either parametrically or flexibly semiparametrically.

In our example and in other such exercises in radiation dosimetry, the estimation of an individual's dose is the result of a complex modeling process including physical transport systems, biological processes, and direct measurements. It is typical to assign to the dose a total uncertainty, which is in effect the sum of the Berkson error variance and the classical error variance. This total uncertainty is known nominally at

the individual level, but the relative contribution of Berkson and classical errors is unknown. However, in these cases, is it reasonable to suppose that the proportion p of the error variance that is due to Berkson error lies within a defined interval on $[0, 1]$. Our Bayesian methods place a prior distribution on the relative contribution p , being uniformly distributed on the predefined interval.

This article is structured as follows. In Section 2, we describe the Nevada test-site data in detail. In Section 3, we describe parametric and semiparametric models for the dose–response. Of particular note in this section is that we develop a semiparametric approach that makes the dose–response monotonic, using a mixture-of-beta cumulative distribution functions (CDFs) approach.

In both these sections, we show that the likelihood function depends on the distribution of a latent variable and may thus be sensitive to misspecification of this distribution. Indeed, in our example, we present evidence that the latent variable in the natural log-dose scale is far from normally distributed. Section 4 describes our Bayesian approach in detail. Of particular note here is that, instead of specifying a distribution for the latent variable, we model the latent variable semiparametrically, using Pólya trees. Section 5 contains the reanalysis of the Nevada test-site data. Section 6 contains the results of a small simulation study. Section 7 has concluding remarks. An appendix gives brief details of the priors and Metropolis–Hastings proposals used in our calculations.

2. Berkson/Classical Errors

Stevens et al. (1992) describe a study of thyroid disease in relation to fallout from the Nevada test site (NTS). Similar statistical issues arise in the Hanford Thyroid Disease Study (Davis et al., 1998) and the Oak Ridge Radiation Study (Ostrouchov, Frome, and Kerr, 1998). In the Nevada study, 2473 individuals who were exposed to radiation as children were examined for thyroid disease. The primary radiation exposure came from milk and vegetables. Dosimetry calculations were based on age at exposure, gender, residence history, x-ray history, whether as a child the individual was breast fed, and a diet questionnaire filled out by the parent focusing on milk consumption and vegetables. The data were then fed into a complex model and, for each individual, the point estimate of thyroid dose and an associated standard error were reported. Unfortunately, only the summary statistics are available in the data file.

A statistically significant relationship between dose and neoplasms developed was obtained when fitting a logistic regression model with stratum-specific intercepts, adjustments for confounders, and a term for dose of the form $\log(1 + \theta \text{dose})$ (see below for more details). In one such analysis (Stevens et al., p. 208), the estimate of θ more than doubled after accounting for dose uncertainty by assuming a classical error model, i.e., if all the error were classical and error is ignored, relative risks are underestimated. In Section 3.1, we show that assuming that all the error is Berkson and ignoring classical error overestimates relative risk.

It is helpful to consider the model used to calculate dose to the thyroid of a specified individual from a single milk source contaminated by a single Nevada test-site event. This model has the following form (Stevens et al., p. 85):

$$W = C \times DCF \times I \times TD \times FP, \quad (1)$$

where W = reported dose to thyroid of the subject; C = time-integrated radioiodine concentration of milk; DCF = ingestion dose conversion factor; I = individual milk intake rate in liters per day, measured by a food frequency questionnaire; FP = frequency of purchase correction factor; and TD = time-delay factor. A detailed elicitation of the error structure for each component is not possible because of space limitations. The following is a brief summary. Milk intake (I), information on frequency of purchase (FP), and the sources of milk used to compute the time-delay factor (TD) come from a food frequency questionnaire (FFQ) filled out by the parent. As such, the error here is probably best thought of as mainly classical. The ingestion dose conversion factors (DCF) are specific for age and isotope. Uncertainties associated with DCF are probably best modeled as a mixture of Berkson and classical types. The time-integrated radioiodine concentration of milk (C) is specific not to individuals but to producers. One would ordinarily think of C as of Berkson type, but there is a major component of it that is classical, namely the deposition of I^{131} (Kerber et al., 1993; Simon et al., 1995) across the regions under study. Thus, the error structure for estimated dose to the thyroid has a mixture of classical and Berkson error.

This brief outline is simplistic. For example, the mass interception of I^{131} on vegetation and the transfer of iodine from feed to cow's milk are important components of the DCF . Their distribution is estimated by a combination of data from a literature review and expert judgment, thus combining classical and Berkson error in complex ways.

Reeves et al. (1998) consider data with a mixture of Berkson and classical error, although in a context far different from ours. At a formal mathematical level, their model is applicable to the Utah study; our approaches to analysis are far different. Let Y be the indicator of disease and let Z denote a vector of covariates measured without error, e.g., age, sex, and state. We will take logarithms in (1) and assign Berkson and classical error formulas to the pieces as described above. Denote true dose by X and observed dose by W . Then there is a latent variable L , which we call the latent intermediate variable, such that

$$\log(X) = \log(L) + U_b, \quad (2)$$

$$\log(W) = \log(L) + U_c. \quad (3)$$

The terminology latent intermediate variable is suggestive because L is intermediate between X and W .

We assume that (W, L) are conditionally independent of the response Y given (X, Z) and that the Berkson and classical errors are independent. Here U_b is Berkson error with variance σ_b^2 , U_c is classical error with variance σ_c^2 , and $\log(L)$ has mean μ_L and variance σ_L^2 . With a change in notation, these models are the same as model (4) in Reeves et al. There are covariates Z measured without error, so as is standard in the measurement error problem, e.g., μ_L may be allowed to depend on a linear function of Z and σ_L^2 is understood to be a conditional variance given Z .

The Utah study data file provides the sum of the Berkson and classical error variances for each individual but does not provide the relative contribution of each to the sum. Our approach is to allocate the total error variance across the two

types, where we allow for a fixed proportion p of an individual's error to come from each source, and vary this source in our Bayesian analysis by placing an informative prior on the fixed proportion.

3. Dose–Response Modeling

In this section, we provide a discussion of model fitting when the distribution of the latent intermediate variable L in (2)–(3) is specified. For convenience, for now we assume that $\log(L)$ is normally distributed conditional on Z and $X > 0$, where Z consists of the patient age at exposure, sex, and state of residence (Utah, Nevada, Arizona). In state s ,

$$[\log(L) \mid Z, \text{state} = s] \sim \text{normal} \left(\alpha_{s0} + Z^T \alpha_{s1}, \sigma_s^2 \right). \quad (4)$$

We denote by \mathcal{A} the collection of these parameters.

In general, we consider four types of modeling efforts: (a) all dose uncertainties are ignored; (b) error purely of Berkson type; (c) error purely of classical type; and (d) error a mixture of Berkson and classical errors, with the fraction of variance due to the Berkson part being p , i.e., $\sigma_b^2 / (\sigma_b^2 + \sigma_c^2) = p$. In the latter case, we face an identifiability issue: while the total uncertainty $\sigma_b^2 + \sigma_c^2$ is given in the data base, p itself is not identifiable. For our Bayesian analysis, we handle this issue by using an informative prior for p . Based on previous considerations, it seems reasonable to balance the classical and Berkson errors, with a substantial fraction being of each type. Thus, we gave p a uniform prior on the interval $[0.2, 0.8]$, creating a form of model mixing.

3.1 Parametric Dose–Response Models

The model used by Stevens et al. (1992) in their dose–response was defined as follows. For numerical convenience, we rescaled dose to be dose/max(dose) = dose/0.461 Gy (Gray). The model is

$$\text{logit} \{ \text{pr}(Y = 1 \mid Z, X) \} = \beta_0 + Z^T \beta_1 + \log(1 + \theta X). \quad (5)$$

Let $\mu_{L|Z}$ and $\sigma_{L|Z}^2$ be the conditional mean and variance of $\log(L)$ given Z . Define $\lambda_{x|w,\ell} = \sigma_{L|Z}^2 / (\sigma_{L|Z}^2 + \sigma_c^2)$. Using standard calculations, assuming that the nonzero L 's are log normal, and making the usual exponential approximation to the logistic function appropriate for rare events, it can be shown that, for the observed data,

$$\begin{aligned} \text{logit} \{ \text{pr}(Y = 1 \mid Z, W) \} \\ \approx \beta_0 + Z^T \beta_1 + \log(1 + \theta \gamma W^{\lambda_{x|w,\ell}}), \\ \gamma = \exp \{ (1 - \lambda_{x|w,\ell}) (\mu_{L|Z} + \sigma_{L|Z}^2 / 2) + \sigma_b^2 / 2 \}. \end{aligned} \quad (6)$$

In the Berkson case, $\lambda_{x|w,\ell} = 1$ and $\gamma = \exp(\sigma_b^2 / 2) > 1$, so that the right-hand side of (6) reduces to $\beta_0 + Z^T \beta_1 + \log(1 + \theta \gamma W)$, meaning that an analysis that ignores Berkson errors overestimates the dose–response parameter by the factor γ , thus falsely inflating the effect of dose. Indeed, when regressing Y against (Z, W) , if one assumes Berkson error, then W should be replaced by γW , where γ varies among individuals; essentially, such an approach was used by Stevens et al. (1992).

3.1.1 MCMC calculations. In terms of our MCMC calculations, we make the following comments. When the measurement error in the dose is incorporated, (X, L) are treated as latent variables, i.e., augmented data, and observations are

sampled from their complete conditionals. Let $f(L \mid Z, \mathcal{A})$ be the density of L depending on Z and the parameter \mathcal{A} . Remember that the data base gives us the value of $\sigma_b^2 + \sigma_c^2$, with the possible unknown being $p = \sigma_b^2 / (\sigma_b^2 + \sigma_c^2)$. Let the prior density be $\pi(\beta_0, \beta_1, \theta, \mathcal{A}, p)$. Then the complete distribution for an observation is written as

$$\begin{aligned} f(Y \mid X, Z, \beta_0, \beta_1, \theta) f(X, W \mid L, \sigma_b^2, \sigma_c^2, p) f(L \mid Z, \mathcal{A}) \\ \times \pi(\beta_0, \beta_1, \theta, \mathcal{A}, p). \end{aligned} \quad (7)$$

In (7), when $\sigma_b^2 + \sigma_c^2 = 0$, we have $L = X = W = 0$.

All priors were chosen to be proper but noninformative, with the exception of that for θ . For θ , the prior was chosen to be normal, truncated at zero, with prior mean being the empirical Bayes estimator ignoring measurement error but with a large variance.

Due to the logistic model framework, the complete conditional distributions for $(\beta_0, \beta_1, \theta, X)$ are nonstandard. We used a Metropolis step to generate observations from these nonstandard distributions. The complete conditionals for L and \mathcal{A} are standard.

3.2 Monotonic, Semiparametric Dose–Response

Here we replace the term $\log(1 + \theta X)$ in (5) by a more flexible semiparametric form, namely

$$\text{logit} \{ \text{pr}(Y = 1 \mid Z, X) \} = \beta_0 + Z^T \beta_1 + g(X). \quad (8)$$

Following (5), in (8) it makes sense to have $g(\cdot)$ be an unknown but strictly monotone function, with the property that $g(0) = 0$. For a general function $g(x)$, modeling in such a circumstance has been considered previously by many authors, e.g., using regression splines. These methods do not guarantee monotonicity of the dose–response. We thus use instead the approach of Mallick and Gelfand (1994), which has three steps: (a) monotonically transform the range of the function to the unit interval, (b) note that then modeling g is equivalent to modeling an unknown distribution function, and (c) model this distribution function as a mixture of beta distribution functions.

Thus, for some function $\mathcal{T}(\cdot)$, we assume that $g(\cdot)$ satisfies

$$\mathcal{T}\{g(x)\} = \sum_{\ell=1}^r \omega_\ell \text{IB}[\mathcal{T}\{g_0(x)\}, c_\ell, d_\ell]. \quad (9)$$

In (9), \mathcal{T} is a monotonic transformation from the real line to $[0, 1]$. In addition, $\text{IB}(u; c, d)$ denotes the incomplete beta function, associated with a beta density in standard form having parameters c and d but evaluated at u . In (9), r denotes the number of mixands and ω_ℓ denotes the mixing weights, with the constraints that $\omega_\ell \geq 0$ and $\sum_{\ell=1}^r \omega_\ell = 1$. Finally, g_0 is a centering function for g . The data will revise g_0 to an estimator of the unknown function g revealing the extent of departure from g_0 . In our application, it is natural to set $g_0(x) = 1 + \theta x$, where in our example θ is the posterior mean estimate of the corresponding parametric model. Because X and hence $G(x)$ are nonnegative, we slightly modify Mallick and Gelfand's suggestion by choosing $\mathcal{T}(v) = v / (1 + v)$.

In viewing g as unknown, we might think of r , $(\omega_1, \dots, \omega_r)$, and the (c_ℓ, d_ℓ) as unknown. In practice, we have found that assuming r is unknown gains little compared with, say, $r = 6$.

Given r , it is mathematically easier to assume that the component beta densities are specified but that the weights are unknown. Following Mallick and Gelfand (1994), we take $c_\ell = \ell$, $d_\ell = r + 1 - \ell$, providing a collection of densities that blanket the unit interval. Hence, specification of g is equivalent to specification of the ω 's. In addition to the constraints that $\omega_\ell \geq 0$ and $\sum_{\ell=1}^r \omega_\ell = 1$, (9) and the condition $g(0) = 0$ implies that $k_1/(1 + k_1) = \sum_{\ell=1}^r \omega_\ell \text{IB}\{k_1/(1 + k_1), c_\ell, d_\ell\}$, i.e., the ω_ℓ satisfy an additional linear constraint.

For the Bayesian analysis, we need to specify a prior distribution for the ω 's, noting this is a distribution on the r -dimensional simplex. We chose for this distribution the Dirichlet($\gamma = 1$) (Berger, 1985, p. 561). The intuition behind this choice is as follows. If g_0 is a baseline function for g , then we might choose $f(\omega)$ such that, *a priori*, g is centered around g_0 . The data would then revise this prior in terms of the support for g_0 . Centering g around g_0 corresponds to centering $\sum_{\ell=1}^r \omega_\ell \text{IB}(u; c_\ell, d_\ell)$ around u . If we center using the mean, as is typically done in the case of Dirichlet processes, we obtain

$$\sum_{\ell=1}^r \omega_\ell \text{E}(w_\ell) \text{IB}(u; c_\ell, d_\ell) = u. \quad (10)$$

Then (9) requires $r^{-1} \sum \text{IB}(u; c_\ell, d_\ell) = u$. If we use c_ℓ and d_ℓ as in previous the paragraph and take r even, expansion of the terms in this summation about $1/2$ yields, to a first order approximation, an average that is u .

4. Intermediate Variable Distribution

We next propose a flexible parametric model for $\log(L)$. There are many ways to specify a flexible, skewed, heavy-tailed distribution for $\log(L)$. Possibilities include the skewed-normal distribution, the mixture of normals distribution, or models such as those used by Davidian and Gallant (1993). These methods are easy to write down, but the MCMC calculations involving them are not entirely straightforward since they require Metropolis steps.

In contrast, our method is to assume $\log(L)$ has an unknown distribution and impose a Pólya-tree prior (Lavine, 1992; Walker and Mallick, 1996). The method allows considerable flexibility in the model for $\log(L)$ as well as great ease of calculation. The flexibility and ease of calculation are bought at the price of difficult notation.

We give here a brief description of the methodology used. Within each state s , we assumed that the distribution function of $\log(L)$ for nonzero doses was $F_s(x - \alpha_{s0} - Z^T \alpha_{s1})$, where $F_s(\cdot)$ is the realization of a random distribution function. The prior for $F_s(\cdot)$ is a Pólya tree distribution, defined as follows.

We start with a base distribution function G , the normal distribution function (with a large standard deviation, in this case 40). We then partition the real line. At stage $m = 1$, the first partition is (B_0, B_1) , where $B_0 = (-\infty, G^{-1}(1/2))$. At stage $m = 2$, we partition B_0 and B_1 separately into (B_{00}, B_{01}) and (B_{10}, B_{11}) , respectively, where $B_{00} = (-\infty, G^{-1}(1/4))$ and $B_{10} = [G^{-1}(1/2), G^{-1}(3/4))$. We continue in this way so that, at stage $m + 1$, we partition B_{i_1, \dots, i_m} into $B_{i_1, \dots, i_m, 0}$ and $B_{i_1, \dots, i_m, 1}$. At any stage m , order the $j = 1, \dots, 2^m$ partitions into B_j^* and note that

$B_j^* = [G^{-1}\{(j-1)/2^m\}, G^{-1}(j/2^m))$. In our calculations, we continued with $m = 1, \dots, M = 8$ levels of partitioning.

The Pólya tree prior for F_s is defined on the sets B_j^* for $j = 1, \dots, 2^M$ as follows. At stage $m = 1$, let C_0 be the realization of a beta random variable with indices (γ_0, ζ_0) . Then $F_s(B_0) = C_0$, and of course $F_s(B_1) = C_1 = 1 - C_0$. At stage $m = 2$, let C_{00} and C_{10} be realizations of beta random variables with indices $(\gamma_{00}, \zeta_{00})$ and $(\gamma_{10}, \zeta_{10})$, respectively. Then $F_s(B_{00}) = C_0 C_{00}$, $F_s(B_{01}) = C_0(1 - C_{00})$, $F_s(B_{10}) = C_1 C_{10}$, $F_s(B_{11}) = C_1(1 - C_{10})$. We continue in this way for $m = 3, \dots, M$, thus defining F_s on the sets B_j^* for $j = 1, \dots, 2^M$. This defines a Pólya tree distribution with partition $\Omega = (B_j^*)_{j=1}^{2^M}$ and parameters $\mathcal{A} = (\gamma_0, \zeta_0, \gamma_{00}, \zeta_{00}, \dots)$, which we denote as $\text{PT}(\Omega, \mathcal{A})$. For our prior, at stage m , we set the γ 's and the ζ 's all equal to $c_{\text{polya}} m^2$, where $c_{\text{polya}} = 0.5$, although we experimented with different values $0.1 \leq c_{\text{polya}} \leq 1.0$ and the results changed hardly at all.

We have now defined the Pólya tree prior for F_s . Given observations L_{is} from state s , the posterior of F_s is also a Pólya tree distribution with the same set partition Ω . The parameters are updated as follows. First, at stage $m = 1$, γ_0 is updated to $\gamma_0 + n_{0s}$, where n_{0s} is the number of L 's in state s that fall into the set B_0 . At stage $m = 2$, γ_{00} and γ_{10} are updated to $\gamma_{00} + n_{00}$ and $\gamma_{10} + n_{10}$, where n_{j0} is the number of L 's in state s that fall in B_{j0} . Further levels of the γ 's are generated in the same way.

In the MCMC calculations, suppose that the complete conditional for F_s is $\text{PT}(\Omega, \mathcal{A}_{s*})$. We generate observations from state s as follows. First generate F_s . In state s , the distribution function for the L 's is $F_s(x - \alpha_{s0} - Z^T \alpha_{s1})$. Observations from this distribution function are easily generated by a Metropolis–Hastings step. Conditioned on F , the regression parameters α_{s0} and α_{s1} are also generated by a Metropolis–Hastings step. See the Appendix for details.

5. Analysis of the Nevada Test-Site Data

5.1 Model Fitting

This section provides our reanalysis of the Nevada test-site data, where we illustrate the methods we have developed. In what follows, we will refer to the parametric dose–response model (5) and the semiparametric dose–response model (8). We will also refer to four error structures: (a) none, i.e., ignoring measurement error; (b) Berkson, i.e., when all measurement error is Berkson; (c) classical, i.e., when all measurement error is classical; and (d) mixture, i.e., when the fraction p of the measurement error variance is Berkson and p is uniformly distributed on the interval $[0.2, 0.8]$. We will also refer to models for the latent intermediate variable L , namely, the parametric normal model (4) and the semiparametric latent intermediate variable model described in Section 4.

In our analyses, the response Y was the period prevalence (1985–1986) of thyroid neoplasms. There were only 19 such neoplasms in the data set, although the effect of dose is statistically significant when ignoring measurement error and performing a likelihood ratio test.

Table 1 gives results for the parametric dose–response model (5) for the cases that measurement error is ignored, is purely Berkson, is purely classical, or is a mixture of

Table 1
 Posterior means and credible sets for the parameter θ in model (5)
 and for the relative risk at true dose 1 Gy (100 rad)

Error model	Latent variable	Posterior mean θ	Lower 95% credible bound	Upper 95% credible bound	RR at dose = 1 Gy	Lower 95% credible bound	Upper 95% credible bound
No error		38.90	16.28	58.98	9.43	4.53	13.79
Classical	Normal	74.06	34.52	108.55	17.06	8.48	24.54
Classical	Semi	68.19	30.91	102.15	15.79	7.70	23.16
Berkson		31.90	13.09	48.00	7.92	3.84	11.41
Mixture	Normal	56.11	18.58	101.98	13.17	5.03	23.12
Mixture	Semi	45.60	12.15	94.99	10.89	2.63	22.60

classical and Berkson error. In the first two cases, no latent intermediate variable model is assumed, while for the other cases, we allow for the parametric or semiparametric latent intermediate variable model. This table gives results both for the estimate of θ as well as for the relative risk at true dose 1 Gy = 100 rad.

Note that, as expected from the theory, ignoring the measurement error leads to a slight overestimate of the dose-response rate as compared with a pure Berkson error analysis. In contrast, if all the measurement error were classical, ignoring measurement error would lead to a substantial underestimate of risk. This is in agreement with the calculations of Stevens et al. (1992). In results not reported here, we computed the maximum likelihood estimate for θ via numerical integration, the estimated value being almost the same as the posterior mean. As might be expected from these considerations, the mixture error model gives risk estimates between the no-error and 100%-classical error estimates.

Figure 1 illustrates the lack of normality of $\log(L)$ in Utah. Specifically, we computed a posterior mean Pólya tree distribution by averaging the MCMC probability values for each partition. We then generated 5000 observations from this posterior mean Pólya tree. As seen in Figure 1, the result is skew, pointing out the need for more flexible latent intermediate variable modeling in the log scale. Coupled with this plot indicating the need for a flexible distribution for the latent intermediate variable, we will present evidence in Section 5.2 in support of the need for a flexible dose-response function.

Table 2 gives the general results and compares the parametric and semiparametric dose-response models (5) and (8). Here we restrict attention to estimating the relative risk at true dose 1 Gy = 100 rad. In our discussion, we specifically want to contrast two analyses: (a) Berkson error model with the dose-response function (5), an analysis fairly close to that done in Stevens et al. (1992), and (b) the mixture of Berkson and classical errors with semiparametric dose-response and latent intermediate variable functions. Note that the latter model suggests a near doubling of the posterior mean relative risk from 7.92 to 14.23.

Perhaps the more interesting result is the comparison between the uncertainties in these posterior means as exhibited through 95% credible intervals. It is well known in measurement error models that correction for measurement error affects both parameter estimation and precision of inference.

In our case, the Berkson error model suggests a lower bound on the relative risk of 3.84, while the mixture semiparametric model suggests a lower bound of 1.68. Corresponding large differences are seen in the upper 95% credible bounds, not too surprising given the extra flexibility in our modeling approach.

5.2 Model Selection

In selecting among the models described in Section 5.1, customary Bayesian model screening selects the model with the largest value of the marginal density of the data evaluated at the observations. In the present case, we will use the deviance information criterion (*DIC*) as in Spiegelhalter, Best, and Carlin (unpublished manuscript) to do this calculation. Let \bar{D} be the posterior expectation of the deviance of the model and P_D be the effective number of parameters in the model, defined as $P_D = \bar{D} - D(\bar{\eta})$, where η contains all the parameters of the model and $\bar{\eta}$ is its posterior expectation. Then $DIC = \bar{D} + P_D$ and can be calculated easily using the MCMC samples and using the sample means of the simulated values of D and the plug-in estimates of the deviance using the sample means of the simulated values of all the parameters η .

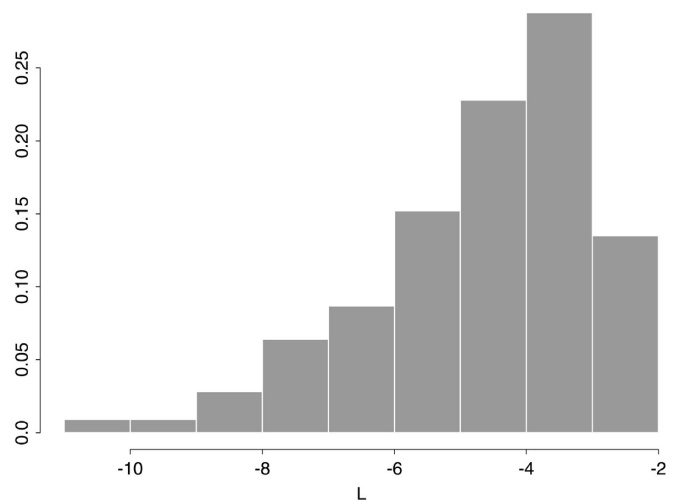


Figure 1. A histogram illustrating the posterior distribution of nonzero values of $\log(L)$ for Utah when modeled by a Pólya tree distribution.

Table 2
 Posterior means and credible sets for the relative risk at true dose 1 Gy (100 rad)

Error model	Regression model	Latent variable model	Relative risk at dose 1 Gy	Lower 95% credible interval	Upper 95% credible interval
No error	Parametric	—	9.43	4.53	13.79
	Semiparametric	—	13.77	2.46	18.91
Berkson	Parametric	—	7.92	3.84	11.41
	Semiparametric	—	9.95	3.10	13.20
Classical	Parametric	Parametric	17.06	8.48	24.54
		Semiparametric	15.79	7.70	23.16
	Semiparametric	Parametric	21.54	9.41	36.66
		Semiparametric	18.98	7.98	32.69
Mixture	Parametric	Parametric	13.17	5.03	23.12
		Semiparametric	10.89	2.63	22.60
	Semiparametric	Parametric	16.42	2.19	34.75
		Semiparametric	14.23	1.68	33.61

For logistic regression with probabilities p_i , the deviance is $D = 2 \sum_i Y_i \log(Y_i/p_i) + (Y_i - 1) \log\{(1 - Y_i)/(1 - p_i)\}$. *DIC* for no dose effect is 229.2. *DIC* for parametric models without error, with Berkson errors, with classical errors, and with mixture errors are 218.5, 214.6, 213.9, and 211.4, respectively. *DIC* for semiparametric models without error, with Berkson errors, with classical errors, and with mixture errors are 216.3, 211.7, 210.4, and 207.2, respectively. This gives some support for the need to use the semiparametric regression model (8) coupled with the semiparametric dose-response model (Section 4).

6. Simulations

We performed a small simulation study to understand the relative performance of our methods. The sample size was the same as in the data set, with two logistic functions in true dose X : (a) $\text{logit}\{\text{pr}(Y = 1 | X)\} = \log(1 + 0.6X)$ and (b) $\text{logit}\{\text{pr}(Y = 1 | X)\} = 2 - 1/(1 + X^2)$. We assumed that half the measurement error was classical and half was Berkson and that the measurement error was relatively large. Specifically, $\log(L) = \text{normal}(-0.3466, 0.8408^2)$, $\log(X) = \log(L) + \text{normal}(0, 0.8408^2)$, and $\log(W) = \log(L) + \text{normal}(0, 0.8408^2)$. These specifications means that $X < 0.10$ with probability 0.05 and $X < 5$ with probability 0.95, not too far from what appears to actually happen in the actual data with dose divided by 0.461Gy. We evaluated the relative risk on the interval 0 to 5. The main difference between the simulation and the data is that the former has many more observations with $Y = 1$.

We generated a single data set. Figures 2 and 3 compare the true relative risk function (thicker solid line), the estimated relative risk when error is ignored (thin solid line), and the fit via our semiparametric dose-response and latent intermediate variable model (dashed line). The figures demonstrate the superiority of our methods in these two cases.

In addition, we computed the *DIC* for these two simulated data sets. As expected, in the first simulation, the parametric model (dose-response and latent intermediate variable) had the lowest *DIC*, while in the second, the semiparametric model had the lowest *DIC*.

7. Discussion

7.1 Summary Comments

We have constructed Bayesian methods for analysis of data when predictors have a combination of classical and Berkson measurement error. We applied our methods to an important data set in radiation epidemiology. The methods allow for a semiparametric yet monotonic regression function along with a semiparametric latent intermediate variable model. The methods are easily extended to any generalized linear model. In our example, the combination of the two semiparametric approaches yielded the smallest deviance information criterion. It also yielded a much larger relative risk at relatively high doses than suggested by a Berkson error model with parametric dose-response function, albeit with much wider uncertainties in the estimate of this relative risk.

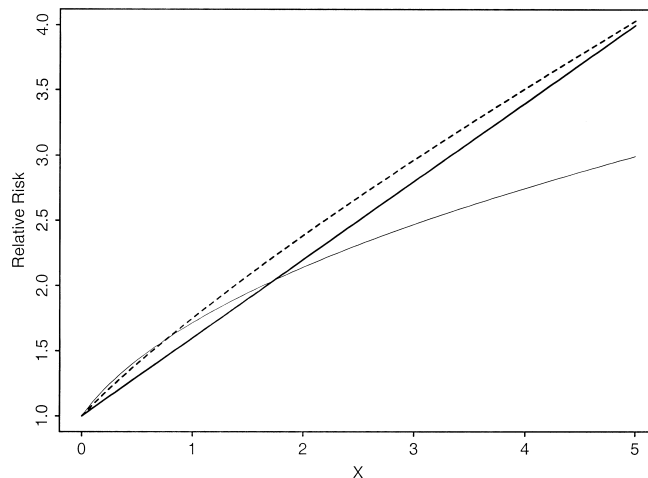


Figure 2. Results for the simulated data set with $\text{logit}\{\text{pr}(Y = 1 | X)\} = \log(1 + 0.6X)$. The true relative risk is the thicker solid line, the estimated relative risk ignoring measurement error is the thin solid line, and our semiparametric estimate is the dashed line.

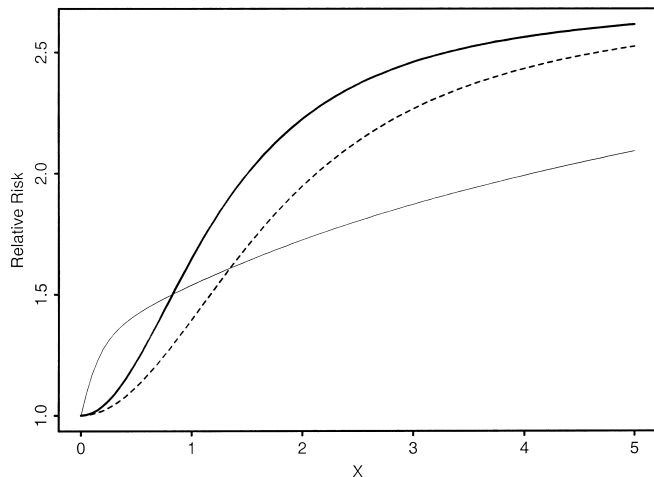


Figure 3. Results for the simulated data set with $\text{logit}\{\text{pr}(Y = 1 | X)\} = 2 - 1./\{1 + (X^2)\}$. The true relative risk is the thicker solid line, the estimated relative risk ignoring measurement error is the thin solid line, and our semiparametric estimate is the dashed line.

In our example, the total error variance, i.e., the sum of the Berkson and classical error variances, was assumed to be known for each individual but the individual error variances was unknown. In the context of the example, it was impractical for us to redo the dosimetry construction and to thus untangle the relative contributions to the total error variance. Our solution to this dilemma was to assume that a fraction p of each variance was of Berkson type, and we placed a uniform distribution on p within a well-defined interval that is reasonable in the context of the example. Of course, if the dosimetry could have been redone, then this information could be incorporated naturally into the Bayesian framework.

The preceding paragraph makes clear, and it is worth reemphasizing, that we have been forced to assume that the data set accurately specified the total error variance. This is clearly a major limitation of any analysis of dose uncertainties. It is also probable that the percentage p of total error that is classical varies from individual to individual; we have chosen to make p fixed across individuals, although at least in principle we could have allowed it to vary according to some specified distribution. These two pieces of unavoidable roughness in the data base means that the Nevada test-site data should best be thought of as an illustration of the general methodology.

In other examples, the total error variance may not be known. Our methods are, in principle, easily extended to this case, although issues of identifiability become more complex.

Whenever there is a component of classical error, a distribution for a latent intermediate variable must be specified. At least in principle, it is possible that misspecifying the distribution of the latent intermediate variable could cause biases in regression modeling. We assumed separate regression models for a categorical variable and considered both fully parametric and flexible semiparametric distributions, the latter based on the Pólya tree distribution. Clearly, there is nothing magical about the Pólya tree distribution, and other flexible semiparametric distributions could be used.

We also examined the form of the dose–response function, allowing the dose to be modeled semiparametrically. In the context of the example, it was reasonable to assume a monotonic function, and our semiparametric approach incorporates the monotonicity naturally.

7.2 Shared Uncertainties

Finally, we comment on our assumption that the Berkson and classical errors are independent across individuals. This is almost certainly not the case, and thus our data analysis may thus be best thought of as an illustration of methodology. The radioiodine concentration of milk, C , in (1) includes the deposition of I^{131} by region, its mass interception on vegetation, the effective half-life of I^{131} in the vegetation, the consumption of vegetation by cows, and the milk transfer coefficient (abbreviated here as MTC). While similar issues apply to the mass interception and the dose conversion factor, consider for example the MTC for a child in a particular region whose milk comes from a backyard cow: the problem we now discuss is probably even greater for children consuming milk from commercial dairies. As we understand it, as part of the modeling process, the Utah study generated a distribution for the MTCs as log-normally distributed with mean μ_{MTC} and variance σ_{MTC}^2 . If these parameters were known, then the error structure for the MTC would be primarily Berkson. However, these parameters are not known and are instead estimated by a combination of historical data and literature review. This means that the error in estimating the coefficients is the same, hence shared, by all the children in the region with a backyard cow.

Understanding how such shared uncertainties affect parameter estimation and inference is an open problem worth considerable study. We have performed one preliminary calculation. We consider the parametric dose–response model, the parametric (normal) latent intermediate variable (L) model, and the mixture of Berkson and classical error structure. We allowed for shared uncertainties in the Berkson error model (2). Specifically, for the six groups formed by the combinations of states and genders, the Berkson errors for individuals within each group were assumed to have common correlation ρ , which we varied from 0.0, 0.2, 0.4, and 0.6. The posterior mean estimates of the parameter θ for each of these situations were 56.11, 65.85, 84.12, and 95.06, respectively. The credible intervals were (18.58, 101.98), (21.44, 120.21), (30.41, 142.95), and (38.23, 151.69), respectively. The fairly large changes in parameter estimates and credible intervals suggest the need in future for data to be gathered that can account for the possibility of shared uncertainties.

ACKNOWLEDGEMENTS

This research was supported by a grant from the National Cancer Institute (CA-57030) and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106). We are deeply grateful to Duncan Thomas and Richard Kerber and to a referee for many useful suggestions.

RÉSUMÉ

Nous construisons des méthodes Bayésiennes pour une modélisation semi-paramétrique d'une fonction de régression

monotone quand les prédicteurs sont mesurés avec une erreur classique, une erreur de Berkson ou les deux. De telles méthodes demandent une distribution pour le prédicteur non observé (variable latente), distribution que nous modélisons aussi de façon semi-paramétrique. De telles combinaisons de méthodes semi-paramétriques pour la réponse à des doses aussi bien que pour la distribution de la variable latente n'ont pas été étudiées dans la littérature sur les erreurs de mesure, quelle que soit la forme de l'erreur de mesure. De plus, nos méthodes proposent une nouvelle approche des problèmes où l'erreur de mesure combine une composante de Berkson et une composante classique. Ces méthodes sont générales mais nous les développons autour d'une application particulière qui est l'étude des maladies de la thyroïde en relation avec les radiations venant du site de test du Nevada. Nous utilisons ces données pour illustrer nos méthodes, lesquelles suggèrent une estimation ponctuelle (moyenne a posteriori) du risque relatif à des doses près de deux fois plus élevée que celle obtenue par les analyses précédentes, mais suggère aussi une bien plus grande incertitude sur le risque relatif.

REFERENCES

- Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd edition. New York: Springer.
- Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*. London: Chapman and Hall.
- Carroll, R. J., Maca, J. D., and Ruppert, D. (1999). Nonparametric regression with errors in covariates. *Biometrika* **86**, 541–554.
- Carroll, R. J., Roeder, K., and Wasserman, L. (1999). Flexible parametric measurement error models. *Biometrics* **55**, 44–54.
- Davidian, M. and Gallant, A. R. (1993). The nonlinear mixed effects model with a smooth random effects density. *Biometrika* **80**, 475–488.
- Davis, S., Kopecky, K. J., Hamilton, T. E., and Amundson, B. (1998). *Hanford Thyroid Disease Study Draft Final Report*. Fred Hutchinson Cancer Research Center, Seattle, Washington.
- Kerber, R. L., Till, J. E., Simon, S. L., Lyon, J. L., Thomas, D. C., Preston-Martin, S., Rollison, M. L., Lloyd, R. D., and Stevens, W. (1993). A cohort study of thyroid disease in relation to fallout from nuclear weapons testing. *Journal of the American Medical Association* **270**, 2076–2083.
- Lavine, M. (1992). Some aspects of Pólya tree distributions for statistical modeling. *The Annals of Statistics* **20**, 1222–1235.
- Mallick, B. K. and Gelfand, A. E. (1994). Generalized linear models with unknown link functions. *Biometrika* **81**, 237–245.
- Müller, P. and Roeder, K. (1997). A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika* **84**, 523–537.
- Ostrouchov, G., Frome, E. L., and Kerr, G. D. (1998). *Dose estimation from daily and weekly dosimetry data: Final draft*. Technical Report. Health Sciences Research Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee.
- Reeves, G. K., Cox, D. R., Darby, S. C., and Whitley, E. (1998). Some aspects of measurement error in explanatory variables for continuous and binary regression models. *Statistics in Medicine* **17**, 2157–2177.
- Roeder, K., Carroll, R. J., and Lindsay, B. G. (1996). A nonparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association* **91**, 722–732.
- Schafer, D. W. (2001). Semiparametric maximum likelihood for measurement error model regression. *Biometrics* **57**, 53–61.
- Schafer, D. W., Lubin, J. H., Ron, E., Stovall, M., and Carroll, R. J. (2002). Thyroid cancer following scalp irradiation: A reanalysis accounting for uncertainty in dosimetry. *Biometrics* **57**, 689–697.
- Simon, S. L., Till, J. E., Lloyd, R. D., Kerber, R. L., Thomas, D. C., Preston-Martin, S., Lyon, J. L., and Stevens, W. (1995). The Utah Leukemia case-control study: Dosimetry methodology and results. *Health Physics* **68**, 460–471.
- Stevens, W., Till, J. E., Thomas, D. C., et al. (1992). *Assessment of leukemia and thyroid disease in relation to fallout in Utah: Report of a cohort study of thyroid disease and radioactive fallout from the Nevada test site*. Technical Report. University of Utah, Salt Lake City.
- Walker, S. and Mallick, B. K. (1999). Semiparametric accelerated life time models. *Biometrics* **55**, 477–483.

Received April 2001. Revised August 2001.

Accepted August 2001.

APPENDIX

MCMC Details

The prior for the β 's were independent normals with mean zero and variance 1000. The prior for θ was a truncated normal with mean 40 and variance 100. The prior for p was uniform[0.2, 0.8]. The prior for the ω 's was Dirichlet($\gamma = 1$) (Berger, 1985, p. 561). The priors for the state-level parameters (α_{s0}, α_{s1}) was independent normals with mean zero and variance 100. The Pólya tree prior is as specified in Section 4.

The Metropolis proposals were as follows. The subscript "old" means the current values of the parameters. For the β 's, the Metropolis proposals were normal($\beta_{old}, 1.0$, and similarly for the θ 's. For the ω 's, the following considerations apply. To accommodate the constraint to the simplex, it is rescaled to $r - 1$ dimensional Euclidean space using a logit transformation. Suppressing subscripts, let $z_\ell = \log(\omega_\ell)$. The Jacobian from the ω -space to the z -space is $\prod_{\ell=1}^r \omega_\ell$, whence the complete conditional distribution for z , up to a constant of proportionality, is readily obtained. A normal proposal density with mean the current value and standard deviation 0.5 is used for z , and starting values $\omega_\ell = 1/r$ worked well. The proposals for the state-level variables α_{s0}, α_{s1} were independent normals with the current values as the mean and standard deviation 0.5. The prior is used as the proposal for p .