

# A Variable Selection Approach to Monotonic Regression with Bernstein Polynomials

S. McKay Curtis

`smcurtis@stat.washington.edu`

University of Washington, Department of Statistics

Box 354322

Seattle, WA 98195-4322, USA

Sujit K. Ghosh

`ghosh@stat.ncsu.edu`

North Carolina State University, Department of Statistics

February 8, 2010

## Abstract

One of the standard problems in statistics consists of determining the relationship between a response variable and a single predictor variable through a regression function. Background scientific knowledge is often available that suggests the regression function should have a certain shape (e.g., monotonically increasing or concave) but not necessarily a specific parametric form. Bernstein polynomials have been used to impose certain shape restrictions on regression functions. The Bernstein polynomials are known to provide a smooth estimate over equidistant knots. Bernstein polynomials are used in this paper due to their ease of implementation, continuous differentiability, and theoretical properties. In this work, we demonstrate a connection between the monotonic regression problem and the variable selection problem in the linear model. We develop a Bayesian procedure for fitting the monotonic regression model by adapting currently available variable selection procedures. We demonstrate the effectiveness of our method through simulations and the analysis of real data.

**KEYWORDS:** Markov Chain Monte Carlo; shape-restricted inference; stochastic search; nonparametric regression; Gibbs sampling; generalized linear models

# 1 Introduction

Much of applied statistics consists of determining the relationship among several variables. This relationship is often represented by assuming a response variable is a stochastic function of one or more predictor variables. Substantive background information often exists that dictates the functional relationship between the response and predictors should exhibit certain shape restrictions. Examples of this can be found in nearly all areas of applied statistics. In biomedical applications, dose response models are assumed to be non-decreasing with the possibility that the dose-response relationship is flat over certain regions. In distance sampling (Buckland et al., 2001), the probability that an observer collecting data will detect a hidden object is assumed to be monotonically decreasing in the distance between the object and the observer. In survival analysis, the survivor function is non-increasing. In reliability and survival analysis, the hazard rate and the failure rate are often assumed to have a “U” shape (Reboul, 2005). In actuarial studies, the mean residual life function  $m(x)$  must satisfy the shape restrictions  $m'(x) + 1 \geq 0$  and  $m(x) \geq 0$ . In microeconomics, the assumption of diminishing marginal returns of factor inputs restricts the production possibilities frontier to be concave. Similarly, in the theory of the consumer, the assumption of diminishing marginal rate of substitution restricts indifference curves to be convex (Nicholson, 1992). In microarray analysis, the time-course expression of a virus gene is thought to follow a unimodal shape with an initial expression of zero, followed by a period of increasing expression, and then a period of decreasing expression (Chien et al., 2009).

The general regression model with shape restriction can be described as follows:

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \dots, n,$$

where  $(x_i, y_i)$  denotes a pair of independently observed predictor and response variables and the regression function  $f(x) = E[Y|X = x]$  is assumed to belong to a class of (smooth) func-

tions having some restrictions on their shapes. The errors are assumed to be independently and identically distributed (iid) with mean 0 and fixed but unknown variance not depending on  $x$ .

Shape restricted regression has a long history in the literature. Studies on shape-restricted regression begin with Hildreth (1954) who proposes a method for estimating points along a concave function and Brunk (1955) who presents maximum likelihood estimators for estimating monotone restrictions on parameters. Barlow et al. (1972) propose the pool-adjacent-violators (PAV) algorithm for fitting an isotonic regression function, which averages over adjacent observations that do not satisfy the isotonicity restraint. Gallant and Golub (1984) explore the use of Fourier flexible forms to impose shape restrictions. Friedman and Tibshirani (1984) combine methods from a running average smoother and the PAV algorithm to obtain a smooth isotonic regression estimator. Ramsay (1988) derives a basis for a monotone regression spline by integrating  $M$ -splines. Mukerjee (1988) proposes kernel methods for smoothing the isotonic estimator of Brunk (1958). Mammen et al. (2001) propose a projection method for constraining smoothers and demonstrate their method on a remarkable array of examples. Hall and Huang (2001) develop a method that monotonizes standard kernel smoothers by the inclusion of a discrete probability distribution on the values of the predictor variable. More recently, Wang and Li (2008) use cubic smoothing splines for isotonic regression. They derive conditions on the parameters of cubic smoothing splines that enforce isotonicity and provide a second-order cone programming algorithm to estimate the parameters. However, one major limitation of these aforementioned methods is that there is no universal method to obtain confidence bands (even based on asymptotic theory) around the regression curve estimate that preserve the shape restriction. One of the advantages of our proposed method is that confidence bands are readily obtained using the posterior quantiles.

The approach we take in this paper uses Bernstein polynomials to approximate the un-

known regression function. Stadtmüller (1986) first used Bernstein polynomials to approximate an unknown regression function, and Tenbusch (1997) generalized this method to the case of more than one predictor. Brown and Chen (1999) adapt the weights of a Bernstein polynomial smoother using kernels of a beta density to obtain a regression function estimate.

Tenbusch (1997) proved many desirable theoretical properties of nonparametric regression function with Bernstein polynomials. These properties include pointwise consistency, asymptotic normality, and uniform consistency of the Bernstein regression estimator. In particular, Tenbusch (1994, 1997) showed that the mean squared error of Bernstein polynomials is comparable to so-called kernel estimators in density estimation and nonparametric regression, respectively. The Bernstein estimate differs from kernel estimate in that the shape of the weight function (or kernel, in this case) changes with the value of  $x$ . Ghosal (2001) has proved theoretical convergence rates of Bernstein polynomials in the related problem of nonparametric density estimation.

The Bernstein polynomial is a natural choice for shape-restricted regression. Consider a Bernstein polynomial of degree  $M$  for the continuous function  $f(\cdot)$

$$B_M(x) = \sum_{k=0}^M f(k/M) \binom{M}{k} x^k (1-x)^{M-k},$$

for  $x \in [0, 1]$ . If we let  $\beta_k = f(k/M)$ , then it is easy to see that restricting the coefficients  $\beta_k$  is equivalent to restricting the values of the function  $f$  at points  $(k/M)$  ( $k = 0, \dots, M$ ) in the domain of  $f$ . (In regression settings where the predictor variable can take values outside the interval  $[0, 1]$ , the predictor variables may be transformed to lie in the interval  $[0, 1]$ . Thus, for the rest of this paper, we assume that predictor variables have been suitably transformed.)

One of the many remarkable properties of Bernstein polynomials is that all of the derivatives possess the same convergence properties. Also, the form of the derivatives of  $B_M(\cdot)$  can

be useful in determining appropriate shape restrictions for a particular application. The  $\ell^{\text{th}}$  derivative of  $B_M(\cdot)$  can be written

$$B_M^{(\ell)}(x) = M(M-1)\cdots(M-\ell+1) \sum_{k=0}^{M-\ell} (\nabla^{(\ell)}\beta_k) \binom{M-\ell}{k} x^k (1-x)^{M-\ell-k}$$

where  $\nabla^{(1)}\beta_k = \beta_{k+1} - \beta_k$ ,  $\nabla^{(2)}\beta_k = \nabla^{(1)}\beta_{k+1} - \nabla^{(1)}\beta_k = \beta_{k+2} - 2\beta_{k+1} + \beta_k$ , and  $\nabla^{(\ell)}\beta_k = \nabla^{(\ell-1)}\beta_{k+1} - \nabla^{(\ell-1)}\beta_k$ . The first and second derivatives simplify to

$$B_M'(x) = M \sum_{k=0}^{M-1} (\beta_{k+1} - \beta_k) \binom{M-1}{k} x^k (1-x)^{M-1-k}$$

$$B_M''(x) = M(M-1) \sum_{k=0}^{M-2} (\beta_{k+2} - 2\beta_{k+1} + \beta_k) \binom{M-2}{k} x^k (1-x)^{M-2-k},$$

respectively. Because the quantities  $\binom{M}{k} x^k (1-x)^{M-k}$  are non-negative for any  $x \in [0, 1]$ , it is easy to see how to construct restrictions on the coefficients that will impose restrictions on the first and second derivatives.

Chak et al. (2005) introduce the idea of using Bernstein polynomials for shape-restricted regression (although Chang et al., 2005, use Bernstein polynomials to model the cumulative hazard function in survival analysis). They give restrictions on the coefficients  $\beta_k$  that enforce the following shapes:

- Nonnegativity:  $\beta_k \geq 0$  for all  $k$
- Isotonicity:  $\beta_k \leq \beta_{k+1}$  for  $k = 0, \dots, M-1$
- Concavity:  $\beta_{k+1} - 2\beta_k + \beta_{k-1} < 0$  for  $k = 1, \dots, M-1$

In addition to monotonicity, Chang et al. (2007) give restrictions on the coefficients that enforce the following shapes:

- Unimodality:  $\beta_0 = \dots = \beta_{r_1} < \beta_{r_1+1} \leq \dots \leq \beta_{r_2}$  and  $\beta_{r_2} \geq \beta_{r_2+1} \geq \dots \geq \beta_{r_3} > \beta_{r_3+1} = \dots = \beta_M$  for  $0 \leq r_1 < r_2 < r_3 \leq M$
- Unimodal concavity:  $\beta_1 - \beta_0 > 0$ ,  $\beta_M - \beta_{M-1} < 0$ , and  $\beta_{k+1} - 2\beta_k + \beta_{k-1} \leq 0$  for  $k = 1, \dots, M - 1$

Unlike the B-splines procedure (which may require quadratic constraints on the coefficients), the Bernstein polynomial procedure can impose monotonicity and other constraints on  $f$  by a simple linear constraint of the form  $\mathbf{D}\boldsymbol{\beta} \geq c$ , where  $\mathbf{D}$  is a suitable difference matrix and  $c$  is a vector of constants. Furthermore, the Bernstein basis has optimal shape preserving properties of all polynomials of the same degree (c.f., Carnicer and Peña, 1993).

We take a Bayesian approach to estimating a shape-restricted regression function. Other authors have developed Bayesian techniques. Lavine and Mockus (1995) propose a nonparametric Bayesian estimator of an isotonic regression function based on the Dirichlet process. Holmes and Heard (2003) use an unconstrained, piecewise-constant function with random knots to model the regression function. To enforce monotonicity, they develop a reversible-jump MCMC algorithm (Green, 1995) to sample from the posterior and throw away posterior draws that do not satisfy the constraint. Such an approach may require a huge amount of sampling as the constraint can often be violated making the algorithm inefficient. Bornkamp and Ickstadt (2009) use the nonparametric Bayesian priors of Ongaro and Cataneo (2004) to put a prior on a monotone dose-response function.

The Bayesian method we present in this paper is similar to the method of Neelon and Dunson (2004). Their approach uses a piecewise linear regression function with constraints on the slope parameters to ensure the regression function is isotonic. A piecewise linear function, however, may be a crude approximation to the unknown regression function. In addition, the MCMC algorithm they propose requires a large number of simulations to explore the posterior; they report only being able to keep every 100<sup>th</sup> iteration of their

Markov chains to reduce the autocorrelation in their posterior sample (Neelon and Dunson, 2004, page 401).

The approach that we take in this paper is to construct a prior distribution for the coefficients  $\beta_0, \dots, \beta_M$  in  $g$  that satisfies the requisite shape restrictions. In this paper we present a prior on the Bernstein coefficients for the case of monotonic regression, although the idea can be extended to other shape restrictions and to generalized linear models (e.g. logistic or probit regression, and Poisson regression). We show that through a simple reparametrization of the  $\beta_k$  coefficients in the Bernstein polynomial expansion, the problem of monotonic regression is equivalent to the problem of variable selection (Section 2). By adapting the variable selection model of Geweke (1996), we obtain a monotonic regression fit through Gibbs sampling (Sections 3 and 4). We demonstrate the usefulness of this prior in several simulated examples (Section 6). We conclude the paper with an analysis of two real data sets and some final comments (Sections 7 and 8).

## 2 Prior on Bernstein Coefficients

### 2.1 A simple prior

Consider, again, the Bernstein polynomial approximation to the unknown regression function

$$B_M(x, \boldsymbol{\beta}) = \sum_{k=0}^M \beta_k \binom{M}{k} x^k (1-x)^{M-k} = \sum_{k=0}^M \beta_k b_M(x, k). \quad (1)$$

Our task is to define a prior on  $\boldsymbol{\beta} = (\beta_0, \dots, \beta_M) \in \mathbb{R}^{M+1}$  such that  $\beta_0 \leq \dots \leq \beta_M$ . A first attempt (see Chang et al., 2007) would be to define a continuous prior (e.g., a normal distribution) on auxiliary variables  $U_0, \dots, U_M$ , then set  $\beta_0 = U_{(0)}, \dots, \beta_M = U_{(M)}$ , where  $U_{(0)}, \dots, U_{(M)}$  are order statistics for the  $U_i$ 's. This approach, however, does not give positive probability to events such as  $\beta_k = \beta_{k+1}$ —events which can be crucial when modeling flat

portions of a regression function.

This difficulty is easily demonstrated using a simulated example. Consider the model

$$Y_i = f(X_i) + \varepsilon_i,$$

where  $i = 1, \dots, 50$ ,  $f(x) = 1$  for all  $x \in [0, 1]$ ,  $X_i \sim \text{Unif}(0, 1)$ , and  $\varepsilon_i \sim \text{N}(0, 0.1^2)$ . We simulated one data set from this model and fitted the model

$$Y_i | \boldsymbol{\beta}, \sigma^2 \sim \text{N}(B_M(x_i, \boldsymbol{\beta}), \sigma^2)$$

$$\sigma^{-2} \sim \text{Gam}(0.01, 0.01)$$

$$\beta_0 = U_{(0)} \cdots \beta_M = U_{(M)}$$

$$U_0, \dots, U_M \sim \text{N}(\alpha, \nu^2)$$

$$\alpha \sim \text{N}(0, 1)$$

$$\nu^{-2} \sim \text{Gam}(1.0, 1.0)$$

to the data, where  $\text{Gam}(a, b)$  is a gamma distribution with mean  $\frac{a}{b}$  and  $\text{N}(m, s^2)$  is the normal distribution with mean  $m$  and variance  $s^2$ , and  $M = 25$ . The model was fitted using an MCMC algorithm constructed by the WinBUGS software (Spiegelhalter, Thomas, and Best, 1999) with 3 chains, 5,000 draws per chain, after 1,000 draws of burn-in, and a plot of the model fit is displayed in Figure 1. The thick, black, solid line is the posterior median from the above model and the thin, black, solid lines are pointwise 95% credible bands. For reference, the solid white line is the true regression function. Also, the dashed, thick, black line is the posterior median and the gray shaded area represent pointwise credible bands from the model we describe in this paper. It is clearly apparent that the order-statistic prior is inadequate to capture the flatness of the regression function. The prior has essentially

forced the estimates of the regression function to have positive slope, whereas our method captures the true regression function in the credible bands.

[Figure 1 about here.]

## 2.2 A Reparametrization

In order to define a prior for shape-restricted regression that allows for events  $\beta_{k+1} = \beta_k$  or, more generally, events  $\beta_{k_1} = \beta_{k_1+1} = \dots = \beta_{k_2}$  for  $0 \leq k_1 \leq k_2 \leq M$ , we reparametrize  $B_M(x, \boldsymbol{\beta})$  by taking the differences between adjacent coefficients. In other words, we obtain new parameters  $\mathbf{u} = (u_0, u_1, \dots, u_M)$  where  $u_0 = \beta_0$  and  $u_k = \beta_k - \beta_{k-1}$  for  $k = 1, \dots, M$ . Using this reparametrization we obtain

$$B_M(x, \boldsymbol{\beta}) = \sum_{j=0}^M \beta_j b_M(x, j) = \sum_{j=0}^M \left( \sum_{k=0}^j u_k \right) b_M(x, j) = \sum_{k=0}^M u_k w_M(x, k) = B_M(x, \mathbf{u})$$

where  $b_M(x, j) = \binom{M}{j} x^j (1-x)^{M-j}$  and  $w_M(x, k) = \sum_{j=k}^M b_M(x, j)$ . Notice that  $w_M(x, 0) = \sum_{j=0}^M b_M(x, j) = 1$  for an  $x \in [0, 1]$ . It easily follows that, under the reparametrization, the Bernstein polynomial regression function has simplified to a linear combination of the parameters  $u_0, \dots, u_M$  and ‘‘covariates’’  $w_M(x, 1), \dots, w_M(x, M)$ , where  $w_M(x, 1), \dots, w_M(x, M)$  can be computed with the incomplete beta function

$$w_M(x, k) = \sum_{j=k}^M b_M(x, j) = F_{B(k, M-k+1)}(x)$$

where  $F_{B(k, M-k+1)}(\cdot)$  is the cdf of a Beta random variable with parameters  $k$  and  $M - k + 1$ . Under this parametrization, if  $u_k = 0$  then  $\beta_k = \beta_{k-1}$ , and if  $u_k > 0$  then  $\beta_k > \beta_{k-1}$  and vice versa.

We note that, technically, Bernstein polynomials cannot be perfectly flat on any interval  $(a, b)$ , except in the trivial case where  $\beta_k = \beta_{k+1}$  for  $k = 0, \dots, M - 1$ . In practice, how-

ever, this is not a serious problem. We can measure the total variation of the Bernstein polynomial regression function over an interval  $(a, b)$  by  $S_{(a,b)}(\boldsymbol{\beta}) = B_M(b, \boldsymbol{\beta}) - B_M(a, \boldsymbol{\beta}) = \int_a^b B'_M(x, \boldsymbol{\beta}) dx$ , which can be written as

$$\begin{aligned} \int_a^b B'_M(x, \boldsymbol{\beta}) dx &= \int_a^b M \sum_{k=0}^{M-1} (\beta_{k+1} - \beta_k) \binom{M-1}{k} x^k (1-x)^{M-1-k} dx \\ &= M \sum_{k=0}^{M-1} (\beta_{k+1} - \beta_k) B(k+1, M-k) [F_{B(k+1, M-k)}(b) - F_{B(k+1, M-k)}(a)] \quad (2) \end{aligned}$$

If the true regression function were flat on the interval  $(a, b)$ , then the sum in (2) will be small. To see this, note that for  $k \in (Ma, Mb)$  we would expect  $\beta_k \approx \beta_{k+1}$  if the underlying true regression function were flat on  $(a, b)$  making the terms  $(\beta_{k+1} - \beta_k)$  small. For  $k$  not in  $(Ma, Mb)$  the terms  $(\beta_{k+1} - \beta_k)$  need not be small, but for  $k < Ma$  we would expect that  $F_{B(k+1, M-k)}(b) \approx F_{B(k+1, M-k)}(a) \approx 1$ , and for  $k > Mb$  we would expect that  $F_{B(k+1, M-k)}(b) \approx F_{B(k+1, M-k)}(a) \approx 0$  making the terms  $[F_{B(k+1, M-k)}(b) - F_{B(k+1, M-k)}(a)]$  small.

We compute (2) in our simulations of Section 6 and show that when the true regression function is known to be flat over an interval the total variation of the Bernstein polynomial regression function is not substantially different from zero.

## 2.3 Variable Selection and Monotonic Regression

The reparametrization above suggests some form of variable selection may be used to determine the fit of the monotonic regression function. A variable selection prior on the  $u$ 's would induce a prior on the  $\beta$ 's with positive probability of adjacent  $\beta$ 's being exactly equal, which would satisfy the monotonicity constraint while allowing for relatively flat portions of the regression function.

The literature contains several studies on Bayesian variable selection (e.g., Mitchell and

Beauchamp (1988), George and McCulloch (1993), George and McCulloch (1997), George and Foster (2000), Cripps et al. (2006), Casella and Moreno (2006)). In a situation similar to our own, Dunson (2005) reparametrizes the regression function for count data by using ratios of adjacent function values. Dunson then uses a mixture of a point-mass at one and a gamma distribution truncated to be greater than one to impose a monotonic constraint to the regression function.

The approach of Geweke (1996) is particularly germane to the problem of monotonic regression as outlined in the previous section. To perform variable selection in the linear model, Geweke uses a prior for each regression coefficient that is a mixture of a point mass at zero and a truncated normal distribution. This mixture prior is a natural fit for our situation in which the prior for  $u_k$  must satisfy two requirements—events  $u_k = 0$  must have positive probability and  $u_k \geq 0$  with probability one.

### 3 Model Specification

By adapting the variable selection approach of Geweke (1996), we can specify the complete Bayesian model. The sampling density for the vector  $\mathbf{y}$  of  $n$  observations is

$$p(\mathbf{y}|\mathbf{u}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{\sum_{i=1}^n (y_i - \mathbf{w}'_i \mathbf{u})^2}{2\sigma^2} \right\}.$$

where  $\mathbf{w} = (w_M(x_i, 0), \dots, w_M(x_i, M))'$ . We give the parameter  $u_0$ , which is not restricted to be positive or zero, a normal prior with mean  $m_0$  and variance  $s_0^2$

$$p(u_0) = (2\pi s_0^2)^{-1/2} \exp \left\{ -\frac{(u_0 - m_0)^2}{2s_0^2} \right\}.$$

As has been done by previous authors, we condition the prior for  $u_1, \dots, u_M$  on binary latent indicator variables  $\gamma_1, \dots, \gamma_M$ . We define the  $k^{\text{th}}$  indicator variable to be one if  $u_k = 0$  and

zero otherwise. The prior for  $u_k$  conditioned on the latent indicator  $\gamma_k$  can then be written

$$p(u_k|\tau^2, \gamma_k) = \gamma_k \mathbf{1}_{\{0\}}(u_k) + (1 - \gamma_k) 2(2\pi\tau^2)^{-1/2} \exp\left\{-\frac{u_k^2}{2\tau^2}\right\} \mathbf{1}_{(0,\infty)}(u_k)$$

where the indicator function  $\mathbf{1}_A(\cdot)$  is one if its argument is in the set  $A$  and zero otherwise.

Each  $\gamma_k$  is given a Bernoulli prior with parameter  $p_\gamma$

$$p(\gamma_k) = p_\gamma^{\gamma_k} (1 - p_\gamma)^{1-\gamma_k},$$

and  $p_\gamma$  is given a uniform prior on  $(0, 1)$ .

To complete the prior specification, we give the variance parameters  $\sigma^2$  and  $\tau^2$  conditionally conjugate inverse gamma priors

$$\begin{aligned} p(\sigma^2) &\propto (\sigma^2)^{-(a_\sigma+1)} \exp\{-b_\sigma/\sigma^2\}, \\ p(\tau^2) &\propto (\tau^2)^{-(a_\tau+1)} \exp\{-b_\tau/\tau^2\}. \end{aligned}$$

The final step in the model specification requires choosing values for  $a_\sigma$ ,  $b_\sigma$ ,  $a_\tau$ ,  $b_\tau$ , and  $M$ . For values of  $a_\sigma$  and  $b_\sigma$ , we use small, “noninformative” values (e.g., 1 or 0.1). Similarly, for  $a_\tau$  and  $b_\tau$  we recommend setting both hyperparameters equal to 1. In practice, the values of  $a_\tau$  and  $b_\tau$  have very little effect on the quality of the fit of the monotonic regression.

In theory, the polynomial order  $M$  should be selected as a function of the sample size  $n$ . It can be shown that the order  $M$  can be chosen to be on the order  $O(n^\kappa)$  where  $\kappa = 1/3$  or  $1/5$  depending on the differentiability of the regression function  $f$  (see McLain and Ghosh, 2009, theorem 2 or Tenbusch, 1997).

To choose a value for  $M$  in finite samples, we take an approach similar to the approach taken by Neelon and Dunson (2004) for choosing the number of knots. Their recommenda-

tion was to choose the number of knots to be equal to the unique values of the predictor variable and let the fitting procedure remove redundant terms in their monotone regression function. Similarly, we recommend choosing a large value for  $M$ —possibly as large as the number of unique values of the predictor variable—and letting the variable selection procedure remove redundant columns of the Bernstein expansion. This approach makes the procedure adaptive (in a loose sense) in that the procedure “chooses” a value of  $M$  by removing redundant columns of the Bernstein basis expansion. This approach is also similar to Smith and Kohn (1996) who used variable selection on the coefficients in a basis expansion to fit an unconstrained, nonparametric regression curve.

In the child-growth example of Section 7, we used  $M = 80$ , which is nearly equal to the number of distinct  $x$ -values in the data set ( $n = 83$ ). For the simulation study in Section 6, we deviated from our recommended approach by choosing  $M = 40$  even though the number of distinct  $x$ -values in the simulated data sets was 100. This choice of  $M = 40$  was made for practical reasons (to keep simulation times manageable); however, the results of the simulation study show that our method performs favorably even when the choice of  $M$  was small relative to the sample size.

## 4 Posterior Sampling

A Markov Chain Monte Carlo approach can be used to sample from the joint posterior distribution of the parameters in the regression model (Tierney, 1994). A Gibbs sampling scheme (Geman and Geman, 1984; Gelfand and Smith, 1990) can proceed by following the algorithm specified in Geweke (1996), except that our algorithm includes two additional Gibbs updates for parameters  $\tau^2$  and  $p_\gamma$ . The Geweke algorithm can be extended to include the case of multiple predictors. Computational details, including an outline of the Gibbs sampler, are included in the appendices.

## 5 Extension to Multiple Predictors

It is relatively straightforward to extend the approach to include multiple predictors. Let  $\mathbf{x} = (x_1, \dots, x_r, x_{r+1}, \dots, x_p)$  be the vector of predictors, where we assume that the regression function  $g(\mathbf{x})$  is additively monotone in the the variables  $x_1, \dots, x_r$ . In other words,

$$f(\mathbf{x}) = \sum_{j=1}^r f_j(x_j) + h(x_{r+1}, \dots, x_p)$$

where each of the  $f_j$ 's are nondecreasing functions. Then each of the  $f_j$ 's can be modeled using the Bernstein polynomials as described in Section 2.

Let  $r$  be the number of predictors with monotone relationship to the mean of the response variable (which we will call monotone predictors), and let  $p - r$  be the number of predictors with a simple, linear relationship with the mean of the response (which we will call linear predictors). We use  $\mathbf{x}_j = (x_{1j}, \dots, x_{nj})$  to denote the  $j^{\text{th}}$  monotone predictor and  $\mathbf{v}_i = (v_{i1}, \dots, v_{is})$  to denote a vector of linear predictor values for the  $i^{\text{th}}$  observation. The conditional mean for the response  $y_i$  can then be written as

$$u_0 + \sum_{j=1}^r \sum_{k=1}^M u_{jk} w_M(x_{ij}, k) + \boldsymbol{\alpha}' \mathbf{v}_i$$

where  $\boldsymbol{\alpha}$  is a vector of regression coefficients and  $u_{j1}, \dots, u_{jM}$  are the transformed basis coefficients for the  $j^{\text{th}}$  monotone predictor. Each set of  $u_{j1}, \dots, u_{jM}$  values are assigned a prior that is a mixture of a point mass at zero and a truncated normal distribution. Thus, the prior for each set of  $u_{j1}, \dots, u_{jM}$  variables has corresponding  $\gamma_{j1}, \dots, \gamma_{jM}$  and  $\tau_j^2$  parameters. The regression coefficients  $\boldsymbol{\alpha}$  for the linear predictors can be given virtually any prior, although it is convenient computationally to assign a conjugate, multivariate normal prior with mean vector  $\mathbf{m}_\alpha$  and covariance matrix  $\mathbf{S}_\alpha$ .

## 6 A Simulation Study

To test our monotonic regression method, we ran several simulations and compared the performance of our approach with competing monotonic regression methods. We generated 50 data sets from each of four models  $y_i = h_s(x_i) + \epsilon_i$  (for  $s = 1, \dots, 4$  and  $i = 1, \dots, 100$ ), where  $\epsilon_i \sim \mathbf{N}(0, 0.1^2)$ ,  $x_i \sim \text{Unif}(0, 1)$ . The regression function  $h_s(\cdot)$  took on the following shapes

$$\begin{aligned} \text{Flat:} & \quad h_1(x) = 1 \\ \text{Linear:} & \quad h_2(x) = x \\ \text{Flat and Nonlinear:} & \quad h_3(x) = I(x > 0.5)\sqrt{2x - 1} \\ \text{Wavy:} & \quad h_4(x) = \frac{\sin(3\pi x) + 3\pi x}{3\pi}. \end{aligned}$$

All functions were chosen to have a range of one, and the standard deviation of  $\epsilon_i$  was chosen to be one-tenth of the range.

We compared our method with three other methods that are accessible as functions in the R statistical software (R Development Core Team, 2007; Ihaka and Gentleman, 1996). The first is the local regression method as implemented by the `loess` function (Cleveland, Grosse, and Shyu, 1992). The second is the isotonic regression estimator of Barlow et al. (1972) as implemented in the `isoreg` function. The third is the monotonic regression estimator of Dette et al. (2006) as implemented in the `monreg` package (Pilz and Titoff, 2005) for R. This method proceeds by first fitting a regression smoother to the data, then integrating a kernel density estimate of the smoother. The final estimate of the monotone regression function is obtained by taking the inverse of this integrated density estimate.

Because there is no publicly available software that implements the method of Chang et al. (2007), we have not included this method in our simulation studies. We note, however, that Chang et al. (2007) have shown good performance of their method relative to Dette

et al. (2006) in simulation studies. (See Chang et al., 2007, section 4.1.)

The `loess` method requires the user to specify one tuning parameter. We used five-fold cross-validation to choose the value for this parameter. The `monreg` function requires the specification of two tuning parameters— $\lambda_d$  and  $\lambda_r$ . As per the recommendation of (Dette et al., 2006, section 4.1), we set  $\lambda_d = \lambda_r^2$ . We then used five-fold cross-validation to choose the value of  $\lambda_r$ .

To fit our Bayesian model to each data set, we used a Bernstein polynomial expansion of order 40 (i.e.,  $M = 40$ ). We ran the MCMC algorithm for 110,000 iterations, discarded the first 10,000 iterations as burn-in, and kept every 10th iteration of the resulting chain (for a final chain length of 10,000).

After fitting each regression method to a simulated data set, we evaluated the quality of the fit by computing fitted values for each method over a grid of 100 equally spaced “ $x$ -values” on the unit interval. The absolute difference between the true function and the fitted value was standardized (by dividing by the true standard deviation of 0.1), and the mean over the grid of 100  $x$ -values was recorded for each method.

The results of the simulation are contained in Table 1 and summarized graphically in Figure 2. In general, the Bayesian method performed very well relative to the other methods. In three of the four simulations, the Bayesian method had the lowest mean standardized absolute deviation from the true function. As expected, the Bayesian method performs substantially better than the other three methods when the true regression function is flat, as indicated by the box plot. Using our method, an analyst may compute the posterior probability of a flat regression curve  $p_{\text{FLAT}} = P(u_1 = \dots = u_M = 0 | \text{data})$  by computing the proportion of the posterior draws in which  $u_1 = \dots = u_M = 0$ . We recorded these values for each simulated data set. When the true regression function was flat, the five number summary of the  $p_{\text{FLAT}}$  values was

Min.    1st Qu.    Median    3rd Qu.    Max.

0.572 0.931 0.972 0.978 0.986

Thus, for most simulated data sets, our method returns  $p_{FLAT}$  values greater than 0.90 when the true curve is flat. When the true curve was other than flat, the  $p_{FLAT}$  values for all simulations were zero.

The Bayesian method also outperforms the other methods when the underlying function is the flat/square root function or the sine function, although the `monreg` function is a close competitor. Predictably, the Bayesian method did not perform as well when the true regression function had no flat portions (i.e., was linear), however, it was not the worst method in this case.

For simulated data sets from flat/square root function, we calculated the average slope (2) of the fitted Bernstein polynomial curve over the interval (0.0,0.5)—the flat portion of the true regression curve. The five number summary of the average slopes from the simulated data sets was

Min.	1st Qu.	Median	3rd Qu.	Max.
0.0014	0.0020	0.0024	0.0027	0.0039

To put these numbers in perspective, consider that the average slope of the non-flat (i.e. square root) portion of the true regression function is 0.24, which is 100 times larger than the median of the average slopes from the Bernstein polynomial fits over the flat portion of the curve. Thus, the median of the slopes of the Bernstein regression curve was only trivially different from zero when the true regression function was flat over the same region.

[Table 1 about here.]

[Figure 2 about here.]

## 7 Illustrations Using Real Data Sets

We try our method on two real data sets. The first example uses data with a continuous response. The second example uses data with a binomial response.

### 7.1 Monotone Regression Model for Continuous Response

The first example uses data from Ramsay (1998) on the growth of a 10-year-old boy. The data contain 83 height measurements over a period of 312 days. Growth curves should exhibit monotonicity and allow for relatively flat regions consistent with the belief that growth occurs in spurts rather than strictly increasing over time (Thalange, Foster, Gill, Price, and Clayton, 1996).

To fit the model with our Bayesian procedure, we rescaled the time (day) variable to the interval  $[0, 1]$ . We fit the Bayesian model using Bernstein polynomials of order 80. We ran the MCMC algorithm of Section 4 three times in order to assess convergence using the Gelman-Rubin (GR) diagnostics (Gelman and Rubin, 1992). Each MCMC run consisted of 160,000 iterations with 10,000 iterations discarded as burn-in and only every tenth iteration kept for plotting and analysis (i.e., the final MCMC size was 15,000 per MCMC run). The GR diagnostics for each parameter were all equal to one (up to two decimal places), except for a handful of parameters which had GR diagnostics very nearly one (the maximum GR diagnostic from this group was 1.02). Figure 3 contains the fit from the Bayesian model along with fits from `isoreg`, `loess`, and `monreg`. All fits (with the exception of `isoreg`) appear relatively smooth. A small portion of the `monreg` fit is not covered by the pointwise 95% credible bands from the Bayes method.

[Figure 3 about here.]

## 7.2 Monotone Regression Model for Discrete Response

The Bayesian variable selection method for monotonic regression can easily be extended to generalized linear models. For example, the sampling probability of a binomial observation is

$$p(y_i | n_i, p_i) = \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}$$

where  $n_i$  is the number of trials and  $p_i$  is the probability of success on given trial. Using the logit transformation to model  $p_i$  we have

$$\log \left( \frac{p_i}{1 - p_i} \right) = \mathbf{w}'_i \mathbf{u}$$

where  $\mathbf{w}$  and  $\mathbf{u}$  are as previously defined in Section 2.2.

A Gibbs sampler for this model can be constructed using the WinBUGS software. We fit this model to the Downs syndrome data of Geyer (1991). The data consist of the number of Downs syndrome births out of all births recorded by the British Columbia Health Surveillance Registry according to the mother's age category. For these data, it is not unreasonable to assume that the probability of a Downs syndrome birth increases monotonically with the age of the birth mother. The fit of the model to this data is plotted in Figure 4 with 95% posterior bands.

[Figure 4 about here.]

## 8 Conclusion

In this paper, we have presented a connection between monotonic regression and variable selection. This connection opens up new ways to fit monotonic regression models. We presented one Bayesian approach in this paper and were able to exploit the already existing MCMC techniques in variable selection to derive a Gibbs sampler for our monotonic regression method. We have demonstrated the effectiveness of our method in a simulation comparison of other competing methods. Finally, we have demonstrated the flexibility of our method by fitting our model to two very different data sets consisting of continuous and discrete-valued response variables. One immediate advantage of our proposed Bayesian method is that posterior interval bands can be obtained in a straightforward manner without appealing to any sort of asymptotic-based inference for both continuous and discrete-valued responses.

## Appendices

### A Computational Details

For speed, we coded our MCMC algorithm in a C function that we called from R. For example, the child-growth example, in which we set  $M = 80$ , takes approximately 8 minutes to run 160,000 iterations (10,000 burn in, 150,000 iterations after burn in, thinned by 10 for a final MCMC sample of 15,000) on a Windows Vista laptop with 1G of RAM and a 1.73 GHz, Intel Centrino Duo processor. The fit changes little when using substantially smaller  $M$ , however. The child-growth example using  $M = 40$  and the same MCMC settings takes only 2 minutes to run on the same laptop. Computation times would be substantial longer if the MCMC algorithm were coded directly in R.

We compiled our C function using the GNU C compiler as found in `Rtools` “package” cre-

ated by Brian Ripley and maintained by Duncan Murdoch (at <http://www.murdoch-sutherland.com/Rtools/>). Pseudo-random truncated normals were computed using the algorithm of Robert (1995). MCMC diagnostics and plots were prepared using the `coda` package in R (Plummer, Best, Cowles, and Vines, 2007). For the Downs syndrome example, we used the `R2WinBUGS` package (Sturtz, Ligges, and Gelman, 2005) to conveniently call `WinBUGS` from R.

An R package with a function that implements the method described in this paper is available from the first author’s website.

## B MCMC algorithm

The following describes the Gibbs sampler for computing posterior draws from the Bayesian monotonic regression procedure with one monotone predictor. The complete conditionals are as in Geweke (1996), except that Geweke’s formulation implicitly relies on  $\gamma_k$  indicator variables for whether the corresponding  $u_k$  is zero. In our formulation, we make this dependence explicit. We also add updates for the additional parameters  $\tau^2$  and  $p_\gamma$ .

Begin the MCMC algorithm by initializing all parameters— $u_0^{(0)}$ ,  $\tau^{2(0)}$ ,  $p_\gamma^{(0)}$ ,  $\sigma^{2(0)}$ ,  $\boldsymbol{\gamma}^{(0)} = (\gamma_1^{(0)}, \dots, \gamma_M^{(0)})$ , and  $\mathbf{u}^{(0)} = (u_1^{(0)}, \dots, u_M^{(0)})$  (conditional on the  $\boldsymbol{\gamma}^{(0)}$ ). For  $t = 1, \dots, n_{iter}$  simulate in turn from each of the following distributions:

1.  $\sigma^{2(t)} \sim \text{InvGam}\left(a_\sigma + \frac{n}{2}, b_\sigma + \sum_{i=1}^n (y_i - u_0^{(t-1)} - \mathbf{w}'_i \mathbf{u}^{(t-1)})^2 / 2\right)$
2.  $u_0^{(t)} \sim \text{N}\left(\frac{m_0/s_0^2 + n\bar{z}/\sigma^2}{1/s_0^2 + n/\sigma^2}, (1/s_0^2 + n/\sigma^2)^{-1}\right)$ , where  $z_i = y_i - \mathbf{w}'_i \mathbf{u}^{(t-1)}$  and  $\bar{z} = \sum_{i=1}^n z_i / n$ .
3. For  $k = 1, \dots, M$ , simulate jointly  $\gamma_k^{(t)}$  and  $u_k^{(t)}$  by first simulating

$$\gamma_k^{(t)} \sim \text{Bern}\left(\frac{\varpi_1}{\varpi_1 + \varpi_0}\right),$$

and then setting  $u_k^{(t)} = 0$  if  $\gamma_k^{(t)} = 1$  or otherwise simulating

$$u_k^{(t)} \sim \mathbf{N}_{(0,\infty)}(m_*, \sigma_*^2)$$

where

$$\begin{aligned} \varpi_1 &= \exp \left\{ - \sum_{i=1}^n \frac{z_i^2}{2\sigma^{2(t)}} \right\} p_\gamma^{(t-1)}, \\ \varpi_0 &= 2(\sigma_*^2/\tau^{2(t-1)})^{1/2} \exp \left\{ - \frac{\sum_{i=1}^n (z_i - b_k w_{ik})^2}{2\sigma^{2(t)}} - \frac{b_k^2}{2\sigma_k^2} + \frac{m_*^2}{2\sigma_*^2} \right\} \times \\ &\quad \left[ 1 - \Phi \left( - \frac{m_*}{\sigma_*} \right) \right] (1 - p_\gamma^{(t-1)}), \\ z_i &= y_i - u_0^{(t)} - w_{i,1}u_1^{(t)} - \dots - w_{i,k-1}u_{k-1}^{(t)} - w_{i,k+1}u_{k+1}^{(t-1)} - \dots - w_{i,M}u_M^{(t-1)}, \\ m_* &= \frac{\tau^{2(t-1)}b_k}{\tau^{2(t-1)} + \sigma_k^2}, \\ \sigma_*^2 &= \frac{\sigma_k^2\tau^{2(t-1)}}{\tau^{2(t-1)} + \sigma_k^2}, \\ \sigma_k^2 &= \frac{\sigma^{2(t)}}{\sum_{i=1}^n w_{ik}^2}, \\ b_k &= \frac{\sum_{i=1}^n z_i w_{ik}}{\sum_{i=1}^n w_{ik}^2}, \end{aligned}$$

$\mathbf{N}_{(a,b)}(\xi, \nu^2)$  is a truncated normal distribution on the interval  $(a, b)$ , and  $\Phi(\cdot)$  is the cdf of the standard normal distribution.

4.  $\tau^{2(t)} \sim \text{InvGam} \left( a_\tau + \frac{M - \|\gamma^{(t)}\|}{2}, b_\tau + \frac{\sum_{j=1}^M (u_j^{(t)})^2}{2} \right)$
5.  $p_\gamma^{(t)} \sim \text{Beta}(a_p + \|\gamma^{(t)}\|, b_p + M - \|\gamma^{(t)}\|)$ , where values  $a_p = 1$  and  $b_p = 1$  are used for all examples and simulations in this paper.

## References

- Barlow, R., Bartholomew, D., Bremner, J., and Brunk, H. (1972), *Statistical Inference under Order Restrictions*, Wiley.
- Bornkamp, B. and Ickstadt, K. (2009), “Bayesian nonparametric estimation of continuous monotone functions with applications to dose-response analysis,” *Biometrics*, 65, 198–205.
- Brown, B. M. and Chen, S. X. (1999), “Beta-Bernstein smoothing for regression curves with compact support,” *Scandinavian Journal of Statistics*, 26, 47–59.
- Brunk, H. D. (1955), “Maximum likelihood estimates of monotone parameters,” *Annals of Mathematical Statistics*, 26, 607–616.
- (1958), “On the estimation of parameters restricted by inequalities,” *Annals of Mathematical Statistics*, 29, 437–454.
- Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., and Thomas, L. (2001), *Introduction to distance sampling: estimating abundance of biological populations*, Oxford University Press.
- Carnicer, J. M. and Peña, J. M. (1993), “Shape preserving representations and optimality of the Bernstein basis,” *Advances in Computational Mathematics*, 1, 173–196.
- Casella, G. and Moreno, E. (2006), “Objective Bayesian variable selection,” *Journal of the American Statistical Association*, 101, 157–167.
- Chak, P. M., Madras, N., and Smith, B. (2005), “Semi-nonparametric estimation with Bernstein polynomials,” *Economics Letters*, 89, 153–156.
- Chang, I., Chien, L., Hsiung, A. H., Wen, C., and Wu, Y.-J. (2007), “Shape restricted

- regression with random Bernstein polynomials,” *IMS Lecture Notes—Monograph Series*, 54, 187–202.
- Chang, I., Hsiung, C. A., Wu, Y.-J., and Yang, C.-C. (2005), “Bayesian survival analysis using Bernstein polynomials,” *Scandinavian Journal of Statistics*, 32, 447–466.
- Chien, L., Chang, I., Jiang, S. S., Gupta, P. K., Wen, C., Wu, Y., and Hsiung, C. (2009), “Profiling time course expression of virus genes: An illustration of Bayesian inference under shape restrictions,” *Annals of Applied Statistics*, (to appear), 1–31.
- Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992), “Local regression models,” in *Statistical Models in S*, eds. Chambers, J. M. and Hastie, T. J., Chapman & Hall / CRC, pp. 309–376.
- Cripps, E., Kohn, R., and Nott, D. (2006), “Bayesian subset selection and model averaging using a centred and dispersed prior for the error variance,” *Australian & New Zealand Journal of Statistics*, 48, 237–252.
- Dette, H., Neumeier, N., and Pilz, K. F. (2006), “A simple nonparametric estimator of a monotone regression function,” *Bernoulli*, 12, 469–490.
- Dunson, D. (2005), “Bayesian semiparametric isotonic regression for count data,” *Journal of the American Statistical Association*, 100, 618–627.
- Friedman, J. and Tibshirani, R. (1984), “The monotone smoothing of scatterplots,” *Technometrics*, 26, 243–250.
- Gallant, A. R. and Golub, G. H. (1984), “Imposing curvature restrictions on flexible functional forms,” *Journal of Econometrics*, 26, 295–321.
- 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. (1992), “Inference from iterative simulation using multiple sequences,” *Statistical Science*, 7, 457–511.
- Geman, S. and Geman, D. (1984), “Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images,” *IEEE Transactions on Pattern Analysis and Machine*, 6, 721–741.
- George, E. I. and Foster, D. P. (2000), “Calibration and empirical Bayes variable selection,” *Biometrika*, 87, 731–747.
- George, E. I. and McCulloch, R. E. (1993), “Variable selection via Gibbs sampling,” *Journal of the American Statistical Association*, 88, 881–889.
- George, E. I. and McCulloch, R. E. (1997), “Approaches for Bayesian variable selection,” *Statistica Sinica*, 7, 339–373.
- Geweke, J. (1996), “Variable selection and model comparison in regression,” in *Bayesian Statistics 5*, eds. Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., Oxford University Press, pp. 609–620.
- Geyer, C. J. (1991), “Constrained maximum likelihood exemplified by isotonic convex logistic regression,” *Journal of the American Statistical Association*, 86, 717–724.
- Ghosal, S. (2001), “Convergence rates for density estimation with Bernstein polynomials,” *Annals of Statistics*, 29, 1264–1280.
- Green, P. J. (1995), “Reversible jump Markov chain Monte Carlo computation and Bayesian model determination,” *Biometrika*, 82, 711–732.
- Hall, P. and Huang, L. (2001), “Nonparametric kernel regression subject to monotonicity constraints,” *Annals of Statistics*, 29, 624–647.

- Hildreth, C. (1954), "Point estimate of ordinates of concave functions," *Journal of the American Statistical Association*, 49, 598–619.
- Holmes, C. C. and Heard, N. A. (2003), "Generalized monotonic regression using random change points," *Statistics in Medicine*, 22, 623–638.
- Ihaka, R. and Gentleman, R. (1996), "R: A language for data analysis and graphics," *Journal of Computational and Graphical Statistics*, 5, 299–314.
- Lavine, M. and Mockus, A. (1995), "A nonparametric Bayes method for isotonic regression," *Journal of Statistical Planning and Inference*, 46, 235–248.
- Mammen, E., Marron, J. S., Turlach, B. A., and Wand, M. P. (2001), "A general projection framework for constrained smoothing," *Statistical Science*, 16, 232–248.
- McLain, A. C. and Ghosh, S. K. (2009), "Estimation of time transformation models with Bernstein polynomials," Institute of Statistics Mimeo Series #2625, North Carolina State University, Raleigh, North Carolina.
- Mitchell, T. J. and Beauchamp, J. J. (1988), "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, 83, 1023–1032.
- Mukerjee, H. (1988), "Monotone nonparametric regression," *Annals of Statistics*, 16, 741–750.
- Neelon, B. and Dunson, D. B. (2004), "Bayesian isotonic regression and trend analysis," *Biometrics*, 60, 398–406.
- Nicholson, W. (1992), *Microeconomic Theory*, The Dryden Press.
- Ongaro, A. and Cataneo, C. (2004), "Discrete random probability measures: A general framework for nonparametric Bayesian inference," *Statistics & Probability Letters*, 67, 33–45.

- Pilz, K. and Titoff, S. (2005), *monreg: Nonparametric monotone regression*, R package version 0.1. Earlier developments by Holger Dette and Kay Pilz.
- Plummer, M., Best, N., Cowles, K., and Vines, K. (2007), *coda: Output analysis and diagnostics for MCMC*, R package version 0.12-1.
- R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
- Ramsay, J. O. (1988), “Monotone regression splines in action,” *Statistical Science*, 3, 425–461.
- (1998), “Estimating smooth monotone functions,” *Journal of the Royal Statistical Society, Series B*, 60, 365–375.
- Reboul, L. (2005), “Estimation of a function under shape restrictions. Applications to reliability,” *Annals of Statistics*, 33, 1330–1356.
- Robert, C. P. (1995), “Simulation of truncated normal variables,” *Statistics and Computing*, 5, 121–125.
- Smith, M. and Kohn, R. (1996), “Nonparametric regression using Bayesian variable selection,” *Journal of Econometrics*, 75, 317–343.
- Spiegelhalter, D. J., Thomas, A., and Best, N. G. (1999), *WinBUGS Version 1.2 User Manual*, MRC Biostatistics Unit.
- Stadtmüller, U. (1986), “Asymptotic properties of nonparametric curve estimates,” *Periodica Mathematica Hungarica*, 17, 83–108.
- Sturtz, S., Ligges, U., and Gelman, A. (2005), “R2WinBUGS: A Package for Running WinBUGS from R,” *Journal of Statistical Software*, 12, 1–16.

- Tenbusch (1994), “Two-dimensional Bernstein polynomial density estimators,” *Metrika*, 41, 233–253.
- Tenbusch, A. (1997), “Nonparametric curve estimation with Bernstein estimates,” *Metrika*, 45, 1–30.
- Thalange, N. K. S., Foster, P. J., Gill, M. S., Price, D. A., and Clayton, P. E. (1996), “Model of normal prepubertal growth,” *Archives of Disease in Childhood*, 75, 427–431.
- Tierney, L. (1994), “Markov chains for exploring posterior distributions,” *Annals of Statistics*, 22, 1701–1728.
- Wang, X. and Li, F. (2008), “Isotonic smoothing spline regression,” *Journal of Computational and Graphical Statistics*, 17, 21–37.

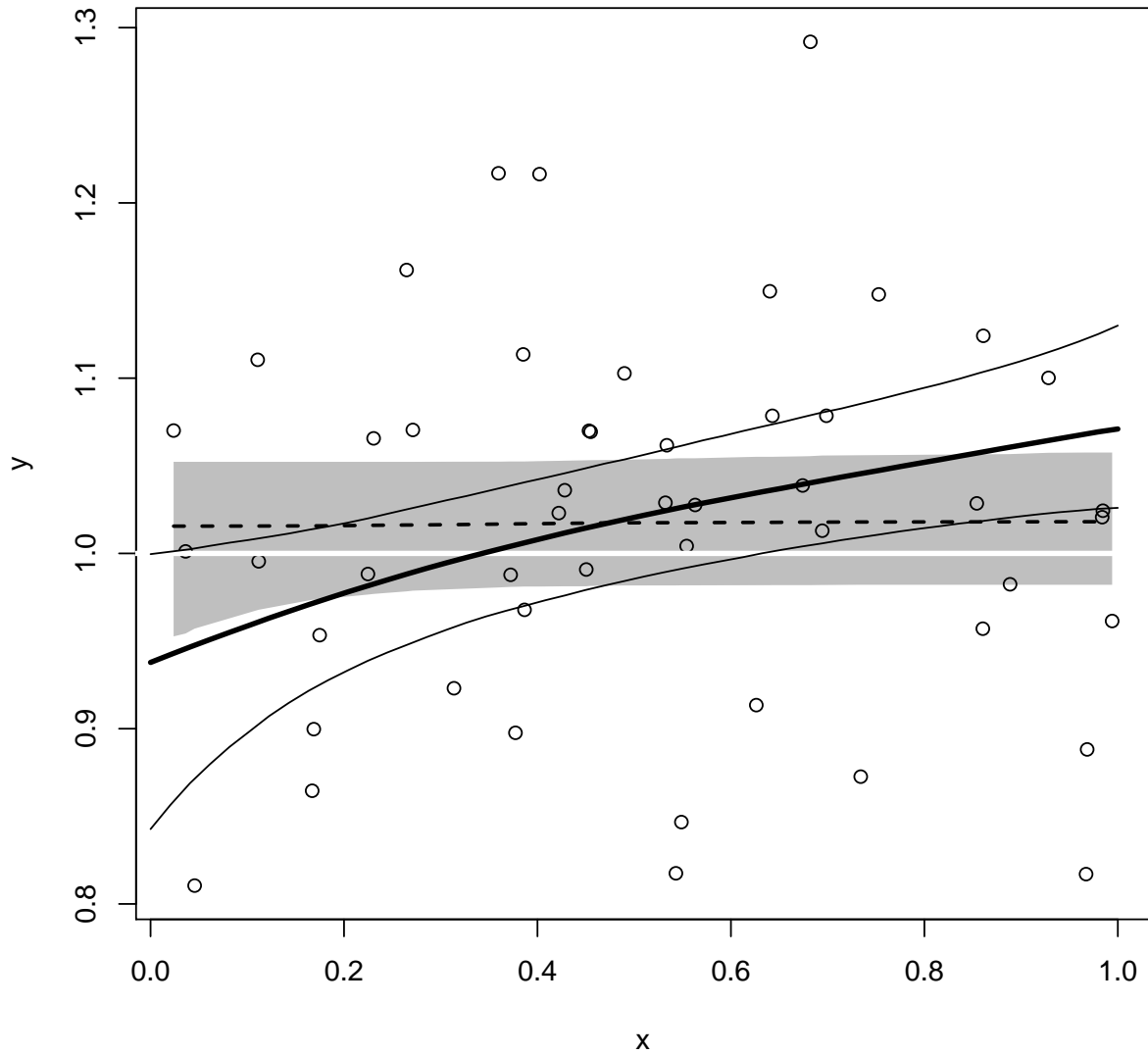


Figure 1: Plot of simulated data from a uniformly flat regression function with a Bayesian model that does not allow for flat portions of the regression curve.

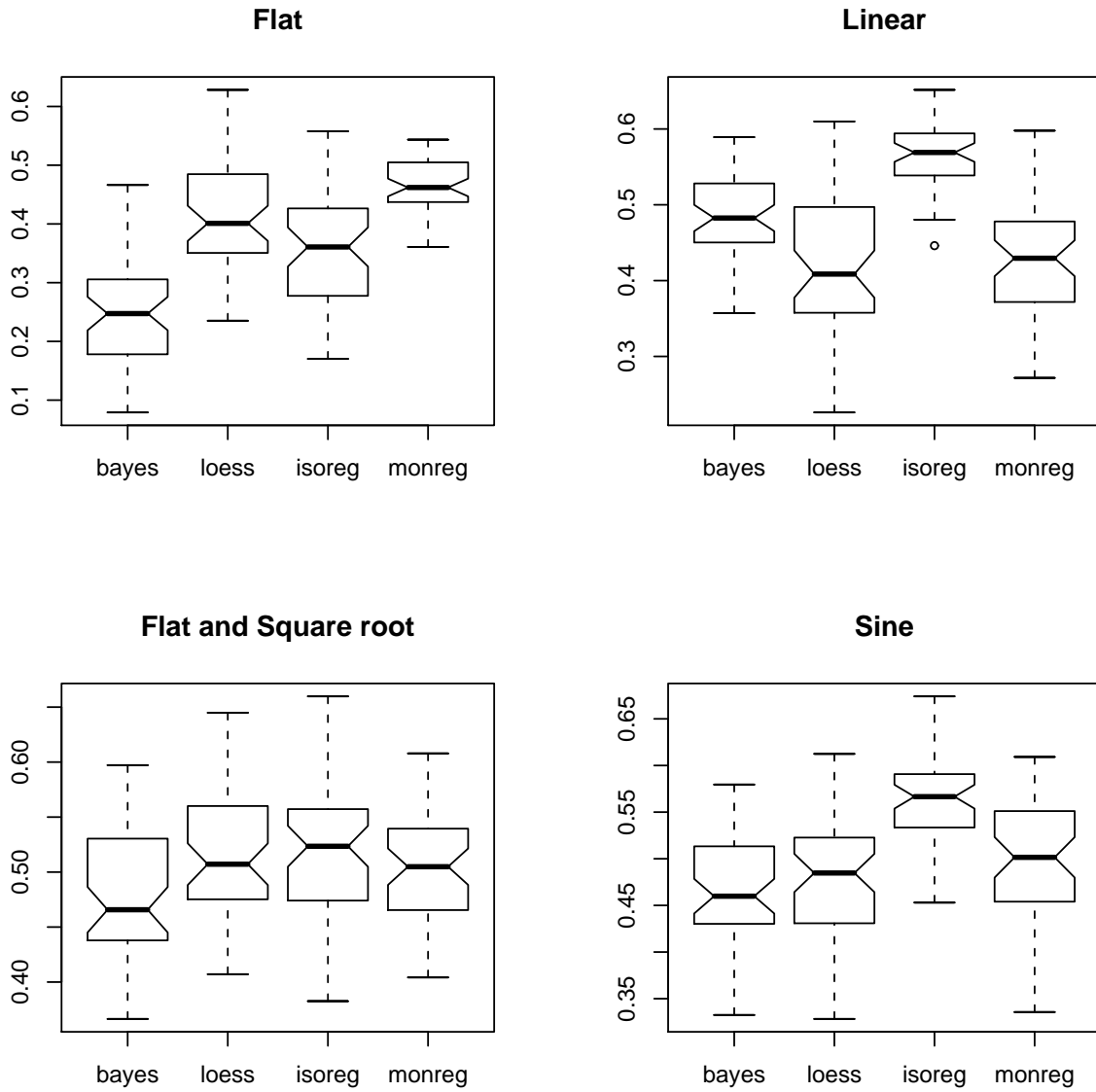


Figure 2: Boxplots of the mean of the standardized, absolute deviations of fitted values from the true function values at an equally-spaced grid of 100 points.

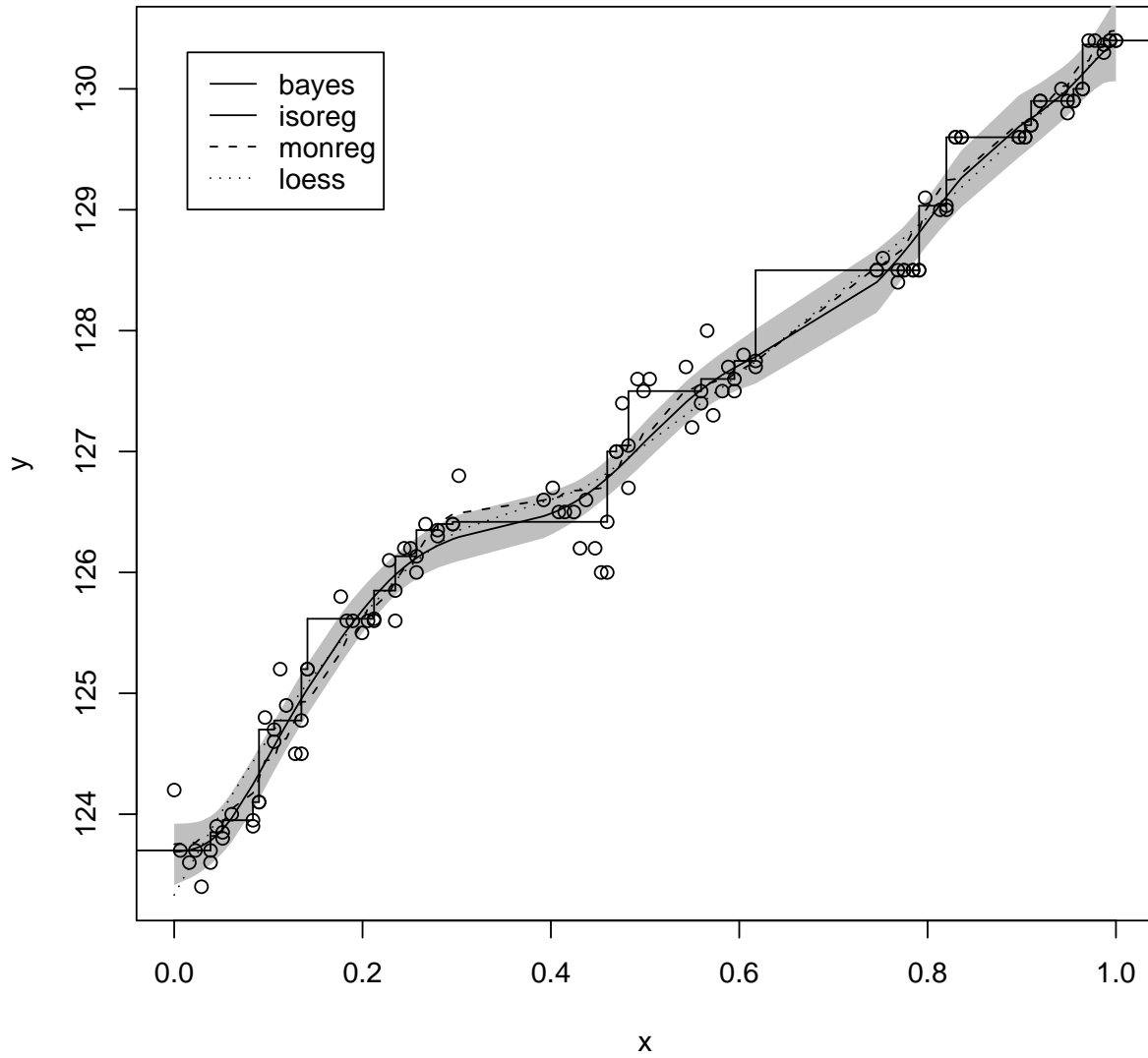


Figure 3: Plot of child height versus day (standardized to fall in the range  $[0, 1]$ ). The smooth, solid line is the Bayesian fit, the step function solid line is the `isoreg` fit, the dashed line is the `monreg` fit and the dotted line is the `loess` fit. The shaded region indicates 95% pointwise credible (confidence) region for the Bayesian fit.

### Downs syndrome data

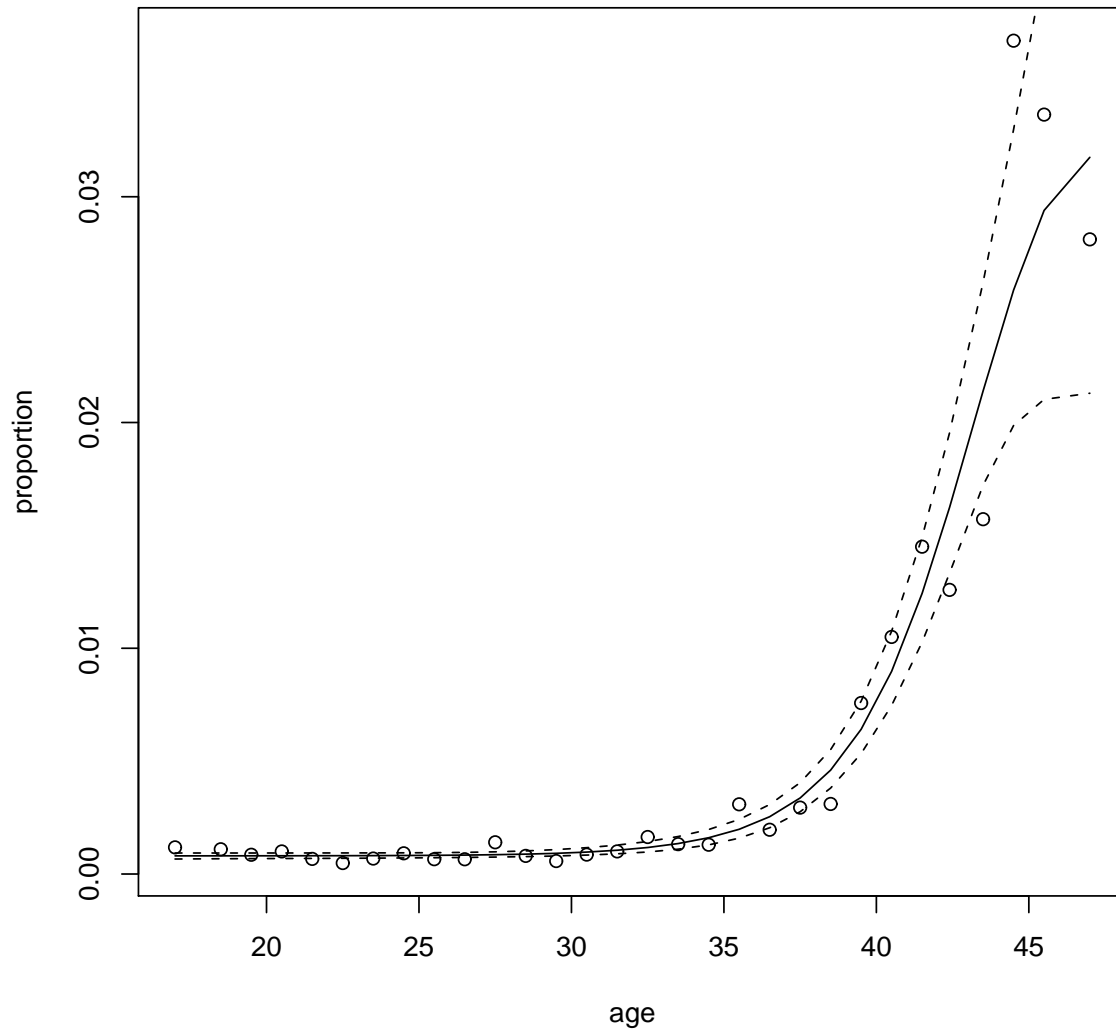


Figure 4: Plot of proportion of Down's syndrome births versus age of birth mother. The solid line is the posterior median and the dashed lines are 95% credible bands.

Table 1: Simulation results for comparison of the Bayesian procedure and competing monotonic regression methods. Values in the table are means of mean absolute standardized deviation of fitted values from the true function values at an equally-spaced grid of 100 points. Standard errors are in parentheses.

	bayes	loess	isoreg	monreg
Flat	0.25 (0.014)	0.41 (0.013)	0.36 (0.014)	0.47 (0.007)
Linear	0.48 (0.008)	0.42 (0.012)	0.57 (0.006)	0.43 (0.010)
Flat/Sqrt	0.48 (0.008)	0.51 (0.008)	0.52 (0.009)	0.50 (0.007)
Sine	0.46 (0.008)	0.48 (0.010)	0.56 (0.007)	0.50 (0.009)