Stat Comput (2013) 23:261–270
DOI 10.1007/s11222-011-9307-3

On conditional variance estimation in nonparametric regression

Siddhartha Chib · Edward Greenberg

Received: 9 February 2011 / Accepted: 28 November 2011 / Published online: 10 January 2012

Abstract In this paper we consider a nonparametric regression model in which the conditional variance function is assumed to vary smoothly with the predictor. We offer an easily implemented and fully Bayesian approach that involves the Markov chain Monte Carlo sampling of standard distributions. This method is based on a technique utilized by Kim, Shephard, and Chib (in Rev. Econ. Stud. **65**:361–393, 1998) for the stochastic volatility model. Although the (parametric or nonparametric) heteroscedastic regression and stochastic volatility models are quite different, they share the same structure as far as the estimation of the conditional variance function is concerned, a point that has been previously overlooked. Our method can be employed in the frequentist context and in Bayesian models more general than those considered in this paper. Illustrations of the method are provided.

Keywords Heteroscedastic errors · Cubic splines · Conditional variance functions · Nonparametric regression · Semiparametric regression

1 Introduction

In nonparametric regression, $y_i = \beta_{0g} + g(w_i) + u_i$, where y_i is the metric response, w_i is the predictor variable, $\mathbb{E}(y_i|w_i) = \beta_{0g} + g(w_i)$ the conditional mean, and u_i the Gaussian random error, the central objective is the estimation of the unknown function $g(\cdot)$. Inference about $g(\cdot)$ is usually conducted under homoscedasticity, the assumption that the conditional error variance $\sigma^2(w) = \text{var}(u|w)$ is constant. For some problems, however, this assumption is restrictive. It can also lead to misleading interval estimates of $g(w)$ if the assumption is incorrect. Besides, there are cases when one has direct interest in the conditional variance function as, for example, in applications where the second moment is a proxy for risk. For these reasons, models in which the conditional variance function varies smoothly with w, or another predictor z, are an important version of nonparametric regression.

From the frequentist perspective, inference in this more general nonparametric regression model has been considered by, for example, Ruppert et al. (2003, Chap. 14), Yu and Jones (2004), and Wasserman (2006, p. 87–89). In each instance, inference is based on a similar two-step procedure. In the first step, the mean function is estimated by a standard nonparametric method, such as local polynomial local likelihood (Yu and Jones 2004), local linear estimation (Wasserman 2006), or spline smoothing (Ruppert et al. 2003). The squared residuals around the mean function are then regressed on the covariates, again non-parametrically, to estimate the conditional variance function.

On the Bayesian side, the full posterior distribution of both functions has been subjected to analysis by Markov chain Monte Carlo (MCMC) methods by Chan et al. (2006) under a radial basis expansion of the unknown functions and Leslie et al. (2007) under a Dirichlet process mixture error distribution. A tuned Metropolis–Hastings (M–H) step is included in the sampling procedure to sample the non-standard conditional posterior distribution of the variance function. The proposal distribution in this step must be care-

S. Chib
Olin Business School, Washington University in St. Louis, Campus Box 1133, 1 Bookings Drive, St. Louis, MO 63130, USA
e-mail: chib@wustl.edu

E. Greenberg (✉)
Department of Economics, Washington University in St. Louis, Campus Box 1208, 1 Bookings Drive, St. Louis, MO 63130, USA
e-mail: edg@wustl.edu

fully tuned when a large number of knots are used in the basis expansion.

The aim of this paper is to provide a technique for sampling the posterior distribution without the involvement of an M–H step. Our method is based on the accurate approximation of the distribution of $\log(y - \beta_{0g} - g(w))^2$ by a mixture of normal distributions, as in the approach of Kim et al. (1998) in stochastic volatility models. Apparently, the use of this approach in the heteroscedastic regression setting is new. We note that our approach has potential applications in the frequentist setting and could be beneficially embedded in the procedures of Chan et al. (2006) and Leslie et al. (2007) when the number and placements of knots is not fixed at the outset. We defer consideration of these possibilities to future work in order to keep the focus on our main contribution.

The rest of the paper is organized as follows. In Sect. 2 we present the model and describe the inferential procedure. Section 3 has the examples and Sect. 4 some summary comments. The Appendix contains further details related to the method.

2 Model and inference

2.1 General procedure

Suppose that $y = (y_1, \ldots, y_n)$ is an independent sample from the model

$$y_i = \beta_{0g} + g(w_i) + \sigma(z_i)\varepsilon_i, \tag{2.1}$$

where β_{0g} is the intercept, w_i is a univariate predictor, $g(\cdot)$ and $\sigma(\cdot) > 0$ are general smooth functions of the predictor, z_i is either w_i or a different predictor, and $\varepsilon_i | w_i, z_i \sim \mathcal{N}(0, 1)$. Let

$$\log \sigma^2(z_i) = \beta_{0h} + h(z_i) \tag{2.2}$$

denote the conditional variance function, where $h(\cdot)$ is a general smooth function. It is possible to entertain multiple predictors in the model by, for example, assuming an additive form. With two predictors, we could suppose that

$$g(w) = \beta_{0g} + g_1(w_1) + g_2(w_2),$$

$$\log \sigma^2(z) = \beta_{0h} + h_1(z_1) + h_2(z_2).$$

Nonetheless, to minimize the notational burden and to highlight our main point, we do not consider multiple predictors until our examples.

Our method to estimate the variance function is straightforward: Given $\{\beta_{0g}, g(w_i)\}$, the model can be written as

$$\log\big(y_i - \beta_{0g} - g(w_i)\big)^2 = \beta_{0h} + h(z_i) + \varepsilon_i^*, \tag{2.3}$$

Table 1 Parameters (q_j, m_j, v_j^2) of the component normal distributions given by Omori et al. (2007) to approximate the log χ^2 distribution with 1 degree of freedom

j	q_j	m_j	v_j^2
1	0.00609	1.92677	0.11265
2	0.04775	1.34744	0.17788
3	0.13057	0.73504	0.26768
4	0.20674	0.02266	0.40611
5	0.22715	−0.85173	0.62699
6	0.18842	−1.97278	0.98583
7	0.12047	−3.46788	1.57469
8	0.05591	−5.55246	2.54498
9	0.01575	−8.68384	4.16591
10	0.00115	−14.65000	7.33342

where $\varepsilon_i^* = \log \varepsilon_i^2$ is distributed as the log of a chi-squared random variable with one degree of freedom and density

$$p(\varepsilon_i^*) = \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{\varepsilon_i^* - \exp(\varepsilon_i^*)}{2}\right\}, \quad \varepsilon_i^* \in R.$$

As in the stochastic volatility model of Kim et al. (1998), where one also encounters the same distribution, the difficulties of dealing with the log-chisquared distribution can be alleviated by an approximating mixture of seven normal distributions. We adopt a further, even more accurate, approximation to the log-chisquared distribution that has been calculated by Omori et al. (2007, p. 428). This approximation has ten normal components. In hierarchical form, with the introduction of latent component indicator variables $s_i \in \{1, 2, \ldots, 10\}$, this mixture distribution can be expressed as

$$\varepsilon_i^* | s_i \sim \mathcal{N}(m_{s_i}, v_{s_i}^2) \tag{2.4}$$

$$\Pr(s_i = j) = q_j, \quad j \le 10, \tag{2.5}$$

where the values of $(q_1, \ldots, q_{10})$, $(m_1, \ldots, m_{10})$, and $(v_1^2, \ldots, v_{10}^2)$ are given in Table 1.

The main point is that under this accurate approximation the posterior distribution of $(\beta_{0g}, \beta_g, \{s_i\}, \beta_{0h}, \beta_h)$ can be simulated through the sampling of standard distributions.

2.2 Data generating process

To describe the details, let

$$g(w) = (g(w_1), \ldots, g(w_n))' : n \times 1$$

and

$$h(z) = (h(z_1), \ldots, h(z_n))' : n \times 1$$

be the unknown function ordinates. We model these functions with cubic splines (see, for example, Green and Silverman 1994; Denison et al. 2002; Ruppert et al. 2003; Congdon 2007), where the knots are set at the equally spaced quantiles of the predictors and the number of knots is determined by comparing the marginal likelihood of models with different numbers of knots.

In our work we have favored the cubic spline basis that is given in Lancaster and Šalkauskas (1986, Sects. 3.7 and 4.2). We favor this basis because, as shown in Chib and Greenberg (2010), the identified basis coefficients are equal to the value of the unknown function at the corresponding knot. This connection with the function ordinates is helpful in formulating a prior distribution on the basis coefficients. In our prior we smooth the neighboring basis coefficients. The prior we use works rather well across problems, even when there are a large number of knots (and hence a large number of basis coefficients).

Under our basis functions, we can express the functions as

$$g(w) = B_g \beta_g, \tag{2.6}$$

$$h(w) = B_h \beta_h, \tag{2.7}$$

where B_f is the $n \times (M_f - 1)$ basis matrix and β_f are the cubic spline parameters, $f = g, h$. Additional details are given in the appendix.

If we now let

$$X_f = (i_n, B_f) \quad \text{and} \quad \beta_f^* = (\beta_{0f}, \beta_f')',$$

where i_n is an n-vector of ones, it follows that, independently across observations, the data generating process is

$$y_i = x_{gi}' \beta_g^* + \exp(0.5 x_{hi}' \beta_h^*) \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.8}$$

where x_{fi}' denotes the ith row of X_f. Equivalently, in vector-matrix notation we can write

$$y | \beta_g^*, \beta_h^* \sim \mathcal{N}_n(X_g \beta_g^*, D), \tag{2.9}$$

where

$$D = \operatorname{diag}(\exp(x_{h1}' \beta_h^*), \ldots, \exp(x_{hn}' \beta_h^*)).$$

2.3 Prior distribution

As mentioned above, the identified spline parameters in the basis are equal to the values of the unknown function at the corresponding knot. We utilize this connection in the prior to smooth the neighboring basis coefficients. In particular, we suppose that

$$\beta_f | \lambda_{fe}^2, \lambda_{fd}^2 \sim \mathcal{N}_{M_f - 1}(0, \Delta_f^{-1} T_f \Delta_f^{-1\prime}), \quad f = g, h,$$

where the quantities that appear in this expression are defined in the Appendix, and $(\lambda_{fe}^2, \lambda_{fd}^2)$ are hyperparameters that control the extent of smoothness. Independently, we also assume that the intercepts follow the prior distribution $\beta_{0f} \sim \mathcal{N}(a_{0f}, A_{00f})$, which leads to the joint distribution

$$\beta_f^* | \lambda_{fe}^2, \lambda_{fd}^2 \sim \mathcal{N}_{M_f}(a_{0f}^*, A_{0f}^*), \tag{2.10}$$

where

$$a_{0f}^* = \begin{pmatrix} a_{0f} \\ 0 \end{pmatrix} \quad \text{and} \quad A_{0f}^* = \begin{pmatrix} A_{00f} & 0 \\ 0 & \Delta_f^{-1} T_f \Delta_f^{-1\prime} \end{pmatrix}.$$

We further assume that the unconditional prior distribution of the smoothing parameters is inverse-gamma,

$$\begin{aligned}(\lambda_{fe}^2, \lambda_{fd}^2) \sim\ & \text{inv gamma}\left(\frac{\alpha_{fe0}}{2}, \frac{\delta_{fe0}}{2}\right) \\ & \times \text{inv gamma}\left(\frac{\alpha_{fd0}}{2}, \frac{\delta_{fd0}}{2}\right),\end{aligned} \tag{2.11}$$

for given values of the hyperparameters. Under these assumptions, therefore, the prior distribution has the form

$$\pi(\theta) = \prod_{f=g,h} p(\lambda_{fe}^2, \lambda_{fd}^2) p(\beta_f^* | \lambda_{fe}^2, \lambda_{fd}^2), \tag{2.12}$$

where $\theta_f = (\beta_f^*, \lambda_{fe}^2, \lambda_{fd}^2)$ and $\theta = (\theta_g, \theta_h)$.

We fix the hyperparameter values in the prior by an iterative simulation-based approach. Specifically, we draw parameters from the prior and then draw outcomes given the parameters. If the distribution of these simulated outcomes is not reasonable on a priori grounds, the hyperparameters are adjusted and the process repeated until there is a match between the distribution of the simulated data and the a priori judgements.

2.4 Posterior and MCMC sampling

We now describe the MCMC approach for sampling the posterior distribution given by the preceding prior distribution and the likelihood function of (2.9). The MCMC sampling is composed of four steps, each requiring the simulation of a tractable known distribution.

1. $\beta_g^* | y, \theta_{-\beta_g^*}$. Standard calculations show that

$$\beta_g^* | y, \theta_{-\beta_g^*} \sim \mathcal{N}\big(\hat{\beta}_g^*, A_g\big), \tag{2.13}$$

where

$$\hat{\beta}_g^* = A_g (A_{0g}^{*-1} a_{0g}^* + X_g' D^{-1} y),$$

$$A_g = (A_{0g}^{*-1} + X_g' D^{-1} X_g)^{-1}.$$

2. $\Pr(s_i = j|y_i, \theta), i \le n$. Now let $y^* = (y_1^*, \ldots, y_n^*)$, where $y_i^* = \log(y_i - x_{gi}'\beta_g^*)^2$, and let $s_i \in \{1, \ldots, 10\}$ denote the mixture component indicators. Conditioned on the parameters, it can be seen that the conditional posterior distribution of the mixture indicators is

$$\Pr(s_i = j|y, \theta) \propto q_j \mathcal{N}(y_i^*|m_j + x_{hi}'\beta_h^*, v_j^2), \quad j = 1, \ldots, 10, \quad i \le n. \tag{2.14}$$

3. $\beta_h^*|y, \theta_{-\beta_h^*}, \{s_i\}$. If we now let $s = (s_1, \ldots, s_n)$ denote the sampled values of the mixture indicators, then

$$y^*|\theta, s \sim \mathcal{N}(m_s + X_h\beta_h^*, V_s),$$

where $m_s = (m_{s_1}, \ldots, m_{s_n})$ and $V_s = \operatorname{diag}(v_{s_1}^2, \ldots, v_{s_n}^2)$. Therefore, it follows that

$$\beta_h^*|y, \theta_{-\beta_h^*} \sim \mathcal{N}\left(\hat{\beta}_h^*, A_h\right), \tag{2.15}$$

where

$$\hat{\beta}_h^* = A_h(A_{0h}^{*-1}a_{0h}^* + X_h'V_s^{-1}(y^* - m_s)),$$
$$A_h = \left(A_{0h}^{*-1} + X_h'V_s^{-1}X_h\right)^{-1}.$$

4. $\lambda_{fe}^2|y, \theta_{-\lambda_{fe}^2}$ and $\lambda_{fd}^2|y, \theta_{-\lambda_{fd}^2}$. Finally, with the prior distributions of β_f from (2.10) and the variances from (2.11), we get that

$$\lambda_{fe}^2|y, \theta_{-\lambda_{fe}^2} \sim \text{inv gamma}\left(\frac{\alpha_{fe0} + 2}{2}, \frac{\delta_{fe0} + \beta_f'\Delta_f'D_{f0}\Delta_f\beta_f}{2}\right),$$
$$\lambda_{fd}^2|y, \theta_{-\lambda_{fd}^2} \sim \text{inv gamma}\left(\frac{\alpha_{fd0} + M_f - 3}{2}, \frac{\delta_{fd0} + \beta_f'\Delta_f'D_{f1}\Delta_f\beta_f}{2}\right), \tag{2.16}$$

where

$$D_{f0} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0_{M_f-3} & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and}$$
$$D_{f1} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & I_{M_f-3} & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

2.5 Marginal likelihood

In practice, one would be interested in comparing models, for example, models with and without heteroscedasticity, or models with and without a nonparametric regression function. One can also compare models with different number of knots. When the number of contending models is not large, the comparisons can be in terms of the marginal likelihoods and pair-wise Bayes factors. The method of Chib (1995) can be readily employed for this purpose. We suppress the details in the interest of space. In other cases, one can also make use of model jump methods, such as those of Carlin and Chib (1995) and Green (1995). Leslie et al. (2007), for example, demonstrate the use of model jump methods to select both the number and placement of knots.

3 Examples

Example 1 As our first illustration, we consider data generated from a model in Yu and Jones (2004) with $n = 750$:

$$y_i = g(w_i) + \sigma(w_i)\varepsilon_i,$$

where

$$g(w) = w + 2\exp(-16w^2),$$
$$h(w) = \log(0.5) - w^2/8,$$

and the w_i are independently drawn from a standard normal distribution truncated to $[-2, 2]$. In our fitting, we estimate and compare several models in which g and h are estimated with varying number of knots. For each of these models, regardless of the number of knots, we suppose that the prior distribution on the smoothing parameters is given by the hyperparameters

$$(\alpha_{fe0}, \delta_{fe0}, \alpha_{fd0}, \delta_{fd0}) = (4.125, 2.005, 4.125, 2.005).$$

We employ the same hyperparameter values in the other examples below. Clearly, therefore, these values have not been chosen with a particular model in mind. Our MCMC simulations in this and the remaining examples consist of 20,000 iterations following a burn-in of 2000 iterations. We use the method of Chib (1995) to calculate the marginal likelihoods and select the most-favored model. In this example, the selected model is found to have 11 knots for g and 4 knots for h. The fitted functions are reproduced in Fig. 1. The figure depicts the true and estimated functions along with the 95% posterior credibility intervals. Since this example contains only one covariate, we plot $E(y|w) = \beta_{0g} + g(w)$ and $\sigma(w) = \exp(0.5[\beta_{0h} + h(w)])$. The figure shows that our fitting procedure recovers both functions quite accurately although there is greater posterior uncertainty surrounding the conditional standard deviation function. Calculation of the latter uncertainty is straightforward from our Bayesian MCMC perspective: we simply calculate the credibility intervals point-wise on the support of the

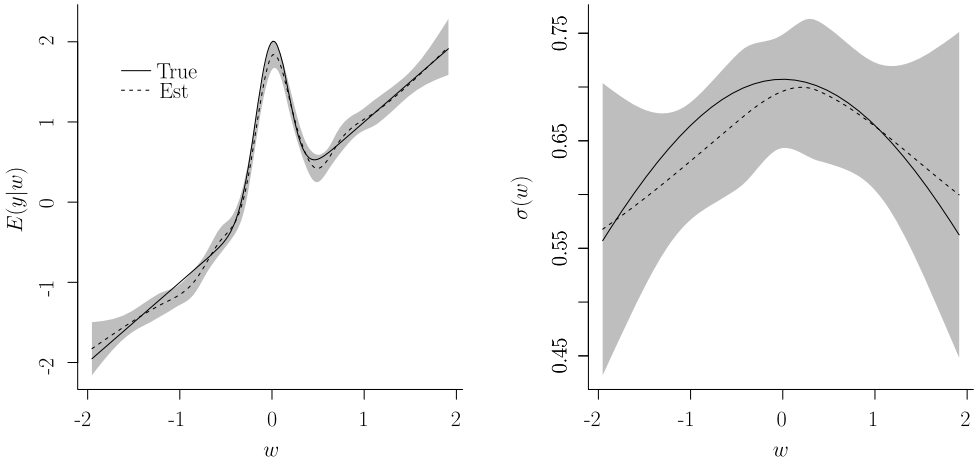

Fig. 1 True and estimated functions and 95% posterior density bands for Example 1

function from the posterior quantiles of the simulated function values. In the frequentist literature, the corresponding interval estimates of the conditional standard deviation function are almost never reported.

Example 2 In this example, we demonstrate how our fitting procedure performs when there is more than one predictor in the conditional mean and conditional standard deviation functions. We consider $n = 500$ observations from an additive model based on functions used by Chan et al. (2006) and Chib and Greenberg (2010):

$$g_1(w_1) = 1.5w_1,$$
$$g_2(w_2) = \{\mathcal{N}(w_1; 0.2, 0.004) + \mathcal{N}(w_1; 0.6, 0.1)\}/2,$$
$$g_3(w_3) = 1 + \sin(2\pi w_3),$$
$$g_4(w_4) = -w_4,$$
$$h_1(z_1) = z_1 + \sin(4\pi z_1),$$
$$h_2(z_2) = -1.5 - z_2 + \exp(-50(z_2 - 0.5)^2),$$
$$h_3(z_3) = -1.2 + z_3,$$

where the w_i are independently drawn from standard uniform distributions, $z_1 = w_1$, $z_2 = w_2$ and $z_3 = w_4$. We fit this model with the prior distribution for the smoothness parameters that was used in Example 1. The results of the function estimation summarized in Fig. 2 are based on equally-spaced knots of $(4, 12, 12, 4)$ for the mean function and $(18, 12, 4)$ for the variance function. We plot deviations from the mean for both true and estimated functions because the level of the functions are not identified when there is more than one covariate. These knots were selected by fitting models with different number of knots and finding the model that was best supported on the marginal likelihood criterion. As in the previous example, the estimates of the unknown functions are close to the true functions.

Example 3 For an example with real-world data, we consider data generated from the LIDAR (light detection and ranging) technique that is used to monitor the distribution of atmospheric pollutants. In these data, which are discussed by Ruppert et al. (2003, pp. 47–49, 264–266), Wasserman (2006, pp. 77–78) and Chan et al. (2006), the outcome y is the logarithm of the ratio of received light from two laser sources; the frequency of one source is equal to the resonance of mercury, and the second has a different frequency. The covariate 'range,' $w = z$, is the distance traveled before the light is reflected back to its source. The sample consists of 221 observations. The results from the fitting of our nonparametric regression model, under the same prior as in the previous examples and from the same MCMC sample size of 20,000, are presented in Fig. 3. These results are based on 5 knots for the mean function and 4 for the variance function. The observed scatter in the left panel of the plot suggests that the dispersion increases with the predictor. This feature is picked up in the estimated standard deviation function, which is plotted in the right panel of the plot. Note that the latter plot includes the posterior credibility interval of the standard deviation function. In contrast, the corresponding figure in Wasserman (2006, p. 78) does not have these bands, and the point estimates, though similar in shape, are larger from our analysis for larger values of the predictor.

4 Conclusions

Our purpose in this paper is to provide an easily applied method to estimate additive nonlinear conditional mean and conditional variance functions. By taking advantage of an accurate approximation to the $\log(\chi^2_1)$ distribution, simulation of the posterior distribution requires only the sampling of distributions that are easy to sample. We illustrate the method with proper prior distributions and utilize marginal likelihoods to compare specifications and to select the number of knots. The approach we have described is easily im-

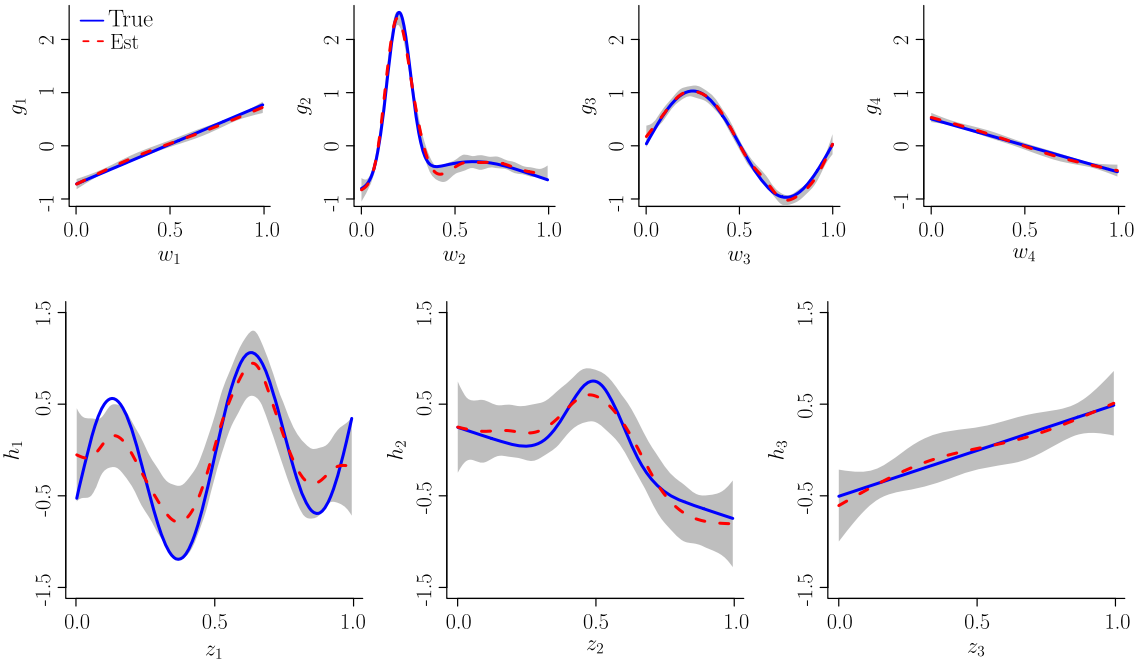

Fig. 2 True and estimated functions and 95% posterior density bands for Example 2

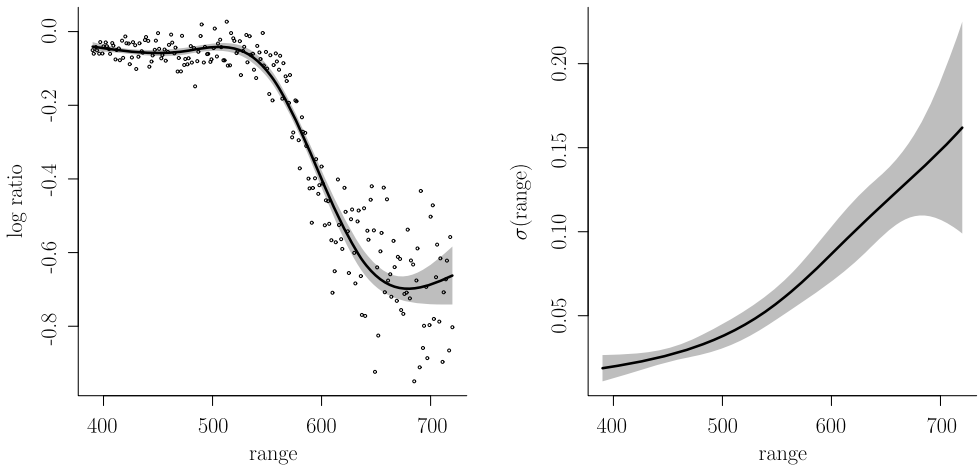

Fig. 3 Estimated functions, observed data, and 95% posterior density bands for LIDAR data

plemented and should facilitate the fitting of such models in practice. The parameters of the model are easy to interpret, and all the Bayesian inferential tools can be employed straightforwardly. The approach could be used in conjunction with nonparametric Bayesian models that incorporate variable selection and free knot points.

Acknowledgements The authors gratefully thank the editor and two referees for constructive comments and suggestions that improved the paper.

Appendix: Construction of the cubic spline basis matrix

Let $f(w) = (f(w_1), \ldots, f(w_n))$ denote the unknown function values at each of the sample values of the covariate $w = (w_1, \ldots, w_n)$, where f stands for either g or h, and let $\tau = (\tau_1, \ldots, \tau_{M_f})$ denote the set of knots, located between $\tau_1 = \min(w)$ and $\tau_{M_f} = \max(w)$. In our implementation, the knots are placed at equally spaced quantiles, and $f(w)$ is expressed in terms of a natural cubic spline basis as

$$f(w) = B_f \beta_f,$$

where B_f is a $n \times (M_f - 1)$ basis matrix and β_f is a $(M_f - 1)$ vector of cubic spline parameters. We now describe how this basis matrix is constructed (further details can be found in Chib and Greenberg 2010).

The basis functions are given by Φ_m and Ψ_m, $m = 1, \ldots, M_f$ which have compact support and are defined as

$$\Phi_m(a) = \begin{cases} 0, & a < \tau_{m-1}, \\ -(2/h_m^3)(a-\tau_{m-1})^2(a-\tau_m-0.5h_m), & \\ \quad \tau_{m-1} \le a < \tau_m, & \\ (2/h_{m+1}^3)(a-\tau_{m+1})^2(a-\tau_m+0.5h_{m+1}), & \\ \quad \tau_m \le a < \tau_{m+1}, & \\ 0, & a \ge \tau_{m+1}, \end{cases}$$

$$\Psi_m(a) = \begin{cases} 0, & a < \tau_{m-1}, \\ (1/h_m^2)(a-\tau_{m-1})^2(a-\tau_m), & \\ \quad \tau_{m-1} \le a < \tau_m, & \\ (1/h_{m+1}^2)(a-\tau_{m+1})^2(a-\tau_m), & \\ \quad \tau_m \le a < \tau_{m+1}, & \\ 0, & a \ge \tau_{m+1}, \end{cases}$$

where $h_m = \tau_m - \tau_{m-1}$ is the spacing between the $(m-1)$st and mth knots.

Next, we evaluate the basis functions for each element of w and each knot and arrange them in the $n \times M_f$ matrices Φ_f and Ψ_f

$$\Phi_f = \begin{pmatrix} \Phi_1(w_1) & \cdots & \Phi_{M_f}(w_1) \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \Phi_1(w_n) & \cdots & \Phi_{M_f}(w_n) \end{pmatrix},$$

$$\Psi_f = \begin{pmatrix} \Psi_1(w_1) & \cdots & \Psi_{M_f}(w_1) \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \Psi_1(w_n) & \cdots & \Psi_{M_f}(w_n) \end{pmatrix}.$$

Now let $\omega_m = h_m/(h_m + h_{m+1})$, $\mu_m = 1 - \omega_m$, and define the $(M_f \times M_f)$ tri-diagonal matrix A_f with 2 on the principal diagonal,

$$(\omega_2, \omega_3, \ldots, \omega_{M_f-1}, 1)$$

on the first sub-diagonal, and

$$(1, \mu_2, \mu_3, \ldots, \mu_{M_f-1})$$

on the first super-diagonal. Also define the $(M_f \times M_f)$ matrix C_f equal to 3 times a tri-diagonal matrix that has

$$\left(-\frac{1}{h_2}, \frac{\omega_2}{h_2} - \frac{\mu_2}{h_3}, \ldots, \frac{\omega_{M_f-1}}{h_{M_f-1}} - \frac{\mu_{M_f-1}}{h_{M_f}}, \frac{1}{h_{M_f}}\right)$$

on the principal diagonal,

$$\left(-\frac{\omega_2}{h_2}, -\frac{\omega_3}{h_3}, \ldots, -\frac{\omega_{M_f-1}}{h_{M_f-1}}, -\frac{1}{h_{M_f}}\right)$$

on the first sub-diagonal, and $(\frac{1}{h_2}, \frac{\mu_2}{h_3}, \ldots, \frac{\mu_{M_f-1}}{h_{M_f}})$ on the first super-diagonal. Let

$$B_f^\dagger = \Phi_f + \Psi_f A_f^{-1} C_f \equiv \left(b_1^\dagger, \ldots, b_{M_f}^\dagger\right), \tag{5.1}$$

where $b_m^\dagger \in \Re^n$ is the mth column of $B_f^\dagger$. To allow for more than one covariate, we impose the identification condition $\sum \beta_{fi} = 0$, which implies that $B_f^\dagger \beta^\dagger = B_f \beta_f$, where $B_f = (b_2^\dagger - b_1^\dagger, \ldots, b_{M_f}^\dagger - b_1^\dagger)$ and $\beta_f = (\beta_{f2}, \ldots, \beta_{fM_f})$.

An attractive property of this basis is that each component of β_f, $f = g, h$ is the value of the unknown function at the corresponding knot, i.e.,

$$\beta_f = \begin{pmatrix} f(\tau_2) \\ \vdots \\ f(\tau_{Mf}) \end{pmatrix}.$$

This property of the spline coefficients is particularly helpful in the Bayesian context because it can be used in the formulation of the prior distribution. An assumption of smoothness can be incorporated by assigning a prior mean of zero to both the differences in the ordinates at the end knots and the differences in slopes between adjacent knots at the interior knots. In particular, for the end knots, one can assume

$$\begin{aligned} &\frac{f(\tau_2) - f(\tau_1)}{h_2} \\ &\quad = \frac{2f(\tau_2) + f(\tau_3) + \ldots + f(\tau_{M_f})}{h_2} \sim \mathcal{N}(0, \lambda_{fe}^2), \\ &\frac{f(\tau_{M_f}) - f(\tau_{M_f-1})}{h_{M_f}} \sim \mathcal{N}(0, \lambda_{fe}^2), \end{aligned}$$

and for the interior knots that

$$\begin{aligned} &\frac{f(\tau_{m+1}) - f(\tau_m)}{h_{m+1}} - \frac{f(\tau_m) - f(\tau_{m-1})}{h_m} \sim \mathcal{N}(0, \lambda_{fd}^2), \\ &\quad m = 3, \ldots, M_f - 1. \end{aligned} \tag{5.2}$$

The parameters λ_{fe}^2 and λ_{fd}^2 are smoothness parameters in the sense that small variances smooth the function because the differences in coefficients are presumed to be small, while large variances have the opposite effect.

The foregoing assumptions imply that

$$\Delta_f \beta_f \mid \lambda_{fe}^2, \lambda_{fd}^2 \sim \mathcal{N}_{M_f-1}(0, T_f),$$

or that

$$\beta_f \mid \lambda_{fe}^2, \lambda_{fd}^2 \sim \mathcal{N}_{M_f-1}(0, \Delta_f^{-1} T_f \Delta_f^{-1\prime}), \tag{5.3}$$

where

$$\Delta_f = \begin{pmatrix} \frac{2}{h_2} & \frac{1}{h_2} & \cdots & \cdots & \cdots & \frac{1}{h_2} \\ \frac{1}{h_3} & -(\frac{1}{h_3}+\frac{1}{h_4}) & \frac{1}{h_4} & 0 & \cdots & 0 \\ 0 & \frac{1}{h_4} & -(\frac{1}{h_4}+\frac{1}{h_5}) & \frac{1}{h_5} & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{1}{h_{M_f-1}} & -(\frac{1}{h_{M_f-1}}+\frac{1}{h_{M_f}}) & \frac{1}{h_{M_f}} \\ 0 & \cdots & 0 & 0 & -\frac{1}{h_{M_f}} & \frac{1}{h_{M_f}} \end{pmatrix},$$

and $T_f = \mathrm{diag}(\lambda_{fe}^2, \lambda_{fd}^2 I_{M-3}, \lambda_{fe}^2)$.

We complete this Appendix by providing R-code for calculating the quantities mentioned in this Appendix. The code works with a single predictor. One can suitably loop over this code when there is more than one predictor.

```
makesplinedat = function(w,K,t2,A0inv_) {
    # w is a n*1 matrix; K is the M*1 matrix of knots
    # t2 is 1*2 matrix of t2e and t2d, and A0inv_ is a k0*k0 matrix
    n = dim(w)[1];
    k0 = dim(A0inv_)[1];
    M = dim(K)[1];
    k1 = M - 1;
    Klag = matrix(0,nr = M);
    Klag[2:M] = K[1:(M-1)];
    h = K - Klag;
    B0inv_ = matrix(0,(k0+k1),(k0+k1));
    B0inv_[1:k0,1:k0] = A0inv_;
    Phi = matrix(0,n,M);
    Psi = matrix(0,n,M);
    k = 1;
    hk11 = h[k+1];
    Kk = K[k];
    Kk11 = K[k+1];
    for (i in 1:n) {
        wi = w[i];
        if (wi >= Kk & wi < Kk11) {
            Phi[i,k] = (2.0/(hk11*hk11*hk11))*(wi-Kk+hk11/2.0)*(wi-Kk11)*(wi-Kk11);
            Psi[i,k] = (1.0/(hk11*hk11))*(wi-Kk)*(wi-Kk11)*(wi-Kk11);
            }
    }
    for (k in 2:(M-1)) {
        hk = h[k];
        hk11 = h[k+1];
        Kk1 = K[k-1];
        Kk = K[k];
        Kk11 = K[k+1];
        for (i in 1:n) {
            wi = w[i];
            if (wi >= Kk1 & wi < Kk) {
                Phi[i,k] = -(2.0/(hk*hk*hk))*(wi-Kk1)*(wi-Kk1) * (wi-Kk-hk/2.0);
                Psi[i,k] = (1.0/(hk*hk))*(wi-Kk1)*(wi-Kk1) * (wi-Kk);
            }
            if  (wi >= Kk & wi < Kk11) {
                Phi[i,k] = (2.0/(hk11*hk11*hk11))*(wi-Kk+hk11/2)*(wi-Kk11)*(wi-Kk11);
                Psi[i,k] = (1.0/(hk11*hk11))*(wi-Kk)*(wi-Kk11)*(wi-Kk11);
            }
        }
    }
    k = M;
    hk = h[k];
    Kk1 = K[k-1];
```

```
Kk = K[k];
for (i in 1:n) {
    wi = w[i];
    if (wi <= Kk & wi > Kk1) {
        Phi[i,k] = -(2.0/(hk*hk*hk))*(wi-Kk1)*(wi-Kk1) * (wi-Kk-hk/2.0);
        Psi[i,k] = (1.0/(hk*hk))*(wi-Kk1)*(wi-Kk1) * (wi-Kk);
    }
}

A = matrix(0,M,M);
A[1,1] = 2.0;
A[1,2] = 1.0;
A[M,M-1] = 1.0;
A[M,M] = 2.0;
for (r in 2:(M-1)) {
    lamk = h[r]/( h[r]+h[r+1] );
    muk = 1.0 - lamk;
    A[r,r-1] = lamk;
    A[r,r] = 2.0;
    A[r,r+1] = muk;
}

C = matrix(0,M,M);
C[1,1] = -1.0/h[2];
C[1,2] = 1.0/h[2];
C[M,(M-1)] = -1.0/h[M];
C[M,M] = 1.0/h[M];
k = 2;
for (r in 2:(M-1)) {
    lamk = h[k]/( h[k]+h[k+1] );
    muk = 1.0 - lamk;
    C[r,(r-1)] = -lamk/h[k];
    C[r,r] = lamk/h[k] - muk/h[k+1];
    C[r,(r+1)] = muk/h[k+1];
    k = k + 1;
}
C = (3.0)*C;

Z = matrix(0,n,k1+1);
B1 = matrix(0,n,k1);
c = 1;
cx = 1;
Z[,c:(c+M-1)] = Phi + Psi %*% solve(A) %*% C;
B1[,cx:(cx+M-2)] = Z[,(c+1):(c+1+M-2)] - Z[,c];

D = matrix(0,M-1,M-1);
D[1,1] = 2.0/h[2];
D[1,2:(M-1)] = (1/h[2])*rep(1,(M-2));
D[(M-1),(M-2)] = -1.0/h[M];
D[(M-1),(M-1)] = 1.0/h[M];
for (r in 2:(M-2)) {
    lamr = -1/h[(r+1)] - 1/h[(r+2)];
    D[r,(r-1)] = 1/h[(r+1)];
    D[r,r] = lamr;
    D[r,(r+1)] = 1/h[(r+2)];
}

j = 1;
r = k0+1;
c = k0+1;
Tinvj = (1/t2[j,2])*diag((M-1));
Tinvj[1,1] = 1/t2[j,1];
Tinvj[(M-1),(M-1)] = 1/t2[j,1];
```

```
    B0invj = t(D) %*% Tinvj %*% D;
    B0inv_[r:(r+M-2),c:(c+M-2)] = B0invj;

    return(list(B1 = B1, Delta = D, B0inv_ = B0inv_));
}
```

References

Carlin, B.P., Chib, S.: Bayesian model choice via Markov chain Monte Carlo methods. J. R. Stat. Soc. B **57**, 473–484 (1995)

Chan, D., Kohn, R., Nott, D., Kirby, C.: Locally adaptive semiparametric estimation of the mean and variance functions in regression models. J. Comput. Graph. Stat. **15**(4), 915–936 (2006)

Chib, S.: Marginal likelihood from the Gibbs output. J. Am. Stat. Assoc. **90**(432), 1313–1321 (1995)

Chib, S., Greenberg, E.: Additive cubic spline regression with Dirichlet process mixture errors. J. Econom. **156**, 322–336 (2010)

Congdon, P.: Bayesian Statistical Modelling, 2nd edn. Wiley Series in Probability and Statistics. Wiley, Chichester (2007)

Denison, D.G.T., Holmes, C.C., Mallick, B.K., Smith, A.F.M.: Bayesian Methods for Nonlinear Classification and Regression. Wiley Series in Probability and Statistics. John Wiley, Chichester (2002)

Green, P., Silverman, B.W.: Nonparameteric Regression and Generalized Linear Models: A Roughness Penalty Approach. Chapman & Hall, London (1994)

Green, P.J.: Reversible jump MCMC computation and Bayesian model determination. Biometrika **82**(4), 711–732 (1995)

Kim, S., Shephard, N., Chib, S.: Stochastic volatility: Likelihood inference and comparison with ARCH models. Rev. Econ. Stud. **65**, 361–393 (1998)

Lancaster, P., Šalkauskas, K.: Curve and Surface Fitting: An Introduction. Academic Press, San Diego (1986)

Leslie, D.S., Kohn, R., Nott, D.J.: A general approach to heteroscedastic linear regression. Stat. Comput. **17**(2), 131–146 (2007)

Omori, Y., Chib, S., Shephard, N., Nakajima, J.: Stochastic volatility with leverage: Fast and efficient likelihood inference. J. Econom. **140**, 425–449 (2007)

Ruppert, D., Wand, M.P., Carroll, R.J.: Semiparametric Regression. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, New York (2003)

Wasserman, L.: All of Nonparametric Statistics. Springer Texts in Statistics. Springer, New York (2006)

Yu, K., Jones, M.C.: Likelihood-based local linear estimation of the conditional variance function. J. Am. Stat. Assoc. **99**(465), 139–144 (2004)