J. R. Statist. Soc. B (1999)
61, *Part* 4, *pp.* 863–879

Bayesian inference for semiparametric regression using a Fourier representation

Peter J. Lenk

University of Michigan, Ann Arbor, USA

[Received July 1996. Final revision February 1999]

Summary. This paper presents the Bayesian analysis of a semiparametric regression model that consists of parametric and nonparametric components. The nonparametric component is represented with a Fourier series where the Fourier coefficients are assumed *a priori* to have zero means and to decay to 0 in probability at either algebraic or geometric rates. The rate of decay controls the smoothness of the response function. The posterior analysis automatically selects the amount of smoothing that is coherent with the model and data. Posterior probabilities of the parametric and semiparametric models provide a method for testing the parametric model against a non-specific alternative. The Bayes estimator's mean integrated squared error compares favourably with the theoretically optimal estimator for kernel regression.

Keywords: Markov chain Monte Carlo methods; Model choice; Smoothing

1. Introduction

This paper considers the Bayesian analysis of the semiparametric additive model

$$y_i = d_i'\beta + g(x_i) + \epsilon_i \qquad \text{for } a \leqslant x_i \leqslant b \text{ and } i = 1, \ldots, n \tag{1}$$

where $\{d_i\}$ are fixed known p-vectors, β is a p-vector of parameters that includes the y-intercept, g is an unparameterized function defined on the interval $[a, b]$ and $\{\epsilon_i\}$ are a random sample from a normal distribution with mean 0 and variance σ^2. This paper will refer to $d'\beta$ as the 'parametric component' and to g as the 'nonparametric component'. Speckman (1988) reviewed penalized least squares methods of estimating these models and proposed a kernel method based on the residuals.

The parametric component frequently represents the researcher's beliefs about the correct model before analysing the data. The nonparametric component allows the posterior regression estimator to deviate from the parametric model. The semiparametric model permits testing the adequacy of the parametric model relative to non-specific alternatives in x with Jeffreys's hypothesis testing framework (see Kass and Raftery (1995)), namely, the posterior probabilities of parametric and semiparametric models are used in a decision theoretic framework.

This paper represents the nonparametric component with a Fourier series:

$$g(x) = \sum_{k=1}^{\infty} \theta_k \, \phi_k(x) \qquad \text{for } a \leqslant x \leqslant b, \tag{2}$$

Address for correspondence: Peter J. Lenk, University of Michigan Business School, Ann Arbor, MI 48109-1234, USA.
E-mail: plenk@umich.edu

© 1999 Royal Statistical Society 1369–7412/99/61863

$$\phi_k(x) = \left(\frac{2}{b-a}\right)^{1/2} \cos\left\{\pi k\left(\frac{x-a}{b-a}\right)\right\} \qquad \text{for } k = 1, 2, \ldots \text{ and } a \leqslant x \leqslant b, \quad (3)$$

$$\theta_k = \int_a^b g(x)\,\phi(x)\,\mathrm{d}x.$$

The cosine functions form an orthonormal basis for the piecewise continuous functions on $[a, b]$ (Kreider *et al.* (1966), pages 349–353). Because the constant function is included in the parametric component, the model has the implicit constraint that g integrates to 0.

The cosine terms have a natural ordering that relates to the smoothness of g: smooth g will not have high frequency components. In fact, if g is m times differentiable and $g^{(m)} \in L^1[a, b]$, then $\theta_k = o(k^{-m})$ (Katznelson (1976), p. 24). With this in mind, *a priori* $\{\theta_k\}$ are mutually independent and decay to 0 in probability:

$$\theta_k \sim N\{0, \tau^2 \exp(-\gamma c_k)\},$$

the normal distribution with mean 0 and variance $\tau^2 \exp(-\gamma c_k)$,

$$c_k = \begin{cases} k & \text{for the 'geometric smoother',} \\ \log(k) & \text{for the 'algebraic smoother',} \end{cases} \quad (4)$$

$$\gamma > c = \begin{cases} 0 & \text{for the geometric smoother,} \\ 1 & \text{for the algebraic smoother.} \end{cases} \quad (5)$$

The condition in equation (5) is needed for the Fourier series in equation (2) to converge in probability. The support for the prior distribution of g is the space of piecewise continuous functions on $[a, b]$ where the decay of the coefficients controls the number of derivatives.

The variances of $\{\theta_k\}$ have two forms in equation (4) which differ qualitatively. The algebraic smoother concentrates its prior mass on functions with γ or fewer derivatives, whereas the geometric smoother has prior mass on analytical functions. The parameter τ^2 can be viewed as the 'global' uncertainty about g: it determines the trade-off between the prior and the likelihood. The parameter γ determines the rate of decay of the Fourier coefficients, and thus the smoothness of g. Clearly, other basis functions or other priors could be used to model g. The representation and prior in this paper work together to express smoothness. This representation of g is less useful for expressing other types of prior knowledge, such as g is larger in one region than in another.

The parametric component should frequently include parametric functions of x. For example, a researcher may be interested in testing the adequacy of a hypothesized model or may have prior information, such as beliefs about a changepoint, that cannot be expressed by the nonparametric component. Also, under the cosine representation the odd derivatives of g are 0 at the end points, which may inadequately represent the phenomenon. The model specified by equations (1) and (2) is unidentified if the parametric component contains a function, say h, of x. Then the Fourier coefficients for g and h are indeterminate. However, the prior distribution for $\{\theta_k\}$ has the identifying restriction that their means are 0. Hence, the non-zero coefficients of h cannot be added to those of g without violating the prior specification.

To complete the model's specification, β, σ^2, τ^2 and γ are mutually independent of each other and of $\{\theta_k\}$, and

$$\beta \sim N(b_0, B_0),$$

the normal distribution with mean b_0 and covariance matrix B_0,

$$\sigma^2 \sim \mathrm{IG}(r_0/2,\, s_0/2),$$

the inverse gamma distribution with shape $r_0/2$ and scale $s_0/2$,

$$\tau^2 \sim \mathrm{IG}(u_0/2,\, v_0/2),$$
$$\gamma \sim E(w_0),$$

the exponential distribution with mean $1/w_0$, for the geometric smoother, and

$$\gamma \sim 1 + E(w_0)$$

for the algebraic smoother. The priors for β and σ, though standard, were selected for mathematical expediency. Other priors could be used without affecting the analysis of g, provided that β and σ are independent of $\{\theta_k\}$.

Smoothing splines are a closely related alternative model for g. Kimeldorf and Wahba (1970) demonstrated the correspondence between smoothing splines and a Bayesian estimator by using Gaussian process priors, and Wahba (1978) derived polynomial splines as the limiting case of the posterior mean of an integrated Wiener process. She also introduced a cross-validation method of choosing the smoothing parameter.

A number of recent papers specify g to be either polynomial splines or local polynomial functions where the coefficients of the basis functions and, to a varying extent, the number of knots or their locations are unknown, i.e. g is *a priori* a spline or local polynomial function. For example, Smith and Kohn (1996) began with a large number of knots and used Bayesian variable selection to find the most significant knots. Wong and Kohn (1996) used a similar method for an additive semiparametric model that allows for outliers and Smith *et al.* (1998) extended the model to autocorrelated errors. Denison *et al.* (1998) used a reversible jump Markov chain to determine the number and location of knots for local polynomial regression estimators. One difference between Wahba's approach and the more recent methods is that she begins with a prior distribution with support on a function space and derives the polynomial spline as the limit of posterior means, whereas the later papers specify flexible functional forms for g.

A natural extension of local polynomials is wavelets (Antoniadis *et al.*, 1994), which form a local basis. Chipman *et al.* (1997) and Clyde *et al.* (1998) have provided Bayesian methods for selecting the number and location of the basis functions.

The specification for g in this paper is closely related to that proposed in the empirical section of Wahba (1983). There, the representation of g has both sines and cosines, which restricts the analysis to periodic functions with $g(a) = g(b)$. She obtained a penalized maximum likelihood estimator and selected the smoothing parameter $\lambda = \tau^{-2}$ by cross-validation for fixed γ in the algebraic smoother. Gallant and Monahan (1985) also used sines and cosines with independent, doubly exponentially distributed Fourier coefficients. Their analysis was conditional on the smoothing parameters. In contrast, this paper treats both τ and γ as unknown, implements a fully Bayesian procedure and illustrates Bayesian hypothesis testing to confirm the adequacy of a parametric model. In related work, Lenk (1991, 1993) implemented a Bayesian, nonparametric density estimator by modelling the density with a logistic transform of g.

The next section shows that, conditionally on β, σ, τ and γ, the mean integrated squared error (MISE) of the Bayes estimator can converge to 0 at a rate that is favourable to optimal kernel estimators. Section 3 describes the posterior inference unconditionally on the smoothing parameters, and Section 4 presents a simulation study and an application. The application

illustrates a use of the semiparametric model and posterior probabilities to confirm the adequacy of a parametric model to non-specific alternatives. Section 5 concludes with a discussion.

The data and program that were used for the analysis in this paper can be obtained from

`http://www.blackwellpublishers.co.uk/rss/`

2. Conditional inference

In applications the infinite series in equation (2) is truncated: $g_K = \Sigma_{k=1}^{K} \theta_k \phi_k$ for $K < n$. Conceptually, Bayesian inference can allow more than $n - 1$ terms; however, this choice leads to aliasing, which complicates the computations without improving the analysis. For a sample of size n the model in equations (1) and (2), using the truncated representation g_K, can be written as

$$Y = D\beta + \Phi\Theta + \epsilon \tag{6}$$

where Y is the vector of n observations, D is an $n \times p$ design matrix, Φ is the $n \times K$ matrix with (i, k) entry $\phi_k(x_i)$, Θ is the vector of Fourier coefficients and ϵ is $N(0, \sigma^2 I)$. Define $\mathrm{var}(\Theta) = \tau^2\Psi$ where Ψ is a diagonal matrix with $\exp(-\gamma c_k)$ for the (k, k) entry. The design is orthogonal when the x_i are fixed at

$$x_i = a + (b - a)\frac{2i - 1}{2n} \qquad \text{for } i = 1, \ldots, n,$$

so $\Phi'\Phi = \{n/(b - a)\}I$ if $K \leqslant n$.

The prior specification for Θ induces a prior distribution for g_K, the truncated representation. It is a zero-mean Gaussian process with covariance function

$$\mathrm{cov}\{g_K(u), g_K(x)\} = \tau^2 \sum_{k=1}^{K} \exp(-\gamma c_k)\, \phi_k(u)\, \phi_k(x).$$

The covariance function is symmetric in its arguments, is non-stationary and integrates to 0. Figs 1(a) and 1(c) graph the covariance function on the unit square for the geometric and algebraic smoother respectively, with $K = 49$, $\tau^2 = 2.5$ and $\gamma = 1$. Both models result in a saddle shape with their maxima at (0, 0) and (1, 1). The parameter τ^2 controls the height of the saddle, and γ controls its width. The algebraic smoother has a sharp ridge along the diagonal whereas the geometric smoother is rounder; thus, the algebraic smoother produces a rougher estimator.

Conditionally on β, σ, τ and γ, the posterior distribution of Θ is normal with variance $\Psi_n = (\sigma^{-2}\Phi'\Phi + \tau^{-2}\Psi^{-1})^{-1}$ and mean $\nu_n = \sigma^{-2}\Psi_n\Phi'(Y - D\beta)$. Further, assuming an orthogonal design on $[a, b]$ results in mutually independent Fourier coefficients with

$$\mathrm{var}(\theta_k | Y) \equiv \psi_{n,k} = \frac{(b - a)\sigma^2\tau^2 \exp(-\gamma c_k)}{(b - a)\sigma^2 + n\tau^2 \exp(-\gamma c_k)},$$

$$E(\theta_k | Y) \equiv \nu_{n,k} = \sigma^{-2}\psi_{n,k} \sum_{i=1}^{n} (y_i - d_i'\beta)\, \phi_k(x_i).$$

The conditional series estimator for orthogonal designs is

$$\tilde{g}(x) = E\{g(x) | Y, \beta, \sigma, \tau, \gamma\} = \sum_{k=1}^{K} w_{n,k}\hat{\theta}_k\, \phi_k(x), \tag{7}$$

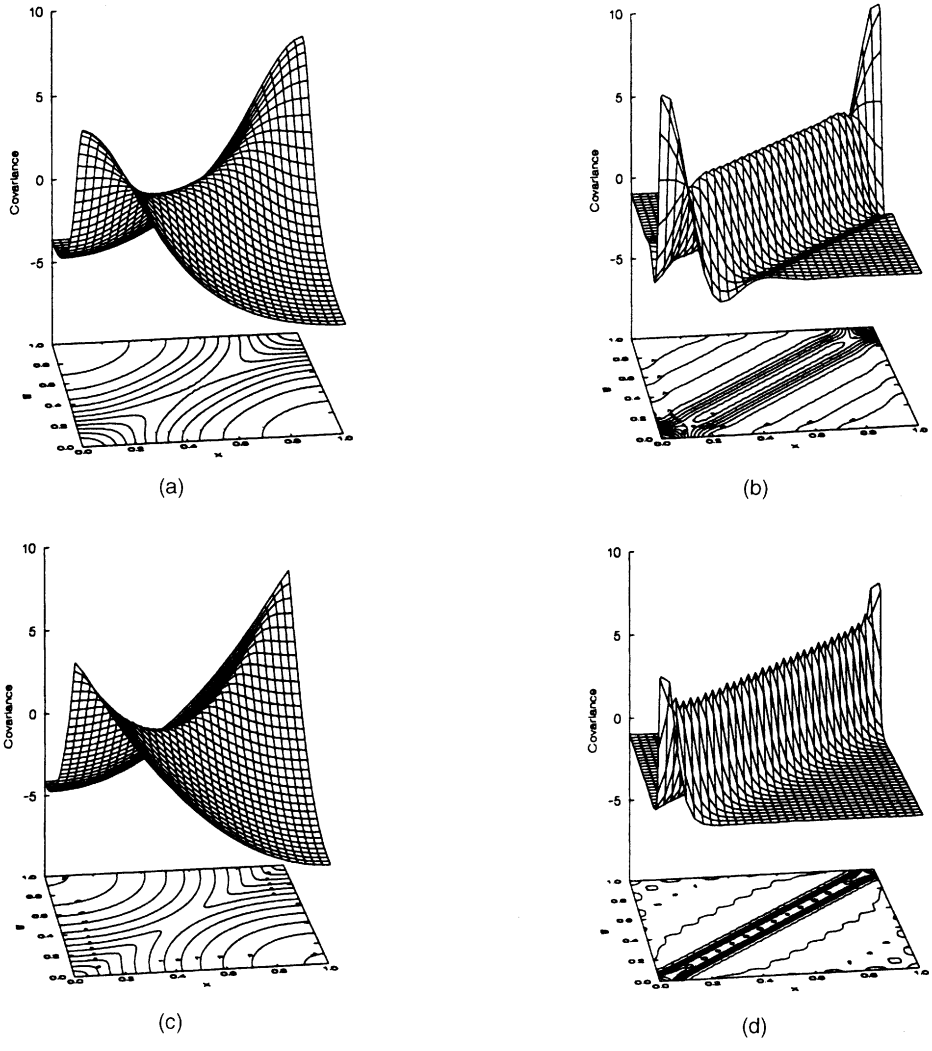

Fig. 1. Covariance function of the nonparametric component and the implied smoothing kernel: (a) geometric smoother; (b) kernel for the geometric smoother after 50 observations; (c) algebraic smoother; (d) kernel for the algebraic smoother after 50 observations

$$w_{n,k} = \frac{n}{\sigma^2}\,\psi_{n,k} = \frac{n(b-a)\tau^2\,\exp(-\gamma c_k)}{(b-a)\sigma^2 + n\tau^2\,\exp(-\gamma c_k)}, \tag{8}$$

$$\hat{\theta}_k = n^{-1}\sum_{i=1}^{n}(y_i - d_i'\beta)\,\phi_k(x_i). \tag{9}$$

This estimator differs from the standard series estimator $\Sigma_{k=1}^{K'}\hat{\theta}_k\,\phi_k(x)$ (Härdle (1990), pages 50–54) by weighting the estimated Fourier coefficients. Rutkowski (1982) showed that the standard series estimator is consistent if $K' = O(n^{1-s})$ for some $0 < s < 1$. When n is small, relatively few basis functions are used with the standard estimator, thus restricting its form. In contrast, the Bayes estimator in equation (7) can include as many basis functions as possible,

up to aliasing, but shrinks the maximum likelihood estimators of the Fourier coefficients towards 0 (see Hall and Titterington (1987)).

Most nonparametric regression procedures are either explicitly or implicitly kernel or approximate kernel methods (see Silverman (1985)). The conditional series estimator can also be expressed as a kernel estimator

$$\tilde{g}(x) = n^{-1} \sum_{i=1}^{n} (y_i - d_i' \beta)\, C_{K,n}(x_i, x)$$

with kernel

$$C_{K,n}(u, x) = \sum_{k=1}^{K} w_{n,k}\, \phi_k(u)\, \phi_k(x)$$

which is proportional to the posterior covariance function of g. $C_{K,n}$ is graphed in Figs 1(b) and 1(d) for the geometric and algebraic smoothers respectively, for $n = 50$. Compared with the prior covariance, the seat of the saddle is taller and narrower. The kernel function puts more weight on observations near the end points, thus ameliorating end effects. Unlike standard kernels, $C_{K,n}(u, x)$ is not stationary, takes negative values and integrates to 0 instead of to 1. However, $\{C_{K,n}\}$ is a delta sequence, as are standard kernels.

The following theorem provides a bound for the conditional series estimator's sampling variance and bias, the two components of the MISE:

$$E\left[\int_a^b \{\tilde{g}(x) - g(x)\}^2\, \mathrm{d}x\right] = \int_a^b \mathrm{var}\{\tilde{g}(x)\}\, \mathrm{d}x + \int_a^b [E\{\tilde{g}(x)\} - g(x)]^2\, \mathrm{d}x.$$

Theorem 1 has three interesting features. First, Bayesian estimators need not be consistent in infinite dimensional parameter spaces (Diaconis and Freedman, 1986a, b). The theorem shows that the MISE of the posterior mean of g converges to 0 when β, σ^2, τ^2 and γ are known. When they are unknown, their likelihood satisfies the usual regularity conditions, so their posterior distribution converges to a point mass at their true values at a faster rate than that for the MISE. Second, the MISE converges to 0 for fixed τ^2 and γ, unlike kernel regression, which requires the smoothing parameter to be a function of the sample size. Third, if the true function is analytical, then the MISE for the geometric smoother can obtain the rate $O\{n^{-1} \log(n)\}$, which is faster than the theoretically optimal rate of $O(n^{-4/5})$ for kernel estimators (see Stone (1982)).

Theorem 1. Consider the estimator $\tilde{g}$ given in equations (7)–(9). Assume the following.

(a) $\{x_i\}$ has an orthogonal design on $[a, b]$. Without loss of generality, assume that $a = 0$, $b = 1$ and $x_i = (2i - 1)/2n$ for $i = 1, \ldots, n$.
(b) $K = n - 1$.
(c) g is Lipschitz continuous: there is a $C_0 > 0$ such that $|g(x) - g(y)| < C_0|x - y|$ for all x and y in $[0, 1]$.
(d) There are positive C, M and ρ such that $|\theta_k| \leqslant C \exp(-\rho c_k)$ for all $k > M$. For the algebraic smoother, $\rho > 2$.
(e) $\gamma > 0$ for the geometric smoother, and $\gamma > 1$ for the algebraic smoother.

Then the integrated variance is bounded by

$$\int_0^1 \mathrm{var}\{\tilde{g}(x)\}\, \mathrm{d}x \leqslant \begin{cases} O\{\log(n)/n\} & \text{for the geometric smoother,} \\ O(n^{-(1-1/\gamma)}) & \text{for the algebraic smoother.} \end{cases}$$

The bias is uniformly bounded in x by

$$|E\{\tilde{g}(x) - g(x)\}| \leqslant \begin{cases} O\{\log(n)/n\} & \text{for the geometric smoother when } \rho \geqslant \gamma, \\ O(n^{-\rho/\gamma}) & \text{for the geometric smoother when } \rho < \gamma, \\ O(n^{-(\gamma-1)/\gamma}) & \text{for the algebraic smoother when } \rho \geqslant \gamma, \\ O(n^{-(\rho-1)/\gamma}) & \text{for the algebraic smoother when } \rho < \gamma. \end{cases}$$

The proof of theorem 1 is in Appendix A.

3. Posterior inference

The posterior inference is performed via Markov chain Monte Carlo (MCMC) sampling (Gelfand and Smith, 1990). The sequence of draws from the full conditional distributions for β, Θ, σ^2 and τ^2 follow standard, normal–inverse gamma development for linear models with normally distributed errors and will not be presented here. The only non-standard distribution is for the smoothing parameter γ. Its random variables are generated by using 'slice sampling' (Damien *et al.*, 1999; Polson, 1996), which is a Markov chain method to decompose a complex density into the product of uniform and exponential densities. Ignoring constants, the full conditional density of γ is

$$\exp\left\{-w_K\gamma - \frac{1}{2\tau^2}\sum_{k=1}^{K}\theta_k^2\exp(\gamma c_k)\right\} I[c \leqslant \gamma] \tag{10}$$

where $w_K = w_0 - 0.5\,\Sigma_{k=1}^{K} c_k$ and c is given in equation (5). Slice sampling introduces K uniform random variables V_k such that their joint density with γ is

$$f(V_1, \ldots, V_K, \gamma) \propto \sum_{k=1}^{K} I[0 \leqslant V_k \leqslant a_k]\exp(-w_K\gamma)\, I[c \leqslant \gamma], \tag{11}$$

$$a_k = \exp\left\{-\frac{\theta_k^2}{2\tau^2}\exp(c_k\gamma)\right\}. \tag{12}$$

One can readily verify that the marginal density of γ from equation (11) is the full conditional in equation (10). Let $\gamma^{(i)}$, $\tau^{(i)}$ and $\{\theta_k^{(i)}\}$ be draws on iteration i of the Markov chain, and let $a_k^{(i)}$ be the corresponding value of a_k in equation (12) with these values substituted for the parameters. Generate $v_k^{(i)} = u_k a_k^{(i)}$ where u_k is a standard uniform random deviate. The conditional distribution of γ given $\{v_k^{(i)}\}$ is

$$f(\gamma|\{v_k^{(i)}\}, \gamma^{(i)}) \propto \exp(-w_K\gamma) \qquad \text{for } c \leqslant \gamma \leqslant d = \min_{k=1\ldots K}(d_k),$$

$$d_k = \gamma^{(i)} + \frac{1}{c_k}\ln\left\{1 - 2\left(\frac{\tau^{(i)}}{\theta_k^{(i)}}\right)^2\exp(-c_k\gamma^{(i)})\ln(u_k)\right\},$$

which has cumulative distribution function

$$F(\gamma) = \frac{\exp(-w_K c) - \exp(-w_k\gamma)}{\exp(-w_K c) - \exp(-w_k d)} \qquad \text{for } c \leqslant \gamma \leqslant d.$$

Inverting this cumulative distribution function, generate

$$\gamma^{(i+1)} = -\frac{1}{w_K}\ln[\exp(-w_k c) - u\{\exp(-w_K c) - \exp(-w_k d)\}]$$

where u is a standard uniform random deviate.

The adequacy of the parametric *versus* the semiparametric models can be tested by computing Bayes factors or the posterior probability of the models. This paper tests

$$H_1\colon\ Y_i = d_i'\beta + \epsilon_i \quad \textit{versus} \quad H_2\colon\ Y_i = d_i'\beta + g(x_i) + \epsilon_i \tag{13}$$

by computing the marginal density of the data via the method of Gelfand and Dey (1994). This procedure requires running independent chains for the two models. Let Ω_j be the unknown parameters for model H_j, let $f_j(Y|\Omega_j)$ be the density of the data given Ω_j under H_j, let p_j be the prior density of Ω_j and let h_j be an arbitrary density on the support of Ω_j. The Gelfand and Dey approximation of the marginal density of the data under H_j is

$$\tilde{f}_j(Y) = \left\{ \frac{1}{U-B} \sum_{u=B+1}^{U} \frac{h_j(\Omega_j^{(u)})}{f_j(Y|\Omega_j^{(u)})\, p_j(\Omega_j^{(u)})} \right\}^{-1}$$

where $\Omega_j^{(u)}$ is the value of Ω_j on the iteration u of the Markov chain, and the last $U-B$ iterations of U iterations are used.

The computations for the semiparametric model are facilitated by integrating Θ from equation (6) to obtain

$$\left.\begin{aligned} Y &= D\beta + \epsilon^*, \\ \epsilon^* &\sim N(0,\ \Sigma), \\ \Sigma &= \sigma^2 I + \tau^2 \Phi \Psi \Phi'. \end{aligned}\right\} \tag{14}$$

An efficient and numerically stable procedure uses the identities

$$\Sigma^{-1} = \sigma^{-2} I - \frac{\tau^2}{\sigma^4}\, \Phi \Psi^{1/2} \left(I + \frac{\tau^2}{\sigma^2}\, \Psi^{1/2} \Phi' \Phi \Psi^{1/2} \right)^{-1} \Psi^{1/2} \Phi',$$

$$\det(\Sigma) = \sigma^{2n} \det\left(I + \frac{\tau^2}{\sigma^2}\, \Psi^{1/2} \Phi' \Phi \Psi^{1/2} \right).$$

The remaining parameters are $\Omega_1 = (\beta',\ \ln(\sigma^2))'$ under H_1 and $\Omega_2 = (\beta',\ \ln(\sigma^2),\ \ln(\tau^2),\ \ln(\gamma - c))'$ under H_2 where c is given in equation (5). In both cases, the arbitrary density h_j in the Gelfand and Dey approximation was taken to be multivariate normal. Their means and variance matrices were estimated by using the generated parameter values from the Markov chain.

4. Simulation study and example

The simulation study generated data sets from $Y_i = f(x_i) + 2z_i + \epsilon_i$ for $i = 1, \ldots, n$ where $\{\epsilon_i\}$ is a random sample from a standard normal distribution, $\{x_i\}$ is an orthogonal design on $[0, 1]$ and $z_i = 2(x_i - 0.5) + \delta_i$ where $\{\delta_i\}$ is a random sample from a normal distribution with mean 0 and standard deviation 0.5.

Four simulations used different combinations of f and parametric models. Simulations A and B used a cubic polynomial with $f(0) = -1$, $f(1) = 3$ and roots at 0.2 and 0.7. The parametric component was linear in x in simulation A and cubic in x in simulation B. Simulation C generated data from the 'double-normal' function

$$f(x) = 5 - 10x + 8\, \exp\{-100(x - 0.3)^2\} - 8\, \exp\{-100(x - 0.7)^2\}$$

and used a linear parametric model in x. Simulation D generated data from the 'linexp' function

$$f(x) = 2 - 5x + \exp\{5(x - 0.6)\},$$

and used a quadratic parametric model in x.

Each of the four simulations consisted of 50 data sets. The simulations were performed twice: first with 25 observations then with 250 per data set. Table 1 reports the results of the simulations averaged over the 50 data sets. Only the geometric smoother is reported because the algebraic smoother gave similar quantitative results. The parametric model H_1 and semiparametric model H_2 in equation (13) were fitted to the data. The coefficients of the

Table 1. Simulation study†

	Results for the following simulations and sample sizes:							
	A		*B*		*C*		*D*	
	True model							
	Cubic		*Cubic*		*Double normal*		*Linexp*	
	Prior model							
	Linear		*Cubic*		*Linear*		*Quadratic*	
	25	*250*	*25*	*250*	*25*	*250*	*25*	*250*
$\ln\{P(Y\|H_1)\}$	−68.06	−498.98	−62.37	−386.82	−89.38	−702.95	−60.99	−399.10
	(0.47)	(1.22)	(0.55)	(1.29)	(0.21)	(0.62)	(0.60)	(1.26)
$\ln\{P(Y\|H_2)\}$	−61.68	−396.28	−62.93	−390.51	−72.79	−410.91	−61.88	−389.43
	(0.54)	(1.35)	(0.53)	(1.31)	(0.43)	(1.32)	(0.55)	(1.33)
Selected H_1‡	0	0	37	49	0	0	32	0
Selected H_2‡	50	50	13	1	50	50	18	50
$E(\sigma^2\|Y, H_1)$	2.614	2.600	1.001	0.992	13.739	13.253	1.143	1.121
	(0.094)	(0.025)	(0.042)	(0.010)	(0.229)	(0.064)	(0.050)	(0.011)
$E(\sigma^2\|Y, H_2)$	0.739	0.991	0.848	0.987	0.631	0.988	0.763	0.993
	(0.044)	(0.011)	(0.043)	(0.011)	(0.039)	(0.011)	(0.041)	(0.011)
HPD coverage H_1	0.617	0.159	0.969	0.927	0.638	0.165	0.932	0.443
	(0.011)	(0.002)	(0.014)	(0.018)	(0.002)	(0.002)	(0.013)	(0.010)
HPD coverage H_2	0.947	0.957	0.973	0.967	0.926	0.973	0.958	0.952
	(0.010)	(0.006)	(0.009)	(0.009)	(0.011)	(0.005)	(0.010)	(0.008)
RISE under H_1	1.304	1.261	0.425	0.140	3.517	3.494	0.538	0.376
	(0.007)	(0.001)	(0.026)	(0.008)	(0.004)	(0.001)	(0.021)	(0.004)
RISE under H_2	0.695	0.234	0.496	0.161	0.835	0.244	0.577	0.189
	(0.026)	(0.006)	(0.028)	(0.008)	(0.029)	(0.006)	(0.026)	(0.007)
Optimal kernel RISE	0.459	0.183	0.459	0.183	0.675	0.269	0.412	0.164
RISE favours H_1§	0	0	33	38	0	0	10	0
RISE favours H_2§	3	4	25	35	8	37	10	17
R^2 under H_1	0.255	0.380	0.740	0.766	0.531	0.534	0.719	0.757
	(0.014)	(0.004)	(0.010)	(0.003)	(0.005)	(0.002)	(0.012)	(0.002)
R^2 under H_2	0.898	0.775	0.817	0.769	0.992	0.967	0.868	0.790
	(0.009)	(0.003)	(0.011)	(0.003)	(0.001)	(0.000)	(0.009)	(0.002)
$E(\tau^2\|Y, H_2)$	0.789	0.692	0.471	0.268	5.420	6.057	0.346	0.374
	(0.041)	(0.019)	(0.018)	(0.003)	(0.247)	(0.108)	(0.007)	(0.009)
$E(\gamma\|Y, H_2)$	0.331	0.397	0.755	0.789	0.347	0.455	0.442	0.658
	(0.012)	(0.007)	(0.048)	(0.017)	(0.009)	(0.004)	(0.028)	(0.010)

†HPD, highest posterior density; RISE, root integrated squared error. H_1 is the parametric model and H_2 is the semiparametric model. 50 data sets were used in each simulation. The averages over the 50 data sets are reported. (Simulation standard errors are given in parentheses.)

‡The number of data sets where the posterior probability of the model exceeded 0.5 when the models had equal prior probabilities.

§The number of data sets where the RISE is less than the optimal asymptotic kernel RISE.

covariate z are estimated with equivalent accuracy under models H_1 and H_2 and are not reported.

Table 1 first reports the logarithm of the marginal density of Y and the results from selecting the model with the largest posterior probability when they are *a priori* equally likely. The selection procedure unambiguously and correctly selects model H_2 in simulations A and C. In simulation D, the linexp function is well approximated by a quadratic function. With only 25 observations, the procedure incorrectly selects model H_1 in 64% of the data sets. With 250 observations, it correctly selects H_2 in all 50 data sets. In simulation B, the posterior probabilities correctly select H_1 in 54% of the data sets with 25 observations and in 98% of the data sets with 250 observations.

As we would expect, when model H_1 is misspecified (simulations A, C and D), the posterior mean of the error variance tends to be too large under H_1 and accurate under H_2. The approximate 95% highest posterior density (HPD) intervals for the mean of Y (see Angers and Delampady (1992) or Wecker and Ansley (1983)) have the wrong coverage under H_1 and the correct coverage under H_2. The root integrated squared error (RISE) of the response function under H_1 is larger than under H_2. The squared correlation R^2 between the actual and predicted observations is smaller under H_1 than under H_2. When the parametric model is correctly specified in simulation B, these performance measures are equivalent for H_1 and H_2.

The RISE in Table 1 can be compared with the optimal asymptotic MISE for the kernel estimator (Gasser and Müller, 1984):

$$\frac{5}{4}\left(n^{-1}\sigma^2 \int K^2\right)^{4/5}\left\{\left(\int u^2 K\right)^2 \int f''^2\right\}^{1/5}$$

if f'' is uniformly continuous, and K is the kernel. Following the lead of Härdle *et al.* (1988), Table 1 reports the optimal asymptotic RISE for the kernel

$$K(x) = (15/8)(1 - 4x^2)^2\, \mathbf{1}_{[-0.5,0.5]}(x).$$

In simulation C with 250 observations the average RISE under model H_2 is less than the optimal kernel RISE, and it is between the optimal kernel RISE and the average RISE under model H_1 for the other simulations when H_1 is false. Even so, the RISE under H_2 often is better than the optimal kernel RISE for some of the data sets.

The last section of Table 1 reports the estimated smoothing parameters τ^2 and γ. Comparing simulations A and B is instructive because the parametric component is incorrect in A and correct in B. τ^2 tends to be significantly smaller in simulation B than in A so the Fourier coefficients have greater shrinkage to their prior mean of 0 in simulation B. Also, γ tends to be larger in simulation B so high frequency terms have less effect. Together, the parametric component plays a substantially reduced role in simulation B.

The empirical example uses monthly traffic accidents for the 108 months from January 1st, 1979, to December 31st, 1987, in the state of Michigan. The data are a 0.1% random sample of all accidents and were obtained from the University of Michigan Transportation Research Institute. Lenk and Rao (1995) analysed these data and found them to have seasonal effects. Evans and Graham (1987), among others, postulated that traffic accidents increase with traffic density. In turn, traffic density tends to be positively correlated with economic activity. A surrogate for economic activity is non-institutional unemployment rates for Michigan, obtained from Streff *et al.* (1988). The correlation between log(number of monthly accidents) and log(unemployment rate) is -0.563, and a graph of the two indicates a linear relationship. A standard assumption for the number of accidents is that they are Poisson random

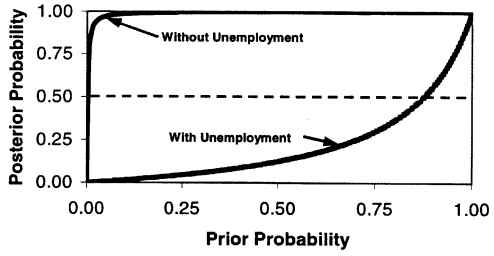

Fig. 2. Posterior probabilities for the semiparametric models with and without unemployment for monthly automobile accidents

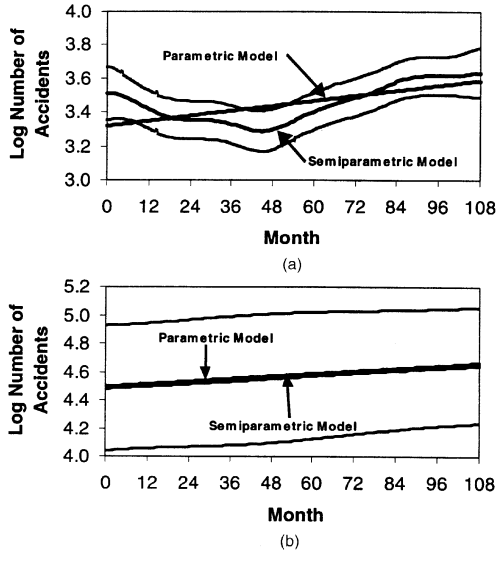

Fig. 3. Posterior analysis of seasonally adjusted log(number of automobile accidents) (a) without log(unemployment) and (b) with log(unemployment): the straight lines are the posterior means of the linear relationship with months from the parametric model, and the centred curved lines are the posterior means of the semiparametric relationship with months from the semiparametric model; the bands are 95% HPD intervals for the semiparametric model

variables. Because the monthly counts are large, the log-counts have approximately a normal distribution.

Two different parametric components were fitted to the data. The first had seasonal effects, coded as 0–1 variables, and a trend, and the second added log(unemployment). Fig. 2 presents the posterior probabilities under the semiparametric model H_2 for the two parametric specifications. When log(unemployment) is not in the model, the posterior probabilities of H_2 are nearly 1. However, H_1 is more probable if log(unemployment) is added. Fig. 3(a) graphs the posterior means of the response functions after adjusting for seasonal effects, and Fig. 3(b) graphs them after adjusting for seasonal effects and log(unemployment). Fig. 3 also includes 95% HPD intervals under H_2. In Fig. 3(a) the 95% HPD intervals do not always include the posterior mean under H_1, thus indicating that the parametric model is inadequate. In Fig. 3(b) they do include the posterior mean under H_1, thus verifying that the nonparametric component is not needed when log(unemployment) is in the model.

Table 2. Monthly traffic accidents†

	Results for the following models:			
	H_1 without unemployment	*H_2 without unemployment*	*H_1 with unemployment*	*H_2 with unemployment*
$\ln\{P(Y)\}$	−26.952	−20.462	−19.026	−20.972
R^2	0.278	0.508	0.514	0.515
Error variance	0.046	0.036	0.034	0.034
	(0.006)	(0.005)	(0.005)	(0.005)
Constant	3.319	3.330	4.496	4.485
	(0.055)	(0.108)	(0.120)	(0.234)
Month	0.0025	0.0022	0.0015	0.0015
	(0.0007)	(0.0019)	(0.0006)	(0.0011)
Spring	−0.184	−0.185	−0.197	−0.187
	(0.058)	(0.053)	(0.051)	(0.050)
Summer	−0.208	−0.205	−0.226	−0.227
	(0.059)	(0.051)	(0.051)	(0.050)
Autumn	−0.067	−0.058	−0.109	−0.110
	(0.058)	(0.052)	(0.051)	(0.052)
Unemployment	—	—	−0.464	−0.458
	—	—	(0.076)	(0.089)

†H_1 is the parametric model and H_2 is the semiparametric model. Posterior means are reported for the parameters, and posterior standard deviations are given in parentheses.

In Table 2 the parametric model H_1 with log(unemployment) has the largest log-density of the data and equivalent R^2 and error variances compared with the semiparametric models H_2. The coefficients for the parametric components are nearly equivalent under H_1 and H_2. There is a small positive trend. The number of accidents is substantially lower in spring and summer, while the effect of autumn is not significant. Unemployment has a significant negative coefficient. In conclusion, a parametric model with linear trend, seasonal effects and unemployment provides an adequate description of the data among the class of models with time factors and unemployment.

5. Discussion

Further research is needed in several areas. All nonparametric procedures require model choices — whether the shape of the kernel for kernel regression, the form of the penalty term for smoothing splines or the geometry of the problem for nearest neighbour methods — that potentially affect the analysis. Bayesian nonparametric models use more flexible likelihood functions than do parametric models and introduce more information or structure into the prior distributions. A natural question is the effect of the prior specification on the posterior analysis.

This paper represents the nonparametric component with a Fourier series and uses a hierarchical specification of the prior distribution of the Fourier coefficients. The first stage of the prior describes the rate of decay of the Fourier coefficients, and the second stage represents prior beliefs about the hyperparameters of the first stage. The first stage works together with the Fourier representation to specify the smoothness of the nonparametric component.

Other choices of basis may require a different specification of the prior to express smoothness. For example, the number and location of knots in polynomial splines determine the amount of detail in the estimator, whereas the prior specification for the regression coefficients is relatively less important, as long as it is not too informative. Conversely, if we use the Fourier representation, a careless choice of prior distribution for the Fourier coefficients will

lead to inappropriate inferences. For example, using a spherical distribution for the Fourier coefficients results in Fourier series that are not absolutely convergent. The prior distribution of the Fourier coefficients was selected to work in concert with the Fourier representation. However, the specification of the prior distribution for the hyperparameters tends to be less critical. The prior distributions for the parametric model, regression coefficients and error variance are standard distributions in the Bayesian literature, and the posterior analysis is relatively insensitive to their specification.

Further work is needed in comparing this paper's model with polynomial splines and local polynomials. The Fourier model's support is the piecewise continuous functions, which contains the support of polynomial splines and local polynomials. Whether this theoretical advantage has practical implications requires further study.

The cosine basis used in this paper is global: adding another basis element affects the estimator over its entire support. In contrast, wavelets (Antoniadis *et al.*, 1994) and local polynomials (Fan and Gijbels, 1996) are localized: adding another wavelet or polynomial to the estimator will only affect the estimator in a compact region of its support. Future research is needed to contrast the two approaches.

Appendix A: Proof of theorem 1

The proof of theorem 1 repeatedly uses an integral approximation to series: if $f(x)$ is non-negative, non-increasing on $[u, v]$ with $0 \leqslant u \leqslant v$ being integers, then

$$\sum_{k=u}^{v} f(k) \leqslant \int_{u}^{v} f(z)\,\mathrm{d}z + f(u). \tag{15}$$

It also relies on the integrals

$$\int \frac{\exp(-\gamma x)}{\sigma^2 + n\tau^2\,\exp(-\gamma x)}\,\mathrm{d}x = \frac{-1}{n\tau^2\gamma}\ln\{\sigma^2 + n\tau^2\,\exp(-\gamma x)\}, \tag{16}$$

$$\int \frac{\exp(-2\gamma x)}{\{\sigma^2 + n\tau^2\,\exp(-\gamma x)\}^2}\,\mathrm{d}x = \frac{-1}{n^2\tau^4\gamma}\left[\ln\{\sigma^2 + n\tau^2\,\exp(-\gamma x)\} + \frac{\sigma^2}{\sigma^2 + n\tau^2\,\exp(-\gamma x)}\right], \tag{17}$$

$$\int_0^\infty \frac{x^a\,\mathrm{d}x}{(m + x^b)^c} = m^{(a+1-bc)/b}\,\frac{\Gamma\{(a+1)/b\}\,\Gamma\{c - (a+1)/b\}}{b\,\Gamma(c)}$$

$$\text{for } a > -1,\ b > 0,\ m > 0 \text{ and } c > (a+1)/b. \tag{18}$$

For an orthogonal design on [0, 1], the conditional series estimator given β, σ, τ and γ is given in equations (7)–(9). To make the dependence on k explicit, define $w_n(k) = w_{n,k}$ (equation (8)), and $c_k = c(k)$ (equation (4)). The MISE of the conditional estimator depends on the sampling distribution of $\{\hat{\theta}_k\}$ (equation (9)), which has the following means and variances for orthogonal designs:

$$E(\hat{\theta}_k) = n^{-1}\sum_{i=1}^{n} g(x_i)\,\phi_k(x_i),$$

$$\mathrm{var}(\hat{\theta}_k) = n^{-2}\sigma^2\sum_{i=1}^{n}\phi_k^2(x_i) = \sigma^2/n,$$

$$\mathrm{cov}(\hat{\theta}_k, \hat{\theta}_{k'}) = n^{-2}\sigma^2\sum_{i=1}^{n}\phi_k(x_i)\,\phi_{k'}(x_i) = 0$$

for $k \neq k'$. Using the orthogonality of the basis functions, the integrated variance term in the MISE can be written as

$$\int_0^1 \mathrm{var}\{\tilde{g}(x)\}\,\mathrm{d}x = \sum_{k=1}^{K} w_{n,k}^2\,\frac{\sigma^2}{n}\int_0^1 \phi_k^2(x)\,\mathrm{d}x \equiv \sum_{k=1}^{K} f(k)$$

where f is defined as

$$f(k) = \frac{n\sigma^2\tau^4 \exp\{-2\gamma c(k)\}}{[\sigma^2 + n\tau^2 \exp\{-\gamma c(k)\}]^2}.$$

For fixed k, each $f(k)$ is $O(n^{-1})$, but there are $K = n - 1$ terms in the sum by assumption (b) of theorem 1. Using the integral approximation for series in equation (15) and applying the integral in equation (17) results in the following upper bound for the geometric smoother when $c(z) = z$:

$$\int_0^1 \mathrm{var}\{\tilde{g}(x)\}\,\mathrm{d}x \leqslant f(1) + \frac{\sigma^2}{\gamma n}\left[\log\left\{\frac{\sigma^2 + n\tau^2 \exp(-\gamma)}{\sigma^2 + n\tau^2 \exp(-\gamma K)}\right\} + \frac{\sigma^2}{\sigma^2 + n\tau^2 \exp(-\gamma)} - \frac{\sigma^2}{\sigma^2 + n\tau^2 \exp(-\gamma K)}\right]$$
$$\leqslant O\left\{\frac{\log(n)}{n}\right\}.$$

For the algebraic smoother, $c(z) = \log(z)$, use equation (18) to obtain the following upper bound for $\gamma > 1$:

$$\int_0^1 \mathrm{var}\{\tilde{g}(x)\}\,\mathrm{d}x \leqslant f(1) + \left(\frac{\sigma^2}{n}\right)^{1-1/\gamma} \tau^{2/\gamma} \frac{\Gamma(2 - 1/\gamma)\,\Gamma(1/\gamma)}{\gamma\,\Gamma(2)} \leqslant O(n^{1/\gamma - 1}).$$

The large sample behaviour of the bias is somewhat more delicate than that of the variance. The Fourier expansion of g can be written as

$$g(x) = \sum_{k=1}^{K} w_{n,k}\theta_k\,\phi_k(x) + \sum_{k=1}^{K}(1 - w_{n,k})\theta_k\,\phi_k(x) + \sum_{k=K+1}^{\infty}\theta_k\,\phi_k(x).$$

Because the basis functions are bounded, $\sup_{0\leqslant x\leqslant 1}|\phi_k(x)| \leqslant 2^{1/2}$, the bias can be bounded by three terms:

$$|E\{\tilde{g}(x)\} - g(x)| = \left|\sum_{k=1}^{K} w_{n,k}\,E(\hat{\theta}_k)\,\phi_k(x) - \sum_{k=1}^{\infty}\theta_k\,\phi_k(x)\right|$$
$$\leqslant 2^{1/2}\sum_{k=1}^{K} w_{n,k}|E(\hat{\theta}_k) - \theta_k| + 2^{1/2}\sum_{k=1}^{K}|w_{n,k} - 1||\theta_k| + 2^{1/2}\sum_{k=K+1}^{\infty}|\theta_k|. \tag{19}$$

By assumption (d) of theorem 1 the last sum is $O\{\exp(-\rho n)\}$ for the geometric smoother and $O(n^{-\rho+1})$ for the algebraic smoother.

The convergence of the first term to 0 is controlled by the bias of $\{\hat{\theta}_k\}$. The following standard argument shows that $|E(\hat{\theta}_k) - \theta_k| = O(n^{-1})$. Let A_i be the interval $[(i-1)/n, i/n]$ so that $x_i = (2i-1)/2n$ is its midpoint, and its length is $1/n$. Then we can write

$$E(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} g(x_i)\,\phi_k(x_i) = \sum_{i=1}^{n}\int_{A_i} g(x_i)\,\phi_k(x_i)\,\mathrm{d}x.$$

Applying assumption (c) of theorem 1 produces

$$|E(\hat{\theta}_k) - \theta_k| \leqslant \sum_{i=1}^{n}\int_{A_i}|g(x_i)\,\phi_k(x_i) - g(x)\,\phi_k(x)|\,\mathrm{d}x$$
$$\leqslant \sup|\phi_k|\sum_{i=1}^{n}\int_{A_i}|g(x_i) - g(x)|\,\mathrm{d}x$$
$$\leqslant 2^{1/2}\sum_{i=1}^{n} C_0\int_{A_i}|x_i - x|\,\mathrm{d}x = 2^{-3/2}C_0 n^{-1},$$

where C_0 is the constant in the Lipschitz continuity condition. Define

$$f(k) = \frac{C_0}{2}\frac{\tau^2 \exp\{-\gamma c(k)\}}{\sigma^2 + n\tau^2 \exp\{-\gamma c(k)\}}.$$

For the geometric smoother, an upper bound for the first term in equation (19) can be obtained from the series approximation in equation (15) and the integral in equation (16):

$$2^{1/2}\sum_{k=1}^{K} w_{n,k}|E(\hat{\theta}_k)-\theta_k| \leqslant O(n^{-1}) + \frac{C_0}{2\gamma n}\log\left\{\frac{\sigma^2+n\tau^2\,\exp(-\gamma)}{\sigma^2+n\tau^2\,\exp(-\gamma K)}\right\}$$
$$\leqslant O\left\{\frac{\log(n)}{n}\right\}.$$

For the algebraic smoother with $\gamma > 1$, applying equation (18) results in the upper bound

$$2^{1/2}\sum_{k=1}^{K} w_{n,k}|E(\hat{\theta}_k)-\theta_k| \leqslant f(1) + \frac{C_0\tau^2}{2\sigma^2}\int_0^{\infty}\left(\frac{n\tau^2}{\sigma^2}+x^{\gamma}\right)^{-1}\mathrm{d}x$$
$$\leqslant f(1) + n^{1/\gamma-1}\,\frac{C_0}{2\gamma}\left(\frac{\tau^2}{\sigma^2}\right)^{1/\gamma}\Gamma\left(1-\frac{1}{\gamma}\right)\Gamma\left(\frac{1}{\gamma}\right)$$
$$\leqslant O(n^{1/\gamma-1}).$$

The second sum in equation (19) can be bounded by using the rate of decay on the coefficients in assumption (d) of theorem 1. For fixed k,

$$|w_{n,k}-1| = \frac{\sigma^2}{\sigma^2+n\tau^2\,\exp(-\gamma c_k)} = O\left(\frac{1}{n}\right).$$

Let M be the constant in assumption (d). For fixed M,

$$2^{1/2}\sum_{k=1}^{M}|w_{n,k}-1||\theta_k| = O\left(\frac{1}{n}\right),$$

and the first M terms can be ignored in the limit. If $M < K$, then

$$2^{1/2}\sum_{k=1}^{K}|w_{n,k}-1||\theta_k| \leqslant O\left(\frac{1}{n}\right) + 2^{1/2}C\sigma^2\int_{M+1}^{K}\frac{\exp\{-\rho\,c(x)\}}{\sigma^2+n\tau^2\,\exp\{-\gamma\,c(x)\}}\,\mathrm{d}x.$$

If $\rho \geqslant \gamma$, then

$$\int_{M+1}^{K}\frac{\exp\{-\rho\,c(x)\}}{\sigma^2+n\tau^2\,\exp\{-\gamma\,c(x)\}}\,\mathrm{d}x \leqslant \int_{M+1}^{K}\frac{\exp\{-\gamma\,c(x)\}}{\sigma^2+n\tau^2\,\exp\{-\gamma\,c(x)\}}\,\mathrm{d}x.$$

Using the integral in equation (16) for the geometric smoother,

$$2^{1/2}\sum_{k=1}^{K}|w_{n,k}-1||\theta_k| \leqslant O(n^{-1}) + \frac{2^{1/2}C\sigma^2}{n\tau^2\gamma}\log\left[\frac{\sigma^2+n\tau^2\,\exp\{-\gamma(M+1)\}}{\sigma^2+n\tau^2\,\exp(-\gamma K)}\right]$$
$$\leqslant O\left\{\frac{\log(n)}{n}\right\}.$$

Using the integral in equation (18) for the algebraic smoother,

$$2^{1/2}\sum_{k=1}^{K}|w_{n,k}-1||\theta_k| \leqslant O(n^{-1}) + 2^{1/2}C\sigma^2\int_0^{\infty}\frac{x^{-\gamma}}{\sigma^2+n\tau^2x^{-\gamma}}\,\mathrm{d}x$$
$$\leqslant O(n^{-1}) + 2^{1/2}C\gamma^{-1}\left(\frac{\sigma^2}{n\tau^2}\right)^{1-1/\gamma}\Gamma\left(1-\frac{1}{\gamma}\right)\Gamma\left(\frac{1}{\gamma}\right)$$
$$\leqslant O(n^{1/\gamma-1}).$$

Combining the three terms in equation (19), when $\rho \geqslant \gamma$ the bias is $O\{\log(n)/n\}$ for the geometric smoother. For the algebraic smoother, the first two terms for the bias are $O(n^{-(1-1/\gamma)})$, and the third term is $O(n^{-(\rho-1)})$. Because $\rho \geqslant \gamma$, the slower rate is $O(n^{-(1-1/\gamma)})$.

Next, consider the case where $\rho < \gamma$. For the geometric smoother,

$$
\begin{aligned}
2^{1/2}\sum_{k=1}^{K}|w_{n,k}-1||\theta_k| &\leqslant O(n^{-1})+2^{1/2}C\sigma^2\int_{M+1}^{K}\frac{\exp(-\rho x)}{\sigma^2+n\tau^2\exp(-\gamma x)}\,\mathrm{d}x\\
&\leqslant O(n^{-1})+2^{1/2}C\frac{\sigma^2}{n\tau^2\rho}\int_0^{\infty}\left(\frac{\sigma^2}{n\tau^2}+u^{\gamma/\rho}\right)^{-1}\mathrm{d}u\\
&\leqslant O(n^{-1})+2^{1/2}C\left(\frac{\sigma^2}{n\tau^2}\right)^{\rho/\gamma}\gamma^{-1}\,\Gamma\left(\frac{\rho}{\gamma}\right)\Gamma\left(1-\frac{\rho}{\gamma}\right)\\
&\leqslant O(n^{-\rho/\gamma}).
\end{aligned}
$$

Combining the three terms in equation (19) when $\rho<\gamma$, the bias for the geometric smoother is $O(n^{-\rho/\gamma})$. For the algebraic smoother,

$$
\begin{aligned}
2^{1/2}\sum_{k=1}^{K}|w_{n,k}-1||\theta_k| &\leqslant O(n^{-1})+2^{1/2}C\sigma^2\int_{M+1}^{K}\frac{x^{-\rho}}{\sigma^2+n\tau^2x^{-\gamma}}\,\mathrm{d}x\\
&\leqslant O(n^{-1})+2^{1/2}C\int_0^{\infty}\frac{x^{\gamma-\rho}}{n\tau^2/\sigma^2+x^{\gamma}}\,\mathrm{d}x\\
&\leqslant O(n^{-1})+2^{1/2}C\left(\frac{\sigma^2}{n\tau^2}\right)^{(\rho-1)/\gamma}\gamma^{-1}\,\Gamma\left(\frac{\rho-1}{\gamma}\right)\Gamma\left(1-\frac{\rho-1}{\gamma}\right)\\
&\leqslant O(n^{-(\rho-1)/\gamma})
\end{aligned}
$$

for $1<\rho<\gamma$. In equation (19), the first term is $O(n^{-(1-1/\gamma)})$, the second term is $O(n^{-(\rho-1)/\gamma})$ and the third term is $O(n^{-(\rho-1)})$. When $1<\rho<\gamma$, the slowest rate is $O(n^{-(\rho-1)/\gamma})$.

References

Angers, J. F. and Delampady, M. (1992) Hierarchical Bayesian curve fitting and smoothing. *Can. J. Statist.*, **20**, 35–49.

Antoniadis, A., Gregoire, G. and McKeague, I. W. (1994) Wavelet methods for curve estimation. *J. Am. Statist. Ass.*, **89**, 1340–1353.

Chipman, H. A., Kolaczyk, E. D. and McCulloch, R. E. (1997) Adaptive Bayesian wavelet shrinkage. *J. Am. Statist. Ass.*, **92**, 1413–1421.

Clyde, M., Parmigiani, G. and Vidakovic, B. (1998) Multiple shrinkage and subset selection in wavelets. *Biometrika*, **85**, 391–401.

Damien, P., Wakefield, J. and Walker, S. (1999) Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *J. R. Statist. Soc.* B, **61**, 331–344.

Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998) Automatic Bayesian curve fitting. *J. R. Statist. Soc.* B, **60**, 333–350.

Diaconis, P. and Freedman, D. (1986a) On the consistency of Bayes estimates. *Ann. Statist.*, **14**, 1–26.

———(1986b) On inconsistent Bayes estimates of location. *Ann. Statist.*, **14**, 68–87.

Evans, W. and Graham, J. D. (1987) *Traffic Fatalities and the Business Cycle*. Boston: New England Injury Prevention Research Center.

Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall.

Gallant, A. R. and Monahan, J. F. (1985) Explicitly infinite-dimensional Bayesian analysis of production technologies. *J. Econometr.*, **30**, 171–201.

Gasser, T. and Müller, H. G. (1984) Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.*, **11**, 171–185.

Gelfand, A. E. and Dey, D. K. (1994) Bayesian model choice: asymptotics and exact calculations. *J. R. Statist. Soc.* B, **56**, 501–514.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.

Hall, P. and Titterington, D. M. (1987) Common structure of techniques for choosing smoothing parameters in regression problems. *J. R. Statist. Soc.* B, **49**, 184–198.

Härdle, W. (1990) *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

Härdle, W., Hall, P. and Marron, J. S. (1988) How far are automatically chosen regression smoothing parameters from their optimal? *J. Am. Statist. Ass.*, **83**, 86–95.

Kass, R. E. and Raftery, A. E. (1995) Bayes factors. *J. Am. Statist. Ass.*, **90**, 773–795.

Katznelson, Y. (1976) *An Introduction to Harmonic Analysis*. New York: Dover Publications.
Kimeldorf, G. S. and Wahba, G. (1970) A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist*., **41**, 495–502.
Kreider, D. L., Kuller, R. G., Ostberg, D. R. and Perkins, F. W. (1966) *An Introduction to Linear Analysis*. Reading: Addison-Wesley.
Lenk, P. (1991) Towards a practicable Bayesian nonparametric density estimator. *Biometrika*, **78**, 531–543.
———(1993) A Bayesian nonparametric density estimator. *J. Nonparam. Statist*., **3**, 53–69.
Lenk, P. and Rao, A. (1995) Transition times: distributions arising from time heterogeneous Poisson processes. *Mangmnt Sci*., **41**, 1117–1129.
Polson, N. G. (1996) Convergence of Markov Chain Monte Carlo algorithms. In *Bayesian Statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. M. F. Smith), pp. 297–312. Oxford: Oxford University Press.
Rutkowski, L. (1982) On line identification of time varying system by nonparametric techniques. *IEEE Trans. Autom. Control*, **27**, 228–230.
Silverman, B. W. (1985) Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *J. R. Statist. Soc*. B, **47**, 1–52.
Smith, M. and Kohn, R. (1996) Nonparametric regression using Bayesian variable selection. *J. Econometr*., **75**, 317–344.
Smith, M., Wong, C.-M. and Kohn, R. (1998) Additive nonparametric regression with autocorrelated errors. *J. R. Statist. Soc*. B, **60**, 311–331.
Speckman, P. (1988) Kernel smoothing in partial linear models. *J. R. Statist. Soc*. B, **50**, 413–436.
Stone, C. J. (1982) Optimal global rates of convergence for nonparametric regression. *Ann. Statist*., **10**, 1040–1053.
Streff, F. M., Schultz, R. H. and Wagenaar, A. C. (1988) Changes in police-reported injuries associated with Michigan's safety belt law: 1988 update. *Report UMTRI-89-5*. University of Michigan Transportation Research Institute, Ann Arbor.
Wahba, G. (1978) Improper prior, spline smoothing and the problem of guarding against model errors in regression. *J. R. Statist. Soc*. B, **40**, 364–372.
———(1983) Bayesian 'confidence intervals' for the cross-validation smoothing spline. *J. R. Statist. Soc*. B, **45**, 133–150.
Wecker, W. E. and Ansley, C. F. (1983) The signal extraction approach to nonlinear regression and spline smoothing. *J. Am. Statist. Ass*., **78**, 81–89.
Wong, C. M. and Kohn, R. (1996) A Bayesian approach to additive semiparametric regression. *J. Econometr*., **74**, 209–235.