J. R. Statist. Soc. B (2003)
65, *Part* 4, *pp.* 901–919

Free-knot polynomial splines with confidence intervals

Wenxin Mao and Linda H. Zhao

University of Pennsylvania, Philadelphia, USA

[Received November 2001. Revised March 2003]

Summary. We construct approximate confidence intervals for a nonparametric regression function, using polynomial splines with free-knot locations. The number of knots is determined by generalized cross-validation. The estimates of knot locations and coefficients are obtained through a non-linear least squares solution that corresponds to the maximum likelihood estimate. Confidence intervals are then constructed based on the asymptotic distribution of the maximum likelihood estimator. Average coverage probabilities and the accuracy of the estimate are examined via simulation. This includes comparisons between our method and some existing methods such as smoothing spline and variable knots selection as well as a Bayesian version of the variable knots method. Simulation results indicate that our method works well for smooth underlying functions and also reasonably well for discontinuous functions. It also performs well for fairly small sample sizes.

Keywords: *B*-splines; Confidence intervals; Free knots; Nonparametric regression; Piecewise polynomials

1. Introduction

The nonparametric regression model

$$y_i = f(x_i) + \sigma\varepsilon_i, \qquad i = 1, \ldots, n, \qquad \varepsilon_i \overset{\text{IID}}{\sim} N(0, 1), \tag{1}$$

with σ^2 unknown, has been studied extensively. We are interested in constructing practical estimates that are accompanied by confidence intervals for the underlying function values $f(x)$.

There are several procedures to estimate f. Kernel-type methods include kernel regression (Nadaraya, 1964) and local polynomial fitting (Fan and Gijbels, 1996; Loader, 1999). Confidence bands based on kernel estimators can be derived with bootstrap methods (Härdle and Marron, 1991) or bias correction methods (Eubank and Speckman, 1993; Xia, 1998). Wavelets are now also widely used and some recent literature has begun to investigate confidence intervals based on wavelet estimators (Picard and Tribouley, 2000).

Spline models provide another popular method for estimating f. Wahba (1983) discussed confidence intervals based on smoothing spline estimators. For a detailed description of smoothing splines, see Wahba (1990).

Because of their conceptual simplicity, polynomial spline methods have been widely used to construct estimators. In these $f(x)$ is estimated by a piecewise mth order ($(m-1)$th degree) polynomial connecting smoothly at points $t_1 < \ldots < t_r$, which are referred to as interior knots. It

Address for correspondence: Linda H. Zhao, Department of Statistics, University of Pennsylvania, Philadelphia, PA 19104, USA.
E-mail: lzhao@wharton.upenn.edu

 1369–7412/03/65901

is important to choose appropriately the number of knots r and their locations. In the current variable knot selection literature, the possible knots come from a predetermined set such as the design points or grid points in the range. A final set of knots is then chosen from these. Depending on the approach the choice of knots may involve linear regression model selection or a Bayesian model. The estimation of regression coefficients given the knots is via linear least squares. Some references which used this approach are Friedman and Silverman (1989), Friedman (1991) and Stone *et al.* (1997). Some more recent, effective variations are in Smith and Kohn (1996), Denison *et al.* (1998), Lindstrom (1999), DiMatteo *et al.* (2001) and Zhou and Shen (2001). All the work cited above concerns the estimation of f. Zhou *et al.* (1998) have provided confidence intervals for f.

Following the appearance of a preprint version of the present paper Kooperberg and Stone (2002) used a closely related free-knot construction of confidence intervals for nonparametric density estimates.

We use free-knot polynomials, i.e. both the knot locations and the regression coefficients are considered to be unknowns to be estimated. This provides flexibility to allow inhomogeneous smoothness of $f(x)$ which can then be fully estimated by the data. Asymptotic confidence intervals can be constructed through a simple classical idea.

We emphasize that model selection is used only to choose the optimal number of knots, not to choose knot locations among a large set of possible locations, as is done in the existing variable knot schemes that were cited above. Partly because of this minimal use of model selection we can expect that the confidence intervals that we construct will have coverage probabilities that are close to their nominal values. Numerical results and some comparisons with smoothing splines and variable knots schemes including a Bayesian version are presented in Section 4.

We now briefly introduce the method. We may view the set of order m splines with r interior knots as a given family of piecewise polynomial functions $\{f(\boldsymbol{\theta}, x): \boldsymbol{\theta}\}$. The $(2r+m)$-dimensional parameter vector $\boldsymbol{\theta}$ describes the r knot locations along with the $r+m$ necessary polynomial coefficients. The functions $f(\boldsymbol{\theta}, x)$ are piecewise polynomials of $(m-1)$th degree. If the knots are distinct then they have $m-2$ everywhere continuous derivatives. The function $f(\boldsymbol{\theta}, x)$ may have a lower degree of smoothness at locations where a multiplicity of knots occurs. For convenience we fix $m = 4$ throughout.

The idea is to view model (1) as if it were a parametric non-linear regression model:

$$y_i = f(\boldsymbol{\theta}, x_i) + \sigma\varepsilon_i, \qquad i = 1, \ldots, n. \tag{2}$$

The estimation part of the statistical analysis involves first fixing r and estimating $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_r$ by maximum likelihood within model (2). Then an estimated best value of r, $r_{\min}$, is chosen through a generalized cross-validation (GCV) model selection device. The function $\hat{f} = f(\hat{\boldsymbol{\theta}}_{r_{\min}}, x)$ is our estimate of f.

The description of this estimator also makes feasible the construction of appealing confidence intervals for $f(x)$. In addition to their heuristic and asymptotic motivation these intervals perform well in various simulations. To understand the primary methodology note that when f is itself a polynomial spline with r knots we can write asymptotically

$$\hat{\boldsymbol{\theta}}_r \sim N\{\boldsymbol{\theta}, \sigma^2 \mathbf{I}_{(1)}^{-1}(\boldsymbol{\theta})\} \qquad \text{as } n \to \infty,$$

where $\mathbf{I}_{(1)}$ denotes the appropriate information matrix given in equations (15)–(18) in Section 3.3.

Let

$$\mathbf{d}^{\mathrm{T}} = \frac{\partial f}{\partial \boldsymbol{\theta}} = \left(\frac{\partial f}{\partial \theta_1}, \ldots, \frac{\partial f}{\partial \theta_s} \right),$$

which depends on x and $\boldsymbol{\theta}$. Here $s = 2r + 4$ is the number of relevant parameters. The variance of $f(\hat{\boldsymbol{\theta}}, x)$ given σ^2 can be approximated by the delta method as $\sigma^2 \mathbf{d}^{\mathrm{T}} \mathbf{I}_{(1)}^{-1}(\boldsymbol{\theta}) \mathbf{d}$, because

$$f(\hat{\boldsymbol{\theta}}, x) \approx f(\boldsymbol{\theta}, x) + \frac{\partial f}{\partial \boldsymbol{\theta}} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}).$$

In addition, let $\tilde{\sigma}^2$ be an estimator of σ^2. Then we can derive an approximate $100(1 - \alpha)\%$ confidence interval for $f(\boldsymbol{\theta}, x)$ as

$$f(\hat{\boldsymbol{\theta}}, x) \pm z_{\alpha/2} \sqrt{\{\tilde{\sigma}^2 \, \mathbf{d}^{\mathrm{T}}(x) \, \mathbf{I}_{(1)}^{-1}(\boldsymbol{\theta}) \, \mathbf{d}(x)|_{\hat{\boldsymbol{\theta}}}\}}, \tag{3}$$

where $P(Z > z_{\alpha/2}) = \alpha/2$ for a standard normal variable Z.

The estimation idea following model (2) seems to be very natural. Indeed, it has been mentioned frequently in the literature. Early references for splines are de Boor and Rice (1968) and de Boor (1978). Free-knot splines have not been adopted widely by statisticians, partly because of their computational difficulty. Jupp (1978) subsequently addressed this problem which has itself become a subject with an extensive history, but this is not the primary topic of this paper. Recent developments, both in computing power and methodology, have made the idea feasible. Section 3.1 gives a brief review of the history.

Our method is locally adaptive to variable smoothness in f because the procedure automatically places more knots in regions where f is not smooth. Furthermore, the family $\{f(\boldsymbol{\theta}, x)\}$ contains functions that have discontinuous derivatives or are themselves discontinuous. These appear naturally as splines having repeated knots at the locations of discontinuities. Because of this, the method that we propose can reasonably effectively deal with functions f having isolated discontinuities or discontinuous derivatives.

This paper is organized as follows. In Section 2 we give some background about B-splines. In Section 3 we give details of our method. In Section 4 we apply this method to simulated data. Section 5 discusses free-knot methodology and reports empirical results that support this as a confidence set methodology. It also describes some alternative types of confidence bands.

2. *B*-splines

An mth-order polynomial spline on $[a, b]$ with r ordered interior knots $\mathbf{t} = (t_1, \ldots, t_r)$ is a piecewise polynomial (of degree $m - 1$). When $a < t_1 < \ldots < t_r < b$ these piecewise polynomials connect at each knot with continuous $(m - 2)$th derivatives. The space of all such functions, $S_{m,r,\mathbf{t}}$, is a linear space of dimension $m + r$. In particular, when $m = 4$, $S_{4,r,\mathbf{t}}$ consists of all cubic splines. Throughout we shall use only the value $m = 4$, which produces plots that are visually smooth, but our methodology can easily be applied with other values of m. The commonly used bases for $S_{4,r,\mathbf{t}}$ are the truncated basis and the B-spline basis. The truncated basis has a simple form and is easy to understand, but it is less stable computationally (Dierckx, 1993). Because of analytical and computational advantages the standard B-spline basis is used below, though its natural spline version could be used instead (Eubank, 1988; Greville, 1969).

The B-splines for $m = 4$ are completely determined by the interior knots $\mathbf{t}$. They can be conveniently defined by using divided difference notation. For an arbitrary ordered set of points $\tau_1 \leqslant \ldots \leqslant \tau_k$, $k \geqslant 2$, inductively define the operator

$$\left.\begin{aligned}
[\tau_1,\tau_2]\,g(\cdot) &= g'(\tau_1), && \text{if } \tau_1=\tau_2,\\
[\tau_1,\tau_2]\,g(\cdot) &= \frac{g(\tau_2)-g(\tau_1)}{\tau_2-\tau_1}, && \text{if } \tau_1\neq\tau_2,\\
[\tau_1,\ldots,\tau_k]\,g(\cdot) &= g^{(k-1)}(t), && \text{if } \tau_1=\ldots=\tau_k=t,\\
[\tau_1,\ldots,\tau_k]\,g(\cdot) &= \frac{(\tau_2,\ldots,\tau_k)\,g(\cdot)-(\tau_1,\ldots,\tau_{k-1})\,g(\cdot)}{\tau_k-\tau_1}, && \text{otherwise.}
\end{aligned}\right\} \tag{4}$$

Now let $t_{-3}=\ldots=t_0=a<t_1\leqslant\ldots\leqslant t_r<b=t_{r+1}=\ldots=t_{r+4}$. The B-spline $N_i(x,\mathbf{t})$ is defined to be

$$N_i(x,\mathbf{t})=(t_i-t_{i-4})(t_{i-4},\ldots,t_i)(\cdot-x)_+^3. \tag{5}$$

From equations (4) and (5), we can see that a recursive relationship can be used to describe B-splines; it provides a very stable numerical algorithm.

B-splines are non-zero only on an interval which covers no more than $m+1=5$ knots. Equivalently at any point x there are no more than $m=4$ B-splines that are non-zero.

For a function that is representable by a B-spline basis with a given set of knots, the degree of smoothness at a point is related to the number of repeating knots at that point by

$$\text{number of stacked knots} + \text{degree of smoothness} = \text{order}.$$

For example, if $t_k=t_{k+1}=t_{k+2}$ is used three times in constructing the B-splines, then at $t=t_k$ the degree of smoothness is $4-3=1$, which means that $f(x)$ is continuous at $t=t_k$ but $f'(x)$ is discontinuous at $t=t_k$.

The derivatives of $N_i(x,\mathbf{t})$ with respect to $\mathbf{t}$ will be needed in the next section. We take these from Schumaker (1981), page 132.

Lemma 1. When $i\leqslant j\leqslant i+4$, we have

$$\frac{\partial N_i(x,\mathbf{t})}{\partial t_j}=\begin{cases}
(t_{i+4}-t_i)(t_i,\ldots,t_j,t_j,\ldots,t_{i+4})(\cdot-x)_+^3, & i<j<i+4,\\
-(t_i,t_i,\ldots,t_{i+3})(\cdot-x)_+^3, & j=i,\\
(t_{i+1},\ldots,t_{i+3},t_{i+4},t_{i+4})(\cdot-x)_+^3, & j=i+4,\\
0, & \text{otherwise.}
\end{cases} \tag{6}$$

3. Methodology

Given the number of knots r we model the mean function to lie in $S_{4,r,\mathbf{t}}$. Thus we treat the data as if they came from the regression model

$$y_i=\sum_{j=1}^{r+4}\beta_j N_j(x_i,\mathbf{t})+\sigma\varepsilon_i, \qquad i=1,\ldots,n, \tag{7}$$

where $\varepsilon_i\sim^{\text{IID}}N(0,1)$, $\boldsymbol{\theta}=(\boldsymbol{\beta},\mathbf{t})$, σ^2 and r are unknown parameters with $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_{r+4})^{\mathrm{T}}$ and $\mathbf{t}=(t_1,\ldots,t_r)^{\mathrm{T}}$. We first estimate $\boldsymbol{\theta}$ and σ^2 conditional on r, which will be chosen later through the GCV criteria described in Section 3.4.

3.1. Estimation of f

Conditional on r we shall use the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_r$ to estimate $\boldsymbol{\theta}$. Because of the normal errors in model (7) it is easy to see that $\hat{\boldsymbol{\theta}}_r$ solves the non-linear least squares problem

$$\min_{\theta}\left[\sum_{j=1}^{n}\left\{y_j - \sum_{i=1}^{r+4}\beta_i N_i(x_j,\mathbf{t})\right\}^2\right], \tag{8}$$

which yields

$$\hat{f}_r(x) = \sum_{i=1}^{r+4}\hat{\beta}_i N_i(x,\hat{\mathbf{t}}). \tag{9}$$

The basic idea for solving problem (8) is the following. Given $\mathbf{t}$, let

$$F(\boldsymbol{\beta},\mathbf{t}) = \sum_{j}\left\{y_j - \sum_{i=1}^{r+4}\beta_i N_i(x_j,\mathbf{t})\right\}^2. \tag{10}$$

The linear least squares solution of $\boldsymbol{\beta}$ is produced, i.e. $G(\mathbf{t}) = \min_{\beta}\{F(\boldsymbol{\beta},\mathbf{t})\}$, and then we search for the minimum of $G(\mathbf{t})$. This non-linear optimization problem needs to be treated carefully. Given a starting value $\mathbf{t}^*$, a local optimum can be obtained from the Newton–Raphson algorithm. If G were strictly concave, the true minimum would be unique and could be easily found.

Jupp (1978) pointed out that this simple method is not foolproof in free-knot spline regression. There are too many saddlepoints and minima on the least squares surface. For certain examples the chance of finding the global minimum on the basis of a few sets of initial knots may be very small with the original parameterization and the Newton–Raphson algorithm has an appreciable chance of converging to local minima.

Several programs are available to calculate $\min\{G(\mathbf{t})\}$ beginning from an initial choice of knots. We use the International Mathematical and Statistical Libraries' routine DBSVLS (double-precision B-spline variable knots least squares); see de Boor (1998) for a reference. We have found this very fast and stable and its computational speed makes feasible the use of several repetitions in the search for a minimum, beginning from varied initial knot locations. This is an important step to help to eliminate falsely identifying local minima. The statistical performance of our procedure is not overly sensitive to the final local minimum; see Sections 3.6 and 5.

3.2. *Estimating* σ^2

In the case of a linear model, the usual estimator of σ^2 is $\hat{\sigma}^2 = \mathrm{SSE}/(n-k)$, where SSE is the sum of squared residuals and k is the number of regression coefficients. It is natural to extend this estimator to our non-linear regression as

$$\tilde{\sigma}^2 = \frac{\mathrm{SSE}}{n-(2r+4)}, \tag{11}$$

since $2r+4$ is the number of free parameters in our model. This estimator is approximately unbiased when f is a spline and works well in our simulations. It agrees with the general suggestion for non-linear least squares models in references such as Hastie and Tibshirani (1990) or Bates and Watts (1988).

In our simulations we have also investigated other methods of estimating σ^2 directly from the data (Rice, 1984). One possibility is

$$\tilde{\sigma}_1^2 = \frac{1}{n-2}\sum_{i=1}^{n-2}(0.809y_i - 0.5y_{i+1} - 0.309y_{i+2})^2. \tag{12}$$

as proposed in Hall *et al.* (1990). Our simulations (Mao, 2000) indicate that this does not perform as well as $\tilde{\sigma}^2$ in our setting. For a review on this and other difference-based variance estimators, see Dette *et al.* (1998).

3.3. Estimation of var($\hat{f}$)

Standard results for asymptotic efficiency of maximum likelihood estimators are then used to assess the variability of $\hat{f}(x)$. The relevant formulae are summarized below.

To proceed, we write model (7) in matrix form as

$$\mathbf{Y} = \mathbf{f}(\mathbf{t}, \boldsymbol{\beta}, \mathbf{X}) + \sigma^2 \varepsilon, \tag{13}$$

where $\mathbf{Y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\mathbf{X} = (x_1, \ldots, x_n)^{\mathrm{T}}$,

$$\mathbf{f}(\mathbf{t}, \boldsymbol{\beta}, \mathbf{X}) = \begin{pmatrix} N_1(x_1, \mathbf{t}) & \ldots & N_{r+4}(x_1, \mathbf{t}) \\ \vdots & & \vdots \\ N_1(x_n, \mathbf{t}) & \ldots & N_{r+4}(x_n, \mathbf{t}) \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_{r+4} \end{pmatrix} = \begin{pmatrix} \sum N_j(x_1, \mathbf{t}) \beta_j \\ \vdots \\ \sum N_j(x_n, \mathbf{t}) \beta_j \end{pmatrix} \tag{14}$$

and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$, where $\varepsilon \sim N(0, I)$.

Let

$$\begin{aligned} \mathbf{D}_{n \times (2r+4)} &= \left(\frac{\partial \mathbf{f}}{\partial \mathbf{t}}, \frac{\partial \mathbf{f}}{\partial \boldsymbol{\beta}} \right) \\ &= \begin{pmatrix} \sum_{i=1}^{r+4} \beta_i \frac{\partial N_i(x_1, \mathbf{t})}{\partial t_1} & \ldots & \sum_{i=1}^{r+4} \beta_i \frac{\partial N_i(x_1, \mathbf{t})}{\partial t_r} & N_1(x_1, \mathbf{t}) & \ldots & N_{r+4}(x_1, \mathbf{t}) \\ \vdots & \ldots & \ldots & \ldots & \ldots & \vdots \\ \sum_{i=1}^{r+4} \beta_i \frac{\partial N_i(x_n, \mathbf{t})}{\partial t_1} & \ldots & \sum_{i=1}^{r+4} \beta_i \frac{\partial N_i(x_n, \mathbf{t})}{\partial t_r} & N_1(x_n, \mathbf{t}) & \ldots & N_{r+4}(x_n, \mathbf{t}) \end{pmatrix}. \end{aligned} \tag{15}$$

It is straightforward to check that the information matrix for $(\boldsymbol{\theta}, \sigma)$ is

$$\mathbf{I}(\boldsymbol{\theta}, \sigma) = \begin{pmatrix} \dfrac{\mathbf{D}^{\mathrm{T}}\mathbf{D}}{\sigma^2} & \mathbf{0} \\ \mathbf{0} & \dfrac{n}{\sigma^2} \end{pmatrix}. \tag{16}$$

Let

$$\begin{aligned} \mathbf{d}^{\mathrm{T}} = \frac{\partial f}{\partial \boldsymbol{\theta}} &= \left(\frac{\partial f}{\partial \theta_1}, \ldots, \frac{\partial f}{\partial \theta_{2r+4}} \right) \\ &= \left(\sum_{i=1}^{r+4} \frac{\partial N_i(x, \mathbf{t})}{\partial t_1}, \ldots, \sum_{i=1}^{r+4} \frac{\partial N_i(x, \mathbf{t})}{\partial t_r}, N_1(x, \mathbf{t}), \ldots, N_{r+4}(x, \mathbf{t}) \right). \end{aligned} \tag{17}$$

Standard results on asymptotic normality of maximum likelihood estimators (Lehmann (1999), theorems 7.5.1 and 5.4.6) yield the following theorem.

Theorem 1. If f is a spline with r knots, as in equation (13), define $\mathrm{var}\{\hat{f}_r(x)\}$ by

$$\mathrm{var}\{\hat{f}_r(x)\} = \sigma^2 \, \mathbf{d}(x)^{\mathrm{T}} \, \mathbf{I}_{(1)}^{-1}(\boldsymbol{\theta}) \, \mathbf{d}(x), \tag{18}$$

where $\mathbf{I}_{(1)}^{-1} = (\mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1}$. Assume that $n \operatorname{var}\{\hat{f}_r(x)\} = O(1)$, as $n \to \infty$. Then

$$\frac{\hat{f}_r(x) - f(x)}{\sqrt{\operatorname{var}\{\hat{f}(x)\}}} \longrightarrow N(0,1) \qquad \text{in distribution, as } n \to \infty. \tag{19}$$

Here $\hat{f}_r$ is given in equation (9).

The variance of $\hat{f}_r(x)$ is then estimated as

$$\widehat{\operatorname{var}}\{\hat{f}_r(x)\} = \tilde{\sigma}^2\, \mathbf{d}(x)^{\mathrm{T}}(\mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1}\, \mathbf{d}(x)|_{\hat{\theta}_r}, \tag{20}$$

where $\hat{\theta}_r$ is obtained in problem (8) and $\tilde{\sigma}^2$ is described in equation (11).

An asymptotic pointwise $100(1-\alpha)\%$ confidence interval for $f(x)$ is

$$\hat{f}_r(x) \pm z_{\alpha/2} \sqrt{\widehat{\operatorname{var}}\{\hat{f}_r(x)\}}. \tag{21}$$

If the number of degrees of freedom $d = n - (2r+4)$ is not large, then it may be desirable to use the corresponding $t_{\alpha/2}$, the upper $\alpha/2$-quantile from a t-distribution, in place of $z_{\alpha/2}$.

Theorem 1 concerns the situation where f is a spline with a known number of knots. If f is not such a spline then more general theory can be applied to yield precise local asymptotic results. What is most relevant from our current perspective is that $\tilde{\sigma}^2$ calculated from equation (11) tends to overestimate the true value of σ^2. At the same time, the estimate $\hat{f}_r$ of equation (9) would be a somewhat biased estimate of f. When confidence intervals are formed as in expression (21) these two effects partially compensate for each other. Mao and Zhao (2003) better explains the asymptotic effect of this behaviour when combined with a simple model selection device such as that in Section 3.4 below. Our present study concentrates on fixed sample properties where, as we shall demonstrate, confidence intervals built from a foundation of equations (9) and (20) appear to perform well in a variety of situations, even when functions are not splines.

If the knot locations are fixed, then equations (15) and (17) reduce to

$$\begin{aligned} \mathbf{d}_*^{\mathrm{T}} &= (N_1(x,\mathbf{t}), \ldots, N_{r+4}(x,\mathbf{t})), \\ \mathbf{D}_* &= \begin{pmatrix} N_1(x_1,\mathbf{t}) & \ldots & N_{r+4}(x_1,\mathbf{t}) \\ \vdots & \ldots & \vdots \\ N_1(x_n,\mathbf{t}) & \ldots & N_{r+4}(x_n,\mathbf{t}) \end{pmatrix}, \end{aligned} \tag{22}$$

and equation (18) reduces to $\sigma^2 \mathbf{d}_*^{\mathrm{T}}(\mathbf{D}_*^{\mathrm{T}}\mathbf{D}_*)^{-1}\mathbf{d}_*$. It follows that

$$\mathbf{d}^{\mathrm{T}}(\mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1}\mathbf{d} > \mathbf{d}_*^{\mathrm{T}}(\mathbf{D}_*^{\mathrm{T}}\mathbf{D}_*)^{-1}\mathbf{d}_*, \tag{23}$$

since the model underlying equations (22) is more restrictive than that underlying our method. In most situations involving knot selection or variable knot locations, statements based on equations (22) should tend to undercover the true values noticeably, unless they somehow compensate by overestimating σ^2, or perhaps by including more knots than $r_{\min}$.

3.4. Optimal number of knots

The number of knots, r, is usually unknown and must be estimated. A modified criterion is used, defined as

$$\mathrm{GCV}(r) = \frac{\sum_{i=1}^{n} \{y_i - \hat{f}(x_i)\}^2}{\{n - (2r+4)\}^2/n}, \tag{24}$$

where $2r+4$ is the number of relevant parameters in the model.

We then use $r_{\min}$, which minimizes GCV(r) over a range of values of r. Because of the computational overheads for each fit, we calculate GCV(r) only for $r \leqslant r_{\max} = \min(n/3, 20)$.

In preliminary studies we investigated some other popular model selection estimates for r, such as the Akaike information criterion, the corrected Akaike information criterion and the Bayes information criterion. We found that the GCV criterion generally produced somewhat better results. Results by using the corrected Akaike information criterion (Hurvich and Tsai, 1989) were overall comparable with those by using GCV.

3.5. Algorithm

In summary, our automatic procedure can be described as follows.

(a) For $1 \leqslant r \leqslant r_{\max}$, solve the non-linear least squares problem (8). This yields estimates $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{t}}$, $\tilde{\sigma}^2$ and $\hat{f}_r(x)$ as functions of r and the given data. An efficient solution of this problem requires the use of fast robust routines such as routine DBSVLS. Care must be taken to start from several initial sets of knots to increase the chance of finding the global minimum. See Section 3.6.

(b) Calculate GCV(r), defined in equation (24). Find $r_{\min}$ to minimize this over $1 \leqslant r \leqslant r_{\max}$. Use the values of $\hat{f}_r$ corresponding to $r_{\min}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{t}}$ as the estimated function.

(c) Use the corresponding sum of squared errors to construct the estimate $\tilde{\sigma}^2$ that is defined in equation (11).

(d) Calculate $\mathbf{D}$ and $\mathbf{d}$ defined in equations (15) and (17) and consequently $\widehat{\text{var}}\{\hat{f}_r(x)\}$ in equation (20) for $r_{\min}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{t}}$. Then calculate confidence intervals for f as in expression (21).

3.6. Multiple local minima

The least squares likelihood surface for fixed r may have several distinct local minima. Consequently, different initial choices of knot locations may lead to different local minima as apparent solutions when using an algorithm such as DBSVLS.

For our purposes the problem of multiple minima is not as serious as might at first be feared. The knot locations corresponding to different apparent local least squares minima can be different. But from our experience the corresponding estimates and confidence intervals appeared qualitatively very similar apart from occasional local perturbations. This was also confirmed by simulation of coverage probabilities and the squared estimation error.

Nevertheless for occasional examples we have noticed that an unfortunate choice of initial knots may lead to drastically inappropriate local minima that would give misleading estimates and confidence sets. For this reason we recommend that a careful use of our algorithm involves repeated attempts to identify the global minimum by beginning from varied initial knot locations. One possibility is to begin with initial knot locations involving independent uniform choices for the knots. Another that we found to be more efficient and entirely satisfactory in our simulations was as follows: begin by dividing $[a, b]$ into q equal adjacent subintervals $I_1, \ldots, I_q$. Throughout the paper, all simulations were carried out by using $q = 2$, which usually sufficed. Place m_i equidistant initial knots at the interior of I_i, $i = 1, \ldots, q$, such that

$$\sum_{i=1}^{q} m_i = r, \qquad 0 \leqslant m_i \leqslant r, \quad i = 1, \ldots, q.$$

Repeat the calculation for all possible choices of $m_1, \ldots, m_q$; there are in all

$$\binom{r+q-1}{q-1}$$

such choices.

Pittman (2002) described recent research into alternative numerical methods that may alleviate the problems of local minima. As noted we have used DBSVLS only because we found it to be convenient, fast and computationally stable.

4. Simulation studies

4.1. Coverage probability

We begin with a simulation investigation of coverage probabilities. We present results for three regression functions representing a varied selection among those that we have studied. We shall return later to present other results for some of these functions.

The first function g_1 is very well behaved from the perspective of our methodology. It is a two-knot spline on $[0, 1]$ with interior knots at 0.25 and 0.8 and B-basis coefficients $\{5, 1, 3, 0, -2, -8\}$. The top of Fig. 1(a) shows a plot of this function along with a typical scatterplot for a sample of size $n = 200$ and $\sigma = 0.76$. This value of σ corresponds to a signal-to-noise ratio of 3 and thus represents moderately noisy data; the signal-to-noise level is σ_g/σ where

$$\sigma_g = \sqrt{\left[\int \{g(x) - \bar{g}\}^2 \, dx\right]}.$$

Fig. 1 also reports results with $n = 200$ and $\sigma = 0.45$, corresponding to a signal-to-noise ratio of 5.

We first take $n = 200$ design points to be equidistant on $[0, 1]$. 200 design points generated from a normal distribution with mean 0.5 and standard deviation 0.25 are also investigated. The simulation reports the results from 1000 replications. The top of Figs 1(b)–1(e) shows empirical conditional coverage probabilities ECCP for 95% and 90% confidence intervals from our procedure on x. The empirical coverage is close to the nominal levels. If $C(x_k)$ are the confidence intervals then the true conditional coverage probability at x_k is defined as

$$\text{CCP}(x_k) = P\{f(x_k) \in C(x_k)\}, \tag{25}$$

and we define the average coverage probability as

$$\text{ACP} = \frac{1}{n} \sum_{k=1}^{n} \text{CCP}(x_k). \tag{26}$$

These probabilities of course depend on n, f and σ. The empirical estimates of these quantities are denoted by ECCP and EACP.

The second function is typical of several that we looked at involving data that were moderately awkward to model. It is taken from Wand (2000) where it was used to investigate the accuracy of function estimates. The function is

$$g_2(x) = 1.5\,\varphi\left(\frac{x - 0.35}{0.15}\right) - \varphi\left(\frac{x - 0.8}{0.04}\right), \qquad 0 \leqslant x \leqslant 1,$$

where φ denotes the standard normal density.

The middle row in Fig. 1 shows this function along with typical samples having $n = 200$ and signal-to-noise ratios 5 and 3 ($\sigma = 0.054$ and $\sigma = 0.09$) (Fig. 1(a)), and ECCP plots for 95% and 90% intervals for both types of design points (Figs 1(b)–1(e)) based on 1000 simulations.

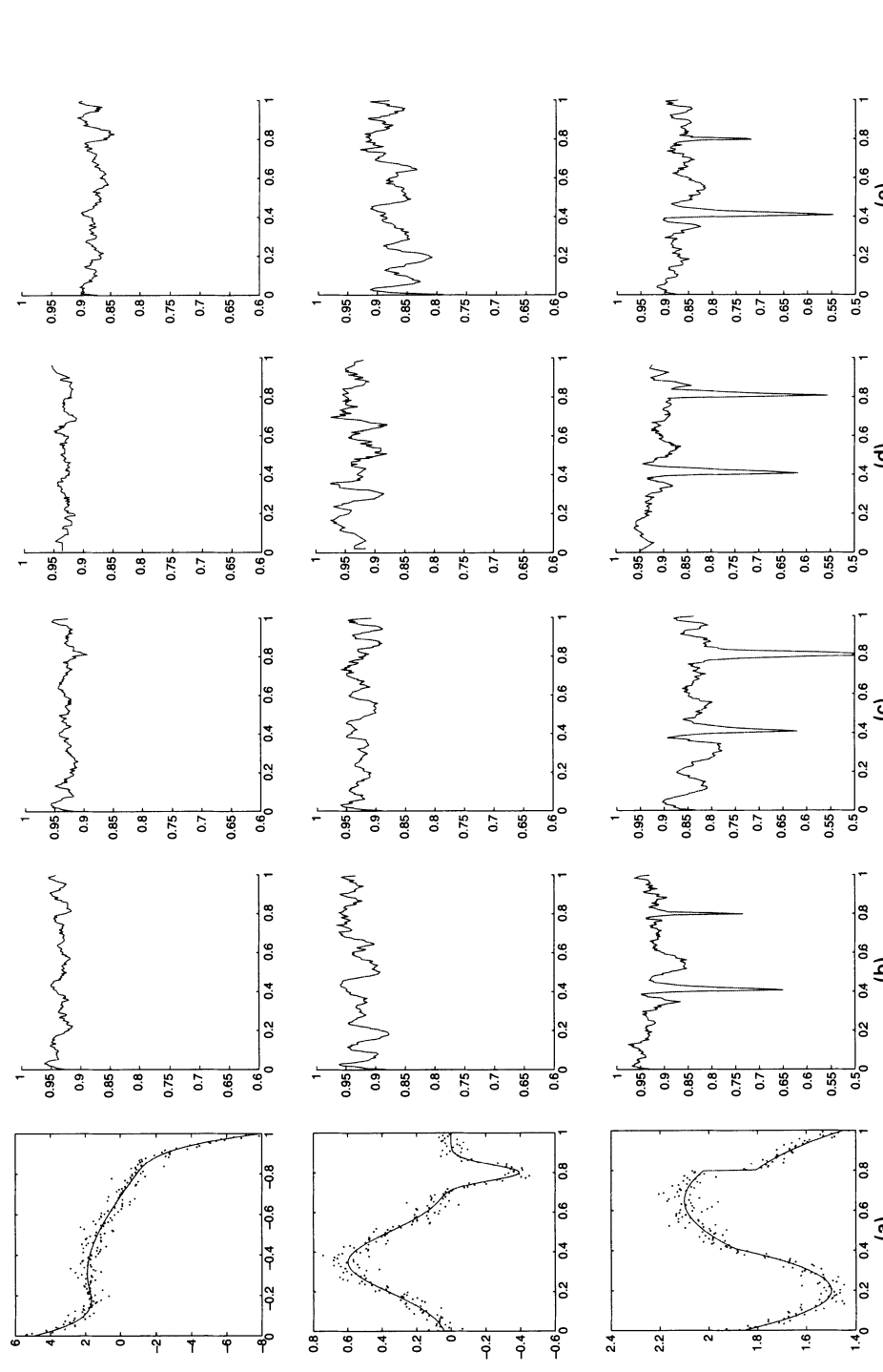

Fig. 1. ECCP plots of various functions at different nominal levels and different distributions for design plots (the top row corresponds to function g_1, the middle row to g_2 and the bottom row to g_3): (a) scatterplots when the signal-to-noise ratio is 3; (b) 95% ECCP plots at a signal-to-noise ratio of 5 when X is equidistant; (c) 95% ECCP plots at a signal-to-noise ratio of 3 when X is equidistant; (d) 95% ECCP plots at a signal-to-noise ratio of 5 when X is normally distributed as $N(1/2, 1/16)$; (e) 90% ECCP plots at a signal-to-noise ratio of 5 when X is equidistant (the sample size $n = 200$ and 1000 replications were used throughout the simulation)

Table 1. EACPs for g_1, g_2 and g_3 at nominal level 95%†

n	*EACPs for the following functions:*		
	g_1	g_2	g_3
50	0.8881 (0.8770, 0.9030)	0.8809 (0.8490, 0.9170)	0.8002 (0.7700, 0.8900)
	0.9175 (0.9110, 0.9230)	0.9161 (0.9130, 0.9320)	0.8142 (0.7600, 0.8900)
100	0.9186 (0.9140, 0.9240)	0.9148 (0.9115, 0.9280)	0.8109 (0.7850, 0.8900)
	0.9305 (0.9255, 0.9370)	0.9327 (0.9215, 0.9440)	0.9038 (0.8900, 0.9350)
200	0.9300 (0.9260, 0.9360)	0.9276 (0.9200, 0.9380)	0.8797 (0.8690, 0.9140)
	0.9348 (0.9300, 0.9410)	0.9306 (0.9175, 0.9440)	0.9139 (0.9060, 0.9370)

†Signal-to-noise ratios 3 (top figure in each cell) and 5 (bottom figure); $n = 50$, 100, 200. The numbers in parentheses are 25% and 75% quantiles of ECCP based on 1000 simulations.

The third function is an order 3 spline with seven knots and a point of discontinuity at $x = 0.8$ and another discontinuity in its derivative at $x = 0.408$; it is

$$g_3(x) = \begin{cases} 3\{3(x-0.2)^2 + 0.5\}, & 0 \leqslant x < 0.4079, \\ 3\{-1.2(x-0.65)^2 + 0.7\}, & 0.4079 \leqslant x < 0.8, \\ 3\{-1.2(x-0.65)^2 + 0.7 - 0.07\}, & 0.8 \leqslant x \leqslant 1. \end{cases}$$

The third row of Fig. 1 shows results for this function. The choice of this function emphasizes that free-knot spline methodology can be appropriate for functions that have discontinuities. Nevertheless such functions can be very difficult to fit on the basis of noisy data. This is reflected in fairly narrow downward spikes in coverage probability in the neighbourhood of the discontinuities. We know of no other standard general procedure that is designed to produce confidence bands for such a situation having possibly discontinuous noisy data. Hence we have no suitable comparison to know whether our procedure has done reasonably well or poorly for this case.

Table 1 summarizes our results by giving values of EACP for g_1, g_2 and g_3, for sample sizes 50, 100, 200 and signal-to-noise ratios 5 and 3. It turns out that the values of $\mathrm{ECCP}(x_k), k = 1, \ldots, n$, are heavily skewed to the left for g_3. To give a better idea of the empirical distribution of $\mathrm{CCP}(x_k)$, Table 1 gives the lower and upper quantiles of $\{\mathrm{ECCP}(x_k) : k = 1, \ldots, n\}$.

We have also investigated the performance of one-sided intervals constructed by the same logic, and we have found generally good behaviour similar to that reported above for two-sided intervals. As might be expected there is a mild tendency for left and right errors in one-sided coverage at given x-values to cancel, so two-sided intervals have somewhat more stable behaviour across values of x than do one-sided intervals.

4.2. Comparison with smoothing spline confidence intervals

Smoothing splines provide important standard methodology for nonparametric regression confidence intervals. Wahba (1983) and Nychka (1988) showed that smoothing splines are Bayes estimators corresponding to a particular Gaussian prior and

$$\hat{\mathbf{f}} = \mathbf{A}_{\hat{\lambda}}\mathbf{Y},$$

$$\mathrm{var}(\hat{\mathbf{f}}|Y) = \sigma^2 \mathbf{A}_{\hat{\lambda}},$$

where $\mathbf{A}_{\hat{\lambda}}\mathbf{Y}$ is the smoothing spline estimator evaluated at $(x_1, \ldots, x_n)^{\mathrm{T}}$ and $\hat{\lambda}$ is the smoothing parameter chosen by minimizing the GCV estimator. Correspondingly, they proposed an

approximate $100(1-\alpha)\%$ confidence interval of the form

$$\hat{f}(x_i) \pm z_{\alpha/2}\hat{\sigma}\sqrt{\mathbf{A}_{ii}},$$

where σ^2 is estimated by $\hat{\sigma}^2 = \mathrm{SSE}/\{n - \mathrm{tr}(\mathbf{A}_{\hat{\lambda}})\}$.

We use Wahba's setting by taking her three smooth functions with one, two and three humps. They are

$$f_1(t) = \frac{1}{3}\beta_{10,5}(t) + \frac{1}{3}\beta_{7,7}(t) + \frac{1}{3}\beta_{5,10}(t),$$

$$f_2(t) = \frac{6}{10}\beta_{30,17}(t) + \frac{4}{10}\beta_{3,11}(t),$$

$$f_3(t) = \frac{1}{3}\beta_{20,5}(t) + \frac{1}{3}\beta_{12,12}(t) + \frac{1}{3}\beta_{7,30}(t),$$

where

$$\beta_{p,q}(t) = \frac{\Gamma(p+q)}{\Gamma(p)\,\Gamma(q)}t^{p-1}(1-t)^{q-1}, \qquad 0 \leqslant t \leqslant 1,$$

is the β density function.

The five noise levels σ are 0.0125, 0.025, 0.05, 0.1 and 0.2 as in Wahba (1990). The three sample sizes n are 32, 64 and 128. The signal-to-noise values corresponding to $\sigma = 0.1$ for the three functions are 6.88, 9.6 and 5.4. Values of $\sigma \leqslant 0.5$ correspond to a larger signal-to-noise ratio. We feel that such values are of less interest for statistical applications, especially when $n = 64$ and $n = 128$, but we have nevertheless reported results for them because they are included in Wahba's study. Some simulation results are reported in Fig. 2.

Table 2 reports values of EACP for our method and for Wahba's method; it is based on 100 replications at each level. Wahba ran simulations involving only 10 replicates. To obtain suitable accuracy we ran simulations for her examples again to produce Bayesian smoothing spline confidence intervals. For this we used the software FUNFITS provided by Nychka *et al.* (1996). Our method appears to produce values of EACP that are acceptably close to the nominal level of 95%. All except five of the 45 values for our method exceed 90%. The two lowest values for our method (86.8% and 86.39%) differ somewhat from the overall pattern and could possibly be underestimates of the true value attributable to random variation. By contrast 20 of the 45 results for FUNFITS fall below 90%. For the largest sample size here, $n = 128$, both methods appear to have acceptable average coverage probabilities at the noise levels that are reported here, as chosen by Wahba.

4.3. Comparison of mean-squared errors with other polynomial spline procedures

Along with its confidence bands our procedure of course also produces estimates of the regression function. A wide range of existing methods produces such estimates. In this section we compare the estimates from our procedure with those from two other popular related methods: the adaptive knot selection procedure POLYMARS that was developed by Stone *et al.* (1997) and the variable knots Bayesian spline procedure `br` that was developed by Smith and Kohn (1996). It should be noted that POLYMARS is piecewise linear and it was developed to apply also in higher dimensional problems. Thus it might not be expected to be competitive as an estimator in our situation.

The average root-mean-square error RMSE will be used to judge accuracy. It is defined as

$$\mathrm{RMSE} = \sqrt{\left[\frac{1}{n}\sum_{i=1}^{n}\{\hat{f}(x_i) - f(x_i)\}^2\right]}.$$

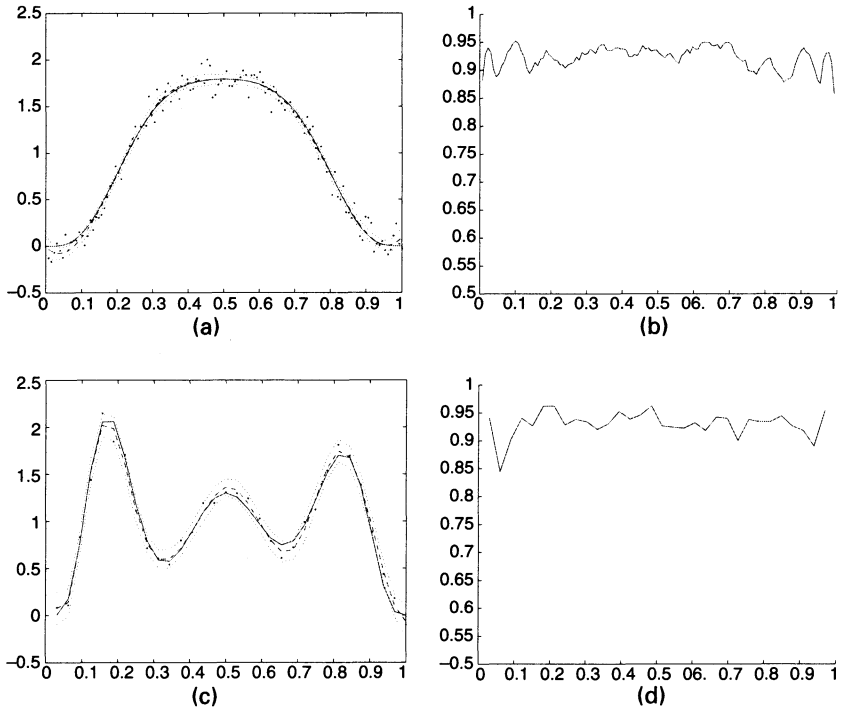

Fig. 2. (a) Typical sample under f_1 when $n = 128$ and $\sigma = 0.1$ (·····, 95% confidence band; -------, estimate; ———, true function); (b) empirical coverage probability as a function of x, under f_1, when $n = 128$ and $\sigma = 0.1$; (c) similar to (a) but with testing function f_3 when $n = 32$ and $\sigma = 0.1$; (d) similar to (b) but with testing function f_3 when $n = 32$ and $\sigma = 0.1$

We give results for the three functions defined in Section 4.1 with the same simulation set-up. The box plots in Fig. 3 summarize our results. These are box plots of the values of $\log_{10}(\text{RMSE})$ for 1000 Monte Carlo replications. It appears that `bs` and our method are generally competitive as estimation procedures, and both improve on POLYMARS. The only major difference in performance appears in Fig. 3(a) for the function that is most difficult to fit, g_3, at a signal-to-noise ratio of 5, where the free-knot method is better.

5. Discussion

This section investigates two aspects of the free-knot methodology as we have applied it to a statistical setting. First we examine the practical effect of the two steps of our method that are only justified by asymptotic criteria. Second we discuss the use of our procedure for other objectives.

5.1. Non-linearity and model selection

Part of the justification for our methodology is its ability to provide suitable estimates and confidence intervals when the true regression function is a polynomial spline. In this section we examine in detail the performance of our procedure when the true regression is the two-knot spline g_1 of Section 4.1.

If the knot locations of g_1 were known then the problem would involve an ordinary Gaussian linear model. The accuracy of estimation would be optimal in several accepted senses and the

Table 2. EACPs of our method (OURS) and the smoothing spline confidence intervals (FUNFITS) for testing functions f_1–f_3†

σ	Case	*EACPs from the following methods and values of n:*					
		$n = 32$		$n = 64$		$n = 128$	
		OURS	*FUNFITS*	*OURS*	*FUNFITS*	*OURS*	*FUNFITS*
0.0125	1	90.84	85.94	94.98	90.69	93.16	92.94
	2	86.87	82.59	93.96	90.31	91.96	92.16
	3	94.21	53.06	94.56	88.52	94.20	93.05
0.025	1	91.50	88.31	88.92	91.14	93.79	92.61
	2	90.56	79.41	86.39	88.70	94.18	92.65
	3	95.34	57.25	91.59	88.80	94.05	92.29
0.05	1	95.93	86.59	93.04	91.08	92.82	93.02
	2	91.46	82.19	93.68	90.56	94.72	92.72
	3	95.40	68.91	91.42	89.98	92.01	92.88
0.1	1	95.28	85.31	94.34	91.08	94.96	92.09
	2	94.12	86.16	94.51	90.59	91.02	91.15
	3	95.25	78.63	95.32	91.31	89.96	92.30
0.2	1	92.62	84.25	89.67	88.02	94.30	92.02
	2	95.21	84.81	90.67	90.72	92.71	92.73
	3	92.59	84.09	93.51	90.53	94.18	91.95

†The nominal level is 95%.

confidence coverage would be exact. The expected root-mean-square error would agree exactly with the theoretical value

$$\mathrm{RMSE}_1 = \left\{ \frac{\sigma^2}{n} \sum_{i=1}^{n} \mathbf{d}_*^{\mathrm{T}}(x_i)(\mathbf{D}_*^{\mathrm{T}}\mathbf{D}_*)^{-1}\mathbf{d}_*(x_i) \right\}^{1/2} \tag{27}$$

that is obtained from the right-hand side of inequality (23).

If the function were assumed to be a two-knot spline then it could be fitted by the non-linear least squares procedure in problem (8) with r fixed at $r = 2$. The asymptotic average root-mean-square error is then given by the left-hand side of inequality (23) as

$$\mathrm{RMSE}_2 = \left\{ \frac{\sigma^2}{n} \sum_{i=1}^{n} \mathbf{d}^{\mathrm{T}}(x_i)(\mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1}\mathbf{d}(x_i) \right\}^{1/2}. \tag{28}$$

This value need not be attained in practice since the theory leading to equation (28) is only asymptotic. For the same reason, the expected average coverage of confidence intervals constructed in this way need not achieve the nominal value 95%.

Finally, we are mainly interested in the practical situation where r is unknown, and the modelled value of r is chosen via GCV. In this case the estimation and confidence interval performance can be adversely affected by an incorrect choice of r as well as by the various stochastic errors that are discussed above.

Table 3 gives values of equations (27) and (28) and various empirical simulation results including average coverage probabilities as well as average confidence interval widths based on 500 simulations at each level. Table 3 includes results for $n = 50$ and $n = 200$ and for signal-to-noise

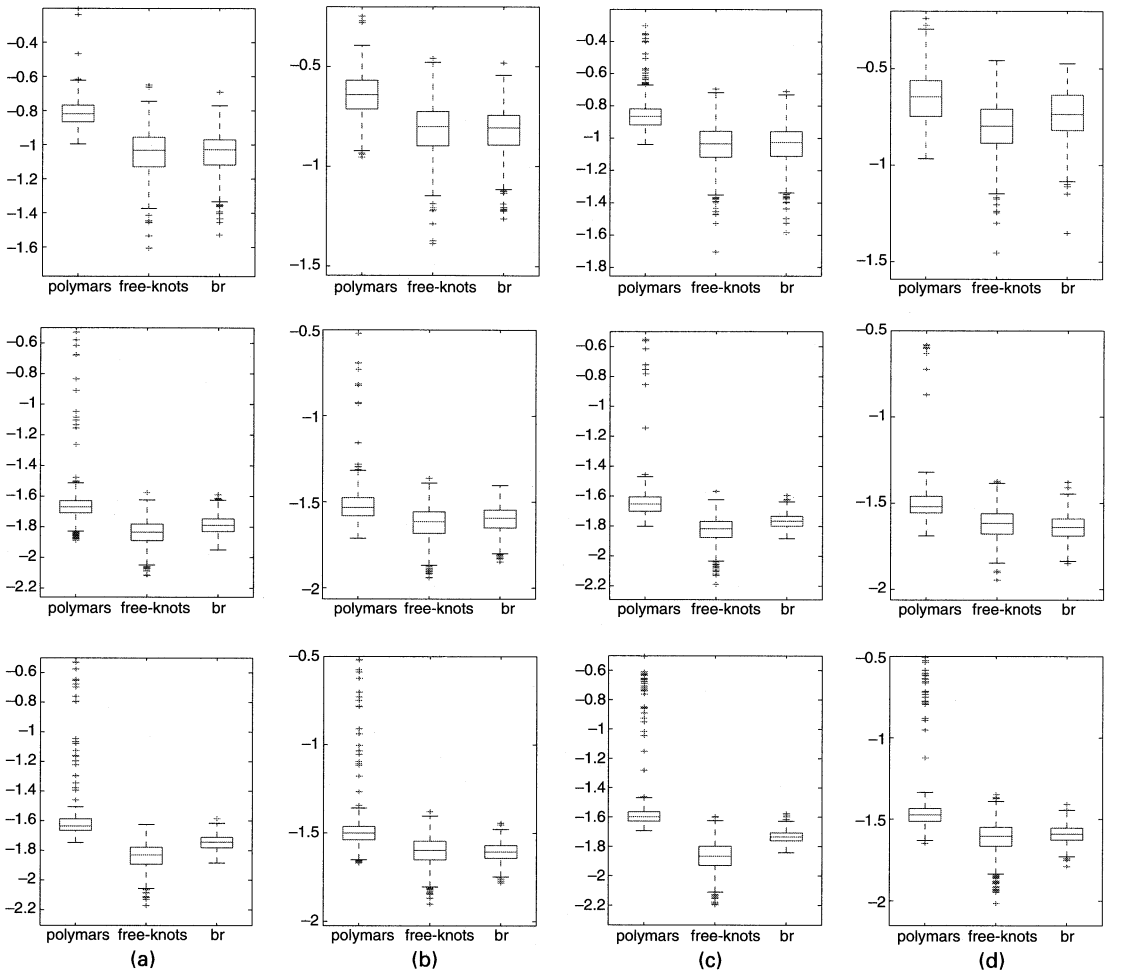

Fig. 3. Box plots of $\log_{10}$(RMSE) for functions $g_1(x)$ (top row), $g_2(x)$ (middle row) and $g_3(x)$ (bottom row): (a) equidistant design points, signal-to-noise ratio 5; (b) equidistant design points, signal-to-noise ratio 3; (c) normally distributed design points, signal-to-noise ratio 5; (d) normally distributed design points, signal-to-noise ratio 3

Table 3. Comparison of performance under three scenarios for g_1†

Parameter	*Results for the following values of n and signal-to-noise ratios:*					
	n = 50			*n = 200*		
	1	*3*	*5*	*1*	*3*	*5*
$RMSE_1$	0.6789	0.0754	0.0272	0.3926	0.1309	0.0785
$ERMSE_1$	0.6829	0.0759	0.0273	0.3815	0.1272	0.0763
$RMSE_2$	0.8717	0.1000	0.0362	0.4523	0.1511	0.0907
$ERMSE_2$	0.8906	0.1107	0.0398	0.4535	0.1519	0.0889
$ERMSE_3$	1.0222	0.1379	0.0466	0.5320	0.1651	0.0971
$EACP_1$	0.9421	0.9421	0.9421	0.9462	0.9462	0.9462
$EACP_2$	0.9377	0.9201	0.9284	0.9299	0.9341	0.9434
$EACP_3$	0.9050	0.8870	0.9133	0.8811	0.9257	0.9316
$EAWidth_1$	1.5193	0.5064	0.3039	0.7294	0.2431	0.1459
$EAWidth_2$	1.7404	0.5812	0.3507	0.8411	0.2819	0.1692
$EAWidth_3$	1.5614	0.5799	0.3565	0.7961	0.2903	0.1745

†Scenario 1, the two knot locations are given; scenario 2, only the number of knots is given; scenario 3, there is no assumption on the number of knots. See the text for complete descriptions.

levels 1, 3 and 5. Entries with subscript 1 refer to fitting with the correct knot locations, those with subscript 2 refer to fitting with two knots at free locations and those with subscript 3 refer to our scheme with GCV for the choice of knots. Entries beginning with 'E' are empirical simulation results; the others are theoretical, as described above.

Fig. 4 shows the histogram of the number of knots chosen by our GCV criterion in these simulations. At higher signal-to-noise values and larger sample sizes the GCV method virtually never underfits by choosing too small a number of knots. It sometimes mildly overfits, but such mild overfitting does not have serious negative consequences for the various performance criteria.

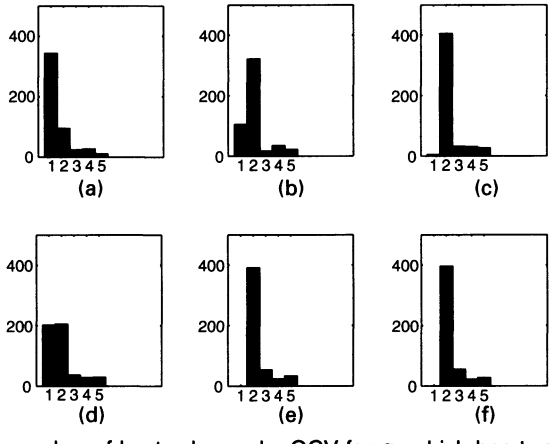

Fig. 4. Histogram of the number of knots chosen by GCV for g_1 which has two knots: (a) $n = 50$, signal-to-noise ratio 1; (b) $n = 50$, signal-to-noise ratio 3; (c) $n = 50$, signal-to-noise ratio 5; (d) $n = 200$, signal-to-noise ratio 1; (e) $n = 200$, signal-to-noise ratio 3; (f) $n = 200$, signal-to-noise ratio 5

The values of ERMSE_2 are close to their theoretical values, RMSE_2. Hence the asymptotic values are fairly close to the actual values. Next, ERMSE is somewhat larger than RMSE_2. This describes the estimation penalty for not knowing how many knots g_1 has.

The three values of EACP decrease somewhat, but not too drastically, as we progress from the precise, correct model to our free-knot model with r to be chosen by GCV. The values of EACP_1 are constant at the three noise levels because the same set of simulated values of ε_i were used for the given n at all three noise levels. The theoretical value of EACP_1 is 0.95, and the observed deviation is attributable to the random simulation effect.

Finally, EAWidth_1 is generally smaller than EAWidth_2 as we should expect. However, the values of EAWidth_2 and of EAWidth are comparable in spite of the fact that the free-knot model is less precise than the two-knot model of EAWidth_2. This juxtaposition suggests that the free-knot confidence intervals may be somewhat too narrow and that better fidelity to nominal coverage values would be obtained by increasing their width somewhat. Such an increase could be motivated by taking into account that the free-knot method involves 'estimation' of the true value of r as well as of the $2r + 4$ co-ordinates of $\boldsymbol{\theta}$. But our methodology does not make an upward adjustment in the length of interval because of estimation of r. Although we could do so in an *ad hoc* fashion we do not know of a statistical principle that would prescribe the magnitude of such an adjustment.

5.2. *Other objectives*

As noted, our primary objective is to produce regression estimates accompanied by two-sided confidence intervals for $f(x)$. These confidence intervals $\text{CI}(x)$ have as a goal the nominal property

$$P\{f(x) \in \text{CI}(x)|x\} \geqslant 1 - \alpha, \tag{29}$$

and consequently

$$E\left[\frac{1}{n}\sum_{i=1}^{n} I_{\text{CI}(x)}\{f(x_i)\}\right] \geqslant 1 - \alpha. \tag{30}$$

Of course, our algorithm is not exact, and so the degree to which inequalities (29) and (30) hold in particular examples needs to be investigated numerically. Section 4 reports some typical investigations. Generally in our examples involving signal-to-noise ratios between 3 and 5 and sample sizes 50–200 we found (as expected) noticeable variability in inequality (29) as a function of x, especially for inhomogeneous f and higher noise levels, but there was only a mild tendency for undercoverage on average with values of inequality (30) for nominal $1 - \alpha = 0.95$ ranging from the mid-80% range to nearly 0.95, depending on the example. For signal-to-noise ratios of 1 or less we found noticeable degradation in the coverage performance of our intervals, as well as of the few existing alternative methods that we have tried.

We have concentrated only on confidence interval criterion (25) because we feel that this is the one that is most often useful in practice. However, our algorithm can easily be adapted to other confidence objectives. We can, for example, produce bands with nominal simultaneous coverage of $1 - \alpha$, i.e. with the goal

$$P\{f(x) \in \text{CI}(x)\} \geqslant 1 - \alpha, \qquad \text{for every } x. \tag{31}$$

For this purpose we could replace the value $z_{1-\alpha/2}$ in expression (21) by $\{(2r + 4)F_{1-\alpha}\}^{1/2}$ where $F_{1-\alpha}$ denotes the upper α cut-off point of an F-distribution with $2r + 4$ and $n - (2r + 4)$ degrees of freedom. This simultaneous confidence band would nominally be asymptotically

conservative. We might hope to reduce this conservativeness by using, for example, methods of Johansen and Johnstone (1990), but we have so far been unable to implement these methods in the current non-linear setting.

We could alternatively desire prediction intervals of the usual sort instead of confidence intervals for $f(x)$. For this we one would replace $\sqrt{\widehat{\text{var}}\{\hat{f}(x)\}}$ in expression (21) by $\sqrt{[\tilde{\sigma}^2 + \widehat{\text{var}}\{\hat{f}(x)\}]}$.

There is heuristic reason to believe that the performance of our methods for these objectives would be even better than that for our primary confidence interval objectives (1) and (2). This will be reported elsewhere.

Acknowledgements

The authors thank Lawrence Brown for his continuous support and numerous useful suggestions. Charles Stone's encouragement also played an important role. We also thank the Joint Editor and the referees for their helpful comments.

References

Bates, D. M. and Watts, D. G. (1988) *Nonlinear Regression Analysis and Its Applications*. New York: Wiley.
de Boor, C. (1978) *A Practical Guide to Splines*. New York: Springer.
de Boor, C. (1998) *Spline Toolbox for Use with Matlab User's Guide*. Natick: Mathworks.
de Boor, C. and Rice, J. R. (1968) Least squares cubic spline approximation II—variable knots. *Technical Report 21*. Computer Science Department, Purdue University, West Lafayette. (Available from `http://www.cs.wisc.edu/~deboor/`.)
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.
Dette, H., Munk, A. and Wagner, T. (1998) Estimating the variance in nonparametric regression—what is a reasonable choice? *J. R. Statist. Soc.* B, **60**, 751–764.
Dierckx, P. (1993) *Curve and Surface Fitting with Splines*. New York: Oxford Science.
DiMatteo, I., Genovese, C. R. and Kass, R. E. (2001) Bayesian curve-fitting with free-knot splines. *Biometrika*, **88**, 1055–1071.
Eubank, R. L. (1988) *Spline Smoothing and Nonparametric Regression*. New York: Dekker.
Eubank, R. L. and Speckman, P. L. (1993) Confidence bands in nonparametric regression. *J. Am. Statist. Ass.*, **88**, 1287–1301.
Fan, J. and Gijbels, I. (1996) *Local Polynomial Modeling and Its Applications*. New York: Chapman and Hall.
Friedman, J. H. (1991) Multivariate adaptive regression splines (with discussion). *Ann. Statist.*, **19**, 1–141.
Friedman, J. H. and Silverman, B. W. (1989) Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics*, **31**, 3–39.
Greville, T. N. E. (1969) Introduction to spline functions. In *Theory and Applications of Spline Functions*, pp. 1–35. New York: Academic Press.
Hall, P., Kay, J. W. and Titterington, D. M. (1990) Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, **77**, 521–528.
Härdle, W. and Marron, J. S. (1991) Bootstrap simultaneous error bars for nonparametric regression. *Ann. Statist.*, **19**, 778–796.
Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*. New York: Chapman and Hall.
Hurvich, C. M. and Tsai, C. L. (1989) Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.
Johansen, S. and Johnstone, I. (1990) Hotelling's theorem on the volume of tubes: some illustrations in simultaneous inference and data analysis. *Ann. Statist.*, **18**, 652–684.
Jupp, D. L. B. (1978) Approximation to data by splines with free knots. *SIAM J. Numer. Anal.*, **15**, 328–343.
Kooperberg, C. and Stone, C. (2002) Confidence intervals for logspline density estimation. In *Proc. Mathematical Sciences Research Institute Wrkshp Nonlinear Estimation and Classification*, pp. 285–296. New York: Springer.
Lehmann, E. L. (1999) *Elements of Large Sample Theory*. New York: Springer.
Lindstrom, M. J. (1999) Penalized estimation of free-knot splines. *J. Comput. Graph. Statist.*, **8**, 333–352.
Loader, C. (1999) *Local Regression and Likelihood*. New York: Springer.
Mao, W. (2000) Free knot polynomial spline confidence intervals. *PhD Thesis*. University of Pennsylvania, Philadelphia.
Mao, W. and Zhao, L. H. (2003) Coverage probability. *Technical Report*. University of Pennsylvania, Philadelphia.
Nadaraya, E. A. (1964) On estimating regression. *Theory Probab. Applic.*, **9**, 141–142.

Nychka, D. (1988) Bayesian confidence intervals for smoothing splines. *J. Am. Statist. Ass.*, **83**, 1134–1143.
Nychka, D., Bailey, B., Ellner, S., Haaland, P. and O'Connell, M. (1996) FUNFITS: data analysis and statistical tools for estimating functions. *Mimeo Ser. 2289*. North Carolina Institute of Statistics, Rayleigh. (Available from `http://www.cgd.ucar.edu/stats/Software/Funfits`.)
Picard, D. and Tribouley, K. (2000) Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.*, **28**, 298–335.
Pittman, J. (2002) Adaptive splines and genetic algorithms. *J. Comput. Graph. Statist.*, **11**, 1–24.
Rice, J. (1984) Bandwidth choice for nonparametric regression. *Ann. Statist.*, **12**, 1215–1230.
Schumaker, L. (1981) *Spline Functions Basic Theory*. New York: Wiley.
Smith, M. and Kohn, R. (1996) Nonparametric regression using Bayesian variable selection. *J. Econometr.*, **75**, 317–344.
Stone, C. J., Hansen, M. H., Kooperberg, C. and Truong, Y. K. (1997) Polynomial splines and their tensor products in extended linear modeling (with discussion). *Ann. Statist.*, **25**, 1371–1470.
Wahba, G. (1983) Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. R. Statist. Soc.* B, **45**, 133–150.
Wahba, G. (1990) *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.
Wand, M. P. (2000) A comparison of regression spline smoothing procedures. *Comput. Statist.*, **15**, 443–462.
Xia, Y. (1998) Bias-corrected confidence bands in nonparametric regression. *J. R. Statist. Soc.* B, **60**, 797–811.
Zhou, S. and Shen, X. (2001) Spatially adaptive regression splines and accurate knot selection schemes. *J. Am. Statist. Ass.*, **96**, 247–259.
Zhou, S., Shen, X. and Wolfe, D. A. (1998) Local asymptotics for regression splines and confidence regions. *Ann. Statist.*, **26**, 1760–1782.