1. INTRODUCTION

Consider a set of data \( (x_i, y_i), i = 1, 2, \ldots, n \) with \( 0 < x_1 < x_2 < \ldots < x_n \leq 1 \) 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 \( \sigma^2 \). \( f \) 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 \( \text{span} \{ \phi_j : 1 \leq j \leq m \} \), 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 \( \text{span} \{ \phi_j : 1 \leq j \leq m \} \) 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 \( \text{span} \{ \phi_j : 1 \leq j \leq m \} \) as a subset. We minimise the Res SS plus a penalty corresponding to a measure of the distance of \( f \) from \( \text{span} \{ \phi_j : 1 \leq j \leq m \} \). 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 \( H^{(2)} = \{ g : [0,1] \to \mathbb{R} | g, g' \text{ are absolutely continuous and } \int_{0}^{1} g''(x)^2 dx < \infty \} \). The presence of the penalty imposes smoothness on the estimator. If we cannot make any smoothness ass-
umptions 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 $\int_{0}^{1} y_1(x) dx$ and fidelity to the data as measured by

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

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

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - g(x_i))^2 + c \int_{0}^{1} g^{(m)}(x)^2 dx \quad 2.1$$

A unique solution to this minimisation problem exists in the space

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

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

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

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

Let $\sigma_i$ 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

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


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

$$\sum_{i=1}^{n} (y_i - \sum_{r=1}^{n} a_r \sigma_r(x_i))^2 + \text{constant}$$

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

$$(\mathbf{y} - \hat{\mathbf{f}})^T (\mathbf{y} - \hat{\mathbf{f}}) + c \mathbf{g}^T \Pi \mathbf{g}$$

where

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

$$\hat{\mathbf{f}} = (\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_n))^T,$$

$$\Pi = (w_{rs}), \quad r,s = 1, 2, \ldots, n.$$
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 \( \sum (y_I - \hat{F}(x_I))^2 \).

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

\[ F(x) = k_1 x^{10}(1-x)^8 + k_2 x^3(1-x)^4 \]

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 \( \infty \). 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: \( F(x) = k_1 x^3 (1-x) + k_2 x^3 (1-x)^{1.6} \)
3. THE BAYESIAN CONNECTION

Wahba (1978) proved the following theorem:

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

\[
Z(x) = \sum_{j=1}^{m} (x-u_{j})^{m-1} \frac{d\omega(u)}{(m-1)!} + \nu \frac{1}{2} \int_{0}^{1} Z(x) \mathrm{d}x
\]

where \( \Omega = (\theta_1, \theta_2, \ldots, \theta_m) \sim N(\theta, \nu I), \nu > 0 \) is fixed and \( Z(x) \) is an \( m \)-fold integrated Wiener process.

Thus the choice of \( m \) and \( \nu \) 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

\[
R(c) = \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - \hat{f}(x_i; c))^2
\]

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 \( ER(c) = 0 \) at a fast rate, where expectation is with respect to the normal distribution on the errors.

Theorem
\[
ER(c) \leq c \int_{0}^{1} \frac{f(m)(x)^2 \mathrm{d}x + K(n/c)^{2m/n}}{\nu}
\]

where

\[
K = \nu \max \left( e^{(x_{i+1} - x_i) \nu}, e^{(1+x^{2m})} \right)^{1/2} = \int_{0}^{m} \frac{\mathrm{d}x}{\nu (1+x^{2m})^{1/2}}
\]

and \( \nu \) is the error variance.

Proof
See Wahba (1978).

Corollary

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

This is to be compared with \( ER(3) = O(n) \) (since \( E(y_i - f(x_i))^2 = \nu) \), and shows clearly the benefit of smoothing.

NOTE:

If \( \int_{0}^{1} f(m)(x)^2 \mathrm{d}x = 0 \) then \( ER(c) = O(1) \).

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

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

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

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

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

$$\hat{f} = A(c)x$$

and $a_{kk}(c)$ is the $k$th 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)

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