DEPARTMENT OF STATISTICS

University of Wisconsin
1210 W. Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 722

September 1983

CROSS VALIDATED SPLINE METHODS FOR THE ESTIMATION OF MULTIVARIATE FUNCTIONS FROM DATA ON FUNCTIONALS

by

Grace Wahba

This research was sponsored by the Office of Naval Research under Contract No. N00014-77-C-0675.

To appear in the Proceedings of the Iowa State University Statistical Laboratory 50th Anniversary Conference, H.A. David, ed.

ABSTRACT

The relation between spline estimation, ridge regression, and Bayes estimation is reviewed. Then a description of multivariate thin plate (TP) smoothing splines in two and several dimensions is given. Several generalizations and related methods are described, including spline estimation where the data are noisy observations on nonlinear functions; when families of linear inequalities, such as positivity, are known a priori; and when measurement error is contaminated Gaussian. Penalized log likelihood density estimation is seen to result in spline estimation. Some recent related work on penalized GLIM methods, which also result in spline estimates, is mentioned. Some numerical results using smoothing multivariate TP splines which are constrained to be between 0 and 1 are given. Finally we describe multivariate partial TP splines which are TP splines in some directions and low degree polynomials in others and are suitable for semiparametric modelling in 4 or more dimensions.

Key Words: cross validated splines, thin plate splines, inequality constrained smoothing splines.

1. Introduction

Smoothing spline estimates for the smooth nonparametric estimation of functions of one or several variables are frequently thought of as a useful descriptive tool. However, in addition to being esthetically pleasing, they enjoy many nice theoretical properties, including best obtainable integrated mean square convergence properties. With the advent of modern software such as EISPACK and LINPACK, recent developments in numerical analysis and computational statistics, and the availability of fourth generation computing equipment, such previously intractable problems as the computation of multidimensional smoothing splines subject to linear inequality constraints becomes feasible,

In the remainder of this introduction we will briefly review the major ideas behind cross validated smoothing spline estimates, for functions of one variable, with Gaussian measurement errors. We describe their relationship to ridge regression, Bayesian estimation and roughness penalty methods. In Section 2 we describe the multivariate thin plate smoothing splines (TPSS). In Section 3 we mention some extensions of spline estimation to non-linear nonparametric regression, to problems where side information such as nonnegativity is available, and to robust smoothing. We note that penalized maximum likelihood density estimation leads to splines. Then we describe some recent work of O'Sullivan on penalized GLIM methods, which also fit into the spline framework. In Section 4 we present some numerical results obtained by Villalobos, on the estimation of posterior probabilities in the classification problem, for bivariate observations, using constrained cross validated TPSS. These results demonstrate the feasibility of computing nonnegative smoothing splines in two dimensions. Finally in Section 4, we propose and describe a new approach using partial thin plate spline models for the partially nonparametric smooth estimation of functions of more than 3 or 4 variables. In the partial spline models some of the variables enter in a specified parametric form while only a "smooth" relationship is assumed for the others.

Smoothing spline estimates for a smooth function are simultaneously i) solutions to an optimization problem; ii) Bayes estimates and iii) the output of a generalized low pass filter. As such, they enjoy nice properties of all of the above. These estimates may also be thought of as ridge regression estimates in Hilbert space. We will illustrate the dual Bayesian-optimization nature of these estimates by a ridge regression example. Let

$$y = X\beta + \varepsilon$$

where $y = (y_1,\ldots,y_n)'$, X is $n{\times}p$,

$$\varepsilon \sim N(0,\sigma^2 I_{n\times n}), \ \beta \sim N(0,b\textstyle\sum_{p\times p}),$$

$\sum$ is known and b and σ^2 are unknown. Let $\lambda = \sigma^2/nb$. Let

$$\beta_\lambda = E(\beta|y) = (X'X+n\lambda\textstyle\sum)^{-1}X'y.$$

Let $I_\lambda(\beta)$ be defined by

$$I_\lambda(\beta) = \frac{1}{n}||y-X\beta||_n^2 + \lambda\beta'\textstyle\sum^{-1}\beta. \qquad (1.1)$$

Here $||y-X\beta||_n^2$ is the residual sum of squares, and $\lambda\beta'\sum^{-1}\beta$ is the "roughness" penalty". We have the following easy to check but important

Theorem:

β_λ is the minimizer of $I_\lambda(\cdot)$.

Thus, the minimizer of $I_\lambda(\cdot)$ is also a Bayes estimate.

The Hilbert space version of the duality between optimization problems and Bayes estimates goes as follows: Let τ be some index set, for example

$\tau = 1,2,\ldots,p$ (corresponds to ridge regression)

$\tau = [0,1]$

$\tau =$ circle

$\tau = E^d$, Euclidean d-space

$\tau = S$, the sphere

$\tau = S \otimes [0,1]$, the atmosphere, etc.

We suppose that it is desired to estimate $f(t), t\varepsilon\Omega\subset\tau$, where f is a real-valued function on τ. Suppose that $f(t)$, $t\varepsilon\tau$ is a zero mean Gaussian stochastic process with covariance $Ef(s)f(t) = bR(s,t)$, where $R(\cdot,\cdot)$ is known. We let H_R be the reproducing kernel Hilbert space with reproducing kernel (r.k) $R(\cdot,\cdot)$. (For more on r.k. spaces, see Aronszajn (1950), Parzen (1962), Hajek (1962), Kimeldorf and Wahba (1970)). Let

$$y_i = L_i f + \varepsilon_i, \ i = 1,2,\ldots,n, \tag{1.2}$$

where L_i is a bounded linear functional on H_R and $\varepsilon \sim N(0,\sigma^2 I)$. Reproducing kernel spaces are exactly those Hilbert spaces for which the evaluation functionals $L_i f = f(t_i)$ are bounded linear functionals and hence are the natural Hilbert spaces in which to work if one either observes noisy values of the function or wishes to estimate functional values. Under weak assumptions

$$L_i f = \int_\tau K(t_i,s)f(s)ds$$

is also a bounded linear functional on H_R and in some r.k. spaces $L_i f = f^{(k)}(t_i)$ is a bounded linear functional for $k = 1,2,\ldots,m-1$. Furthermore, if $f(\cdot)$ is a zero mean Gaussian stochastic process with covariance $bR(\cdot,\cdot)$, it can be shown that $L_i f$ is a random variable well defined in quadratic mean if and only if L_i is a bounded linear functional on H_R. It is appropriate to think of $L_i f$ as the Hilbert space generalization of $L_i(\beta) = \sum_j x_{ij}\beta_j$. Now let $\lambda = \sigma^2/nb$, and let

$$f_\lambda(t) = E\{f(t)|y\}, \quad t\varepsilon\tau .$$

It is easy to write down an explicit formula for $f_\lambda(t)$ for fixed t using standard formulas from multivariate analysis once the $(n+1)\times(n+1)$ covariance matrix for $\{f(t), y_1,\ldots,y_n\}$ is known. However it is known that

$$Ef(t)L_i f = bL_{i(s)}R(s,t)$$

$$EL_i fL_j f = bL_{i(s)}L_{j(t)}R(s,t)$$

where $L_{i(s)}$ means that the linear functional L_i is to be applied to the subsequent expression considered as a function of s, so that this covariance matrix is usually easy to write down. For example, if $L_i f = \int K(t_i,s)f(s)ds$

$$EL_i fL_j f = b\int\int K(t_i,s)R(s,u)K(t_j,u)dsdu.$$

Let $I_\lambda(f)$ be defined by

$$I_\lambda(f) = \frac{1}{n}\sum_{i=1}^{n}(L_i f-y_i)^2 + \lambda||f||_R^2 \qquad (1.3)$$

where $||\cdot||_R$ is the norm in H_R. $\sum_{i=1}^{n}(L_i f-y_i)^2$ is the residual sum of squares, and $||f||_R^2$, the smoothness penalty, is the analogue of $\beta'\sum^{-1}\beta$. We have

Theorem: (Kimeldorf and Wahba (1971))

f_λ is the minimizer of $I_\lambda(\cdot)$.

Thus, the minimizer of I_λ is a Bayes estimate.

We would like to replace $||f||_R^2$ by $J(f)$, where $J(f)$ is a seminorm in H_R with a finite dimensional null space. This is the Hilbert space analogue of the case where $\sum^{-1}$ is not of full rank. This rank deficiency will come about if, for example, we let one or more eigenvalues of $\sum$ tend to infinity, thus endowing some linear functional (s) of β with an improper prior. The most celebrated example is the polynomial smoothing spline. In this example, $\tau = [0,1]$, $\mathcal{H}_R$ is the Sobolev space W_2^m (Hilbert space of functions with m-1 continuous derivatives and mth derivative in $\mathcal{L}_2[0,1]$, see Adams (1975)), and $L_i f = f(t_i)$.

$$I_\lambda(f) = \frac{1}{n}\sum_{i=1}^{n} (f(t_i)-y_i)^2 + \lambda\int_0^1 (f^{(m)}(u))^2 du. \tag{1.4}$$

Here $J(f) = \int_0^1 (f^{(m)}(u))^2 du$ has a null space in W_2^m, spanned by the polynomials of degree less than m. If there are at least m distinct values of the $\{t_i\}$, then I_λ has a unique minimizer for each $\lambda > 0$. The minimizer, call it f_λ, is a polynomial of degree m-1 in $[0,t_1]$ and in $[t_n,1]$ and is a polynomial of degree 2m-1 in each interval $[t_i,t_{i+1}]$, $i = 1,2,\ldots,n-1$, with the pieces joined so that f_λ has 2m-2 continuous derivatives. The Bayesian model corresponding to I_λ is

$$f(t) \sim \sum_{\nu=1}^{m} \theta_\nu t^\nu + b\int_0^t \int_0^{u_{m-1}} \cdots \int_0^{u_1} dW(u)$$

where $E\theta_\nu = \infty$, $\int\ldots\int$ is the m-fold integrated Weiner process and $\lambda = \sigma^2/nb$. Details may be found in Wahba (1978, 1983a). The bandwidth parameter λ controls the tradeoff between fidelity to the data and smoothness of the solution. As $\lambda \to \infty$, f_λ tends to the polynomial of degree m-1 best fitting the data in a least squares sense, and as $\lambda \to 0$, f_λ tends to that element in W_2^m which minimizes $\int_0^1 (f^{(m)}(u))^2 du$ subject to $f_\lambda(t_i) = y_i$, $i = 1,2,\ldots,n$. If $\lambda \to 0$ as $n \to \infty$ but not "too fast", then $f_\lambda \to f$ in various ways, and much is known about convergence rates, see Craven and Wahba (1979), Cox (1983), Ragozin (1981), Rice and Rosenblatt (1981, 1983), Utreras (1981b) and Wahba (1975).

The "bandwidth" parameter λ as well as the "shape" parameter m can be estimated from the data by the method of generalized cross validation (GCV). The GCV estimate $\hat{\lambda}$ for λ is the minimizer of

$$V(\lambda) = \frac{\frac{1}{n}||(I-A(\lambda))y||^2}{(\frac{1}{n}\text{Trace}(I-A(\lambda)))^2} \tag{1.5}$$

where $A(\lambda)$ is the n×n influence matrix which satisfies

$$\begin{pmatrix} L_1 f_\lambda \\ \vdots \\ L_n f_\lambda \end{pmatrix} = A(\lambda)y.$$

$\hat{\lambda}$ estimates the λ which minimizes the predictive mean square error

$$R(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(L_i f - L_i f_\lambda)^2 \tag{1.6}$$

under a wide range of assumptions on f. Properties of $\hat{\lambda}$ and related estimators are an area of active research, see Chow, Geman and Wu (1983), K. Li (1983a,b), Speckman (1982), Utreras (1981b), Wahba (1983b). A computer program which computes $\hat{\lambda}$ and $f_{\hat{\lambda}}$ for the m = 2 case is available in IMSL (1981).

2. Spline estimates of functions of several variables

Several generalizations of spline functions to more than one variable have appeared in the recent approximation theory literature. If one wants an isotropic penalty functional J(.) on Euclidean d-space, with the polynomials of total degree less than m in the d variables for the null space, one is lead to the thin plate (TP) splines. The original results characterizing thin plate splines were provided by Duchon (1976) and Meinguet (1979) and later elaborated on by Wahba and Wendelberger (1980). Numerical methods for computing these splines, with the GCV estimate of λ, may be found in Bates and Wahba (1983), Wahba (1980b, 1984) and Wendelberger (1981, 1982).

Integrated mean square error convergnece rates for TPS were conjectured in Wahba (1979) and a number of rigorous results provided in Cox (1982). The facts given in the rest of this section may be found in the above references.

For d = 2 dimensions and m = 2 the smoothness penalty J(f) associated with the TPS is

$$J(f) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{x_1x_1}{}^2 + 2f_{x_1x_2}{}^2 + f_{x_2x_2}{}^2 dx_1 dx_2 \qquad (2.1)$$

and for m > 2

$$J(f) = \sum_{\nu=0}^{m}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\binom{m}{\nu}\left[\frac{\partial^m f}{\partial x_1^{m-\nu}\partial x_2^{\nu}}\right]^2 dx_1 dx_2 \qquad (2.2)$$

and, in general in d-dimensions

$$J(f) = \sum_{\sum\alpha_i=m} \frac{m!}{\alpha_1!\ldots\alpha_d!} \int\ldots\int \left[\frac{\partial^m f}{\partial x_1^{\alpha_1}\ldots\partial x_d^{\alpha_d}}\right]^2 dx_1\ldots dx_d. \tag{2.3}$$

We want to estimate $f(x_1,\ldots,x_d) = f(t)$, given data

$$y_i = L_i f + \varepsilon_i, \quad i = 1,2,\ldots,n \tag{2.4}$$

with $\varepsilon \sim N(0,\sigma^2 I)$. The thin plate smoothing spline estimate of f is the minimizer (in an appropriate function space X) of

$$\frac{1}{n}\sum_{i=1}^{n} (L_i f - y_i)^2 + \lambda J(f). \tag{2.5}$$

X is a space of real valued functions on Euclidean d-space E^d, all of whose derivatives of total order m are square integrable, see Meinguet (1979). The null space of J(f) in this space is the $M = M(d,m) = \binom{d+m-1}{d}$ polynomials of total degree less than m. For example, if $d = 2$, $m = 3$, then $M = 6$ and these 6 polynomials are

$$\begin{array}{lll} \phi_1(x_1,x_2) = 1 & \phi_2(x_1,x_2) = x_1 & \phi_3(x_1,x_2) = x_2 \\ \phi_4(x_1,x_2) = x_1^{\,2} & \phi_5(x_1,x_2) = x_1x_2 & \phi_6(x_1,x_2) = x_2^{\,2} \end{array}.$$

It is necessary that $2m-d>0$ in order that X be a reproducing kernel space, that is, in order that the evaluation functionals be bounded. Expression (2.5) will have a unique minimizer if $t_1,\ldots,t_n$ are such that the $n\times M$ matrix T with $i\nu$th entry $L_i\phi_\nu$ is of rank M. An explicit representation of the minimizer of (2.5) is given by

$$f_\lambda(t) = \sum_{\nu=1}^{M} d_\nu\phi_\nu(t) + \sum_{i=1}^{n} c_i\xi_i(t) \tag{2.6}$$

where

$$\xi_i(t) = L_{i(s)}E_m(t,s), \qquad i = 1,2,\ldots,n. \tag{2.7}$$

Here $L_{i(s)}$ means the linear functional L_i applied to what follows considered as a function of s, and $E_m(t,s)$ is given by $E_m(t,s) = E(|t-s|)$, where, if $t = (x_1,\ldots,x_d)$, $s = (u_1,\ldots,u_d)$, then $|t-s| = (\sum_{i=1}^{d}(x_i-u_i)^2)^{1/2}$ and

$$\begin{aligned} E(|\tau|) &= \theta_{m,d}|\tau|^{2m-d}\ell n|\tau| && \text{d even} \\ &= \theta_{m,d}|\tau|^{2m-d} && \text{d odd} \end{aligned} \tag{2.8}$$

$$\theta_{m,d} = \frac{(-1)^{d/2+1+m}}{2^{2m-1}\pi^{d/2}(m-1)!(m-d/2)!}, \text{ d even}$$

$$\theta_{m,d} = \frac{\Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!}, \text{ d odd.}$$

Letting K be the n × n matrix with ijth entry $E_m(t_i,t_j)$, then $c = (c_1,\ldots,c_n)'$ and $d = (d_1,\ldots,d_M)'$ satisfy

$$(K+n\lambda I)c + Td = y \tag{2.9}$$

$$T'c = 0. \tag{2.10}$$

We remark that in the case $d = 1$ and $L_i f = f(t_i)$ the expression (2.6) reduces to a function which is a polynomial of degree 2m-1 between t_i and t_{i+1}, for each $i = 1,2,\ldots,n-1 (t=x_1)$, and by virtue of (2.10), is a polynomial of degree m-1 for $t \le t_1$ and $t \ge t_n$. Also, symbolically, we have

$$D^{2m}f_\lambda = \sum_{i=1}^{n} c_i\delta_i$$

where δ_i is the delta function at t_i. This relationship generalizes to d dimensions as follows: Let Δ be the Laplacian operator in d dimensions - $\Delta f = f_{x_1x_1} + f_{x_2x_2} + \ldots + f_{x_dx_d}$. Now E is the fundamental solution of the m-th iterated Laplacian in d dimensions (see. e.g. Schwartz (1966)), $\Delta^m E = \delta_0$, so that, in general when $L_i f = f(t_i)$,

$$\Delta^{2m} f_\lambda = \sum_{i=1}^{n} c_i \delta_i .$$

The influence matrix $A(\lambda)$ satisfies

$$I - A(\lambda) = R(RKR'+n\lambda I)^{-1}R' \qquad (2.11)$$

where R is any $n \times n - M$ matrix satisfying $R'R = I_{n-M}$, $T'R = 0_{M \times n-M}$, and the GCV estimate $\hat{\lambda}$ of λ may be computed as the minimizer of (1.5). Wendelberger and others have successfully computed $f_{\hat{\lambda}}$ for n as large as 350, using EISPACK and LINPAC. The numerical work in computing f_λ and $\hat{\lambda}$ is roughly independent of d but is $O(n^3)$. However, it is evident that for large d, it is necessary for n to be large to obtain a good estimate of f. It is also possible to choose m by GCV (Wahba and Wendelberger (1980), and to choose at least one scale parameter in the d = 3 case by GCV. (See Hutchinson et al, (1983), Wendelberger (1982)). It is desirable to do this when, e.g. x_1 and x_2 are space variables and x_3 is time - the units of x_3 relative to x_1 and x_2 must be selected in some reasonable manner.

Several recent developments allow the computation of f for d = 2 and 3 for n of the order of 1000. Instead of computing the minimizer of (2.5) in X exactly via (2.6-2.10) one selects a convenient N dimensional subspace $X_N = \{\text{span } B_\ell,\ \ell = 1,2,\dots,N\}$ of X, where the $\{B_\ell\}$ are known to have good approximation theoretic properties. Then one seeks the minimizer of (2.5) in X_N, and if N is large enough. under suitable conditions this will be a good approximation to f_λ, the minimizer in X. The influence matrix $A(\lambda)$ may be found for this approximate method, and $\hat{\lambda}$ obtained by minimizing (1.5).

For d = 1, the natural choice for X_N is a space of so-called B-splines of degree 2m+1 (see deBoor (1978), Wahba (1980a)) which are hill functions formed from smoothly joined piecewise polynomials. In two dimensions one may use a space of tensor product B-splines (see deBoor (1978)) or use the thin plate basis functions (TPBF's) suggested in Wahba (1980b). These basis functions are obtained as follows: Choose $s_1,\dots,s_N$, N points in E_d, spread around in an appropriate manner (and such that the N×M matrix with $\ell\nu$th entry $\phi_\nu(s_\ell)$ is of rank M) and let

$$B_\ell = \phi_\ell,\ \ell = 1,2,\dots,M$$

$$\text{span}\{B_{M+\ell}\}_{\ell=1}^{N-M} = \text{span } \{\sum_{r=1}^{N} c_{\ell r} E_m(\cdot - s_r),\ \ell = 1,\dots,N-M\}$$

where the $c_\ell = (c_{\ell 1}\dots c_{\ell N})'$, $\ell = 1,\dots,N-M$ are N-M linearly independent N-vectors which are "generalized divided differences of order m", that is

$$\sum_{r=1}^{N} c_{\ell r}\phi_\nu(s_r) = 0,\ \nu = 1,2,\dots,M,\ \ell = 1,2,\dots,N-M.$$

This space is equivalent to a space of B-splines if d = 1, and for d > 1, also spans a space of hill functions (see Dyn and Levin (1981)). The A matrix can be easily written down. Hutchinson et al. (1983)have successfully used these TPBF's for estimating Australian solar radiation as a function of latitude, longitude and rainfall index. $f_{\hat{\lambda}}$ can be computed for even larger data sets by combining TPBS with a truncated singular value decomposition, see Bates and Wahba (1983).

The intrinsic random functions of Matheron (1973) provide one description of the stochastic process model behind the TPS. Some of the procedures behind kriging used in mining engineering are related to Matheron's intrinsic random functions, and hence to TPS. The connection between kriging models and spline smoothing also appears in Kimeldorf and Wahba (1971), and Duchon (1976).

3. Generalizations

(1) Nonlinear functionals. In numerous physical problems one observes mildly nonlinear functionals of the function one desires to estimate. For example, the relationship between vertical atmospheric temperature T(p) as a function of pressure p, and satellite observeable upwelling radiance R_ν and wavenumber ν is given by

$$R_\nu(T) = B_\nu(T(p_0))\tau_\nu(p_0) - \int_0^{p_0} B_\nu(T(p)]\tau_\nu'(p)dp,$$

where p_0 is the surface pressure, 0 is the pressure at the top of the atmosphere, $\tau_\nu(p)$ is the transmittance of the atmosphere above pressure p at wavenumber ν, and B_ν is Planck's function, given by

$$B_\nu(T(p)) = c_1\nu^3/[\exp(c_2\nu/T(p)-1],$$

where c_1 and c_2 are constants, see Fritz et al. (1972). Satellite radiance measurements are modelled by

$$y_{\nu_i} = R_{\nu_i}(T) + \varepsilon_i, \qquad i = 1,2,\ldots,n,$$

where R_{ν_i} (T) is the above nonlinear functional of T. Letting N_i be a general nonlinear functional, O'Sullivan (1983) considers the estimation of f given data

$$y_i = N_i f + \varepsilon_i, \qquad i = 1,2,\ldots,n$$

$\varepsilon \sim N(0,\sigma^2 I)$, by the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}(N_i f - y_i)^2 + \lambda J(f). \tag{3.1}$$

O'Sullivan provides existence theorems and proposed and tested a numerical algorithm for obtaining minimizer(s) of (3.1). The algorithm consists of solving a sequence of linear problems where the kth problem consists of finding the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}(y_i^k - L_i^k f)^2 + \lambda J(f) \tag{3.2}$$

where

$$y_i^k = y_i - N_i(f_\lambda^{k-1}) + L_i^k f_\lambda^{k-1}$$

$$L_i^k = N_i'(f^{k-1}).$$

Here f_λ^{k-1} is the minimizer for the k-1st problem, and $N_i'(f_\lambda^{k-1})$ is the Frechet derivative of N_i evaluated at f_λ^{k-1}. ($N_i'(f)$) is a continuous linear functional on H_R for each $f \varepsilon H_R$ under suitable assumptions). This is a Gauss-Newton method based on linearizing $N_i f$ as

$$N_i f \simeq N_i f_\lambda^{k-1} + N_i'(f_\lambda^{k-1})(f - f_\lambda^{k-1}) \tag{3.3}$$

(A good starting guess for T(p) is generally available in the satellite radiance problem). There are two advantages of this iteration. Firstly, discretization of the problem occurs late or not at all, since the minimizer of (3.2) in function space can be expressed explicitly in terms of a finite number of unknowns. Secondly, the software developed for the linear problem may be called as a subroutine to minimize (3.2).

In solving ill posed problems (as this one is) numerically, it is generally important to do any necessary discretization as late as possible.

Using the generalized cross validation function for nonlinear problems in Wahba (1980a), namely

$$V(\lambda) = \frac{\frac{1}{n}RSS(\lambda)}{[\frac{1}{n}Tr(I-A(\lambda))]^2} \tag{3.4}$$

where

$$A_{ij}(\lambda) = \frac{\partial N_i f_\lambda}{\partial y_j}\Bigg|_y$$

O'Sullivan observes that an approximation to $A(\lambda)$ is obtained at the final step of the iterative procedure described above as the A matrix for the final linear problem of (3.2).

(2) Linear inequality constraints. Returning to the model

$$y_i = L_i f + \varepsilon_i, \quad i = 1,2,\ldots,n,$$

suppose it is known that f is in some closed convex set $C \subseteq H_R$. For example $C = \{f: f(t) \geq 0, t\varepsilon\Omega\}$, is a closed convex set in any reproducing kernel space if Ω is closed. C is closed convex since, in any r.k. space H_R, there exists $\delta_t \varepsilon H_R$ such that $\langle\delta_t, f\rangle_R = f(t)$, where $\langle\cdot,\cdot\rangle_R$ is the inner product in H_R. Then $C = \{f: \langle\delta_t, f\rangle \geq 0, t\varepsilon\Omega\}$ is the intersection of a family of (closed, convex) half spaces, and so is closed and convex. Letting $\phi_1,\ldots,\phi_M$ span the null space of $J(\cdot)$, if the n×M matrix T with iνth entry $L_i\phi_\nu$ is of rank M, then it can be shown that the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}(L_i f-y_i)^2 + \lambda J(f) \qquad (3.5)$$

in any closed convex C, exists and is unique. Frequently C can be well approximated by C_L, the intersection of L half spaces, e.g.

$$C_L = \{f: \langle\psi_\ell, f\rangle_R \geq \alpha_\ell,\ \ell = 1,2,\ldots,L\}.$$

For example $C = \{f: f(t) \geq 0, t\varepsilon[0,1]$ may be approximated by

$$C_L = \{f: f(\tfrac{\ell}{L}) \geq 0,\ \ell = \tfrac{1}{2}, \tfrac{3}{2}, \ldots, \tfrac{2L-1}{2}\ \}$$

for sufficiently large L. Then the computation of the minimizer of (3.5) subject to $f\varepsilon C_L$ can be reduced to the solution of a quadratic programming problem subject to a finite family of linear inequality constraints, see Kimeldorf and Wahba (1971), Wahba (1973, 1980a, 1982), Villalobos and Wahba (1983). Wegman (1982) and Wegman and Wright (1983) have also discussed the imposition of side conditions in spline smoothing.

A generalization of GCV for constrained problems, was proposed in Wahba (1980a) and is defined as follows:

$$V(\lambda) = \frac{\frac{1}{n}RSS(\lambda)}{(1-\frac{1}{n}\sum_{k=1}^{n} a_{kk}{*}(\lambda))^2} \qquad (3.6)$$

where

$$a_{kk}{*}(\lambda) = \frac{L_k f_\lambda(\underset{\sim}{y}+\underset{\sim}{\delta}_k)-L_k f_\lambda(\underset{\sim}{y})}{\delta_k} \qquad (3.7)$$

Here, $\delta_k = L_k f_\lambda^{[k]}-y_k$, where $f_\lambda^{[k]}$ is the minimizer of (3.5) subject to $f\varepsilon C_L$ with the kth data point omitted, $\underset{\sim}{\delta}_k = (0,\ldots,0,\delta_k,0,\ldots,0)$ (δ_k in the kth position), $f_\lambda(\underset{\sim}{y}+\underset{\sim}{\delta})$ is the minimizer of (3.5) subject to $f \varepsilon C_L$ with the data $\underset{\sim}{y} + \underset{\sim}{\delta}$, and $f_\lambda(\underset{\sim}{y}) = f_\lambda$. $a_{kk}{*}(\lambda)$ is thus a divided difference of $L_k f_\lambda$ considered as a function of y_k. This definition is motivated by the identity below for the ordinary "leaving out one" function $V_0(\lambda)$:

$$V_0(\lambda) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}(L_k f_\lambda^{[k]} - y_k)^2 \equiv \frac{1}{n}\sum_{i=1}^{n}(L_k f_\lambda - y_k)^2/(1-a_{kk}{*}(\lambda))^2, \tag{3.8}$$

which becomes (3.6) in the generalized version of cross validation. For a proof of (3.8) see Craven and Wahba (1979). $V(\lambda)$ is generally preferable to $V_0(\lambda)$ in spline smoothing problems because of its theoretical properties as well as its relative computational ease.

The denominator $(1-\frac{1}{n}\sum_{i=1}^{n} a_{kk}{*}(\lambda))$ is fairly expensive to compute. However, a frequently reasonable approximation to $a_{kk}{*}(\lambda)$ is

$$a_{kk}{*}(\lambda) \simeq \frac{\partial L_k f_\lambda}{\partial y_k} . \tag{3.9}$$

Here (as with nonlinear problems) one may obtain V easily by observing that, if f_λ is the minimizer of (3.5) subject to $\langle\psi_\ell, f\rangle \geq \alpha_\ell$, and the constraints involving $\psi_{\ell 1}, \ldots, \psi_{\ell_{\max}}$ are active, then f_λ is also the minimizer of (3.5) subject to the equality constraints $\langle\phi_\ell, f\rangle = \alpha_\ell$, $\ell = \ell_1, \ldots, \ell_{\max}$. Now the problem: Minimize (3.5) subject to $\langle\psi_\ell, f\rangle = \alpha_\ell$, $\ell = \ell_1, \ldots, \ell_{\max}$ is linear in the data, and the $A(\lambda)$ matrix of this problem has for its diagonal entries the right hand side of (3.9). See Villalobos (1983), Wahba (1982). Here $TrA(\lambda)$ and hence $V(\lambda)$ are not continuous functions of λ (jumps may occur at critical points where new constraints become active) but this appears not to be a serious problem in most of the examples tried. We shall return to spline smoothing with linear inequality constraints in Section 4.

(3) Robust smoothing. If, for example, the measurement errors are contaminated Gaussian, the residual sum of squares $\sum_{i=1}^{n}(L_i f-y_i)^2$ may be replaced by a sum of robust functionals $\sum_{i=1}^{n}\rho(L_i f-y_i)$. This approach has been suggested by Huber (1979), Lenth (1977), and Utreras (1981). Various combinations of the setup in (1), (2) and (3) are possible.

(4) Penalized likelihood methods.

(i) Density Estimation. A number of authors have discussed maximum penalized likelihood estimates for a density. Let $X_1,\ldots,X_n$ be a random sample from a density f, then one may estimate f as the minimizer of

$$-\log \prod_{i=1}^{n} f(X_i) + \lambda J(f).$$

See Tapia and Thompson and references cited there. Silverman (1982) has proposed estimating g = log f as the minimizer of

$$\sum_{i=1}^{n} g(X_i) + \lambda J(g)$$

subject to $\int_0^\infty e^{g(t)}dt = 1$. These estimates will be splines if the J's of the preceding sections are used.

(ii) Penalized GLIM methods. O'Sullivan (1983) has suggested generalizing the GLIM method of Nelder and Wedderburn (1972) to penalized GLIM methods. In the GLIM method one has independent observations $y_1,\ldots,y_n$ where each observation y_i has a one-parameter exponential density Φ_i of the form

$$\Phi_i(y_i) = \exp\{[y_i\theta_i - b(\theta_i)]/a_i(\phi) + c(y_i,\phi)\} \tag{3.10}$$

for suitable choices of a_i, b and c. The mean and variance of y_i can be expressed in terms of θ_i and ϕ as

$$E(y_i) = \mu_i = b'(\theta_i),\ Var(y_i) = b''(\theta_i)a_i(\phi).$$

The model specification is completed by supplying a linearizing transformation $g(\mu_i)$ of the mean, $g(\mu_i) = L_i f$, and g relates μ_i to the function f it is desired to estimate. O'Sullivan suggests a maximum penalized likelihood estimate of f as the minimizer, in an appropriate function space of

$$I_\lambda(f) = -\sum_{i=1}^{n} \log\Phi_i(y_i) + \lambda J(f) \tag{3.11}$$

where $\Phi_i(y_i) = \Phi_i(y_i,f)$, and he discusses numerical algorithms.

Since the normal distribution has a density of the form (3.10), the spline estimates of the preceeding chapters are a special case, with $g(\mu_i) = \mu_i = L_i f$, $\log\Phi_i(y_i,f) = \text{const.}(y_i-L_i f)^2$. Two other interesting cases are the Poisson distribution and the Binomial distribution.

Let

$$y_i \sim P\ (\mu_i)$$

where P (μ) is the Poisson distribution with mean μ. Let $g(\mu_i) = \log\ \mu_i\ = L_i f$. By using a penalized GLIM model one is assuming that f is "smooth"-if $L_i f = f(t_i)$ and $\mu_i = \log \mu(t_i)$ this is the same as assuming that $\log \mu(t)$ is a smooth function of t. We have

$$\Phi_i(y_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}\ , \quad \log \mu_i = L_i f$$

$$\log\Phi_i(y_i,f) = y_i L_i f - e^{L_i f} + \log(y_i!)$$

and the maximum penalized likelihood estimate of f is obtained by minimizing

$$I_\lambda(f) = \sum_{i=1}^{n} (y_i L_i f - e^{L_i f} + \log(y_i!)) + \lambda J(f) \tag{3.12}$$

over a suitable space of functions.

In the binomial case, let

$$y_i \sim B\,(1,p_i)$$

and suppose one is interested in estimating the logit, $\log[p_i/(1-p_i)]$. Thus we may choose the link function $g(p_i) = \log[p_i/(1-p_i)] = L_i f$. Then

$$\Phi_i(y_i) = p_i^{y_i}(1-p_i)^{1-y_i}$$

$$\begin{aligned}\log\Phi_i(y_i) &= y_i \log p_i + (1-y_i)\log(1-p_i)\\ &= y_i \log\frac{p_i}{1-p_i} + \log(1-p_i)\end{aligned}$$

and

$$\log\Phi_i(y_i,f) = y_i L_i f - \log[1+e^{L_i f}]$$

and the maximum penalized likelihood estimate of f is the minimizer of

$$I_\lambda(f) = -\sum_{i=1}^{n} \{y_i L_i f - \log[1+e^{L_i f}]\} + \lambda J(f). \qquad (3.13)$$

This approach provides a nonparametric form of logistic regression.

This approach can also be used to provide an estimate of the log likelihood ratio in the classification problem. Suppose one is going to have available n/2 observations $X_{11},\ldots,X_{1,n/2}$ from some density h_1 and n/2 observations $X_{21},\ldots,X_{2,n/2}$ from some other density h_2. From this data one wishes to estimate the log likelihood ratio $f(t) = \log[h_1(t)/h_2(t)]$. Relabeling the n observations as $t_1,\ldots,t_n$, one may define a random variable $y_i = y(t_i)$ which takes the value 1 if t_i was from population 1 and 0 if t_i was from population 2. Conditional on there being an observation at t_i, the posterior probability that it is from population 1 is $p_i = h_1(t_i)/[(h_1(t_i) + h_2(t_i)]$, and we may treat y_i as though

$$y_i \sim B\,(1,p_i).$$

Since $f(t_i) = \log(h_1(t_i)/h_2(t_i)) = \log[p_i/(1-p_i)]$, we may estimate f by minimizing (3.13) with $L_i f = f(t_i)$.

This estimate for the likelihood ratio was proposed by Silverman (1978), see also Raynor and Bates (1983), Villalobos (1983), Villalobos and Wahba (1983).

O'Sullivan has proposed a further generalization of GCV for estimating λ in the penalized GLIM setup. Letting

$$Q(y,f) = \text{const} \sum_{i=1}^{n} \log\Phi_i(y_i,f)$$

(which becomes $\frac{1}{n}\sum_{i=1}^{n}(y_i-L_i f)^2$ in the normal errors case) O'Sullivan's proposal for estimating λ in (3.11) reduces to the minimization of

$$V_{GLIM}(\lambda) = \frac{Q(y,f_\lambda)}{(\frac{1}{n}Tr(I-A(\lambda))^2} \qquad (3.14)$$

where here the entries $a_{ij}(\lambda)$ of $A(\lambda)$ are approximations to $\frac{\partial L_i f_\lambda}{\partial y_j}\Big|_y$ which are available as the A matrix at the last step of an iterative algorithm similar to the procedure for non-linear functionals. He argues that the minimizer of $V(\lambda)$ of (3.14) in the binomial case is an estimate of the weighted predictive mean square error $R_{GLIM}(\lambda) = \sum_{i=1}^{n} p_i(1-p_i)(f(t_i)-f_\lambda(t_i))^2$.

Silverman (1982) has made the elegant observation that if $\lambda = \infty$ and $m = 3$, since the null space of J is the span of the polynomials of degree 2 or less, then the estimate of $f = \log(h_1/h_2)$ will in fact correspond to h_1, h_2, normal densities - this holds so long as $2m-d>0$, i.e. $d = 1,2,\ldots,5$.

4. Constrained TPSS estimates for posterior probabilities in the classification problem

Simultaneous with some of the preceeding work on the estimation of the log likelihood ratio $f = \log(h_1/h_2)$ via a penalized GLIM method, a study was made of the use of bivariate constrained thin plate smoothing splines for the estimation of the posterior probability $p = h_1/(h_1+h_2)$. See Villalobos (1983), Villalobos and Wahba (1983). $p(t)$ is the posterior probability that an observation t came from population 1 if each population is a priori equally likely. (If the prior probability of population 1 is q then the posterior probability $p_q(t)$ is $qp(t)/[(qp(t)+(1-q)(1-p(t))]$. These posterior probabilities are frequently of direct interest, for example in estimating the probability of a heart attach, given x_1 = blood pressure and x_2 = cholesterol level.

The estimate p_λ of p under study is the minimizer of

$$\frac{1}{n}\sum(p(t_i)-y_i)^2 + \lambda J(p) \qquad (4.1)$$

subject to

$$0 \le p(s) \le 1, \quad s\varepsilon\tau$$

where the y_i's are 1's and 0's, according to whether the observation at t_i was from population 1 or not. Although the motivation for this estimate is heuristic, the numerical results were excellent and demonstrate the feasibility and effectiveness of bivariate constrained smoothing splines, even with non-Gaussian data. (A constrained penalized

GLIM method for p would consist in minimizing

$$\sum_{i=1}^{n} \{y_i \log(p(t_i)/(1-p(t_i)) + \log(1-p(t_i))\} + \lambda J(p)$$

subject to $0 \leq p(s) \leq 1$, a harder numerical problem than the one under study.)

In general, to compute p_λ, one will first normalize the x_1 and x_2 scales so that the data falls within a square and the region of interest for estimating p_λ is the same square. In the work that follows, the region of interest was discretized by a 15×15 = 225 array of equally spaced points $s_1,\ldots,s_{225}$ and the constraints discretized as

$$0 \leq p(s_\ell) \leq 1, \ \ell = 1,2,\ldots,225 \ . \tag{4.2}$$

The problem of minimizing (4.1) subject to (4.2) can then be reduced to a finite dimensional quadratic programming problem subject to the inequality constraints (4.2). (See Villalobos and Wahba (1983)). A quadratic programming algorithm due to Gill, Gould, Murray, Sanders and Wright (1982) (GGMSW) was used to solve the quadratic program. In the present problem, the unconstrained problem is first solved and the (unconstrained) cross validation estimate of λ found. A substantial number of the s_ℓ's can be eliminated as identifying possibly active constrains when the unconstrained solution at s_ℓ is sufficiently inside the interior of [0,1]. Of the remaining possible constaints a good starting guess for those which will be active can also be made. If the possible set and the starting guess set are chosen reasonably well, the GGMSW algorithm converges rapidly. This is done for the unconstrained cross validation estimate $\hat{\lambda}$ of λ and then λ is changed slightly until the minimizer of the constrained cross validation function is found. The active constraints for the most recent value of λ are used as the starting guess for the new value of λ. See Villalobos (1983) for more details.

The figures below were obtained by M. Villalobos. Figure 1 gives a plot of the test example $p(x_1,x_2) = h_1(x_1,x_2)/[(h_1(x_1,x_2)+h_2(x_1,x_2)]$, where

$$h_1 \sim N(\begin{smallmatrix}0\\0\end{smallmatrix}, (\begin{smallmatrix}1&0\\0&1\end{smallmatrix}))$$

$$h_2 \sim \tfrac{1}{2}N(\begin{smallmatrix}1\\-2.5\end{smallmatrix},(\begin{smallmatrix}1&0\\0&1\end{smallmatrix})) + \tfrac{1}{2}N(\begin{smallmatrix}1\\+2.5\end{smallmatrix},(\begin{smallmatrix}1&0\\0&1\end{smallmatrix})).$$

Figure 2 gives a plot of a pseudo random sample of $n/2 = 70$ observations from h_1 (crosses) and 70 observations from h_2 (circles). Figure 3 presents the estimate $p_{\hat{\lambda}}(x_1,x_2)$. Figure 4 presents a plot of the level curves of $p_{\hat{\lambda}}$ superimposed on the pseudo data. Figure 5 presents a plot of the level curves of $p_{\hat{\lambda}}$ along with those of the true p (corresponding to Figure 1).

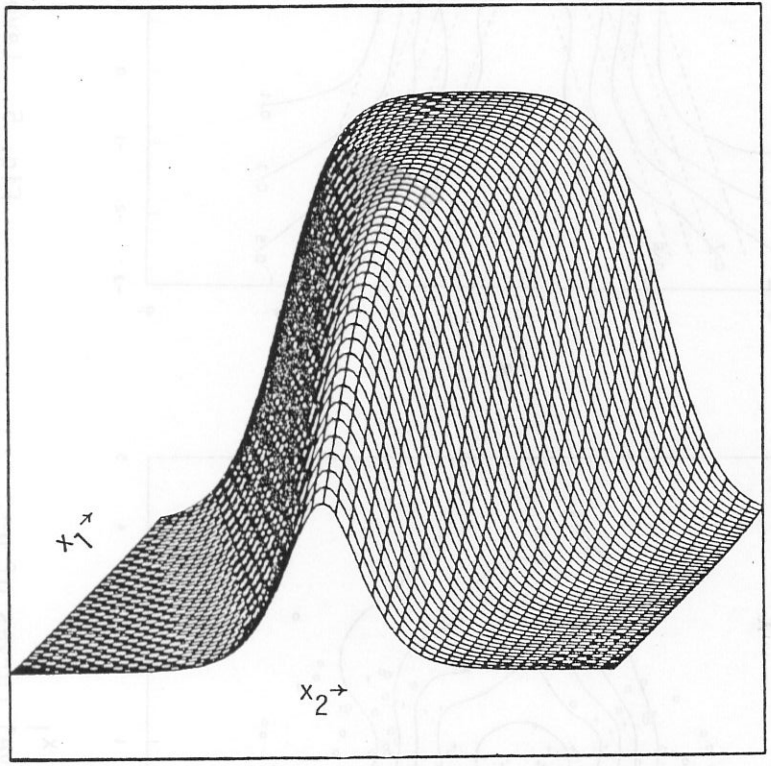

Fig. 1. The true $p(x_1,x_2)$.

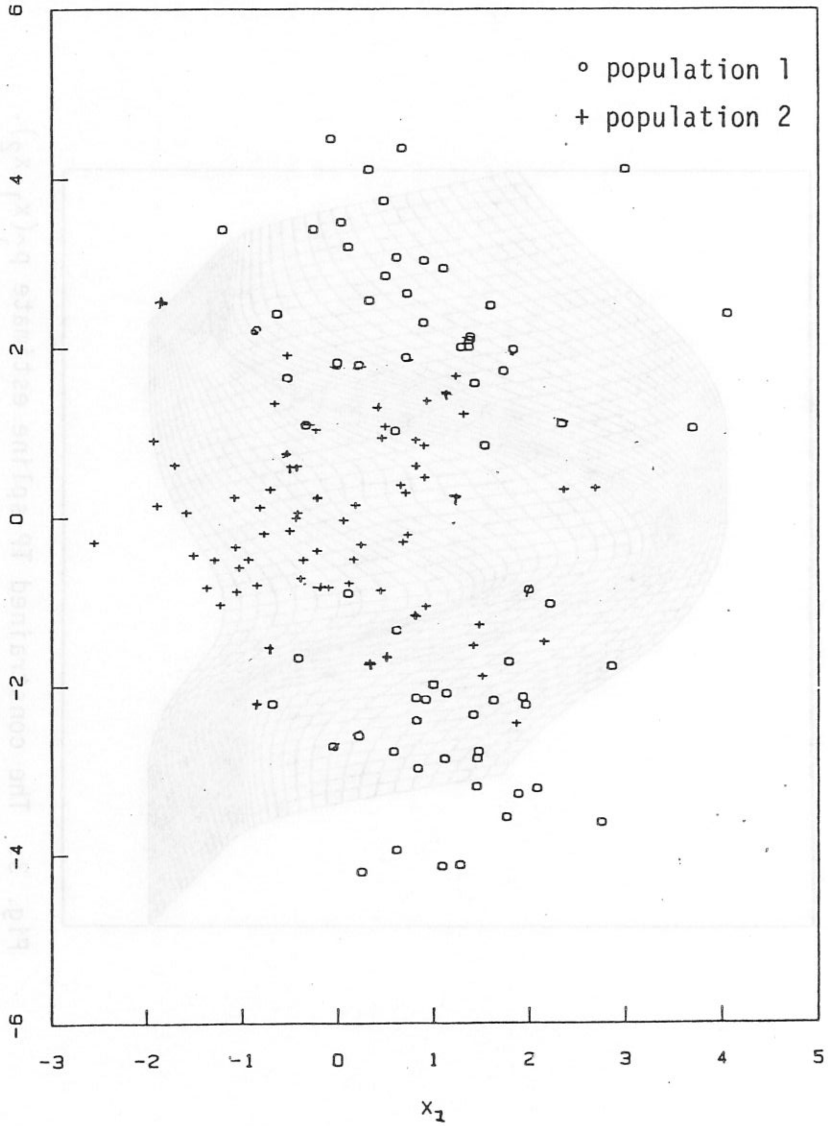

Fig. 2. The pseudo data.

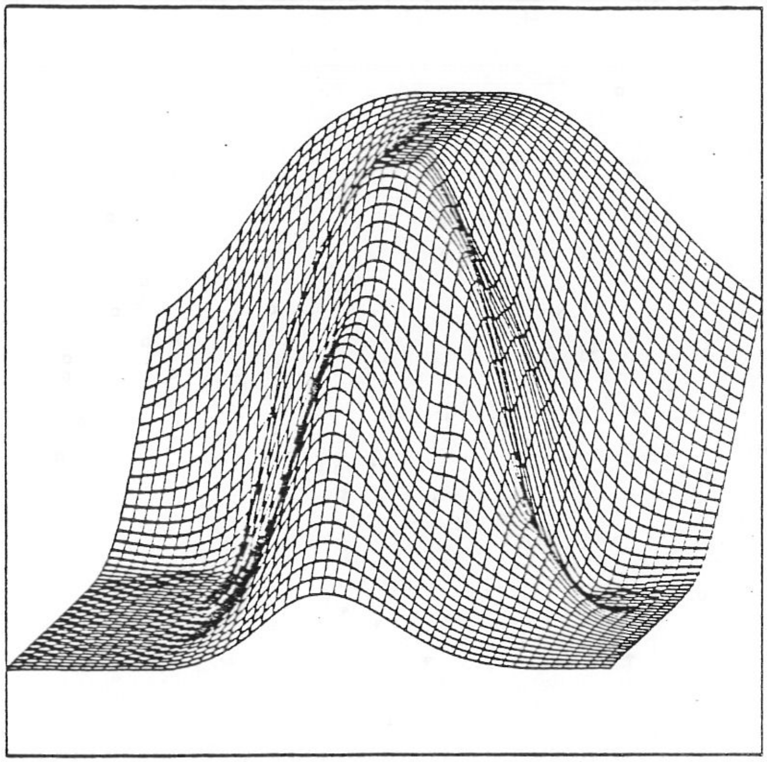

Fig. 3. The constrained TP spline estimate $p_{\hat{\lambda}}(x_1,x_2)$.

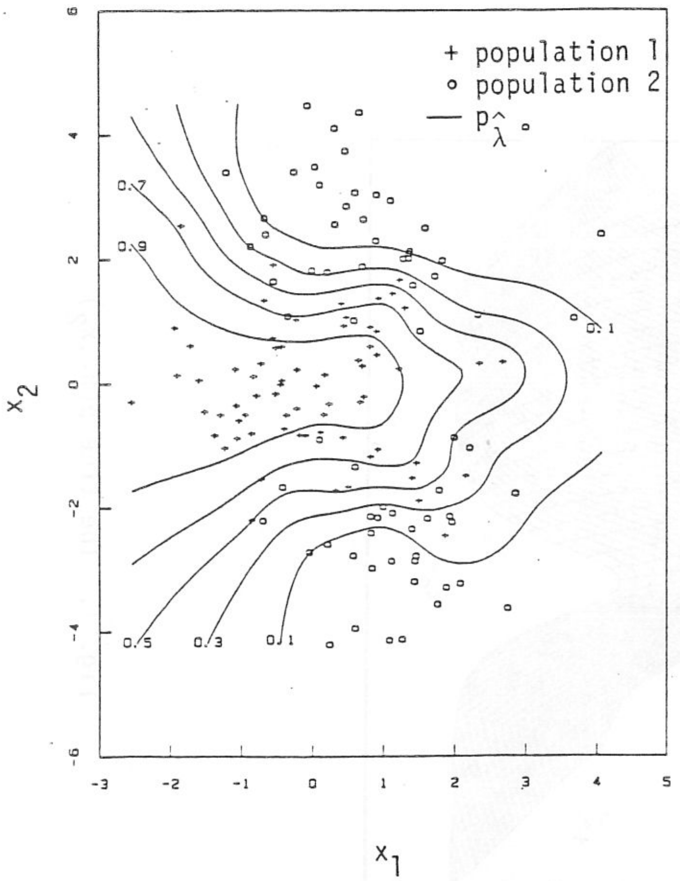

Fig. 4. Level curves of $p_{\hat{\lambda}}$,
and the pseudo data.

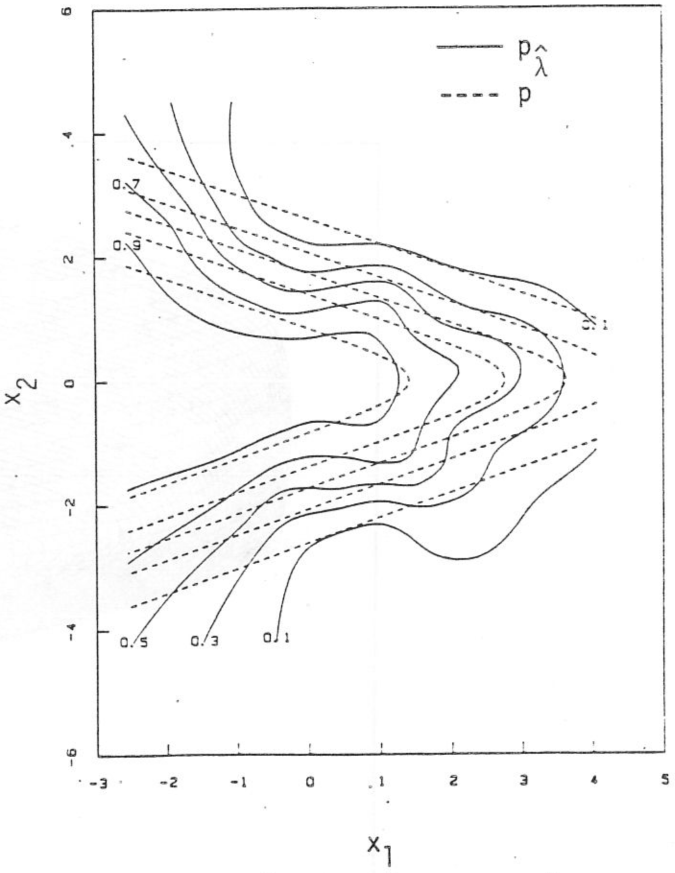

Fig. 5. Level curves of $p_{\hat{\lambda}}$
and the true p.

5. Partial TP spline models in many dimensions

Thin plate splines have been successfully used in three dimensions, and with the advent of more powerful computers and the availability of larger data sets, will probably be useful in somewhat higher dimensions. But frequently one has a function f of interest which depends on more than three or four variables. One way of combining parametric models and TP spline models for use in higher dimensions to get "semiparametric" models is as follows.

Let $f(x_1,\dots,x_k,x_{k+1},\dots,x_d)$ be modelled as follows: For fixed $x_{k+1},\dots,x_d$, f is a polynomial of degree m-1 or less in $x_1,\dots,x_k$, and, for any fixed $x_1,\dots,x_k$, f considered as a function of the d-k variables $x_{k+1},\dots,x_d$ is in the Hilbert space X of real valued functions of d-k variables all of whose derivatives of total order m are square integrable, that is,

$$J(f;x_1,\dots,x_k) = \sum_{\sum_{i=1}^{d-k}\alpha_i=m} \frac{m!}{\alpha_1!\dots\alpha_{d-k}!}\int\dots\int\left[\frac{\partial^m f(x_1,\dots,x_k;x_{k+1}\dots x_d)}{\partial x_{k+1}^{\alpha_1}\dots\partial x_d^{\alpha_{d-k}}}\right]^2 dx_{k+1}\dots dx_d<\infty$$

for each $x_1,\dots,x_k$. (We assume $2m-(d-k)>0$).

We may construct a Hilbertspace $\mathcal{H}$ of functions of d variables with the desired properties and $J(f;0,0,\dots,0)$ as seminorm, as follows.

Let

$$\mathcal{H} = \mathcal{H}_\pi \oplus X$$

where $\mathcal{X}$ is a Hilbert space of real valued functions of the d-k variables $x_{k+1},\ldots,x_d$ all of whose derivatives of total order m are square integrable. ($\mathcal{X}$ is equivalent to the space X of functions of d-k variables of Section 2.)

We would like $\mathcal{H}$ to contain the $M_d = \binom{d+m-1}{d}$ polynomials of total degree less than m in the variables $x_1,\ldots,x_d$. The space $\mathcal{X}$ already contains those $\binom{d-k+m-1}{d-k}$ polynomials which do not contain any $x_1,\ldots,x_k$ as factors, so we will let $\mathcal{H}_\pi$ consist of those $M_0 = \binom{d+m-1}{d} - \binom{d-k+m-1}{d-k}$ polynomials of total degree less than m in the variables $x_1,\ldots,x_d$, which contain at least one $x_1,\ldots,x_k$ as a factor.

For example, if m = 3, d = 5 and k = 3, $\mathcal{H}_\pi$ consists of the M_0 functions

$$x_1,\ x_2,\ x_3,\ x_1^2,\ x_2^2,\ x_3^2,\ x_1x_2,\ x_2x_3,\ x_1x_3$$

$$x_1x_4,\ x_1x_5,\ x_2x_4,\ x_2x_5,\ x_3x_4,\ x_3x_5,$$

whereas $\mathcal{X}$ contains

$$1,\ x_4,\ x_5,\ x_4^2,\ x_5^2,\ x_4x_5.$$

Thus $\mathcal{H}$ consists of all functions of the form

$$f = f_\pi + f_{\mathcal{X}}$$

where f_π is a polynomial in $(x_1,\ldots,x_d)\varepsilon\mathcal{H}_\pi$, and $f_{\mathcal{X}} = f_{\mathcal{X}}(x_{k+1},\ldots,x_d)$ is an element of $\mathcal{X}$.

It is now trivial to check that $J(f;0,0...0)$ is a seminorm on H, with the M_d polynomials of total degree less than m in the variables $x_1,\ldots,x_d$ spanning its null space.

Now, let

$$y_i = L_i f + \varepsilon_i, \qquad i = 1,\ldots,n$$

where the L_i are bounded linear functionals on H. We can define the partial TPSS as the minimizer in H of

$$\frac{1}{n}\sum_{i=1}^{n}(L_i f - y_i)^2 + \lambda J(f;0,\ldots,0).$$

Geometrically, this problem is essentially the same as that considered in Section 2. There will be a unique solution provided the $M_d \times n$ matrix T with $i\nu$th entry $L_i\phi_\nu$ (ϕ_ν is the νth polynomial, of the M_d polynomials of total degree less than m) is of rank M_d, and the minimizer can be represented as a linear combination of the ϕ_ν's and the functions $\xi_i(x_{k+1},\ldots,x_d)$ of (2.7), where the equations for the coefficients and the influence matrix are analogous to (2.9, 2.10, 2.11). The numerical methods in Bates and Wahba (1983) can be used to compute f_λ.

If the behavior of f as a function of $x_1,\ldots,x_k$ really is polynomial-like, then this approach should provide a good estimate for f with fewer data points than a spline estimate in all directions. (Other parametric constructions for H_π are possible, for example, fewer or more polynomials.) Engle et al. (1983) have recently carried out this construction for the case d - k = 1, with first degree polynomials in the non-splined variables.

In exploratory data analysis it would be of interest to know how to separate the variables $x_1,\ldots,x_d$ into two subsets, so that f will be modelled by polynomials on one subset and a TPS on the other. More generally, let Γ be a $d \times d$ orthogonal matrix, and let $(\tilde{x}_1,\ldots,\tilde{x}_d)' = \Gamma(x_1,\ldots,x_d)$, and model f as a polynomial in $(\tilde{x}_1,\ldots,\tilde{x}_k)$ and as a TPS in $(\tilde{x}_{k+1},\ldots,\tilde{x}_d)$. One may then ask the question: How can Γ be chosen? If the goal is model building then some predictive mean square error criteria is appropriate. In principle the GCV function can be computed as a function of Γ, however, it is an open question at this time whether the use of GCV for that purpose is a good procedure, particularly if λ is chosen by GCV. Analogous to an idea of Huber (1983) one may ask: What are the interesting directions? One might then define the interesting directions as those for which the dependency is not well modelled by a polynomial of degree less than m. In this case one has found the Γ producing the "interesting" directions if $J(f_{\hat{\lambda}})$ using $(\tilde{x}_1,\ldots,\tilde{x}_d)' = \Gamma(x_1,\ldots,x_d)$ is larger than $J(f_{\hat{\lambda}})$ obtained with any other orthogonal transformation. These "interesting" directions are not necessarily the same as those that would be selected by a predictive mean square error criteria but might be the ones one chooses to examine visually. For fixed Γ, $J(f_\lambda)$ is a monotone decreasing function of λ. In some sense, λ measures "deviation from polynomialness". Thus, it may be useful to look at $\hat{\lambda}$ as a function of Γ, or at an estimate $\hat{b}$ of the parameter b which appeared in Section 2, as a function of Γ. (One such estimate is $\hat{b} = J(f_{\hat{\lambda}})/TrA(\hat{\lambda})$.) Numerous questions remain as to the choice of an "optimality" or "projection pursuit" criteria for Γ, and, how to find the optimal Γ numerically,

REFERENCES

Adams, R. (1975). *Sobolev Spaces*. New York: Academic Press.

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Am. Math. Soc.*, *68*, 337-404.

Bates, D. and Wahba, G. (1983). *A truncated singular value decomposition and other methods for generalized cross-validation.* University of Wisconsin-Madison Statistics Dept. TR 715, submitted.

Chow, Y., Geman, S., and Wu, L. (1983). Consistent crossvalidated density estimation. *Ann. Math. Statist.*, *11*, 25-38.

Cox, D. (1982). "Convergence rates for multivariate smoothing spline functions." Mathematics Research Center TSR 2437, University of Wisconsin-Madison.

Cox, D. (1983). Asymptotics for M type smoothing splines. *Ann. Math. Statist.*, *11*, 530-551.

Craven, P. and 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*, New York: Springer-Verlag.

Duchon, J. (1975). "Fonctions splines et vecteurs aleatoires." Seminaire d'Analyse Numerique No. 213, Grenoble.

Duchon, J.(1976). Splines minimizing rotation-invariant semi-norms in Sobolev Spaces, pp. 85-100. In *Constructive Theory of Functions of Several Variables*, K. Zeller (ed.), Springer.

Dyn, N. and Levin, D. (1981). A procedure for generating diagonal dominance in ill conditioned systems originating from integral equations and surface interpolation. *School of Mathematical Sciences, Tel Aviv University TR 81-18*.

Engle, R., Granger, C., Rice, J., and Weiss, A.(1983). "Nonparametric estimates of the relation between weather and electricity demand." Discussion paper 83-17, Dept. of Economics, University of California, San Diego.

Fritz, S., Wark, D., Fleming, J., Smith, W., Jacobowitz, H., Hilleary, D., and Alishouse, J. (1972). "Temperature sounding from satellites." NOAA Tech. Rept. NESS 59, National Oceanic and Atmospheric Administration, Washington, D. C.

Gill, P., Gould, N., Murray, W., Saunders, M., and Wright, M. (1982). "Range-space methods for convex quadratic programming." Systems Optimization Laboratory TR SOL 82-14, Department of Operations Research, Stanford University.

Hajek, J. (1962). On linear statistical problems in stochastic processes. *Czech. Math. J.*, *87*, 404-444.

Huber, P. (1979). Robust smoothing. In *Robustness in Statistics*, G. Wilkinson (ed.), Academic Press.

Huber, P. (1983). "Projection pursuit." Mathematical Sciences Research Institute MSRI 009-83, Berkeley, CA.

Hutchinson, M., T.Booth, J. McMahon and Nix, M., (1983). Estimating monthly mean values of daily total solar radiation for Australia, to appear, Solar Energy.

IMSL (International Mathematical and Statistical Library (1981), Subroutine ICSSCV. Houston, TX.

Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. J. Math. Anal. Applic., 33, 82-95.

Lenth, R., (1977). Robust splines, manuscript.

Li, K. (1983a). From Stein's unbiassed risk estimates to the method of generalized cross validation, Purdue University, Department of Statistics TR 83-34.

Li, K. (1983b). Cross-validation in non-parametric regression. *Bull. I.M.S.*, *12*, 148.

Matheron, G. (1973). The intrinsic random functions and their applications. *Adv. Appl. Prob.*, *5*, 439-468.

Meinguet, J. (1979). Multivariate interpolation at arbitrary points made simple. *J. App. Math. Phys.* (*ZAMP*), *30*, 292-304.

Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *J. Roy. Stat. Soc. A*, *135*, 370-384.

O'Sullivan, F. (1983). *The analysis of some penalized likelihood estimation schemes*. Department of Statistics, University of Wisconsin-Madison, thesis.

Parzen, E. (1962). An approach to time series analysis. *Ann. Math. Statist.*, *32*, 951-989.

Ragozin, D. (1981). "Error bounds for derivative estimates based on spline smoothing of exact or noisy data." Department of Mathematics, University of Washington GN-50.

Raynor, W. and Bates, D., (1983). Smoothing generalized linear models with applications to logistic regression, manuscript.

Rice, J. and Rosenblatt, M. (1981). Integrated mean squared error of a smoothing spline. *J. Approx. Thy.*, *33*, 353-369.

Rice, J. and Rosenblatt, M. (1983). Smoothing splines: regression, derivatives and deconvolution. *Ann. Math. Statist.*, *11*, 141-156.

Schwartz, L. (1966). *Theorie des Distributions*. Paris: Hermann.

Silverman, B. (1978). Density ratios, empirical likelihood and cot death. *J. Roy. Stat. Soc. C*, *27*, 26-33.

Silverman, B. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.*, *10*, 795-810.

Speckman, P. (1982). "Efficient nonparametric regression with cross-validated smoothing splines." Dept. of Statistics, University of Missouri-Columbia, manuscript.

Tapia, R. and Thompson, J. (1978). *Nonparametric Probability Density Estimation*. Baltimore: Johns Hopkins University Press.

Utreras, F. (1978). "Quelques resultats d'optimalite pour la methode de validation croissee." Seminaire d'Analyse Nuumerique, 301, Grenoble.

Utreras, F. (1980). Sur le choix du parametre d'adjustement dans le lissage par fonctions spline. *Numer. Math.*, *34*, 15-28.

Utreras, F. (1981a). On computing robust splines and applications. *SIAM J. Sci. Stat.Comput.*, *2*, 153-163.

Utreras, F. (1981b). Optimal smoothing of noisy data using spline functions. SIAM J. Sci. Stat. Comput., 2, 349-362.

Villalobos, M. (1983). *Estimation of posterior probabilities using multivariate smoothing splines and generalized cross-validation*. Department of Statistics, University of Wisconsin-Madison, thesis.

Villalobos, M. and Wahba, G. (1983). Multivariate thin plate spline estimates for the posterior probabilities in the classification problem. *Commun. Statist.*; *12*, 1449-1480.

Wahba, G. (1973). On the minimization of a quadratic functional subject to a continuous family of linear inequality constraints. *SIAM J. Control*, *11*, 64-79.

Wahba, G. (1975). Smoothing noisy data by spline functions. *Numer. Math.*, *24*, 383-393.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J.R. Statist. Soc. B*, *40*, 364-372.

Wahba, G. (1979). Convergence rates of "Thin Plate" smoothing splines when the data are noisy. pp. 232-246. In *Smoothing Techniques for Curve Estimation*, T. Gasser and M. Rosenblatt (ed.), Springer-Verlag.

Wahba, G.(1980a). "Ill posed problems: Numerical and statistical methods for mildly, moderately, and severely ill posed problems with noisy data." University of Wisconsin-Madison Statistics Department Technical Report No. 595. (To appear in the Proceedings of the International Conference on Ill Posed Problems, M.Z. Nashed, ed.

Wahba, G.(1980b). Spline bases, regularization, and generalized cross validation for solving approximation problems with large quantities of noisy data. pp. 905-912. In *Approximation Theory III*, W. Cheney (ed.), Academic Press.

Wahba, G. and Wendelberger, J. (1980). Some new mathematical methods for variational objective analysis using splines and cross-validation. *Monthly Weather Review*, *108*, 36-57.

Wahba, G. (1982). Constrained regularization for ill posed linear operator equations, with applications in meteorology and medicine. pp. 383-418. In *Statistical Decision Theory and Related Topics III*, *Vol. 2*, S.S. Gupta and J.O. Berger (eds.), Academic Press.

Wahba, G. (1983a). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Stat. Soc. B.*, *45*, 133-150.

Wahba, G. (1983b). "A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem." University of Wisconsin-Madison Statistics Dept. TR No. 712.

Wahba, G., (1984). Surface fitting with scattered, noisy data on Euclidean d-spaces and on the sphere, to appear, *Rocky Mountain J. Math.*, 14, 1, 281-299.

Wegman, E. (1982). Optimal nonparametric function estimation. In *Proceedings of the Conference on Nonparametrics*, Z. Govindarajulu (ed.), Lexington: University of Kentucky.

Wegman, E. and Wright, I. (1983). Splines in statistics. *J.A.S.A.*, *78*, 351-366.

Wendelberger, J.(1981). "The computation of Laplacian smoothing splines with examples." University of Wisconsin-Madison TR No. 648.

Wendelberger, J. (1982). *Smoothing noisy data with multidimensional splines and generalized crossvalidation*. University of Wisconsin-Madison Statistics Dept. PhD. Thesis.

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report No. 722	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) CROSS VALIDATED SPLINE METODS FOR THE ESTIMATION OF MULTIVARIATE FUNCTIONS FROM DATA ON FUNCTIONALS		5. TYPE OF REPORT & PERIOD COVERED Scientific Interim 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Grace Wahba		8. CONTRACT OR GRANT NUMBER(s) ONR N00014-77-C-0675
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics, University of Wisconsin 1210 W. Dayton St. Madison, WI 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research 800 N. Quincy St. Arlington, VA		12. REPORT DATE September 1983 13. NUMBER OF PAGES
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

cross validated splines; thin plate splinesl inequality constrained smoothing splines

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The relation between spline estimation, ridge regression, and Bayes estimation is reviewed. Then a description of multivariate thin plate (TP) smoothing splines in two and several dimensions is given. Several generalizations and related methods are described, including spline estimation where the data are noisy observations on nonlinear functions; when families of linear inequalities, such as positivity, are known a priori; and when measurement error is contaminated Gaussian. Penalized log likelihood density estimation is seen to result in spline estimation. Some recent related work on penalized GLIM methods, which also result in spline estimates, is mentioned. Some numerical results using smoothing multivariate TP splines which are constrained to be between 0 and 1 are given. Finally we describe multivariate partial TP splines which are TP splines in some directions and low degree polynomials in others and are suitable for semiparametric modelling in 4 or more dimensions.