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

TECHNICAL REPORT NO. 1024

September 20, 2000

(Smoothing) Splines in Nonparametric Regression [1]

by

Grace Wahba
http://www.stat.wisc.edu/~wahba

[1] Prepared for the Encyclopedia of Environmetrics with the title 'Splines in Nonparametric Regression', within the space and style limitations of the Encyclopedia. Based partly on research supported by NSF Grant DMS9704758, NIH Grant RO1 EY09946, and NASA Grant NAG5-3769

(Smoothing) Splines in Nonparametric Regression [2]

Grace Wahba [3]
Department of Statistics, University of Wisconsin, Madison WI

September 20, 2000

1 What is a spline function?

The classical definition of a spline function on $[0, 1]$ is a function which, given n *knots* $0 \leq x_1 < x_2 < \cdots < x_n \leq 1$, is defined as a polynomial in each of the intervals $[0, x_1), \cdots (x_j, x_{j+1}), \cdots (x_n, 1]$. Somewhat more specifically, the pieces of the polynomial are frequently assumed to be joined in such a way that the function is continuous, possess some specified number of continuous derivatives, and, possibly, satisfies some boundary conditions. See deBoor (1978) for a detailed discussion of splines as piecewise polynomials. The name 'spline' was given to these functions functions by Iso Schoenberg, who observed that certain ones approximately reproduced curves that were drawn by shipbuilders using a tool called a spline, which consisted of weights connected to a flexible strip. A univariate spline is still thought of as a piecewise polynomial, but some functions which piecewise satisfy some differential equation, are also called spline functions. (Kimeldorf & Wahba (1971), Ramsay & Dalzell (1991)). It is a celebrated result of Schoenberg (1964a), Schoenberg (1964b) that a polynomial spline satisfying certain boundary and continuity properties is the solution to a variational problem which minimizes the sum of two terms, the first being the residual sum of squares and the second the square integral of the mth derivative. There are several generalizations of spline functions to higher dimensions, and to other domains, for example to the sphere. All of them called splines, but not all of them are represented as piecewise polynomials. They generalize the univariate polynomial spline function in various ways. We will briefly note some of the generalizations which are piecewise polynomials, but we will go into much greater detail for those generalizations which are obtained as solutions to variational problems. The various kinds of spline functions can be used to interpolate data, and to smooth data. Interpolating splines in two and three dimensions are very popular in computer-aided design, but in this article we will be interested in splines as tools for visualizing and analyzing noisy observational data, and so will restrict ourselves to smoothing splines and regression splines, which generally do not interpolate the data. We will first describe the univariate polynomial smoothing spline, which may be thought of as the granddaddy of spline functions used in data analysis. Then we describe cross validation and Generalized Cross Validation (GCV) for choosing the smoothing parameter. After briefly describing regression splines, we then describe a number of generalizations of the univariate smoothing spline to various domains, which are obtained via the solution of a variational problem. These include the thin plate spline, the histospline, splines on the sphere, vector splines on the sphere, hybrid splines, partial splines, and smoothing spline ANOVA models on complex domains. We end with some remarks on computing. Publicly available software is mentioned along the way.

[2] Author address: Prof. Grace Wahba, Department of Statistics, University of Wisconsin, 1210 W. Dayton St., Madison WI 53706.©Grace Wahba 2000. Figures ©SIAM, IMS, ASA, IFNA = Society for Industrial and Applied Math, Institute of Mathematical Statistics, American Statistical Association, Interface Foundation of North America, as noted at the figures, all rights reserved by the respective copyright holders.

[3] Research supported in part by NIH Grant EY09946, NSF Grant DMS9704758 and NASA Grant NAG5-3769.

2 The univariate polynomial smoothing spline

The univariate polynomial smoothing spline is the solution of the following variational problem: Given abscissae $x = (x_1, \cdots, x_n)$, which we will assume, without loss of generality, satisfy $0 < x_1 < x_2 < \cdots < x_n < 1$ and ordinates $y = (y_1, \cdots, y_n)$; find f in an appropriately defined collection of functions[4] to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(t_i))^2 + \lambda \int_0^1 (f^{(m)}(u))^2 du. \tag{1}$$

For any $\lambda > 0$ this problem will have a unique solution provided that the least squares regression of the data on the polynomials of degree $m-1$ is unique, and then the solution has the following properties

$$f(x) \in \begin{cases} \pi^{m-1} & \text{for } x \in [0, x_1] \\ \pi^{2m-1} & \text{for } x \in [x_j, x_{j+1}] \\ \pi^{m-1} & \text{for } x \in [x_n, 1] \\ C^{2m-2} & \text{for } x \in [0, 1], \end{cases} \tag{2}$$

here, π^k are polynomials of degree k, and C^k are functions with k continuous derivatives. Thus the solution of this variational problem is a piecewise polynomial with the pieces joined so that the resulting function has $2m-2$ continuous derivatives, and it can be shown that f satisfies boundary conditions $f^{(k)}(1) = f^{(k)}(0) = 0$ for $k = m, m+1, \cdots, 2m-1$. The solid curves in Figure 1 show three different smoothing spline fits (with $m = 2$) to the same data. The data were generated according to the model

$$y_i = f(x_i) + \epsilon_i, \quad , i = 1, \cdots, n,$$

where $f(x) = 4.26(e^{-x} - 4e^{-2x} + 3e^{-3x})$ is given by the dashed lines in each plot, the x_i are $n = 100$ equally spaced abscissae, and the ϵ_i came from a random number generator simulating independent, identically distributed $\mathcal{N}(0, \sigma^2)$ random variates with $\sigma = .2$. The top plot was obtained using a value of λ that was too small, and the middle plot was obtained using a value of λ too large, to recover the underlying curve. The bottom plot was obtained using a value of λ which was estimated from the data using the method of Generalized Cross Validation (GCV), which will be described later. It can be shown, under the same conditions as stated to guarantee a unique solution, that, as λ tends to 0, the curve will come closer and closer to interpolating the data, and as λ becomes larger and larger, the curve will approach the polynomial of degree $m-1$ which best fits the data in a least squares sense. For example, the solid line in middle plot would eventually flatten out to a straight line if λ were made much larger.

[4] To be very specific, $\mathcal{H}$ is the Sobolev Hilbert space of functions with $m-1$ absolutely continuous derivatives, and $\int_0^1 (f^{(m)}(u))^2 du < \infty$

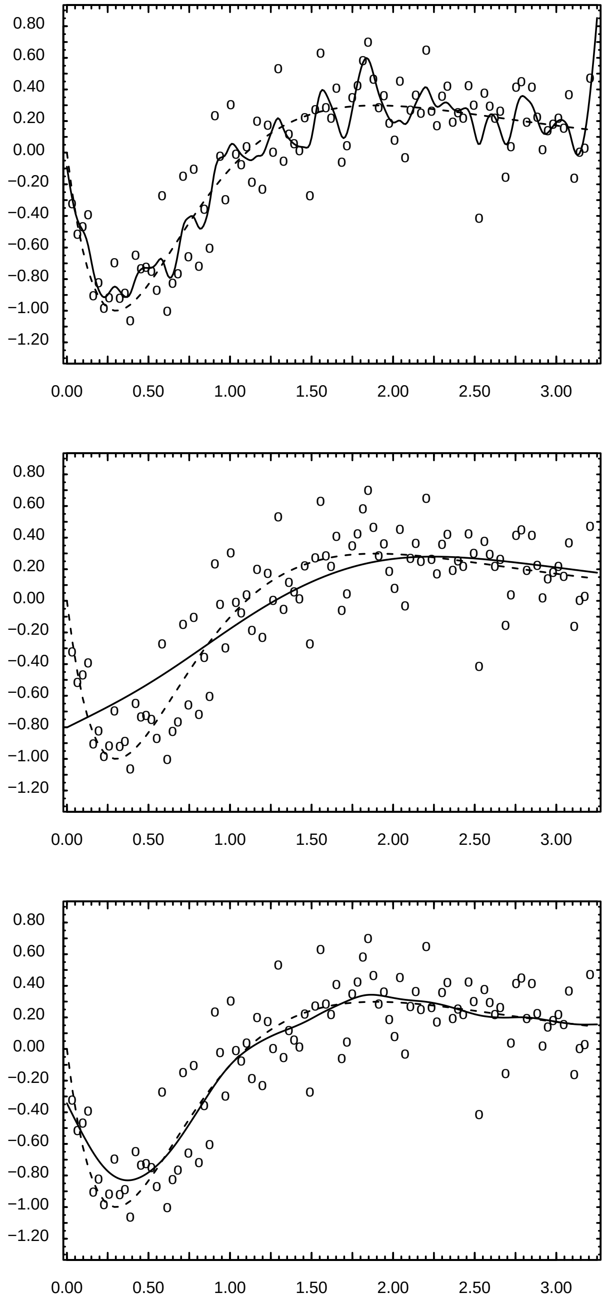


Figure 1: The polynomial smoothing spline ©SIAM 1990

3 Choosing the smoothing parameter

3.1 Ordinary Cross Validation, or, Leaving-out-one

Let $f_\lambda^{[k]}$ be the minimizer of

$$\frac{1}{n}\sum_{\substack{i=1\\i\neq k}}^{n}(y_i - f(x_i))^2 + \lambda\int_0^1 (f^{(m)}(u))^2 du, \tag{3}$$

the variational problem with the kth data point left out. Then the "ordinary cross validation function" $V_0(\lambda)$ is defined as

$$V_0(\lambda) = \frac{1}{n}\sum_{k=1}^{n}\left(y_k - f_\lambda^{[k]}(x_k)\right)^2, \tag{4}$$

and the leaving out one estimate of λ is the minimizer of $V_0(\lambda)$. To proceed, we need to describe the influence matrix. It is not hard to show that, for fixed λ and each x_k that $f_\lambda(x_k)$ is a linear combination of the components of $y = (y_1, \cdots, y_n)'$, and so there exists a matrix $A(\lambda)$ satisfying

$$\begin{pmatrix} f_\lambda(x_1) \\ \vdots \\ f_\lambda(x_n) \end{pmatrix} = A(\lambda)y. \tag{5}$$

The Leaving-Out-One Lemma (Craven & Wahba (1979)) gives us a very useful mathematical identity, which will not be proved here, but is:

$$(y_k - f_\lambda^{[k]}(x_k)) \equiv (y_k - f_\lambda(x_k))/(1 - a_{kk}(\lambda)) \tag{6}$$

where $a_{kk}(\lambda)$ is the kkth entry of $A(\lambda)$. By substituting (6) into (4) we get a simplified form for V_0, which, again, is a mathematical identity:

$$\frac{1}{n}\sum_{k=1}^{n}(y_k - f_\lambda^{[k]}(x_k))^2 \equiv V_0(\lambda) \equiv \frac{1}{n}\sum_{k=1}^{n}(y_k - f_\lambda(x_k))^2/(1 - a_{kk}(\lambda))^2. \tag{7}$$

The right hand version of (7) is easier to compute than the left, however, the GCV, described next, is even easier.

3.2 Generalized Cross Validation

Generalized Cross Validation (GCV) is a method for choosing the smoothing parameter, which is based on leaving-out-one, but it has two advantages: Firstly, it is easier to compute, and, secondly, it possess some important theoretical properties that would be impossible to prove for leaving-out-one, although in many examples the GCV and leaving-out-one estimates will give answers that are close. Theoretical properties of the GCV may be found in Craven & Wahba (1979), Golub, Heath & Wahba (1979), Li (1986). The GCV function $V(\lambda)$ is obtained by replacing $a_{kk}(\lambda)$ in $V_0(\lambda)$ by $\bar{a}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} a_{ii}(\lambda) = \frac{1}{n}trA(\lambda)$. The GCV function $V(\lambda)$ is defined by

$$V(\lambda) = \frac{1}{n}\sum_{k=1}^{n}(y_k - f_\lambda(x_k))^2/(1 - \bar{a}_{kk}(\lambda))^2 \equiv \frac{\frac{1}{n}\|(I - A(\lambda))y\|^2}{[\frac{1}{n}tr(I - A(\lambda))]^2}. \tag{8}$$

$V(\lambda)$ may be viewed as a weighted version of $V_0(\lambda)$, since $V(\lambda) = \frac{1}{n}\sum_{k=1}^{n}\left(y_k - f_\lambda^{[k]}(x_k)\right)^2 w_{kk}(\lambda)$ where $w_{kk}(\lambda) = (1-a_{kk}(\lambda))^2/(1-\bar{a}(\lambda))^2$. If $a_{kk}(\lambda)$ does not depend on k, then $V_0(\lambda) \equiv V(\lambda)$. The λ used to obtain the bottom plot of Figure 1 was chosen via the *GCV* method. FORTRAN freeware for computing the smoothing spline with λ chosen by *GCV* may be found in the codes `sbart` (O'Sullivan) and `gcvspl` (Woltring) in the `gcv` directory of netlib `http://www.netlib.org/gcv/`. The R freeware system `http://r-project.org/`, contains the smoothing spline code `pspline` (Ramsay), and the Splus commercial package contains the code `smooth.spline()`. The smoothing spline with *GCV* may also be found as a special case in some more general codes described later.

4 Regression splines

Given a large data set, a modest number of basis functions which are themselves spline functions may be generated, and used as a basis set for regression. Popular choices for regression splines are the truncated power functions $\phi_j(x) = (x - x_j)_+^{2m-1}$, $j = 1, 2, ..L$, augmented by low degree polynomials, and the B-splines. B-splines are polynomial splines which satisfy (2), and have minimal support, that is, they have the fewest possible number of knots. Simple examples of B-splines are obtained as a (scaled, shifted) convolution of k uniform densities. The B-spline will be a piecewise polynomial of degree $k - 1$ and for k bigger than 3 the B-splines obtained this way will tend to look very much like normal curves over most of their domain. See deBoor (1978) for more on B-splines. Using a set of scaled, shifted B-splines for regression may be visualized as approximating the desired (smooth) function by regressing on a set of shifted hill-functions. The 'wiggliness' of the result will depend on the scale and number of the B-splines. Truncated power functions and tensor products of them in higher dimensions provide the basis functions for the popular MARS algorithm (Friedman (1991)). Ridge regression is sometimes carried out with a set of splines for regression functions. In this case the fit is chosen in the span of the basis functions to minimize the residual sum of squares plus a penalty functional on the coefficients or on the fit. Such fits are sometimes known as hybrid spline fits.

5 The thin plate spline

The thin plate spline is a natural generalization of univariate polynomial spline to two or more dimensions via a generalization of the variational problem in (1). The variational problem in Euclidean two-space is: Find f in an appropriate space $\mathcal{X}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_1(i), x_2(i)))^2 + \lambda \sum_{\nu=0}^{m}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\begin{array}{c} m \\ \nu \end{array}\right)\left(\frac{\partial^m f}{\partial x_1^\nu \partial x_2^{m-\nu}}\right)^2 dx_1 dx_2. \tag{9}$$

Note that the limits on the integral are ±∞. If a finite boundary is specified, then a boundary value problem must be solved numerically. The definition of $\mathcal{X}$ is beyond the scope of this article, see Duchon (1977). With the limits at infinity, the minimizer has a representation:

$$f_\lambda(t) = \sum_{\nu=1}^{\binom{m+1}{2}} d_\nu \phi_\nu(t) + \sum_{i=1}^{n} c_i E_m(t, t(i)), \tag{10}$$

where $t = (x_1, x_2)$, $t(i) = (x_1(i), x_2(i))$, the ϕ^ν are the $\binom{m+1}{2}$ monomials $1, x_1, x_2, x_1x_2, ...$ of total degree less than m, and $E_m(t, t(i)) = |t - t(i)|^{2m-2}\ln|t - t(i)|$, with $|t - t(i)| = [(x_1 - x_1(i))^2 +$

$(x_2 - x_2(i))^2]^{1/2}$. When f (of the form (10)) is a minimizer of (9) it is known that the $\{c_i\}$ must satisfy a certain condition under which (9) is finite and has a known closed form expression as a quadratic form in the $\{d_\nu\}$ and $\{c_i\}$. Details may be found in Wahba & Wendelberger (1980) or Wahba (1990).

Figures 2, 3 and 4, from Wahba (1990) show, respectively, a test surface, noisy data from the test surface, the ($m = 2$) thin plate spline fit to this data with λ too large, the fit with λ too small and the fit with λ estimated by GCV. (Equations (5) and (8) serve to define the GCV estimate here and more generally). With λ too large the surface flattens out, and as λ tends to infinity, the surface would flatten to the least squares plane best fitting the data. As λ tends to 0 the surface would tend to interpolate the data. In general, a unique solution to the variational problem always exists for any non-negative λ if the least squares fit to the polynomials of degree $m - 1$ or less exists uniquely. Generalizations to three and higher dimensions are available, see Wahba (1990) and references cited there.

The Fortran freeware `GCVPACK` (Bates, Lindstrom, Wahba and Yandell) for the thin plate spline may be found in the `gcv` directory of netlib noted earlier. The `funfits` code `http://www.cdg.ucar.edu/stats/software.shtml` (Nychka) at NCAR also contains thin plate spline freeware. Commercial code may be found in SAS (`tpspline`), and `ANUSPLIN` (`http://cres20.anu.edu/au/software/anusplin.html`).

6 The thin plate histospline

Frequently one is interested in obtaining a graphical representation of a geographically distributed quantity when only its averages or integrals over a region are given. In an example discussed in Wahba (1981a), 1970-1975 standardized age adjusted female lung cancer rates in Wisconsin are given by county for the 72 counties. Let y_i by the rate for the ith county, Ω_i be the ith county, and $|\Omega_i|$ be the area of the ith county. The volume-smoothing histospline is given as the solution to the following variational problem: Find f in $\mathcal{X}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n} w_i(y_i - \frac{1}{|\Omega_i|}\int_{\Omega_i} f(u)du)^2 + \lambda\sum_{\nu=0}^{m}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\begin{array}{c} m \\ \nu \end{array}\right)\left(\frac{\partial^m f}{\partial x_1^\nu \partial x_2^{m-\nu}}\right)^2 dx_1 dx_2. \quad (11)$$

where the w_i are some weights. The solution is known to have a representation of the form

$$f_\lambda(t) = \sum_{\nu=1}^{\binom{m+1}{2}} d_\nu \phi_\nu(t) + \sum_{i=1}^{n} c_i \frac{1}{|\Omega_i|}\int_{\Omega_i} E_m(t,u)du, \quad (12)$$

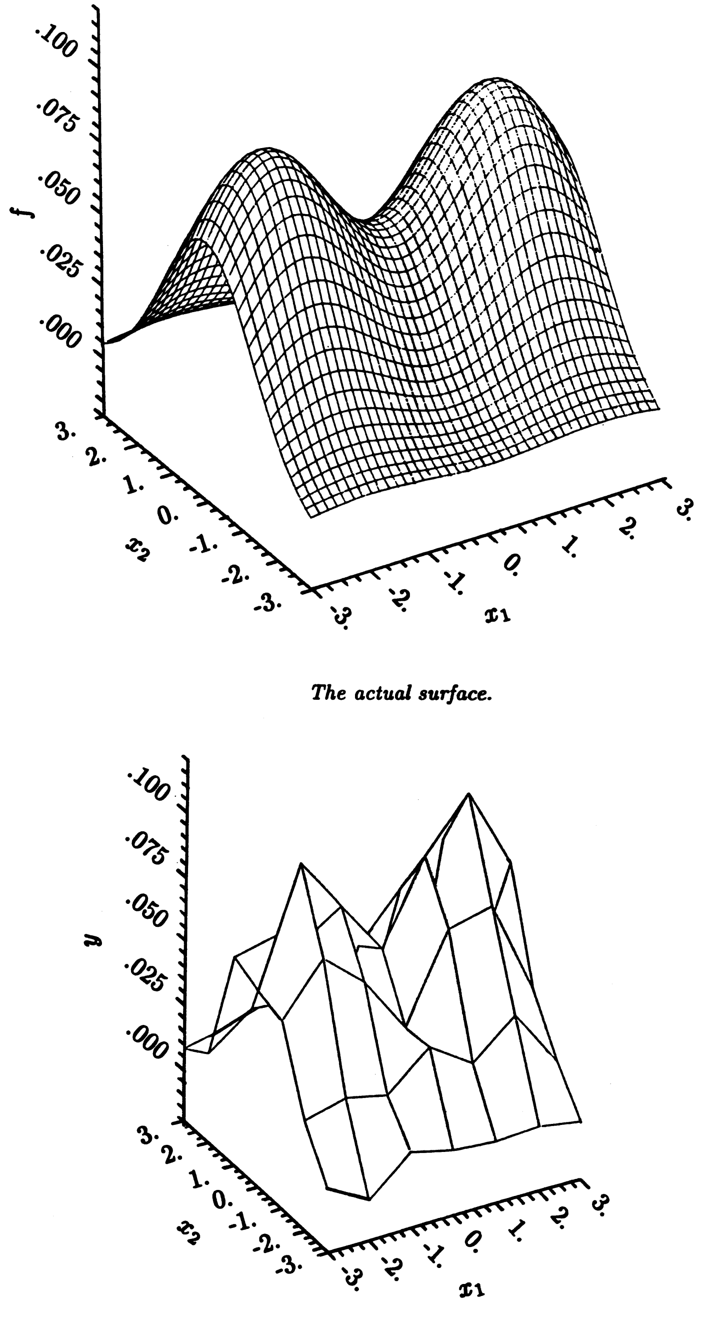
where the ϕ^ν and E_m are as before. The minimizer exists and will be unique if the weighted least squares fit to the ϕ_ν is unique. As $\lambda \to 0$ the minimizer will tend to the histospline that minimizes

$$\sum_{\nu=0}^{m}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\begin{array}{c} m \\ \nu \end{array}\right)\left(\frac{\partial^m f}{\partial x_1^\nu \partial x_2^{m-\nu}}\right)^2 dx_1 dx_2. \quad (13)$$

subject to the volume-matching property:

$$\frac{1}{|\Omega_i|}\int f dt = y_i, i = 1, \cdots, n. \quad (14)$$

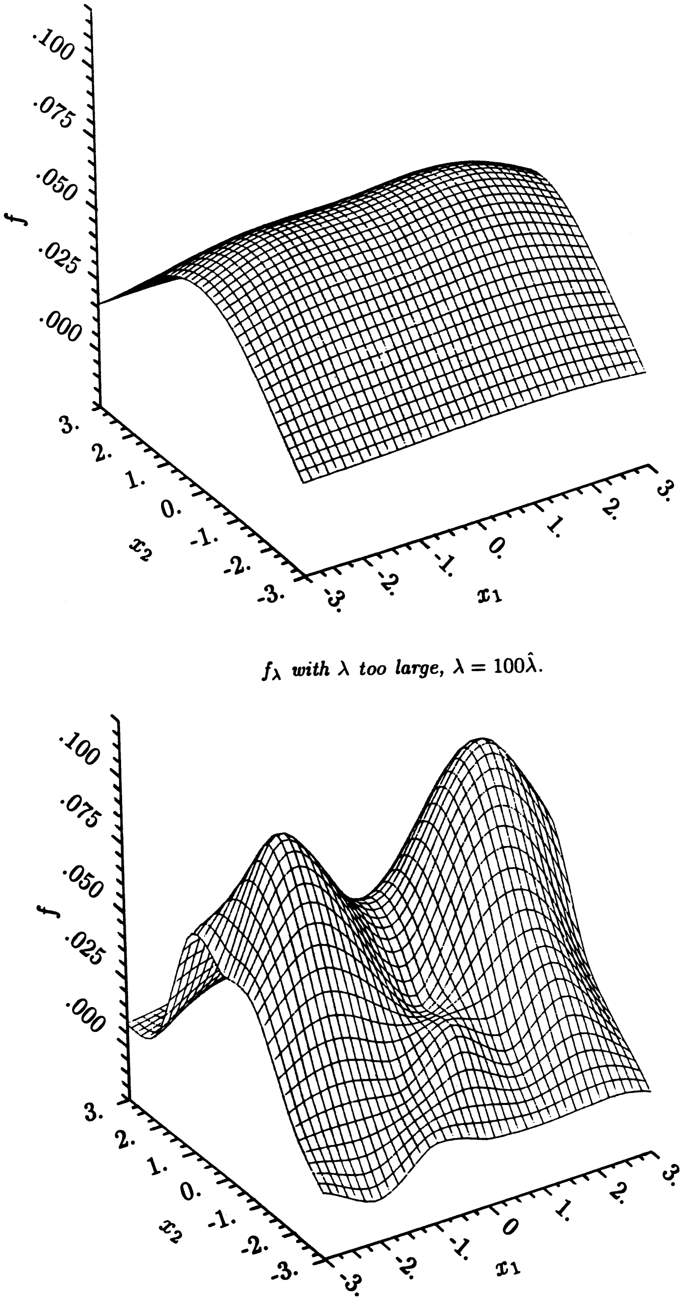
Theoretical results are in Dyn & Wahba (1982), see also Dyn, Wahba & Wong (1979). The trick in carrying out the calculations of the thin plate histospline is to approximate the functions of t defined by $\int_{\Omega_i} E_m(t,u)du$ as a sum of functions of t given by $\frac{1}{\#(\Omega_i)}\sum_{u_k \in \Omega_i} E_m(t, u_k), i = 1, \cdots, n,$



The actual surface.

The data.

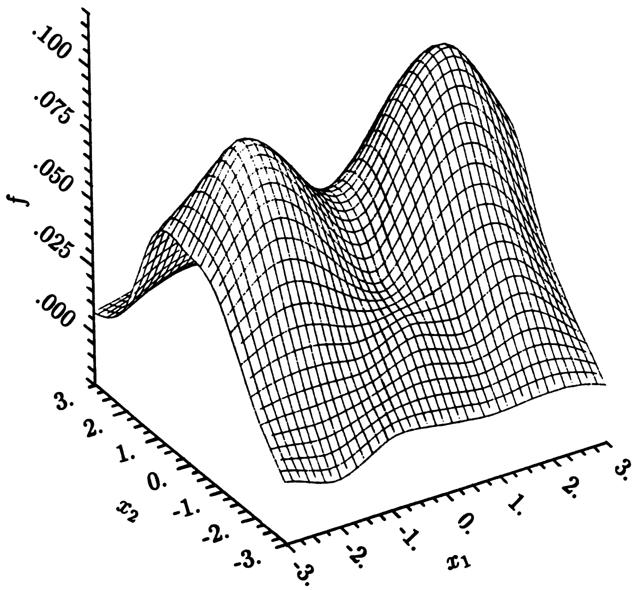
Figure 2: Thin plate spline demonstration. Top: True surface. Bottom: Noisy observations.©SIAM 1990



f_λ with λ too large, $\lambda = 100\hat{\lambda}$.

f_λ with λ too small, $\lambda = .01\hat{\lambda}$.

Figure 3: Thin plate spline estimates. Top: f_λ with λ too large. Bottom: f_λ with λ too small.©SIAM 1990



f_λ with λ estimated by GCV.

Figure 4: Thin plate spline estimate. $f_{\hat{\lambda}}$ with λ estimated by GCV. ©SIAM 1990

where the u_k are a fine rectangular grid of points over $\Omega = \cup \Omega_i$ and $\#(\Omega_i)$ is the number of grid points in county i. With this approximation, then, as in the preceeding thin plate spline example, a closed form expression for (11) as a quadratic form in $\{d_\nu, c_i\}$ is known, and may be minimized by solving a linear system. See Wahba (1981a) for computational details.

7 Splines on the sphere

To define splines on the sphere, it is useful to discuss the spherical harmonics, and the (surface) Laplacian on the sphere. Let θ be longitude, $(0 \le \theta \le 2\pi)$ and ϕ be latitude $(-\frac{\pi}{2} \le \phi \le \frac{\pi}{2})$. The spherical harmonics are defined by

$$Y_{\ell s} = \begin{cases} \theta_{\ell s} \cos(s\lambda) P_{\ell s}(\sin \phi) & 0 < s \le \ell \\ \theta_{\ell s} \sin(s\lambda) P_{\ell |s|}(\sin \phi) & -\ell \le s < 0 \\ \theta_{\ell 0} P_\ell(\sin \phi), & s = 0 \end{cases}$$

$$\ell = 0, 1, 2, \ldots ,$$

where the $\theta_{\ell s}$ are constants not reproduced here, P_ℓ are the Legendre polynomials and $P_{\ell,s}$ are the Legendre functions, see, for example Sansone (1959).

The spherical harmonics are the eigenfunctions of the (horizontal) Laplacian Δ on the sphere:

$$\Delta f = \frac{1}{a^2} \left[\frac{1}{\cos^2 \phi} f_{\theta\theta} + \frac{1}{\cos \phi} (\cos \phi f_\phi)_\phi \right],$$

specifically,

$$\Delta Y_{\ell s} = -\ell(\ell+1)Y_{\ell s}$$

where the subscripts refer to derivatives with respect to θ and ϕ, i. e. longitude and latitude. The spherical harmonics play the same role on the sphere as sines and cosines on the circle. Note, for example that the sines and cosines are eigenfunctions of D^2 the second derivative: $D^2 \sin(2\pi x) = -2\pi \sin(2\pi x)$. Just as the sines and cosines form a complete orthonormal basis for square integrable functions on the circle, the spherical harmonics form a complete orthonormal basis for the square integrable functions on the sphere.

Letting $P = (\theta, \phi)$, a (smoothing) spline f_λ on the sphere may be defined as the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f_\lambda(P_i))^2 + \int_{\mathcal{S}}(\Delta^{m/2}f)^2 dP \tag{15}$$

where $\mathcal{S}$ is the sphere. (Fractional powers of Δ can be defined in terms of the action of Δ on the Fourier coefficients of f in its expansion in spherical harmonics). It can be shown that the minimizer has a representation of the form

$$f_\lambda(P) = d + \sum_{i=1}^{n} c_i R(P, P_i) \tag{16}$$

where

$$R(P, P') = \sum_{\ell=1}^{\infty}\sum_{s=-\ell}^{\ell}\left[\frac{1}{\ell(\ell+1)}\right]^m Y_{\ell s}(P)Y_{\ell s}(P'). \tag{17}$$

Unfortunately, closed form expressions for $R(P, P')$ are not available in general, but infinite series expansions which have similar behavior as R of (17), and which have closed form expressions, have appeared in Wahba (1981*b*) and Wahba (1981a), for a range of m's.

We next illustrate splines on the sphere, using as the example a hybrid adaptive spline, from Luo & Wahba (1997). Figure 5 gives a contour plot of the average December-January-February surface temperature for 1980-81, based a hybrid adaptive spline on the sphere using data from $n = 725$ temperature observing stations. (See Luo & Wahba (1997) for further details of the observational data.) Dots indicate the location of the observing stations. The hybrid adaptive spline was obtained as follows: Let $Q(P, P')$ be the approximation to R for $m = 2$. Then a basis for the (approximate) smoothing spline is the functions $1, Q(\cdot, P_i), i = 1, \cdots, n$ where "$\cdot$" indicates a function of P. A subset of these basis functions is chosen by a forward stepwise selection procedure, using as a stopping criteria a GCV criteria with an inflated degrees of freedom factor to account for the fact that the basis functions are chosen adaptively. Then, the selected basis functions are used as a basis for a penalized least squares regression, with the GCV estimate of λ. However, little smoothing has been done at this last stage, because the limited number of basis functions has already reduced the degrees of freedom. The net effect is that there will be relatively more basis functions where the data has more structure, or is denser, and fewer elsewhere.

Figure 6 is the enlarged European part of Figure 5. It can be seen that the surface is generally quite smooth, but there is a sharp minimum over the Alps, where the surface temperature is colder. This adaptive procedure allows the retention of some 'rough' features while not undersmoothing the smooth areas.

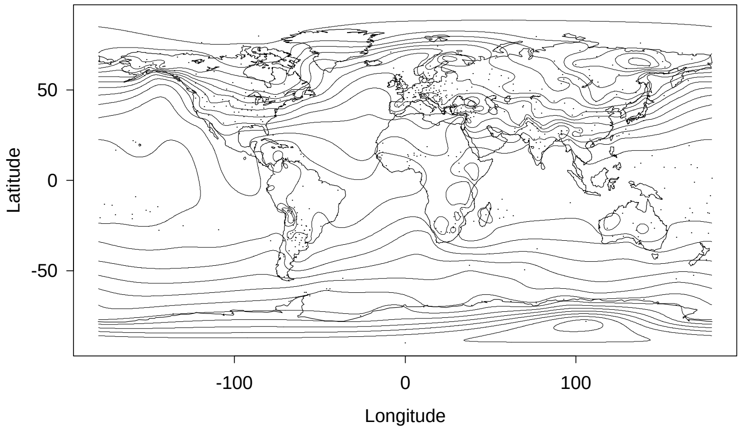


Figure 5: December-January-February 1980-81 average winter surface temperature. ©ASA 1997

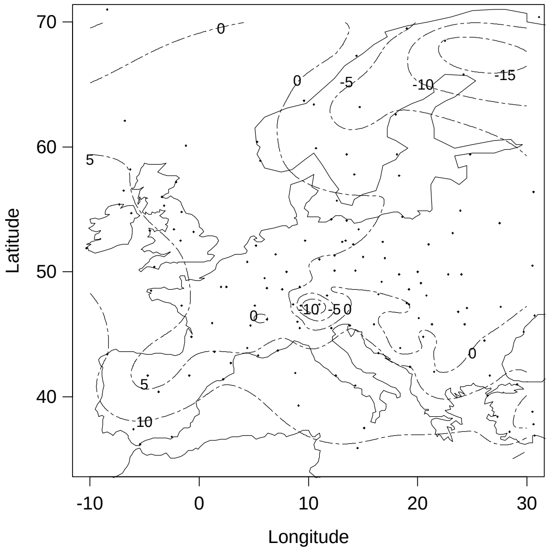


Figure 6: European area of Figure 5 enlarged. ©ASA 1997

Vector splines on the sphere can be used to estimate wind fields from observational data, and to model transformations from the sphere to itself as a tool in modeling anisotropic covariance functions on the sphere, via the technique of using the vector spline to define a transformation from the sphere to itself, and then using an isotropic covariance in the transformed coordinate system. Let the desired vector field be $\mathbf{V}(P) = (U(P), V(P))$ where $U(P)$ is the eastward component and $V(P)$ is the northward component at P. By Helmholtz' theorem, there exist two functions Ψ and Φ defined on $\mathcal{S}$, called the stream function and the velocity potential respectively, such that

$$U = \frac{1}{a}\left(-\frac{\partial\Psi}{\partial\phi} + \frac{1}{\cos\phi}\frac{\partial\Phi}{\partial\theta}\right), \qquad V = \frac{1}{a}\left(\frac{1}{\cos\phi}\frac{\partial\Psi}{\partial\theta} + \frac{\partial\Phi}{\partial\phi}\right), \tag{18}$$

where a is the radius of the sphere. The vorticity ζ and the divergence D of $\mathbf{V}$ are defined by

$$\zeta = \frac{1}{a\cos\phi}\left[-\frac{\partial}{\partial\phi}(U\cos\phi) + \frac{\partial V}{\partial\theta}\right], \qquad D = \frac{1}{a\cos\phi}\left[\frac{\partial U}{\partial\theta} + \frac{\partial}{\partial\phi}(V\cos\phi)\right] \tag{19}$$

and $\zeta = \Delta\Psi, D = \Delta\Phi$. Given data (U_i, V_i) from the model $U_i = U(P_i) + \epsilon_i^U$, $V_i = V(P_i) + \epsilon_i^V$ where the ϵ_i^U and ϵ_i^V are random errors or discrepancies, a smoothing spline vector field based on this data is defined as $\mathbf{V}_\lambda = (U_\lambda, V_\lambda)$ where $U_\lambda = \frac{1}{a}\left(-\frac{\partial\Psi_\lambda}{\partial\phi} + \frac{1}{\cos\phi}\frac{\partial\Phi_\lambda}{\partial\theta}\right)$, $V_\lambda = \frac{1}{a}\left(\frac{1}{\cos\phi}\frac{\partial\Psi_\lambda}{\partial\theta} + \frac{\partial\Phi_\lambda}{\partial\phi}\right)$ and $\Psi_\lambda, \Phi_\lambda$ are the minimizers of

$$\frac{1}{n}\sum_{i=1}^{n}\left(U_i - \frac{1}{a}\left[-\frac{\partial\Psi}{\partial\phi}(P_i) + \frac{1}{\cos\phi}\frac{\partial\Phi}{\partial\theta}(P_i)\right]\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(V_i - \frac{1}{a}\left[\frac{1}{\cos\phi}\frac{\partial\Psi}{\partial\theta}(P_i) + \frac{\partial\Phi}{\partial\phi}(P_i)\right]\right)^2$$

$$+\left[\lambda_1\int_{\mathcal{S}}(\Delta^{m/2}\Psi)^2 dP + \lambda_2\int_{\mathcal{S}}(\Delta^{m/2}\Phi)^2 dP\right].$$

See Wahba (1982b) for details.

8 Partial Splines

Partial splines are data fits involving some variables that are modeled parametrically and some modeled as splines. In Gu & Wahba (1993*b*) lake acidity (pH) as a function of calcium content (t_1) and latitude and longitude ($t_2 = (x_1, x_2)$) in the Blue Ridge mountains was modeled as a partial spline. The model was

$$y_i = f_1(t_1) + f_2(t_2) + \epsilon_i, \quad i = 1, \cdots, n, \tag{20}$$

where f_1 is linear in t_1 and f_2 is a thin plate spline. Side conditions, amounting to requiring the average of the thin plate part over the observation points to be 0, are built into the thin plate spline so that the constant term is identifiable. Possible dependence on location is of interest here since the rectangular lake region considered contains the crest of the Blue Ridge mountains running from SE to NW diagonally across it. The model was fitted as the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - d_0 - d_1 t_1(i) - f_2(t_2(i))^2 + \lambda\sum_{\nu=0}^{2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\begin{pmatrix}2\\ \nu\end{pmatrix}\left(\frac{\partial^2 f_2}{\partial x_1^\nu \partial x_2^{2-\nu}}\right)^2 dx_1 dx_2. \tag{21}$$

where $t_k(i)$ is the value of $t_k, k = 1, 2$ at the ith observation point. GCV was used to choose λ. The side condition on the thin plate spline is enforced using the theory of smoothing spline ANOVA (mentioned later), but see Gu & Wahba (1993*b*) for details. This kind of model can be fitted using the code RKPACK (Gu), which may be found in the GCV directory of netlib, or in the R library, where it is part of a suite called gss.

Figure 7 gives the result. The geographic plot contains the states VA, TN, NC, SC and GA. The dotted lines are the state boundaries and the solid lines are the geography main effect. A comparison with an elevation contour plot of the region suggests that the geography effect is related to elevation. The highest peak of the mountain ridge lies slightly west of the center of the geography plot.

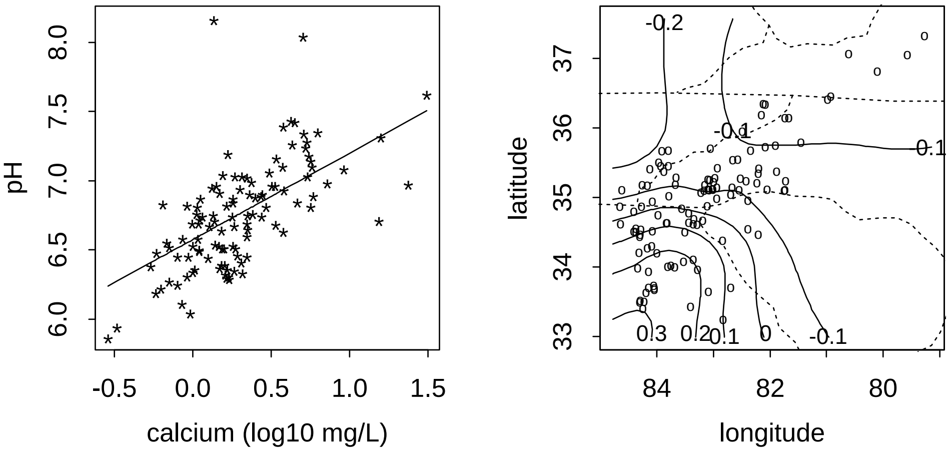


Figure 7: Lake acidity as a function of calcium content and location. The constant term is incorporated into the calcium main effect, and the geography plot is scaled so that the average over the observation points is 0. ©ASA,IMS,IFNA 1993

9 Smoothing Spline ANOVA

Let $t = (t_1, \cdots, t_d) \in \mathcal{T}^{(1)} \otimes \cdots \otimes \mathcal{T}^{(d)} = \mathcal{T}$, where the $\mathcal{T}^{(\alpha)}$ are measurable spaces of rather general form. The examples of $\mathcal{T}^{(\alpha)}$ we have seen in this article so far are the unit interval, the plane and the sphere, but more general domains are possible, including categorical and ordered categorical data. Given data $\{y_i, t(i), i = 1, \cdots, n\}$, a smoothing spline ANOVA model is a fit, by penalized likelihood, to a function of the form $f(t) = C + \sum_\alpha f_\alpha(t_\alpha) + \sum_{\alpha<\beta} f_{\alpha\beta}(t_\alpha, t_\beta) + \cdots$, where, in interesting cases, the functions are splines of various kinds. The components of the decomposition satisfy side conditions which generalize the usual side conditions for parametric ANOVA, and thus the components will all be identifiable. If $\mathcal{E}_\alpha$ is some (given) averaging operator over functions in $\mathcal{T}^{(\alpha)}$ then these conditions are of the form $\mathcal{E}_\alpha f_\alpha = 0$, $\mathcal{E}_\alpha f_{\alpha,\beta} = \mathcal{E}_\beta f_{\alpha,\beta} = 0$, etc. Examples of averaging operators include (weighted) integrals, and averages over observation points, which have the property that the average of the constant function 1 is 1. Details may be found in Wahba, Wang, Gu, Klein & Klein (1995*b*), where it is observed that the components are orthogonal projections of f onto subspaces of an appropriately defined reproducing kernel Hilbert space. The estimate f_λ is obtained as the minimizer, in an appropriate function space, of $\mathcal{L}(y, f) + \sum_\alpha \lambda_\alpha J_\alpha(f_\alpha) +$

$\sum_{\alpha<\beta}\lambda_{\alpha\beta}J_{\alpha\beta}(f_{\alpha\beta})+\cdots$, where $\mathcal{L}(y,f)$ is the negative log likelihood of $y=(y_1,\cdots,y_n)'$ given f. The $J_\alpha, J_{\alpha\beta},\cdots$ are quadratic penalty functionals (essentially, like those we have seen, or tensor products of them, one (oversimplified) example being $\int\int\left(\partial^{2m}f/\partial t_\alpha^m\partial t_\beta^m\right)^2 dt_\alpha dt_\beta)$, but various generalizations are possible. The ANOVA decomposition is terminated in some manner. A popular special case is the main effects model, with only sums of functions of one variable. The main effects models are studied in some detail in Hastie & Tibshirani (1990). Here λ stands for the collection $\lambda_\alpha,\lambda_{\alpha\beta}$ etc. So far we have considered only Gaussian data, that is, the log likelihood is a multiple of the residual some of squares $\sum_{i=1}^n(y_i-f_\lambda(t(i))^2$, but this quantity may be replaced by densities from the exponential family of the form $h(y_i,f_i)=exp[y_if_i-b(f_i)+c(y_i)]$, where we are letting $f_i=f(t(i))$ and b and c are given functions with b is twice continuously differentiable and bounded away from 0. In particular the example we will consider below involves Bernoulli data ($y_i\in\{0,1\}$), then $f_i=ln[p_i/(1-p_i)]$ with p_i the probability that $y_i=1$, $b(f)=ln(1+expf)$ and $c=0$. For $\mathcal{T}^{(\alpha)}$ and penalty functionals including those we have considered so far, the solution to to the variational problem is a linear combination of $n+M$ known basis functions satisfying M conditions, where M is the dimension of the unpenalized part of f. For example, in the polynomial spline on the unit interval with penalty functional $\int(f^{(m)})^2$ the polynomials of degree $m-1$ or less are unpenalized, and then $M=m$. In the spline ANOVA cases here the basis functions considered as a function of one variable with the other variables fixed, are splines of various kinds. This property of the solution holds whether the likelihood term is residual sum of squares, or, whether it is some other convex functional in f. However, if the likelihood term is not quadratic, then a variational problem for the $M+n$ unknown coefficients has to be solved numerically by some iterative method, typically by a Gauss-Newton iteration.

Wahba et al. (1995*b*) examined the four year risk of progression of diabetic retinopathy in a selected subgroup of 669 subjects in the Wisconsin Epidemiological Study of Diabetic Retinopathy (WESDR) as a function of the baseline predictor variables `dur`, `gly`, and `bmi`, respectively duration of diabetes, glycosylated hemoglobin and body mass index. Various model selection activities took place before these variables were selected from a larger set, similarly, the model given below was arrived at after some study, described in the paper. The model was

$$f(\texttt{dur},\texttt{gly},\texttt{bmi})=\mu+f_1(\texttt{dur})+a_2\cdot\texttt{gly}+f_3(\texttt{bmi})+f_{13}(\texttt{dur},\texttt{bmi}) \tag{22}$$

The penalty functionals for the main effects f_1 and f_3 were, after scaling the problem to the unit cube, based on $\int_0^1(f^{(2)}(u))^2du$. The interaction term actually has three parts, which we describe heuristically here. They are a term which is the product of a linear term in f_1 and a 'smooth' term in f_3, a corresponding term with 1 and 3 reversed, and a term that is smooth in both variables and is penalized by the tensor product penalty described above. Thus there are actually 5 smoothing parameters in this model. In models with more variables it may be necessary to enforce relationships between some of the λ's. The smoothing parameters were chosen by an iteratively computed version of the unbiased risk estimate (Mallows (1973), Craven & Wahba (1979)) proposed by Gu (1990). The Gaussian version of this model can be fit using `RKPACK`. The Bernoulli case here was fit using `GRKPACK`(Wang), available in `netlib`, which calls `RKPACK`. See also `gss`. The left panel in Figure 8 gives a plot of the data as a function of `bmi` and `gly`. Solid circles represent subjects with progression, and open circles those without progression. The contours are contours of posterior standard deviation, which are not described here, but are a guide to the region where the fit may be considered reliable. The right panel gives the estimated probability of progression as a function of `dur` and `bmi` with `gly` fixed at its median.

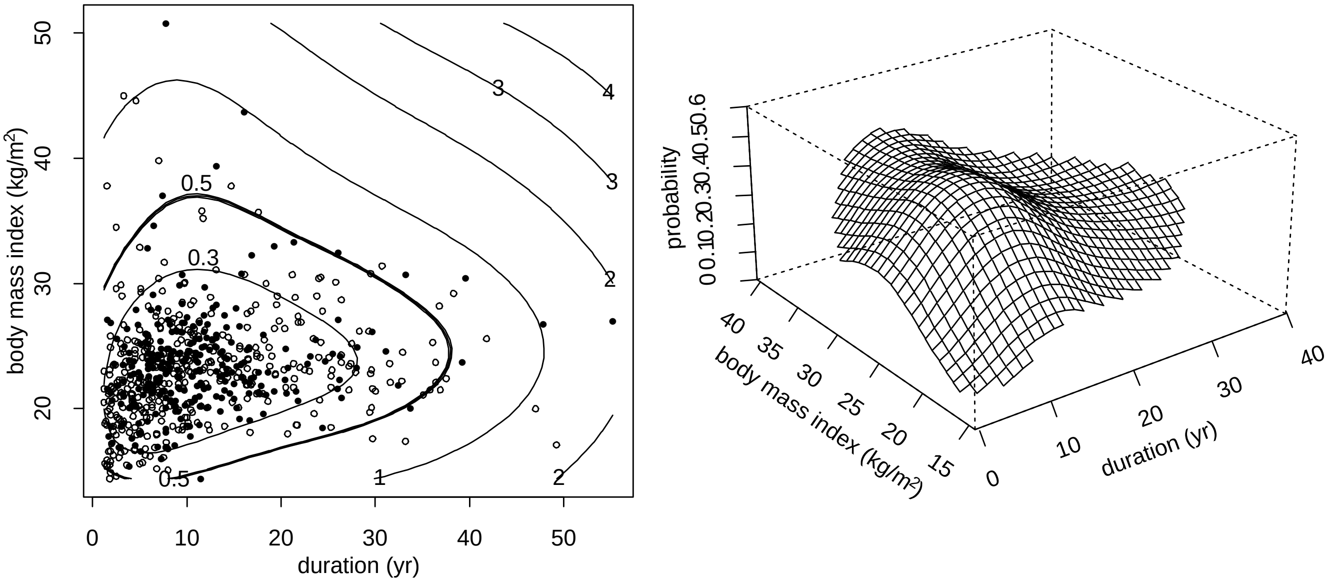


Figure 8: Left: Data and contours of constant posterior standard deviation. Right: Estimated probability of progression as a function of duration and body mass index for glycosylated hemoglobin fixed at its median. ©IMS 1995

A time and space ANOVA spline model, with time $t = 1, 2, \cdots, 30$ years, and space $P = lat., long.$ was studied in Chiang, Wahba, Tribbia & Johnson (1999) for the purpose of examining space-time patterns of global warming, based on $n = 23,119$ observations over the thirty year period 1961-90, from 1,000 observing stations. It is

$$f(t, P) = d_1 + d_2\phi(t) + f_1(t) + f_2(P) + f_{\phi,2}(P)\phi(t) + f_{12}(t, P), \tag{23}$$

where $\phi(t) = (t - 31/2)$. The component functions of f satisfied the following moment conditions

$$\begin{aligned} \sum_{t=1}^{n_1} f_1(t) &= \sum_{t=1}^{n_1} f_1(t)\phi(t) = \sum_{t=1}^{n_1} f_{12}(t, P) = \sum_{t=1}^{n_1} f_{12}(t, P)\phi(t) = 0 && (24) \\ \int_{\mathcal{S}} f_2(P)dP &= \int_{\mathcal{S}} f_{\phi,2}(P)dP = \int_{\mathcal{S}} f_{12}(t, P)dP = 0 && (25) \end{aligned}$$

for all t and P. The penalty functionals were built up from the penalty functional of (15) on the sphere, and the sum of squares second differences on $\{1, 2, \cdots, 30\}$. See Wahba (1990), Gu & Wahba (1993*a*) and Luo, Wahba & Johnson (1998) for details about formulating such ANOVA models. These components are climatologically meaningful. d_1 is the grand mean of average winter temperature; d_2 is the linear trend coefficient of the global average winter temperature; $d_1 + d_2\phi(t) + f_1(t)$ is the global average winter temperature history; $d_1 + f_2(P)$ is the average winter temperature at location P; $d_2 + f_{\phi,2}(P)$ is the linear trend coefficient of the average winter temperature at location P. $f_{\phi,2}(P)$ represents a pattern of the anomaly (deviation from the average) of the linear rate of change of temperature as a function of space, and is of interest in comparing with similar patterns generated by climate models that are forced with greenhouse gases.

All of the multivariate models we have described so far are generalizations of univariate splines as solutions to variational problems. Generalizations as piecewise polynomials are, of course possible.

We just mention two of the most popular statistical models in higher dimensions based on piecewise polynomials - they are MARS (Friedman (1991)), and POLYCLASS (Kooperberg, Bose & Stone (1997)).

10 Remarks on Computations

A characteristic of all of the spline models considered here which are obtained as solutions of a variational problem is that they are known to reside in a space of $n + M$ basis functions satisfying M conditions, where M is the dimension of the unpenalized part of the solution. In the case of the univariate spline, a set of basis functions with compact support is known, which leads to the requirement that a linear system with banded matrix be solved to obtain the coefficients of the spline. This result is utilized in the univariate spline software cited. In general, the matrix of the linear system to be solved is not sparse, however, which limits the size of the data set that can be handled, unless some tricks can be employed. One trick is to observe that, with very large data sets, if a sufficiently large representative or stratified sample of N basis functions are used, then very accurate approximations to the exact solution using all the basis functions can be obtained, at the cost of solving a linear system of dimension N. This procedure is included in `ANUSPLIN` and has been employed elsewhere. Note that the basis functions selected by the forward stepwise procedure in Luo & Wahba (1997) are not a representative or stratified sample and were chosen as they were for a different reason than to obtain a cheap approximation to the original variational problem. Backfitting algorithms have been employed in a descent algorithm to fit additive and other spline ANOVA models, see Hastie & Tibshirani (1990), Wahba et al. (1995*b*). Conjugate gradient algorithms for solving the quadratic minimization problem have also been found to be useful. Very large problems can be solved using a conjugate gradient algorithm that is not iterated to convergence, but is stopped after a carefully chosen number of iterations. Heuristically speaking, the conjugate gradient algorithm tends to project the solution onto smooth subspaces first, so that stopping the iteration has a smoothing effect. The *GCV* has been used to choose the number of iterations along with a smoothing parameter, see Wahba, Johnson, Gao & Gong (1995*a*). In the time and space model cited in Section 9, if every one of the 1000 stations had 30 years of observations, then the matrices involved in the linear system to be solved would have a tensor product structure as the tensor product of a 1000×1000 matrix and a 30×30 matrix. Then solving the linear systems could loosely speaking be reduced to solving a 1000×1000 system and a 30×30 system. To analyze the $n = 23,119$ observations in Chiang et al. (1999), an imputation trick adapted from Luo et al. (1998) was used to fill in the missing observations, so that the tensor product structure could be used to advantage. Imputing the data iteratively, it was shown there that the solution there converged to the actual solution given the original $23,119$ observations.

A major advance in calculating splines with the smoothing parameter(s) chosen by *GCV* occurred with the publication of the randomized trace estimate of $trA(\lambda)$, see Girard (1989), Hutchinson (1989). It is based on the observation that if z is a vector of n zero mean independent random variables with variance 1 then the expected value of $z'Az$ is trA. Thus, if we let $f_\lambda^y = \begin{pmatrix} f_\lambda(x_1) \\ \vdots \\ f_\lambda(x_n) \end{pmatrix} = A(\lambda)y$ and $f_\lambda^{y+z} = A(\lambda)(y+z)$ then $z'[f_\lambda^{y+z} - f_\lambda^y]$ is an estimate of $trA(\lambda)$. This means that the GCV function with large data sets can be evaluated with the additional cost of computing the fit with data $y + z$ in addition to the fit with data y. This argument is quite independent of the particular numerical approximations that have been used to obtain the solution

- whatever method is used, the randomized trace technique is estimating the trace of the effective influence matrix. This method, or variants of it have proved to be very accurate, see Girard (1998), and it has been used in the Bernoulli case with large complex sets employing the GACV (Xiang & Wahba (1996)) for choosing smoothing parameters, not discussed here, see Gao, Wahba, Klein & Klein (2001).

References

Chiang, A., Wahba, G., Tribbia, J. & Johnson, D. (1999), A quantitative study of smoothing spline-ANOVA based fingerprint methods for attribution of global warming, Technical Report 1010, Department of Statistics, University of Wisconsin, Madison WI.

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

deBoor, C. (1978), *A Practical Guide to Splines*, Springer-Verlag, New York.

Duchon, J. (1977), Splines minimizing rotation-invariant semi-norms in Sobolev spaces, *in* 'Constructive Theory of Functions of Several Variables', Springer-Verlag, Berlin, pp. 85–100.

Dyn, N. & Wahba, G. (1982), 'On the estimation of functions of several variables from aggregated data', *SIAM J. Math. Anal.* **13**, 134–152.

Dyn, N., Wahba, G. & Wong, W. (1979), 'Comment on "Smooth pychnophylactic interpolation for geographical regions by W. Tobler', *J. Am. Statist. Assoc.* **74**(367), 530–535.

Friedman, J. (1991), 'Multivariate adaptive regression splines', *Ann. Statist* **19**, 1–141.

Gao, F., Wahba, G., Klein, R. & Klein, B. (2001), 'Smoothing spline ANOVA for multivariate Bernoulli observations, with applications to ophthalmology data', *Ann. Statist.* **96**, xx–xx.

Girard, D. (1989), 'A fast 'Monte-Carlo cross-validation' procedure for large least squares problems with noisy data', *Numer. Math.* **56**, 1–23.

Girard, D. (1998), 'Asymptotic comparison of (partial) cross-validation, GCV and randomized GCV in nonparametric regression', *Ann. Statist.* **126**, 315–334.

Golub, G., Heath, M. & Wahba, G. (1979), 'Generalized cross validation as a method for choosing a good ridge parameter', *Technometrics* **21**, 215–224.

Gu, C. (1990), 'Adaptive spline smoothing in non-Gaussian regression models', *J. Amer. Statist. Assoc.* **85**, 801–807.

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

Gu, C. & Wahba, G. (1993*b*), 'Smoothing spline ANOVA with component-wise Bayesian "confidence intervals"', *J. Computational and Graphical Statistics* **2**, 97–117.

Hastie, T. & Tibshirani, R. (1990), *Generalized Additive Models*, Chapman and Hall, 335pp.

Hutchinson, M. (1989), 'A stochastic estimator for the trace of the influence matrix for Laplacian smoothing splines', *Commun. Statist.-Simula.* **18**, 1059–1076.

Kimeldorf, G. & Wahba, G. (1971), 'Some results on Tchebycheffian spline functions', *J. Math. Anal. Applic.* **33**, 82–95.

Kooperberg, C., Bose, S. & Stone, C. (1997), 'Polychotomous regression', *Journal of the American Statistical Association* **92**, 117–127.

Li, K. C. (1986), 'Asymptotic optimality of C_L and generalized cross validation in ridge regression with application to spline smoothing', *Ann. Statist.* **14**, 1101–1112.

Luo, Z. & Wahba, G. (1997), 'Hybrid adaptive splines', *J. Amer. Statist. Assoc.* **92**, 107–114.

Luo, Z., Wahba, G. & Johnson, D. (1998), 'Spatial-temporal analysis of temperature using smoothing spline ANOVA', *J. Climate* **11**, 18–28.

Mallows, C. (1973), 'Some comments on C_p', *Technometrics* **15**, 661–675.

Ramsay, J. O. & Dalzell, C. J. (1991), 'Some tools for functional data analysis (disc: P561-572)', *Journal of the Royal Statistical Society, Series B, Methodological* **53**, 539–561.

Sansone, G. (1959), *Orthogonal Functions*, Interscience, New York.

Schoenberg, I. (1964a), 'Spline functions and the problem of graduation', *Proc. Nat. Acad. Sci. U.S.A.* **52**, 947–950.

Schoenberg, I. (1964b), 'On interpolation by spline functions and its minimum properties', *Int. Ser. Numer. Anal.* **5**, 109–129.

Wahba, G. (1981a), Cross validation and constrained regularization methods for mildly ill posed problems, Technical Report 629, Statistics Dept., University of Wisconsin, Madison, WI.

Wahba, G. (1981*a*), 'Numerical experiments with the thin plate histospline', *Commun. Statist.-Theor. Meth.* **A10**, 2475–2514.

Wahba, G. (1981*b*), 'Spline interpolation and smoothing on the sphere', *SIAM J. Sci. Stat. Comput.* **2**, 5–16.

Wahba, G. (1982b), Vector splines on the sphere, with application to the estimation of vorticity and divergence from discrete, noisy data, *in* W. Schempp & K. Zeller, eds, 'Multivariate Approximation Theory, Vol.2', Birkhauser Verlag, pp. 407–429.

Wahba, G. (1990), *Spline Models for Observational Data*, SIAM. CBMS-NSF Regional Conference Series in Applied Mathematics, v. 59.

Wahba, G. & Wendelberger, J. (1980), 'Some new mathematical methods for variational objective analysis using splines and cross-validation', *Monthly Weather Review* **108**, 1122–1145.

Wahba, G., Johnson, D., Gao, F. & Gong, J. (1995*a*), 'Adaptive tuning of numerical weather prediction models: randomized GCV in three and four dimensional data assimilation', *Mon. Wea. Rev.* **123**, 3358–3369.

Wahba, G., Wang, Y., Gu, C., Klein, R. & Klein, B. (1995*b*), 'Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy', *Ann. Statist.* **23**, 1865–1895. Neyman Lecture.

Xiang, D. & Wahba, G. (1996), ‘A generalized approximate cross validation for smoothing splines with non-Gaussian data’, *Statistica Sinica* **6**, 675–692.

List of Figures