Bayesian Isotonic Regression and Trend Analysis

Brian Neelon[1] and David B. Dunson[2,*]

October 10, 2003

[1] Department of Biostatistics, CB #7420, University of North Carolina at Chapel Hill,
Chapel Hill, NC 27599

[2] Biostatistics Branch, MD A3-03, National Institute of Environmental Health Sciences
P.O. Box 12233, Research Triangle Park, NC 27709

[*] *email*: dunson1@niehs.nih.gov

SUMMARY. In many applications, the mean of a response variable can be assumed to be a non-decreasing function of a continuous predictor, controlling for covariates. In such cases, interest often focuses on estimating the regression function, while also assessing evidence of an association. This article proposes a new framework for Bayesian isotonic regression and order restricted inference. Approximating the regression function with a high dimensional piecewise linear model, the non-decreasing constraint is incorporated through a prior distribution for the slopes consisting of a product mixture of point masses (accounting for flat regions) and truncated normal densities. To borrow information across the intervals and smooth the curve, the prior is formulated as a latent autoregressive normal process. This structure facilitates efficient posterior computation, since the full conditional distributions of the parameters have simple conjugate forms. Point and interval estimates of the regression function and posterior probabilities of an association for different regions of the predictor can be estimated from a single MCMC run. Generalizations to categorical outcomes and multiple predictors are described, and the approach is applied to an epidemiology application.

KEY WORDS: Additive model; Autoregressive prior; Constrained estimation; Monotonicity; Order restricted inference; Smoothing; Threshold model; Trend test.

1. Introduction

In many applications, the mean of a response variable, Y, conditional on a predictor, X, can be characterized by an unknown isotonic function, $f(\cdot)$, and interest focuses on (i) assessing evidence of an overall increasing trend; (ii) investigating local trends (e.g., at low dose levels); and (iii) estimating the response function, possibly adjusted for the effects of covariates, $\mathbf{Z}$. For example, in epidemiologic studies, one may be interested in assessing the relationship between dose of a possibly toxic exposure and the probability of an adverse response, controlling for confounding factors. In characterizing biologic and public health significance, and the need for possible regulatory interventions, it is important to efficiently estimate dose response, allowing for flat regions in which increases in dose have no effect.

In such applications, one can typically assume *a priori* that an adverse response does not occur less often as dose increases, adjusting for important confounding factors, such as age and race. It is well known that incorporating such monotonicity constraints can improve estimation efficiency and power to detect trends (Robertson, Wright, and Dykstra, 1988), and several frequentist approaches have been proposed for smooth monotone curve estimation (Mammen, 1991; Ramsay, 1998). Motivated by the lack of a single framework for isotonic dose response estimation and trend testing, accounting for covariates and flat regions, this article proposes a Bayesian approach.

Consider a regression model, where a response Y is linked to a vector of covariates $\mathbf{X} = (x_1, \ldots, x_p)'$ through an additive structure:

$$Y = \alpha + \sum_{l=1}^{p} f_l(x_l) + \epsilon, \tag{1}$$

where α is an intercept parameter, $f_l(\cdot)$ is an unknown regression function for the lth covariate, and ϵ is a zero-mean error residual. Additive models are appealing since they reduce the problem of estimating a function of the p-dimensional predictor $\mathbf{X}$ to the more manageable problem of estimating p univariate functions $f_l(\cdot)$, one for each covariate x_l.

There is a well developed literature on frequentist approaches for fitting additive models, using a variety of methods to obtain smoothed estimates of each $f_l(\cdot)$ (Hastie and Tibshirani, 1990). In the Bayesian setting, methods have been proposed for curve estimation using piecewise linear or polynomial splines (Denison, Mallick, and Smith, 1998; Holmes and Mallick, 2001). A prior distribution is placed on the number and location of knots and estimation proceeds via reversible jump Markov chain Monte Carlo (MCMC) (Green, 1995). However, these methods do not incorporate monotonicity or shape restrictions on the regression functions.

In the setting of estimating a potency curve, Gelfand and Kuo (1991) and Ramgopal, Laud, and Smith (1993) proposed nonparametric Bayesian methods for dose response estimation under strict constraints. Lavine and Mockus (1995) considered related methods for continuous response data, allowing nonparametric estimation of the mean regression curve and residual error density. These methods focus on estimation subject to strict monotonicity constraints, and cannot be used directly for inferences on flat regions of the dose response curve.

To address this problem, Holmes and Heard (2003) recently proposed an approach for Bayesian isotonic regression using a piecewise constant model with unknown numbers and locations of knots. Posterior computation is implemented using a reversible jump MCMC algorithm, which proceeds without considering the constraint. To assess evidence of a monotone increasing dose response function, they compute Bayes factors based on the proportions of draws from the unconstrained posterior and unconstrained prior for which the constraint is satisfied. The resulting tests are essentially comparisons of the hypothesis of monotonicity to the hypothesis of any other dose response shape.

This article proposes a fundamentally different approach, based on a piecewise linear model with prior distributions explicitly specified to have restricted support and to allow flat regions of the dose response curve. We choose the number of potential knots to be large (e.g., corresponding to the unique values of the predictor observed in the data set), and then specify a prior that allows for autocorrelation in the interval-specific slopes while adaptively dropping out knots that are not needed. In addition to estimation of the regression function, our focus is on comparing the null hypothesis of a flat regression function to the alternative that there is at least one increase. Our prior for the slope parameters takes the form of a product mixture of point masses at zero, which allow for flat regions, and truncated normal densities. This structure is related to priors for Bayesian variable selection in linear regression (Geweke, 1996; George and McCulloch, 1997). However, motivated by the problem of simplifying computation while smoothing the regression function, we propose a novel latent Markov process formulation.

Section 2 proposes the model, prior structure, and MCMC algorithm. Section 3 presents the results from a simulation study. Section 4 applies the methods to data from an epidemiologic study of pesticide exposure and preterm birth, and Section 5 discusses the results.

2. The Model

2.1 *Piecewise Linear Isotonic Regression*

We focus initially on the univariate normal regression model,

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{2}$$

where $f \in \Theta^+$ is an unknown isotonic regression function, with

$$\Theta^+ = \{f : f(x_1) \leq f(x_2) \; \forall (x_1, x_2) \in \Re^2 : x_1 < x_2\}$$

denoting the space of non-decreasing functions, and $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$ an error residual for the ith subject. Modifications for $f \in \Theta^-$, the space of non-increasing functions, are straightforward.

We approximate $f(\cdot)$ using a piecewise linear model,

$$f(x_i) \approx \alpha + \sum_{j=1}^{k} w_j(x_i)\beta_j = \alpha + \sum_{j=1}^{k} w_{ij}\beta_j = \boldsymbol{w}_i'\boldsymbol{\theta}, \quad j = 1, \ldots, k, \quad i = 1, \ldots, n, \tag{3}$$

where α is an intercept parameter, $w_{ij} = w_j(x_i) = \min(x_i, \gamma_j) - \gamma_{j-1}$ if $x_i \geq \gamma_{j-1}$ and $w_{ij} = 0$ otherwise, $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_k)'$ are knot locations (with $x_i \in [\gamma_0, \gamma_k]\, \forall i$), β_j is the slope within interval $(\gamma_{j-1}, \gamma_j]$, $\boldsymbol{w}_i = (1, w_{i1}, \ldots, w_{ik})'$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$, and $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}')'$.

The conditional likelihood of $\mathbf{y} = (y_1, \ldots\, y_n)'$ given $\boldsymbol{\theta}$, σ^2 and $\boldsymbol{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ can therefore be written as

$$L(\mathbf{y}|\boldsymbol{\theta}, \sigma^2, \boldsymbol{x}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \boldsymbol{w}_i'\boldsymbol{\theta})^2 \Big\}. \tag{4}$$

This likelihood can potentially be maximized subject to the constraint $\beta_j \geq 0$, for $j = 1, \ldots, k$, to obtain estimates satisfying the monotonicity constraint. However, the resulting restricted maximum likelihood estimates of the regression function will not be smooth, and there can be difficulties in performing inferences, since the null hypothesis of no association falls on the boundary of the parameter space.

2.2 *Autoregressive Mixture Prior*

We instead follow a Bayesian approach, choosing a prior distribution that (1) restricts the curve $f(\cdot)$ to be non-decreasing; and (2) allows for flat regions with positive probability. We assume that $\boldsymbol{\gamma}$ is a high dimensional vector of potential knot locations, possibly corresponding to the unique values of X in the data set or to a tightly-spaced grid. To specify a prior distribution for $\boldsymbol{\beta}$ with the appropriate properties, we first let $\beta_j = 1_{(\beta_j^* \geq \delta)} \beta_j^*$, where β_j^* is a latent slope variable and δ is a small positive constant, which can be viewed as a threshold below which the latent slope β_j^* is effectively zero. We then choose an autoregressive prior for $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_k^*)'$ by letting

$$\pi(\boldsymbol{\beta}^*) = \pi(\beta_1^*) \prod_{j=2}^{k} \pi(\beta_j^* \,|\, \beta_{j-1}^*) = \mathrm{N}(\beta_1^*; E_{01}; V_{01}) \prod_{j=2}^{k} \mathrm{N}(\beta_j^*; \beta_{j-1}^*, \lambda^{-1}),$$

where E_{01} is an investigator-specified best guess for the average slope of $f(x)$ across the range of X, V_{01} measures uncertainty in this guess, and λ is a hyperparameter controlling the degree of smoothing.

Since δ is positive, each of the slopes $\beta_j = 1_{(\beta_j^* \geq \delta)} \beta_j^*$ will be non-negative, ensuring that property (1) holds. In addition, by assigning $\beta_j = 0$ when the latent β_j^* has a value less than δ, we place a point mass at $\beta_j = 0$ which ensures property (2). Letting $E_{0j} = \beta_{j-1}^*$ and $\mathrm{V}_{0j} = \lambda^{-1}$ for $j = 2, \ldots, k$, the prior density for $(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$ can be expressed as $\pi(\boldsymbol{\beta}, \boldsymbol{\beta}^*) =$

$$\prod_{j=1}^{k} \pi(\beta_j \,|\, \beta_j^*) \pi(\beta_j^* \,|\, \beta_{j-1}^*) = \prod_{j=1}^{k} \Big\{ 1_{(\beta_j = 0)} 1_{(\beta_j^* < \delta)} + 1_{(\beta_j = \beta_j^*)} 1_{(\beta_j^* \geq \delta)} \Big\} \mathrm{N}(\beta_j^*; E_{0j}, V_{0j}). \tag{5}$$

For $j = 2, \ldots, k$, the conditional prior density for β_j given β_{j-1}^* is

$$\pi(\beta_j \,|\, \beta_{j-1}^*) = 1_{(\beta_j = 0)} F(\delta; \beta_{j-1}^*, \lambda^{-1}) + 1_{(\beta_j \geq \delta)} \mathrm{N}(\beta_j; \beta_{j-1}^*, \lambda^{-1}), \tag{6}$$

where $F(\cdot, \mu, \sigma^2)$ is the cumulative normal distribution function.

Density (6) is the mixture of a point mass at $\beta_j = 0$ and a truncated normal distribution. Related mixture priors have been proposed for variable selection in linear regression (Geweke, 1996; George and McCulloch, 1997; Raftery, Madigan, and Hoeting, 1997; Chipman, George and McCulloch, 2001). Our problem is different from the typical variable selection problem in that the point mass at $\beta_j = 0$ corresponds to the case in which the regression function $f(x)$ is flat across the jth interval, $(\gamma_{j-1}, \gamma_j]$. The special case where $\beta_1 = \ldots = \beta_k = 0$ corresponds to the global null hypothesis, $\mathrm{H}_0 : f(x)$ is constant for all $x \in [\gamma_0, \gamma_k]$, and we will be interested in assessing evidence of H_0 versus $\mathrm{H}_1 : f(\gamma_0) < f(\gamma_k)$. Motivated by this problem, we have incorporated autocorrelation to smooth the regression function and borrow information across adjacent intervals. In contrast, in the variable selection literature, the most common strategy is to assume *a priori* independence and set the point mass probabilities equal to 0.5.

The hyperparameter λ measures the degree of autocorrelation in the elements of $\boldsymbol{\beta}$, and hence controls the degree of smoothing. In contrast, the hyperparameter δ controls the point mass probabilities, with values of δ close to zero assigning relatively low probability to flat regions of $f(x)$ compared with higher value of δ. To limit sensitivity of inferences to the specification of these hyperparameters, we choose hyperprior distributions to allow uncertainty in their elicitation: $\pi(\lambda, \delta) = \pi(\lambda)\pi(\delta)$, where $\pi(\lambda) = \mathcal{G}(\lambda; c_1, d_1)$ and $\pi(\delta) = \mathcal{G}(\delta; c_2, d_2)$ are independent gamma densities. There is information in the data about λ and δ, and the posterior distributions can differ substantially from the priors. Issues in elicitation and sensitivity to hyperprior specification will be discussed in Sections 3 and 4. We complete a Bayesian specification of the model with a normal conjugate prior for the intercept: $\pi(\alpha) = \mathrm{N}(\alpha; \alpha_0, \sigma_\alpha^2)$, and a gamma conjugate prior for the error precision: $\pi(\sigma^{-2}) = \mathcal{G}(\sigma^{-2}; a, b)$, where $\alpha_0, \sigma_\alpha^2, a$ and b are investigator-specified hyperparameters.

2.3 *Posterior Computation*

The joint posterior density of the parameters $\{\boldsymbol{\theta}, \sigma^2\}$, hyperparameters $\{\lambda, \delta\}$, and latent variables $\{\boldsymbol{\beta}^*\}$ conditional on the data $\{\mathbf{y}, \mathbf{x}\}$ is proportional to

$$L(\mathbf{y} \mid \boldsymbol{\theta}, \sigma^2, \mathbf{x}) \Big[\prod_{j=1}^{k} \Big\{ 1_{(\beta_j = 0)} 1_{(\beta_j^* < \delta)} + 1_{(\beta_j = \beta_j^*)} 1_{(\beta_j^* \geq \delta)} \Big\} \mathrm{N}(\beta_j^*; E_{0j}, V_{0j}) \Big] \pi(\alpha, \sigma^{-2}, \lambda, \delta). \qquad (7)$$

The full conditional posterior distributions for α, σ^{-2}, (β_j, β_j^*), and λ all have simple conjugate forms, which can be derived from (7) following standard algebraic routes (refer to Appendix A). However, the posterior of δ is difficult to sample from directly. If δ were known, posterior computation could proceed via a simple Gibbs sampling algorithm. To allow for unknown δ, we propose to use an MCMC algorithm which iterates between the following steps after assigning initial values to the parameters and latent variables:

1. Sample α, σ^{-2} and λ from their conjugate full conditional distributions.

2. Sample a candidate value for δ from a proposal density centered on the previous value of δ (e.g., normal).

3. Sample candidate values for $(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$ given the candidate δ by sampling sequentially from the full conditional distributions shown in Appendix A.

4. Accept or reject the candidate for δ and $(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$ in one block with the typical Metropolis-Hastings (Hastings, 1970) probability.

5. Repeat steps 1-4 until apparent convergence and calculate posterior summaries based on a large number of additional iterations.

We have observed good computational efficiency for this algorithm in simulated and real data examples, with good rates of convergence and low to moderate levels of autocorrelation in the samples. Since the unknown curve $f(\cdot)$ is characterized by a high dimensional vector of constrained parameters $\boldsymbol{\beta}$, efficient posterior computation is very challenging. The prior structure proposed in subsection 2.2 was carefully chosen not only for its appealing theoretical properties but also to facilitate simple and efficient computation.

The posterior probability of H_0 can be estimate directly from the output of the Gibbs sampler by using $\widehat{\pi} = \frac{1}{S}\sum_{s=1}^{S} 1(\beta_1^{(s)} = \beta_2^{(s)} = \cdots = \beta_k^{(s)} = 0)$, where $s = 1, \ldots, S$ denotes the iterations collected after burn-in, and $\boldsymbol{\beta}^{(s)}$ is the value of $\boldsymbol{\beta}$ at iteration s. The approach of averaging model indicators to estimate posterior model probabilities was described by Carlin and Chib (1995). Similar approaches can be used to obtain pointwise credible intervals for $f(x)$ and to estimate posterior probabilities for local null hypotheses (e.g., $\mathrm{H}_{0j} : \beta_j = 0$). A quantity which is often of interest is the dose level corresponding to the first increase in $f(x)$. Among the observed values of $x \in \mathbf{x}$, this threshold dose is $\tau = \min_i \{x_i : x_i \in (\gamma_{j-1}, \gamma_j], \beta_j > 0\}$. A posterior distribution for τ can be estimated directly from $\tau^{(s)}$, $s = 1, \ldots, S$.

2.4 Extensions to Multiple Predictors

The piecewise linear model can easily be extended to accommodate multiple predictors through the additive structure outlined in equation (1). In particular, let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ denote a $p \times 1$ vector of predictors with corresponding regression functions $f_1(x_{i1}), \ldots, f_p(x_{ip})$, and let $z_{i1}, \ldots, z_{iq}$ denote a $q \times 1$ vector of additional covariates, which are assumed to have linear effects. Note that some of the regression functions $f_l(\cdot)$ $(l = 1, \ldots, p)$ may be unconstrained while others are assumed monotonic. If a particular regression function is unconstrained, we can fit the piecewise linear approximation described above without incorporating the constraint.

The mean regression function for the ith response, y_i, can be expressed in terms of the piecewise linear approximation (3) through an additive structure:

$$\begin{aligned} E(y_i \,|\, \boldsymbol{\theta}, \boldsymbol{x}_i, \boldsymbol{z}_i) &= \alpha_0 + \sum_{h=1}^{q} \alpha_h z_{ih} + \sum_{l=1}^{p} f_l(x_{il}) = \boldsymbol{z}_i'\boldsymbol{\alpha} + \sum_{l=1}^{p}\sum_{j=1}^{k} w_j(x_{il})\beta_{jl} \\ &= \boldsymbol{z}_i'\boldsymbol{\alpha} + \sum_{l=1}^{p} \boldsymbol{w}_{il}'\boldsymbol{\beta}_l = \boldsymbol{w}_i'\boldsymbol{\theta}, \quad i = 1, \ldots, n, \end{aligned} \tag{8}$$

where $\boldsymbol{z}_i = (1, z_{i1}, \ldots, z_{iq})'$, $\boldsymbol{w}_{il} = (w_1(x_{il}), \ldots, w_k(x_{il}))'$, $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_q)'$, $\boldsymbol{\beta}_l = (\beta_{l1}, \ldots, \beta_{lk})'$, $\boldsymbol{w}_i = (\boldsymbol{z}_i', \boldsymbol{w}_{i1}', \ldots, \boldsymbol{w}_{ip}')'$ and $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_p')'$. Posterior computation can proceed by direct extension of the MCMC algorithm outlined in the previous subsection. In particular, letting λ_l, δ_l denote the values of λ, δ for the lth unknown regression function, for $l = 1, \ldots, p$, the parameters $\boldsymbol{\alpha}$, σ^{-2}, and λ_l can be sampled directly from their full conditional posterior distributions, while the parameters δ_l and $(\boldsymbol{\beta}_l, \boldsymbol{\beta}_l^*)$ can be updated in blocks using Metropolis-Hastings steps.

2.5 *Probit Models for Categorical Data*

It is straightforward to extend this approach for categorical y_i by following the approach of Albert and Chib (1993). Suppose that $y_1, \ldots, y_n$ are independent Bernoulli random variables, with

$$\Pr(y_i \,|\, \boldsymbol{\theta}, \boldsymbol{x}_i, \boldsymbol{z}_i) = \Phi\Big(\boldsymbol{z}_i'\boldsymbol{\alpha} + \sum_{l=1}^{p} \boldsymbol{w}_{il}'\boldsymbol{\beta}_l\Big) = \Phi(\mathbf{w}_i'\boldsymbol{\theta}). \tag{9}$$

As noted by Albert and Chib (1993), model (9) is equivalent to assuming that $y_i = 1_{(y_i^* > 0)}$, with $y_i^* \sim \mathrm{N}(\mathbf{w}_i'\boldsymbol{\theta}, 1)$ denoting independent and normally distributed random variables underlying y_i, for $i = 1, \ldots, n$. Posterior computation can proceed via an MCMC algorithm that alternates between (i) sampling from the conditional density of y_i^*, $\pi(y_i^* \,|\, y_i, \boldsymbol{\theta}, \boldsymbol{x}_i, \boldsymbol{z}_i) \overset{d}{=}$

$\mathrm{N}(\mathbf{w}_i'\boldsymbol{\theta}, 1)$ truncated below (above) by 0 for $y_i = 1$ ($y_i = 0$), and (ii) applying the algorithm of subsections 2.3-2.4.

3. Simulation Study

To study the behavior of the proposed procedure, we conducted a simulation study, applying the method to data generated from three $N(f(x), 1)$ models, with $f(x)$ chosen to be (i) constant, (ii) piecewise constant with a threshold and (iii) sinusoidal. In each case, $n = 200$ covariate values were generated from a uniform distribution over the range $(0, 10)$. For each scenario, we applied the Ramsay (1998) smooth monotone function estimator, and our proposed Bayesian approach. We chose somewhat vague priors for the intercept (α), the error variance (σ^2), and the initial latent slope (β_1^*) by letting $\alpha_0 = 0.0$, $\sigma_\alpha^2 = 10$, $a = b = 0.1$, $E_{01} = 0.0$, and $V_{01} = 10$. To specify a prior for λ and δ, we let $c_1 = k/25$, $d_1 = 1$, $c_2 = 1.25$, and $d_2 = 25$. The prior for λ was chosen so that the expected level of autocorrelation in the latent slopes was moderately high corresponding to a smooth curve, and this autocorrelation increased with the number of knots. In the analyses, we set k equal to $n - 1 = 199$ and placed a knot at each data point. The prior for δ was chosen to have $\mathrm{E}(\delta) = 0.05$, which was considered to be effectively 0 in the context of the data application of Section 4. The variance was chosen to allow a moderate degree of uncertainty in this choice.

We ran the MCMC algorithm for 100,000 iterations with a burn-in of 1500, and retained every 100th sample to reduce autocorrelation. Standard MCMC diagnostics, such as trace plots, suggested rapid convergence and efficient mixing. For each simulated data set, we calculated (1) the estimated posterior probability of H_0: no change in $f(x)$ across the range of x_i; and (2) pointwise posterior means and 95% credible intervals at each of the observed values of the predictor.

The estimates are presented in Figures 1 - 3 for the flat, threshold, and sinusoidal cases, respectively. In each case, the posterior mean was very close to the true curve and the estimated posterior probability of the null hypothesis assigned high probability to the correct hypothesis. For the flat curve shown in Figure 1, our Bayesian estimate was flat, the true curve was contained in the 95% credible interval, and we estimated $\Pr(\mathrm{H}_0 \mid \mathrm{data}) = 0.76$. In contrast, the Ramsay (1998) estimator exhibited a slight increasing trend, and no formal hypothesis test statistic was available. For the threshold curve shown in Figure 2, we estimated $\Pr(\mathrm{H}_0 \mid \mathrm{data}) < 0.01$, and our Bayesian estimator again provided a close approximation to the true function. Although our estimator and the Ramsay estimator both smoothed out the discontinuous jump at $x = 8$ as expected, our estimator did better at capturing this shape. For the sinusoidal-type curve shown in Figure 3, we estimated $\Pr(\mathrm{H}_0 \mid \mathrm{data}) < 0.01$, and our estimator again provided an excellent approximation to the true function. Although the Ramsay (1998) estimator also did well, there was some over smoothing.

In summary, our proposed approach did very well in each of the simulated cases we considered, including a variety of cases not presented here (linear and quadratic true curves, different sample sizes, different choices of prior for λ and δ, binary response). Based on these analyses, the point and interval estimators for the curve appear to be quite robust to the prior specification. This robustness is most likely due to the incorporation of hyperprior distributions for λ and δ that allow the data to inform about the degree of smoothing and flatness of the curve. Analyses that assumed known λ and/or δ were not as robust.

4. Application: Effect of DDE on Preterm Birth

To motivate our approach, we re-analyzed data from a recent study by Longnecker et al. (2001) examining the effect of DDE exposure on the risk of preterm birth. DDE is a metabolite of DDT, a pesticide still commonly used in many developing countries. The sample comprised 2380 pregnant women who were enrolled as part of the US Collaborative Perinatal Project, a multi-site prospective study of infant health conducted in the 1960s. Out of the 2380 pregnancies, there were 361 preterm births. Serum DDE concentrations (in $\mu g/l$) were collected for each woman in the sample, along with potentially confounding maternal characteristics, such as cholesterol and triglyceride levels, age, BMI and smoking status (yes or no).

The aim of our analysis was to incorporate a non-decreasing constraint on the regression function relating level of DDE to the probability of preterm birth in order to improve efficiency in estimating the function and to increase the power to detect a trend without assuming linearity. In addition, study investigators had a strong *a priori* belief that the curve should be non-decreasing, and wanted an estimate consistent with this assumption.

For comparative purposes, we first conducted a frequentist regression analysis using an unconstrained probit generalized additive model (GAM), with preterm birth as the response and with the predictors consisting of a DDE level x_i and the control variables mentioned above. Specifically, letting $y_i = 1$ indicate preterm birth for ith woman, we assumed that:

$$\Pr(y_i = 1 \,|\, \boldsymbol{\theta}, \boldsymbol{x}_i, \boldsymbol{z}_i) = \Phi\Big(\alpha_0 + \sum_{l=1}^{5} z_{il}\alpha_l + f(x_i)\Big) = \Phi(\boldsymbol{z}_i'\boldsymbol{\alpha} + f(x_i)), \tag{10}$$

where z_{i1}, z_{i2}, z_{i3}, z_{i4} and z_{i5} are covariate values representing, respectively, cholesterol level, triglyceride level, age, BMI, and smoking status (yes or no), $f(x)$ is an unconstrained, nonparametric regression function relating DDE level x_i to the response y_i, $\boldsymbol{z}_i = (1, z_{i1}, \ldots, z_{i5})'$ and $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_5)'$. All continuous variables were normalized prior to analysis.

We then implemented our proposed Bayesian approach to analysis of model (10), incorporating a non-decreasing constraint on $f(x)$. To specify the prior, we first chose diffuse $N(0, 10)$ priors for the regression coefficients for the covariates $\mathbf{z}_i$. Priors for the remaining parameters and details on computational implementation are as outlined in Section 3. Again, standard MCMC convergence diagnostics were conducted to verify apparent convergence and adequate mixing.

The estimated regression function relating DDE to the probability of preterm birth, adjusting for covariate effects, is presented in Figure 4. The estimated posterior mean under our proposed Bayesian approach is similar to the unconstrained frequentist estimate from the GAM analysis. Both estimates show a non-linear increase in the probability of preterm birth with DDE dose, with the curve gradually flattening out as dose increases. Although a frequentist trend test from a linear-logistic model analysis was significant, the frequentist GAM analysis produced a non-significant p-value of 0.23. In contrast, our Bayesian analysis showed clear evidence of an increasing dose response curve, with the estimated $\Pr(\mathrm{H}_0 \,|\, \mathrm{data}) < 0.01$. In addition, using our Bayesian approach, we can estimate the minimum dose at which there is an increase in the function relative to the value at the lowest dose observed: $\widehat{\tau} = 7.03$, 95% credible interval $= [3, 21]$.

The flattening out pattern in the dose response curve is interesting from both public health and biological perspectives. First, it suggests that modest reductions in the level of exposure for highly exposed women may have little impact on the risk of preterm birth for these women, while similar reductions for women with more moderate exposures may have a substantial impact. There are several plausible biological explanations for the flattening out pattern. First, women may vary in their sensitivity due to genetic factors (e.g., involved in DDE metabolism) so that a subset of more robust women may not have a preterm birth even if they are exposed to very high levels of DDE. Second, women who were susceptible to high DDE exposures may have had an early loss or spontaneous abortion rather than a preterm birth (Longnecker et al., 2003). Although such survival effects can even cause a downturn in the dose response function in extreme cases, it is very unlikely that the magnitude of the DDE effect on pregnancy loss observed in this study was large enough to cause such a downturn. Finally, there may have been a biological threshold effect in which the system was saturated when DDE dose achieved a certain level.

5. Discussion

This article has proposed a Bayesian approach for inference on an unknown regression function, which is known *a priori* to be non-decreasing but may have flat regions across which increases in the predictor have no effect. This approach has several advantages over previous methods for estimation of smooth monotone functions. First, by allocating positive prior probability to flat regions instead of forcing the function to be strictly increasing, we potentially reduce over-estimation bias, which can occur for isotonic regression estimators even when the true curve is strictly increasing, particularly when the curve increases gradually and sample sizes are small to moderate. A primary motivation for the prior structure is that by allowing flat regions, we assign positive probability to null hypotheses of no effect of the predictor overall and within particular regions (e.g., low doses). Autocorrelation is incorporated to borrow information across the different regions and smooth the curve. Due to the simple conditionally-conjugate structure, it is straightforward to implement an MCMC algorithm to efficiently generate samples from the posterior distribution. Posterior probabilities of null and alternative hypotheses, locations of thresholds (e.g., corresponding to the first increase in the regression function), and point and interval estimates of the function for any value of the predictor can be calculated directly from these samples.

The DDE application has illustrated some of the appealing features of the approach, and the simulation study suggests that the procedure is robust to different prior specifications, due to the incorporation of hyperpriors to account for uncertainty in prior elicitation. It is straightforward to apply the methodology to general hierarchical models, and extensions can be made to allow the smoothness of the curve to vary across different regions of the predictor. This may be useful when prior information is available about specific regions across which rapid changes are more likely to occur. However, in the absence of such information, the incorporation of multiple smoothing parameters may lead to problems with over-parameterization.

Another important extension would be to cases with more complex order restrictions. For example, one may want to account for the possibility of a downturn in the regression function at higher values of a predictor. Such downturns may occur when toxicity occurring at high doses leads to a reduction in the response of interest, either directly or through an intermediate outcome. For example, an adverse chemical exposure may actually decrease cancer risk at high doses due to reduced survival, lowered body weight, or a direct effect of the chemical on killing cancer cells. An important area of future research is the development of efficient methods for posterior computation when unknown changepoints are incorporated in the regression function.

Acknowledgements

The authors would like to thank Zhen Chen and Sean O'Brien for helpful comments and discussions. Thanks also to Matthew Longnecker for providing the data, and an anonymous referee for important suggestions.

References

Albert, J.H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**, 669-679.

Carlin, B.P. and Chib, S. (1995). Bayesian model choice via Markov-chain Monte-Carlo methods. *Journal of the Royal Statistical Society Series B* **57**, 473-484.

Chipman, H., George, E.I., and McCulloch, R.E. (2001). The practical implementation of Bayesian model selection. *IMS Lecture Notes - Monograph Series* **38**.

Denison, D.G.T., Mallick, B.K., and Smith, A.F.M. (1998). Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society Series B* **60**, 333-350.

Gelfand, A.E. and Kuo, L. (1991). Nonparametric Bayesian bioassay including ordered polytomous response. *Biometrika* **78**, 657-666.

George, E.I. and McCulloch, R.E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339-373. item Geweke, J. (1996). Variable selection and model comparison in regression. In *Bayesian Statistics* **5**, J.O. Berger, J.M. Bernardo, A.P. Dawid, and A.F.M. Smith (eds.). Oxford University Press, 609-620.

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, **82**, 711-732.

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models.* London: Chapman & Hall.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-109.

Holmes, C.C. and Heard, N.A. (2003). Generalised monotonic regression using random change points. A revised version has appeared in *Statistics in Medicine* **22**, 623-638.

Holmes, C.C. and Mallick, B.M. (2001). Bayesian regression with multivariate linear splines. *Journal of the Royal Statistical Society, Series B* **63**, 3-17.

Lavine, M. and Mockus, A. (1995). A nonparametric Bayes method for isotonic regression. *Journal of Statistical Planning and Inference* **46**, 235-248.

Longnecker, M.P., Klebanoff, M.A., Zhou, H. and Brock, J.W. (2001). Association between maternal serum concentration of the DDT metabolite DDE and preterm and small-for gestational-age babies at birth. *The Lancet* **258**, 110-114.

Longnecker M.P., Klebanoff M.A., Dunson D.B., Guo X., Chen Z., Zhou H., Brock J.W.

(2003). Maternal serum level of the DDT metabolite DDE in relation to fetal loss in previous pregnancies. *Environmental Research.* In press.

Mammen, E. (1991). Estimating a smooth monotone regression function. *Annals of Statistics* **19**, 724-740.

Raftery, A.E., Madigan, D., and Hoeting, J.A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92**, 179-191.

Ramgopal, P., Laud, P.W. and Smith, A.F.M. (1993). Nonparametric Bayesian bioassay with prior constraints on the shape of the potency curve. *Biometrika* **80**, 489-498.

Ramsey, J.O. (1998). Estimating smooth monotone functions. *Journal of the Royal Statistical Society B* **60**, 365-375.

Robertson, T., Wright, F. and Dykstra, R. (1988) *Order Restricted Statistical Inference.* New York: Wiley.

Appendix A

Conditional Posterior Distributions

The conditional distribution of the error precision σ^{-2} follows a conjugate form

$$\pi(\sigma^{-2}\,|\,\boldsymbol{\theta}, \boldsymbol{\beta}^*, \lambda, \delta, \mathbf{y}, \mathbf{x}) = \mathcal{G}\Big(\sigma^{-2}; a + \frac{n}{2}, b + \frac{1}{2}(\mathbf{y} - \boldsymbol{W}\boldsymbol{\theta})'(\mathbf{y} - \boldsymbol{W}\boldsymbol{\theta})\Big),$$

where $\boldsymbol{W} = (\boldsymbol{w}_1', \ldots, \boldsymbol{w}_n')'$. The full conditional for α also follows a conjugate form:

$$\pi(\alpha\,|\,\boldsymbol{\beta}, \boldsymbol{\beta}^*, \sigma^2, \lambda, \delta, \mathbf{y}, \mathbf{x}) = \mathrm{N}(\alpha; \widehat{\alpha}, \widehat{\sigma}_\alpha^2),$$

where $\widehat{\sigma}_\alpha^2 = (\sigma_0^{-2} + n\sigma^{-2})^{-1}$ and $\widehat{\alpha} = \widehat{\sigma}_\alpha^2\Big\{\sigma_0^{-2}\alpha_0 + \sigma^{-2}\sum_{i=1}^n(y_i - \mathbf{w}_{i(-1)}'\boldsymbol{\beta})\Big\}$, with $\mathbf{w}_{i(-1)} = (w_{i1}, \ldots, w_{ik})'$. The full conditional posterior distribution of λ is proportional to

$$\pi(\lambda\,|\,\boldsymbol{\theta}, \boldsymbol{\beta}^*, \sigma^{-2}, \delta, \mathbf{y}, \mathbf{x}) = \mathcal{G}\Big(c_1 + \frac{k-1}{2}, d_1 + \frac{1}{2}\sum_{j=2}^k(\beta_j^* - \beta_{j-1}^*)^2\Big),$$

The full conditional posterior distribution of β_j and β_j^* is proportional to

$$\Big[\prod_{i=1}^n \mathrm{N}(y_{ij}^*; w_{ij}\beta_j, \sigma^2)\Big]\Big\{1_{(\beta_j=0)}1_{(\beta_j^*\le\delta)} + 1_{(\beta_j=\beta_j^*)}1_{(\beta_j^*\ge\delta)}\Big\}\mathrm{N}(\beta_j^*; E_{0j}, V_{0j})\mathrm{N}(\beta_{j+1}^*; \beta_j^*, \lambda^{-1}), (A.1)$$

where $y_{ij}^* = y_i - \mathbf{w}_{i(-j)}'\boldsymbol{\theta}_{(-j)}$, and $\mathbf{w}_{i(-j)}$ is the subvector of $\mathbf{w}_i$ with w_{ij} excluded. After straightforward but extensive algebra, one can show that (A.1) is proportional to

$$1_{(\beta_j=0)}1_{(\beta_j^*\le\delta)}\Big(\frac{\mathrm{N}(\beta_j^*; \widetilde{E}_{0j}, \widetilde{V}_{0j})}{\mathrm{N}(0; \widetilde{E}_{0j}, \widetilde{V}_{0j})}\Big) + 1_{(\beta_j=\beta_j^*)}1_{(\beta_j^*\ge\delta)}\Big(\frac{\mathrm{N}(\beta_j^*; \widehat{E}_j, \widehat{V}_j)}{\mathrm{N}(0; \widehat{E}_j, \widehat{V}_j)}\Big),$$

where $\widetilde{V}_{0j} = (V_{0j}^{-1} + \lambda)^{-1}$, $\widetilde{E}_{0j} = \widetilde{V}_{0j}(V_{0j}^{-1}E_{0j} + \lambda\beta_{j-1}^*)$, $\widehat{V}_j = \Big(\widetilde{V}_{0j} + \sigma^{-2}\sum_{i=1}^n w_{ij}^2\Big)^{-1}$, and $\widehat{E}_j = \widehat{V}_j\Big(\widetilde{V}_{0j}^{-1}\widetilde{E}_{j0} + \sigma^{-2}\sum_{i=1}^n w_{ij}y_{ij}^*\Big)$. The normalizing constant for this distribution is

$$C = \frac{F(\delta; \widetilde{E}_{j0}, \widetilde{V}_{j0})}{\mathrm{N}(0; \widetilde{E}_{j0}, \widetilde{V}_{j0})} + \frac{1 - F(\delta; \widehat{E}_j, \widehat{V}_j)}{\mathrm{N}(0, \widehat{E}_j, \widehat{V}_j)} = A + B.$$

It follows directly that the full conditional distribution of β_j^* is a mixture of a $\mathrm{N}(\beta_j^*; \widetilde{E}_{j0}, \widetilde{V}_{j0})$ density truncated above by δ (with probability A/C) and a $\mathrm{N}(\beta_j^*; \widehat{E}_j, \widehat{V}_j)$ density truncated below by δ (with probability B/C).

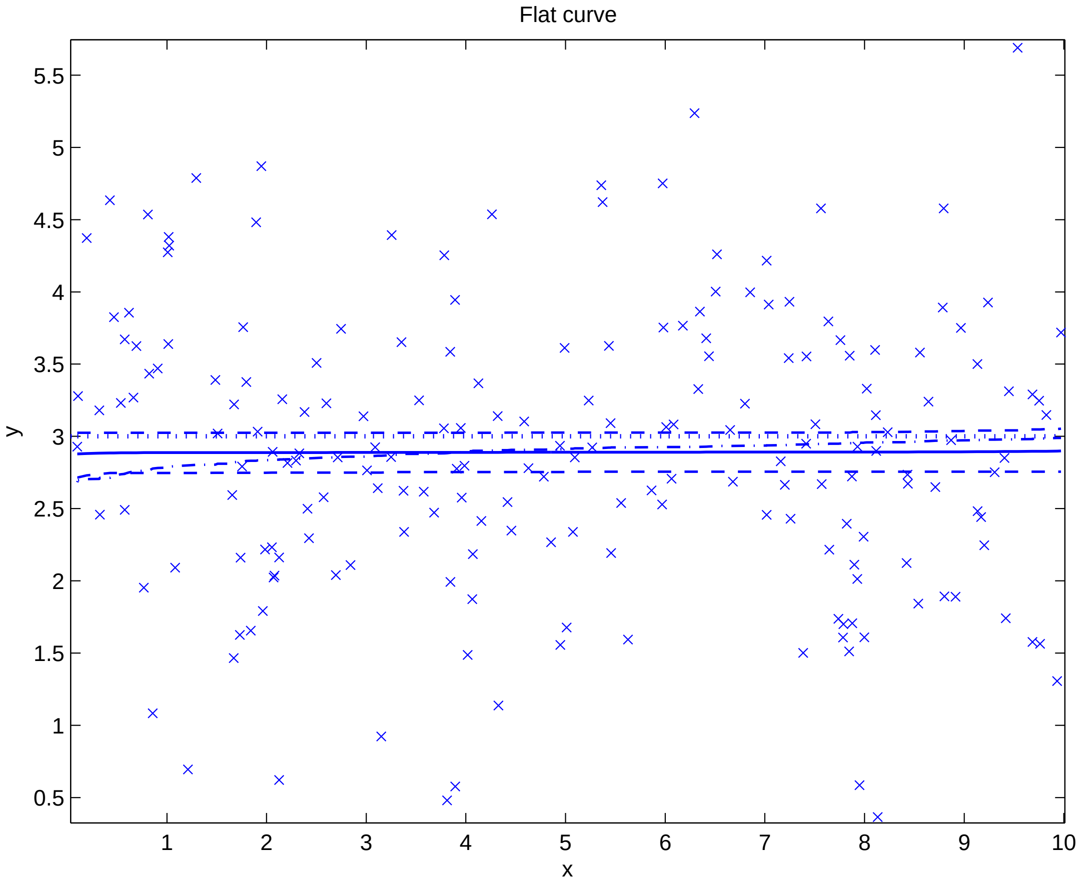
Figure Captions:

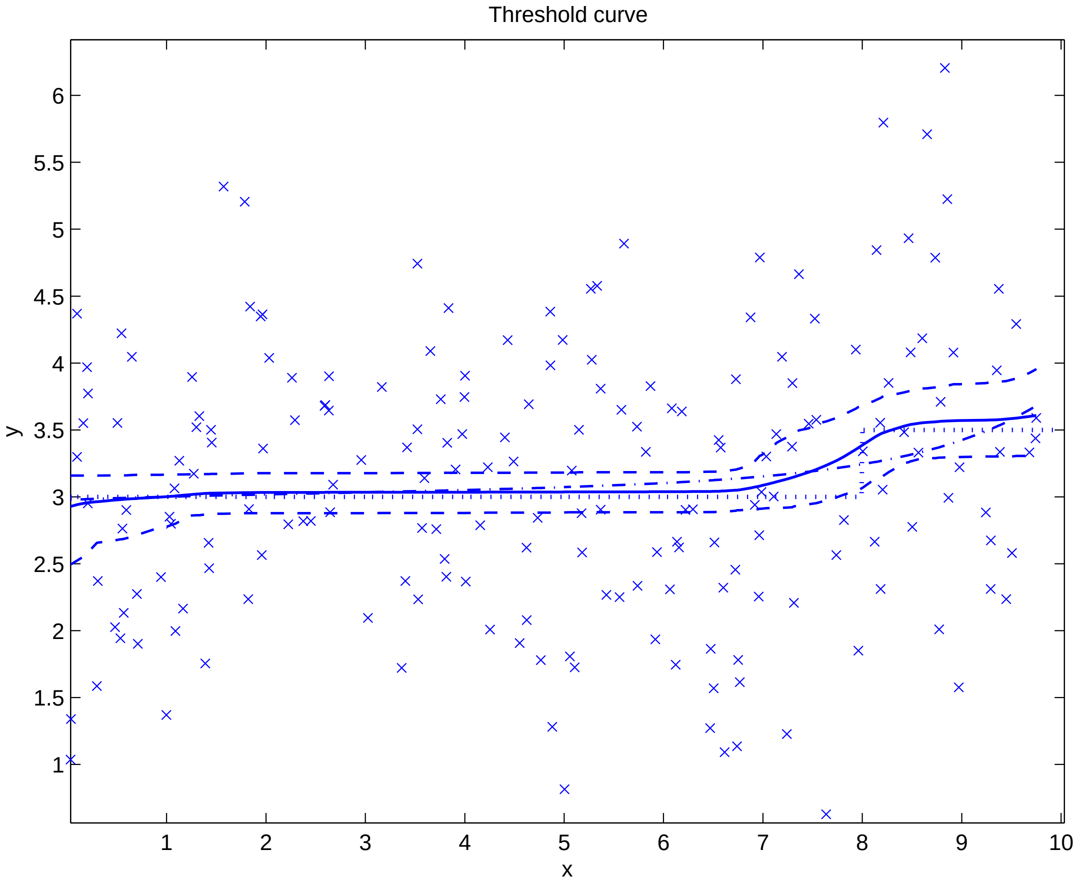
Figure 1. Estimates for data simulated under $f(x) = 3$. Solid line is posterior mean, dotted line is true curve, dashed lines are 95% pointwise credible intervals, and dashed-dotted lines are the Ramsay (1998) estimator.

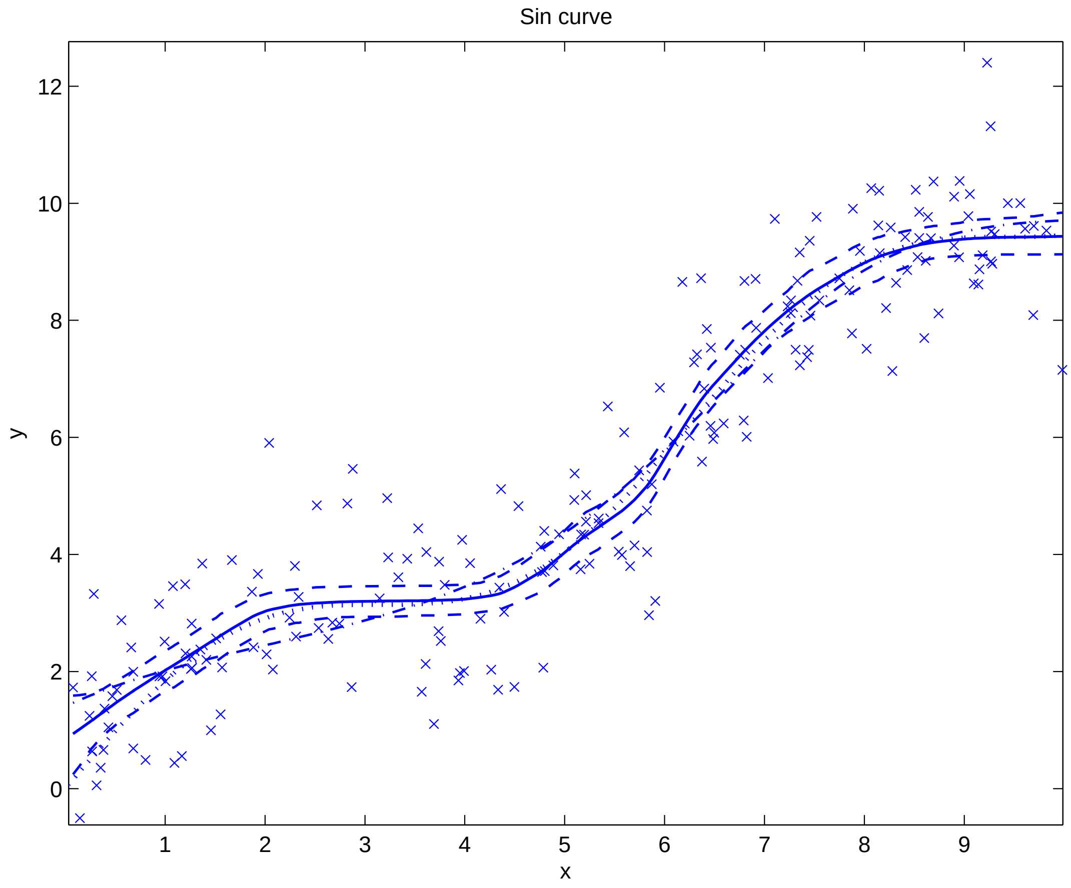
Figure 2. Estimates for data simulated under $f(x) = 3 + .5 \times I_{(x>8)}$. Solid line is posterior mean, dotted line is true curve, dashed lines are 95% pointwise credible intervals, and dashed-dotted lines are the Ramsay (1998) estimator.

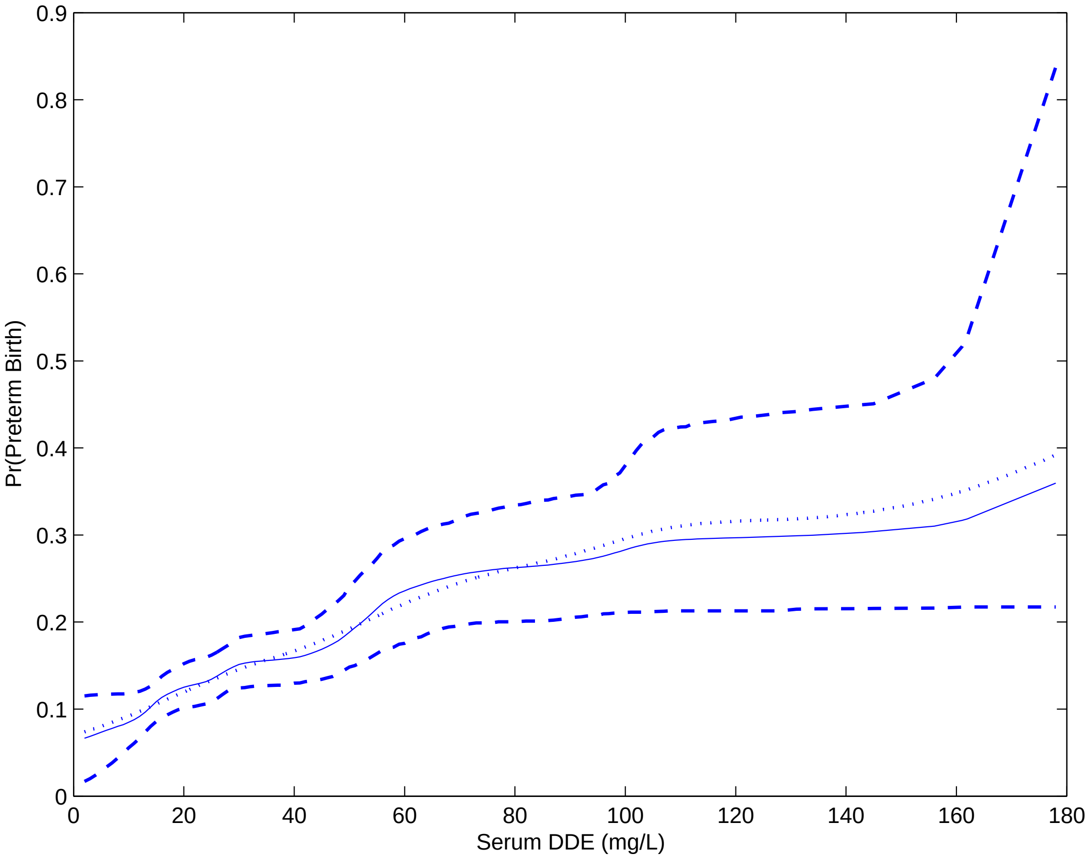
Figure 3. Estimates for data simulated under $f(x) = x + sin(x)$. Solid line is posterior mean, dotted line is true curve, dashed lines are 95% pointwise credible intervals, and dashed-dotted lines are the Ramsay (1998) estimator.

Figure 4. Estimated probability of preterm birth as a function of dde dose. The solid line is the posterior mean for the proposed Bayesian approach, dashed lines are 95% pointwise credible intervals, and the dotted line is an unconstrained frequentist GAM estimator.


Flat curve
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
y
1
2
3
4
5
6
7
8
9
10
x


Threshold curve
y
x
6
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
1
2
3
4
5
6
7
8
9
10


Sin curve
12
10
8
6
4
2
0
y
1
2
3
4
5
6
7
8
9
x


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Pr(Preterm Birth)
0
20
40
60
80
100
120
140
160
180
Serum DDE (mg/L)