

# Spatially Adaptive Bayesian Penalized Regression Splines (P-splines)

Veerabhadran BALADANDAYUTHAPANI,  
Bani K. MALLICK, and Raymond J. CARROLL

In this article we study penalized regression splines (P-splines), which are low-order basis splines with a penalty to avoid undersmoothing. Such P-splines are typically not spatially adaptive, and hence can have trouble when functions are varying rapidly. Our approach is to model the penalty parameter inherent in the P-spline method as a heteroscedastic regression function. We develop a full Bayesian hierarchical structure to do this and use Markov chain Monte Carlo techniques for drawing random samples from the posterior for inference. The advantage of using a Bayesian approach to P-splines is that it allows for simultaneous estimation of the smooth functions and the underlying penalty curve in addition to providing uncertainty intervals of the estimated curve. The Bayesian credible intervals obtained for the estimated curve are shown to have pointwise coverage probabilities close to nominal. The method is extended to additive models with simultaneous spline-based penalty functions for the unknown functions. In simulations, the approach achieves very competitive performance with the current best frequentist P-spline method in terms of frequentist mean squared error and coverage probabilities of the credible intervals, and performs better than some of the other Bayesian methods.

**Key Words:** Additive models; Bayesian methods; Locally adaptive smoothing parameters; Markov chain Monte Carlo; Mixed models; Variance models.

## 1. INTRODUCTION

Regression splines are approximations to functions typically using a low-order number of basis functions. Such splines, like all splines, are subject to a lack of smoothness and various strategies have been proposed to attain this smoothness. A particularly appealing class are the regression P-splines (Eilers and Marx 1996), which achieve smoothness by penalizing the sum of squares or likelihood by a single penalty parameter. The penalty

---

Veerabhadran Baladandayuthapani is a Graduate Student, Bani K. Mallick is Professor, and Raymond J. Carroll is Distinguished Professor, Department of Statistics, Texas A&M University, College Station, TX 77843-3143 (E-mail: addresses: veera@stat.tamu.edu; and bmallick@stat.tamu.edu; carroll@stat.tamu.edu).

©2005 American Statistical Association, Institute of Mathematical Statistics,  
and Interface Foundation of North America

*Journal of Computational and Graphical Statistics*, Volume 14, Number 2, Pages 378–394

DOI: 10.1198/106186005X47345

parameter and the fit using P-splines are easy to compute using mixed model technology (see Robinson 1991; Coull, Ruppert, and Wand 2001; Rice and Wu 2001, among others), and are not sensitive to knot parameter selection (Ruppert 2002).

Despite these advantages, P-splines with a single penalty parameter are not suitable for spatially adaptive functions that can oscillate rapidly in some regions and are rather smooth in other regions (Wand 2000). Rather than using a global penalty parameter, Ruppert and Carroll (2000) proposed a local penalty method wherein the penalty is allowed to vary spatially so as to adapt to the spatial heterogeneity in the regression function. The Web site <http://orie.cornell.edu/~davidr> contains the MATLAB code for computing this spatially adaptive estimator.

The purpose of this article is to construct a Bayesian version of the local penalty method. We do this by modeling the penalty as another regression P-spline, in effect a variance function, in a hierarchical structure. The method is relatively simple to compute and implement, and the MATLAB code for it is given at <http://www.stat.tamu.edu/~veera>. The advantage of using a Bayesian approach to P-splines is that it allows for simultaneous estimation of the function and the underlying penalty curve in addition to providing uncertainty intervals for the estimated curve. We show that our method achieves competitive performance with that of Ruppert and Carroll (2000) in terms of *frequentist* mean squared error and coverage probabilities of credible intervals. The Bayesian credible intervals obtained for the estimated curve are shown to have pointwise frequentist coverage probabilities close to nominal. In simulations our method outperforms, sometimes substantially, many other Bayesian methods existing in literature.

The article is structured as follows: Section 2 introduces the Bayesian model used, along with the prior and distributional assumptions on the random variables and parameters. Section 3 is devoted to the MCMC setup for the calculations. Section 4 discusses the simulation study undertaken and the results of our findings. We extend the univariate ideas to additive models in Section 5. Technical details are collected into an Appendix.

## 2. MODEL FORMULATION

Given data  $(X_i, Y_i)$ , where  $X_i$  is univariate, our nonparametric model is defined by

$$Y_i = m(X_i) + \epsilon_i,$$

where  $m(\bullet)$  is an unknown function, the  $\epsilon_i$ 's are independent conditional on  $X_i$  and normally distributed with mean zero and variance  $\sigma_Y^2$ .

To estimate  $m(\bullet)$  we use regression P-splines. As the basis functions, here we use piecewise polynomial functions whose highest order derivative takes jumps at fixed "knots." Other basis functions such as B-splines (De Boor 1978) could also be used. With this basis, the functional form of the regression spline of degree  $p \geq 1$  is given by

$$m(X) = \alpha_{Y_0} + \alpha_{Y_1}X + \cdots + \alpha_{Y_p}X^p + \sum_{j=1}^{M_Y} \beta_{Y_j}(X - \kappa_{Y_j})_+^p,$$

where  $(\alpha_{Y0}, \dots, \alpha_{Yp}, \beta_{Y1}, \dots, \beta_{YM_Y})$  is a vector of regression coefficients and  $(a)_+^p = a^p I(a \geq 0)$ , and  $\kappa_{Y1} < \dots < \kappa_{YM_Y}$  are fixed knots.

To model the unknown smooth function  $m(\bullet)$ , we illustrate the theory using regression splines of degree 1, so that

$$m(X) = \alpha_{Y0} + \alpha_{Y1}X + \sum_{j=1}^{M_Y} \beta_{Yj}(X - \kappa_{Yj})_+. \tag{2.1}$$

Of course, changes to polynomials of higher degree are trivial. We take  $M_Y$ , the number of knots, to be large but much less than  $n$ , the number of data points. Unlike knot-selection techniques we retain all candidate knots. In this article we take the knots to be the equally spaced sample quantiles of  $X$ , although one could just as easily take the knots to be equally spaced.

The number of knots here is specified by the user. Although the choice is not crucial (Ruppert 2002), a minimum number of knots are needed to capture the spatial variability in the data. The choice of knots is discussed in detail later in the article (Section 4.2 and Section 6).

We can interpret (2.1) as a Bayesian linear model. Rewrite (2.1) as

$$Y = Z_Y \Omega_Y + \epsilon_Y, \tag{2.2}$$

where  $Y_{n \times 1} = (Y_1, \dots, Y_n)^T$ ,  $\Omega_Y = (\alpha_{Y0}, \alpha_{Y1}, \beta_{Y1}, \dots, \beta_{YM_Y})^T$  is a  $(M_Y + 2) \times 1$  vector of regression coefficients,  $\epsilon_Y = (\epsilon_1, \dots, \epsilon_n)^T$  is  $n \times 1$  error vector, and the design matrix  $Z_Y$  is defined as

$$Z_Y = \begin{bmatrix} 1 & X_1 & (X_1 - \kappa_{Y1})_+ & \dots & (X_1 - \kappa_{YM_Y})_+ \\ 1 & X_2 & (X_2 - \kappa_{Y1})_+ & \dots & (X_2 - \kappa_{YM_Y})_+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_n & (X_n - \kappa_{Y1})_+ & \dots & (X_n - \kappa_{YM_Y})_+ \end{bmatrix}.$$

Suppose that  $\epsilon_1, \dots, \epsilon_n$  are independent and identically distributed  $\text{Normal}(0, \sigma_Y^2)$ . The parameters in  $(\alpha_{Y0}, \alpha_{Y1})$  can be considered as fixed effects in the model. We put a normal prior on  $(\alpha_{Y0}, \alpha_{Y1})$  with 0 mean and large variance (say 100). This effectively acts as a noninformative uniform prior on the fixed effects. The random variables in  $\{\beta_{Yj}\}_{j=1}^{M_Y}$ , are assumed a priori independent and normally distributed, that is,  $\beta_{Yj} \sim \text{Normal}\{0, \sigma_j^2(\kappa_{Yj})\}$ , where  $j = 1, \dots, M_Y$ . Note here that  $\sigma_j^2(\kappa_{Yj})$  is the smoothing parameter (shrinkage or ridge parameter).

In the usual regression P-spline formulation with a global smoothing parameter, the  $\sigma_j^2(\kappa_{Yj})$  are all constant as a function of  $j$ , so that the smoothing is not spatially adaptive. We next describe how we extend the model to allow for spatially adaptive smoothing.

To develop a spatially adaptive technique we need to model  $\sigma_j^2(\kappa_{Yj})$ . This is crucial to capturing the spatial heterogeneity of the data because different smoothing parameters lend different amounts of smoothing in different regions. Allowing the smoothing parameter to be spatially adaptive also helps improve the mean squared error (MSE) of the fits, as well as

the accuracy of inference (Ruppert and Carroll 2000; Wand 2000). In this spirit, we develop a hierarchical model for  $\sigma_j^2(\kappa_{Yj})$ , where  $\sigma^2(\bullet)$  is a function evaluated at the knots ( $\kappa_{Yj}$ ). The functional form of  $\sigma^2(\bullet)$  is taken to be another linear regression spline, for example, for a linear spline

$$-\log\{\sigma^2(X)\} = \alpha_{s_0} + \alpha_{s_1}X + \sum_{k=1}^{M_s} \beta_{s_k}(X - \kappa_{s_k})_+, \quad (2.3)$$

where again  $\kappa_1 < \dots < \kappa_{M_s}$  are fixed knots. The number of subknots  $M_s$  is again user specified and is typically far less than  $M_Y$ , the number of knots in the original spline. The knots  $\{\kappa_k\}_{k=1}^{M_s}$  are again taken to be equally spaced quantiles of  $X$ . We now write (2.3) as a Bayesian linear model:

$$\rho = Z_s \Omega_s, \quad (2.4)$$

where  $\rho = [-\log\{\sigma^2(\kappa_1)\}, \dots, -\log\{\sigma^2(\kappa_{M_s})\}]^T$ ,  $\Omega_s = (\alpha_{s_0}, \alpha_{s_1}, \beta_{s_1}, \dots, \beta_{s_{M_s}})^T$  is an  $(M_s + 2) \times 1$  vector and  $Z_s$  is the design matrix, identical to that for (2.2) except for the change in the knots.

The random variables in the above equation are again assumed a priori independent and normally distributed, that is,  $\beta_{s_k} \sim \text{Normal}(0, \xi^2)$ , where  $k = 1, \dots, M_s$  and the parameters  $(\alpha_{s_0}, \alpha_{s_1})$  are again independent and normally distributed with zero mean and large variance.

As described in Section 3, although the motivation as a variance function to achieve spatially adaptive smoothing is clear, we will actually use a slight modification of (2.3)–(2.4) in order to avoid  $\Omega_s$  having to be sampled by a complex Metropolis–Hastings step.

### 3. IMPLEMENTATION VIA MARKOV CHAIN MONTE CARLO SIMULATION

This section sets up the framework to carry out the Markov chain Monte Carlo (MCMC) calculations. The prior distributions of the variance ( $\sigma_Y^2$ ) of the error vector  $\epsilon_Y$ , and  $\xi^2$ , the variance of the  $\beta_{s_k}$ 's, are taken to be a conjugate inverse gamma distribution with parameters  $(a_Y, b_Y)$  and  $(a_s, b_s)$ , respectively, that is,  $\sigma_Y^2 \sim \text{IG}(a_Y, b_Y)$  and  $\xi^2 \sim \text{IG}(a_s, b_s)$ , where  $\text{IG}(\bullet)$  is the inverse gamma distribution.

The parameters and random variables to be estimated in the model are  $\Omega_Y, \Omega_s, \xi^2$ , and  $\sigma_Y^2$ . With the above model and prior setup all the conditional distributions turn out to be of known standard forms except that of  $\Omega_s$ , which is a complex multivariate density. Hence we need a multivariate Metropolis–Hastings (MH) step to generate the samples. Because this involves searching over a  $(M_s + 2)$ -dimensional space for convergence, we noticed during the simulations that the movement of the MH step was very slow.

Hence we resort to the following device to reduce the dimension, thereby making the moves faster. We add an error term ( $\epsilon_u$ ) to the functional form of  $\sigma^2(X)$  in (2.3)–(2.4), leading to the model

$$\rho = Z_s \Omega_s + \epsilon_u, \quad (3.1)$$

where  $\epsilon_u = \text{Normal}(0, \sigma_u^2 I)$ . We fix the value of  $\sigma_u^2$  for our simulations to  $= .01$  because this variance is unidentified in the model. This device reduces the computational costs by reducing the MH step to one dimension to generate each of  $\sigma_j^2(\kappa_{Yj})$ 's, which are now conditionally dependent only on  $\Omega_s$  and conditionally independent of the rest of the parameters. This in effect makes the movement of the MCMC samples across the model space extremely fast and also improves the acceptance rate of MH moves. In our simulations we found that the choice of the value of  $\sigma_u^2$  does not have great influence on the performance of the MCMC. The complete conditional posteriors are derived in the Appendix.

## 4. SIMULATIONS

This section presents simulation studies primarily to evaluate the performance of our methodology and to compare it with other related approaches in literature. Section 4.1 compares the Bayesian P-spline approach to the frequentist local penalty approach of Ruppert and Carroll (2000) and with a variety of recent Bayesian approaches, in particular with the BARS (Bayesian Adaptive Regression Splines) method proposed by DiMatteo, Genovese, and Kass (2001). Section 4.2 discusses the issue of the choice of knots in the implementation of our algorithm.

### 4.1 COMPARISON WITH OTHER METHODS

We compare our Bayesian approach with the frequentist penalized splines approach (RC, Ruppert and Carroll 2000), through the following simulation study. The  $X$ 's were equally spaced on  $[0, 1]$ ,  $n = 400$ ,  $\sigma_u^2 = .01$  and the  $\epsilon_i$ 's were  $\text{Normal}(0, .04)$ . First, we use the regression function as in RC whose spatial variability was controlled by parameter  $j$ ,

$$m(x) = \sqrt{x(1-x)} \sin \left[ \frac{2\pi(1 + 2^{(9-4j)/5})}{x + 2^{(9-4j)/5}} \right], \quad (4.1)$$

where  $j = 3$  gives low spatial variability and  $j = 6$  gives severe spatial variability; see Figure 1 panels (a) and (b). The fits obtained by our algorithm using a truncated power basis function of degree 2 are shown in panels (c) and (d) along with associated 95% credible intervals. The credible intervals are estimated by computing the respective quantiles of the sampled function evaluations.

In this article, we allow the smoothing/penalty parameter to be a function of the independent variable  $\mathbf{X}$  as in (2.3). As mentioned earlier, this is important in capturing the spatial heterogeneity in the data by allowing different amounts of smoothing in different regions. We plot the underlying penalty function,  $\sigma^2(X)$  in Figure 1 panels (e) and (f). We would expect the value of  $\sigma^2(X)$  to be large if the regression curve has rapid changes in curvature, so that the second derivative of the fitted spline can take jumps large enough to accommodate these changes. Conversely, if the curvature changes slowly, then we would expect  $\sigma^2(X)$  to be small. Observe that the penalty curve adapts to the spatial heterogeneity

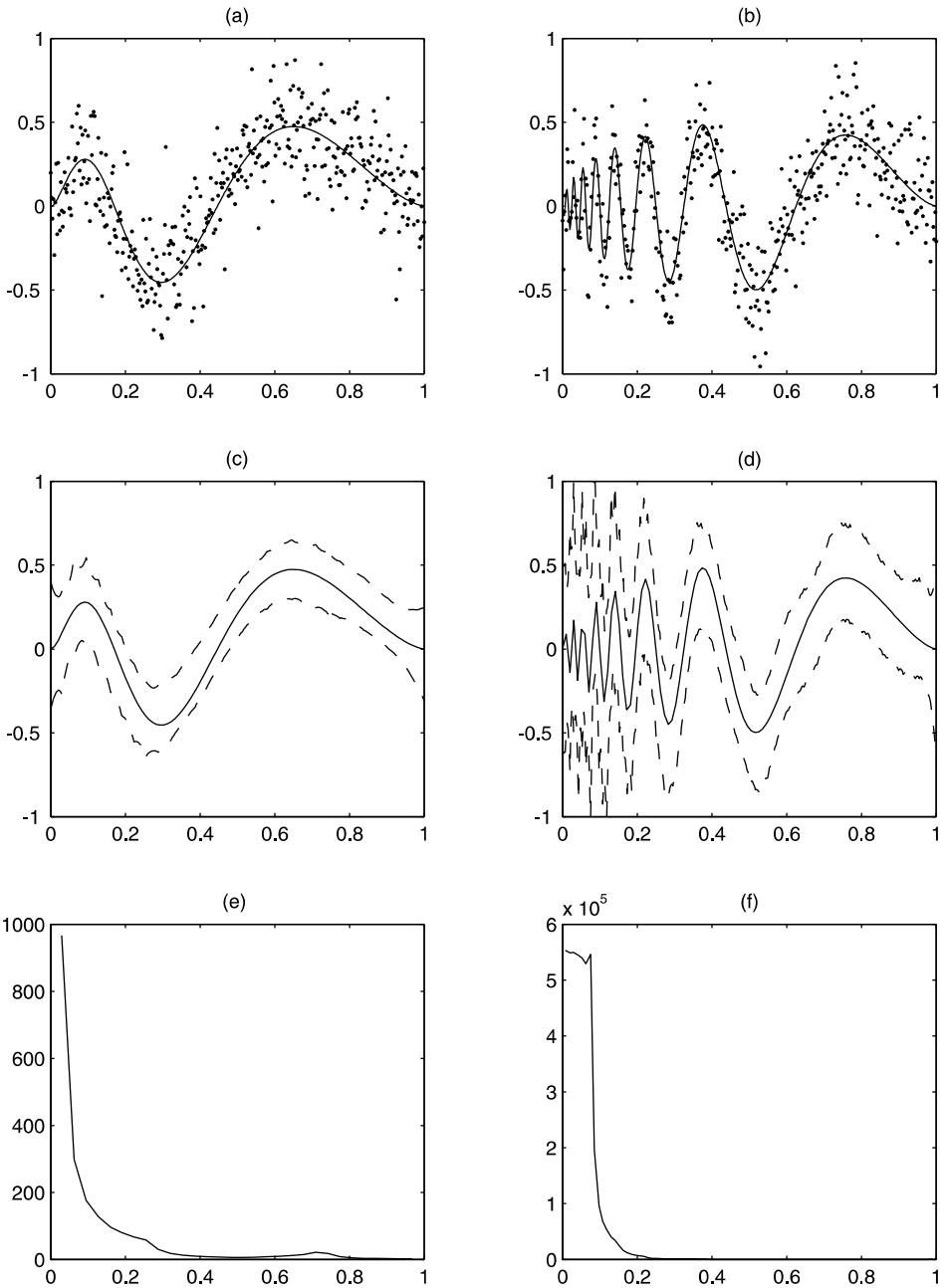


Figure 1. (a) Plot of the regression function (solid) and one sample (dots). The spatial variability of this curve is controlled by the parameter  $j$ . In this case  $j = 3$  gives low spatial variability. (b) Same as (a) with  $j = 6$  (severe spatial variability). (c) Plot of the estimated regression function along with the 95% credible intervals. The number knots ( $M_Y$ ) is 30 and number of subknots ( $M_s$ ) is 5. (d) Same as (c) but  $j = 6$  with  $M_Y = 90$  and  $M_s = 15$ . (e) The estimated penalty function,  $\sigma^2(X)$  for  $j = 3$ . (f) Same as (c) but  $j = 6$ .

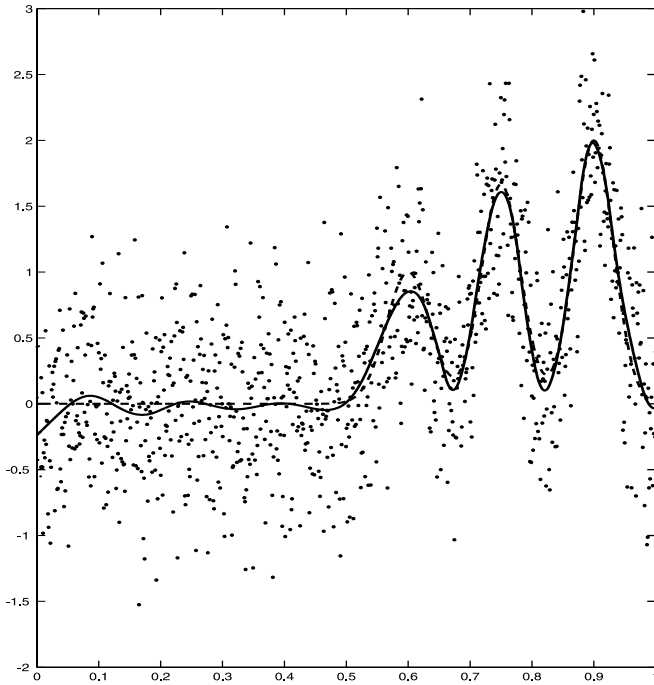


Figure 2. A copy of simulated data for comparing coverage probabilities. Shown are dots = data, dashed curve = true function, and solid curve = Bayesian P-spline estimate.

of the underlying regression function with large values in the regions where the curve is nonsmooth and small values in the smooth regions.

In order to compare the performance of fit we compute the averaged mean squared error (AMSE), which is given by

$$AMSE = n^{-1} \sum_{i=1}^n \{\hat{m}(x_i) - m(x_i)\}^2. \tag{4.2}$$

Our estimated AMSE for  $j = 3$  and  $j = 6$  is .0006 and .0027 respectively, which is comparable to those obtained by RC on the same example, which were .0007 and .0026, respectively.

We also compared the frequentist coverage properties of the Bayesian credible intervals with the frequentist local penalty confidence intervals of RC and with BARS. BARS employs free-knot splines, where the number and location of knots are random, and uses reversible jump MCMC (Green 1995) for implementation. We consider a spatially heterogeneous regression function

$$m(X) = \exp\{-400(X-.6)^2\} + \frac{5}{3} \exp\{-500(X-.75)^2\} + 2 \exp\{-500(X-.9)^2\}. \tag{4.3}$$

The  $X$ 's are equally spaced on  $[0, 1]$ , the sample size was  $n = 1,000$ , and the  $\epsilon_i$  were normally distributed with  $\sigma = .5$ . We use truncated power basis function of degree 2 with

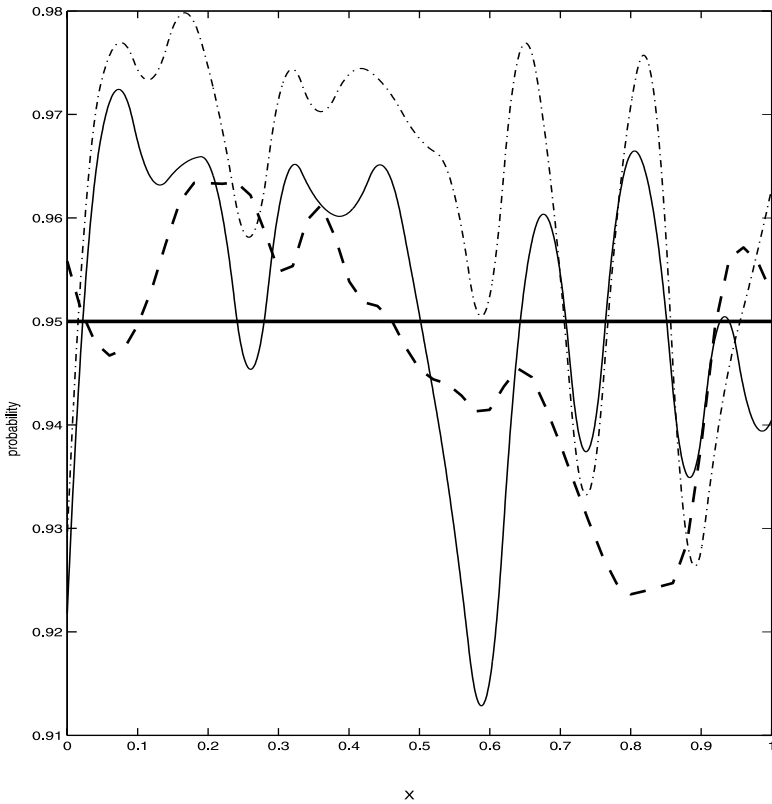


Figure 3. Comparison of pointwise coverage probabilities using 95% credible intervals of three methods: Bayesian P-splines (solid), adjusted local penalty confidence interval of Ruppert and Carroll (dots and dashes), and BARS (dashes). The coverage probabilities shown have been smoothed using P-splines to remove the Monte Carlo variability.

$M_Y = 40$  knots and  $M_s = 4$ . We again set  $\sigma_u = .01$ . The BARS program was graciously provided by the authors of DiMatteo et al. (2001). The BARS estimates are based on a Poisson prior with mean 6 for the number of knots, and the MCMC chain was run for 10,000 iterations with a burn-in period of 1,000.

Figure 2 shows a typical dataset with the true and fitted function plotted. To compare the coverage probabilities of the Bayesian credible intervals, we compute the frequentist coverage probabilities of the 95% credible intervals over 500 simulated datasets. Figure 3 shows the pointwise coverage probabilities of the 95% Bayesian credible intervals along with the “adjusted” local penalty confidence intervals of RC and those obtained by BARS. The adjustment used by RC is to multiply the pointwise posterior variances of the local-penalty estimates by a constant so that the average pointwise posterior variance of the estimate is the same for the global and local penalty estimate (Ruppert and Carroll 2000, sec. 4). The coverage probabilities shown have been smoothed using P-splines to remove the Monte Carlo variability. The average coverage probability obtained by the three methods (Bayesian P-splines, RC, BARS) are (95.22%, 96.28%, 94.72%). The coverage probabilities

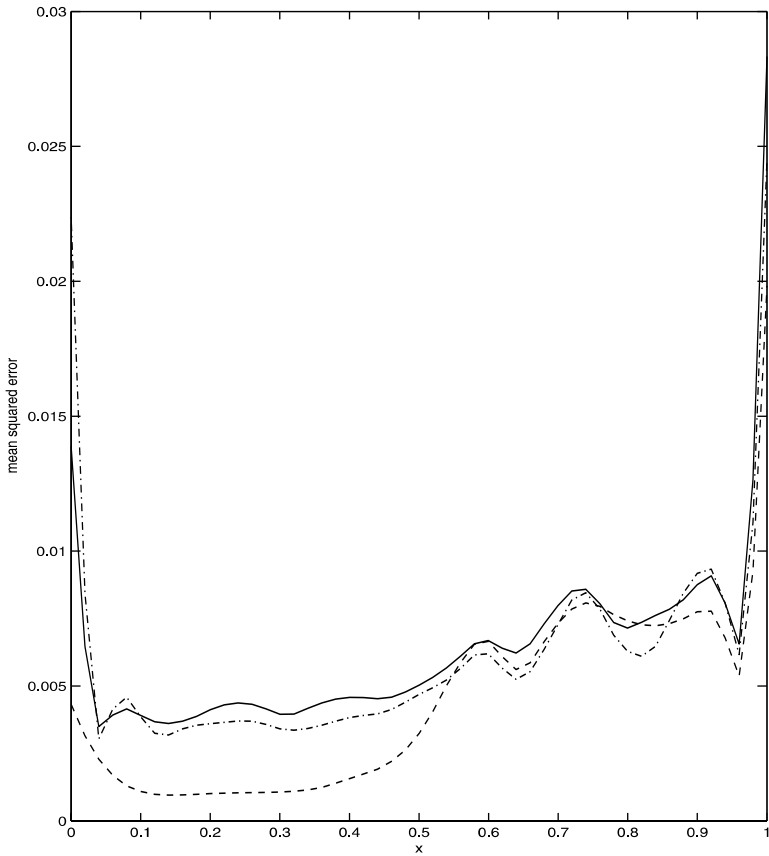


Figure 4. Comparison of pointwise mean squared errors (MSE) of three methods: Bayesian P-splines (solid), adjusted local penalty confidence interval of Ruppert and Carroll (dots and dashes), and BARS (dashes). The coverage probabilities shown have been smoothed using P-splines to remove the Monte Carlo variability.

for both Bayesian P-splines and BARS are slightly closer to the nominal coverage of 95% than the more conservative local penalty intervals of RC. Figure 4 shows the pointwise AMSE using the three methods. The average MSE for BARS (.0043) is somewhat smaller than the MSE for the Bayesian P-spline (.0061) and RC (.0065). Thus, our results are competitive to BARS in terms of frequentist coverage probabilities but BARS does seem to do a slightly better job than our method in terms of overall MSE.

We also compared our method with two other Bayesian approaches: the “Automatic Bayesian Curve Fitting” method proposed by Denison, Mallick, and Smith (1998), and the wavelet methods described by Donoho and Johnstone (1994). Further discussion of some potential problems with the method of Denison et al. was given by DiMatteo et al. (2001). We used the four test curves: “Heavisine”, “Blocks”, “Bumps”, and “Doppler” as in Donoho and Johnstone. The  $X$ ’s were again equally spaced on  $[0, 1]$ ,  $n$  was 2,048, and the  $\epsilon_i$ ’s were Normal(0, 1). The values of  $(M_Y, M_s)$  are  $\{(60, 10), (300, 30), (90, 15), (250, 80)\}$  for Heavisine, Blocks, Bumps, and Doppler, respectively. Denison et al. (1998) reported the

Table 1. Average Mean Squared Error (AMSE) Comparison From 10 Replications for Different Example Curves Across Different Methods: Wavelet Threshold Methods (Donoho and Johnstone 1994), Automatic Bayesian Curve Fitting (Denison et al. 1998), and Bayesian P-splines

Function	Wavelet threshold	Wavelet threshold	Automatic curve fitting	
	$\lambda_n^*$	$\{2\log(n)\}^{1/2}$	Denison et. al. (1998)	Bayesian P-spline
Heavisine	.060	.083	.033	.028
Blocks	.427	.905	.170	.137
Bumps	.499	1.080	.167	.098
Doppler	.151	.318	.135	.024

average MSE from 10 replications using the above examples and compared the results with those obtained by Donoho and Johnstone. Table 1 compares our results to those obtained by Denison et al. and Donoho and Johnstone. Here  $\lambda_n^*$  is the optimal wavelet threshold chosen specifically for each dataset, while  $\{2\log(n)\}^{1/2}$  is a universal threshold proposed by Donoho and Johnstone. As noted by Denison et al., the wavelet results are obtained with  $\sigma^2$  known and, for ease of computation, require the number of data points to be a power of 2. Specifically, we take  $n = 2,048$  and  $\sigma_u^2 = .01$  to compare our results with that of Denison et. al. and Donoho and Johnstone. Our method performs markedly better than the wavelet threshold methods in all the examples considered. Our results are comparable with those obtained by Denison et. al., for the Heavisine, Blocks, and Bumps functions but is much better for the Doppler example.

## 4.2 CHOICE OF KNOTS

This article presents a penalty approach that is similar in spirit to smoothing splines, but with fewer knots. In P-splines the crucial parameter for controlling the amount of smoothness is the penalty; that is, in our case  $\sigma^2(\kappa)$ . Once a certain minimum number of knots is reached, further increase in the number of knots causes little change to the fit given by P-spline (Ruppert 2002; Ruppert, Wand, and Carroll 2003). To this effect we ran an analysis with different number of knots, but the same selection for each method. The  $X$ 's were equally spaced on  $[0, 1]$ ,  $n = 400$ ,  $\sigma_u^2 = .01$ , and the  $\epsilon_i$ 's were Normal(0, .04). We again used the regression function as in (4.1) with  $j = 3$  (low spatial variability) and  $j = 6$  (severe spatial variability). We used five different sets of knots for the regression curve and the penalty curve, that is,  $\{(20, 3), (40, 4), (60, 6), (90, 9), (120, 15)\}$ . To compare the performance of fit across the different sets of knots we computed the AMSE as in (4.2).

Table 2 shows the AMSE for the test cases described earlier. For  $j = 3$  there is essentially no improvement on the fit of the curve on increasing the number knots. For the severe spatially variable case ( $j = 6$ ) the AMSE improves appreciably by increasing the number of knots from (20, 3) to (40, 4) but marginally by increasing the knots further. In all the examples we consider, there is evidence that there is a minimum necessary number of knots to be reached to fit the features in the data, and a further increase in the number of knots does not have any appreciable effect on the fit. Thus, if enough knots are specified, adaptive P-splines will be able to track the sudden changes in the underlying function, and

Table 2. Average Mean Squared Error (AMSE) Comparison Using Different Sets of Knots ( $M_Y, M_s$ ). Shown are the AMSE obtained for two test cases of a simulation example curve (4.1) (see text) where  $j = 3$  gives low spatial variability and  $j = 6$  gives severe spatial variability.

<i>Knot set</i>	$j = 3$	$j = 6$
(20,3)	.0007	.0094
(40,4)	.0007	.0048
(60,6)	.0008	.0036
(90,9)	.0009	.0028
(120,15)	.0012	.0027

where the underlying function is smooth the penalty will shrink the jumps at those knots to 0.

For the penalty curve  $\sigma^2(X)$ , the number of subknots  $M_s$  is taken to be much smaller than  $M_Y$ , the number of knots for the original regression spline. We tried a variety of choices for  $M_s$  in our simulation examples and found that the choice of  $M_s$  has relatively little effect on the fit. We keep the value of  $M_s$  large enough for the penalty curve to be spatially variable and small enough to reduce the computational cost.

The variance of the error term in all the simulation examples was taken so that we could mimic the simulation setup of the methods to which we compare our method. To study the performance of our estimator in the presence of increased noise we ran a further simulation study. We took the same simulation curve as in (2.1) with  $j = 3$ , the  $X$ 's were equally spaced on  $[0, 1]$ ,  $n = 400$  and  $\sigma_u^2 = .01$ . The variance of the error term ( $\sigma_Y^2$ ) was taken to be at three different levels: (.04, .1, .5). The average MSE over 25 simulated datasets was found to be (.0015, .0055, .0070), respectively, showing that the fitted curve can estimate the underlying regression function well even under increased noise.

## 5. EXTENSION TO ADDITIVE MODELS

### 5.1 AN ALGORITHM FOR ADDITIVE MODELS

To this point, we have confined our attention to univariate cases only. The methodology developed previously can be easily extended to additive models. The general additive model problem is to find functions  $m_j$  such that

$$Y = \alpha + \sum_{j=1}^p m_j(X_j) + \epsilon, \tag{5.1}$$

where the  $X_j$  are the predictor variables,  $E(\epsilon|X_1, \dots, X_p) = 0$  and  $\text{var}(\epsilon|X_1, \dots, X_p) = \sigma_Y^2$ . Thus, the overall regression function is a sum of  $p$  univariate functions, or curve fits. The univariate functions can be modeled with univariate splines, as we shall assume here. A model of the general type in (5.1) is known as an additive model. Hastie and Tibshirani (1990) provided an extensive account of these models.

The extension to Bayesian additive models is straightforward when we use a basis function representation (such as P-splines) for the individual curves. That is, we can write  $g(X) = \sum_{j=1}^p m_j(X_j)$  in (5.1) as a linear combination of P-splines and regression coefficients as:

$$g(X) = \alpha_0 + \sum_{j=1}^p \alpha_{1j} X_j + \sum_{j=1}^p \sum_{i=1}^{M_j} \beta_{ji} (X_j - \kappa_{ji})_+, \tag{5.2}$$

where again  $X_j$  is the  $j$ th predictor in  $X$  and  $M_j$  is the number of knots for the  $j$ th curve. Each one-dimensional function is again described by the parameters  $\beta_{ji}$  (the coefficients) and  $\kappa_{ji}$  (the knots).

As in previous sections we can use the same Bayesian linear model results, to make posterior inference for additive models. Thus, the fact that a general set of predictors is now a vector, rather than just a scalar, is of little consequence. In matrix notation we again write

$$Y = B\beta + \epsilon,$$

with  $\epsilon \sim \text{Normal}(0, \sigma^2 I)$ ,  $\beta = (\alpha_0, \alpha_1, \dots, \alpha_p, \beta_1, \dots, \beta_p)^T$  with  $\beta_j = (\beta_{j,1}, \dots, \beta_{j,M_j})$  and

$$\begin{aligned} \mathbf{B} &= \begin{bmatrix} 1 & X_1 & B_{1,1}(X_1) & \dots & B_{1,M_1}(X_1) & B_{2,1}(X_1) & \dots & B_{p,M_p}(X_1) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_n & B_{1,1}(X_n) & \dots & B_{1,M_n}(X_n) & B_{2,1}(X_n) & \dots & B_{p,M_p}(X_n) \end{bmatrix} \\ &= [1 \ X \ \mathbf{B}_1 \ \dots \ \mathbf{B}_p], \end{aligned}$$

where

$$\mathbf{B}_j = \begin{pmatrix} (X_{j1} - \kappa_{j,1})_+ & \dots & (X_{j1} - \kappa_{j,M_j})_+ \\ \vdots & \ddots & \vdots \\ (X_{jn} - \kappa_{j,1})_+ & \dots & (X_{jn} - \kappa_{j,M_j})_+ \end{pmatrix}.$$

With the above formulation, we adopt the same methodology as discussed in the previous sections for the univariate case. The distributional assumptions and prior structure on the random variables and parameters respectively are exactly the same as described in Section 2. The functional form of the variance of  $\beta$  is again a linear regression spline as in (2.3).

### 5.2 SIMULATIONS OF AN ADDITIVE MODEL

We take a slightly modified example from Hastie and Tibshirani (1990, pp. 247–251). We simulated from the functions  $m_1$  and  $m_2$  for the model,

$$Y_i = m_1(X_i) + m_2(Z_i) + \epsilon_i, \quad i = 1, \dots, 100,$$

where

$$m_1(X) = \begin{cases} -2X & \text{for } X < .6, \\ -1.2 & \text{otherwise,} \end{cases}$$

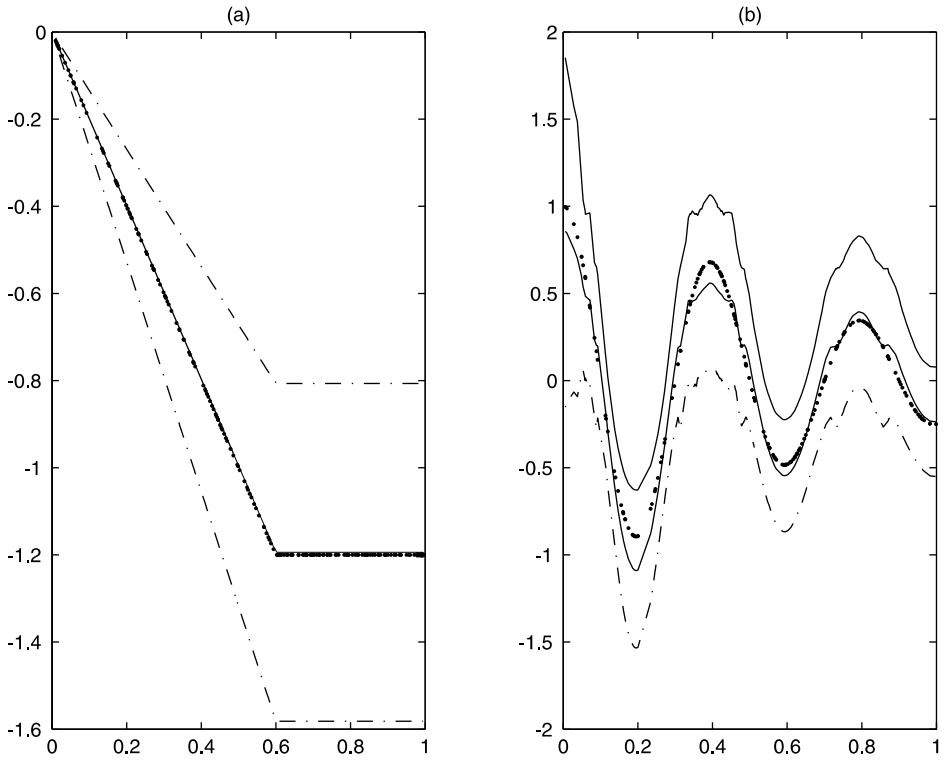


Figure 5. Additive model example. The dotted line represents the true function and the solid line the estimate regression function. Also shown are the 95% credible intervals. (a)  $m_1(X)$ ; (b)  $m_2(Z)$ .

$$m_2(Z) = \frac{\cos(5\pi Z)}{1 + 3Z^2},$$

with  $X_i$  and  $Z_i$  generated independently from the Uniform  $(0, 1)$  distribution and  $\epsilon_i$  from an Normal  $(0, 0.25)$  distribution. Figure 5 shows the estimates for functions  $m_1(X)$  and  $m_2(Z)$ , along with the credible intervals. The fits are better in terms of AMSE than the estimates provided by Denison et al. (1998). Also Denison et al. used plug-in estimates for the regression coefficients ( $\beta$ 's), and thus underestimate the uncertainty. We perform a full Bayesian analysis where in we draw the regression coefficients from the sampler, and hence obtain standard Bayesian credible intervals.

## 6. DISCUSSION AND CONCLUSION

We have presented an automated Bayesian method to fit spatially adaptive curves. We also provided an extension to additive models, wherein we obtained estimates of regression function and uncertainty intervals. We use regression P-splines to model the unknown smooth function, but we allow spatial adaptivity in the penalty function by modeling it as another regression P-spline. Our simulations indicate that our method is very competitive to

that of Ruppert and Carroll (2000) in terms of frequentist mean squared error and coverage probabilities of the credible intervals. We also provide Bayesian uncertainty intervals: the intervals had pointwise coverage close to the frequentist nominal level. Other methods, such as those of Donoho and Johnstone (1994) and Denison et al. (1998) appear to be no better than ours, and in some cases worse. Simulations indicate that our methods are roughly comparable to the BARS method of DiMatteo et al. (2001).

The reviewers of this article raised the most common issue regarding regression P-splines, that is, how many knots should one choose? We have relied on the work of Ruppert (2002) in simulated datasets, the analyses of data examples in Ruppert et al. (2003), and the work of Eilers and Marx (1996, 2002) as evidence that a large number of knots is unnecessary within the P-spline paradigm, with recommendations of between 10 and 50 knots for most situations of nonspatially adaptive smoothing. We believe that choosing a far smaller number of knots for spatially adaptive smoothing, as we have done, makes intuitive sense, and clearly works well in our examples, and in the examples of Ruppert and Carroll (2000). Inevitably, however, there will be interest in letting the data select the number of knots, even in the P-spline paradigm. Indeed, this has been done in the frequentist approach; see Ruppert and Carroll (2000) and Ruppert et al. (2003, chap. 17), where the number of knots and subknots is effectively chosen by GCV. We conjecture that the following method will work in our Bayesian context: it is based on the work of Kass and Wasserman (1995) and was illustrated in a model-averaging context by Carroll, Roeder, and Wasserman (1999). Specifically, choose a set of numbers of knots and subknots: Ruppert et al. used combinations of (10, 20, 40, 80, 120) for the former and (3, 4, 5, 6) for the latter. Then run our method for each of the combinations, and compute BIC for each at the posterior mean of median of the parameters. Finally, either select the combination on the basis of BIC, or average the fits using BIC as a model averaging device. We conjecture that this approach will work comparably to the frequentist approaches.

The other obvious issue here is the general comparison between regression spline methods. There are basically three approaches: (a) the P-spline approach as advocated here needs little more introduction; (b) knot selection methods such as BARS; and (c) regression splines without penalization but where the number of knots are selected using devices such as BIC and AIC; see Rice and Wu (2001) for an illustration. All these methods have value. Approach (c) generally tends to choose a smallish number of knots and is essentially available only in a frequentist context. Our own view is that in the frequentist context of regression splines, penalization with a fixed number of knots is a more natural approach than selecting the number of knots, especially when inference is of interest, since inference after model selection is extremely difficult in the frequentist context. The knot selection methods (free-knot splines) are clearly geared to handle problems where a high degree of spatial adaptation is necessary: that the P-spline approach does reasonably well in comparison to say BARS may in fact be seen as somewhat surprising. One nice feature of P-splines is that being little more than mixed models methods, they are readily adapted to new problems without great effort. Ciprian Crainiceanu (personal comm.) has recently shown how to implement our methods in WinBUGS, which gives some sense of its ease of application.

An interesting question concerns extension of these methods to non-Gaussian data. Indeed, BARS for example was motivated originally for the treatment of Poisson data, where it is extremely effective. Bayesian versions of penalized regression splines that are not spatially adaptive are easy to develop for generalized linear models, either via brute-force (Ruppert et al. 2003, chap. 16) or via latent variable methods such as those of Albert and Chib (1993) for binary data, and of Holmes and Mallick (2003) for binary and count data as special cases. These latter approaches essentially place one back into the Gaussian framework after the latent variables are computed, and these devices should allow spatially adaptive smoothing to be developed readily. Another interesting question is spatially adaptive smoothing in the presence of heteroscedasticity: frequentist P-splines are readily developed in this case (Ruppert et al. 2003, chap. 14), and adding Bayesian spatially adaptive smoothing should be possible.

### APPENDIX: DETAILS OF THE SAMPLER

The algorithm for the MCMC is the following:

- Give initial values to all parameters:  $\Omega_Y$ ,  $\Omega_s$ ,  $\xi^2$ ,  $\{\sigma^2(\kappa_j)\}_{j=1}^{M_Y}$ , and  $\sigma_Y^2$ .
- Start the MCMC sampler and iterate.
- *Updating* ( $\Omega_Y, \sigma_Y^2$ )

Conditional on the rest of the parameters, using Bayesian linear model theory with conjugate priors, the conditional posterior distribution of  $(\Omega_Y, \sigma_Y^2)$  is,

$$[\Omega_Y, \sigma_Y^2] \sim \text{Normal}(m_Y, \Sigma_Y)\text{IG}(\tilde{a}_Y, \tilde{b}_Y),$$

where  $m_Y = (1/\sigma_Y^2)(\Sigma_Y Z_Y^T Y)$ ,  $\Sigma_Y = [(Z_Y^T Z_Y / \sigma_Y^2 + \Lambda_Y^{-1})^{-1}]$ ,  $Z_Y$  is the regression spline design matrix and  $\Lambda_Y = \text{diag}\{100, \dots, 100, \sigma^2(\kappa_1), \dots, \sigma^2(\kappa_{M_Y})\}$  is the prior variance on  $\Omega_Y$ . Here  $\text{IG}(\bullet)$  is the inverse gamma distribution with shape parameter,  $\tilde{a}_Y = [(n/2) + a_Y]$  and scale parameter,  $\tilde{b}_Y = [(1/2)\{(Y - Z_Y \Omega_Y)^T (Y - Z_Y \Omega_Y)\} + (1/b_Y)]^{-1}$ .

- *Updating* ( $\Omega_s, \xi^2$ )
- With conjugate priors on  $(\Omega_s, \xi^2)$ , the conditional posterior distribution is

$$[\Omega_s, \xi^2] \sim \text{Normal}(m_s, \Sigma_s)\text{IG}(\tilde{a}_s, \tilde{b}_s),$$

where  $m_s = (1/\sigma_u^2)(\Sigma_s Z_s^T \rho)$ ,  $\Sigma_s = [(1/\sigma_u^2)(Z_s^T Z_s) + \Lambda_s^{-1}]^{-1}$  and  $\rho$  denotes the vector  $[-\log\{\sigma^2(\kappa_1)\}, \dots, -\log\{\sigma^2(\kappa_{M_Y})\}]^T$ .  $\Lambda_s = \text{diag}\{100, 100, \xi^2, \dots, \xi^2\}$  is the prior variance on  $\Omega_s$ . The posterior inverse gamma parameters are  $\tilde{a}_s = [(M_Y/2) + a_s]$  and  $\tilde{b}_s = [(1/2)\{\sum_{j=1}^{M_s} \beta_{j_s}^2\} + (1/b_s)]^{-1}$ . We set  $(a_s, b_s)$  to be  $(1, 1)$  for all the examples.

- *Updating*  $\{\sigma^2(\kappa_j)\}_{j=1}^{M_Y}$
- The penalty parameters,  $\{\sigma^2(\kappa_j)\}_{j=1}^{M_Y}$ , conditional on the current model parameters does not have an explicit form. Thus we resort to Metropolis-Hastings procedure

with a proposal density  $T[\sigma^{2*}(\kappa_j), \sigma^2(\kappa_j)]$  that generates the moves from the current state  $\sigma^{2*}(\kappa_j)$  to a new state  $\sigma^2(\kappa_j)$ . The proposed updates are then accepted with probabilities,

$$\alpha = \min \left\{ 1, \frac{p[\sigma^{2*}(\kappa_j)|\text{rest}]T[\sigma^2(\kappa_j), \sigma^{2*}(\kappa_j)]}{p[\sigma^2(\kappa_j)|\text{rest}]T[\sigma^{2*}(\kappa_j), \sigma^2(\kappa_j)]} \right\},$$

otherwise the current model is retained. It is convenient to take the proposal distribution  $T[\sigma^{2*}(\kappa_j), \sigma^2(\kappa_j)]$  to be a symmetric distribution (e.g., Gaussian) with mean equal to the old value  $\sigma^2(\kappa_j)$  and a prespecified standard deviation. Because the density involves exponential terms, the likelihood values calculated during the implementation of the algorithm are typically very large, hence we worked on a log scale. A common problem encountered in the implementation is the nonmobility of the MH step. If we start at bad starting values it may take a large number of iterations or even worse may not converge. To circumvent the problem we use frequentist estimates as starting values for the MCMC run, and in particular use the estimates of the smoothing parameter that minimize the generalized cross-validation (GCV) statistic

$$\text{GCV} = \frac{\|Y - Z_Y \Omega_Y(\sigma^2(\kappa))\|}{[(1 - \text{df}(\sigma^2(\kappa)))/n]^2}$$

where

$$\text{df}(\sigma^2(\kappa)) = \text{tr}\{(Z_Y^T Z_Y + \Lambda_Y)^{-1}(Z_Y^T Z_Y)\},$$

is the degree of freedom of the smoother which is defined to be the trace of the smoother matrix (Hastie and Tibshirani, 1990 sec. 3.5).

## ACKNOWLEDGMENTS

Our research was supported by a grant from the National Cancer Institute (CA-57030), the National Science Foundation (NSF DMS-0203215) and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106). We are deeply indebted to Rob Kass and Garrick Wallstrom for their gracious help in the proper computation of the BARS procedure.

*[Received December 2002. Revised July 2004.]*

## REFERENCES

- Albert, J. H., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669-679.
- Carroll, R. J., Roeder, K., and Wasserman, L. (1999), "Flexible Parametric Measurement Error Models," *Biometrics*, 55, 44-54.
- Coull, B. A., Ruppert, D., and Wand, M. P. (2001), "Simple Incorporation of Interactions Into Additive Models," *Biometrics*, 57, 539-545.

- de Boor, C. (1978), *A Practical Guide to Splines*, New York: Springer-Verlag.
- Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998), "Automatic Bayesian Curve Fitting," *Journal of Royal Statistical Society, Ser. B*, 60, 333–350.
- DiMatteo, I., Genovese, C. R., and Kass, R. E. (2001), "Bayesian Curve Fitting with Free-knot Splines," *Biometrika*, 88, 1055–1071.
- Donoho, D. L., and Johnstone I. M. (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425–455.
- Eilers, P. H. C., and Marx, B. D. (1996), "Flexible Smoothing With B-splines and Penalties" (with discussion), *Statistical Science*, 11, 89–121.
- (2002), "Generalized Linear Additive Smooth Structures," *Journal of Computational and Graphical Statistics*, 11, 758–783.
- Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.
- Hastie, W., and Tibshirani, R. (1990), *Generalized Additive Models*, London: Chapman and Hall.
- Holmes, C., and Mallick, B. K. (2003), "Generalized Nonlinear Modeling with Multivariate Smoothing Splines," *Journal of the American Statistical Association*, 98, 352–368.
- Kass, R. E., and Wasserman, L. (1995), "A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion," *Journal of the American Statistical Association*, 90, 928–934.
- Rice, J. A., and Wu, C. (2001), "Nonparametric Mixed Effects Models for Unequally Sampled Noisy Curves," *Biometrics*, 57, 253–269.
- Robinson, G. K. (1991), "That BLUP is a Good Thing: The Estimation of Random Effects" (with discussion), *Statistical Science*, 6, 15–51.
- Ruppert, D. (2002), "Selecting the Number of Knots for Penalized Splines," *Journal of Computational and Graphical Statistics*, 11, 735–757.
- Ruppert, D., and Carroll, R. J. (2000), "Spatially-Adaptive Penalties for Splines Fitting," *Australian and New Zealand Journal of Statistics*, 42, 205–223.
- Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, New York: Cambridge University Press.
- Wand, M. P. (2000), "A Comparison of Regression Spline Smoothing Procedures," *Computational Statistics [formerly Computational Statistics Quarterly]*, 15, 443–462.