Bayesian Curve Fitting with Free-Knot Splines

Ilaria DiMatteo, Christopher R. Genovese, and Robert E. Kass

The authors are in the Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213. The work of the first and third author was partially supported by NSF grants DMS-9303557 and DMS-9803433, and NIH grant RO1-CA54852-04. The work of the second author was partially supported by NSF grants DMS-9705034.

Summary

Reversible-jump Markov chain Monte Carlo may be used to fit scatterplot data with cubic splines having unknown numbers of knots and knot locations. Key features of the implementation investigated here are (i) a fully Bayesian formulation that puts priors on the spline coefficients and (ii) Metropolis-Hastings proposal densities that attempt to place knots close to one another. Simulation results indicate this methodology can produce fitted curves with substantially smaller mean squared-error than competing methods. The reversible-jump implementation requires ratios of marginal densities for the data (integrated likelihood ratios). We approximate these using the Bayes Information Criterion and thereby obtain a general approach to Bayesian nonparametric regression for arbitrary response-variable distributions. We illustrate with an application to Poisson nonparametric regression modeling of neuron firing patterns.

Key Words: BIC; Generalized linear models; Nonparametric regression; Reversible-jump MCMC; Smoothing; Unit-information prior.

1 Introduction

Over the past 30 years much progress has been made in devising and understanding methods of fitting curves to data, including those based on splines. Smoothing splines, in particular, are appealing because they provide computationally efficient estimation, often do a good job of smoothing noisy data, and have both frequentist and Bayesian justifications (e.g., Hastie and Tibshirani, 1990; Wahba, 1990). Two shortcomings of smoothing splines, however, are the need to choose a global smoothness parameter and the inability of linear procedures (conditional on the choice of smoothness) to adapt to spatially heterogeneous functions. This has led to investigations of curve-fitting using free-knot splines, that is, splines in which the number of knots and their locations are determined from the data (Denison, Mallick, and Smith, 1998; Lindstrom, 1999; Zhou and Shen, 1998; Genovese, 2000; Eilers and Marx, 1996). In this paper we use reversible-jump Markov chain Monte Carlo (MCMC) technology (Green, 1995) to fit curves to data using free-knot splines. The implementation we investigate substantially improves and generalizes that of Denison, Mallick, and Smith (DMS), incorporating a key observation made by Zhou and Shen (ZS).

1.1 The Normal Regression Model

We begin by treating the basic nonparametric regression model

$$Y_i = f(x_i) + \varepsilon_i, \tag{1}$$

where the unknown f is a real-valued function on $[a, b]$ and where we assume that ε_i for $i = 1, \ldots, n$ are independent draws from $N(0, \sigma^2)$ with σ unknown. In order to estimate f, we will assume it is a cubic spline with an unknown number k of knots at unknown locations $\xi = (\xi_1, \cdots, \xi_k)^T$, where $a < x_{(1)} < \xi_1 \leq \cdots \leq \xi_k < x_{(n)} < b$ and where $x_{(1)}$ and $x_{(n)}$ denote respectively the minimum and maximum sampling points. In practical terms, we are assuming that f may be well-approximated by such a cubic spline.

We let $b_j(x)$ denote the j-th function in a cubic B-spline basis with natural boundary constraints (linear outside $[a,b]$) and let B_ξ be the matrix with (i,j)th component equal to $b_j(x_i)$. We use the subscript ξ to express the dependency of the matrix on the number and location of the knots. Because we use the natural cubic B-splines, there are $(k+2)$ such basis functions. Under our assumptions, we can write f as a linear combination

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

We write the vector $(f(x_i))_{i=1}^n = B_\xi \beta$ where $\beta = (\beta_1, \ldots, \beta_{k+2})$.

We complete the Bayesian formulation of the problem by introducing a prior on the unknown quantities (β, ξ, k, σ) and estimate the curve using the pointwise posterior mean $\hat{f}(x) = E(f(x)|y)$. The remaining challenge, then, is to choose (i) priors and (ii) MCMC proposal distributions in such a way that the computational scheme is tractable and the resulting curve $\hat{f}$ is a good estimator of f.

Green (1995) pointed out that reversible-jump MCMC is well-suited to estimating the number and locations of change-points in fitting step functions. DMS made the observation that the analogous solution to the closely related problem of estimating number and locations of knots for free-knot splines should open up a potentially very powerful nonparametric regression method. Instead of a fully Bayesian approach to the problem, however, DMS adopted a quasi-Bayesian solution that avoided the use of a prior on β. (They used least-squares estimates.) As we illustrate in Section 4, and discuss briefly in Section 7, this makes it impossible to learn from the data about likely values of k and tends to result in severe overfitting. Furthermore, it undercuts the power of the Bayesian approach by eliminating the posterior distribution on β, the vehicle by which inferences are most naturally made about characteristics of the function f, such as the location of its maximum (which is important in the example we treat in Section 6).

In the present paper, we use a particular conjugate Normal prior on β that Kass and Wasserman (1995) called the *unit-information prior*, together with the prior $\pi_\sigma(\sigma) = 1/\sigma$. With these choices, β and σ may be integrated from the posterior in order to obtain a Markov chain for sampling from the marginal posterior on (ξ, k). Specifically, the integral

$$p(y|\xi, k) = \int p(y|\beta, \xi, k, \sigma)\pi(\beta, \sigma|\xi, k)d\beta d\sigma \tag{3}$$

may be evaluated analytically. Reversible-jump MCMC uses a Metropolis-Hastings acceptance probability $\alpha = \min\{1, \text{likelihood ratio} \times \text{prior ratio} \times \text{asymmetry correction}\}$. Here,

$$\text{likelihood ratio} = \frac{p(y|\xi^c, k^c)}{p(y|\xi, k)}, \tag{4}$$

where (ξ, k) and (ξ^c, k^c) are respectively the current state of the chain and the Metropolis-Hastings candidate state. From equation (3) for the marginal density, we are able to evaluate the likelihood ratio, and the MCMC implementation becomes straightforward. We then estimate the curve in the form

$$\hat{f}(x) = E(f(x)|y) = E(E(f(x)|y, \xi, k)) \tag{5}$$

by averaging $E(f|y, \xi, k)$ over the MCMC-sampled values of (ξ, k).

While it is analytically and computationally tractable to use the posterior mean of $f(x)$, using a posterior mode may have some competing advantages. For instance, if the scatterplot is fit comparably well by several functions with distinct spatial structure, averaging over those high-posterior cases may tend to oversmooth. Choosing the mode produces an estimate of appropriate smoothness, although at the cost of ignoring the other structural variation. It is straightforward to use the mode with our method whenever the posterior mode conditional on (ξ, k) can be easily determined, as in the model (1).

Kass and Wasserman (1995) and Pauler (1998) showed that ratios of marginal densities such as those occurring in (4) when the unit-information prior is used may be approximated very accurately in terms of the Bayes information criterion (BIC). Although not of great relevance to our treatment of the basic problem in (1), the ability to compute (4) with an accurate approximation allows the methodology to be generalized.

1.2 A General Nonparametric Regression Model

Our approach applies under a more general framework with little additional computational expense. Suppose for values $x_1, \ldots, x_n$ of an explanatory variable x, we have data $Y_1, \ldots, Y_n$ that satisfiy

$$Y_i \sim p(y|f(x_i), \zeta) \tag{6}$$

for some distribution $p(y|\theta, \zeta)$ with θ real-valued function of x (so that $\theta = f(x)$), and optional ζ vector-valued. An important special case occurs when Y_i follows a distribution from an exponential family and ζ is a possible dispersion parameter; with a particular choice of link function, this becomes Bayesian nonparametric regression for a generalized linear model. Using a spline basis, introducing the unit-information prior, and approximating with the BIC, we are able to employ essentially the same MCMC implementation with model (6) as with model (1).

In Section 2, we provide some further details about our choice of priors and our approximation to the likelihood ratios that appear in the reversible-jump MCMC algorithm and in Section 3 discuss some details of posterior simulation. We give simulation results for three elementary test functions in Section 4 and in Section 5, we analyze fMRI data. In Section 6, we analyze some Poisson data to illustrate both the generalization we have proposed and the importance of being able to make inferences about characteristics of the function f. In Section 7, we discuss possible further refinements and extensions.

2 Choice of Priors and the Likelihood Ratio Calculation

Let us begin by treating model (1). It is convenient, though not essential as we show below, to use a prior for which (3) may be obtained analytically. We decompose the prior as follows $\pi(\beta, \xi, k, \sigma) = \pi_\beta(\beta|\xi, k, \sigma)\pi_\xi(\xi|k)\pi_k(k)\pi_\sigma(\sigma)$ where $\pi_\sigma(\sigma) = 1/\sigma$ and

$$\beta \,|\, \xi, k, \sigma \sim N_{k+2}\left(0, \sigma^2 n (B_\xi^T B_\xi)^{-1}\right). \tag{7}$$

The remaining priors on ξ and k could be chosen to express knowledge about these parameters or, equivalently, to force some desired behavior in the posterior. In our simulations and application below we have adopted a uniform prior (Dirichlet$(1, 1, \ldots, 1)$) on ξ (given k) and either a Poisson prior or Uniform$\{1, \ldots, K_0\}$ prior on k. A negative binomial prior on k is another attractive option. 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, priors of the form (7) have been used by many authors (see Pauler, 1998). Kass and Wasserman (1995) have called these "unit-information" priors because the amount of information 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 (7) was used by Smith and Kohn (1996) in a different but related context of spline knot selection, where instead of n in (7) 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 are also able to obtain analytically the conditional posterior distribution of β given (ξ, k). This means that we can sample from the marginal posterior of β by drawing a value of β from this conditional posterior along with every sampled value of (ξ, k) in our chain. Furthermore, we obtain analytically the conditional posterior expectation of the unknown function f evaluated at the given values $x_1, \ldots, x_n$ from

$$E(f|\xi, k, y) = \frac{n}{n+1} B_\xi (B_\xi^T B_\xi)^{-1} B_\xi^T y \approx B_\xi \hat{\beta}. \tag{8}$$

The posterior expectation of $f(x)$ for any x can then be computed by averaging this conditional expectation over (ξ, k) samples. In the normal model, the posterior distribution of β given ξ and k can be computed analytically, making it easy to assess the uncertainty in β *after* a simulation on ξ and k alone. In the more general model, some additional steps might be required. When making inferences about functionals of f, this extra uncertainty cannot be ignored.

With the prior (7), the likelihood ratio (4) can be computed explicitly but also offers a close approximation based on BIC. The MCMC implementation has steps involving addition of a knot, deletion of a knot, and relocation of a knot. In the case of addition of a knot, for example, if the current state is (ξ, k) and the candidate state is (ξ^c, k^c) with $k^c = k + 1$, then (4) becomes

$$\begin{aligned}
\frac{p(y|\xi^c,k^c)}{p(y|\xi,k)} &= \frac{1}{\sqrt{n+1}}\left(\frac{y^T(I_n-\frac{n}{n+1}B_\xi(B_\xi^TB_\xi)^{-1}B_\xi^T)y}{y^T(I_n-\frac{n}{n+1}B_{\xi^c}(B_{\xi^c}^TB_{\xi^c})^{-1}B_{\xi^c}^T)y}\right)^{n/2} \\
&\approx \frac{1}{\sqrt{n}}\left(\frac{(y-B_\xi\hat{\beta})^T(y-B_\xi\hat{\beta})}{(y-B_{\xi^c}\hat{\beta}^c)^T(y-B_{\xi^c}\hat{\beta}^c)}\right)^{n/2} \\
&= \exp(-BIC/2),
\end{aligned} \tag{9}$$

where $\hat{\beta}$ and $\hat{\beta}^c$ are the respective least-squares estimates of β for the current and candidate states and where $BIC = \log p(y|\hat{\beta},\xi,k) - \log p(y|\hat{\beta}^c,\xi^c,k^c) - \frac{1}{2}\log n$. Because the least-squares fit is easy to compute, the exact form can be used directly in the normal model. In this case, we have used $n/(n+1)\approx 1$ and $1/\sqrt{n+1}\approx 1/\sqrt{n}$ in (9), thus obtaining the $\exp(-BIC/2)$ approximation which incurs an error of order $O(n^{-1/2})$.

The use of $\hat{\beta}$, i.e., integrating out the coefficients in the chain, is a key feature of both our method and the method of DMS. 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 $\hat{\beta}$. Second, 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).

The primary contrast between our method and that of DMS is that DMS do not use a prior on β but instead replace the likelihood ratio with $p(y|\hat{\beta}^c,\xi^c,k^c)/p(y|\hat{\beta},\xi,k)$, which omits the consequential factor $1/\sqrt{n}$ in (9). That is, it omits the term that BIC includes to penalize the likelihood ratio for dimensionality.

When we move to the more general framework of model (6), we use the same priors. The BIC approximation is particularly convenient here. An alternative would be to use a Laplace approximation to evaluate the marginal density (3) for other choices of priors.

3 Posterior Simulation

3.1 Reversible-jump MCMC

As we indicated in the Introduction, the framework we have adopted produces a Markov chain with the marginal posterior on (ξ, k) as stationary distribution. We have already discussed the likelihood factor in the Metropolis-Hastings acceptance probability. The prior factor is simply $\pi_{\xi,k}(\xi^c, k^c)/\pi_{\xi,k}(\xi, k)$ (where $\pi_{\xi,k}(\xi, k) = \pi_\xi(\xi|k)\pi_k(k)$). The proposal factor involves the ratio of proposal densities $q(\xi, k|\xi^c, k^c)/q(\xi, k|\xi^c, k^c)$. For the proposals we follow the general scheme used by Green (1995) and DMS, using steps that add, delete, or relocate knots. In contrast to DMS, where new knots are generated "far" from existing knots, our chain uses the locality heuristic devised in Zhou and Shen (1998) 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 parameterized by $(k, \xi_1, \cdots, \xi_k)$. The addition, deletion, and relocation steps of the reversible-jump sampler are attempted, respectively, with the following probabilities:

$$\begin{aligned} b_k &= c \cdot \min\{1, p(k+1)/p(k)\} \\ d_k &= c \cdot \min\{1, p(k-1)/p(k)\} \\ \eta_k &= 1 - b_k - d_k. \end{aligned}$$

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

Birth Step Let us 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 centered at ξ_{j^*} with some spread parameter τ_B having density $h_B(\xi_{cand}|\xi, \tau_B)$. In this case the jump probability is given by:

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

(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 to be deleted 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

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

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

Let $\xi = (\xi_1, \ldots, \xi_{j^*-1}, \xi_{j^*}, \xi_{j^*+1}, \ldots, \xi_k)^T$ be the current sequence of knots, and $\xi^c = (\xi_1, \ldots, \xi_{j^*-1}, \xi_{cand}, \xi_{j^*+1}, \ldots, \xi_k)^T$ 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 follow:

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

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, there remain values of parameters c, τ_B and τ_R to be selected. In principle, these may be considered tuning parameters, adjusted to produce a chain having good acceptance probabilities.

Here, we choose the constant c to be 0.4 as in DMS (a limited study of our own suggests this is a good value), and we take both the birth and relocation proposals to be Beta distributions with parameters $\xi_{j^*}\nu$ and $(1-\xi_{j^*})\nu$. We choose the parameter ν to be 50 in our examples. We obtain 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 Section 3.1 produces a chain on (ξ, k). As we indicated in Section 2, under model (1) 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 (ξ, k) for each sampled value of (ξ, k). Under model (6), however, it is typically infeasible to sample from this conditional posterior directly. Therefore, we resort to importance reweighting.

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

$$E(g(\beta,\xi,k)|y) \quad = \quad \frac{\int\cdot\cdot\int g(\beta,\xi,k)\frac{q(\beta|y,\xi,k)}{\hat{q}(\beta|y,\xi,k)}\hat{q}(\beta|y,\xi,k)p(\xi,k|y)d\beta d\xi dk}{\int\cdot\cdot\int \frac{q(\beta|y,\xi,k)}{\hat{q}(\beta|y,\xi,k)}\hat{q}(\beta|y,\xi,k)p(\xi,k|y)d\beta d\xi dk}$$

$$\approx \quad \frac{\sum_l g(\beta^{(l)}, \xi^{(l)}, k^{(l)}) w(\beta^{(l)}, \xi^{(l)}, k^{(l)})}{\sum_l w(\beta^{(l)}, \xi^{(l)}, k^{(l)})} \tag{10}$$

where $w(\beta^{(l)}, \xi^{(l)}, k^{(l)}) = \frac{q(\beta^{(l)}|y, \xi^{(l)}, k^{(l)})}{\hat{q}(\beta^{(l)}|y, \xi^{(l)}, k^{(l)})}$, $(\xi^{(l)}, k^{(l)})$ is the pair accepted by the reversible-jump sampler (i.e. sampled from $p(\xi, k|y)$), and $\beta^{(l)}$ is sampled from a suitable approximation $\hat{q}$ to the conditional posterior of β given (ξ, k). In fact, we may approximate the likelihood function on β given (ξ, k) rather than the full conditional posterior, which is typically easier under model (6), yielding weights of the form $w(\beta^{(l)}, \xi^{(l)}, k^{(l)}) = \frac{p(y|\beta^{(l)}, \xi^{(l)}, k^{(l)})}{\hat{p}(y|\beta^{(l)}, \xi^{(l)}, k^{(l)})}$. A standard choice of $\hat{p}$ would be a multivariate t density (e.g., Evans and Swartz, 1995). Verification that the importance weights are correct when $q/\hat{q}$ is replaced by $p/\hat{p}$ is straightforward; the argument is given in Appendix B. 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 (3) and, second, we use the locality heuristic of ZS to place new knots. Both of these choices may be contrasted with the implementation of DMS. In our simulation study we compute mean squared-error (MSE) for our Bayesian Adaptive Regression Splines (BARS) and compare with SARS (due to ZS, using their software) and with the DMS method (using software available at `http://www.ma.ic.ac.uk/~dgtd`). We also investigate the relative importance of the two implementation changes by comparing with what we call Modified-DMS, which includes the BIC approximation but not the change in candidate knot locations. (We computed the Modified-DMS by inserting the required factor $1/\sqrt{n}$ into their code and recompiling.) The

BARS estimates are obtained from (5), with expectations estimated from our MCMC 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 (spatially inhomogeneously smooth) and a function with a discontinuity, to which noise is added in generating the data. The functions together with samples of data are shown in Figure 1:

- Test Function 1: The true function is a spline with three internal knots at $(0.2, 0.6, 0.7)^T$ and coefficients $\beta = (20, 4, 6, 11, 6)^T$. The function is evaluated on a regular grid of 101 points, and a zero-mean Normal noise is added to the true function with $\sigma = 0.9$, so that the signal-to-noise ratio, $SD(f)/\sigma$, is 3.

- Test Function 2: The true function is

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

  evaluated at 101 regularly spaced points, and the standard deviation of the noise is chosen to be $\sigma = 0.3$, so that again the signal-to-noise ratio is 3.

- Test Function 3: The true function is a spline with five knots located at $(0.4, 0.4, 0.4, 0.4, 0.7)$ 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 $\sigma = 0.55$.

We compare our BARS estimates with SARS, DMS, and our Modified-DMS using mean squared-error

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

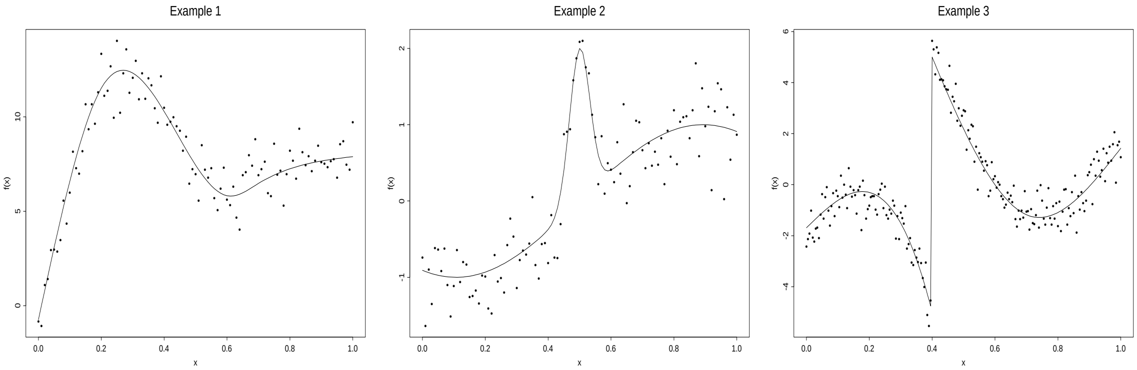

Figure 1: The three true functions used in the simulation study together with one sample.

The average MSE, together with standard errors, based on 10 samples of data is reported in Table 1. The Bayesian estimates in the Table 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 MSE 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 BARS produces MSEs that are smaller than DMS and SARS across all three test functions. The Modified-DMS works well for Example 2 and always improves on the original DMS. However in Examples 1 and 3, BARS provides substantial further improvement, in part due to the locality heuristic for generating new knots. In Figure 2, we can see the true function of Example 3 together with its estimates obtained using different procedures. The figures suggest that the success of BARS is due to its avoiding overfitting and its ability to adapt to sharp jumps in the curves.

For both the functions in Example 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.

Average MSE (s.e.)				
	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)

Table 1: Average MSE based on 10 sample obtained using different procedure.

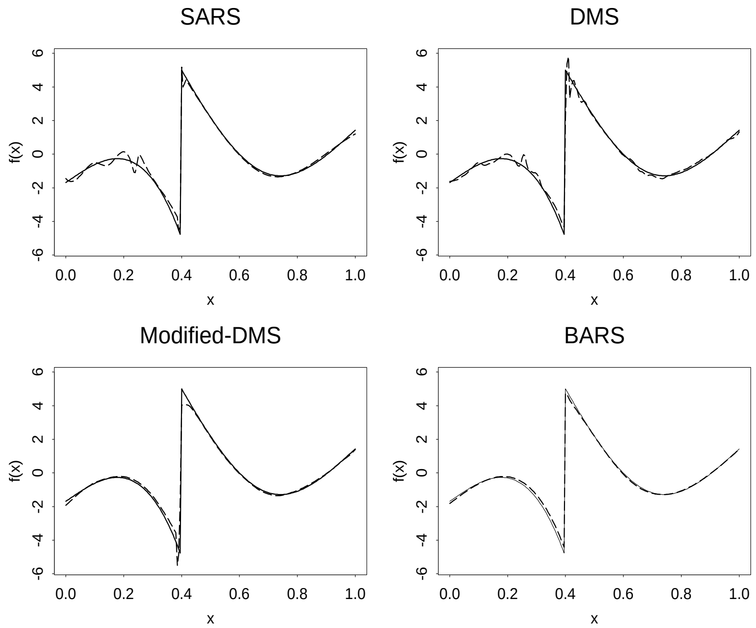

Figure 2: Comparisons of the estimates of the discontinuous function of Example 3: true curve (—), estimate of the curve (- - -)

5 Functional MRI Example

In a functional Magnetic Resonance Imaging (fMRI) experiment, a subject is placed in a Magnetic Resonance (MR) scanner and asked to perform a sequence of behavioral 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 localized physiological response that is detectable in the images as a small, localized signal change. An analysis of fMRI data attempts to identify and characterize these task-related signal changes amidst a complicated noise process and other nuisance sources of variation. See Genovese (2000) for more details.

In studying the performance of BARS, we applied the method to signal time-courses associated with a small volume in the brain. These time courses are taken from two experiments. The first is a simple, unpublished fMRI experiment in which the subject maintains visual fixation on a cross in the center of the visual field while alternating 8 seconds periods of rest and tapping thumb and index finger together at a constant frequency. In the second experiment, each period is 42 seconds long and the task is to note the location of a light flash somewhere in the visual field. When the light flashes, the subject's eyes move involuntarily to the flash location and back to center. In the first experiment the MR signal corresponds to a voxel in the primary motor cortex, while in the second experiment it corresponds to a voxel in the frontal eye fields.

The task-related signal changes are visible in the data as periodic fluctuations in the signal. A standard analysis of these data would account for the known timing of the tasks and fit a particular form to the task-related signal changes and other sources of variation. Nonetheless, we attempted a fully nonparametric fit without any of the experimental information to examine how the adaptive-knot splines can capture highly variable functions, and in the second experiment we investigate the possibility of using semiparametric fit.

The results are shown in Figures 3 and 4. Figure 3 compares the BARS, and SARS fits in the first experiment, where the task-related changes are large with respect to the noise

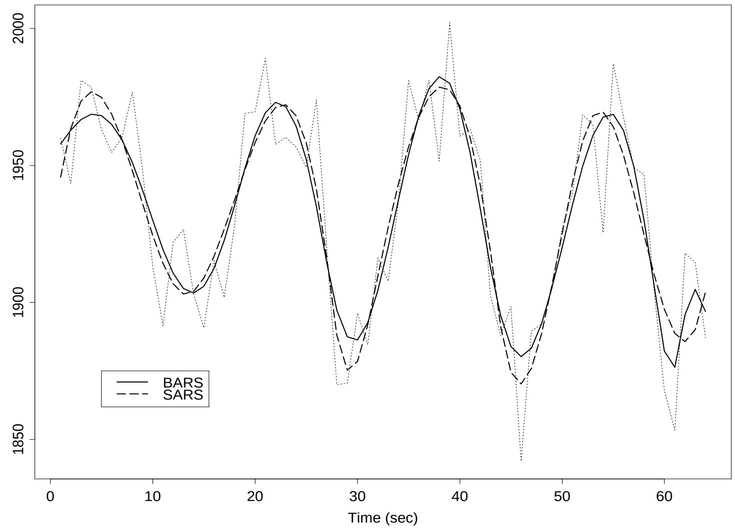

Figure 3: The time course shows the MR signal for the first experiment measured in one volume element over time in "local MR units" that depend on the scanner and pre-processing used. Superimposed on the signal (thin dotted line) are the BARS fit (solid line) and the SARS fit (– – –).

level. Both fits capture the task-related changes quite well.

Figure 4 compares the fits obtained using SARS and BARS in the second experiment, where the signal changes are smaller relative to the noise and where there is a notable nuisance signal drift. SARS detects some of the task-related signal changes but smooths over others near the drift change point. BARS, on the other hand, depending on the prior on the smoothness of the function can produce a smooth trend that averages out the individual perturbations, and it can produce a function that catches the fluctations in the data. The former case is obtained by assuming a Poisson prior with mean 3 for the number of knots while the latter is obtained by assuming a Poisson prior with mean 20. In both cases BARS estimates consistently the curve by catching all the fluctuations or smoothing them

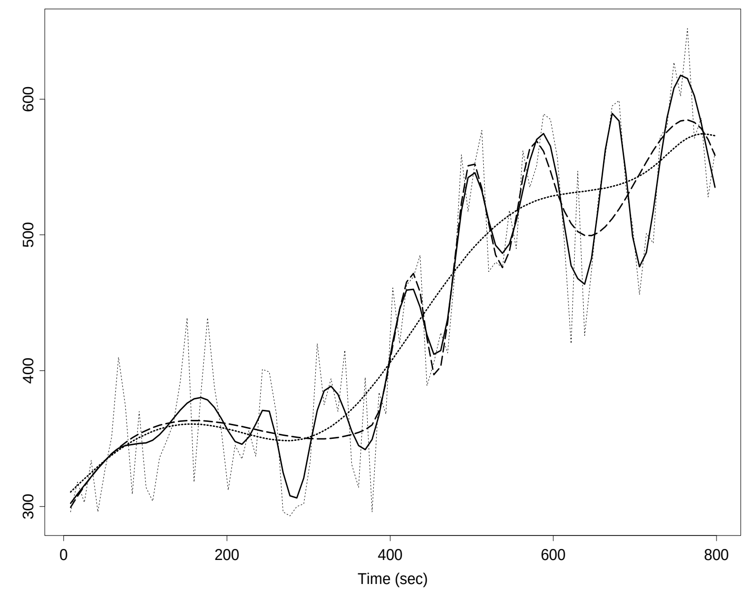

Figure 4: The time course shows the MR signal for the second experiment. Superimposed on the signal (thin dotted line) are the SARS fit (– – –), the BARS fits using a Poisson(20) prior (solid line) and a Poisson(3) prior (dotted line) on the number of knots.

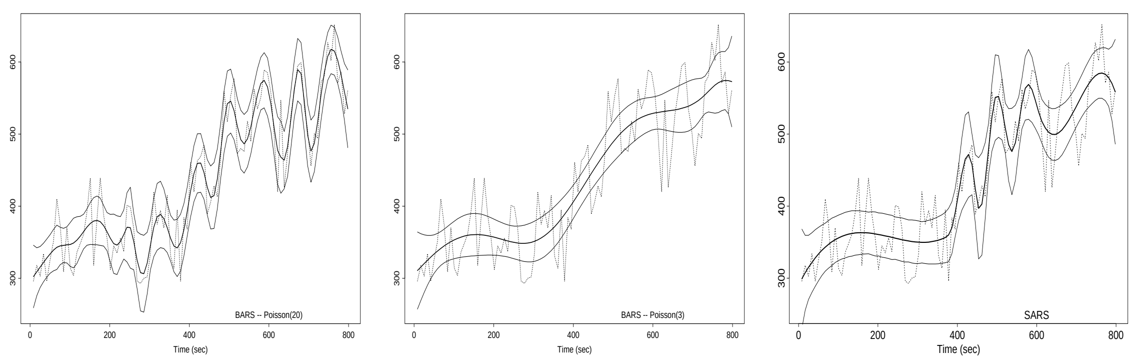

Figure 5: 95% confidence intervals (thin solid line) for the curve estimate (solid line) superimposed on the signal (thin dotted line) for the estimates of the curve using BARS with Poisson prior with mean 20 and 3 (first two figures from the left respectively) and SARS (third figure from the left).

all out. The 95% confidence bands for the fits, as shown in Figure 5, suggest that there is a non-trivial structure in the data captured by BARS, a signal drift and task related signal changes. The experimental task is the same in every corresponding block and we expect to see a task related change in every block, although it is possible that some may be missing. This suggests that SARS over smoothes some of the blocks (the confidence band for SARS is based on 1000 bootstrap samples).

This example is emblematic of the flexibility of our model to estimate a function. BARS can be used to estimate the whole function, combining signal drift and task-related fluctuations, with an assessment of uncertainty that enables comparison of experimental conditions. This automatically incorporates variation in the magnitude and shape of the task response across blocks within condition. Alternatively, BARS can estimate the signal drift changes while a more constrained parametric method can be used to estimate the task-related fluctuations. For example, it is possible to include in our model information on the experimental conditions represented by a periodic component. Figure 6 shows the semiparametric fit obtained by adding a periodic parametric component to our model. Through the back-fitting algorithm (Hastie and Tibshirani 1990) we fit an additive model in which the function is decomposed into a periodic parametric function and a smooth non-parametric function. Figure 6 shows the estimate of the curve together with the estimate of the non parametric component.

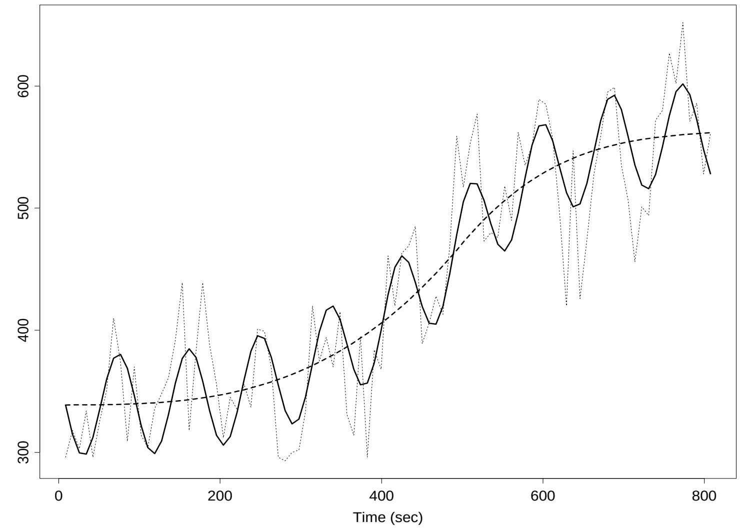

Figure 6: The time course shows the MR signal for the second experiment. Superimposed on the signal (thin dotted line) is a semiparametric fit (solid line) together with the nonparametric estimate of the linear trend (dashed line).

6 A Poisson Model for Neuron Firing Data

In a recent experiment, the firing of individual neurons in the supplementary eye field of a macaque monkey was recorded while he watched images appear on a screen in front of him (Olson, et al., 1999). The problem was to quantify the apparent tendency for some neurons to have a more delayed maximal firing rate in one experimental condition, here called the "pattern" task, compared to the other condition, called the "spatial" task. We consider data from one of the neurons, with spike counts from multiple trials binned into 10 millisecond intervals. Preliminary examination of the data indicated that an assumption that the counts in each bin were Poisson was very reasonable (and, furthermore, the effect of binning was negligible). The model we use, therefore, for the counts $\{Y_i,\ i = 1, \cdots, n\}$

at times $\{x_i,\ i = 1, \cdots, n\}$ is as follows:

$$
\begin{cases}
p(Y_i|\beta, \xi, k) = \text{Poison}(\lambda_i) \\
\quad \log(\lambda_i) = \mathbf{B}(x_i)\beta \\
\pi_\beta(\beta|\xi, k) = N(0, D) \\
p(\xi, k)
\end{cases}
\tag{11}
$$

where D is the matrix from the unit-information prior. We do not have to write down D explicitly because, as explained in Sections 2 and 3, we use the BIC-based reversible-jump MCMC scheme together with importance reweighting. As our importance function (for the posterior on β) 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 is in fact a good approximation.

The resulting fits are given together with the raw counts in Figure 7. Table 2 gives the posterior means and posterior standard deviations for the location of each maximal firing rate $g(\beta, \xi, k) = \text{argmax}_t f(t) \approx \text{argmax}_{t_i} \mathbf{B}\beta$. Also shown are the posterior means and standard deviations for the quantity of interest, the delay in milliseconds in maximal firing rate for the pattern condition compared to the spatial condition. These results were obtained by implementing the procedure of Sections 2 and 3 in S-PLUS.

The smooth estimates of the firing intensities are consistent with intuition about the slowly-varying nature of the firing rate on this time scale, and are preferable to the more wiggly kernel density estimates also shown in Figure 7. The approach quantifies the delay in maximal response as being roughly 186 ($\pm$ 100) ms. These results are consistent with those obtained by more detailed parametric modeling of the intensity function (Ventura, Carta, et al., 1999).

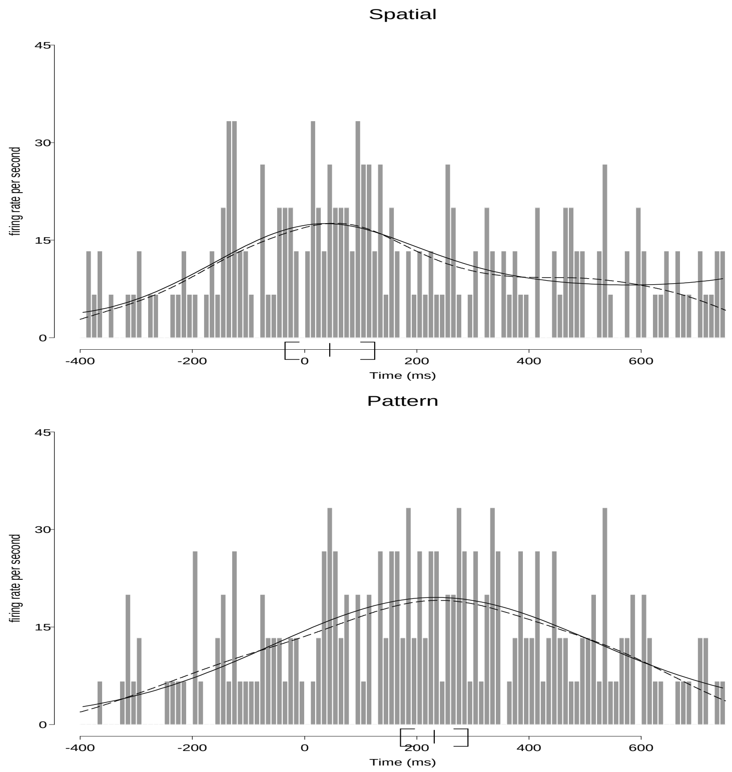

Figure 7: Peristimulus histogram in the Spatial and Pattern tasks. The solid line represent the Bayesian estimate and the dotted line is a kernel estimate. On the x-axis the posterior mean for the argmax$f(x)$ are marked with a bar together with ± 1 posterior standard deviation.

Posterior mean ($\pm$ posterior standard deviation)		
Spatial	Pattern	Delay (pattern-spatial)
45 ($\pm$80)	231 ($\pm$60)	186 ($\pm$100)

Table 2: Estimates of the $g =$argmax$_x f(x)$ for the Spatial and Pattern tasks together with the estimate of the delay in maximal firing rate. In parenthesis the posterior standard deviations are given.

7 Conclusions

In their paper, ZS reported dramatically better performance of their SARS method compared to DMS. We have shown here that the reversible-jump procedure is greatly improved by using (i) a fully Bayesian formulation and (ii) a locality heuristic in the generation of candidate knots. The simulation results, including a modified-DMS approach that incorporated (i) but not (ii), showed that both features are important. This implementation makes the Bayesian approach as good as and, in some cases, much better than, the SARS algorithm. While the test cases show that BARS can fit smooth functions with localized singularities, the fMRI example suggests that BARS can also fit functions with global variations as well.

We also showed how the method may be generalized to models outside the Normal regression framework. For the approach we articulated, one must be able to compute the MLE from the likelihood on β conditioning on (ξ, k). For general curve-fitting, the MLE would be needed to obtain the BIC values. For inferences about characteristics of the function f, the MLE would also be used in importance reweighting. Alternatives approximations (using versions of Laplace's method that expand around the posterior mode rather than the MLE, for instance), could also be introduced, but the procedure we have employed appears to be very effective.

There are many possibilities for computational improvements, including possible local updating schemes or other ideas for choosing knot locations. One might include multiple knots at the same location, to allow for discontinuities of derivatives or of the function itself. However, as our Example 3 showed, the method we have used can be effective in fitting such non-smooth functions: it allows for knots to be placed so close together that they are nearly on top of each other.

The methodology is likely to extend in various directions, including application to additive models, hierarchical models, survival models, and to spectral estimation. We intend to pursue these extensions.

Appendix A: Detailed Balance

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

$$\pi(M_k)Pr(M_{k-1}|M_k) = \pi(M_{k-1})Pr(M_k|M_{k-1}) \tag{12}$$

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

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

$Pr(M_{k-1}|M_k)$ is a Markov transition kernel: the transition probability of going from M_k to M_{k-1}. Let us define the following:

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

The set of knots in the two space differs only in the j^*-th element. We can now write the transition probabilities as follows:

$$\begin{aligned} Pr(M_{k-1}|M_k) &= \underbrace{Pr(k-1|k)}_{d_k} \cdot \underbrace{Pr(\text{delete } \xi_{j^*}|k)}_{1/k} \cdot \underbrace{(\text{acceptance probability})}_{\alpha_d} \\ &= d_k \frac{1}{k} \min\{1, A\} \end{aligned}$$

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

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

We can now show (12). If $A < 1$, then $\alpha_d = A$ and $\alpha_b = 1$, therefore rewriting (12) we have:

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

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

Appendix B: Importance Sampling

We want to determine the weight for our problem. Let $g(\beta, \xi, k)$ be the functional of interest. we then need to compute:

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

where

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

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

similarly

$$
B = \frac{\hat{p}(y)}{p(y)} \int \cdot \cdot \int \frac{p(y|\beta, \xi, k)}{\hat{p}(y|\beta, \xi, k)} \hat{q}(\beta|y, \xi, k)p(\xi, k|y)d\beta d\xi dk
$$

therefore

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

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

Reference

Denison D.G.T., Mallick B.K. and Smith A.F.M. (1998). Automatic Bayesian curve fitting. *J. R. Statist. Soc. B.* **60**, Part 2, pp 330-350.

Donoho D. and Johnstone I.M. (1995). Adaptive to Unknown Smoothness via Wavelet Shrinkage. *JASA*. **90**, pp 1200-1224.

Eilers P.H.C. and Marx B.D. (1996). Flexible Smoothing with B-splines and Penalties. *Statistical Science*. **11**, pp 89-121.

Evans, M. and Swartz, T. (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statist. Sci*, **10**, 254-272.

Genovese, C. R. (2000). A Bayesian Time-Course Model for Functional Magnetic Resonance Imaging Data. *JASA*, to appear.

Green P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika.* **82**, 4, pp 711-732.

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

Kass R.E. and Wasserman L. (1995). A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion. *JASA*. **90**, No. 431, pp 928-934.

Lindstrom M.J. (1999). Penalized Estimation of Free-Knot Splines. *Journal of Computational and Graphical Statistics.* **8**, No. 2, pp 333-352.

Olson, C.R., Gettner, S.N., Ventura, V., Carta, R. and Kass, R.E. (1999). Neuronal activity in Macaque supplementary eye field during selection of targets under exogenous and endogenous control.

Pauler D.K. (1998). The Schwarz criterion and related methods for normal linear models. *Biometrika.* **85**, 1, pp 13-27.

Smith M. and Kohn R. (1996). Nonparametric regression using Bayesian variable selection. *J. of Econometrics.* **75**, pp 317-343.

Ventura, V., Carta, R., Kass, R.E., and Olson, C. (1999). Statistical analysis of temporal evolution in single-neuron firing rates.

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

Zhou S. and Shen X. (1998). Spatially adaptive regression splines and accurate knot selection schemes. *Tech. Report, Dept. of Stat. Ohio State Univ.*