Hybrid Adaptive Splines

Zhen LUO and Grace WAHBA

An adaptive spline method for smoothing is proposed that combines features from both regression spline and smoothing spline approaches. One of its advantages is the ability to vary the amount of smoothing in response to the inhomogeneous "curvature" of true functions at different locations. This method can be applied to many multivariate function estimation problems, which is illustrated by an application to smoothing temperature data on the globe. The method's performance in a simulation study is found to be comparable to the wavelet shrinkage methods proposed by Donoho and Johnstone. The problem of how to count the degrees of freedom for an adaptively chosen set of basis functions is addressed. This issue arises also in the MARS procedure proposed by Friedman and other adaptive regression spline procedures.

KEY WORDS: Inflated degrees of freedom; Regression spline; Smoothing on the sphere; Smoothing spline; Spatial adaptability; Stepwise regression.

1. INTRODUCTION

Spatially adaptive smoothing, or function estimation, methods that can handle a wide variety of shapes and spatial inhomogeneity have interested statisticians for a long time. Recently, Donoho and Johnstone (1994, 1995, and with Kerkyacharian and Picard [1995]) introduced a group of wavelet shrinkage methods shown to have desirable spatial adaptability by both theoretical arguments and simulation study. Traditionally, two techniques have been used to address this problem of spatial adaptability. One technique uses local variable smoothing parameters (or bandwidths) in common smoothing methods, such as smoothing spline and kernel methods. The other technique is to place knots adaptively in a regression spline method (or, equivalently, adaptively choose a set of spline basis functions for regression). Recent examples in the first category include work of Abramovich and Steinberg (1995) and Fan and Gijbels (1995) (see also Wahba 1995). Much other recent research in this area is noted in the lengthy discussion to the work of Donoho, Johnstone, Kerkyacharian, and Picard (1995).

In the second category, subsequent to Smith's early (1982) work on using statistical variable selection techniques to fit splines, there have been quite a few works along this direction in the adaptive regression spline literature (Friedman and Silverman 1989 [TURBO], Friedman 1991 [MARS], Stone, Hansen, Kooperberg, and Troung 1995). The idea of placing knots adaptively using some kind of variable selection technique has become the primary choice in the development of regression spline methodology. This technique tends to choose more basis functions in data-dense regions where the underlying true function has more structure, which is what we want.

In this article we combine some of the features of adaptive regression splines and traditional smoothing splines to obtain a hybrid smoothing procedure termed hybrid adaptive splines (HAS), which may be implemented with large datasets and displays a desirable form of spatial adaptability when the underlying function is spatially inhomogeneous in its degree of complexity. The basis functions chosen as a subset of the basis functions occurring naturally in smoothing splines are selected one basis function at a time, using a forward stepwise regression procedure. A generalized cross-validation (GCV) criterion with an inflated degrees of freedom (IDF) factor to account for the fact that the basis functions are chosen adaptively is used as a stopping criterion, similar to MARS procedure. Then, instead of a backwards deletion, the selected basis functions are used in a penalized regression derived from the original smoothing spline method. We explain the procedure in Section 2, and provide some theoretical results on the appropriate IDF factor. The choice of IDF factor arises in MARS and other regression spline procedures, and we discuss how our results might apply to such procedures.

Hastie (1989), in his discussion to Friedman and Silverman (1989), suggested a similar scheme. He suggested that an overparameterized model constructed by TURBO or by other regression spline methods could be regularized by a ridge regression step to reduce the variability due to the adaptive selection of basis functions. In a way, HAS can be viewed as a modification of Hastie's suggestion. But HAS differs from his suggestion in two respects. First, we use the basis functions derived naturally from the corresponding smoothing spline methods; that is, those based on the reproducing kernels. The same is true with the penalty term of the final ridge regression step. These are some of the smoothing spline features incorporated in HAS, besides the obvious penalization feature. In the univariate case, as a referee suggested, we may use other equivalent basis functions, such as truncated power basis functions, or B splines if we are concerned with computational efficiency. However, basis functions based on the reproducing kernels can be extended naturally to multivariate smoothing spline methods (e.g., see Sec. 4). This of course does not mean that other basis functions cannot be generalized to multivariate cases; however, then they will not be equivalent. Second,

Zhen Luo was a graduate student, Department of Statistics, University of Wisconsin, Madison, WI 53706, during the work on this article and is now Assistant Professor, Department of Statistics, The Pennsylvania State University, University Park, PA 16802. Grace Wahba is Bascom Professor, Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706. This research was supported in part by National Science Foundation grant DMS 9121003 and National Aeronautics and Space Administration grant NAGW-2961. The authors thank the referees and the associate editor for their helpful comments.

Journal of the American Statistical Association
March 1997, Vol. 92, No. 437, Theory and Methods

our main purpose in proposed HAS is for its spatial adaptability, over-parameterization is not desirable in this case, since it will reduce this spatial adaptability.

The HAS procedure is not the same as choosing a random, or representative, or systematic sample of the basis functions that occur naturally in spline smoothing. This latter procedure (which does not use the response vector as part of its selection method) has been suggested and implemented by various authors as a numerical tool for efficiently calculating a good approximation to the original smoothing spline variational problem (see, e.g., Hutchinson, Kalma, and Johnson 1984, and Wahba 1980). If the underlying function is highly spatially inhomogeneous, then the HAS selection of basis functions is not expected to be a representative sample of the naturally occurring basis functions. It could be argued that the HAS estimate will then be a solution to a slightly different (weighted) variational problem, although we offer no theoretical argument to back this up.

Several features of this procedure are worth mentioning. First, the procedure is well suited to highly unequally spaced data. Second, it extends in a straightforward way to the general penalized likelihood setup as discussed by, for example, Wahba (1990), and in particular to the smoothing spline analysis of variance (ANOVA) setup of Gu and Wahba (1993a,b). Some examples were given by Luo (1994). The procedure can be used in the context of splines on the sphere, which has the potential for wide application in meteorological and environmental studies. We use an application to the interpolation and smoothing of global winter surface temperature to illustrate this application in Section 4. Finally, based on simulated examples, including the four used by Donoho and Johnstone (1994), it seems fair to say that this procedure, in terms of both mean squared error (MSE) and visual appearance, is comparable to the wavelet simulation results. Moreover, in our examples when the signal does not have much spatial inhomogeneity (in which case nonadaptive smoothing methods, such as smoothing splines with a global smoothing parameter, perform better on the average than adaptive methods), HAS's performance is close, whereas wavelet methods seem to need further refinement to obtain close results. We provide these comparisons in Section 3.

2. HYBRID ADAPTIVE SPLINES

2.1 Smoothing Splines

Let

$$y_i = f(x_i) + e_i, \qquad i = 1, 2, \ldots, n,$$

where $x_i \in [0,1]$, the $\{e_i\}$ are iid $N(0, \sigma^2)$, and f is "smooth". More precisely, suppose that f is in the Sobolev space $\mathcal{W}_2[0,1] = \{f\colon f, f' \text{ absolutely continuous}, f'' \in \mathcal{L}_2\}$. The traditional cross-validated cubic smoothing spline estimate of f is the solution to this problem: Find $f \in \mathcal{W}_2[0,1]$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_0^1 (f''(x))^2\,dx, \tag{1}$$

where λ is chosen by GCV (see Wahba 1990). It can be shown (Wahba 1990, pp. 11–12) that the minimizer f_λ of (1) has a representation

$$f_\lambda(x) = d_1\phi_1(x) + d_2\phi_2(x) + \sum_{i=1}^{n} c_i R(x; x_i), \tag{2}$$

where $\phi_1(x) = 1, \phi_2(x) = k_1(x)$, and $R(x;x') = k_2(x)k_2(x') - k_4([x-x'])$. Here $k_1(x) = x - 1/2, k_2(x) = (k_1^2(x) - 1/12)/2$, and $k_4(x) = (k_1^4(x) - k_1^2(x)/2 + 7/240)/24$. Furthermore, $\int_0^1 (d^2/dx^2(\sum_{i=1}^n c_i R(x;x_i)))^2\,dx = \sum_{i,j=1}^n c_i c_j R(x_i;x_j)$.

Plugging the right side of (2) into (1), the original variational problem becomes a quadratic optimization problem,

$$\underset{\mathbf{c},\mathbf{d}}{\operatorname{argmin}} \frac{1}{n}\|\mathbf{y} - (\mathbf{T}\mathbf{d} + \boldsymbol{\Sigma}\mathbf{c})\|^2 + \lambda \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c}, \tag{3}$$

where $\mathbf{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{\Sigma}$ is the $n \times n$ matrix with (i,j)th entry $R(x_i;x_j)$, $\mathbf{T}$ is the $n \times 2$ matrix with (i,ν)th entry $\phi_\nu(x_i)$, $\mathbf{d} = (d_1, d_2)'$, and $\mathbf{c} = (c_1, c_2, \ldots, c_n)'$.

2.2 Hybrid Adaptive Spline Procedure

Instead of minimizing (1) in $\mathcal{W}_2[0,1]$, we minimize it in span ϕ_1, ϕ_2 plus a specially selected subset of the n basis functions $R(\cdot; x_i), i = 1, 2, \ldots, n$, chosen in a forward stepwise manner. Having chosen ϕ_1, ϕ_2 and $R(\cdot; x_{i_l}), l = 1, 2, \ldots, k-3$, we choose the kth basis function to maximize the reduction in the residual sum of squares (RSS). We did this rapidly, as follows. We need to compute the RSS of the least squares fit of $\mathbf{y}$ on the selected basis vectors (basis functions evaluated at data points), with and without the vector corresponding to a new candidate basis function, denoted by $\mathbf{u}$. Suppose that there are k basis vectors in the subset already, and that they are stored in a n by k matrix $\mathbf{X}$. We first do a QR decomposition of $\mathbf{X}$: $\mathbf{Q}'\mathbf{X} = (\mathbf{R}_{k\times k}\mathbf{0}_{k\times(n-k)})'$; we then multiply both $\mathbf{y}$ and $\mathbf{u}$ by $\mathbf{Q}$: $\mathbf{Q}'\mathbf{y} = (\mathbf{z}_{1\times k}\mathbf{t}_{1\times(n-k)})'$ and $\mathbf{Q}'\mathbf{u} = (\mathbf{v}_{1\times k}\mathbf{s}_{1\times(n-k)})'$. It is easy to verify that RSS($\mathbf{y}$ regressed on $\mathbf{X}$) $= \mathbf{t}\mathbf{t}'$ and RSS($\mathbf{y}$ regressed on $\mathbf{X}$ and $\mathbf{u}$) $= \mathbf{t}\mathbf{t}' - (\mathbf{s}\mathbf{t}')^2/\mathbf{s}\mathbf{s}'$. Each time among those unselected, the basis function making the largest RSS deduction (i.e. the largest $(\mathbf{s}\mathbf{t}')^2/\mathbf{s}\mathbf{s}'$) will be the next one to enter the subset. We chose the Householder reflection method to do the QR decomposition because it, along with the multiplication of $\mathbf{y}$ and $\mathbf{u}$ by $\mathbf{Q}$, can be easily updated every time a new basis vector is appended to $\mathbf{X}$ (see Seber 1977, pp. 312–314, 338–341).

The number of basis functions is chosen by minimizing a similar GCV score as implemented in MARS. The GCV score for k selected basis functions is $\text{GCV}(k) = \text{RSS}/(1 - (2 + (k-2)\text{IDF})/n)^2$, where IDF is applied to each of the $k-2$ adaptively selected basis functions to account for the added flexibility due to the fact that they have been selected adaptively. The GCV score is minimized over $k = 2, 3, \ldots, q$ for some (safe) upper limit q. Obviously, IDF should be larger than 1. Friedman (1991) recommended 3 as a generally appropriate choice (default) in MARS and most IDF's chosen by cross-validation, and simulation studies show that this is an appropriate choice in MARS. In HAS,

however, simulations show that 1.2, or at least a number less than 2, is a better choice. Some theoretical explanations for this difference are discussed in Section 2.3. For all of the examples in this article, the IDF is fixed at 1.2.

The final step of penalized regression is done by subroutine dsnsm in GCVPACK developed by Bates, Lindstrom, Wahba, and Yandell (1989). dsnsm is a routine to do ridge regressions with smoothing parameters chosen by a GCV criterion (which is independent of the GCV criterion that we use to choose the number of basis functions). Note that using only a subset of basis functions in representation (2),

$$f_\lambda(x) = d_1\phi_1(x) + d_2\phi_2(x) + \sum_{l=1}^{k} c_l R(x; x_{i_l}), \qquad (4)$$

the quadratic optimization problem, derived from (1) by plugging (4) in it, is the same as (3) except that the two Σ's are replaced by their corresponding submatrices; that is, the first Σ is replaced by $(R(x_i; x_{i_l})), i = 1, 2, \ldots, n$, and $l = 1, 2, \ldots, k$ and the second Σ is replaced by $(R(x_{i_l}; x_{i_{l'}})), l, l' = 1, 2, \ldots, k$.

It took less than 10 minutes to get a HAS fit on our Alpha DEC3000/M400 machine for the simulated Examples 1–5 of Section 3 with sample size 2,048 and $q = 150$, less than 2 seconds for Examples 6 and 7 with sample size 256 and $q = 60$, and about 3 minutes for the example in Section 4 with sample size 725 and $q = 500$. (Note that this procedure can be applied to any of the spline models in Wahba 1990.)

2.3 Inflated Degrees of Freedom for an Adaptively Selected Basis Function

In this section, we are going to investigate how large the IDF should be. The IDF ultimately controls the number of basis functions put in our final model. The larger the IDF, the fewer basis functions we put in. Another reason for studying this problem is to explain why Friedman chose 3 as a general good choice in MARS, but our experience shows that a IDF below 2 is better in HAS.

Consider a simplified version of our problem, the regression model

$$y_i = aR(x_i, t) + e_i, \qquad i = 1, 2, \ldots, n, \qquad (5)$$

where $x_i = i/n, a$ and t are two parameters, and $\{e_i\}$ are iid $N(0, 1)$. The function R is a known reproducing kernel used for smoothing spline estimates as in (2). In this section we consider only two reproducing kernels corresponding to periodic linear and cubic splines on [0, 1] (Wahba 1990, pp. 21–22):

$$R_1(s,t) = \left(\left(|s-t| - \frac{1}{2}\right)^2 - \frac{1}{12}\right) \Big/ 2 \qquad (6)$$

and

$$R_2(s,t) = \left(-\left(|s-t| - \frac{1}{2}\right)^4 + \frac{1}{2}\left(|s-t| - \frac{1}{2}\right)^2 - \frac{7}{240}\right) \Big/ 24. \qquad (7)$$

Note that R_2 is only part of R used in (2). These periodic forms are chosen because the stochastic processes derived later will be stationary; hence some existing results about stationary processes can be used. Note that even though R_1 is corresponding to periodic linear splines, it is not piecewise linear itself.

The difference between the residual sum of squares (RSS) of the least squares fit under the null model, H_0: $y_i = e_i$, and under (5) is defined as the model sum of squares of (5) as in linear model theory. The expectation of this difference, when the data are drawn from H_0, is the degrees of freedom of the model (5). Because there is only one basis function in (5), this can also be interpreted as the degrees of freedom for one basis function.

If t is fixed and known, then this is just an ordinary linear regression problem. Let

$$S(a,t) = \sum_{j=1}^{n} (y_j - aR(x_j, t))^2.$$

Then the model sum of squares of (5) is

$$\text{RSS(model } H_0) - \text{ RSS(model(5))} = \mathbf{y}'\mathbf{y} - \min_a S(a,t) = \frac{(\sum R(x_j,t)y_j)^2}{\sum R(x_j,t)^2}.$$

Denote this difference under H_0 by $V^2(t)$; that is, $V(t) = (\sum_j R(x_j,t)e_j)/\sqrt{\sum_j R(x_j,t)^2}$. We know that $V^2(t)$ is distributed as χ_1^2 and that $E(V^2(t)) = 1$ for each t.

According to our adaptive procedure, we choose as t the x_i minimizing $\min_a S(a, x_i)$ among all x_i. Hence now the model sum of squares under H_0 is

$$\mathbf{e}'\mathbf{e} - \min_{t\in\{x_1,\ldots x_n\},a} S(a,t) = \max_{t\in\{x_1,\ldots x_n\}} (\mathbf{e}'\mathbf{e} - \min_a S(a,t)) = \max_i V^2(x_i),$$

which is greater than or equal to any of $V^2(x_i)$. Therefore, its expectation is greater than or equal to that of $V^2(x_i)$; that is, 1.

On the other side,

$$\max_i V^2(x_i) \le \max_{t\in[0,1]} V^2(t) = \mathbf{e}'\mathbf{e} - \min_{a,t} S(a,t),$$

which is the model sum of squares under H_0 of the nonlinear regression model (5), which has two parameters a and t. If R is a reproducing kernel corresponding to cubic splines (i.e., R_2, which is twice continuously differentiable on $[0,1]^2$), then by the standard nonlinear regression asymptotic theory (see, e.g., Gallant 1987), we know that this model sum of squares is asymptotically distributed as χ_2^2. Therefore, the degrees of freedom of model (5) or an adaptively chosen cubic spline basis function should be between 1 and 2.

Note that R_1 is only continuous, not differentiable. Therefore, the standard asymptotic theory does not apply. The simulation study done by Hinkley for a similar simple change point model, used by Friedman and Silverman

Table 1. Specifications of Simulated Examples

Example	f	σ	Sample size (n)	SD(f)/σ	Number of replicates
1	DJ(1994)'Blocks*3.5	1.0	2,048	6.92	31
2	DJ(1994)'Bumps*4.5	1.0	2,048	6.93	31
3	DJ(1994)'Heavisine*2.2	1.0	2,048	6.54	31
4	DJ(1994)'Doppler*22	1.0	2,048	6.36	31
5	DJ(1994)'Doppler*22	exp(x)/1.648	2,048	6.36	31
6	$\sin(2(4x-2)) + 2\exp(-16x^2)$	.3	256	2.80	400
7	$(4x-2) + 2\exp(-16x^2)$	.4	256	3.16	400

(1989) for the purpose of deciding how many extra degrees of freedom should be given to an adaptively chosen basis function, indicates that the model sum of squares then is approximately distributed as χ^2_3. This is also supported by Owen (1991)'s theoretical argument.

Another way to investigate the degrees of freedom for an adaptively chosen basis function is to consider a centered Gaussian processes, Z_n, defined by

$$Z_n(t) = (1 - nt + [nt])V_{[nt]} + (nt - [nt])V_{[nt]+1},$$

for

$$t \in [0, 1],$$

where $V_i = V(x_i)$ for $i = 1, 2, \ldots, n$ and $V_0 = 0$. Z_n is just a process joining V_i at i/n by straight lines.

It is clear that $\{V_i\}$ are multinormal distributed, $E(V_i) = 0$, $\text{var}(V_i) = 1$, and

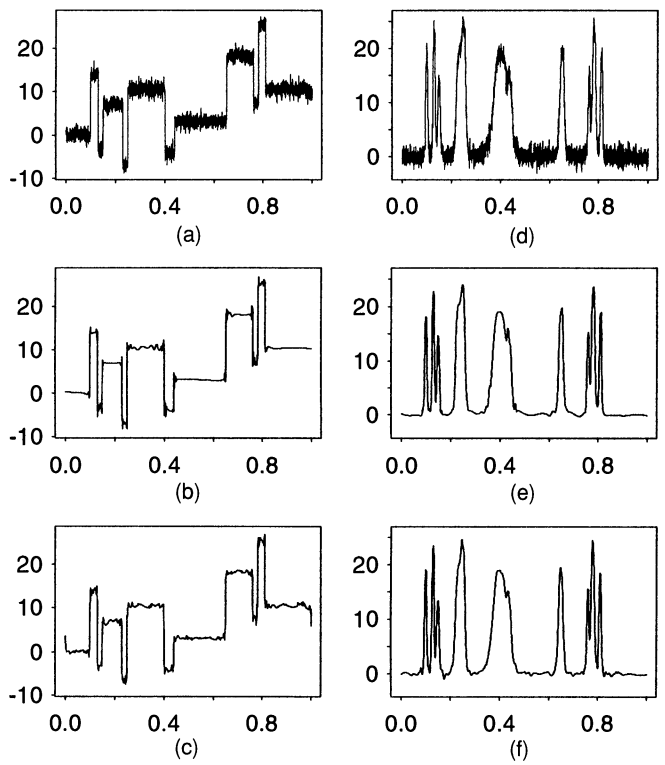

Figure 1. Example 1 (Blocks, a,b,c) and Example 2 (Bumps, d,e,f). (a) One copy of data; (b) HAS fit with median MSE; (c) SUREShrink fit with median MSE; (d) one copy of data; (e) HAS fit with median MSE; (f) SUREShrink fit with median MSE.

$$\text{cov}(V_i, V_k) = \frac{\sum_j R(x_j, x_i)R(x_j, x_k)}{\sqrt{\sum_j R(x_j, x_i)^2}\sqrt{\sum_j R(x_j, x_k)^2}}.$$

It can be proven that for the reproducing kernel, R_1 and R_2, the process Z_n converges weakly to a centered Gaussian process Z with covariance function

$$G(s,t) = \frac{\int_0^1 R(u,s)R(u,t)\,du}{\sqrt{\int R(u,s)^2}\sqrt{\int R(u,t)^2}}, \qquad s,t \in [0,1]. \quad (8)$$

Therefore, the model sum of squares of (5) with adaptively chosen t under H_0,

$$\max_i V^2(x_i) = \max_i V_i^2 = \sup_{t\in[0,1]} Z_n(t)^2,$$

converges to $\sup_{t\in[0,1]} Z(t)^2$ in distribution.

Proposition 2.3.1. For the reproducing kernels R_1 and R_2 in (6) and (7), the corresponding $Z_n \Rightarrow Z$, a zero-mean stationary Gaussian process on [0, 1] with respective covariance functions

$$G_1(s;t) = 1 - 30(t-s)^2 + 60(t-s)^3 - 30(t-s)^4 \quad (9)$$

and

$$G_2(s;t) = 1 - 20(t-s)^2 + 70(t-s)^4 - 140(t-s)^6 + 120(t-s)^7 - 30(t-s)^8. \quad (10)$$

Proposition 2.3.1 tells us that the asymptotic distribution of the model sum of squares of (5), $\max_i V^2(x_i)$, and hence the degrees of freedom of an adaptively chosen basis function, is decided by the function R, which determines the basis function family. In particular, cubic spline basis and linear spline basis can be expected to have different IDF's for an adaptively chosen basis function. Considering that Friedman used linear spline basis functions in MARS instead of cubic spline basis functions that we use, it is not a surprise that our experience is different than his.

The convergence results in Proposition 2.3.1 may be proved in a more general case. But for our purpose the current form is enough to show our point, and it also has the following nice corollary based on the existing theory of extreme value of stationary Gaussian processes.

Proposition 2.3.2. For the Gaussian processes Z defined in Proposition 2.3.1,

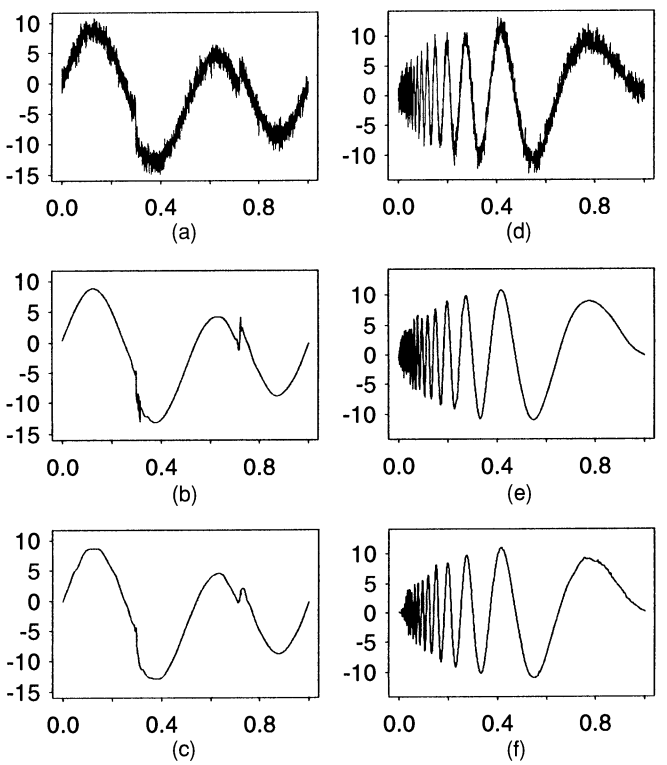

Figure 2. Example 3 (Heavisine, a,b,c) and Example 4 (Doppler, d,e,f). (a) One copy of data; (b) HAS fit with median MSE; (c) SUREShrink fit with median MSE; (d) one copy of data; (e) HAS fit with median MSE; (f) SUREShrink fit with median MSE.

$$\lim_{u\to\infty} \exp\left(\frac{1}{2}u^2\right) \Pr\left\{\sup_{0\le t\le h} |Z(t)| > u\right\} = h\sqrt{2C_i/\pi}$$

for each $0 < h < 1, i = 1, 2$, where $C_i = 30$ or 20, corresponding to covariance function G_1 or G_2 given in Proposition 2.3.1.

In some sense this result tells us the tail probability of $\sup_{t\in[0,1]} |Z(t)|$, but because h must be less than 1 in the proposition, this probability is not exactly what we want. However, if we restrict our searching for t in a smaller area $[0, 1-\varepsilon]$ instead of [0, 1], as suggested by Owen (1991) as a way to reduce the cost (degrees of freedom) of an adaptively chosen basis function, then the model sum of squares will converge in distribution to $\sup_{t\in[0,1-\varepsilon]} |Z(t)|^2$, whose tail probability $P\{\sup_{t\in[0,1-\varepsilon]} |Z(t)|^2 > u\}$ by Proposition 2.3.2 can be approximated by $(1-\varepsilon)\exp(-u/2)\sqrt{2C/\pi}$. Because the process corresponding to linear spline basis has a larger C (which is 30) than the process corresponding to cubic spline basis, the model sum of squares has a larger tail probability, and hence larger variation as well. This means that more degrees of freedom should be given to an adaptively chosen linear spline basis function than to an adaptively chosen cubic spline basis function. This partially justifies the choices of 1.2 for HAS and 3 for MARS.

3. SIMULATION STUDY

In this section we use simulated examples to examine the performance of HAS compared to the MARS, wavelet shrinkage, and smoothing splines SS procedures.

The first five examples, which show strong spatial inhomogeneity, were taken from Donoho and Johnstone (1994). We also include two examples from Fan and Gijbels (1995) that do not have such strong spatial inhomogeneity. To enable comparison with wavelet methods, all designs in these examples are chosen as equally spaced, although the other three methods can apply to non–equally spaced designs as well. Gaussian noise is added such that $\mathrm{SD}(f)/\sigma$ as an approximate measure of signal-to-noise ratio is about 7 for Examples 1–5 and 3 for Examples 6 and 7. Example 5 does not have a common standard deviation, so σ in the ratio is replaced by the median standard deviation. More information about these examples is given in Table 1. We used the pseudostandard normal random number generator `rnor`, a Fortran subroutine from CMLIB.

Of all the wavelet shrinkage methods proposed by Donoho and Johnstone, we chose the SUREShrink method (Donoho and Johnstone 1995) for our comparisons, because it has a level-dependent threshold feature and is better on the average than RiskShrink (Donoho and Johnstone 1994) in our experience. We chose the "primary resolution level" as 5, as used by Donoho and Johnstone (1994). We performed the computation with the software *wavethresh,* developed by Nason and Silverman (1994) in S-PLUS; that the family of wavelets was *DaubLeAsymm* with *filter number* 8. The S-PLUS commands that we used are given in Appendix A.

Among adaptive regression spline methods, we chose MARS for comparison with HAS, because it is among the most widely used adaptive regression spline programs. Other methods include that of Stone, Hansen, Kooperberg,

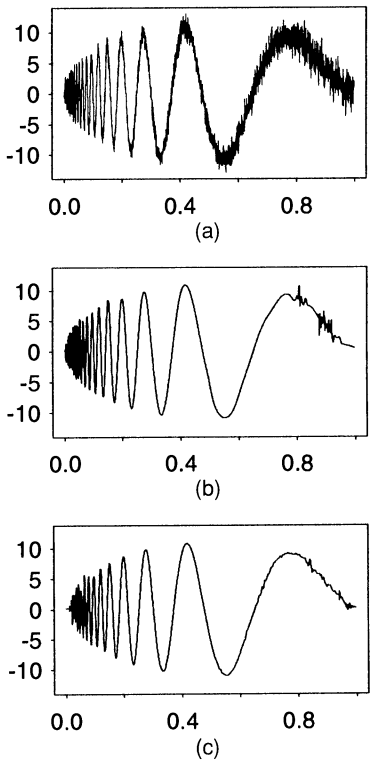

Figure 3. Example 5 (Doppler2). (a) One copy of data; (b) HAS fit with median MSE; (c) SUREShrink fit with median MSE.

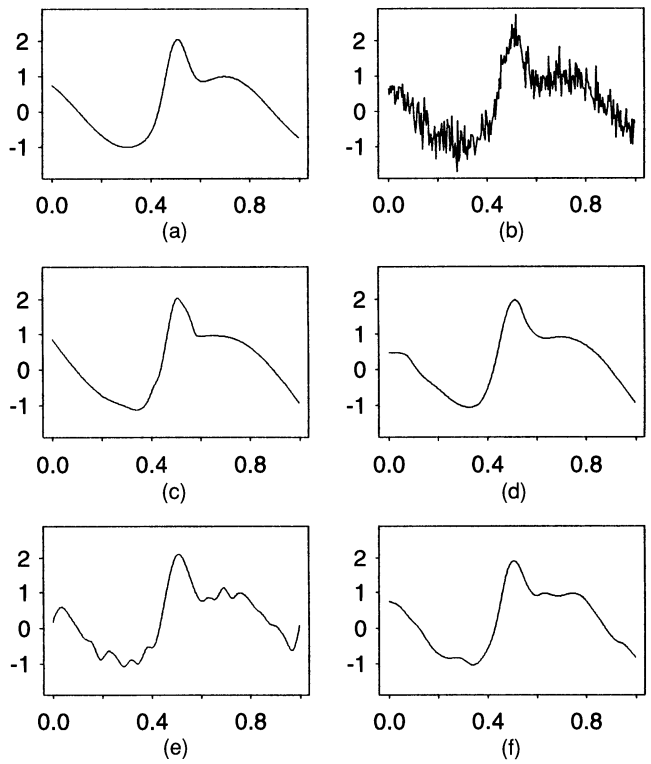

Figure 4. Example 6. (a) True function; (b) one copy of data; (c) HAS fit with median MSE; (d) MARS fit with median MSE; (e) SUREShrink fit with median MSE; (f) SS fit with median MSE.

and Truong (1995), which is similar to MARS, especially in the regression case (see their Section 6).

We set the maximum number of basis functions (q) in both HAS and MARS at 150 for Examples 1–5 (except for Example 1, where q in HAS is set at 250) and 60 for Examples 6 and 7. (The number of basis functions finally used in HAS fits was about 190 for Example 1, 120 for Example 2, 50 for Example 3, 90 for Example 4, 120 for Example 5, and 13 for Examples 6 and 7.) We set the minimum span parameter in MARS at zero (the default value) in all examples. We tried other choices as well, including one that allows just one observation between two consecutive knots. We also tried other choices of IDF in MARS (it is called df there), including the one chosen by ten-fold cross-validation. All of the results were similar to those we report here.

The Fortran routines used to compute HAS estimates are available on request from the first author. We computed the SS estimates using the code GCVSPL in Fortran by Woltring, with the smoothing parameters chosen by GCV. Woltring's code combines different people's programs, including Hutchinson and de Hoog (1985). For complete references, please see the GCVSPL's documentation. The codes *mars3.5* for MARS, *wavethresh*, and CMLIB can be obtained from statlib. GCVPACK and GCVSPL can be obtained from netlib.

The median performances in terms of mean squared error (MSE), defined as $\sum_{i=1}^{n}(\hat{f}(x_i)-f(x_i))^2/n$, of SUREShrink and HAS for Examples 1–5 are shown in Figures 1–3. For Examples 6 and 7, the median performances of HAS, MARS, SUREShrink, and SS are shown in Figures 4 and 5. The medians and the differences of first and third quartiles (as a measure of variation in the results) of MSE for all these examples are given in Table 2.

In Examples 1–5, both HAS and SUREShrink exhibit spatial adaptability, whereas HAS has smaller median MSE than SUREShrink. Notice that SUREShrink has about the same MSE in Example 5 as in Example 4, even though the noise variance in Example 5 is not homogeneous in x; but the same cannot be said about HAS. Visually, however, both methods are sensitive to the unequal variance in the noise.

SS's relatively inferior performance in Examples 1–5 is no doubt due to the use of a single smoothing parameter across the entire design space, which makes it either follow the high-frequency signal without smoothing out much of noise or smooth out the noise with the signal degraded at the same time. However, it has the smallest variation in MSE. This is not surprising, given the fact that all of the other methods are trying to do different amounts of smoothing at different locations, and hence are trying to estimate more than a single smoothing parameter. MARS essentially did not give sensible answers in Examples 1, 2, 4, and 5. The smallest of the missing entries of Table 2 was greater than 6. Of course MARS was specifically designed for high-dimensional problems, not one-dimensional problems with (pathological) discontinuities. On the other hand, in the spatially more homogeneous Examples 6 and 7, SS was best, with HAS and MARS close behind and SUREShrink further behind. However, a lower "primary resolution level" (we chose level 5 because it gave the best performance over-

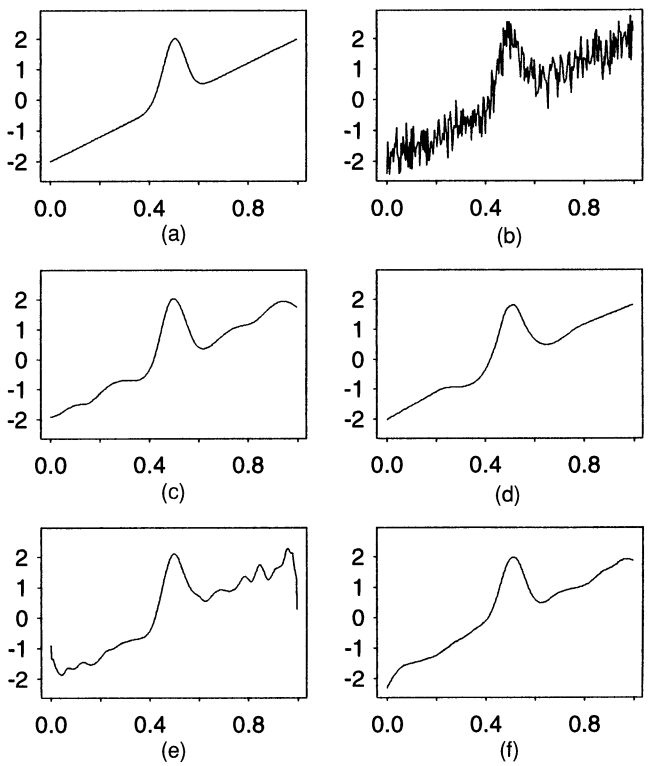

Figure 5. Example 7. (a) True function; (b) one copy of data; (c) HAS fit with median MSE; (d) MARS fit with median MSE; (e) SUREShrink fit with median MSE; (f) SS fit with median MSE.

Table 2. Median of MSE and the Difference of the First and Third Quartiles of MSE (in Parentheses)

Example	HAS	SS	SUREShrink	MARS
1	.137(.018)	.546(.023)	.398(.049)	
2	.087(.021)	.124(.010)	.167(.015)	
3	.039(.013)	.075(.005)	.062(.007)	.150(.014)
4	.068(.015)	.205(.011)	.145(.013)	
5	.100(.072)	.232(.014)	.149(.013)	
6	.007(.006)	.006(.003)	.018(.004)	.007(.004)
7	.012(.011)	.010(.005)	.042(.012)	.012(.007)

all) might give better results in these two examples, which have their energy at lower frequencies (see Fan and Gijbels 1995 for further discussion).

Notice that HAS has a larger MSE variability than SS, particularly in Examples 3, 5, 6, and 7. We do not know whether this is due to variability in the stepwise selection procedure or to variability in the GCV criterion that we used to decide the number of basis functions. We compared the results to those obtained with an ideally chosen number of basis functions, using the same stepwise selection but deciding k by looking at the MSE with respect to the truth. This "ideal" procedure had much less variation, suggesting that the source of the variability may be the latter.

4. APPLICATION TO A SMOOTHING PROBLEM ON THE SPHERE

We now illustrate HAS's applicability to multivariate problems using a smoothing problem in meteorology. From a global monthly surface temperature data archive developed by Jones et al. (1991), we extracted all of the 1981 winter temperature records with the locations (longitude and latitude) of the recording stations. The winter temperature is defined as the average of December 1980 and January and February 1981 monthly temperatures. The total of 725 stations with such records are distributed very irregularly on the sphere.

A spline on the sphere estimate was defined by Wahba (1981) as the solution of the following optimization problem:

$$\operatorname*{argmin}_f \frac{1}{n}\sum_{i=1}^{n}(y_i - f(P_i))^2 + \lambda \int_S (\Delta^{m/2} f)^2 \, dP, \quad (11)$$

where P = (*latitude, longitude*) is a point on the sphere S, P_i is the location of the ith station, Δ is the Laplacian on the sphere, and f is in the Sobolev space $\mathcal{H}_m(S) = \{f\colon f \in \mathcal{L}_2(S), \Delta^{m/2} f \in \mathcal{L}_2(S)\}$. Wahba showed that the minimizer of (11) has a representation of the form

$$f(P) = d + \sum_{i=1}^{n} c_i Q_m(P; P_i),$$

where $Q_m(P; P')$ $(m = 1, 2, \ldots)$ are a family of reproducing kernels related to Green's functions for Δ^m, for which closed-form expressions are not known. A family $R_m(P; P')$ of reproducing kernels approximating the Q_m and with closed-form expressions were given by Wahba (1981, eqs. 3.3 and 3.4). We use $R_2(P; P')$ from that article, denoted by $R(P; P')$ in what follows.

HAS can be directly applied to this situation. The only difference is that the collection of candidate basis functions now is $\{\phi_1, R(\cdot; P_i)$, for $i = 1, 2, \ldots, n\}$, where $\phi_1(P) \equiv 1$.

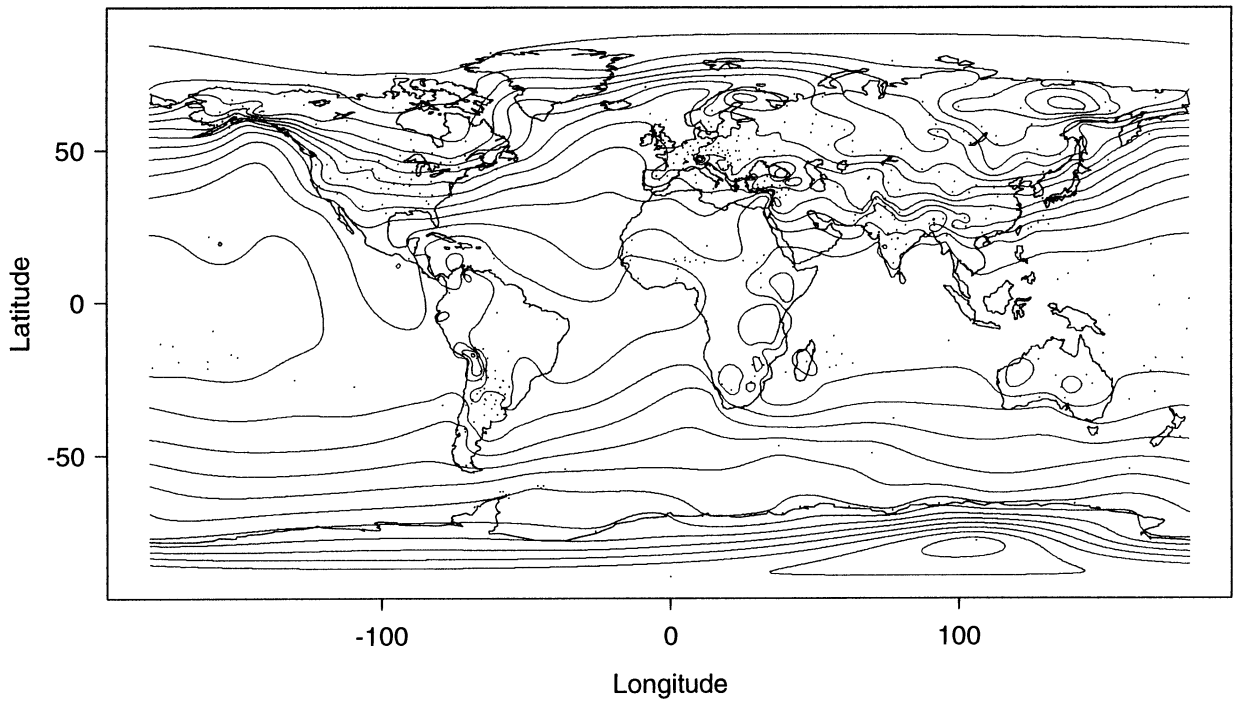

Figure 6. Example in Section 4: Contour Plot of HAS Fit of 1981 Global Winter Temperature. The dots indicate the locations of the recording stations.

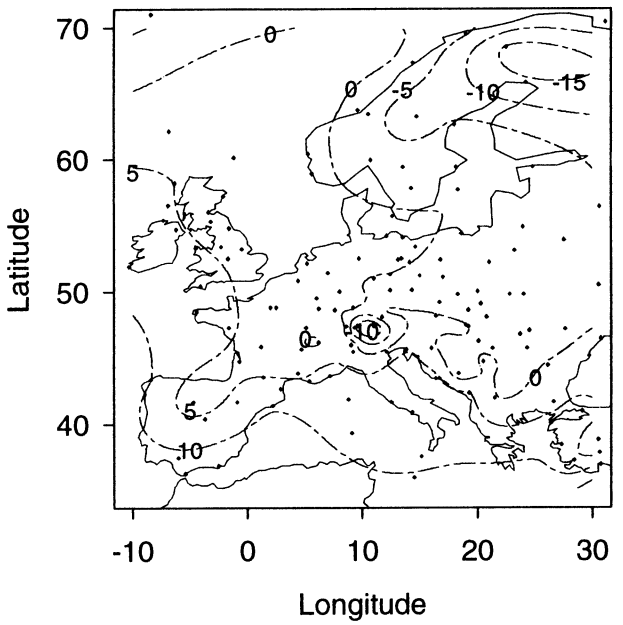

Figure 7. Enlarged European Part of Figure 6. The dots indicate the locations of the recording stations. The dashed lines are the contour lines of HAS fit.

A fit by HAS on the whole globe is shown in Figure 6, and an enlarged European part is shown in Figure 7. The maximum number of basis functions was set at 500, and the final number of basis functions chosen by GCV was 425. We can see that without disturbing those areas with little data or without much structure, the fine details at places where there sufficient data exist are kept when smoothing is done. Figure 7 shows the colder surface temperature measured in the Alps. Similar detail can be seen in the Andes. Some structure is obtained over the Himalayas, but there are few stations there. On the other hand, the surface temperature (generally observed on islands) is seen to be quite smooth over the oceans. The smoothing spline on the sphere (using the full set of basis functions) gave a similar picture; however, the interesting features over the mountain ranges were considerably smoother. Simulated examples that we have studied show the same kind of spatial adaptability as the one-dimensional examples. We remark we chose $R = R_2$ because we suspected that it probably represented a good general-purpose low-pass filter on the sphere. In practice we might wish to optimize this choice, either over m (see Wahba and Wendelberger 1980), or by considering other families of reproducing kernels (see, e.g., Weber and Talkner 1993). Gao (1993) and Wahba (1982) have used reproducing kernels on the sphere based on historical meteorological information.

A referee suggested that a more complete and accurate description of global temperature would incorporate a surface elevation component in the model. That would remove a substantial amount of the spatial heterogeneity. Here would be a nice opportunity to agree. Work in this direction, particularly in a space-time–type modeling of historical global temperature, is in progress (Luo 1996).

5. DISCUSSION AND CONCLUSIONS

HAS may be applied to ANOVA in functions spaces (Gu and Wahba 1993a,b; Wahba 1990; Wahba, Wang, Gu, Klein, and Klein 1994). Because all of the estimates there have representations like (2), the extension is immediate. Extensions to global temperature as a function of year and space are under study.

Finally we comment further on why we want a penalized regression step in HAS. Basically, choosing the number of basis functions via GCV has done most of the work for balancing between bias and variance, hence the penalized regression step here is primarily a refinement of the results from the regression step. In our experience it does usually reduce the MSE, albeit just minimally (a few percentage point reduction). This actually confirms a theoretical result of von Golitschek and Schumaker (1990) which says that unless the truth is in the span of the basis functions, in average some smoothing will always be better than just regression in terms of MSE. But there is a still more important reason why we want to do a penalized regression—namely, for numerical stability. As is well known, when the number of basis functions (regressors) increases, the regression problem becomes more and more ill-conditioned, which makes its numerical computation less and less stable. The basis functions that we used in the simulations—cubic spline basis functions—have larger correlations among them than linear spline basis functions, as shown in Section 2.3. Hence the ill-conditioning problem is more serious here. The penalized regression step acts as a remedy for this.

APPENDIX A: S-PLUS COMMANDS FOR SURESHRINK

This is the list of S-PLUS commands that we used to compute SUREShrink estimates with Nason and Silverman (1994)'s *wavethresh* package.

```
sureth<-function(d,x){# based on (6) and (7) in DJ(1995)
  sure<-d-2*(1:d)+cumsum(sort(abs(x))^2)
  x[order(sure)[1]]}
J<-7   # corresponding to the sample size 256
j0<-5  # the primary resolution level
ywd<-wd(ynoise,filter.number=8,family="DaubLeAsymm",verbose=F)
sigma<-median(abs(accessD(ywd,J)-median(accessD(ywd,J))))/.6745
ywd<-wd(ynoise/sigma,filter.number=8,family="DaubLeAsymm",verbose=F)
th<-numeric( )
for(i in j0:J){
  z<-accessD(ywd,i)
```

```
  s2<-(sum(z^2)-2^i)/2^i
  if(s2<=i^1.5/sqrt(2^i))th[i]<-sqrt(2*log(2^i))
  else                   th[i]<-sureth(2^i,z)}
yrecon<-wr(threshold.wd(ywd, levels=j0:J, policy="manual",value=th[j0:J],
  type="soft", boundary=T))*sigma
```

APPENDIX B: MATHEMATICAL PROOFS

To prove Proposition 2.3.1, we need the following result.

Theorem (Theorem 10.3.1 of Berman 1992). Let $\{Z_n(t)\}$ be a family of real separable Gaussian processes with mean 0, and let $\{Z(t)\}$ be a similar process such that the finite-dimensional distributions of $\{Z_n(t)\}$ converge to those of $\{Z(t)\}$ for $n \to \infty$. If for some $t \in [0,1], \sup_n EZ_n^2(t) < \infty$, and $\lim_{h \downarrow 0} \sup_n Q_n(h)(\log h^{-1})^{1/2} = 0$, where $Q_n(h) = \varphi_n(h) + (2+\sqrt{2}) \int_1^\infty \varphi_n(hp^{-y^2})\, dy, \varphi_n(h) = \max_{|s-t| \le h, s,t \in [0,1]} [E(Z_n(t) - Z_n(s))^2]^{1/2}$ and p is an integer not smaller than 2, then the measure on $C[0,1]$ induced by $\{Z_n(t)\}$ converges weakly, for $n \to \infty$, to that induced by $\{Z(t)\}$.

Proof of Proposition 2.3.1

It is easy to verify that

$$|R_i(s,t) - R_i(s,t')| \le C|t - t'|,$$

for

$$s,t,t' \in [0,1], \qquad i = 1,2 \tag{A.1}$$

and some constant C; that is, both R_1 and R_2 are Lipschitz continuous.

Using this, it is easy to prove that $\operatorname{cov}(V_{[ns]}, V_{[nt]})$ converges to $G(s,t)$, and hence the covariance function of Z_n also converges to G, the covariance function of Z. Because Z_n and Z are Gaussian processes, all of the finite-dimension distributions of Z_n converge to those of Z as well.

Again using (A.1), it can be shown that

$$|1 - \operatorname{cov}(V_{[ns]}, V_{[nt]})| \le \frac{C|[ns] - [nt]|}{n}, \quad \text{for} \quad s,t \in [0,1],$$

where the constant C does not depend on n and is not necessarily the same as in (A.1). Then it can be verified that

$$E((Z_n(s) - Z_n(t))^2) \le C|s - t|.$$

Therefore, $\varphi_n(h) = \max_{|s-t| \le h}(E(Z_n(s) - Z_n(t))^2)^{1/2} \le C\sqrt{h}$, where C is independent of n as well. The rest is just an application of the theorem of Berman, G_1 and G_2 were obtained by plugging R_1 and R_2 of (6) and (7) into (8), through some tedious algebraic manipulation with the help of a symbolic computation code.

Proof of Proposition 2.3.2

Note that $G_i(s,t) = 1 - C_i(s-t)^2 + o((s-t)^2)$ for $i = 1,2$. The conclusion is a direct corollary of theorem 9 of Albin (1990).

[Received June 1995. Revised May 1996.]

REFERENCES

Abramovich, F., and Steinberg, D. (1996), "Improved Inference in Nonparametric Regression Using L_k-Smoothing Splines," *Journal of Statistical Planning Inference*, 49, 327–341.

Albin, J. M. P. (1990), "On Extremal Theory for Stationary Processes," *Annals of Probability*, 18, 92–128.

Bates, D., Lindstrom, M., Wahba, G., and Yandell, B. (1987), "GCVPACK-Routines for Generalized Cross-Validation," *Communication in Statistics, Part B—Simulation and Computation*, 16, 263–297.

Berman, S. M. (1992), *Sojourns and Extremes of Stochastic Processes*, Pacific Grove, CA: Wadsworth & Brooks/Cole.

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

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

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995), "Wavelet Shrinkage: Asymptopia?" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 57, 301–370.

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*, Ser. 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," *Technometrics*, 31, 3–39.

Gallant, R. A. (1987), *Nonlinear Statistical Models*, New York: Wiley.

Gao, F. (1993), "On Combining Data From Multiple Sources With Unknown Relative Weights (Thesis)," Technical Report 902, University of Wisconsin at Madison, Dept. of Statistics.

von Golitschek, M., and Schumaker, L. (1990), "Data Fitting by Penalized Least Squares," in *Algorithms for Approximation II*, eds. J. C. Mason and M. G. Cox, New York: Chapman and Hall, pp. 210–227.

Gu, C., and Wahba, G. (1993a), "Semiparametric Analysis of Variance With Tensor Product Thin Plate Splines," *Journal of the Royal Statistical Society*, Ser. B, 55, 353–368.

——— (1993b), "Smoothing Spline ANOVA With Component-Wise Bayesian Confidence Intervals," *Journal of Computational and Graphical Statistics*, 2, 97–117.

Hastie, T. (1989), Discussion of "Flexible Parsimonious Smoothing and Additive Modeling" by J. H. Friedman and B. W. Silverman, *Technometrics*, 31, 25.

Hutchinson, M. F., Kalma, J., and Johnson, M. (1984), "Monthly Estimates of Wind Speed and Wind Run for Australia," *Journal of Climatology*, 4, 311–324.

Hutchinson, M. F., and de Hoog, F. R. (1985), "Smoothing Noisy Data With Spline Functions," *Numerische Mathematik*, 47, 99–106.

Jones, P. D., Raper, S. C. B., Cherry, B. S. G., Goodess, C. M., Wigley, T. M. L., Santer, B., Kelly, P. M., Bradley, R. S., and Diaz, H. F. (1991), "An Updated Global Grid Point Surface Air Temperature Anomaly dataset: 1851–1988," Environmental Sciences Division Publication No. 3520, U.S. Department of Energy.

Luo, Z. (1996), "Backfitting in Smoothing Spline ANOVA With an Application to a Global Historical Temperature dataset," unpublished Ph.D. dissertation, University of Wisconsin at Madison, Dept. of Statistics.

Owen, A. (1991), Discussion of "Multivariate Adaptive Regression Splines" by J. Friedman, *The Annals of Statistics*, 19, 102–112.

Seber, G. A. F. (1977), *Linear Regression Analysis*, 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.

Stone, C. J., Hansen, M., Kooperberg, C., and Truong, Y. K. (1995), "Polynomial Splines and Their Tensor Products in Extended Linear Modeling," Technical Report 437, University of California, Berkeley, Dept. of Statistics.

Wahba, G. (1980), "Spline Bases, Regularization, and Generalized Cross-Validation for Solving Approximation Problems With Large Quantities of Noisy Data," in *Approximation Theory III*, ed. W. Cheney, Orlando, FL: Academic Press, pp. 905–912.

——— (1981), "Spline Interpolation and Smoothing on the Sphere," *SIAM Journal of Scientific Statistical Computing*, 2, 5–16; Erratum (1982), 3, 385–386.

——— (1982), "Vector Splines on the Sphere, With Application to the Estimation of Vorticity and Divergence From Discrete, Noisy Data," in *Multivariate Approximation Theory*, Vol. 2, eds. W. Schempp and K. Zeller, Basel: Birkhauser Verlag, pp. 407–429.

——— (1990), *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59, Philadelphia: Society of Industrial and Applied Mathematics.

——— (1995), Discussion of "Wavelet Shrinkage: Asymptopia?" by D. L. Donoho et al., *Journal of the Royal Statistical Society*, Ser. B, 57, 360–361.

Wahba, G., Wang, Y., Gu, C., Klein, R., and Klein, B. (1994), "Smoothing Spline ANOVA for Exponential Families, With Application to the Wisconsin Epidemiological Study of Diabetic Retinopathy," Technical Report 940, University of Wisconsin at Madison, Dept. of Statistics.

Weber, R., and Talkner, P. (1993), "Some Remarks on Spatial Correlation Function Models," *Monthly Weather Review*, 121, 2611–2617.