Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://amstat.tandfonline.com/loi/uasa20

Spatially Adaptive Regression Splines and Accurate Knot Selection Schemes

Shanggang Zhou & Xiaotong Shen

To cite this article: Shanggang Zhou & Xiaotong Shen (2001) Spatially Adaptive Regression Splines and Accurate Knot Selection Schemes, Journal of the American Statistical Association, 96:453, 247-259, DOI: 10.1198/016214501750332820

To link to this article: http://dx.doi.org/10.1198/016214501750332820

Spatially Adaptive Regression Splines and Accurate Knot Selection Schemes

Shanggang ZHOU and Xiaotong SHEN

Spline procedures have proven effective in estimating smooth functions. However, spline procedures based on stepwise addition and/or deletion have some drawbacks. They suffer from the knot compounding problem, making their performance suboptimal. Furthermore, due to computational complexity, spline procedures may not achieve their full potential. In this article, we propose a novel knot selection algorithm for regression spline estimation in nonparametric regression. The algorithm includes three new components: knot relocation, guided search, and local fitting. The local properties of the spline functions are used to efficiently implement the algorithm. Extensive simulation studies are performed to demonstrate the improvement of the new knot selection algorithm over the stepwise addition and deletion scheme, and the advantages of the spline procedure with the new knot selection scheme over alternative adaptive methods. In the simulations, our procedure achieves very competitive performance with alternative methods and has substantial advantage in nonsmooth functions. Finally, the usefulness of the proposed method is illustrated by an application to signal recovery in speech signal processing.

KEY WORDS: Knot relocation; Regression spline; Spatial adaptation; Stepwise Knot selection.

1. INTRODUCTION

Consider the problem of estimating regression function $f(\cdot)$ from iid noisy samples,

$$y_i = f(x_i) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where the design points $x_i \in [0, 1]$ can be either fixed or random and ϵ_i is random noise with mean 0 and unknown variance σ^2. Our goal is to estimate $f(\cdot)$ with small mean squared error as well as good visual quality without assuming that the degree of smoothness of f is known.

In many practical circumstances, $f(\cdot)$ is inhomogeneous, and the degree of smoothness is unknown in advance. Such a problem often occurs in signal recovery from noisy background in engineering. Figure 1 displays the picture of a speech signal of the word "greasy," obtained from the signal processing lab at The Ohio State University, sampled at 16 KHz. The speech signal exhibits nonstationary characteristics with variable structure in time. In this situation, spatial adaptation clearly is necessary for recovering the signal.

There has been considerable recent work on the development and application of procedures that can adapt to inhomogeneous smoothness of the unknown function f. The work includes an adaptive regression spline method (e.g., Friedman 1991), an adaptive smoothing spline method by Luo and Wahba (1997), an adaptive kernel method by Fan and Gijbels (1995), and wavelet shrinkage methods by Donoho and Johnstone (1994, 1995). In the literature of spline estimation, algorithms for adaptive estimation have been proposed by Friedman and Silverman (TURBO; 1989); Friedman (MARS; 1991); Stone, Hansen, Kooperberg, and Troung (1997); and Luo and Wahba (1997). These methods, using the traditional stepwise forward and/or backward knot selection (e.g., Smith 1982), have proven effective in estimating relatively smooth regression functions. But these spline procedures equipped with a stepwise selection scheme have several drawbacks: They suffer from knot compounding problem (as discussed in Sec. 3), and they may not achieve their full potential for spatial adaptation due to computational cost in stepwise selection. As demonstrated in the simulation studies in Section 6, their performance is indeed suboptimal. Usually, good approximations of inhomogeneous functions by spline functions usually require a set of highly unevenly distributed knots. Hence examining a large set of candidate knots is necessary, causing tremendous computational burden when the sample size is large. On the other hand, an only partial search for knots may cause poor performance of the spline procedures, as reported by Luo and Wahba (1997). To overcome these problems, we seek an accurate and efficient knot selection scheme and further develop the potential of the regression spline method.

Ideally, knot selection should be performed jointly instead of marginally, to obtain the global optimal knots. Unfortunately, this task is infeasible even for small datasets, because the number of numerical evaluations grows exponentially with the number of knots. Therefore, it is necessary to restrict the knot search in a subset, which leads to the undesirable knot confounding problem. To overcome this difficulty and reduce computational cost, we use certain special local properties of the spline estimators:

1. Knot addition at a single point essentially does not change the values of the spline estimate outside a local neighborhood of the point.
2. The value of the spline estimate at a particular point can be well approximated by a locally fitted spline using the observations in its neighborhood whose size is determined by the spline used and the associated knots; see Section 5 for more details.

To our knowledge, this aspect has not been explored in the existing literature. In comparison with the existing knot selection schemes, our scheme differs in three aspects:

1. Knot relocation is used to reduce the aforementioned knot confounding problem.

Shanggang Zhou is Assistant Professor, Department of Statistics, Kansas State University, Manhattan, KS 66506 (E-mail: *zhou@stat.ksu.edu*). Xiaotong Shen is Associate Professor, Department of Statistics, The Ohio State University, Columbus, OH 43210 (E-mail: *xshen@state.ohio-state.edu*).
The project was partially supported by National Science Administration grant MDA904-98-1-0030. The authors thank the referee and editors for their helpful comments.

2. Knot search focuses in areas where more knots are necessary.

3. The locally fitted splines are used, which significantly eases computation burden.

To evaluate the performance of our procedure, we perform simulations using well-known examples from the literature. In the simulations, our procedure performs very competitively with the other adaptive procedures in all of these examples and has substantial advantage in nonsmooth examples. Moreover, our procedure does not suffer from the knot confounding problem and thus yields more accurate estimates. Furthermore, our procedure locates the optimal knots quite fast even for a large dataset. This is due to the built-in search scheme and the use of local spline fitting instead of the global fitting used by the conventional splines. Finally, the proposed procedure is applied to recover the speech signal in Figure 1 from a noisy background.

The article is organized as follows. Section 2 introduces the regression spline. Section 3 discusses the new idea in our knot selection, and Section 4 describes the knot selection scheme in detail. Section 5 discusses the effects of knot addition and local spline approximation. Section 6 summarizes simulation results, and Section 7 presents the speech recovery example. Section 8 provides some concluding remarks, the Appendix gives technical proofs.

2. SPLINE REGRESSION AND MODEL SELECTION

In estimation, the true regression function $f(\cdot)$ in (1) is approximated by a polynomial spline,

$$\tilde{f}(x; \mathbf{a}, \mathbf{t}) = \sum_{i=0}^{m-1} a_i x^i + \sum_{j=1}^{k} a_{j+m-1}(x - t_j)_+^{m-1},$$

where $(x - t_j)_+^{m-1} = \max(0, (x - t_j)^{m-1})$ and $\mathbf{t} = (0 = t_0 < t_1 < \cdots < t_{k+1} = 1)$ is the knot sequence. A regression spline estimator $\hat{f}(x; \mathbf{t})$ is then defined as $\tilde{f}(x; \hat{\mathbf{a}}, \mathbf{t})$, where $\hat{\mathbf{a}}$ is the solution of the least squares problem

$$\min_{\mathbf{a}} \sum_{i=1}^{n} (y_i - \tilde{f}(x_i; \mathbf{a}, \mathbf{t}))^2. \tag{2}$$

In computation, the corresponding equivalent B-spline representation for $\tilde{f}(x; \mathbf{a}, \mathbf{t})$ is used in (2) because B-splines are relatively well conditioned and yield an estimate that is numerically more stable than the power series representations. (See, e.g., de Boor 1978, pp.129–131 for an inductive construction of B-splines.) Let $\mathbf{N}(x) = (N_1(x), \ldots, N_{k+m}(x))'$ be the normalized B-spline functions associated with knot sequence $\mathbf{t}$, where

$$N_i(x) = (t_i - t_{i-m})[t_{i-m}, \ldots, t_i](t - x)_+^{m-1}.$$

Here $[t_{i-m}, \ldots, t_i]g$ denotes the mth order divided difference of the function g and $t_i = t_{\min(\max(i,0),k+1)}$ for any $i = 1 - m, \ldots, k + m$. The regression spline $\hat{f}(x; \mathbf{t})$ can be written explicitly as

$$\hat{f}(x; \mathbf{t}) = \mathbf{N}(x)\mathbf{G}^{-1}\mathbf{X}\mathbf{Y}, \tag{3}$$

where $\mathbf{G} = n\mathbf{X}\mathbf{X}'$, $\mathbf{X} = n^{-1}(\mathbf{N}(x_1), \ldots, \mathbf{N}(x_n))$, and $\mathbf{Y} = (y_1, \ldots, y_n)'$.

Knowing the number of knots and knot locations is critical for precise spline estimation. Such knowledge is typically unavailable in practice however. Therefore, a model selector is required for selecting the optimal knots which provide the best trade-off between the estimation and approximation errors. In

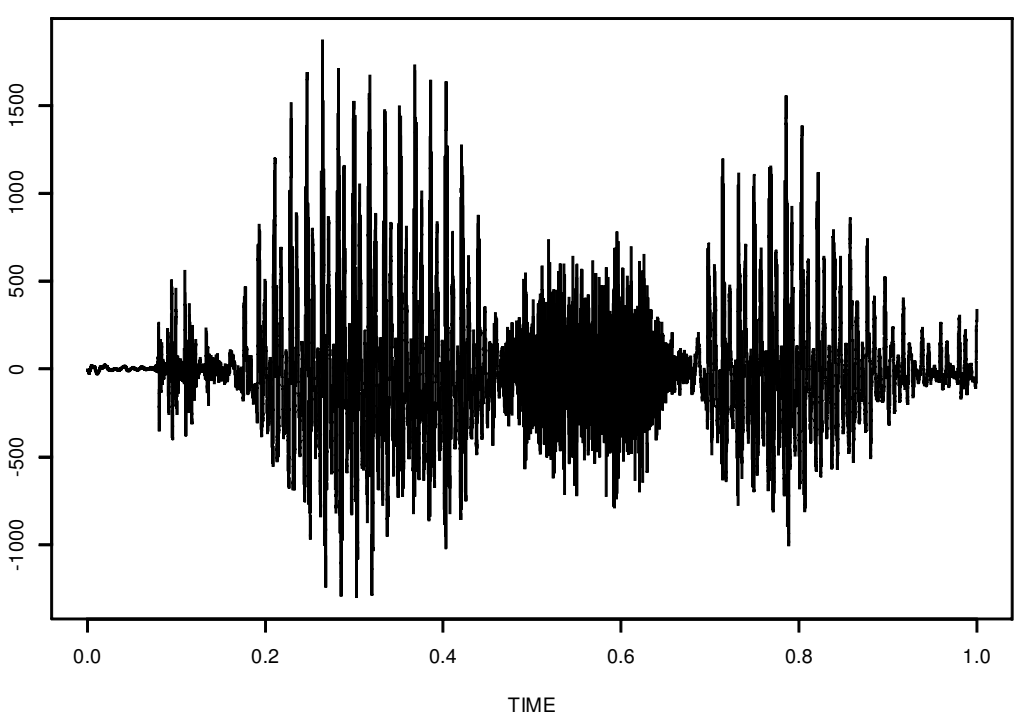

Figure 1. Speech Recording for the Word "Greasy."

our procedure, Stein's unbiased risk estimate (Stein 1981) is used as a model selector, which takes the form of

$$R(\hat{f}) = \sum_{i=1}^{n}(y_i - \hat{f}(x_i; \mathbf{t}))^2/n + C(k+m)\sigma^2/n, \qquad (4)$$

where $C = 2$. When σ^2 is unknown, σ is estimated by a robust estimator median$\{|y_{2i} - y_{2i-1}|/(.6745\sqrt{2}) : i = 1, \ldots, n/2\}$. Because evaluation of $R(f)$ requires solving the least squares problem (2), it is computationally expensive when the sample size n or the number of knots is large. Therefore, we use a local spline approximation to reduce the computational cost in implementation. The details are deferred to Section 5.

In the literature, commonly used model selectors include Stein's unbiased risk estimate, the Akaike information criterion (AIC), and generalized cross validation (GCV) (Craven and Wahba 1979). In spline estimation, several researchers suggested using different coefficient C instead of $C = 2$ in (4) for selecting an "optimal" set of knots. For example, Friedman (1991) recommended $C = 3$ for MARS, which tends to yield a smaller model, and Luo and Wahba (1997) used $C = 1.2$ for model selection, which tends to yield a larger model.

In implementation of spline procedures, such a choice of C becomes important for the accuracy of spline estimation. An inflated coefficient C may yield poor performance for less-smooth or nonsmooth functions, although it may do well for smooth functions, and the effect of a deflated coefficient C may be opposite. This aspect has been confirmed in a simulation study by Luo and Wahba (1997). Indeed, as shown in our simulations in Section 6, the choice of $C = 2$ is appropriate for both smooth and nonsmooth functions.

To explain this phenomenon, we note that stepwise knot selection procedures may fail to select the "right" knots and may yield more knots when $C = 2$ is used; see Section 3. Therefore, use of an inflated coefficient C (e.g., $C = 3$) may yield fewer knots. Such a practice may appear sensible for smooth functions but may yield poor performance for nonsmooth functions, where more knots are necessary. On the other hand, a similar problem occurs when a deflated coefficient C is used. Thus it is necessary to deal with the knot confounding problem directly rather than choose a subjective coefficient C for spline procedures to work well for both smooth and nonsmooth functions.

3. NEW IDEAS IN KNOT SELECTION

3.1 Knot Confounding and Relocation

As discussed in Section 1, the traditional knot selection methods based on stepwise forward addition and backward deletion may fail to locate the appropriate knots. Here we illustrate this phenomenon with a simple numerical example. Figure 2(a) displays simulated data from (1), where the ϵ_i are iid. N(0, 1) and the test function $f(x)$ is a cubic spline with four internal knots at .5, .7, .8, and .9, which are of course the optimal knot locations for this function. In Figure 2(a), the knots selected by the traditional stepwise procedure are shown at the bottom, where the order of the knot addition is also numbered. Apparently, the stepwise selection method has not located the optimal knots and has selected more knots

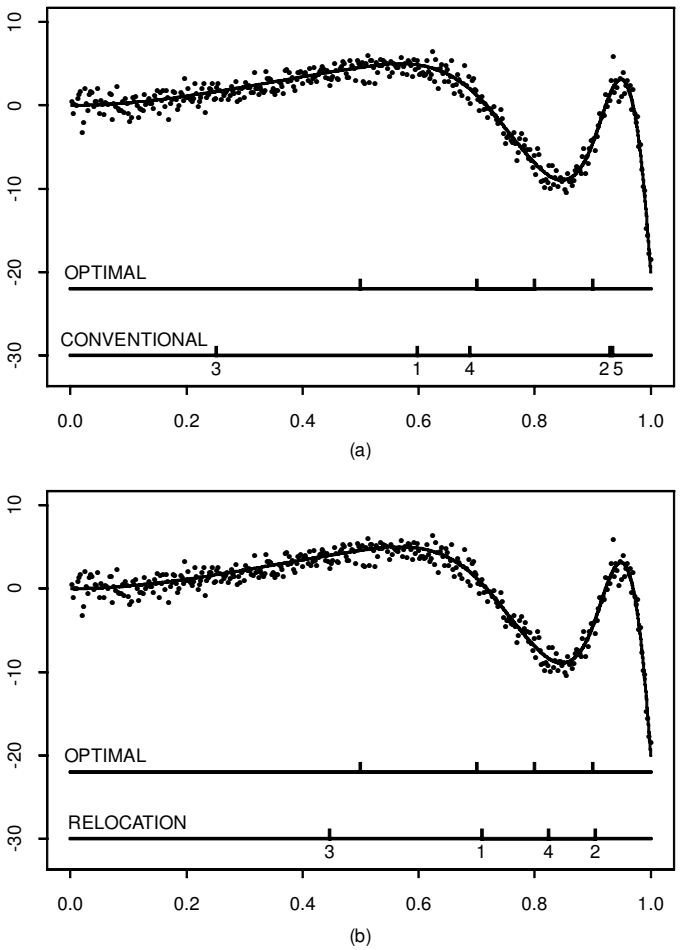

Figure 2. Comparison Between the Conventional Knot Selection Scheme and Our Knot Selection Method Scheme. (a) Knots selected by stepwise addition and deletion; (b) knots selected by our knot selection method with knot relocation.

than necessary, because knot 5 is actually redundant. In the knot addition process, knot 1 is optimal among all possible single knots, but it is no longer optimal when more knots are used. In other words, the first several selected knots are most likely "incorrect." In the knot deletion process, most of these "incorrect" knots cannot be deleted, because their effects on the model selector $R(\hat{f})$ are significant. As noted by Friedman and Silverman (1989), the improvement due to knot deletion is essentially small. In summary, the stepwise procedure tends to select unnecessary knots and place knots at wrong locations. Intuitively, the knot confounding problem is due mainly to the use of marginal searches that do not take into account the association between knots. To overcome this difficulty, we propose a knot relocation scheme that efficiently uses the association between adjacent knots to adjust knot positions in the process of knot addition and deletion. This scheme leads to a set of knots that are much closer to the global optimal knots than the traditional stepwise knot selection scheme. Our simulation results in Section 6 support the utility of this scheme, although we have not proven that the scheme gives the global optimal set of knots in general.

The knot relocation process can become computationally intensive if one considers relocating all existing knots. But this is unnecessary, because the addition of a new knot may affect

only the adjacent knots, as stated in Theorem 1 of Section 5. Hence it suffices to relocate the knots in neighborhoods of the newly added knots.

For illustration, we examine the numerical example in Figure 2(a) using the knot relocation scheme. The selected knots are given at the bottom of Figure 2(b). Clearly, the knot relocation scheme has significantly reduced the knot confounding problem.

In knot selection, it is necessary to consider computational cost, especially for large sample sizes, because the amount of computation may increase exponentially in both sample size and number of knots. Obviously, knot relocation increases the computational cost. Next we discuss an idea to improve computational efficiency of the knot selection.

3.2 Guided Knot Search

For a spatially inhomogeneous regression function that is smooth in one region and not smooth in another region, we need to insert more knots in areas where the regression function is less smooth. This motivates the use of a guided knot search, which makes the search more efficient. In our procedure, the intervals in which more knots are necessary are searched more frequently. In other words, the likelihood of searching an area is determined by whether knots have been added in the area in the previous search step (i.e., step 2 in Sec. 4). When an interval has not had knot addition after being searched in the previous step, the likelihood for more knot addition in the current and later stages becomes smaller, and thus lower priority should be assigned to this interval. In the guided search, intervals close to a knot added in the previous search step are visited with higher priority. More precisely, (t_i, t_{i+1}) is chosen to be searched if and only if at least one of the knots $t_{i-L}, t_{i-L+1}, \ldots, t_i, t_{i+1}, \ldots, t_{i+L+1}$ $(L > 0)$ was a knot added in the previous search step, where $\mathbf{t} = (t_0, \ldots, t_{k+1})$ is the current knot sequence. In implementation, we choose an appropriate L to save computation and to ensure that the search is thorough. Based on our limited experience, $L = 2$ seems sufficient.

4. IMPLEMENTATION

4.1 Algorithm

We now describe the algorithm based on the ideas described in Sections 3.

Step 1 (Knot Initialization). This step provides a set of good initial knots, which speeds up the knot search process, although it is not necessary. The details are given in Section 4.2.

Step 2 (Knot Search). Let the existing knot sequence be $\mathbf{t} = (t_0 = 0 < t_1 < \cdots < t_k < 1 = t_{k+1})$. If $\mathbf{t}$ is the set of initial knots, then for $i = 0, \ldots, k$, $[t_i, t_{i+1}]$ is checked for possible knot addition. Otherwise, the guided knot search, as discussed in Section 3, is performed as follows: for $i = 0, \ldots, k$, search for a potential knot addition in $[t_i, t_{i+1}]$ if at least one of the knots $t_{i-L}, t_{i-L+1}, \ldots, t_i, t_{i+1}, \ldots, t_{i+L+1}$ is among the knots added in the previous knot addition, where $L = 2$ is used in our procedure for computational consideration. This gives the subintervals with the most recent knot additions and their close neighborhoods a high likelihood of being searched.

More precisely, let $\mathbf{T}$ be a working knot sequence, which bookkeeps the updated knots, and let $\phi \subset \mathbf{t}$ be the set of knots added in the previous knot addition. The knot search is conducted recursively, as described in the following pseudocode:

```
Let T = t;
from i = 0 to i = k do
    if there is t_j ∈ φ such that |i − j| ≤ L, then
      search in (t_i, t_{i+1}) for a potential knot t* which
      minimizes R(f̂); add t* to T if R(f̂) with
      t* is smaller than R(f̂) without t*;
    end;
end.
Update t; that is, t = T.
```

Now $\mathbf{t}$ becomes the current knot sequence. If no knots have been added, then go to step 4.

Step 3 (Knot Relocation and Deletion). Let the current knot sequence be $\mathbf{t}_a = (t_0 = 0 < t_1 < \cdots < t_{k_a} < 1 = t_{k_a+1})$. Let ϕ be the set of knots added in step 2. The knots that are left or right adjacent to any knot in ϕ are checked for relocation or deletion, where the relocation and deletion are performed using the model selector defined in (4) of Section 2.

More precisely, let $\mathbf{T}$ be as defined in step 2. The knot adjustment (relocation or deletion) is conducted recursively, as described in the following pseudocode:

```
Let T = t_a;
from i = 1 to i = k_a do
    if there is t_j ∈ φ such that |i − j| = 1, then
      relocate t_i in [t_{i−1}, t_{i+1}] or delete t_i using
      the model selector in (4); update T;
    end;
end.
```

Readjust the knots in ϕ that were added in step 2 using the same procedure. Then update $\mathbf{t}_a$; that is, $\mathbf{t}_a = \mathbf{T}$. Go to step 2.

Step 4 (Refinement). When no knots can be added in step 2 for the first time, adjust (relocate or delete) all the selected knots. Then label all the existing knots as initial knots and go to step 2 again. Stop when no knots can be added in step 2 for the second time.

In the knot search in step 2 and knot relocation in step 3, we use the Fibonacci search (see, e.g., Gill, Murray, and Wright 1981) to expedite the process.

4.2 Knot Initialization

Using an appropriate set of initial knots speeds up the knot search process, especially when a large number of knots is involved, mainly because it permits use of local spline approximation in an early stage of the search process to reduce the computational cost, and it provides guidance for possible locations of knot additions at later stages. To reduce the computational complexity in knot initialization, we derive a simple procedure assuming that the design points are equally spaced. Such an assumption is not critical, because precise initial knots are not necessary and incorrect locations can be adjusted in step 3 of the knot selection algorithm in Section 4.1.

We now begin our discussion on the selection of initial knots. Intuitively, a need for knot insertion at any design point

x_i implies that a local mth-order polynomial does not yield an adequate fit for the data in the neighborhood $[x_i - b, x_i + b]$ for some $b > 0$. To test the adequacy of fit for a given b, we fit the data in $[x_i - b, x_i + b]$ by the mth-order polynomial plus an extra term and solve the following least squares problem:

$$\min_{\alpha_{l,b},\beta} \sum_{x_j \in [x_i - b,\, x_i + b]} \Bigg(y_j - \alpha_{l,b}(x_i)(x_j - x_i)_+^l - \sum_{v=0}^{m-1} \beta_v (x_j - x_i)^v \Bigg)^2, \quad (5)$$

where $0 \le l < m$ is a fixed integer. In (5), the coefficient $\alpha_{l,b}(x_i)$ that corresponds to the added term is used to check the adequacy of fit. Some simple linear algebra yields

$$\hat{\alpha}_{l,b}(x_i) = \sum_{x_j \in [x_i - b,\, x_i + b]} w_{l,b}(i - j) y_j, \quad (6)$$

where $w_{l,b}(\cdot)$ is a weight function and $x_i \in [b, 1 - b]$. A significant nonzero estimated coefficient $\hat{\alpha}_{l,b}(x_i)$ for some $l = 0, \ldots, m - 1$ and $b > 0$ indicates a lack of fit around x_i. Therefore, knot insertion at x_i is necessary. More precisely, for any l and $b > 0$, let $r(x_i)$ be a local maxima in the sense that

$$r(x_i) = \left| \hat{\alpha}_{l,b}(x_i) \Big/ \sqrt{\operatorname{Var}(\hat{\alpha}_{l,b}(x_i))} \right| = \max_{x_u \in [x_i - b,\, x_i + b]} \left(\left| \hat{\alpha}_{l,b}(x_u) \Big/ \sqrt{\operatorname{Var}(\hat{\alpha}_{l,b}(x_u))} \right| \right),$$

where $\hat{\alpha}_{l,b}(x_u)$ is obtained by replacing x_i and i in (6) by x_u and u. By assumption, $\operatorname{Var}(\hat{\alpha}_{l,b}(x_i)) = \operatorname{Var}(\hat{\alpha}_{l,b}(x_u)) = \sigma^2 \sum_{x_j \in [x_i - b,\, x_i + b]} w_{l,b}^2(i - j)$. Because σ^2 is unknown, it is estimated by a robust estimator such as $\operatorname{median}\{|y_{2i} - y_{2i-1}|/(.6745\sqrt{2}) : i = 1, \ldots, n/2\}$. This leads to an estimate of $\operatorname{Var}(\hat{\alpha}_{l,b}(x_i))$ and $r(x_i)$. The estimated value of $r(x_i)$ decides whether the design point x_i should be selected as an initial knot; that is if $r(x_i) > C_l$ for some $0 \le l < m$, $C_l > 0$ and b, then x_i is to be selected as an initial knot. In implementation, the critical value C_l is chosen to be 5 for $l = 0$ and 1.

The foreging selection procedure for initial knots is performed iteratively. For each iteration, the newly identified knots are added to a set, say Δ, and in the next iteration the selection is performed only for the design points $\{x_i\}$ that are not close to the knots currently in Δ. The algorithm is as follows.

Let $b = (x_n - x_1)/n$.

Step 1. Compute $\hat{\alpha}_{l,b}(x_i)$ for $l = 0$ and $x_i \in \Omega(b, \Delta)$, where $\Omega(b, \Delta) = \{z : z \in \{x_i\}_{i=1}^n, \inf_{x \in \Delta} |z - x| > b\}$. Identify the initial knots and add them to Δ.

Step 2. Repeat step 1 for $l = 1$ and replace b by rb for some $r > 1$. Go to step 1 if $b < (x_n - x_1)/5$. Otherwise, stop.

In this algorithm, b is chosen from a geometric grid with $r > 1$. The choice of r controls the computational cost for knot initialization and the selected initial knots. A larger value of r yields smaller computational cost but usually fewer initial knots. Our limited experience shows that $r = 1.2$ provides a satisfactory compromise.

5. EFFECTS OF KNOT ADDITION AND LOCAL SPLINE APPROXIMATION

5.1 Effects of Knot Addition

In this section we provide some theory to make the statement about the effect of adding a new knot in Section 3 more precise. For simplicity, we consider only the case of $x_i = (i - 1)/n$ in Theorems 1 and 2.

Theorem 1. If $h_n = n \min_{0 \le i \le k+1} |t_i - t_{i-m}| \to \infty$ as $n \to \infty$, then there exist constants $c_m > 0$ and $\gamma_m \in (0, 1)$ (depending only on m) such that for any new knot $t^* \in (t_{i^*}, t_{i^*+1})$, $i^* = 0, \ldots, k$,

$$|\hat{f}(x; \mathbf{t}) - \hat{f}(x; \mathbf{t}^*)| \le c_m \gamma_m^{|i^* - j|} g_n(i^*; \mathbf{t}) \quad \text{if } x \in [t_j, t_{j+1}], \quad (7)$$

where $\mathbf{t}^* = (t_0, \ldots, t_{i^*}, t^*, t_{i^*+1}, \ldots, t_{k+1})$ and

$$g_n(i^*; \mathbf{t}) = h_n^{-1}(t_{i^*+1} - t_{i^*})^{m-2} \max(D_{\hat{f}}(t_{i^*}), D_{\hat{f}}(t_{i^*+1}))$$

with

$$D_{\hat{f}}(x) = \left| \hat{f}^{(m-1)}(x+; \mathbf{t}^*) - \hat{f}^{(m-1)}(x-; \mathbf{t}^*) \right|,$$

and $\hat{f}^{(m-1)}(x-; \mathbf{t}^*)$ and $\hat{f}^{(m-1)}(x+; \mathbf{t}^*)$ are the left and right $(m-1)$th degree derivatives of $\hat{f}(x; \mathbf{t}^*)$.

From (7), we see that $|\hat{f}(x; \mathbf{t}) - \hat{f}(x; \mathbf{t}^*)|$ decreases exponentially in $|i^* - j|$, the number of knots between x and t^*, which implies that the difference becomes small when there are more knots between x and the knot t^*. In other words, the addition of knot t^* has little effect on the value of the spline estimate at a distant point x in the sense that the number of knots between them is not small. Furthermore, the quantity $g_n(i^*; \mathbf{t}) = O(h_n^{-1}(t_{i^*+1} - t_{i^*})^{m-2}) \to 0$ as $n \to \infty$, which implies that the difference becomes even smaller when the sample size n is large.

5.2 Local Spline Regression

The evaluation of the model selector $R(\hat{f})$ given in (4) for any potential knot refinement (knot addition, relocation, or deletion) requires refitting the entire spline curve. But this is redundant, as each knot refinement changes essentially only a small part of the original fitted curve. Hence we use local spline regression in our procedure to approximate $R(\hat{f})$, to reduce computational cost. The local spline estimator $f(x; \tilde{t})$ $(x \in [t_{i_0}, t_{i_1}])$ is defined as the solution to the least squares problem

$$\min_{s(x;\tilde{t}) \in S(m,\tilde{t})} \sum_{x_j \in [t_{i_0},\, t_{i_1}]} (y_j - s(x_j; \tilde{t}))^2,$$

where $\tilde{t} = (t_{i_0}, \ldots, t_{i_1})$ $(i_0 < i_1)$ is a knot subsequence of $\mathbf{t}$ and

$$S(m, \tilde{t}) = \{s(x; \tilde{t}) \in C^{m-2}[t_{i_0}, t_{i_1}] : s(x; \tilde{t}) \text{ is a polynomial of degree } (m-1) \text{ on each subinterval } [t_i, t_{i+1}]; \; i = i_0, \ldots, i_1 - 1\}.$$

To examine the accuracy of the approximation of $R(\hat{f})$, one needs to know the difference between globally and locally

defined splines. Theorem 2 provides an upper bound for the difference between $\hat{f}(x;\mathbf{t})$ and the local fitted spline $\hat{f}(x;\mathbf{t}_{i_0})$, where $\mathbf{t}_{i_0} = (t_{i_0-M}, \ldots, t_{i_0+M+1})'$ for any $0 \le i_0 \le k$.

Theorem 2. Under the same conditions as in Theorem 1, for a fixed integer $M > 0$ and any $0 \le j \le M$, we have

$$|\hat{f}(x;\mathbf{t}) - \hat{f}(x;\mathbf{t}_{i_0})| < c_m^* \mu_{i_0+j,m} \gamma_m^{|j-M|}, \quad \text{if } x \in [t_{i_0+j}, t_{i_0+j+1}] \cup [t_{i_0-j}, t_{i_0-j+1}],$$

where $c_m^* > 0$ and $\gamma_m \in (0,1)$ are constants, and

$$\mu_{i_0+j,m} = \left(\frac{\max(t_{i_0-M} - t_{i_0-M-m+1}, t_{i_0+M+m} - t_{i_0+M+1})}{\min_{1\le u\le m}(t_{i_0+j+u} - t_{i_0+j+u-m})} \right)^{1/2} \times \max(H_L, H_R)$$

with

$$H_L = \max_{i_0-M\le i<i_0-M+m} \left| (n(t_{i_0-M} - t_{i-m}))^{-1} \times \sum_{x_j \in [t_{i_0-M},\, t_{i-m}]} N_i(x_j)(y_j - \hat{f}(x_j;\mathbf{t})) \right|$$

and

$$H_R = \max_{i_0+M+1<i\le i_0+M+m+1} \left| (n(t_i - t_{i_0+M}))^{-1} \times \sum_{x_j \in [t_i,\, t_{i_0+M}]} N_i(x_j)(y_j - \hat{f}(x_j;\mathbf{t})) \right|.$$

Theorem 2 says that the difference $\hat{f}(x;\mathbf{t}_{i_0}) - \hat{f}(x;\mathbf{t})$ decreases exponentially in $|j - M|$, the number of knots between x and the nearest boundary point of $[t_{i_0-M}, t_{i_0+M+1}]$. Hence for sufficiently large M, the difference is small if x stays away from the boundary of the interval $[t_{i_0-M}, t_{i_0+M+1}]$ in the sense that both $[t_{i_0-M}, x]$ and $[x, t_{i_0+M+1}]$ contain some knots. Although it seems artificial, $\mu_{i_0+j,m}$ depends on a mesh ratio, as in a similar conjecture made by de Boor (1979). $\max(H_L, H_R)$ will become smaller when more knots are added because the residuals $y_j - \hat{f}(x_j;\mathbf{t})$ tend to be more locally orthogonal to spline functions under least squares projection. Consequently, the approximation will be more precise as knot selection process continues.

Here γ_m can be computed using a result of Huang (1990) when the knots are equally spaced and $m = 2, 3, 4$. The result for a general m can be derived similarly. Without any restriction on the knot sequence $\mathbf{t}$, an upper bound on γ_m can be obtained, but it may be close to 1.

5.3 Local Approximation of the Model Selector

By Theorems 1 and 2, the model selector can be approximated using locally fitted splines to reduce the computational cost. After some simple algebra, we obtain

$$R(\hat{f}(\cdot;\mathbf{t})) - R(\hat{f}(\cdot;\mathbf{t}^*)) = \sum_{i=1}^{n} (\hat{f}(x_i;\mathbf{t}) - \hat{f}(x_i;\mathbf{t}^*))^2/n - 2\sigma^2/n,$$

which is used to decide whether t^* should be added. In traditional knot selection schemes, both $R(\hat{f}(\cdot;\mathbf{t}))$ and $R(\hat{f}(\cdot;\mathbf{t}^*))$ are computed, and hence they are computationally expensive for large k. By Theorem 1, we know that $\hat{f}(x;\mathbf{t}) - \hat{f}(x;\mathbf{t}^*)$ is small when x is not close to t^*, and this

$$R(\hat{f}(\cdot;\mathbf{t})) - R(\hat{f}(\cdot;\mathbf{t}^*)) \approx \sum_{x_i \in I'_{i^*}} (\hat{f}(x_i;\mathbf{t}) - \hat{f}(x_i;\mathbf{t}^*))^2/n - 2\sigma^2/n,$$

where $I'_{i^*} = [t_{i^*-M'}, t_{i^*+M'+1}]$ and $M' > 0$ is a constant. By Theorem 2, for any $x \in [t_{i^*-M'}, t_{i^*+M'+1}]$ and $M > M'$, $\hat{f}(x;\mathbf{t})$ and $\hat{f}(x;\mathbf{t}^*)$ can be approximated by the locally fitted splines $\hat{f}(x;\mathbf{t}_{i^*})$ and $\hat{f}(x;\mathbf{t}^*_{i^*})$, where $\mathbf{t}^*_{i^*} = (t_{i^*-M}, \ldots, t_{i^*}, t^*, t_{i^*+1}, \ldots, t_{i^*+M+1})'$. Therefore,

$$R(\hat{f}(\cdot;\mathbf{t})) - R(\hat{f}(\cdot;\mathbf{t}^*)) \approx \sum_{x_i \in I'_{i^*}} (\hat{f}(x_i;\mathbf{t}_{i^*}) - \hat{f}(x_i;\mathbf{t}^*_{i^*}))^2/n - 2\sigma^2/n.$$

Furthermore, by Theorem 1, $\hat{f}(x;\mathbf{t}_{i^*}) - \hat{f}(x;\mathbf{t}^*_{i^*})$ is small if $x \notin I'_{i^*}$. Hence

$$R(\hat{f}(\cdot;\mathbf{t})) - R(\hat{f}(\cdot;\mathbf{t}^*)) \approx \sum_{x_i \in [t_{i^*-M},\, t_{i^*+M+1}]} (\hat{f}(x_i;\mathbf{t}_{i^*}) - \hat{f}(x_i;\mathbf{t}^*_{i^*}))^2/n - 2\sigma^2/n,$$

which says that the locally fitted splines can be used to decide whether knots should be added. Therefore, the amount of computation is significantly reduced when $k \gg M$.

In implementation, M must to be chosen as small as possible to ease computation and to ensure the accuracy of the local spline approximation. Our limited experience suggests that $M = 7$ suffices for a cubic spline ($m = 4$).

6. SIMULATIONS

In this section we investigate the performance of our spline procedure and the new knot selection scheme via simulations. In particular, we compare the new knot selection scheme with the stepwise addition and deletion scheme in regression spline estimation, and compare our spline procedure with five competing function estimation procedures: hybrid adaptive splines (HAS) proposed by Luo and Wahba (1997); smoothing splines (e.g, Wahba 1990); —— adaptive regression splines (MARS), considered by Friedman (1991); and the wavelet shrinkage method of Donoho and Johnstone (1995). To make a fair comparison, we consider examples used by these researchers to illustrate the effectiveness of their methods. In these examples, the test functions have various smoothness properties, and the degrees of smoothness of the functions are considered unknown. In the simulations, cubic splines ($m = 4$) are used and the root mean squared error (MSE), $\sum_{i=1}^{n}(f(x_i) - \hat{f}(x_i))^2/n$, and the square root of MSE (RMSE) are used to measure the performance. The simulations use the C language, and the corresponding figures are generated from the S package. For convenience, here and in the sequel we call our procedure *spatially adaptive regression splines* (SARS).

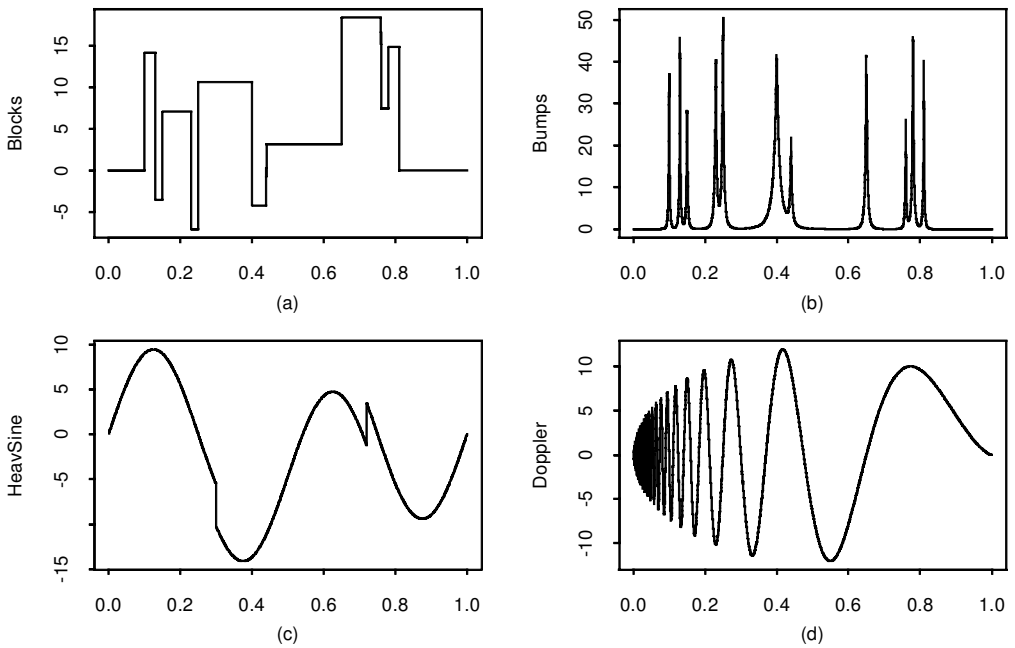

Figure 3. Four Inhomogeneous Functions: (a) Blocks; (b) Bumps; (c) HeaviSine; (d) Doppler.

Simulation 1: Comparison With the Stepwise Knot Addition and Deletion Scheme. Many well-known spline procedures based on the stepwise knot addition and/or deletion, such as MARS, restrict the possible knot locations to ease the computational burden, which may dramatically affect the spline procedure's performance in spatial adaptation. Hence, to locate knots more accurately when using the stepwise scheme, we relax the restriction and allow knots to be placed at any design point.

To compare our new knot selection scheme with the stepwise scheme, we use four examples (Blocks, Bumps, HeaviSine, and Doppler) of Donoho and Johnstone (1995) with sample sizes $n = 2^i$, $i = 7, \ldots, 11$, $\sigma = 1$ and the signal-to-noise ratio $\mathrm{SD}(f)/\sigma - 7$. Figures 3 and 4 display the test functions and their noisy versions. Table 1 summarizes the simulation results, with the MSEs based on 100 runs.

On average, SARS has smaller MSE than the stepwise spline procedure in all examples except the Doppler example,

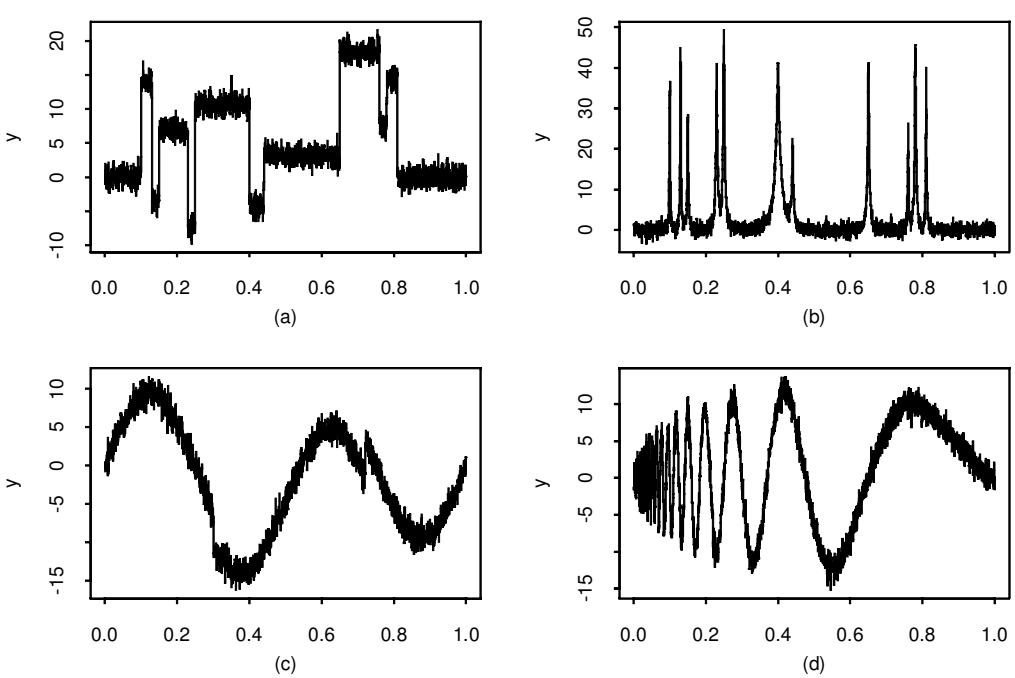

Figure 4. Noisy Version of the Test Functions ($n = 2{,}048$, $\sigma = 1$, Signal-to-Noise Ratio SD(f)/$\sigma = 7$): (a) Blocks; (b) Bumps; (c) HeaviSine; (d) Doppler.

Table 1. Average MSE of the Stepwise Spline Procedure and SARS Based on 100 Simulation Runs

	n	Stepwise	SARS
Blocks	128	.9102	.7683
	256	.6506	.4620
	512	.4254	.2638
	1,024	.2487	.1505
	2,048	.1343	.0824
Bumps	128	1.799	1.189
	256	1.032	.4601
	512	.2845	.2670
	1,024	.1949	.1637
	2,048	.1219	.0918
HeaviSine	128	.3652	.3356
	256	.2289	.1721
	512	.1255	.0948
	1,024	.0649	.0453
	2,048	.0363	.0254
Doppler	128	.5919	.4517
	256	.5138	.4362
	512	.2350	.2186
	1,024	.1297	.1270
	2,048	.0641	.0622

in which the MSEs are close to one another. Detailed examination reveals that adding redundant knots in the neighborhood of origin may actually improve a procedure's performance which explains why the stepwise scheme performs reasonably well. As a matter of fact, the stepwise procedure performs better than HAS, smoothing splines, and the wavelets in these examples, as to be seen from Table 3. However, the disadvantage of the stepwise spline procedure is that its computational complexity grows more than quadratically with the sample size. In contrast, the computational complexity of SARS is of order $n \cdot \log^2(n)$, based on a rough analysis of our algorithm. This analysis agrees with our numerical experience with SARS, although we have not yet rigorously proved the rate. For instance, on average, the computational time needed for a single run of SARS in these examples is approximately $n/1{,}000$ minutes for $n \leq 2^{17} = 131{,}072$ on a Sun Ultra10 workstation.

Simulation 2: Comparison With Other Spline Methods. To compare SARS with the HAS of Luo and Wahba (1997), smoothing splines, and MARS, we use examples considered by Luo and Wahba (1997, table 1); see Table 2 for the particular functions. We did not include example 2 from Luo and Wahba, because the true function in that example does not agree with the one of Donoho and Johnstone (1995) as claimed.

Table 3, reports the simulation results for SARS based on 31 runs for Examples 1, 3–5 and based on 400 runs for Examples 6–7. Clearly, SARS performed at least as well as the spline competitors in all of these examples, and significantly better than the spline competitors in Examples 1, 3, 4, and 5, where the test functions are not smooth. Furthermore, SARS is numerically more stable than HAS in all six examples.

In Examples 6 and 7, the test functions are smooth (Figs 5 and 5) and therefore the exact locations of knots are much less important. As a result, SARS did not perform much better than these spline competitors in these examples.

Table 2. Formulas for Test Functions Used in Simulation 2

Example	f(x)	σ	n	No. of runs
1	Blocks*3.5	1.0	2,048	31
3	HeaviSine*2.2	1.0	2,048	31
4	Doppler*22	1.0	2,048	31
5	Doppler*22	exp(x)/1.648	2,048	31
6	$\sin(2(4x-2)) + 2\exp(-256(x-.5)^2)$	.3	256	400
7	$(4x-2) + 2\exp(-256(x-.5)^2)$	.4	256	400

Note that in Examples 1, 3, 4, and 5, MARS gives unsatisfactory results. Its poor performance is not due to the stepwise knot addition and deletion method, because the stepwise spline procedure in Simulation 1 performed very well in these examples. We believe that the reason is the restriction on knot placement, which significantly reduces the spatial adaptability of MARS.

Simulation 3: Comparison With Smoothing Splines in a Simple Example. In Simulation 2, the MSEs of SARS and smoothing splines are very close in Examples 6 and 7, where the test functions are quite smoother than the ones in Examples 1–5. This is not surprising, because spatial adaptation is much less important in estimating these smooth functions. In such a situation, SARS may not have the advantage if no spatial adaptation is necessary. For illustration, we consider the test function

$$f(x) = 4 * \sin(2 * \pi * x)$$

and $\sigma = 1$. Figure 7 suggests that a nonadaptive smoothing procedure performs well in this example. Table 4 shows that smoothing splines performs slightly better than SARS based on 400 runs with sample sizes $n = 2^i$, $i = 7, \ldots, 14$. Here the selection error due to knot selection affects performance slightly when knot selection is less important, which is the price that must be paid for spatial adaptation.

Simulation 4: Comparison With the Wavelet Shrinkage Method. To compare with the wavelet shrinkage method of Donoho and Johnstone (1995), we use the same examples as in Simulation 1. The simulation results based on 100 runs are summarized in Figure 8 and 9.

Table 3. Median of MSE and the Interquartile Range of MSE (in parentheses)

Example	SARS	HAS	SS	MARS
1	.081(.010)	.137(.018)	.546(.023)	
3	.025(.011)	.039(.013)	.062(.007)	.150(.014)
4	.064(.013)	.068(.015)	.205(.011)	
5	.049(.013)	.100(.072)	.232(.014)	
6	.005(.003)	.007(.006)	.006(.003)	.007(.004)
7	.009(.006)	.012(.011)	.010(.005)	.012(.007)

NOTE: The results for HAS, SS, and MARS are reproduced from Table 2 of Luo and Wahba (1997). The missing entries in the column MARS are greater than 6.

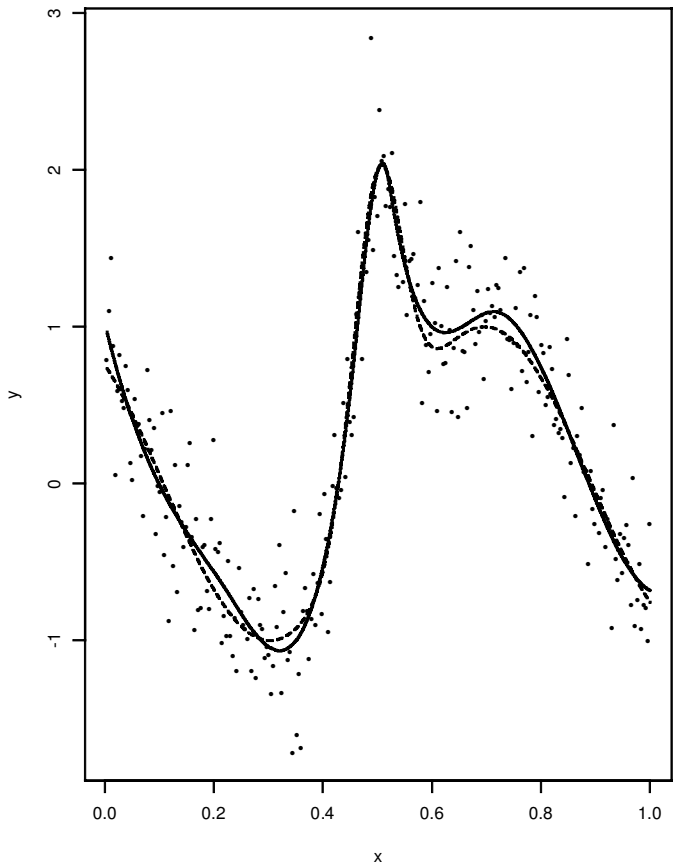

Figure 5. Example 6: Estimated Function With Median MSE (solid line), True Curve (dotted line), and a Copy of Simulated Data (scatterplot).

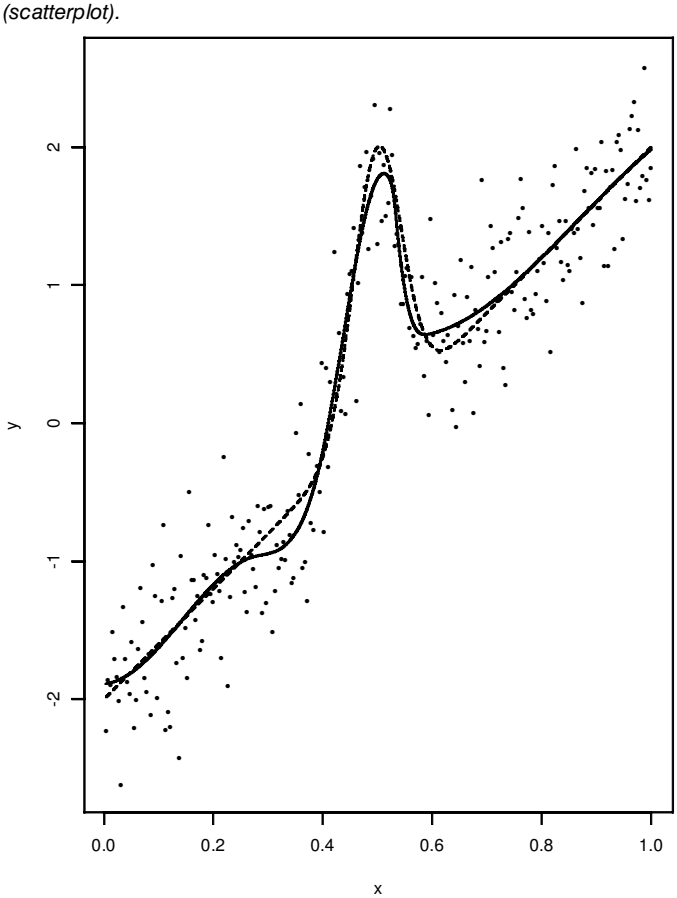

Figure 6. The Same as in Figure 5, But for Example 7.

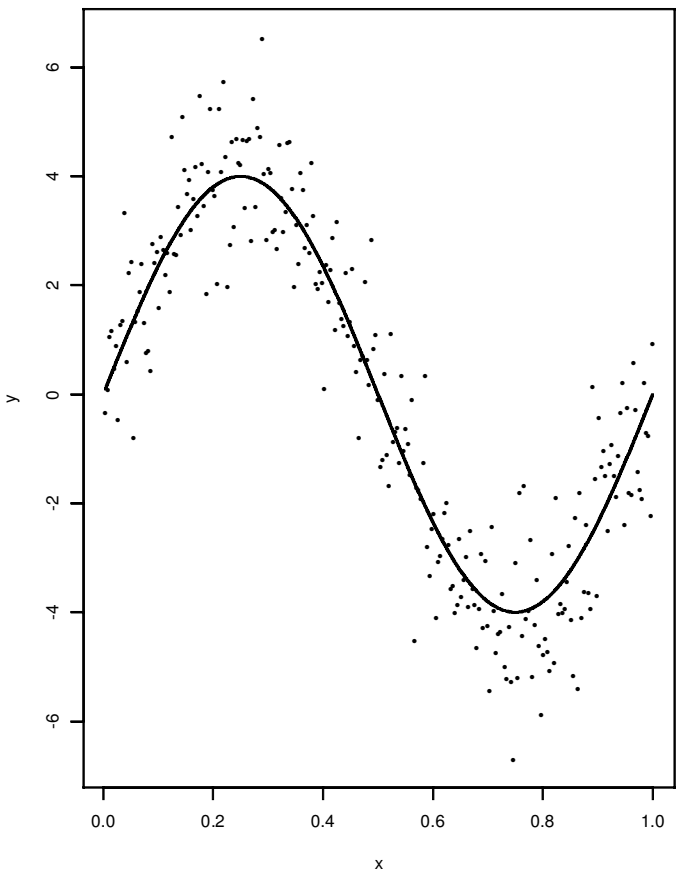

Figure 7. Test Function in Simulation 3 and a Copy of Simulated Data.

In all of these examples, for sample size $n > 2^7 = 128$, SARS outperformed, (in terms of RMSE), the best performance of wavelet shrinkage estimates based on a class of wavelets (Haar, DB4, Ciof3, Symm8, ThrC3, ThrS8, and VisS8) using various thresholding techniques (SureShrink, RiskShrink, and VisuShrink) reported in table 2 of Donoho and Johnstone (1995). In fact, Figure 8 shows that SARS yields a much smaller RMSE relative to those of wavelet shrinkage estimates as the sample size n increases. This suggests that, at least for these four types of functions, the rate of convergence for SARS is actually faster than that of the wavelet estimate, although Donoho and Johnstone (1995) showed that in these cases, the wavelet shrinkage estimate converges at an optimal rate in a minimax sense. In addition, boxplots of RMSEs of SARS in Figure 9 shows that SARS is numerically stable, and Figure 6 suggests that the visual quality of SARS is also higher than that of the wavelet estimates.

Table 4. The RMSE of SARS and Smoothing Splines Based on 400 Runs in Simulation 3

Sample size (n)	*128*	*256*	*512*	*1,024*	*2,048*	*4,096*	*8,192*	*16,384*
SARS	.27	.20	.15	.109	.076	.055	.039	.0287
Smooth Splines	.25	.18	.13	.094	.067	.050	.036	.0266

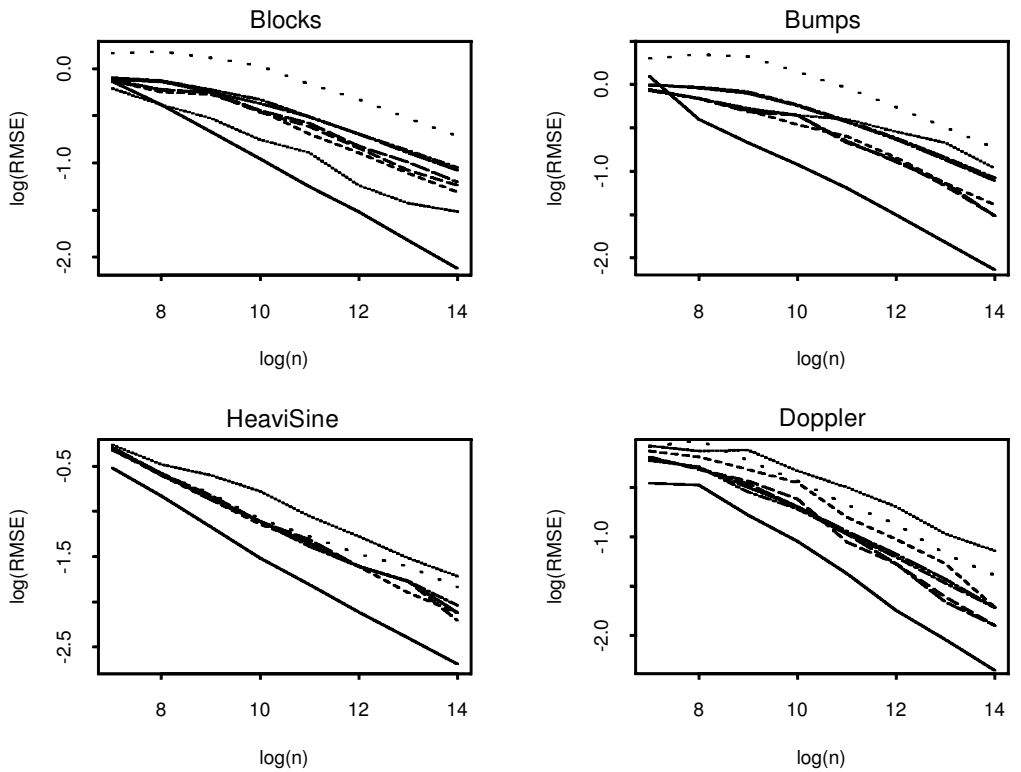

Figure 8. Comparison of the Logarithm of the Root Mean Squared Estimate (RMSE), of SARS (solid line). Based on 100 runs, with those of wavelet shrinkage estimates (Haar, DB4, Ciof3, Symm8, ThrC3, ThrS8, VisS8) reported in table 2 of Donoho and Johnstone (1995). Plotted against $\log_2(n)$ for $n = 2^i$, $7 \le i \le 14$.

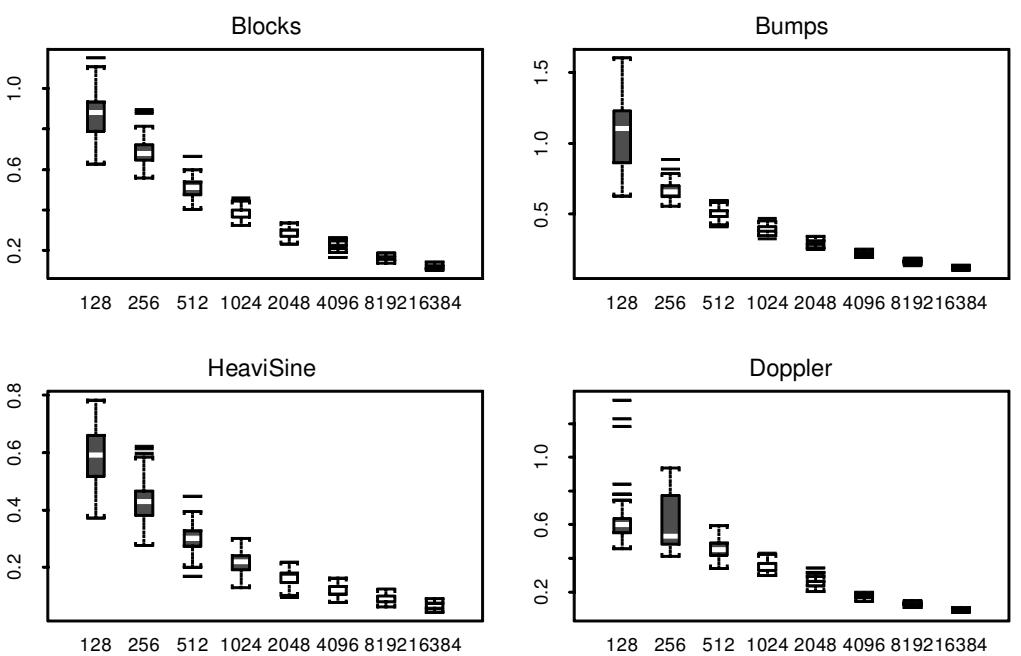

Figure 9. Boxplots of RMSEs of SARS Based on 100 Runs for Sample Sizes $n = 2^i$, $7 \le i \le 14$.

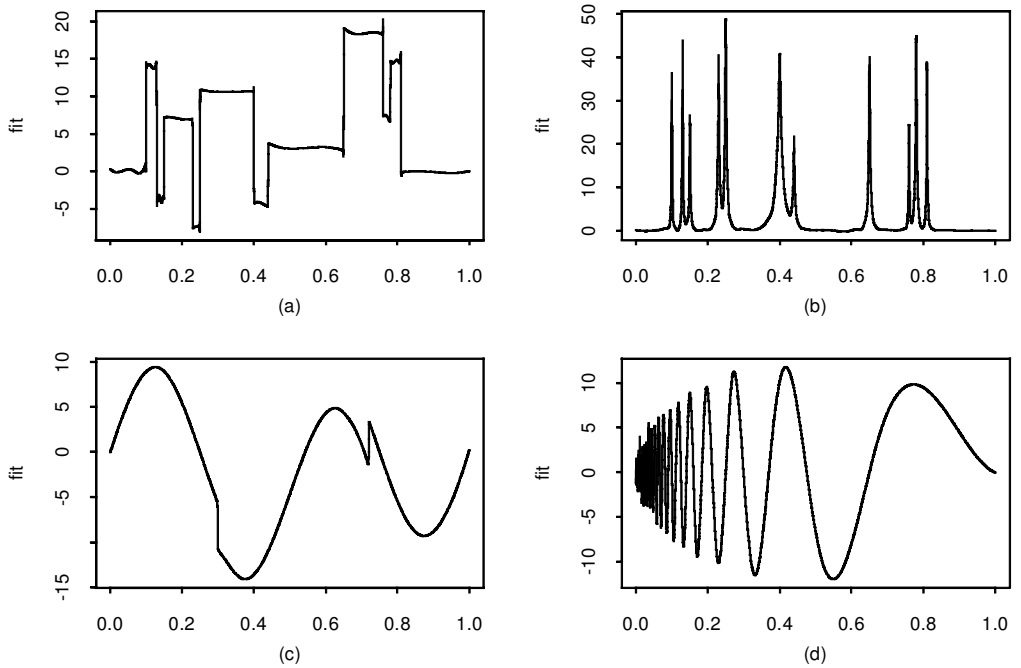

Figure 10. Four Estimated Functions: (a) Blocks; (b) Bumps; (c) HeaviSine; (d) Doppler.

7. APPLICATION TO SPEECH SIGNAL RECOVERY

To illustrate an application of the proposed method, we consider the speech signal presented in Figure 1. Physical modeling suggests that many recorded voices are characterized by harmonic and additive noise signals. In this situation, our goal is to recover the true voice signal $f(x)$ from its noisy background using recorded voice signals $\{Y_i\}_{i=1}^n$ and the corresponding sampling times $\{T_i\}_{i=1}^n$. The sample size for the signal recording is $n = 7,000$, and it takes about 3 minutes for SARS to produce the recovered signal. The number of selected knots is about 1,400. Figures 11(a) and 11(b) present the noisy version of the signal and the recovered signal using SARS. By visually comparing these two figures, we conclude that the proposed procedure does well and the recovered signal captures the basic features of the true speech signal. Meanwhile, we point out that SARS produced an oversmoothed curve in an interval around .7 due to the fact that the noise level in the interval is very high relative to the signal in the interval.

8. CONCLUSION

This article proposed an adaptive spline procedure based on a new knot selection scheme for nonparametric regression. The proposed procedure uses certain special local properties of spline function in knot selection and thus overcomes the knot confounding problem and high computational complexity in adaptive estimation encountered by spline procedures based on the traditional knot addition and deletion method. Simulation studies indicate that this new procedure has high spatial adaptability and is computationally efficient. Hence it is potentially useful in a variety of applications. In addition, the proposed knot selection methodology may be applicable to other spline estimation problems, such as density estimation and logistic regression.

APPENDIX: PROOFS

We first prove two useful technical lemmas. Lemma A.1 says that the spline estimate $\hat{f}(x; \mathbf{t}^*)$ with the new knot t^* can be well approx-

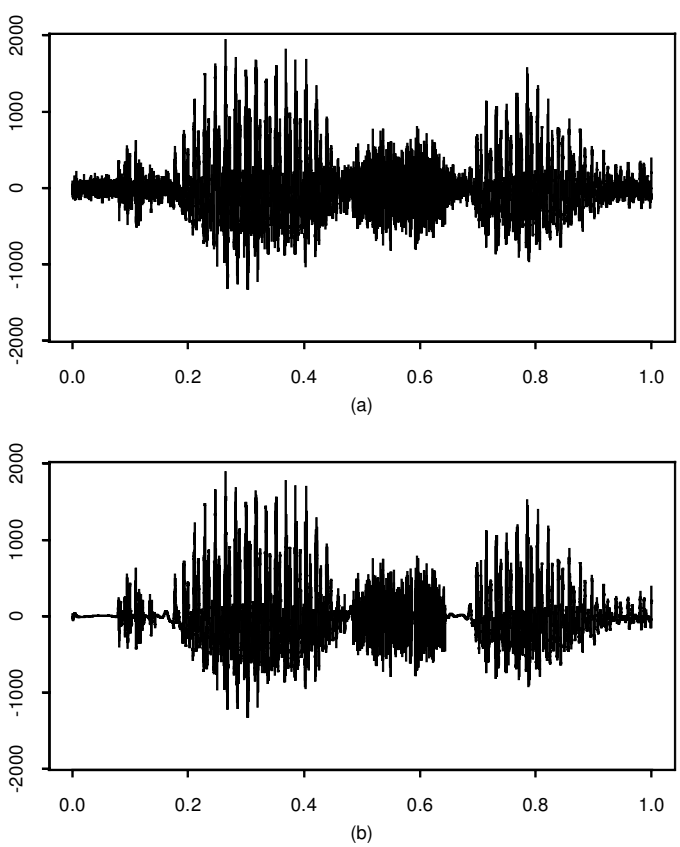

Figure 11. Noisy Version of Speech Signal and Recovered Speech Signal. (a) Noisy speech signal with the white noise, with a the signal-to-noise ratio of SD(f)/$\sigma = 5$; (b) recovered speech signal.

imated by a spline function $s(x)$ without it. The approximation error is 0 outside an interval $[t_{i^*}, t_{i^*+1}]$ and small inside the interval, which implies that the difference between $\hat{f}(x; \mathbf{t})$ and $\hat{f}(x; \mathbf{t}^*)$ is small for x outside the interval. Lemma A.2 says that its off-diagonal elements decay exponentially, which characterizes elements of the inverse of the Gram matrix $\mathbf{G}$. On the basis of these lemmas and the property of projection estimators, Theorem A.2 is proved after some direct calculations.

Lemma A.1. There exists an $s(x) \in S(m, \mathbf{t})$ such that $s(x) = \hat{f}(x; \mathbf{t}^*)$ for $x \notin (t_{i^*}, t_{i^*+1})$. Consequently,

$$\left\|s(x) - \hat{f}(x; \mathbf{t}^*)\right\|_{L_\infty[t_{i^*}, t_{i^*+1}]} \le c_0 (t_{i^*+1} - t_{i^*})^{m-1} \times \max(D_{\hat{f}}(t_{i^*}), D_{\hat{f}}(t_{i^*+1})),$$

where c_0 is a constant and $D_{\hat{f}}(\cdot)$ is defined in Theorem 1.

Proof of Lemma A.1. Let $\mathbf{N}(x) = (N_1(x), \ldots, N_k(x))'$ and $\mathbf{N}^*(x) = (N_1^*(x), \ldots, N_{k+1}^*(x))'$ be the normalized B-spline functions associated with knot sequences $\mathbf{t}$ and $\mathbf{t}^*$. Because $\hat{f}(x; \mathbf{t}^*) \in S(m, \mathbf{t}^*)$, $\hat{f}(x; \mathbf{t}^*)$ can be written as $(\mathbf{a}^*)'\mathbf{N}^*(x)$ for some $\mathbf{a}^* = (a_1^*, \ldots, a_{k+1}^*)'$. To construct $s(x) = \mathbf{a}'\mathbf{N}(x)$, we let $a_i = a_{i+1}^*$ for $i = 1, \ldots, i^* - m$, and let $a_i = a_i^*$ for $i = i^* + m + 1, \ldots, k$. Furthermore, for $i = i^* - m + 1, \ldots, i^* + m$, a_i is chosen such that $s(x)$ interpolates $\hat{f}(x; \mathbf{t}^*)$ at t_i, which can be obtained by solving a set of linear equations.

To show that $s(x)$ is the desired spline function, we use the fact that B-splines are locally defined; for example, the support of $N_i(x)$ is $[t_{i-m}, t_i]$. It then follows that $N_i(x) = N_i^*(x)$ for $i = 1, \ldots, i^* - m$ and $N_i(x) = N_{i+1}^*(x)$ for $i = i^* + m + 1, \ldots, k$. Hence $s(x)$ is such that $s(x) = \hat{f}(x; \mathbf{t}^*)$ for any $x \notin (t_{i^*-m}, t_{i^*+m})$.

Now it suffices to show that $s(x) = \hat{f}(x; \mathbf{t}^*)$ for $x \in [t_{i^*-m}, t_{i^*}] \cup [t_{i^*+1}, t_{i^*+m}]$. Because both $s(x)$ and $\hat{f}(x; \mathbf{t}^*)$ are $(m-1)$th degree polynomials in $[t_{i^*-m}, t_{i^*-m+1}]$, and moreover, $s^{(j)}(t_{i^*-m}) = \hat{f}^{(j)}(t_{i^*-m}; \mathbf{t}^*)$ for any $j = 0, 1, \ldots, m-2$, and $s(t_{i^*-m+1}) = \hat{f}(t_{i^*-m+1}; \mathbf{t}^*)$ by the definition of $s(x)$, it follows that $s(x) = \hat{f}(x; \mathbf{t}^*)$ for $x \in [t_{i^*-m}, t_{i^*-m+1}]$. Similarly, we can obtain $s(x) = \hat{f}(x; \mathbf{t}^*)$ for $x \in [t_{i^*-m+1}, t_{i^*-m+2}]$. By induction, $s(x) = \hat{f}(x; \mathbf{t}^*)$ for any $x \in [t_{i^*-m}, t_{i^*}] \cup [t_{i^*+1}, t_{i^*+m}]$. This completes the proof of Lemma A.1.

Lemma A.2. There exist constants c_1 and $\gamma_m \in (0, 1)$ (depending only on m) such that

$$|\alpha_{i,j}| \le c_1 \gamma_m^{|i-j|} \{(t_i - t_{i-m})(t_j - t_{j-m})\}^{-1/2},$$

where $\mathbf{G}$ is the gram matrix defined in (3) and $\alpha_{i,j}$ is the (i, j)th element of $\mathbf{G}^{-1}$.

Proof of Lemma A.2. For any $\mathbf{a} \in R^k$, after some simple algebra we have that

$$\mathbf{a}'\mathbf{G}\mathbf{a} = \int_0^1 \left(\mathbf{a}'\mathbf{N}(x)\right)^2 dQ_n(x) = \int_0^1 s^2(x)\, dQ_n(x),$$

where $s(x) = \mathbf{a}'\mathbf{N}(x)$ and $Q_n(x)$ is the empirical distribution function of x_i. Because $\{x_i\}_{i=1}^n$ are equally spaced, $Q_n(x)$ converges to the uniform distribution on [0, 1]. Hence, after straightforward calculations, we obtain

$$\int_0^1 s^2(x)\, dQ_n(x) = (1 + o(1)) \int_0^1 s^2(x)\, dx.$$

Let $\tilde{\mathbf{G}} = \mathbf{E}^{1/2}\mathbf{G}\mathbf{E}^{1/2}$, where $\mathbf{E}$ is a diagonal matrix with (i, i)th element given by $m/(t_i - t_{i-m})$. By lemma 1 of de Boor (1976), there exist constants c_2 and $\gamma_m \in (0, 1)$ such that the (i, j)th element of $\tilde{\mathbf{G}}^{-1}$ is bounded above by $c_2 \gamma_m^{|i-j|}$. The desired result then follows from the fact that $\mathbf{G}^{-1} = \mathbf{E}^{-1/2}\tilde{\mathbf{G}}^{-1}\mathbf{E}^{-1/2}$.

Proof of Theorem 1

By the orthogonal property of least squares estimation, we have

$$\sum_{i=1}^n (y_i - \hat{f}(x_i; \mathbf{t}))^2 = \sum_{i=1}^n (y_i - \hat{f}(x_i; \mathbf{t}^*))^2 + \sum_{i=1}^n (\hat{f}(x_i; \mathbf{t}^*) - \hat{f}(x_i; \mathbf{t}))^2.$$

Because $\hat{f}(x; \mathbf{t})$ is the least squares estimator among all spline functions in $S(m, \mathbf{t})$, it follows after some simple algebra that

$$\hat{f}(x; \mathbf{t}) = \mathbf{N}(x)\mathbf{G}^{-1}\mathbf{X}\hat{\mathbf{f}}_n(\mathbf{t}^*), \tag{A.1}$$

where $\hat{\mathbf{f}}_n(\mathbf{t}^*) = (\hat{f}(x_1; \mathbf{t}^*), \ldots, \hat{f}(x_n; \mathbf{t}^*))'$. By Lemma A.1, there exists an $s(x) \in S(m, \mathbf{t})$ such that $s(x) = \hat{f}(x; \mathbf{t}^*)$ for any $x \notin (t_{i^*}, t_{i^*+1})$. By (A.1),

$$\hat{f}(x; \mathbf{t}) - s(x) = \mathbf{N}(x)\mathbf{G}^{-1}\mathbf{X}(\hat{\mathbf{f}}_n(\mathbf{t}^*) - \mathbf{s}_n) = \mathbf{N}(x)\mathbf{G}^{-1}\mathbf{D}, \tag{A.2}$$

where $\mathbf{s}_n = (s(x_1), \ldots, s(x_n))'$ and $\mathbf{D} = (D_1, \ldots, D_k)$ with

$$D_i = \sum_{j=1}^n N_i(x_j)(\hat{f}(x_j; \mathbf{t}^*) - s(x_j))/n.$$

By Lemma A.1, we have that $D_i = 0$ for $i = 1, \ldots, i^* - m$, $i^* + m, \ldots, k$ and

$$|D_i| \le c_0 n^{-1} (t_{i^*+1} - t_{i^*})^{m-2} \max(D_{\hat{f}}(t_{i^*}), D_{\hat{f}}(t_{i^*+1})),$$

for any $i = i^* - m + 1, \ldots, i^* + m$. Note that the support of $N_j(x)$ is $[t_{j-m}, t_j]$ and $N_j(x) \le 1$ for any j and x. By (A.2) and Lemma A.1, we obtain that for any $j \ne i^*$,

$$|\hat{f}(x; \mathbf{t}) - \hat{f}(x; \mathbf{t}^*)| \le \left(\sum_{l=0}^m \sum_{u=0}^m |\alpha_{j+l, i^*+u}|\right) c_0 n^{-1} (t_{i^*+1} - t_{i^*})^{m-2} \times \max(D_{\hat{f}}(t_{i^*}), D_{\hat{f}}(t_{i^*+1})),$$

where $\alpha_{l,u}$ is the (l, u)th element of $\mathbf{G}^{-1}$. Then the desired result follows from Lemma A.2.

The following lemma is necessary for the proof of Theorem 2.

Lemma A.3. For any function $s_{i_0}(x) \in S(m, \mathbf{t}_{i_0})$, there exists an $s_{i_0}^*(x) \in S(m, \mathbf{t})$ such that $s_{i_0}^*(x) = s_{i_0}(x)$ for $x \in I_{i_0} = [t_{i_0-M}, t_{i_0+M+1}]$ and $s_{i_0}^*(x) = \hat{f}(x; \mathbf{t})$ for $x \notin \tilde{I}_{i_0}$.

Proof of Lemma A.3. Let $\mathbf{a}$ be the vector such that $\hat{f}(x; \mathbf{t}) = \mathbf{a}\mathbf{N}(x)$ and let $\mathbf{e}_i = (e_1, \ldots, e_{2M+m})$ be such that $s_{i_0}(x) = \sum_{u=1}^{2M+m} e_u N_{u+i_0-M}(x)$. We now construct $s_{i_0}^*(x) = (\mathbf{a}^*)\mathbf{N}(x)$ by setting $a_i^* = a_i$ for $i = 1, \ldots, i_0 - M$, $i_0 + M + m + 1, \ldots, k$, and $a_i^* = e_{i-i_0+M}$ for $i = i_0 - M + 1, \ldots, i_0 + M + m$. The desired result then follows from the fact that the support of $N_i(x)$ is $[t_{i-m}, t_i]$ for any i.

Proof of Theorem 2

For any $s_{i_0}(x) \in S(m, \mathbf{t}_{i_0})$, let $s_{i_0}^*(x) \in S(m, \mathbf{t})$ be the spline function as defined in Lemma A.3. Since $\hat{f}(x; \mathbf{t})$ is the least squares estimator, we have

$$\sum_{i=1}^n (y_i - \hat{f}(x_i; \mathbf{t}))^2 = \sum_{i=1}^n (y_i - s_{i_0}^*(x))^2 - \sum_{i=1}^n (\hat{f}(x_i; \mathbf{t}) - s_{i_0}^*(x))^2.$$

Let $\Delta = \sum_{x_i \in I_{i_0}} (y_i - s_{i_0}(x_i))^2$. It then follows from Lemma A.3 that

$$\begin{aligned}
\Delta &= \sum_{x_i \in I_{i_0}} (y_i - s^*_{i_0}(x_i))^2 = \sum_{i=1}^{n} (y_i - s^*_{i_0}(x_i))^2 - \sum_{x_i \notin I_{i_0}} (y_i - s^*_{i_0}(x_i))^2 \\
&= \sum_{x_i \in \tilde{I}_{i_0}} \{(\hat{f}(x_i; \mathbf{t}) - s^*_{i_0}(x_i))^2 + (y_i - \hat{f}(x_i; \mathbf{t}))^2\} - \sum_{x_i \in I^o_{i_0}} (y_i - s^*_{i_0}(x_i))^2 \\
&= \sum_{x_i \in \tilde{I}_{i_0}} \{s_f^2(x_i) + (y_i - \hat{f}(x_i; \mathbf{t}))^2\} - \sum_{x_i \in I^o_{i_0}} (y_i - \hat{f}(x_i; \mathbf{t}) + s_f(x_i))^2 \\
&= \sum_{x_i \in \tilde{I}_{i_0}} (y_i - \hat{f}(x_i; \mathbf{t}))^2 + \sum_{x_i \in I_{i_0}} s_f^2(x_i) \\
&\quad - \sum_{x_i \in I^o_{i_0}} s_f(x_i)(y_i - \hat{f}(x_i; \mathbf{t})), \qquad \text{(A.3)}
\end{aligned}$$

where $s_f(x) = \hat{f}(x; \mathbf{t}) - s^*_{i_0}(x)$ and $\tilde{I}_{i_0} = [t_{i_0 - M - m}, t_{i_0 + M + m + 1}]$. Because $\hat{f}(x; \mathbf{t}_{i_0})$ is the minimizer of $\sum_{x_i \in I_{i_0}} (y_i - s_{i_0}(x_i))^2$, $s_f(x)$ is the minimizer of (A.3) when $s^*_{i_0}(x) = \hat{f}(x; \mathbf{t}_{i_0})$. It then follows after some simple algebra that

$$\hat{f}(x; \mathbf{t}) - \hat{f}(x; \mathbf{t}_{i_0}) = \mathbf{N}'_{i_0}(x) \mathbf{G}^{-1}_{i_0} \mathbf{D}_{i_0}, \qquad \text{(A.4)}$$

where $\mathbf{N}'_{i_0}(x) = (N_{i_0 - M + 1}(x), \ldots, N_{i_0 + M + m}(x))'$; $\mathbf{G}_{i_0}$ is a $(2M + m) \times (2M + m)$ matrix with the (i, j)th element given by

$$\sum_{x_l \in I_{i_0}} N_{i + i_0 - M}(x_l) N_{j + i_0 - M}(x_l)/n;$$

and $\mathbf{D}_{i_0} = (D_{i_0 - M + 1}, \ldots, D_{i_0 + M + m})'$ with

$$D_i = \sum_{x_j \in (\tilde{I}_{i_0} - I_{i_0})} N_i(x_j)(y_j - \hat{f}(x_j; \mathbf{t}))/n \times \begin{cases} \le (t_{i_0 - M} - t_{i - m}) H_L & \text{if } i < i_0 - M + m \\ \le (t_i - t_{i_0 + M}) H_R & \text{if } i > i_0 + M + 1 \\ = 0 & \text{otherwise.} \end{cases}$$

Furthermore, it can be shown that the result in Lemma A.2 continues to hold for the elements of $\mathbf{G}^{-1}_{i_0}$. The desired result then follows from (A.4) and some simple algebra.

[Received July 1999. Revised February 2000.]

REFERENCES

Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data With Spline Functions: Estimating the Correct Degree of Smoothing by the Generalized Cross-Validation," *Numerische Mathematik.*, 31, 337–403.

de Boor, C. (1976), "A Bound on the L_∞-Norm of a L_2-Approximation by Splines in Terms of a Global mesh ratio," *Mathematrics of Computation*, 30, 765–771.

de Boor, C. (1978), *A Practical Guide to Splines*, New York: Springer.

Denison, D. G. T, Mallick, B. K., and Smith, A. F. M. (1998), "Automatic Bayesian Curve Fitting," *Journal of the Royal Statistical Society*, Series B, 60, 333–350.

Donoho, D., and Johnstone, I. M. (1994), "Ideal Spatial Adaption by Wavelet Shrinkage," *Biometrika*, 81, 425–455.

——— (1995), "Adapting to Unknown Smoothness via Wavelet Shrinkage," *Journal of the American Statistical Association*, 90, 1200–1224.

Fan, J. and Gijbels, I. (1995), "Data-Driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation," *Journal of the Royal Statistical Society*, Series B, 57, 371–394.

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

Friedman, J. H., and Silverman, B. W. (1989), "Flexible Parsimonious Smoothing and Additive Modeling" (with discussion), *Technometrics*, 31, 3–39.

Gill, P. E., Murray, W., and Wright, M. H. (1981), *Practical Optimization*, London: Academic Press.

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

Huang, S. Y. (1990), "Density Estimation Using Spline Projection Kernels," Unpublished doctoral thesis, Purdue University.

Lepski, O. V., Mammen, E., and Spokoiny, V. G. (1997), "Optimal Spatial Adaptation to Inhomogeneous Smoothness: An Approach Based on Kernel Estimates With Variable Bandwidth Selections," *The Annals of Statistics*, 25, 929–947.

Luo, Z., and Wahba, G. (1997), "Hybrid Adaptive Splines," *Journal of the American Statistical Association*, 92, 107–115.

Schumaker, L. L. (1981), *Spline Functions: Basic Theory*, New York: Wiley.

Smith, P. L. (1982), "Curve Fitting and Modeling With Splines Using Statistical Variable Selection Techniques," Report NASA 166034, Langley Research Center, Hampton, VA.

Stein, C. (1981), "Estimation of the Mean of a Multivariate Normal," *The Annals of Statistics*, 9, 1135–151.

Stone, C. J., Hansen, M., Kooperberg, C., and Truong, Y. K. (1997), "Polynomial Splines and Their Tensor Products in Extended Linear Modeling," *The Annals of Statistics*, 25, 1371–1470.

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