Biometrika (2001), **88,** 4, *pp.* 1055–1071

Printed in Great Britain

Bayesian curve-fitting with free-knot splines

BY ILARIA DIMATTEO, CHRISTOPHER R. GENOVESE AND ROBERT E. KASS

Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, U.S.A.

dimatteo@stat.cmu.edu genovese@stat.cmu.edu kass@stat.cmu.edu

SUMMARY

We describe a Bayesian method, for fitting curves to data drawn from an exponential family, that uses splines for which the number and locations of knots are free parameters. The method uses reversible-jump Markov chain Monte Carlo to change the knot configurations and a locality heuristic to speed up mixing. For nonnormal models, we approximate the integrated likelihood ratios needed to compute acceptance probabilities by using the Bayesian information criterion, BIC, under priors that make this approximation accurate. Our technique is based on a marginalised chain on the knot number and locations, but we provide methods for inference about the regression coefficients, and functions of them, in both normal and nonnormal models. Simulation results suggest that the method performs well, and we illustrate the method in two neuroscience applications.

Some key words: BIC; Generalised linear model; Nonparametric regression; Reversible-jump Markov chain Monte Carlo; Smoothing; Unit-information prior.

1. INTRODUCTION

Smoothing splines are often appealing tools for curve estimation because they provide computationally efficient estimation. They tend to do a good job in smoothing noisy data, and they have both frequentist and Bayesian interpretations (Hastie & Tibshirani, 1990; Wahba, 1990). However, in practice, smoothing splines have two shortcomings: they require specification of a global smoothness parameter; and, conditionally on the choice of smoothness, they are linear estimators and thus have difficulty adapting to functions that are heterogeneous over their domains. The first problem has been addressed through various data-driven methods, such as crossvalidation, for choosing the smoothness parameter, but such methods are not convincing in small samples and they offer no measure of uncertainty in the estimated smoothness. The second problem is more fundamental. Whereas smoothing splines use many knots located at the data, an alternative that has been explored is to use fewer knots that are well placed (Denison et al., 1998; Lindstrom, 1999; Zhou & Shen, 2001; Biller, 2000; Hansen & Kooperberg, 2000; Halpern, 1973; Genovese, 2000; Eilers & Marx, 1996; Smith & Kohn, 1996). This approach is often called curve-fitting with free-knot splines because the number of knots and their locations are determined from the data.

In this paper, we describe a fully Bayesian method for curve-fitting with free-knot splines for data drawn from an exponential family distribution, which we call Bayesian adaptive regression splines. Our implementation is based on reversible-jump Markov chain Monte

Carlo (Green, 1995) and incorporates a key observation made by Zhou & Shen (2001). We compare our method's performance to both the Bayesian method of Denison et al. (1998) and the frequentist, iterative spatially adaptive regression spline method of Zhou & Shen (2001). Our method gives more accurate estimates of our test function than either of the others.

Our method applies to independent data $(X_1, Y_1), \ldots, (X_n, Y_n)$ that satisfy the following model:

$$Y_i | X_1, \ldots, X_n \sim p\{y | f(X_i), \sigma\} \quad (i = 1, \ldots, n), \tag{1}$$

where f is a real-valued function on $[a, b]$, and σ is an optional and potentially vector-valued nuisance parameter. We think of the X_i's here as observed explanatory variables. The goal is to estimate the unknown function f from these data under the assumption that f lies in some fixed, and usually infinite-dimensional, class of functions.

We focus on the special case when $p(y|\theta, \sigma)$ is an exponential family distribution with dispersion parameter σ. In particular, when $p(.)$ is a $N(\theta, \sigma^2)$ distribution, we obtain the nonparametric regression model

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

where the ε_i are independent draws from $N(0, \sigma^2)$ and $\sigma > 0$ is unknown.

Our method implicitly assumes that f is well approximated between a and b by a cubic spline with some number of knots. In practice, we will assume that f is such a spline. This class of cubic splines is quite large and approximates any locally smooth function arbitrarily well.

We will denote knot configurations by pairs (k, ξ), where the number of knots k is a nonnegative integer and the knot locations are given by the k-vector $\xi = (\xi_1, \ldots, \xi_k)$, for $a < x_{(1)} < \xi_1 \leqslant \ldots \leqslant \xi_k < x_{(n)} < b$. Let $b_j(x)$, for $j = 1, \ldots, k+2$, denote the jth function in a cubic B-spline basis with natural boundary constraints, i.e. linear outside $[a, b]$, and let $B_{k,\xi}$ be the matrix whose i, j component is $b_j(x_i)$. The subscript k,ξ expresses the dependence of the matrix $B_{k,\xi}$ on the number and locations of knots. Under our assumptions, we can write f as a linear combination

$$f(x) = \sum_{j=1}^{k+2} \beta_j b_j(x) \tag{3}$$

for some vector $\beta = (\beta_1, \ldots, \beta_{k+2})$. We have the linear relation $B_{k,\xi}\beta = f(X) \equiv (f(X_1), \ldots, f(X_n))$ at the observed design points.

To complete the Bayesian formulation of the model, we must specify a prior on the unknown quantities β, σ, k and ξ. In this paper, we use uniform or Poisson priors on k and a uniform prior on ξ induced by the uniform prior over the standard k-simplex by rescaling ξ to $[a, b]$. Given k and ξ, we use a particular conjugate Normal prior on β that Kass & Wasserman (1995) called the unit-information prior and, independently, the improper prior $\pi_\sigma(\sigma) = 1/\sigma$. With these choices, the posterior under the Normal model (2) can be computed analytically. For example, β and σ can be integrated out of the posterior in order to obtain a Markov chain for sampling from the marginal posterior on (k, ξ):

$$p(y|k, \xi) = \int p(y|\beta, k, \xi, \sigma)\pi(\beta, \sigma|k, \xi)\, d\beta\, d\sigma. \tag{4}$$

In the general model (1), we rely on an approximation for ratios of marginal likelihoods (4) in terms of the Bayesian information criterion, BIC. Kass & Wasserman (1995) and

Pauler (1998) showed these approximations to be accurate when the unit-information prior on β is used.

Since the parameter space in the model (1) is a disjoint union of spline spaces, sampling from the posterior benefits from the reversible jump Markov chain Monte Carlo technique introduced by Green (1995) and shown by him to be effective for estimating step functions with variable number and locations of the break points. Denison et al. (1998) generalised this approach to higher-order free-knot splines, producing a potentially powerful nonlinear regression method. However, Denison et al. (1998) avoided specifying a prior on β, preferring instead to plug in its least-squares estimator at each stage. This quasi-Bayesian solution affects how the method penalises dimensionality and often leads to severe overfitting.

We use a reversible-jump Metropolis–Hastings Markov chain Monte Carlo simulation on the (k, ξ) pairs, with β and σ marginalised out. Since we use a fully Bayesian formulation, inferences on β and σ can be included with additional post hoc simulations as desired. We can use the results to estimate f with a mean of the posterior sample from $f(x)$ which is a function of β. The mode can also be useful in some cases; while the mean is analytically and computationally tractable, the mode avoids averaging over disparate structures when there are many qualitatively different functions in regions of high posterior density. By using a spline basis, introducing the unit-information prior and approximating with the BIC, we are able to employ essentially the same Markov chain Monte Carlo implementation with the general model (1) as with the Normal model (2).

In § 2, we provide further details about our choice of priors and our approximation to the likelihood ratios. In § 3, we discuss further details of our posterior simulation. In § 4, we show the results of simulations for three elementary test functions. In §§ 5 and 6, we apply the method to two real datasets. The former uses the Normal model (2) to analyse functional magnetic resonance imaging data; the latter uses a Poisson model based on (1) to estimate the time-intensity function of neuronal firing in a monkey's brain. Finally, in § 7, we discuss several possible refinements and extensions of our method.

2. Choice of priors and the likelihood ratio approximation

We begin by treating model (2). It is convenient, though not essential as we show below, to use a prior for which (4) may be obtained analytically. We decompose the prior as follows:

$$\pi(\beta, k, \xi, \sigma) = \pi_\beta(\beta \,|\, \xi, k, \sigma)\pi_\xi(\xi \,|\, k)\pi_k(k)\pi_\sigma(\sigma),$$

where $\pi_\sigma(\sigma) = 1/\sigma$ and

$$\beta \,|\, k, \xi, \sigma \sim N_{k+2}\{0, \sigma^2 n(B_{k,\xi}^{\mathrm{T}} B_{k,\xi})^{-1}\}. \tag{5}$$

The remaining priors on ξ and k could be chosen to express knowledge about these parameters or, equivalently, to force some desired behaviour in the posterior. In our simulations and applications below, we have adopted a prior on ξ given k induced by a $\mathrm{Dir}(1, 1, \ldots, 1)$ prior on the k-simplex by scaling $[a, b]$ to $[0, 1]$. For k we also adopted a Poisson prior or Uniform prior on $\{1, \ldots, K_0\}$. In many applications, the results are unlikely to be very sensitive to the precise specification of the prior on k.

For linear regression models $Y = X\beta + \varepsilon$, with the more general design matrix X replacing $B_{k,\xi}$, priors of the form (5) have been used by many authors (Pauler, 1998). Kass & Wasserman (1995) have called these 'unit-information' priors because the amount of infor-

mation in the prior, represented by the covariance matrix, is equal to the amount of information in one observation, as represented by the Fisher information matrix. A prior very similar to (5) was used by Smith & Kohn (1996) in a different but related context of spline knot selection, where instead of n in (5) they used a constant between 10 and 1000 which they judged to work well for the data they examined.

With these choices of priors on β and σ, we can compute analytically the marginal posterior for (ξ, k) via equation (4). This makes it easy to compute the likelihood ratios $p(y|\xi^c, k^c)/p(y|\xi, k)$, that are used in the reversible jump algorithm to determine whether or not to move from state (k, ξ) to candidate state (k^c, ξ^c). For example, one type of move in our Markov chain Monte Carlo implementation involves the addition of a knot. If the current state is (k, ξ) and the candidate state is $(k^c = k + 1, \xi^c)$, then the likelihood ratio becomes

$$\frac{p(y|k^c, \xi^c)}{p(y|k, \xi)} = \frac{1}{\sqrt{(n+1)}} \left(\frac{y^{\mathrm{T}}\{I_n - n(n+1)^{-1} B_{k,\xi}(B_{k,\xi}^{\mathrm{T}} B_{k,\xi})^{-1} B_{k,\xi}^{\mathrm{T}}\} y}{y^{\mathrm{T}}\{I_n - n(n+1)^{-1} B_{k,\xi^c}(B_{k,\xi^c}^{\mathrm{T}} B_{k,\xi^c})^{-1} B_{k,\xi^c}^{\mathrm{T}}\} y} \right)^{n/2}. \tag{6}$$

Similarly, we can obtain analytically the conditional posterior expectation

$$E\{f(x)|k, \xi, y\} = \frac{n}{n+1} B_{k,\xi}(B_{k,\xi}^{\mathrm{T}} B_{k,\xi})^{-1} B_{k,\xi}^{\mathrm{T}} y \doteq B_{k,\xi}\hat{\beta}, \tag{7}$$

for any x. The posterior expectation $E\{f(x)|y\}$ can then be computed by averaging this conditional expectation over (k, ξ) samples. This expectation is the Bayes estimator $\hat{f}(x)$ for $f(x)$ under squared-error loss.

When we are making inferences about functionals of f, the uncertainty in β cannot be ignored. With our choice of priors in the normal model, $p(\beta|y, \xi, k)$ can be computed analytically, making it easy to assess the uncertainty in β after a simulation on ξ and k alone. To do this, we draw a value from this posterior for each (k, ξ) sample from our chain.

In the more general model (1), we use the same priors. However, it is often infeasible in this case to obtain analytical expressions such as those above. With the unit information prior (5) on β, the likelihood ratio $p(y|\xi^c, k^c)/p(y|\xi, k)$ in the Markov chain Monte Carlo can be approximated using the BIC with an error of $O(n^{-\frac{1}{2}})$, and this produces a posterior distribution on (k, ξ) that also has an error of $O(n^{-\frac{1}{2}})$; see Appendix 3. Examples in Kass & Wasserman (1995) show that BIC often produces a very good approximation to the unit-information posterior in practice. Implementation requires maximum likelihood estimators $\hat{\beta}$ under each spline model, which are often easily computed with standard software. In particular, conditionally on ξ and k and when the data are drawn from an exponential family distribution, our model in equation (1) becomes a generalised linear model (McCullagh & Nelder, 1989).

The use of $\hat{\beta}$, that is integrating out the coefficients in the chain, is a key feature of both our method and the method of Denison et al. (1998). This approach has two advantages. First, it speeds up the simulation by reducing the dimensionality of the parameter space with minimal additional cost to compute $\hat{\beta}$. Secondly, it facilitates the jumps between spline models because such moves no longer require a delicate re-balancing of the coefficients when knots are added or deleted (Genovese, 2000). However, it is essential that the uncertainty in β be accounted for in the final inferences. In the normal model, the simulation on (k, ξ) and (β, σ) can be decoupled because we have analytical expressions for the marginalised expectations. Thus, we can draw samples of (β, σ) at each step as

needed after the original simulation is completed. In the general model, additional work is needed. We describe one approach in § 3·2 below.

While our method and the spatially adaptive regression spline method of Zhou & Shen (2001) share many features, the primary contrast between the two is that ours is a fully Bayesian simulation method while spatially adaptive regression spline is a frequentist, iterative method. The primary contrast between our method and that of Denison et al. (1998) is that the two Markov chains have different equilibrium distributions. Denison et al. (1998) do not use a prior on β but instead replace the likelihood ratio $p(y|\xi^c, k^c)/p(y|\xi, k)$ with $p(y|\hat{\beta}^c, \xi^c, k^c)/p(y|\hat{\beta}, \xi, k)$. This plug-in approximation with the least-squares estimator produces a version of the marginal density $p^{\mathrm{DMS}}(y|\xi, k)$ that is monotonically increasing in k; we are more specific in Appendix 3. As a consequence, unlike that based on BIC, the equilibrium distribution produced by the chain with the plug-in approximation does not have the desired properties: the data cannot be informative about the number of knots, the procedure will tend to overfit, and the effect of the prior on k will not vanish as the dataset gets large, since the likelihood will become roughly constant in k as k increases. Indeed, experimentation using the software kindly provided by Denison et al. (1998) displays extreme sensitivity of the posterior on k to the choice of prior on k. Incidentally, we can also see this in the likelihood ratio approximation for the normal model in equation (6) by

$$\frac{p(y|k^c, \xi^c)}{p(y|k, \xi)} \eqsim \frac{1}{\sqrt{n}} \left(\frac{(y - B_{k,\xi}\hat{\beta})^{\mathrm{T}}(y - B_{k,\xi}\hat{\beta})}{(y - B_{k,\xi^c}\hat{\beta}^c)^{\mathrm{T}}(y - B_{k,\xi^c}\hat{\beta}^c)} \right)^{n/2} = \exp(-\mathrm{BIC}/2), \tag{8}$$

where $\hat{\beta}$ are the least-squares estimates. The method of Denison et al. (1998) omits the consequential factor $1/\sqrt{n}$ in (8), which BIC includes to penalise the likelihood ratio for dimensionality.

3. Posterior simulation

3·1. *Reversible-jump Markov chain Monte Carlo*

As we indicated in § 1, the framework we have adopted produces a Markov chain with the marginal posterior on (k, ξ) as stationary distribution. The Metropolis–Hastings acceptance probability combines the likelihood ratio discussed earlier, a prior ratio $\pi_{k,\xi}(k^c, \xi^c)/\pi_{k,\xi}(k, \xi)$, where $\pi_{k,\xi}(k, \xi) = \pi_{\xi}(\xi|k)\pi_k(k)$, and an asymmetry correction $q(k, \xi|k^c, \xi^c)/q(k^c, \xi^c|k, \xi)$, where q is the proposal density (Tierney, 1994). We use the general scheme used by Green (1995) and Denison et al. (1998), which involves moves that add, delete and relocate knots. In contrast to Denison et al. (1998), where new knots are generated 'far' from existing knots, our chain uses the locality heuristic devised in Zhou & Shen (2001), which is based on the idea that more knots are needed where the curve changes rapidly. The heuristic holds that a new knot is more likely to be needed in regions where knots have recently been added.

Let M_k represent a model parameterised by $(k, \xi_1, \ldots, \xi_k)$. The addition, deletion and relocation steps of the reversible-jump sampler are attempted, respectively, with the following probabilities:

$$b_k = c \min\{1, p(k+1)/p(k)\}, \quad d_k = c \min\{1, p(k-1)/p(k)\}, \quad \eta_k = 1 - b_k - d_k.$$

These probabilities ensure that $b_k p(k) = d_{k+1} p(k+1)$. Appendix 1 contains a proof of detailed balance for the following proposal scheme.

Birth step. Suppose that the current model M_k contains k knots located at $\xi_1, \ldots, \xi_k$. To

generate a new candidate knot we first choose one knot uniformly from the set of existing knots. Let ξ_{j^*} be such a knot. The candidate new knot, ξ_{cand}, is generated by a distribution centred at ξ_{j^*} with some spread parameter τ_B having density $h_B(\xi_{\text{cand}}|\xi, \tau_B)$. In this case the jump probability is given by

$$q(M_{k+1}|M_k) = b_k \frac{1}{k} \sum_i h_B(\xi_{\text{cand}}|\xi_i);$$

in the expression $q(M_{k+1}|M_k)$ there is a mixture of densities because the new knot ξ_{cand} can be generated by any of k different distributions.

Death step. The candidate knot for deletion is chosen uniformly from the set of existing knots. Let M_k be the current model. Then the jump probability of going from M_k to M_{k-1} is

$$q(M_{k-1}|M_k) = d_k \frac{1}{k}.$$

Relocation step. We first choose a candidate knot ξ_{j^*} uniformly from the set of existing knots $\{\xi_1, \ldots, \xi_k\}$. The candidate new location, ξ_{cand}, for the knot ξ_{j^*} is generated by a distribution centred at ξ_{j^*} with spread parameter τ_R and having density $h_R(\xi^c|\xi_{j^*}, \tau_R)$.

Let $\xi = (\xi_1, \ldots, \xi_{j^*-1}, \xi_{j^*}, \xi_{j^*+1}, \ldots, \xi_k)^{\mathrm{T}}$ be the current sequence of knots, and let $\xi^c = (\xi_1, \ldots, \xi_{j^*-1}, \xi_{\text{cand}}, \xi_{j^*+1}, \ldots, \xi_k)^{\mathrm{T}}$ be the candidate new sequence of knots, which differs from ξ only in the replacement of knot ξ_{j^*} by knot ξ_{cand}. Note that the candidate new location does not have to be the j^*th element. The jump probability is computed as follows:

$$q(M_{\text{cand}}|M_{\text{curr}}) = \eta_k \frac{1}{k} h_R(\xi_{\text{cand}}|\xi_{j^*}).$$

Candidate distributions. One convenient choice for the birth and relocation proposal distributions, with densities h_B and h_R, would be Beta distributions. The precise form, however, is not likely to make much difference. Once these are selected, it remains to choose values of parameters c, τ_B and τ_R. In principle, these may be regarded as tuning parameters, adjusted to produce a chain having good acceptance probabilities.

Here, we choose the constant c to be 0·4 as in Denison et al. (1998); a limited study of our own suggests that this is a good value. We take both the birth and relocation proposals to be Beta distributions with parameters $\xi_{j^*}v$ and $(1-\xi_{j^*})v$, and we choose $v = 50$ in our examples. We obtained essentially the same results using for the birth and relocation densities h_B and h_R a Normal distribution with mean ξ_{j^*} and variance τ^2, truncated to $[\xi_{j^*-2}, \xi_{j^*+2}]$.

3·2. *Importance reweighting*

The reversible-jump scheme described in § 3·1 produces a chain on (k, ξ). As we indicated in § 2, under model (2) we can obtain draws from the marginal posterior on β, to make inferences about characteristics of f, by also drawing a value of β from the conditional posterior of β given (k, ξ) for each sampled value of (k, ξ). Under model (1), however, it is usually infeasible to calculate this distribution directly, so additional simulation is required. To avoid a lengthy Markov chain Monte Carlo simulation at each knot configuration from the original chain, we use importance reweighting. This allows more

efficient sampling from an approximate distribution that can be specified directly. Our method is as follows.

Denote by $g(\beta, k, \xi)$ some feature of the curve, such as the location of its maximum, that we wish to estimate. Let $q(\beta \mid y, \xi, k) \propto p(y \mid \beta, k, \xi)\pi_\beta(\beta \mid k, \xi)$. The posterior expectation of $g(\beta, k, \xi)$ given y may be computed from

$$E\{g(\beta, \xi, k) \mid y\} = \frac{\int \dots \int g(\beta, \xi, k) \dfrac{q(\beta \mid y, k, \xi)}{\hat{q}(\beta \mid y, \xi, k)} \hat{q}(\beta \mid y, k, \xi) p(k, \xi \mid y)\, d\beta\, d\xi\, dk}{\int \dots \int \dfrac{q(\beta \mid y, k, \xi)}{\hat{q}(\beta \mid y, \xi, k)} \hat{q}(\beta \mid y, k, \xi) p(k, \xi \mid y)\, d\beta\, d\xi\, dk}$$

$$\simeq \frac{\sum_l g(\beta^{(l)}, \xi^{(l)}, k^{(l)}) w(\beta^{(l)}, \xi^{(l)}, k^{(l)})}{\sum_l w(\beta^{(l)}, \xi^{(l)}, k^{(l)})}, \tag{9}$$

where

$$w(\beta^{(l)}, \xi^{(l)}, k^{(l)}) = \frac{q(\beta^{(l)} \mid y, \xi^{(l)}, k^{(l)})}{\hat{q}(\beta^{(l)} \mid y, \xi^{(l)}, k^{(l)})},$$

$(\xi^{(l)}, k^{(l)})$ is the pair accepted by the reversible-jump sampler, i.e. sampled from $p(k, \xi \mid y)$, and $\beta^{(l)}$ is sampled from a suitable approximation $\hat{q}$ to the conditional posterior of β given (k, ξ). In fact, we may approximate the likelihood function on β given (k, ξ) rather than the full conditional posterior, which is typically easier under model (1), yielding weights of the form

$$w(\beta^{(l)}, \xi^{(l)}, k^{(l)}) = \frac{p(y \mid \beta^{(l)}, \xi^{(l)}, k^{(l)})}{\hat{p}(y \mid \beta^{(l)}, \xi^{(l)}, k^{(l)})}.$$

A standard choice for $\hat{p}$ would be a multivariate t density (Evans & Swartz, 1995). Verification that the importance weights are correct when $q/\hat{q}$ is replaced by $p/\hat{p}$ is straightforward; see Appendix 2. From this method of computing posterior expectations we may also obtain posterior variances and posterior interval probabilities.

4. Simulation studies

Our implementation has two key features: first, we use a fully Bayesian approach, together with a BIC approximation to the marginal density (4) and, secondly, we use the locality heuristic of Zhou & Shen (2001) to place new knots. Both of these choices may be contrasted with the implementation of Denison et al. (1998). In our simulation study we compute mean squared error for our Bayesian adaptive regression splines and compare with spatially adaptive regression splines, using the software of Zhou & Shen (2001), and with the Denison et al. method, using software available at http://www.ma.ic.ac.uk/$\sim$dgtd. We also investigate the relative importance of the two implementation changes by comparing with what we call the modified Denison et al. method, which includes the BIC approximation but not the change in candidate knot locations; we computed the modified Denison et al. method by inserting the required factor $1/\sqrt{n}$ into their code and recompiling. The Bayesian adaptive regression spline estimates of $E\{f(x) \mid y\} = E[E\{f(x) \mid y, k, \xi\}]$ are found from our Markov chain Monte Carlo with runs of length 10 000 following burn-ins of 1000.

In this section we consider three functions: a slowly-varying smooth function, a function

with a sharp peak, that is spatially inhomogeneously smooth, and a function with a discontinuity. Noise is added to each in generating the data. The functions together with samples of data are shown in Fig. 1.

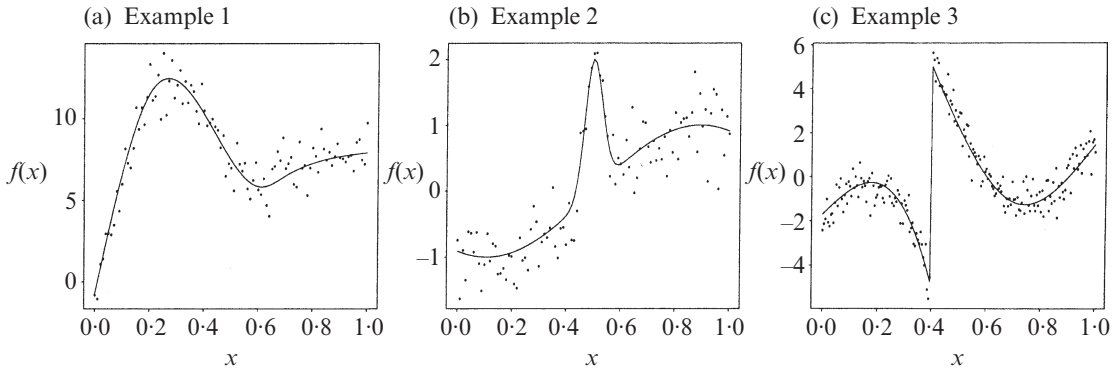

Fig. 1. The three true functions used in the simulation study together with one sample.

Example 1. The true function is a spline with three internal knots at $(0{\cdot}2, 0{\cdot}6, 0{\cdot}7)^{\mathrm{T}}$ and coefficients $\beta = (20, 4, 6, 11, 6)^{\mathrm{T}}$. The function is evaluated on a regular grid of 101 points, and a zero-mean Normal noise is added to the true function with $\sigma = 0{\cdot}9$, so that the signal-to-noise ratio, $\mathrm{SD}(f)/\sigma$, is 3.

Example 2. The true function is

$$f(x) = \sin(x) + 2\exp(-30x^2), \quad x \in [-2, 2],$$

evaluated at 101 regularly spaced points, and the standard deviation of the noise is chosen to be $\sigma = 0{\cdot}3$, so that again the signal-to-noise ratio is 3.

Example 3. The true function is a spline with five knots located at $(0{\cdot}4, 0{\cdot}4, 0{\cdot}4, 0{\cdot}4, 0{\cdot}7)$ and coefficients $(2, -5, 5, 2, -3, -1, 2)$. The function is evaluated on a regular grid of 201 points in $[0, 1]$, and zero-mean Normal noise is added to the true function with $\sigma = 0{\cdot}55$.

We compare our Bayesian adaptive regression splines estimates with spatially adaptive regression splines, Denison et al. (1998), and our modified Denison et al. estimates using mean squared error

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\{\hat{f}(t_i) - f(t_i)\}^2.$$

The average mean squared error, together with standard errors, based on 10 samples of data is reported in Table 1. The Bayesian estimates in Table 1 are all based on a Poisson prior with mean 5 for the number of knots, k. However, when we used a Uniform prior on $1, \ldots, 20$ or a Poisson with mean ranging in value between 1 and 20, the mean squared error never changed by more than 25% across these examples, and these changes do not alter the basic ordering found.

We see from Table 1 that Bayesian adaptive regression splines produces values of mean squared error that are smaller than those from Denison et al. (1998) and spatially adaptive regression splines across all three test functions. The modified Denison et al. method works well for Example 2 and always improves on the original Denison et al. (1998).

Table 1. *Simulation study. Average mean squared errors with estimated standard errors in brackets based on* 10 *samples obtained using four different procedures*

	SARS	DMS	Modified-DMS	BARS
Example 1	0·144 (0·030)	0·206 (0·029)	0·103 (0·019)	0·066 (0·007)
Example 2	0·015 (0·001)	0·025 (0·002)	0·012 (0·001)	0·008 (0·001)
Example 3	0·044 (0·006)	0·106 (0·007)	0·091 (0·004)	0·019 (0·003)

Methods: SARS, spatially adaptive regression splines; DMS, Denison et al. (1998); Modified-DMS, modified Denison et al.; BARS, Bayesian adaptive regression splines.

However, in Examples 1 and 3, Bayesian adaptive regression splines provides substantial further improvement, in part as a result of the locality heuristic for generating new knots. In Fig. 2, we see the true function of Example 3 together with its estimates obtained using different procedures. Figure 2 suggests that the success of Bayesian adaptive regression splines results from its avoiding overfitting and its ability to adapt to sharp jumps in the curves.

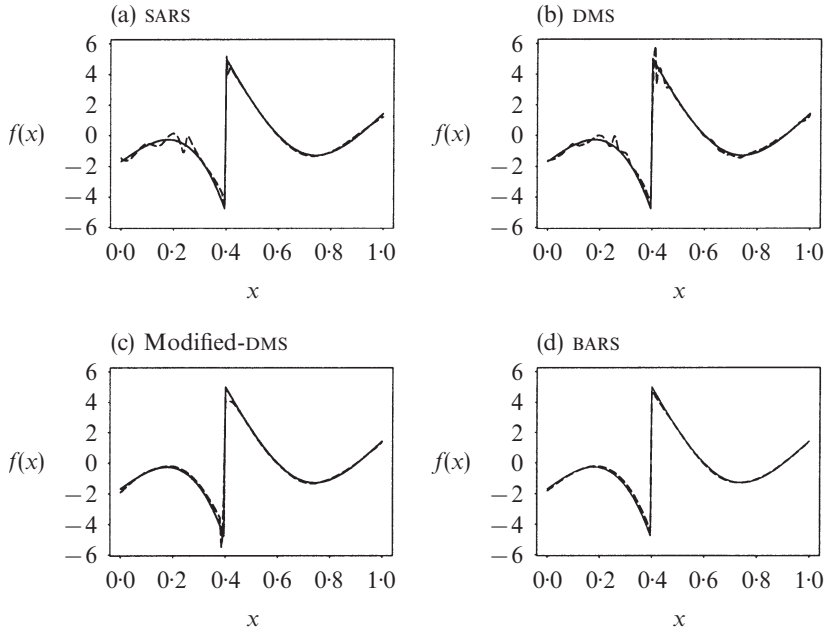

Fig. 2. Simulation study. Comparisons of the estimates of the discontinuous function of Example 3: solid lines, true curves; dashed lines, estimates of the curve. Methods: SARS, spatially adaptive regression splines; DMS, Denison et al. (1998); Modified-DMS, modified Denison et al.; BARS, Bayesian adaptive regression splines.

For the functions in both Examples 1 and 3 the posterior mode for the number of knots is the true number of knots, which is three and five respectively, and the conditional posterior of the locations of the knots given that the modal number of knots is concentrated around the true locations of the knots.

5. Functional magnetic resonance imaging example

In a functional magnetic resonance imaging experiment, a subject is placed in a magnetic resonance scanner and asked to perform a sequence of behavioural tasks while three-dimensional images of the subject's brain are acquired at regular intervals. Concentrated neural firing in the brain gives rise to a localised physiological response that is detectable in the images as a small, localised signal change. An analysis of functional magnetic resonance imaging data attempts to identify and characterise these task-related signal changes amidst a complicated noise process and other nuisance sources of variation; see Genovese (2000) for more details.

We consider two simple experiments in which the subject maintains visual fixation on a cross in the centre of the visual field while alternating S-second periods of rest and an experimental task. In Experiment 1, $S = 8$ and the task is to tap the thumb and forefinger together. In Experiment 2, $S = 42$ and the task is to note the location of a flash of light which appears at a random location in the visual field. Figure 3 shows magnetic resonance signal time-courses for the two experiments. The signals are taken from small volumes in the brain that are activated by the respective experimental tasks; the task-related signal changes in response to performing the experimental task are visible in both cases.

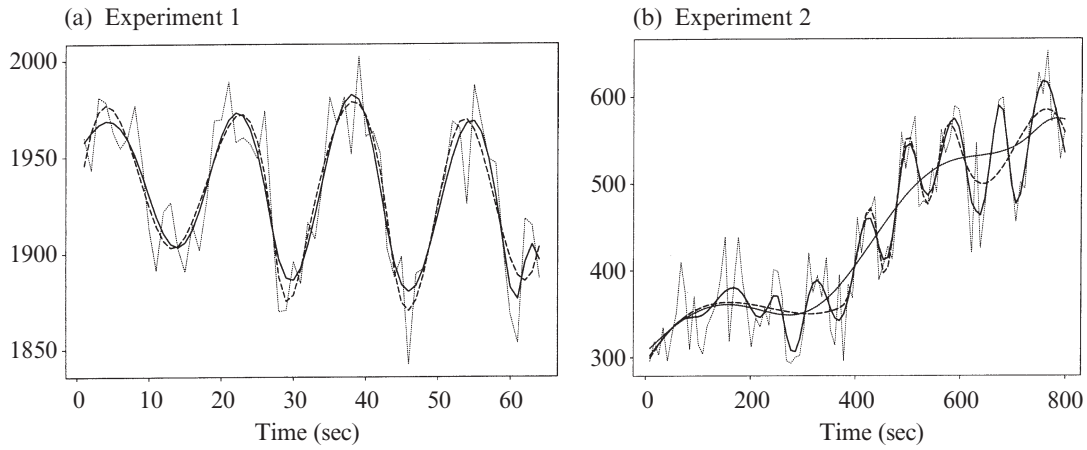

Fig. 3. Magnetic resonance example. The time-courses show the magnetic resonance signal as a thin dotted line. (a) shows the signal for Experiment 1 measured in one volume element over time in 'local magnetic resonance units' that depend on the scanner and pre-processing used. Superimposed on the signal are the Bayesian adaptive regression splines fit, solid line in (a), and the spatially adaptive regression spline fit, dashed line. (b) shows the signal for Experiment 2. Superimposed are the spatially adaptive regression spline fit (dashed line), the Bayesian adaptive regression spline fits using a Po(20) prior (solid line) and a Po(3) prior (dotted line) on the number of knots.

Bayesian adaptive regression splines can be useful in functional magnetic resonance imaging in many different roles. We discuss two here: (i) a flexible denoiser for magnetic resonance time-courses, where all smooth sources of variation are combined into the function being estimated, and (ii) a component in a semiparametric model that explicitly parameterises the task-related signal changes while treating nuisance variation such as drifts flexibly. The first approach can serve as a front-end to spatial and regional analyses and group comparisons, automatically incorporating variations in response shape and magnitude across the replicated task blocks in the experiment. The second approach can serve as a component in a hierarchical model for the data and can be used to characterise

task-related signal changes. A key advantage of our Bayesian formulation in both cases is that it easily provides a useful assessment of uncertainty.

To illustrate the method's role as a flexible denoiser, Fig. 3(a) compares the denoised time-courses given by Bayesian adaptive regression splines and spatially adaptive regression splines for Experiment 1. Both methods give similar results and appear to capture the gross signal changes quite well. Spatially adaptive regression splines better captures the small, sharp undershoot dips after the main response peak, but Bayesian adaptive regression splines appears more stable near the boundaries of the interval. Both methods give sharper activation peaks than the data seem to suggest by eye. Figure 3(b) presents a similar comparison for Experiment 2, where the signal changes are smaller relative to the noise and where there is a notable nuisance signal drift. Spatially adaptive regression splines detects some of the task-related signal changes but smooths over others near the drift changepoint. Bayesian adaptive regression splines, on the other hand, captures all of the responses reasonably. Figure 4 displays pointwise 95% high posterior density and confidence intervals for these estimates; the spatially adaptive regression splines confidence intervals were generated with 1000 samples from a parametric bootstrap treating the noise at each time point as independent and identically distributed normals.

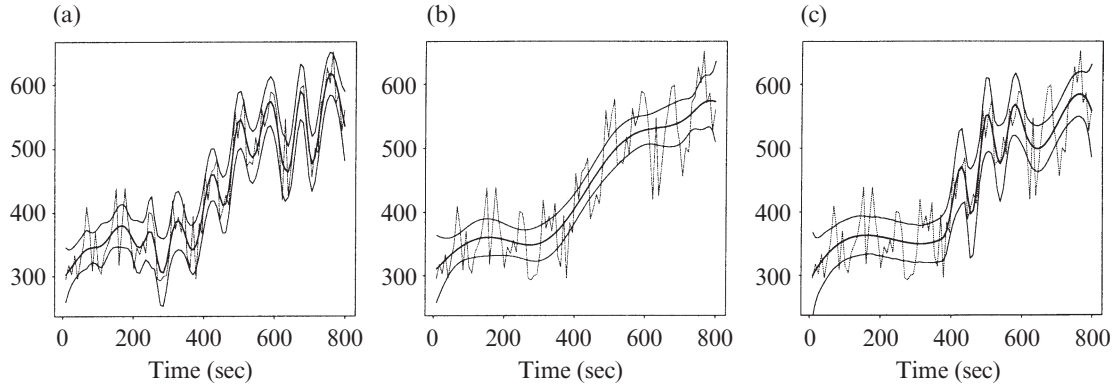

Fig. 4. Magnetic resonance example for Experiment 2. (a) and (b) show 95% high posterior density Bayesian adaptive regression splines, and (c) shows 95% confidence spatially adaptive regression splines intervals (all as thin solid lines), for the curve estimate (solid line) superimposed on the signal (thin dotted line) for the estimates of the curve using Bayesian adaptive regression splines with Poisson prior in (a) with mean 20, and (b) with mean 3, and using spatially adaptive regression splines in (c).

To illustrate the method's role as a semiparametric model component, we use Bayesian adaptive regression splines as part of an additive model with a flexible component for signal drift and a parametric component for task-related signal changes. For example, if we set the prior on the number of knots to a smooth setting, for example Po(3), we obtain the estimate in Fig. 3(b) and Fig. 4(b). Figure 5 shows the semiparametric fit obtained by adding a periodic parametric component to our model. Through the back-fitting algorithm (Hastie & Tibshirani, 1990, Ch. 4) we fit an additive model in which the function is decomposed into a sinusoid of the same period as the experimental design and a smooth component as just described. Figure 5 shows the estimate of the function and the extracted signal drift component, and Fig. 4 shows corresponding 95% high posterior density and confidence intervals. We could also use Bayesian adaptive regression splines for the task-related component by fitting each task block with a separate additive term, though at

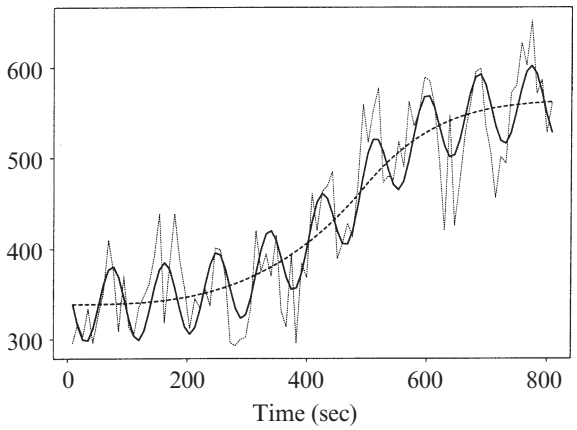

Fig. 5. Magnetic resonance example. The time-course shows the magnetic resonance signal for Experiment 2. Superimposed on the signal, thin dotted line, is a semiparametric fit, solid line, together with the nonparametric estimate of the linear trend, dashed line.

more computational expense. Even better would be to cast the additive structure in a Bayesian hierarchical model.

6. A Poisson model for neuron firing data

In a recent experiment, the firing of individual neurons in the inferotemporal cortex of a macaque monkey were recorded while he watched images appear on a screen in front of him (Olson & Rollenhagen, 1999). In one experimental condition, Condition 1, a black patterned object was displayed as the stimulus for 600 milliseconds. In a second condition, Condition 2, prior to the display of the stimulus a pair of blue rectangles were displayed and these remained illuminated while the stimulus was displayed. The typical inferotemporal neuronal response to the stimulus was roughly damped-oscillatory firing. In the second condition, however, the oscillation tended to be more pronounced, with higher drop from peak to trough. We use the methodology developed in §§ 2 and 3 to fit the data for one neuron and quantify the comparison of initial peak-to-trough drop in firing rates.

The data consist of neuronal spike counts from 193 repeated trials binned into 10-millisecond intervals. As discussed in a related context by Kass & Ventura (2001) and Ventura et al. (2001), such count data may be expected to be very nearly Poisson, and preliminary examination of the data indicated that this assumption was very reasonable. The model we use, therefore, for the counts $\{Y_i, i = 1, \ldots, n\}$ at time $\{x_i, i = 1, \ldots, n\}$ is as follows:

$$(Y_i|\beta, k, \xi) \sim \mathrm{Po}(\lambda_i), \quad \log(\lambda_i) = B(x_i)\beta, \quad \beta|k, \xi \sim N(0, D), \quad p(k, \xi), \tag{10}$$

where D is the matrix from the unit-information prior. We do not have to write down D explicitly because, as explained in §§ 2 and 3, we use the BIC-based reversible-jump Markov chain Monte Carlo scheme together with importance reweighting. As our importance function for the posterior on β, we have used a Normal approximation based on the maximum likelihood estimates and the observed information matrix. Comparison of the results before and after importance reweighting indicates that the Normal distribution is in fact a good approximation. We did all our computations in S-Plus.

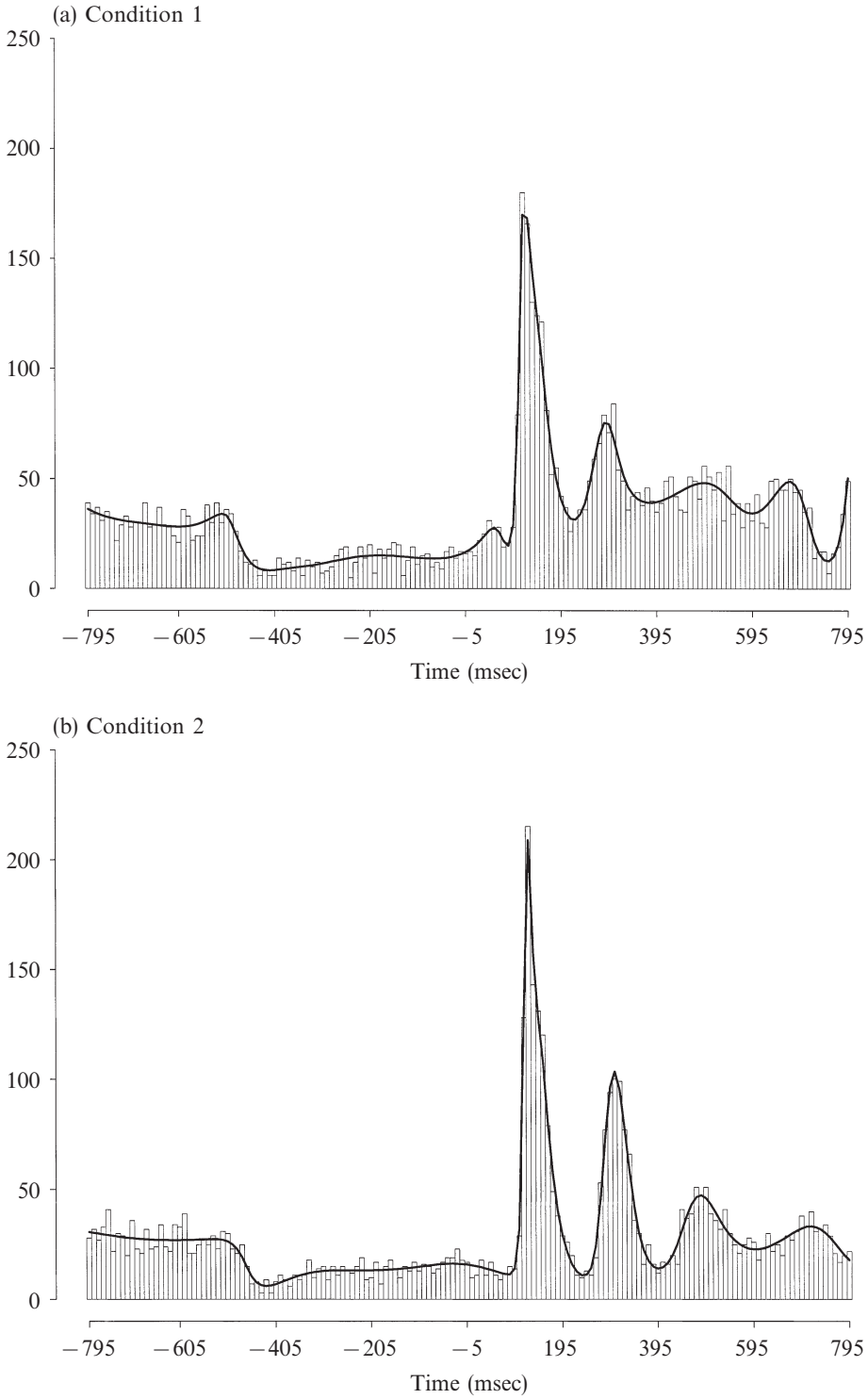

Fig. 6. Neuronal firing example. Counts in 10 millisecond bins together with the fitted curves, representing posterior means $E\{f(t)|y\}$.

The resulting Bayesian adaptive regression splines fits for the posterior means $E\{f(t)|\, y\}$ are given together with the raw counts in Fig. 6. Bayesian adaptive regression splines nicely adapts to the changing irregularities of the intensity functions producing estimates that are consistent with intuition: the intensities may change sharply on time scales of about 50 milliseconds, but are quite smooth on finer time scales. Table 2 gives the posterior means and posterior standard deviations for the quantities of interest. The maximal firing rate was, for example, defined as $g(\beta, k, \xi) = \arg\max_t f(t) \doteq \arg\max_{t_i} B\beta$. The substantive conclusion is that the drops from the first, highest peak to the following trough for Conditions 1 and 2 were $131{\cdot}8 \pm 4{\cdot}4$ and $181{\cdot}8 \pm 20{\cdot}3$ spikes per second; Condition 2 had a more pronounced drop, estimated to be $50{\cdot}0 \pm 20{\cdot}8$ spikes per second greater than that for Condition 1, with 95% probability interval (8·4, 91·7).

Table 2. *Neuronal firing example. Posterior means of maximal firing rate, local minimal firing rate just after the maximal firing rate, and the drop, i.e. the difference between these two firing rates, for each condition. Posterior standard deviations are shown in parentheses*

Firing rate	Condition 1	Condition 2
Maximum	166·5 (5·2)	193·0 (20·6)
Local min.	34·8 (1·9)	11·5 (1·5)
Difference	131·8 (4·4)	181·8 (20·4)

7. Discussion

Bayesian adaptive regression splines is a fully Bayesian, flexible spline model suitable for both normal and nonnormal data. It provides a mechanism for deriving useful uncertainties in function estimates and can easily be inserted as a component in a larger hierarchical model, as we have demonstrated here in § 5. Balancing against this advantage is the additional computational cost of the simulation: spatially adaptive regression splines is notably faster than Bayesian adaptive regression splines. However, this should not be a serious handicap in applications involving small or moderately large datasets. Key advantages of the method adopted here that distinguish it from the closely related approach applied by Biller (2000) are the placement of knots by a continuous proposal distribution and the introduction of the unit-information prior as a default, so that the chain simulates the approximate marginal posterior of (k, ξ) after integrating β; this increases efficiency (Liu et al., 1994).

Bayesian adaptive regression splines results and performance depend to some extent on the choice of knot priors. Thus, user input on the expected number of knots is needed as a kind of smoothing parameter. We used this to our advantage in the functional magnetic resonance imaging example to adapt Bayesian adaptive regression splines to different tasks. However, for large signal-to-noise ratios, or large samples, Bayesian adaptive regression splines will correctly find the appropriate number of knots regardless of the prior.

This paper has focused on estimates and standard errors, but one big advantage of a Bayesian formulation is the ability to estimate a wide range of features for the function of interest. We intend to explore Bayesian adaptive regression splines' effectiveness as a component of Bayesian hierarchical models in future work.

ACKNOWLEDGEMENT

This research was partially supported by grants from the U.S. National Science Foundation and National Institutes of Health.

APPENDIX 1

Detailed balance

In order to prove that detailed balance holds for this chain, we have to show that

$$\pi(M_k)\,\mathrm{pr}(M_{k-1}|M_k) = \pi(M_{k-1})\,\mathrm{pr}(M_k|M_{k-1}), \tag{A1}$$

where M_k denotes the parameters of the model with k knots: $M_k = \{k, \xi_1, \ldots, \xi_k\}$, for $k = 1, 2, \ldots$ and $\xi_i \in (0, 1)$. The $\pi(M_k)$ density is the target from which we want to draw observations; in our case $\pi(M_k)$ is the posterior distribution of M_k, namely

$$\pi(M_k) = \frac{p(y|\xi_1, \ldots, \xi_k)p(\xi_1, \ldots, \xi_k|k)p(k)}{p(y)}.$$

The formula $\mathrm{pr}(M_{k-1}|M_k)$ is a Markov transition kernel, the transition probability of going from M_k to M_{k-1}. Let

$$M_k = \{k, \xi_1, \xi_2, \ldots, \xi_{j^*-1}, \xi_{j^*}, \xi_{j^*+1}, \ldots, \xi_k\},$$
$$M_{k-1} = \{k-1, \xi_1, \xi_2, \ldots, \xi_{j^*-1}, \xi_{j^*+1}, \ldots, \xi_k\}.$$

The sets of knots in the two spaces differ only in the j^*th element. We can now write the transition probabilities as follows:

$$\begin{aligned}
\mathrm{pr}(M_{k-1}|M_k) &= \underbrace{\mathrm{pr}(k-1|k)}_{d_k} \times \underbrace{\mathrm{pr}(\text{delete } \xi_{j^*}|k)}_{1/k} \times \underbrace{(\text{acceptance probability})}_{\alpha_d} \\
&= d_k \frac{1}{k} \min(1, A), \\
\mathrm{pr}(M_k|M_{k-1}) &= \underbrace{\mathrm{pr}(k|k-1)}_{b_{k-1}} \times \underbrace{\mathrm{pr}(\text{add } \xi_{j^*}|k-1)}_{1/(k-1)\sum_i h_B(\xi_{j^*}|\xi_i)} \times \underbrace{(\text{acceptance probability})}_{\alpha_b} \\
&= b_{k-1} \frac{1}{k-1} \sum_i h_B(\xi_{j^*}|\xi_i) \min(1, B),
\end{aligned}$$

where

$$A = \frac{\pi(M_{k-1})}{\pi(M_k)} \frac{b_{k-1}(k-1)^{-1}\sum_i h_B(\xi_{j^*}|\xi_i)}{d_k k^{-1}}, \quad B = \frac{\pi(M_k)}{\pi(M_{k-1})} \frac{d_k k^{-1}}{b_{k-1}(k-1)^{-1}\sum_i h_B(\xi_{j^*}|\xi_i)} = 1/A.$$

We can now verify (A1). If $A < 1$, then $\alpha_d = A$ and $\alpha_b = 1$, and therefore rewriting (A1) we have that

$$\begin{aligned}
\pi(M_k)\,\mathrm{pr}(M_{k-1}|M_k) &= \pi(M_k) d_k \frac{1}{k} A = \pi(M_k) d_k \frac{1}{k} \frac{\pi(M_{k-1})}{\pi(M_k)} \frac{b_{k-1}(k-1)^{-1}\sum_i h_B(\xi_{j^*}|\xi_i)}{d_k k^{-1}} \\
&= \pi(M_{k-1}) b_{k-1} \frac{1}{k-1} \sum_i h_B(\xi_{j^*}|\xi_i) = \pi(M_{k-1})\,\mathrm{pr}(M_k|M_{k-1}).
\end{aligned}$$

The case when $A > 1$ is now obvious. Also the proof of the detailed balance condition when we move from M_k to M_k', a relocation step, is straightforward.

Appendix 2

Importance sampling

We wish to determine the weight for our problem. If $g(\beta, k, \xi)$ is the functional of interest, we need to compute

$$E\{g(\beta, \xi, k)|y\} = \frac{\int \dots \int g(\beta, \xi, k) \dfrac{q(\beta|y, k, \xi)}{\hat{q}(\beta|y, \xi, k)} \hat{q}(\beta|y, k, \xi) p(k, \xi|y)\, d\beta\, d\xi\, dk}{\int \dots \int \dfrac{q(\beta|y, k, \xi)}{\hat{q}(\beta|y, \xi, k)} \hat{q}(\beta|y, k, \xi) p(k, \xi|y)\, d\beta\, d\xi\, dk} = \frac{A}{B},$$

say, where

$$\begin{aligned}
A &= \int \dots \int g(\beta, k, \xi) q(\beta|y, k, \xi) p(k, \xi|y)\, d\beta\, d\xi\, dk \\
&= \int \dots \int g(\beta, k, \xi) \frac{q(\beta|y, k, \xi)}{\hat{q}(\beta|y, k, \xi)} \hat{q}(\beta|y, k, \xi) p(k, \xi|y)\, d\beta\, d\xi\, dk \\
&= \int \dots \int g(\beta, k, \xi) \frac{p(y|\beta, k, \xi)\pi_\beta(\beta|k, \xi)}{\hat{p}(y|\beta, \xi, k)\pi_\beta(\beta|\xi, k)} \frac{\hat{p}(y)}{p(y)} \hat{q}(\beta|y, k, \xi) p(k, \xi|y)\, d\beta\, d\xi\, dk \\
&= \frac{\hat{p}(y)}{p(y)} \int \dots \int g(\beta, \xi, k) \frac{p(y|\beta, k, \xi)}{\hat{p}(y|\beta, \xi, k)} \hat{q}(\beta|y, k, \xi) p(k, \xi|y)\, d\beta\, d\xi\, dk, \\
B &= \frac{\hat{p}(y)}{p(y)} \int \dots \int \frac{p(y|\beta, k, \xi)}{\hat{p}(y|\beta, \xi, k)} \hat{q}(\beta|y, k, \xi) p(k, \xi|y)\, d\beta\, d\xi\, dk.
\end{aligned}$$

Therefore

$$\begin{aligned}
E\{g(\beta, \xi, k)|y\} &= \frac{\int \dots \int g(\beta, \xi, k) \dfrac{p(y|\beta, k, \xi)}{\hat{p}(y|\beta, \xi, k)} \hat{q}(\beta|y, k, \xi) p(k, \xi|y)\, d\beta\, d\xi\, dk}{\int \dots \int \dfrac{p(y|\beta, k, \xi)}{\hat{p}(y|\beta, \xi, k)} \hat{q}(\beta|y, k, \xi) p(k, \xi|y)\, d\beta\, d\xi\, dk} \\
&\doteq \frac{\sum_l g(\beta^{(l)}, \xi^{(l)}, k^{(l)}) w(\beta^{(l)}, \xi^{(l)}, k^{(l)})}{\sum_l w(\beta^{(l)}, \xi^{(l)}, k^{(l)})},
\end{aligned}$$

where

$$w(\beta^{(l)}, \xi^{(l)}, k^{(l)}) = \frac{p(y|\beta^{(l)}, \xi^{(l)}, k^{(l)})}{\hat{p}(y|\beta^{(l)}, \xi^{(l)}, k^{(l)})},$$

$(\xi^{(l)}, k^{(l)})$ is the pair accepted by the reversible-jump sampler, i.e. is sampled from $p(k, \xi|y)$, and $\theta^{(l)}$ is sampled from $\hat{q}(\beta|y, \xi^{(l)}, k^{(l)})$.

Appendix 3

Posterior approximations

First, we elaborate on the essential property of the BIC-based approximation we are using. Let $\hat{p}(y|k, \xi)$ be the approximation to $p(y|k, \xi)$ and assume that $k \leqslant K$ for some fixed K. Then, from Laplace's method, $\hat{p}(y|k, \xi) = p(y|k, \xi)\{1 + O_p(n^{-1/2})\}$ uniformly in (k, ξ). Here, O_p refers to the sampling distribution of the data. Let us use Pr to denote probabilities under this sampling distribution and let Ω denote the space of (k, ξ) values. It follows by integration that, for any arbitrarily small positive η, there exists a bound M such that, for all measurable subsets $A \subseteq \Omega$ and for all

sufficiently large n, we have

$$\Pr\{|\hat{P}(A|y) - P(A|y)| < M/\sqrt{n}\} > 1 - \eta,$$

where $P(A|y)$ and $\hat{P}(A|y)$ denote posterior and approximate posterior probabilities of A. This is the formal sense in which the posterior using BIC approximates the correct posterior.

Secondly, we provide details for our statement that the marginal density $p^{\mathrm{DMS}}(y|k, \xi)$ is monotonically increasing in k. Let $k' \geqslant k$ and, given a basis matrix $B_{k,\xi}$, generate another, $B'_{k'\xi'}$, by adding knots. Then

$$\mathrm{span}(B_{k,\xi}) \subseteq \mathrm{span}(B'_{k',\xi'}), \quad \max_{\beta} p(y|\beta, k, \xi) \leqslant \max_{\beta'} p(y|\beta', k', \xi').$$

Therefore, for each (k, ξ) there exists (k', ξ') such that $p^{\mathrm{DMS}}(y|k, \xi) \leqslant p^{DMS}(y|k', \xi')$.

References

Biller, C. (2000). Adaptive Bayesian regression splines in semiparametric generalized linear models. *J. Comp. Graph. Statist.* **9**, 122–40.

Denison, D. G. T., Mallick, B. K. & Smith, A. F. M. (1998). Automatic Bayesian curve fitting. *J. R. Statist. Soc.* B **60**, 330–50.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with Discussion). *Statist. Sci.* **11**, 89–121.

Evans, M. & Swartz, T. (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statist. Sci.* **10**, 254–72.

Genovese, C. R. (2000). A Bayesian time-course model for functional Magnetic Resonance Imaging data. *J. Am. Statist. Assoc.* **95**, 691–719.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–32.

Halpern, E. F. (1973). Bayesian spline regression when the number of knots is unknown. *J. R. Statist. Soc.* B **35**, 347–60.

Hansen, M. H. & Kooperberg, C. (2001). Spline adaptation in extended linear models. *Statist. Sci.* To appear.

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

Kass, R. E. & Ventura, V. (2001). A spike train probability model. *Neural Comp.* **13**, 1713–20.

Kass, R. E. & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Statist. Assoc.* **90**, 928–34.

Lindstrom, M. J. (1999). Penalized estimation of free-knot splines. *J. Comp. Graph. Statist.* **8**, 333–52.

Liu, J. S., Wong, W. H. & Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81**, 27–40.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.

Olson, C. R. & Rollenhagen, J. E. (1999). Low-frequency oscillations in Macaque IT cortex during competitive interactions between stimuli. *Soc. Neurosci. Abstr.* **25**, 916.

Pauler, D. K. (1998). The Schwarz criterion and related methods for normal linear models. *Biometrika* **85**, 13–27.

Smith, M. & Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Economet.* **75**, 317–43.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with Discussion). *Ann. Statist.* **22**, 1701–62.

Ventura, V., Carta, R., Kass, R. E., Gettner, S. & Olson, C. R. (2001). Statistical analysis of temporal evolution in single-neuron firing rates. *Biostatistics* **2**. To appear.

Wahba, G. (1990). *Spline Models for Observational Data.* Philadelphia: Society for Industrial and Applied Mathematics.

Zhou, S. & Shen, X. (2001). Spatially adaptive regression splines and accurate knot selection schemes. *J. Am. Statist. Assoc.* **96**, 247–59.

[*Received February* 2000. *Revised May* 2001]