Spatially Adaptive Bayesian Penalized Splines  
With Heteroscedastic Errors

Ciprian M. CRAINICEANU, David RUPPERT, Raymond J. CARROLL, Adarsh JOSHI, and Billy GOODNER

Penalized splines have become an increasingly popular tool for nonparametric smoothing because of their use of low-rank spline bases, which makes computations tractable while maintaining accuracy as good as smoothing splines. This article extends penalized spline methodology by both modeling the variance function nonparametrically and using a spatially adaptive smoothing parameter. This combination is needed for satisfactory inference and can be implemented effectively by Bayesian MCMC. The variance process controlling the spatially adaptive shrinkage of the mean and the variance of the heteroscedastic error process are modeled as log-penalized splines. We discuss the choice of priors and extensions of the methodology, in particular, to multivariate smoothing. A fully Bayesian approach provides the joint posterior distribution of all parameters, in particular, of the error standard deviation and penalty functions. MATLAB, C, and FORTRAN programs implementing our methodology are publicly available.

Key Words: Heteroscedasticity; MCMC; Multivariate smoothing; Regression splines; Spatially adaptive penalty; Thin-plate splines; Variance functions.

1. INTRODUCTION

Penalized splines (Eilers and Marx 1996; Marx and Eilers 1998; Ruppert, Wand, and Carroll 2003) have become a popular nonparametric tool. Their success is due mainly to the use of low rank bases, which makes computations tractable. Also, penalized splines can be viewed as mixed models and fit with widely available statistical software (Ngo

Ciprian M. Crainiceanu is Assistant Professor, Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe St, E3636 Baltimore, MD 21205 (E-mail: ccrainic@jhsph.edu). David Ruppert is Andrew Schultz Jr. Professor of Engineering and Professor of Statistical Science, School of Operational Research and Industrial Engineering, Cornell University, Rhodes Hall, NY 14853 (E-mail: dr24@cornell.edu). Raymond J. Carroll is Distinguished Professor of Statistics and Professor of Nutrition and Toxicology, Department of Statistics, 447 Blocker Building, Texas A&M University, College Station, TX 77843–3143 (E-mail: carroll@stat.tamu.edu). Adarsh Joshi is a Graduate student, Department of Statistics, 447 Blocker Building, Texas A&M University, College Station, TX 77843–3143 (E-mail: adarsh@stat.tamu.edu). Billy Goodner is a Graduate student, Department of Statistics, Texas A&M University College Station, 447 Blocker Building, TX 77843–3143 (E-mail: bgoodner@stat.tamu.edu).

© 2007 American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of North America
Journal of Computational and Graphical Statistics, Volume 16, Number 2, Pages 265–288
DOI: 10.1198/106186007X208768
and Wand 2004; Crainiceanu, Ruppert, and Wand 2005; Wood 2006). Bayesian penalized splines (Ruppert, Wand, and Carroll 2003; Lang and Brezger 2004; Crainiceanu, Ruppert, and Wand 2005) use a stochastic process model as a prior for the regression function. It is typical to assume that both this process and the errors are homoscedastic.

The penalized spline methodology has been extended to heteroscedastic errors using an iterative frequentist approach (Ruppert, Wand, and Carroll 2003). We develop the corresponding Bayesian inferential methodology using a simple, previously unknown, solution. Spatially adaptive penalty parameters have also been used before (Ruppert and Carroll 2000; Eilers and Marx 2002; Baladandayuthapani, Mallick, and Carroll 2005; Lang and Brezger 2004; Jullion and Lambert 2006; Pintore, Speckman, and Holmes 2006). However, this is the first article to combine these features. We show that this combination is important and that spatially varying penalties can adapt to spatial heterogeneity of both the mean function and the error variance. Implementation of this extension was not straightforward, but we developed and implemented an algorithm that works well in practice.

Our methodology can be extended to almost any penalized spline model—for example, additive, varying coefficient, interaction, and multivariate smoothing models. To illustrate this we also study multivariate low-rank thin-plate splines. As in the univariate case we model the mean, the variance, and the smoothing functions nonparametrically and estimate them from the data using a fully Bayesian approach. The computational advantage of low rank over full rank smoothers becomes even greater in more than one dimension.

Section 2 provides a quick introduction to penalized splines. Section 3 discusses the choice of priors for penalized spline regression. Section 4 describes how to obtain simultaneous credible bounds for mean and variance functions and their derivatives. Section 5 compares our proposed methodology to the one proposed by Baladandayuthapani et al. (2005) for adaptive univariate penalized spline regression. Section 6 describes the extension to multivariate smoothing and Section 7 compares our multivariate methodology with other adaptive surface fitting methods. Section 8 discusses an example of bivariate smoothing with heteroscedastic errors and spatially adaptive smoothing. Section 9 presents the full conditional distributions and discusses our implementation of models in MATLAB, FORTRAN, and C. Section 11 provides our conclusions.

2. PENALIZED REGRESSION SPLINES

Consider the regression equation \( y_i = m(x_i) + \epsilon_i, \ i = 1, \ldots, n \), where the \( \epsilon_i \) are independent and the mean is modeled as

\[
m(x) = m(x, \theta_X) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{k=1}^{K_m} b_k (x - \kappa^m_k)^+,
\]

where \( \beta = (\beta_0, \ldots, \beta_p)^T \), \( b = (b_1, \ldots, b_{K_m})^T \), \( \theta_X = (\beta^T, b^T)^T \), \( \kappa^m_1 < \kappa^m_2 < \cdots < \kappa^m_{K_m} \) are fixed knots, and \( a_{\alpha}^m \) denotes \( \max(\alpha, 0))^p \). Following Gray (1994) and Ruppert (2002), we take \( K_m \) large enough (e.g., 20) to ensure the desired flexibility. To avoid overfitting the \( b_k \sim N(0, \sigma_b^2(\kappa^m_k)) \) and \( \epsilon_i \sim N(0, \sigma^2(x_i)) \) are shrunk towards zero by an amount controlled by \( \sigma_b^2(\cdot) \) and \( \sigma^2(\cdot) \), which vary across the range of \( x \).
The penalized spline model can be written as

\[
\begin{align*}
  y_i &= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \sum_{k=1}^{K_m} b_k (x_i - \kappa_m^k)^p + \epsilon_i; \\
  \epsilon_i &\sim N(0, \sigma^2_{e}(x_i)), \quad i = 1, \ldots, n; \\
  b_k &\sim N(0, \sigma^2_{b}(\kappa_m^k)), \quad k = 1, \ldots, K_m,
\end{align*}
\]  

(2.1)

where \( \epsilon_i \) and \( b_k \) are mutually independent given \( \sigma^2_{e}(\cdot) \), \( \sigma^2_{b}(\cdot) \). In Sections 2.1 and 2.2 \( \sigma^2_{e}(x_i) \) and \( \sigma^2_{b}(\kappa_m^k) \) are modeled using log-spline models. Denote by \( X \) the \( n \times (p + 1) \) matrix with the \( i \)th row equal to \( (1, x_i, \ldots, x_i^p) \) and by \( Z \) the \( n \times K_m \) matrix with \( i \)th row equal to \( (x_i - \kappa_m^1)^p, \ldots, (x_i - \kappa_m^{K_m})^p) \), \( Y = (y_1, \ldots, y_n)^T \) and \( \epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \). Model (2.1) can be rewritten in matrix form as

\[
Y = X\beta + Zb + \epsilon, \quad \mathbb{E}(b) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \text{cov}(b) = \begin{pmatrix} \Sigma_b & 0 \\ 0 & \Sigma_\epsilon \end{pmatrix},
\]

(2.2)

where the joint conditional distribution of \( b \) and \( \epsilon \) given \( \Sigma_b \) and \( \Sigma_\epsilon \) is assumed normal. The \( \beta \) parameters are treated as fixed effects. The covariance matrices \( \Sigma_b \) and \( \Sigma_\epsilon \) are diagonal with the vectors \( \{\sigma^2_{b}(\kappa_m^1), \ldots, \sigma^2_{b}(\kappa_m^{K_m})\} \) and \( \{\sigma^2_{e}(x_1), \ldots, \sigma^2_{e}(x_n)\} \) as main diagonals respectively. For this model \( E(Y) = X\beta \) and \( \text{cov}(Y) = \Sigma_\epsilon + ZZ' \Sigma_\epsilon Z' \).

### 2.1 Error Variance Function Estimation

Ignoring heteroscedasticity may lead to incorrect inferences and inefficient estimation, especially when the response standard deviation varies over several orders of magnitude. Moreover, understanding how variability changes with the predictor may be of intrinsic interest (Carroll 2003). Transformations can stabilize the response variance when the conditional response variance is a function of the conditional expectation (Carroll and Ruppert 1988), but in other cases one needs to estimate it by modeling the variance as a function of the predictors.

We model the variance function as a loglinear model

\[
\begin{align*}
  \log[\sigma^2_{e}(x_i)] &= \gamma_0 + \cdots + \gamma_p x_i^p + \sum_{s=1}^{K_\epsilon} c_s (x_i - \kappa_{\epsilon}^s)^p \\
  c_s &\sim N(0, \sigma^2_{c}(\kappa_{\epsilon}^s)), \quad s = 1, \ldots, K_\epsilon,
\end{align*}
\]  

(2.3)

where the \( \gamma \) parameters are fixed effects and \( \kappa_{\epsilon}^1 < \cdots < \kappa_{\epsilon}^{K_\epsilon} \) are knots. The normal distribution of the \( c_s \) parameters shrinks them towards 0 and ensures stable estimation.

### 2.2 Smoothly Varying Local Penalties

Following Baladandayuthapani et al. (2005), we model \( \sigma^2_{b}(\cdot) \) as

\[
\begin{align*}
  \log[\sigma^2_{b}(x)] &= \delta_0 + \cdots + \delta_d x^d + \sum_{s=1}^{K_\beta} d_s (x - \kappa_{\beta}^s)^q \\
  d_s &\sim N(0, \sigma^2_{d}(\kappa_{\beta}^s)), \quad s = 1, \ldots, K_\beta,
\end{align*}
\]  

(2.4)

where the \( \delta \)'s are fixed effects, the \( d_s \)'s are mutually independent, and \( \{\kappa_{\beta}^1, \ldots, \kappa_{\beta}^{K_\beta}\} \) are knots. The particular case of a global smoothing parameter corresponds to the case when the spline function is a constant, that is \( \log[\sigma^2_{b}(\kappa_m^k)] = \delta_0 \).
Ruppert and Carroll (2000) proposed a "local penalty" model similar to (2.4). Baladandayuthapani, Mallick, and Carroll (2005) developed a Bayesian version of Ruppert and Carroll’s (2000) estimator and showed that their estimator is similar to Ruppert and Carroll’s in terms of mean square error and outperforms other Bayesian methods. Lang, Fronk, and Fahrmeir (2002) considered locally adaptive dynamic models and find that their method is roughly comparable to the method of Ruppert and Carroll (2000) in terms of mean square error (MSE) and coverage probability. Krivobokova, Crainiceanu, and Kauermann (2007) used a model similar to (2.4) for low rank thin plate spline and apply Laplace approximations instead of Bayesian MCMC.

In Lang and Brezger’s (2004) prior for the $b_k$, the nonconstant variance is independent from knot to knot, since the variances of the $b_k$ are assumed to be independent inverse-gammas. In contrast, our model allows dependence so that if the variance is high at one knot, then it is high at neighboring knots. Stated differently, the Lang and Brezger model is one of random heteroscedasticity and ours of systematic heteroscedasticity. Both types of priors are sensible and will have applications, but in any specific application it is likely that one of the two will be better. For example, Lang and Brezger found that for “Doppler type” functions their estimator is not quite as good as Ruppert and Carroll’s and that the coverage of their credible intervals is not so accurate either. It is not hard to understand why. Doppler-type functions that oscillate more slowly going from left to right are consistent with a prior variance for the $b_k$ that is monotonically decreasing, exactly the type of prior included in our model but not Lang and Brezger’s.

### 2.3 Choice of Spline Basis, Number and Spacing of Knots, and Penalties

For concreteness, we have made some specific choices about the spline bases, number and knot spacings, and form of the penalty. In terms of model formulation, the choice of spline basis in the model is not important since an equivalent basis gives the same model and the basis used in computations need not be the same as the one used to express the model. However, spline bases are very different in terms of computational stability. For example, cubic thin plate splines (Ruppert, Wand, and Carroll 2003, pp. 72–74) have better numerical properties than the truncated polynomial basis. In more than one dimension, the tensor product of truncated polynomials proved unstable and we preferred using low rank radial smoothers (see Section 6).

The penalty we use is somewhat different from that of Eilers and Marx (1996) and also somewhat different from the penalties used by smoothing splines. We believe that the penalty parameter is the crucial choice, so we have concentrated on spatially adaptive modeling of the penalty parameter.

In this article we use knots to model the mean, variance, and spatially adaptive smoothing parameter. The methodology described here does not depend conceptually on the number of knots. For example, one could use a knot at each observation for each of the three functions. Although we use quantile knot spacings, the best type of knot spacings is controversial. Eilers and Marx (2004) stated that “Equally spaced knots are always to be preferred.” In contrast, Ruppert, Wand, and Carroll (2003) used quantile-spacing in all of their examples,
though they did not make any categorical statement that quantile-spacing is always to be preferred. Since knot spacing is not the main focus of this article and our results were not sensitive to it we do not discuss this issue here.

In this article we have used quantile knot spacing for the univariate case and space filling design (Nychka and Saltzman 1998) for the multivariate case, which is based on the maximal separation principle and generalizes the quantile knot spacing in more than one dimension. For the multivariate case we also used equally spaced knots in simulations to make our methods more comparable to other published methods.

3. PRIOR SPECIFICATION

Any smoother depends heavily on the choice of smoothing parameter, and for penalized splines the smoothing parameter is the ratio of the error variance to the prior variance on the mean (Ruppert, Wand, and Carroll 2003). The smoothness of the fit depends on how these variances are estimated. For example, Crainiceanu and Ruppert (2004) showed that, in finite samples, the (RE)ML estimator of the smoothing parameter is biased towards oversmoothing and Kauermann (2002) obtained corresponding asymptotic results for smoothing splines.

In Bayesian models, the estimates of the variance components are known to be sensitive to the prior specification, for example, see Gelman (2006). To study the effect of this sensitivity upon Bayesian penalized splines, consider model (2.1) with one smoothing parameter and homoscedastic errors so that \( \sigma^2_b \) and \( \sigma^2_\epsilon \) are constant. In terms of the precision parameters \( \tau_b = 1/\sigma^2_b \) and \( \tau_\epsilon = 1/\sigma^2_\epsilon \), the smoothing parameter is \( \lambda = \tau_\epsilon / \tau_b = \sigma^2_b / \sigma^2_\epsilon \) and a small (large) \( \lambda \) corresponds to oversmoothing (undersmoothing).

3.1 PRIORS ON THE FIXED EFFECTS PARAMETERS

It is standard to assume that the fixed effects parameters, \( \beta_i \), are apriori independent, with prior distributions either \( [\beta_i] \propto 1 \) or \( \beta_i \propto N(0, \sigma^2_\beta) \), where \( \sigma^2_\beta \) is very large. In our applications we used \( \sigma^2_\beta = 10^6 \), which we recommend if \( x \) and \( y \) have been standardized or at least have standard deviations with order of magnitude one.

For the fixed effects \( \gamma \) and \( \delta \) used in the log-spline models (2.3) and (2.4) we also used independent \( N(0, 10^6) \) priors. When this prior is not consistent with the true value of the parameter, a possible strategy is to fit the model using a given set of priors and obtain the estimators \( \hat{\gamma} \), \( \hat{\delta} \), and \( \tilde{\gamma} \). We could then use independent priors \( N(\hat{\gamma}, 10^6 \hat{\sigma}^2_\gamma) \) and \( N(\hat{\delta}, 10^6 \hat{\sigma}^2_\delta) \) for \( \gamma \) and \( \delta \), respectively.

3.2 PRIORS ON THE PRECISION PARAMETERS

As just mentioned, the priors for the precisions \( \tau_b \) and \( \tau_\epsilon \) are crucial. We now show how critically the choice of \( \tau_b \) may depend upon the scaling of the variables. The gamma family of priors for the precisions is conjugate. If \( [\tau_b] \sim \text{Gamma}(A_b, B_b) \) and, independently of
\( \tau_b, [\tau_\epsilon] \sim \text{Gamma}(A_\epsilon, B_\epsilon) \) where Gamma(A, B) has mean A/B and variance A/B^2, then

\[
[\tau_b|\mathbf{Y}, \mathbf{\beta}, \mathbf{b}, \tau_\epsilon] \sim \text{Gamma} \left( A_\epsilon + \frac{K_m}{2}, B_\epsilon + \frac{||\mathbf{b}||^2}{2} \right),
\]

and

\[
[\tau_\epsilon|\mathbf{Y}, \mathbf{\beta}, \mathbf{b}, \tau_\epsilon] \propto \text{Gamma} \left( A_\epsilon + \frac{n}{2}, B_\epsilon + \frac{||\mathbf{Y} - \mathbf{X}\mathbf{\beta} - \mathbf{Z}\mathbf{b}||^2}{2} \right).
\]

Also,

\[
E(\tau_b|\mathbf{Y}, \mathbf{\beta}, \mathbf{b}, \tau_\epsilon) = \frac{A_\epsilon + K_m/2}{B_\epsilon + ||\mathbf{b}||^2/2}, \quad \text{var}(\tau_b|\mathbf{Y}, \mathbf{\beta}, \mathbf{b}, \tau_\epsilon) = \frac{A_\epsilon + K_m/2}{(B_\epsilon + ||\mathbf{b}||^2/2)^2},
\]

and similarly for \( \tau_\epsilon \).

The prior does not influence the posterior distribution of \( \tau_\epsilon \) when both \( A_\epsilon \) and \( B_\epsilon \) are small compared to \( K_m/2 \) and \( ||\mathbf{b}||^2/2 \), respectively. Since the number of knots is \( K_m \geq 1 \) and in most problems considered \( K_m \geq 5 \), it is safe to choose \( A_\epsilon \leq 0.01 \). When \( B_\epsilon < < ||\mathbf{b}||^2/2 \) the posterior distribution is practically unaffected by the prior assumptions. When \( B_\epsilon \) increases compared to \( ||\mathbf{b}||^2/2 \), the conditional distribution is increasingly affected by the prior assumptions. \( E(\tau_b|\mathbf{Y}, \mathbf{\beta}, \mathbf{b}, \tau_\epsilon) \) is decreasing in \( B_\epsilon \) indicating that large \( B_\epsilon \) compared to \( ||\mathbf{b}||^2/2 \) corresponds to undersmoothing. Since the posterior variance of \( \tau_b \) is also decreasing in \( B_\epsilon \), a poor choice of \( B_\epsilon \) will likely result in underestimating the variability of the smoothing parameter \( \lambda = \tau_\epsilon/\tau_b \) causing confidence intervals for the mean function to be too narrow. The condition \( B_\epsilon < < ||\mathbf{b}||^2/2 \) shows that the “noninformativeness” of the gamma prior depends essentially on the scale of the problem.

To show the potential severity of these effects we consider the model (2.1) with a global smoothing parameter and homoscedastic errors. We used quadratic splines with 30 knots and the light detection and ranging (LIDAR) data for illustration. The LIDAR data was obtained from atmospheric monitoring of pollutants and consists of the range, or distance traveled before light is reflected back to a source and log ratio, the logarithm of the ratio of received light from two laser sources. For more details see Ruppert, Wand, Holst, and Hössier (1997).

Figure 1 shows the effect of four mean-one Gamma priors for the precision of the truncated polynomial parameters. The variances of these priors are 10, 10^3, 10^6, and 10^10, respectively. Obviously, the first two inferences provide severely undersmoothed, almost indistinguishable, posterior means. The third graph is much smoother but still exhibits roughness especially in the right-hand side of the plot, while the fourth graph displays a pleasing smooth pattern, consistent with our frequentist inference. Using either prior distribution one obtains that the posterior mean of \( ||\mathbf{b}||^2/2 \) is of order \( 10^{-6} \) to \( 10^{-5} \). This explains why values of \( B_\epsilon \) larger than \( 10^{-6} \) proved inappropriate for this problem.

The size of \( ||\mathbf{b}||^2/2 \) depends upon the scaling of the \( x \) and \( y \) variables and in the case of the LIDAR data \( ||\mathbf{b}||^2/2 \) is small because the standard deviation of \( x \) is much larger than the standard deviation of \( y \). If \( y \) is rescaled to \( a_y y \) and \( x \) to \( a_x x \), then the regression function becomes \( a_y m(a_x x) \) whose \( p \)th derivative is \( a_y a_x^p m^{(p)}(a_x x) \) so that \( ||\mathbf{b}||^2/2 \) is
rescaled by the factor $a_i^2a_{2p}^2$. Thus, $||b||^2/2$ is particularly sensitive to the scaling of $x$. The size of $||b||^2/2$ also depends on $K_m$. Because the integral of $m(p)$ over the range of $x$ will be approximately $\sum_{k=1}^{K_m} b_k \approx \sqrt{K_m}\sigma_b$, we can expect that $\sigma_b^2 \propto (K_m)^{-1}$ and the smoothing parameter should be proportional to $K_m$. For the LIDAR data, the GCV chosen smoothing parameter is 0.0095, 0.0205, 0.0440, and 0.0831 for $K_m$ equal to 15, 30, 60, and 120, respectively, so as expected the smoothing parameter approximately doubles as $K_m$ doubles.

Figure 2 presents the same type of results as Figure 1 for Gamma priors for the precision parameter $\tau_b$ with the mean held fixed at $10^{-6}$ and variances equal to 10, $10^3$, 10$^6$, and 10$^{10}$, respectively. These prior distributions have a much smaller effect on the posterior mean of the regression function. The fit seems to be undersmooth when the variance is 10. Clearly, when the variance increases the fit becomes smooth indicating that a value of the variance larger than $10^3$ will produce a reasonable fit.

A similar discussion holds true for $\tau_\epsilon$ but now large $B_\epsilon$ corresponds to oversmoothing and $\tau_\epsilon$ does not depend on the scaling of $x$. In applications it is less likely that $B_\epsilon$ is comparable in size to $||Y - X\beta - Zb||^2$, because the latter is an estimator of $n\sigma_\epsilon^2$. If $\hat{\sigma}_\epsilon^2$ is an estimator of $\sigma_\epsilon^2$ a good rule of thumb is to use values of $B_\epsilon$ smaller than $n\hat{\sigma}_\epsilon^2/100$. This
Figure 2. LIDAR data with unstandardized covariate: effect of four Gamma priors with mean $10^{-6}$ for the precision of the truncated polynomial parameters. The variances of these priors are $10$, $10^3$, $10^6$, and $10^{10}$, respectively.

Rule should work well when $\hat{\sigma}_2^2$ does not have an extremely large variance.

Alternative to gamma priors are discussed by, for example, Natarajan and Kass (2000) and Gelman (2006). These have the advantage of requiring less care in the choice of the hyperparameters. However, we find that with reasonable care, the conjugate gamma priors can be used in practice. Nonetheless, exploration of other prior families for penalized splines would be well worthwhile, though beyond the scope of this article.

4. SIMULTANEOUS CREDIBLE BOUNDS

Simultaneous credible intervals have been discussed before in the literature, for example, Besag, Green, Higdon, and Mengersen (1995). The idea can be used to produce simultaneous credible bounds for functions estimated smoothly. To show this, let $f(\cdot)$ be either $m(\cdot)$, $\log(\sigma_x^2(\cdot))$, $\log(\sigma_b^2(\cdot))$, or a derivative of order $q$, $1 \leq q \leq p$, of one of these functions. It is straightforward to use MCMC output to construct simultaneous credible bounds on $f$ over an arbitrary finite interval $[x^1, x^N]$. Typically, $x^1$ and $x^N$ would be the smallest and largest observed values of $x$.

Let $x^1 < x^2 < \cdots < x^N$ be a fine grid of points on this interval. Let $E\{f(x^i)\}$ and $SD\{f(x^i)\}$ be the posterior mean and standard deviation of $f(x^i)$ estimated from a MCMC sample. Let $M_\alpha$ be the $(1 - \alpha)$ sample quantile of $\max_{1 \leq i \leq N} \left|\frac{f(x^i) - E\{f(x^i)\}}{SD\{f(x^i)\}}\right|$ computed from the realizations of $f$ in the MCMC sample. Then $I(x^i) = E\{f(x^i)\} \pm M_\alpha SD\{f(x^i)\}, \ i = 1, \ldots, N$, are simultaneous credible intervals. For $N$
large, the upper and lower limits of these intervals can be connected to form simultaneous credible bands. In the remainder of this article we only report simultaneous credible bounds for the functions and their derivatives. In the examples considered in the following these bounds tend to be roughly 30–50% wider than pointwise credible bounds.

5. COMPARISON WITH OTHER UNIVARIATE SMOOTHERS

Baladandayuthapani et al. (2005) presented a comprehensive simulation study of their Bayesian spatially adaptive penalized spline model, which is obtained as a particular case of our full model with constant error variance. The results of this study indicate that the method is comparable to or better than several other spatially adaptive Bayesian methods on a variety of datasets. Results also showed that the method produced practically indistinguishable inferences from those obtained using the fast frequentist method of Ruppert and Carroll (2000). Since we are unaware of any other estimation methodology that estimates jointly the nonparametric adaptive mean and nonparametric variance models, we will compare our methodology with that of Ruppert and Carroll (2000). We will compare the frequentist properties of our Bayesian methodology, such as MSE and coverage probabilities calculated by simulating from existent benchmark models.

Consider the regression model

\[
y_i = m(x_i) + \epsilon_i,
\]

where \(\epsilon_i \sim N(0, \sigma^2_\epsilon)\) are independent errors. The first function considered was the Doppler function, described by Donoho and Johnstone (1994)

\[
m(x) = \sqrt{x(1-x)} \sin \left\{ \frac{2\pi(1 + 0.05)}{x + 0.05} \right\}.
\]

For the Doppler function we considered \(n = 2048\) equally spaced \(x\)’s between \([0, 1]\) and \(\sigma^2_\epsilon = 1\).

Ruppert and Carroll (2000) proposed the following family of models related to the Doppler but allowing for more shape flexibility

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

where \(j = 3\) and \(j = 6\) correspond to low and severe spatial heterogeneity, respectively. For this function we considered \(n = 400\) equally spaced \(x\)’s on \([0, 1]\) and \(\sigma^2_\epsilon(x) = 0.04\). We will refer to this as the RC \(j\) function, where \(j\) controls the amount of spatial heterogeneity.

Ruppert and Carroll (2000) and Baladandayuthapani et al. (2005) also considered the following function

\[
m(x) = \exp\{-400(x - 0.6)^2\} + \frac{5}{3} \exp\{-500(x - 0.75)^2\} + 2 \exp\{-500(x - 0.9)^2\},
\]

(5.1)
Table 1. Average mean square error (AMSE) for different mean functions based on 100 simulations. The column “True variance” indicates whether the data-generating process is homoscedastic or heteroscedastic. The columns “Fitted error variance” refer to the type of assumption used for fitting the data: “Homoscedastic” uses the Ruppert and Carroll (2000) fitting algorithm while “Heteroscedastic” uses the Bayesian inferential algorithm described in this article. The mean and log-variance functions were modeled using degree 2 splines with $K_m = K_\epsilon = 40$ equally spaced knots. The shrinkage process was modeled using degree 1 splines with $K_b = 4$ subknots.

<table>
<thead>
<tr>
<th>Function</th>
<th>True variance</th>
<th>Fitted error variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doppler</td>
<td>Constant</td>
<td>0.0197</td>
</tr>
<tr>
<td>RC$_{j=3}$</td>
<td>Constant</td>
<td>0.0011</td>
</tr>
<tr>
<td>RC$_{j=6}$</td>
<td>Constant</td>
<td>0.0061</td>
</tr>
<tr>
<td>3 Bumps</td>
<td>Constant</td>
<td>0.0054</td>
</tr>
<tr>
<td>3 Bumps</td>
<td>Variable</td>
<td>0.0031</td>
</tr>
</tbody>
</table>

which is roughly constant between $[0, 0.5]$ and has three sharp peaks at 0.6, 0.75 and 0.9. For this function we considered $n = 1,000$, equally spaced $x$’s on $[0, 1]$ and two different scenarios for the standard error of the error process: (a) $\sigma_\epsilon(x) = 0.5$ for homoscedastic errors and (b) $\sigma_\epsilon(x_i) = 0.5 - 0.8x + 1.6(x - 0.5)_+$. We will refer to this as the “three-bump” function.

We used 100 simulations from these models and fit a spatially adaptive penalized spline that uses a constant variance. For the three-bump function we also fitted a model with nonconstant variance modeled as a log-penalized-spline. For the mean and log-variance functions we used quadratic penalized splines with $K_m = K_\epsilon = 40$ knots. For the log-variance corresponding to the shrinkage process we used $K_b = 4$ knots.

We calculated the mean square error for each simulated dataset as

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

where $\hat{m}(x)$ is the posterior mean of $m(x)$ using a given model. Table 1 provides a summary of the average MSE (AMSE) results over 100 simulations comparing the method that allows for heteroscedastic errors with the one that does not. Remarkably, our method that estimates a nonconstant variance performed well compared to methods that assume a constant variance when the variance is constant. In contrast, when the true variance process is not constant our method performed better.

Our method was slightly outperformed by the Ruppert and Carroll (2000) method in the Doppler case. This is probably due to the fact that the Ruppert and Carroll method assumes a constant variance for the Doppler function when the variance used for simulations was $\sigma_\epsilon^2 = 1$. However, both methods outperformed the method of Baladandayuthapani et al. (2005) who reported an AMSE of 0.024 for the Doppler function.

For the three-bump function with constant variance, $\sigma_\epsilon(x) = 0.5$, the two models produced practically indistinguishable MSEs. Note that our AMSE over all $x$’s using 100 simu-
Figure 3. Comparison of pointwise coverage probabilities of the 95% credible intervals in 500 simulations. Data was simulated from the model described in Section 5 with heteroscedastic error variance. Estimation was done using two methods: Bayesian adaptive penalized splines with homoscedastic errors (dashed), and Bayesian adaptive penalized splines with heteroscedastic errors (solid). The coverage probabilities have been smoothed to remove Monte Carlo variability.

lations was AMSE = 0.0054, which is smaller than 0.0061 reported by Baladandayuthapani et al. (2005). The coverage probabilities of the 95% credible intervals were very similar for the two models for each value of the covariate, with a slight advantage in favor of the model using homoscedastic errors. The average coverage probability was 94.7% for the model with homoscedastic errors and 93.5% for the model with heteroscedastic errors. These coverage probabilities are very similar to the ones reported by Baladandayuthapani et al. (2005) in their Figure 3.

For case (b), $\sigma_\epsilon(x_i) = 0.5 - 0.8x + 1.6(x - 0.5)_+$, the heteroscedastic model substantially outperformed the homoscedastic model both in terms of AMSE and coverage probabilities. In this situation it would be misleading to compare only the average coverage probability for the 95% credible intervals. Indeed, these averages are very close, 94.1% for the heteroscedastic and 93.0% for the homoscedastic method, but they are obtained from very different sets of pointwise coverage probabilities. Figure 3 displays the coverage probabilities for these two methods. Note that for the heteroscedastic method the coverage probability in the $[0, 0.5]$ interval is close to 95%. In the same interval the coverage probability for the homoscedastic method starts from around 0.8 and increases until it crosses the 95% target around 0.2. Moreover, in the interval $[0.3, 0.5]$ this coverage probability is estimated to be 1. This is due to the fact that on this interval $\sigma_\epsilon(x)$ decreases linearly from 0.5 to 0.1. Since the homoscedastic method assumes a constant variance, the credible intervals will tend to be shorter than nominal in regions of higher variability ($x$ close to zero), thus producing lower coverage probabilities. In regions of smaller variability ($x$ close to 0.5) the credible intervals will tend to be much wider, thus producing extremely large
coverage probabilities.

In the interval [0.55, 0.65] the homoscedastic method slightly outperforms the heteroscedastic method. This seems to be the effect of a “lucky” combination of two factors. As we discussed, the size of the credible intervals for the homoscedastic method in a neighborhood of 0.5 is much larger than nominal. However, at the same point the mean function changes from a constant to a rapidly oscillating function and the coverage probabilities drop roughly at the same rate. In the interval [0.8, 1] one can notice a phenomenon very similar to the one described for the interval [0, 0.5]. While the function oscillates more rapidly in this region, the heteroscedastic adaptive method produces credible intervals with coverage probabilities close to the nominal 95%. However, the homoscedastic method does not take into account the increased variability and produces credible intervals that are too short.

For these examples both methods performed roughly similar on simulated data sets that did not require variance estimation. However, when the error variance is not constant the method using log-penalized-splines to estimate the error variance considerably outperforms the method that does not.

6. LOW RANK MULTIVARIATE SMOOTHING

In this section we generalize the ideas in Section 2 to multivariate smoothing while preserving the appealing geometric interpretation of the smoother. Consider the following regression $y_i = m(x_i) + \epsilon_i$ where $m(\cdot)$ is a smooth function of $L$ covariates. We will use radial basis functions which have the advantage of being more numerically stable than the tensor product of truncated polynomials in more than one dimension. Suppose that $x_i \in \mathbb{R}^L$, $1 \leq i \leq n$ are vectors of covariates and $\kappa_k \in \mathbb{R}^L$, $1 \leq k \leq K_m$ are $K_m$ knots. Consider the following distance function

$$C(r) = \begin{cases} \frac{||r||^{2M-L}}{2M-L \log ||r||} & \text{for } L \text{ odd} \\ \frac{||r||^{2M-L}}{2M-L} & \text{for } L \text{ even} \end{cases},$$

where $|| \cdot ||$ denotes the Euclidean norm in $\mathbb{R}^L$, the integer $M$ controls the smoothness of $C(\cdot)$, $X$ the matrix with $i$th row $X_i = [1 \ x_i^T]$, $Z_{K_m} = \{C(||x_i - \kappa_k||)\}_{1 \leq i \leq n, 1 \leq k \leq K_m}$, $\Omega_{K_m} = \{C(||\kappa_k - \kappa_{k'}||)\}_{1 \leq k \leq K, 1 \leq k' \leq K}$ and define $Z = Z_K \Omega_K^{-1/2}$, where $\Omega_K^{-1/2}$ is the principal square root of $\Omega_K$. With these notations the low rank approximation of thin plate spline regression can be obtained as the BLUP in the LMM (Kammann and Wand 2003; Ruppert, Wand, and Carroll 2003)

$$Y = X\beta + Zb + \epsilon, \quad E\left( \begin{array}{c} b \\ \epsilon \end{array} \right) = \left( \begin{array}{c} 0 \\ 0 \end{array} \right), \quad \text{cov} \left( \begin{array}{c} b \\ \epsilon \end{array} \right) = \left( \begin{array}{cc} \sigma_b^2 \mathbf{1}_{K_m} & 0 \\ 0 & \sigma_\epsilon^2 \mathbf{1}_n \end{array} \right).$$

This model contains only one variance component, $\sigma_b^2$, for controlling the shrinkage of $b$, which is equivalent to one global smoothing parameter $\lambda = \sigma_b^2/\sigma_\epsilon^2$ and implicitly assumes homoscedastic errors. To relax these assumptions, we consider a new set of knots $\{\kappa^*_1, \ldots, \kappa^*_{K_b}\}$ and define $X^*$ and $Z^*$ similarly with the corresponding definition of matrices
Consider the following model for the response

\[
\begin{align*}
    y_i &= \beta_0 + \sum_{j=1}^L \beta_j x_{i,j} + \sum_{k=1}^m b_k z_{i,k} + \epsilon_i; \\
    \epsilon_i &\sim N(0, \sigma^2_\epsilon(x_i)), \quad i = 1, \ldots, n; \\
    b_k &\sim N(0, \sigma^2_b(\kappa_k)), \quad k = 1, \ldots, K_m, \quad (6.2)
\end{align*}
\]

where the error variance, \(\log(\sigma^2_\epsilon(x_i))\), is modeled as a low rank log-thin plate spline

\[
\begin{align*}
    \log(\sigma^2_\epsilon(x_i)) &= \gamma_0 + \sum_{j=1}^L \gamma_j x_{i,j} + \sum_{k=1}^m c_k z_{i,k}; \\
    c_k &\sim N(0, \sigma^2_c), \quad k = 1, \ldots, K_c, \quad (6.3)
\end{align*}
\]

and the random coefficient variance, \(\log(\sigma^2_b(\kappa_k))\), is modeled as another low rank log-thin plate spline

\[
\begin{align*}
    \log(\sigma^2_b(\kappa_k)) &= \delta_0 + \sum_{j=1}^L \delta_j x^*_{k,j} + \sum_{j=1}^s d_j z^*_{k,j}; \\
    d_j &\sim N(0, \sigma^2_d), \quad j = 1, \ldots, K_b, \quad (6.4)
\end{align*}
\]

where \(x^*_{k,j}\) and \(z^*_{k,j}\) are the entries of \(X^*\) and \(Z^*\) matrices, respectively, and \(c_k\) and \(d_j\) are assumed mutually independent.

In the case of low rank smoothers the set of knots for the covariates and subknots for modeling the shrinkage process have to be chosen. One possibility is to use equally spaced knots and subknots. We use this approach in our simulation study in Section 7 to make our method comparable from this point of view with previously published methods. Another possibility is to select the knots and subknots using the space filling design (Nychka and Saltzman 1998), which is based on the maximal separation principle. This avoids wasting knots and is likely to lead to better approximations in sparse regions of the data. The \texttt{cover.design()} function from the \texttt{R} package \texttt{Fields} (Fields Development Team 2006) provides software for space filling knot selection. We use this approach in Section 8 in the Noshiro elevation example.

### 7. COMPARISONS WITH OTHER ADAPTIVE SURFACE FITTING METHODS

Lang and Brezger (2004) compared their adaptive surface smoothing method with several methods used in a simulation study by Smith and Kohn (1997): MARS (Friedman 1991), “loocf” (Cleveland and Grosse 1991), “tps” (bivariate cubic thin plate splines with a single smoothing parameter), tensor product cubic smoothing splines with five smoothing parameters, and a parametric linear interaction model. We will use the following regression functions, also used by Lang and Brezger

- \(m_1(x_1, x_2) = x_1 \sin(4\pi x_2)\) where \(x_1\) and \(x_2\) are distributed independently uniform on \([0, 1]\).
- \(m_2(x_1, x_2) = 1/5 \exp(-8x_1^2) + 3/5 \exp(-8x_2^2)\) where \(x_1\) and \(x_2\) are distributed independently normal with mean 0.5 and variance 0.1.
Function \( m_1(\cdot) \) contains complex interactions and moderate spatial variability while \( m_2(\cdot) \) has only main effects and is much smoother. We used a sample size \( n = 300 \), \( \sigma = 1/4 \text{ range}(f) \) and 250 simulations from each model.

We will compare the performance of our method with that of Lang and Brezger’s, which performed at least as good or better than other adaptive methods in a variety of examples. To avoid dependence of results on knot choice we used the same choice as the one described by Lang and Brezger: 12 × 12 equally spaced knots for the mean function. However, we used thin-plate splines instead of tensor products of B-splines. We modeled the smoothing parameter using a low rank log-thin plate spline with an equally spaced 5 × 5 knot grid. We considered two estimators, one with constant error variance (CEV) and one with estimated error variance using a log-thin plate spline (EEV) with the same 12 × 12 knot grid as for the mean function. The CEV estimator is the one that can directly be compared with the estimators considered by Lang and Brezger (2004). Because Lang and Brezger presented only comparative results using boxplots we will use their results as read from their Figure 5. The performance of all estimators was measured by the log of the mean squared error, \( \log(\text{MSE}) \), where \( \text{MSE}(\hat{f}) = 1/n \sum_{i=1}^{n}(f(x_i) - \hat{f}(x_i))^2 \).

For the CEV estimator and function \( f_1(\cdot) \) corresponding to moderate spatial heterogeneity we obtained a median \( \log(\text{MSE}) \) of \(-3.67 \) with an interquartile range \([-3.80, -3.53]\) and a range \([-4.21, -3.13]\). These values are better than the best values reported by Lang and Brezger’s methods in their Figure 5(b). Lang and Brezger reported that their method, denoted “P-spline” outperforms all other methods in this case and achieved a median \( \log(\text{MSE}) \) of \(-3.50 \). Remarkably, the EEV estimator performed almost as well as the CEV estimator in this case and also outperformed the other methods. The median \( \log(\text{MSE}) \) for EEV was of \(-3.59 \) with an interquartile range \([-3.74, -3.43]\) and a range \([-4.14, -3.02]\). These results are consistent with univariate results reported by Ruppert and Carroll (2000) and Baladandayuthapani, Mallick, and Carroll (2005) who found that allowing the smoothing parameter to vary smoothly outperforms other adaptive techniques.

For function \( f_1(\cdot) \) Figure 4 displays the coverage probabilities for the 95% credible intervals calculated over a 20 × 20 equally spaced grid in \([0, 1]^2\). For one grid point we calculated the frequency with which the 95% pointwise credible interval covers the true value of the function at the grid point. As expected, these coverage probabilities show strong spatial correlation with lower coverage probabilities along the ridges of the sinus function. Coverage probability is lowest when \( x_1 \) is in the \([0.2, 0.5]\) range. The signal-to-noise in this region is about half the signal-to-noise ratio in the region corresponding to \( x_1 \) close to 1. This explains why the coverage probability is smaller in this region. Interestingly, the zeros of the true function appearing at \( x_2 = 0.25, 0.50, 0.75 \) are covered with at least 95% probability. Another interesting feature is the lower coverage probabilities near the north, south, and east boundaries. These features of the pointwise coverage probability map are consistent with the frequentist analysis of Krivobokova et al. (2007) but are not consistent with the coverage probabilities reported by Lang and Brezger for their Bayesian P-spline method.

For our CEV estimator and function \( f_2(\cdot) \) corresponding to very low spatial heterogeneity we obtained a median \( \log(\text{MSE}) \) of \(-5.95 \) with an interquartile range \([-6.12, -5.66]\)
Figure 4. Coverage probability of the pointwise 95% credible intervals for the mean function for function $f_1(x_1, x_2) = x_1 \sin(4\pi x_2)$ with a constant error standard deviation $\sigma = \text{range}(f)/4$. Probabilities are computed on a $20 \times 20$ equally spaced grid of points in $[0, 1]^2$. The model used was a thin plate spline with $K = 100$ knots for the mean function, a log-thin plate spline with $K = 100$ knots for the variance function, and a log-thin plate spline with $K^* = 16$ knots for the spatially adaptive shrinkage parameter. The parameter controlling the degree of smoothness of the thin-plate spline basis was $m = 2$.

and a range $[-6.51, -5.41]$. The results of Lang and Brezger in their Figure 5(a) indicate that their “P-spline” method achieved a median log(MSE) of $-6.00$. Our method performs roughly similar to the Bayesian P-spline method of Lang and Brezger and to the cubic thin plate spline (tps), being outperformed only by Lang and Brezger’s adaptive P-spline method with two main effects, “P-spline, main.” We could also add two main effects to our bivariate adaptive smoother to improve MSE for this example, but this is not our concern in this article. Again, the EEV estimator performed very similarly to our CEV estimator.

The last simulation study was done using the function $f_1(x_1, x_2)$ where $x_1, x_2$ are independent uniformly distributed in $[0, 1]$ with an error standard deviation function

$$
\sigma_e(x_1, x_2) = \frac{r}{32} + \frac{3r}{32} x_1^2,
$$

where $r = \text{range}(f_1)$. We compared our CEV and EEV estimators described at the beginning of this section and, as expected, the EEV estimator outperformed the CEV both in terms of log(MSE) and coverage probabilities. More precisely for log(MSE) we obtained a median of $-5.26$ for CEV and $-5.35$ for EEV, an interquartile range $[-5.37, -5.15]$ for CEV and $[-5.48, -5.24]$ for EEV, and a range $[-4.65, -5.72]$ for CEV and $[-4.84, -5.84]$ for EEV. For both estimators the general patterns for coverage probabilities were the ones displayed in Figure 4. EEV outperformed CEV in terms of coverage probabilities. For example, the regions where the nominal level is exceeded for both estimators have a roughly similar
shape and location, but the EEV coverage probabilities tend to be closer to their nominal level.

Note that our model incorporating nonparametric variance estimation is more general than the models considered in the literature. However, its performance for function $m_1(\cdot, \cdot)$ with moderate spatial variability and constant variance was better than all methods considered by Lang and Brezger in their simulation study (see their Figure 5, panel b). Our method has good performance for the function $m_2(\cdot, \cdot)$ being the second best when compared to methods consider by Lang and Brezger (see their Figure 5, panel a). For the case when variance is not constant our adaptive method that estimates the variance nonparametrically outperformed our adaptive method that estimates a constant variance.

8. THE NOSHIRO EXAMPLE

Noshiro, Japan, was the site of a major earthquake. Much of the damage from the quake was due to soil movement. Professor Thomas O’Rourke, a geotechnical engineer, was investigating the factors that might help predict soil movement during a future quake. One factor thought to be of importance was the slope of the land. To estimate the slope, Ruppert (1997) used a dataset with 799 observations where $x$ was longitude and latitude and $y$ was elevation at locations in Noshiro. The object of primary interest is the gradient, at least its magnitude and possibly its direction.

We used the thin-plate spline models (6.2) for the mean of $y$ with $K_m = 100$ knots selected by the maximum separation criterion implemented in the function `cover.design()` of the R package Fields (Nychka, Haaland, O’Connell, and Ellner 1998). The sampling
Figure 6. Posterior mean of the mean regression function for the Noshiro example. The model used was a thin plate spline with \( K = 100 \) knots for the mean function, a log-thin plate spline with \( K = 100 \) knots for the variance function, and a log-thin plate spline with \( K^* = 16 \) knots for the spatially adaptive shrinkage parameter. The parameter controlling the degree of smoothness of the thin plate spline basis was \( m = 2 \).

Figure 6 displays the posterior mean of the mean response function. There is a rather sharp peak around \((0.42, 0.4)\) with a slow decay from the peak in the general south direction and a much faster decay in all other directions. The function becomes smoother towards the boundary. This is exactly the type of function for which adaptive spatial smoothing can substantially improve the fit.

Figure 7 shows that in large areas of the map \( \sigma_e(\cdot) \) is smooth and with small values. However, in a neighborhood of the peak of the mean function, \( \sigma_e(\cdot) \) displays two relatively sharp maxima located NNW and SSE of the peak. We also performed a two-stage frequentist analysis. In the first stage we used a low rank thin plate spline for the mean function and the results from this regression were used to obtain residuals. We then fitted another low rank thin plate spline to the absolute values of these residuals. While this map was not identical to the one in Figure 7, it did exhibit the same patterns of heteroscedasticity.

The process \( \log(\sigma^2_b(\kappa_k)) \) controlling the shrinkage of the \( b_k \) is displayed in Figure 8.
Figure 7. Posterior mean of the standard deviation, $\sigma_{\epsilon}(x_i)$, of the error process function for the Noshiro example. The model used was a thin plate spline with $K = 100$ knots for the mean function, a log-thin plate spline with $K = 100$ knots for the variance function, and a log-thin plate spline with $K^* = 16$ knots for the spatially adaptive shrinkage parameter. The parameter controlling the degree of smoothness of the thin-plate spline basis was $m = 2$.

Figure 8. Posterior mean of the shrinkage process $-\log(\sigma_b^2(\kappa))$. The model used was a thin plate spline with $K = 100$ knots for the mean function, a log-thin plate spline with $K = 100$ knots for the variance function, and a log-thin plate spline with $K^* = 16$ knots for the spatially adaptive shrinkage parameter. The parameter controlling the degree of smoothness of the thin-plate spline basis was $m = 2$. 
Smaller values of this function correspond to less shrinkage of \( b_k \) towards zero and more local behavior of the smoother. Figure 8 indicates that in a neighborhood of the peak of the mean function the shrinkage is smaller to allow the mean function to change rapidly. Away from the peak, the shrinkage is larger corresponding to a smoother mean.

9. IMPLEMENTATION USING MCMC

We will now provide some details for MCMC simulations of model (2.1), where the variances \( \sigma^2_e(x_i) \) and \( \sigma^2_k(k) \) are modeled by Equations (2.3) and (2.4). The implementation for other models, for example, for the multivariate smoothing in Section 6, is similar. Consider independent normal priors for the coefficients of the monomials: \( b_i \sim N(0, \sigma^2_{0,b}) \), \( \gamma_i \sim N(0, \sigma^2_{0,y}) \), \( i = 0, \ldots, p \), \( \delta_i \sim N(0, \sigma^2_{0,k}) \), \( i = 0, \ldots, q \), and independent inverse Gamma priors for the variance components: \( \sigma^2_e \sim \text{IGamma}(a_e, b_e) \) and \( \sigma^2_k \sim \text{IGamma}(a_d, b_d) \). Using these priors many full conditionals of the posterior distribution are easy to derive, while a few have complex multivariate forms. Our implementation of the MCMC using multivariate Metropolis-Hastings steps proved to be unstable with poor mixing properties. A simple and reliable solution was to change the model by adding error terms to the log-spline models, that is mixing properties. A simple and reliable solution was to change the model by adding error terms to the log-spline models, that is

\[
\log \{ \sigma^2_e(x_i) \} = \gamma_0 + \cdots + \gamma_p x_i^p + \sum_{s=1}^{K_e} c_s(x_i - \kappa_s^e)_+ + u_i
\]

\[
\log \{ \sigma^2_k(k^m) \} = \delta_0 + \cdots + \delta_q (k_q^m)_+^q + \sum_{s=1}^{K_k} d_s (k_s^m - \kappa_s^b)_+ + v_k
\]

(9.1)

where \( u_i \sim N(0, \sigma^2_e) \) and \( v_k \sim N(0, \sigma^2_k) \). This idea was also used for \( \sigma^2_b \) by Baladandayuthapani, Mallick, and Carroll (2005). We fixed the values of \( \sigma^2_e = \sigma^2_v = 0.01 \), as these variances appear nonidentifiable and a standard deviation of 0.1 is small on a log-scale. This device makes computations feasible by replacing sampling from complex full conditionals by simple univariate MH steps.

Define as before by \( \Sigma_e, \Sigma_b, \theta_X \), and define \( \theta_e = (y^T, c^T)^T \), \( \theta_b = (\delta^T, d^T)^T \), \( C_X = (X Z), C_e = (X_e Z_e), C_b = (X_b Z_b) \), where \( X, X_e, X_b \) contain the monomials and \( Z, Z_e, Z_b \) contain the truncated polynomials of the spline models for \( m \), \( \log(\sigma^2_e) \), and \( \log(\sigma^2_b) \) respectively. Also, denote by

\[
\Sigma_{0X} = \begin{bmatrix} \sigma^2_{0,\beta} I_{p+1} & 0 \\ 0 & \Sigma_b \end{bmatrix}
\]

\[
\Sigma_{0e} = \begin{bmatrix} \sigma^2_{0,y} I_{p+1} & 0 \\ 0 & \sigma^2_e I_{K_e} \end{bmatrix}
\]

\[
\Sigma_{0b} = \begin{bmatrix} \sigma^2_{0,k} I_{q+1} & 0 \\ 0 & \sigma^2_k I_{K_m} \end{bmatrix}
\]

The full conditionals of the posterior are detailed below

1. \( [\theta_X] \sim N \left( M_X C_X^{-1} Y, M_X \right) \), where \( M_X = \left( C_X^{-1} \Sigma_e^{-1} C_X + \Sigma_{0X}^{-1} \right)^{-1} \).

2. \( [\theta_e] \sim N \left( M_e C_e^{-1} Y_e / \sigma^2_e, M_e \right) \), where \( Y_e = [\log(\sigma^2_{e,1}), \ldots, \log(\sigma^2_{e,n})]^T \) and \( M_e = (C_e^{-1} C_e / \sigma^2_e + \Sigma_{0e}^{-1})^{-1} \).
Figure 9. Computation time (in minutes) versus sample size for 10,000 simulations from the joint posterior distribution given the data. The mean function is modeled as a penalized spline with \( K_m \) knots with the log of the variance of the shrinkage process modeled as another penalized spline with \( K_b = 4 \) subknots. The log of the error variance is modeled as another penalized spline with \( K_\epsilon = K_m \) knots. Different symbols correspond to increasing number of knots: dots \((K_m = 10)\), asterisks \((K_m = 20)\), circles \((K_m = 30)\), x’s \((K_m = 40)\).

3. \[ \theta_b \sim N \left( M_b C^T_b Y_b / \sigma^2_c, M_b \right) \]
   \[ Y_b = \left[ \log(\sigma^2_1), \ldots, \log(\sigma^2_{K_m}) \right]^T \]
   \[ M_b = \left( C^T_b C_b / \sigma^2_c + \Sigma_{0b}^{-1} \right)^{-1}. \]

4. \[ \sigma^2_c \sim IGamma \left( a_c + K_c / 2, b_c + ||c||^2 / 2 \right). \]

5. \[ \sigma^2_\epsilon \sim IGamma \left( a_\epsilon + K_\epsilon / 2, b_\epsilon + ||\epsilon||^2 / 2 \right). \]

6. \[ \sigma^2_{\epsilon,i} \propto \sigma^2_{\epsilon,i} \exp \left\{ -(y_i - \mu_i)^2 / (2\sigma^2_{\epsilon,i}) \right\} \left\{ \log(\sigma^2_{\epsilon,i}) - \eta_i \right\}^2 / (2\sigma^2_\epsilon), \]
   where \( \mu_i \) and \( \eta_i \) are the \( i \)-th components of \( C_X \theta_X \) and \( C_\epsilon \theta_\epsilon \) respectively.

7. \[ \sigma^2_k \propto \sigma^{-3}_k \exp \left\{ -b_k^2 / \left( 2\sigma^2_k \right) \right\} \left\{ \log(\sigma^2_k) - \xi_k \right\}^2 / (\sigma^2_\epsilon), \]
   where \( \xi_k \) is the \( k \)-th component of \( C_b \theta_b \).

All the above conditionals have an explicit form with the exception of the \( n + K_m \)
one-dimensional conditionals from 6 and 7. For these distributions we use the Metropolis- 
Hastings algorithm with a normal proposal distribution centered at the current value and 
small variance.

10. IMPLEMENTATION AND COMPUTATIONAL PERFORMANCE

Our MCMC simulation algorithm provided reliable results and is reasonably fast. To il-
istrate this, Figure 9 displays computation time (in minutes) versus sample size for 10,000.
Figure 10. MCMC histories for eight parameters of interest including the log of standard deviation of the error process and log of standard deviation of the shrinkage process, as well as some parameters modeling the different functions. The model uses 40 equally spaced knots for the mean and error variance functions and 4 subknots for the adaptive shrinkage process. Chains correspond to the Doppler function example using quadratic truncated polynomials to fit the mean and log variance function and linear truncated polynomial basis for fitting the shrinkage process. For clarity of presentation only every 20th simulated observation is displayed.

Simulations from the joint posterior distribution given the data. The mean function is modeled as a penalized spline with \( K_m \) knots with the log of the variance of the shrinkage process modeled as another penalized spline with \( K_b = 4 \) subknots. The log of the error variance is modeled as another penalized spline with \( K_\epsilon = K_m \) knots. Different symbols correspond to increasing number of knots: dots (\( K_m = 10 \)), asterisks (\( K_m = 20 \)), circles (\( K_m = 30 \)), x’s (\( K_m = 40 \)). Computation time increases roughly linearly with the number of observations with the same intercept, 0, and larger slopes for larger number of knots. Note that even with \( n = 1,000 \) observations and \( K_m = K_\epsilon = 40 \) simulation time is about two minutes.

For the models and datasets presented in this article, 10,000 simulations proved sufficient for accurate and reliable inference. Some parameters, such as the mean and variance functions tended to exhibit much better mixing properties than the variance components, \( \sigma_c^2 \) and \( \sigma_d^2 \), controlling the shrinkage processes. This is probably due to the reduced amount of information in the data about these parameters. The MCMC convergence and mixing properties were assessed by visual inspection of the chain histories of many parameters of interest. Since we started with good starting values of the parameters convergence is fast. Figure 10 displays the histories of eight parameters of interest for the model fitted to the
Doppler function described in Section 5. For clarity we display only every 100th sampled valued of the chain, but the same patterns would be observed in the full chain history. Similar good mixing properties have been noted in all the other examples presented in this article.

In the search for a good software platform we implemented our algorithms in MATLAB, C, and FORTRAN. Programs can be downloaded from [www.biostat.jhsph.edu/~ccrainic](http://www.biostat.jhsph.edu/~ccrainic). For our application and implementation there were no serious differences between C and FORTRAN, which have performed better than MATLAB by reducing computation time two to three times. The reported times in Figure 9 are for our implementation in C.

Given the complexity of the model, we were interested in improving the mixing properties of the Markov Chains. The computational trick described in Section 9 was the *sine qua non* method that allowed us to replace an intractable multivariate MH step with several tractable univariate MH steps. We used normal random walk proposal distributions centered at the current value and variance tuned to provide acceptance rates between 30% and 50%.

11. COMMENTS

Our joint models for heterogeneous mean and nonconstant variance are designed to extract the maximum amount of information from complex data and suggest simpler parametric models. Bayesian inference based on MCMC offers a natural and powerful framework for model fitting. However, MCMC in this context is not straightforward and implementation depends essentially on the ability of transforming intractable simulations from high dimensional distributions into tractable one dimensional updates, as described in Section 9. Our implementation provides accurate, fast and reproducible results while software is available. Actual simulation time increases only linearly with the number of observations and mixing properties are adequate, provided that proposal distribution variances are tuned for optimal acceptance rates.

ACKNOWLEDGMENTS

Carroll’s research was supported by a grant from the National Cancer Institute (CA-57030) 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).

[Received September 2004. Revised October 2006.]

REFERENCES


