DEPARTMENT OF STATISTICS
University of Wisconsin
1210 West Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 947

June 5, 1995

Hybrid Adaptive Splines[1]

by

Zhen Luo and Grace Wahba

[1] This research was supported in part by National Science Foundation Grant DMS 9121003 and National Aeronautics and Space Administration Grant NAGW-2961. E-mail zhen@stat.wisc.edu, wahba@stat.wisc.edu

Hybrid Adaptive Splines

Zhen Luo, Grace Wahba

Department of Statistics, University of Wisconsin-Madison

June 5, 1995

Abstract. An adaptive spline method for smoothing is proposed which 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 in this paper by an application to smoothing temperature data on the globe. The performance of this method 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 and phrases: Smoothing, spatial adaptability, splines, stepwise regression, the inflated degrees of freedom for an adaptively chosen basis function, MARS.

1 Introduction

Spatially adaptive smoothing, or function estimation, methods which 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 which were shown to have desirable spatial adaptability by both theoretical arguments and simulation study. Traditionally, there have been two techniques to address this problem of spatial adaptability. One 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, (equivalently, adaptively choose a set of spline basis functions for regression). Recent examples in the first category include Fan and Gijbels (1995), Abramovich and Steinberg (1995), see also Wahba (1995). Much other recent research in this area is noted in the lengthy discussion to Donoho, et. al. (1995).

Probably the best known method in the second category is the MARS procedure (Friedman (1991)). MARS was designed primarily for estimating multivariate functions efficiently. Its procedure for choosing basis functions adaptively will tend to choose more basis functions in data-dense regions where the underlying true function has more structure.

In this paper we combine some of the features of MARS and of traditional smoothing splines to obtain a hybrid smoothing procedure (Hybrid Adaptive Splines, HAS) which may be implemented with large data sets and displays a desirable form of spatial adaptability when the underlying function is spatially inhomogeneous in its degree of complexity. The basis functions which are chosen as a subset of the basis functions occuring naturally in smoothing splines, are selected one basis function at a time, using a forward stepwise regression procedure. A 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. The procedure is explained in Section 2, along with some theoretical results on the appropriate IDF factor. The choice of IDF factor arises in MARS and other regression spline procedures and we discusses how our results might apply to MARS.

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, for example Wahba(1980), Hutchinson, Kalma and Johnson (1984). 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 occuring 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.

There are several features of this procedure which we think worth mentioning. First, it is well suited to highly unequally spaced data. Second, it extends in a straightforward way to the general penalized likelihood setup as discussed in, for example, Wahba(1990), and in particular to the Smoothing Spline ANOVA setup as in Gu and Wahba(1993a,b). Some examples are given in Luo(1994). It can be used in the context of splines on the sphere, which has the potential for wide application in meteorological and environmental studies. An application to the interpolation and smoothing of global winter surface temperature is used to illustrate this application in Section 4. Finally, based on simulated examples including the four used in Donoho and Johnstone (1994), it seems fair to say that this procedure, in terms of both mean square error 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 cases non-adaptive smoothing methods such as smoothing splines with a global smoothing parameter perform better on the average than adaptive methods, HAS's performance is close, while wavelet methods seem to need further refinement in order to get close results. These comparisons are shown in Section 3.

2 Hybrid Adaptive Splines

2.1 Smoothing Splines

Let

$$y_i = f(x_i) + e_i, i = 1, 2, ..., n$$

where $x_i \in [0, 1]$, and the $\{e_i\}$ are i.i.d. $N(0, \sigma^2)$, and f is "smooth", more precisely, suppose f is in the Sobolev space $\mathcal{W}_2[0, 1] = \{f : f, f'$ absolutely continuous, $f'' \in \mathcal{L}_2\}$. The traditional cross-validated cubic smoothing spline estimate of f is the solution to the 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 generalized cross-validation (GCV), see Wahba (1990). It is well known 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, $\int_0^1 (\frac{d^2}{dx^2}(\sum_{i=1}^n c_i R(x; x_i)))^2 dx = \sum_{i,j=1}^n c_i c_j R(x_i; x_j)$.

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

$$\arg\min_{c,d} \frac{1}{n}||\mathbf{y} - (\mathbf{Td} + \mathbf{\Sigma c})||^2 + \lambda \mathbf{c}'\mathbf{\Sigma c} \tag{3}$$

where $\mathbf{y} = (y_1, ..., y_n)'$, $\mathbf{\Sigma}$ 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 $\mathbf{c} = (c_1, c_2, ..., c_n)'$.

2.2 Hybrid Adaptive Spline procedure

Instead of minimizing (1) in $\mathcal{W}_2[0, 1]$, we will minimize it in span ϕ_1, ϕ_2 plus a specially selected subset of the n basis functions $R(\,\cdot\,; x_i), i = 1, 2, ..., n$, chosen in a forward stepwise manner. Having chosen ϕ_1, ϕ_2 and $R(\,\cdot\,; x_{i_l}), l = 1, 2, ..., k - 3$, we chose the k-th basis function to maximize the reduction in the residual sum of squares (RSS). This is done 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 there are k basis vectors in the subset already and they are stored in a n by k matrix $\mathbf{X}$. We first do a QR decomposition of $\mathbf{X}$: $\mathbf{Q}'\mathbf{X} = (\mathbf{R}_{k\times k}\ \mathbf{0}_{k\times(n-k)})'$, 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 $\mathbf{X}) = \mathbf{tt}'$, and $RSS(\mathbf{y}$ regressed on $\mathbf{X}$ and $\mathbf{u}) = \mathbf{tt}' - (\mathbf{st}')^2/\mathbf{ss}'$. Each time among those unselected the basis function making largest RSS deduction, i.e. the largest $(\mathbf{st}')^2/\mathbf{ss}'$, 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, and 338-341).

The number of basis functions will be chosen by minimizing a similar GCV score as implemented in MARS. The GCV score for k selected basis functions is $GCV(k) = RSS/(1-(2+(k-2)\,\text{IDF})/n)^2$ where IDF is the inflated degrees of freedom factor 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 will be minimized over $k = 2, 3, ..., q$ for some (safe) upper limit q. Obviously IDF should be larger than 1. Friedman recommends 3 as a generally appropriate choice (default) in MARS and most IDF's chosen by cross validation and simulation studies show that it 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 explanation for this difference are discussed in Section 2.3. For all the examples in this paper it is fixed at 1.2.

The final step of penalized regression is done by subroutine `dsnsm` in *GCVPACK* developed by Bates, et. al. (1989). Note using only a subset of basis functions in representation (2)

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

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 sub-matrices, i.e., the first Σ replaced by $(R(x_i; x_{i_l}))$, $i = 1, 2, ..., n$, and $l = 1, 2, ..., k$, and the second Σ replaced by $(R(x_{i_l}; x_{i_{l'}}))$, $l, l' = 1, 2, ..., k$.

For the simulated Examples 1-5 of Section 3 with sample size 2048, and $q = 150$, it took about 10 minutes to get a HAS fit on our DEC Alpha machine. For the example in Section 4 with sample size 725 and $q = 500$, it took about 3 minutes.

Note that this procedure can be applied to any of the spline models in Wahba (1990).

2.3 The 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 IDF is, the fewer basis functions we will put in. Another reason we want to study this problem is to explain why Friedman chose 3 as a general good choice in MARS while our experience shows a IDF less than 2 is better in HAS.

Let us consider a simplified version of our problem, a regression model:

$$y_i = aR(x_i, t) + e_i, i = 1, 2, ..., n. \tag{5}$$

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

$$\begin{aligned} R_1(s,t) &= ((|s-t| - \frac{1}{2})^2 - \frac{1}{12})/2 & (6)\\ R_2(s,t) &= (-(|s-t| - \frac{1}{2})^4 + \frac{1}{2}(|s-t| - \frac{1}{2})^2 - \frac{7}{240})/24 & (7)\end{aligned}$$

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.

The difference between the residual sum of squares (RSS) of the least square 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; and the expectation of this difference, when the data are drawn from H_0, is the degrees of freedom of the model (5). Since 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, 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

$$RSS(modelH_0) - RSS(model(5)) = \mathbf{y}'\mathbf{y} - \min_a S(a,t) = \frac{(\sum R(x_j,t)y_j)^2}{\sum R(x_j,t)^2}$$

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

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

which is greater than or equal to any of $V^2(x_i)$, therefore its expectation is greater than or equal to that of $V^2(x_i)$, i.e. 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 (say, Gallant (1987)), we know 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 in Friedman and Silverman (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_3^2. This is supported also 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

$$Z_n(t) = (1 - nt + [nt])V_{[nt]} + (nt - [nt])V_{[nt]+1}, \text{ for } t \in [0,1]$$

where $V_i = V(x_i)$ for $i = 1,2,...,n$, 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 multi-normal distributed, $E(V_i) = 0$, $Var(V_i) = 1$, and

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

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

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

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.

In order to prove the weak convergence of processes Z_n, we need the following result.

Theorem (Theorem 10.3.1 in Berman (1992)) Let $\{Z_n(t)\}$ be a family of real separable Gaussian processes with mean 0, and $\{Z(t)\}$ 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 $t \in [0,1]$, $\sup_n E\ Z_n^2(t) < \infty$, and $\lim_{h\downarrow 0}\sup_n Q_n(h)(\log\ h^{-1})^{1/2} = 0$, where $Q_n(h) = \varphi_n(h) + (2+\sqrt{2})\int_1^\infty \varphi_n(h\,p^{-y^2})dy$, $\varphi_n(h) = \max_{|s-t|\le h,\ s,t\in[0,1]}[E(Z_n(t)-Z_n(s))^2]^{1/2}$ 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 $\{Z(t)\}$.

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 covariance functions,

$$\begin{aligned} G_1(s;t) &= 1 - 30(t-s)^2 + 60(t-s)^3 - 30(t-s)^4 && (9)\\ G_2(s;t) &= 1 - 20(t-s)^2 + 70(t-s)^4 - 140(t-s)^6 + 120(t-s)^7 - 30(t-s)^8 && (10)\end{aligned}$$

respectively.

Proof: It is easy to verify that

$$|R_i(s,t) - R_i(s,t')| \le C|t-t'|,\ \text{for } s,t,t' \in [0,1], i = 1,2 \tag{11}$$

and some constant C, i.e, both R_1 and R_2 are Lipschitz continuous.

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

Again using (11), it can be shown that

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

where the constant C does not depend on n and is not necessarily the same as in (11). 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 too. The rest is just an application of the above theorem of Berman. G_1 and G_2 were gotten 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. ■

Proposition 2.3.1 tells us that the asymptotic distribution of the model sum of squares of (5), $\max_i V^2(x_i)$, 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 uses linear spline basis functions in MARS instead of cubic spline basis functions which 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 has also 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,

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

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.

Proof: 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 in Albin (1990). ■

In some sense this result tells us the tail probability of $\sup_{t \in [0,1]} |Z(t)|$, but since h has to 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 $[0, 1-\epsilon]$ 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 $\sup_{t \in [0,1-\epsilon]} |Z(t)|^2$ whose tail probability $P\{\sup_{t \in [0,1-\epsilon]} |Z(t)|^2 > u\}$ by Proposition 2.3.2 can be approximated by $(1-\epsilon)exp(-u/2)\sqrt{2C/\pi}$. Since the process corresponding to linear spline basis has a larger C (which is 30) than the one corresponding to cubic spline basis, the model sum of squares has a larger tail probability, hence larger variation too. That means 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 choice of 1.2 for HAS and 3 for MARS.

3 Simulation study

In this section, simulated examples are used to examine the performance of HAS compared with other procedures (MARS, Wavelet Shrinkage, and smoothing splines (SS)).

The first five examples, which show strong spatial inhomogeneity, are taken from Donoho and Johnstone (1994). We also include two examples from Fan and Gijbels (1995) which 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 too. Gaussian noise are added such that $SD(f)/\sigma$ as an approximate measure of signal-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 that σ has been replaced by the median standard deviation. More information about these examples is given in Table 1. The pseudo standard normal random number generator we used is `rnor`, a Fortran subroutine from *CMLIB*.

Of all the Wavelet Shrinkage methods proposed by Donoho and Johnstone the SUREShrink method (Donoho and Johnstone (1995)) is chosen for our comparisons, since it has a level-dependent threshold feature and is better on the average than RiskShrink (Donoho and Johnstone (1994)) in our experience. The "primary resolution level" is chosen as 5 as used by Donoho and Johnstone (1994). The computation is done by the software *wavethresh* developed by Nason and Silverman (1994) in S-PLUS. The family of wavelets is chosen as *DaubLeAsymm* with *filter number* 8. The S-PLUS commands we used are given in the appendix.

Example	f	σ	sample	$SD(f)/\sigma$	number of
			size		replicates
1	DJ(1994)'Blocks * 3.5	1.0	2048	6.92	31
2	DJ(1994)'Bumps * 4.5	1.0	2048	6.93	31
3	DJ(1994)'Heavisine * 2.2	1.0	2048	6.54	31
4	DJ(1994)'Doppler * 22	1.0	2048	6.36	31
5	DJ(1994)'Doppler * 22	$exp(x)/1.648$	2048	6.36	31
6	$sin(2(4x-2)) + 2exp(-16x^2)$	0.3	256	2.80	400
7	$(4x-2) + 2exp(-16x^2)$	0.4	256	3.16	400

Table 1: *Summary of examples.*

The maximum number of basis functions (q) in both HAS and MARS is set at 150 for Examples 1-5, and 60 for Examples 6-7. The minimum span parameter in MARS is set at zero in all examples. The Fortran routines to compute HAS estimates are available upon request from Zhen Luo.

Example	HAS	SS	SUREShrink	MARS
1	.266(.023)	.546(.023)	.398(.049)	-
2	.082(.012)	.124(.010)	.167(.015)	-
3	.036(.011)	.075(.005)	.062(.007)	.150(.014)
4	.060(.011)	.205(.011)	.145(.013)	-
5	.051(.023)	.232(.014)	.149(.013)	-
6	.007(.006)	.006(.003)	.018(.004)	.007(.004)
7	.012(.011)	.010(.005)	.042(.012)	.012(.007)

Table 2: *Median of MSE and difference of first and third quartiles of MSE (in paratheses).*

SS estimates are computed using the code *GCVSPL* in Fortran by Woltring with the smoothing parameters chosen by GCV. The codes *mars3.5* for MARS, *wavethresh*, and *CMLIB* can be obtained from `statlib`. *GCVPACK, GCVSPL* can be obtained from `netlib`.

The median performances in terms of mean square 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 the first five examples, both HAS and SUREShrink exhibit spatial adaptability, while 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, and the same remark holds for HAS.

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 the other methods are trying to do different amounts of smoothing at different locations, hence trying to estimate more than a single smoothing parameter. MARS essentially did not give sensible answers in the four Examples 1,2,4,5. The smallest of the missing entries of Table

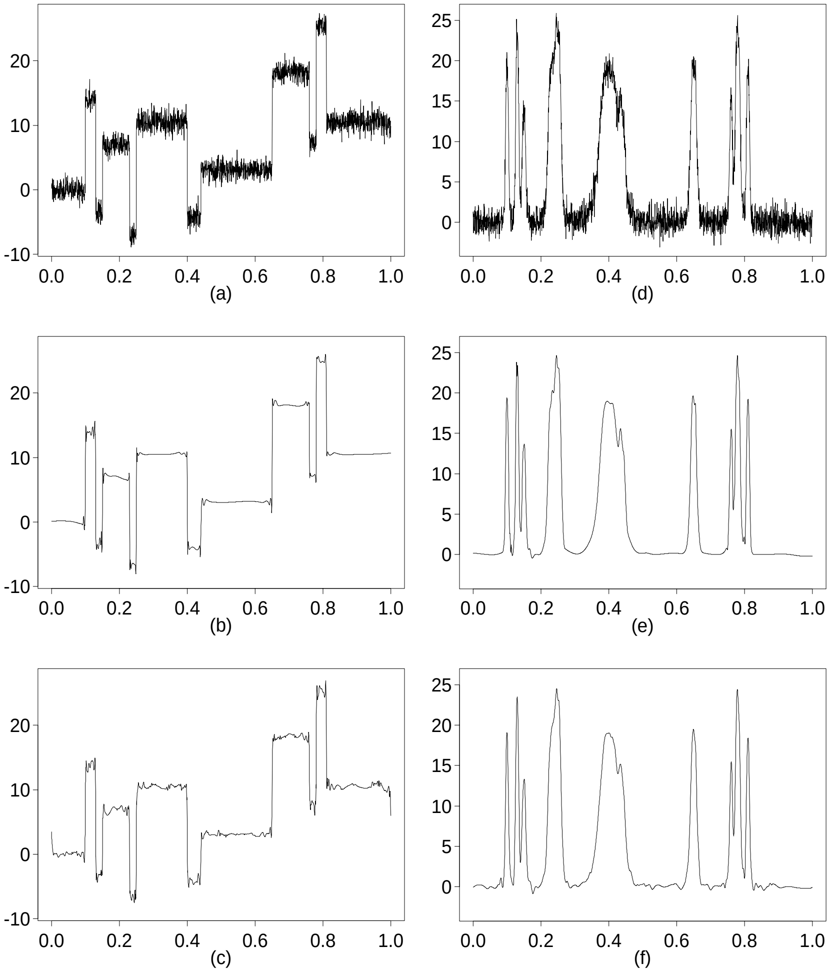

Figure 1: *Example 1, Blocks. (a) one copy of data, (b) HAS fit with median MSE, (c) SUREShrink fit with median MSE. Example 2, Bumps. (d) one copy of data, (e) HAS fit with median MSE, (f) SUREShrink fit with median MSE.*

2 was over 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. SUREShrink was further behind. However a lower "primary resolution level" (we chose level 5 because it gave the best performance overall) 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 variability in the MSE as compared to SS, particularly in Examples 3, 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 we use 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.

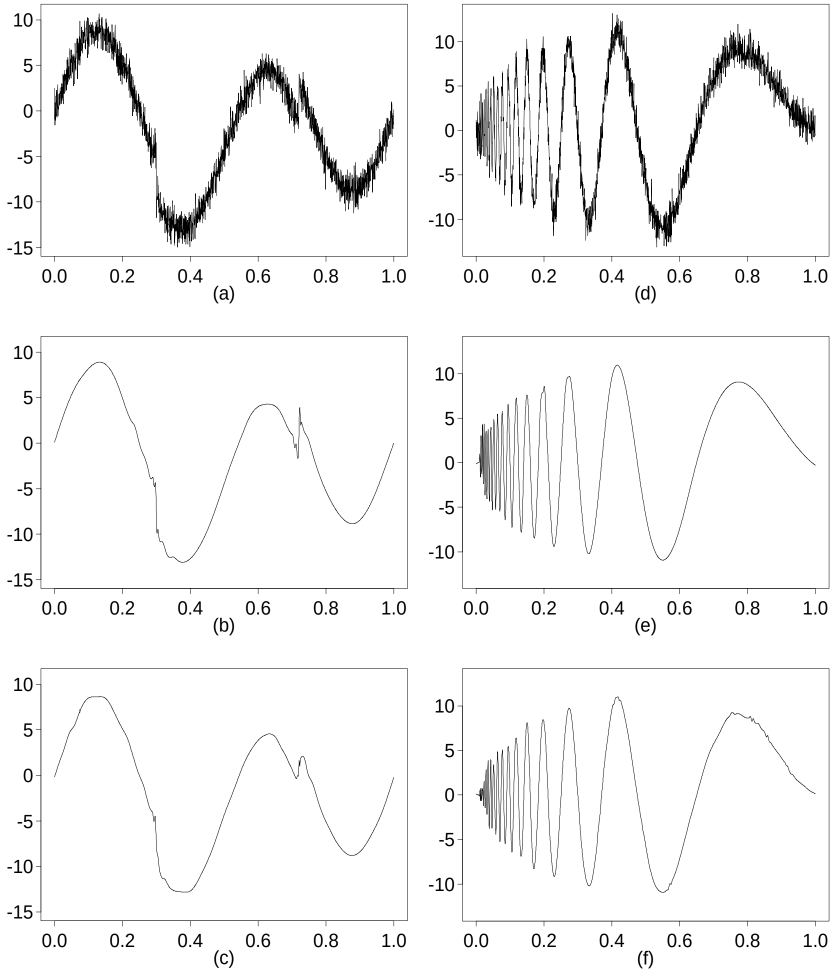

Figure 2: *Example 3, Heavisine. (a) one copy of data, (b) HAS fit with median MSE, (c) SUREShrink fit with median MSE. Example 4, Doppler. (d) one copy of data, (e) HAS fit with median MSE, (f) SUREShrink fit with median MSE.*

4 Application to a smoothing problem on the sphere

We illustrate HAS's applicability to multivariate problems in this section using a smoothing problem in meteorology. From a global monthly surface temperature data archive developed by Jones et. al. (1991), we extracted all 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 of the previous year, January and February monthly temperatures. The total of 725 stations with such records are distributed very irregularly on the sphere.

A spline on the sphere estimate is defined in Wahba (1981) as the solution of the following

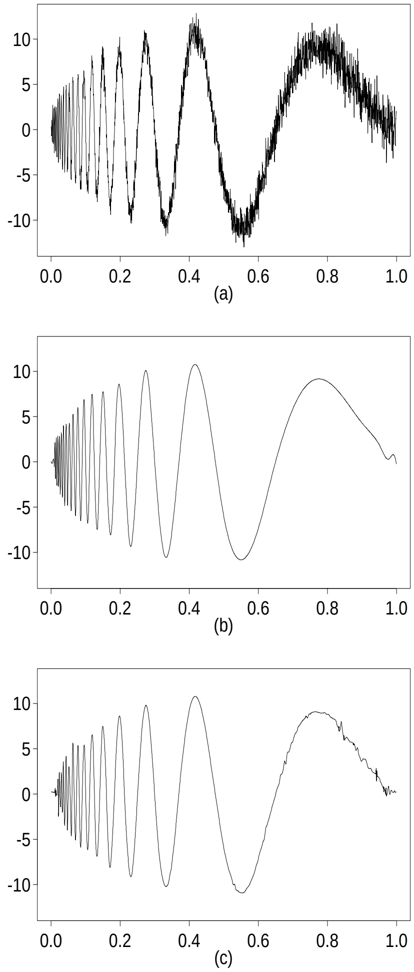

Figure 3: *Example 5, Doppler2. (a) one copy of data, (b) HAS fit with median MSE, (c) SUREShrink fit with median MSE.*

optimization problem:

$$\arg\min_f \frac{1}{n}\sum_{i=1}^{n}(y_i - f(P_i))^2 + \lambda \int_S (\Delta^{m/2} f)^2 dP \tag{12}$$

where $P = (latitude, longitude)$ is a point on the sphere S, P_i is the location of the i-th station, Δ is the Laplacian on the sphere and f is in the Sobolev space $\mathcal{H}_m(S) = \{f : f \in \mathcal{L}_2(S), \Delta^{m/2} f \in \mathcal{L}_2(S)\}$. It is shown in that paper that the minimizer of (12) has a representation of the form

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

where $Q_m(P; P')$ $(m = 1, 2, ...)$ 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

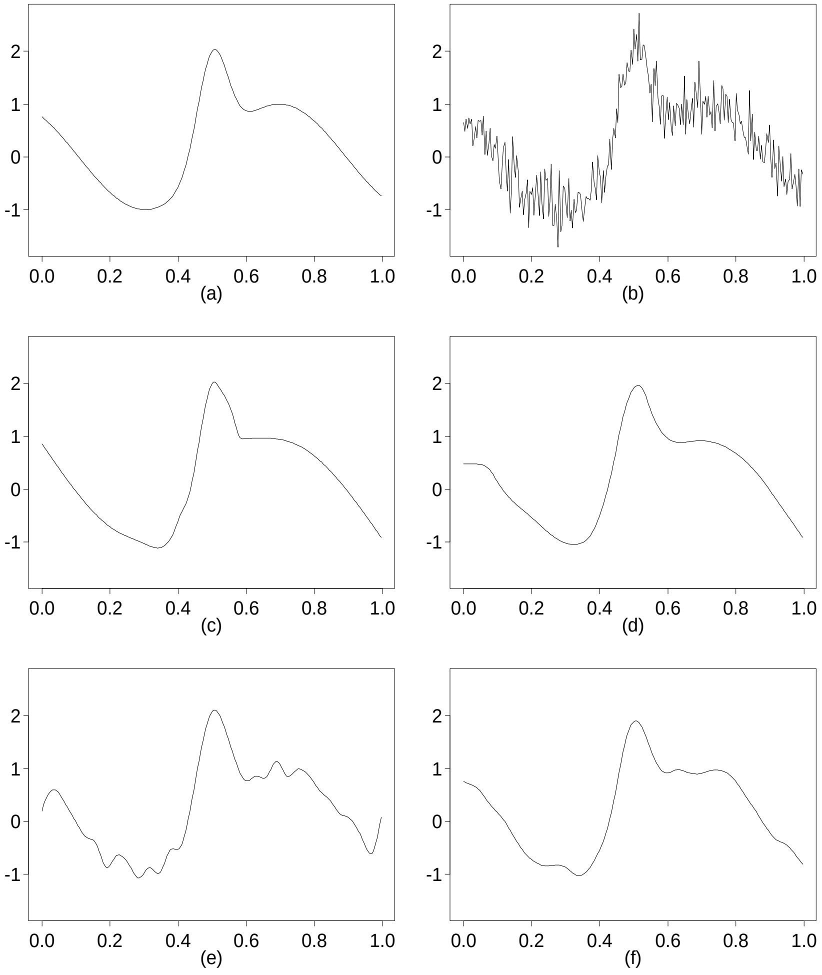

Figure 4: *Example 6, (a) true function, (b) one copy of data, (c) HAS fit with median MSE, (d) MARS fit with median MSE, (e) SUREShrink fit with median MSE, (f) SS fit with median MSE.*

approximating the Q_m and with closed form expressions, are given in Wahba(1981), Equation (3.3) and (3.4). We will use $R_2(P;P')$ from that paper, 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 $\{\phi_1, R(\,\cdot\,;P_i),\ \text{for}\ i = 1, 2, ..., n\}$, where $\phi_1(P) \equiv 1$.

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 while 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 exist enough data 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

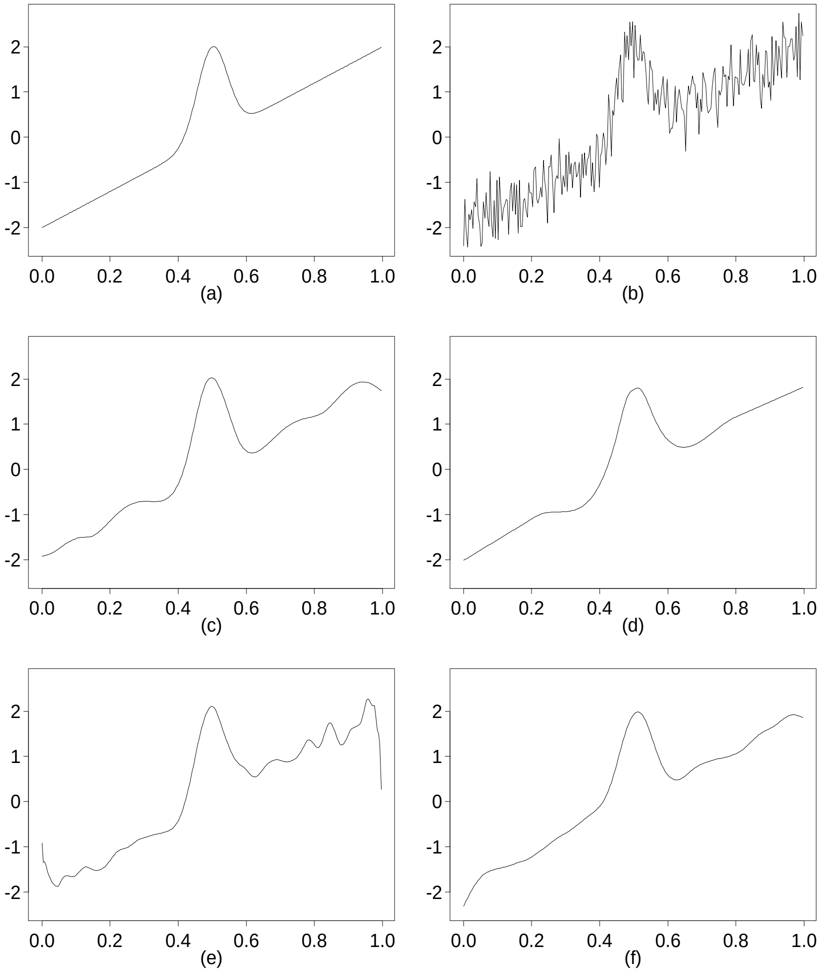

Figure 5: *Example 7, (a) true function, (b) one copy of data, (c) HAS fit with median MSE, (d) MARS fit with median MSE, (e) SUREShrink fit with median MSE, (f) SS fit with median MSE.*

ranges were considerably smoother. Simulated examples we have studied show the same kind of spatial adaptability as the one-dimensional examples. We remark that our choice of $R = R_2$ was made because we suspected that it was 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, for example, as given in Weber and Talkner (1993). Gao (1993) and Wahba (1982) have used reproducing kernels on the sphere based on historical meteorological information. Work in this direction is in progress.

5 Discussion

HAS may be applied to ANOVA in functions spaces (Wahba (1990), Gu and Wahba (1993a,b), Wahba, Wang, Gu, Klein, and Klein (1994)). Since their estimates have a representation (2), the extension is immediate. Extensions to global winter temperature as a function of year and space

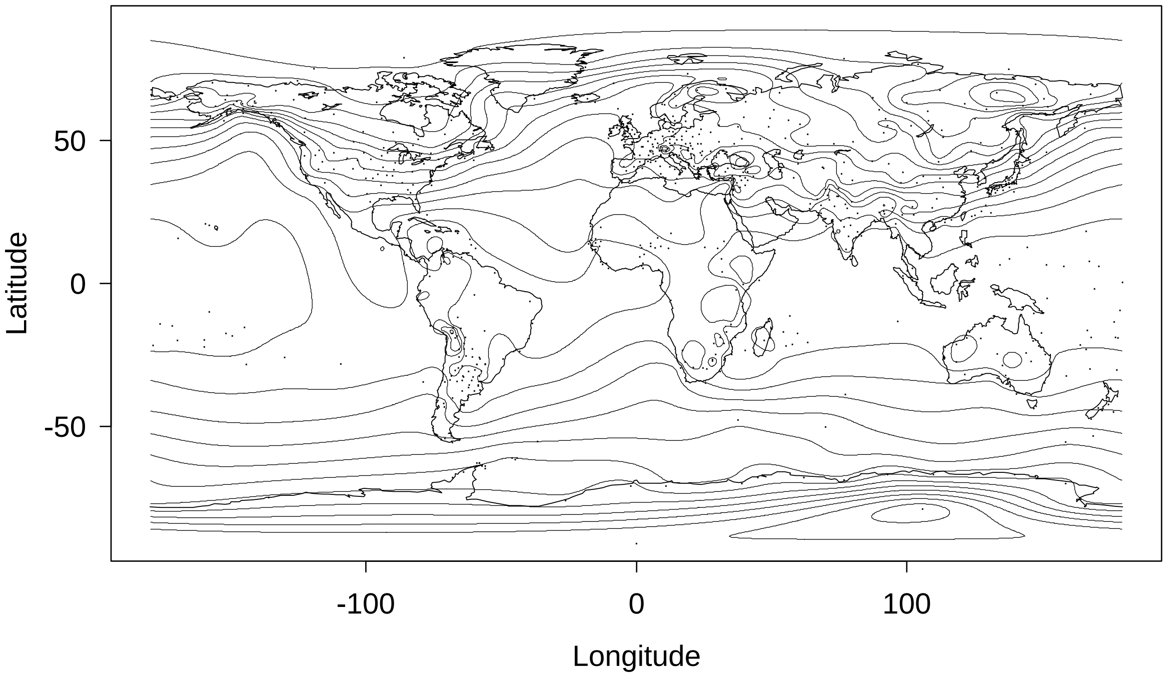

Figure 6: *Example in Section 4. Contour plot of HAS fit of 1981 global winter temperature. Dots are the locations of the recording stations.*

are under study.

Finally a further comment on the reasons why we want a penalized regression step in HAS. Basically, choosing the number of basis functions by 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 generally improve the MSE, though usually just a little, in all the seven simulated examples of Section 3. 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) gets bigger, the regression problem becomes more and more ill-conditioned, which makes its numerical computation less and less stable. The basis functions 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 as a remedy for this.

Appendix: The list of S-PLUS commands used to compute SUREShrink estimates

```
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)
```

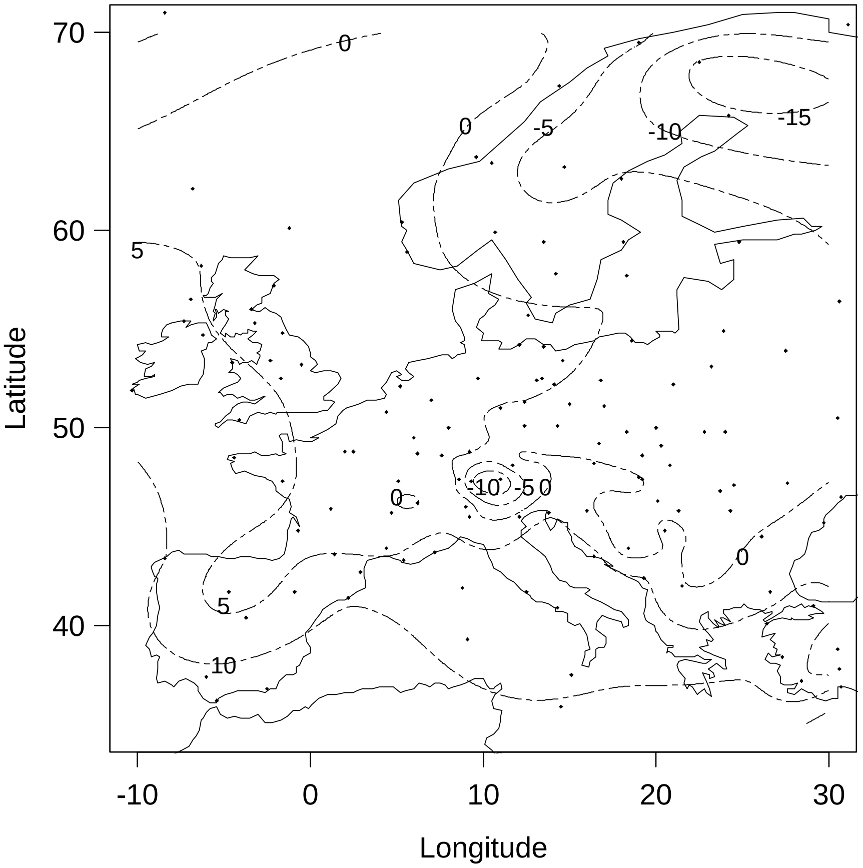

Figure 7: *Enlarged European part of Figure 6. Dots are the locations of the recording stations. Dashed lines are the contour lines of HAS fit.*

```
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
```

References

Abramovich, F. and D. Steinberg (1995). Improved inference in nonparametric regression using

L_k-smoothing splines. *J. Statistical Planning Inference*, to appear.

Albin, J.M.P. (1990). On extremal theory for stationary processes. *Annals of Probability*, Vol. 18, No. 1, 92-128.

Bates, D., Lindstrom, M., Wahba, G., and Yandell, B. (1987). GCVPACK-routines for generalized cross validation. *Comm. Statist. B—Simulation Comput.*, Vol. 16, 263-297.

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

Donoho, David L. and Johnstone, Iain M. (1994). Ideal spatial adaptation by Wavelet Shrinkage. *Biometrika*, Vol. 81, No. 3, 425-455.

Donoho, David L. and Johnstone, Iain M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Ass.*, to be published.

Donoho, David L., Johnstone, Iain M., Kerkyacharian, G. and Picard, D. (1995). Wavelet Shrinkage: asymptopia? (with discussion) *J. Royal Statistical Soc. Ser. B*, Vol. 57, No. 2, 301-370.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Royal Statistical Soc. Ser. B*, Vol. 57, No. 2, 371-394.

Friedman, Jerome H. (1991). Multivariate Adaptive Regression Splines. (with discussion) *Annals of Statistics*, Vol. 19, No. 1, 1-141.

Friedman, Jerome H. and Silverman, Bernard W. (1989). Flexible Parsimonious Smoothing and Additive Modeling. *Technometrics*, Vol. 31, No. 1, 3-39.

Gallant, Ronald A. (1987). *Nonlinear Statistical Models.* John Wiley & Sons, New York.

Gao, F. (1993). On combining data from multiple sources with unknown relative weights (Thesis). *Technical Report No. 902*, Dept. of Statistics, University of Wisconsin at Madison.

Gu, C. and Wahba, G. (1993a). Semiparametric analysis of variance with tensor product thin plate splines. *J. Royal Statistical Soc. Ser. B*, Vol. 55, 353-368.

Gu, C. and Wahba, G. (1993b). Smoothing spline ANOVA with component-wise Bayesian confidence intervals. *Journal of Computational and Graphical Statistics*, Vol. 2, 97-117.

Hutchinson, M.F., J. Kalma and M. Johnson (1984). Monthly estimates of wind speed and wind run for Australia. *J. Climatology*, Vol. 4, 311-324.

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

Luo, Z. (1994). Hybrid Adaptive Splines. The paper submitted for the Ph.D preliminary examination, Department of Statistics, University of Wisconsin at Madison.

Owen, A. (1991). Discussion to J. Friedman, Multivariate Adaptive Regression Splines. *Annals of Statistics*, Vol. 19, No. 1, 102-112.

Seber, G.A.F. (1977). *Linear Regression Analysis.* John Wiley & Sons, New York.

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. by W. Cheney, 905-912, Academic Press.

Wahba, G. (1981). Spline interpolation and smoothing on the sphere. *SIAM J. Sci. Stat. Comput.*, Vol.2, No.1, 5-16. (Erratum. (1982), *SIAM J. Sci. Stat. Comput.*, Vol.3, No.3, 385-386.)

Wahba, G. (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, Ed. by W. Schempp and K. Zeller, 407-429, Birkhauser Verlag.

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

Wahba, G. (1995). Discussion to D. L. Donoho, et. al., Wavelet Shrinkage: asymptopia? *J. Royal Statistical Soc. Ser. B*, Vol. 57, No. 2, 360-361.

Wahba, G., Y. Wang, C. Gu, R. Klein and B. Klein (1994). Smoothing Spline ANOVA for Exponential Families, with Application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Technical Report No. 940*, Department of Statistics, University of Wisconsin at Madison.

Weber, R. and P. Talkner (1993). Some remarks on spatial correlation function models. *Monthly Weather Review*, Vol. 121, 2611-2617.