ELSEVIER

Journal of Statistical Planning and Inference 49 (1996) 327–341

journal of statistical planning and inference

Improved inference in nonparametric regression using L_k-smoothing splines

Felix Abramovich[1], David M. Steinberg*

Department of Statistics and Operations Research, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Ramat Aviv 69978, Israel

Received 29 March 1993; revised 21 February 1995

Abstract

Smoothing splines are one of the most popular approaches to nonparametric regression. Wahba (*J. Roy. Statist. Soc. Ser. B* **40** (1978) 364–372; **45** (1983) 133–150) showed that smoothing splines are also Bayes estimates and used the corresponding prior model to derive interval estimates for the regression function. Although the interval estimates work well on a global basis, they can have poor local properties. The source of this problem is the use of a global smoothing parameter. We introduce the notion of L_k- smoothing splines. These splines allow for a variable smoothing parameter and can substantially improve local inference.

AMS Subject Classification: 62G05, 62G15

Keywords: Bayesian linear model; Confidence interval; L-spline; Variable smoothing parameter

1. Introduction

Consider the standard nonparametric regression setting

$$y(t_i) = g(t_i) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1.1}$$

where $0 \leqslant t_1 < \cdots < t_n \leqslant 1$, $\{\varepsilon_i\}$ are i.i.d. normal variables with zero mean and variance σ^2 and we wish to estimate the response function $g(\cdot)$ from the data without assuming any particular parametric form for g.

When $g(\cdot)$ is assumed to be 'reasonably smooth', an effective estimation method is spline smoothing, formulated by Schoenberg (1964) and Reinsch (1967) and developed

* Corresponding author.

[1]Current Address: School of Mathematics, University of Bristol, Bristol, BS8 1TW, UK.

SSDI 0378-3758(95)00021-6

by many authors (see Eubank, 1988; Wahba, 1990 for a detailed survey). The basic idea of spline smoothing is to find an estimate which fits the data but at the same time is a 'smooth' function. A standard measure of an estimate's goodness-of-fit to the data is the residual sum of squares, $\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{g}(t_i))^2$, while a natural measure of its smoothness is $\int \hat{g}^{(m)}(t)^2 \mathrm{d}t$.

A smoothing spline estimate $\hat{g}$ is defined formally as a minimizer of a weighted sum of these two usually contradictory criteria

$$\min \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(t_i))^2 + k^2 \int_0^1 (f^{(m)}(t))^2 \mathrm{d}t \right\} \tag{1.2}$$

over all functions from $W_2^m = \{f: f \in C^{m-1}[0, 1],\ f^{(m)} \in L_2[0, 1]\}$. The smoothing parameter k^2 controls the trade-off between fidelity to the data and the smoothness of the estimate. It is well known that the solution of (1.2) is unique and is a natural polynomial spline of degree $(2m - 1)$ with knots $\{t_i\}$.

The idea of using a smoothing spline as an estimate of the unknown response function $g(\cdot)$ originated from research in the theory of function approximation. Wahba (1978, 1983) demonstrated an attractive *statistical* interpretation for $\hat{g}$. She proved that $\hat{g}$ may be viewed as a Bayes estimate of g with respect to a certain prior on the class of possible response functions. The Bayesian approach allows one not only to estimate the unknown function, but also to provide error bounds by constructing the corresponding Bayesian point-wise probability intervals (Wahba, 1983; Wecker and Ansley, 1983; Silverman, 1985; Ansley et al., 1993).

Wahba (1983) studied several examples by Monte Carlo simulation and found that the Bayesian intervals for smoothing splines had valid coverage properties as confidence intervals for the unknown function. However, in her discussion of Silverman (1985), Wahba noted that the coverage probabilities do not hold at each individual point, but rather are valid when averaged across the entire curve. She found that the true coverage could fall far short of the nominal level at points where 'there is an unusual large local curvature, or worse, a kink in the curve'. Nychka (1988) reanalyzed Wahba's examples and showed that bias, although generally the modest part of the mean squared error, increases significantly in these regions. The increase in bias occurs because of the *global* value of the smoothing parameter k^2, which is appropriate on the average across all the points, but does not adapt to the local behavior of the function in regions of high curvature, where the polynomial spline with global k^2 tends to *oversmooth*. This feature is especially problematic in interval estimation. Use of a global smoothing parameter leads to intervals whose widths do not depend on the degree of local curvature and thus fail to correctly reflect such bias. The coverage probabilities are correct on average because they are too low at points of high curvature and too high at points of low curvature.

Our major goal in this paper is to modify the smoothing spline approach so that we can find interval estimates that do not suffer from the problems noted above. We consider the use of L_k-splines, which employ a *variable* smoothing parameter. The

L_k-spline estimates can reduce the bias in regions of high curvature by reducing the penalty for lack of smoothness; where curvature is low, the estimate emphasizes smoothness and reduces the variance. Moreover, the differing emphases on smoothness are reflected in the Bayesian interval estimates, which are now wider in regions of high curvature and greatly improve pointwise coverage properties. From a Bayesian standpoint, the L_k-spline can be seen as a vehicle for a more accurate expression of prior belief about g.

In Section 2 we generalize the definition (1.2) to the case of a variable smoothing parameter and discuss basic properties of the proposed estimator, which turns out to be a natural L-spline for a certain differential operator L_k. Section 3 is devoted to a Bayesian model that leads to the L_k-smoothing spline as the posterior mean of the true function. This section helps clarify the motivation for using a variable smoothing parameter, extends Wahba's (1978, 1983) model for polynomial smoothing splines, and provides intuition for the subsequent smoothing algorithm. In Section 4 we describe our L_k-smoothing spline procedure and consider two examples that illustrate the effectiveness of the proposed approach for inference in nonparametric regression.

2. Derivation of the estimator and its basic properties

Consider the model (1.1). We propose to estimate $g(t)$ as in the smoothing spline approach (see (1.2)) but with a variable smoothing parameter. Thus we define our estimator $\hat{g}(\cdot)$ as the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(t_i))^2 + \int_0^1 k^2(t)(f^{(m)}(t))^2\,dt, \tag{2.1}$$

over all functions f from $\Omega_k^m = \{f: f \in C^{m-1}[0,1],\ kf^{(m)} \in L_2[0,1]\}$, where the smoothing parameter $k^2(\cdot) \in L_2[0,1]$ and is strictly positive.

This definition is a practically interesting particular case of a more general problem where the integral term in (2.2) is replaced by $\int(Lf)^2$, where L is some mth order differential operator (Kimeldorf and Wahba, 1971; Wahba, 1985; Kohn and Ansley, 1983, 1988).

We derive the estimator in Theorem 1 below. First, though, we need to introduce some notation. Define the function $Q(s,t)$ by

$$Q(s,t) = \int_0^{\min(s,t)} \frac{(t-u)^{m-1}(s-u)^{m-1}}{(m-1)!\,(m-1)!}\,\psi^2(u)\,du,$$

where the *roughness function* $\psi^2(t) = 1/k^2(t)$. Let T denote the $n \times m$ matrix with $T_{ij} = t_i^{j-1}/(j-1)!$, Q_n the $n \times n$ matrix with $Q_{ij} = Q(t_i, t_j)$ and let $M = Q_n + nI$.

The following theorem gives the explicit derivation of $\hat{g}(t)$ from (2.1):

Theorem 1. *The unique element of* Ω_k^m *which minimizes* (2.1) *is*

$$\hat{g}(t) = \sum_{r=1}^{m} \alpha_r \frac{t^{r-1}}{(r-1)!} + \sum_{j=1}^{n} d_j Q(t_j, t), \tag{2.2}$$

where

$$\alpha = (T'M^{-1}T)^{-1}T'M^{-1}y, \qquad \boldsymbol{d} = M^{-1}(I - T(T'M^{-1}T)^{-1}T'M^{-1})y.$$

The proof follows directly from the general result established in Kimeldorf and Wahba (1971, Theorem 5.1) for generalized splines.

In particular, the vector of fitted values at the data points $\{t_i\}$ is

$$\hat{\boldsymbol{g}} = (\hat{g}(t_1), \ldots, \hat{g}(t_n))' = Ay,$$

where $A = T(T'M^{-1}T)^{-1}T'M^{-1} + Q_nM^{-1}(I - T(T'M^{-1}T)^{-1}T'M^{-1})$.

Repeating Wahba's (1978, 1983) calculations one gets a more convenient representation of the matrix A:

$$A = I_n - nB'(BQ_nB' + nI_n)^{-1}B,$$

where B is any $(n-m)\times n$ matrix whose $n-m$ rows are orthonormal and orthogonal to the columns of T, that is $BB' = I_{(n-m)}$ and $BT = 0_{(n-m)\times m}$.

Kimeldorf and Wahba (1970) showed that the solution of (2.1) with the general integral term $\int (Lf)^2 dt$ is a natural L-spline with knots $\{t_i\}$. Thus, $\hat{g}(t)$ is a natural L-spline for the differential operator $L_k = k(t)D^m$ with the set of knots $\{t_i\}$, that is

$$L_k^* L_k[\hat{g}] = (-1)^m (k^2(t)\hat{g}^{(m)}(t))^{(m)} = 0 \tag{2.3}$$

everywhere except, maybe, the data-points $\{t_i\}$ and $\hat{g}(t)$ satisfies the following natural boundary conditions:

$$(k^2(t)\hat{g}^{(m)}(t))^{(p)}|_{t=0} = (k^2(t)\hat{g}^{(m)}(t))^{(p)}|_{t=1} = 0, \quad p = 0, \ldots, m-1.$$

From the definition of an L-spline, it follows that an L_k-spline consists of piecewise solutions of the $2m$th order linear differential equation $(k^2(t)\hat{g}^{(m)}(t))^{(m)} = 0$ joined at the knots to provide $\hat{g} \in C^{m-1}$ and $k^2\hat{g}^{(m)} \in C^{m-2}$. Thus, in particular, if $k^2 \in C^q$, then $\hat{g} \in C^p$, where $p = \min(m+q;\ 2m-2)$, which implies a smoother solution than the original requirement of being just in C^{m-1} (see the definition (2.1) of $\hat{g}$).

For constant k^2, $L_k^*L_k$ coincides with D^{2m}, the differentiation operator of order $2m$, and (2.3) yields a piecewise-polynomial solution of order $2m-1$, which together with the conditions on knots leads to a 'standard' polynomial spline.

We end this section by mentioning briefly what happens when we relax the requirement that $k^2(t)$ in (2.1) is strictly positive and consider the case when $k^2(t) \equiv 0$ on some interval $[\tau_1, \tau_2]$. Let $\hat{g}_*(\cdot)$ be the minimizer of (2.1) for this case. Obviously, each function from Ω_k^m that coincides with $\hat{g}_*(\cdot)$ outside $[\tau_1, \tau_2]$ and interpolates the

data within it cannot increase the value of the functional which is minimized by $\hat{g}_*(t)$ and, hence, it is also a solution of (2.1). Thus, there will be an *infinite* set of solutions of (2.1), which consists of functions that differ only within the interval $[\tau_1, \tau_2]$ for nondesign points. One of the curves in that set is the partially interpolant (in $[\tau_1, \tau_2]$) L_k-spline.

3. Bayesian interpretation

Wahba (1978, 1983), Wecker and Ansley (1983), and Kohn and Ansley (1988) showed that smoothing splines have a natural interpretation as Bayes estimators of the unknown response function. In this section, we show that our L_k-spline is also a Bayes estimate. The difference between the Bayesian models is that the polynomial smoothing spline corresponds to a prior distribution that treats the mth derivative of the response function as homoscedastic white noise while the prior for the L_k-spline allows the possibility of heteroscedasticity. Thus, for example, if we knew in advance that $g(\cdot)$ was nearly linear in some region, but might have high curvature in another region, we could reflect this by adopting $m = 2$ and a prior with lower variance in the former region. Not surprisingly, posterior interval estimates will tend to be wider in those regions where prior uncertainty was high. Thus improved inference should be possible if one can assign a prior that more closely reflects belief about the local non-linearity of the response function.

The basic idea behind the Bayesian models is to assign a prior distribution to the space of possible response functions. The prior defines a stochastic process on [0, 1] and the response function can be treated as a *sample function* from that process. Typically the stochastic process will be the sum of a parametric regression function, whose coefficients are unknown, and a random error function that expresses the deviation (bias) of the true function from the parametric one. This idea was first suggested by Blight and Ott (1975). Related models are considered in O'Hagan (1978) and Steinberg (1990). The analogous Bayesian interpretation for spline estimators was proposed by Wahba (1978, 1983) and has been further discussed by Silverman (1985) and Kohn and Ansley (1988). Some roots of the idea can be found in Kimeldorf and Wahba (1970, 1971). Steinberg (1983, Section 2.4) shows that all of these models have the common structure described above, differing only in the choice of prior distribution for the bias component.

The following theorems define the Bayesian model that corresponds to an L_k-spline and naturally extend Wahba's (1978, 1983) results for the polynomial splines arising in the *constant* smoothing parameter case:

Theorem 2. *Consider the model* (1.1)

$$y_i = g(t_i) + \varepsilon_i.$$

Now assume that $g(t)$ has a prior distribution given by the stochastic process

$$X_\xi(t) = \sum_{j=1}^{m} \theta_j \phi_j(t) + \gamma^{1/2} Z(t), \quad t \in [0, 1], \tag{3.1}$$

where $\phi_j(t) = t^{j-1}/(j-1)!$, $\theta = (\theta_1, \ldots, \theta_m)' \sim \mathcal{N}(\mathbf{0}, \xi I_m)$, γ is a positive constant and $Z(\cdot)$ is a Gaussian stochastic process with zero mean and covariance function $Q(s, t)$ defined in Section 2. We assign a diffuse prior to θ by allowing ξ to tend to infinity.

Then, for $\gamma = \sigma^2/n$, the posterior mean of $g(t)$ is precisely the L_k-spline $\hat{g}(\cdot)$, which is the minimizer of (2.1):

$$\hat{g}(t) = \lim_{\xi \to \infty} E_\xi \{ g(t) | \boldsymbol{y} \}.$$

Here E_ξ is the posterior mean of $g(\cdot)$ for a fixed, finite value of ξ and the posterior mean for the diffuse prior is obtained in the limit as $\xi \to \infty$.

The proof follows directly from Theorem 2 of Wahba (1978) for generalized splines.

The model yields

$$y(t) = \sum_{j=1}^{m} \theta_j \phi_j(t) + \gamma^{1/2} Z(t) + \varepsilon. \tag{3.2}$$

The first term in (3.2) is a polynomial of degree $m-1$, the last one is the 'standard' random error, while the second term represents prior belief regarding the deviation of the true model from the polynomial one in the sense of Blight and Ott.

Our prior distribution for $g(t)$ and its relation to the prior assumed by Wahba (1978, 1983) can be more easily understood by considering the local magnitude of the $(m-1)$st derivative of $g(t)$. Let $\Delta_h^m g(t) = g^{(m-1)}(t+h) - g^{(m-1)}(t)$ for some fixed, small increment h. By (3.1), the prior distribution for $\Delta_h^m g(t)$ is that of $\Delta_h^m X(t) = X^{(m-1)}(t+h) - X^{(m-1)}(t) = \gamma^{1/2}[Z^{(m-1)}(t+h) - Z^{(m-1)}(t)]$. Calculating the corresponding partial derivatives of $Q(s, t)$, it is easy to verify that $\Delta_h^m X(t)$ is a normal variable with zero mean and variance $\gamma \int_t^{t+h} \psi^2(u)\,\mathrm{d}u = \gamma \psi^2(t)\, h + o(h)$. Thus the roughness function $\psi^2(t)$ in the Bayesian model should reflect prior belief about the magnitude of $\Delta_h^m g(t)$. In Wahba's model, $\psi^2(t)$ is assumed to be constant, which corresponds to prior belief that $\Delta_h^m g(t)$ has about the same magnitude for all t. The heteroscedasticity of the modified prior proposed here allows us to take advantage of presumed or evident differences in the local behavior of the $(m-1)$st derivative of the unknown response function.

The mth derivative of the prior, $X^{(m)}(\cdot)$, can be defined formally as a process that satisfies the stochastic differential equation $\mathrm{d}^m X(t)/\mathrm{d}t^m = \gamma^{1/2} \psi(t) \mathrm{d}W(t)/\mathrm{d}t$, where $W(t)$ is a Wiener process with $\mathrm{Var}\{W(1)\} = 1$. When $\psi^2(t)$ is assumed to be constant, as in Wahba's model, this defines a 'white noise' process. When $\psi^2(t)$ is not constant, we may, by analogy, refer to $X^{(m)}(\cdot)$ as *generalized* 'white noise'.

Theorem 3. *The posterior variance of $g(t)$ for the prior of Theorem 2 is*

$$\mathrm{Var}\{g(t)|\,y\} = \gamma(Q(t,t) + \phi_t' W_1 \phi_t - 2\phi_t' W_2 q_t - q_t' W_3 q_t),$$

where $\phi_t = (\phi_1(t), \ldots, \phi_m(t))'$, $q_t = (Q(t,t_1), \ldots, Q(t,t_m))'$,

$$W_1 = (T'M^{-1}T)^{-1}, \qquad W_2 = (T'M^{-1}T)^{-1}T'M^{-1}$$

and

$$W_3 = M^{-1}(I - T(T'M^{-1}T)^{-1}T'M^{-1}).$$

The proof is analogous to that of Theorem 2 of Wahba (1983) for constant smoothing parameter.

Theorem 3 allows us to derive Bayesian posterior intervals for $g(\cdot)$. A $(1-\alpha)$-level Bayesian interval for the unknown function $g(\cdot)$ at a particular point t is given by

$$\hat{g}(t) \pm z_{1-\alpha/2}\sqrt{\mathrm{Var}(g(t|\,y))}. \tag{3.3a}$$

If we focus just on the observed values of t and denote $g = (g(t_1), \ldots, g(t_n))'$, then

$$g|\,y \sim \mathcal{N}(Ay, \sigma^2 A).$$

The interval estimate for $g(t_i)$ will then be

$$\hat{g}(t_i) \pm z_{1-\alpha/2}\sigma\sqrt{a_{ii}}. \tag{3.3b}$$

Generally σ^2 is unknown and must be estimated from the data. By analogy with linear regression the commonly used estimates for σ^2 are based on RSS and are of the form:

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-p}, \tag{3.4}$$

where $n-p$ is often called 'the equivalent degrees of freedom' (Wahba, 1983; Green and Silverman, 1994). The most popular extension for the equivalent degrees of freedom for nonparametric regression, which is based on the fact that $\sigma^2 A$ is the posterior covariance matrix of g, is $p = \mathrm{tr}\,A$ (Wahba, 1983; Silverman, 1985; Green and Silverman, 1994). Some related definitions are discussed in Buja et al. (1989). Carter and Eagleson (1992) have investigated whether (3.4) can be improved (in the constant smoothing parameter case) by a different choice of denominator. They show that an improvement in terms of MSE is possible by using $\hat{\sigma}^2 = \mathrm{RSS}/\mathrm{tr}\{(I-A)^2\}$. The relative difference in MSE between the two estimates can be rather significant for small samples but 'decreases rapidly with increasing sample size' (see Fig. 4. of Carter and Eagleson, 1992).

One can also assign a prior distribution to σ^2. Assuming the standard improper prior density $f(\sigma^2) \propto 1/\sigma^2$, the posterior distribution of σ^2 is, for fixed k^2, a scaled inverse χ^2_{n-m} with scale factor

$$s^2 = y'[M^{-1} - M^{-1}T(T'M^{-1}T)^{-1}T'M^{-1}]y/(n-m).$$

The proof is analogous to that of Steinberg (1990). Thus a Bayesian estimate of σ^2 is $\hat{\sigma}^2 = E(\sigma^2|y, k^2) = s^2$, which is the mean of a weighted residual sum of squares. A complete Bayesian approach would involve declaring a prior distribution for k^2 and then averaging the conditional estimate above with respect to the psoterior distribution for k^3.

If σ is unknown, $\hat{\sigma}$ is used in interval estimates (3.3) instead of σ.

Remark. Practically all necessary computations can be carried out effectively using the procedure of Hutchinson and de Hoog (1986) and also by the Kalman filter approach of Wecker and Ansley (1983) in $O(n)$ operations.

4. L_k-smoothing spline procedure and examples

To apply L_k-spline smoothing it is necessary to specify, or to estimate, the smoothing parameter $k^2(t)$. Our approach will be based on the Bayesian model described in Section 3. That model implies that the roughness function $\psi^2(t)$ should reflect the local variability of $g^{(m-1)}(t)$. Since the mth derivative of $g(t)$ is the natural analytic description of this local variability, it seems reasonable to choose $\psi^2(t)$ proportional to $(g^{(m)}(t))^2$. In some problems, prior knowledge about $g^{(m)}(t)$ might serve as a guide to choosing $\psi^2(t)$. Typically, though, we think that it will be necessary to estimate $\psi^2(t)$ from the data.

Our approach is as follows. First, we fit a polynomial smoothing spline $\hat{g}_0(\cdot)$, with a *constant* smoothing parameter k_0^2 chosen by generalized cross-validation (GCV). Then we estimate $g^{(m)}(\cdot)$ by the mth derivative of $\hat{g}_0$, which is a piecewise polynomial of degree $m-1$ and may be easily obtained from (2.2):

$$\hat{g}_0^{(m)}(t) = \psi_0^2 \sum_{j=1}^{n} d_{j0} \frac{(t_j - t)_+^{m-1}}{(m-1)!}.$$

At the second step we fit an L_k-smoothing spline according to (2.2) with $\psi^2(t) = \rho \hat{g}_0^{(m)}(t)^2$ where the coefficient ρ is also derived by GCV. An analogous two-step scheme was proposed by Müller and Stadtmüller (1987) for kernel estimation with bandwidth dependent on the mth derivative of the unknown function.

Since $\hat{g}_0^{(m)}(\cdot)$ may behave poorly and exibit large random fluctuations for small and medium samples, we use a truncated variant of the above estimate of $\psi^2(\cdot)$:

$$\psi^2(t) = \max(\tau\psi_0^2, \rho \hat{g}_0^{(m)}(t)^2). \tag{4.1}$$

The similar truncated estimate was used by Müller and Stadtmüller (1987) for the variable smoothing parameter in kernel estimation.

Running several examples, we studied different values for τ by visual analysis of the resulting plot of $\psi^2(t)$: we decreased the value of τ untill the basic features of the plot began to become distorted due to the random noise, which strengthens with

decreasing τ. We found that values 0.5–0.8 perform quite satisfactorily for most cases and may be recommended as appropriate values for τ. For 'smoother' functions even smaller values of τ may be successfully used (see Example 1, where the original function was linear almost everywhere). An 'optimal' choise of τ depends, of course, also on the number of observations. Theoretically, τ may enter the GCV-criterion as a second parameter (together with ρ); however that complicates the procedure and in our opinion is superfluous.

Ideally, we would like to modify the error intervals to account for the extra uncertainty that arises from the need to estimate $\psi(t)$. However, even in the simple case, where $\psi(t)$ is assumed constant, no procedures have been suggested for solving this problem and the error intervals in the literature have all treated the estimated smoothing parameter as if it had been known in advance, rather in the spirit of empirical Bayes procedures (see Wahba, 1983; Silverman, 1985; Ansley et al., 1993). The effects of the additional uncertainty in estimating $\psi(t)$ are likely to be more serious for our approach due to the extra flexibility that we allow in the form of the smoothing parameter. However, the examples considered below show that even in this case the corresponding plug-in intervals are quite reasonable.

We carried out several Monte Carlo experiments to verify the effectiveness of L_k-spline smoothing for inference in nonparametric regression. We considered the case $m = 2$ which is used most often in practice.

Example 1. We reanalyse here Case 4 of Wahba (1983):

$$g(t) = \begin{cases} 0, & 0 \leqslant t \leqslant \frac{1}{3}, \\ 36(t - \frac{1}{3}), & \frac{1}{3} < t \leqslant \frac{1}{2}, \\ 36(\frac{2}{3} - t), & \frac{1}{2} < t \leqslant \frac{2}{3}, \\ 0, & \frac{2}{3} < t \leqslant 1. \end{cases}$$

In this example, the cubic spline is a good global estimator of g, but does not give acceptable results at the three points where $g'(t)$ is discontinuous.

Formally, since $g(t)$ is a triangular function and has a discontinuity in its first derivative, the spline estimate with $m = 2$ may not be applicable. But the corresponding results turned out to be quite good (see also Wahba's paper), so this example may serve as a good illustration for L_k-spline smoothing.

We drew 50 random samples of size $n = 128$ by adding to $g(t_i)$ a normal random error with zero mean and $\sigma = 0.1$. Data points $\{t_i\}$ were equally spaced with $t_i = (i - 1)/n$, and σ^2 was estimated by $\hat{\sigma}^2$ from (3.4) with $p = \operatorname{tr} A$.

The results for cubic spline smoothing were very similar to those of Wahba. Among the 95% pointwise Bayesian probability intervals (3.3b), 95.09% covered the true points, but for the most 'interesting' points near $t = 0.33, 0.5, 0.66$, where the original function has sharp peaks, the corresponding proportions were 0.62, 0.08 (!) and 0.70 respectively (see Fig. 1). It is obvious that the prior belief of homogeneity of the second derivative of the true function, on which the cubic spline estimator is based, is not reasonable at all for these points. The bias for them is extremely large (see Fig. 2).

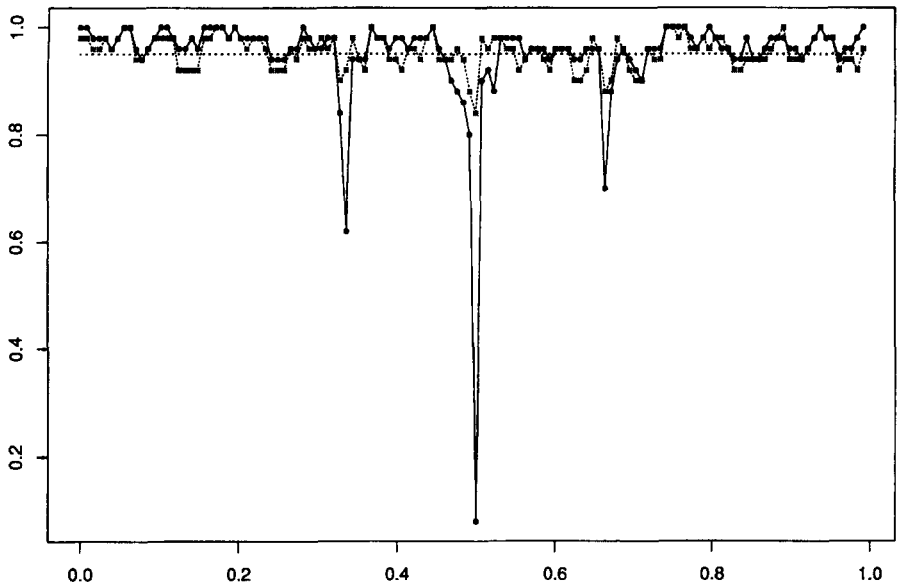

Fig. 1. Pointwise coverage probabilities of Bayesian intervals estimated from 50 trials: cubic spline (solid line) and L-spline (dashed line).

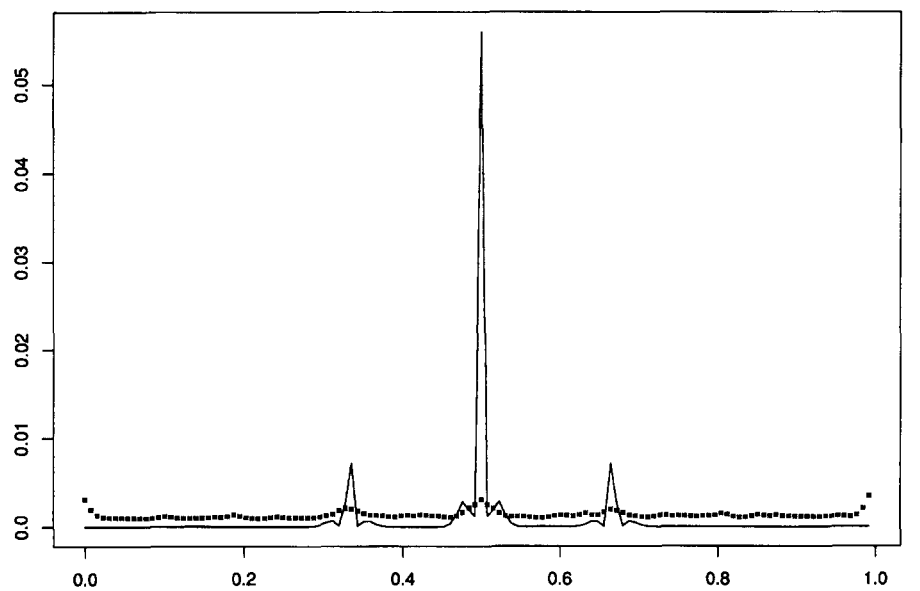

Fig. 2. Squared bias of spline estimates: cubic spline (solid line) and L-spline (asterisks) (based on one trial).

The corresponding results for L_k-spline smoothing are also shown in Figs. 1 and 2. The pointwise coverage probabilities for the problematic points are significantly improved and became 0.92, 0.86 and 0.88, respectively. The bias of the estimates at these points is drastically decreased.

It turned out that the variance estimate (3.4) for L_k-smoothing splines generally underestimates the real value of σ^2. This is not so surprising, since now we have to estimate more parameters from the same data-set, which leads to underestimating the variance. To avoid this we used the variance estimate obtained from the initial constant smoothing parameter spline, which was quite satisfactory. The use of Carter and Eagleson's (1992) variance estimate (see Section 3) might be helpful here but needs further study.

An example of the estimated roughness function $\psi^2(t)$ is given in Fig. 3. τ in (4.1) was taken as 0.1. One sees that this estimate is quite close to what might be used by someone who knew the true function $g(t)$.

Figure 4 shows the diagonal elements of the hat matrix A which are proportional to the posterior variance at the data points and, therefore, to the squared length of the corresponding Bayesian intervals (3.3b). Neglecting the boundary effects, the variation in a_{ii} for the cubic spline is very small, so all the posterior probability intervals are approximately of the same length. For the L_k-smoothing spline, however, the diagonal elements of A vary significantly, increasing when the real function has rapid local changes and decreasing in smooth regions. This yields the broader error intervals at problematic points and narrower intervals where the function is smooth. In the linear regions of $g(t)$ the L_k-spline intervals were generally about 30% narrower than the corresponding cubic spline intervals with no significant drop in coverage probabilities. This ratio also depends on the chosen value of τ in (4.1). Thus the L_k-spline

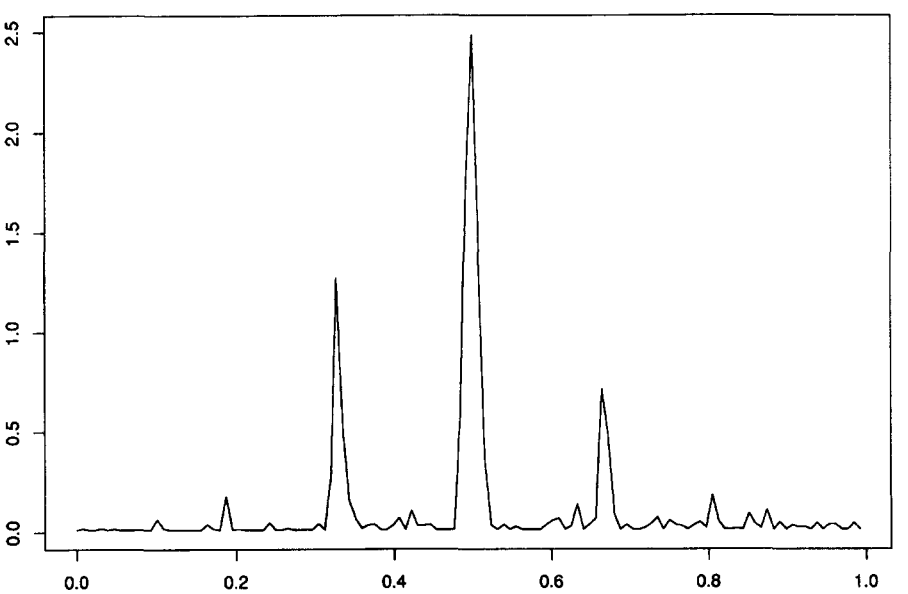

Fig. 3. Truncated estimate of roughness function x 1e-9 (based on one trial).

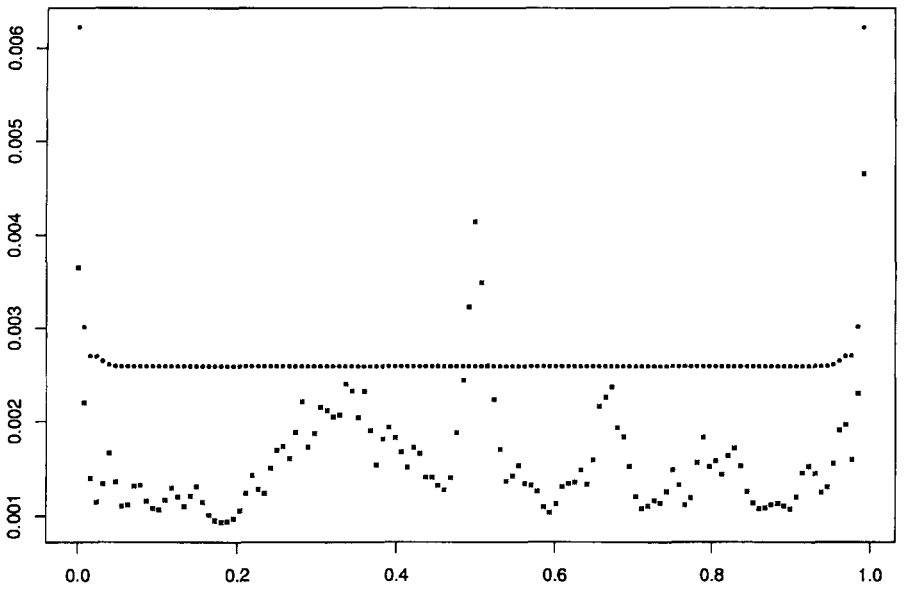

Fig. 4. Diagonal elements of hat-matrix A vs. $t(i) = (i - 1)/n$ for cubic spline (points) and L-spline (asterisks) (based on one trial).

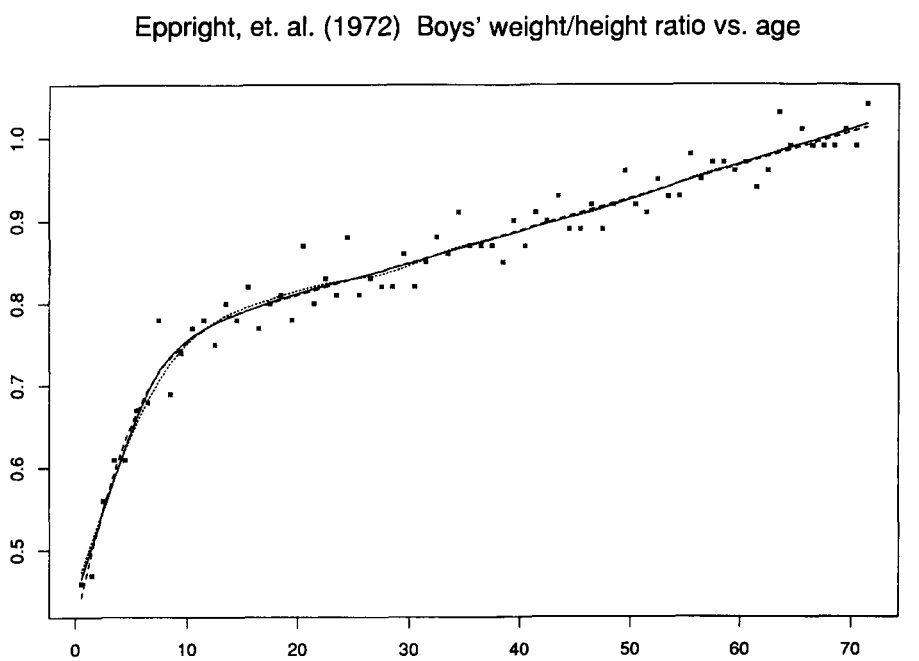

Fig. 5. Data, quadratic-quadratic-linear (dashed line) and spline estimates: cubic (dotted line) and L (solid line).

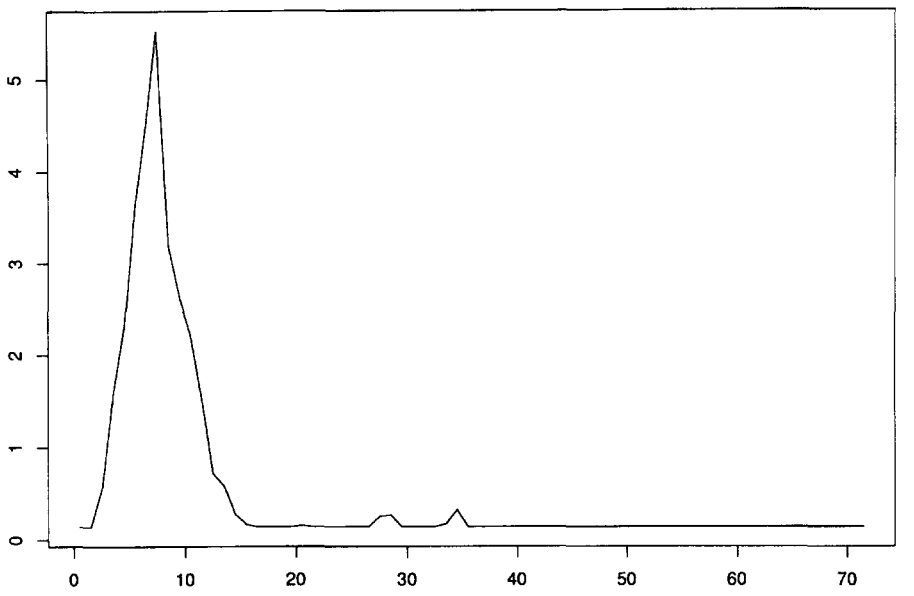

Fig. 6. Truncated roughness function.

provides much better coverage properties at the points of nonlinearity, and narrower intervals in the regions of linearity.

Sometimes spline smoothing is a catalyst for better understanding of a model's structure and even may hint at an appropriate parametric model. The following example may serve as a useful illustration.

Example 2. The data show weight/height ratio (y) against age (x) for pre-school boys sampled from families in north central states of the USA (see Fig. 5) and are given in Eppright et al. (1972). Different approaches for fitting these data were proposed in Gallant and Fuller (1973), who studied piecewise-polynomial parametric models, and in Eubank (1988), who used the data to illustrate the cubic spline smoothing technique.

We applied the L_k-spline smoothing procedure described in this paper (with $\tau = 0.6$) to estimate the response function (see Fig. 5). Analysis of the estimated roughness function (see Fig. 6) hints at a three-segmented structure of the data, where the third segment (x greater than 13–15) seems to be linear ($\psi^2(\cdot)$ is close to zero), while for small x a two-segmented polynomial model of some low degree with the knot about 7–8 (the peak of $\psi^2(\cdot)$) seems to be reasonable for the data. Such a *parametric* segmented polynomial regression model with unknown knots and a requirement of continuity for the estimate's derivative was originally proposed for these data by Gallant and Fuller (1973).

They studied a quadratic-quadratic-linear model with two unknown knots α_1 and α_2 that can be written as

$$g(x) = \theta_1 + \theta_2 x + \theta_3(\alpha_1 - x)_+^2 + \theta_4(\alpha_2 - x)_+^2 .$$

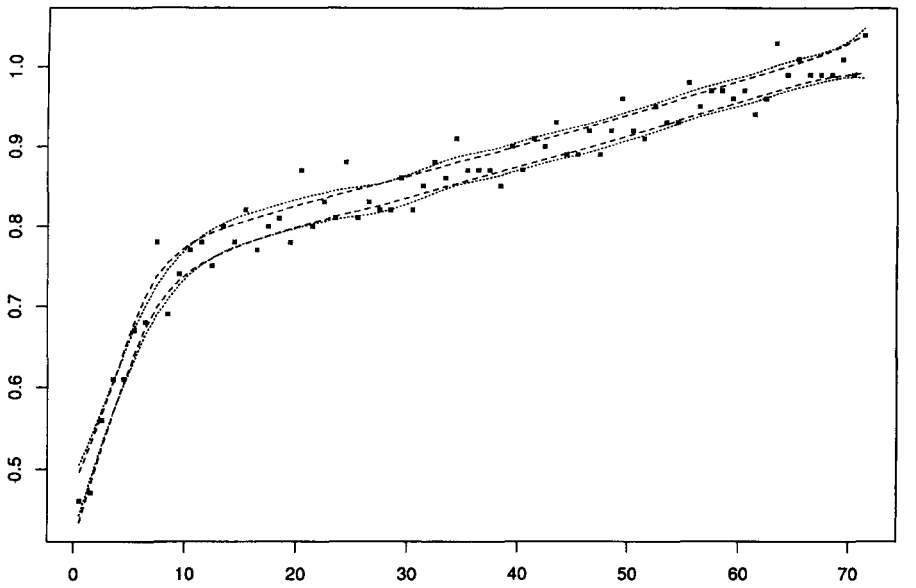

Fig. 7. Bayesian intervals for cubic spline (dotted lines) and L-spline (dashed lines).

Their estimated quadratic–quadratic-linear model fits the data quite satisfactorily (see Fig. 5) and is very close to the L_k-spline estimate. The estimated knots were $\hat{\alpha}_1 \approx 8.3$, $\hat{\alpha}_2 \approx 14.8$, which nicely matches our preliminary considerations.

Eubank (1988) used polynomial (cubic) spline smoothing to estimate the response function for the weight/height ratio data. Comparison of the cubic spline estimate with the L_k-spline shows that both estimates are very similar and are close to the quadratic-quadratic-linear estimate (see Fig. 5); however in the 'linear' regions the L_k-spline's probability intervals are about 30% narrower (see Fig. 7).

Acknowledgement

We would like to thank Bernard Silverman and the anonymous referee for their helpful comments.

References

Ansley, C.F., R. Kohn and C.-H. Wong (1993). Nonparametric spline regression with prior information. *Biometrika* **80**, 75–88.

Blight, B.J.N. and C. Ott (1975). A Bayesian approach to model inadequacy for polynomial regression. *Biometrika* **62**, 79–88.

Buja, A., T. Hastie and R. Tibshirani (1989). Linear smoothers and additive models. *Ann. of Statist.* **17**, 453–555 (with discussion).

Carter, C.K. and G.K. Eagleson (1992). A comparison of variance estimators in nonparametric regression. *J. Roy. Statist. Soc. Ser. B* **54**, 773–780.

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

Eppright, E.S., H.M. Fox, B.A. Fryer, G.H. Lamkin, V.M. Vivian and E.S. Fuller (1972). Nutrition of infants and preschool children in the north central region of the United States of America. *World Rev. Nutrition Dietetics* **14**, 269–332.

Eubank, R.L. (1988). *Spline Smoothing and Nonparametric Regression.* Dekker, New York.

Gallant, A.R. and W.A. Fuller (1973). Fitting segmented polynomial regression models whose joint points have to be estimated. *J Am. Statist. Assoc.* **68**, 144–147.

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

Hutchinson, M. and F. de Hoog (1985). Smoothing data with spline functions. *Numerische Mathematik* **47**, 99–106.

Kimeldorf, G.S. and G. Wahba (1970). A correspondance between Bayesian estimation of stochastic processes and smoothing by splines. *Ann. Math. Statist.* **41**, 495–502.

Kimeldorf, G.S. and G. Wahba (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* **33**, 82–95.

Kohn, R. and C.F. Ansley (1983). On the smoothness properties of the best linear unbiased estimate of a stochastic process observed with noise. *Ann. Statist.* **11**, 1011–1017.

Kohn, R. and C.F. Ansley (1988). Equivalence between Bayesian smoothness priors and optimal smoothing for function estimation. In: J.C. Spall, Ed., *Bayesian Analysis of Time Series and Dynamic Models.* Marcel Dekker, New York.

Müller, H.G. and V. Stadtmüller (1987). Variable bandwidth kernel estimators of regression curves. *Ann. Statist.* **15**, 182–201.

Nychka, D.W. (1988). Bayesian confidence intervals for a smoothing spline. *J. Am. Statist. Assoc.* **83**, 1134–1143.

O'Hagan, A. (1978). Curve fitting and optimal design for prediction. *J. Roy. Statist. Soc. Ser. B* **40**, 1–41 (with discussion).

Reinsch, C. (1967). Smoothing by spline functions. *Numerische Mathematik* **10**, 177–183.

Schoenberg, I.J. (1964). Spline functions and the problem of graduation. *Proc. National Academy Sci. USA.* **52**, 947– 950.

Silverman, B.W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *J. Roy. Statist. Soc. Ser. B* **47**, 1–52 (with discussion).

Steinberg, D.M. (1983). Bayesian models for response surfaces and their implications for experimental design. Ph.D. Thesis, University of Wisconsin–Madison.

Steinberg, D.M. (1990). A Bayesian approach to flexible modeling of multivariable response functions. *J. Multivariate Anal.* **34**, 157–172.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B* **40**, 364–372.

Wahba, G. (1983). Bayesian confidence intervals for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B* **45**, 133–150.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.* **13**, 1378–1402.

Wahba, G. (1990). *Spline Models for Observational Data*, SIAM, Philadelphia.

Wecker, W.E. and Ansley, C.F. (1983). The signal extraction approach to nonlinear regression and spline smoothing. *J. Am. Statist. Assoc.* **78**, 81–89.