1,3

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

TECHNICAL REPORT NO. 872

December 1990

Multivariate Model Building With Additive, Interaction and Tensor Product Thin Plate Splines [1]

by

Grace Wahba

[1] Prepared for the Proceedings of the International Conference on Curves and Surfaces, Chamonix Mont-Blanc, France, June 1990, Larry Schumaker and Pierre-Jean Laurent, Eds.

MULTIVARIATE MODEL BUILDING WITH ADDITIVE, INTERACTION AND TENSOR PRODUCT THIN PLATE SPLINES

GRACE WAHBA

December 1990

Abstract

We review some recent work, primarily with Chong Gu, on multivariate model building using tensor products and sums of polynomial and thin plate smoothing splines. The goal of the work is to provide a family of predictive response models suitable for use with multidimensional empirical scattered, noisy response data from medical, economic, demographic, geophysical and other sources. We will discuss construction of the models, based on elementary properties of reproducing kernel Hilbert spaces, and mention some practical computational problems and existing software. We will describe several possible model selection methods, whereby the observational data is used to decide on an appropriate level of complexity of the model. Some approaches to making accuracy statements are also noted.

1 INTRODUCTION

Suppose we observe

$$y_i = f(t(i)) + \epsilon(i), \qquad i = 1, \cdots, n$$

where $t(i) \in R^d$ and the $\epsilon(i)$'s are independent normally distributed random variables with mean 0 and common unknown variance σ^2. We want to estimate $f \in \mathcal{H}$ from the data. Here $\mathcal{H}$ is a reproducing kernel Hilbert space, that is, a Hilbert space in which all the evaluation functionals are bounded, see [1]. We suppose that $\mathcal{H}$ has a decomposition $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0$ has dimension $M < n$ and satisfies some other conditions. What might be called a generalized smoothing spline estimates f as the minimizer in $\mathcal{H}$ of

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(t(i))^2 + \lambda \parallel P_1 f \parallel^2 \qquad (1)$$

where P_1 is the orthogonal projection in $\mathcal{H}$ onto $\mathcal{H}_1$. If $\parallel P_1 f \parallel^2$ can be thought of as a roughness penalty, then the smoothing parameter λ controls the tradeoff between the goodness-of-fit and the roughness of the solution. If $\mathcal{H}$ is $W_2^m[0,1]$, $\mathcal{H}_0$ is the span of the polynomials of degree less than m, and $\parallel P_1 f \parallel^2 = \int (f^{(m)}(t))^2 dt$, we get that the minimizer f_λ of (1) is the celebrated polynomial spline of degree $2m - 1$.

Table 1: Terms in tensor product space $\mathcal{H}$.

$\mathcal{H}_{c,c}$	$\mathcal{H}_{c,\pi}$	$\mathcal{H}_{c,s}$
$\mathcal{H}_{\pi,c}$	$\mathcal{H}_{\pi,\pi}$	$\mathcal{H}_{\pi,s}$
$\mathcal{H}_{s,c}$	$\mathcal{H}_{s,\pi}$	$\mathcal{H}_{s,s}$

We next consider the tensor product of two reproducing kernel spaces (which may be different). Let each one have an orthogonal decomposition of the form $\mathcal{H} = \mathcal{H}_c \oplus \mathcal{H}_\pi \oplus \mathcal{H}_s$ where $\mathcal{H}_c$ is the one dimensional space of constant functions, $\mathcal{H}_\pi$ will be a finite dimensional subspace perpendicular to $\mathcal{H}_c$, (which will be spanned by polynomials in the examples condsidered in this paper), and $\mathcal{H}_s$ is everything else, chosen so that the squared norm on $\mathcal{H}_s$ represents "roughness". If we take the tensor product of these two spaces, and expand out all the terms, we get the 9 terms given in Table 1, where $\mathcal{H}_{\alpha,\beta} = \mathcal{H}_\alpha^{(1)} \otimes \mathcal{H}_\beta^{(2)}$, $\alpha, \beta = c, \pi, s$. Here the superscripts (1) and (2) refer to the two spaces we started with. The four terms in the upper left corner of Table 1 are finite dimensional, and are generally not penalized in applications. Any element in the resulting $\mathcal{H}$ has a unique decomposition of the form $f(\boldsymbol{t}) = \mu + f_1(t^{(1)}) + f_2(t^{(2)}) + f_{12}(t^{(1)}, t^{(2)})$, where $f_1 \in \mathcal{H}_{\pi,c} \oplus \mathcal{H}_{s,c}$, $f_2 \in \mathcal{H}_{c,\pi} \oplus \mathcal{H}_{c,s}$ and $f_{12} \in \mathcal{H}_{\pi,\pi} \oplus \mathcal{H}_{\pi,s} \oplus \mathcal{H}_{s,\pi} \oplus \mathcal{H}_{s,s}$. Of course this decomposition will depend on the norms in the original subspaces. We can take the tensor product of d of these subspaces, and we get a 3^d table analogous to 1, and a unique decomposition of the form

$$f(\boldsymbol{t}) = \mu + \sum_{\gamma} f_\gamma(t^{(\gamma)}) + \sum_{\gamma_1 < \gamma_2} f_{\gamma_1,\gamma_2}(t^{(\gamma_1)}, t^{(\gamma_2)}) + \cdots + f_{1,\cdots,d}(t^{(1)}, \cdots, t^{(d)}). \tag{2}$$

The f_γ's are the *main effects*, the f_{γ_1,γ_2}'s are the *two-factor interactions*, and so on. If this model resulted from a d tensor sum of spaces $W_2^{m(\gamma)}$, where the $m(\gamma)$'s may be different, then each function has a marginal square integrable $m(\gamma)$th derivative. There is a growing literature on various forms of these additive and interaction spline models. See [2,3,4,12,13,15,16,17,18,19,20,31]. Generally only the main effects and lower order interactions will be included in the model.

There are three main parts to the methodology, which we will describe in turn. First, we choose norms on the original $\mathcal{H}_0$ spaces for which we can construct reasonable and convenient reproducing kernels. Second, we lump all of the subspaces that are candidates for inclusion into one of two grand spaces which play the roles of $\mathcal{H}_0$ and $\mathcal{H}_1$ in (1). Given the reproducing kernels for all of the component subspaces, we construct the reproducing kernel for the grand space $\mathcal{H}_1$ as sums and products of the original reproducing kernels. Multiple smoothing parameters can be incorporated into this representation by rescaling norms of the various projections. Third, this allows us to use what might be considered as a canonical form for this optimization problem, and to use some publicly available software[11] to solve it. In Section 2 we will review these results for the so-called *additive and interaction* splines based on $\otimes^d W_2^{m(\gamma)}[0,1]$ This will set the stage for the generalization which is the subject of this report.

In dealing with geographical variables, like latitude and longitude, or, when studying, e. g. properties of the ocean, latitude, longitude and depth, or, even latitude, longitude, depth and time, it may be more appropriate to partition the d variables that one is dealing with into homogenous groups, so that, within a group of variables the penalty functional is invariant under within-group variable rotations. For concreteness, suppose that $\boldsymbol{t} = (\boldsymbol{t}^{(1)}, \boldsymbol{t}^{(2)})$, where $\boldsymbol{t}^{(\cdot)}$ has $d(\cdot)$ components.. We will consider taking the tensor product of two Hilbert spaces, the first consisting of functions of $\boldsymbol{t}^{(1)}$, and the second consisting of functions of $\boldsymbol{t}^{(2)}$. Momentarily letting $\boldsymbol{t}^{(1)} = t_1, ..., t_d$, say, we will take $\| P_1 f \|^2$ in the space of functions of $\boldsymbol{t}^{(1)}$ as J_m^d, where $m = m(1), d = d(1)$ and

$$J_m^d(f) = \sum_{\alpha_1+\cdots+\alpha_d=m} \frac{m!}{\alpha_1!\cdots\alpha_d!} \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\frac{\partial^m f}{\partial t_1^{\alpha_1}\cdots\partial t_d^{\alpha_d}}\right)^2 dt_1\cdots dt_d. \tag{3}$$

and similarly for $\boldsymbol{t}^{(2)}$. J_m^d is the penalty functional for the thin plate splines. A technical requirement is $2m > d$ for each (m, d) pair. If we allow $\boldsymbol{t} = \boldsymbol{t}^{(1)}, \cdots \boldsymbol{t}^{(\Gamma)}$, where each $\boldsymbol{t}^{(\gamma)}$ now has $d(\gamma)$ components, we will then have a decomposition into main effects, two factor interactions, just as in (2), except that each $t^{(\gamma)}$ is replaced by $\boldsymbol{t}^{(\gamma)}$, where $\boldsymbol{t}^{(\gamma)}$ has $d(\gamma)$ components, and there are all together $\sum_\gamma d(\gamma)$ variables. In Sections 3 and 4 we will describe how this can be done. In Section 5 we describe an example. In Sections 6 and 7 we discuss some approaches for using the data to decide which subspaces should be included, and make some remarks concerning methods for making practical accuracy estimates of the solution.

2 MULTIPLE SMOOTHING PARAMETERS, SPLINES BASED ON W_2^m

Let $\mathcal{H}$ be a reproducing kernel space with a decomposition of the form

$$\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$$

where $\mathcal{H}_0$ is span $\{\phi_1, \ldots, \phi_M\}$ and $\mathcal{H}_1$ is the direct sum of p orthogonal subspaces $\mathcal{H}^1, \ldots, \mathcal{H}^p$,

$$\mathcal{H}_1 = \sum_{\beta=1}^{p} \oplus \mathcal{H}^\beta. \tag{4}$$

We wish to find $f \in \mathcal{H}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{t}(i))^2 + \lambda \sum_{\beta=1}^{p} \theta_\beta^{-1} \| P^\beta f \|^2 \tag{5}$$

where P^β is the orthogonal projection in $\mathcal{H}$ onto $\mathcal{H}^\beta$ and $\theta_\beta > 0$. The following result is well known, see [32]. Suppose that the $n \times M$ matrix T with $i\nu$th entry $\phi_\nu(\boldsymbol{t}(i))$ is of rank M, and suppose that the reproducing kernel for $\mathcal{H}^\beta$ is $R^\beta(\cdot,\cdot)$. Then the minimizer $f_{\lambda,\theta}$ of (5) is given by

$$f_{\lambda,\theta}(\cdot) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(\cdot) + \sum_{i=1}^{n} c_i \sum_{\beta=1}^{p} \theta_\beta R^\beta(\cdot, \boldsymbol{t}(i)). \tag{6}$$

where the coefficient vectors $c = (c_1, ..., c_n)$ and $d = (d_1, ..., d_M)$ satisfy

$$(\Sigma + n\lambda I)c + Td = y \tag{7}$$

$$T'c = 0. \tag{8}$$

Here

$$\Sigma = \theta_1 \Sigma_1 + \theta_2 \Sigma_2 + \ldots + \theta_p \Sigma_p, \tag{9}$$

where the ijth entry of Σ_β is $R^\beta(\boldsymbol{t}(i), \boldsymbol{t}(j))$

The smoothing parameters $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_p)$, where $\lambda_\beta = \lambda\theta_\beta^{-1}, \beta = 1, ..., p$, can be chosen by by the method of generalized cross validation (GCV)[6]. This entails finding $\boldsymbol{\lambda}$ to minimize $V(\boldsymbol{\lambda})$ given by

$$V(\boldsymbol{\lambda}) = \frac{\frac{1}{n}\|(I - A(\boldsymbol{\lambda}))y\|^2}{[\frac{1}{n}tr(I - A(\boldsymbol{\lambda}))]^2}. \tag{10}$$

where $A(\boldsymbol{\lambda})$ is the influence matrix, which satisfies

$$\begin{pmatrix} f_{\boldsymbol{\lambda}}(\boldsymbol{t}(1)) \\ \vdots \\ f_{\boldsymbol{\lambda}}(\boldsymbol{t}(n)) \end{pmatrix} = A(\boldsymbol{\lambda})y. \tag{11}$$

It is known [32] that V can be rewritten as

$$V(\lambda, \theta_1, ..., \theta_p) = \frac{z'(\tilde{\Sigma} + n\lambda I)^{-2} z}{(tr(\tilde{\Sigma} + n\lambda I)^{-1})^2} \tag{12}$$

where $z = F'y, \tilde{\Sigma} = F'\Sigma F$ and F is any $n \times (n-M)$ matrix satisfying $F'T = 0_{n-M \times M}$. F can be obtained from the QR decomposition. The general purpose code RKPACK[11], which is available through netlib@research.att.com, will obtain c, d, and the GCV estimate of $\boldsymbol{\lambda}$, given the input data, a basis for $\mathcal{H}_0$, and the relevant reproducing kernels. It is not trivial to carry out the minimization. RKPACK uses the truncated Householder transformation with fixed θ's to optimize with respect to λ as described in [15] alternating with a Newton iteration in the logs of the θ's. See [18] for details. Girard [10] has provided a Monte Carlo method for evaluating trA.

For $f \in W_2^m[0,1]$, let

$$M_\nu f = \int_0^1 f^{(\nu)}(x)dx, \ \nu = 0, 1, \ldots, m-1, \tag{13}$$

and let $k_\nu(t) = B_\nu(t)/\nu!$, where B_ν is the ν-th Bernoulli polynomial. We have $M_\nu B_\mu = 1, \nu = \mu$, and 0 otherwise. Letting

$$\parallel f \parallel_{W_m}^2 = \sum_{\nu=0}^{m-1} (M_\nu f)^2 + \int_0^1 (f^{(m)}(u))^2 du, \tag{14}$$

and letting $\mathcal{H}_\pi = span\{k_\nu\}$ we have that the reproducing kernels for $\mathcal{H}_\pi$ and $\mathcal{H}_s$ are, respectively

$$R_\pi(t,t') = \sum_{i=1}^{m-1} k_\nu(t) k_\nu(t') \tag{15}$$

and

$$R_s(t,t') = k_m(t)k_m(t') + (-1)^{m-1} k_{2m}([t-t']) \tag{16}$$

where $[\tau]$ is the fractional part of τ. To fit an additive and interaction spline model in d variables, one decides which subspaces (from the 3^d table) one wishes to include. The reproducing kernel for each $\mathcal{H}^\beta$ is obtained as the relevant sums and products of R_π and R_s of the constituent spaces.

An example of the successful recovery of a four dimensional additive function, observed over an irregular set of points on $[0,1]^4$ with noise, is shown in [18]. In that example, a four dimensional additive model was selected for fitting. We will discuss later the problem of deciding which subspaces to include, when apriori information is not necessarily available.

3 THIN PLATE SPLINES

A few references to thin plate splines are Duchon[8], Meinguet[24], Wahba and Wendelberger[33], Utreras[28] and references cited there, Thomas-Agnan[27], and numerous interesting applications by Hutchinson and collaborators to climatological data, see, for example Hutchinson, Kalma, and Johnson[21]. A one-dimensional special case of the trick we use to define reproducing kernels below goes back at least to deBoor and Lynch[7].

The only thing needed to extend the preceeding results to include the thin plate splines is to determine suitable reproducing kernels for $\mathcal{H}_c, \mathcal{H}_\pi$, and $\mathcal{H}_s$. where $\mathcal{H}_s$ is a suitable space of functions for which $J_m^d < \infty$. Discussion of these spaces may be found in the works of Duchon and Meinguet. For our purposes we need only to produce a reproducing kernel, which will then determine the space uniquely. The following appears in [16,17].

Let $\phi_0, ..., \phi_{M-1}$ span the $\binom{m+d-1}{d}$ polynomials of total degree less than m in R^d, that is the null space of J_m^d. Given m and d, a set of $N+1$ points $\boldsymbol{w}_0, ..., \boldsymbol{w}_N$ in R^d is said to be unisolvent, if least squares regression on $\phi_0, \cdots, \phi_{M-1}$ given data at these points is unique. Let $\{\boldsymbol{w}_k\}_{k=1}^N$ be unisolvent set of N points in R^d. Define $(f,g)_N = (1/N)\sum_{k=1}^N f(\boldsymbol{w}_k)g(\boldsymbol{w}_k)$, then $(f,g)_N$ can be used as a the inner product over $\mathcal{H}_0 = span\{\phi_\nu\}$. Let $(f,g)_N =< f,g >_0$ We will assume that the $\phi_\nu' s$ are an orthonormal basis for $\mathcal{H}_0$ with this inner product, satisfying $\phi_0 = 1$, and $< \phi_\nu, \phi_\mu >_0 = 1$ or 0 according as $\mu = \nu$ or not. (Given any spanning set, the

ϕ_ν's can be found using the QR decomposition.) Then it is easy to see that the reproducing kernel for $\mathcal{H}_c$ is 1, and the reproducing kernel for $\mathcal{H}_\pi$ is $R_\pi(\boldsymbol{t},\boldsymbol{s}) = \sum_{\nu=1}^{M-1}\phi_\nu(\boldsymbol{t})\phi_\nu(\boldsymbol{s})$

A set of (unisolvent) points $\boldsymbol{w}_0,\cdots,\boldsymbol{w}_N$ and associated weights $h_0,\cdots,h_N$ is called a generalized divided difference if

$$\sum_{k=0}^{N} h_k\phi_\nu(\boldsymbol{w}_k) = 0, \quad \nu = 0,1,\cdots,M.$$

F is said to be m-conditionally positive definite if $\sum_{j,k} h_j h_k F(\|\boldsymbol{w}_j - \boldsymbol{w}_k\|) \geq 0$ whenever $\{h_k; \boldsymbol{w}_k\}$ is a generalized divided difference with respect to the polynomials of total degree less than m, on R^d, see Micchelli[25] for examples. Let $E_m^d(\cdot)$, the functions used to construct thin plate splines, be defined by

$$E_m^d(\cdot) = \begin{cases} C_m\{(\cdot)^{2m-d}\log(\cdot)\}, & d \text{ even,} \\ & C_m = (-1)^{d/2+m+1} \,/\, \{2^{2m-1}\pi^{d/2}(m-1)!(m-d/2)!\} \\ C_m\{(\cdot)^{2m-d}\}, & d \text{ odd,} \\ & C_m = \Gamma(d/2-m) \,/\, \{2^{2m}\pi^{d/2}(m-1)!\} \end{cases}$$

It is well known that the E_m^d are m-conditionally positive definite[23].

Let the bilinear form $< f,g >_*$ be defined, wherever it exists, by

$$< f,g >_* = \sum_{\alpha_1+\cdots+\alpha_d=m} \frac{m!}{\alpha_1!\cdots\alpha_d!} \int\cdots\int \frac{\partial^m f}{\partial t_1^{\alpha_1}\cdots\partial t_d^{\alpha_d}} \frac{\partial^m g}{\partial t_1^{\alpha_1}\cdots\partial t_d^{\alpha_d}} dt_1\cdots dt_d.$$

and let $E(\boldsymbol{t},\boldsymbol{s}) = E_m^d(\|\boldsymbol{t}-\boldsymbol{s}\|)$, $E_{\boldsymbol{t}}(\cdot) = E(\boldsymbol{t},\cdot)$. It is known (see, e. g. [24]), that, if $\{h_k, \boldsymbol{w}_k\}$ is a generalized divided difference, that

$$< \sum_{j=0}^{N} h_j E_{\boldsymbol{w}_j}(\cdot), \sum_{k=0}^{N} h_k E_{\boldsymbol{w}_k}(\cdot) >_* = \sum_j\sum_k h_j h_k E(\boldsymbol{w}_j,\boldsymbol{w}_k).$$

For any f continuous, we can define P_0 by

$$(P_0 f)(\cdot) = \sum_{\nu=0}^{M-1} (\phi_\nu, f)_N \phi_\nu(\cdot)$$

Let $P_{0(\boldsymbol{t})}$ be P_0 applied to what follows considered as a function of $\boldsymbol{t}$. We take $R_{\boldsymbol{s}}$ as

$$R_{\boldsymbol{s}}(\boldsymbol{t},\boldsymbol{s}) = (I - P_{0(\boldsymbol{t})})(I - P_{0(\boldsymbol{s})})E(\boldsymbol{t},\boldsymbol{s}).$$

Then $\mathcal{H}_{\boldsymbol{s}}$ will be a collection of functions for which $< f,f >_* = J_m^d(f) = \| f \|^2 < \infty$, and satisfying $P_0 f = 0$. To see this, let

$$h_k(\boldsymbol{t}) = \sum_{\nu=0}^{M-1} \phi_\nu(\boldsymbol{t})\phi_\nu(\boldsymbol{w}_k).$$

Then, it can be shown that $(1, -h_1(\boldsymbol{t}), \cdots, -h_N(\boldsymbol{t}); \boldsymbol{t}, \boldsymbol{w}_1, \cdots, \boldsymbol{w}_N)$ is a generalized divided difference for any $\boldsymbol{t}$. It is then straightforward to compute that $R_s(\boldsymbol{t}, \boldsymbol{s})$ is (unconditionally) positive (that is, nonnegative) definite, and furthermore, R_s has the reproducing property

$$< R_s(\boldsymbol{s}, \cdot), R_s(\boldsymbol{t}, \cdot) >_* = R_s(\boldsymbol{s}, \boldsymbol{t}).$$

Construction of a thin plate spline via Equation(6), with this reproducing kernel will give the usual thin plate spline. In the application discussed in [17] it is reasonable to let the $\boldsymbol{w}_k$ be the data points. One is then decomposing the data vector in a natural way into its components along and perpendicular to $\mathcal{H}_0$.

In what follows we let $\mathcal{H}^{m,d}$ be the space with reproducing kernels R_c, R_π, and R_s of this section.

4 GRAND MODELS

We note for completeness, that the $d = 1$ case for the thin plate splines will reduce to ordinary polynimial splines of degree $2m - 1$. The norm on the null space of $\int f^{(m)}$ is different from that of Section 2, however, which will make a difference when taking tensor products. In the discussion below it is being assumed that the thin plate version is taken whenever $d = 1$. It is now immediate how one obtains a decomposition of the form (2), with the $t^{(\gamma)}$ replaced by $\boldsymbol{t}^{(\gamma)}$. Given d and m, let

$$\mathcal{H}^{m,d} = \mathcal{H}_c^{m,d} \oplus \mathcal{H}_\pi^{m,d} \oplus \mathcal{H}_s^{m,d}. \tag{17}$$

For Γ given pairs (d, m) let $\mathcal{H}$ be the big space, on $E^{d_1+d_2+\ldots+d_\Gamma}$ given by

$$\otimes_\gamma \mathcal{H}^{m(\gamma),d(\gamma)} = \otimes_\gamma \left[\mathcal{H}_c^{m(\gamma),d(\gamma)} \oplus \mathcal{H}_\pi^{m(\gamma),d(\gamma)} \oplus \mathcal{H}_s^{m(\gamma),d(\gamma)}\right] \tag{18}$$

Here

$$\boldsymbol{t} = (t_{11}, \ldots, t_{1,d_1}; t_{2,1}, \ldots, t_{2,d_2};, \ldots, ; t_{\Gamma,1}, \ldots, t_{\Gamma,d_\Gamma}) = (\boldsymbol{t}^{(1)}, \ldots, \boldsymbol{t}^{(\Gamma)}), \tag{19}$$

and we can consider main effects, two factor interactions, etc, in the variables $\boldsymbol{t}^{(\gamma)}$.

5 AN EXAMPLE

As an example of the use of these methods, Gu and Wahba[17] looked at pH measurements (y) from two groups of lakes, in the Blue Ridge Mountains $(n = 159)$, and in Florida $(n = 112)$. The data set contained geographical position (latitude, longitude), which was set equal to $(\boldsymbol{t}^{(1)})$, log calcium content $(t^{(2)})$, and other information for each lake. It was desired to make a descriptive model of the relation of pH to geographical position and calcium content for each group of lakes. A model of the form

$$y_i = \mu + f_1(\boldsymbol{t}^{(1)}(i)) + f_2(t^{(2)}(i)) + f_{12}(\boldsymbol{t}^{(1)}(i), t^{(2)}(i)) + \epsilon(i) \tag{20}$$

was considered. The model was fitted using RKPACK with GCV to estimate the multiple smoothing parameters. The regression diagnostics proposed in Gu[13], which we will describe in a moment, were used to decide which subspaces to retain. For the Florida data, these diagnostics suggested that there was no observable geographic effect and so f_1 and f_{12} were dropped from the model, and the model was refitted using only f_2. In this example all m's were set to 2. The diagnostics for the Blue Ridge data suggested retaining all terms. In fact the Florida area containing the lakes was flat, while the Blue Ridge lakes ran along a mountain range, and so this result was not surprising since pH is known to depend on elevation. A plot of the main effects for geography (f_1) had a rough visual correlation with the location of the mountain range. A simple model replacing $\boldsymbol{t}^{(1)}$ = (latitude, longitude) with $t^{(1)}$ = elevation was then fitted. The diagnostics gave some suggestion that the latitude-longitude model was slightly better, thus suggesting that there might be some geographic factor other than altitude involved.

In the next section we describe several diagnostic procedures for model building of this sort. At the present time, these diagnostic procedures do not have the mathematical rigor to be found in the hypothesis testing and model-building literature for parametric models. Research is proceeding that will hopefully provide some rigor. However, exact probabilistic statements like those obtainable for parametric models will probably not be available. However, we believe that these models have the potential for building descriptive predictive models for complicated data sets from various medical, economic, demographic and other sources, and that some of the procedures to be described warrant further studies as to their properties. Furthermore, we have some reason to believe, that at least with large data sets, cross validation and similar methods can be used to draw plausible, if not rigorous inferences from the models.

6 DIAGNOSTICS, MODEL SELECTION

Diagnostics naturally into three classes, which may be called preanalysis, analysis and post-analysis. These ideas have been clarified by C. Gu (personal communication). We discuss these three in turn.

Preanalysis diagnostics examine the so-called design points $\boldsymbol{t}(i), i = 1, ..., n$, and the models being entertained, that is the choices of $\mathcal{H}_0$ and the $\mathcal{H}^{\beta}$'s to see if the design points have the potential for "carrying" the model. To do this one should examine the matrix T and the matrices $\tilde{\Sigma}_{\beta} = F'\Sigma_{\beta}F$ of Section 2. It is necessary that T be of full column rank. If T is poorly conditioned, then an immediate rethinking of the model is necessary - in that case even least squares regression on $\mathcal{H}_0$ is inadvisable. If $\tilde{\Sigma}_{\beta_1}$ and $\tilde{\Sigma}_{\beta_2}$ are "close", it will be difficult to identify θ_{β_1} and θ_{β_2}, and the components $P^{\beta_1}f$ and $P^{\beta_2}f$ will probably not be independently meaningful. It would be better to combine the two subspaces, or delete one of them. One natural measure of the "closeness" of two matrices with the same dimensions is $cos(A, B) = trAB'/(trAA')^{1/2}(trBB')^{1/2}$. Ideally, the $p \times p$ (correlation) matrix of $cos(\tilde{\Sigma}_{\beta_1}, \tilde{\Sigma}_{\beta_2})$ should be well conditioned. Purely objective criteria for well-conditionedness in this context remain to be identified.

Analysis diagnostics are those used to delete components from a tentative model, and, possibly, compare two different models. At the outset, in some applications one might be interested in seeing whether ordinary least squares regression onto $\mathcal{H}_0$ "explains" the data. Tests for doing this are described in Cox et al [5] and Wahba [32]. Assuming that a multivariate spline model will be fitted, we note that there is more than one way to group components for deciding whether to include them in the model or not. One could group by the terms in the general form of the decomposition of Equation(2) or, by the components in each $\mathcal{H}^\beta$. We will do the former, and use the model of Section 5 as an example.

After fitting the model of Section 5, one has a decomposition of the data vector as $\boldsymbol{y} = \boldsymbol{\mu} + \tilde{\boldsymbol{f}}_1 + \tilde{\boldsymbol{f}}_2 + \tilde{\boldsymbol{f}}_{12} + \tilde{\boldsymbol{\epsilon}}$ where $\tilde{\boldsymbol{f}}_{(\cdot)}$ is the n dimensional vector of the fitted $f_{(\cdot)}(\boldsymbol{t}(i))$. After projecting each component onto $\{\mathbf{1}\}^\perp$ we get $\boldsymbol{z} = \boldsymbol{f}_1 + \boldsymbol{f}_2 + \boldsymbol{f}_{12} + \boldsymbol{\epsilon}$. Since the $\boldsymbol{f}_{(\cdot)}$ are supposed to predict $\boldsymbol{z}$, $cos(\boldsymbol{z}, \boldsymbol{f}_{(\cdot)})$ near 0 makes $\boldsymbol{f}_{(\cdot)}$ suspect. $R^2 = \parallel \boldsymbol{z} - \boldsymbol{\epsilon} \parallel_n^2 / \parallel \boldsymbol{z} \parallel_n^2$ is a measure of the overall goodness of the model. Since R^2 depends on the smoothing parameters, for R^2 to be meaningful, the smoothing parameters must have been chosen well. Again, further research is required to develop objective criteria for what constitutes large and small in this fairly non-standard context.

In the estimation of the $\lambda_1, ..., \lambda_p$, a λ_β of ∞ indicates that $\mathcal{H}^\beta$ is being removed. However, a very large but finite λ_β could be generated by noise. The statistic

$$v = inf_{\boldsymbol{\lambda}} V(\boldsymbol{\lambda}) / inf_{\lambda_\beta = \infty} V(\boldsymbol{\lambda})$$

is a measure of the benefit of including $\mathcal{H}^\beta$. A v near 1 suggests that $\mathcal{H}^\beta$ should be removed. Again, further work is needed to develop objective criteria for "near 1"

In [26,30] Bayesian "confidence intervals" were established, based on the stochastic model behind splines. The intervals have reasonable properties for $f \in \mathcal{H}$ even though they were based on an assumption that f is a sample function from a stochastic process (and hence not in $\mathcal{H}$ with probability 1). They have to be interpreted differently than the usual confidence intervals, however. Roughly the result (for one smoothing parameter) goes as follows: Let I_i be the interval

$$I_i = f_{\hat{\lambda}}(\boldsymbol{t}(i)) \pm 1.96\sigma\sqrt{a_{ii}(\hat{\lambda})} \qquad (21)$$

where $a_{ii}(\hat{\lambda})$ is the iith entry of $A(\hat{\lambda})$ and $\hat{\lambda}$ is the GCV estimate of λ. Let $I_i(x)$ be 1 or zero according as the interval I_i contains x or not. Then, roughly, and under some assumptions

$$E\frac{1}{n}\sum_{i=1}^{n} I_i(f(\boldsymbol{t}(i)) \approx .95 \qquad (22)$$

The noise variance σ^2 can be estimated from the data as

$$\hat{\sigma^2} = \|(I - A(\hat{\lambda}))\boldsymbol{y}\|^2 / [tr(I - A(\hat{\lambda}))].$$

Chong Gu and I are exmining the possiblity of using a component wise version of Equation(22) to decide if a component $f_{(\cdot)}$ should remain in the model. One would then look at $\frac{1}{n}\sum I_i^{(\cdot)}(0)$ and if this is large, then the $(\cdot)$ component would be deleted.

7 POST ANALYSIS DIAGNOSTICS, or, ACCURACY STATEMENTS FOR SMOOTHING SPLINES

We emphasize that research in this area is in its infancy. Quantitative empirically determined statements concerning possible accuracy of nonparametric models of varous types constitute an area of active research in modern statistics. Bootstrap and cross validation methods appear to be popular. Further references can be found in [9,16,17]. We will only discuss some possibilities based on extensions of the results here and in [14,26,30].

The Bayesian model[29] behind smoothing splines goes as follows:

$$y_i = f(\boldsymbol{t}(i)) + \epsilon(i), \qquad i = 1, \cdots, n$$

where the $\epsilon(i)$'s are as before, but now f is a zero mean Gaussian stochastic process, with a representation

$$f((\boldsymbol{t}) = \sum_{i=1}^{M} \tau_\nu \phi_\nu(\boldsymbol{t}) + bZ(\boldsymbol{t}), \tag{23}$$

here, the τ_ν are independent Gaussian random variables with variance w, and $Z(\boldsymbol{t})$ is a zero mean Gaussian stochastic process with covariance

$$EZ(\boldsymbol{s})Z(\boldsymbol{t}) = R(\boldsymbol{s}, \boldsymbol{t}) = \sum_{\beta=1}^{p} \theta_\beta R^\beta(\boldsymbol{s}, \boldsymbol{t}).$$

The τ_ν's and $Z(\boldsymbol{t})$ are independent and $w \to \infty$. Here as before the ϕ_ν span $\mathcal{H}_0$ and $R(\cdot, \cdot)$ is the reproducing kernel for $\mathcal{H}_1$. With this model, $f_\lambda(\boldsymbol{t})$ is the posterior mean of $f(\boldsymbol{t})$ given the data $\boldsymbol{y}$, for the particular choice of $\lambda = \sigma^2/nb$. We remark that it is known that f of the form (23) is, with probability 1 not in $\mathcal{H}$. If $f \in \mathcal{H}$, then, although the meaning of σ^2 is the same as for f as in (23), the meaning of λ is as a variable smoothing parameter, and it is not necessarily clear how to interpret b. σ^2 and λ are estimated from the data.

The posterior covariance $C(\boldsymbol{s}, \boldsymbol{t})$ of f_λ, given the data $\boldsymbol{y}$

$$C(\boldsymbol{s}, \boldsymbol{t}) = E(f(\boldsymbol{s}) - f_\lambda(\boldsymbol{s}), f(\boldsymbol{t}) - f_\lambda(\boldsymbol{t}))|\boldsymbol{y})$$

is given in [30], Theorem 2. We will not reproduce that formula here. However, the formula there can be rearranged several ways, so that its terms have some meaning for $f \in \mathcal{H}$. (More details will appear in Gu and Wahba, in preparation). Let $Q_{\boldsymbol{t}}(\cdot)$ be the representer of evaluation in $\mathcal{H}$, that is,

$$Q_{\boldsymbol{t}}(\cdot) = \sum_{i=1}^{M} \phi_\nu(\boldsymbol{t})\phi_\nu(\cdot) + R(\boldsymbol{t}, \cdot).$$

Let B_n be the map from $\mathcal{H}$ to $\mathcal{H}$ defined by $B_n f \to g$ where g is the solution to $min \parallel P_1 g \parallel$ subject to $g(\boldsymbol{t}(i)) = f(\boldsymbol{t}(i)), i = 1, ..., n$, and let B_n^* be its adjoint in $\mathcal{H}$, that is, $< B_n u, v > =$

$< u, B_n^* v >$, any u, v in $\mathcal{H}$. Let $a_{ij}(\lambda)$ be the ijth entry of $A(\lambda)$, and, finally, let δ_{in} be that element in $\mathcal{H}$ which minimizes $\| P_1 \delta \|$ subject to $\delta_{in}(\boldsymbol{t}(j)) = 1, i = j$ and 0 otherwise. Then, it can be shown, that

$$C(\boldsymbol{s},\boldsymbol{t}) = bC_1(\boldsymbol{s},\boldsymbol{t}) + \sigma^2 C_2(\boldsymbol{s},\boldsymbol{t}) \tag{24}$$

where

$$\begin{aligned} C_1(\boldsymbol{s},\boldsymbol{t}) &= < Q_{\boldsymbol{s}} - B_n^* Q_{\boldsymbol{s}}, Q_{\boldsymbol{t}} - B_n^* Q_{\boldsymbol{t}} > \\ C_2(\boldsymbol{s},\boldsymbol{t}) &= \sum_{i,j} \delta_{in}(\boldsymbol{s}) \delta_{jn}(\boldsymbol{t}) a_{ij}(\lambda). \end{aligned}$$

Suppose we are interested in a particular component, say $P^\gamma f$, where $\mathcal{H}^\gamma$ is a subspace of $\mathcal{H}$. Then it can be shown that the posterior covariance of $P^\gamma f_\lambda(\boldsymbol{s})$ and $P^\gamma f_\lambda(\boldsymbol{t})$ given $\boldsymbol{y}$ is given by

$$E(P^\gamma f(\boldsymbol{s}) - P^\gamma f_\lambda(\boldsymbol{s}))(P^\gamma f(\boldsymbol{t}) - P^\gamma f_\lambda(\boldsymbol{t}))|\boldsymbol{y} = P^\gamma_{(\boldsymbol{s})} P^\gamma_{(\boldsymbol{t})} C(\boldsymbol{s},\boldsymbol{t}) \tag{25}$$

where $P^\gamma_{(\boldsymbol{s})}$ means P^γ applied to what follows, considered as a function of $\boldsymbol{s}$. Note that C_1 is 0 for $\boldsymbol{s}$ or $\boldsymbol{t} = \boldsymbol{t}(i)$, any i, and that $\{C_2(\boldsymbol{t}(i), \boldsymbol{t}(j))\} = A(\lambda)$. We remark that this decomposition of C into C_1 and C_2 represents a decomposition of the uncertainty due to the unobservable part of $f, (C_1)$ plus the uncertainty due to the propagation of the observational errors given contaminated values of $f(\boldsymbol{t}(i)), (C_2)$. Note that since $P_0(I - B_n) = 0$, we can obtain, for any $f \in \mathcal{H}$, the following *hypercircle inequality*:

$$(f(\boldsymbol{t}) - B_n f(\boldsymbol{t}))^2 \leq \| P_1(I - B_n) f \|^2 C_1(\boldsymbol{t},\boldsymbol{t}).$$

In practice however, $\| P_1(I - B_n) f \|$ or a reasonable bound on it may not be known. However, it can be expected to be small as the data points become dense.

Unfortunately the computation of C_1 and C_2 is apparently unstable, since the matrix $\tilde{\Sigma}^{-1}$, which may be ill conditioned, enters in. Set $b = \frac{\sigma^2}{n\lambda} + b_0$ in (24), then (24) can be rewritten as

$$C(\boldsymbol{s},\boldsymbol{t}) = \frac{\sigma^2}{n\lambda}\{C_1(\boldsymbol{s},\boldsymbol{t}) + n\lambda C_2(\boldsymbol{s},\boldsymbol{t})\} + b_0 C_1(\boldsymbol{s},\boldsymbol{t}). \tag{26}$$

Hopefully, $b_0 C_1$ is ignorable, since we have estimates of σ^2 and λ from the data, but no obvious way to interpret b rigorously in the $f \in \mathcal{H}$ case. There is a computationally stable version of the term in brackets, which goes as follows: Let $\tilde{C} = C_1 + n\lambda C_2$, let $B_n(\lambda)$ be the map from $\mathcal{H}$ to $\mathcal{H}$ defined by $B_n(\lambda) f \rightarrow g$ where g is the solution to *min*

$$\frac{1}{n}\sum_{i=1}^n (f(\boldsymbol{t}(i) - g(\boldsymbol{t}(i))^2 + \lambda \| P_1 g \|^2$$

and let $B_n^*(\lambda)$ be its adjoint. Let ϵ_λ be the minimizer in $\mathcal{H}$ of

$$\frac{1}{n}\sum_{i=1}^n (\epsilon(i) - \epsilon(\boldsymbol{t}(i))^2 + \lambda \| P_1 \epsilon \|^2$$

where the $\epsilon(i)$ are as before. Upon rearranging terms, it can also be shown that $\tilde{C}(\boldsymbol{s},\boldsymbol{t}) = C_3(\boldsymbol{s},\boldsymbol{t}) + n\lambda C_4(\boldsymbol{s},\boldsymbol{t})$ where

$$\begin{aligned} C_3(\boldsymbol{s},\boldsymbol{t}) &= < Q_{\boldsymbol{s}} - B_n^*(\lambda)Q_{\boldsymbol{s}}, Q_{\boldsymbol{t}} - B_n^*(\lambda)Q_{\boldsymbol{t}} > \\ C_4(\boldsymbol{s},\boldsymbol{t}) &= \frac{1}{\sigma^2}E\epsilon_\lambda(\boldsymbol{s})\epsilon_\lambda(\boldsymbol{t}). \end{aligned}$$

where E is expected value. The calculation of $\tilde{C}(\boldsymbol{s},\boldsymbol{t})$ is stable, since it involves $(\tilde{\Sigma} + n\lambda I)^{-1}$ rather than $\tilde{\Sigma}^{-1}$. Gu[14] has given a formula which can be shown to be equivalent to $\tilde{C}$ and a stable computational algorithm for it. See [14] for computational details.

Now for $f \in \mathcal{H}$ we have $Ef_\lambda(\boldsymbol{t}) = B_n(\lambda)f(\boldsymbol{t})$ so that

$$f(\boldsymbol{t}) - f_\lambda(\boldsymbol{t}) = (I - B_n(\lambda)f(\boldsymbol{t}) + \epsilon(\boldsymbol{t}).$$

It can be shown that $P_0(I - B_n^*(\lambda)) = 0$ so that we can write

$$\begin{aligned} E(f(\boldsymbol{t}) - f_\lambda(\boldsymbol{t}))^2 &= < (I - B_n(\lambda))f, Q_{\boldsymbol{t}} >^2 + E\epsilon(\boldsymbol{t})^2 \\ &\leq \parallel P_1 f \parallel^2 C_3(\boldsymbol{t},\boldsymbol{t}) + \sigma^2 C_4(\boldsymbol{t},\boldsymbol{t}). \end{aligned}$$

We remark that, when dealing with experimental data in more than 1 dimension, the data may be quite irregular, and extrapolation of f_λ far from the data is generally not defensible. One could use the size of $\tilde{C}(\boldsymbol{t},\boldsymbol{t})$ as a criteria - one only trusts f_λ for values of $\boldsymbol{t}$ for which it is not too large.

References

[1] N. Aronszajn. Theory of reproducing kernels. *Trans. Am. Math. Soc.*, 68:337–404, 1950.

[2] D. Barry. Nonparametric Bayesian regression. *Ann. Statist.*, 14:934–953, 1986.

[3] A. Buja, T. Hastie, and R. Tibshirani. Linear smoothers and additive models. *Ann. Statist.*, 17:453–555, 1989.

[4] Z. Chen, C. Gu, and G. Wahba. Comments to "Linear Smoothers and Additive Models, by Buja, Hastie and Tibshirani. *Ann. Statist.*, 17:515–521, 1989.

[5] D. Cox, E. Koh, G. Wahba, and B. Yandell. Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. *Ann. Statist.*, 16:113–119, 1988.

[6] P. Craven and G. Wahba. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, 31:377–403, 1979.

[7] C. deBoor and R.E. Lynch. On splines and their minimum properties. *J. Math. Mech.*, 15:953–969, 1966.

[8] J. Duchon. Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In *Constructive Theory of Functions of Several Variables*, pages 85–100. Springer-Verlag, Berlin, 1977.

[9] J. Friedman and B. Silverman. Flexible parsimonious smoothing and additive modeling. *Technometrics*, 3:3–21, 1989.

[10] D. Girard. Asymptotic optimality of the fast randomized versions of GCV and C_L in ridge regression and regularization. Technical Report RR793 M, Institute IMAG, Grenoble, 1990.

[11] C. Gu. RKPACK and its applications: Fitting smoothing spline models. Technical Report 857, Dept. of Statistics, University of Wisconsin, Madison, WI, 1989. code available through netlib.

[12] C. Gu. Adaptive spline smoothing in non-Gaussian regression models. *J. Amer. Statist. Assoc.*, 85:801–807, 1990.

[13] C. Gu. Diagnostics for nonparametric additive models. Technical Report 92, University of British Columbia, Vancouver, Canada, 1990.

[14] C. Gu. Penalized likelihood regression: a Bayesian analysis. manuscript, Statistics Dept., Purdue University, 1990.

[15] C. Gu, D.M. Bates, Z. Chen, and G. Wahba. The computation of GCV functions through householder tridiagonalization with application to the fitting of interaction spline models. *SIAM J. Matrix Anal.*, 10:457–480, 1989.

[16] C. Gu and G. Wahba. Comments to "Multivariate Adaptive Regression Splines, by J. Friedman. *Ann. Statist.*, to appear, 1991.

[17] C. Gu and G. Wahba. Semiparametric ANOVA with tensor product thin plate splines. Technical Report 90-61, Dept. of Statistics, Purdue University, Lafayette, IN, 1990.

[18] C. Gu and G. Wahba. Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Statist. Comput.*, 12, 1991, to appear.

[19] T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1:297–318, 1986.

[20] T. Hastie and R. Tibshirani. *Generalized Additive Models.* Chapman and Hall, 1990.

[21] M. Hutchinson, J. Kalma, and M. Johnson. Monthly estimates of windspeed and wind run for Australia. *J. Climatology*, 4:311–324, 1984.

[22] G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic.*, 33:82–95, 1971.

[23] G. Matheron. The intrinsic random functions and their applications. *Adv. Appl. Prob.*, 5:439–468, 1973.

[24] J. Meinguet. An intrinsic approach to multivariate spline interpolation at arbitrary points. In B. N. Sahney, editor, *Polynomial and Spline Approximation*, pages 163–190. Reidel, 1979.

[25] C. Micchelli. Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2:11–22, 1986.

[26] D. Nychka. Confidence intervals for smoothing splines. *J. A. S. A.*, 83:1134–1143, 1988.

[27] C. Thomas-Agnan. Smoothing noisy data by two equivalent techniques. In E. Diday, editor, *Data Analysis, Learning Symbolic and Numeric Knowledge*, Nova Science Publishers, 247–256, 1989.

[28] F. Utreras. On generalized cross-validation for multivariate smoothing spline fucntions. *SIAM J. Sci. Stat. Comput.*, 8:630–643, 1988.

[29] G. Wahba. Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Stat. Soc. Ser. B*, 40:364–372, 1978.

[30] G. Wahba. Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Stat. Soc. Ser. B*, 45:133–150, 1983.

[31] G. Wahba. Partial and interaction splines for the semiparametric estimation of functions of several variables. In T. Boardman, editor, *Computer Science and Statistics: Proceedings of the 18th Symposium*, pages 75–80. American Statistical Association, Washington, DC, 1986.

[32] G. Wahba. *Spline Models for Observational Data.* SIAM, 1990. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59.

[33] G. Wahba and J. Wendelberger. Some new mathematical methods for variational objective analysis using splines and cross-validation. *Monthly Weather Review*, 108:1122–1145, 1980.

Supported by NSF Grant DMS-9002566, AFOSR Grant AFOSR-90-0103 and NASA Grant NAG5-316.