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∗∗∗

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 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, it is reasonable to suppose that the proportion 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 mixtures-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 Polya 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 (Osborn, Froome, 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, sex, 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 molespans developed was obtained when fitting a logistic regression model with stratum-specific intercepts, adjustments for confounders, and a term for dose of the form $\log(1+\text{dose})$ (see below for more details). In one such analysis (Stevens et al., p. 208), the estimate of $p$ more than doubled after accounting for dose uncertainty by assuming a classical error model, i.e., if all the error is classical and error is ignored, relative risks are underestimated. In Section 3.1, we show that assuming 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):

$$W = C \times DCF \times I \times TD \times FP, \tag{1}$$

where $W$ = reported dose to thyroid of the subject; $C$ = time-integrated radiiodine 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 radiiodine concentration of milk ($C$) is specific 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 $^{131}$I (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 intercept of $^{131}$I 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

$$\log(X) = \log(L) + U_3. \tag{2}$$

$$\log(W) = \log(L) + U_4. \tag{3}$$

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 $U_3$ is Berkson error with variance $\sigma_3^2$, $U_4$ is classical error with variance $\sigma_4^2$, and $\log(L)$ has mean $\mu_L$ and variance $\sigma_L^2$. 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., $\mu_L$ may be allowed to depend on a linear function of $Z$ and $\sigma_L^2$ 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
Semicparametric Regression Modeling with Error

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, 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 (Uda, Nevada, Arizona). In so a,

\[
\log(\lambda(L) | Z, \text{state} = s) \sim \text{normal}(\alpha_s + Z^T \alpha_1, \sigma_1^2).
\]

We denote by \( A \) 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 and classical errors, with the fraction of variance due to the Berkson part being \( \sigma_2^2 / \sigma_1^2 \) 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 latent variables, i.e., augmented data, and observations are identifiable. For our Bayesian analysis, we make the following comments. When the measure-}

3.1.1 MCMC calculations. In terms of our MCMC calcula-

3.1.2 Parametric Dose–Response Models

The model used by Stevens et al. (1992) in their dose–response}

3.2 Monotonic, Semicparametric Dose–Responses

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

\[
\logit(p(Y = 1 | Z, X)) = \beta_1 + Z^T \beta_0 + g(X).
\]

Following (5), in (8) it makes sense to have \( g(0) \) is a centering function for \( r \) and \( \sigma \), making the usual exponential approximation to the logistic function appropriate for rare events, it can be shown that, for the observed data,

\[
\logit(p(Y = 1 | Z, X)) \approx \beta_0 + Z^T \beta_1 + \log(1 + \exp(\gamma - \gamma_1)),
\]

where \( \gamma = \exp((1 - b X)/c + \sigma_2^2/2 + \sigma_2^2/2) \).

In the Berkson case, \( b_0 = -1 \) and \( \gamma = \exp(\sigma_2^2/2) > 1 \), so that the right-hand side of (6) reduces to \( \beta_0 + Z^T \beta_1 + \log(1 + \exp(W)) \), meaning that an analysis that ignores Berkson errors overestimates the dose–response parameter by the factor \( \gamma \), 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 \( W \), where \( \gamma \) varies among individuals; essentially, such an approach was used by Stevens et al. (1992).

In (9), \( T \) is a monotonic transformation from the real line to \([0,1]\). In addition, \( \mathcal{H}(v, c, d) \) denotes the incomplete beta function, associated with a beta density in standard form having parameters \( c, d \) and \( b \) evaluated at \( v \). In (9), \( \rho \) denotes the number of mixands and \( \omega \) denotes the mixing weights, with the constraints that \( \omega_1 \geq 0 \) and \( \sum_{i=1}^{\rho} \omega_i = 1 \). Finally, \( \gamma_0 \) is a centering function for \( g \). The data will revise \( \gamma_0 \) to an estimate of the unknown function \( g \) revealing the extent of departure from \( g \). In our application, it is natural to set \( g(0) = 1 + \gamma_0 \), where in our example \( \gamma_0 \) 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 \( T(v) = \sqrt{1 + v} \).

In viewing \( g \) as unknown, we might think of \( \omega_1, \ldots, \omega_{\rho} \); the \( (c_i, d_i) \) 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
\[
  \epsilon_{x_{1}} = \epsilon_{x}, \quad \epsilon_{x_{1}(1)} = r - 1 - \epsilon, \text{ providing a collection of densities that blanket the unit interval. Hence, specification of } g \text{ is equivalent to specification of the } \omega \text{'s}. \]
In addition to the constraints that \( \omega_{x} \geq 0 \) and \( \Sigma_{1, x_{1}} \omega_{x} = 1, \) (9) and the condition \( y(0) = 0 \) implies that \( k_{1}(1 + k_{2}) = \Sigma_{1, x_{1}} \omega_{x} = 1 \), i.e., the \( \omega \) satisfy an additional linear constraint.

For the Bayesian analysis, we need to specify a prior distribution for the \( \omega \)'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 \( f(s) \) such that, a posteriori, \( g \) 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 \( \Sigma_{1, x_{1}} \omega_{x}(c_{1}, d_{1}) \) around \( u \). If we center using the mean, as is typically done in the case of Dirichlet processes, we obtain
\[
  \sum_{j=1}^{r} E(x_{1} | B_{u j}(c_{1}, d_{1})) = u. \tag{10}
\]
Then, (9) requires \( r \Sigma_{1} \Sigma_{2} + \Sigma_{2}^{2} = n \). If we use \( c_{1} \) and \( d_{1} \) 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 \( log(Y) \).

There are many ways to specify a flexible, skewed, heavy-tailed distribution for \( log(Y) \). Possibilities include the skewed normal distribution, the mixed versions of normal distributions, 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 \( log(Y) \) 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 \( log(Y) \) as well as greater 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 \( \log(Y) \) for numerous doses was \( F_{s}(x - \alpha_{s})^{2} + \beta_{s} \), where \( F_{s}(.) \) is the realization of a random distribution function. The prior for \( F_{s}(.) \) 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 \( s = 1 \), the first partition is \( (B_{00}, B_{11}) \), where \( B_{00} = (-\infty, G^{-1}(1/2)) \). At stage \( s = 2 \), we partition \( B_{00} \) and \( B_{11} \) separately into \( (B_{00}, B_{11}) \) and \( (B_{01}, B_{10}) \), respectively, where \( B_{00} = G^{-1}(1/4), B_{10} = G^{-1}(1/2), G^{-1}(3/4). \)

We continue in this way so that, at stage \( s = 1 \), we partition \( B_{00}, \ldots, B_{00} \) into \( B_{00}, \ldots, B_{00} \) and \( B_{10}, \ldots, B_{10} \). At any stage \( m \), order the \( j = 1, \ldots, 2^{m} \) partitions into \( B_{j}^{*} \) and note that
\[
  B_{j}^{*} = \left( G^{-1}\left\{1 - \left(1 - \frac{2^{i}}{2^{m}}\right)^{2}\right\}\right)^{2} \tag{11}
\]
In our calculations, we continued with \( m = 1, \ldots, M = 8 \) levels of partitioning.

The Pólya tree prior for \( F_{s} \) is defined on the sets \( B_{j}^{*} \) for \( j = 1, \ldots, 2^{M} \). At stage \( m = 1 \), this is the realization of a beta random variable with indices \( (\gamma_{0}, \gamma_{0}) \). Then \( F_{s}(B_{00}) = C_{00} \), and of course \( F_{s}(B_{11}) = C_{11} = 1 - C_{00} \). At stage \( m = 2 \), we perform realizations of\( \beta \) random variables with indices \( (\gamma_{0}, \gamma_{0}) \) and \( (\gamma_{0}, \gamma_{0}) \), respectively. Then \( F_{s}(B_{01}) = C_{01}C_{10}, F_{s}(B_{1}) = C_{10}(1 - C_{01}) \). We continue in this way for \( m = 3, \ldots, M \), thus defining \( F_{s} \) on the sets \( B_{j}^{*} \) for \( j = 1, \ldots, 2^{M} \). This defines a Pólya tree distribution with partition \( H = \{B_{j}^{*}\}^{2^{M}} \) and parameters \( \gamma = \{C_{00}, C_{10}, C_{11}, (\gamma_{0}, \gamma_{0}), (\gamma_{0}, \gamma_{0})\} \), which we denote as \( PT(\Omega, A) \). For our prior, at stage \( m, \) we set \( \gamma \) and \( \gamma \) as equal to \( \gamma_{0}, \gamma_{0} \) and \( \gamma_{0}, \gamma_{0} \), although we experimented with different values \( 0 \leq \gamma_{0}, \gamma_{0} \leq 0.5 \) and the results changed hardly at all.

We have now defined the Pólya tree prior for \( F_{s} \). Given observations \( L_{s} \) from state \( s \), the posterior of \( F_{s} \) is also a Pólya tree distribution with the same set partition \( \Omega \).

The parameters are updated as follows. First, at stage \( s = 1 \), \( \gamma_{0} \) is updated to \( \gamma_{0} + \frac{L_{s}}{n_{0}}, \) where \( n_{0} \) is the number of \( L_{s} \) in state \( s \) that fall into the set \( B_{00} \). At stage \( m = 2, \gamma_{0} \) and \( \gamma_{0} \) are updated to \( \gamma_{0} + \frac{L_{s}}{n_{0}} \) and \( \gamma_{0} + \frac{L_{s}}{n_{0}} \), where \( n_{0} \) is the number of \( L_{s} \) in state \( s \) that fall in \( B_{01} \). Further levels of the \( \gamma \)'s are generated in the same way.

In the MCMC calculations, suppose that the complete conditional for \( F_{s} \) is \( PT(\Omega, A_{s}) \). We generate observations from state \( s \) as follows. First generate \( F_{s} \). In state \( s \), the distribution function for the \( L_{s} \) is \( F_{s}(x - \alpha_{s} / 2^{3}, \alpha_{s}) \). Observations from this distribution function are easily generated by a Metropolis–Hastings step. Conditioned on \( F_{s} \), the regression parameters \( \beta_{0} \) and \( \alpha_{s} \) are 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 shall 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 is classical; and (d) mixture, i.e., when the fraction \( p \) of the measurement error variance is Berkson and \( 1 - p \) is uniformly distributed on the interval \([0, 0.5] \). We shall also refer to models for the latent intermediate variable \( L_{s} \), 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 18 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 $\theta$ in model (5) and for the relative risk at true dose $1$ Gy (100 rad)

<table>
<thead>
<tr>
<th>Error model</th>
<th>Latent variable</th>
<th>Posterior mean $\theta$</th>
<th>Lower 95% credible bound</th>
<th>Upper 95% credible bound</th>
<th>Lower 95% credible bound at dose $= 1$ Gy</th>
<th>Upper 95% credible bound at dose $= 1$ Gy</th>
</tr>
</thead>
<tbody>
<tr>
<td>No error</td>
<td></td>
<td>38.90</td>
<td>16.28</td>
<td>58.98</td>
<td>9.43</td>
<td>4.53</td>
</tr>
<tr>
<td>Classical</td>
<td>Normal</td>
<td>74.06</td>
<td>34.52</td>
<td>108.55</td>
<td>17.06</td>
<td>8.48</td>
</tr>
<tr>
<td>Classical</td>
<td>Semi</td>
<td>68.19</td>
<td>30.91</td>
<td>102.15</td>
<td>15.79</td>
<td>7.70</td>
</tr>
<tr>
<td>Berkson</td>
<td>Normal</td>
<td>51.90</td>
<td>18.58</td>
<td>94.99</td>
<td>10.89</td>
<td>2.63</td>
</tr>
<tr>
<td>Mixeure</td>
<td>Semi</td>
<td>45.60</td>
<td>12.15</td>
<td>94.99</td>
<td>10.89</td>
<td>2.63</td>
</tr>
</tbody>
</table>

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 $\theta$ 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 7.84, 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 $\hat{D}$ be the posterior expectation of the deviance of the model and $P_D$ be the effective number of parameters in the model, defined as $P_D = \hat{D} - \hat{D}(\eta)$, where $\eta$ contains all the parameters of the model and $\hat{\eta}$ is its posterior expectation. Then $\text{DIC} = \hat{D} + P_D$ 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 $\eta$.
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 logit(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.
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 radiiodine 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 \(\rho_{MTC}\) and variance \(\sigma^2_{MTC}\). 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 combination \(\theta\) for each of these situations were 56.11, 65.85, 84.12, and 95.06, respectively. The credible intervals were (18.59, 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-57038) and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09186). 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 ainsi 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 pour l'erreur de mesure plus élevée que celle obtenue par les analyses précédentes, mais suggère aussi une bien plus grande incertitude sur le risque relatif.

Références


