Aust. N. Z. J. Stat. **47**(2), 2005, 193–202

ADDITIVE MODELS WITH PREDICTORS SUBJECT TO MEASUREMENT ERROR

BHASWATI GANGULI[1], JOHN STAUDENMAYER[2*] AND M.P. WAND[3]

Indian Institute of Management, University of Massachusetts and the University of New South Wales

Summary

This paper develops a likelihood-based method for fitting additive models in the presence of measurement error. It formulates the additive model using the linear mixed model representation of penalized splines. In the presence of a structural measurement error model, the resulting likelihood involves intractable integrals, and a Monte Carlo expectation maximization strategy is developed for obtaining estimates. The method's performance is illustrated with a simulation study.

Key words: Metropolis–Hastings; mixed models; Monte Carlo expectation maximization; nested EM; penalized splines; restricted maximum likelihood.

1. Introduction

Additive models are multi-predictor regression models that assume little more than additivity. Their advantage over ordinary regression models is that the functional form of continuous predictors is not restricted to be in a parametric family and therefore is more flexible.

For regression models in general it is well established that measurement error in a predictor distorts its estimated relationship to the response (e.g. Fuller, 1987 Chapter 1). A number of studies have devised schemes for redressing this situation for parametric regression models (e.g. Carroll, Ruppert & Stefanski, 1995; Spiegelman, Rosner & Logan, 2000) and single-predictor non-parametric models (Fan & Truong, 1993; Carroll, Maca & Ruppert, 1999; Berry, Carroll & Ruppert, 2002; Staudenmayer & Ruppert, 2004). However, to our knowledge, there is no available methodology for handling measurement error in additive models.

Carroll *et al.* (1999) formulated and compared several estimators that have small, persistent, asymptotic biases but perform quite well in simulations. These estimators address the covariate measurement error with either simulation extrapolation or extended regression calibration, and they use penalized splines or local linear regression for the non-parametric estimation. Berry *et al.* (2002) proposed Bayesian estimation methods for this problem, and their simulations indicate improvements over the methods of Carroll *et al.* (1999).

Received August 2003; revised January 2004; accepted January 2004.

*Author to whom correspondence should be addressed.

[1] Indian Institute of Management, Calcutta, India.

[2] Dept of Mathematics and Statistics, Lederle Graduate Research Tower, University of Massachusetts, Amherst MA, 01003-9305, USA. e-mail: jstauden@math.umass.edu

[3] Dept of Statistics, The University of New South Wales, Sydney NSW 2052, Australia.

Acknowledgments. This work was partially supported by NIH grant T32 ES07142-18 and NSF-DMS 0306227. The authors thank the editor, an associate editor, and an anonymous referee for helpful suggestions and comments.

In this paper we exploit connections between penalized spline smoothing and likelihood-based mixed models (e.g. Ruppert, Wand & Carroll, 2003). With this approach the additive model extension is straightforward, so we work at this level of generality. The measurement error structure involves adjustment of the likelihood, although the resulting likelihood function is intractable. We overcome this problem using a nested Monte Carlo expectation maximization (MCEM) algorithm (Wei & Tanner, 1990; Smyth, 1996; van Dyk, 2000).

Although our paper focuses on smoothing, our methods generally apply to linear mixed models of the form $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\varepsilon}$ where $\boldsymbol{b}$ and $\boldsymbol{\varepsilon}$ are multivariate normally distributed random vectors and the matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ both have mismeasured components. We believe that these are the first methods to generally address the case when the random effect's covariates ($\boldsymbol{Z}$) have measurement error. The case when only $\boldsymbol{X}$ is allowed to be mismeasured is addressed in Buonaccorsi, Demidenko & Tosteson (2000) for instance.

In Section 2 we give the model formulation and likelihood function, and review briefly the connections between penalized spline smoothing and likelihood based mixed models. Section 3 describes the estimation scheme. Section 4 illustrates our approach, through a simulation study and an application of the method to air pollution or mortality data. We conclude with some discussion and mention of extensions in Section 5.

2. Formulation

Let $\boldsymbol{y}$ be an $n \times 1$ vector of responses, and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d$ be $n \times 1$ vectors of predictors for which the Gaussian additive model,

$$\boldsymbol{y} = \beta_0 \mathbf{1} + \sum_{j=1}^{d} m_j(\boldsymbol{x}_j) + \boldsymbol{\varepsilon}\,, \qquad \boldsymbol{\varepsilon} \stackrel{\mathrm{d}}{=} \mathrm{N}_n(\mathbf{0}, \sigma_\varepsilon^2 \boldsymbol{I})\,,$$

is reasonably assumed. The m_j are general smooth functions, modelled through the mixed model formulation of penalized regression splines: $m_j(\boldsymbol{x}_j) = \beta_j \boldsymbol{x}_j + \sum_{k=1}^{K_j} b_k^j (\boldsymbol{x}_j - \kappa_k^j \mathbf{1})_+$, $\boldsymbol{b}_j \stackrel{\mathrm{d}}{=} \mathrm{N}_{K_j}(\mathbf{0}, \sigma_{\boldsymbol{b}_j}^2 \boldsymbol{I})$, independent $(j = 1, \ldots, d)$, where $\boldsymbol{b}_j = (b_1^j, \ldots, b_{K_j}^j)$ (e.g. Eilers & Marx, 1996; Wand, 2003). Note that $\sigma_{\boldsymbol{b}_j}^2 = 0$ corresponds to linearity of m_j. A fixed number of knots are placed on a grid in each dimension $j = 1, \ldots, d$. See Ruppert (2002) for a discussion of knot (κ_k^j) selection.

As an aside, a link between penalized splines and mixed models is reviewed in Ruppert *et al.* (2003 Section 4.9). In the context of a scatterplot smoother $(d = 1)$, for ease of explanation, the penalized spline fitting criterion finds $\boldsymbol{\beta}$ and $\boldsymbol{b}$ to minimize

$$\frac{1}{\sigma_e^2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i - \sum_{k=1}^{K} b_k (x_i - \kappa_k)_+ \right)^2 + \frac{\lambda^2}{\sigma_e^2} \sum_{k=1}^{K} b_k^2\,.$$

When $\boldsymbol{b}$ is treated as a random variable with mean $\mathbf{0}$ and covariance $\sigma_u^2 \boldsymbol{I}$ where $\sigma_u^2 = \sigma_e^2/\lambda^2$, this criterion is equivalent to the best linear unbiased predictor (BLUP) criterion in a linear mixed model. Additional connections between mixed models and splines have been explored and exploited by a number of other authors including Verbyla *et al.* (1999) who drew upon Green & Silverman (1994 Section 3.8) and Speed (1991).

In our situation, some of the $\boldsymbol{x}_j$ are not directly observable, and instead we observe

$$\boldsymbol{w}_j = \boldsymbol{x}_j + \boldsymbol{u}_j\,, \quad \text{where } \boldsymbol{u}_j \stackrel{\mathrm{d}}{=} \mathrm{N}_n(\boldsymbol{0}, \sigma^2_{\boldsymbol{u}_j}\boldsymbol{I}) \text{ independent } (j = 1, \ldots, d)\,,$$

where the $\sigma^2_{\boldsymbol{u}_j}$ are known ($\sigma^2_{\boldsymbol{u}_j} = 0$ corresponds to $\boldsymbol{x}_j$ being observed). Let $f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d; \boldsymbol{\theta})$ denote the joint density of the predictors, where f is determined by the parameter vector $\boldsymbol{\theta}$; this is sometimes called a structural assumption. We also assume that the dependence of $\boldsymbol{y}$ on the predictors is non-differential:

$$(\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_d, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_d) \stackrel{\mathrm{d}}{=} (\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_d)\,.$$

With these specifications,

$$(\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_d) \stackrel{\mathrm{d}}{=} \mathrm{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V}_{\boldsymbol{x}})\,,$$

where
$$\boldsymbol{V}_{\boldsymbol{x}} = \sum_{j=1}^{d} \sigma^2_{\boldsymbol{b}_j} \boldsymbol{Z}_j \boldsymbol{Z}_j^{\mathrm{T}} + \sigma^2_{\boldsymbol{\epsilon}} \boldsymbol{I}\,, \qquad \boldsymbol{\beta} = (\beta_0, \ldots, \beta_d)\,,$$

$$\boldsymbol{X} = [\boldsymbol{1} \;\; \boldsymbol{x}_1 \;\; \cdots \;\; \boldsymbol{x}_d] \quad \text{and} \quad \boldsymbol{Z}_j = [(\boldsymbol{x}_j - \kappa^j_1 \boldsymbol{1})_+ \;\; \cdots \;\; (\boldsymbol{x}_j - \kappa^j_{K_j} \boldsymbol{1})_+]\,.$$

We also use the notation

$$\boldsymbol{x} = [\boldsymbol{x}_1 \;\; \cdots \;\; \boldsymbol{x}_d]\,, \quad \boldsymbol{b} = [\boldsymbol{b}_1 \;\; \cdots \;\; \boldsymbol{b}_d]\,, \quad \boldsymbol{w} = [\boldsymbol{w}_1 \;\; \cdots \;\; \boldsymbol{w}_d] \quad \text{and} \quad \boldsymbol{Z} = [\boldsymbol{Z}_1 \;\; \cdots \;\; \boldsymbol{Z}_d]\,.$$

Further, let $\boldsymbol{x}_{-j}$ represent $\boldsymbol{x}$ without its jth component. Finally, let $\boldsymbol{\sigma}^2_b = (\sigma^2_{\boldsymbol{b}_1}, \ldots, \sigma^2_{\boldsymbol{b}_d})$, $\boldsymbol{\sigma}^2_{\boldsymbol{u}} = (\sigma^2_{\boldsymbol{u}_1}, \ldots, \sigma^2_{\boldsymbol{u}_d})$, and $\boldsymbol{\psi} = (\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2_b, \sigma^2_\varepsilon)$. The likelihood is

$$\begin{aligned}
\mathrm{L}(\boldsymbol{\psi}) &= f(\boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\psi}) \\
&= \int_{\mathbb{R}^n} \cdots \int_{\mathbb{R}^n} f(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{\sigma}^2_b, \sigma^2_{\boldsymbol{\epsilon}})\, f(\boldsymbol{w} \mid \boldsymbol{x}; \boldsymbol{\sigma}^2_{\boldsymbol{u}})\, f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d; \boldsymbol{\theta})\, d\boldsymbol{x} \\
&= \int_{\mathbb{R}^n} \cdots \int_{\mathbb{R}^n} \varphi_{\boldsymbol{V}_{\boldsymbol{x}}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\, f(\boldsymbol{w} \mid \boldsymbol{x}; \boldsymbol{\sigma}^2_{\boldsymbol{u}})\, f(\boldsymbol{x}; \boldsymbol{\theta})\, d\boldsymbol{x}\,,
\end{aligned}$$

where
$$\varphi_{\boldsymbol{\Sigma}}(\boldsymbol{x}) = (2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} \exp(-\tfrac{1}{2} \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{x})$$

denotes the $\mathrm{N}_n(\boldsymbol{0}, \boldsymbol{\Sigma})$ density.

To better appreciate the complexity that measurement error brings to this likelihood, suppose that

$$\boldsymbol{x}_j \stackrel{\mathrm{d}}{=} \mathrm{N}(\mu_j \boldsymbol{1}, \sigma^2_j \boldsymbol{I})\,, \;\; \text{independent } (j = 1, \ldots, d)\,.$$

Then the likelihood is

$$\mathrm{L}(\boldsymbol{\psi}) = \int_{\mathbb{R}^n} \cdots \int_{\mathbb{R}^n} \varphi_{\boldsymbol{V}_{\boldsymbol{x}}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \prod_{j=1}^{d} \varphi_{\sigma^2_{u_j} \boldsymbol{I}}(\boldsymbol{w}_j - \boldsymbol{x}_j) \prod_{j=1}^{d} \varphi_{\sigma^2_j \boldsymbol{I}}(\boldsymbol{x}_j - \mu_j \boldsymbol{1})\, d\boldsymbol{x}\,,$$

and, despite the fact that everything is normal, the integrals cannot be reduced due to the presence of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d$ in $\boldsymbol{V}_{\boldsymbol{x}}$. While a more flexible structural model is often desirable (e.g. Carroll *et al.*, 1984), we present the case of a normal structural model to illustrate the method.

This problem appears to be similar to the one encountered in maximum likelihood estimation in the generalized linear mixed model (e.g. McCulloch & Searle, 2001 Chapter 8). We emphasize that the integrals here are a result of the measurement error and not the mixed model. This observation highlights an interesting connection between the somewhat separate statistical subfields of mixed models and measurement error.

Apart from maximization of $\mathrm{L}(\boldsymbol{\psi})$ for estimation of the parameters in $\boldsymbol{\psi}$, prediction of $\boldsymbol{b}$ is required for estimation of the function $m_j(\cdot)$, $1 \le j \le K$. The best predictor (BP) of $\boldsymbol{b}$ is

$$\hat{\boldsymbol{b}} = \mathrm{E}(\boldsymbol{b} \mid \boldsymbol{y}, \boldsymbol{x}) .$$

A commonly used estimator is the empirical best predictor (eBP), $\hat{\boldsymbol{b}}^j$, the BP calculated with $\boldsymbol{\sigma}_b^2$ and σ_ε^2 fixed at their current estimates, $\hat{\boldsymbol{\sigma}}_b^2$ and $\hat{\sigma}_{\boldsymbol{\epsilon}}^2$, which we estimated by REML. Algebraically, the eBP is

$$\hat{\boldsymbol{b}} = (\boldsymbol{Z}^\top \boldsymbol{Z} + \boldsymbol{D})^{-1} \boldsymbol{Z}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) , \quad \text{where } \boldsymbol{D} = \hat{\sigma}_{\boldsymbol{\epsilon}}^2 \, \mathrm{blockdiag}_{1 \le j \le d} \left(\frac{1}{\hat{\sigma}_{\boldsymbol{b}_j}^2} \boldsymbol{I}_{K_j} \right) .$$

In the case of normal $\boldsymbol{\varepsilon}$ and $\boldsymbol{b}$, the eBP is also unbiased. As a result of this and its form, the eBP is also the empirical best linear unbiased estimator (eBLUP) in this case (e.g. McCulloch & Searle, 2001 Chapter 9).

3. Estimation

As discussed in Section 2, since $\mathrm{L}(\boldsymbol{\psi})$ involves intractable integrals, direct determination of $\hat{\boldsymbol{\psi}}$ to maximize $\mathrm{L}(\boldsymbol{\psi})$ is very difficult. In the somewhat similar case of the generalized linear mixed model (e.g. McCulloch & Searle, 2001 Chapters 8, 10), a tractable approach uses the nested Monte Carlo Expectation Maximization algorithm (nested MC–EM algorithm). Section 3.2 introduces the nested MC–EM algorithm and describes how to generate the required Monte Carlo sample. Section 3.3 then enumerates our algorithm in detail.

3.1. EM algorithm

Denote the observed data as $\boldsymbol{Y}_{\mathrm{obs}} = (\boldsymbol{y}, \boldsymbol{w})$ with log-likelihood $\ell_{\mathrm{obs}}(\boldsymbol{\psi})$. Let $\boldsymbol{Y}_{\mathrm{comp}} = (\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{x}, \boldsymbol{b})$ be an augmentation of $\boldsymbol{Y}_{\mathrm{obs}}$, which we refer to as the complete data; let the associated log-likelihood be denoted by $\ell_{\mathrm{comp}}(\boldsymbol{\psi})$. The EM algorithm maximizes $\ell_{\mathrm{obs}}(\boldsymbol{\psi})$ in $\boldsymbol{\psi}$ using an iterative process comprising two steps. Let the current estimates of the parameters be $\boldsymbol{\psi}^{(t)}$ where (t) indexes iterations. The expectation or E-step consists of computing the 'best guess' (conditional expectation with the expectation indexed by the most recent parameter estimates) of the complete data log-likelihood given the observed data,

$$Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(t)}) = \mathrm{E}\big(\ell_{\mathrm{comp}}(\boldsymbol{\psi}) \mid \boldsymbol{Y}_{\mathrm{obs}}; \boldsymbol{\psi}^{(t)}\big) .$$

Next, the maximization or M-step consists of calculating

$$\boldsymbol{\psi}^{(t+1)} = \operatorname*{argmax}_{\boldsymbol{\psi}} Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(t)}) .$$

When possible, the procedure can be streamlined by computing only the conditional expectations of the sufficient statistics for the complete data problem. Under regularity conditions, $\boldsymbol{\psi}^{(t)} \to \hat{\boldsymbol{\psi}}_{\mathrm{MLE}}$ as $t \to \infty$.

3.2. MCEM algorithm and nesting

One extension to the EM algorithm is Monte Carlo (MC) estimation of the E-step (Wei & Tanner, 1990). This extension replaces analytic (or numerical) determination of the required expectation with a Monte Carlo estimate. We use this method since our E-step does not have a closed form.

The E-step could be implemented by sampling from the distribution of $(\boldsymbol{x}, \boldsymbol{b} \mid \boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\psi}^{(t)})$, using a Gibbs sampler that alternates between samples from $(\boldsymbol{b} \mid \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\psi}^{(t)})$ (multivariate normal) and $(\boldsymbol{x}_j \mid \boldsymbol{b}, \boldsymbol{x}_{-j}, \boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\psi}^{(t)})$, $j = 1, \ldots, d$ (see Fact 1 below), but this approach would probably be quite inefficient. Smyth (1996) and van Dyk (2000) suggest a more efficient approach called the nested EM algorithm. The idea behind that approach is to alternate one of the Gibbs chain elements with a 'nested' EM algorithm that holds Monte Carlo estimates from the Gibbs output fixed. van Dyk (2000) likens the approach to an EM version of blocked Gibbs sampling.

In our case, this approach consists of two steps for each EM iteration (indexed by t) and each additive component (indexed by j). The first step is the Monte Carlo step. It consists of generating a sample from $(\boldsymbol{x}_j \mid \boldsymbol{b}, \boldsymbol{x}_{-j}, \boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\psi}^{(t)})$ and using this sample to compute Monte Carlo estimates of the expectations of the complete data sufficient statistics for $\boldsymbol{\psi}^{(t)}$ that involve $\boldsymbol{x}_j$. The second step is followed by the nested EM step, where we fix the estimates of the complete data sufficient statistics at their current values and compute $\boldsymbol{\psi}^{(t+1)}$ from several iterations of an EM algorithm that treats only $\boldsymbol{b}$ as latent. van Dyk (2000) shows that convergence of the nested EM in each nested step is not necessary for convergence of the complete algorithm. The benefit of this procedure is that it makes more efficient use of computationally expensive Metropolis–Hastings samples. Since the nested step is aimed at fitting a linear mixed model, another method such as Newton–Raphson could be used in the nested step instead of EM.

The following fact shows one way of drawing from the conditional densities of

$$(\boldsymbol{x}_j \mid \boldsymbol{b}, \boldsymbol{x}_{-j}, \boldsymbol{y}, \boldsymbol{w}, \boldsymbol{\psi}^{(t)}) \qquad (j = 1, \ldots, d)\,.$$

This method was also used by Berry *et al.* (2002). It is a direct consequence of the likelihoods that specify our model.

Fact 1. *The density of* $(\boldsymbol{x}_j \mid \boldsymbol{b}, \boldsymbol{x}_{-j}, \boldsymbol{y}, \boldsymbol{w}, \boldsymbol{\psi}^{(t)})$ *is proportional to*

$$\exp\left(-\frac{1}{2\sigma_\varepsilon^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b}) - \sum_{j=1}^{d}\frac{1}{2\sigma_{\boldsymbol{x}_j}^2}(\boldsymbol{x}_j - \boldsymbol{\mu}_j)^{\mathrm{T}}(\boldsymbol{x}_j - \boldsymbol{\mu}_j)\right.$$
$$\left. - \sum_{j=1}^{d}\frac{1}{2\sigma_{\boldsymbol{u}_j}^2}(\boldsymbol{w}_j - \boldsymbol{x}_j)^{\mathrm{T}}(\boldsymbol{w}_j - \boldsymbol{x}_j)\right).$$

As a result, generating from the conditional distribution of $\boldsymbol{x}_j$ can be done using the Metropolis–Hastings algorithm; see e.g. Casella & Robert (1999 Chapter 6).

3.3. Complete algorithm

Now we specify the complete algorithm.

(i) Start with an initial $\boldsymbol{\psi}^{(0)}$. To establish the starting values, we fit a fixed degree-of-freedom additive model using $\mathrm{E}(\boldsymbol{x}_j \mid \boldsymbol{w}_j)$ in place of the observed covariates. These conditional expectations require estimates of $\mu_{\boldsymbol{x}_j}$ and $\sigma^2_{\boldsymbol{x}_j}$, and we estimate them using method of moments estimators: $\hat{\mu}_{\boldsymbol{x}_j} = \mathbf{1}_n^\top \boldsymbol{w}_j / n$ and $\hat{\sigma}^2_{\boldsymbol{x}_j} = (s^2_{\boldsymbol{w}_j} - \sigma^2_{\boldsymbol{u}_j})/s^2_{\boldsymbol{w}_j}$.
With $j = 1, \ldots, d$ indexing the additive components, the $(t+1)$th parameter update involves step (ii).

(ii) Sample from $(\boldsymbol{x}_j \mid \boldsymbol{b}, \boldsymbol{x}_{-j}, \boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\psi}^{(t)})$ using the Metropolis–Hastings algorithm and Fact 1.

(iii) Let

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{X} & \boldsymbol{X}^\top \boldsymbol{Z} \\ \boldsymbol{Z}^\top \boldsymbol{X} & \boldsymbol{Z}^\top \boldsymbol{Z} + \boldsymbol{D} \end{bmatrix}.$$

With the results from step (ii), compute Monte Carlo estimates of

$$\boldsymbol{P},\ \boldsymbol{W}^\top \boldsymbol{y},\quad \text{and} \quad \boldsymbol{W}^\top \boldsymbol{W},\quad \text{where } \boldsymbol{W} = [\boldsymbol{X} \quad \boldsymbol{Z}].$$

We denote estimates of these quantities by

$$\hat{\boldsymbol{P}},\ \hat{\boldsymbol{W}}^\top \boldsymbol{y},\quad \text{and} \quad \hat{\boldsymbol{W}}^\top \hat{\boldsymbol{W}}.$$

(iv) Holding the estimates from step (iii) fixed, run several iterations of an update scheme. For instance, using the standard EM algorithm to compute REML estimates (e.g. Dempster, Rubin & Tsutakawa, 1981), the $(k+1)$th nested updates are:

(a) $\boldsymbol{b}^{(k+1)} = (\boldsymbol{\beta}^{(k+1)}, \boldsymbol{b})$ and $\boldsymbol{b}^{(k+1)} = \hat{\boldsymbol{P}}^{-1} \hat{\boldsymbol{W}}^\top \boldsymbol{y}$.

(b) Letting $\boldsymbol{P}^{-1}_{[\boldsymbol{b}_j]}$ correspond to the block submatrix of $\boldsymbol{P}^{-1}$ associated with $\boldsymbol{b}_j$,

$$\sigma^{2(k+1)}_{\boldsymbol{b}_j} = \frac{1}{K_j} = \left(\boldsymbol{b}_j^\top \boldsymbol{b}_j + \mathrm{tr}(\sigma^{2(k)}_\varepsilon \hat{\boldsymbol{P}}^{-1}_{[\boldsymbol{b}_j]})\right) \qquad (j = 1, \ldots, d).$$

(c) $\sigma^{2(k+1)}_\varepsilon = \dfrac{1}{n}\left(\boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{y}^\top \hat{\boldsymbol{W}} \boldsymbol{b}^{(k+1)} + \boldsymbol{b}^{(k+1)^\top} \hat{\boldsymbol{W}}^\top \hat{\boldsymbol{W}} \boldsymbol{b}^{(k+1)} + \sigma^{2(k)}_\varepsilon \mathrm{tr}(\hat{\boldsymbol{W}}^\top \hat{\boldsymbol{W}} \hat{\boldsymbol{P}}^{-1})\right)$.

(v) Calculate $\mu^{(t+1)}_{\boldsymbol{x}_j}$ and $\sigma^{2(t+1)}_{\boldsymbol{x}_j}$ using standard point estimates based on the Monte Carlo data in step (iii).

We terminate the algorithm after plots of the current estimates of the additive components $(\hat{m}_j^{(t+1)}(\cdot))$ appear to stabilize (Wei & Tanner, 1990).

4. Illustrations

4.1. Simulation

Figure 1 contains the results of a simulation study to illustrate the performance of our method on scatterplot data and explore its sensitivity to different levels of regression and measurement error variance. The simulation design was a 2×2 factorial on measurement error variance and regression error variance. The measurement error variance levels were determined by setting a scale-free measure of measurement error (the reliability ratio $\lambda =$

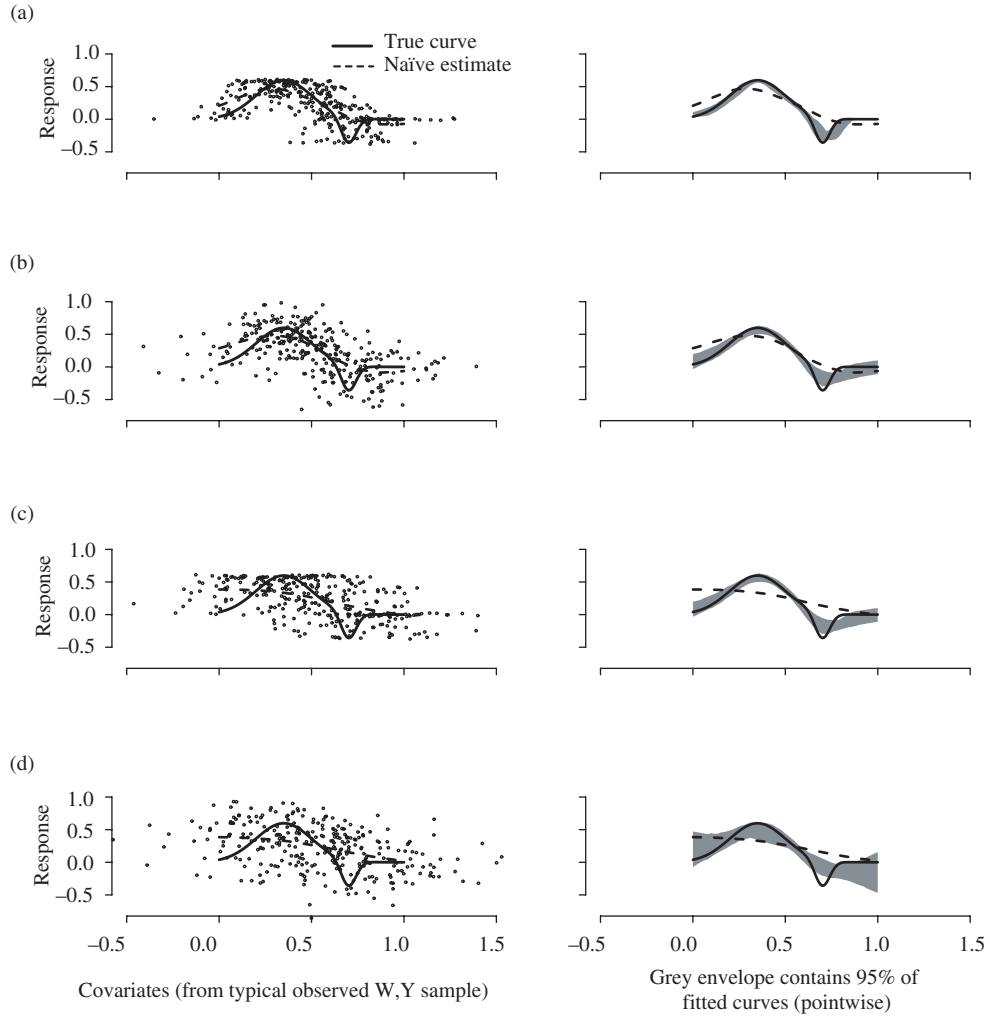

Figure 1. Output from a Monte Carlo simulation study to illustrate the performance of the method. The test cases consist of a 2×2 factorial combination of two levels of measurement error and regression error; 200 replicates; sample size = 300. The left column of panels shows the true curve, the mean estimate that ignores measurement error (naïve estimate), and the observed data. The right column shows the naïve estimate, true curve, and a shaded envelope that covers 95% (pointwise) of the EM-spline estimates. (a) Moderate measurement error and low regression error, (b) moderate measurement error and high regression error, (c) high measurement error and low regression error, and (d) high measurement error and high regression error.

$\sigma_x^2/(\sigma_u^2 + \sigma_x^2)$) to moderate = 0.8 and high = 0.5. The regression error variance was set to a very small amount, $\sigma_\varepsilon^2 = 0.01^2$, and a higher amount, $\sigma_\varepsilon^2 = 0.2^2$. In each case, true $m(x) = 1.5\varphi((x - 0.35)/0.15) - \varphi((x - 0.7)/0.04)$ where φ denotes the normal probability density function, and the sample size was 300. We ran 200 Monte Carlo iterations, and we computed average pointwise squared error for each fitted curve to assess the estimates. Within each simulation case, the 5th, 50th, or 95th percent best fit in Figure 1 is the curve associated with the empirical 5th, 50th, or 95th percent smallest average pointwise squared error.

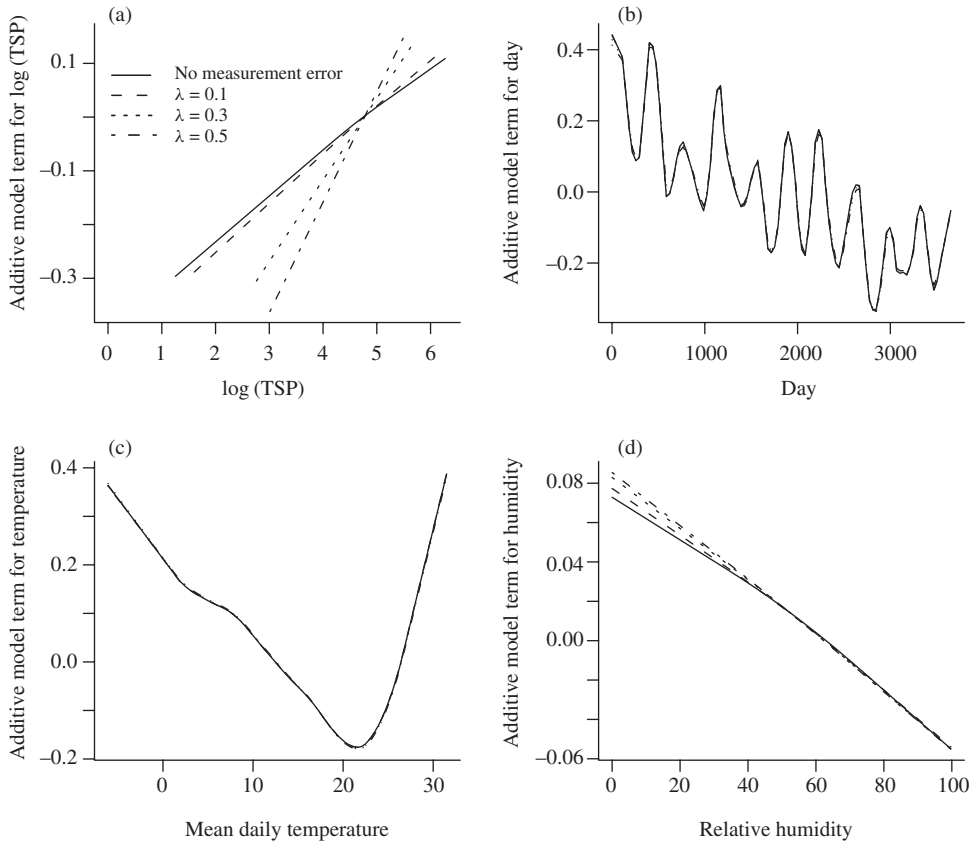

Figure 2. The sensitivity of the fits to four levels of additive covariate measurement error in log(TSP), for an additive model that describes the relationship between mortality and the factors: (a) total suspended particles (log(TSP)), (b) day, (c) temperature, and (d) humidity in Milan, Italy 1980–1989 (Zanobetti *et al.*, 2000)

The figures illustrate that our proposed method identifies features of the true curve that were missed by the estimator that ignores covariate measurement error (the naïve estimator). Even in the presence of 'moderate' measurement error, the naked eye does not immediately identify the true curve from the observed scatterplot data.

4.2. Air pollution example

Next, we fit a simple additive model to an air pollution/mortality dataset recorded in Milan, Italy, from 1980 to 1989 (Zanobetti *et al.*, 2000) and analyse its sensitivity to measurement error in total suspended particles. Given the lack of validation data, our purpose is to show that models such as this one can be sensitive to covariate measurement error. See Dominici & Zeger (2000) for a discussion of measurement error in air pollution/mortality studies.

Let i index day and TSP_i, day_i, temp_i, humid_i, and mort_i denote the measured total suspended particles, sequential day number, average temperature, average relative humidity, and mortality (from natural causes) count on day i. Our additive model regression equation is:

$$\mathrm{E}\big(\ln(\mathrm{mort}_i)\big) = \beta_0 + m_1\big(\ln(\mathrm{TSP}_i)\big) + m_2(\mathrm{day}_i) + m_3(\mathrm{temp}_i) + m_4(\mathrm{humid}_i)\,.$$

To assess the sensitivity of estimates to measurement error in TSP_i, we fit the model assuming that the reliability ratio (see Section 4.1) for $\ln(\mathrm{TSP}_i)$ is 1.0 (no measurement error), 0.9, 0.8, and 0.7.

Figure 2 shows that the estimate of the function that relates ln(TSP) to mortality appears to be sensitive to measurement error. This suggests that an analysis of these data would benefit from measurement error modelling and the collection of validation data.

5. Discussion and extensions

We have described an effective mechanism for additive modelling when predictors are subject to measurement error. Unlike Berry *et al.* (2002), our approach is relative frequentist, and our simulations point to evidence that good performance can be achieved without the specification of priors. We have also considered multipredictor additive models rather than single predictor scatterplot smoothing.

Three straightforward augmentations to our method include calculating approximate standard errors for the curve estimates, incorporating validation data, and allowing a more flexible model for the structural distribution of $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d)$. Since the curve estimates are linear combinations of estimates of $\boldsymbol{\beta}$ and $\boldsymbol{b}$, it is sufficient to estimate the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}$. This can be done using the method described in Section 5 of Chan & Kuk (1997). Validation data can be incorporated, using the method described in Spiegelman *et al.* (2000) for instance. The distribution of the unobserved covariates can be modelled with a mixture of normals, as described in Carroll *et al.* (1999) for instance.

References

BERRY, S., CARROLL, R.J. & RUPPERT, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *J. Amer. Statist. Assoc.* **97**, 160–169.

BUONACCORSI, J., DEMIDENKO, E. & TOSTESON, T. (2000). Estimation in longitudinal random effects models with measurement error. *Statist. Sinica* **10**, 885–903.

CARROLL, R.J., MACA, J.D. & RUPPERT, D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika* **86**, 541–554.

CARROLL, R.J., RUPPERT, D. & STEFANSKI, L.A. (1995). *Measurement Error in Nonlinear Models*. London: Chapman and Hall.

CARROLL, R.J., SPIEGELMAN, C.H., LAN, K.K.G., GAILEY, K.T. & ABBOTT, R.D. (1984). On errors in variables for binary regression models. *Biometrika* **71**, 19–25.

CASELLA, G. & ROBERT, C.P. (1999). *Monte Carlo Statistical Methods*. New York: Springer.

CHAN, J.S.K. & KUK, A.Y.C. (1997). Estimation for probit–linear mixed models with correlated random effects. *Biometrics* **53**, 86–97.

DEMPSTER, A.P., RUBIN, D.B. & TSUTAKAWA, R.K. (1981). Estimation in covariance components models. *J. Amer. Statist. Assoc.* **76**, 341–353.

DOMINICI, F. & ZEGER S.L. (2000). A measurement error model for time–series studies of air pollution and mortality. *Biostatistics* **1**, 157–175.

EILERS, P.H.C. & MARX, B.D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statist. Sci.* **89**, 89–121.

FAN, J. & TRUONG, Y.K. (1993). Nonparametric regression with errors in variables. *Ann. Statist.* **21**, 1900–1925.

FULLER, W.A. (1987). *Measurement Error Models*. New York: John Wiley & Sons.

GREEN, P.J. & SILVERMAN, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall.

MCCULLOCH, C.E. & SEARLE, S.R. (2001). *Generalized, Linear, and Mixed Models*. New York: Wiley.

RUPPERT, D. (2002). Selecting the number of knots for penalized splines. *J. Comput. Graph. Statist.* **11**, 735–757.

RUPPERT, D., WAND, M.P. & CARROLL, R.J. (2003). *Semiparametric Regression.* Cambridge, UK: Cambridge University Press.

SMYTH, G.K. (1996). Partitioned algorithms for maximum likelihood and other non-linear estimation. *Statist. Comput.* **6**, 201–216.

SPEED, T. (1991). Discussion of "That BLUP is a good thing: the estimation of random effects" by G. Robinson. *Statist. Sci.* **6**, 42–44.

SPIEGELMAN, D., ROSNER, B. & LOGAN, R. (2000). Estimation and inference for logistic regression with covariate misclassification and measurement error, in main study/validation study designs. *J. Amer. Statist. Assoc.* **95**, 51–61.

STAUDENMAYER, J. & RUPPERT, D. (2004). Local polynomial regression and simulation extrapolation. *J. Roy. Statist. Soc. Ser. B* **66**, 17–30.

VAN DYK, D.A. (2000). Nesting EM algorithms for computational efficiency. *Statist. Sinica* **10**, 203–226.

VERBYLA, A.P., CULLIS, B.R., KENWARD, M.G. & WELHAM, S.J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines (with discussion). *J. Roy. Statist. Soc. Ser. C* **48**, 269–300.

WAND, M.P. (2003). Smoothing and mixed models. *Comput. Statist.* **18**, 223–249.

WEI, G.C.G. & TANNER, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.* **85**, 699–704.

ZANOBETTI, A., WAND, M.P., SCHWARTZ, J. & RYAN, L.M. (2000). Generalized additive distributed lag models: quantifying mortality displacement. *Biostatistics* **1**, 279–292.