Bayesian Smoothing and Regression Splines for Measurement Error Problems

Scott M. BERRY, Raymond J. CARROLL, and David RUPPERT

In the presence of covariate measurement error, estimating a regression function nonparametrically is extremely difficult, the problem being related to deconvolution. Various frequentist approaches exist for this problem, but to date there has been no Bayesian treatment. In this article we describe Bayesian approaches to modeling a flexible regression function when the predictor variable is measured with error. The regression function is modeled with smoothing splines and regression P-splines. Two methods are described for exploration of the posterior. The first, called the iterative conditional modes (ICM), is only partially Bayesian. ICM uses a componentwise maximization routine to find the mode of the posterior. It also serves to create starting values for the second method, which is fully Bayesian and uses Markov chain Monte Carlo (MCMC) techniques to generate observations from the joint posterior distribution. Use of the MCMC approach has the advantage that interval estimates that directly model and adjust for the measurement error are easily calculated. We provide simulations with several nonlinear regression functions and provide an illustrative example. Our simulations indicate that the *frequentist* mean squared error properties of the fully Bayesian method are better than those of ICM and also of previously proposed frequentist methods, at least in the examples that we have studied.

KEY WORDS: Bayesian methods; Efficiency; Errors in variables; Functional method; Generalized linear models; Kernel regression; Measurement error; Nonparametric regression; P-splines; Regression splines; SIMEX method; Smoothing splines; Structural modeling.

1. INTRODUCTION

In this article we present a fully Bayesian approach to the problem of nonparametric regression when the independent variables are measured with error. This is known to be an extremely difficult problem in terms of global rates of convergence. Fan and Truong (1993) showed that for additive normally distributed measurement error, the optimal rate of convergence is $\{\log(n)\}^2$ for a globally consistent estimator when making no assumptions other than the existence of two continuous derivatives. They constructed an estimator using kernel methods that achieve this very slow rate of convergence.

Carroll, Maca, and Ruppert (1999) relaxed the assumption of global consistency. They suggested two estimators: (a) a semiparametric estimator based on the SIMEX method of Cook and Stefanski (1994), which makes no assumptions about the unknown and unobserved covariates, and (b) a more parametric estimator that assumes that the unobserved covariates follow a mixture of normals distribution. Their methods are based on fixed-knot regression splines (or polynomial splines; see Eilers and Marx 1996; Ruppert and Carroll 2000; Sec. 2.2). In simulations, they showed that their methods are generally far superior to those of Fan and Truong (1993), with only moderate bias and smaller variability. The more parametric model tended to have by far the best performance in their simulation study.

There is at least one major difficulty with the efficient but more parametric approach of Carroll et al (1999). Let X be the true covariate and let W be the measured covariate. Basically, they propose fitting a modified polynomial spline. They start with the power function basis in X described in Section 2.2. Then the modified polynomial spline has as its basis functions the regression of the true basis functions in X on the observed covariate W. The resulting modified polynomial basis functions of W are very highly correlated. Their method, like ours and any other nonparametric method, requires the choice of smoothing parameters. As described by them, standard approaches to smoothing parameter estimation, such as generalized cross-validation (GCV) cannot be used, because GCV occasionally does no smoothing. This matters because in such cases their formulation leads to unusually great instability of function estimation. They developed two ad hoc methods for handling this problem: put a positive lower bound on the smoothing parameter, and use an entirely different method based on estimating the mean squared error. However, they gave evidence that shows nonetheless that their method remains numerically unstable if there are more than 15 knots.

One way to deal with this numerical instability is to use a different set of basis functions in X that are nearly orthogonal, for example, the B-spline basis. One would then conjecture that when the B-spline basis functions in X are regressed on the observed covariate W, they will not be highly correlated, and the method of Carroll et al. (1999), will be more stable. In practice, this conjecture remains to be proven.

Rather than trying to tweak the method of Carroll et al. in this way, we set out to do something radically different. Specifically, we conjectured that a fully Bayesian approach had the potential to achieve large gains in efficiency of estimation compared to previously proposed methods. One additional assumption is necessary—namely, the error distribution of the response Y about its mean was specified up to parameters.

In this article we propose a new method for nonparametric function estimation when the covariate is measured with error. Our procedure can be looked at as the natural fully Bayesian extension of the techniques of Carroll et al. (1999). It can also

Scott M. Berry is Statistical Scientist, Berry Consultants, Sycamore, IL 60178 (E-mail:*scott@berryconsultants.com*). Raymond J. Carroll is Distinguished Professor, Department of Statistics, Texas A&M University, College Station TX 77843 (E-mail: *carroll@stat.tamu.edu*). David Ruppert is Professor, School of Operations Research & Industrial Engineering, Cornell University, Ithaca, New York 14853 (E-mail: *davidr@orie.cornell.edu*). Carroll's research was supported by National Cancer Institute grant CA-57030, and by the Texas A&M Center for Environmental and Rural Health via National Institute of Environmental Health Sciences grant P30-ES09106. Ruppert's research was supported by National Science Foundation grant DMS-9804058. The authors thank the associate editor and two referees for many helpful comments.

be viewed as the extension to measurement error models of the Markov chain Monte Carlo (MCMC) technique of Hastie and Tibshirani (1998) or, viewed more broadly, the entire Bayesian formulation of smoothing splines (e.g., Wahba 1978, 1983; Nychka 1988, 1990).

The methodology that we present is new in two respects. First, the adjustment for bias due to measurement error comes automatically from the Bayesian machinery. In contrast, other methods explicitly analyze the bias and devise a correction in a more ad hoc fashion. Second, and perhaps more importantly, the smoothing parameter selector, which also comes automatically from the Bayesian approach, is designed for the measurement error problem. Earlier work either did not propose a smoothing parameter selector (Fan and Truong 1993) or applied a smoothing parameter selector that ignores the effects of measurement error. However, measurement error has large effects on both bias and variance, and a smoothing parameter that is optimal for correctly measured covariates may be far from optimal in the presence of measurement error.

In Section 2 we describe some background information on smoothing and regression P-splines that is necessary for our development. In Section 3 we present our methodology. Two approaches are used. One is straightforward from a calculation standpoint, but estimates only the conditional mode. The second method uses the fully Bayes approach and finds the entire posterior distribution. In Section 4 we presents simulations of these two algorithms. The results indicate that even as a frequentist estimator, our fully Bayesian method is at least competitive with that of Carroll et al. (1999), and sometimes is much better. In Section 5 we provide an illustrative example and in Section 6 we present a discussion of the results.

2. SMOOTHING AND REGRESSION P-SPLINES

Here we present a brief introduction to smoothing and P-splines. For additional information, see the work of Wahba (1978, 1990), Green and Silverman (1994), Hastie and Tibshirani (1998), and Eubank (1999) on smoothing splines and Eilers and Marx (1996) and Ruppert and Carroll (2000) on P-splines.

2.1 Smoothing Splines

Assume that $Y_i = m(X_i) + \epsilon_i$, where ϵ_i has mean 0 and variance σ_ϵ^2. Let $[a, b]$ be the interval for which an estimate of m is sought. Let g be the best natural cubic spline (NCS) approximator of m, that is, the NCS that minimizes $\sum_{i=1}^n \{m(X_i) - g(X_i)\}^2$. If m is smooth, then the error in approximating m by g typically is negligible compared to the estimation error, so we assume that $m = g$.

A *smoothing spline* is defined as the minimizer over g of the penalized sum of squares,

$$S(g) = \sum_{i=1}^{n} \{Y_i - g(X_i)\}^2 + \alpha \int_a^b \{g''(x)\}^2 \, dx, \tag{1}$$

for $\alpha > 0$. This minimizer is a NCS with knots at the distinct X_i values. The integral term of (1) is a roughness penalty, and α is the smoothing parameter. Let $\mathbf{g} = \{g(X_1), g(X_2), \ldots, g(X_n)\}^\mathsf{T}$. The penalty term can be written as $\alpha \int_a^b \{g''(x)\}^2 \, dx = \alpha \mathbf{g}^\mathsf{T} \mathbf{K} \mathbf{g}$, where $\mathbf{K}$ is an $n \times n$-dimensional matrix of rank $n-2$, defined by Eubank (1999). The smoothing spline minimizing $S(g)$ is $\hat{\mathbf{g}} = \mathbf{A}(\alpha)\mathbf{Y}$, where $\mathbf{A}(\alpha) = (\mathbf{I} + \alpha \mathbf{K})^{-1}$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)^\mathsf{T}$. The vector $\hat{\mathbf{g}}$ uniquely defines the smoothing spline.

The Bayesian approach to smoothing splines gives the vector $\mathbf{g}$ a prior density proportional to the "partially improper" Gaussian process,

$$(\alpha / 2\sigma_\epsilon^2)^{M/2} \exp\left\{ -(\alpha/\sigma_\epsilon^2) \mathbf{g}^\mathsf{T} \mathbf{K} \mathbf{g} \right\}, \tag{2}$$

where $M = n - 2$ and $\mathbf{K}$ is defined as before. Although both $\mathbf{K}$ and $\mathbf{g}$ depend on the knot locations, because $\mathbf{g}'\mathbf{K}\mathbf{g} = \int_a^b \{g''(x)\}^2 \, dx$, this prior is independent of the knot locations. If the observations Y_i are independent and normally distributed with mean $g(X_i)$ and variance σ_ϵ^2, then the posterior distribution for $\mathbf{g}$ is multivariate normal with mean $\hat{\mathbf{g}} = \mathbf{A}(\alpha)\mathbf{y}$ and covariance matrix $\sigma_\epsilon^2 \mathbf{A}(\alpha)$.

2.2 Regression P-Splines

Smoothing splines become less practical when n is large, because they use n knots. A more general approach to spline fitting is *penalized splines*, or simply *P-splines*, a term borrowed from Eilers and Marx (1996). Let $\mathbf{B}(x) = \{B_1(x), \ldots, B_N(x)\}^\mathsf{T}$, $N \le n$ be a spline basis. The P-spline model specifies that for some N-dimensional $\boldsymbol{\beta}$, $g(x) := \mathbf{B}(x)^\mathsf{T}\boldsymbol{\beta}$. Let $\mathbf{D}$ be a fixed, symmetric, positive semidefinite $N \times N$ matrix and let α be a smoothing parameter. The penalized least squares estimator $\hat{\boldsymbol{\beta}}(\alpha)$ minimizes

$$\sum_{i=1}^{n} \left\{ Y_i - \mathbf{B}(X_i)^\mathsf{T} \boldsymbol{\beta} \right\}^2 + \alpha \boldsymbol{\beta}^\mathsf{T} \mathbf{D} \boldsymbol{\beta}.$$

Let $\mathcal{B}$ be the $n \times N$ matrix with ith row equal to $\mathbf{B}(X_i)^\mathsf{T}$. Then the penalized least squares estimator is

$$\hat{\boldsymbol{\beta}}(\alpha) = (\mathcal{B}^\mathsf{T}\mathcal{B} + \alpha \mathbf{D})^{-1} \mathcal{B}^\mathsf{T} \mathbf{Y}.$$

The choice of k has been discussed by Ruppert (2000) who found that the exact value of k is not important, provided that k is at least a certain minimum value. Generally, $k = 20$ suffices for the types of regression functions found in practice, and $k = 40$ provides a margin of safety. Of course, there will be exceptions where more knots are required, for example, a long periodic time series. Also, functions whose higher derivatives are large may not be well approximated by, say, quadratic splines; see Figure 3.

Here we use the term "P-splines" to refer to both P-splines and smoothing splines as a special case. Convenient classes of P-splines are the penalized B-splines of Eilers and Marx (1996) and the closely related splines of Ruppert and Carroll (2000). The latter are the pth-degree polynomial splines with k fixed knots, $t_1, \ldots, t_k$. The knots could be equally spaced on the range of the X_i's, although we prefer to select them at the quantiles of the X's. A convenient basis is $\mathbf{B}(x) = (1, x, x^2, \ldots, x^p, (x - t_1)_+^p, \ldots, (x - t_k)_+^p)^\mathsf{T}$. Then $\beta_{2+p}, \ldots, \beta_N$ are the sizes of the jumps in the pth derivative of $g(x) = \mathbf{B}(x)^\mathsf{T}\boldsymbol{\beta}$ at the knots. Ruppert and Carroll (2000) penalize these jumps by letting $\mathbf{D}$ be the $N \times N$ diagonal matrix with $p+1$ 0's followed by k 1's along the diagonal. Then $\boldsymbol{\beta}^\mathsf{T}\mathbf{D}\boldsymbol{\beta} = \sum_{j=1}^k \beta_{1+p+j}^2$ is the sum of the squared jumps.

2.2.1 Bayesian P-Splines. We partition $\boldsymbol{\beta}$ into the coefficients of the monomial basis functions of the truncated power basis functions by letting $\boldsymbol{\beta}^{\mathsf{T}} = (\boldsymbol{\beta}_1^{\mathsf{T}}, \boldsymbol{\beta}_2^{\mathsf{T}})$ where $\boldsymbol{\beta}_1$ is of length $p+1$ and $\boldsymbol{\beta}_2$ is of length k.

The penalized least squares estimator is the mean of the posterior distribution of $\boldsymbol{\beta}$ when $\boldsymbol{\beta}_1$ has an improper uniform (on $\mathbf{R}^{p+1}$) prior density and $\boldsymbol{\beta}_2$ has a proper prior proportional to $\gamma^{k/2}\exp\{-(\gamma/2)\boldsymbol{\beta}_2^{\mathsf{T}}\boldsymbol{\beta}_2\}$ where $\gamma = \alpha/\sigma_\epsilon^2$ and, as before, k is the number of knots. This prior on $\boldsymbol{\beta}$ induces a prior on $g(\cdot)$ and $\mathbf{g}$ because $g(x) = \mathbf{B}(x)^{\mathsf{T}}\boldsymbol{\beta}$.

The posterior of $\boldsymbol{\beta}$, conditional on σ_ϵ^2 and α, is $N\{(\mathcal{B}^{\mathsf{T}}\mathcal{B} + \alpha\mathbf{D})^{-1}\mathcal{B}^{\mathsf{T}}\mathbf{Y}, \sigma_\epsilon^2(\mathcal{B}^{\mathsf{T}}\mathcal{B} + \alpha\mathbf{D})^{-1}\}$. Let $\mathbf{A}(\alpha) = \mathcal{B}(\mathcal{B}^{\mathsf{T}}\mathcal{B} + \alpha\mathbf{D})^{-1}\mathcal{B}^{\mathsf{T}}$. Then the posterior distribution of $\mathbf{g} = \mathcal{B}\boldsymbol{\beta}$, conditional on $(\alpha, \sigma_\epsilon^2)$, is $N\{\mathbf{A}(\alpha)\mathbf{Y}, \sigma_\epsilon^2\mathbf{A}(\alpha)\}$, the same result obtained for smoothing splines. Often $\mathbf{D}$ is singular but $\mathcal{B}^{\mathsf{T}}\mathcal{B} + \alpha\mathbf{D}$ is nonsingular, so that the prior is improper but the posterior is proper.

3. GENERAL MODEL

We consider the measurement error model

$$Y_i = m(X_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{3}$$

where the ϵ_i are the independent normal random variables with mean 0 and variance σ_ϵ^2. The X's are not observable (i.e., they are latent variables), but W that are surrogates for the Xs are observed,

$$W_{ij} = X_i + U_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m_i, \tag{4}$$

where the U_{ij} are independent normal errors with mean 0 and variance σ_u^2.

Model (4) is more general that it first appears. It can be interpreted as stating that a *known* function of the observed covariates (W) is the same function of the latent variables, plus independent, homoscedastic, normally distributed measurement errors. We have written (4) as if the function were the identity function, but it could be anything (e.g., the logarithm). The reason for this generality is that in (3), the function $m(\cdot)$ is unknown, so that, for example, if $m_*(v) = m\{\exp(v)\}$, then $m_*\{\log(x)\} = m(x)$.

The mean function m is modeled as a P-spline with smoothing parameter α. We use the notation $[A]$ and $[A|B]$ to represent prior densities and conditional densities.

Denote $\boldsymbol{\theta} = (\mathbf{g}, \mathbf{X}, \sigma_\epsilon^2, \sigma_u^2, \alpha)$. The posterior density is

$$[\boldsymbol{\theta}|\mathbf{Y}, \mathbf{W}] \propto [\mathbf{Y}|\mathbf{g}, \mathbf{X}, \sigma_\epsilon^2][\mathbf{W}|\mathbf{X}, \sigma_u^2][\mathbf{g}|\alpha][\sigma_u^2][\sigma_\epsilon^2][\mathbf{X}][\alpha]. \tag{5}$$

The form of (5) has a structure that is common in latent variable models. We exploit an important feature of such models, namely that in the Gibbs sampler or other Monte Carlo computational approaches, once $X_1, \ldots, X_n$ are generated from the posterior, estimation of g becomes a standard problem for which much software exists. A basic modeling issue and computational problem is how to generate $X_1, \ldots, X_n$.

Two approaches to the estimation of $\mathbf{g}$ are taken. The first method, which is "quick and dirty," estimates the posterior mode of $\mathbf{g}$ by the iterative conditional modes (ICM) algorithm (Besag 1986). The calculation is fast and easy to blend with a program that calculates a P-spline. The ICM method also serves to create starting values for the second method, which is fully Bayesian but involves more complex and time-consuming calculation. The former is described in the next section; the latter, in Section 3.2.

3.1 The Iterative Conditional Modes Algorithm

In this section we describe an iterative method of estimating g in the presence of measurement error by finding the mode of the posterior in (5). This methodology sacrifices the philosophical advantages of a fully Bayesian analysis for computational ease. In the first step, we estimate the three variance components $\sigma_\epsilon^2, \sigma_u^2$, and α; these estimates are held fixed for the rest of the procedure. We first describe the estimation of these variance components.

If $m_i = 1$ for all $i = 1, \ldots, n$, then the user must supply an estimate for σ_u^2; otherwise, the usual pooled sample variance of the W's from analysis of variance calculations is used. If s_i^2 is the sample variance of $W_{i1}, \ldots, W_{im_i}$, then $\hat{\sigma}_u^2 = \sum_{i=1}^n (m_i - 1)s_i^2 / \sum_{i=1}^n (m_i - 1)$.

The initial estimate for the $\mathbf{X}$ parameter is $\mathbf{X}^{(0)} = (\overline{W}_1, \ldots, \overline{W}_n)$, where $\overline{W}_i = \sum_{j=1}^{m_i} W_{ij}/m_i$. A *naive* smoothing spline, $\hat{\mathbf{g}}^{(0)}$, is estimated by assuming $\mathbf{X} = \mathbf{X}^{(0)}$ and fitting the standard nonmeasurement error smoothing spline. The smoothing parameter for the naive estimator, $\hat{\alpha}$, is fit using cross-validation (CV) or GCV; we use CV in our numerical work in this article. We use the standard estimate of σ_ϵ^2 (see Green and Silverman 1994),

$$\hat{\sigma}_\epsilon^2 = \sum_{i=1}^n \{Y_i - \hat{g}^{(0)}(\hat{X}_i^{(0)})\}^2 / \operatorname{trace}\{\mathbf{I} - \mathbf{A}(\hat{\alpha})\}.$$

A normal prior distribution is used for each X_i, with mean μ_x and variance σ_x^2, where μ_x and σ_x are constants. In the algorithm, we replace these by the mean and standard deviation of the Ws.

Conditional on $\hat{\sigma}_\epsilon^2$, $\hat{\sigma}_u^2$, and $\hat{\alpha}$, the posterior distribution is proportional to

$$\exp\left[-\frac{1}{2\hat{\sigma}_\epsilon^2}\sum_{i=1}^n\{Y_i - g(X_i)\}^2 - \frac{1}{2\hat{\sigma}_u^2}\sum_{i=1}^n\sum_{j=1}^{m_i}(W_{ij} - X_i)^2 - \frac{1}{2\sigma_x^2}\sum_{i=1}^n(X_i - \mu_x)^2 - \frac{\hat{\alpha}}{2\hat{\sigma}_\epsilon^2}\mathbf{g}^{\mathsf{T}}\mathbf{K}\mathbf{g}\right]. \tag{6}$$

We use the ICM approach described by Besag (1986) to find the posterior mode of (6). This approach can be also phrased as generalized EM (Meng and Rubin 1993). The approach is to find the posterior mode by sequentially finding the mode of the complete conditional distributions. This is in contrast to Gibbs sampling, where observations are iteratively *drawn* from the complete conditionals. In summary, ICM iteratively updates the parameters with their modal values.

ICM Algorithm

1. Find the estimates $\hat{\sigma}_\epsilon^2$, $\hat{\sigma}_u^2$, $\hat{\alpha}$, and the naive spline $\mathbf{g}^{(0)}$. Set $\mathbf{i} = 1$.

2. Fixing $\mathbf{K}$, find the vector $\mathbf{X}$ that maximizes (6) conditional on $\mathbf{g}^{(i-1)}$. These are labeled $\mathbf{X}^{(i)}$. There is no analytic solution available. We use a grid search to find the maximum for each component of $\mathbf{X}$. The complete conditional for each component of $\mathbf{X}$ is independent of the other components of $\mathbf{X}$, because we have fixed $\mathbf{K}$. Therefore, this maximization can be done individually. We maximize each component using a uniform grid evaluation, followed by a finer grid used in the region of the maximum value from the first grid.

3. Find the vector $\mathbf{g}^{(i)}$ that maximizes (6) conditional on $\mathbf{X}^{(i)}$. This is the usual P-spline for a nonmeasurement error problem, which is described in Section 2.

4. Set $i = i+1$ and repeat steps 2 and 3 until the estimate $\mathbf{g}^{(i)}$ converges.

Figure 1 demonstrates the ICM method with simulated data. The data consist of 100 X's generated from a standard normal distribution. The responses, Y, were generated from a normal distribution with a standard deviation of .3 and a mean function of

$$m(x) = \frac{\sin(\pi x/2)}{1+2x^2\{\text{sign}(x)+1\}}. \tag{7}$$

Each $m_i = 2$ and the W_{ij} for $j = 1, 2$, are normally distributed with a mean of X_i and a standard deviation of .8. Figure 1 shows the (x, y) pairs with open diamonds and the $(\overline{W}, y)$ pairs with solid diamonds. The regression function $m(x)$ and the naive regression spline, $g^{(0)}$, are presented, and the estimated $(X^{(1)}, Y)$ pairs from the first ICM iteration are shown by the "1" symbols. The naive spline, the estimated curve from one iteration of the ICM procedure, $g^{(1)}$, and the curve judged to have converged, $g^{(\infty)}$, are presented.

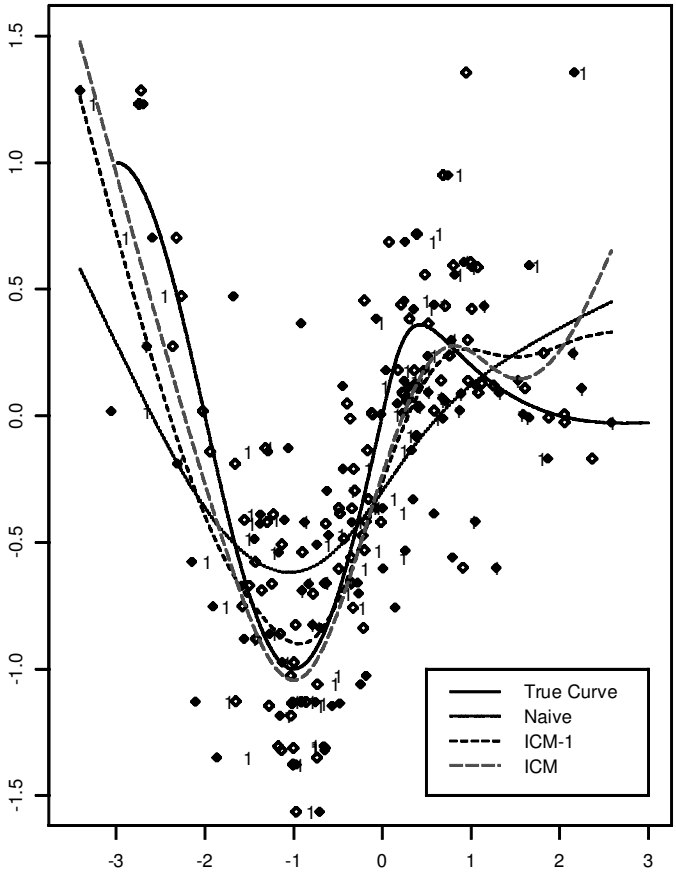

Figure 1. The (x,y) Pairs Are Shown With Open Diamonds Whereas the $(\overline{W},Y)$ Observations Are Shown With Solid Diamonds. The true regression function is shown with the solid line. The naive spline estimate, the ICM spline from one iteration (ICM-1), and the converged ICM spline (ICM) are also shown. The 1's represent the $(X^{(1)},Y)$ pairs.

The motivation behind the ICM approach is that it uses some of the strengths of the Bayesian approach but is very easy to program and fast to compute. An S-PLUS function for the ICM approach is available from the first author.

The major difficulty with the ICM method as a general method for measurement error models can be seen most clearly by considering parametric models $g(X) = g(X, \beta)$ of known form but with an unknown parameter β. If in (6) we set $\alpha = 0$, delete the term $\sum_{i=1}^{n}(X_i - \mu_x)^2/(2\sigma_x^2)$, replace $g(X_i)$ by $g(X_i, \beta)$, and maximize in the unknown parameters $(X_1, \ldots, X_n, \beta, \sigma_u^2, \sigma_\epsilon^2)$, then we are computing what is known in the literature as the *functional maximum likelihood estimate* (Fuller 1987). Functional models assume that $X_1, \ldots, X_n$ are fixed parameters to be estimated. In contrast, in a *structural* model, the X_i's are latent variables from a distribution depending on structural parameters; one integrates the X_i's out of the likelihood and maximizes simultaneously over the structural and other parameters. In linear regression, the functional approach is known to yield consistent estimates of regression parameters. In nonlinear regression, this need not be the case. Fuller (1987) and Amemiya and Fuller (1988) studied parametric nonlinear regression problems as $n \to \infty$ and $\sigma_u^2 \to 0$ in such a way that $\sigma_u^2 \propto n^{-1/2}$. They found that in the terms of rate of convergence, the functional approach is no better than the naive approach, because both have bias of order $O(\sigma_u^2) = O(n^{-1/2})$; see eq. (2.11) of Amemiya and Fuller (1988).

This similarity with functional modeling suggests that although the ICM approach is computationally simple, it need not yield consistent estimates of the regression function, not even in parametric problems. The next section describes the fully Bayesian approach, which is computationally more difficult but has the benefits of the Bayesian machinery. The fully Bayesian approach is structural, and with diffuse priors, one gets essentially the structural maximum likelihood estimate (MLE), a consistent estimator. In this context, the ICM method largely serves to produce starting values for the full Bayesian approach.

3.2 Fully Bayesian Approach

In this section we develop the fully Bayesian approach to this problem. The ICM approach estimates the variance components and keeps them fixed, allowing the smoothing spline estimate and the X's to fluctuate. In this section, prior distributions are placed on all parameters, including the structural parameters (μ_x, σ_x^2) and the variance components $(\sigma_\epsilon^2, \sigma_u^2)$, and the joint posterior distribution is calculated. One of the benefits of this approach is that observations of the smoothing spline are generated from the posterior, and thus we estimate the entire posterior distribution of g, not just its mode. Thus calculation of the various forms of "error bars" is straightforward. These credible sets take into account the measurement

error of the independent variables and the use of a data-based smoothing parameter.

The method is as follows. Without loss of generality, we replace α/σ_ϵ^2 by γ. The prior distributions for σ_ϵ^2 and σ_u^2 are inverse-gamma distributions, and the prior distribution for γ is a gamma distribution: $\sigma_\epsilon^2 \sim \text{IG}(A_\epsilon, B_\epsilon)$, $\sigma_u^2 \sim \text{IG}(A_U, B_U)$, and $\gamma \sim \text{G}(A_\gamma, B_\gamma)$. We use the definitions of the inverse-gamma and gamma distributions (respectively) from Berger (1985):

$$f(x|A,B) = \frac{1}{\Gamma(A)B^A x^{A+1}} \exp\left(-\frac{1}{Bx}\right) I_{(0,\infty)}(x)$$

and

$$f(x|A,B) = \frac{1}{\Gamma(A)B^A} x^{A-1} \exp\left(-\frac{x}{B}\right) I_{(0,\infty)}(x).$$

There is no reasonable prior distribution for the X's, which eases the computational burden. This prior distribution also can easily change from application to application. In some examples, a flat reference prior may be reasonable, whereas in others, a normal hierarchical distribution may be appropriate. A mixture of normals is a flexible approach that has some intuitive appeal (see Carroll, Roeder, and Wasserman 1999). A difficulty is that the X_is continually change throughout the MCMC algorithm, and updating this mixture at every iteration is chronically slow. We leave the choice of prior distribution for X an open choice for the particular application. For the simulations and examples in this article, we use a hierarchical Bayes approach. A normal distribution with mean μ_x and variance σ_x^2 is used, where $\mu_x \sim \text{normal}(d_x, t_x^2)$ and $\sigma_x^2 \sim \text{IG}(A_x, B_x)$.

The hyperparameters that are fixed a priori and thus are "tuning constants" are denoted by Roman fonts. These are $A_Y, B_Y, A_U, B_U, A_\gamma, B_\gamma, d_x, t_x^2, A_x$, and B_x.

Assuming the hierarchical normal structure for $[X]$, the joint posterior is proportional to

$$\begin{aligned}
&\exp\left\{-\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}\{Y_i - g(X_i)\}^2 - \frac{1}{2\sigma_u^2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(W_{ij}-X_i)^2 \right.\\
&\qquad \left. - \frac{1}{2\sigma_x^2}\sum_{i=1}^{n}(X_i-\mu_x)^2 - \frac{1}{2t_x^2}(\mu_x - d_x)^2\right\}\\
&\times \exp\left\{-(\gamma/2)\mathbf{g}^\top\mathbf{K}\mathbf{g} - \frac{1}{B_\epsilon\sigma_\epsilon^2} - \frac{1}{B_U\sigma_u^2} \right.\\
&\qquad \left. - \frac{\gamma}{B_\gamma} - \frac{1}{B_x\sigma_x^2}\right\}\\
&\times \sigma_\epsilon^{-2(n/2+A_\epsilon+1)}\sigma_u^{-2(1/2\sum_{i=1}^n m_i + A_U+1)}\\
&\times \sigma_x^{-2(n/2+A_X+1)}\gamma^{(A_\gamma+M/2-1)}, \qquad (8)
\end{aligned}$$

where $M = n-2$ as in (2). The sampling is done using a successive substitution algorithm (Gelfand and Smith 1990). The complete conditional distributions for the parameters are

$$\begin{aligned}
&\mathbf{g}|\mathbf{X},\gamma,\sigma_\epsilon^2,\mathbf{Y},\mathbf{W} \sim \text{normal}\left\{\mathbf{A}(\sigma_\epsilon^2\gamma)\mathbf{y}, \sigma_\epsilon^2\mathbf{A}(\sigma_\epsilon^2\gamma)\right\},\\
&[X_i|\mathbf{W}_i,\mathbf{g},\sigma^2,\sigma_u^2,\mathbf{Y},\mathbf{W}]\\
&\quad \propto \exp\left(-\frac{1}{2\sigma_u^2}\sum_{j=1}^{m_i}(W_{ij}-X_i)^2 \right.\\
&\qquad \left. -\frac{1}{2\sigma_\epsilon^2}\{Y_i - g(X_i)\}^2 - \frac{1}{2\sigma_x^2}(X_i-\mu_x)^2\right), \qquad (9)
\end{aligned}$$

$$\begin{aligned}
&\sigma_\epsilon^2|\mathbf{g},\mathbf{X},\mathbf{Y},\mathbf{W}\\
&\qquad \sim \text{IG}\left(A_\epsilon + n/2,\ [1/B_\epsilon + (1/2)\sum_{i=1}^{n}\{Y_i - g(X_i)\}^2]^{-1}\right),\\
&\sigma_u^2|\mathbf{X} \sim \text{IG}\left(A_U + (1/2)\right.\\
&\qquad \left.\sum_{i=1}^{n} m_i,\ \left[1/B_U + (1/2)\sum_{i=1}^{n}\sum_{j=1}^{m_i}(W_{ij}-X_i)^2\right]^{-1}\right),\\
&\qquad \gamma|\mathbf{g},\mathbf{X} \sim \text{G}\left(A_\gamma + \frac{M}{2},\ \left[1/B_\gamma + \frac{1}{2}\mathbf{g}^\top\mathbf{K}\mathbf{g}\right]^{-1}\right),\\
&\mu_x|\mathbf{X} \sim \text{normal}\{(n\overline{X}t_x + d_x\sigma_x^2)/(nt_x^2+\sigma_x^2), \sigma_x^2 t_x^2/(nt_x^2+\sigma_x^2)\},\\
&\qquad \sigma_x^2|\mathbf{X} \sim \text{IG}\left[A_x + n/2,\ \left\{B_x^{-1} + (1/2)\sum_{i=1}^{n}(X_i-\mu_x)^2\right\}^{-1}\right].
\end{aligned}$$

The estimates $\hat{\sigma}_\epsilon^2$, $\hat{\sigma}_u^2$, $\hat{\gamma}$, $\mathbf{x}^{(\infty)}$, and $\mathbf{g}^{(1)}$ from the ICM approach are used as starting values for the MCMC algorithm. Observations from each of the complete conditionals are drawn iteratively in the order just presented. The generation of an observation of $\mathbf{g}$ is computationally difficult for smoothing splines, because they have n knots. Because the values of $\mathbf{X}$, $\mathbf{K}$ (for smoothing splines), and γ are continually changing in the algorithm, the matrix $\mathbf{A}(\sigma_\epsilon^2\gamma)$ (which is $n \times n$) and its inverse must be recomputed for each iteration of the MCMC algorithm. Hastie and Tibshirani (1998) discussed an algorithm for generating observations of $\mathbf{g}$ in $O(n)$ operations. Computations can be reduced by using P-splines with fewer than n knots, with no real loss of precision (Ruppert 2000).

The complete conditionals for the X_i's require a Metropolis–Hastings step. This is done by generating a candidate observation of X_i from a normal distribution with a mean of the current value of X_i and a standard deviation of $2\sigma_u^{(i)}/\sqrt{m_i}$, where $\sigma_u^{(i)}$ is the current value of σ_u in the MCMC algorithm. Using $\overline{W}_i$ as an estimate of X_i, without the information in the regression function, has a standard error of $\sigma_u^{(i)}/\sqrt{m_i}$. Using the rule of thumb of a candidate value with a standard deviation twice the standard deviation of the marginal posterior provides a conservative candidate distribution for X_i. In terms of efficiency of the Metropolis–Hastings step, in our experience it is better to overestimate this standard deviation than to underestimate it. The evaluation of the complete conditional for X_i is computationally straightforward.

Generating observations from each of the other complete conditionals is straightforward and fast. Because the position of $\mathbf{X}$ changes throughout the algorithm, when using smoothing splines, we keep track of the value of g at a uniformly distributed grid of points. For each realization of g in the sampler, the value of g for each grid point is recorded. This enables us to keep track of pointwise moments and percentiles. For fixed-knot P-splines, g is defined by $\boldsymbol{\beta}$, the coefficients of the basis functions. Because the basis functions stay fixed as $\mathbf{X}$ varies, there is no need to record the values of g on a grid. Rather, one keeps track of the realizations of $\boldsymbol{\beta}$. For any realization of $\boldsymbol{\beta}$ there is a corresponding realization of $\mathbf{g}$ given by $\mathbf{g} = \mathcal{B}\boldsymbol{\beta}$. Details of implementation are given in the Appendix.

Having observations from the joint posterior distribution provides a powerful tool for inference. The pointwise mean

curve is a natural estimate of the regression mean function m. Pointwise credible intervals can also be calculated very easily from the observations of $\mathbf{g}$. Functions (linear or nonlinear) of the regression function can also be estimated, along with standard errors. This is the approach used by Wahba (1983) in nonmeasurement error cases and by Hastie and Tibshirani (1998) in nonmeasurement error semiparametric models. Wahba's work predated the revolution in Bayesian computations, and she treated the smoothing parameter as fixed. Hastie and Tibshirani used the Gibbs sampler to adjust the credible sets for uncertainty in the variance components that define the smoothing parameter. In this article, use of the Gibbs sampler also adjusts the credible sets for measurement error.

Although the regression function and its functionals are the main focus of this article, inferences about the mismeasured X_i's can also be made. Posterior means and credible intervals can easily be constructed for each of the individual X_i's. The variance components may also be of interest, and likewise constructing estimates and credible intervals for them is straightforward.

For an example of this method, we use the same data from the ICM example and use smoothing splines. A pointwise posterior mean curve is used for the estimate of m. Credible curves are calculated by interpolating the pointwise $100(1-\alpha)\%$ credible intervals. A burn-in time of 500 observations is used with 1,000 observations from the posterior. Figure 2 shows the estimate of g and the 90% pointwise credible curves.

There are at least two possible methods for choosing the smoothing parameter for a smoothing spline. We place a prior distribution on γ; Hastie and Tibshirani (1998) used an identical procedure within semiparametric models. It is worth noting that by placing a continuous density prior on γ, we have automatically given zero prior probability to the possibility of doing no smoothing at all. This is an automatic way of avoiding the possibility of gross undersmoothing that caused so much trouble for the methods of Carroll et al. (1999).

An alternative to this method that we recommend is to choose the smoothing parameter using a criterion such as CV or GCV during each iteration of the successive substitution sampling. This method loses some of the Bayesian interpretation, is computationally expensive, and does not account for the effects of measurement error. However, a referee asked that we evaluate this procedure; see Section 4.

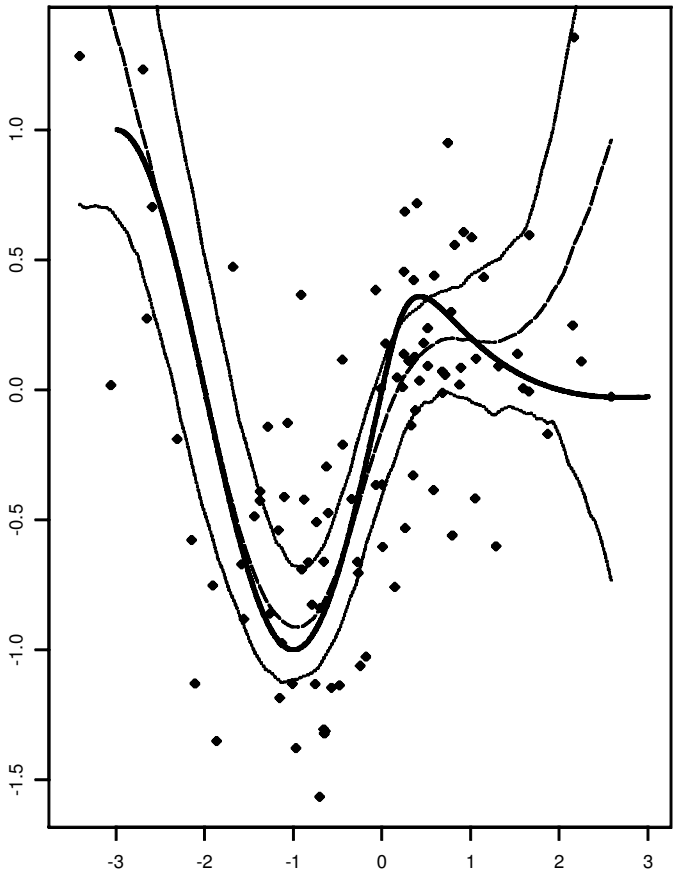

Figure 2. An Example of the Fully Bayesian Spline. The solid curve is the true regression function. The dashed middle curve is the mean of the posterior of the regression function. The dotted error bars represent the piecewise 90% credible intervals.

4. SIMULATIONS

4.1 Basic Simulations

We performed a series of simulations to compare our methods with those of Carroll et al. (1999). The results presented here are based on smoothing splines, but checks show that P-splines with 30 knots give much the same results. In each case, 200 simulated datasets were generated, with X_i generated as independent normal random variables with mean μ_x and variance σ_x^2, with two replicates ($m_i = 2$), and with ϵ_i also normally distributed. In each simulation, for the fully Bayesian method, the following prior distributions are used: $\sigma_\epsilon^2 \sim$ IG(1, 1), $\sigma_u^2 \sim$ IG(1, 1), $\gamma \sim$ G(3, 1000), $\mu_x \sim N(0, 10^2)$, and $\sigma_x^2 \sim$ IG(1, 1). These priors were selected because of their relative flexibility. They are all proper, yet they are not strong, in the sense of bringing a lot of information to the problem. We found the results insensitive to moderate modifications of these priors. The flexibility of these priors is demonstrated by their success in the different regression functions used in the simulations.

For purposes of bias and mean squared error calculations, the smoothing spline estimates of g were computed on a grid of 101 points in the interval $[a, b]$, the interval chosen to contain most of the distribution for X. The mean squared biases and mean squared errors were computed over this grid.

It is impossible to assess convergence of the MCMC chain for all simulated data sets. Instead, for a few selected datasets, we used tests of convergence (Gelman and Rubin 1992) separately for each parameter and also for the estimated function on a few selected grid points.

The first five cases considered were as follows:

Case 1: The regression function, m, is given in (7), with $n = 100$, $a = -2.0$, $b = 2.0$, $\sigma_\epsilon^2 = .3^2$, $\sigma_u^2 = .8^2$, $\mu_x = 0$, and $\sigma_x^2 = 1$.

Case 2: Same as case 1, except $n = 200$.

Case 3: A modification of case 1 except that $n = 500$.

Table 1. The Mean Squared Bias and MSE for the Simulation

Method	*Case 1*	*Case 2*	*Case 3*	*Case 4*	*Case 5*	*Case 6*	*Case 7*	*Case 8*
Mean squared bias $\times 10^2$								
Naive	5.59	4.92	5.21	1108	3733	4.83	4.80	15.27
ICM	2.98	2.22	2.04	629	1541	2.20	1.94	8.51
Bayes	.78	.38	1.04	17.4	468	1.74	1.68	6.20
Structural(5)	1.38	.62	.46	3.7	838	1.47	1.48	12.82
Structural(15)	1.44	.60	.66	3.3	226	1.75	1.70	12.36
MSE $\times 10^2$								
Naive	6.91	5.57	5.38	1155	3793	5.77	5.84	16.48
ICM	5.93	3.87	3.29	751	1948	4.36	3.93	12.38
Bayes	**2.84**	**1.56**	**1.47**	**195**	1031	**2.69**	**2.49**	**7.41**
Structural(5)	8.17	3.82	1.73	217	2032	7.27	7.91	16.84
Structural(15)	9.90	5.40	1.85	237	**799**	6.94	9.91	20.22

NOTE: The regression functions are $m(x) = \sin(\pi x/2)/[1+2x^2\{\text{sign}(x)+1\}]$ (cases 1, 2, 3 and 6), $m(x) = 1000x_+^3(1-x)_+^3$ (case 4), and $m(x) = 10\sin(4\pi x)$ (case 5). Case 7 is same as case 1 except that X is a normalized chi-squared(4) random variable, and ϵ is generated as a Laplace random variable. Case 8 is same as case 1 except that $m(x) = H(100x) + H\{-100(x-.5)\}$, where $H(x) = \{1+\exp(-x)\}^{-1}$. This function is poorly fit by a regression P-spline with 35 knots. "Naive" is the naive smoothing spline, "ICM" is the fully iterated ICM method, "Bayes" is the fully Bayesian method, and "Structural(m)" is the structural regression P-spline of Carroll et al. (1999) with m knots. In each column, the smallest MSE values is in boldface.

Case 4: Case 1 of Carroll et al. (1999), so that $m(x) = 1000x_+^3(1-x)_+^3$, $x_+ = xI(x>0)$, with $n = 200$, $a = .1$, $b = .9$, $\sigma_\epsilon^2 = .0015^2$, $\sigma_u^2 = (3/7)\sigma_x^2$, $\mu_x = .5$, and $\sigma_x^2 = .25^2$.

Case 5: A modification of case 4 of Carroll et al. (1999), so that $m(x) = 10\sin(4\pi x)$, with $n = 500$, $a = .1$, $b = .9$, $\sigma_\epsilon^2 = .05^2$, $\sigma_u^2 = .141^2$, $\mu_x = .5$, and $\sigma_x^2 = .25^2$.

The methods compared were the following:

- Naive smoothing spline fit ignoring measurement error
- Fully iterated ICM approach
- Fully Bayesian approach
- Structural method (Carroll et al. 1999), 5 knots
- Structural method, 15 knots.

Table 1 presents summary results for mean squared bias and mean squared error (MSE). The SIMEX method discussed by Carroll et al. (1999) using a 40-knot quadratic P-spline and a quadratic extrapolant was also computed, with results better than the naive estimator but generally inferior to the others. The striking feature of this table is that our Bayesian estimator has at least as good *frequentist* properties as the frequentist methods. In cases 1 and 2 it clearly dominates, having less than half of the MSE of the other methods. In case 3, it MSE efficiency is 20% greater than the structural spline with 15 knots, whereas in case 5 it is only 25% less efficient. The improvement of the fully Bayesian method over the frequentist methods is especially large for smaller sample sizes, cases 1 and 6 for example, where $n = 100$.

Clearly, even this limited simulation suggests that our Bayesian method is at least competitive with other methods proposed previously in the literature.

4.2 Robustness to Priors

Our priors are proper yet not particularly informative. However, as suggested by a referee, it is interesting to compare our results when different priors are used. Here we focus on case 1 in the simulation, with the priors modified as follows: $\sigma_\epsilon^2 \sim \text{IG}(3, 1)$, $\sigma_u^2 \sim \text{IG}(3, 1)$, $\gamma \sim \text{G}(2, 2000)$, $\mu_x \sim \text{N}(0, 100^2)$, and $\sigma_x^2 \sim \text{IG}(3, 1)$. Compared with Table 1, when we ran the simulation using these priors, the MSE of the Bayesian approach changed from 2.84 to 2.53, a minimal change. We have run selected exercises on datasets with different priors, and in all cases there were only minimal changes.

4.3 Distributional Robustness and Model Misspecification

The method that we have developed assumes that X and ϵ are normally distributed, and that the function $m(x)$ is adequately represented by a spline. We ran a limited number of simulations to study violations of these assumptions.

Case 6: The same as case 1 except that X is a normalized chi-squared(4) random variable. Squared bias and MSE are evaluated on $[-1.25, 2.00]$.

Case 7: The same as case 1 except that X is a normalized chi-squared(4) random variable and ϵ is generated as a Laplace random variable. Squared bias and MSE are evaluated on $[-1.25, 2.00]$.

Case 8: The same as case 1 except that $m(x) = H(100x) + H\{-100(x-.5)\}$, where $H(x) = \{1+\exp(-x)\}^{-1}$. This function is poorly fit by a regression P-spline with 35 knots; see Figure 3. The results are also displayed in Table 1.

In case 6, where the distribution from X is far from the normal distribution, the Bayes method is still more efficient than the other methods. For case 7, where in addition the distribution for ϵ is nonnormal, we see that the Bayes method is still best, although as expected with a small loss of efficiency. We are not sufficiently bold or naive to suggest that the Bayes method will retain distributional robustness in all cases, but the results are at least encouraging in the case of small model deviations.

In case 8, it is the spline representation of the function $m(x)$ that fails. Because all of the methods in Table 1 are based on splines, it was difficult to guess a priori what would happen in the simulation, although one perhaps would have expected that all the methods would be equally bad, although as it turns out the Bayes method had the smallest bias and MSE.

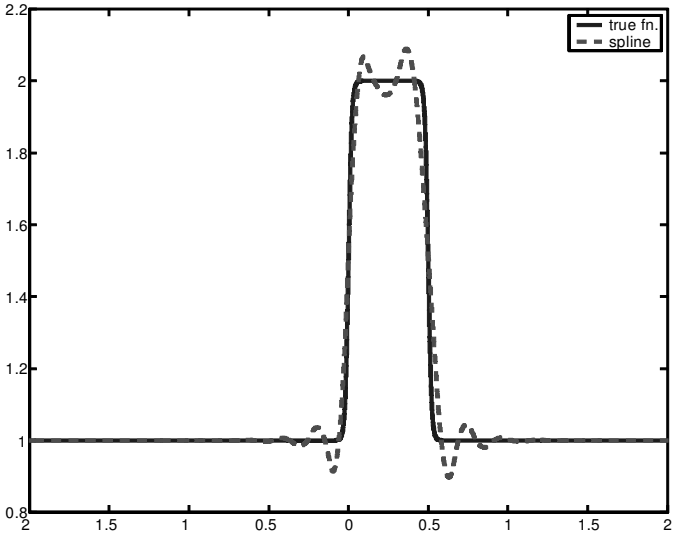

Figure 3. The Function $m(x) = H(100x) + H\{-100(x - .5)\}$, Where $H(x) = \{1 + \exp(-x)\}^{-1}$ (——), and the Best-Fitting Quadratic P-Spline With 35 Knots (– – – –).

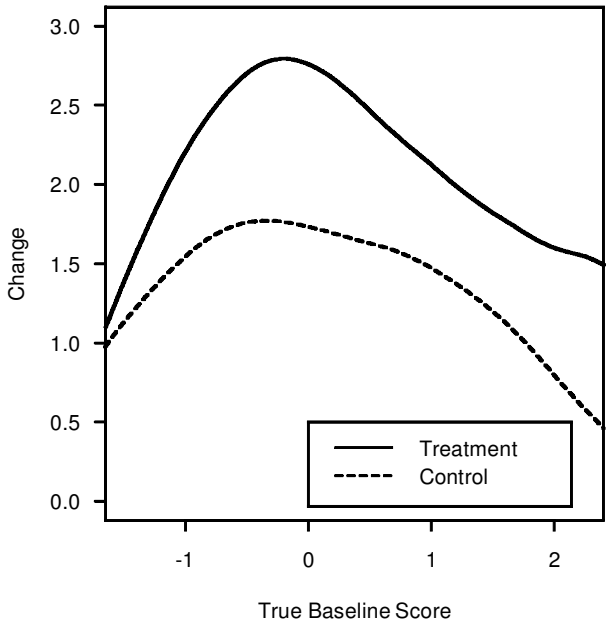

Figure 4. Estimate of the Function $\Delta(x) = m(x) - x$ for the Control Group (- - - -) and the Treatment Group (——) in the Example.

5. EXAMPLE

This article was partially motivated by the analysis of a real dataset. Unfortunately, we do not have permission to discuss the study details here, or to make the data available to the public. The data that we have have been transformed and rescaled, and random noise has been added.

Essentially, there is a treatment group and a control group, which are evaluated using a scale at baseline (W) and at the end of the study (Y). Smaller values of both indicate a more severe disease. The scale itself is subject to considerable error, because it is based on a combination of self-report and clinic interview. The study investigators estimate that in their transformed and rescaled form, the measurement error variance is approximately $\sigma_u^2 = .35$.

A preliminary Wilcoxon test applied to the observed change from baseline, $Y - W$, indicated a highly statistically significant difference between the two groups.

In the notation of (3), the main interest focuses on the population mean change from baseline $\Delta(X) = m(X) - X$ for the two groups and, most importantly, on the difference between these two functions.

Preliminary nonparametric regression analysis of the data ignoring measurement error indicates possible nonlinearity in the data. A quadratic regression is marginally statistically significant in the control group ($p \approx .03$) and marginally nonstatistically significant in the treated group ($\approx .07$). When we corrected the quadratic fits for the measurement error (Cheng and Schneeweiss 1998) and bootstrapped the resulting parameter estimates, both p values exceeded .20, although the fitted functions had substantial curvature. Thus the evidence for a linear model is mixed. We are interested in understanding the nature of the statistically significant difference between the two groups as evidenced by the Wilcoxon test.

Figure 4 plots $\Delta(X)$ for the placebo and treatment group using the fully Bayesian method. The functions are fairly similar in shape, with the treated group having higher values, essentially uniform in the range $[-1.50, 2.25]$ and covering most of the distribution of X. Both fitted functions exhibit curvature, although for those with true baseline score exceeding 0, the fitted functions are fairly linear.

Figure 5 shows the differences between the two functions, along with the 90% pointwise credible interval using the fully Bayesian method. The upper part of this interval is below 0 from approximately -1 to 2, and thus is in agreement with the Wilcoxon analysis. Interestingly, there is no evidence that the treatment is particularly effective for those who have the most severe disease at baseline (i.e., those with a true baseline score less than -1).

6. DISCUSSION

The Bayesian approach to measurement error, modeling the mismeasured variables as latent random variables and integrating them out, is a powerful one. In this article we have developed a Bayesian method for nonparametric regression in the presence of measurement error. By modeling a smoothing spline from a Bayesian standpoint, we create algorithms to calculate the posterior distribution of the regression function. The resulting estimate accounts for the effects of measurement error both on the estimator and on the smoothing parameter. The resulting smoothing parameter selector appears to be the first to adjust for the effects of measurement error.

Two algorithms are presented. The first algorithm is a quick-and-dirty method to find a posterior mode. The technique, based on the ICM procedure, is easy to combine with a program that calculates splines and is very fast. The fully Bayesian procedure is based on an MCMC algorithm. The fully Bayesian procedure is computationally more difficult but benefits from the modeling of each unknown and exploring the posterior, rather than finding the mode.

The simulations demonstrate the flexibility of the fully Bayesian approach, and even its efficiency in the frequentist

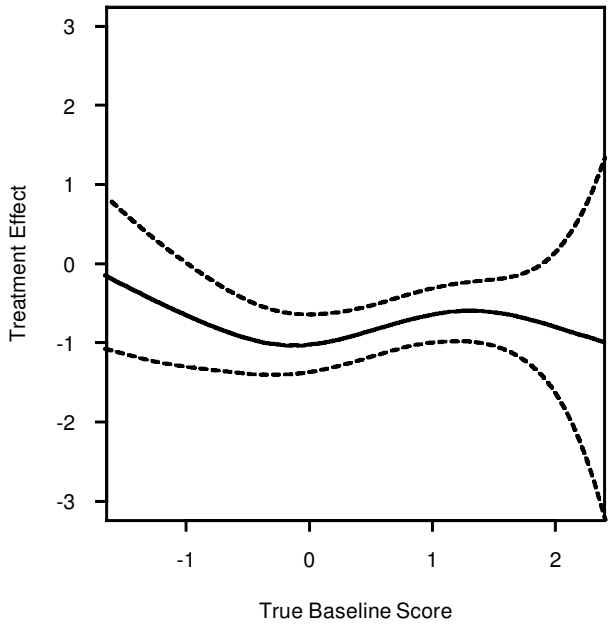

Figure 5. Estimate (solid line) of the Difference of the Function $\Delta(x) = m(x) - x$ Between the Treatment Group and the Control Group in the Example (control-treatment) With 90% Pointwise Credible Intervals (dashed lines).

sense, at least in the examples we have investigated. The fully Bayesian approach also enables inference on more than just the regression function.

We believe that the fully Bayesian approach works better than the previous proposals is because: often one can estimate the unknown X_i significantly more accurately using *all* information in the data about X_i rather than using just the $W_{ij}, j = 1, \ldots, n_i$. It is possible that a likelihood-based frequentist spline approach can also take advantage of this information, but such work is clearly outside the scope of this article. Many errors-in-variables techniques in parametric problems, and the technique of Carroll et al. (1999) in the nonparametric problems, and estimate X_i using only the W_{ij}. However, there is information in Y_i about X_i, and the fully Bayesian approach extracts this information. This fact can be seen in (9) for the conditional density of X_i given the other parameters.

As an illustration, we generated data from the following model. The sample size was $n = 201$, the X_i were normal(1, 1), $m_i \equiv 2$, and $m(x) = \sin(2x)$. Also $\sigma_\epsilon = .15$ and $\sigma_u = 1$. The fully Bayesian estimate of m is the spline fit with the imputed X's, averaged over the Gibbs sample. Thus the crucial quantity is how close the imputed $m(X_i)$ are, on average over the Gibbs samples, to the actual value of this quantity. In examples such as this one where m is nonmonotonic, for some cases the imputed X_i might not be close to the actual X_i but the imputed $m(X_i)$ might be close to the true $m(X_i)$. For estimation of $m(\cdot)$, the latter is good enough. Therefore, we compared various estimates of the X_i by using the norm $\|m(\mathbf{X}) - m(\widehat{\mathbf{X}})\|$, where $\mathbf{X} = (X_i, \ldots, X_n)^T$ is the vector of true X_i and similarly $\widehat{\mathbf{X}}$ is the vector of predicted X_i. Note that $m(\cdot)$ here is the true regression function. We are *not* comparing estimates of $m(\cdot)$, only estimates of the X_i. Consider two estimators of $\mathbf{X}$. The first estimator is the conditional expectation of X_i given $\overline{W}_i$. Because $(X_i, \overline{W}_i)$ is jointly normal, this is the optimal estimator given only $\overline{W}_i$. The second estimator uses the X_i from the Gibbs output and thus is a sample from the distribution of X_i given the Y_i and the W_{ij}. We calculated $\|m(\mathbf{X}) - m\{E(\mathbf{X}|\mathbf{W})\}\|$ and $\|m(\mathbf{X}) - \text{ave}\{m(\mathbf{X})\}\|$ where "ave" means average over the MCMC output and is thus a Monte Carlo estimate of conditional expectation given the Y_i and the W_{ij}. For six samples, the ratios of these norms were 3.1, 3.7, 3.8, 2.7, 2.1, and 4.2. These values are consistently well above 1, showing that the fully Bayesian approach is giving more information about X_i than what is available from $\overline{W}_i$ alone. The latter also uses the full structural model for the marginal distribution of the X_i, so it is not the structural assumption that is giving extra information about the X_i to the Bayesian estimator; rather, it is the regression model relating Y_i to X_i that provides this information. Clearly, this leaves open the possibility that with highly nonnormal errors, or highly heteroscedastic ones, the misspecified information from our simple model will lead to bias or other deleterious behavior in the Bayesian method.

We study the case where the measurement error and natural error are normally distributed. In most of the cases in the measurement error literature, the results are robust to the assumption of normality, once the additivity of errors in (4) is satisfied, possibly by transformation. Extending the normality of the ϵ's in (3), the natural error, to other distributions adds a level of complexity to the problem, because the normality of the complete conditional distribution for the spline will no longer hold. The methods presented in this article could be naturally combined with the work of Hastie and Tibshirani (1998) for modeling measurement error in semiparametric models.

While this manuscript was undergoing a final revision, an interesting unpublished manuscript by Ganguli, Staudenmayer, and Wand appeared. These authors extend our model by assuming multiple covariates and an additive regression function. They also estimate fixed effects and the variance components (of the random coefficients in the spline) by maximum likelihood. Because the smoothing parameters are ratios of variance components, these are also chosen by maximum likelihood. The likelihood involves a high-dimensional integral, so computation of the MLEs is not trivial, and the author uses the nesting EM algorithm of van Dyk (2000). Simulations indicate that the MLEs behave satisfactorily, but the MLEs were not compared to other estimators.

APPENDIX: BAYESIAN IMPLEMENTATION FOR REGRESSION P–SPLINES

For fixed-knot P-splines, $g(x) = \mathbf{B}^\intercal(x)\boldsymbol{\beta}$, $\mathbf{g} = \mathcal{B}\boldsymbol{\beta}$, and the penalty matrix is $\mathbf{D}$. Apportion $\mathbf{B}(x) = \{\mathbf{B}_1^\intercal(x), \mathbf{B}_2^\intercal(x)\}^\intercal$, where $\mathbf{B}_1^\intercal(x)$ is the first $p+1$ elements of $\mathbf{B}^\intercal(x)$. Apportion $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\intercal, \boldsymbol{\beta}_2^\intercal)^\intercal$ similarly. Then let the prior for $\boldsymbol{\beta}_1$ be normal$(0, \delta\Sigma)$ for a fixed covariance matrix Σ and δ "large," and the prior for $\boldsymbol{\beta}_2$ be normal$\{0, \gamma^{-1}\mathbf{I}\}$ ($\mathbf{I}$ is the identity matrix). Define the matrix $\mathbf{D}_* = \sigma_\epsilon^2 \text{diag}(\Sigma^{-1}/\delta, \gamma\mathbf{I})$. In the limit as $\delta \to \infty$, $\mathbf{D}_* \to \sigma_\epsilon^2\gamma\mathbf{D}$. With these conventions, the joint

posterior for $\boldsymbol{\beta}$ becomes

$$\boldsymbol{\beta}|\mathbf{Y}, \mathbf{X}, \mathbf{W} = \text{normal}(\mathbf{QH}, \mathbf{Q}),$$

$$\mathbf{H} = \sigma_\epsilon^{-2} \sum_{i=1}^{n} \mathbf{B}(X_i) Y_i = \sigma_\epsilon^{-2} \mathcal{B}^T \mathbf{Y},$$

and

$$\mathbf{Q} = \sigma_\epsilon^2 \left\{ \sum_{i=1}^{n} \mathbf{B}(X_i)\mathbf{B}^\mathsf{T}(X_i) + \mathbf{D}_* \right\}^{-1} = \sigma_\epsilon^2 (\mathcal{B}^T \mathcal{B} + \mathbf{D}_*)^{-1}.$$

Similarly, the complete conditional for $\mathbf{X}$ is

$$X_i|Y_i, W_i \propto \exp\left\{ -\frac{1}{2\sigma_\epsilon^2} \sum_{i=1}^{n} \{Y_i - \mathbf{B}^\mathsf{T}(X_i)\boldsymbol{\beta}\}^2 - \frac{1}{2\sigma_u^2} \sum_{i=1}^{n} \sum_{j=1}^{m_i} (W_{ij} - X_i)^2 - \frac{1}{2\sigma_x^2} \sum_{i=1}^{n} (X_i - \mu_x)^2 \right\}.$$

As described in the text, Metropolis–Hastings steps can be used to generate new values of the X_i's.

The full conditional for γ given $\boldsymbol{\beta}$ is $\mathrm{G}(A_\gamma + k/2, \{B_\gamma^{-1} + \boldsymbol{\beta}_2^\mathsf{T}\beta_2/2\}^{-1})$. The full conditionals for σ_ϵ^2, σ_u^2, μ_x, and σ_x^2 are the same as for smoothing splines as given in Section 3.2.

Unlike smoothing splines, P-splines need not have their knots at the values of the X_i. Instead, we place the P-spline knots at fixed quantiles of the $\overline{W}_i$, so that the knots are fixed throughout the MCMC iterations. More specifically, if there are k knots, then we take $k+2$ values of p equally spaced on $[0, 1]$, delete the first and last (0 and 1), and then place a knot at the pth quantile of the $\overline{W}_i$ for each of these k values of p.

[Received January 2000. Revised October 2000.]

REFERENCES

Amemiya, Y., and Fuller, W. A. (1988), "Estimation for the Nonlinear Functional Relationship," *The Annals of Statistics*, 16, 147–160.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis* (2nd ed.), New York: Springer-Verlag.

Besag, J. (1986), "On the Statistical Analysis of Dirty Pictures" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 48, 259–279.

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.

Cheng, C. L., and Schneeweiss, H. (1998),"Polynomial Regression With Errors in Variables," *Journal of the Royal Statistical Society,* Ser. B, 60, 189–200.

Cook, J. R., and Stefanski, L. A. (1994), "Simulation–Extrapolation Estimation in Parametric Measurement Error Models," *Journal of the American Statistical Association*, 89, 1314–1328.

Eilers, P. H. C., and Marx, B. D. (1996), "Flexible Smoothing with B-Splines and Penalties" (with discussion), *Statistical Science*, 11, 89–102.

Eubank, R. L. (1999), *Nonparametric Regression and Spline Smoothing* (2nd ed.), New York: Marcel Dekker.

Fan, J., and Truong, Y. K. (1993), "Nonparametric Regression with Errors in Variables," *The Annals of Statistics*, 21, 1900–1925.

Fuller, W. A. (1987), *Measurement Error Models*, New York: Wiley.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling–Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A., and Rubin, D. B. (1992), "Inference From Iterative Simulation Using Multiple Sequences," *Statistical Science*, 7, 457–472.

Green, P. J., and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, London: Chapman and Hall.

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, New York: Chapman and Hall.

——— (2000),"Bayesian Backfitting," (with discussion), *Statistical Science*, 15, 193–223.

Meng, X. L., and Rubin, D. B. (1993), "Maximum Likelihood Estimation via the ECM Algorithm: A General Framework," *Biometrika*, 80, 267–278.

Nychka, D. (1988), "Bayesian Confidence Intervals for a Smoothing Spline," *Journal of the American Statistical Association*, 83, 1134–1143.

——— (1990), "The Average Posterior Variance of a Smoothing Spline and a Consistent Estimate of the Average Squared Error," *The Annals of Statistics*, 18, 415–428.

Ruppert, D. (2000), "Selecting the Number of Knots for Penalized Splines," preprint (available at www.orie.cornell.edu/~davidr/papers).

Ruppert, D., and Carroll, R. J. (2000), "Spatially Adaptive Penalties for Spline Fitting," *Australia and New Zealand Journal of Statistics*, 42, 205–223.

van Dyk, D. A. (2000), "Nesting EM Algorithms for Computational Efficieny," *Statistica Sinica*, 10, 203–226.

Wahba, G. (1978), "Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression," *Journal of the Royal Statistical Society,* Ser. B, 40, 364–372.

——— (1983),"Bayesian 'Confidence Intervals' for the Cross-Validated Smoothing Spline," *Journal of the Royal Statistical Society,* Ser. B, 45, 133–150.

——— (1990), *Spline Models for Observational Data*, Providence: SIAM Press.

This article has been cited by:

1. Raymond J. Carroll, Aurore Delaigle, Peter Hall. 2009. Nonparametric Prediction in Measurement Error ModelsNonparametric Prediction in Measurement Error Models. *Journal of the American Statistical Association* **104**:487, 993-1003. [Abstract] [PDF] [PDF Plus] [Supplementary material]
2. Yu-Jen Cheng, Ciprian M. Crainiceanu. 2009. Cox Models With Smooth Functional Effect of Covariates Measured With ErrorCox Models With Smooth Functional Effect of Covariates Measured With Error. *Journal of the American Statistical Association* **104**:487, 1144-1154. [Abstract] [PDF] [PDF Plus] [Supplementary material]
3. Aurore Delaigle , , Jianqing Fan , , Raymond J. Carroll . 2009. A Design-Adaptive Local Polynomial Estimator for the Errors-in-Variables ProblemA Design-Adaptive Local Polynomial Estimator for the Errors-in-Variables Problem. *Journal of the American Statistical Association* **104**:485, 348-359. [Abstract] [PDF] [PDF Plus]
4. Duchwan Ryu , Debajyoti Sinha , Bani Mallick , Stuart R . Lipsitz , Steven E . Lipshultz . 2007. Longitudinal Studies With Outcome-Dependent Follow-upModels and Bayesian RegressionLongitudinal Studies With Outcome-Dependent Follow-up. *Journal of the American Statistical Association* **102**:479, 952-961. [Abstract] [PDF] [PDF Plus]
5. N. D . Pearce , M. P . Wand . 2006. Penalized Splines and Reproducing Kernel MethodsPenalized Splines and Reproducing Kernel Methods. *The American Statistician* **60**:3, 233-240. [Abstract] [PDF] [PDF Plus]
6. Panu Erästö , Lasse Holmström . 2005. Bayesian Multiscale Smoothing for Making Inferences About Features in ScatterplotsBayesian Multiscale Smoothing for Making Inferences About Features in Scatterplots. *Journal of Computational and Graphical Statistics* **14**:3, 569-589. [Abstract] [PDF] [PDF Plus]