A Bayesian approach to hybrid splines non-parametric regression

Ronaldo Dias
Departamento de Estatística.
Universidade Estadual de Campinas. São Paulo, Brasil
E-mail address: `dias@ime.unicamp.br`

and

Dani Gamerman
Instituto de Matemática
Universidade Federal do Rio de Janeiro. Rio de Janeiro, Brasil
E-mail address: `dani@im.ufrj.br`

Abstract

A Bayesian approach is considered to estimate the number of basis functions and the smoothing parameter of the hybrid splines non-parametric regression procedure. The method used to obtain the estimate of the regression curve and its Bayesian confidence intervals is based on the reversible jump MCMC (Green1995). Illustrations with simulated data are provided and show good performance of the proposed approach over the existing methods.

1 Introduction

Several methods were suggested to estimate non-parametrically an unknown regression curve g by using splines since the pioneer work of Craven and Wahba (1979). Kimeldorf and Wahba (1970) and Wahba (1983) gave an attractive Bayesian interpretation for an estimate $\hat{g}$ of the unknown curve g. They showed that $\hat{g}$ can be viewed as a Bayes estimate of g with respect to a certain prior on the class of all smooth functions. The Bayesian approach allows one not only to estimate the unknown function, but also to provide error bounds by constructing the corresponding Bayesian confidence intervals (Wahba1983). In addition, Wahba noted that the coverage probabilities do not hold at each individual point, but rather are valid when averaged across the entire curve. Moreover, the true coverage could fall far short of the nominal level at points where there is unusual large local curvature. Nychka (1988) pointed out that the bias increases in those regions and the increase in bias is due to the global value of the smoothing parameter, which is appropriate on the average across all the points, but does not adapt to the local behavior of the function in regions of high curvature, where the polynomial spline with global smoothing parameter tends to over-smooth. Traditionally, there have been two techniques to address the problem of spatial adaptability. One uses local variable smoothing parameters (or bandwidths) in common smoothing methods, e.g., smoothing spline and kernel methods. In the smoothing splines techniques, in general, the number of basis functions is set to be as large as the number of observations and a smoothing parameter is chosen to control the tradeoff between adaptability to data and smoothness enforced by a penalty term. Another technique is to place knots adaptively in a regression spline method which is equivalent to adaptively choose a set of spline basis functions for regression. Abramovich and Steinberg (1996), Luo and Wahba (1997) and Dias (1999) proposed more adaptive methods to obtain good estimates of the regression curves. The h-splines method introduced by Dias (1998) in the case of non-parametric density estimation, and by Luo and Wahba (1997) and Dias (1999) in the context of non-parametric regression, combines ideas from regression splines and smoothing spline methods by finding the number of basis functions and the smoothing parameter iteratively. One advantage of the h-splines method is the ability to vary the amount of smoothing in response to the inhomogeneous curvature of the true functions at different locations.

In this work we allow two smoothing parameters as Luo and Wahba (1997) and Dias (1999) did. However, instead of going through the difficulties of specifying them precisely in an ad-hoc manner, they are allowed to vary according to prior information. In this way, the procedure becomes more capable of providing an adequate fit.

Donoho and Johnstone (1994) give a frequentist approach to the spatial adaptability problem providing an ideal adaptation procedure. They assume a particular case where the design points are equally spaced. In addition, the smoothing parameter is not obtained by data-driven methodology. In fact, their procedure represents an ideal that they cannot expect to attain but it is the target they believe should be followed. The approach used by Denison, Mallick and Smith (1998) which can be viewed as a Bayesian approach to regression splines only allows for a single smoothing parameter controlling the number of basis functions used. The extra flexibility in their approach is achieved by allowing variation in the positions of the knots. This is likely to cause instability in the fit (see Bates and Wahba (1982)) and as pointed out by Hastie and Tibshirani (1990, pages 25–27) the problem with having only the number of basis functions as the smoothing parameter is that the smoothness cannot be varied continuously as function of a single smoothing parameter. These difficulties are naturally avoided in our case through the introduction of an extra smoothing parameter, ensuring that instabilities associated with ill-posed problem will not occur. Therefore we are combining the superior advantages of h-splines method over traditional smoothing splines by taking into account prior information and we solve the problem of numerical stability by using an extra smoothing parameter and reducing the number of basis functions.

The set up for splines non-parametric regression assumes that an unknown function g of one or more variables and a set of observations are given such that:

$$y(t_i) = g(t_i) + \varepsilon_i \quad \text{for} \quad i = 1, \ldots, n,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are measurement errors usually assumed to be independent, with common distribution normal with mean zero and unknown variance σ^2.

Straightforward least square fitting is often appropriate but it produces a function which is not sufficiently smooth for some data fitting problems. In such cases, it may be better to look for a function which minimizes a criterion that involves a combination of goodness of fit and an appropriate measure of smoothness. Such criterion is the well known penalized least square problem defined as the following: Finding the minimizer of the penalized least square criterion which is,

$$L_\lambda(g) = \sum_{i=1}^{n}(y_i - g(t_i))^2 + \lambda J(g), \tag{1.1}$$

where $J(g)$ is the penalty term usually taken as $\int(g'')^2$ with $g \in \mathcal{W}_2^2 = \{g : g'$ absolutely continuous and $\int(g'')^2 < \infty\}$ and λ is the smoothing parameter which controls the trade-off between fidelity to the data and smoothness.

It is well known that the minimizer $\hat{g}$ is necessarily a natural cubic spline with knots at t_i (see, for example, Silverman and Green (1994) and Craven and Wahba (1979)). Note that the roughness penalty $\int(g'')^2$ has the property of reducing the problem of choosing g from an infinite-dimensional class of functions to a finite class of functions since $\hat{g}$ can be written as linear combination of basis functions. Although this fact might lead one to think that the non-parametric regression problem becomes a parametric problem, the number of parameters can be as large as the number of observations, and there may be difficulties in interpreting a curve or surface g. Moreover, if the number of observations is large, the system of linear equations for exact solution is too expensive to solve.

In smoothing techniques, the number of basis functions is chosen to be as large as the distinct design points and then the smoothing parameter is chosen to control the flexibility of the fitting (Bates and Wahba1982). The h-splines method (Luo and Wahba (1997), Dias (1998) and Dias (1999)) combines ideas from regression splines and smoothing splines methods by finding the number of basis functions and the smoothing parameter iteratively. By taking the penalty term $J(g)$ as $\int (g'')^2$, $\mathbf{y} = (y_1, \ldots, y_n)^T$ and $\mathbf{g} = (g(t_1), \ldots, g(t_n))^T$, the penalized least square criterion (1.1) becomes,

$$L_\lambda(g) = ||\mathbf{y} - \mathbf{g}||^2 + \lambda \int (g'')^2, \tag{1.2}$$

Assume that $g \approx g_{K,\theta} = \sum_{i=1}^{K} \theta_i B_i = X_K \theta$ so that $g_{K,\theta} \in \mathcal{H}_K$, where $\mathcal{H}_K$ denotes the space of natural cubic splines (NCS) spanned by the basis functions $\{B_i\}_{i=1}^K$ and X_K is a $n \times K$ matrix with entries $(X_K)_{\{i,j\}} = B_i(t_j)$, for $i = 1, \ldots, K$ and $j = 1, \ldots, n$. Then, the numerical problem is to find a vector $\theta = (\theta_1, \ldots, \theta_K)^T$ that minimizes,

$$L_\lambda^*(\theta) = \|\mathbf{y} - X_K \theta\|^2 + \lambda \theta^T \Omega \theta, \tag{1.3}$$

where Ω is $K \times K$ matrix with entries $\Omega_{ij} = \int B_i''(t) B_j''(t) dt$. Standard calculations de Boor (1978) provide θ as a solution of the following linear system $(X^T X + \lambda\Omega)\theta_\lambda = X^T y$. Note that the linear system now involves $K \times K$ matrices instead of using $n \times n$ matrices which is the case of smoothing splines. Both K and λ controls the trade off between smoothness and fidelity to the data. In particular, when $\lambda = 0$ we have the regression spline case, where K is the parameter that controls the flexibility of the fitting. To exemplify the action of K on the estimated curve, let us consider an example by simulation with $y(t) = \exp(-t)\sin(\pi t/2)\cos(\pi t) + \varepsilon$ with $\varepsilon \sim N(0, .05)$. The curve estimates were obtained by least square method with four different numbers of basis functions which are the cubic B-splines.

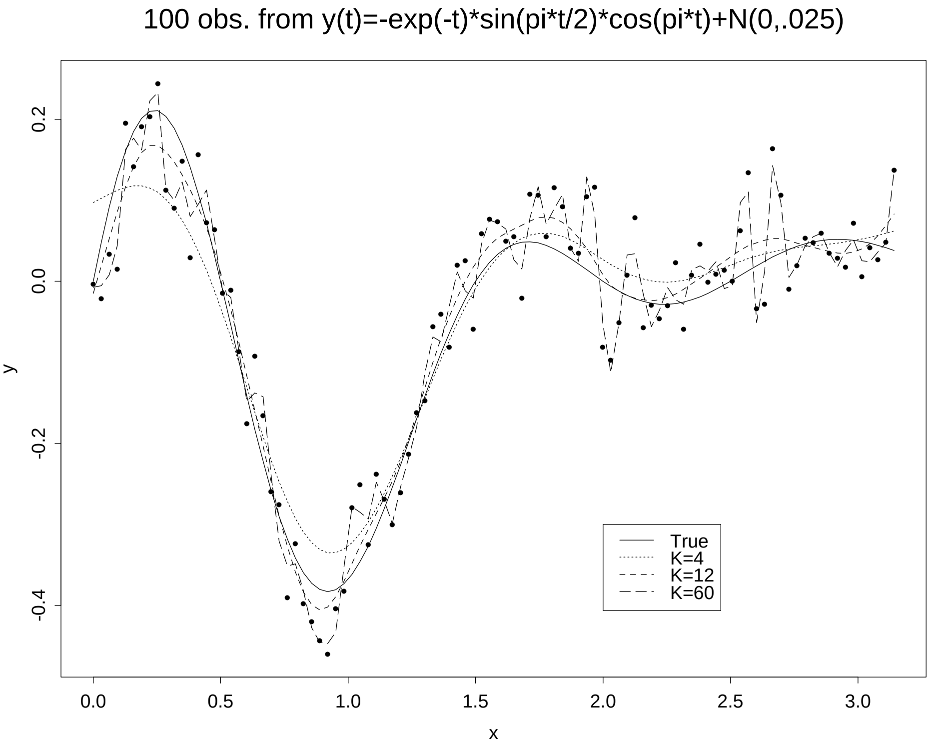

Figure 1.1: Spline least square fittings for different values of K

Figure 1.1 shows the effect of varying the number of basis functions on the estimation of the true curve. Note that the number of basis functions is the same as the number of knots since it is assumed that we are dealing with natural cubic splines space. Observe that small values of K make smoother the estimate and hence over smoothing may occur. Large values of K may cause under-smoothing. In section 2 we describe the set up to a Bayesian approach to the h-splines non-parametric regression. In section 3 we present some illustrations with simulated data and we compare the performance of the proposed approach over the existing methods. In section 4 we summarize our work with our concluding remarks.

2 Bayesian approach

Suppose we have the following regression model,

$$y_i = g(t_i) + \varepsilon_i \quad i = 1, \ldots, n.$$

where ε_i's are uncorrelated with a $N(0, \sigma^2)$. Moreover, assume that the parametric form of the regression curve g is unknown. Then the likelihood of g given the observations $\mathbf{y}$ is,

$$l_{\mathbf{y}}(g) \propto (\sigma^2)^{-n/2} \exp\{-\frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{g}||^2\}. \tag{2.1}$$

The Bayesian justification of penalized maximum likelihood is to place a prior density proportional to $\exp\{-\frac{\lambda}{2}\int(g'')^2\}$ over the space of all smooth functions. One may see that with this prior the optimization problems (1.2) and (2.1) are equivalent, (see details in Silverman and Green (1994) and Kimeldorf and Wahba (1970)). However, an infinite dimensional case has a paradox alluded to by Wahba (1983). Silverman (1985) proposed a finite dimensional Bayesian formulation to avoid the paradoxes and difficulties involved in the infinite dimensional case. For this, let $g_{K,\theta} = \sum_{i=1}^{K} \theta_i B_i^x = X_K \theta_K$ with a knot sequence x placed at order statistics. A complete Bayesian approach would assign prior distribution to the coefficients of the expansion, to the knot positions, to the number of knots and for σ^2. Here, we are suggesting a completely different procedure. The h-splines method combines some features of regression splines and of traditional smoothing splines to obtain a hybrid smoothing procedure which is usually implemented with large data sets and displays a desirable form of spatial adaptability when the underlying function is spatially inhomogeneous in its degree of complexity. Basically, choosing the number of basis functions, for instance by GCV criterion, will do most of the work for balancing between bias and variance. But there is a more important reason why we want to do a penalized regression, namely numerical stability. It is well known that as the number of basis functions increases, the regression problem becomes more ill-conditioned, which makes the numerical computation less stable. The basis functions used, in general, are the cubic spline basis which have larger correlations among them than linear spline basis, hence the ill-conditioning problem is more serious. The penalized regression step acts as a remedy for this.

A Bayesian approach to hybrid splines non-parametric regression assigns priors for g_K, K, λ and σ^2. Given a realization of K the interior knots are placed at order statistics. This well known procedure in non-parametric regression reduces the computational cost substantially and avoids trying to solve a difficult problem of optimizing the knot positions. Any other procedure has to take into account the fact that changes in the knot positions might cause considerable change in the function g (see details in Wahba (1982) for ill-posed problems in splines non-parametric regression). Moreover, in theory the number of basis functions (which is a linear function of the number of knots) can be as large as the sample size. But then one has to solve a system of n equations instead of K. An attempt to keep the computational cost down one might want to have K small as possible and hence over-smoothing may occur. For any K large enough λ keeps the balance between over-smoothing and under-smoothing. Thus the penalized likelihood is (1.1) with $g = g_K$, where $g_K = g_{K,\theta} = X_K\theta_K$. The reason why we suppress the subindex θ in $g_{K,\theta}$ will be explained later in this section. Note that maximization of the penalized likelihood is equivalent to minimization of (1.3). For this proposed Bayesian set up we have a prior for $(g_K, K, \lambda, \sigma^2)$ of the form,

$$
\begin{aligned}
p(g_K, K, \lambda, \sigma^2) &= p(g_K|K, \lambda)p(K, \lambda)p(\sigma^2) \qquad (2.2)\\
&= p(g_K|K, \lambda)p(\lambda|K)p(K)p(\sigma^2),
\end{aligned}
$$

where

$$
p(g_K|K, \lambda) \propto \exp\{-\frac{\lambda}{2}\int (g_K'')^2\},
$$

$$
p(\sigma^2) \propto \frac{1}{(\sigma^2)^{(u+1)}} \exp\{-v\sigma^2\},
$$

for $u > 0, v > 0$,

$$
p(K) = \frac{\exp\{-a\}a^K/K!}{1-\exp\{-a\}(1+q^*)}, \quad K = 1, \ldots, K^*
$$

where $q^* = \sum_{j=K^*+1}^{\infty} a^j/j!$, and

$$
p(\lambda|K) = \psi(K)\exp\{-\psi(K)\lambda\},
$$

with ψ any smooth function of K. It is well known that, for $\lambda > 0$, when the number of basis functions increases the smoothing parameter decreases to a point and then it increases with K. That is, for large values of K the smoothing parameter λ becomes larger to enforce smoothness. (See details in Dias (1999).). Therefore, functions ψ

that satisfy these requirements are recommended. In particular, a flexible class is given by $\psi(K) = K^{-b}\exp(-cK)$ for suitably chosen hyperparameters b and c. However, undesirably large values of K can be excluded through fixing K^* appropriately or be made unlikely by fixing a accordingly. These large values of K can be controlled by the hyperparameters a and K^* of the prior $p(K)$.

The choice of a is governed by the prior expectation of the structure of the underlying curve such as maxima, minima, inflection points etc. We suggest the reader to follow some of the rules recommended by Wegman and Wright (1983). These recommendations are based on the assumption of fitting a cubic spline, the most popular case and are summarized below.

1. Extrema should be center in intervals and inflection points should be located near knot points.

2. No more than one extremum and one inflection point should fall between knots (because a cubic could not fit more).

3. Knot points should be located at data points.

Note that $g_K = X_K\theta_K$ is completely determined if we know the coefficients θ_K. Hence, the overall parameter space Ξ can be written as countable union of subspaces $\Xi_K = \{\xi : \xi = (\theta_K, \phi) \in (\mathbb{R}^K \times \{1, \ldots, K^*\} \times [0, \infty]^2)\}$ with $\phi = (K, \lambda, \sigma^2)$. Thus, the posterior is given by

$$\pi(\xi|\mathbf{y}) \propto l_p(\xi|\mathbf{y})p(\xi). \tag{2.3}$$

In order to sample from the posterior $\pi(\xi|\mathbf{y})$ we have to consider the variation of dimensionality of this problem. Hence one has to design move types between subspaces Ξ_K. However, assigning a prior to g_K, equivalently to the coefficients θ_K, leads to a serious computational difficulty pointed out by Denison et al. (1998) where a comparative study was developed. They suggest that the least square estimates for the vector θ_K leads to a non-significant deterioration in performance for overall curve estimation. Similarly, given a realization of (K, λ, σ^2), we solve the penalized least square objective function (1.3) to obtain the estimates, $\hat{\theta}_K = \hat{\theta}_K(\mathbf{y})$, for the vector θ_K and consequently we have an estimate $\hat{g}_K = X_K\hat{\theta}_K$. Thus, there is no prior assigned for this vector of parameters, and so, we write $g_K = g_{K,\theta}$. Having got $\hat{\theta}_K$, we approximate the marginal posterior $\pi(\phi|\mathbf{y})$ by the conditional posterior

$$\pi(\phi|\mathbf{y}, \hat{\theta}_K) \propto l_p(\phi|\mathbf{y}, \hat{\theta}_K)p(\phi), \tag{2.4}$$

with

$$p(\phi) = p(K, \lambda, \sigma^2) = p(\lambda|K)p(K)p(\sigma^2).$$

Note that if one assigns independent normal distributions to the parameters θ_K it will not be difficult to obtain the marginal posterior $\pi(\phi|\mathbf{y})$ and approximations will not be necessary. However, the results will be very similar.

We used reversible jump methodology Green (1995) to sample from $\pi(\phi|\mathbf{y}, \hat{\theta}_K)$. Let the transitions be,

(I) movement of the smoothing parameter λ,

(II) addition of a basis function and

(III) deletion of a basis function.

Note that only the steps (II) and (III) change the dimension of the model. At each transition, an independent random choice is made between each of the three move types. These have probabilities η_K for step (I), b_K for step (II) and d_K for step (III). Naturally, $b_K + d_K + \eta_K = 1$ for all $K = 1, \ldots, K^*$ with $b_0 = 1$ $d_0 = 0$, $\eta_0 = 0$ and $b_{K^*} = 0$. These probabilities were chosen so that

$$b_K = c \min\{1, p(K+1)/p(K)\}, \quad d_{K+1} = c \min\{1, p(K)/p(K+1)\}$$

and so, $\eta_K = 1 - (b_K + d_K)$. In the simulation we set $c = 0.4$, but other values are valid. Following the notation of Green (1995) we define the acceptance probability as

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

The move type (I) is on moving from λ to λ' according to the proposal density $\mathcal{U}[0, c]$ where c is a tuning parameter which depends on the range of the observations. Then the acceptance probability is given by

$$\alpha(\lambda, \lambda') = \min\Big\{1, \exp\{-\frac{\theta_K^T \Omega \theta_K}{2}(\lambda' - \lambda)\} \times \frac{p(\lambda'|K)}{p(\lambda|K)}\Big\}.$$

Note that the proposal ratio equals to 1 since we sample λ' from the same distribution of λ.

The step (II) changes the dimension of the parameter space. That is, $K \to K+1$ and the new basis are chosen from a proposal density which is the $K+1$-order statistics. Specifically, let $\mathbf{x} = (x_1, \ldots, x_K)$ be the vector of interior knots such that $x_1 < \ldots < x_K$. Then the joint density of $\mathbf{x}$ is $p(\mathbf{x}|x_1 < \ldots < x_K) \propto (K!)^{-1}$. The acceptance probability is

$$\begin{aligned}\alpha(K, K+1) &= \min\Big\{1, \exp\{-(1/2\sigma^2)^{n/2}||\mathbf{y} - X_K\theta_{K+1}|| - ||\mathbf{y} - X_K\theta_K||\} \\ &\quad \times \frac{p(K+1)}{p(K)} \frac{d_K}{b_K} \frac{K}{K+1}\Big\}.\end{aligned}$$

The acceptance probability for the corresponding deletion of a basis function has the same form with the appropriate changes and the ratio terms inverted.

In order to produce an estimate based on the proposed method we develop the following algorithm to summarizing the description above.

Algorithm 2.1

1. *Initialize by setting the hyper parameters for the prior* $p(\phi)$.
2. *Generate* u *from a uniform* $[0, 1]$.
3. *Go to the move type*

a) *if* $u \leq b_K$ *then go to the addition step (II);*

b) *else if* $b_K \leq u \leq b_K + d_K$ *then go to deletion step (III);*

c) *else go to step (I).*

4. *Draw* σ^2 *by Gibbs step.*
5. *Return to step (2) for a new draw.*

The algorithm was performed in several different types of test functions. Examples of the typical sets of test functions used in our simulation are: $\exp\{-t^2/2\}\cos(4\pi t)$, for $t \in [0, \pi]$, $t + 2\exp(-16t^2)$, for $t \in [-2, 2]$. Following, Gamerman (1997), we start with a single long chain and convergence was ascertained by considering small ($K \leq 12$) and large curvatures ($K > 12$). For functions with small curvature we have the first 1000 sample curves as the burn-in step. As it was expected, more structured functions (larger curvature) required many more iterations for their burn-in step and then we decided to have 5000 sample curves. After the burn-in step we took the next 100 sample curves as coming from the posterior distribution. The final curve estimate is the average of those 100 sample curves.

3 Simulation

In this section we present several typical examples of this method. In all examples, instead of inverse gamma for σ^2, we used a vague but proper Gamma prior for σ^{-2}, namely, $Gamma(10^{-3}, 10^{-3})$. The computer program is based on the algorithm 2.1.

In figures 3.1, 3.4, we present some simulated examples to verify the estimates provided by this approach and in Figures 3.2, 3.5, are the histograms of the marginal posteriors of the smoothing parameters and the mean square error produced by the 100 curve estimates. The final estimate is,

$$\hat{y}_+(t_i) = \frac{1}{100}\sum_{j=1}^{100} \hat{y}_j(t_i).$$

Figures 3.3, 3.6, exhibit approximate Bayesian confidence intervals for the true curve regression f and they were computed as following. Let $y(t_i)$ and $\hat{y}(t_i)$ be a particular model and its estimate provided by this proposed method with $i = 1, \ldots, n$ where n is the sample size. For each $i = 1, \ldots, n$ the fitting vectors $(\hat{y}_1(t_i), \hat{y}_2(t_i), \ldots, \hat{y}_{100}(t_i))^T$ form random samples and from those vectors the lower and upper quantiles were computed in order to obtain the confidence intervals.

Figure 3.1 exhibits an example of how useful a Bayesian approach to hybrid splines non-parametric regression can be. It describes a situation where a prior information tells that the underlying curve has large curvature and the variance of the error measurements is not too small and the traditional methods of smoothing, e.g. smoothing splines, might not be able to capture all the structure of the true regression curve. By using vague but proper priors to the smoothing parameters K and λ and for the variance σ^2 this Bayesian approach provides a much better fitting than the traditional smooth splines approach does.

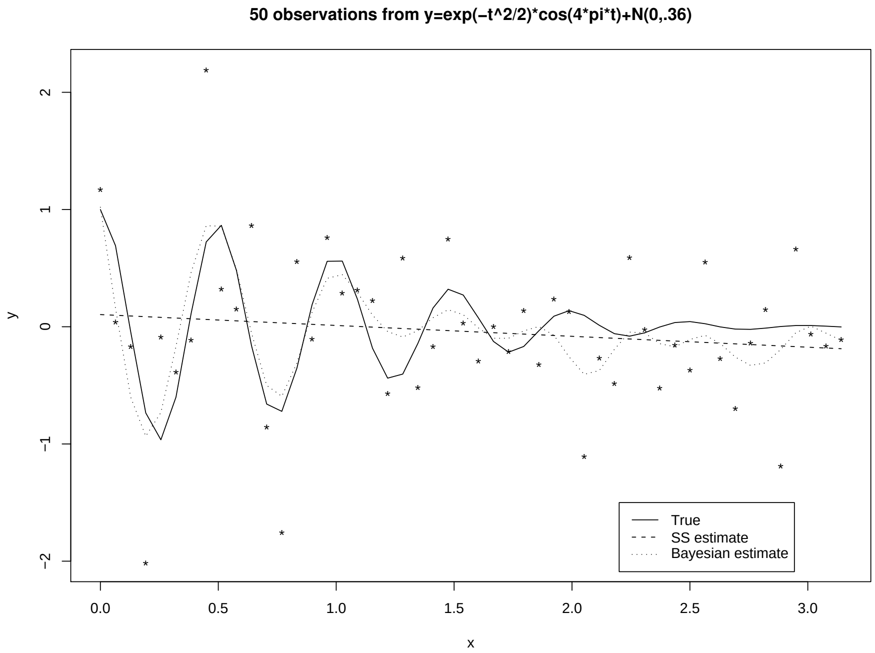

Figure 3.1: Estimation results: a) Bayesian estimate with $a = 17$ and $\psi(K) = K^3$ (dotted line); b) (SS) smoothing splines estimate (dashed line). The true regression function is also plotted (solid line). The SS estimate was computed using the R function `smooth.spline` from which 4 degrees of freedom were obtained and λ was computed by GCV.

Figure 3.2 shows histograms suggesting that the marginal posteriors are unimodal and for K we have a symmetric distribution around 19 (prior mean is 17). The histogram of the posterior distribution of the smoothing parameter λ indicates that its mode is 0.0001 (prior mean 0.0002). The relationship between K and λ has been previously studied in a frequentist setting by Dias (1999). Here, we observe in figure 3.2 (c) a confirmation in the posterior of the near log-linear relation between K and λ. This relationship was specified in the prior, providing further reassurance that this was a sensible choice.

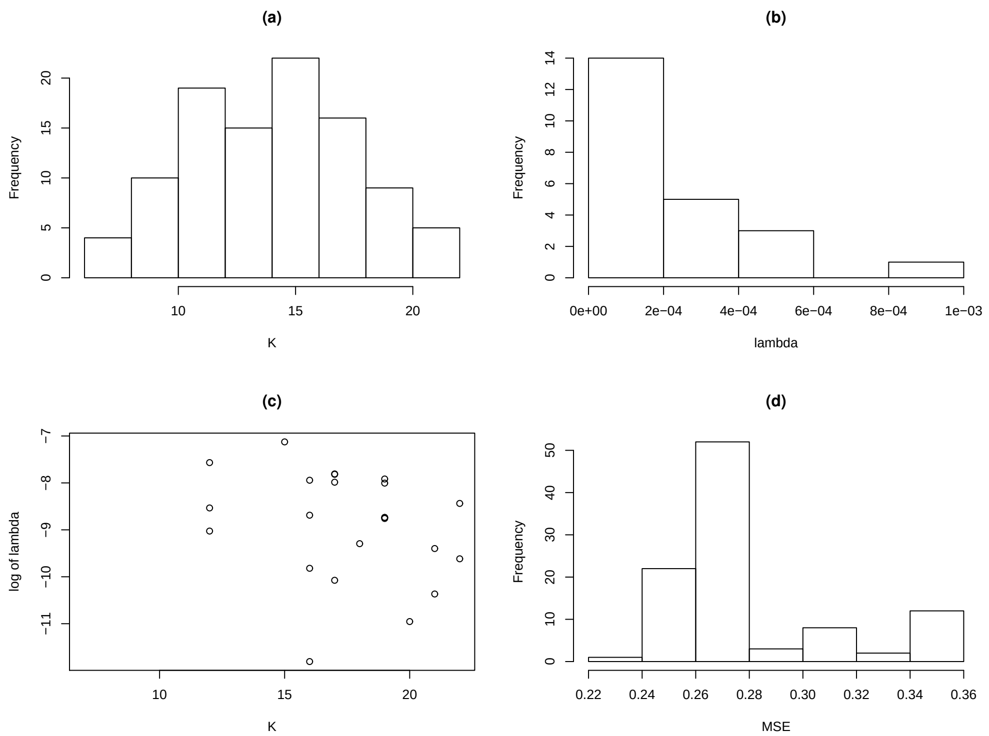

Figure 3.2: Histograms of the samples from the marginal posteriors of the smoothing parameters: (a)-K; (b)-λ; (c)- λ (in log scale) $\times K$; (d)- mean square error of the 100 curve estimates of the function $\exp(-t^2/2)\cos(4\pi t)$with $t \in [0, \pi]$.

Figure 3.3 shows one hundred curves sampling from the posterior (after burn-in) and approximate 95% Bayesian confidence interval for the regression curve $g(t) = \exp(-t^2/2)\cos(4\pi t)$ with $t \in [0, \pi]$. On the right panel of this figure we see the curve estimate which is an approximation for the posterior mean and the percentiles curves 2.5% and 97.5%.

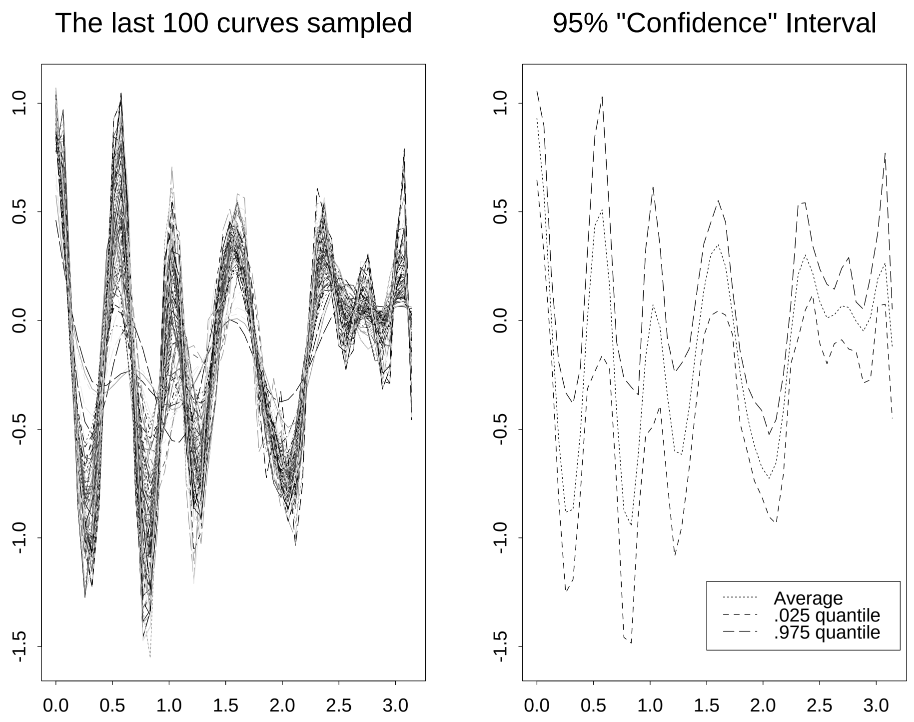

Figure 3.3: One hundred estimates of the curve 3.1 and a Bayesian confidence interval for the regression curve $g(t) = \exp(-t^2/2)\cos(4\pi t)$ with $t \in [0, \pi]$.

Among several other examples, the next two test functions were taken from Denison et al. (1998) and a comparison with traditional smoothing splines estimate is shown. One may see that in situations where the underlying regression curve has high curvature but is not spatially inhomogeneous in its degree of complexity this Bayesian approach also provides good estimates for the true regression curve.

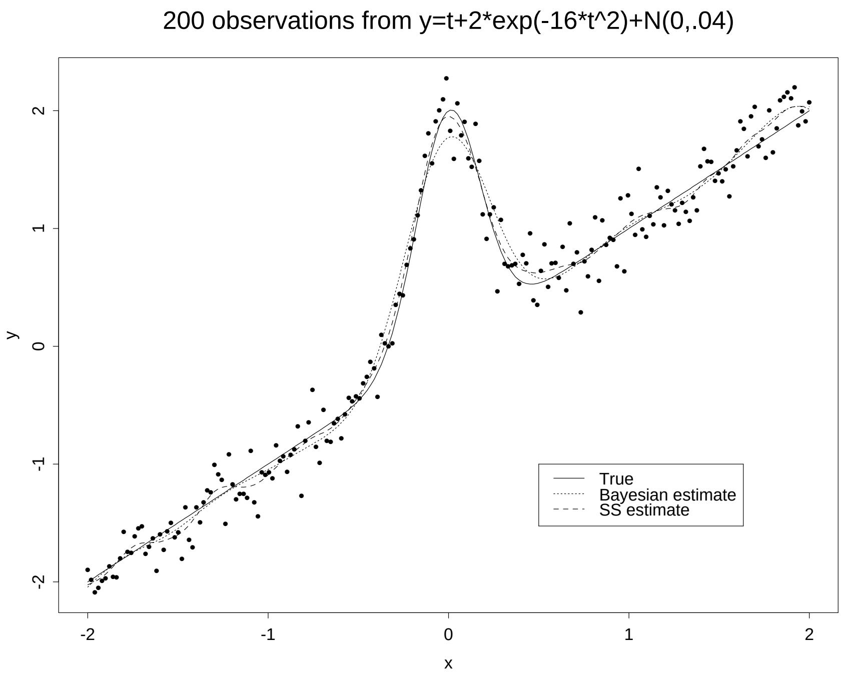

Figure 3.4: Estimation results: (a) Bayesian estimate with $a = 10$ and $\psi(K) = K^2$ (dotted); (b) (SS) smoothing splines estimate (dashed). The true regression function is also plotted (solid line). The SS estimate was computed using the R function `smooth.spline` and λ was obtained by GCV.

In this particular example in figure 3.4, one may observe that this procedure produces a better estimate than smoothing splines does even with such vague priors.

The histograms shown in figure 3.5 also indicate a symmetric single mode distribution for the posterior distribution of the parameter K and a single mode distribution for the smoothing parameter λ. Note that from the histogram of the posterior of K the value of the mode and the mean is 11 (prior mean is 10) and for λ we have mode equals to 0.0005 (prior mean is 0.01).

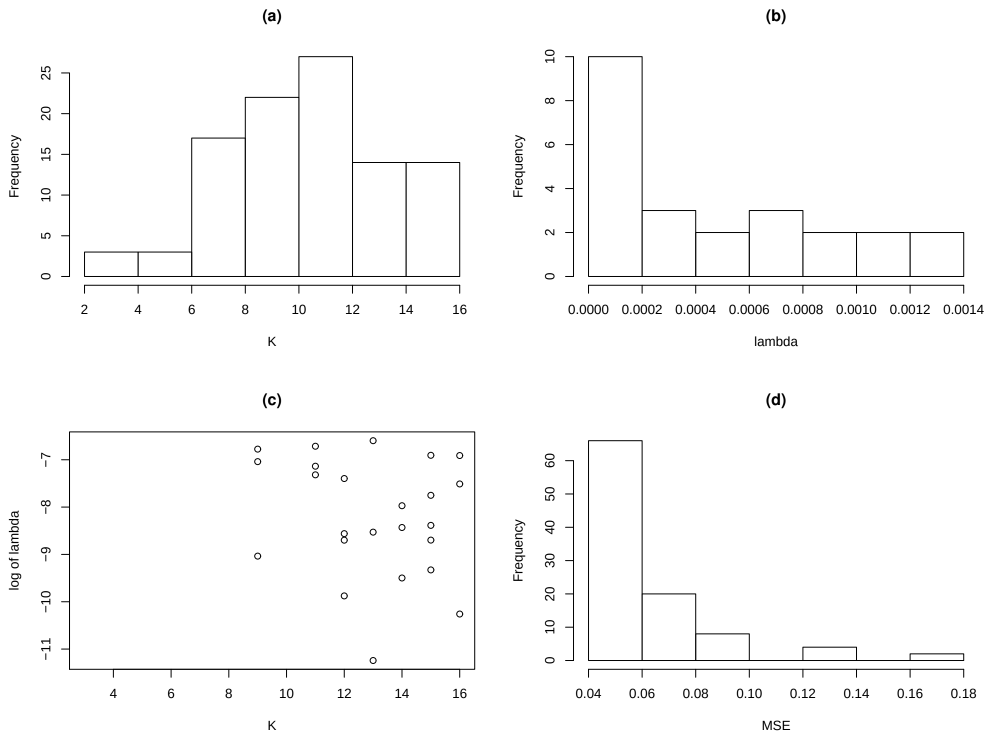

Figure 3.5: Histograms of the samples from the marginal posteriors of the smoothing parameters: (a)-K; (b)-λ; (c)- λ (in log scale) $\times K$; (d)- mean square error of the 100 curve estimates of the function $t + 2\exp(-16t^2)$.

Figure 3.6 shows one hundred curves sampling from the posterior (after burn-in) and approximate 95% Bayesian confidence interval for the regression curve $g(t) = t + 2\exp(-16t^2)$. On the right corner of this figure we see the curve estimate which is an approximation for the posterior mean and the percentiles curves 2.5% and 97.5%.

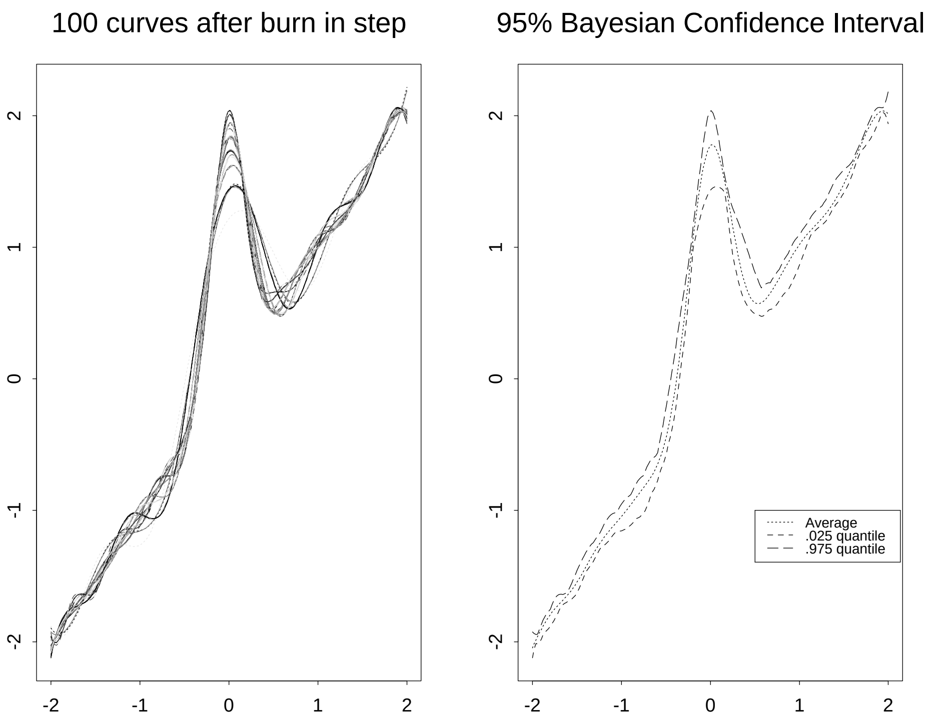

Figure 3.6: One hundred estimates of the curve 3.4 and a Bayesian Confidence interval for the true regression curve.

An investigation was carried out here to evaluate the impact of the specification of the function ψ in the estimation results. Different choices of ψ were made and the corresponding residual sum of squares are given in Table 1. The estimated curves are very similar across the choices of ψ we made. The main reason for the difference in the values of the residual sum of squares is provided by the fit near to the end points.

Table 1. Residual sum of squares computed for each form of ψ.

ψ	K	$K + \log K$	K^2	K^3
RSS	0.44	0.43	0.33	0.38

4 Conclusion

The Bayesian approach to h-splines non-parametric regression is very adaptive, easy to understand and fast to compute. Our simulations showed that a good guess for the hyper-parameters for the priors of K and $(\lambda|K)$ makes the procedure amazingly fast for this type of MCMC algorithm. As an example, the curve estimate in Figure 3.1 took, in average, less than 3 minutes using SUN ultra-2 (200MHz and 64MB) of the University of California, Berkeley. In practice, we suggest 3 basis functions for each bump in the regression curve. If one knows this information a priori the computation can be reduced substantially. Otherwise, one may start the procedure with one basis function. An example of having the suggested prior based on the number of bumps, for the same regression curve we have got an estimate in half the time when we started with one basis function. For other more complex functions it might take longer than the time mentioned here. The natural way of thinking of future work for this procedure is to extend it to high dimensions (surfaces) and the penalized likelihood method for density estimation problems.

ACKNOWLEDGMENT

We would like to thank the Associate Editor and the referees for the comments and suggestions that made this work better and clearer. Also, the first author would like to thank Prof. Michael Newton and Prof. Mary Lindstrom for all suggestion made during his visit to department of Statistic of University Wisconsin, Madison;

FAPESP grant 95/4996-3, FAEP 259/98 and CNPq grant 450560/98. The second author has benefited from grants from FAPERJ, CNPq and Ministério da Ciência e Tecnologia (PRONEX). Part of this work was done while the first author was visiting scholar at Department of Statistics, University of California, Berkeley with financial support by FAPESP grant no. 99/00261-0 and CNPq grant no. 300644/94-9.

References

Abramovich, F. and Steinberg, D. M. (1996). Improved inference in nonparametric regression using L_k-smoothing splines, *J. Statist. Plann. Inference* **49**(3): 327–341.

Bates, D. and Wahba, G. (1982). *Computational Methods for Generalized Cross-Validation with large data sets*, Academic Press, London.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions, *Numerische Mathematik* **31**: 377–403.

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

Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998). Automatic bayesian curve fitting, *Journal of the Royal Statistical Society B* **60**: 363–377.

Dias, R. (1998). Density estimation via hybrid splines, *Journal of Statistical Computation and Simulation* **60**: 277–294.

Dias, R. (1999). Sequential adaptive non parametric regression via h-splines, *Communications in Statistics: Computations and Simulations* **28**: 501–515.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage, *Biometrika* **81**(3): 425–455.

Gamerman, D. (1997). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, Chapman and Hall: London.

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

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

Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines, *The Annals of Mathematical Statistics* **41**: 495–502.

Luo, Z. and Wahba, G. (1997). Hybrid adaptive splines, *Journal of the American Statistical Association* **92**: 107–116.

Nychka, D. (1988). Bayesian confidence intervals for smoothing splines, *J. Amer. Statist. Assoc.* **83**(404): 1134–1143.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting, *Journal of the Royal Statistical Society, Series B, Methodological* **47**: 1–21.

Silverman, B. W. and Green, P. J. (1994). *Nonparametric Regression and Generalized Linear Models*, Chapman and Hall (London).

Wahba, G. (1982). Constrained regularization for ill posed linear operator equations, with applications in meteorology and medicine, *in* S. S. Gupta and J. O. Berger (eds), *Statistical Decision Theory and Related Topics III, in two volumes*, Vol. 2, Academic:NY:Lnd, pp. 383–418.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline, *JRSS-B, Methodological* **45**: 133–150.

Wegman, E., J. and Wright, I. W. (1983). Splines in statistics, *Journal of the American Statistical Association* **78**: 351–365.