Centrum voor Wisjunde en Informatica,
Library,
P.O. Box 4079,
1009 AB Amsterdam,
The Netherlands.

Thank you very much in advance for your cooperation.

Yours sincerely,

Ms Joke Sterringa

Library

CURVE SMOOTHING USING SPLINES

Don Barry

1. INTRODUCTION

Consider a set of data (x_i, y_i), $i = 1,2, \ldots, n$ with $0 \leqq x_1 < x_2 < \ldots < x_n \leqq 1$ and $y_i = F(x_i) + e_i$, $i = 1,2, \ldots, n$ where F is a well behaved function of x and the errors $\{e_i\}$ are independently and identically distributed each with mean zero and variance v. F is known as the regression function of Y on X and its estimation from a finite set of observations is one of the central problems in statistics.

The usual parametric approach to regression estimation assumes F to lie in span $\{\phi_j : 1 \leqq j \leqq m\}$, the set of linear combinations of the basis functions $\phi_1, \phi_2, \ldots, \phi_m$ and then estimates F by the function $\hat{F}$ in span $\{\phi_j : 1 \leqq j \leqq m\}$ which minimises the residual sum of squares given by

$$\text{Res. SS} = \sum_{i=1}^{n} (y_i - \hat{F}(x_i))^2.$$

Classical polynomial regression uses as basis functions

$$\phi_j(x) = x^{j-1}, \quad j = 1,2, \ldots, m.$$

We wish to choose $\hat{F}$ in a larger class of functions containing span $\{\phi_j : 1 \leqq j \leqq m\}$ as a subset. We minimise the Res SS plus a penalty corresponding to a measure of the distance of F from span $\{\phi_j : 1 \leqq j \leqq m\}$. For example we might choose $\hat{F} \in H^{(2)}$ to minimise

$$\sum_{i=1}^{n} (y_i - g(x_i))^2 + c\int_0^1 g''(x)^2 dx$$

where $H^{(2)} = \{g : [0,1] \to R \mid g, g' \text{ are absolutely continuous and } \int_0^1 g''(x)^2 dx < \infty\}$. The presence of the penalty imposes smoothness on the estimator. If we cannot make any smoothness assumptions regarding F, then y_i contains information about $F(x_i)$ and none about $F(x)$ for $x \neq x_i$. This makes estimation of F impossible. The constant c is chosen by the user and controls the trade-off between roughness as measured by $\int_0^1 g''(x)^2 dx$ and fidelity to the data as measured by

$$\sum_{i=1}^{n} (y_i - g(x_i))^2.$$

In what follows we describe the use of roughness penalties in more detail, examine the relationship between choice of penalty and choice of Bayesian analysis, briefly describe the large sample properties of such estimators and finally consider a method for using the data to guide the choice of c.

2. POLYNOMIAL SMOOTHING SPLINES

Consider choosing a function g to minimise

$$\sum_{i=1}^{n} (y_i - g(x_i))^2 + c\int_0^1 g^{(m)}(x)^2 dx \qquad 2.1$$

A unique solution to this minimisation problem exists in the space

$$H^{(m)} = \{g : [0,1] \to \mid g, g', \ldots, g^{(m-1)} \text{ are absolutely continuous and } \int_0^1 g^{(m)}(x)^2 dx < \infty\}$$

and we choose our regression estimate $\hat{F}$ to be that solution.

Schoenberg (1964) has shown that $\hat{F}$ lies in the linear space S_m of polynomial splines of degree 2m-1. S_m consists of all functions $g : [0,1] \to R$ such that

(i) g is a polynomial of degree 2m-1 in each interval $[x_i, x_{i+1}]$, $i = 1,2, \ldots, n-1$.

(ii) g is a polynomial of degree m-1 in the intervals $[0, x_1]$, $[x_n, 1]$.

(iii) g is continuously differentiable up to order 2m-2.

It can be shown that given $a_1, a_2, \ldots, a_n$ there exists one and only one function $s \in S_m$ such that

$$s(x_i) = a_i \qquad i = 1,2, \ldots, n \qquad 2.2$$

Let σ_i denote the only element of S_m satisfying

$$\sigma_i(x_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

It is easy to see that $\{\sigma_1, \sigma_2, \ldots, \sigma_n\}$ is a basis for S_m and using this basis the element s of 2.2 is given by

$$s(x) = \sum_{i=1}^{n} a_i \sigma_i(x).$$

See Rice (1969) Chapter 10 for details.

Using this basis we can rewrite the minimisation problem 2.1 as: choose $a_1, a_2, \ldots, a_n$ to minimise

$$\sum_{i=1}^{n} (y_i - \sum_{r=1}^{n} a_r \sigma_r(x_i))^2 + c\Sigma\Sigma a_r a_s w_{rs} \qquad 2.3$$

where $w_{rs} = \int_0^1 \sigma_r^{(m)}(x)\sigma_s^{(m)}(x)dx$. This is now a finite problem and the optimal values for $a_1, a_2, \ldots, a_n$ are precisely the values $\hat{F}(x_1), \ldots, \hat{F}(x_n)$. We can write 2.3 as

$$(\underline{y} - \underline{\hat{F}})^T(\underline{y} - \underline{\hat{F}}) + c\underline{\hat{F}}^T\Omega\underline{\hat{F}}$$

where

$$\underline{y} = (y_1, y_2, \ldots, y_n)^T,$$

$$\underline{\hat{F}} = (\hat{F}(x_1), \hat{F}(x_2), \ldots, \hat{F}(x_n))^T,$$

$$\Omega = (w_{rs}), \quad r,s = 1, 2, \ldots, n.$$

The minimising value for $\hat{F}$ is $\underline{\hat{F}} = A\underline{y}$, where $A = (I + c\Omega)^{-1}$.

The value of c must be chosen by the user and is of vital importance. For $c = \infty$ we must have $\int_0^1 F^{(m)}(x)^2 dx = 0$ which, together with the absolute continuity requirements, implies that F must be a polynomial of degree m-1, indeed the usual least squares polynomial of degree m-1. For c = 0 we can make 2.1 equal to zero by choosing $\hat{F} \in S_m$ satisfying

$$\hat{F}(x_i) = y_i$$

Intermediate values for c involve a trade-off between the smoothness of the estimate as measured by $\int_0^1 F^{(m)}(x)^2 dx$ and the fidelity to the data as measured by $\Sigma(y_i - \hat{F}(x_i))^2$.

Figure 1 shows some data generated by adding normal errors to the function

$$F(x) = K_1 x^{10}(1 - x)^6 + K_2 x^3 (1 - x)^{10}$$

where K_1, K_2 are positive constants. The 50 x-values are equally spaced in [0,1]. Figure 2 shows the fitted curve obtained using the m = 1 roughness penalty and a value for c chosen by cross validation (see Section 5). The true curve is shown in Figure 3 and indicates that the estimate in this case behaves very well.

Figures 4-7 refer to some real data concerning computer repair times. Here X is the number of units to be repaired and Y is the length of the call in minutes. The figures show the fitted curve obtained using m = 2 as the value for c increases from 0 to ∞. It can clearly be seen how increasing c causes the estimate to become smoother and to follow the data less closely.

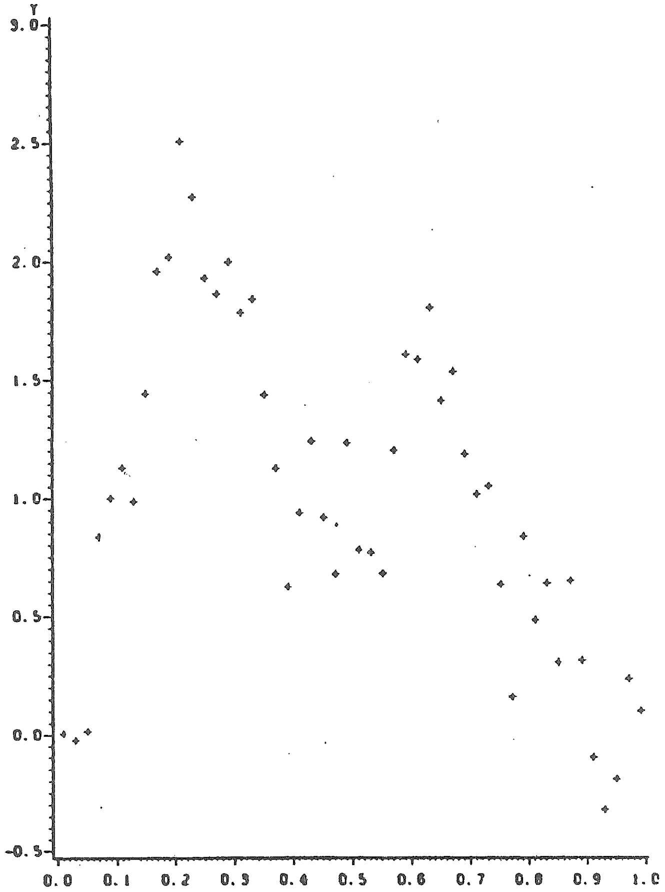

FIGURE 1: Raw Data

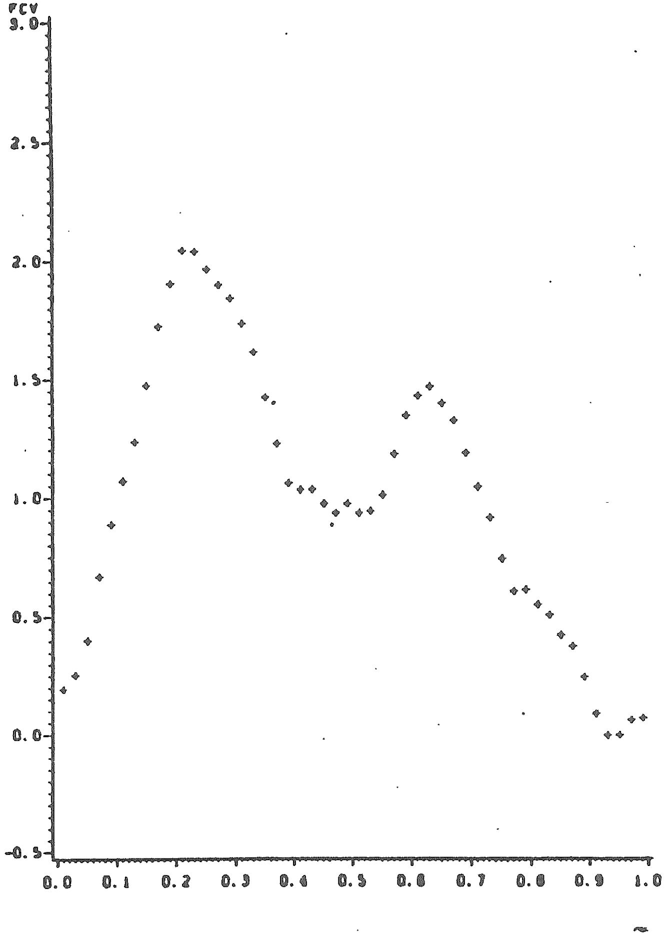

FIGURE 2: Fitted Values

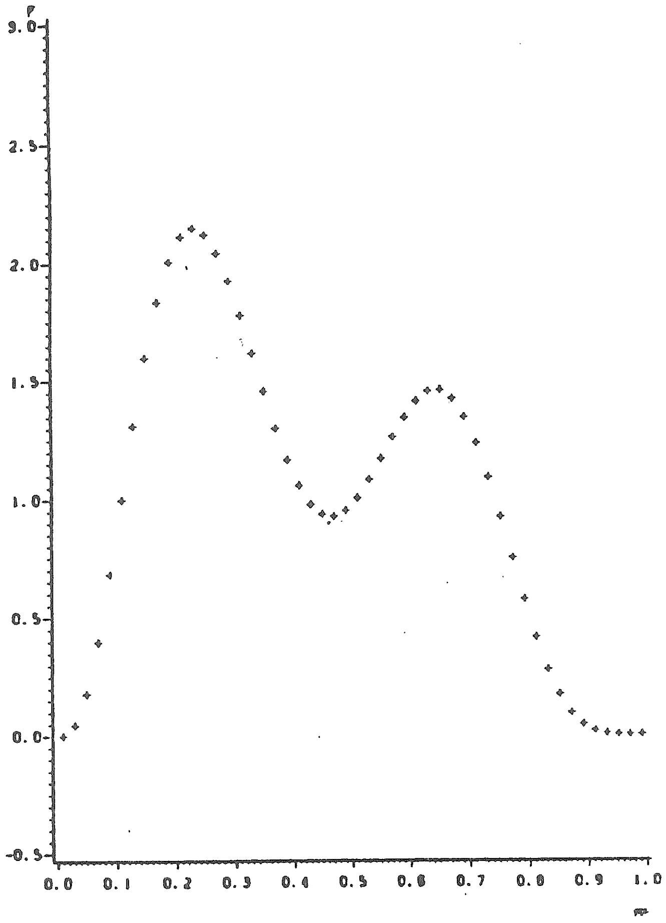

FIGURE 3: $F(x) = k_1 x^{10}(1-x)^6 + k_2 x^3 (1-x)^{10}$

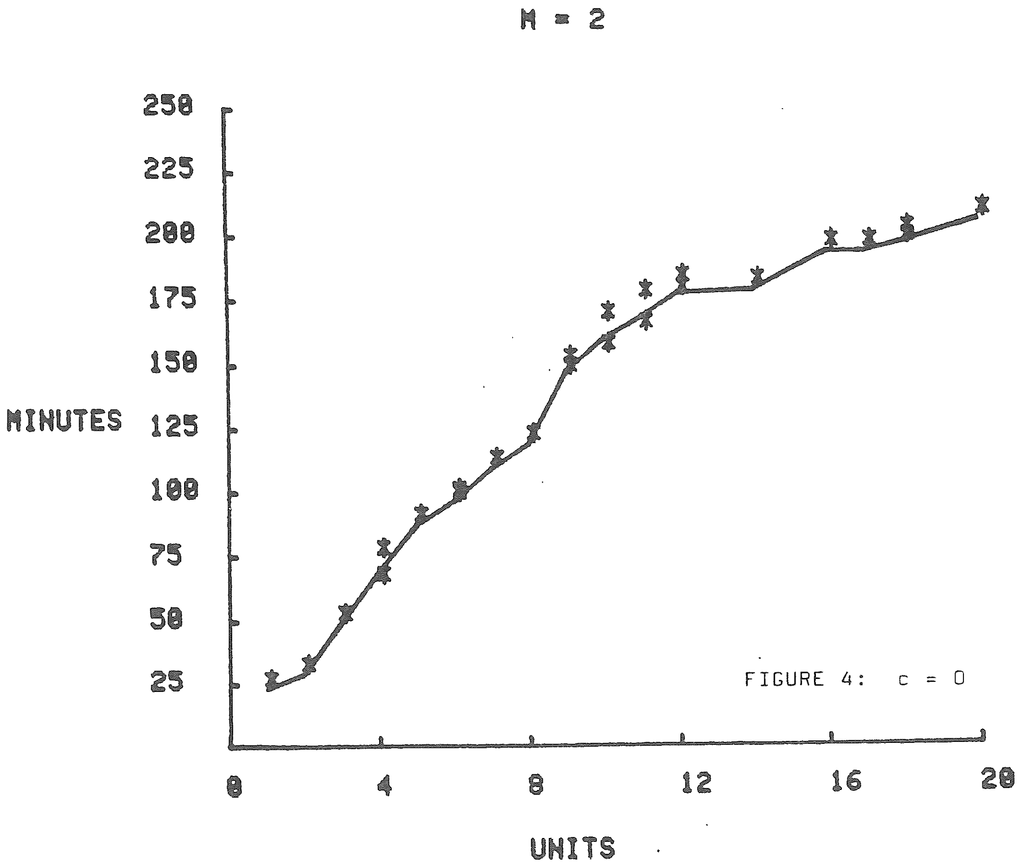

FIGURE 4: c = 0

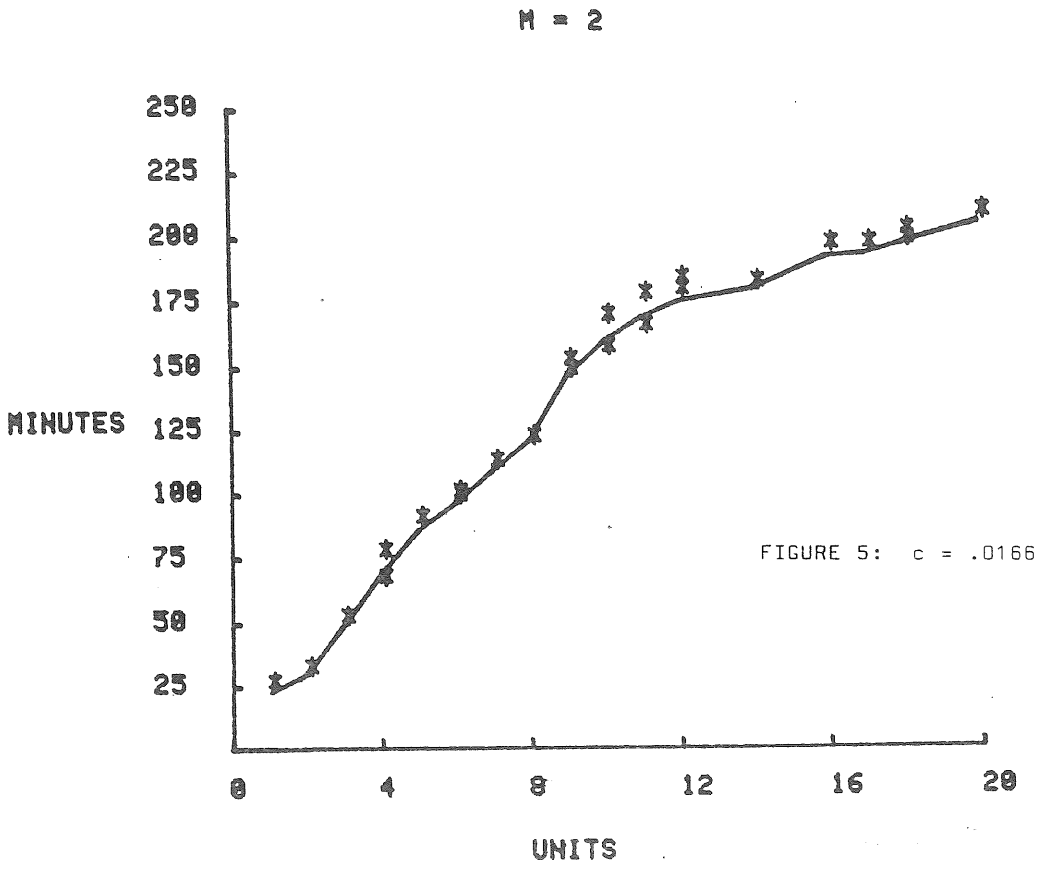

FIGURE 5: c = .0166

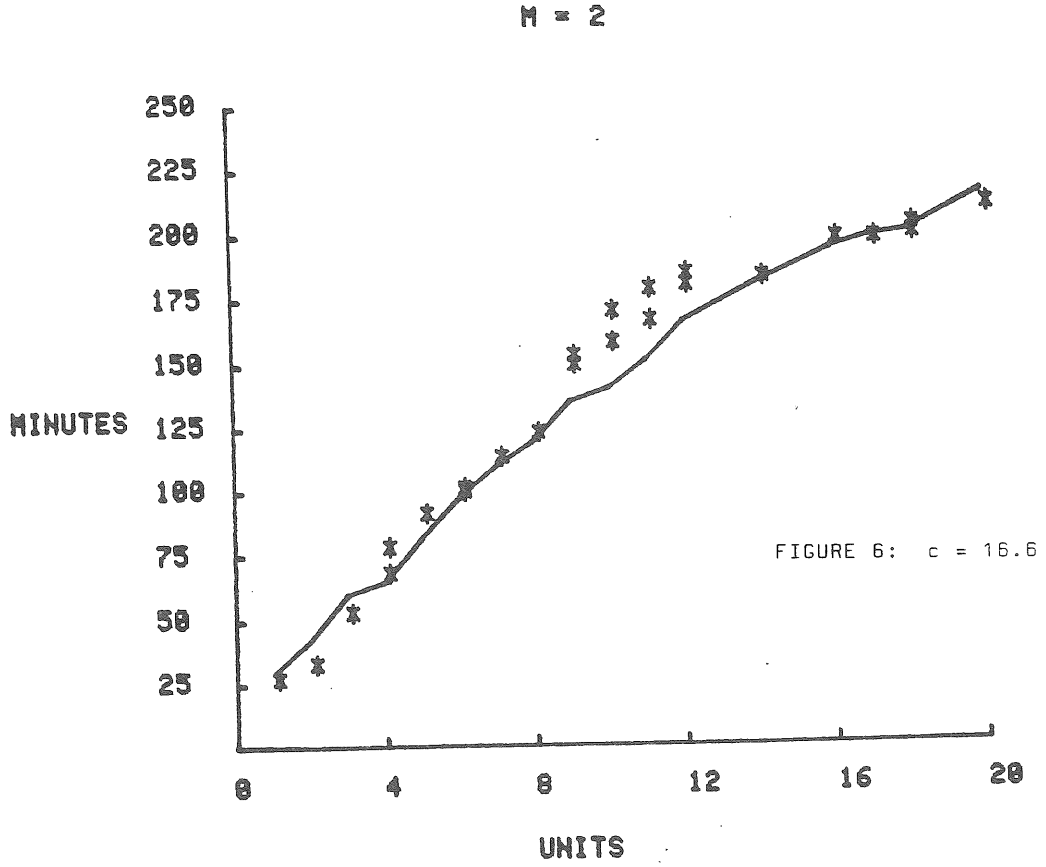

FIGURE 6: c = 16.6

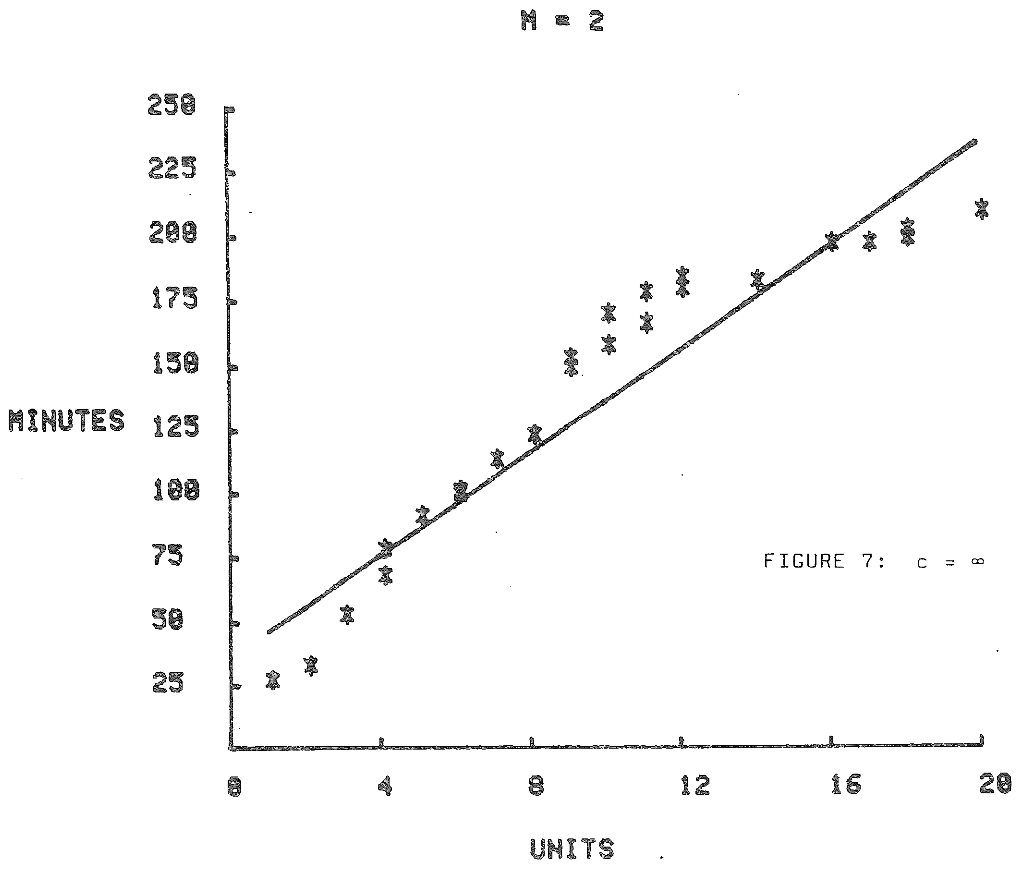

FIGURE 7: c = ∞

3. THE BAYESIAN CONNECTION

Wahba (1978) proved the following theorem:

Theorem

Assume that $y_i = F(x_i) + e_i$, $i = 1,2, \ldots, n$ where $\{e_i\}$ are independent and identically distributed as $N(0,v)$. Let the prior distribution of $F(x)$, $x \in [0,1]$ be that of the stochastic process

$$\sum_{j=1}^{m} \theta_j x^{j-1} + v_1^{\frac{1}{2}} Z(x)$$

where $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_m) \sim N[0, v_0 I]$, $v_1 > 0$ is fixed and $Z(x)$ is an m-fold integrated Wiener process

$$Z(x) = \int_0^1 \frac{(x-u)^{m-1}}{(m-1)!} \, dw(u)$$

where W is a Brownian motion (see Shepp, 1966). Then $\hat{F}(x\ ;\ c)$ the minimiser of 2.1 has the property that

$$\hat{F}(x\ ;\ c) = \lim_{v_0 \to \infty} E\{F(x) \mid y_1, y_2, \ldots, y_n\}$$

with $v_1 = v/c$ where E is expectation over the posterior distribution of $F(x)$ generated by the above probability model.

Thus the choices of m and c are closely related to the choice of Bayesian prior.

4. ASYMPTOTIC PROPERTIES

Let $\hat{F}(x\ ;\ c)$ be the estimator corresponding to a particular choice of c. Define

$$R(c) = \sum_{i=1}^{n} (\hat{F}(x_i\ ;\ c) - F(x_i))^2$$

i.e. the sum of squared errors if c is used as a smoothing constant. R(c) is a random variable since different y-values produce a different function $\hat{F}(x\ ;\ c)$ and hence a different

value for R(c). The following theorem shows that if we allow c to increase with n, but not too quickly then $ER(c) \to 0$ at a fast rate, where expectation is with respect to the normal distribution on the errors.

Theorem

$$ER(c) \leqq c \int_0^1 F^{(m)}(x)^2 dx + K(n/c)^{\frac{1}{2m}}$$

where

$$K = v[n.\max(x_{i+1} - x_i)]^{\frac{1}{2m}} \int_0^\infty \frac{dx}{(1+x^{2m})^2}$$

and v is the error variance.

Proof

See Wahba (1978).

Corollary

For $c = O(n^{\frac{1}{2m+1}})$ we have $ER(c) = O(n^{\frac{1}{2m+1}})$.

This is to be compared with $ER(0) = O(n)$ (since $E(y_i - F(x_i))^2 = v$), and shows clearly the benefit of smoothing.

NOTE:

If $\int_0^1 F^{(m)}(x)^2 dx = 0$ then $ER(\infty) = O(1)$.

5. THE CHOICE OF c

Many attempts have been made to use the data to guide the choice of c. We shall describe one such attempt known as cross-validation. The idea underlying cross-validation is that a value of c good for the whole data set should also be good if a single point is removed and that performance can be judged by seeing how well the dropped point is estimated using c as smoothing parameter on the remaining n-1 points. Leaving out each point in turn we would choose c to minimise

$$V_0(c) = \sum_{k=1}^{n} (y_k - \hat{F}^{[k]}(x_k \, ; \, c))^2$$

where $\hat{F}^{[k]}(x \, ; \, c)$ is the estimate of F based on all the data except the kth point. Craven and Wahba (1979) propose that instead of minimising $V_0(c)$ a weighted sum of the form

$$V(c) = \sum_{k=1}^{n} (y_k - \hat{F}^{[k]}(x_k \, ; \, c))^2 w_k(c)$$

should be used. They suggest

$$w_k(c) = \left[\frac{1 - a_{kk}(c)}{\frac{1}{n}\text{Trace } (I-A(c))} \right]^2$$

where A(c) is the matrix such that

$$\underline{\hat{F}} = A(c)\underline{y}$$

and $a_{kk}(c)$ is the kth diagonal element of A(c). $a_{kk}(c)$ is the weight given y_k when estimating $F(x_k)$. If it is close to one, the point x_k has few close neighbours and so the error in estimating $F(x_k)$ is unavoidably large and should be down-weighted in measuring the worth of a particular choice for c. In their 1979 paper Craven and Wahba support their advocacy of cross-validation both by theoretical arguments and by means of a simulation study.

REFERENCES

CRAVEN, P. and WAHBA, G. (1979)
"Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalised Cross-Validation", *Numer. Math.*, 31 (377-403).

RICE, J.R. (1969)
'The Approximation of Functions', Vol. 2 (Addison-Wesley).

SCHOENBERG, I.J. (1964)
"Spline Functions and the Problem of Graduation", *Proc. Nat. Acad. Sci. (USA)*, 52 (947-950).

SHEPP, L. (1966)
"Radon-Nikodym Derivatives of Gaussian Measures", *Ann. Math. Stat.*, 37 (321-354).

WAHBA, G. (1978).
"Improper Priors, Spline Smoothing and the Problem Guarding against Model Errors in Regression", *Jour. Royal Stat. Soc.*, *'B'*, 40 (364-372).

Department of Statistics,
University College,
Cork.