A Bayesian Approach to Non-Parametric Monotone Function Estimation

Author(s): Thomas S. Shively, Thomas W. Sager and Stephen G. Walker

Source: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, Jan., 2009, Vol. 71, No. 1 (Jan., 2009), pp. 159-175

Published by: Wiley for the Royal Statistical Society

Stable URL: http://www.jstor.com/stable/20203882

J. R. Statist. Soc. B (2009)
71, *Part* 1, *pp.* 159–175

A Bayesian approach to non-parametric monotone function estimation

Thomas S. Shively and Thomas W. Sager

University of Texas at Austin, USA

and Stephen G. Walker

University of Kent, Canterbury, UK

[Received June 2005. Final revision May 2008]

Summary. The paper proposes two Bayesian approaches to non-parametric monotone function estimation. The first approach uses a hierarchical Bayes framework and a characterization of smooth monotone functions given by Ramsay that allows unconstrained estimation. The second approach uses a Bayesian regression spline model of Smith and Kohn with a mixture distribution of constrained normal distributions as the prior for the regression coefficients to ensure the monotonicity of the resulting function estimate. The small sample properties of the two function estimators across a range of functions are provided via simulation and compared with existing methods. Asymptotic results are also given that show that Bayesian methods provide consistent function estimators for a large class of smooth functions. An example is provided involving economic demand functions that illustrates the application of the constrained regression spline estimator in the context of a multiple-regression model where two functions are constrained to be monotone.

Keywords: Asymptotic properties; Markov chain Monte Carlo sampling scheme; Mixture prior distributions; Regression splines; Small sample properties

1. Introduction

Monotone functions arise naturally in economics and many other disciplines. Often, it is known or theoretically plausible that the relationship $y = f(x)$ between a dependent variable y and an independent variable x should be monotone. Examples include demand and supply curves, Phillip's curves, functions relating the probability of firm insolvency to holdings of risky assets and functions representing children's growth patterns through time. However, actual observations may violate monotonicity on account of measurement error, chance and/or the cumulative effects of all variables that are excluded from the model. The problem is to estimate the monotone functional form $f(x)$, given observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ with $x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n$, but where the ys may not all be monotone.

Frequently, it is also known or theoretically plausible that the monotone form $y = f(x)$ is smooth. Usually, 'smooth' means that derivatives of $f(x)$ exist up to a specified order. Ramsay (1998) has provided a useful characterization of smooth monotone functions that simplifies the estimation problem. His characterization permits the replacement of estimation subject to monotonicity by unconstrained estimation. We consider two approaches to non-parametric

Address for correspondence: Thomas S. Shively, Department of Information, Risk, and Operations Management, Mail Code B6500, University of Texas at Austin, Austin, TX 78712, USA.
E-mail: Tom.Shively@mccombs.utexas.edu

 1369–7412/09/71159

monotone function estimation. First, we adopt a modification of Ramsay's characterization that allows for unconstrained estimation, but we imbed our model in a hierarchical Bayes framework to utilize the power of Bayesian methods. Second, we use a Bayesian regression spline model (Smith and Kohn, 1996) with a mixture distribution of constrained normal distributions as the prior for the regression coefficients to ensure the monotonicity of the resulting function estimate. The prior places positive probability on the boundary of the constrained parameter space so the resulting function estimates will do well when significant portions of the function being estimated are flat.

The predominant focus of the theoretical literature on monotone function estimation has been on the methodology of order-restricted inference, which is sometimes also known as isotonic regression (see Barlow *et al.* (1972) for theoretical background and a comprehensive framework). The raw isotonic solution to the problem is the order-constrained least squares estimate. But this solution is a monotone step function, with a small bias, especially near end points of the domain of x. Because of the attractiveness of smooth estimates, several researchers have subsequently explored the combination of isotonic regression with smoothing considerations (Wright and Wegman (1980), Friedman and Tibshirani (1984) and Mammen (1991)—see Ramsay (1998) for a brief review of these and related extensions). Holmes and Heard (2003) and Neelon and Dunson (2004) developed methods for non-parametric monotone function estimation that are based on a Bayesian analysis of the isotonic regression model and order-restricted inference.

Simulation results in Section 4 show that the constrained regression spline method has good small sample properties relative to the methods that were proposed by Holmes and Heard (2003) and Neelon and Dunson (2004) across a wide range of functions. The asymptotic results in Section 5 show that Bayesian methods such as those which are developed in this paper provide consistent function estimators for a large class of smooth functions.

The paper is organized as follows. Section 2 develops the Bayesian method for monotone function estimation by using Ramsay's characterization of a monotone function as well as a Markov chain Monte Carlo (MCMC) sampling scheme to implement the method. Section 3 develops the regression spline method for monotone function estimation and the associated MCMC sampling scheme. Section 4 presents simulation results to show the small sample properties across a range of functions for the function estimators that are developed in Sections 2 and 3 as well as the methods that were proposed by Holmes and Heard (2003) and Neelon and Dunson (2004). Section 5 discusses the asymptotic properties of Bayesian estimators such as those in Sections 2 and 3. Section 6 presents a real data example in a multiple-regression context involving economic demand functions to illustrate the constrained regression spline estimator that is developed in Section 3.

2. Constrained estimation using a Wahba-type spline model

2.1. The model

Suppose that $y_1, y_2, \ldots, y_n$ are observed according to the model

$$y_i = f_0(x_i) + \varepsilon_i \qquad i = 1, 2, \ldots, n \tag{1}$$

where $x_0 \leqslant x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n$ and $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent, identically distributed $N(0, \sigma^2)$ random variables. Without loss of generality, take $x_0 = 0$ and $x_n = 1$. Further suppose that

$$f_0(x) = \alpha + \int_0^x \exp\{\beta + \tau\, w(t)\}\, \mathrm{d}t \tag{2}$$

where α, β and τ are parameters, and $w(\cdot)$ is a standardized continuous integrable function on the real line with $w(0) = 0$. We shall specify in Section 2.3 a Wiener process as the prior

distribution for $w(\cdot)$. The method of standardization will be specified later. Thus, $w(\cdot)$ will be a realization of a Brownian motion, and hence $w(0)=0$. For each value s, $w(s)$ plays the role of a model parameter so we conceive of $w(\cdot)$ as an infinite dimensional parameter. The complete parameterization of the model therefore involves α, β, τ, $w(\cdot)$ and σ^2. τ amplifies or dampens the oscillations that are induced by the standard $w(\cdot)$ function and thus plays the role of a tuning parameter. In the extreme case of $\tau=0$, $f_0(x)$ reduces to a linear function $f_0(x)=\alpha+x\exp(\beta)$, which may be considered the smoothest form of all. We note that $f_0(x)$ is necessarily monotone increasing. To avoid repetition, we confine our attention to monotone increasing functions. All our results apply equally to monotone decreasing functions after obvious modifications, such as changing the sign in front of the integral in equation (2) from plus to minus. This representation of $f_0(x)$ provides a large and flexible class of semiparametric forms that are especially suitable for fitting the data by Bayesian methods.

2.2. Class of functions represented by the model

It is of some interest to investigate the extent to which the model form (2) can represent the class of smooth monotone functions. To do this we use a modification of Ramsay's (1998) characterization of smooth monotone functions. Consider the class of functions $C=\{f(x); f'(x)$ exists, $f'(x)$ continuous, $f'(x)>0\}$. Also, let $C(\Re)$ represent the class of continuous integrable functions on the real line (restricted to [0,1] without loss of generality). The following theorem is similar to a theorem that was given by Ramsay. We provide a more elementary proof than Ramsay that does not require using the theory of differential equations.

Theorem 1. $f(x)\in C$ if and only if there exists $u(x)\in C(\Re)$ such that $f(x)=a+\int_0^x \exp\{b+u(t)\}\,dt$.

Proof. If $f(x)\in C$, then let $u(x)=\log\{f'(x)\}$. Since $f'(x)$ is continuous on [0,1], then $\log\{f'(x)\}$ is also, so $u(x)$ is bounded and thus integrable. Set $a=f(0)$ and $b=0$ to yield the representation. However, if $f(x)=a+\int_0^x \exp\{b+u(t)\}\,dt$ and $u\in C(\Re)$ then, by the fundamental theorem of calculus, $f'(x)=\exp\{b+u(x)\}>0$. Thus, $\log\{f'(x)\}=b+u(x)$ is continuous, so $f'(x)=\exp[\log\{f'(x)\}]$ is continuous.

To apply this characterization to the function in equation (2), we associate a with α, b with β and $u(t)$ with $\tau\, w(t)$. Since $w(\cdot)$ is assumed continuous and integrable, it is easy to check that all functions of the form (2) are in the class C. Conversely, given a continuous integrable function $u(x)$ and $\tau>0$, define $w(x)=\tau^{-1}\{u(x)-u(0)\}$ (so $u(x)=u(0)+\tau\, w(x)$). Thus, if $f(x)$ is in the monotone class C, then $f(x)$ has the form

$$a+\int_0^x \exp\{b+u(t)\}\,dt = a+\int_0^x \exp\{u(0)+\tau\, w(t)\}\,dt$$

which matches the form (2) if we set $a=\alpha$ and $u(0)=\beta$. We conclude that the class of functional forms that is given by equation (2) coincides with the class C of monotone functions with a continuous first derivative on the unit interval—a rich class of smooth models from which to choose.

2.3. Prior distributions

To complete the specification of the Bayesian model, we assume independent prior distributions for α, β, τ and σ^2. In particular, we use $N(0, 10^2)$ priors for α and β, and flat priors on $[0,10^3]$ for τ^2 and σ^2.

To provide a prior distribution for the infinite dimensional parameter $w(\cdot)$, we specify a Wiener process $W(\cdot)$ with variance τ^2 on [0,1] and take $w(\cdot)$ to be the standardized component of $W(\cdot)$, i.e. $w(\cdot) \equiv W(\cdot)/\tau$ can be viewed as a realization of Brownian motion with unit variance. For estimation, it is actually not necessary to distinguish τ from $w(\cdot)$ as a separate parameter, but it is useful to do so to have a measure of the smoothness of the path function $W(\cdot) \equiv \tau\, w(\cdot)$. If we do distinguish τ, then it is necessary to ensure its identifiability. We do this by defining τ to be a measure of the variability of the identifiable path function $W(\cdot)$ and then divide $W(\cdot)$ by that measure to extract the 'standardized' $w(\cdot)$. The Bayes interpretation is that τ^2 is a realization of a prior distribution on the variance of a non-unit variance Wiener process.

With probability 1, sample paths of a Wiener process $W(\cdot)$ are continuous and integrable, so the class of model forms (2) coincides almost surely with the class C of smooth monotone functions. Note that model (2) without the 'exp'-term corresponds to Wahba's (1978) spline model that was used by Wong and Kohn (1996) and others for unconstrained non-parametric function estimation.

2.4. Bayesian estimation of f(x)

In Bayesian inference, the estimator of an unknown quantity $f_0(x)$ is its posterior mean, given the data: $E\{f_0(x)|y_1, \ldots, y_n\}$. The expectation is taken with respect to all sources of parametric uncertainty, namely $\alpha, \beta, \tau, w(\cdot)$ and σ^2. The usual procedure to obtain the estimate is to produce the joint distribution of the data and parameters by combining the likelihood function of the data with the prior distributions of the parameters, and then to derive the posterior distribution and its mean from the joint distribution.

Unfortunately, it is difficult to compute the posterior mean $E\{f_0(x)|y_1, \ldots, y_n\}$ even by using MCMC methods. For this reason we estimate $f_0(x)$ at a finite set of evenly spaced nodes. This simplification collapses the infinite dimensional prior into a finite dimensional prior. For each integer m and the given continuous function $w(\cdot)$, define the function

$$\gamma_m(t) = \tau\, w\left(\frac{[mt]}{m}\right)$$

where $[s]$ is the greatest integer less than or equal to s. $\gamma_m(\cdot)$ is a step function approximation to $\tau\, w(\cdot)$. For $0 \leqslant t \leqslant 1$, $\gamma_m(t)$ has only $m+1$ distinct values, determined at the node points $0, 1/m, 2/m, \ldots, 1$. We treat the m values $\gamma_m(1/m), \ldots, \gamma_m(1)$ as (derived) parameters (note that $\gamma_m(0) = 0$) and endow them with the prior distributions that are determined by the Wiener process prior of $\tau\, w(\cdot)$. For the finitely parameterized case, the standardization of the path function $W(\cdot) = \tau\, w(\cdot)$ can be achieved by setting $\tau^2 = \tau_m^2$, the variance of the m quantities $\{\sqrt{m}[W(i/m) - W\{(i-1)/m\}], i = 1, \ldots, m\}$. For a Wiener process with variance τ^2, the variance of each independent increment $\sqrt{m}[W(i/m) - W\{(i-1)/m\}]$ is τ^2. Our finitely parameterized approximation to f_0 is f_m defined by

$$f_m(x) = \alpha + \int_0^x \exp\{\beta + \gamma_m(t)\}\, \mathrm{d}t \tag{3}$$

and the approximating model that corresponds to model (1) is

$$y_i = f_m(x_i) + \varepsilon_i. \tag{4}$$

It is difficult in practice to estimate $f_m(x)$. However, a piecewise linear approximation to $f_m(x)$ can be used in conjunction with an MCMC sampling scheme to provide an estimator in practice for $f_0(x)$. Its finite sample properties are shown via simulation in Section 4. To keep the notation simple, we shall assume that n is a multiple of m. If $1/m$ is small, then for $(k-1)/m < x < k/m, k = 1, \ldots, m-1$,

$$f_m(x) \cong f_m\{(k-1)/m\} + f_m'\{(k-1)/m\}\{x-(k-1)/m\}.$$

Let $\gamma_{m,k} = \gamma_m(k/m)$. Then using equation (3) and the fundamental theorem of calculus we have $f_m'\{(k-1)/m\} = \exp(\beta + \gamma_{m,k-1})$ and for $(k-1)/m < x < k/m$

$$f_m(x) \cong f_m(x_0) + \sum_{j=1}^{k-1} f_m'\{(j-1)/m\}(1/m) + f_m'\{(k-1)/m\}\{x-(k-1)/m\}$$

$$= \alpha + \sum_{j=1}^{k-1} \exp(\beta + \gamma_{m,j-1})(1/m) + \exp(\beta + \gamma_{m,k-1})\{x-(k-1)/m\}.$$

This approximation allows model (4) to be written

$$y = Z\eta + \varepsilon$$

where $y = (y_1, \ldots, y_n)'$, $Z = (z_1, \ldots, z_{m+1})$ with z_j, $j = 1, \ldots, m+1$, defined appropriately, $\eta = (\alpha, \exp(\beta), \exp(\beta + \gamma_{m,1}), \ldots, \exp(\beta + \gamma_{m,m-1}))'$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$.

Using this approximation, the MCMC sampling scheme that is described below is used to carry out function estimation. For a discussion of Bayesian inference using MCMC methods see Gelfand and Smith (1990), Casella and George (1992) and Tierney (1994).

For notational purposes, let $\gamma_m = (\gamma_{m,1}, \ldots, \gamma_{m,m-1})'$, and $\gamma_{m,(-j)} = \gamma_m$ without the jth element. We then have the following sampling scheme. Start with some initial values $\alpha^{[0]}$, $\beta^{[0]}$, $\gamma_m^{[0]}$, $(\tau_m^2)^{[0]}$ and $(\sigma^2)^{[0]}$.

Step 1: generate α conditional on y, β, γ_m and σ^2.
Step 2: generate β conditional on y, α, γ_m and σ^2.
Step 3: generate $\gamma_{m,j}$ conditional on y, α, β, $\gamma_{m,(-j)}$, τ_m^2 and σ^2, $j = 1, \ldots, m-1$.
Step 4: generate τ_m^2 conditional on γ_m.
Step 5: generate σ^2 conditional on y, α, β and γ_m.

The sampling scheme is irreducible and aperiodic because it is readily checked that in one step the sampling scheme can reach any point in the parameter space from any other point. Therefore, the sampling scheme converges to the posterior distribution by Tierney (1994).

Steps 1–5 are repeated many times and in two stages. The first stage is the warm-up period and it is assumed that at the end of this period the sampling scheme generates iterates from the posterior distribution. The second stage is the sampling period and iterates that are generated from this stage are used for inference.

Let $\alpha^{[l]}$, $\beta^{[l]}$ and $\gamma^{[l]}$ be the iterates of α, β and γ in the sampling period. Then an estimate of the posterior mean of $f_m(x)$ and therefore an estimate of $f_0(x)$ is

$$\hat{f}_m(x) = (1/L) \sum_{l=1}^{L} \left[\alpha^{[l]} + \sum_{j=1}^{k-1} \exp(\beta^{[l]} + \gamma_{m,j-1}^{[l]})(1/m) + \exp(\beta^{[l]} + \gamma_{m,k-1}^{[l]})\{x-(k-1)/m\} \right].$$

Steps 1, 4 and 5 are straightforward to implement. Steps 2 and 3 use sampling techniques that were given in Damien *et al.* (1999) and Damien and Walker (2001). The details are available from the authors on request.

3. Constrained estimation using a regression spline model

3.1. The model

Consider model (1) and a finitely parameterized approximation to $f_0(x)$ given by

$$f_m(x) = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 (x - \tilde{x}_1)_+^2 + \ldots + \beta_{m+2}(x - \tilde{x}_m)_+^2 \qquad (5)$$

where $\tilde{x}_1, \ldots, \tilde{x}_m$ are m 'knots' placed along the domain of the independent variable x such that $0 < \tilde{x}_1 < \ldots < \tilde{x}_m < 1$ and $(z)_+ = \max(0, z)$. The resulting approximating model that we use in this section is $y_i = f_m(x_i) + \varepsilon_i$. The constraints on the β_js that are required to ensure that the function is non-decreasing (isotonic) are discussed below. We note here that model (5) along with the constraints that are discussed below allow for functions where $f_0'(x)$ can take on the value 0 whereas model (2) restricts $f_0'(x)$ to be positive. Quadratic regression splines are used instead of the more typical cubic regression splines because they impose a degree of smoothness on the function but the constraints that are required to ensure isotonicity are more tractable than for cubic splines. We show in Section 4 that the resulting function estimator has good small sample properties.

Following Smith and Kohn (1996), who considered non-parametric function estimation using regression splines without a monotonicity constraint, we initially place the knots at m prespecified values and then choose the location of the knots to be left in the model by using variable selection. Using the m knots, equation (5) can be written in matrix notation as

$$y = \iota\alpha + X\beta + \varepsilon$$

where y is an $n \times 1$ vector of observations, $\iota = (1, \ldots, 1)'$, X is an $n \times (m+2)$ matrix, $\varepsilon \sim N(0, \sigma^2 I_n)$ and $\beta = (\beta_1, \ldots, \beta_{m+2})'$. Let I be the $(m+2) \times 1$ vector of indicator variables with the jth element I_j such that $I_j = 0$ if $\beta_j = 0$ and $I_j = 1$ if $\beta_j \neq 0$.

Given I, we can impose constraints on the β_js to ensure that the resulting function is non-decreasing. For example, if $I_j = 1$ for all j then the constraints are $\beta_1 \geqslant 0$, $\beta_1 + 2\tilde{x}_1\beta_2 \geqslant 0$ and $\beta_1 + 2\tilde{x}_{j+1}\beta_2 + 2\sum_{k=1}^{j}(\tilde{x}_{j+1} - \tilde{x}_k)\beta_{k+2} \geqslant 0$, $j = 1, \ldots, m$ (with $\tilde{x}_{m+1} = 1$). In general, if β_I consist of the elements of β corresponding to those elements of I that are equal to 1 then the linear restrictions on the elements of β_I that are required to ensure that the function is non-decreasing can be written as $\gamma_I = L_I\beta_I$, where L_I is a lower triangular matrix that depends on I and the $\tilde{x}_j$, and each element of γ_I must be greater than or equal to 0.

3.2. Prior distributions

Given I and σ^2, the prior distribution for β_I is a mixture distribution of a constrained normal distribution $N(0, c\sigma^2\Omega_I)$, where the elements of β_I are constrained to the portion of the β_I-space that guarantees isotonicity, and probability distributions on the boundaries of the constrained parameter space. Using a mixture distribution for the prior is important because incorporating the probability distributions on the boundary provides good function estimates when the function being estimated has significant flat portions. Neelon and Dunson (2004) also used a mixture prior in their method for estimating isotonic functions non-parametrically. If the boundary distributions are not included then the prior implies a large positive derivative of the unknown function. c is a positive scale factor that is specified by the user. For our model, we let c equal the number of observations n. We set $\Omega_I = L_I^{-1}(L_I')^{-1}$ to allow for the feasible calculation of the constants that makes the mixture prior integrate to 1. The properties of the prior on β_I are such that the resulting function estimator has good small sample properties across a wide range of functions (see Section 4).

To motivate the ideas underlying the construction of the prior for β, particularly the boundary distributions, we first consider a simple model with no knots (i.e. $m = 0$) so $f(x) = \alpha + \beta_1 x + \beta_2 x^2$. The ideas that are used to construct the prior in this case can be generalized to a model with m knots. First, consider the case $I = (1, 1)$. The values of $\beta_I = (\beta_1, \beta_2)$ that give monotone increasing functions fall in the region that is defined by $\beta_1 > 0, \beta_1 + 2\beta_2 > 0$, whereas values that give functions with $f'(0) = 0$ fall along the ray that is defined by $\beta_1 = 0$ and $\beta_2 \geqslant 0$

(i.e. the non-negative β_2-axis), and values that give $f'(1)=0$ fall along the ray that is defined by $\beta_1 \geqslant 0$ and $\beta_1+2\beta_2=0$. To construct the prior for the constrained parameters β_I, we simply define a prior for the unconstrained parameters β_I^*, map β_I^* into β_I and derive the resulting distribution for β_I. If β_I^* satisfies the constraint, the map is the identity; if β_I^* does not satisfy the constraint (there are several cases to consider), we project β_I^* into the boundary of the constrained region. The resulting distribution of β_I is a mixture of singular distributions on the boundaries of the constrained region and a continuous distribution on the interior. In more detail, let $\beta_I^*=(\beta_1^*, \beta_2^*)$ where $\beta_I^* \sim N(0, c\sigma^2\Omega_I)$ is unconstrained and consider the model $y=f_*+\varepsilon$ where $f_*=\iota\alpha+X_I\beta_I^*$ and the ith row of X_I is (x_i, x_i^2). The basic idea is the following. For β_I^*-vectors that give a function f_* satisfying the isotonicity constraints we set $\beta_I=\beta_I^*$ so the function $f=\iota\alpha+X_I\beta_I$ also satisfies the constraints. Conversely, if the β_I^*-vector gives a function f_* that does not satisfy the isotonicity constraints then we set β_I equal to a boundary point of the constrained region, such that the resulting function $f=\iota\alpha+X_I\beta_I$ is 'close' to the function f_* but satisfies the constraints. The distribution for β_I^* and the method of setting β_I will define the prior for β_I and, in particular, the probability distributions on the boundary of the constrained parameter space. We consider four cases for β_I^*.

(a) *Case 1*: $\beta_1^* \geqslant 0$ and $\beta_1^*+2\beta_2^* \geqslant 0$. Then $f_*'(0)=\beta_1^* \geqslant 0$ and $f_*'(1)=\beta_1^*+2\beta_2^* \geqslant 0$, so we set $\beta_1=\beta_1^*$ and $\beta_2=\beta_2^*$. Therefore, for $\beta_1>0$ and $\beta_1+2\beta_2>0$, the prior for β_I is $h\{\beta_1, \beta_2 | I=(1,1)\}=(2\pi)^{-1}|c\sigma^2\Omega_I|^{-1/2}\exp\{-\frac{1}{2}\beta_I'(c\sigma^2\Omega_I)^{-1}\beta_I\}$.

(b) *Case 2*: $\beta_1^* \geqslant 0$ and $\beta_1^*+2\beta_2^*<0$. Then $f_*'(0)=\beta_1^* \geqslant 0$ but $f_*'(1)=\beta_1^*+2\beta_2^*<0$. Therefore, the derivative of f_* becomes negative between 0 and 1. To set the function f so that it is non-decreasing in this interval but still close to f_* we set $\beta_1=\beta_1^*$ and $\beta_2=-\beta_1/2=-\beta_1^*/2$. This gives $f'(0)=\beta_1=\beta_1^* \geqslant 0$ and $f'(1)=\beta_1+2\beta_2=\beta_1^*+2(-\beta_1^*/2)=0$. Note that f and f_* have the same derivative at $x=0$ whereas $f'(1)=0$ is as close to $f_*'(1)$ as possible while still satisfying the isotonicity constraint. The resulting (β_1, β_2) vector is on the boundary of the parameter space that guarantees a non-decreasing function. To obtain $h\{\beta_1, \beta_2 | I=(1,1)\}$ on the boundary, the probability that is associated with a specific β_1^*-value and all values of β_2^* such that $\beta_1^*+2\beta_2^*<0$ is accumulated and placed on the boundary at $(\beta_1, -\beta_1/2)$. Therefore,

$$h\{\beta_1, -\beta_1/2 | I=(1,1)\}=\int_{-\infty}^{-\beta_1^*/2} h(\beta_1^*, \beta_2^*)\, d\beta_2^*=\frac{1}{2}(2\pi)^{-1/2}(c\sigma^2)^{-1/2}\exp\left(\frac{-1}{2c\sigma^2}\beta_1^2\right), \qquad \beta_1 \geqslant 0.$$

(c) *Case 3*: $\beta_1^*<0$ and $\beta_1^*+2\beta_2^* \geqslant 0$. Then $f_*'(0)=\beta_1^*<0$ and $f_*'(1)=\beta_1^*+2\beta_2^* \geqslant 0$. In this case, we set $\beta_1=0$ and $\beta_2=(\beta_1^*+2\beta_2^*)/2$. This gives $f'(0)=\beta_1=0$ and $f'(1)=\beta_1+2\beta_2=0+2(\beta_1^*+2\beta_2^*)/2=\beta_1^*+2\beta_2^*$. Therefore, on the non-negative β_2-axis we have

$$h\{0, \beta_2 | I=(1,1)\}=\frac{1}{2}(2\pi)^{-1/2}\left(\frac{c\sigma^2}{4}\right)^{-1/2}\exp\left\{\frac{-1}{2}\beta_1\left(\frac{c\sigma^2}{4}\right)^{-1}\beta_1\right\}, \qquad \beta_2 \geqslant 0.$$

(d) *Case 4*: $\beta_1^*<0$ and $\beta_1^*+2\beta_2^*<0$. Then $f_*'(0)=\beta_1^*<0$ and $f_*'(1)=\beta_1^*+2\beta_2^*<0$. In this case f_* is decreasing over the entire interval [0, 1]. To force the function to be non-decreasing we set $(\beta_1, \beta_2)=(0,0)$. The probability that is associated with all values of (β_1^*, β_2^*) in the third quadrant is accumulated and placed at the origin, i.e. $h\{0,0|I=(1,1)\}=\frac{1}{4}$.

Similar reasoning is used to construct the mixture priors for the $I=(1,0)$ and $I=(0,1)$ cases. For $I=(0,0)$, $(\beta_1, \beta_2)=(0,0)$ with probability 1.

The above technique for constructing priors can be generalized to handle any number of knots m. However, the boundary value distributions become tedious to derive when m is large. To simplify the prior we make the transformation $\gamma_I = L_I \beta_I$. For the $m = 0$ case that was considered above, the constrained parameter space that guarantees non-decreasing functions is the first quadrant and the boundaries are the non-negative γ_1- and γ_2-axes. For $I = (1, 1)$, the prior using the γ-parameterization is $N(0, c\sigma^2 I_2)$ for $\gamma_1 > 0$ and $\gamma_2 > 0$, where I_2 is the 2×2 identity matrix, $0.5\, N(0, c\sigma^2)$ for $\gamma_1 = 0$ and $\gamma_2 > 0$, $0.5\, N(0, c\sigma^2)$ for $\gamma_1 > 0$ and $\gamma_2 = 0$, and probability $\frac{1}{4}$ at $\gamma_1 = 0$ and $\gamma_2 = 0$.

The prior using the γ-parameterization can be generalized in a straightforward manner to handle any number of knots m. In general, the portion of the parameter space that guarantees a monotone increasing function is the multidimensional generalization of the first quadrant where each element of $\gamma_I = L_I \beta_I$ is positive whereas the boundaries are the hyperplanes that border this quadrant. The density function for a value on a specific hyperplane boundary can be obtained by integrating $f(\gamma_I) = (2\pi)^{-m_I/2} (c\sigma^2)^{-m_I/2} \exp\{-1/(2c\sigma^2)\gamma_I' \gamma_I\}$ over the γ-values that take on the value 0 on the hyperplane, where m_I is the sum of the elements of I. This corresponds to accumulating the probability in a multidimensional quadrant that does not produce a non-decreasing function and placing it on the boundary of the constrained parameter space. It is possible, although tedious, to transform the mixture prior for γ_I back to a mixture prior for β_I. However, as discussed in Section 3.3, this is not necessary for constructing the sampling scheme.

We also need to specify priors for α, σ^2 and I_j. The prior for α is $N(0, 10^{10})$ and the prior for σ^2 is a flat prior on $[0, 10^3]$. The I_j are assumed to be *a priori* independent with $\text{pr}(I_j = 1) = p_j$ for $j = 1, \ldots, m+2$. The values used for p_j are discussed in Section 4.

3.3. *Markov chain Monte Carlo sampling scheme*

It is difficult to construct an MCMC sampling scheme working directly with β because the isotonicity constraints on the elements of β change as I changes (i.e. as variables are added or removed from the regression) and the boundary distributions for β are difficult to construct for m large. For this reason, we use an alternative representation of the function $f_m(x)$ in equation (5) that has the same probability structure as f_m but is easier to work with in the construction of the MCMC sampling scheme. More specifically, let $\gamma = (\gamma_1, \ldots, \gamma_{m+2})'$ and, for a given I, define $f_{m,\gamma} = \alpha\iota + W_I \gamma_I$ where $W_I = X_I L_I^{-1}$, X_I consists of the columns of X corresponding to those elements of I that are equal to 1, γ_I consists of the elements of γ corresponding to those elements of I that are equal to 1 and γ_I has the prior that is defined in Section 3.2. It is straightforward to show that, for all I, f_m and $f_{m,\gamma}$ have the same probability distribution. The $f_{m,\gamma}$ representation makes the MCMC sampling scheme easier to construct because the constraints on γ that ensure isotonicity do not change as I changes and the boundary distributions are easy to work with.

For notational purposes, let $I_{(-j)} = I$ without the jth element and $\gamma_{(-j)} = \gamma$ without the jth element. Then the sampling scheme is as follows. Start with some initial values $\gamma^{[0]}$, $I^{[0]}$ and $(\sigma^2)^{[0]}$.

Step 1: generate α conditional on y, I, γ and σ^2.
Step 2: generate σ^2 conditional on y, α, I and γ.
Step 3: generate (I_j, γ_j) jointly conditional on y, α, $I_{(-j)}$, $\gamma_{(-j)}$ and σ^2 by generating

(a) I_j conditional on y, α, $I_{(-j)}$, $\gamma_{(-j)}$ and σ^2, and
(b) γ_j conditional on I_j, y, α, $I_{(-j)}$, $\gamma_{(-j)}$ and σ^2.

Let $\alpha^{[l]}$, $\gamma^{[l]}$ and $I^{[l]}$ be the iterates of α, γ and I in the sampling period. If $w_{I,i}$ represents the ith row of W_I, then an estimate of the posterior mean of the ith element of $f_{m,\gamma}$, and therefore an estimate of $f(x_i)$, is

$$\frac{1}{L}\sum_{l=1}^{L}(\alpha^{[l]} + w_{I^{[l]},i}\gamma_{I^{[l]}}^{[l]}).$$

Steps 1 and 2 are straightforward. To generate I_j conditional on y, α, $I_{(-j)}$, $\gamma_{(-j)}$ and σ^2 in step 3(a), we note that $\text{pr}(I_j=0|y,\alpha,\sigma^2,I_{(-j)},\gamma_{(-j)}) = c^* g(y|I_j=0,\alpha,\sigma^2,I_{(-j)},\gamma_{(-j)})\,\text{pr}(I_j=0)$ and $\text{pr}(I_j=1|y,\alpha,\sigma^2,I_{(-j)},\gamma_{(-j)}) = c^* g(y|I_j=1,\alpha,\sigma^2,I_{(-j)},\gamma_{(-j)})\,\text{pr}(I_j=1)$, where c^* is a constant that does not depend on I_j or γ_j. $g(y|I_j=0,\alpha,\sigma^2,I_{(-j)},\gamma_{(-j)})$ is straightforward to compute and $g(y|I_j=1,\alpha,\sigma^2,I_{(-j)},\gamma_{(-j)})$ can be obtained by integration. Except for a single calculation of the standard normal cumulative distribution function at a specified value the integration can be done analytically.

To generate γ_j in step 3(b) we note that if $I_j=0$ then γ_j is 0 with probability 1. If $I_j=1$, then γ_j is generated from a mixture distribution of a point mass and a constrained normal distribution where the parameters of the mixture distribution have been computed in step 3(a).

4. Small sample properties of the function estimators

This section reports the results of a simulation experiment that was used to compare the small sample properties of the monotone function estimators that were developed in Sections 2 and 3 and the estimators that were proposed by Holmes and Heard (2003) and Neelon and Dunson (2004). Using model (1) the simulation experiment sets $\sigma = 1.0$, considers two sample sizes, $n=100$ and $n=200$, and the following eight functions for $f(x)$:

(a) $f_1(x)=3, x\in(0,10]$ (flat function);
(b) $f_2(x)=0.32\{x+\sin(x)\}, x\in(0,10]$ (sinusoidal function);
(c) $f_3(x)=3$ if $x\leqslant 8$ and $f_3(x)=6$ if $x>8, x\in(0,10]$ (step function);
(d) $f_4(x)=0.3x, x\in(0,10]$ (linear function);
(e) $f_5(x)=0.15\exp(0.6x-3), x\in(0,10]$ (exponential function);
(f) $f_6(x)=3/\{1+\exp(-2x+10)\}, x\in(0,10]$ (logistic function);
(g) $f_7(x)=3\exp[\{-1/2(0.02)^2\}(0.1x-1)^2], x\in(0,10]$ (half-normal function);
(h) $f_8(x)=6\,F(0.1x), x\in(0,10]$, where $F(\cdot)$ is the distribution function for a mixture of $N(0.25,0.004^2)$ and $N(0.75,0.04^2)$ distributions (bimodal distribution function).

The n x-values are equally spaced on (0, 10]. Functions (a)–(c) were considered by Neelon and Dunson (2004) whereas functions (d)–(f) were considered by Holmes and Heard (2003). The half-normal function is an example of a function with a sharp increase at the edge of the x-range whereas the bimodal distribution function is an example of a function with two 'hills'. All the functions except the flat and bimodal distribution functions have a range of approximately 3 to make comparisons across functions easier. For the bimodal distribution function, each hill has a height of 3.

For the constrained Wahba-type spline model that was developed in Section 2 we set $m=20$ equally spaced knots. For the constrained regression spline estimator in Section 3 we set $m=33$ equally spaced knots and $p_j=\text{pr}(I_j=0)=0.8$. p_j is set to 0.8 to counteract the tendency of the model to incorporate too many knots when $m=33$ (which induces function estimates that are too 'unsmooth'). Using $m=33$ knots and $p_j=0.8$ works well empirically as indicated by the simulation results in Table 1. For the Wahba-type spline estimator the MCMC sampling

Table 1. Root-mean-square errors for four monotone function estimators†

Function	*RMSEs for the following estimators:*			
	Regression spline	*Wahba-type spline*	*Neelon and Dunson (2004)*	*Holmes and Heard (2003)*
n = 100				
Flat	0.097	0.092	0.095	0.088
Sinusoidal	0.229	0.223	0.214	0.309
Step	0.285	0.335	0.362	0.173
Linear	0.240	0.220	0.237	0.315
Exponential	0.213	0.225	0.208	0.306
Logistic	0.194	0.209	0.190	0.299
Half-normal	0.165	0.201	0.250	0.255
Bimodal	0.296	0.349	0.362	0.288
n = 200				
Flat	0.062	0.062	0.064	0.059
Sinusoidal	0.194	0.178	0.175	0.254
Step	0.215	0.312	0.317	0.104
Linear	0.181	0.152	0.180	0.264
Exponential	0.150	0.154	0.149	0.245
Logistic	0.139	0.159	0.138	0.228
Half-normal	0.134	0.130	0.191	0.181
Bimodal	0.247	0.314	0.310	0.246

†All results are based on 50 simulation runs. The functions are defined in Section 4.

scheme was run for a warm-up period of 200000 iterations and a sampling period of 500000 iterations and for the constrained regression spline estimator it was run for a warm-up period of 50000 iterations and a sampling period of 100000 iterations.

If $\hat{f}_0(x_i)$ is the estimate of $f_0(x_i)$, we use the root-mean-square error

$$\text{RMSE} = \sqrt{\left[\frac{1}{n}\sum_{i=1}^{n}\{f_0(x_i) - \hat{f}_0(x_i)\}^2\right]}$$

to quantify the accuracy of $\hat{f}_0(x_i)$, where the x_i are the n equally spaced x-values. Table 1 gives RMSE for the four function estimators.

The simulation results in Table 1 indicate that the regression spline method does considerably better than the Wahba-type spline method for the step and bimodal distribution functions, slightly better for the exponential and logistic functions, about the same for the flat and half-normal functions, and slightly worse for the sinusoidal and linear functions. On the basis of the simulation evidence for this set of functions, it appears that the regression spline method is a good robust choice since it is always competitive with the Wahba method and does considerably better for functions that change direction sharply, such as the step and bimodal distribution functions.

The regression spline method does considerably better than the Neelon and Dunson (2004) method for the step, half-normal and bimodal distribution functions. It does about the same for the flat, linear, exponential and logistic functions whereas the Neelon and Dunson method does slightly better for the sinusoidal function. The same pattern holds for both sample sizes. In general, the regression spline method does better when the function being estimated changes

direction sharply and both methods have similar estimation properties for smooth functions, including the flat function.

The Holmes and Heard (2003) method does considerably better than the regression spline method for the step function and about the same for the flat and bimodal distribution functions. However, their method does considerably worse than the regression spline method for all other functions, including the half-normal function which also changes direction sharply. It is not surprising that the Holmes and Heard method does well for a step function because this function is a special case of their more general model with $M=2$ (i.e. one changepoint) and two values of μ (i.e. the function values of each of the two flat portions of the step function).

5. Asymptotic properties of the function estimators

This section shows that Bayesian estimators of the type given in Sections 2 and 3 are consistent estimators. Consistency of Bayes estimators holds under very general conditions when the number of parameters is finite. However, the statistical literature (e.g. Freedman (1999) and Cox (1993)) has provided numerous surprises and anomalies when the number of parameters is infinite. In our problem, essentially each value of the function on the interval $0 \leqslant x \leqslant 1$ is a parameter so it is necessary to exercise care.

Consider the model

$$y_i|x_i \sim N\{\theta(x_i), \sigma^2\}$$

where σ is assigned a prior distribution $\pi(\sigma)$ on $(0, \infty)$ and θ has prior distribution $\Pi(\mathrm{d}\theta)$, a probability measure on the space of non-decreasing continuous functions on [0,1]. For $\xi = (\theta, \sigma) \in (\Theta, (0, \infty))$, $\hat{\xi}_n$ exists and, under certain conditions,

$$n^{-1} \log\{l_n(\hat{\xi}_n)/l_n(\xi_0)\} \to 0 \qquad \text{almost surely } P_0^{\infty}, \tag{6}$$

where $\theta_0 \in \Theta$ is the true value of θ and σ_0 is the true value of σ. (The notation is slightly different from that in previous sections with $\theta_0(x)$ now representing the function to be estimated instead of $f_0(x)$. The letter f is used below to represent a density function.) Here P_0 is the true distribution of (x, y), given by $x \sim G_0$ and $y|x \sim N\{\theta_0(x), \sigma_0^2\}$, and $l_n(\xi)$ is the likelihood function, given by

$$l_n(\xi) = \prod_{i=1}^{n} N\{y_i|\theta(x_i), \sigma^2\}.$$

That result (6) is true for the space of functions that is considered in this paper follows from Robertson and Wright (1975), corollary 2.4, and is shown in Appendix A.

The posterior distribution for a set A is given by

$$\Pi_n(A) = \Pi\{A|(x_1, y_1), \ldots, (x_n, y_n)\} = \frac{\int_A \{l_n(\xi)/l_n(\xi_0)\}\, \Pi(\mathrm{d}\theta)\, \pi(\sigma)\, \mathrm{d}\sigma}{\int \{l_n(\xi)/l_n(\xi_0)\}\, \Pi(\mathrm{d}\theta)\, \pi(\sigma)\, \mathrm{d}\sigma}$$

and we shall consider

$$A_\varepsilon = \{\xi : d(\xi, \xi_0) > \varepsilon\},$$

where

$$d(\xi_1, \xi_2) = \int d_{\mathrm{H}}[N\{\theta_1(x), \sigma_1^2\}, N\{\theta_2(x), \sigma_2^2\}] G_0(\mathrm{d}x)$$

and d_{H} is a version of the Hellinger distance, given by

$$d_{\mathrm{H}}(f_1, f_2) = 1 - \int \sqrt{(f_1 f_2)}$$

for densities f_1 and f_2. Our aim is to show that $\Pi_n(A_\varepsilon) \to 0$ almost surely P_0^∞ for all $\varepsilon > 0$.

For this, consider

$$\begin{aligned} J_n(A_\varepsilon) &= \int_{A_\varepsilon} \{l_n(\xi)/l_n(\xi_0)\}\, \Pi(\mathrm{d}\theta)\, \pi(\sigma)\, \mathrm{d}\sigma \\ &\leqslant \{l_n(\hat{\xi}_n)/l(\xi_0)\}^{1/2} \int_{A_\varepsilon} \{l_n(\xi)/l_n(\xi_0)\}^{1/2}\, \Pi(\mathrm{d}\theta)\pi(\sigma)\, \mathrm{d}\sigma. \end{aligned}$$

With result (6) holding for $\hat{\xi}_n$, we have that

$$\{l_n(\hat{\xi}_n)/l(\xi_0)\}^{1/2} \leqslant \exp(nd)$$

almost surely for all large n for any $d > 0$. So now consider

$$K_n(A_\varepsilon) = \int_{A_\varepsilon} \{l_n(\xi)/l_n(\xi_0)\}^{1/2}\, \Pi(\mathrm{d}\theta)\, \pi(\sigma)\, \mathrm{d}\sigma$$

and note that

$$\begin{aligned} E\{K_n(A_\varepsilon)\} &= \int_{A_\varepsilon} \prod_{i=1}^{n} \left\{ \int (1 - d_{\mathrm{H}}[N\{\theta(x_i), \sigma^2\}, N\{\theta_0(x_i), \sigma_0^2\}]) G_0(\mathrm{d}x_i) \right\} \Pi(\mathrm{d}\theta)\pi(\sigma)\, \mathrm{d}\sigma \\ &= \int_{A_\varepsilon} \{1 - d(\xi, \xi_0)\}^n\, \Pi(\mathrm{d}\theta)\pi(\sigma)\, \mathrm{d}\sigma \\ &\leqslant (1 - \varepsilon)^n. \end{aligned}$$

Therefore, $P\{K_n(A_\varepsilon) > \exp(-n\eta)\} < \exp(n\eta)\exp[-n\{-\log(1-\varepsilon)\}]$ and hence, from the Borel–Cantelli lemma, $K_n(A_\varepsilon) < \exp(-n\eta)$ almost surely for all large n for any $\eta < -\log(1-\varepsilon)$. Hence $J_n(A_\varepsilon) < \exp\{-n(\eta - d)\}$ almost surely for all large n for any $\eta < -\log(1-\varepsilon)$ and we shall obviously choose $d < \eta$. This technique, of using a combination of the consistency of the maximum likelihood estimator and a Bayesian component, was first introduced in Walker and Hjort (2001) and is applicable whenever the maximum likelihood estimator exists, which it does in isotonic regression.

We now look at the denominator

$$I_n = \int \{l_n(\xi)/l_n(\xi_0)\}\, \Pi(\mathrm{d}\theta)\, \pi(\sigma)\, \mathrm{d}\sigma,$$

and merely note that if

$$\Pi\{\xi : d_K(\xi, \xi_0) < \delta\} > 0 \tag{7}$$

for all $\delta > 0$, where

$$d_K(\xi, \xi_0) = \int D[N\{\theta(x), \sigma^2\}, N\{\theta_0(x), \sigma_0^2\}]\, G_0(\mathrm{d}x),$$

and D denotes the Kullback–Leibler divergence, then $I_n > \exp(-nc)$ almost surely for all large n for any $c > 0$. This is a well-known result; see for example Schwartz (1965).

Putting together the result for $J_n(A_\varepsilon)$ and I_n we have that $\Pi_n(A_\varepsilon) \to 0$ almost surely. Therefore, Bayesian consistency follows for all priors Π for which ξ_0 is in the Kullback–Leibler support of Π (i.e. for all ξ_0 satisfying inequality (7)), and the conditions under which result (6) holds for $\hat{\xi}_n$.

In the paper, finite approximations to an infinite dimensional prior are used, which is a common practice when using finite dimensional models. However, consistency is typically still proven for the infinite dimensional model, since this is where any problem can go wrong. If the dimension is fixed then consistency holds for any fixed m, but only for ξ_0 in a smaller set of possible true values. It is possible to allow m to increase with the sample size so that $m_n \to \infty$ as $n \to \infty$ and consistency holds however fast this happens and for all ξ_0 in the Kullback–Leibler support of Π.

From the convexity of d, it follows that

$$d(\bar{\xi}_n, \xi_0) \leqslant \int d(\xi, \xi_0)\, \Pi\{\mathrm{d}\xi | (x_1, y_1), \ldots, (x_n, y_n)\},$$

where $\bar{\xi}_n$ is the Bayes estimate of ξ. Now, letting Π_n denote the posterior,

$$\begin{aligned}\int d(\xi, \xi_0)\, \Pi_n(\mathrm{d}\xi) &= \int_{A_\varepsilon} d(\xi, \xi_0)\, \Pi_n(\mathrm{d}\xi) + \int_{A_\varepsilon^c} d(\xi, \xi_0)\, \Pi_n(\mathrm{d}\xi) \\ &\leqslant \Pi_n(A_\varepsilon) + \varepsilon\end{aligned}$$

since $d(\xi, \xi_0) \leqslant 1$. This is true for all $\varepsilon > 0$ and hence from the consistency result for Π_n it is seen that the Bayes estimate is consistent almost surely P_0^∞ with respect to d.

An explicit expression for the Hellinger distance between two normal distributions yields

$$\sqrt{\left\{\frac{2\sqrt{(\lambda_n \lambda_0)}}{\lambda_n + \lambda_0}\right\}} E\left[\exp\left\{\frac{-Z_n(x)\lambda_n\lambda_0}{4(\lambda_n + \lambda_0)}\right\}\right] \to 1 \qquad \text{almost surely } P_0^\infty,$$

where $Z_n(x) = \{\bar{\theta}_n(x) - \theta_0(x)\}^2$, $\lambda_n = \bar{\sigma}_n^{-2}$, $\lambda_0 = \sigma_0^{-2}$ and the expectation is with respect to the distribution of x. This ensures that $\lambda_n \to \lambda_0$ almost surely P_0^∞ and also that

$$G_0\{x : Z_n(x) > \varepsilon\} \to 0 \qquad \text{almost surely } P_0^\infty \text{ for all } \varepsilon > 0.$$

By exploiting the special conditions that are assumed to hold in this paper—namely, normal distributions for the data and continuous monotone functions for the function space—it is possible to prove consistency of the Bayes estimator in the sup-metric as well as in the Hellinger metric. For this, consider

$$B_\varepsilon = \{\xi = (\theta, \sigma^2) : \sup_x |\theta(x) - \theta_0(x)| > \varepsilon \text{ or } |\sigma/\sigma_0 - 1| > \varepsilon\}.$$

We aim to show first that $\Pi_n(B_\varepsilon) \to 0$ almost surely $[P_0^\infty]$ for all $\varepsilon > 0$ and then that $\max\{\sup_x |\bar{\theta}_n(x) - \theta_0(x)|, |\bar{\sigma}_n/\sigma_0 - 1|\} \to 0$, almost surely $[P_0^\infty]$ where $(\bar{\theta}_n, \bar{\sigma}_n^2)$ is the Bayes estimate of (θ_0, σ_0^2).

We begin by making a two-part observation about the total variation distance between $N(\mu, \sigma^2)$ and $N(\mu_0, \sigma_0^2)$ distributions, which is $\frac{1}{2}\int |f_{N(\mu,\sigma^2)} - f_{N(\mu_0,\sigma_0^2)}|\,\mathrm{d}x$ in terms of densities. First, we observe that

$$\frac{1}{2}\int |f_{N(\mu,\sigma^2)} - f_{N(\mu_0,\sigma_0^2)}|\,\mathrm{d}x \geqslant \frac{1}{2}\left\{\frac{1}{2} - P\left(Z > \frac{|\mu - \mu_0|}{\sigma_0}\right)\right\},$$

regardless of σ^2, where Z is $N(0, 1)$. To see this, suppose that $\mu_0 \leqslant \mu$ without loss of generality. The area between densities is larger than the area between tails on either the right or the left:

$$\int |f_{N(\mu,\sigma^2)} - f_{N(\mu_0,\sigma_0^2)}|\,\mathrm{d}x \geqslant \max\left\{\int_{-\infty}^{\mu_0}(\cdot), \int_{\mu}^{\infty}(\cdot)\right\}$$
$$= \max\left[\frac{1}{2} - P\{N(\mu,\sigma^2) \leqslant \mu_0\}, \frac{1}{2} - P\{N(\mu_0,\sigma_0^2) \geqslant \mu\}\right]$$
$$= \max\left\{\frac{1}{2} - P\left(Z \geqslant \frac{\mu-\mu_0}{\sigma}\right), \frac{1}{2} - P\left(Z \geqslant \frac{\mu-\mu_0}{\sigma_0}\right)\right\}.$$

So, if $|\mu - \mu_0| > \varepsilon_{\sup}$, then

$$\frac{1}{2}\int |f_{N(\mu,\sigma^2)} - f_{N(\mu_0,\sigma_0^2)}|\,\mathrm{d}x > \frac{1}{2}\left\{\frac{1}{2} - P\left(Z > \frac{\varepsilon_{\sup}}{\sigma_0}\right)\right\}.$$

Since the total variation and Hellinger metrics are equivalent, therefore $d_{\mathrm{H}}\{N(\mu,\sigma^2), N(\mu_0,\sigma_0^2)\} > \varepsilon_{\mathrm{H}}$, regardless of σ^2, whenever $|\mu - \mu_0| > \varepsilon_{\sup}$.

Second, we can similarly verify that, if $|\sigma/\sigma_0 - 1| > \varepsilon_{\sup}$, then $\int |f_{N(\mu,\sigma^2)} - f_{N(\mu_0,\sigma_0^2)}|\,\mathrm{d}x$ is bounded below by

$$(2\pi)^{-1/2}\left|\frac{\min(\sigma_0,\sigma)}{\max(\sigma_0,\sigma)} - 1\right| \min\left(1, \sqrt{\left[2\log\left\{\max\left(\frac{\sigma_0}{\sigma}, \frac{\sigma}{\sigma_0}\right)\right\}\right]}\right)$$
$$\geqslant (2\pi)^{-1/2}\frac{\varepsilon_{\sup}}{2}\min[1, \sqrt{\{2\log(1+\varepsilon_{\sup})\}}]$$

regardless of μ, for $\varepsilon_{\sup}$ sufficiently small. (This lower bound can be obtained as the area of an isosceles triangle inscribed between the $N(\mu,\sigma^2)$ and $N(\mu_0,\sigma_0^2)$ densities, within the range where the taller density is concave.) Since the total variation and Hellinger metrics are equivalent, therefore $d_{\mathrm{H}}\{N(\mu,\sigma^2), N(\mu_0,\sigma_0^2)\} > \varepsilon_{\mathrm{H}}$, regardless of σ^2, whenever $|\sigma/\sigma_0 - 1| > \varepsilon_{\sup}$.

Next, we show that $\Pi_n(B_\varepsilon) \to 0$ almost surely $[P_0^\infty]$ by using the continuity and monotonicity of our functions. Since θ_0 is uniformly continuous on [0,1] by assumption, then given $\varepsilon_{\sup}$ we have $|\theta_0(x+\delta) - \theta_0(x)| < \varepsilon_{\sup}/2$ uniformly in x for some δ. Now, if θ is a monotone function such that $\sup_x |\theta(x) - \theta_0(x)| > \varepsilon_{\sup}$, then $|\theta(x) - \theta_0(x)| > \varepsilon_{\sup}/2$ for all x in some interval $(x_0, x_0+\delta)$, where x_0 may depend on θ. By the first part of our observation above, we have $d_{\mathrm{H}}[N\{\theta(x),\sigma^2\}, N\{\theta_0(x),\sigma_0^2\}] > \varepsilon_{\mathrm{H}_1}$ for all x in $(x_0, x_0+\delta)$. Likewise, if θ is a monotone function such that $|\sigma/\sigma_0 - 1| > \varepsilon_{\sup}$, then by the second part of our observation we have $d_{\mathrm{H}}[N\{\theta(x),\sigma^2\}, N\{\theta_0(x),\sigma_0^2\}] > \varepsilon_{\mathrm{H}_2}$ for all x in [0,1]. In either case,

$$d(\xi,\xi_0) = \int d_{\mathrm{H}}[N\{\theta(x),\sigma^2\}, N\{\theta_0(x),\sigma_0^2\}]G_0(\mathrm{d}x)$$
$$\geqslant \varepsilon_{\mathrm{H}}\{G_0(x_0+\delta) - G_0(x_0)\}$$
$$\geqslant \varepsilon_{\mathrm{H}} \inf_{0\leqslant x_0\leqslant 1-\delta}\{G_0(x_0+\delta) - G_0(x_0)\},$$

with $\varepsilon_{\mathrm{H}} = \max(\varepsilon_{\mathrm{H}_1}, \varepsilon_{\mathrm{H}_2})$. G_0 assigns positive probability to every interval by assumption, and the lower bound does not depend on θ. Therefore, the set $B_{\varepsilon_{\sup}}$ of monotone functions with large sup-distances is contained in some set $A_{\varepsilon_{\mathrm{H}}}$ of monotone functions with large Hellinger distances. Thus, $\Pi_n(B_{\varepsilon_{\sup}}) \leqslant \Pi_n(A_{\varepsilon_{\mathrm{H}}})$. We have already shown that the latter converges to 0 almost surely $[P_0^\infty]$ for all $\varepsilon_{\mathrm{H}} > 0$. Therefore, $\Pi_n(B_{\varepsilon_{\sup}}) \to 0$ almost surely $[P_0^\infty]$ for all $\varepsilon_{\sup} > 0$.

Now we turn to showing the consistency of the Bayes estimate $\bar{\xi}_n = (\bar{\theta}_n, \bar{\sigma}_n^2)$ in the sup-metric. Define $D_n(\bar{\xi}_n, \xi_0) = \max\{\sup_x |\bar{\theta}_n(x) - \theta_0(x)|, |\bar{\sigma}_n/\sigma_0 - 1|\}$. Consider $E_\varepsilon = \{(x_1,y_1), \ldots, (x_n,y_n), \ldots : D_n(\bar{\xi}_n, \xi_0) > \varepsilon \text{ for infinitely many } n\}$. Our aim is to show that $P_0^\infty(E_\varepsilon) = 0$. If $D_n(\bar{\xi}_n, \xi_0) > \varepsilon$ for some realization $(x_1,y_1), \ldots, (x_n,y_n), \ldots$, then by repeating the preceding argument we have

$$d(\bar{\xi}_n, \xi_0) = \int d_{\mathrm{H}}[N\{\bar{\theta}_n(x), \bar{\sigma}_n^2\}, N\{\theta_0(x), \sigma_0^2\}]G_0(\mathrm{d}x)$$
$$\geqslant \varepsilon_{\mathrm{H}} \inf_{0 \leqslant x_0 \leqslant 1-\delta} \{G_0(x_0+\delta) - G_0(x_0)\},$$

where the lower bound does not depend on $\bar{\xi}_n$. Thus

$$E_\varepsilon \subset \{(x_1, y_1), \ldots, (x_n, y_n), \ldots : d(\bar{\xi}_n, \xi_0) > \varepsilon_{\mathrm{H}} \text{ for infinitely many } n\}.$$

But the latter set has P_0^∞ probability 0 since we previously showed that $d(\bar{\xi}_n, \xi_0) \to 0$ almost surely $[P_0^\infty]$. This completes the argument.

6. Application—market response functions

The constrained regression spline method that was developed in Section 3 can be extended to estimate multiple functions in a model where each function is constrained to be either monotonically non-decreasing or non-increasing. An example of such a situation is the estimation of market response functions that model the change in demand for a product due to a change in the product's own price and the prices of competing products. In this section we apply our methodology to estimate the response function for the weekly sales of stick margarine at a large grocery chain. The following model is used:

$$\log\{\mathrm{Sales}(t)\} = f_1\{p_1(t)\} + f_2\{p_2(t)\} + \delta_1 A(t) + \delta_2 S_2(t) + \delta_3 S_3(t) + \delta_4 S_4(t) + \varepsilon(t)$$

where Sales(t) represents the sales of a specific brand of margarine in week t, $p_1(t)$ is the price of the margarine (in cents) in week t, $p_2(t)$ is the price of the brand's major competitor in the category, $A(t)$ is a dummy variable that indicates whether the brand under consideration was featured in a retail flyer in week t and $S_q(t)$ is a quarterly dummy variable indicating whether period t is from the qth quarter, $q = 2, 3, 4$. The four seasons are defined as March–May, June–August, September–November and December–February. March–May was designated as the base season and $S_2(t)$, $S_3(t)$ and $S_4(t)$ correspond to the other three seasons respectively. The non-parametric estimation of market response functions in this model without monotonicity constraints was considered by Kalyanam and Shively (1998).

Economic theory suggests that the demand function f_1 is a monotonically non-increasing function, i.e. demand for a product decreases or remains flat as its price increases, all other things equal. Therefore, f_1 should be constrained to be monotonically non-increasing. Conversely, demand for a product increases or remains flat as the price of a competing product increases so the cross-price function f_2 should be constrained to be monotonically non-decreasing. Although economic theory indicates that the functions f_1 and f_2 are monotonic it does not suggest a specific functional form. Therefore, a non-parametric estimation technique is appropriate that allows the data to specify the functional forms of the relationships subject to the constraint of monotonicity.

Fig. 1(a) provides the estimated demand function for Sales which was obtained by using the constrained regression spline method (full curve) and Smith and Kohn's unconstrained regression spline method (broken curve). The demand functions are obtained by estimating f_1, f_2 and the δ_j-coefficients, setting p_2 to 69 cents and the four dummy variables to 0 (which implies that the product is not featured in a retail flyer and the week is in the March–May period), and allowing p_1 to vary over its range. Similarly, Fig. 1(b) provides the cross-price response functions for both methods.

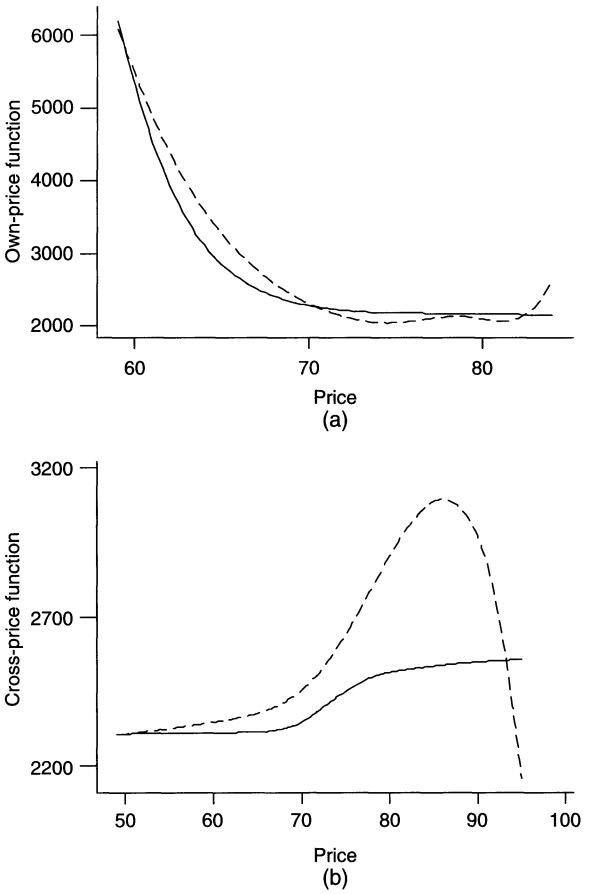

Fig. 1. Estimated market response functions obtained by using the constrained regression spline method (———) and Smith and Kohn's (1996) unconstrained regression spline method (– – –): (a) own price response functions; (b) cross-price response functions

Figs 1(a) and 1(b) indicate that the unconstrained estimated response functions are not monotonic, particularly the cross-price function. For both functions, it is difficult to interpret and explain the non-monotonicity by using economic theory. Also, the cross-price functions in Fig. 1(b) are substantially different. For the demand function in Fig. 1(a) there is a considerable amount of data at the highest prices ($p_1 = 84$ cents for 13 weeks, 79 cents for 5 weeks and 75 cents for 10 weeks) so the non-monotonicity cannot be easily explained by a lack of data. For the cross-price function there are 3 weeks when the price p_2 is 95 cents and 13 weeks when the price is 89 cents.

Appendix A

To establish that result (6) holds for continuous monotone functions, we utilize Robertson and Wright (1975). Maximization of l_n subject to $\theta(x)$ being monotone results in

$$\hat{\theta}(x_i) = \max_{j \leqslant i} \min_{k \geqslant i} \left\{ \frac{1}{k-j+1} \sum_{\nu=j}^{k} y_\nu \right\},$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \{y_i - \hat{\theta}(x_i)\}^2,$$

where, without loss of generality, we assume that $x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n$. In the notation of Robertson and Wright (1975) our $\hat{\theta}(x_i)$ is their $\hat{m}_n(t_i)$. The sample mean of a normal distribution is well known to satisfy the exponential probability bound (2.6) of Robertson and Wright (1975). It is easy to verify the other conditions of corollary 2.4 of Robertson and Wright (1975), from which we have $P\{\max_i |\hat{\theta}(x_i) - \theta_0(x_i)| \geqslant \varepsilon\} \leqslant C_1 \rho_1^n$ for $\rho_1 < 1$. Hence,

$$P\left[\frac{1}{n}\sum\{\hat{\theta}(x_i) - \theta_0(x_i)\}^2 \geqslant \varepsilon^2\right] \leqslant \sum P\{|\hat{\theta}(x_i) - \theta_0(x_i)| \geqslant \varepsilon\} \leqslant nC_1\rho_1^n,$$

which is summable, and therefore $(1/n)\Sigma\{\hat{\theta}(x_i) - \theta_0(x_i)\}^2 \to 0$ almost surely $[P_0^\infty]$. Now

$$\frac{1}{n}\sum\{y_i - \hat{\theta}(x_i)\}^2 \leqslant \frac{1}{n}\sum\{y_i - \theta_0(x_i)\}^2$$

by maximization of $l_n(\xi)$. This inequality and the triangle inequality yield

$$\begin{aligned}\frac{1}{n}\sum\{y_i - \theta_0(x_i)\}^2 &\leqslant \frac{1}{n}\sum\{y_i - \hat{\theta}(x_i)\}^2 + \frac{1}{n}\sum\{\hat{\theta}(x_i) - \theta_0(x_i)\}^2 \\ &\leqslant \frac{1}{n}\sum\{y_i - \theta_0(x_i)\}^2 + \frac{1}{n}\sum\{\hat{\theta}(x_i) - \theta_0(x_i)\}^2.\end{aligned}$$

Now $(1/n)\Sigma\{y_i - \theta_0(x_i)\}^2 \to \sigma_0^2$ almost surely $[P_0^\infty]$. So, by taking the limit in the preceding inequality, we have $(1/n)\Sigma\{y_i - \hat{\theta}(x_i)\}^2 \to \sigma_0^2$ almost surely $[P_0^\infty]$. Therefore,

$$\frac{1}{n}\log\left\{\frac{l_n(\hat{\xi})}{l_n(\xi_0)}\right\} = \frac{1}{2\hat{\sigma}^2}\left[\frac{1}{n}\sum\{y_i - \hat{\theta}(x_i)\}^2 - \frac{1}{n}\sum\{y_i - \theta_0(x_i)\}^2\right] \to 0 \qquad \text{almost surely } [P_0^\infty].$$

References

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972) *Statistical Inference under Order Restrictions: the Theory and Application of Isotonic Regression*. London: Wiley.

Casella, G. and George, E. I. (1992) Explaining the Gibbs sampler. *Am. Statistn*, **46**, 167–174.

Cox, D. D. (1993) An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.*, **21**, 903–923.

Damien, P., Wakefield, J. and Walker, S. (1999) Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *J. R. Statist. Soc.* B, **61**, 331–344.

Damien, P. and Walker, S. G. (2001) Sampling truncated normal, beta, and gamma densities. *J. Comput. Graph. Statist.*, **10**, 206–215.

Freedman, D. A. (1999) On the Bernstein-von Mises theorem with infinite dimensional parameters. *Ann. Statist.*, **4**, 1119–1140.

Friedman, J. and Tibshirani, R. (1984) The monotone smoothing of scatterplots. *Technometrics*, **26**, 243–250.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.

Holmes, C. C. and Heard, N. A. (2003) Generalized monotonic regression using random change points. *Statist. Med.*, **22**, 623–638.

Kalyanam, K. and Shively, T. S. (1998) Estimating irregular pricing effects: a stochastic spline approach. *J. Marktng Res.*, **35**, 16–29.

Mammen, E. (1991) Estimating a smooth monotone regression function. *Ann. Statist.*, **19**, 724–740.

Neelon, B. and Dunson, D. B. (2004) Bayesian isotonic regression and trend analysis. *Biometrics*, **60**, 398–406.

Ramsay, J. O. (1998) Estimating smooth monotone functions. *J. R. Statist. Soc.* B, **60**, 365–375.

Robertson, T. and Wright, F. T. (1975) Consistency in generalized isotonic regression. *Ann. Statist.*, **3**, 350–362.

Schwartz, L. (1965) On Bayes procedures. *Z. Wahrsch. Ver. Geb.*, **4**, 10–26.

Smith, M. and Kohn, R. (1996) Nonparametric regression using Bayesian variable selection. *J. Econometr.*, **75**, 317–343.

Tierney, L. (1994) Markov chains for exploring posterior distributions. *Ann. Statist.*, **22**, 1701–1762.

Wahba, G. (1978) Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. R. Statist. Soc.* B, **40**, 364–372.

Walker, S. and Hjort, N. L. (2001) On Bayesian consistency. *J. R. Statist. Soc.* B, **63**, 811–821.

Wong, C. and Kohn, R. (1996) A Bayesian approach to additive semiparametric regression. *J. Econometr.*, **74**, 209–235.

Wright, I. and Wegman, E. (1980) Isotonic, convex, and related splines. *Ann. Statist.*, **8**, 1023–1035.