J. R. Statist. Soc. B (1998)
60, *Part* 2, *pp.* 333–350

Automatic Bayesian curve fitting

D. G. T. Denison, B. K. Mallick and A. F. M. Smith†

Imperial College of Science, Technology and Medicine, London, UK

[Received December 1995. Final revision June 1997]

Summary. A method of estimating a variety of curves by a sequence of piecewise polynomials is proposed, motivated by a Bayesian model and an appropriate summary of the resulting posterior distribution. A joint distribution is set up over both the number and the position of the knots defining the piecewise polynomials. Throughout we use reversible jump Markov chain Monte Carlo methods to compute the posteriors. The methodology has been successful in giving good estimates for 'smooth' functions (i.e. continuous and differentiable) as well as functions which are not differentiable, and perhaps not even continuous, at a finite number of points. The methodology is extended to deal with generalized additive models.

Keywords: Additive models; Back-fitting algorithm; Least squares regression; Piecewise polynomials; Reversible jump Markov chain Monte Carlo method; Splines

1. Introduction

Regression techniques are among some of the most widely used methods in applied statistics. Given a response Y and explanatory variable X the problem is to estimate an assumed functional relationship between Y and X, and to predict further responses for new values of the covariate. The basic regression model with bivariate observations $(x_1, y_1), \ldots, (x_n, y_n)$ has the form

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

where the ϵ_i are zero-mean random errors and f is an unknown regression function that we wish to estimate. The value $f(X)$ is the conditional expectation of Y given the value X, so it can be used to predict the future values of Y for different measured values of X.

Both parametric and nonparametric techniques are commonly used to find the regression function f. Polynomial regression is a familiar parametric approach but it suffers from various drawbacks, in particular that individual observations can exert an influence, in unexpected ways, on remote parts of the curve. Also, owing to the global nature of polynomial fitting there are problems in estimating wiggly curves. Nonparametric techniques include smoothing splines and kernel smoothers (see Hastie and Tibshirani (1990) for an overview). Adaptive techniques are also available. Parametric examples include spline fitting with adaptive knot placement (Smith, 1982; Friedman and Silverman, 1989; Friedman, 1991) and nonparametric examples include variable bandwidth kernel methods (Müller and Stadtmüller, 1987; Fan and Gijbels, 1995).

One general approach to curve fitting is to allow f to be a *piecewise polynomial* function made up of low order pieces that are non-zero only between adjacent *knot* points. We provide

†*Address for correspondence*: Department of Mathematics, Huxley Building, Imperial College of Science, Technology and Medicine, 180 Queen's Gate, London, SW7 2BZ, UK.
E-mail: a.smith@ic.ac.uk

1369–7412/98/60333

a Bayesian version which models f by a piecewise polynomial with an unknown number of knots at unknown locations, all treated as parameters to be inferred. Between each knot we fit a fixed low order polynomial. The flexible structure resulting from the Bayesian knot selection technique allows us to estimate any type of unknown curve, smooth or wiggly, and compares favourably with spline approximations with many more knots. The method benefits from not overparameterizing the parameter space by choosing too many intervals. It also combats underparameterization leading to a smooth curve that does not accurately reflect the data.

The novelty of the Bayesian methodology here is that without choosing a single collection of knots we are mixing over all the possible subsets of a large collection of prespecified candidate knot sites. The resulting mixture distribution covers a rich class of models and has high predictive power. If we needed to pick a single collection of knots this could be the posterior modal subset of the knot locations or the modes in the posterior probability of the candidate knot locations, which could then be used for the spline smoothing.

The problem of routine calculation of posterior distributions for both the number and the location of knots is addressed using the Markov chain Monte Carlo (MCMC) simulation technique of reversible jumps (Green, 1995).

One of the attractive features of our method is that it can be easily extended to the multivariate case by using additive models. Suppose that the observations are of the form $(y_i, \mathbf{x}_i)$, where each $\mathbf{x}_i$ is now an l-vector $(x_{1i}, \ldots, x_{li})$. It is assumed, as before, that the variable Y depends on $\mathbf{X}$ by a relationship of the form $Y = f(\mathbf{X}) + \text{error} = f(X_1, \ldots, X_l) + \text{error}$. For this paper we concentrate on a particular dependence structure where the function f is a sum of functions of the individual components of $\mathbf{X}$,

$$f(\mathbf{X}) = f_1(X_1) + f_2(X_2) + \ldots + f_l(X_l). \tag{2}$$

This approach is known as additive regression or additive modelling (Hastie and Tibshirani, 1990) and replaces the problem of estimating a function f of an l-dimensional vector $\mathbf{X}$ by the problem of estimating l separate one-dimensional functions f_j. Although not completely general, additive models are often effective and easily interpretable.

In Section 2 we shall introduce the Bayesian model and the algorithm. We give examples of curve fitting in one dimension in Section 3 and Section 4 extends our methodology, via additive models, to multidimensional settings. A short discussion is provided in Section 5.

2. Curve fitting in one dimension

2.1. Piecewise polynomials

The basic idea of piecewise polynomials is to replace the single function f, defined over the entire range of X-values, with several generally low order polynomials each defined over a different subinterval, the union of which is the range of X. The points that delineate the subintervals are the knots. The most popular piecewise polynomial fitting procedures are based on splines (de Boor, 1978). These involve setting up basis functions for the estimate which are defined by the positions of the knots. Splines of order q are usually defined in such a way that they are continuous functions with $q-1$ continuous derivatives. For example the commonly used cubic polynomial spline function (Hastie and Tibshirani, 1990) can be represented simply by the truncated power series basis giving us an estimate of the curve in the form

$$\hat{y}(x) = \hat{\alpha}_0 + \hat{\alpha}_1 x + \hat{\alpha}_2 x^2 + \hat{\alpha}_3 x^3 + \sum_{i=1}^{k} \hat{\beta}_i (x - r_i)_+^3 \qquad \text{for } x \in [r_0, r_{k+1}] \tag{3}$$

where $a_+ = \max(0, a)$, the r_i $(i = 1, \ldots, k)$ are the k interior knots, r_0 and r_{k+1} are the boundary knots and the $\hat{\alpha}$s and $\hat{\beta}$s are found by least squares regression. This form ensures that the estimate is continuous and has continuous first and second derivatives. These assumptions are often used to ensure that the estimate looks smooth but may not be an adequate basic model if the true curve is far from smooth.

Many functions do not have properties which make them suitable for estimation with splines. Functions with rapidly varying first and second derivatives and maybe even discontinuous functions cannot be modelled well with functions that are, by design, 'smooth'. Although smoothness is a desirable property when we know that the true functions are smooth or we wish to find derivatives of our estimates it restricts the flexibility of a model. For this reason we do not use splines in this paper and instead concentrate on piecewise polynomials, which can be chosen to be more flexible than splines. In fact we choose to rewrite equation (1) as

$$\begin{aligned} y_i &= f_{k,l}(x_i) + \epsilon_i \\ &= \sum_{n=0}^{l} \beta_{n,0}(x_i - r_0)_+^n + \sum_{m=1}^{k} \sum_{n=l_0}^{l} \beta_{n,m}(x_i - r_m)_+^n + \epsilon_i, \qquad i = 1, \ldots, n, \end{aligned} \tag{4}$$

where we define $a_+^0 = I(a \geqslant 0)$ (where I is the usual indicator function) and the r_m, indexed in ascending order, are the knot points with the boundary knots given by $r_0 = x_1$ and $r_{k+1} = x_n$ leaving k $(\geqslant 0)$ interior knots in the model. The possible location of the knots r_m are the n regular grid points on the range of X. We could have chosen these candidate knot locations on a continuous scale but chose not to (see Section 2.2). Also l $(\geqslant 0)$ is the order of the piecewise polynomials that we use in the model and l_0 $(\geqslant 0)$ gives us the degree of continuity at the knot points. For instance if $l_0 = 0$ then $f_{k,l}$ is not constrained to be continuous at the knots and for other values of l_0 $(\geqslant 1)$ the estimating function is continuous with $l_0 - 1$ continuous derivatives. In fact taking $l = l_0 = 3$ in equation (4) gives us the spline in equation (3).

The main difficulty of working with piecewise polynomials (and splines) is selecting the number and position of the knots, which control the trade-off between smoothness and flexibility of the estimated curve. The simplest approach requires a single parameter, the number of interior knots. The positions are then chosen uniformly over the range of the data (cardinal splines). A slightly more adaptive version places the knots at appropriate quantiles of the predictor variable. It seems clearly preferable to use the data themselves to select the number and position of the knots, an idea apparently first proposed by Smith (1982). Friedman and Silverman (1989) gave an algorithm for optimizing over the number and location of the knots in an adaptive way. The key idea is first to determine the knot positions by using a piecewise linear function and then to convert the piecewise linear functions to piecewise cubic functions by essentially rounding the corners at each knot. These are then used at the chosen knots to compute the piecewise cubic fit.

2.2. *The Bayesian model*

Using the notation of equation (4), in our model we take the number of interior knots, k, random and fix the order of estimating piecewise polynomials, l and l_0. The errors ϵ_i are assumed to come from an $N(0, \sigma^2)$ distribution, where σ is an unknown constant. The inference is then carried out assuming that the 'true' model is unknown but comes from the class of models $M_0, M_1, \ldots$ where M_k denotes the model with exactly k interior knots. The overall parameter space Θ can then be written as a countable union of subspaces $\Theta = \cup_0^\infty \Theta_k$ where Θ_k is a subspace of the Euclidean space $R^{n(k)}$, and $R^{n(k)}$ denotes the $n(k) = (k+1)$-dimensional parameter space corresponding to model M_k. Here $\theta^{(k)} = (r_1, \ldots, r_k, \sigma^2)$. The joint distribution of $(k, \theta^{(k)}, y)$, where $\theta^{(k)}$ denotes a generic element of Θ_k and y the data vector, is then modelled as

$$p(k, \theta^{(k)}, y) = p(k)\, p(\theta^{(k)}|k)\, p(y|k, \theta^{(k)}), \tag{5}$$

i.e. as the product of model probability, parameter prior and likelihood. Inference about k and $\theta^{(k)}$ will be based on the joint posterior $p(k, \theta^{(k)})$ which is known as the *target* distribution.

The full Bayesian model of the joint posterior density for the models and parameters given the data can be written as

$$p(k, \theta|y) = \frac{1}{Z}\, p(k)\, p(\theta^{(k)}|k)\, p(y|k, \theta^{(k)}) \tag{6}$$

where Z is a normalizing constant and is given by

$$Z = \sum_{k=0}^{\infty} p(k) \left\{ \int_{\Theta_k} p(\theta^{(k)}|y)\, p(y|k, \theta^{(k)})\, \mathrm{d}\theta^{(k)} \right\}. \tag{7}$$

We shall generate samples from the joint posterior of $(k, \theta^{(k)})$ (as in these non-trivial cases of changing dimension the standard MCMC theory does not apply) by using a wider class of reversible jump Metropolis–Hastings algorithms (Green, 1995). Full details of these methods can be found in Green (1995). Here, we focus on the essence of the methodology and the particular forms of the algorithms in our current context.

For normal errors the log-likelihood $l_k(\theta|y)$ is

$$l_k(\theta|y) = \log\{L_k(\theta|y)\} = -n\log(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \{y_i - f_{k,l}(x_i)\}^2, \tag{8}$$

where $f_{k,l}$ is defined by equation (4).

The coefficients $\beta_{n,m}$ defined in equation (4) are determined by the data and the parameter vector $\theta^{(k)}$ and are taken to be the least squares estimates which can easily be calculated with standard least squares regression theory. A complete Bayesian approach would assign proper prior distributions to these polynomial coefficients and work with an extended posterior distribution which also included these parameters. However, this leads to a serious additional computational burden, especially when many knots are required to fit the curve adequately, and comparative studies have shown that the least squares estimation approach leads to no significant deterioration in performance for overall curve estimation.

We use a vague, but proper, prior for the error variance σ^2, i.e. $\pi(\sigma^{-2}) = \text{gamma}(10^{-3}, 10^{-3})$. A Poisson distribution (with parameter λ) is used to specify the prior probabilities for each of the models, giving

$$p(k) = \frac{\lambda^k \exp(-\lambda)}{k!}, \qquad k = 0, 1, 2, \ldots. \tag{9}$$

In practice, a Poisson distribution truncated to $k < n$ or to $k < k_{\max}$, for a suitable choice of $k_{\max}$, is adopted. The choice of λ will be discussed later. The r_i are taken to be the order statistics from a uniform random variable with state space the candidate knot sites $\{x_1,$

$x_2, \ldots, x_n\}$. We could have chosen the candidate knot locations as any point in the range of X but this leads to problems in defining the 'modal' collection of the knots which is why we feel that the candidate knot locations that we use are more desirable.

2.3. Methodology

Our aim is to simulate samples from the joint posterior distribution of $p(\theta^{(k)}, k|y)$, since analytic or numerical analyses are totally intractable in this situation.

Owing to the varying dimensionality of our problem we must design move types between the subspaces Θ_k which will combine to form what Tierney (1994) called a hybrid sampler. These will allow the sampler to explore the combined parameter space freely.

For this problem some possible transitions are

(a) the addition of a knot (a birth step),
(b) the deletion of a knot (a death step) and
(c) the movement of a knot.

In steps (a) and (b) we are changing the dimension of the model.

The location of the proposed knot to add in step (a) is found by uniformly choosing one of the x_i which does not contain a knot within $l+1$ design points to its left or right. The proposed knot to delete in step (b) is simply chosen uniformly from the knots that are currently in the model. The movement step (c) is required because of the limited way in which we chose to add knots. If step (c) were not incorporated in the model then, in practice, the space of possible proposal models can become small when many knots are present. Hence, this step chooses a knot uniformly, say x_c, and then chooses another point uniformly from the set

$$C = \{x_i \colon |i-c| \leqslant l+1 \text{ and no other knots are within } l+1 \text{ candidate locations of } x_i\}$$

and proposes that the knot x_c be moved to this new point. Once the knot locations of the proposed model have been established in each of the three move types standard least squares theory is used to obtain the $\beta_{n,m}$ which are defined in equation (4). This gives us the complete proposed model which is in the same form as $f_{k,l}$ in equation (4).

At the end of each iteration, after the transition step has been performed, we generate a new error variance σ^2 by using a Gibbs step (Gelfand and Smith, 1990). Thus, posterior samples of $(k, r_{(k)}, \sigma^2)$ provide the basis for any required posterior inference or model comparison purpose, in particular, estimating the unknown function f by the Monte Carlo posterior mean, i.e. the pointwise average of the functions arising from each of the samples generated.

2.4. Algorithm

Considering the three move types (a)–(c) we can rewrite this set of moves as $m = \{M, 0, 1, 2, \ldots\}$. Here M means the movement of a knot and $m = 0, 1, 2, \ldots$ refers to increasing the number of interior knots from m to $m+1$ or decreasing from $m+1$ to m. Independent move types are randomly chosen with probabilities η_k for $m = M$, b_k for $m = k$ and d_k for $m = k-1$ which satisfy $\eta_k + b_k + d_k = 1$ for all k. In this problem we took

$$b_k = c \min\{1, p(k+1)/p(k)\}$$

and

$$d_{k+1} = c \min\{1, p(k)/p(k+1)\}$$

for $k = 1, 2, \ldots$, which then forces $\eta_k = 1 - b_k - d_k$. For $k = 0$ we put $b_0 = 1$ and $d_0 = m_0 = 0$. The constant c controls the rate at which move types which change dimension are proposed. We take $c = 0.4$ in the forthcoming examples but other values are equally valid, provided that $c \in [0, \frac{1}{2}]$ as, if $c > \frac{1}{2}$, then the sum of the probabilities b_k and d_k could be greater than 1 for some values of k.

Using the notation of Green (1995), the acceptance probability for each of the move types in our problem is

$$\alpha = \min(1, \text{likelihood ratio} \times \text{prior ratio} \times \text{proposal ratio}).$$

For the move step (c) the prior ratio and proposal ratio are both 1 since all collections of the same number of knots have the same prior probability and the proposals are made from the same distribution.

For the birth step (a) the prior ratio is given by

$$\begin{aligned}\text{prior ratio} &= \frac{\text{prior for } k+1 \text{ knots}}{\text{prior for } k \text{ knots}} \, \frac{\text{prior for location of } k+1 \text{ knots}}{\text{prior for location of } k \text{ knots}} \\ &= \frac{p(k+1)}{p(k)} \, \frac{k+1}{n} \qquad (10)\end{aligned}$$

since we know that if $\mathbf{r} = (r_1, \ldots, r_k)$ where the r_i are drawn from a uniform distribution with n points and $r_1 < r_2 < \ldots < r_k$ then $p(\mathbf{r}) = n^{-k}k!$. The corresponding proposal ratio is given by

$$\text{proposal ratio} = \frac{d_{k+1}/(k+1)}{b_k/\{n - Z(k)\}} \qquad (11)$$

where

$$Z(k) = 2(l+1) + k(2l+1)$$

and is the number of candidate knot locations where a new knot cannot be placed so it is not too close to an existing knot (i.e. $l + 1$ data points away). We chose the proposal probabilities for the birth and death steps in such a way as to ensure

$$b_k \, p(k) = d_{k+1} \, p(k+1)$$

so it follows from equations (10) and (11) that the acceptance probability for a birth step is

$$\alpha = \min \left\{ 1, \text{likelihood ratio } \frac{n - Z(k)}{n} \right\}$$

and for the death step it is the same except that the fraction is inverted.

The algorithm that we use is very simple and works quickly. To monitor our results we look at the mean-squared error (MSE) of the models generated by the MCMC algorithm, given by

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \{y_i - f_{k,l}(x_i)\}^2$$

where $f_{k,l}$ is defined by equation (4). We then think of the chain as having 'converged' when this quantity settles down. This ensures that the points in the sample from the posterior distribution all have a similar 'goodness of fit' for the data. Monitoring the log-posterior for

$(k, \theta^{(k)})$, for example, would be inappropriate because of scaling problems introduced by moves across dimensions.

2.4.1. *Algorithm*

Step 1: initialize a configuration by choosing λ knot locations uniformly along the range $[r_0, r_{k+1}]$ at least $l+1$ points away from each other.
Step 2: set k equal to the number of interior knots in the present model.
Step 3: generate u uniformly on [0,1].
Step 4: go to the move type determined by u—
(i) if $u \leqslant b_k$ then go to the *birth* step;
(ii) otherwise if $b_k < u \leqslant b_k + d_k$ then go to the *death* step;
(iii) otherwise go to the *move* step.
Step 5: draw σ^2 by using a Gibbs step.
Step 6: repeat step 2 until there is little change in the mean-squared error of the models.

3. Examples

3.1. *Smooth functions*

We begin with a relatively simple challenge and consider two smooth functions:

(a) $f(x) = x + 2\exp(-16x^2), \qquad x \in [-2, 2],$

and

(b) $f(x) = \sin(2x) + 2\exp(-16x^2), \qquad x \in [-2, 2].$

Simulated data are created as follows. The function is rescaled so that its support is the unit interval and then is evaluated at 200 points in [0, 1] generated from a $U(0, 1)$ distribution. Zero-mean normal noise is added with σ chosen so that the signal-to-noise ratio is 3 as in Fan and Gijbels (1995), i.e. $\sigma = 0.4$ in example (a) and 0.3 in example (b).

We choose $l_0 = 1$ and $l = 2$, so $f_{k,l}$ is a continuous quadratic piecewise polynomial, and put a Poisson prior over k with $\lambda = 1$, to reflect assumed knowledge that the curves are very smooth and hence that a large number of knots is unlikely to be required. We initially place λ knots as described in the algorithm.

In Figs 1 and 2 we display the true functions in examples (a) and (b) together with estimates to these functions by using our method and the adaptive knot selection method of Friedman and Silverman (1989). We used exactly the same model for the function, given in equation (4) with $l = 2$ and $l_0 = 1$, in the adaptive knot selection algorithm. Our estimates are smoother because we display the posterior mean estimate to the true function, obtained by pointwise averaging, whereas the knot selection technique produces a single estimate which is of the form shown in equation (4).

In both examples our model had a posterior modal number of knots similar to the number of knots found by the adaptive knot selection algorithm. In fact we found the same number (3) in example (a) and we found one more (4) in example (b). The MSEs, given by

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \{\hat{f}(x_i) - f(x_i)\}^2,$$

where f is the true function and $\hat{f}$ is our estimate to the true function, are displayed in Table 1

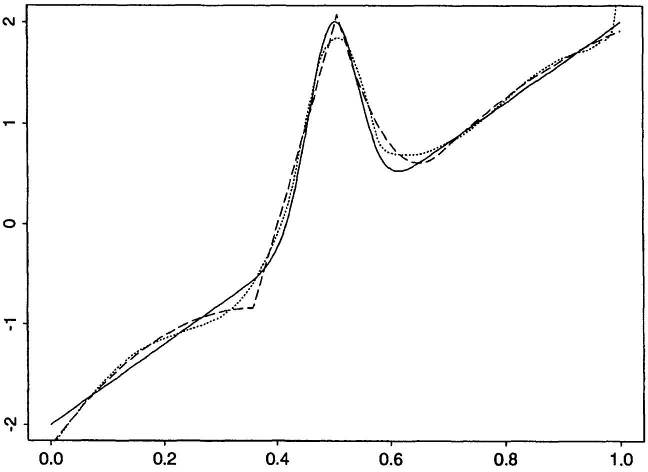


Fig. 1. Example (a), using the continuous piecewise quadratic model: ———, true curve; ··········, Bayes estimate; – – –, adaptive knot selection estimate

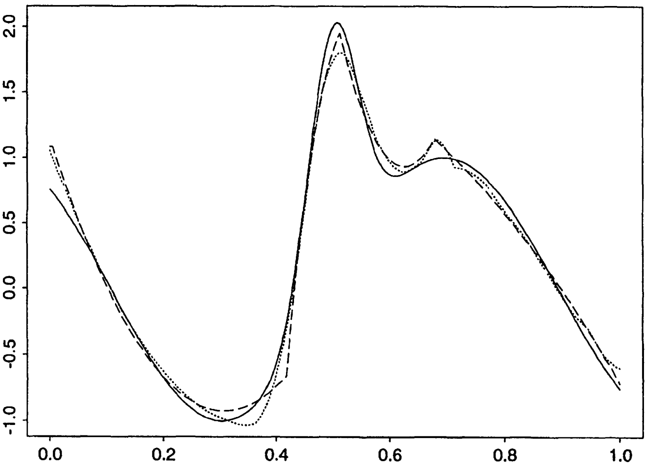


Fig. 2. Example (b), using the continuous piecewise quadratic model: ———, true curve, ··········, Bayes estimate; – – –, adaptive knot selection estimate

and show that our estimates are marginally better. We took the values x_i on a uniform grid with n points, i.e. $x_i = (i-1)/(n-1)$ $(i = 1, \ldots, n)$.

It can be seen from Figs 1 and 2 that the estimates may be displaying unnecessary variance because they are following the data too closely. We believe that this variance is caused more by the high signal-to-noise ratio than because of the extra degree of freedom that we allow at each knot compared with traditional splines for which $l = l_0$. We illustrate this point by running the Bayesian curve fitting and adaptive knot selection algorithms again but this time using linear splines (i.e. $l = l_0 = 1$). The results are shown in Figs 3 and 4. For the trivial example (a) we find a slightly lower MSE for the estimate when using linear instead of quadratic pieces but even for the only marginally more difficult example (b) the situation is

Table 1. MSE of estimates shown in Figs 1–4

Example	*Bayes estimate*		*Friedman estimate*	
	Linear	*Quadratic*	*Linear*	*Quadratic*
(a)	0.0079	0.0097	0.0308	0.0129
(b)	0.0096	0.0087	0.0181	0.0110

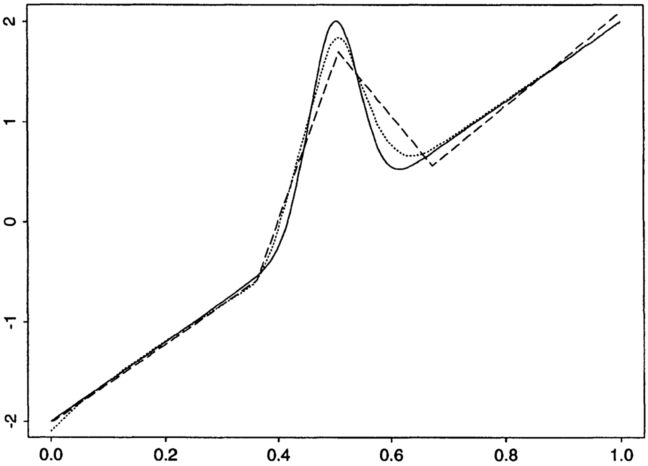


Fig. 3. Example (a), using the continuous piecewise linear model: ———, true curve; ··········, Bayes estimate; – – –, adaptive knot selection estimate

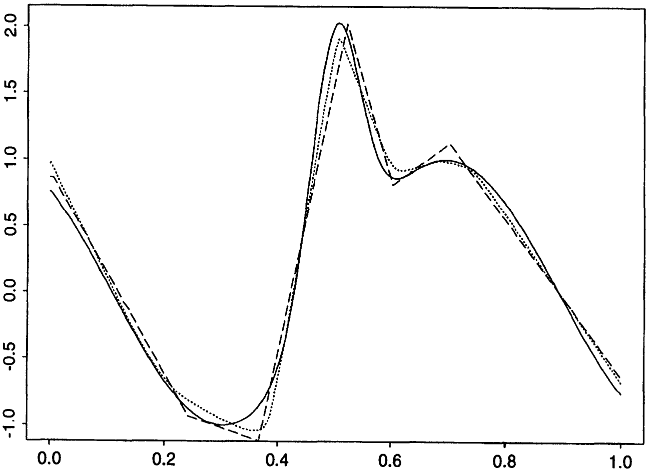


Fig. 4. Example (b), using the continuous piecewise linear model: ———, true curve; ··········, Bayes estimate; – – –, adaptive knot selection estimate

reversed. This suggests that using linear splines is worthwhile only when it is believed that the underlying true curve is very simple and for other cases the piecewise continuous quadratic model works well. In fact, we would advocate using this model all the time so the procedure is 'automatic'. However, as the improvement of the fits in Fig. 3 compared with Fig. 1 seems to be mainly at the edges of the range of the data we could choose to use a different spline basis to offset this such as the *natural* splines (see, for example, de Boor (1978)) which are linear beyond the last interior knot.

3.2. Unsmooth functions

We now illustrate the performance of the methodology with the four simulated examples ('Heavisine', 'Blocks', 'Bumps' and 'Doppler') used as test curves in Donoho and Johnstone (1994) to show the efficacy of our methodology to estimate wiggly curves.

A fixed uniform design $x_i = i/n$ is used. The number of grid points, n, is taken to be 2048 and we set the noise standard deviation to be $\sigma = 1.0$ so that the signal-to-noise ratio is 7. Again we choose $l_0 = 1$ and $l = 2$ and assign a Poisson prior over k but this time we choose $\lambda = 5$.

The choice of a Poisson prior is somewhat arbitrary. In related studies, we have used a negative binomial prior, but the choice of the distribution does not seem important. The choice of the prior mean is a compromise between flexibility and parsimony. A very small value of λ reflects a very strong insistence on smoothness. A very large value (relative to n) causes the fit to follow the data too closely. In this context, the results are relatively insensitive to choices in the range 5–20 (see Table 4 of Section 3.3 and associated discussion).

In Figs 5–8 we display the results with unsmooth functions. The estimates were obtained by ergodic averaging over 50000 iterations after a suitable 'burn-in' period. The length of the burn-in period was chosen to be sufficiently long that the mean-squared errors of the models had settled down. As shown in Fig. 9 this depends on the problem at hand with the MSE taking longer to settle down when the data set required more knots to be well estimated. Our method is seen to work well for all the examples with exactly the same initial choice of parameters.

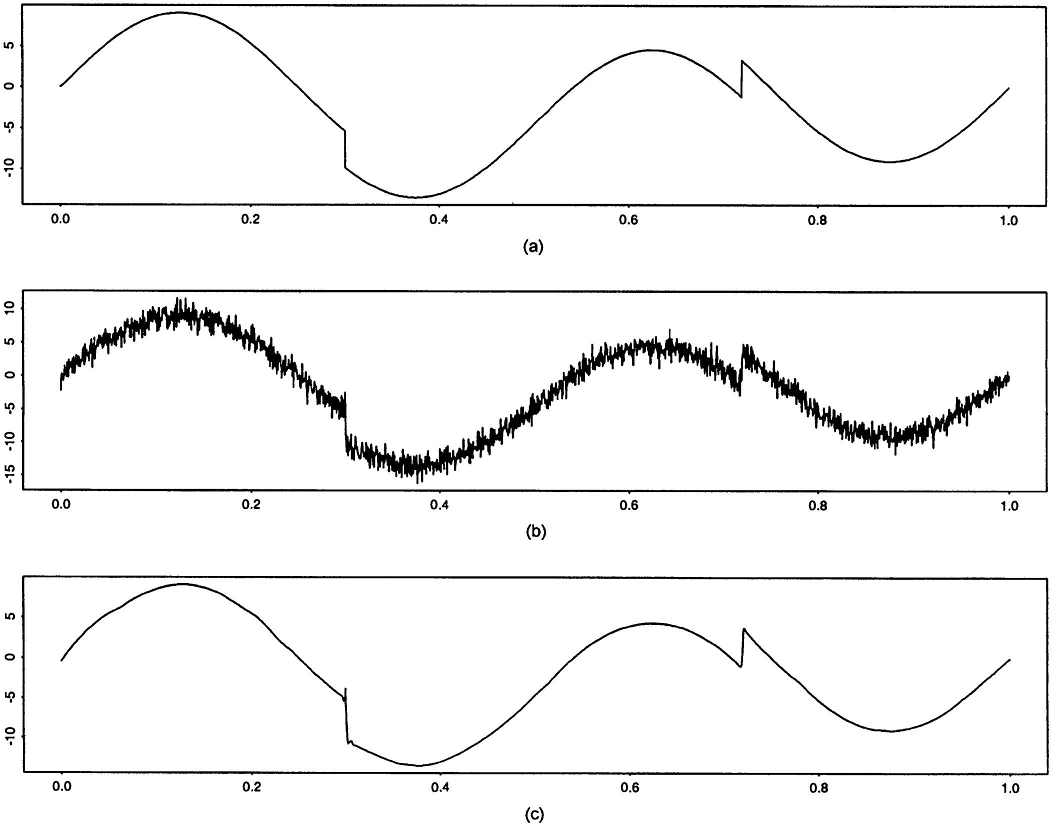
Despite the fixed choice of λ the posterior mode of the number of knots varies widely depending on the data. This again demonstrates the adaptive nature of the model which remains largely independent of the choice of λ. This is shown in Table 2.

Discontinuities in the true curve are found well by our method. This is particularly evident in Fig. 7 in which, because our method places many knots around the spikes, the data are followed closely in these regions. This leads to the estimated heights of the spikes being very close to the true values, something which is invariably lost when we use smoothing techniques on the same problem.

Table 2. Modal posterior for the number of knots†

Function	*Modal posterior number of knots*
Heavisine	17
Blocks	35
Bumps	62
Doppler	37

†$\lambda = 5$.



(a)

(b)

(c)

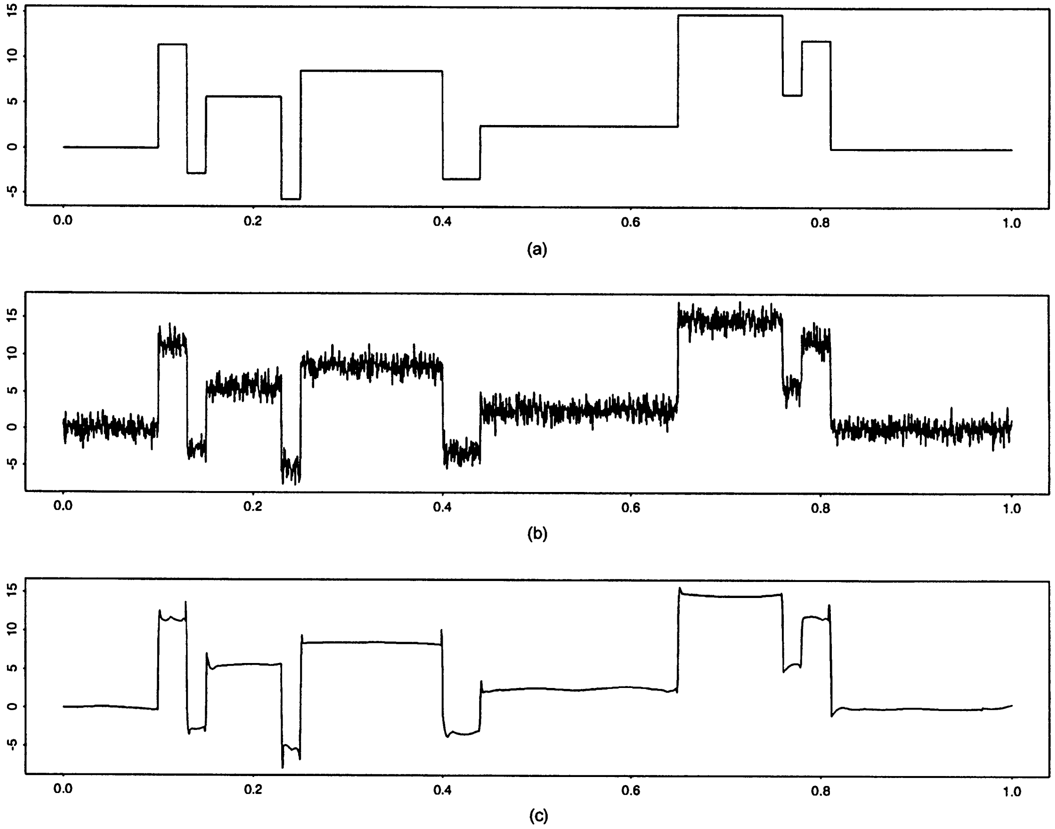
Fig. 5. Heavisine test curve: (a) true function; (b) true function with noise added; (c) estimate of the function

Table 3. Average MSE from 10 replications†

Function	*Wavelet threshold* λ_n^*	*Wavelet threshold* $\{2\log(n)\}^{1/2}$	*Bayes estimate*
Heavisine	0.060	0.083	0.033
Blocks	0.427	0.905	0.170
Bumps	0.499	1.080	0.167
Doppler	0.151	0.318	0.135

†$n = 2048$.

Table 3 compares our results with those obtained by Donoho and Johnstone (1994). Here λ_n^* is the optimal wavelet threshold chosen specifically for each data set, whereas $\{2\log(n)\}^{1/2}$ is a universal threshold which Donoho and Johnstone proposed for all such problems. The Bayes method fares well in comparison with their wavelet thresholding techniques and is markedly better than the wavelet threshold results using $\{2\log(n)\}^{1/2}$ which is also, in some sense, 'automatic'. Note that the wavelet results are obtained with σ^2 known and, for ease of computation, require the number of data points to be a power of 2. We only take $n = 2048$ so that we can compare our results with those of Donoho and Johnstone (1994).



(a)

(b)

(c)

Fig. 6. Blocks test curve: (a) true function; (b) true function with noise added; (c) estimate of the function

3.3. Prior specification

In the examples shown we have mainly used quadratic piecewise polynomials ($l = 2$) and forced continuity at the knot points ($l_0 = 1$). Choosing l too high in our function $f_{k,l}$, when we only make continuity assumptions, will lead to the data being too closely followed and the regression being unnecessarily complicated whereas choosing it too low will give us results which have no bearing in reality when the true curve is complicated. When the underlying curve is simple, as in example (a), it could be beneficial to use linear splines as shown in Fig. 3 but as the main focus of this paper is the adaptive way that complex curves can be found we do not explore such trivial examples further. We only force continuity at the knot points because this allows us the flexibility to model unsmooth curves well (Figs 5–8) while still maintaining a model which estimates smooth curves adequately (Figs 1 and 2). We have found that this model gives better results than other more widely used models. Splines, where $l = l_0$, tend not to have the flexibility required to model all the difficult examples that we attempted, which is why we allow 2 degrees of freedom at each knot point.

The only other parameter which we must choose before we carry out the algorithm, the Poisson parameter λ for the prior number of knots, has remarkably little effect on the results which we obtain. We chose a value of 1 for the smooth examples because the data set was small ($n = 200$) whereas we chose $\lambda = 5$ for the Donoho and Johnstone examples as we had a large data set ($n = 2048$) and wanted our simulation of the target posterior distribution not to

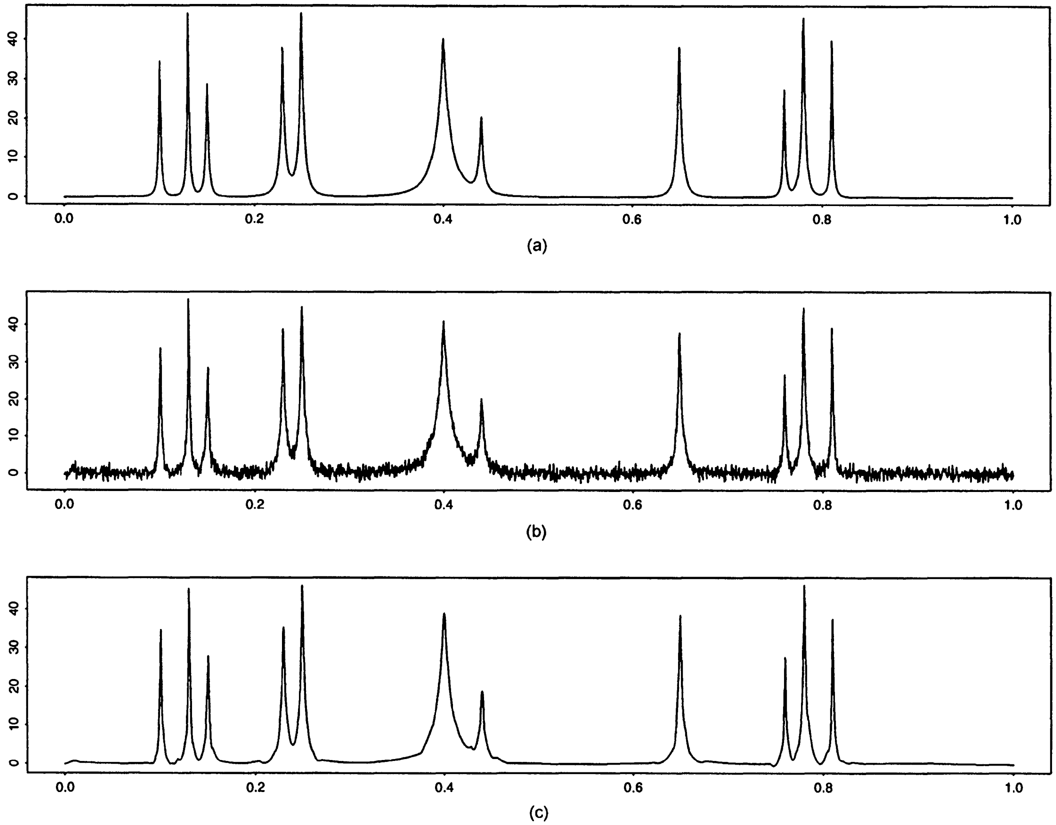


Fig. 7. Bumps test curve: (a) true function; (b) true function with noise added; (c) estimate of the function

Table 4. Average MSE from 10 replications for the Blocks example with varying λ

λ	1	3	5	7	10	15	20
MSE	0.368	0.181	0.170	0.173	0.173	0.174	0.195

be too constricted so that the number of knots in the models could vary widely. In fact the values chosen are almost certainly not the optimum values for the given examples but to justify our title including the word automatic we have restrained from varying λ to improve our results. It seems that the greatest danger would be to set λ too small when we have a large data set as demonstrated in Table 4 for the Blocks function. So, for reasonable choices of λ, the method appears to be robust.

4. Extension to additive models

4.1. Introduction

Regression techniques quickly become unmanageable when we try to extend them to more than two dimensions owing to the interaction terms which have to be accounted for. Other

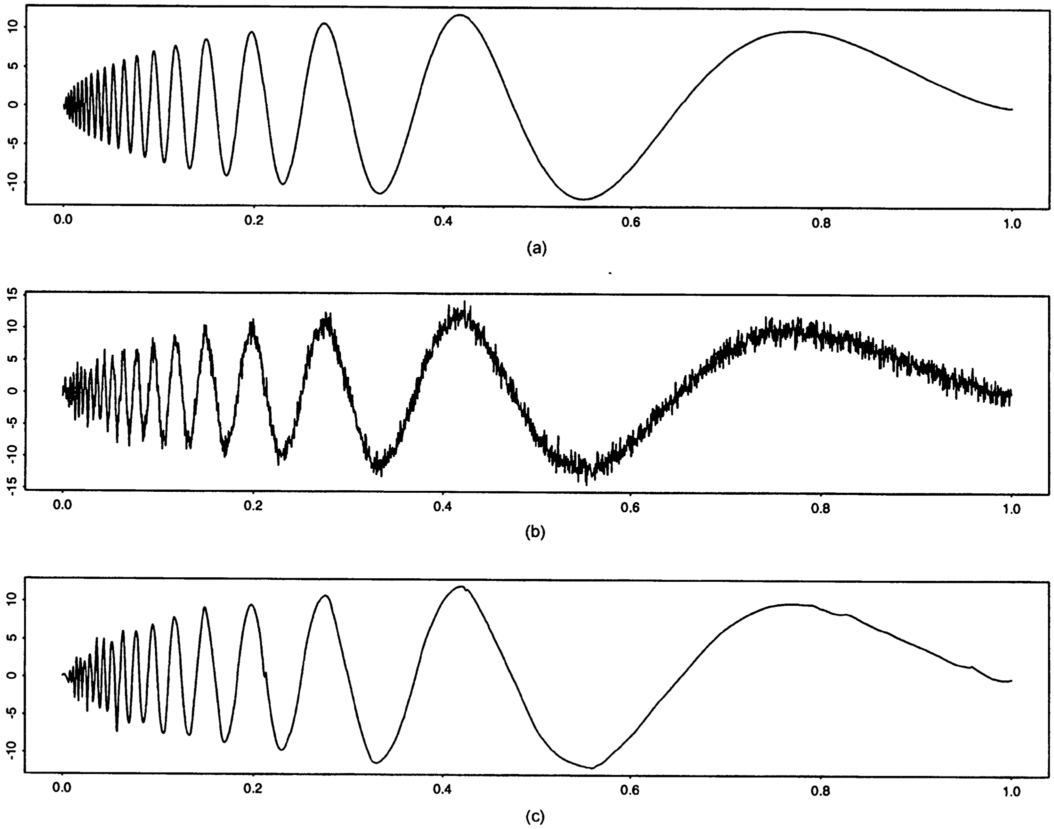


Fig. 8. Doppler test curve: (a) true function; (b) true function with noise added; (c) estimate of the function

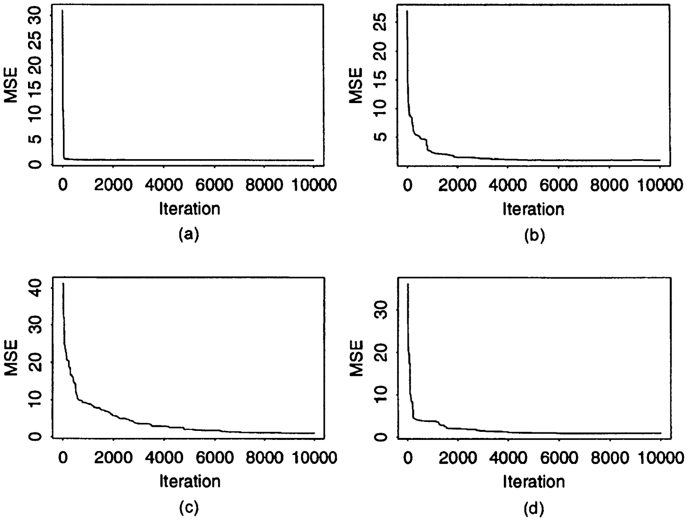


Fig. 9. MSE of the output from the MCMC algorithm: (a) Heavisine test curve; (b) Blocks test curve; (c) Bumps test curve; (d) Doppler test curve (the first 5000 iterations, at least, are discarded as burn-in iterations)

methods for estimating surfaces in two or more dimensions have been used to overcome these problems such as projection pursuit regression (Friedman and Stuetzle, 1981), multivariate adaptive regression splines (Friedman, 1991) and additive models (Hastie and Tibshirani, 1990) among others. Here we concentrate on the additive models approach and provide an extension of our Bayesian curve fitting method. A non-Bayesian analogue can be found in Friedman and Silverman (1989).

The general additive model problem is to find functions f_j such that

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \epsilon \tag{12}$$

where the X_j are the predictor variables, $E(\epsilon) = 0$, $\text{var}(\epsilon) = \sigma^2$ and ϵ is independent of the X_js. The f_j are arbitrary univariate functions and can be found nonparametrically by using scatterplot smoothers and the back-fitting algorithm as described in Hastie and Tibshirani (1990). However, assuming a specific functional form for the f_j, which makes the model parametric, can help us to work out the degrees of freedom and pointwise standard errors of the model, and this is the approach that we shall be using.

The common approach to this problem is either to decide on the number of knots and their locations beforehand or to find suitable knot locations by some preliminary examination of the data. Then a least squares approximation is carried out with the spline basis functions fixed by the knot locations to find the parameters that we require. Instead of assuming that the f_j are splines of some sort with determined knot locations (Hastie and Tibshirani, 1990) we set up the f_j as continuous piecewise quadratic polynomials over random knot locations as used in the previous section. The difference with our method is that we do not prospectively look for knots but apply the well-known 'back-fitting algorithm', as used in nonparametric additive models, to find the knot locations and hence the estimates to the terms in equation (12).

4.2. Methodology

We proceed in the same way as for the general nonparametric additive model problem by using a slightly modified back-fitting algorithm (Hastie and Tibshirani (1990), p. 91). Again we monitor the output via the mean-squared errors of the models.

4.2.1. Algorithm

Step 1: initialize—$\alpha = \text{ave}(y_i)$, $f_j = f_j^0$, $j = 1, \ldots, p$.
Step 2: cycle—$j = 1, \ldots, p, 1, \ldots, p, \ldots$;

$$f_j = (\hat{f}_j | f_1, \ldots, f_{j-1}, f_{j+1}, \ldots, f_p).$$

Step 3: continue step 2 until there is little change in the mean-squared errors of the models.

So instead of using the criterion of cross-validation for selecting our knots we are looking at the posterior probability of the location of the knots given the data.

4.3. Example

We take a slightly modified example from Hastie and Tibshirani (1990), pages 247–251. We try to fit functions f_1 and f_2 for the model

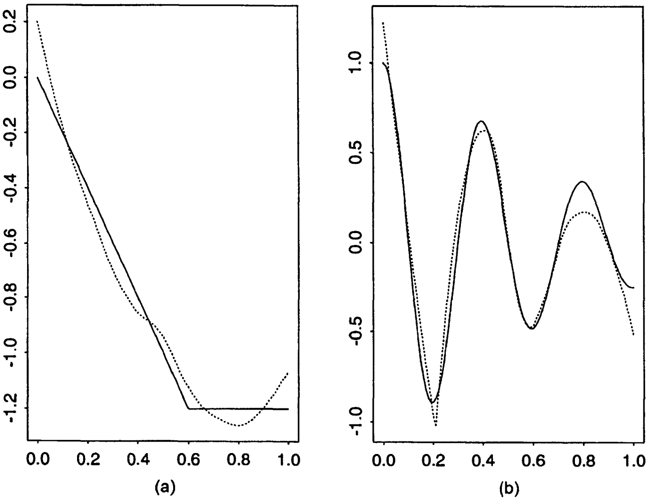


Fig. 10. Additive model example (———, true function; ·········, estimate): (a) $f_1(x)$; (b) $f_2(z)$

$$y_i = f_1(x_i) + f_2(z_i) + \epsilon_i, \qquad i = 1, \ldots, 100,$$

$$f_1(x) = \begin{cases} -2x & \text{for } x < 0.6, \\ -1.2 & \text{otherwise,} \end{cases}$$

$$f_2(z) = \frac{\cos(5\pi z)}{1 + 3z^2},$$

with x_i and z_i generated independently from a $U(0, 1)$ distribution and ϵ_i from an $N(0, 0.25)$ distribution. This gives us a signal-to-noise ratio of approximately 2. Fig. 10 shows the estimates to the functions $f_1(x)$ and $f_2(z)$. The estimate for $f_1(x)$ is not quite as good as that given in Hastie and Tibshirani (1990) but that for $f_2(z)$ is considerably better than the estimate that they gave. We also found strong evidence for knots at $z = 0.21$ and $z = 0.60$ as they were both local maxima in the posterior density. No good knot locations were found for x. However, using these knots we should also obtain a good improvement in the spline fitting approach which is recommended in Hastie and Tibshirani (1990).

5. Discussion

We have presented a method which works well for a wide range of challenging functions. It has been shown to be competitive with the adaptive knot selection algorithm given in Friedman and Silverman (1989) and with wavelet thresholding techniques (Donoho and Johnstone, 1994; Donoho *et al.*, 1995). This method has been shown to be particularly good at approximating rapidly varying curves, which is its main strength. There are many methods to estimate curves which are continuous and smooth and even though this curve fitting approach works well in these cases this was not our primary concern.

We have used a single model throughout; however, we may use prior knowledge in fitting the curve. If we know that a curve is continuous with continuous first and second derivatives we may model it with a cubic spline by taking $l = l_0 = 3$. Also, after a glance at the initial estimate of the Blocks function using $l = 2$ and $l_0 = 1$, we could fit a step function ($l = l_0 = 0$) to the data; this results in Fig. 11. This is an almost perfect reconstruction of the true curve.

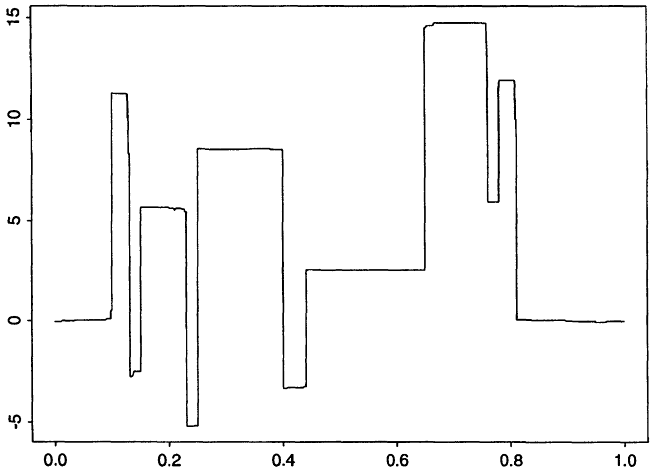


Fig. 11. Estimate of the Blocks function using a step function as the piecewise polynomial

In most curve fitting problems the main aim is prediction, which is why we have focused on posterior mean estimates throughout this paper. However, as a referee pointed out, it is important to look at individual sample curves generated by the chain to look for features which might be hidden by the posterior mean. The posterior mean estimate can smooth out features that are present in individual sample curves and if this appears to be the case the posterior mode may be a more honest estimate of the true curve.

The results shown took between 10 and 30 min to run on a Sun SPARC 5 workstation. The software (written in C) used to produce these results is available from the World Wide Web address `http://www.ma.ic.ac.uk/~dgtd` or by sending an electronic mail message to `d.denison@ic.ac.uk`.

Acknowledgements

We thank James Bennett and the referees for their helpful comments on earlier versions of the paper and Harvard Rue and Chris Holmes for their help in some of the programming. The first author was supported by an Engineering and Physical Sciences Research Council research studentship.

Appendix A

The move types birth, death and move given in the algorithm in Section 2.4 are undertaken as follows. The notation follows that in Section 2.4.

A.1. Birth

Step 1: generate the proposed new knot to add uniformly from one of the $n - Z(k)$ candidate grid points.
Step 2: sort the knots into ascending numerical order. This becomes the model proposed.
Step 3: work out the coefficients β in the proposed model by using least squares.
Step 4: generate u uniformly on [0, 1].
Step 5: work out the acceptance probability α.

Step 6: if $u < \alpha$ accept the model proposed; otherwise keep the current model.
Step 7: return to the main algorithm.

A.2. Death

Step 1: generate the proposed changepoint to delete uniformly from the interior knots present. This is the model proposed.
Step 2: work out the coefficients β in the proposed model by using least squares.
Step 3: generate u uniformly on [0, 1].
Step 4: work out the acceptance probability α.
Step 5: if $u < \alpha$ accept the model proposed; otherwise keep the current model.
Step 6: return to the main algorithm.

A.3. Move

Step 1: generate the proposed knot to move, say x_c, uniformly from the interior knots that are present in the model.
Step 2: generate the proposed new position of the knot x_c uniformly from the set of possible points to move to ($\mathcal{C}$).
Step 3: work out the coefficients β in the proposed model by using least squares.
Step 4: generate u uniformly on [0, 1].
Step 5: work out the acceptance probability α.
Step 6: if $u < \alpha$ accept the model proposed; otherwise keep the current model.
Step 7: return to the main algorithm.

References

de Boor, C. (1978) *A Practical Guide to Splines*. New York: Springer.

Donoho, D. L. and Johnstone, I. M. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, K. and Picard, D. (1995) Wavelet shrinkage: asymptopia (with discussion)? *J. R. Statist. Soc.* B, **57**, 301–369.

Fan, J. and Gijbels, I. (1995) Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. R. Statist. Soc.* B, **57**, 371–394.

Friedman, J. H. (1991) Multivariate adaptive regression splines. *Ann. Statist.*, **19**, 1–141.

Friedman, J. H. and Silverman, B. W. (1989) Flexible parsimonious smoothing and additive modelling (with discussion). *Technometrics*, **31**, 3–39.

Friedman, J. H. and Stuetzle, W. (1981) Projection pursuit regression. *J. Am. Statist. Ass.*, **76**, 817–823.

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

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

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

Müller, H.-G. and Stadtmüller, U. (1987) Variable bandwidth kernel estimators of regression curves. *Ann. Statist.*, **15**, 182–201.

Smith, P. L. (1982) Curve fitting and modeling with splines using statistical variable selection techniques. *Report NASA 166034*. Langley Research Center, Hampton.

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