Free Knot Polynomial Spline Confidence Intervals

Vincent W. Mao
Linda H. Zhao
Department of Statistics
University of Pennsylvania
Philadelphia, PA 19104-6302
lzhao@wharton.upenn.edu

Abstract

We construct approximate confidence intervals for a nonparametric regression function. The construction uses polynomial splines with free knot locations. The number of knots is determined by the GCV criteria. The estimates of knot locations and coefficients are obtained through a nonlinear least square solution that corresponds to the maximum likelihood estimate. Confidence intervals are then constructed based on the asymptotic distribution of the MLE. Average coverage probabilities and accuracy of the estimate are examined via simulation. This includes comparisons between our method and some existing ones 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 seems to work well for smooth underlying functions and also reasonably well for unsmooth (discontinuous) functions. It also performs well for fairly small sample sizes. As a practical example we apply the method to study the productivity of US banks. The corresponding analysis supports certain research hypotheses concerning the effect of federal policy on banking efficiency.

Key words: Nonparametric regression; Confidence intervals; MLE; Piecewise polynomials; Free knots; B-splines.

1 Introduction

The nonparametric regression model

$$y_i = f(x_i) + \sigma\,\epsilon_i,\ \ 1 \leqslant i \leqslant n,\ \ \epsilon_i \overset{iid}{\sim} N(0,1),\ \ \sigma^2 \text{ unknown} \tag{1}$$

has been studied extensively in the literature. We are interested in constructing estimates that are accompanied by confidence intervals for the underlying function values, $f(x)$.

There exist a number of procedures to estimate f. Kernel type methods include kernel regression (Nadaraya (1964)) and local polynomial fitting (Fan and Gijbels (1996) and 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 (Tribouley (2000)).

Spline models provide another popular method for estimating f. Wahba (1983) discusses 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. In these $f(x)$ is estimated by a piecewise m^{th} order ($(m-1)^{th}$ degree) polynomial connecting smoothly at points $t_1 < t_2, \ldots < t_r$, which are referred as interior knots. It is important to appropriately choose the number of knots r and the knot locations. In the current variable knots 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 approach the choice of knots may involve a linear regression model selection scheme or a Bayesian model. The estimation of regression coefficients given the knots is via linear least squares. Some papers using this approach are Friedman and Silverman (1989), Friedman (1991), and Stone, Hansen, Kooperberg and Truong (1997). Some more recent, effective variations are in Smith and Kohn (1996), Denison, Smith and Mallick (1998), Lindstrom (1999). All the work cited above is about the estimation of f. Zhou, Shen and Wolfe (1998) provide confidence intervals along this line.

Following appearance of a preprint version of our manuscript Kooperberg and Stone (2001) used a closely related free knot construction of confidence intervals for nonparametric density estimates.

We use free knots polynomials, i.e., the knot locations are considered to be unknown

parameters as well as the regression coefficients. Doing so provides flexibility to allow $f(x)$ with inhomogeneous smoothness 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 from among $r = 1, 2, \ldots$. It is not used to choose knot locations among a large set of possible locations, as is done in existing variable knots schemes cited above. Partly because of this minimal use of model selection one can anticipate that the confidence intervals we construct will have probabilities of coverage 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. One 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. $f(\boldsymbol{\theta}, x)$ may have a lower degree of smoothness at locations where a multiplicity of knots occurs. For convenience we fix $m = 4$ throughout our treatment here.

Our motivation is to then view (1) as if it were a parametric nonlinear regression model:

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

The estimation part of the statistical analysis involves first fixing r and estimating $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_r$ by maximum likelihood within the model (2). Then an estimated best value of r, called r_{min} is chosen through a 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 asymptotically valid confidence intervals for $f(x)$. To understand the primary methodology note that when f is itself a polynomial spline with r knots we can write

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

where $\mathbf{I}_{(1)}$ denotes the appropriate information matrix given in (15) and (19).

Let $\mathbf{d}^T = \dfrac{\partial f}{\partial \boldsymbol{\theta}} = \left(\dfrac{\partial f}{\partial \theta_1}, \ldots, \dfrac{\partial f}{\partial \theta_s}\right)$. (Note that $\mathbf{d}^T$ depends on x as well as $\boldsymbol{\theta}$.) 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 $\sigma^2 \mathbf{d}^T \mathbf{I}_{(1)}^{-1}(\boldsymbol{\theta})\, \mathbf{d}$ since

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

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

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

One can produce simultaneous band for $f(x)$ based on (3) but replacing $z_{\alpha/2}$ by $((2r+4)F_{1-\alpha})^{1/2}$ where $F_{1-\alpha}$ is the upper α cut off point of an F distribution. One would expect these to be conservative. We are investigating along this line in a separate paper.

The estimation idea following (2) seems to be very natural. Indeed, it has been mentioned frequently in the literature. An early reference for splines is de Boor and Rice (1968). 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 of investigation, 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.

So far as we know confidence intervals like those supplied by (3), although statistically natural, have not previously been investigated in the setting of free knot splines. Section 4 reports simulation evidence relating to our confidence set objective. This suggests that these confidence intervals perform well in terms of coverage and accuracy.

The method here is locally adaptive to variable smoothness in f because the procedure will automatically place 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 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 introduce some background knowledge about B-splines. In Section 3 we give details of our method. In Sections 4 and 5 we apply this method to simulated data and to real data. The real data application in Section 5 involves a study of banking efficiency first reported in Faulhaber (2000). Section 6 has discussions about the free-knot methodology and reports explanatory empirical results that tend to

support this as a confidence set methodology. It also describes some alternative types of confidence bands. We conclude in Section 7 with a brief summary.

2 B-splines

An m^{th}-order polynomial spline on $[a, b]$ with r ordered interior knots $\boldsymbol{t} = (t_1, ..., t_r)$ 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, $S_{m,r,\boldsymbol{t}}$, is a linear space of dimension $m + r$. In particular when $m = 4$, $S_{4,r,\boldsymbol{t}}$ consists of all cubic splines. Throughout we will use only the value $m = 4$. This produces plots that are visually smooth. Our methodology can be easily applied with other values of m. The commonly used bases for $S_{4,r,\boldsymbol{t}}$ 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. (The natural spline version of this basis could be used instead (Eubank (1988) and Greville (1969)).)

The B-splines for $m = 4$ are completely determined by the interior knots $\boldsymbol{t}$. Let $t_{-3} = ...t_0 = a < t_1 \leqslant ... \leqslant t_r < b = t_{r+1} = ... = t_{r+4}$. The B-spline $N_i(x, \boldsymbol{t})$ is defined to be

$$N_i(x, \boldsymbol{t}) \triangleq (t_i - t_{i-4})[t_{i-4}, \ldots, t_i](\cdot - x)^3_+ \tag{4}$$

$$\triangleq [t_{i-3} \ldots, t_i](\cdot - x)^3_+ - [t_{i-4}, \ldots, t_{i-1}](\cdot - x)^3_+, \quad i = 1, \ldots, r+4. \tag{5}$$

Here "[ ]" denotes divided difference,

$$[t_i, t_j]g(\cdot) \triangleq (g(t_j) - g(t_i))/(t_j - t_i),$$

$$[t_1, \ldots, t_n]g(\cdot) \triangleq g^{(n-1)}(t) \text{ if } t_1 = \ldots = t_n = t,$$

$$[t_1, \ldots, t_n]g(\cdot) \triangleq \frac{[t_2, \ldots, t_n]g(\cdot) - [t_1, \ldots, t_{n-1}]g(\cdot)}{t_n - t_1}.$$

Remarks:

1. From (5), we can see that there is a recursive relationship that can be used to describe B-splines. This relationship provides a very stable numerical computation algorithm.

2. One useful property of B-splines is that they 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.

3. For a function 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 as follows:

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

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 $= 4 - 3 = 1$, which means that $f(x)$ is continuous at $t = t_k$ but $f'(x)$ is discontinuous at $t = t_k$.

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

Lemma 2.1 *When $i \leqslant j \leqslant i+4$, we have*

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

$\frac{\partial N_i(x,\boldsymbol{t})}{\partial t_j} = 0$ *otherwise.*

3 Methodology

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

$$y_i = \sum_{j=1}^{r+4} \beta_j \, N_j(x_i, \boldsymbol{t}) + \sigma\epsilon_i, \quad i = 1, \ldots, n. \tag{7}$$

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

3.1 Estimation of f

We will use the MLE $\hat{\boldsymbol{\theta}}$, to estimate $\boldsymbol{\theta}$. Because of the normal errors in the model (7) it is easy to see that $\hat{\boldsymbol{\theta}}$ solves the following nonlinear least square problem:

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

Immediately we have an estimate of $f(x)$:

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

The basic idea of solving (8) is the following: Given $\boldsymbol{t}$, let

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

The linear least squares solution of $\boldsymbol{\beta}$ is produced, i.e. $G(\boldsymbol{t}) = \min_{\boldsymbol{\beta}} F(\boldsymbol{\beta}, \boldsymbol{t})$. Then we search for the minimum of $G(\boldsymbol{t})$.

This nonlinear optimization problem needs to be treated carefully. Given a starting value $\boldsymbol{t}^*$, a local optima 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 fool-proof in free-knot spline regression. There are too many saddle points and minima on the least square surface. For certain examples the chance of finding the global minimum based on 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 the local minima that are distinct from the global minimum.

Several programs are available to calculate $\min G(\boldsymbol{t})$ beginning from an initial choice of knots. We use the IMSL routine DBSVLS. We have found this algorithm very fast and stable. The computational speed of this routine 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 eliminate falsely identifying local minima as global ones. We also note that the statistical performance of our procedure is not overly sensitive to the final local minima found. We discuss this issue more fully in Sections 3.6 and 6.

3.2 Estimating σ^2

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

$$\tilde{\sigma}^2 = \text{SSE}/(n-(2r+4)) \tag{11}$$

since $2r+4$ is the number of relevant free parameters in our model.

This estimator is approximately unbiased and works well in our simulations. It agrees with the general suggestion for non-linear least squares models in many standard 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. 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 indicate that this over estimates σ^2 in our setting, as one might expect. 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 MLEs are then used to assess the variability of $\hat{f}(x)$. The relevant formulas are summarized below in order to concretely describe our procedure.

To proceed, let us write the model (7) in matrix form:

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

here $\mathbf{Y} = (y_1, ..., y_n)^T$, $\mathbf{X} = (x_1, ..., x_n)^T$,

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

and $\epsilon = (\epsilon_1, ..., \epsilon_n)^T$, $\epsilon \sim N(0, I)$.

Let

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

The following lemma gives the information matrix for $(\boldsymbol{\theta}, \sigma)$.

Lemma 3.1

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

Proof: The log likelihood function is

$$l = -\frac{(\mathbf{Y}-\mathbf{f})^T(\mathbf{Y}-\mathbf{f})}{2\sigma^2} - n\,\log\sigma + \text{const.}$$

Taking the derivative, we have

$$\frac{\partial l}{\partial \boldsymbol{\theta}} = \frac{\mathbf{D}^T(\mathbf{Y}-\mathbf{f})}{\sigma^2}, \quad \frac{\partial l^2}{\partial \boldsymbol{\theta}^2} = -\frac{\mathbf{D}^T\mathbf{D}}{\sigma^2}$$

$$\frac{\partial l^2}{\partial \sigma^2} = \frac{-3(\mathbf{Y}-\mathbf{f})^T(\mathbf{Y}-\mathbf{f})}{\sigma^4} + \frac{2n}{\sigma^2}, \quad \frac{\partial l^2}{\partial \boldsymbol{\theta}\partial\sigma} = -\frac{2\mathbf{D}^T(\mathbf{Y}-\mathbf{f})}{\sigma^3}.$$

These lead to (16). □

Let

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

Standard results on asymptotic normality of MLEs, see e.g., Lehmann (1999, Theorems 7.5.1 and 5.4.6), yield

Theorem 3.1

$$\frac{\hat{f}(x) - f(x)}{\sqrt{Var\,(\hat{f}(x))}} \Rightarrow N(0,1). \tag{18}$$

Here, in the limit as $n \to \infty$

$$Var\,(\hat{f}(x)) \sim \sigma^2\mathbf{d}(x)^T(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{d}(x) \triangleq \sigma^2\mathbf{d}(x)^T\mathbf{I}_{(1)}^{-1}(\boldsymbol{\theta})\mathbf{d}(x), \tag{19}$$

here $\mathbf{I}_{(1)}^{-1} = (\mathbf{D}^T\mathbf{D})^{-1}$.

The variance of $\hat{f}(x)$ is then estimated by a plug in method as

$$\hat{\text{Var}}\ (\hat{f}(x)) = \tilde{\sigma}^2\, \mathbf{d}(x)^T(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{d}(x)|_{\hat{\boldsymbol{\theta}}}, \tag{20}$$

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

The following asymptotic pointwise $100(1-\alpha)\%$ confidence interval for $f(x)$ is then obtained :

$$\hat{f}(x) \pm z_{1-\alpha/2}\sqrt{\hat{\text{Var}}\ (\hat{f}(x))}. \tag{21}$$

Remarks:

If the number of degrees of freedom $d = n-(2r+4)$ is not large, then it may be desirable to use the corresponding t-cutoff in place of $z_{1-\alpha/2}$.

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

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

and (19) reduces to $\sigma^2\mathbf{d}_*^T(\mathbf{D}_*^T\mathbf{D}_*)^{-1}\mathbf{d}_*$. It follows that

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

since the model underlying (22) is more restrictive than that underlying our method. In most situations involving knot selection or variable knot locations statements based on (22) should tend to noticeably undercover the true values; 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 needs to be estimated in a separate process. A modified GCV(r) criteria is used. Given r GCV is defined to be

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

Here $2r+4$ is the total 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 computational overhead for each fit, we only calculate GCV(r) for $r \leqslant R_{max}$, which is taken to be $r_{max} = min\{n/3,\ 20\}$. Section 6.1 shows an effect of using this choice of r_{min}.

In preliminary studies we investigated some other popular model selection estimates for r, such as AIC and BIC. We found that the GCV criterion generally produced somewhat better results.

3.5 Algorithm

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

1. For $1 \leqslant r \leqslant r_{max}$ solve the nonlinear least squares problem (8). This yields estimates $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{t}}$, $\tilde{\sigma}^2$ and $\hat{f}(x)$ as functions of r and the given data.

   Efficient solution of this problem requires use of fast robust routines such as the IMSL routine DBSVLS. Care must be taken to start from several initial sets of knots in order to verify that the final solution is sufficiently close to the global minimum and is not merely a possibly unsatisfactory local extremum. See Section 3.6, Section 4 and 6.

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

3. Use the corresponding SSE to construct the estimate $\tilde{\sigma}^2$ defined in (11).

4. Calculate $\mathbf{D}$ and $\mathbf{d}$ defined in (15) and (17) and consequently $\hat{\text{Var}}\,(\hat{f}(x))$ in (20) for r_{min}, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{t}}$. Then calculate confidence intervals for f as in (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 so 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 squared estimation error.

There is some theoretical support for this observed insensitivity relative to our statistical objective. Note first that asymptotic theory supporting the use of the Wald method does not require centering of confidence intervals at the exact maximum likelihood estimates (=

least squares minima). It suffices to center on any sequence of estimates having likelihood ratio relative to the MLE converging to one. Second, our primary goal of satisfactory confidence intervals is somewhat robust with respect to use of formally incorrect local minima. Some situations we observed involved local minima with about 5% larger least squares than the apparent global minima found after repeated numerical solutions for fixed r involving a variety of initial knot locations. However, as noted above the estimating function were visually similar over much of the range of x-values. Further, an increase of say, 5% in the least squares corresponds to an increased width factor of only $\sqrt{1.05}$; i.e. only about a 2.5% increase in width. This helps explain why the average coverage, size and placement of our confidence intervals was not highly sensitive to the existence of local minima with considerably varying knot locations. This insensitivity of final confidence intervals was observed to also carry over to our complete algorithm involving the GCV criterion to select the final number of knots.

Nevertheless, the insensitivity described above is only an empirical observation aided by some heuristic motivation. Furthermore, 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 set. For these reasons we recommend that careful use of our algorithm involve repeated attempts to identify the global minimum by beginning from varied initial knots location. One possibility is to begin with initial knots 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 $I_1, \ldots, I_q$. (Usually $q = 2$ sufficed. Throughout the paper, all simulations were carried out by using $q = 2$.) Place m_i equidistant initial knots at the interior of I_i, $i = 1, \ldots, q$ such that

$$\sum_i m_i = r,\ 0 \leqslant m_i \leqslant r,\ i = 1, \ldots, q.$$

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. (i.e. $r + 1$ when $q = 2$.)

(Pittman (2001) contains recent research into alternative numerical methods that may alleviate the local minima problems. 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 some simulation investigations of coverage probabilities under our methodology. We present results for three regression functions. These functions represent a varied selection of those we have studied. We will 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 $\{5,1,3,0,-2,-8\}$. Figure 1 shows a plot of this function along with typical scatterplots for samples of size $n = 200$ and $\sigma = .45$ and .76, respectively. These two values of σ correspond to signal to noise ratios of 5 and 3, and thus correspond in a context such as this to well modeled data and to moderately noisy data. (The signal to noise level is defined in general as $S/N = \sigma_g/\sigma$ where $\sigma_g = \sqrt{\int (g(x) - \overline{g})^2\, dx}$.)

We take $n = 200$ design points to be equidistant on $[0,1]$. The simulation reports summarize the results from 1000 replications. Figure 2 shows the simulation average conditional coverage probabilities for 95% confidence intervals from our procedure conditional on x. (The true conditional coverage probability at x_k is defined as

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

and we define the average coverage probability as

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

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

The second function is typical among several we looked at involving moderately difficult to model data. It is taken from Wand (1999) where it is used to investigate accuracy of function estimates. The function is

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

Here φ denotes the standard normal density.

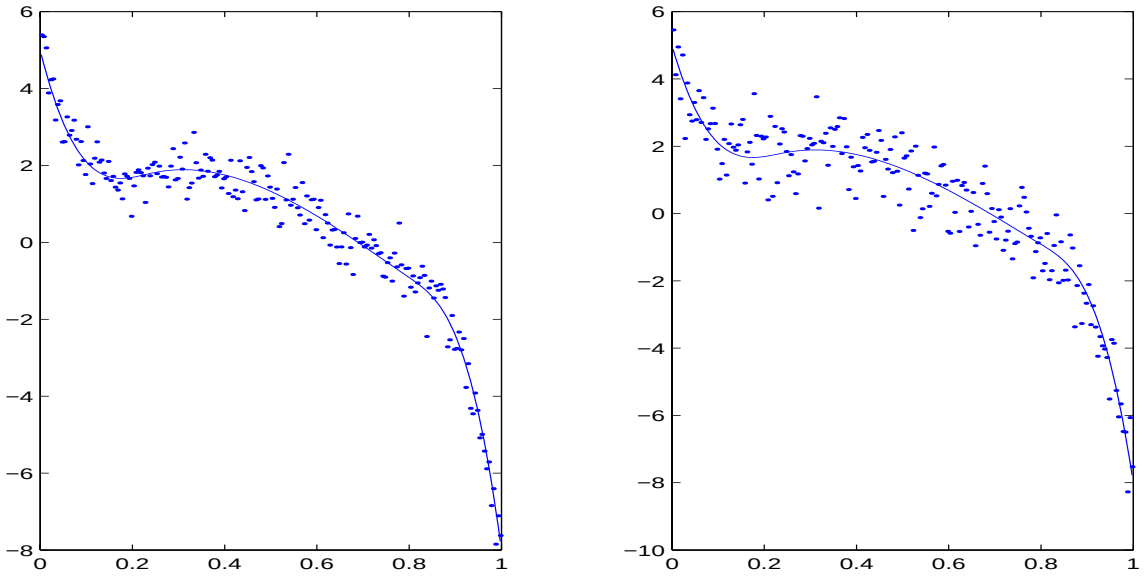

Figure 1: Scatterplots for function g_1 corresponding to S/N=5 (Signal/Noise) (left) and 3 (right).

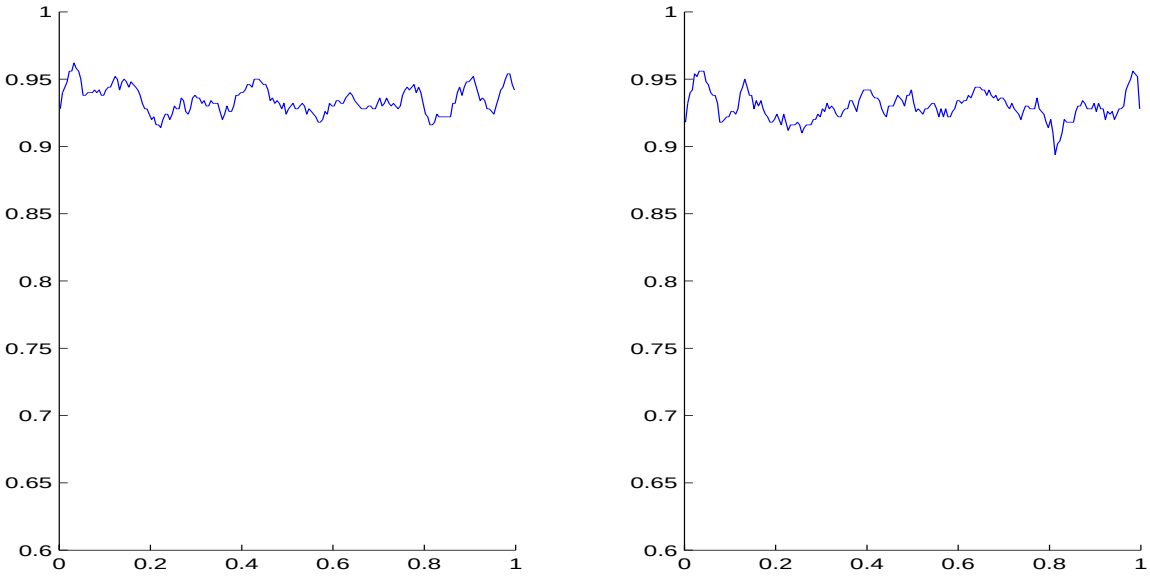

Figure 2: Empirical coverage plots for function g_1 corresponding to S/N =5 (left) and 3 (right).

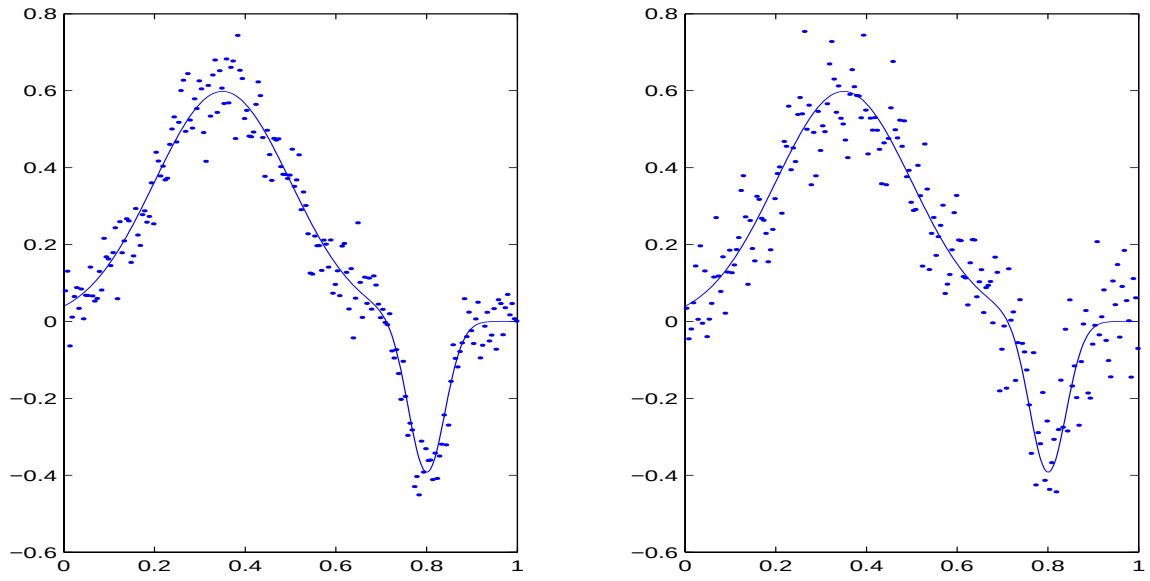

Figure 3: Scatterplots for function g_2 corresponding to S/N=5 (left) and 3 (right).

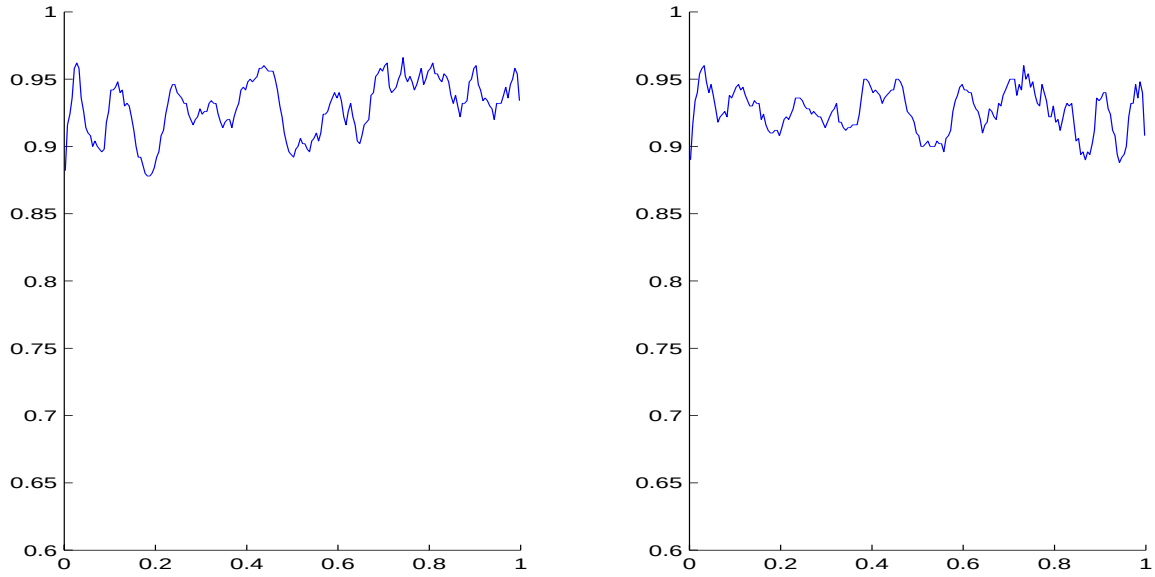

Figure 4: Empirical coverage plots for function g_2 corresponding to S/N=5 (left) and 3 (right).

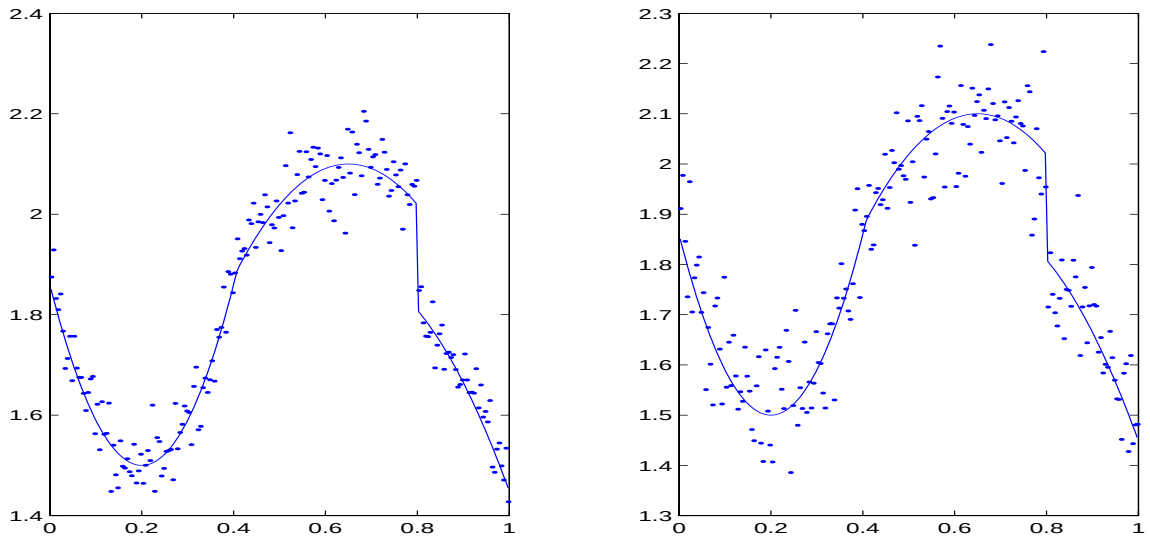

Figure 5: Scatterplots for function g_3 corresponding to S/N=5 (left) and 3 (right).

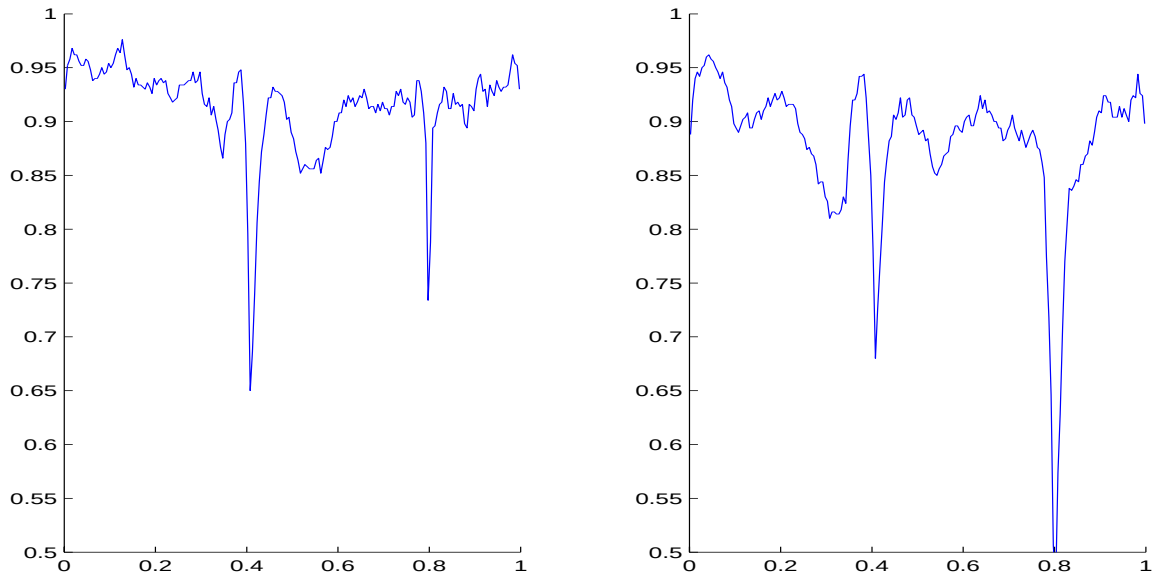

Figure 6: Empirical coverage plots for function g_3 corresponding to S/N=5 (left) and 3 (right).

Table 1: EACP for g_1,g_2,g_3; S/N = 3, 5; and n=50,100,200. The top figure is for S/N=3 and the bottom is for S/N = 5. The numbers in parentheses are 25% and 75% quantiles based on 1000 simulations.

	g_1	g_2	g_3
n = 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)
n = 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)
n = 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)

Figure 3 shows this function along with typical samples having $n = 200$ and S/N = 5, 3 (σ = .054, .09). Figure 4 shows empirical plots of CCP for 95% intervals for this situation based on 1000 simulations.

The third function is chosen by us. It is a hard to model function. It is a third order spline, but has 7 knots with a point of discontinuity at $x = .8$ and another discontinuity in it's derivative at $x = 0.408$.

$$g_3(x) = \begin{cases} 3(3(x-.2)^2+.5) & 0 \leqslant x < .4079 \\ 3(-1.2(x-.65)^2+.7) & .4079 \leqslant x < .8 \\ 3(-1.2(x-.65)^2+.7-.07) & .8 \leqslant x \leqslant 1 \end{cases}$$

Figure 5 and 6 show corresponding results for this function.

Our use of this function is intended to emphasize that free knots spline methodology can be appropriate for functions having discontinuities. Nevertheless such functions can be very hard to fit on the basis of noisy data. This is reflected in fairly narrow downward spikes in coverage probability in the neighborhood of the discontinuities. (We know of no other standard, general procedure 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 ratio 5 and 3. It turns out that the values of ECCP(x_k), $k = 1, \ldots, n$ are heavily skewed to the left for the hard to fit function, g_3. To give a better idea of the empirical distribution of CCP(x_k) we also report in Table 1 the lower and upper

quantiles of $\{\text{ECCP}(x_k) : k = 1, \ldots, n\}$.

4.2 Comparison to smoothing spline confidence intervals

Smoothing splines have been used to provide an important standard methodology for non-parametric regression confidence intervals. Wahba (1983) and Nychka (1988) show that smoothing splines are Bayes estimators corresponding to a particular Gaussian prior and

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

where $\mathbf{A}_{\hat{\lambda}}\mathbf{Y}$ is the smoothing spline estimator evaluated at $(x_1, \ldots, x_n)^T$ and $\hat{\lambda}$ is the smoothing parameter chosen by minimizing generalized cross validation (GCV). Correspondingly, they propose 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 $\hat{\sigma}^2 = \text{SSE}/(n - \text{tr}(\mathbf{A}_{\hat{\lambda}}))$. See Wahba (1990) for more information about smoothing spline techniques.

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

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

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

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

where

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

is the β density function.

The five noise levels are $\sigma = .0125, .025, .05, .1$ and $.2$ as in Wahba (1990). The three sample sizes are $n = 32, 64$ and 128. The S/N values corresponding to $\sigma = .1$ for three functions are 6.88, 9.6 and 5.4, respectively. Values of $\sigma \leqslant .5$ correspond to larger signal to noise ratio. We feel such values are of less interest for statistical applications, especially when $n = 64, 128$, but have nevertheless reported results for them because they are included in Wahba's study.

Figure 7 shows a typical sample from testing function f_1 with $n = 128$ and $\sigma = .1$. The function f_1 is plotted as a solid line. Applying our method we get the fitted line (dashed) and the 95% point wise confidence bands (dotted lines).

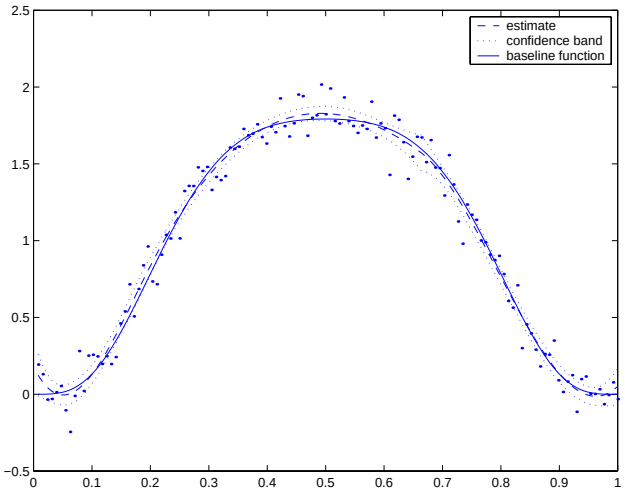

Figure 7: A typical result under f_1 when $n = 128$, $\sigma = .1$.

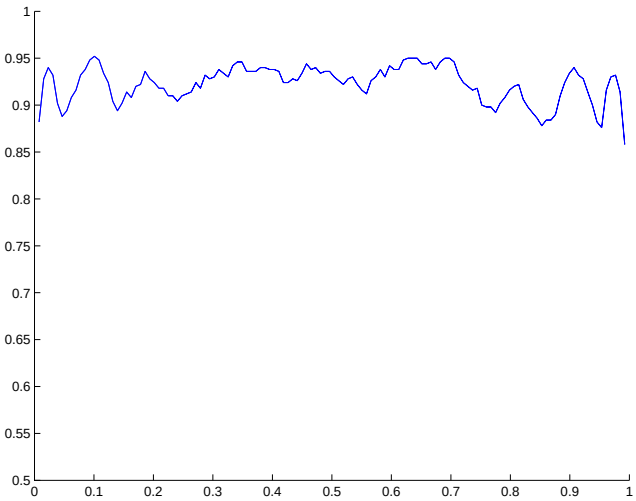

Figure 8: Empirical coverage probability as a function of x, under f_1, $n = 128$, $\sigma = .1$

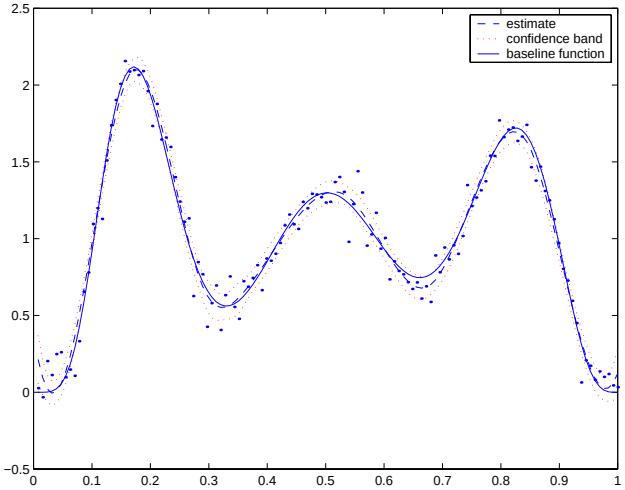

Figure 9: A typical result under f_3 when $n = 32$, $\sigma = .1$

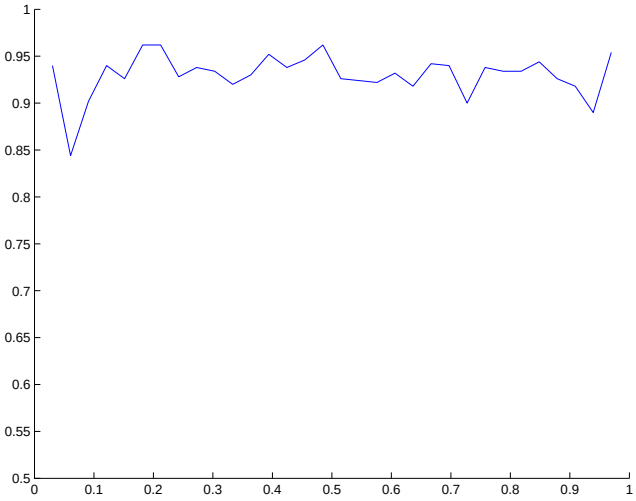

Figure 10: Empirical coverage probability as a function of x, under f_3, $n = 32$, $\sigma = .1$

Table 2: Empirical Average Coverage Probabilities (EACP) for testing function f_1 to f_3. The nominal level is 95%. Unshaded columns are results from FUNFITS; Shaded columns are from our method.

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

Figure 8 reports pointwise empirical coverage probabilities at each x for the same setting as above based on 500 replications. This shows that the coverage probability is fairly close to the nominal level .95.

Figure 9 and Figure 10 are similar to that in Figure 7 and Figure 8 but with testing function f_3 and sample size 32. Even though the sample size is relatively small the performance of our method in terms of the function estimation as well as coverage probability is reasonably satisfactory.

Table 2 reports empirical values of ACP for our method and for Wahba's method. This table is based on 100 replications at each level. (Wahba runs simulations involving only 10 replicates. To get suitable accuracy we re-ran simulations for her examples in order 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 ACP acceptably close to the nominal level of 95%. (All but 5 of the 45 values for our method exceed 90%.) The two lowest values for our method (86.8% and 86.39%) digress 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 ACP's.

4.3 Comparison of MSE with other polynomial spline procedures

Along with its confidence bands our procedure of course also produces estimates of the regression function. There is a wide range of existing methods designed to produce such estimates. Some are mentioned in our introduction. In this section we compare the estimates from our procedure with those from two other popular related methods – the adaptive knot selection procedure POLYMARS developed by Stone, et al (1997) and the variable knots Bayesian spline procedure `br` 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 be not 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

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i) - f(x_i))^2}.$$

We give results for the three functions defined in Section 4.1 with the same simulation

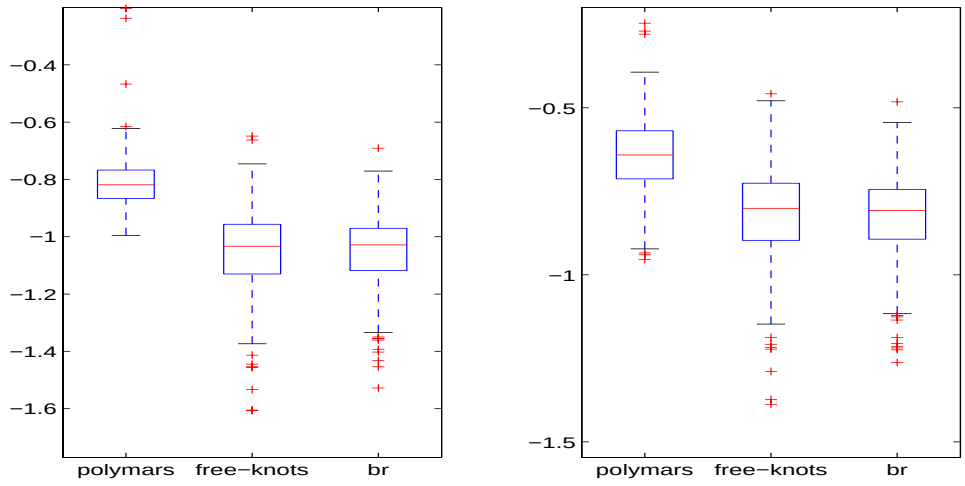

Figure 11: Boxplots of $\log_{10}$(RMSE) for function $g_1(x)$ with S/N=5 (left panel) and 3 (right panel).

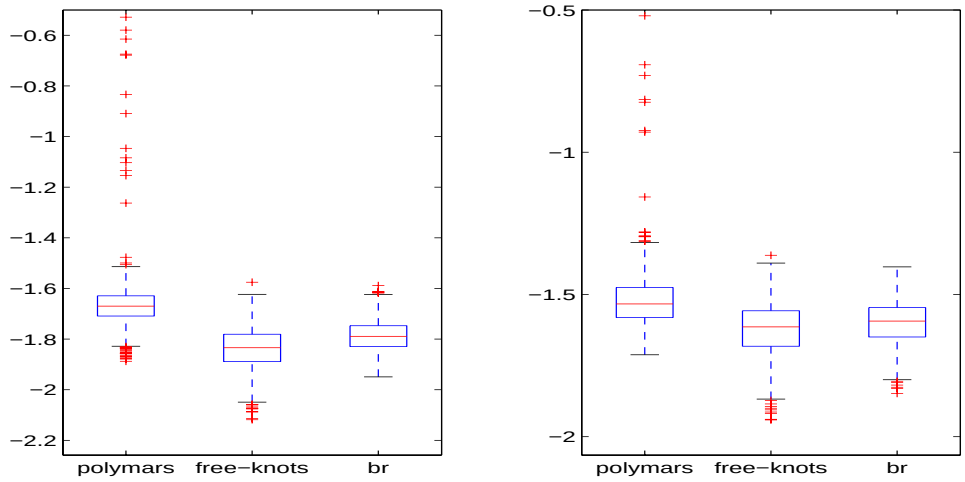

Figure 12: Boxplots of $\log_{10}$(RMSE) for function $g_2(x)$ with S/N=5 (left panel) and 3 (right panel).

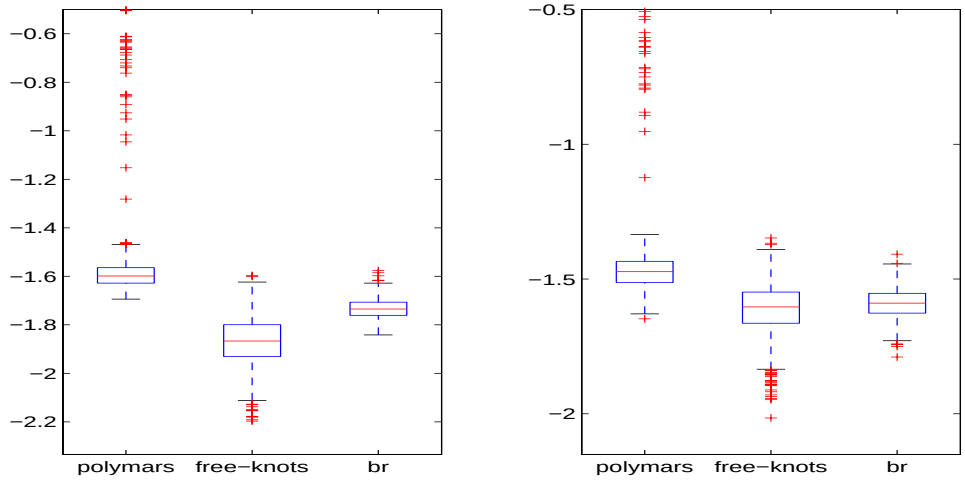

Figure 13: Boxplots of $\log_{10}$(RMSE) for function $g_3(x)$ with S/N=5 (left panel) and 3 (right panel).

setup. The boxplots in Figures 11 - 13 summarize our results. These are boxplots of the values of $\log_{10}$(RMSE) for 1000 Monte-Carlo replications of the problem. In summary, it appears that `br` and our free-knots method are generally competitive as estimation procedures and both improve on POLYMARS. The only major difference in performance appear in the left panel of Figure 13.

5 Analysis of banking data

As an example of our methodology we will reanalyze a data set discussed in Faulhaber (2000). We summarize below the essential features of this data and some of the conclusions it yielded. The original article should be consulted for further details.

The data was collected to study the "productivity" of US banks. Analysis of this data supported certain research hypotheses concerning the effect of federal policy on banking efficiency. These hypotheses are briefly summarized following our analysis of the data.

The data involves quarterly reports from a subset of US banks having assets over $1 billion in 1984. (The study thus involves only "mid-size" to "large" banks.) The period covered is 1984 through 1992.

The independent variable in this regression analysis is the total quarterly revenue for each bank. The y-variable is a measure of the banks quarterly risk to earnings ratio. In all, the data set contains 1483 (x, y) values, each representing the quarterly report from some mid-size or larger bank.

Some of the data points represent reports from the same bank over different quarterly periods. These points should thus exhibit some degree of temporal correlation. Faulhaber (2000) ignored this issue in his analysis and we will also do so, and treat the data as if the random errors are independent. (Neither bank identities nor the quarter number were reported in the paper or in the data communicated to us.)

Figure 14 contains a plot of the data in terms of $x = \log(\text{revenue})$ (with revenue in thousand dollars). We chose to use log(revenue) for this plot rather than revenue itself since the distribution of revenue is more nearly uniform in the logarithmic scale. The analysis on the log scale is thus more stable and more informative. Figure 14 also shows the polynomial spline regression curve produced by our method and the 95% confidence intervals for this regression curve.

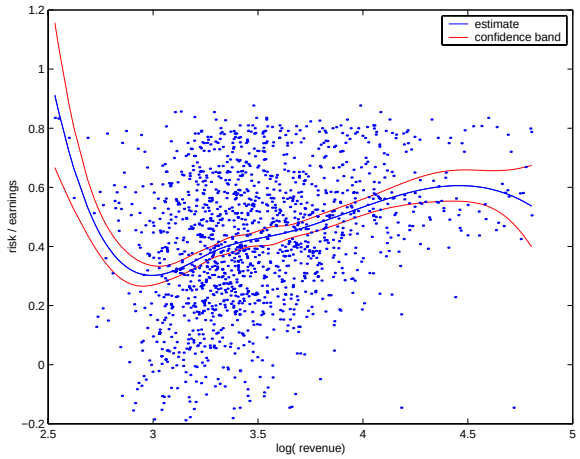

Figure 14: The bank data in log scale with fitted curve (solid) and 95% confidence intervals (dotted).

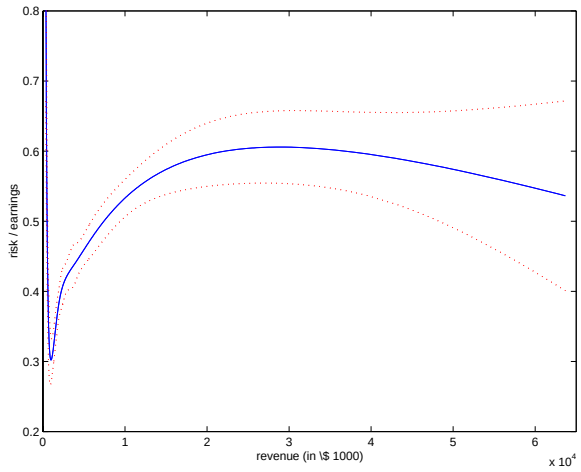

Figure 15: The fitted curve (solid) together with 95% confidence intervals for the bank data; original scale.

It is more conventional to interpret this data in terms of revenue rather than log(revenue). Figure 15 shows on this scale the regression curve and confidence region from Figure 14. Three statistically significant qualitative features are visible on this plot:

1. A sharp decrease in this curve at very small values of revenue.
2. A subsequent increase in the curve.
3. A leveling-off of the curve for larger values of the revenue.

These three features are explained in some detail in Faulhaber (2000) as respectively reflecting the following factors:

1. The decrease at small revenues is consistent with earlier studies done for smaller bank sizes that shows risk/earnings generally decreases with bank size in this range.

2. The subsequent rise in the curve reflects an optimal response to the "too big to fail" hypotheses. That hypothesis holds that for a range of revenues there is a probability that the bank will be bailed out by the government in case of failure. That probability increases with bank size over a range of values of revenue. The bank managers should be more prone to engage in risky behavior as this probability increases.

3. Above a certain revenue the probability of bail-out is nearly one. This explains why the curve levels off at larger revenues.

The plot together with the preceding explanation suggests that the "too big to fail" effect begins to occur for quarterly revenues in the vicinity of $\$1,000,000$ and is nearly complete for quarterly revenues above about $\$10,000,000$. As one may judge from the confidence bands, the slightly concave pattern of our curve above this value is not statistically significant, and may partly be an artifact of our spline methodology.

6 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 address confidence band as an alternative object.

6.1 Nonlinearity 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 subsection 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 estimation accuracy would be optimal in a number of accepted senses and the confidence coverage would be exact. The expected root mean square error will agree exactly with the theoretical value

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

obtained from the right side of (23).

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

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

This value need not be attained in practice since the theory leading to (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 modeled value of r is chosen via GCV. In this case the estimation and confidence performance can be adversely affected by incorrect choice of r as well as by the various stochastic errors discussed above.

Table 3 gives values of (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. The table includes results for $n = 50$ and 200 and for S/N level 1, 3, 5. Entries with subscript 1 refer to fitting with the correct knot locations; with subscript 2 refer to fitting with two knots at free locations; and with no subscript refer to our scheme with GCV as choice of knots. Entries beginning with "E" are empirical simulation results; the others are theoretical, as described above.

Table 3: Theoretical and empirical values ("E") for g_1. See text for complete descriptions.

	n=50			n=200		
S/N=	1	3	5	1	3	5
RMSE_1	0.6789	0.0754	0.0272	0.3926	0.1309	0.0785
"E" RMSE_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
"E" RMSE_2	0.8906	0.1107	0.0398	0.4535	0.1519	0.0889
"E" RMSE	1.0222	0.1379	0.0466	0.5320	0.1651	0.0971
"E" ACP_1	0.9421	0.9421	0.9421	0.9462	0.9462	0.9462
"E" ACP_2	0.9377	0.9201	0.9284	0.9299	0.9341	0.9434
"E" ACP	0.9050	0.8870	0.9133	0.8811	0.9257	0.9316
"E" AWidth_1	1.5193	0.5064	0.3039	0.7294	0.2431	0.1459
"E" AWidth_2	1.7404	0.5812	0.3507	0.8411	0.2819	0.1692
"E" AWidth	1.5614	0.5799	0.3565	0.7961	0.2903	0.1745

Figure 16 shows the histogram of the number of knots chosen by our GCV criterion in these simulations. Note that at higher S/N values and larger sample sizes the GCV virtually never underfits by choosing too small a number of knots. It sometimes does mildly overfit, but such mild overfitting does not have serious negative consequences for the various performance criteria.

Note that the values of "E" RMSE_2 are close to their theoretical values, RMSE_2. Hence the asymptotic values are fairly close to the actual ones. Next, "E" RMSE is somewhat larger than RMSE_2. This describes the estimation penalty for not knowing how many knots g_1 has.

The three values of "E" ACP decrease somewhat, but not too drastically as one progresses from the precise, correct model to our free knot model with r to be chosen by GCV. (The values of "E" ACP_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 "E" ACP_1 is 0.95, and the observed deviation is attributable to the random simulation effect.

Finally, "E" AWidth_1 is generally smaller than "E" AWidth_2 as one should expect. However the values of "E" AWidth_2 and of "E" AWidth are comparable in spite of the fact that the free-knot model is less precise than the two-knot model of "E" AWidth_2. This juxtaposition suggests that the free-knot confidence intervals may be somewhat too narrow and

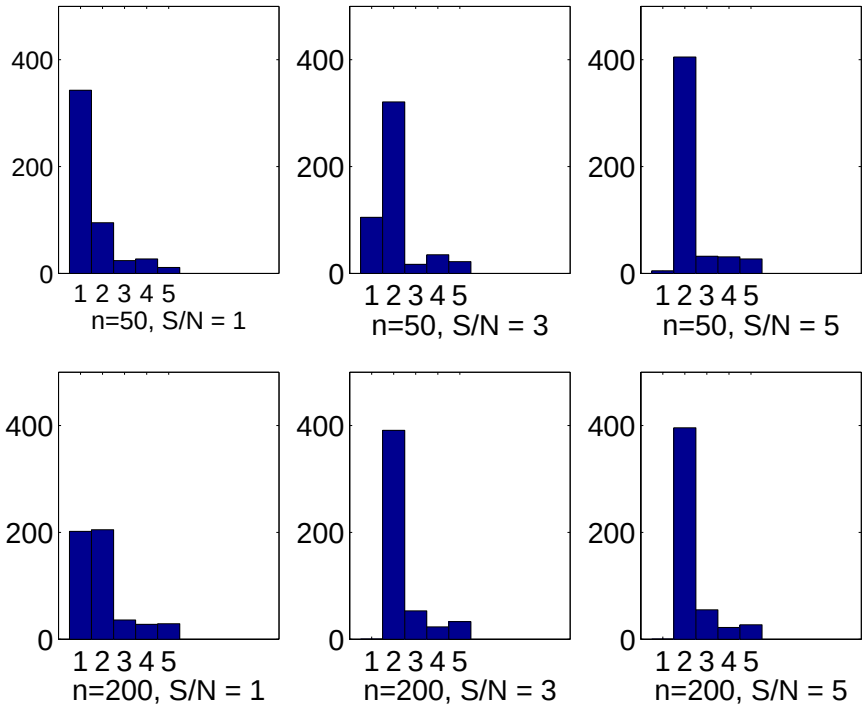

Figure 16: Histogram of the number of knots chosen by GCV in the simulations involving g_1.

that better fidelity to nominal coverage values would be obtained by somewhat increasing their width. 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 $2r+4$ coordinates of $\boldsymbol{\theta}$. But our methodology does not make an upward adjustment in length of interval because of "estimation" of r. While 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.

6.2 Alternate objectives:

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

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

and consequently

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

Of course, our algorithm is not exact, and so the degree to which (29)-(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 to 200 we found (as expected) noticeable variability in (29) as a function of x, especially for inhomogeneous f and higher noise levels; but there was only a mild tendency for under coverage on average with values of (30) for nominal $1-\alpha = .95$ ranging from the mid 80% range to nearly .95, depending on the example. For signal to noise ratios of one or less we found noticeable degradation in the coverage performance of our intervals, as well as of the few existing alternative methods we have tried.

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

$$P(f(x) \in CI(x)\ \forall x) \geqslant 1-\alpha. \tag{31}$$

For this purpose one could replace the value $z_{1-\alpha/2}$ in (21) by $((2r+4)F_{1-\alpha})^{1/2}$ where $F_{1-\alpha}$ denotes the upper α cutoff point of an F - distribution with $(2r+4)$ and $n-(2r+4)$ df. This simultaneous confidence band would nominally be conservative (asymptotically). One might hope to reduce this conservativity 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.

One could alternatively desire prediction intervals of the usual sort instead of confidence intervals for $f(x)$. For this purpose one would replace $\sqrt{\hat{Var}(\hat{f}(x))}$ in (21) by $\sqrt{\tilde{\sigma}^2 + \hat{Var}(\hat{f}(x)}$.

There is heuristic reason to believe that performance of our methods for the above objectives would be even better than that for our primary confidence objectives (1), (2). This will be reported elsewhere. (Zhao (in preparation).)

7 Summary

The method of polynomial splines has been made more flexible by letting knot locations vary freely.

The number of knots is chosen via a model selection (GCV) method, the other basic parameters are estimated via maximum likelihood. (σ^2 is estimated by an asymptotically

unbiased estimator, as is customary in such settings.) The predominant use of maximum likelihood estimation supports the construction of reliable confidence intervals for the underlying regression function. Simulation, as well as general theory, supports the conclusion that these confidence intervals generally have coverage acceptably near their nominal value.

One of the advantage of our method is that it exploits existing algorithms so the coding is relatively flexible and easy. The main program is written in MATLAB using Spline Toolbox (de Boor (1998)) with function calls to IMSL.

Acknowledgment

The authors would like to thank Lawrence Brown for his continuous support and numerous useful suggestions. Charles Stone's encouragement also played an important role.

References

Bates, D.M. and Watts, D.G., (1988) *Nonlinear Regression Analysis and its Applications*, John Wiley & Sons, Inc., New York.

de Boor, C. and Rice, J. R. (1968) Least Squares Cubic Spline Approximation II – Variable Knots, Technical Report 21, Purdue University, Computer Science Department. (available at `http://www.cs.wisc.edu/~deboor/`).

de Boor, C. (1978) *A Practical Guide to Splines* , Springer-Verlag.

de Boor, C. (1998) *Spline Toolbox for Use with Matlab User's Guide*, The Mathworks Inc.

Dierckx, P. (1993) *Curve and Surface Fitting with Splines*, Oxford Science Publications.

Denison, D.G.T., Smith, A.F.M., and Mallick, B.K. (1998) Automatic Bayesian Curve Fitting, *JRSS*, B, 60, Part 2, 333-350.

Dette, H., Munk, A., and Wagner, T. (1998) Estimating the Variance in Nonparametric Regression – What is a Reasonable Choice?, *JRSS*, B, 60, Part 4, 751-764.

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

Eubank, R.L. and Speckman, P.L. (1993) Confidence Bands in Nonparametric Regression,

JASA, Vol. 88, No. 424, 1287-1301.

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

Faulhaber, G.R. (2000) Banking Markets: Productivity, Risk, and Customer Satisfaction, Technical report.

Friedman, J.H. (1991) Multivariate Adaptive Regression Splines (with discussion), *Annals of Statistics*, 19, 1-141.

Friedman, J.H. and Silverman, B.W. (1989) Flexible Parsimonious Smoothing and Additive Modeling (with discussion), *Techometrics*, Vol 31, No. 1, 3-39.

Greville, T.N.E. (1969) Introduction to Spline Functions, *Theory and Applications of Spline functions, Proceedings of Seminar, Math. Research Center, Univ. of Wis., Madision*, New York: Academic Press, 1-35.

Hall, P., Kay, J.W., and Titterington, D.M. (1990) Asymptotically Optimal Difference-based Estimation of Variance in Nonparametric Regression, *Biometrika*, 77, 3, 521-528.

Härdle, W. and Marron, J.S. (1991) Bootstrap Simultaneous Error Bars for Nonparametric Regression, *The Annals of Statistics*, Vol. 19, No. 2, 778-796.

Hastie, T. J. and Tibshirani, R.J. (1990) *Generalized Additive Models*, Chapman & Hall.

Johansen, S. and Johnstone, I. (1990) Hotelling's Theorem on the Volume of Tubes: Some Illustrations in Simultaneous Inference and Data Analysis, *Annals of Statistics*, Vol. 18, No. 2, 652–684.

Jupp, D.L.B (1978) Approximation to Data by Splines with Free Knots, *SIAM Journal of Numerical Analysis*, 15, 328-343.

Kooperberg, C. and Stone, C. (2001) Logspline Density Estimation with Free-knot Splines, Preprint.

Lehmann, E. L. (1999) *Elements of Large Sample Theory*, Springer-Verlag, New York.

Lindstrom, M. J. (1999) Penalized Estimation of Free-knot Splines, *JCGS*, Volume 8, Number 2, 333-352.

Loader, C. (1999) *Local Regression and Likelihood*, Springer-Verlag, New York.

Nadaraya, E.A. (1964) On Estimating Regression, *Theory Probab Appl*, 9, 141-142.

Nychka, D. (1988) Bayesian Confidence Intervals For Smoothing Splines, *JASA*, Vol 83, NO. 404, 1134-1143.

Nychka, D. et al (1996) FUNFITS: Data Analysis and Statistical Tools for Estimating Functions, *North Carolina Institute of Statistics Mimeoseries No. 2289*. (available from `http://www.cgd.ucar.edu/~nychka/`.)

Pittman, J. (2001) Adaptive Splines and Genetic Algorithms with an Application to Classification, Preprint.

Schumaker (1981) *Spline Functions Basic Theory*, John Wiley & Sons.

Smith, M. and Kohn, R. (1996) Nonparametric Regression Using Bayesian Variable Selection, *J. Econometrics*, **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), *Annals of Statistics*, Vol 25, No. 4, 1371-1470.

Tribouley, K. (2000) Adaptive Confidence Interval for the Density, Preprint.

Wahba, G. (1983) Bayesian 'Confidence Intervals' for the Cross-validated Smoothing Spline, *JRSS B*, 45, No.1, 133-150.

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

Wand, M.P. (1999) A Comparison of Regression Spline Smoothing Procedures, Technical Report. (Available at `http://www.biostat.harvard.edu/~mwand`.)

Xia, Y. (1998) Bias-corrected Confidence Bands in Nonparametric Regression, *JRSS B*, 60, Part 4, 797-811.

Zhou, S., Shen, X., and Wolfe, D.A. (1998) Local Asymptotics for Regression Splines and Confidence Regions, *Annal of Statistics*, Vol 26, No 5, 1760-1782.