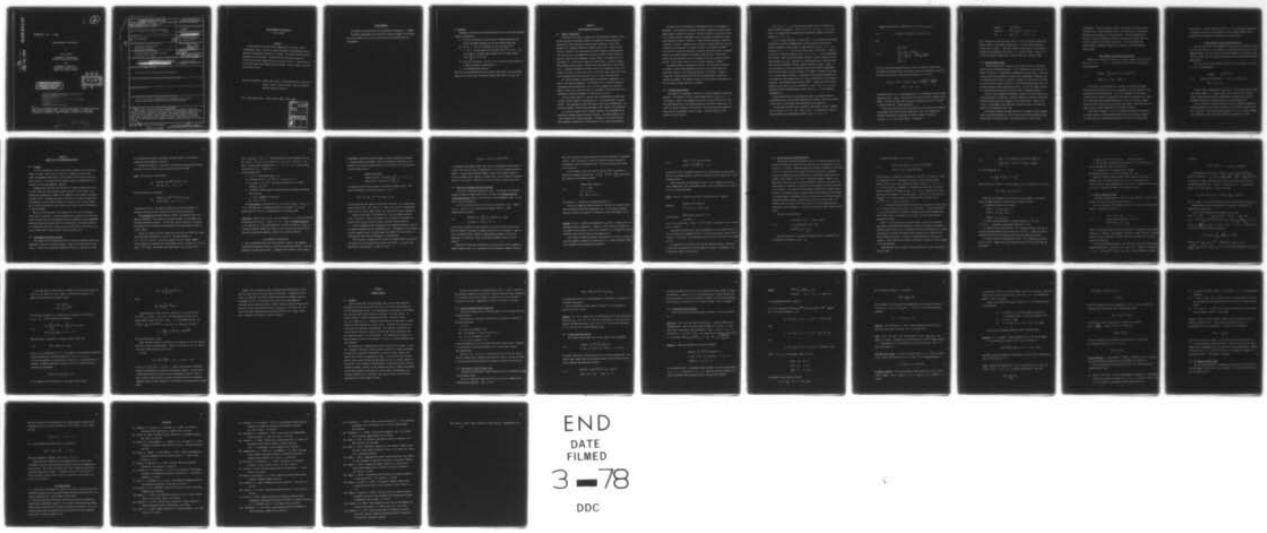
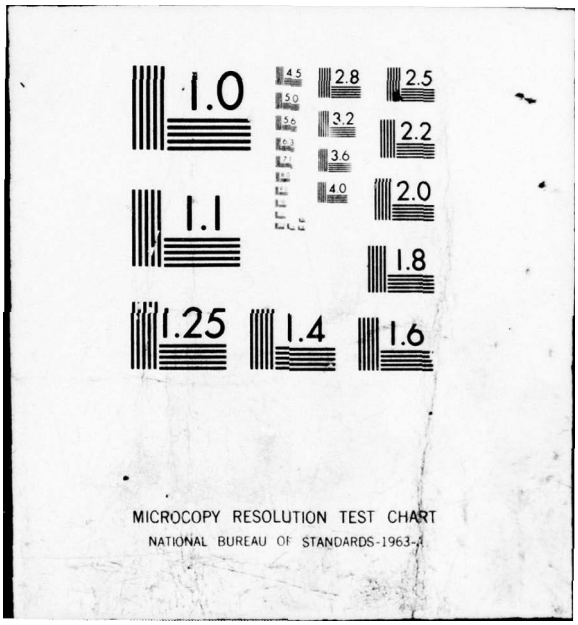
AD-A049 197 NORTH CAROLINA UNIV AT CHAPEL HILL DEPT OF STATISTICS F/G 12/1
SPLINE METHODS IN STATISTICS,(U)
1977 I W WRIGHT AFOSR-75-2840

UNCLASSIFIED AFOSR-TR-77-1307 NL

| OF |
AD
A049197



END
DATE
FILMED
3-78
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

(2)

AFOSR-TR- 77-1307

SPLINE METHODS IN STATISTICS*

by

Ian W. Wright

Department of Mathematics
Manchester University, England

and

Department of Statistics
University of North Carolina
Chapel Hill, North Carolina

*The work was supported in part by the Air Force Office of Scientific Research under Grant No. AFOSR-75-2840. The author is presently on leave from Department of Mathematics, Papua New Guinea, University of Technology.

(See 1473)

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: (18) AFOSR-TR-77-1307 (19)
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): (6) SPLINE METHODS IN STATISTICS,
5. TYPE OF REPORT & PERIOD COVERED: Interim
6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): (10) Ian W. Wright
8. CONTRACT OR GRANT NUMBER(s): (15) AFOSR-75-2840

9. PERFORMING ORGANIZATION NAME AND ADDRESS:
University of North Carolina
Department of Statistics
Chapel Hill, NC 27524
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: (16) 61102F (17) 2304/A5

11. CONTROLLING OFFICE NAME AND ADDRESS:
Air Force Office of Scientific Rsch/NM
Bolling AFB, DC 20332
12. REPORT DATE: (11) 1977
13. NUMBER OF PAGES: ~~45~~ (12) 47p.

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): deleted per AFOSR
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)

spline, cubic spline, interpolating spline, smoothing spline, isotone, isotonic spline, density estimation, spectral density estimation

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

Spline functions are particularly appropriate in fitting a smooth non-parametric model. The use of spline functions in non-parametric density estimation and spectral estimation is surveyed. The requisite spline theory background is also included. Isotonic splines offer great promise in filtering noise from a smooth model with an order structure. Some results on the existence and several applications of general isotonic splines are given in the final section.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

182850

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Spline Methods in Statistics

Ian W. Wright

Abstract

Spline functions are particularly appropriate in fitting a smooth non-parametric model. The use of spline functions in non-parametric density estimation and spectral estimation is surveyed. The requisite spline theory background is also included. Isotonic splines offer great promise in filtering noise from a smooth model with an order structure. Some results on the existence and several applications of general isotonic splines are given in the final section.

Key Words and Phrases: Spline, Cubic Spline, Interpolating Spline, Smoothing Spline, Isotone, Isotonic Spline, Density Estimation, Spectral Density Estimation.

A.M.S. Classification Nos.: 62G05, 62M15, 65D05, 65D10, 06A10

Acknowledgement

This paper was written under the directorship of Professor E. J. Wegman.

The author acknowledges the assistance of Professor Wegman in the conception of the project and at a key juncture and also for general help and encouragement.

§0 Notation

Throughout this thesis the following notation will be used without further comment or explanation.

L_2 - the set of functions on [0,1] which are Lebesgue measurable and square integrable equipped with the usual Banach space norm.

W_m - for $m \geq 1$. The set of functions on [0,1] for which $f^{(j)}$, $j = 0,1,\ldots,m-1$ are absolutely continuous and $f^{(m)}$ is in L_2. This will always be a Hilbert space equipped with the inner product $\langle f,g \rangle = \sum_{j=0}^{m} \int_0^1 f^{(j)} g^{(j)} dt$

C^k - for $k \geq 0$ or infinite. The set of all functions on [0,1] which are k times continuously differentiable.

D - the differentiation operator.

When we are considering functions with domains other than [0,1] the relevant domain will be shown after the function space symbol above, e.g. $W_m(-\infty,\infty)$.

CHAPTER 1

Spline Methods in Statistics

§1.1 General Introduction

Statistics (as we know it) began with fitting parametric models to data. Various principles for making estimates and inferences were developed and refined until their efficiencies reached their (asymptotic) limits with methods such as maximum likelihood, minimum variance, and likelihood ratio. Attention returned to the basic models and there appeared a growing realization that not all of the parametric structure was needed to make inferences. From this idea arose the techniques variously known as distribution free or non-parametric. At the cost of some loss of efficiency in certain instances, these methods prevented model violations from being reflected in false inferences. Although non-parametric methods have had some remarkable success (for example, the theory of rank tests) they all too often ignore useful non-statistical information that may be present, and as a result lose efficiency.

The classical method of incorporating non-statistical information is by means of the Bayesian framework. In the absence of any canonical methods of determining and assessing priors, this has to be regarded with suspicion. Happily, there are at least three ways to incorporate validly non-statistical knowledge in inference procedures. We discuss each in turn.

The first occurs when it is realized that the predominantly normal data contains a certain amount of contamination, i.e. the normal model is roughly correct. This knowledge may come from central limit type considerations. The robustness methods of Huber (1964) and Hampel (1974) exploit this knowledge.

If a system is known to have an order structure, this knowledge may be exploited by methods known as isotonic inference. The book by Barlow, et al (1972) reviews most of these techniques. Knowledge of order often follows from elementary consideration of the structure of the system being modelled.

The third sort of non-statistical knowledge we can use is knowledge of smoothness. The least action principle of dynamics suggests that nature makes things change as smoothly as possible. Our main concern in this thesis will be with a certain measure of smoothness and the consequences of making the underlying function as smooth as possible. A function which optimizes a smoothness criterion is a spline. Perhaps a little ironically spline methods also provide a sound route for Bayesians to re-enter the scene, as Kimeldorf and Wahba (1970 and 1971) show there often exist a Bayesian model which gives a smoothing spline as the posterior mean in that model, given the data. Models which require an order structure as well as smoothness lead us to consider isotonic splines. Some original results in this area are presented in Chapter 3.

The present account is organized as follows: The remainder of Chapter 1 is devoted to those purely mathematical parts of spline theory which have proved most relevant to the recent applications in statistics. Chapter 2 reviews the literature associated with the main applications of splines to statistical problems. The author's results on the existence and characterization of isotonic splines are given in Chapter 3.

§1.2 Classical Spline Theory

The spline is the engineer's solution to a problem frequently concerning engineers. The problem is to fit a curve through points (t_i,y_i) $i=1,2,\ldots,n$ in the plane. However, the engineer often needs to obtain values of the first and second derivatives of the underlying function from this fitted curve. The spline, an optimal solution of this problem, uses the analogy with weightless beams, part of the engineer's stock in trade. A precise account of the engineer's spline follows.

Let $\{(t_i,y_i)\colon i = 1,\ldots,n\}$ be the (error-free) points in the plane for which we seek an interpolant. Call $\Delta = \{\xi_1 (= t_1) < \xi_2 < \xi_3 < \ldots < \xi_N\ (= t_n)\}$ a *mesh*. For computational considerations the mesh will normally just be the numbers $\{t_i\colon i = 1,\ldots,n\}$.

A (cubic) spline with mesh Δ, written $S_\Delta(t)$, is a function with continuous derivatives up to (and including) order 2 which coincides exactly with a (possibly different) cubic function on each interval $[\xi_i,\xi_{i+1}]$ $i = 1,\ldots,N-1$. The points $\{\xi_i\colon i = 2,\ldots,N-1\}$ are called the *knots* of the spline. A spline $S_\Delta(t)$ which in addition satisfies $S_\Delta(t_i) = y_i$, $i = 1,2,\ldots,n$ is an interpolant for the data. Some further restriction is needed in order to make this interpolant unique on $[t_1,t_n]$. Although for certain applications, other end conditions are more convenient, the most natural condition is $S''_\Delta(t_1) = S''_\Delta(t_n) = 0$. This corresponds to giving the analogous beam cantilevered ends (protruding beyond the end points) and also minimizes the "energy of flexion" of the beam (i.e. the mean square curvature). The proofs of these various properties are established in a much general context by various contributors to the theory of L-splines. See Ahlberg, Nilson and Walsh (1967). Notice that if the spline between (t_i,y_i) and (t_{i+1},y_{i+1}) has equation $y = a_i t^3 + b_i t^2 + c_i t + d_i$, the continuity of the lower derivatives ensures that b_i, c_i, d_i are constants, independent of i.

Since curve fitting is a very practical task, it is against numerical considerations that a curve fitting method must be judged. The ease with which a cubic spline is fitted must have contributed to its popularity.

We now sketch the main steps in fitting a cubic spline to data. Suppose data is $\{(t_i,y_i)$, $i = 1,2,\ldots,n\}$. Let $h_i = t_{i+1} - t_i$ and take M_i = value of second derivative of interpolating spline S_Δ at t_i, $i = 1,2,\ldots,n$.

Suppose the polynomial interpolating (t_i, y_i) and (t_{i+1}, y_{i+1}) is

1.2.1 $$y = a_i(t-t_i)^3 + b_i(t-t_i)^2 + c_i(t-t_i) + d_i$$

then

1.2.2 $$\begin{cases} b_i = M_i/2 \\ a_i = (M_{i+1} - M_i)/6h_i \\ c_i = \dfrac{y_{i+1} - y_i}{h_i} - \dfrac{2(h_i M_i + h_i M_{i+1})}{6} \\ d_i = y_i \end{cases}$$

Thus our curve fitting problem reduces to that of finding the values of M_i. The equations relating the M_i are obtained by using the continuity of the first derivative of the spline, along with the relations 1.2.2 to give

$$h_{i-1} M_{i-1} + (2h_{i-1} + 2h_i)M_i + h_i M_{i+1} = 6\left(\frac{y_{i+1}-y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}}\right)$$

$$\text{for } i = 2,3,\ldots,n-1.$$

Our demand that $M_1 = M_n = 0$ leads immediately to a tridiagonal system of linear equations for $M_2,\ldots,M_{n-1}$. This system can be easily solved by Gaussian elimination. So easily in fact, that fitting an interpolating spline to 40 points is feasible with only a simple calculator.

It was soon realized that the cubic spline was the solution of an optimizing problem which could be easily generalized. Let L be a differential operator with constant coefficients of order m (usually D^m) and let $\{(t_i, y_i): i = 1,\ldots,n\}$ be data points. The problem

$$\begin{array}{lll} \text{minimize} & \int_{-\infty}^{\infty} (Lf)^2 \, dt & \\ \text{subject to} & f^{(j)} \in L_2(-\infty,\infty) & j = 0,1,\ldots,m \\ \text{and} & f(t_i) = y_i & i = 1,\ldots,n \end{array}$$

has a solution f(t) which satisfies L*Lf(t) = 0 in the intervals between knot points, where L* is the adjoint operator to L. We call such a solution an "*Interpolating L-Spline*". There are accounts of such splines in the books by Ahlberg, Nilson and Walsh (1967) and T. N. E. Greville (1969). However, for statistical purposes another type of spline turns out to be more useful.

§1.3 Smoothing Spline Theory

For most applications in statistics the smoothing spline is much more useful than the interpolating spline. This is because most real-life data is subject to error be it from sampling, measurement or other sources. There are two main spline fitting methods in common use corresponding to different ways of dealing with the "noise" in the data. Because this data does not constrain the fitted function nearly as firmly as in the interpolating spline case, the fitting requires a genuine optimization routine, not just the simple solution of a linear system of equations as with the cubic interpolating spline.

The first, more frequently used method, parallels the least squares curve fitting procedure by minimizing a criterion depending on squares of deviations from data points and on the "roughness" of the fitted curve. When we have little or no knowledge of the magnitude of possible errors in our data this method is the appropriate one to use.

On the other hand when the data points are, for example, direct readings from a calibrated instrument, we may be able to set fairly narrow 100% confidence limits for each data point. The second method is used in these

circumstances. For this we need to replace the ordinate y in the two dimensional data with a 100% confidence interval, and constrain the fitted spline function to pass through all of these intervals. This is accomplished in practice by using an optimization routine to minimize the (convex) roughness criterion, subject to the linear constraints. It will be noticed that method attaches considerable importance to outliers, rather than largely ignoring them.

First Method of Fitting Smoothing Splines

Suppose the t values of the data lie in a finite interval say [0,1] and we have $0 < t_1 < t_2 < \ldots < t_n < 1$. Fitting the spline leads us to solving the following problem.

$$\text{Minimize} \quad \sum_{i=1}^{n} (f(t_i)-y_i)^2 + \lambda \int_0^1 (f^{(m)})^2 \, dt \qquad 1.3.1$$

$$\text{Subject to} \quad f \in W_m, \quad \lambda \text{ fixed} > 0 \; .$$

The solution is given explicitly in the paper of Kimeldorf and Wahba (1970) and as expected turns out to be a polynomial spline of degree 2m-1 with possible knots at the data points. As so often happens, this theoretical solution cannot be used as an algorithm in any realistic practical case. When the t values are evenly spaced throughout [0,1], and f is periodic, Cogburn and Davis (1974) show how to do the fitting more easily. Apart from this one happy instance, a heavy optimization is invariably required.

Notice that the number $\lambda > 0$ in 1.3.1 controls the amount of smoothing; λ too small results in overfitting and insufficient removal of noise, whereas λ too large results in underfitting and removal of much of the wanted signal

with the noise. Clearly the correct choice of λ is of the greatest importance. A satisfactory solution to this problem is given by Wahba and Wold (1975), although not all theoretical consequences are yet developed.

Second Method of Fitting Smoothing Splines

From what was written in the preamble of this section, the reader will have seen that this spline is a cross between the interpolating spline and our first smoothing spline. For this reason the fitting technique is also known as (the solution procedure for) the Generalized Hermite-Birkhoff Interpolation Problem-(GHB problem).

Let $[\alpha_i,\beta_i]$ be the 100% confidence interval for the ordinate at t_i (with $\alpha_i < \beta_i$). The GHB problem is

$$\text{Minimize} \qquad \int_0^1 (f^{(m)})^2 \, dt$$

1.3.2 $$\text{subject to the constraints} \quad f \in W_m, \quad \alpha_i \leq f(t_i) \leq \beta_i$$

$$\text{for } i = 1,2,\ldots,n.$$

Various recent contributions to the theory of such splines have been made by M. Attéia (1968), P. J. Laurent (1969), and K. Ritter (1969). Because Hilbert space methods are directly applicable, this spline has been more thoroughly theoretically analyzed than the first smoothing spline.

For our results on isotonic splines we are able to adapt the methods of Attéia and Laurent to the isotonic situation, so our isotonic splines are of this second kind. For the record, the solution of 1.3.2 is a spline of degree 2m-1 with knots at those data points where the constraints are active.

§1.4 Some Simplifications

For most practical applications a slightly sub-optimum solution may be preferable to the optimum if it involves a great deal less computation effort. We have already seen that not all of the data points are knot points for the smoothing spline. By using the following guidelines the old statistical virtue of eyeballing the data can be converted to considerable computational advantage.

Knot Point Selection (Cubic smoothing spline).

1) Knot points should be at data points.

2) Try to have at least 4 or 5 data points between knots.

3) Have not more than one extremum and one inflexion point between knots.

4) Have extrema centered in intervals and inflexion point near knots.

For more details see Wold (1974).

The form of the optimal spline function f.

Suppose data points $\{t_{i_1}, t_{i_2}, \ldots, t_{i_J}\}$ have been chosen as knots. Then the optimal spline function will be f such that $f(t) = a_0 + a_1 t + a_2 t^2 + \sum_{j=1} d_j \cdot (t - t_{i_j})_+^3$ where $(t - c)_+^3 \begin{cases} = 0 & t < c \\ = (t-c)^3 & t \geq c \end{cases}$.

§1.5 Bayesian Estimation Again

We are now in a position to make our remarks about the equivalence of smoothness and Bayesian posterior means precise.

Let $L = \sum_{j=0}^{m} a_j D^j$ be a differential operator and let $B = [b_{jk}]$ be a positive definite matrix. Suppose $B^{-1} = [b^{jk}]$.

Problem I

Find $f \in W_m(-\infty,\infty)$ which minimizes

$$\sum_{j,k} (f(t_j) - y_j) b^{jk} (f(t_k) - y_k) + \int_{-\infty}^{\infty} (Lf)^2 \, dt$$

Problem II

Find $f(t)$ with $f(t) = E(x(t)|y(t_1), y(t_2),\ldots,y(t_n))$ where $y_j = x(t_j) + e_j$ with $e_j \sim N(0,B)$ and $x(t)$ is a stationary Gaussian process with mean zero and spectral density $f(\lambda) = \frac{1}{2\pi} \frac{1}{|P(\lambda)|^2}$ where $P(\lambda) = \sum_{j=0}^{m} a_j (i\lambda)^j$

In their paper of 1970, Kimeldorf and Wahba show that the solution f of Problems I and II is the same function.

§1.6 Some General Remarks

(a) A little reflection will convince the reader that smoothing spline methods will be the most useful when

(i) An appropriate parametric model is not known and

(ii) High accuracy is needed and

(iii) A considerable amount of (noisy) data is available.

When these conditions are satisfied, spline methods will give very good value for the computational effort invested.

(b) The (smoothing) spline is very much the child of its age - the 1960's - depending as it does on optimization theory and the medium/large computer for its implementation.

(c) Hilbert space methods associated with L_2 spaces have been used throughout this thesis, in preferences to more general ones in other L_p spaces (where these exist). In the absence of any evidence that L_p is more appropriate than L_2 for the particular application, we believe this to be justifiable. The fact that $L_2[0,1]$ can always be embedded in $L_p[0,1]$ for $1 \leq p \leq 2$ further bolsters this position.

CHAPTER 2
Statistical Spline Methods Literature

§2.1 Preamble

In the introduction we tried to place spline methods in the statistical scheme of things. Splines will be seen to be a departure from the most general non-parametric model back a little towards the (structure rich) parametric situation. The reward for the changed position is improved efficiency coupled still with non-parametric integrity.

Although spline functions were available in a highly refined form by the mid 1960's they were for some years largely ignored by statisticians. This situation was dramatically changed by the appearance of the paper by Kimeldorf and Wahba (1970). Although the author's intention was to show the equivalence of smoothing by splines with the finding of a posterior mean, the real effect was to convince statisticians that splines were effective and not as difficult as they had thought.

When all is said and done, spline methods are just a way of fitting a smooth curve to some data. The curve estimates most studied in the statistical spline literature are for non-parametric density estimation from an independent, identically distributed sample and for the estimation of the spectral density of a stationary time series. Splines have also been used in other areas, but are at present more a curiosity than a serious practical tool.

§2.2 Non-Parametric Density Estimation

Our account will be confined exclusively to density estimation based on an independent identically distributed sample. There are two main routes we may follow: we may use the empirical distribution function in some way or we may

use an appropriate analogue of maximum likelihood adapted to the infinite dimensional (non-parametric) situation.

The empirical density is very easily obtained from the empirical distribution when we use the Sobolev spaces W_m because of the following:

Lemma: The solution f of the problem

$$(A) \qquad \begin{aligned} &\text{Minimize } \int_0^1 (f^{(m)})^2 dt \text{ with } f \in W_m \\ &\text{and } f(t_i) = y_i, \quad i = 1,\ldots,n \end{aligned}$$

and the solution g of the problem

$$(B) \qquad \begin{aligned} &\text{Minimize } \int_0^1 (g^{(m-1)})^2 dt \text{ with } g \in W_{m-1} \\ &\text{and } (D^{-1}g)(t_i) = y_i, \quad i = 1,\ldots,n \end{aligned}$$

are related by $Df = g$. This means the empirical spline fitted density is obtained by differentiating the spline fitted distribution function.

The histosplines described by Boneva, Kendall, and Stefanov (1971) are empirical densities, in the nature of a smooth analogue of a histogram, with pleasant mathematical features. To make their analysis feasible, the authors are prepared to allow densities which sometimes take small negative values in a small region.

Let $W_1(-\infty,\infty)$ denote the set of functions on the real line which are, along with their first derivatives, in $L_2(-\infty,\infty)$. The inner product is $\langle u,v\rangle = \int_{-\infty}^{\infty} (uv + u'v')dt$. Also let ℓ_2 denote the set of square summable (double ended) real sequences with inner product $(\theta u,\theta v)$. Define $\theta: W_1 \to \ell_2$ by $(\theta u)_j = \int_j^{j+1} u(t)dt$. Now define a new inner product $[u,v]$ on $W_1(-\infty,\infty)$ by

$[u,v] = (\theta u, \theta v)_{\ell_2} + \int_{-\infty}^{\infty} u'v'$ which generates the same topology as the old one. Write $Z = \{u \in W_1 : \theta u = 0\}$ and $S = \{s \in W_1 : [s,u] = 0 \text{ for all } u \in Z\}$. Each $\sigma \in W_1$ has unique decomposition $\sigma = s + z$, $s \in S$, $z \in Z$. Thus we obtain the projection $P: W_1 \to S$ where $P\sigma = s$.

Then the authors show

1) θ is a 1-1 bicontinuous map $S \to \ell_2$
2) S consists of all $s \in W_1$: $\int_{-\infty}^{\infty} s'z' = 0$ for all $z \in Z$
3) For given $h \in \ell_2$, $\theta^{-1}h$ is the unique solution of $\theta\sigma = h$ which minimizes $\int_{-\infty}^{\infty} (\sigma')^2$
4) S consists of those functions continuous and continuously differentiable such that
   i) s(t) is quadratic in each cell
   ii) $\int (s^2 + s'^2) < \infty$.

The <u>delta-spline</u> is that function $s_0 \in S$ which has $(\theta s_0)_0 = 1$, $(\theta s_0)_i = 0$, $i \neq 0$. This function is tabulated explicitly in the paper. All of the maneuvering with Z and [ ] is rewarded with the following result.

<u>Proposition</u>: Take $h \in \ell_2$, $h = (h_j)$. For any integer j, let s_j be the translated delta-spline with $(\theta s_j)_j = 1$, $(\theta s_j)_i = 0$, $i \neq j$. Then the <u>unique histospline</u> $\sigma \in W_1$ which has $\theta\sigma = h$ and which minimizes $\int (\sigma')^2 dt$ is given by $\sigma = \sum_{j=-\infty}^{\infty} s_j$. The paper of Boneva, Kendall and Stefanov (1971) also describes another histospline and includes much empirical material on histospline behavior.

Some Remarks on Histosplines

1. Once the tabulated form of the delta-spline is stored in the computer (requiring 39 parameters) there is no explicit optimization required - just the grouping of the data into classes. Consequently this method is well suited to

programmable calculators and mini-computers without optimization routines.

2. Boneva, Kendall and Stefanov (1971) and Schoenberg (1972a and b) also consider the variant histospline defined as the derivative of that function G in $W_2[0,1]$ which solves:

$$\text{2.2.1} \qquad \begin{aligned} &\text{Minimize } \int_0^1 (G'')^2 dt \\ &\text{Subject to } G(0) = 0 \text{ and } G(ih) = \sum_{i=0}^{i-1} h_j \quad i = 1,2,\ldots,\ell \\ &\text{where } (\ell+1)h = 1 \text{ and } G'(0) = G'(1) = 0 \end{aligned}$$

Yet another variant of this problem is considered by Wahba (1975b). This involves replacing the final constraints in 2.2.1 by

$$G'(0) = \hat{a}_1 \quad \text{and} \quad G'(1) = \hat{b}_1 \quad \text{where} \quad \hat{a}_1, \hat{b}_1$$

are calculated from the empirical distribution function. This variant gives better accuracy near 0 and 1 than 2.2.1. When the criterion is minimum mean square error at a point, Wahba (1975b) also shows how to choose h optimally.

Finally we note that if the problem 2.2.1 has the further constraint $G'(t) \geq 0$ for all $t \in [0,1]$ the solution will be isotonic with respect to a natural order on W_2 and will give a more acceptable density function.

3. It must be emphasized that histosplines are interpolating splines based on the sample histogram, and not a smoothing spline. Consequently in the presence of noise (sampling error) we cannot expect this method to be much better at filtering the noise than the histogram it is derived from.

This assertion is supported by the results of Wahba (1975b) who shows for her variant of the histospline that for the true density $f \in W_m$ and f_n the histospline corresponding to a sample of size n

$$E(f_n(t) - f(t))^2 = 0\ (n^{-(2m-1)/2m}).$$

In a companion paper Wahba (1975a) shows that the expected mean square error at a point t has that same order of magnitude for all of the following estimation methods: the polynomial algorithm (Wahba), kernel type estimator (Parzen), certain orthogonal series estimates (Kronmal-Tatar), and the ordinary histogram. However, the constants covered by the 0 may be larger in these latter cases.

§2.3 Densities by Maximum Penalized Likelihood

This area is relatively unexplored to date. The analogy with parametric maximum likelihood estimation gives rise to the hope that Maximum Penalized Likelihood Estimators (MPLEs) may be optimal in some fundamental sense. We now look at some of the details.

Let Ω be an interval (a,b) and let $H(\Omega)$ be a manifold in $L_1(\Omega)$. Suppose $(t_1,t_2,\ldots,t_n)$ is a i.i.d. sample from an unknown density $f \in L_1(\Omega)$. Unfortunately the problem

$$\text{2.3.1} \qquad \text{Maximize } L(v) = \prod_{i=1}^{n} v(t_i) \text{ subject to } v \in H(\Omega),$$
$$\int_\Omega v(t)dt = 1,\ v(t) \geq 0 \quad \forall t \in \Omega$$

will not have a solution for most manifolds of interest (certain isotone manifolds for example of unimodal or monotone functions are an exception). Specifically, any manifold which contains an approximating sequence to any linear combination of δ-functions, admits no maximum likelihood estimator for the density f.

From heuristic Bayesian considerations, Good and Gaskins (1971) suggested adding a penalty term to the likelihood which would penalize unsmooth estimates.

They chose a manifold and penalty function that lead inevitably to polynomial splines. Good's and Gaskin's results were refined and made rigorous by de Montricher, Tapia and Thompson (1975). We can now describe the current state of the art.

It will normally be the case that the manifold $H(\Omega)$ is contained in $W_m(\Omega)$ and the penalty function $\Phi(v) = ||v||_{W_m}^2$. Let $\hat{L}(v) = \prod_{i=1}^{n} v(t_i)\exp(-\Phi(v))$ and consider the optimization problem

2.3.2

$$\text{Maximize } \hat{L}(v) \text{ subject to}$$

$$\text{(i)} \quad v \in H(\Omega)$$

$$\text{(ii)} \quad \int_\Omega v \, dt = 1$$

$$\text{(iii)} \quad v(t) \geq 0, \ \forall t \in \Omega$$

The solution v is the MPLE of the underlying density, f.

The task of computing the MPL Estimate of the density is greatly simplified by knowing the form the optimum must take. The following existence theorem is proved in the paper by de Montricher, Tapia and Thompson (1975).

Theorem: For $m \geq 1$, the MPLE corresponding to W_m exists, is unique, and is a polynomial spline of degree 2m-1. Moreover, if the estimate is positive in the interior of an interval, then in this interval it is of degree 2m-1 and of continuity class 2m-2 with knots at the sample points.

From (Fisher) information-theoretic considerations, as well as a desire to avoid the awkward non-negativity constraint $v(t) \geq 0$, Good and Gaskins (1971) also considered the MPLE problem with manifold

2.3.3
$$H_1(\Omega) = \{v: v^{\frac{1}{2}} \in W_1 \ (-\infty,\infty)\} \text{ and}$$
$$\Phi_1(v) = \alpha \int_{-\infty}^{\infty} \frac{(v')^2}{v}\, dt, \quad \alpha > 0.$$

H_1 consists of those non-negative functions in L_1 whose square roots have derivative in L_2, and the squared L_2 norm of the derivative of the square root is the roughness penalty.

After noting that the reformulation trick 2.3.3 is standard in the literature, deMontricher, Tapia and Thompson (1975) record conditions for its valid use with the following lemma.

Lemma: Let H be a subset of $L_2(\Omega)$ and J a functional on H. Consider

Problem I
$$\text{Maximize } J(v^{\frac{1}{2}}) \text{ subject to}$$
$$v^{\frac{1}{2}} \in H, \int v\, dt = 1, v(t) \geq 0 \ \forall t$$

and Problem II
$$\text{Maximize } J(u) \text{ subject to } u \in H$$
$$\text{and } \int u^2 dt = 1$$

Let u* be a solution of II. Then $v^* = (u^*)^2$ solves I if and only if $|u^*| \in H$ and $J(u^*) = J(|u^*|)$.

The authors (de Montricher, Tapia and Thompson) go on to establish that the price of using the non-negativity trick is to lose the polynomial spline form of solution - the solution is an exponential spline instead, with knots at the sample points.

We regretfully record that on the two most important aspects - how effectively noise is filtered out, and the asymptotic (large n) performance - the literature on MPLEs is quite silent.

§2.4 Noise Filtering by Smoothing Splines

Statisticians and applied mathematicians are continually faced with the problem of recovering a smooth function when only noisy measurements of it are available. In fitting a parametric model the residuals are made up of the noise as well as the deviations of the model from the true function. Smoothing splines are admirably placed to estimate this true function (known only to be smooth) for two reasons. First, they are flexible enough to respond to any real local variation, without allowing pathological behavior, and second, the actual degree of smoothing (= filtering of noise together with rapid variation) is controllable. Even when the correct degree of smoothing is unknown, these features, in conjunction with a technique called cross-validation (to determine the correct degree of smoothing) allow us to remove most of the model deviation component from the residuals, leaving virtually only the (real) noise. We presently give an account of the main features of fitting smoothing spline functions by cross validation - full details are given by Wahba and Wold (1975a and b).

The model we are fitting is

$$Y(t) = f(t) + e(t) \quad t \in [0,1] \text{ where}$$

2.4.1 $$f \in W_m \text{ and } Ee(t) = 0 \text{ all } t \text{ and}$$

$$Ee(s)e(t) = \begin{cases} \sigma^2 & s = t \\ 0 & s \neq t \end{cases}$$

The noise variance σ^2 is generally unknown and $Y(t)$ is observed at (an increasing set of points) $t_1, t_2, \ldots, t_n$.

Consider the problem: Find $f \in W_m$ to

$$\text{Minimize } \left(\frac{1}{n} \sum (Y(t_j) - f(t_j))^2 + \lambda \int_0^1 (f^{(m)})^2 dt\right) \tag{2.4.2}$$

where $\lambda > 0$ is a fixed real number.

The first term is a measure of the fidelity to the data, the second is λ times the "smoothness" of f. The optimum solution is known (Greville (1969), Reinsch (1967)) to be a cubic spline with knots at the t_i, $i = 1,2,\ldots,n$. As $\lambda \to \infty$, the solution $f_{n,\lambda}$ approaches its smoothest possible form - the least squares straight line through the data. As $\lambda \to 0$, $f_{n,\lambda}$ approaches the interpolating spline through all of the data points. Thus we call λ *the degree of smoothing*. It is shown in Wahba (1973 and 1974) that in order to have $f_{n,\lambda} \overset{W_m}{\to} f$ as $n \to \infty$ we must also have $\lambda \to 0$.

If (from previous experience of our particular problem) the correct value of λ is known, we have only to solve 2.4.2 using that λ. Unless the problem 2.4.1 can be converted to the periodic smoothing spline form of Cogburn and Davis (1974) there is no simple way of solving 2.4.2 other than the usual optimization routine.

When λ is not known we can (with much labor) use the Cross Validation Mean Square Error (minimizing) technique to estimate the appropriate degree of smoothing *from the data alone*. The method has been used successfully in various applications by Feinberg and Holland (1972), Hocking (1972), Mosteller and Wallace (1963) and others. In effect the CVMSE method gives the value of parameter which maximizes the internal consistency of the data set with regard to the applied model.

Wahba and Wold find it useful to recast problem 2.4.2 into a form used by Reinsch (1967).

2.4.3 Find $f \in W_2$ to minimize $\int_0^1 (f'')^2 dt$ subject to $\frac{1}{n} \sum_{j=1}^{n} (Y(t_j) - f(t_j))^2 \leq S$, where S is given.

It is well known that if

$$n S \leq \inf_{a,b} \sum (Y(t_j) - a + b\, t_j)^2$$

then there exists a unique $\lambda = \lambda(S)$ such that $f_{n,\lambda}$ is the solution to 2.4.2 and

$$\frac{1}{n} \sum (Y(t_j) - f_{n,\lambda}(t_j))^2 = S.$$

Armed with the appropriate tools from the last paragraph, we now give an account of Wahba and Wold's Cross Validation procedure (1975a).

1) Divide the data set into p groups

Group 1: $t_1, t_{1+p}, t_{1+2p}, \ldots$

Group 2: $t_2, t_{2+p}, t_{2+2p}, \ldots$

...

Group p: $t_p, t_{2p}, t_{3p}, \ldots$

2) Guess a starting value for S (Almost invariably $S = k\sigma^2$ with $0.7 \leq k \leq 1$. A reasonable starting point might be $k = 0.8$).

3) Delete the first group of data. Fit a smoothing spline to the rest of the data using the method of Reinsch with the S of step 2. Compute the sum of squared deviations of this smoothing spline from the deleted data points.

4) Delete instead the second group of data. Fit a smoothing spline with the S of step 2. Compute the sum of squared deviations of the spline from the data points.

5) Repeat Step 4 for the 3rd, 4th, ..., p^{th} group of data.

6) Add the sums of squared deviation from steps 3-5, and divide by n. This is the CVMSE for S and is written CV(S).

7) Determine the $S = S_1$ making CV(S) a minimum.

The smoothing problem 2.4.3 can now be solved with $S = S_1$.

Empirical studies by Wahba and Wold (1975a) indicate that when σ^2 is extremely small, the CVMSE estimate for k in $S = k\sigma^2$ has positive bias, resulting in very slight undersmoothing. This effect is negligible for realistic sized σ^2, although the authors do not present a proof.

§2.5 Periodic Smoothing Splines

The work of Cogburn and Davis (1974) has been referred to several times already. We now describe their results in detail.

Let G be the group of real numbers modulo 2Π with the usual topology and measure. The model to be fitted is

$$h(t) = f(t) + e(t) \qquad \forall t \in G$$

2.5.1 $$\text{with } f \in W_m(G) \text{ and } Ee(t) = 0 \ , \ \forall t \in G$$

$$\text{and } Ee(s)e(t) = \sigma^2 \quad s = t$$
$$= 0 \quad s \neq t$$

where h is observed either on a lattice of points or continuously and the noise variance σ^2 is unknown. The asymptotic solution for large n devised by Cogburn and Davis is very convenient to handle, and easy to compute since it avoids explicit optimization.

First we need some definitions. Let $L_2(G)$ be the set of all square integrable functions on G. Let $W_m(G) = \{g \in L_2(G)$: $g^{(j)}$ is absolutely continuous, $j = 1,2,\ldots, m-1$ and $g^{(m)} \in L_2(G)\}$. Let L be a linear differential operator

of order m

$$L = D^m + \gamma_1 D^{m-1} + \ldots + \gamma_m \quad \text{with } \gamma_j \text{ constant.}$$

In approximating a function $h \in L_2(G)$ by $g \in W_m$ we need the following measure of closeness. $\Delta_{n,\lambda,L}(g,h) = \frac{1}{n}\sum_{k=1}^{2n}(g(\frac{k\Pi}{n}) - h(\frac{k\Pi}{n}))^2 + \frac{1}{\lambda^{2m}\Pi}\int_0^{2\Pi}(Lg)^2dt$. The function $g \in W_m(G)$ minimizing $\Delta(g,h)$ will be called the Periodic Lattice Smoothing Spline (LSS) to h. When h is known for all $t \in G$ the smoothing spline to h is that function $g \in W_m(G)$ minimizing

$$\Delta_{\lambda,L}(g,h) = \frac{1}{\Pi}\int_G (g(t) - h(t))^2dt + \frac{1}{\lambda^{2m}\Pi}\int_G (Lg)^2\,dt.$$

Such a $g \in W_m(G)$ is called the Periodic Continuous Smoothing Spline (CSS) to h.

Cogburn and Davis discuss algorithms for fitting the LSS and CSS. We summarize their final form.

Let P(u) be the characteristic polynomial of L so that $P(u) = u^m + \gamma_1 u^{m-1} + \ldots + \gamma_m$ and let $Q(k) = |P(ik)|^2$, $i = \sqrt{-1}$. Take $q_{n,j} = 1 + Q(j)\,[\sum_{\ell\neq 0} 1/Q(j+2n\ell)]$ and let $a_{n,\lambda,j} = \frac{1}{2n}\left(\frac{\lambda^{2m}}{Q(j) + \lambda^{2m}q_{n,j}}\right)$, $|j| \leq n$ and $a_{n,\lambda,\ell} = \frac{Q(j)}{Q(\ell)}a_{n,\lambda,j}$ for $\ell \in \{j \pm 2n, j \pm 4n, \ldots\}$, $|j| \leq n$. Then the LSS to h is given by the (discrete) convolution

$$h^{*n}S_{n,\lambda}(t) = \sum_{\ell=-n+1}^{n} h(\tfrac{\ell\Pi}{n})S_{n,\lambda}(t - \tfrac{\ell\Pi}{n})$$

where $S_{n,\lambda}(t) = \sum_{k=-\infty}^{\infty} a_{n,\lambda,k}\,e^{ikt}$. Letting $a_{\lambda,\ell} = \lim_{n\to\infty} n\,a_{n,\lambda,\ell} = \frac{\lambda^{2m}}{2(Q(\ell) + \lambda^{2m})}$ and $S_\lambda(t) = \frac{1}{\Pi}\sum a_{\lambda,\ell}\,e^{i\ell t}$ we can closely approximate $n\,S_{n,\lambda}(t)$ by

$\Pi\ S_\lambda(t)$. The CSS to h is given by the convolution

$$h \circledast S_\lambda(t) = \int_G h(y)\ S_\lambda(t - y)dy\ .$$

If it is known that the function f to be estimated from 2.5.1 has derivatives of order m, but no specific operator L is known, a natural choice is $L = D^m$; in which case $S_\lambda(u)$ becomes

$$S_\lambda(u) = \frac{1}{2\Pi} \sum_{\ell=-\infty}^{\infty} \frac{\lambda^{2m}}{\ell^{2m} + \lambda^{2m}}\ e^{i\ell u}\ .$$

Writing $t_\lambda(u) = \frac{1}{2\Pi} \int_{-\infty}^{\infty} \frac{\lambda^{2m}}{\lambda^{2m} + y^{2m}}\ e^{iuy}\ dy$ and $\hat{t}_\lambda(u) = \sum_{k=-\infty}^{\infty} t_\lambda(u+2k\Pi)$, it is shown $h \circledast S_\lambda(u) = h \circledast \hat{t}_\lambda(u) = \int_{-\infty}^{\infty} h(y)t_\lambda(u - y)dy$.

Remarks

1) The asymptotic results of Cogburn and Davis are doubly valuable because: (a) they are computationally easy, and (b) they facilitate further theoretical investigation by giving the LSS and CSS in closed form.

2) The requirement that the data should be periodic and lattice is tailor-made for the usual spectral estimation from periodogram. Unfortunately, any δ-function spikes in the spectrum which cannot be well approximated by functions in $W_m(G)$ will not be well resolved. The next section considers this question.

3) No work seems to have been done on ways to convert non-lattice, non periodic data fitting problems to the easy periodic lattice form of this section.

§2.6 Estimating Spectral Densities

Determining spectral densities gives spline theory not only a great opportunity but also a severe test. When the spectrum is absolutely continuous, spline methods are extremely effective. However, when the spectrum has a discrete component, i.e. δ-function spikes, spline methods based on L_2 have little chance of sharply resolving the spike without a great deal of data in the vicinity of the spike. The heart of the difficulty is that every sequence of functions approximating a δ-function must be unbounded in L_2 norm and so also in W_m norm, and thus the smoothing spline is duty bound to flatten these real spikes out. However, before abandoning the L_2 based W_m spaces, we need to see the problem in perspective. Because of the superficial similarity between a δ-spike and a noisy observation, any attempt to transfer the problem to a space where δ-function approximants are bounded seems doomed to failure because we would be unable to filter out the noise in such a space.

Thus reconciled to remaining in W_m with its pleasant inner product and Fourier Transform structure we may yet be able to find a way out. One strategy may be to proceed as follows. Since further data points are easily obtained from the periodogram, it might be feasible to use repeated applications of the CVMSE method coupled with a procedure to introduce extra data points in regions where the rate of change (of fitted function) is large.

First let us examine the current state of the art for estimating spectral densities with spline methods.

Estimating the Spectral Density of a Stationary Stochastic Process

Let $X_1, X_2, X_3, \ldots$ be a second order stationary stochastic process with $EX_k = 0$, $EX_j X_{j+k} = \rho_k$.

The ρ_k are Fourier coefficients of a symmetric distribution function F on $[-\Pi,\Pi]$, $\rho_k = \frac{1}{\Pi} \int_0^{\Pi} \text{Cos kw d F(w)}$. When F is absolutely continuous, it is completely determined by the spectral density

$$f(w) = (DF)(w)$$
$$f(w) = \sum_{-\infty}^{\infty} \rho_k e^{ikw} .$$

The statistical problem is to estimate f(w) on the basis of observations $X_1, X_2, \ldots, X_n$. Let $\hat{f}(w)$ denote the periodogram

2.6.1
$$\hat{f}(w) = \sum_{-n+1}^{n-1} \hat{\rho}_k e^{ikw} = \hat{\rho}_o + \sum_{k=1}^{n-1} \hat{\rho}_k \text{ Cos kw} \quad \text{where}$$
$$\hat{\rho}_k = \frac{1}{n} \sum_{j=1}^{n-k} X_j X_{j+k}, \quad k = 0,\ldots,n-1.$$

When the process is Gaussian it is shown in Walker (1965) that

2.6.2
$$\hat{f}(w) = f(w)\, U_\varepsilon(w) + \eta_n(w)$$

where $\eta_n \to 0$ in probability as $n \to \infty$ and $U_\varepsilon(j\Pi/n)$ are uncorrelated exponential random variables with a mean and variance of 1 for $j = 0,\ldots,\pm n-1$.
Since the periodogram is an inconsistent estimator of f, some modification is required. Smoothed (consistent) estimators of f(w) are obtained either by smoothing the periodogram

$$f^*(w) = \int_{-\Pi}^{\Pi} \hat{f}(\lambda)\, K(w - \lambda)\, d\lambda$$

or by weighting the covariances by a "lag window" $k_M(r)$ giving

$$f^*(w) = \frac{1}{2\Pi} \sum_{r=-M}^{M} k_M(r) e^{-irw} \hat{\rho}_r$$

where

$$K(w) = \frac{1}{2\Pi} \sum_{-M}^{M} e^{-irw} k_M(r) \ .$$

Cogburn and Davis (1974) consider estimating f by a CSS or LSS to $\hat{f}$: ie. $\hat{f} \circledast^n S_{n,\lambda}$ or $\hat{f} \circledast \hat{t}_\lambda$. They take $L = D^m$ and obtain an estimate of the integrated mean square error for n large. It is $\int_{-\Pi}^{\Pi} MSE(\hat{f}(t))dt \cong \frac{\lambda}{2n} \sigma_m^2 \int_{-\Pi}^{\Pi} f^2 dt + \frac{1}{\lambda^{4m}} \int_{-\Pi}^{\Pi} (f^{(2m)})^2 dt$. The value of λ minimizing the RHS is

$$\lambda_o = \left(\frac{4nm}{\sigma_m^2} \int (f^{(2m)})^2 dt / \int f^2 dt \right)^{\frac{1}{4m+1}}$$

and the resulting MSE is $0(\frac{1}{n})$.

Wahba and Wold (1975b) are concerned with estimating log f(w) for spectral densities f which are bounded below. Taking the logarithm of (2.6.2) converts the curve fitting problem to that of §2.5.

We have

$$Y(j) = \log f\left(\frac{j\Pi}{N}\right) + \gamma + e_j, \quad j = \pm 1, \pm 2, \ldots, \pm n-1$$

with $Ee_j = 0$, $Ee_j^2 = \frac{\Pi^2}{6}$, $\gamma = 0.5772....$ Euler's constant, and a reasonably symmetric distribution of e_j about 0 with constant variance. In their paper (1975b) Wahba and Wold use various results from Cogburn and Davis (1974) en route to their objective which is to show (in principle) that the smoothing parameter chosen by CVMSE converges to the parameter minimizing the mean squared error.

Wegman (1977) records the various advantages and disadvantages of using log f(w) rather than f(w) for the spline fitting model. Although the fit to log f(w) is improved, the kernel interpretation and the attendant results on consistency are lost. For multiple time series (with which Wegman was concerned) various functions such as transfer function, multiple coherence and coherency fit the spline model directly via log $\hat{f}_{XX}(w)$, log $\hat{f}_{XY}(w)$, etc. and each of the above functions may be estimated directly with a spline, rather than as products and quotients of spline estimates.

CHAPTER 3

Isotonic Splines

§3.1 Preamble

There are many model fitting problems where we either have some prior knowledge of the form the solution must take, or else have some insight into the laws governing the system. This knowledge may be equivalent to the requirement that the fitted function preserve some order on the data points. An obvious example is a fitted distribution function - this must satisfy $F(x) \geq F(y)$ whenever $x \geq y$. Functions which preserve (in some sense) an order relation of their arguments are called *isotone*. Knowledge of isotonicity may follow from very elementary considerations. For technical convenience, the class of partial orders will need to be restricted to those compatible with the (additive) group of all real functions. Our framework will certainly be general enough to cater for the monotone, or convex, or positive functions and other families.

The difficulty in compelling an arbitrary function to be isotone, is that this in general involves an infinite number of constraints. While our methods can deal with a continuum of constraints (if they are weakly compact) they cannot save us from the consequences of them. As our results show, the most general isotone spline solving the GHB problem may have an infinite number of knot points. This situation is unsatisfactory to numerical analysts, who need concrete solutions. However, we also propose and solve a slightly less general isotone spline problem, which has only a finite number of boundedness and uniformity condition (of a form that a user can live with), and the fitted spline has only a finite number of knots.

The main justification for smoothing spline theory is that it provides a non-parametric framework for extracting a function from some noisy observations. It is expected that isotonic splines will facilitate an even more drastic filtering of the noise. Regretfully, we offer no evidence on this here.

§3.2 Partially Ordered Groups of Functions

We shall be dealing only with abelian groups of real functions which can be added pointwise.

The relation $\succeq$ on the group G of functions is required to satisfy the following conditions

i) $f \succeq f$

ii) $f \succeq g$, $g \succeq h$ implies $f \succeq h$

iii) $f \succeq g$ implies $f + h \succeq g + h$ for all $h \in G$

iv) $f \succeq 0$, $g \succeq 0$ implies $f + g \succeq 0$

v) $f \succeq g$, $g \succeq f$ implies $f = g$.

A group satisfying (i) to (v) is called a *partially ordered group*. Sometimes it is convenient to drop the insistence on (v) in which case we have a pre-ordered group.

The set $P = \{g \in G: g \succeq 0\}$ is called the *positive cone* of the order $\succeq$. The set P defines and is defined by the (partial) order $\succeq$. We also say that P is just the set of functions which are isotone with respect to the order $\succeq$.

§3.3 Realization of Partial Orders on W_m

We restrict our attention to W_m although the results and methods can apply to more general spaces.

Let F be a continuous linear map of W_{m-1} into L_2 which commutes with differentiation operators. That is to say

$$D(Ff) = F(Df) \text{ for all } f \in W_m \subset W_{m-1} .$$

It is known that such an F can be represented in the form $Ff = S*f$ where S is a Schwartz distribution.

We can now define a partial order $\succeq$ on W_m by $f \succeq 0$ if and only if $(Ff)(t) \geq 0$ for $\forall t \in [0,1]$.

Examples: (i) If F = Identity the set of functions $\succeq 0$ are just the positive functions in W_m. (ii) If F = D, the set of functions $\succeq 0$ is just the set of monotone increasing functions in W_m. (iii) If $F = D^2$ the set of functions $\succeq 0$ is just the set of convex functions in W_m.

§3.4 Isotonic Smoothing Splines

The isotonic splines dealt with in this chapter solve the problem:

3.4.1
$$\text{Minimize } \int_0^1 (f^{(m)})^2 dt \text{ subject to}$$
$$f \in W_m, \quad \alpha_i \leq f(t_i) \leq \beta_i, \quad Ff(t) \geq 0 \quad \forall t \in [0,1].$$

Our results show that a solution exists and is a spline of degree 2m-1. The Hilbert space methods used in this problem cannot be used directly with the other (commoner) smoothing spline problem:

3.4.2
$$\text{Minimize } (\lambda \int_0^1 (f^{(m)})^2 dt + \Sigma(y_i - f(t_i))^2$$
$$\text{subject to } f \in W_m , \quad Ff(t) \geq 0 \quad \forall t .$$

The objective function in 3.4.2 is not a Hilbert space norm squared, and hence the difficulty. However, 3.4.2 is the minimization of a convex function with linear constraints so it can in principle be handled by non-linear programming. The existence of a solution of 3.4.2 is never in doubt - only its character.

§3.5 A Restricted Isotonic Spline

First we solve the restricted problem, then move on to the most general one.

Conditions: Let F be a continuous linear map $W_{m-1} \to L_2$ which commutes with differentiation. Denote the partial order defined by F on W_m by $\succeq$. Assume there exists a function $g \in W_m$ satisfying $\alpha_i < g(t_i) < \beta_i$ for $i = 1,2,,\ldots,n$ with $(Fg)(t) \geq \epsilon > 0$ for $\forall t \in [0,1]$. The continuity of DF implies there exists $d > 0$ such that $||DFf||_{L_2} \leq d||D^m f||_{L_2}$, $\forall f \in W_m$. Take $N \geq \dfrac{d^2||D^m g||_{L_2}^2}{\epsilon^2}$.

Theorem 1: Under the conditions just given the problem:

$$\text{Minimize } \int_0^1 (f^{(m)})^2 \, dt \text{ subject to}$$

$$f \in W_m, \quad f \succeq 0, \quad \alpha_i \leq f(t_i) \leq \beta_i \quad i = 1,2,\ldots,n$$

$$Ff(j/N) \geq \epsilon \quad j = 0,1,\ldots,N$$

has a solution which is a polynomial spline of degree 2m-1 with possible knots at t_i; $i = 1,\ldots,n$ and j/N; $j = 0,1,\ldots,N$. Before we can prove Theorem 1 we need to establish some preliminary results, and two major theorems.

Lemma 1: $||D^m f||_{L_2} \leq ||D^m g||_{L_2}$ and

$Ff(j/N) \geq \epsilon$ for $j = 0,1,\ldots,N$ imply $f \geq 0$.

The straightforward proof is omitted.

Let T be a continuous linear map $W_m \overset{\text{onto}}{\to} L_2$ (in our case $T = D^m$). Suppose Ker T is of finite dimension. Let

$$\Lambda_1 = \{u_i \in W_m;\ i = 1,2,\ldots,n : \langle u_i, f\rangle = -f(t)\ \forall f \in W_m\}$$

and

$$\Lambda_2 = \{v_i \in W_m;\ i = 1,\ldots,n : \langle v_i, f\rangle = f(t_i),\ \forall f \in W_m\}$$

and

$$\Lambda_3 = \{w_i \in W_m;\ i = 0,1,\ldots,N: \langle w_i, f\rangle = -Ff(j/N),\ \forall f \in W_m\}.$$

Take $L = \Lambda_1 \cup \Lambda_2 \cup \Lambda_3$ and define a map $p: L \to R$ by

$$p(u_i) = -\alpha_i \quad u_i \in \Lambda_1$$

$$p(v_i) = \beta_i \quad v_i \in \Lambda_2$$

$$p(w_i) = -\epsilon \quad w_i \in \Lambda_3 .$$

Now define a convex subset C of W_m by

$$C = \{f \in W_m: \quad \forall \ell \in L,\ \langle \ell, f\rangle \leq p(\ell)\}.$$

We are seeking an element $\sigma \in C$ satisfying

$$||T\sigma|| = \operatorname{Min}_{f \in C} ||Tf||.$$

Such an element σ will be called the (generalized) spline function relative to T in the set C. Let $\tilde{C} = \{f \in W_m : \forall \ell \in L \ \langle \ell, f \rangle \leq 0\}$. We need the following assumption

$$H1 \ldots\ldots \tilde{C} \cap (\operatorname{Ker} T) = \{0\}.$$

Theorem 2: (M. Attéia and P. L. Joly). Under assumption H1 there exists at least one spline function relative to T in the convex set C.

Proof: Let $\Omega = T(C)$ and $\tau = T\sigma$. We must have (*) $||\tau|| = \operatorname{Min}_{y \in \Omega} ||y||$. The subset Ω is convex. If moreover Ω is *closed* there will exist a unique element $\tau \in \Omega$ satisfying (*)(it is the projection of the origin onto Ω). Thus $\sigma \in C$ exists such that $T\sigma = \tau$.

Proof that Ω is closed: $\Omega = T(C)$ is closed iff $(\operatorname{Ker} T) + C = T^{-1}(\Omega)$ is closed in W_m. Clearly the restriction $\tilde{T}$ of T to $(\operatorname{Ker} T)^{\perp}$ is a homeomorphism and

$$\Omega = \tilde{T}\left((\operatorname{Ker} T + C) \cap (\operatorname{Ker} T)^{\perp}\right). \qquad \square$$

Dieudonne's Theorem: A,B closed non-empty, convex subsets of a T.V.S. with A locally compact. Let $A_\infty = \bigcap_{\lambda > 0} \lambda(A - a)$ $a \in A$. Then $A_\infty \cap B_\infty = 0$ implies A - B closed.

In the present case Ker T is locally compact (because it is of finite dimension) and we can easily prove that $(\text{Ker } T)_\infty = \text{Ker } T$ and $C_\infty = \tilde{C}$. Thus assumption H1 implies C + Ker T is closed.

In order to characterize the spline relative to T in C we need some more structure axioms.

H2 The subset $L \subset W_m$ is weakly compact and the map p is a continuous map of L (with weak topology) into R.

H3 $C \cap \text{Ker } T$ is empty

H4 $I = \{f \in W_m: \ \forall \ell \in L, \ <\ell, f> < p(\ell)\}$ is not empty.

We can now put everything together with the following result.

Theorem 3: (P. J. Laurent). Under assumptions H1, H2, H3, H4 the element C is a spline function (relative to T in C) if and only if

$$-T^*T\sigma \in \overline{CC}(F_\sigma) \text{ where } F_\sigma = \{\ell \in L: <\ell,\sigma> = p(\ell)\}$$

and $\overline{CC}(F_\sigma)$ denotes the smallest closed convex cone with vertex 0 containing F_σ; and T^* is the adjoint of T.

Proof: Consider the function $f(y) = f_1(y) + f_2(y)$ with $f_1(y) = ||T_y||$ and $f_2(y) = 0$ if $y \in C$, ∞ if $y \notin C$. We wish to characterize σ such that

$$f(\sigma) = \min_{y \in W_m} f(y).$$

The element σ satisfies this iff

$$0 \in \partial f(\sigma)$$

where $\partial f(\sigma)$ denotes the set of subgradients of f at σ. As f_1 is convex and continuous and f_2 convex and lower semicontinuous we have

$$\partial f(\sigma) = \partial f_1(\sigma) + \partial f_2(\sigma).$$

From H3 we have $||T\sigma|| \neq 0$ and $\partial f_1(\sigma)$ is reduced to one element $\partial f_1(\sigma) = \frac{T^*T\sigma}{||T\sigma||}$. Using H2, H3, H4 it is shown that

$$\partial f_2(\sigma) = \overline{CC}\ (F_\sigma) .$$

Finally the condition (*) is satisfied iff there exists $\ell_1 \in \partial f_1(\sigma)$ and $\ell_2 \in \partial f_2(\sigma)$ such that $\ell_1 + \ell_2 = 0$ i.e. if

$$-T^*T\sigma \in \overline{CC}\ (F_\sigma) .$$

□

Proof of Theorem 1: At the optimum σ, $||D^m\sigma||_{L_2} \leq ||D^m g||_{L_2}$ and so by Lemma 1, the spline function satisfying all the constraints of L must be $\succeq 0$.

We can now translate the assumptions H1-2-3-4 into concrete form and confirm that they hold.

H1 Since T is D^m, Ker T is the set of polynomials of degree m-1. Every function in $\tilde{C}$ is zero at all the points t_i, i = 1,...,n. Hence $\tilde{C} \cap$ Ker T = {0} (i.e. H1) holds if n, the number of data points exceeds m-1.

H2 L is finite, therefore compact in any topology. $p: L \to R$ is automatically continuous.

H3 $C \cap \text{Ker } T$ is empty. This is safe to assume since any function in this set would automatically be the optimum solution of Theorem 1 without Theorems 2 and 3.

H4 $I = \{f \in W_m: \forall \ell \in L, <(\ell,f) < p(\ell)\}$ contains the function g of the conditions to Theorem 1 and so I is not empty.

Theorem 3 shows that the minimizing element σ which was shown to exist in Theorem 2 satisfies $-T^*T\sigma \in$ closed convex cone generated by all $\ell \in L$ corresponding to active constraints. In other words the functional

$$T^*T\sigma = -\sum_{\ell \in L} d_\ell \, \ell$$

with all $d_\ell \geq 0$ and $d_\ell = 0$ when ℓ is not an active constraint, and all the $\ell \in L$ are point evaluation (delta) functions. Thus $T^*T\sigma$ is zero at all points except those corresponding to active constraints (which become knots). Between knot points $T^*T\sigma = 0$ i.e. $D^{2m} = 0$ so that σ is a polynomial of degree 2m-1 between knot points. Thus Theorem 1 is proved. □

§3.6 The General Isotonic Spline

We have already mentioned the general isotonic spline, with possibly an infinite number of knots, able to occur anywhere. The following result makes all the details precise.

Theorem 4: Let $\succeq$ be the partial order defined by F: $W_{m-1} \to L_2$. If there exists $g \in W_m$ satisfying

$$Fg(t) > 0 \qquad t \in [0,1]$$
$$\alpha_i < g(t_i) < \beta_i \qquad i = 1,2,\ldots,n .$$

Then the problem

$$\text{Minimize } \int_0^1 (D^m f)^2 dt \text{ subject to}$$
$$\alpha_i \leq f(t_i) \leq \beta_i;\ i = 1,\ldots,n \text{ and}$$
$$f \succeq 0,\quad f \in W_m$$

has a solution which is a polynomial spline of degree 2m-1 with knots at the data points, and in exceptional cases a countable number elsewhere.

Proof: This follows the lines of Theorem 1 when we take $T = D^m$, $\Lambda_3 = \{w_t \in W_m,\ t \in [0,1]: \langle w_t, f\rangle = Ff(t),\ \forall f \in W_m\}$ and $p(w) = 0$ for all $w \in \Lambda_3$. The key requirement that Λ_3 be weakly compact holds in this case, and p is continuous. The rest is routine. □

§3.7 Applications of Theorems 1 and 4

We now present two interesting applications of Theorems 1 and 4, which extend the partial results of Passow (1974), Passow and Roulier (1974).

Proposition 1: If $y_1 < y_2 < \ldots < y_n$ with arbitrarily small error bounds, there exists a globally monotone (cubic) smoothing spline of the form of Theorems 1 and 4, minimizing $\int_0^1 (f'')^2$.

Proof: We only need show there exists a suitable g satisfying the assumptions of Theorem 1. Let

$$\phi(t) = \exp\,(1/(t^2-1)) \quad |t| < 1$$
$$= 0 \qquad\qquad\qquad |t| \geq 1$$

$\phi(t)$ is known to be infinitely differentiable on the whole real line and so in all W_m, $m \geq 1$. Let $\phi_1(t) = \int_{-\infty}^{t} \phi(n)dn$. ϕ_1 is in C^∞ and it is clear that

$$\phi_1(t) = 0, \quad t \leq -1$$
$$\phi_1(t) = k = \int_{-1}^{1}\phi(t)dt \text{ when } t \geq 1$$

and of course $D\phi_1(t) \geq 0$ for all t. When the conditions of this proposition hold, adding together suitably scaled translates of ϕ_1 will give a function $f_1(t) \in C^\infty$ with $Df_1(t) \geq 0$, $\forall t$ which satisfies $f_1(t_i) = y_i$, $i = 1,2,\ldots,n$. If ϵ is the error bound on the y_i, the function

$$g_1(t) = f_1(t) + \epsilon t \quad t \in [0,1]$$

satisfies Theorems 1 and 4 with F = D, m = 2. □

Proposition 2: If $\dfrac{y_{i+1} - y_i}{t_{i+1} - t_i} < \dfrac{y_{i+2} - y_{i+1}}{t_{i+2} - t_{i+1}}$ for $i = 1,2,\ldots,n-2$ and the y_i have arbitrarily small error bounds then there exists a globally convex (quintic) smoothing spline of the form of Theorem 1 and 4, minimizing $\int_0^1 (f''')^2$.

Proof: From the function $\phi(t)$ of Proposition 1 we take

$$\phi_2(t) = \int_{-\infty}^{t} [\int_{-\infty}^{n} \phi(u)du]dn.$$

When the conditions of this proposition hold, adding together suitably scaled translates of $\phi_2(t)$ will give a function $f_2 \in C^\infty$ with $D^2 f(t) \geq 0$ all t which satisfies

$$f_2(t_i) = y_i \qquad i = 1,2,\ldots,n.$$

If ϵ is the maximum error bound on the y_i the function

$$g_2(t) = f_2(t) + \frac{\epsilon t^2}{2} \qquad t \in [0,1]$$

meets the assumption of Theorem 1 and 4 with $F = D^2$, $m = 3$. □

Passow shows the existence of an interpolating spline, isotonic and occasionally antitonic between knot points which are data points augmented by certain others. The results of Propositions 1 and 2 establish necessary and sufficient conditions for global isotonicity and can also be applied to Passow's problem to obtain locally isotonic or antitonic splines without the artifice of extra knots.

Concluding Remarks

1) In connection with Theorem 4 it would be very useful to have necessary and sufficient conditions for a prescribed minimum spacing between knot points or at least to guarantee only a finite number of knot points.

2) Efficient algorithms for fitting isotonic splines would be valuable for converting this theory into practice. An iteration technique based on adding further linear constraints at key points could easily be feasible. The interpolating spline might be computationally more difficult than the smoothing spline when an iterative method is used.

References

[1] Ahlberg, J. H., Nilson, E. N. and Walsh, J. L. (1967), *The Theory of Splines and Their Applications* , Academic Press, New York.

[2] Attéia, M. (1968), "Fonctions (spline) definies sur un ensemble convexe", *Num. Math., 12,* 192-210.

[3] Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972), *Statistical Inference under Order Restrictions* , John Wiley and Sons, New York.

[4] Boneva, L., Kendall, D., and Stefanov, I. (1971), "Spline Transformations: Three new diagnostic aids for the data analyst", *J. Royal Statist. Soc(B), 33,* 1-70.

[5] Cogburn, R. and Davis, H. T. (1974), "Periodic Splines and Spectral Estimation", *Ann. Statist. 2,* 1108-1126.

[6] Feinberg, S. E., and Holland, P. W. (1972), "On the choice of flattening constants for estimating multinomial probabilities", *J. Multivariate Analysis, 2,* 127-134.

[7] Good, I. J. and Gaskins, R. A. (1971), "Non parametric roughness penalties for probability densities", *Biometrica, 58,* 255-277.

[8] Greville, T. N. E. (ed)(1969), *Theory and Application of Spline Functions,* Academic Press, New York.

[9] Hampel, F. R. (1974), "The influence curve and its role in robust estimation", *J. Am. Statist. Assoc., 69,* 383-393.

[10] Hocking, R. R. (1972), "Criteria for selection of a subset regression: Which one should be used?", *Technometrics, 14,* 967-970.

[11] Huber, P. J. (1964), "Robust estimation of a location paramter," *Ann. Math. Statist., 35,* 73-101.

[12] Kimeldorf, G. S. and Wahba G. (1970), "A correspondence between Bayesian Estimation on Stochastic processes and smoothing by splines," *Ann. Math. Statist., 41,* 495-502.

[13] Kimeldorf, G. S. and Wahba, G. (1971), "Some results on Tchebycheffian spline functions", *J. Math. Anal. Appl., 33,* 82-94.

[14] Laurent, P. J. (1969), "Construction of spline functions in a convex set", *Approximation with Special Emphasis on Spline Functions,* (ed. I. J. Schoenberg), 415-446, Academic Press, New York.

[15] deMontricher, G. F., Tapia, R. A. and Thompson, J. R. (1975), "Non-parametric maximum likelihood estimation of probability densities by penalty funtion methods," *Ann. Statist., 3,* 1329-1348.

[16] Mosteller, F. and Wallace, D. L. (1963), "Inference in an authorship problem," *J. Amer. Statist. Assoc., 58,* (302), 275-309.

[17] Passow, E. (1974), "Piecewise Monotone Spline Interpolation", *J. Approx. Theory, 12,* 240-241.

[18] Passow, E. and Roulier, J. A. (1974), "Monotone and Convex Spline Interpolation", Preprint Temple University.

[19] Reinsch, C. H. (1967), "Smoothing by Spline functions I", *Num. Math, 10,* 177-183.

[20] Reinsch, C. H. (1971), "Smoothing by Spline function II", *Num. Math, 16,* 451-454.

[21] Ritter, K. (1969), "Generalized Spline Interpolation and Non-linear Programming", *Approximation with Special Emphasis on Spline Functions,* (I. J. Schoenberg (ed)), 75-118, Academic Press, New York.

[22] Schoenberg, I. J. (ed) (1969), *Approximations with Special Emphasis on Spline Functions,* Academic Press, New York.

[23] Schoenberg, I. J. (1972a), "Notes on spline functions II. On the smoothing histograms", Univ. of Wisconsin M.R.C. Technical Summary Report #1222, Madison.

[24] Schoenberg, I. J. (1972b) "Splines and Histograms", Univ. of Wisconsin M.R.C. Technical Summary Report #1273, Madison.

[25] Wahba, G. (1971), "A Polynomial Algorithm for Density estimation", *Ann. Math. Statist.*, *42*, 1870-1886.

[26] Wahba, G. (1973), "Convergence Properties of the method of regularization for noisy linear operator equations", TSR No. 1132, Math. Res. Center, Univ. of Wisconsin, Madison.

[27] Wahba, G. (1974), "Smoothing Noisy data by spline functions", Tech. Report No. 380, Department of Statistics, University of Wisconsin, Madison.

[28] Wahba, G. (1975a), "Optimal Convergence Porperties of Variable Knot, Kernel and Orthogonal Series Estimates for Density Estimation", *Ann. Statist.*,*3*, 15-29.

[29] Wahba, G. (1975b), "Interpolating Spline Methods for Density Estimation I, Equi-spaced Knots", *Ann. Statist.*, *3*, 130-148.

[30] Wahba, G. and Wold, A. (1975a), "A Completely Automatic French Curve: Fitting Spline Functions by Cross Validation", *Comm. Statist.*, *4*, 1-17.

[31] Wahba, G. and Wold, A. (1975b), "Periodic Splines for Spectral Density Estimation: The Use of Cross Validation for Determining the Degree of Smoothing," *Comm. Statist.*, *4*, 125-141.

[32] Walker, A. M. (1965), "Some asymptotic results for the periodogram of a stationary time series", *J. Austral. Math. Soc.*, *5*, 107-128.

[33] Wegman, E. J. (1977), "Vector splines and the estimation of filter functions", Preprint, Manchester-Sheffield School of Probability and Statistics, Manchester, England.

[34] Wold, S. (1974), "Spline functions in data analysis", *Technometrics, 16,* 1-11.