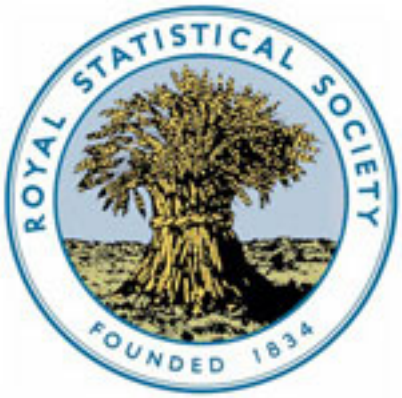

J. R. Statist. Soc. B (1985)
47, *No.* 1, *pp.* 1–52

Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting

By B. W. SILVERMAN

University of Bath

[*Read before the* Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, October 10th, 1984*, Professor J. B. Copas *in the Chair*]

SUMMARY

Non-parametric regression using cubic splines is an attractive, flexible and widely-applicable approach to curve estimation. Although the basic idea was formulated many years ago, the method is not as widely known or adopted as perhaps it should be. The topics and examples discussed in this paper are intended to promote the understanding and extend the practicability of the spline smoothing methodology. Particular subjects covered include the basic principles of the method; the relation with moving average and other smoothing methods; the automatic choice of the amount of smoothing; and the use of residuals for diagnostic checking and model adaptation. The question of providing inference regions for curves—and for relevant properties of curves—is approached via a finite-dimensional Bayesian formulation.

Keywords: ROUGHNESS PENALTY; SMOOTHING; WEIGHT FUNCTION; VARIABLE KERNEL; CROSS-VALIDATION; AUTOMATIC SMOOTHING; RESIDUALS; REGRESSION DIAGNOSTICS; LOCAL REWEIGHTING; CHANGE POINT; MODEL CHOICE; BAYESIAN INFERENCE; EMPIRICAL BAYES; B-SPLINES; GROWTH CURVES; FUNCTIONALS OF CURVES; ROBUST SMOOTHING; GENERALIZED SMOOTHING; SURFACE ESTIMATION

1. INTRODUCTION

Consider the regression problem where we have *observations* Y_i at *design points* t_i, $i = 1, \ldots, n$ and the observations are assumed to satisfy

$$Y_i = g(t_i) + \epsilon_i. \tag{1.1}$$

In this paper the non-parametric estimation of the function g will be discussed. It will be assumed that the design points satisfy $t_1 \leqslant t_2 \leqslant \ldots \leqslant t_n$ and that the errors ϵ_i are uncorrelated with zero mean. At first the variances of the ϵ_i will be assumed to be equal, but later this assumption will be relaxed.

1.1. *Motivation*

Before embarking on any technical details, it is important to consider the reasons why we might be interested in the estimation of the curve g, or, indeed, in any regression technique. Regression, of whatever kind, has two main purposes. Firstly, it provides a way of exploring and presenting the relationship between the design variable and the response variable; secondly, it gives predictions of observations yet to be made. A method for estimating a curve g will also be used for a third purpose, to give estimates of interesting properties of g. For example, in the study of growth curves the maximum rate of growth is an important quantity and so the maximum of the derivative of g will be of interest. This example will be discussed further in Section 7 below.

Especially for the first and third of these purposes, a non-parametric method of estimation is

Present address: School of Mathematics, University of Bath, Bath, BA2 7AY.

 0035–9246/85/47001 $2.00

desirable, because it does not force the model into a rigidly defined class. An initial non-parametric estimate may well suggest a suitable parametric model (such as linear regression) but nevertheless will give the data more of a chance to speak for themselves in choosing the model to be fitted. We shall see, in Section 5.2 below, an example where a non-parametric regression curve is most helpful in suggesting an appropriate parametric model.

The non-parametric regression method described and developed in this paper is the *spline smoothing* approach. The basic idea dates back at least to Whittaker (1923) and the method has been much studied in the past 20 years. Despite this attention by specialists, the method is not as widely known or used among the wider statistical and scientific community as perhaps it should be.

1.2. *Material Covered*

Various aspects of the spline smoothing method are discussed below. It will be seen that spline smoothing provides a natural and flexible approach to curve estimation, which copes well whether or not the design points are regularly spaced. As in almost all non-parametric smoothing methods, there is a smoothing parameter which determines how much the data are smoothed to produce the estimate. The automatic choice of this smoothing parameter will be discussed; the method described is computationally cheap and statistically efficient.

Much of the recent work in regression analysis has been concerned with the use of residuals for model checking. We shall see how some of this work can be adapted to non-parametric regression. A feature frequently highlighted by residual plots is inhomogeneity of the error variance, suggesting the appropriateness of a weighted model, where observations are weighted by the reciprocals of their variances. A procedure for estimating the weights using local variance estimates will be described and illustrated by an example.

Most non-parametric smoothing methods provide only a single estimate for the curve, without giving any indication of the likely estimation accuracy. Spline smoothing can be viewed in a Bayesian context as an inference problem in a high but finite dimensional space. This Bayesian formalism, described in the latter part of the paper, allows inference regions to be found directly for values of the curve itself and for quantities which depend linearly on the function g. A fast method for simulating from the posterior distribution of g will be developed; this makes it straightforward to obtain, by Monte Carlo methods, point and interval estimates for quantities, such as the maximum gradient, which do not depend linearly on g.

The various methods developed in the paper are illustrated in practice by their use on three data sets drawn from different fields of application. All the techniques described have been implemented in Fortran on a mainframe computer; details of the programs are available from the author.

The final section of the paper discusses extensions of the various techniques and gives some additional references to related work. A detailed bibliographic review of previous work on spline smoothing is rendered unnecessary by the excellent survey by Wegman and Wright (1983). Among the many developments since the formulation of spline smoothing in its modern form by Schoenberg (1964) and Reinsch (1967), special mention must be made of the substantial contribution made by Wahba in a series of papers in recent years; a selection of these is given in the references below.

2. THE BASIC IDEA

The most widely used approach to curve fitting is, of course, least squares. If we place no restrictions at all on the curve g then we can reduce the residual sum of squares $\Sigma\{Y_i - g(t_i)\}^2$ to zero by choosing g to be any curve which actually interpolates the data (provided the t_i are all distinct). Such an interpolant would usually be rejected by the statistician on the grounds that its rapid fluctuations were implausible. The most commonly used device for avoiding such "implausible" estimates is to restrict attention to curves g which fall in some parametric class. Another approach is to quantify the competition between the two conflicting aims in curve

estimation, which are to produce a good fit to the data but to avoid too much rapid local variation.

A measure of the rapid local variation of a curve can be given by a *roughness penalty* such as the integrated squared second derivative. Various roughness penalties have been suggested and used (see Good and Gaskins, 1971, and Boneva *et al.*, 1972) but $\int(g'')^2$ is most convenient for our purpose. Using this measure, define the modified sum of squares

$$S(g) = \Sigma\{Y_i - g(t_i)\}^2 + \alpha \int g''(x)^2 dx; \tag{2.1}$$

the *smoothing parameter* α represents the rate of exchange between residual error and local variation. Minimizing $S(g)$ over the class of all (twice-differentiable) functions g will yield an estimate $\hat{g}$ which, for the given value of α, gives the best compromise between smoothness and goodness of fit.

It can be shown (see Reinsch, 1967) that the curve $\hat{g}$ has the following properties:

(i) it is a cubic polynomial in each interval (t_i, t_{i+1});
(ii) at the design points t_i, the curve and its first two derivatives are continuous, but there may be a discontinuity in the third derivative;
(iii) in each of the ranges $(-\infty, t_1)$ and (t_n, ∞) the second derivative is zero, so that $\hat{g}$ is linear outside the range of the data. (2.2)

Any curve which satisfies (i) and (ii) is called a *cubic spline* with *knots* t_i. It should be stressed that the properties (2.2) are not imposed on the estimate, but arise automatically from the choice of roughness penalty $\int(g'')^2$.

One of the useful consequences of the properties of $\hat{g}$ is computational. To find $\hat{g}$ explicitly we need to find the four coefficients which give the polynomial form of $\hat{g}$ in each interval. It turns out that all these coefficients can, essentially, be found by solving a band-limited linear system of size n. Stable and fast numerical algorithms for solving such systems are available: De Boor (1978, Chapter 14) gives a description of the way that $\hat{g}$ can be found and a Fortran implementation. The present author and collaborators have adapted De Boor's programs to yield a method which finds $\hat{g}$ using $35n$ multiplications/divisions for the first value of the smoothing parameter and $25n$ thereafter. Thus the computational burden involved in finding $\hat{g}$ is very small and goes up linearly as the number of data points.

3. WHAT IS SPLINE SMOOTHING ACTUALLY DOING TO THE DATA?

A major conceptual problem with curve estimates like the spline smoother is that they are defined implicitly as the solution to a minimization problem rather than as an explicit formula involving the data values. This difficulty can be resolved, at least approximately, by considering how the estimate behaves on large data sets. The approximation described in this section has a dual purpose: not only to give a deeper understanding of the spline smoothing method but also to provide an ingredient of much of the practical methodology given later in the paper. Exact definitions, statements and proofs of results in this section are given in Silverman (1984a).

It can be shown from the quadratic nature of (2.1) (cf. equation (2.2) of Wahba, 1975) that $\hat{g}$ is linear in the observations Y_i, in the sense that there exists a *weight function* $G(s, t)$ such that

$$\hat{g}(s) = n^{-1} \sum_{i=1}^{n} Y_i G(s, t_i). \tag{3.1}$$

The weight function depends on the design points $t_1, \ldots, t_n$ and also on the smoothing parameter α. We can obtain the asymptotic form of the weight function, and hence an approximate explicit form of the estimate. Suppose that n is large and that the design points have local density $f(t)$, in that the proportion of t_i in an interval of length dt near t is approximately $f(t)dt$.

Provided s is not too near the edge of the interval on which the data lie, and α is not too big or too small, it is the case that, for large n,

$$G(s, t) \doteq \frac{1}{f(t)} \frac{1}{h(t)} \kappa\left(\frac{s-t}{h(t)}\right) \tag{3.2}$$

where the *kernel function* κ is given by

$$\kappa(u) = \tfrac{1}{2} \exp(-|u|/\sqrt{2}) \sin(|u|/\sqrt{2} + \pi/4) \tag{3.3}$$

and the *local bandwidth* $h(t)$ satisfies

$$h(t) = \alpha^{\frac{1}{4}} n^{-\frac{1}{4}} f(t)^{-\frac{1}{4}}. \tag{3.4}$$

A graph of κ is given in Fig. 1. The basic message of these formulae is that the spline smoother is approximately a convolution (or weighted moving average) smoothing method. However, in general, the data are not convolved with a fixed width kernel function, but the scaling parameter h varies across the sample. Several specific conclusions can be drawn.

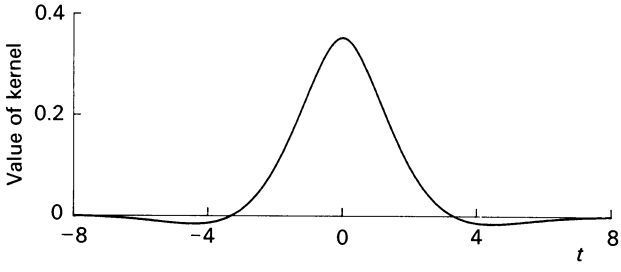

Fig. 1. The effective kernel function κ.

(i) The form of κ implies that the observation at t_i only has an influence on nearby parts of the curve $\hat{g}$. This influence dies away exponentially—a favourable contrast with the behaviour of some other curve fitting methods such as polynomial regression.
(ii) Altering the smoothing parameter α alters the amount of smoothing applied generally. However, it is important to note the one-quarter power dependence of h on α, consequent on the fact that α has dimensions of the cube of length. Thus we should not be surprised to encounter a large variation in the appropriate value of α in different problems, particularly if the scales of the design variables are different.
(iii) The dependence of the local bandwidth on the density f is intermediate between fixed kernel smoothing (no dependence on f) and smoothing based on an average of a fixed number of neighbouring values (effective local bandwidth proportional to $1/f$). Theoretical considerations (see Silverman, 1984a) suggest that such intermediate behaviour is desirable because, in certain senses, moving from fixed kernel to nearest neighbour methods over-compensates for effects caused by variability in density of design points. Under suitable assumptions the ideal local bandwidth would be proportional to $f^{-0.2}$. Given the asymptotic nature of all the arguments used we should conclude only that h proportional to a low negative power of f is appropriate; the dependence in (3.4) shows that the spline smoother adapts automatically to non-uniform data in an excellent way.

Some exact calculations of the weight function G of (3.1) reported in Silverman (1984a) show that the approximate formula (3.2) is very accurate for moderate n except in the extreme tails of the design sample. That paper also contains details of a boundary correction to (3.2) that improves the approximation near the edge of the data.

4. CHOOSING THE SMOOTHING PARAMETER

For many practical purposes, it is probably sufficient to choose the smoothing parameter α subjectively, by plotting out a few curves and choosing the one which "looks best". From the exploratory point of view this exercise is beneficial in that it will draw attention to interesting

features that only show up at certain values of the smoothing parameter. However, there are many reasons why an automatic method of choice is useful. The inexperienced user will doubtless feel happier if the method is fully automatic. An automatic choice can in any case be used as a starting point for subsequent subjective adjustment. Scientists reporting or comparing their results will want to make reference to a standardized method. If the smoothing method is to be used routinely on a large number of data sets or as part of a larger procedure then an automatic method is essential. Against these arguments should, of course, be set the caveat that completely automating statistical methods encourages the user to use methods blindly and not to give enough consideration to prior assumptions (whether or not from a Bayesian point of view). For this reason I prefer to use the word *automatic* rather than *objective* for methods that do not require explicit specification of control parameters. Of course, all these remarks apply equally to the closely-related problem of determining how many parameters to fit when using *parametric* models.

4.1. *Cross-validation and Related Criteria*

Several methods have been proposed for choosing the smoothing parameter. Probably the most attractive class of such methods is *cross-validation* which is popular, generally, for choosing the complexity of statistical models. See, for example, Stone (1974).

The basic principle of cross-validation is to leave the data points out one at a time and to choose that value of α under which the missing data points are best predicted by the remainder of the data. To be precise, let g_α^{-i} be the smoothing spline calculated from all the data pairs except (t_i, Y_i), using the value α for the smoothing parameter. The cross-validation choice of α is then the value of α which minimizes the cross-validation score

$$\text{XVSC}(\alpha) = n^{-1} \Sigma \{Y_i - g_\alpha^{-i}(t_i)\}^2. \tag{4.1}$$

This is identical to the PRESS criterion for model selection in regression generally; see Cook and Weisberg (1982, Section 2.2.3). Define a matrix $A(\alpha)$ by

$$A_{ij}(\alpha) = n^{-1} G(t_i, t_j) \tag{4.2}$$

where G is the weight function of (3.1) above.

A standard argument in regression theory, given by Cook and Weisberg (1982) and for this special case by Craven and Wahba (1979), shows that (4.1) has the easier computational form

$$\text{XVSC}(\alpha) = n^{-1} \sum_{i=1}^{n} \frac{\{Y_i - \hat{g}(t_i)\}^2}{\{1 - A_{ii}(\alpha)\}^2}. \tag{4.3}$$

Craven and Wahba also suggest the use of a related criterion, called *generalized cross-validation*, obtained from (4.3) by replacing $A_{ii}(\alpha)$ by its average value, $n^{-1}\text{tr}\, A(\alpha)$. This gives the score

$$\text{GXVSC}(\alpha) = n^{-1} \text{RSS}(\alpha)/\{1 - n^{-1} \text{tr}\, A(\alpha)\}^2, \tag{4.4}$$

where $\text{RSS}(\alpha)$ is the residual sum of squares $\Sigma \{Y_i - \hat{g}(t_i)\}^2$. In their paper Craven and Wahba (1979) also give theoretical arguments to show that generalized cross-validation should, asymptotically, choose the best possible value of α in the sense of minimizing the average squared error at the design points. This predicted good performance is borne out by published practical examples in various papers by Wahba and co-workers.

To use the formula (4.4) one still needs to find the trace of $A(\alpha)$; unless a suitable numerical device or approximation is used this can be quite a burden computationally.

It is possible to derive such an approximation, which can be calculated extremely quickly. Let $\hat{f}(t)$ be an estimate of the local density of the design points t_i, calculated using the fast algorithm of Silverman (1982a), on a range $[a, b]$ just containing the design points; for definiteness set $a = t_1 - \frac{1}{2} n^{-1}(t_n - t_1)$ and $b = t_n + \frac{1}{2} n^{-1}(t_n - t_1)$. Let

$$c_0 = \pi^4 n^{-1} \left\{ \int_a^b \hat{f}(t)^{1/4}\, dt \right\}^{-4} . \tag{4.5}$$

The constant c_0 is not needed to enormous accuracy and is calculated once only for each design set. It can then be shown that

$$\operatorname{tr} A(\alpha) \doteq 2 + \sum_{i=3}^{n} \{1 + c_0\, \alpha (i-1.5)^4\}^{-1} . \tag{4.6}$$

Substituting (4.6) into (4.4) gives a score function called the *asymptotic generalized cross-validation* (AGXV) *score* which can then be minimized to give an automatic choice of smoothing parameter. Full details of the mathematical justification of the approximation, and further computational remarks, are given in Silverman (1984b).

It requires trivial additional computing time to find the AGXV score once the smoothing spline $\hat{g}$ has been obtained. Minimizing the score to within reasonable accuracy takes, in practice, consideration of about ten values of α. The time taken by the entire procedure goes up roughly linearly as the number of data points and takes under 1 second for 50 data points on a Honeywell Multics mainframe machine.

To see how the method behaves in practice, a simulation study was carried out. Full details are reported in Silverman (1984b). The practical performance of AGXV turns out to be even better than that of generalized cross-validation in that, for non-uniformly spaced data, the proportion of cases giving a bad value of the smoothing parameter is substantially reduced.

5. WEIGHTED OBSERVATIONS AND REGRESSION DIAGNOSTICS

It is often the case that we wish to consider weighted observations, where the sum of squared deviations in (2.1) is replaced by a weighted sum of squares to give a modified weighted sum of squares

$$S_W(g) = \Sigma\, w_i \{ Y_i - g(t_i) \}^2 + \alpha \int g''(x)^2 dx \tag{5.1}$$

for a given sequence of weights w_i. The minimizer of (5.1) will again satisfy the conditions (2.2) and can again be found using De Boor's (1978) algorithms. In this section, various aspects of the weighted smoothing problem will be considered.

5.1. *Choosing the Smoothing Parameter*

The arguments about cross-validation can be extended to the weighted case. It is natural to use a weighted cross-validation score

$$\text{XVSC}\,(\alpha) = n^{-1} \Sigma\, w_i \{ Y_i - g_\alpha^{-i}(t_i) \}^2$$

in place of (4.1). The same arguments as in Section 4 then give a generalized cross-validation score as in (4.4) but with the residual sum of squares replaced by a corresponding weighted sum of squares $\Sigma\, w_i \{ Y_i - \hat{g}(t_i) \}^2$. The approximation (4.6) to $\operatorname{tr} A(\alpha)$ carries through as above, except that the density estimate $\hat{f}$ used in (4.5) is constructed from the weighted design points, that is to say

$$\hat{f}(t) = \sum_i w_i h^{-1}\, \phi\{ h^{-1}(t - X_i) \} / \sum_i w_i, \tag{5.2}$$

where ϕ is the standard normal density function and h is chosen using the analogous formula to (2) of Silverman (1982a). The algorithm of that paper is easily adapted to calculate (5.2). As in the non-weighted case, the overhead required to find the AGXV score is trivial once $\hat{g}$ has been found, and the time taken by the entire procedure is just as before.

5.2. *Regression Diagnostics and Estimation of the Variance*

There has been a great deal of attention paid recently to diagnostic plots for checking the assumptions in regression. An excellent treatment is given by Cook and Weisberg (1982, Chapter 2). Many of the ideas for diagnostics in linear regression can be carried over to the spline smoothing technique, though there are some important differences. The usual, and well-established, methods of plotting residuals directly, against fitted values or against the design points, carry over without modification.

The more sophisticated techniques discussed by Cook and Weisberg mostly require knowledge of the so-called "hat matrix" which maps the vector of observations Y_i into the vector of predicted values $\hat{g}(t_i)$. The hat matrix is then precisely the matrix $A(\alpha)$ considered in Section 4 above, since it is the case that

$$\hat{g}(t_i) = \sum_{j=1}^{n} A_{ij}(\alpha)\, Y_j \quad \text{for each } i; \tag{5.3}$$

to see this, combine equations (3.1) and (4.2).

The basic principle behind the techniques discussed by Cook and Weisberg (1982) is to adjust the residuals to account for biases caused by the estimation. Consider the model $\operatorname{var} \epsilon_i = \sigma^2 w_i^{-1}$, for a given sequence of weights. The *studentized residuals* are then defined by

$$r_i = \frac{w_i^{1/2}\{Y_i - \hat{g}(t_i)\}}{\hat{\sigma}\{1 - A_{ii}(\alpha)\}^{1/2}}, \tag{5.4}$$

where $\hat{\sigma}$ is an estimate of the standard deviation factor in the model.

Motivated by the corresponding formula for standard linear regression, Wahba (1978) has suggested a formula in the unweighted case which generalizes to

$$\hat{\sigma}^2 = \frac{\Sigma\, w_i\{Y_i - \hat{g}(t_i)\}^2}{n - \operatorname{tr} A(\alpha)}. \tag{5.5}$$

In simulation studies Wahba (1983) has found that (5.5) gives a good estimate of σ^2.

It is in the spirit of the argument on cross-validation to replace the r_i of (5.4) by *generalized residuals* defined by

$$r_i^* = \frac{w_i^{1/2}\{Y_i - \hat{g}(t_i)\}}{\hat{\sigma}\{1 - n^{-1} \operatorname{tr} A(\alpha)\}^{1/2}} \tag{5.6}$$

and furthermore to replace $\operatorname{tr} A(\alpha)$ in (5.5) and (5.6) by the approximation (4.6), which will already have been calculated if the smoothing parameter is chosen by AGXV. If this procedure is followed, then both $\hat{\sigma}$ and the r_i^* are available for negligible cost.

It is interesting to note from the definition of r_i^* that the average value of r_i^{*2} is 1. This provides valuable intuition when looking at residual plots and is a sample version of the property $\operatorname{var}(r_i) = 1$ which holds for studentized residuals in ordinary linear regression (Cook and Weisberg, 1982, p. 19). The exact distribution of the r_i^{*2} of (5.6) is a generalization of the doubly non-central beta, and in any case depends on the unknown curve g, and so any inferences to be drawn from residual plots are likely to be qualitative rather than quantitative. The same is of course true of the way that residuals for ordinary regression are actually used in practice the vast majority of the time.

An alternative approach based on the equation (5.4) is possible. The work described in Section 3 can be used to give easily-calculated approximations to the individual diagonal entries of the hat matrix, and hence approximate values of the r_i. We shall not pursue this idea further in the present paper.

5.3. *Two Examples*

To illustrate the techniques developed so far, and to provide motivation for some of the later discussion, we now consider the data presented in Fig. 2. These observations consist of accelerometer readings taken through time in an experiment on the efficacy of crash helmets. The experiment, a simulated motor-cycle crash, is described in detail by Schmidt *et al.* (1981) and this particular data set was very kindly provided by Wolfgang Härdle. For various reasons, the time points are not regularly spaced, and there are multiple observations at some time points. In addition the observations are all subject to error. It is of interest both to discern the general shape of the underlying acceleration curve and to draw inferences about its minimum and maximum values. In this section we shall concentrate on the general shape; for remarks on the

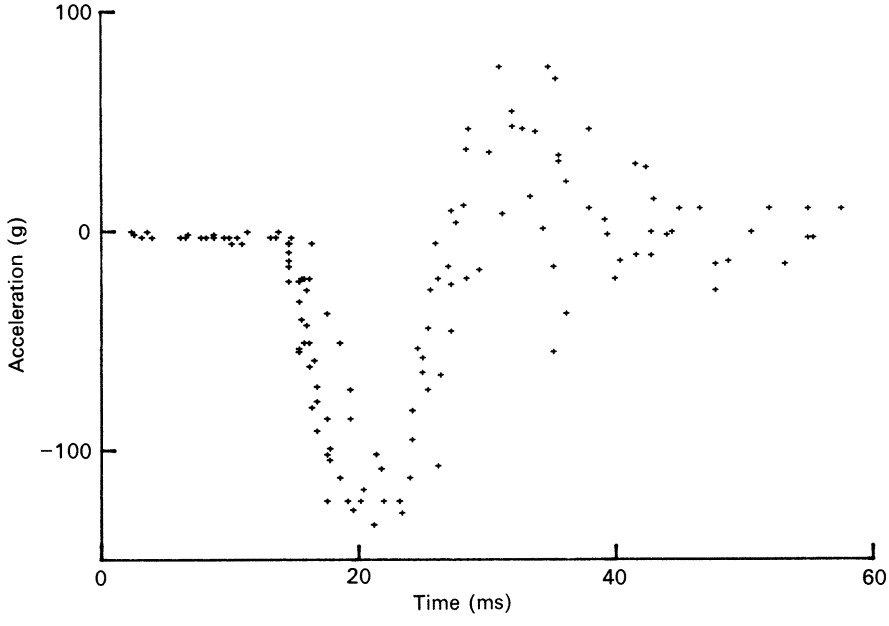

Fig. 2. The motor-cycle impact data.

other question see Section 7 below. Obviously, a more detailed analysis would concentrate on the time series nature of the data, but for illustrative purposes we shall concentrate on the model (1.1) with independent errors.

It is clear from Fig. 2 that the variance of the data is not constant, and a method to cope with this difficulty is discussed in the next section. Applying the smoothing method directly, and using AGXV to choose the smoothing parameter, gives the plot shown in Fig. 3. This gives a clear indication of the general pattern of the data, in particular the fact that the curve rebounds well above its original level before settling back, and gives crude point estimates for the maximum and minimum values. It perhaps does not follow sufficiently accurately the first bend in the curve where the variance is relatively small.

A plot of the absolute values of the generalized residuals, as defined in (5.6), is given in Fig. 4. This makes the inhomogeneity in variance very clear. We shall see in Fig. 7 below that the variance-stabilizing method developed in the next section also copes with the two possible outliers in this plot.

Another example is presented in Fig. 5. The data set is that discussed in Example 2.3.2 of Cook and Weisberg (1982) and is described in detail there. The data relate to 272 eruptions of Old

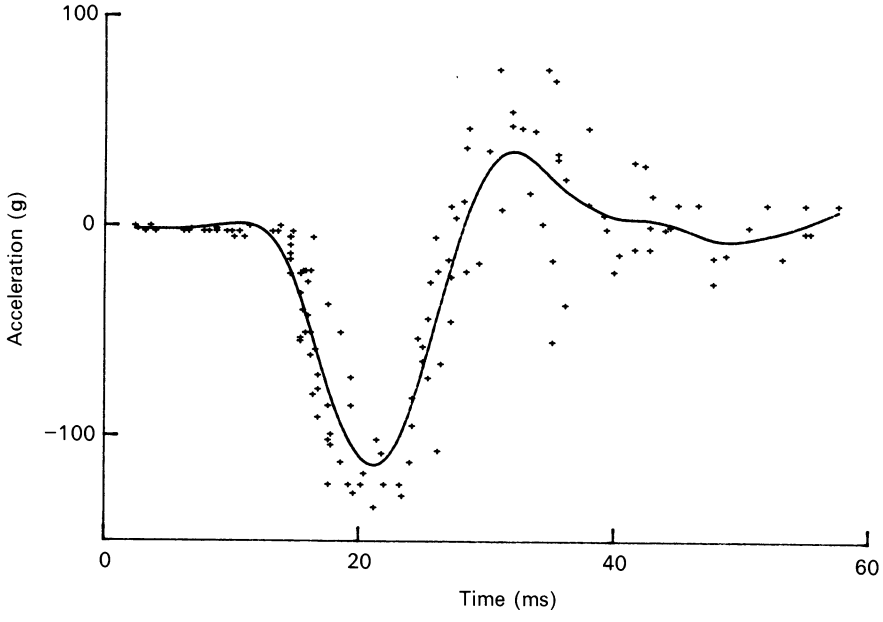

Fig. 3. The motor-cycle impact data with automatically chosen smoothing curve.

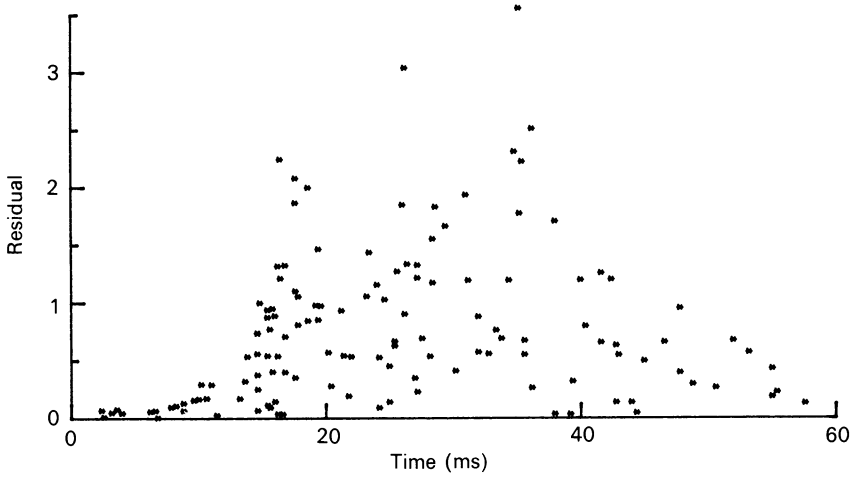

Fig. 4. Absolute values of generalized residuals for motor-cycle impact data. Unweighted case.

Faithful geyser in Yellowstone National Park, USA, and were provided by Roderick A. Hutchinson, the Yellowstone Park geologist. Cook and Weisberg fitted a linear regression to these data but noted some curvature effects in a residual plot; in particular they suggested that for large z-values the linear regression might be giving predicted y-values that are too large. The automatically-fitted spline smoothing curve makes this conclusion immediately clear, and furthermore suggests that an alternative *parametric* model for prediction might be a two-phase linear regression, treating the two data clusters separately. The least squares two-phase regression divides the data at $x = 3$ and is shown in Fig. 5. A likelihood ratio test decisively rejects single linear

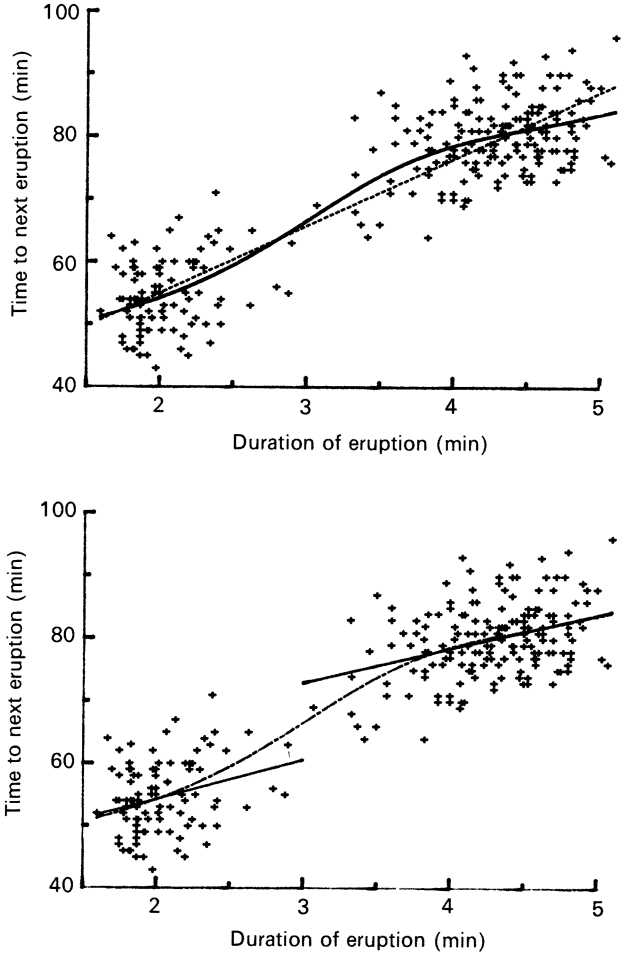

Fig. 5. Old Faithful geyser data, showing linear regression fit, automatic spline smoothing curve, and least squares two-phase linear regression.

regression in favour of two-phase regression; the data give a χ^2 statistic of 33.85 while the maximum of 100 simulated values from the null distribution of the χ^2 statistic is 16.83. (The simulation is necessary because of the remarks of Feder, 1975, that the usual theory does not give the correct null distribution. See also Hinkley, 1971, for remarks on a related problem.)

Even if one prefers to use the more classical two-phase linear regression for the ultimate purposes of explanation and prediction, the spline curve is an important exploratory step towards the final model choice. It would need a very well-trained eye to suggest any model other than ordinary linear regression on the basis of the data plot in Fig. 2.3.4 of Cook and Weisberg (1982).

5.4. *Iterative Estimation of the Weights*

Consider the regression model (1.1) where the errors ϵ_i are no longer of equal variance, but satisfy var $\epsilon_i = w_i^{-1}\sigma^2$, for some weight system w_i. It is rare in weighted least squares regression for the appropriate weights to be given explicitly. A natural and simple technique for obtaining estimates of the weights in non-parametric regression is to use the unweighted estimate of the curve g as an initial estimate to obtain *local* estimates, via local residual sums of squares, of the

error variances. This approach assumes that the true error variance depends smoothly on the design variable.

Since there is no great need to have very accurate values for the weights, a fairly crude procedure may be used. It seems to be satisfactory, in practice, to estimate w_i via a local moving average of squared generalized residuals, by

$$\hat{w}_i^{-1} = (n_i - m_i + 1)^{-1} \sum_{j=m_i}^{n_i} r_j^{*2}, \tag{5.7}$$

where, for some fixed k,

$$m_i = \max(1, i-k) \quad \text{and} \quad n_i = \min(n, i+k). \tag{5.8}$$

Putting $k = 5$ has produced good results for data sets of moderate size.

The estimated weights can then be fed back into the model. The method of Section 5.1 can be used to give an automatic choice of smoothing parameter for the weighted problem, and hence a new estimate of g can be obtained. A plot of the generalized residuals for the weighted problem can now be used to check the adequacy of the estimated weights.

The effect of applying this technique to the data of Section 5.2 is shown in Fig. 6. The dotted curves should be ignored for the moment. This curve is generally rather similar to Fig. 4 but follows the data near the left of the picture rather more closely, suggesting that the curve is constant at first. In addition there is some suggestion of a regular oscillation in the right-hand half of the plot. The plot of generalized residuals shown in Fig. 7 shows behaviour much improved over that of Fig. 4.

It is possible to perform another iteration where the weights are re-estimated; the natural analogue of (5.7) is to define new weight estimates by

$$\hat{w}_i^{-1} = (n_i - m_i + 1)^{-1} w_i^{-1} \sum_{j=m_i}^{n_i} r_j^{*2}, \tag{5.9}$$

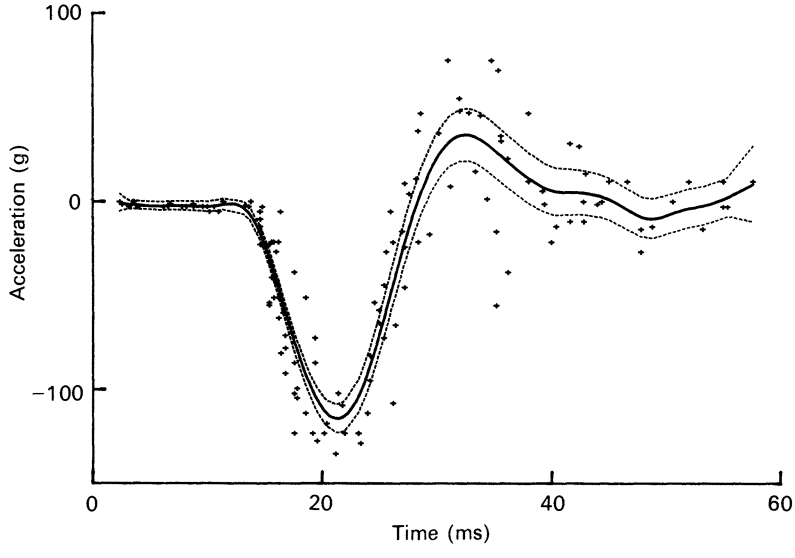

Fig. 6. Reweighted curve, with approximate probability intervals, constructed from motor-cycle impact data.

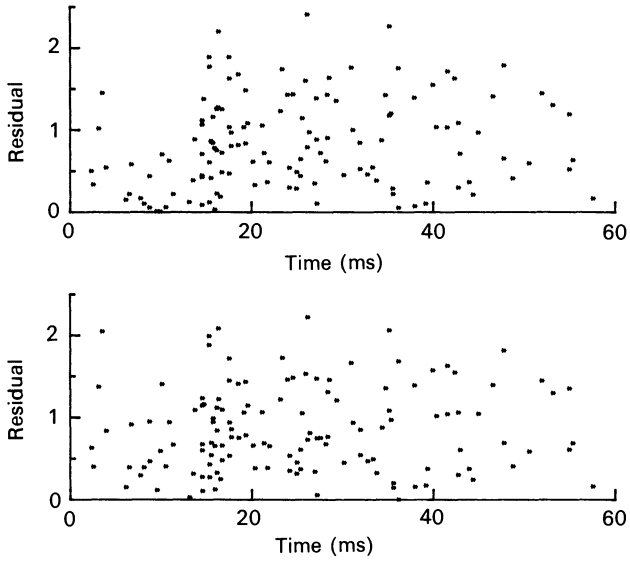

Fig. 7. Absolute generalized residuals for reweighted motor-cycle impact data. Above: first reweighting iteration; below: second reweighting iteration.

where w_j are the old estimates of the weights, and the r_j^* are the new generalized residuals. The process of finding the automatically smoothed curve and plotting residuals can then be repeated. If this is done with the motor-cycle impact data, Fig. 6 is little changed and there is a slight improvement in the residual plot, but at the expense of introducing a pattern of weights that is very non-uniform indeed.

It is, of course, possible to automate the entire procedure so that the first, unweighted, estimate is not plotted out at all, but the reweighting step is carried out as a matter of course. One could even iterate until some sort of convergence occurs. My own preference is to proceed more cautiously, looking at residual plots and the estimated curve at each stage before carrying out another reweighting iteration.

Another possible application of reweighting in the light of local variance estimation arises in ordinary regression in cases where the variance can be taken to depend smoothly on the predicted value of each observation. Here, the natural approach would be to smooth a plot of squared residuals against predicted values, to obtain estimates of the weights. An approach of this kind can be applied quite generally, and provides an alternative to the use of transformations to deal with inhomogeneous variances.

6. ERROR ESTIMATES FOR CURVES

Most non-parametric regression methodology to date has concentrated on the estimation of a curve without paying too much attention to the question of finding a confidence region (or other prediction region) for the curve. Two notable recent exceptions are Wahba (1983) and Wecker and Ansley (1983); see also Clark (1980). In the remainder of this paper, we shall build on, and modify, the ideas of Wahba (1983) to develop a methodology which is relatively simple and which enables inferences to be made not only about values of the curve itself but also about interesting functionals of the curve, such as its maximum value or gradient. The basis for our approach is a finite-dimensional Bayesian formulation of the curve estimation problem.

6.1. *A Bayesian Model*

The idea of viewing non-parametric curve estimation in a Bayesian context dates back at least to Whittle (1958). It is, perhaps, natural to look at the problem in a Bayesian way, because the choice of how much to smooth corresponds to some sort of prior information.

Bayesian models previously suggested in connection with non-parametric smoothing (Kimeldorf and Wahba, 1970; Good and Gaskins, 1971) have involved inference in infinite-dimensional spaces. Apart from causing conceptual difficulties, the use of an infinite-dimensional formulation leads to paradoxes such as the one alluded to by Wahba (1983); although the intention is to choose among curves for which $\int g''^2$ is finite, the posterior distribution is entirely concentrated *outside* the space of such smooth curves. The treatment described in this section requires finite-dimensional spaces only, and avoids such problems. It shares with previous approaches the use of a prior log likelihood equal to a negative multiple of the roughness penalty, but concentrates the prior, and hence the posterior, entirely on the space of spline curves with knots at the design points.

Consider the model (1.1) with the errors ϵ_i having independent normal distributions with mean zero and variances $w_i^{-1}\sigma^2$, where the weights w_i and the variance factor σ^2 are assumed known. Let W be the diagonal matrix with entries w_i. For simplicity, assume that the design points t_i are all distinct; coincident design points are easily dealt with but at some cost in notation. Let Γ be the space of all spline curves g satisfying (2.2), in other words all cubic splines with knots $\{t_i\}$ satisfying the "natural boundary conditions" (2.2) (iii).

For each $i = 1, \ldots, n$, let $\beta_i(t)$ be the so-called *B-spline*, which has the following properties:

$$\begin{aligned} &\beta_i \in \Gamma, \text{ so } \beta_i \text{ satisfies conditions (2.2);} \\ &\beta_i(t_i) > 0 \text{ for each } i; \\ &\beta_i(t) = 0 \text{ if } t \text{ is outside the interval } (t_{i-2}, t_{i+2}). \end{aligned} \tag{6.1}$$

B-splines are well known in the numerical analysis literature and the reader is referred, for example, to De Boor (1978, Chapter 9) for fuller details and historical remarks. The only property we shall need for the moment is the fact that any curve g in Γ can be written uniquely as a linear combination of B-splines; call the coefficients γ_i so that

$$g(t) = \sum_{i=1}^{n} \gamma_i \beta_i(t). \tag{6.2}$$

Thus, to specify a curve $g(t)$ in Γ, we need to specify the n parameters γ_i; write these as a vector γ.

Define $n \times n$ matrices B and Ω by

$$B_{ij} = \beta_j(t_i) \tag{6.3}$$

and

$$\Omega_{ij} = \int_{-\infty}^{\infty} \beta_i''(t)\, \beta_j''(t)\, dt. \tag{6.4}$$

It can then be shown, by adapting the arguments of Utreras (1980), that Ω is a non-negative definite symmetric matrix with two zero eigenvalues. It is easy to see that the residual sum of squares and the roughness penalty can be written in terms of the vector γ and the matrices B and Ω; we have

$$\sum_i w_i \{Y_i - g(t_i)\}^2 = \sum_i w_i \{Y_i - \sum_j \gamma_j \beta_j(t_i)\}^2 = (Y - B\gamma)^{\mathrm{T}}\, W(Y - B\gamma)$$

and

$$\int g''(t)^2 dt = \gamma^{\mathrm{T}} \Omega \gamma.$$

Our Bayesian formalism underlying spline smoothing can now be stated simply. Define

$\lambda = \alpha/\sigma^2$. Using the notation $\stackrel{c}{=}$ to mean "equals up to a constant", take the prior log likelihood over Γ to be

$$l_{\text{prior}}(\gamma) \stackrel{c}{=} -\tfrac{1}{2}\lambda\gamma^T\Omega\gamma = -\tfrac{1}{2}\lambda\int g''(t)^2\,dt. \tag{6.5}$$

Were it not for the two zero eigenvalues of Ω, (6.5) would give a multivariate normal prior structure; as it is, the prior is "partially improper" giving infinite variance to two of the eigenvectors of Ω.

Combining (6.5) with the density of the observations Y_i, given the curve g, gives the posterior log likelihood (by a standard Bayesian manipulation):

$$l_{\text{post}}(\gamma) \stackrel{c}{=} -\tfrac{1}{2}\lambda\gamma^T\Omega\gamma - \tfrac{1}{2}\sigma^{-2}\sum_{i=1}^{n} w_i\{Y_i - g(t_i)\}^2 \tag{6.6}$$

$$\stackrel{c}{=} -\tfrac{1}{2}\gamma^T(\lambda\Omega + \sigma^{-2}B^T WB)\gamma + \sigma^{-2}Y^T WB\gamma \tag{6.7}$$

Thus the posterior distribution of γ is multivariate normal with mean $\hat{\gamma}$ and variance matrix S^{-1}, where

$$S = \lambda\Omega + \sigma^{-2}B^T WB \tag{6.8}$$

and

$$\hat{\gamma} = \sigma^{-2}S^{-1}B^T WY. \tag{6.9}$$

The connections with spline smoothing become clear by noting that (6.6) combined with (6.5) gives

$$-2\sigma^2 l_{\text{post}}(g) \stackrel{c}{=} \alpha\int g''(t)^2 dt + \Sigma w_i\{Y_i - g(t_i)\}^2,$$

so that the posterior log likelihood is a negative multiple of the modified weighted sum of squares (5.1). Thus maximizing $l_{\text{post}}(g)$ and minimizing $S_W(g)$ are identical operations. Two main conclusions can be drawn; these parallel closely those of Wahba (1978, 1983), but they have been obtained in a different framework and by a much simpler argument.

1. The spline smoother $\hat{g}$ is the posterior mean (and the maximum of the posterior likelihood) in the Bayesian formulation described above.
2. From (6.3) the vector of values $g(t_i)$ is equal to $B\gamma$. The posterior mean obtained for γ shows that the hat matrix $A(\alpha)$ satisfies

$$A(\alpha) = \sigma^{-2}BS^{-1}B^T W$$

and hence that the posterior variance/covariance matrix of the vector $g_i = g(t_i)$ is given by

$$\text{var}_{\text{post}}(g) = \text{var}_{\text{post}}(B\gamma) = BS^{-1}B^T = \sigma^2 A(\alpha)W^{-1}. \tag{6.10}$$

The choice of smoothing parameter α coincides with the choice of the inverse scale factor λ in the prior covariance implied by (6.5). We shall discuss this point a little further in the next section.

6.2. *Plotting Inference Regions for Estimated Curves*

It is of great importance to consider how Section 6.1 ties in with earlier discussion about the appropriate choice of smoothing parameter. Pure Bayesians are advised to skip this paragraph since they will have no philosophical need to choose the smoothing parameter automatically!

A possible approach is to estimate the smoothing parameter α (and, if necessary, the variance factor σ^2) from the data, using the techniques of Sections 4 and 5, but then to draw inferences using the Bayesian model. This broad approach is that used by Wahba (1983) for her own prior and cross-validation method, and was found by her to perform well in simulation studies, in that the inference regions obtained have the properties required of frequentist confidence intervals.

It can be viewed as an empirical Bayes approach, since the prior is chosen automatically by the data under consideration.

Concentrate, first, on obtaining inference regions for predicted values $g(t_i)$. To find the exact posterior variance of $g(t_i)$ from (6.10) requires the determination of the diagonal element $A(\alpha)_{ii}$, but the approximations discussed in Section 3 makes it possible to find an approximate inference region very rapidly. For the case where all the weights are one, we have from (4.2) and (3.2) that

$$A(\alpha)_{ii} \doteq \alpha^{-1/4} n^{-3/4} 2^{-3/2} f(t_i)^{-3/4}. \tag{6.11}$$

Replacing $f(t_i)$ by the kernel estimate previously obtained in the calculation of the AGXV score gives the required approximation. A generalization of (6.11) to the weighted case, derived from the results of Silverman (1984a), is given by

$$A(\alpha)_{ii} \doteq \alpha^{-1/4} w_i (\Sigma w_k)^{-3/4} 2^{-3/2} f(t_i)^{-3/4}.$$

Using (6.10) it follows that an approximate 95 per cent probability interval for $g(t_i)$ is given by

$$\hat{g}(t_i) \pm 2\sigma\alpha^{-1/8} (\Sigma w_k)^{-3/8} 2^{-3/4} \hat{f}(t_i)^{-3/8}. \tag{6.12}$$

The formula (6.12), together with the boundary correction mentioned in Section 3, was used to draw the dotted curves in Fig. 6. They make it clear that our confidence in the accuracy of the predicted curves is high in some places, for example near the left-hand end of the interval, and lower near the middle.

Another example, which we will discuss further in the subsequent sections, is given in Fig. 8. Here the data were collected in a microbiological experiment carried out at the ARC Meat Research Institute, Langford, Bristol and were kindly supplied by Dr T. A. Roberts. The measurements are the logarithms (to base 10) of the population count per millilitre of the organism *Staphylococcus aureus* in a heart infusion broth. The real interest in this particular study is not in the value of the size of the colony at any time but in the behaviour of the rate of growth and particularly in the value of the maximum rate of growth. We shall discuss these questions in the next section.

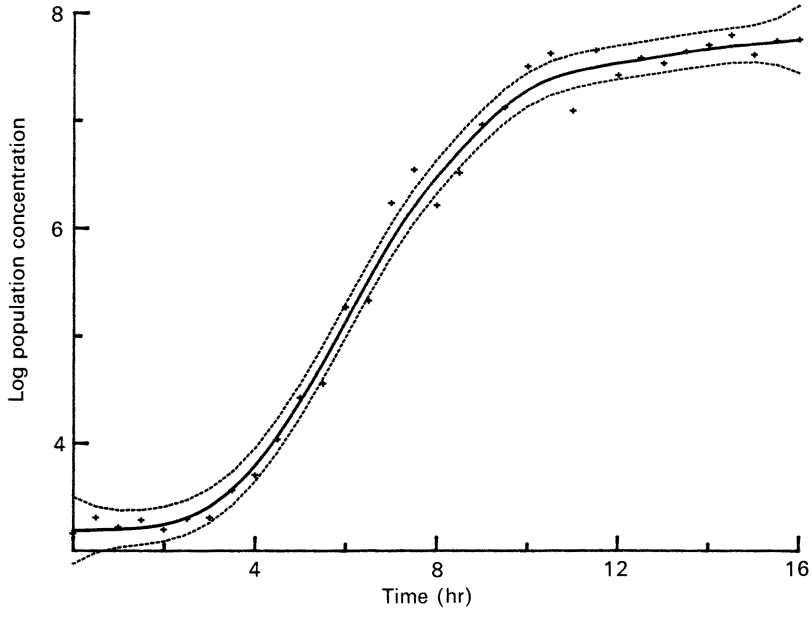

Fig. 8. Microbiological data, with estimated growth curve and probability intervals.

7. ESTIMATING PROPERTIES OF CURVES

Very many of the important questions in curve estimation involve quantities derived from the curve such as its gradient at a particular time or its maximum value. Such numerical properties are called *functionals* of the curve. (A functional $\psi(g)$ is a mapping from the space of curves to the real numbers.)

In discussing the estimation of functionals of curves, we shall adopt the same Bayesian approach as in Section 6, with the same expectation that many readers will choose the smoothing parameter automatically from the data. Before discussing details of particular functionals, it is convenient to distinguish between the two cases of *linear* and *non-linear* functionals, and also to develop the algebraic details of Section 6.1 a little further.

7.1. *Preliminaries*

A functional ψ is called *linear* if, given any curves g_1 and g_2 and numbers λ_1 and λ_2,

$$\psi(\lambda_1 g_1 + \lambda_2 g_2) = \lambda_1 \psi(g_1) + \lambda_2 \psi(g_2).$$

For example, the gradient at time t is a linear functional because, if $g_3 = \lambda_1 g_1 + \lambda_2 g_2$, then

$$g_3'(t) = \lambda_1 g_1'(t) + \lambda_2 g_2'(t).$$

However the maximum value in $[0, 1]$ is a non-linear functional because, in general,

$$\max(\lambda_1 g_1 + \lambda_2 g_2) \neq \lambda_1 \max g_1 + \lambda_2 \max g_2.$$

The estimation of linear functionals is relatively straightforward and will be dealt with in Section 7.2 below. Non-linear functionals, discussed in Section 7.3, require a little more care. In both cases, the B-spline parametrization introduced in Section 6.1 leads to very considerable savings in computer time and space. The key property is the last property of (6.1), which implies that the matrix B of (6.3) is a *band matrix of bandwidth 2* (that is to say that B_{ij} is zero if $|j - i| \geqslant 2$) and that the matrix Ω of (6.4) is a band matrix of band width 4. Since Ω is symmetric and W is diagonal, the inverse covariance matrix S of (6.8) is a symmetric band matrix of bandwidth 4. Let S have Choleski decomposition

$$S = LL^{\mathrm{T}}. \tag{7.1}$$

Then L is a lower-triangular band matrix of bandwidth 4.

The values and all derivatives of all the B-splines at each t_i can be found quickly using the properties given in De Boor (1978). The band nature of all the matrices involved then makes it possible to find the matrices S and L, and the vector $\hat{\gamma}$, in relatively small amounts of time and storage which depend linearly on the number of data points. Once this has all been done, the discussion of the next two sections yields practicable methods for finding estimates and inference regions for both linear and non-linear functionals of the estimated curve.

7.2. *Estimating Linear Functionals*

Suppose $\psi(g)$ is a linear functional of interest, for example $\psi(g) = g'(t)$ for some fixed t. Define a vector ψ_i by

$$\psi_i = \psi(\beta_i). \tag{7.2}$$

Finding the values β_i is straightforward in all cases considered, since the value and derivatives of β_i at each t_j have already been found, and each β_i is a piecewise cubic polynomial. It follows from (6.2) and the linear properties of ψ that

$$\psi(g) = \Sigma \gamma_i \psi_i. \tag{7.3}$$

Since the posterior distribution of γ is multivariate normal with mean $\hat{\gamma}$ and variance matrix S^{-1}, it follows from (7.3) that the posterior distribution of $\psi(g)$ is normal with mean $\psi^{\mathrm{T}}\hat{\gamma}$ and variance

$\sigma_\psi^2 = \psi^T S^{-1} \psi$. Finding σ_ψ^2 is easy because of the band nature of the matrix L; the vector $L^{-1}\psi$ can be found by back-substitution in a small linear amount of time and storage, and, by (7.1), σ_ψ^2 is just $(L^{-1}\psi)^T(L^{-1}\psi)$.

Of course, the mean $\psi^T\hat{\gamma}$ is precisely $\psi(\hat{g})$; this is to say that the point estimate of, for example, $g'(t)$ is given by differentiating the curve estimate $\hat{g}$ at t. The linearity of the functional ψ is vital if this is to be the case.

This methodology is illustrated by giving estimates of the growth rate underlying the data in Fig. 8 above. Fig. 9 shows plots of the estimated derivative $\hat{g}'(t)$ and inference regions calculated by ±2 posterior standard deviations of $g'(t_i)$ at each t_i.

7.3. *Estimating Non-linear Functionals by Simulating from the Posterior*

If the functional ψ is non-linear, then the posterior distribution of $\psi(g)$ will not, in general, be tractable, because it is that of a non-linear function of a high-dimensional multivariate normal distribution. However we shall see in this section that it is easy to simulate from the posterior distribution and hence to make inferences about any functional of interest.

Suppose z is a vector of n independent standard normal random variables. Let $\rho = \hat{\gamma} + (L^T)^{-1}z$. Then ρ has a multivariate normal distribution with mean $\hat{\gamma}$ and variance matrix

$$(L^T)^{-1}\{(L^T)^{-1}\}^T = (L^T)^{-1}L^{-1} = (LL^T)^{-1} = S$$

so that the distribution of ρ is precisely the required posterior distribution of the B-spline coefficients γ of (6.2). Furthermore, finding each realization of ρ involves simulating n normal random variables and solving a single band-limited upper-triangular linear system, a very small computational burden.

Once a posterior realization of γ has been found, it is straightforward, using if necessary the stored values of the B-splines and their derivatives, to find the corresponding value of the functional $\psi(g)$. Again the computer time and storage required will be linear in n. This methodology makes it practicable to generate large numbers of simulated values from the posterior distribution of $\psi(g)$, and hence to obtain Monte Carlo estimates, to any reasonable degree of accuracy, of its characteristics.

Examples of non-linear functionals whose posterior distributions are intractable, but which yield easily to this approach, are the maximum value and the maximum gradient of g in a given interval, the point at which g is maximized, and the number of "bumps" (suitably defined) in g. Another such functional, of interest in developing non-parametric calibration methods, is the point at which g crosses a given level. It should be stressed that the basic technique will work for *any* functional of interest.

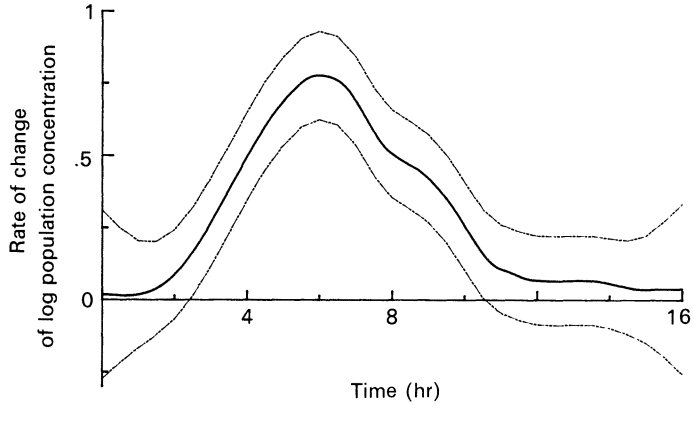

Fig. 9. Estimated growth rate for microbiological data, with probability intervals.

The maximum gradient is an important quantity in the growth curve study described above. Using the value automatically chosen by AGXV for the smoothing parameter, and the estimate (5.5) for σ^2, one hundred realizations of the posterior were generated. The one hundred posterior values obtained for the maximum gradient had sample mean 0.84 and standard deviation 0.06. Furthermore, a probability plot (tested by the Shapiro and Wilk, 1965, method as implemented in the MINITAB statistical package) showed that the posterior distribution of the maximum gradient is approximately normal. It is important, and possibly surprising, to note that the point estimate 0.84 of the maximum gradient is somewhat larger than the maximum value 0.77 of the curve $\hat{g}'$ plotted in Fig. 9. This discrepancy is a consequence of the non-linearity of the functional ψ; in fact it can be shown, using Jensen's inequality, that the posterior mean of the maximum gradient will always be greater than the maximum gradient of $\hat{g}'$.

A similar procedure was applied to obtain estimates and posterior standard deviations for the minimum and maximum of the acceleration curve estimated in Fig. 7. The point estimates were -114.8 and 37.5 and the standard deviations 6.2 and 9.1 respectively.

Stewart (1979) also uses the idea of simulating from a high-dimensional posterior distribution in order to find the distribution of functions of the underlying parameters or curve. Our approach allows realizations of the posterior to be obtained directly, rather than by Stewart's rejection sampling technique.

8. RELATED TOPICS

8.1. *Robust and Generalized Smoothing*

A natural extension of the spline smoothing approach is to devise a robust version of the procedure, by replacing the sum of squared errors in (2.1) by a different function of the errors, to give

$$S_R(g) = \Sigma \rho\{Y_i - g(t_i)\} + \alpha \int g''(x)^2 dx. \tag{8.1}$$

Here the function $\rho(x)$ would usually be a convex function which is less rapidly increasing than x^2. Minimizing $S_R(g)$ then gives a smoothing spline which is robust against or resistant to outliers in the data. This idea has been discussed by Lenth (1977), Huber (1979) and Cox (1983), among others. Huber (1979) points out that the minimization of S_R may be carried out in practice by an iterative reweighting scheme where a sequence of functionals S_W, as defined in (5.1), are minimized successively for weights and data points which are modified at each stage. The basic ideas of iteratively reweighted linear regression (see Green, 1984) carry over to the spline smoothing case, and have been explored and developed by O'Sullivan (1983). Especially when viewed as an iterative reweighting procedure, the robust spline smoothing method has something in common with the local variance estimation method suggested in Section 5.4 above. However, there are obvious differences both in the underlying model and in the calculations actually carried out in practice.

The formulation of (8.1) can be extended further by generalizing the location dependence of the distribution of Y on $g(t)$. This has the flavour of generalized linear models (McCullagh and Nelder, 1983). Suppose that $\rho(y, \theta)$ is a function of a real parameter θ and an observation y, which may now take values in any space. Define

$$S_{GL}(g) = \Sigma \rho\{Y_i, g(t_i)\} + \alpha \int g''(x)^2 dx. \tag{8.2}$$

Usually, $-\frac{1}{2}\rho(y, \theta)$ will be the log likelihood or partial likelihood of θ given y for some parametric family of distributions. In that case, $-\frac{1}{2} S_{GL}(g)$ is a penalized version of the log likelihood function of $\{g(t_i)\}$ as a vector of parameters underlying independent observations Y_i. A general discussion of the idea of penalized likelihood is given by Silverman (1984c). By the same procedure as in generalized linear models, suitable choice of the function ρ gives non-parametric versions of many regression techniques, such as logistic regression (see Silverman, 1978; Anderson and Blair, 1982) and the robust regression methods discussed above.

Provided that S_{GL} has a finite minimum at $\hat{g}$, it can be shown that $\hat{g}$ will be a spline function

satisfying conditions (2.2) above; again, $\hat{g}$ can often be found by an iterative reweighting strategy. It appears to be the case (see Silverman, 1978, 1982b; O'Sullivan, 1983) that $\hat{g}$ will exist provided that $\Sigma \rho\{Y_i, g(t_i)\}$ has a finite minimum in the space of all *linear* functions g.

The cross-validation ideas for choosing α have a natural analogue: the score to be minimized would be

$$\text{XVSC}(\alpha) = \Sigma \rho\{Y_i, g_\alpha^{-i}(t_i)\}, \tag{8.3}$$

where g_α^{-i} is the minimizer of $S_{GL}(g) - \rho\{Y_i, g(t_i)\}$. Since, in the general case, each g_α^{-i} would have to be found by an iterative technique, calculating (8.3) for a range of values of α requires considerable computational effort. O'Sullivan (1983) has considered how generalized cross-validation can be applied in this case. Detailed work on the generalization of the approximations of Section 4 above remains to be done.

Another topic for further investigation is the application of the Bayesian ideas of Sections 6 and 7 in the context of this discussion. Assume that $-\frac{1}{2}\rho$ is a log likelihood function. An easy and natural approach, once the maximum $\hat{g}$ has been found, is to expand the posterior log likelihood $-\frac{1}{2}S_{GL}$ to second order about $g = \hat{g}$. This quadratic approximation (cf. Box and Tiao, 1973, equation (1.3.85)) yields a simple multivariate normal approximation to the posterior distribution. Furthermore, one can find an *equivalent weighted least-squares problem*, based on *pseudo-observations* η_i and *pseudo-weights* ω_i, such that, setting $\sigma^2 = 1$, the normal approximation is precisely the posterior distribution for the equivalent problem. Thus all the methodology of Sections 6 and 7 can be applied, without modification, to find approximate inference regions and to simulate from the approximate posterior, once the estimate $\hat{g}$ has been found. The pseudo-observations and pseudo-weights are defined (using subscript 2 to denote differentiation with respect to the second argument) by

$$\eta_i = \hat{g}(t_i) - \rho_2\{Y_i, \hat{g}(t_i)\}/\rho_{22}\{Y_i, \hat{g}(t_i)\},$$

$$\omega_i = \tfrac{1}{2}\rho_{22}\{Y_i, \hat{g}(t_i)\}.$$

Formulae closely related to these are found in Huber (1981) and McCullagh and Nelder (1983). The accuracy and usefulness of the normal approximation in this precise context would be an interesting subject for future work.

8.2. *The Multivariate Case: Surface Estimation*

Huber (1979) described univariate spline smoothing as the "theoretically cleanest approach to linear smoothing". The attraction of the method is the happy combination of circumstances that the estimate is the solution of a neatly expressed and intuitively attractive minimization, and that it can be calculated and stored easily as a piecewise polynomial. Unfortunately, in the case where the design points are multivariate, the mathematics do not fall out quite as nicely. The function g is now a function of a vector variable; if the design points are bivariate, then g can be viewed as the height of a two-dimensional surface.

One approach is the "thin plate" spline (see Meinguet, 1979) where the roughness penalty $\int g''^2$ is replaced by $\int (g_{11}^2 + 2g_{12}^2 + g_{22}^2)$ and a corresponding formula in higher dimensions; here subscripts denote partial derivatives with respect to the corresponding arguments. Some progress can now be made; roughly speaking, it is possible to find a finite set of functions $\{\beta_i\}$ such that the smoothing surface can, for the given design points, be expressed in the form (6.2). Unfortunately the β_i are not bounded or of bounded support and the counterparts of all the matrix manipulations required to find the coefficients $\hat{\gamma}_i$ require the handling of full, rather than banded, matrices. Wahba and Wendelberger (1980) give further details together with interesting practical examples. The approach of Sections 6 and 7 above could equally be applied to this case, though at considerable computational cost. It is to be hoped that further theoretical and practical work will be done in this area; it would be interesting to be sure that the behaviour of the functions β_i does not lead to numerical instabilities, a problem considered by Dyn and Levin (1983).

Of course, the need for useful computational approximations is even greater in the multivariate case. Another important topic would be to understand the rôle that boundary effects play in the estimation. In the one-dimensional case the boundary condition (2.2) (iii) is applied implicitly in the procedure (see Rice and Rosenblatt, 1983); though the present author has not found this to lead to any practical difficulties, it will certainly be the case that the boundary will be felt much more strongly in the multivariate case, where far more of the points are near the "edge" of the data set.

ACKNOWLEDGEMENTS

I am delighted to acknowledge the helpful comments of the referees and of several colleagues, including A. Baddeley, C. Chatfield, R. Fowler, W. Härdle, J. Rice, A. Robinson and S. Wilson. I am very grateful to G. Watters for computational assistance and to the Science and Engineering Research Council for support.

REFERENCES

Anderson, J. A. and Blair, V. (1982) Penalized maximum likelihood estimation in logistic regression and discrimination. *Biometrika*, **69**, 123–136.

Boneva, L. I., Kendall, D. G. and Stefanov, I. (1971) Spline transformations: three new diagnostic aids for the data analyst (with Discussion). *J. R. Statist. Soc.* B, **33**, 1–70.

Box, G. E. P. and Tiao, G. C. (1973) *Bayesian Inference in Statistical Analysis.* Reading, Mass.: Addison–Wesley.

Clark, R. M. (1980) Calibration, cross-validation and carbon-14. II. *J. R. Statist. Soc.* A, **143**, 177–194.

Cook, R. D. and Weisberg, S. (1982) *Residuals and Influence in Regression.* London: Chapman and Hall.

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

Craven, P. and Wahba, G. (1979) Smoothing noisy data with spline functions. *Numer. Math.*, **31**, 377–403.

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

Dyn, N. and Levin, D. (1983) Iterative solution of systems originating from integral equations and surface interpolation. *SIAM J. Numer. Anal.*, **20**, 377–390.

Feder, P. I. (1975) The log likelihood ratio in segmented regression. *Ann. Statist.*, **3**, 84–97.

Good, I. J. and Gaskins, R. A. (1971) Nonparametric roughness penalties for probability densities. *Biometrika*, **58**, 255–277.

Green, P. J. (1984) Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with Discussion). *J. R. Statist. Soc.* B, **46**, 149–192.

Hinkley, D. V. (1971) Inference in two-phase regression. *J. Amer. Statist. Ass.*, **66**, 736–743.

Huber, P. J. (1979) Robust smoothing. In *Robustness in Statistics* (R. L. Launer and G. N. Wilkinson, eds). New York: Academic Press.

——— (1981) *Robust Statistics.* New York: Wiley.

Kimeldorf, G. and Wahba, G. (1970) A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.*, **41**, 495–502.

Lenth, R. V. (1977) Robust splines. *Commun. Statist.*, **A6**, 847–854.

McCullagh, P. and Nelder, J. A. (1983) *Generalized Linear Models.* London: Chapman and Hall.

Meinguet, J. (1979) Multivariate interpolation at arbitrary points made simple. *ZAMP*, **30**, 292–304.

O'Sullivan, F. (1983) The analysis of some penalized likelihood schemes. Technical Report no. 726, Dept of Statistics, University of Wisconsin–Madison, USA.

Priestley, M. B. and Chao, M. T. (1972) Non-parametric function fitting. *J. R. Statist. Soc.* B, **34**, 385–392.

Reinsch, C. (1967) Smoothing by spline functions. *Numer. Math.*, **10**, 177–183.

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

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

Schmidt, G., Mattern, R. and Schueler, F. (1981) Biomechanical investigation to determine physical and traumatological differentiation criteria for the maximum load capacity of head and vertebral column with and without protective helmet under the effects of impact. EEC Research Program on Biomechanics of Impacts, Final report, Phase III, Project G5, Institut für Rechtsmedizin, University of Heidelberg, West Germany.

Shapiro, S. S. and Wilk, M. B. (1965) An analysis of variance test for normality (complete samples). *Biometrika*, **52**, 591–611.

Silverman, B. W. (1978) Density ratios, empirical likelihood and cot death. *Appl. Statist.*, **27**, 26–33.

——— (1982a) Kernel density estimation using the fast Fourier transform. *Appl. Statist.*, **31**, 93–99.

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

——— (1984a) Spline smoothing: the equivalent variable kernel method. *Ann. Statist.*, **12**, 898–916.
——— (1984b) A fast and efficient cross-validation method for smoothing parameter choice in spline regression. *J. Amer. Statist. Ass.*, **79**, 584–589.
——— (1984c) Penalized maximum likelihood. In *Encyclopedia of Statistical Sciences*, Vol. 6 (S. Kotz and N. L. Johnson, eds). New York: Wiley.
Stewart, L. (1979) Multiparameter univariate Bayesian analysis. *J. Amer. Statist. Ass.*, **74**, 684–693.
Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions (with Discussion). *J. R. Statist. Soc.* B, **36**, 111–147.
Utreras D. F. (1980) Sur le choix du parametre d'ajustement dans le lissage par fonctions spline. *Numer. Math.*, **34**, 15–28.
Villalobos, M. A. and Wahba, G. (1982) Multivariate thin plate spline estimates for the posterior probabilities in the classification problem. Technical report 686, Department of Statistics, University of Wisconsin-Madison, USA.
Wahba, G. (1975) Smoothing noisy data with spline functions. *Numer. Math.*, **24**, 383–393.
——— (1978) Improper priors, spline smoothing, and the problem of guarding against model errors in regression. *J. R. Statist. Soc.* B, **49**, 364–372.
——— (1983) Bayesian confidence intervals for the cross-validated smoothing spline. *J. R. Statist. Soc.* B, **45**, 133–150.
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.
Wecker, W. P. and Ansley, C. F. (1983) The signal extraction approach to nonlinear regression and spline smoothing. *J. Amer. Statist. Ass.*, **78**, 81–89.
Wegman, E. J. and Wright, I. W. (1983) Splines in statistics. *J. Amer. Statist. Ass.*, **78**, 351–365.
Whittaker, E. (1923) On a new method of graduation. *Proc. Edinburgh Math. Soc.*, **41**, 63–75.
Whittle, P. (1958) On the smoothing of probability density functions. *J. R. Statist. Soc.* B, **20**, 334–343.

DISCUSSION OF DR SILVERMAN'S PAPER

Professor P. Whittle (Statistical Laboratory, Cambridge University): Mr Chairman, Colleagues. It is a real pleasure to me to propose the vote of thanks to Dr Silverman. His paper not only addresses a theoretical problem elegantly and effectively, but, I think you will agree, makes a genuine effort to address a practical problem practically. In other words, Dr Silverman recognises that it is good to have a clever idea, even better to have a workable one, and best of all to have a workable clever idea.

It is plain that one cannot discuss these matters without being prepared to consider a Bayesian formulation (which I am understanding in a frequentist non-personal sense) and I am glad that Dr Silverman did so in such a matter-of-fact fashion. Of polemics we have had more than enough over the years, and it should be recognized that, with the exception of a short historical interlude, the approach has always been considered a perfectly natural one.

Dr Silverman's approach is to minimize the form (2.1), sum of squares plus roughness penalty. As he observes, this is an approach which dates back at least to Whittaker. In fact, Whittaker gave the approach in its obvious Bayesian setting, with the roughness penalty interpreted as proportional to the logarithm of a prior density. Whittaker and Robinson (1924, p. 303) are quite explicit on this. Interestingly, they refer to a very early modern view of these matters "by Mr G. King, in the course of a discussion on Dr T. B. Sprague's paper of 1886, *J.I.A.*, **26**, p. 77: 'What is the real object of graduation? Many would reply, to get a smooth curve, but that is not correct. The reply should be, to get the most probable deaths'."

Of course, the problem is also just that familiar as "signal extraction", $g(x)$ being the signal and the sample providing intermittent noisy observation. In this context it is taken for granted that one must specify the joint statistics of signal and noise if one is to deduce reasonable procedures.

The appeal to cross-validation to provide an estimating principle for the smoothing coefficient is of course a plausible one. However, this is an appeal to an additional principle, and one might hope that the principle of maximum likelihood would be sufficient. If one were to give expression (2.1) its full Bayesian interpretation then one would have a negative log-likelihood

$$\frac{1}{\sigma^2}\sum_1^n (Y_i - g(t_i))^2 + \frac{\lambda}{N}\sum_1^N (N^2\Delta^2 g)^2 + n\log\sigma^2 - N\log\left(\frac{\lambda}{N}\right)$$

Here λ, σ^2 are appropriate scale factors and the integral in (2.1) has been approximated by a sum, the differences $\Delta^2 g$ being taken over an appropriate interval. One might now minimize this expression, not merely with respect to g, but also with respect to σ^2 and λ, thus effectively estimating the smoothing coefficient $\alpha = \lambda\sigma^2$ However, the quantity N now appears as a parameter, which one cannot allow to become infinite without degeneracy. This seems then like an exchange of one parameter, the smoothing coefficient, for another, the number of "degrees of freedom of the model". These ideas seem close to those the author sets out in Sections 5 and 6, and I should be interested in his reactions to them.

The multivariate or multidimensional case so often provides a test of the fundamental workability of a procedure, and Section 8 is interesting for this reason. If one applies the author's procedure with p-dimensional x, then the smoothing kernel (under constant observation density) would have Fourier transform

$$\frac{\alpha}{(\sum_j \omega_j^2)^2 + \alpha} \rightarrow \frac{\alpha\,|\,\omega\,|^{p-1}}{|\,\omega\,|^4 + \alpha},$$

where the arrow indicates what the expression becomes under a polar transformation. The integral of this expression diverges if $p \geqslant 4$, indicating that the smoothing kernel is then infinite at the origin. In other words, one is scarcely smoothing the observations at all if $p \geqslant 4$. Rather paradoxically, one must increase the degree of the differentials in the roughness penalty, i.e. *relax* the statistical assumptions on g, if one is to obtain a useful smoothing.

There is much more I could say, but I am now restricted to proposing, with warmth, that the author be awarded a vote of thanks.

Professor D. M. Titterington (University of Glasgow): Tonight we have heard a stylish and persuasive account of the practical capabilities of what is currently a very hot topic. Now, I expect the author will object to my apparent neglect of the theoretical aspects of the paper so I had better clarify the spirit of that first sentence. The recent statistical literature is very well sprinkled with papers on a class of problems, including nonparametric density estimation, to which the present topic belongs. However, it has to be admitted that many of these articles are theoretical and seem to lose track of their fundamentally practical basis. In simple terms, the crucial role of these techniques is to produce a reasonable curve or picture. Of course it is important to establish reassuring, more or less distribution-free, asymptotic properties, but the timing of the present paper is meritorious in prompting us to take stock of the current balance of activity in these areas.

The main practical questions to me, are as follows. (i) Is smoothing worthwhile? (ii) If so, is it worth using sophisticated procedures for choosing the smoothing parameter? (iii) In practical terms, is the method of cross-validation a good one? If the answers are all "No", then many people, myself included, have wasted a lot of time in recent years! I shall now try to argue away, at least partially, from these answers.

In Section 6.1 we encounter both a Bayesian interpretation for the smoothing procedure and a ridge-regression formulation for the recipe (equations (6.8) and (6.9)). In a review of the massive literature on "conventional" ridge regression, Draper and van Nostrand (1979) find that the technique hardly ever improves substantially upon ordinary least squares, unless the Bayesian structure is genuine. The degree of smoothing imposed is typically very small, in that the resulting smoothed estimates do not differ much from the unsmoothed. One has to admit that the Bayesian interpretation of the spline-smoothing prescription is rather contrived but it is clear, from Figs 5, 6 and 8, that the smoothed estimate here differs substantially from the unsmoothed interpolating spline. The crucial factor is that, in most conventional ridge regression contexts, the number of "parameters", k, say, is small, compared with the sample size, n. Here, $k = n$, effectively, and, as a result, the variance of the unsmoothed estimator is uselessly large. Other manifestations of such high parameterizations are the work of Green *et al.* (1983) and an approach to image enhancement which emanates from the following model. The relationship between the true image intensities, γ, and the observed picture, Y, is

$$Y_i = (B\gamma)_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where n is the number of pixels, B describes the resolution properties of the observing instrument, $E(\epsilon) = 0$ and $\mathrm{cov}\,(\epsilon) = \sigma^2 W$, say. Note that, even if B, W and σ^2 are known, this is a saturated

model and is often ill-posed. A common method of producing a stabilized estimate of γ is to use a so-called *regularization* prescription which is, notationally, the same as (6.8) and (6.9). The imposition of a certain amount of smoothing is usually very advantageous, but the formality of choice has varied. Phillips (1962) suggests the *ad hoc* approach of simply choosing a degree of smoothing "that appears to take out the oscillation (instability) without appreciably smoothing the function", but the more formal method of Reinsch (1967) has been popular. In one version of this, for the case $W = \text{diag}\,(w_1, \ldots, w_n)$, α might be chosen so that

$$\sum_i w_i \{ Y_i - (B\hat{\gamma}(\alpha))_i \}^2 = n\sigma^2, \qquad (*)$$

in which σ^2 would have to be estimated. Some interesting questions are as follows.

(a) Is Phillips' very *ad hoc* approach as good as any, particularly if, as in some sections of the present paper, we are only carrying out exploratory analysis?
(b) Is cross-validatory choice feasible in image-enhancement problems in which n can be very large, 2^{18} or more?
(c) Comparison of (*) with (5.5) suggests that (*) imposes more smoothing, and therefore larger biases, than does cross-validation. Is the difference meaningful from a practical point of view and is it possible that cross-validation is less reliable? In Wahba (1981), for instance, examples are given with $n = 32$ and $n - \text{tr}\,A(\alpha)$ below 1, suggesting that cross-validation can severely undersmooth.

At this point, I should like to make a few specific remarks.

(i) I have to say that the curves in Figs 3 and 6 seem very similar, in spite of the differences in sophistication. I wonder if a more interesting residual plot would be one of the signs of the residuals.
(ii) It would be nice if the confidence bands in Fig. 6 allowed simultaneous inference for all values of t. It would then be possible to say useful things about the curve as a whole. For instance, one could use Professor Stone's (other) bootlace (Stone, 1983) to assess the minimum plausible number of inflexions or modes by tightening the lace between the bands. However, I note from p. 139 of Wahba (1983) that "strictly speaking, these curves only have meaning at $t = 1/n$" (the knots) and that they are pointwise intervals.
(iii) The kernel-function representation is fascinating and I wonder if the link will lead to quick ways of calculating a good smoothing parameter in density estimation.

Finally, let me return to my original theme. A recent monograph (Prakasa Rao, 1983) contains an impressively comprehensive review of the theory behind these and other problems, but among over 500 pages, only one is devoted to Figures. It is perhaps not surprising that these Figures originate from the work of tonight's speaker and point to his regard for practical and pragmatic aspects, which complements his theoretical expertise. I hope that tonight's paper will stimulate further investigation of the real practical capabilities of the methods as well as reducing the tendency towards theoretical overkill.

I have much pleasure in seconding the vote of thanks.

The vote of thanks was carried by acclamation.

Dr E. M. Scott (University of Glasgow): I wish to congratulate Dr Silverman on a well-prepared and comprehensive paper, which has surely provided sufficient detail of one form of spline smoothing for many "less expert" users to consider implementing this approach.

My comments take the form of general remarks concerning the relationship between spline smoothing and other methods of non-parametric regression estimation. There are one or two very important points which I think worthy of note.

The first comment concerns the representation of the smoothing spline as a weighted average. However complicated the form, this brings spline functions into line with other methods of non-parametric regression estimation, and so definitions of consistency and convergence given by Stone (1977, 1982) will prove applicable. I suspect the similarity between the various methods in final form will mean that other methods are preferred before splines due to their "less complicated form". There is one other formulation of smoothing splines which computationally proves much simpler, namely regression splines (Wegman and Wright, 1983). The description of splines as piece-

wise polynomials constrained to join and be continuous at knots, means that by the introduction of knot number and position as further parameters, the spline may be fitted using standard multiple regression packages (Smith, 1979; Buse and Lim, 1977). Suggestions concerning the choice of these further parameters are given by Wold (1974). Estimation of the regression function by a variety of methods of non-parametric regression reveal little difference in the final smoothed function. Thus choice of method will often be controlled by computational simplicity and ease of use (Scott, 1984).

Dr Silverman stresses rightly the importance of being able to use the fitted form for inference concerning fitted values etc. He does, however, state that this aspect has been neglected in the literature. If we consider some alternative methods of non-parametric regression estimation we will see that other authors are aware of the importance of this aspect of their work. The Bayesian estimate proposed by O'Hagan (1978) includes the derivation of the posterior expected value and variance of his estimate, while Clark (1980) dealt almost entirely with the problem of providing confidence and prediction bands for his estimate based on convolution smoothing. If we consider kernel regression estimation, confidence and prediction bands for the curve may be obtained through a "conditional density estimation" approach.

My final comments deal with the extension to multivariate function estimation. Thin plate or Laplacian smoothing splines prove practically intractable in greater than 3 dimensions. Kernel regression estimates will extend readily, while for convoluted smoothing a multi-dimensional interpolating function would require to be defined. The multivariate problem may be tackled in a different manner using the ideas of projection pursuit which models the regression surface as a sum of smooth functions of linear combinations of the predictor variables (Friedman and Stuetzle, 1981).

Dr F. H. C. Marriott (University of Oxford): Dr Silverman has given an admirably clear and complete account of spline regression. The examples show how well the method works, and naturally raise the question of possible extensions, particularly to the case of more than one regressor variable.

The paper relies heavily on two results – the relationship between a natural roughness penalty and the cubic spline, and the simplicity of the computation owing to the band structure of the matrices involved. Neither seems to generalize simply to the multiple regression case. This difficulty in extending results beyond the simplest problems seems to apply to nearly all non-parametric methods, and largely accounts for the popularity of linear models and their extensions.

Perhaps the most important application of the method is graphical, making it possible to display a curve fitted more or less objectively, but without preconceptions about its form. An analogous procedure is scarcely possible with more than two or three regressor variables, and it is reasonable to restrict attention to these cases. I have not seen the references in the final section of the paper, but there are obvious difficulties in the complexity of the surfaces and in the computing problems.

There are, of course, other methods of tackling the problem, none of them so elegant or intuitively appealing, but that may nevertheless give a reasonable answer.

(i) Simple interpolation may be applied to subsets of the data, and the resulting surfaces averaged. The degree of smoothing then depends on the size of the subsets.

(ii) Regressor variables may be fitted stepwise, using cubic splines. This is clearly less flexible than a true multidimensional spline, but may be modified by fitting x_1, then fitting x_2 to the residuals for overlapping ranges of x_1 and splicing the resulting surfaces together.

(iii) Kriging methods have already been mentioned in the discussion.

These possibilities are all, in a sense, approximations to generalized spline fitting. I should be interested to know whether my doubts about the feasibility of a true generalization are justified, and if so whether any of them gives a satisfactory working method.

In conclusion, as well as being an excellent account of existing methodology, this paper raises theoretical and practical questions of great importance. I hope that it will stimulate much valuable research.

Dr J .T. Kent (University of Leeds): I have two comments to make on this very lucid and useful paper. First is a statement about some work on *spherical splines* carried out with Peter Jupp at the University of St Andrews. We were interested in fitting smooth paths to data from the line to

the sphere, for example, polar wander paths. It turned out that the most promising approach was to "unwrap" the problem from the sphere onto the plane, thus looking for a smooth path from the line to the plane. Of course, a path to the plane R^2 is just an ordered pair of paths to the line R^1, in our case with a common smoothing parameter. Thus, all of the standard spline techniques, as discussed in tonight's paper, can be used to fit a smooth function and to assess its accuracy. Note that our problem can be considered to be multivariate in that we are concerned with smooth functions from the line to a several-dimensional space. This usage is in contrast to Section 8.2 of the paper which is concerned with smooth real-valued functions of several variables.

Secondly, I have a question. Spline techniques require the explanatory variable x to be known precisely, just as in standard regression analysis. However, if in a linear regression model of y on x, it turns out that x is measured subject to error, then the regression model should be replaced by a linear structural relationship. Is there any extension of spline techniques to this "errors-in-variables" case? In the example of polar wander paths, the times are seldom known very precisely.

Dr Frank Critchley (University of Warwick): I would like to congratulate the author on the clarity, breadth and depth of his paper, to make one comment, to note two related topics and to ask three questions about possible extensions.

Comment Even allowing for their definition as predictors of $g(\cdot)$ itself rather than future values, the approximate probability intervals in Fig. 6 seem perhaps rather too good to be empirically convincing, especially in the region of 15 ms. Presumably their narrowness reflects the strength of (prior) model and/or the convolution nature of the estimation method.

Related topics. (1) Splines are slowly becoming more widely used in multivariate analysis: see, for example, Gifi (1981), De Leeuw (1982), Van Rijckevorsel (1982) and Ramsay (1982). This latter uses M-splines to fit the key distance-dissimilarity function of multidimensional scaling. Note that, if we regard the degree of the polynomials and the number and location of the knots as parameters, we have a special case of the parametric scaling method introduced by Critchley (1978). (2) In recent unpublished work I consider a multivariate variable kernel density estimation method with local smoothing parameter proportional to $s(sh_i)^{-\gamma}$, $0 \leqslant \gamma \leqslant 1$, $i = 1, \ldots, n$, where s denotes a global measure of scale. If we require that the method be self-respecting, i.e. that $\hat{f}(t_i) = h_i$, then for given γ, this reduces to a one-parameter problem. Moreover, the jackknifed likelihood is now an explicitly known smoothed version of the likelihood so that no further computation is needed to evaluate it. See also Abramson (1982). It will be of interest to see how the ideas of tonight's paper relate to this problem, for example taking y_i to be a fixed kernel estimate of $f(t_i)$.

Questions. It is of interest to ask how far the present paper's results extend to: (1) a full quadratic loss function; time series models; (2) the multivariate case in which the data points t_i are *ordered* in some way; (3) cases with constraints such as $g(0) = 0$.

Dr M .C. Jones (University of Birmingham): It might be useful to potential users of nonparametric regression estimates to point out two practical advantages of the spline regression method, as discussed in Dr Silverman's excellent paper, over some other nonparametric regression estimation methods, particularly in comparison with the kernel method. These remarks are illustrated by a practical application to a small dataset ($n = 28$) kindly supplied by my colleague Dr M. J. Faddy.

The data are shown in Fig. D1, along with a (constant bandwidth) kernel estimate (the dotted line); the value of the smoothing parameter has been chosen subjectively to give a reasonable fit to the main body of data corresponding to small and moderate values of the x-variable. Also in Fig. D1, a similar spline regression estimate is shown (the solid line); again, the smoothing parameter is chosen subjectively so that the two estimates are similar over much of the range of values of interest.

These data exhibit a long right-hand tail to the x-distribution. Consequently, the (constant bandwidth) kernel estimate follows the last few y-values very closely; the spline estimate, however, performs rather better (at least for the 4th-, 3rd- and 2nd-last datapoints). This amply illustrates Dr Silverman's comments (Silverman (1984a) and Section 3 of the current paper) on the (asymptotic) equivalence of the spline method to a variable kernel estimate, in which a larger bandwidth would be used in the tail. Indeed, the insistence even of the spline smoother on getting very close to the last y-value reflects the breaking down of that equivalence close to the boundary

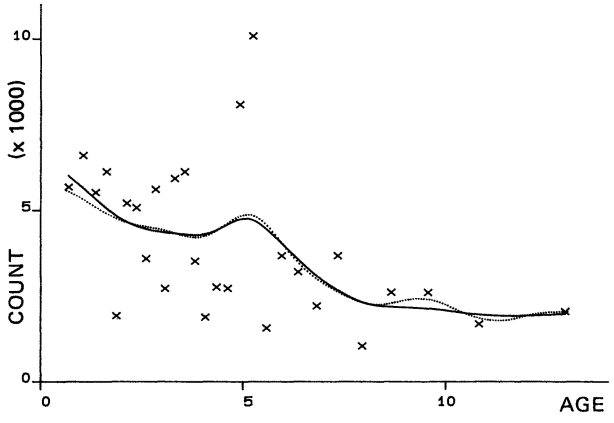

Fig. D1. ..... kernel estimate, —— spline estimate.

of the interval of interest. The spline approach has an advantage even over the variable kernel method in that the variation in the bandwidth is decided implicitly in a sensible way.

The second advantage of spline regression over the kernel approach lies in accommodating constraints—a question raised by Dr Critchley in his contribution. For these data, there are physical reasons why the curve must go through the origin. There is no satisfactory way of doing this in the kernel case; *ad hoc* methods such as adding several points at (0, 0) can give some sort of approximation to this, but are not very good (see, for example, the dotted line in Fig. D2 corresponding to adding 4 points at the origin). Conversely, however, there is a perfectly natural

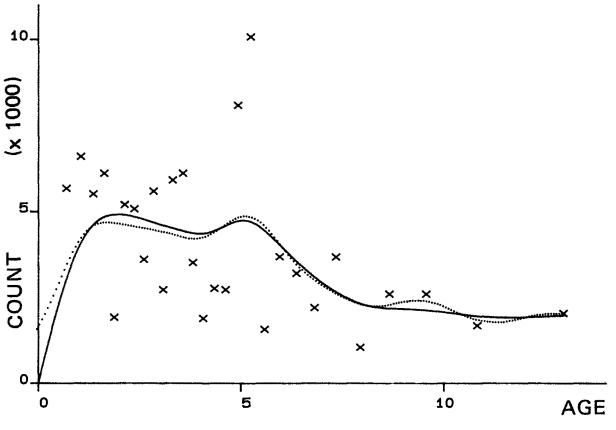

Fig. D2. ..... kernel estimate with 4 additional points at (0, 0), —— spline estimate constrained to go through (0, 0).

extension of spline smoothing to cope with such constraints; this amounts to a combination of spline smoothing with the well-known method of spline interpolation. Computationally, there is no further programming to be done since, as de Boor (1978 p. 274) points out, interpolation to a given point can be achieved by a weighted spline smoothing (of the data plus constraint point) with the weight corresponding to the constraint point set to a very large value. The resulting estimate for this dataset is the solid line of Fig. D2; this does, indeed, give a much more satisfactory curve. Perhaps the first peak still looks a little too low, but this would seem to be a consequence of insisting on fitting a smooth function to these data; it may well be that a sharp peak is appropriate in this case.

Professor A. F. M. Smith (University of Nottingham): In his Bayesian model (Section 6.1), Dr Silverman has specified a class of prior distributions concentrated on the space of natural spline functions with knots at the design points. In this way, he has succeeded in obtaining a finite-dimensional version of the curve-fitting problem with obvious attractions from a computational point of view.

There is, however, a rather drastic implicit restriction of the space of possible curves, as can be seen most clearly by considering the limiting case of zero errors ($\sigma^2 \to 0$, in the author's notation). In this case, the posterior concentrates entirely on the unique natural spline function interpolating the data points, and the posterior variance is zero everywhere along the curve, no matter how far it is extrapolated beyond the data configuration. This elimination of posterior uncertainty makes me feel uneasy, although the problem is obviously less acute if the method is to be used only for interpolatory purposes with large data sets possessing a rather dense set of knots.

With small data sets, even in the case of zero errors we should surely require that the posterior variance for curve values be strictly positive at all non-design points. This is so, for example, if we use the integrated Wiener prior over the completed space of functions with square integrable second derivatives (see Wecker and Ansley, 1983). Moreover, since the prior densities implicit in the two methods are proportional, this difference in qualitative behaviour is due entirely to the restriction placed on the space of functions by the author's method.

A somewhat different approach can be developed by noting that curve-fitting using a $(d-1)$st degree polynomial corresponds to the rather rigid prior belief that the dth and $(d+1)$st derivatives of the function are zero everywhere. It is interesting to consider what happens if this assumption is weakened by specifying instead a class of stochastic processes which reflect beliefs that these derivatives are, say, small, continuous and differentiable everywhere. These kinds of ideas are being explored by my student Martin Upsdell in a Ph.D. thesis being written at Nottingham and he will give a brief account of work in progress.

Mr M. P. Upsdell (University of Nottingham): Consider specifying prior distributions on functions by specifying a function $C(x, y)$, giving the prior covariance between the values of the dth derivatives of the function at the positions x and y. The prior mean of the dth derivative will be taken to be zero. The space of functions considered is that of the reproducing kernel Hilbert space with kernel $C(x, y)$, which consists of the completion of the space of functions of the form

$$f(x) = \Sigma\, a_i C(x, b_i)$$

for some finite collection of constants a_i, b_i. It can be shown that any function f in this space is continuous in mean square if and only if C is continuous, and f is differentiable in mean square if and only if C is differentiable. Thus choosing a covariance function for the dth derivative which is continuous and differentiable will ensure that the probability is concentrated on functions with finite dth and $(d+1)$st derivatives. Any parameters used to specify the function need not be prespecified but can be estimated from the data by the assignment of "vague" hyperpriors. This methodology does not require large data sets with dense knots, nor does it use asymptotic arguments. It also provides intuitively reasonable answers when we pass to the limit case of zero errors.

If the covariance function is Markov, then the Kalman Filter method of Wecker and Ansley can be used to estimate the parameters with order n computations per iteration. However, if the covariance function is not Markov, it requires the inversion of the covariance matrix with order n^3 computations per iteration. We are thus faced with a dilemma. Either we can ensure order n computations, or we can specify a prior which concentrates the probability on functions with $(d+1)$st derivatives, but not both. Perhaps the solution is to obtain access to a parallel processor, since parallel processors can invert the required matrices in order n steps.

As an illustration of a covariance function that might be used, consider the function

$$C(x, y) = \frac{\tau^2}{r} \exp\left(\frac{x-y}{r}\right)^2$$

suggested by O'Hagan (1978). An idea of the relative probability given to different functions can be obtained by considering limiting values of the parameters τ^2 and r (which, in practice, can be estimated from the data, thus providing a form of data-adaptive solution): (i) as $\tau^2 \to 0$, the variance of the dth derivative tends to 0 and the probability becomes concentrated on the space of polynomials of order $d-1$; (ii) as $r \to \infty$, the correlation between the function values tends to

one, the dth derivative tends to a constant and the probability becomes concentrated on the space of polynomials of order d; (iii) as $r \to 0$, the covariance function tends to the Dirac delta function, the probability becomes concentrated on those functions which do not have a dth derivative, and the mean function becomes the natural polynomial spline.

Approaches which consider the Wiener process for the $(d-1)$st derivative correspond to using a covariance function $C(x, y) = \tau^2 \min(x, y)$ for the $(d-1)$st derivative or equivalently, $C(x, y) = \tau^2 \delta(x-y)$ for the dth derivative, where δ is the Dirac delta function (cf. the case $r \to 0$ above). This leads, of course, to the apparent paradox which the author refers to at the beginning of Section 6.1. It might be argued, however, that this is of little practical consequence; in much the same way as the exclusion of the set of rationals by a normal distribution on the real-line causes us no real worry.

Dr P. J. Diggle (University of Newcastle and CSIRO, Canberra): It is a pleasure to add my congratulations to Dr Silverman on his excellent paper. I would like to take up the issue of serially correlated errors in non-parametric regression, which Dr Silverman mentions briefly in his discussion of the motorcycle impact data.

Suppose that we replace (1.1) by an autoregressive formulation,

$$Y_i - g(t_i) = \rho\{Y_{i-1} - g(t_{i-1})\} + \epsilon_i: \; i = 1, \ldots, n, \tag{1}$$

where the design-points t_i form a time-sequence. Consideration of the usual weighted least squares criterion for estimating $g(t)$, conditional on Y_1 and assuming ρ known, suggests replacing RSS(α) in equation (4.4) by

$$\mathrm{RSS}(\alpha, \rho) = \sum_{i=2}^{n} [\{Y_i - g(t_i)\} - \rho\{Y_{i-1} - g(t_{i-1})\}]^2. \tag{2}$$

As an example, I simulated (1) with $n = 100$, $\epsilon_i \sim N(0, 1)$, $\rho = 0.75$, design points t_i equally spaced over the interval (0, 1) and $g(t) = 10(1 + t)$. Setting $\rho = 0.75$ in (2) led to the choice of $\alpha = 0.00044$, whereas optimization of (4.4), which effectively sets $\rho = 0$ in (2), gave $\alpha = 0.000\,00034$. Fig. (D3) shows that the smaller value of α gives an estimate $\hat{g}(t)$ which follows

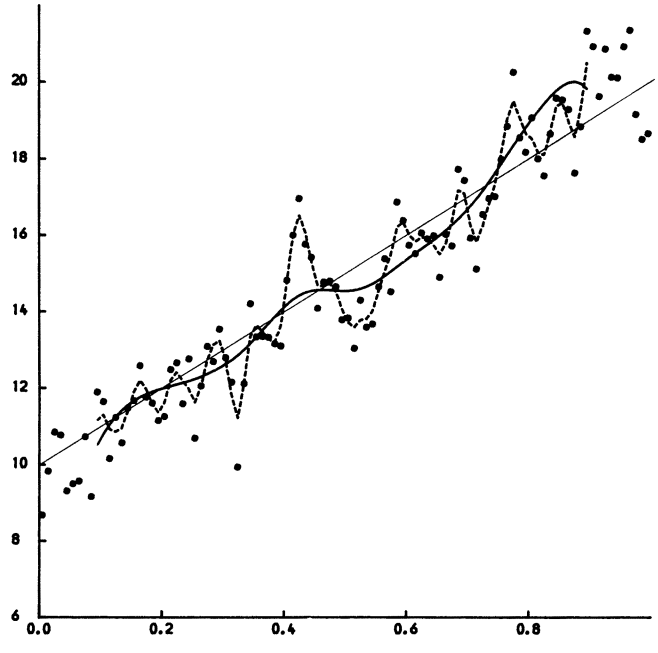

Fig. (D3). Simulation of linear $g(t)$ with autoregressive errors, $\rho = 0.75$. Fine solid line shows $g(t) = 10(1 + t)$. • data; - - - - $\hat{g}(t)$ assuming $\rho = 0$; —— $\hat{g}(t)$ assuming $\rho = 0.75$.

the data rather too closely, whereas the larger value gives a much better approximation to the underlying linear trend.

Obviously, one should not read too much into a single, small-scale example. In particular, the autoregressive equation (1) is but one of many possible formulations of serially correlated error structure. Nevertheless, the example does illustrate the important point that ignoring serial correlation will generally lead to undersmoothing. This seems inevitable since, in a non-parametric framework and without controlled replication, it is impossible to maintain a formal distinction between a deterministic regression curve $g(t)$ and a $g(t)$ which is a realization of a stationary random process.

For essentially the same reason, any attempt to estimate the autocorrelation structure from a single set of data seems bound to run into difficulties. An exception to this would be if physical considerations suggest that $g(t)$ can be assumed constant over some part of the design region – the initial portion of the motorcycle impact data might be a case in point.

Dr P. J. Green (University of Wisconsin-Madison): I congratulate Dr Silverman on an excellent paper that will surely encourage the practical application of spline methods.

The definition of the prior in Section 6.1 remains somewhat implicit, although the author has clarified the issue by reducing the dimensionality from ∞ to n. B-splines have attractive features important for numerical computation, but an alternative basis aids interpretation without altering the prior.

Let ξ denote the vector $(g(t_i))$, then (6.5) is equivalent to a two-stage model in which ξ is Normally distributed with $E(\xi_i \mid \beta) = \beta_0 + \beta_1 t_i$ and var $(\Delta\xi) = \sigma^2 \alpha^{-1} I_{n-2}$, and β_0, β_1 in turn have a vague prior distribution. Here Δ is any $(n-2) \times n$ matrix of rank $(n-2)$ such that $B^{\mathrm{T}} \Delta^{\mathrm{T}} \Delta B = \Omega$, and can be chosen so that it is close to being proportional to the second-difference operator when the $\{t_i\}$ are equally-spaced. It *would* be this operator if the roughness penalty had the discrete form $\Sigma\,(g(t_i) - 2g(t_{i+1}) + g(t_{i+2}))^2$ as in the original proposal by Whittaker, and as used by Green, Jennison and Seheult (1983) in the analysis of agricultural field experiments. It is quite easy to visualize this prior, with its expectation linear in t and independent perturbations $\Delta\xi$ generated almost sequentially with t because of the form of Δ.

It also helps, I think, to make explicit the linear regression that represents the "perfectly smooth" curves that are not penalized, and to which the smoothing method is invariant. Other fixed effects may be included in a semi-parametric model $y_i = \Sigma\, x_{ij}\theta_j + g(t_i) + \epsilon_i$, as in Green, *et al.* (1983), and the partial splines of Wahba (1984).

One could question whether this linear mean is an appropriate prior assumption for the example illustrated by Fig. 3: surely here the context might suggest other "perfectly smooth" curves, from which a special-purpose roughness penalty could be constructed?

If one blurs the distinction in status between the prior and error distributions, this becomes a linear model with two variance parameters. Among various methods for the well-known problem of estimating the variance ratio α, the REML approach (Patterson and Thompson, 1971) seems attractive, and Robin Thompson has pointed out a close parallel between this and GXV.

I have recently compared several criteria using both variety trial data and simulation: automatic choices of α are fairly similar (Green, 1985). What differences there are seem to have negligible influence on the estimation of fixed effects; if extraction rather than identification of the smooth component is the aim, precise choice of the smoothing parameter is much less important.

Dr C. Jennison (University of Durham): I would like to join the other discussants in thanking Dr Silverman for a most interesting paper.

I am, however, worried by the proposed method for estimating the maximum gradient of a curve: it seems very optimistic to hope to treat such a problem in a nonparametric framework. Infinitesimal changes in a curve can produce large changes in its gradient; thus, whilst I do not question conclusions concerning single values of the $g(t)$ fitted by the spline smoothing method, properties of the derivatives of the fitted g do not strike me as reliable.

The method of obtaining Bayesian confidence intervals for the maximum gradient by simulating from the posterior distribution is influenced greatly by the choice of prior. The prior described in Section 6 is concentrated on curves of a particular form and a small perturbation of such a curve, which would be undetectable in practice, can change its maximum gradient considerably

One could argue that, in some experiments, the gradient at a point is only of interest if it is

sustained over a certain short interval of the t-axis. This leads us to consider the maximum average gradient over a small interval, which is proportional to the maximum increase in g over such an interval. Thus, in the microbiological example, one might be interested in the largest increase in log concentration over, say, a one hour, thirty minute or one minute period. Using the spline smoothing approach the maximum such increase can be estimated from the fitted g in an obvious way. In addition, if *simultaneous* confidence intervals for $g(t)$ over the whole range of t-values were formed, a confidence interval for the maximum such increase could be obtained by considering all curves fitting inside this confidence "envelope". I suspect that this method would produce a rather wide interval. In their own way, standard parametric methods automatically smooth over *very* local behaviour and they may well be preferred for these purposes. The fitted spline smoothing curve would still be useful in suggesting a particular parametric model or indicating the region of the t-axis over which one is most interested in fitting a parametric model.

Dr A. O'Hagan (University of Warwick): This is a most impressive paper, full of interesting material, and so there is a number of things I would like to comment on. However, since previous speakers having referred to several of these, I can keep my comments brief. The most exciting aspect of the spline approach to smoothing is its computational advantages, further enhanced by Dr Silverman's work. When I think that my own attempt on smoothing, presented here seven years ago (O'Hagan, 1978), required the inversion of an $n \times n$ matrix, a reduction to order-n computations is quite startling. Furthermore, as a Bayesian I applaud Dr Silverman's use of Bayesian methods, and this makes me even more keen to try his approach.

But this is where the difficulties start. The Bayesian specifies his prior distribution so as to model his own prior beliefs. Convenience and tractability play a part, but Dr Silverman's prior distribution (6.5) seems to have been specified purely for convenience. Its role is to enable him to give a "Bayesian" justification for doing what he had already decided to do. Surely a Bayesian who wishes to use Dr Silverman's model would want to choose the parameters of his prior, in this case Ω, to reflect his own prior information. Yet it seems that if we are allowed an arbitrary Ω then the nice spline algorithms will be unavailable and we must obtain the posterior mean through (6.9). We are back to square one—$n \times n$ matrix inversion, order—n^3 computations.

Nor is this the only difficulty. Specifying the prior information will be difficult because of the unnatural form in which it is required. What do I know about the coefficients of the basis splines of the regression function? I suspect that even if I had far greater knowledge and experience of splines this would be a very difficult question, so what hope has the average practitioner?

In conclusion, this paper tantalizes me. The methods are exciting, but, as a Bayesian, I would very much like to know what the model means and how I can use it. Also, does the model (as opposed to the computational algorithm) extend to higher dimensions as easily as in Section 3 of my paper?

Professor A. P. Dawid (University College London): Minimizing (2.1) is the Lagrange multiplier equivalent of minimizing $\Sigma\{Y_i - g(t_i)\}^2$ subject to $\int g''(x)^2\,dx \leqslant \beta$. In performing cross-validation, it would be possible to use β, rather than α, as the parameter of our prescription (this is very loosely analogous to concentrating on the size of a hypothesis-test, rather than its critical likelihood ratio). For given data, there will be a one-to-one relationship between α and β. However, since this relationship is data-dependent, and thus changes when an observation is omitted, cross-validatory assessment of β might differ considerably from that of α. Are there any rational arguments for choosing between these two approaches?

One could also argue that, to ensure sensible dependence on sample-size n, (2.1) should be replaced by $S(g) = n^{-1}\Sigma\{Y_i - g(t_i)\}^2 + \gamma \int g''(x)^2\,dx$, and that the value for γ chosen from cross-validation on subsets of size $n-1$ should be used. This is equivalent to scaling up the selected value of α by a factor $n/(n-1)$. Might this small correction improve performance appreciably? Note that no such adjustment is required if β is taken as the parameter.

Dr P. Prescott (Southampton University): I should like to join my colleagues in thanking Dr Silverman for an interesting paper, entertainingly presented. However, I was a little surprised to see no direct reference to Cleveland's (1979) work on iterative robust locally weighted regression, although related references are mentioned in Section 8 of the paper. The objectives are the same and there are similarities in the approach to producing a smoothed representation of the data.

Cleveland's method uses an automatically adjusted window passing through the data. An appropriate model is fitted using weighted least squares within this window and each local analysis is used to predict a single observation. In this way an initial set of residuals is produced. The procedure is then repeated, iteratively, using a further set of weights produced by applying robust M-estimator weights based on the residuals produced at each step. Two steps are usually sufficient for convergence of the robust weights. The main difference between the two approaches seems to be the local model employed. Cleveland conjectures that, although any polynomial may be used, a simple straight line is adequate in most practical cases.

Dr Andrew Walden and I (1983) have used this method in some recent work on smoothing annual maximum sea levels at selected English ports in order to adjust the extreme sea levels prior to further analysis. We should be interested in knowing whether any comparisons have been made between these two approaches.

The following contributions were received in writing, after the meeting.

Dr Craig F. Ansley (University of Chicago): Dr Silverman's paper is an excellent review of curve fitting by splines using the cross-validation approach of Craven and Wahba (1979). However, he has ignored a large section of the literature which provides an enlightening alternative view of the problem. I refer to the work of Weinert, Byrd and Sidhu (1980; see also the references therein) who show how a smoothing spline can be expressed as the conditional expectation of a stochastic process observed with error, and, further, that the stochastic process has a simple state space representation that allows conditional expectations and variances to be obtained efficiently through standard filtering and smoothing algorithms. Wahba (1978) also has a stochastic model, although she does not use it as a computational device, and Ansley and Kohn (1984) show that the two approaches are equivalent. A form of Wahba's model with diffuse initial conditions is used by Wecker and Ansley (1983).

An area in which the state space formulation is particularly useful is in establishing error estimates for fitted curves. Dr Silverman follows Wahba's (1983) Bayesian approach which states that the joint prior distribution of the values on the curve at the observed argument values is the same as if the curve were generated by Wahba's (1978) stochastic model. Their prior distribution thus depends on the arguments; if another observation were available, the prior would change. A prior chosen independently of the arguments would state that the joint distribution of the n values of the curve at *any* n arguments is the same as if the curve were generated from Wahba's model. This leads to exactly the results produced by Weinert *et al.* (1980) and Wecker and Ansley (1983). The computational advantages of this approach are considerable. Use of the underlying stochastic model also gives direct results for a number of associated problems: higher order splines (Wecker and Ansley, 1983); smoothness properties (Kohn and Ansley, 1983); derivative estimation and models with concomitant variables (Ansley and Wecker, 1983). It also leads naturally to a fully Bayesian analysis of the model, as discussed in a forthcoming paper.

Professor A. C. Atkinson (Imperial College, London): Like other speakers I thank Dr Silverman for a lucid and interesting paper. I have two technical comments and one query about the data.

1. In Section 5.2 Dr Silverman uses the studentized residual. Belsley, Kuh and Welsch (1980) and Atkinson (1981) give arguments in favour of the *deletion residual* (alas called studentized by Belsley *et al.*) in which the estimate $\hat{\sigma}$ is replaced by the deletion estimate $\hat{\sigma}_{(i)}$. Among other properties this has the advantage that, for the normal theory linear model, the distribution is Student's t. Interestingly, the deletion residual comes from the likelihood ratio test for the presence of a single outlier. The studentized residual is obtained from the score test in which, of course, all parameters are estimated under the null hypothesis.

2. The analysis of the Old Faithful geyser data provides, I feel, a clear demonstration of the exploratory power of smoothers. For parametric modelling I would prefer something which is again smoother, but now in the sense of a curve for which the two straight lines of Fig. 5 form asymptotes. Testing hypotheses in this smooth model might eliminate the problems of separate families of hypotheses mentioned at the end of Section 5.3.

3. Inspection of Fig. 2, particularly in the region 30–40 ms, suggests that the data may not be a single time series but are rather the superposition of, perhaps, three series. An argument in favour of plots of signed residuals, rather than the squared residuals of

Fig. 4, is that this pattern might be more clearly displayed. If the data do consist of a superposition of similar series which are slightly displaced horizontally, one would expect the variance to be highest where the mean value is changing fastest. Would Dr Silverman please be more forthcoming about the structure of his data?

Professor G. A. Barnard (Retired): Dr Silverman's use of "automatic" instead of "objective" to describe his suggested procedure for obtaining an initial value for the smoothing parameter is to be welcomed provided the two words are not taken to be opposites. The mere fact that a procedure can be made part of a computer routine does not, of course, make it objective. But all the procedures Dr Silverman discusses are objective in that they reflect known or checkable properties, qualitative or quantitative, of the data. His Bayesian model of Section 6 embodies known qualitative aspects of the data in that known mechanical properties of (simulated?) flesh and bone justify our assumption, for the motor cycle crash data, that low values for the second derivatives are more plausible than high values. And cross-validation for the smoothing parameter also reflects known homogeneity properties of the data although, in some cases, on looking at the resultant curves we may realize we know of other qualitative features which render the value arrived at unsuitable.

All our statistical procedures should aim at objectivity in the sense that any assumptions we make should be clearly stated and should at least be empirically checkable. It would not be necessary to stress this, were it not for the fact that some "personal-Bayesian" statisticians give the impression that purely personal willingness to lay coherent bets may be all that is needed to justify some assumptions.

Mr P. J. Bates (Scicon Ltd): Dr Silverman's paper is very interesting and practically important. There are several points I would like to make and on which I would welcome Dr Silverman's comments.

Firstly, a hypothetical question—if spline smoothing can be extended to be practicable in the multidimensional case, can this be done in such a way to allow the dependence on one or more of the explanatory variables to be constrained to be of a particular parametric form? Possibly, in the case of linear dependence, some progress could be made by noting that, in one dimension, setting the smoothing parameter α to ∞ corresponds to linear regression. If possible at all in what situations would the combination of parametric and non-parametric regression be viable?

Secondly, the fact that the method can be extended easily to other forms of error structure as shown in Section 8.1 is heartening. An application in which both this and what is envisaged above would be useful arises in a military context. This concerns prediction of the probability of success of a homing weapon activated at a position (x, y, z) relative to its intended target. The dependence of this probability on (x, y, z) may be too complex to rely on a parametric model and there may be other discrete or continuous covariates to take into account.

Finally, Dr Diggle has illustrated the importance of taking into account dependence between the observations when the degree of dependence, in this case the serial correlation coefficient ρ, is known. His example also demonstrates that the value of ρ can drastically affect the amount of smoothing chosen by the cross-validation method. How should one proceed when dependence is suspected, but the value of ρ is unknown and must be estimated? The particular application I have in mind is to Monte Carlo simulation experiments in which the variance reduction technique of using the same random stream for different values of the input variable is employed. Dependence between the observations is deliberately introduced with the aim, usually, of estimating some functional of the curve more precisely, but it is not known how much dependence and so how much it actually improves the precision of the estimate or what the "best" curve is.

One approach may be to make a joint cross-validatory choice of α and $\hat{\rho}$ but I suspect that this may not give sensible results. Also one would then be placing α and $\hat{\rho}$ on the same footing philosophically: would α be regarded as a parameter estimate or $\hat{\rho}$ as a choice? What is the distinction anyway?

An alternative may be to apply the iterative procedure originally suggested by Cochrane and Orcutt (1949) for linear regression with autocorrelated errors. This would involve guessing the value of ρ, selecting (automatically or not) α given this value of ρ, computing the usual estimate of ρ with this α, selecting another α and so on, it is to be hoped, to convergence. Will this work?

Is there some more sensible way of attacking this problem?

Mr J. E. Besag (University of Durham): Dr Silverman's paper is particularly welcome to one who has in the past avoided splines, though I was a little sorry to see the paper conclude perhaps gloomily with the problems of multivariate design points. For univariate curve fitting, my usual non-parametric approach is with a pencil: I recognize that this does not allow statistical inferences to be made but equally I am unconvinced by the somewhat arbitrary Bayesian assumptions in Section 6 of the paper. Of course, in many problems, an automated solution is required. But this did not seem to be the case in Dr Silverman's numerical examples, which I would like to consider further. It may be unfair to treat them too seriously but I am unclear, and this is a general point, at what stage of development one ought to abandon the luxury of purely illustrative examples.

Thus, the motor cycle impact data confirm that simple spline smoothing copes poorly with a discontinuity in derivative and that taking account of very marked heteroscedasticity improves the resulting picture; of course, the performance of simple smoothing is affected also by the length (number of points) of the recorded interval of (virtually) constant speed prior to impact. In the practical context, my tendency would be to adopt a parametric approach, based on a second-order differential equation whose fitted coefficients could be of intrinsic interest; the point of impact might generally require formal estimation but here this does not seem necessary. Residual structure could suggest more detailed analysis, though this would probably require further data. I certainly see no evidence of the strange behaviour depicted in Fig. 6; surely, acceleration returns essentially to zero, even if one accepts the Bayesian pointwise confidence bands?

The geyser data provide another interesting numerical illustration, though I did not initially understand why a regression was being fitted to what seemed to be two blobs: Cook and Weisberg (1982, p. 40) provide a light-hearted(?) explanation and, for practical purposes, the differences between the various predictors is negligible, given the variability in the data. I am intrigued by the paucity of eruptions lasting between 2½ and 3½ minutes and wonder whether Dr Silverman has any further information.

The third example, I would again treat parametrically: here a logit curve would seem the obvious candidate. One might exclude the observation at 11 hours as an outlier, though this seems somewhat arbitrary and, in any case, should make negligible difference, against experimental error, to the estimate of maximum gradient. To use the approach of Section 7 seems to me to pander to the availability of computational complexity and, like Dr Jennison, I am concerned about the effects of local characteristics, whether real or apparent; this problem does not arise for the extrema in the first example.

My conclusion from this very useful paper is not to dismiss spline smoothing from statistical analysis, nor its potential for surface estimation, but to ask for real applications in which its use is more substantial and in which less sophisticated non-parametric smoothers, such as those of Velleman (1980), are insufficient. Comparisons with alternative sophisticated techniques, for example Cleveland (1979), would also be informative, especially where these can be easily adapted to multivariate design points. Examples might require very large data sets, which are difficult to present in a journal, but I wonder whether Dr Silverman can be persuaded to provide an additional application in his reply to the discussion: I feel sure this would be helpful to his cause.

Dr G. Collomb (Université Paul Sabatier, Toulouse): I would like to congratulate the author on his most interesting paper, which I hope will stimulate the use of nonparametric methods in applied statistics: this paper is a good advertisement because it involves a minimum of knowledge in probability theory and asymptotic statistics.

As for the basic idea (Section 2), we note that the classical Nadaraya–Watson estimate leads to estimations which are very rough curves (e.g. Watson, 1964); spline estimations are more beautiful curves. Are these estimates also more accurate or do we have to choose between beauty (smoothness) and accuracy? An answer to these questions can be obtained by the utilization of (3.1) and formulae giving the m.s.e. $E\{\hat{g}(s) - g(s)\}^2$, s fixed, of the kernel estimate: see Gasser and Muller (1979) for the curve fitting problem and Collomb (1977a) for the regression problem. The evaluations of the bias $E\{\hat{g}(s)\} - g(s)$ obtained in these two last papers depend mainly on the convexity of the estimated curve (e.g. Collomb, 1982, p. 176), so that convexities or concavities of the estimation suggest appropriate corrections in the reading of this estimation. This last remark completes Section 3 and also Section 6 which only concerns the case of a bias small with respect to the variance of $\hat{g}(s)$. The inference regions (6.12) suggest the following questions: is the normality of ϵ_i to be supposed, is it possible to "studentize" this result (i.e. estimate σ in 6.12), is

the homoscedasticity condition ($E\,\epsilon_i^2$ does not depend on t_i) necessary? Collomb (1977b) gives such completely studentized pointwise confidence intervals in the heteroscedasticity case for the kernel regression estimate: (3.1) suggests a utilization of this last result, which also concerns the multivariate case. Such intervals are very useful from an exploratory data analysis point of view (e.g. Collomb, 1982, p. 177), however they do not define really a confidence band as obtained by Liero (1982).

The presentation of the robust smoothing (Section 8.1) leads to the following question: can we say that $\hat{g}$ minimizing $S_R(g)$ defined by (8.1) is an estimate of g defined by (1.1)? I think that $\hat{g}$ estimates g_R defined by $g_R(t) = \operatorname{argmin}_g E[\rho\{Y - g(t)\}]$, for all $t \in \mathcal{R}$, and, for usual functions ρ, generally we have $g = g_R$ only if the distribution of ϵ_i is symmetric. Lastly we note that similar nonparametric robust estimates are developed by Collomb and Härdle (1984) in the case of a mixing process: this kind of result shows that the initial condition of independence (Section 1) can be relaxed without alteration of the properties of the considered methods, which can also be used (Collomb, 1983, 1984) in time series analysis and prediction.

Professor Noel Cressie (Iowa State University): Dr Silverman's underlying model is set out in (1.1). His aim is to estimate, nonparametrically, the curve or surface therein notated as g. However, when the given data $\{Y_i; i = 1, \ldots, n\}$ are to be analysed, there is a certain non-identifiability in the model specification. One could equally as well postulate

$$Y_i = a(t_i) + \eta_{t_i},$$

where $\{\eta_{t_i}; i = 1, \ldots, n\}$ is a sample from a stationary process $\{\eta_t; t \in \mathcal{R}\}$ on the real line, or more generally in space. Roughly speaking, the stationary model does not try to explain *all* the large scale variation through a trend surface, but rather leaves some of it with the error (Cressie, 1985). This allows prediction in regions where there is little or no data; linear prediction exploiting such spatial relationships has been called *kriging* in the mining and geostatistical literature (Matheron, 1963). In contrast, the spline approach presented by Dr Silverman, tries to estimate a trend surface smoothly. Which model one chooses, depends on the type of questions one is trying to answer; Watson (1972) observes that a good deal of geological data needs to be analysed using the dependent error model.

Finally, it is worthwhile mentioning that there is a formal relationship between kriging and splines. Watson (1984) is the best source for a clear exposition of the details. Without loss of generality, assume design points satisfy $0 < t_1 < \ldots < t_n < 1$. Add to the minimization of Dr Silverman's (2.1), the conditions $Y_i = g(t_i); i = 1, \ldots, n$, and $g(0) = g'(0) = g(1) = g'(1) = 0$. The former demand on exact interpolator, while the latter are there essentially for convenience, since the solution is directly comparable to simple kriging (i.e. kriging in the absence of drift). Without them, we would have to talk about universal kriging, which is best left for another occasion.

The solution is a cubic spline

$$g(t) = \mathbf{Y}^{\mathrm{T}}(C + \alpha I)^{-1}\, c(t),$$

where $\mathbf{Y}_{\cdot}^{\mathrm{T}} \equiv (Y_1, \ldots, Y_n)$, α is the smoothing parameter in (2.1), $c(t)^{\mathrm{T}} \equiv (c(t, t_1), \ldots, c(t, t_n))$, $C \equiv \{c(t_i, t_j)\}$, and finally the piecewise cubic polynomials are given by

$$c(t, t_j) = \{(t - t_j)\, I(t \geqslant t_j)\}^3/6 - t^3(1 + 2t_j)\,(1 - t_j)^2/6 + t^2 t_j (1 - t_j)^2/2; 0 \leqslant t \leqslant 1.$$

Watson (1984) observes that the matrix C is positive definite, and hence so is $C + \alpha I$.

Thus we could postulate having a zero mean process $\{Y_t; t \in \mathcal{R}\}$, observed at $t_1, \ldots, t_n$, and with covariance function $E(Y_t Y_u) = c(t, u), t \neq u; = c(t, t) + \alpha, t = u$; α is often called the nugget effect in the geostatistics literature. We want to predict Y_t using a linear function of the data $Y_{t_1}, \ldots, Y_{t_n}$; this is precisely kriging. By minimizing the mean squared prediction error, $E(Y_t - \Sigma_{i=1}^n \lambda_i Y_{t_i})^2$, with respect to weights λ, one finds the optimal λ^*

$$\lambda^* = (C + \alpha I)^{-1}\, c(t),$$

and hence the optimal predictor is

$$Y_t^* = \mathbf{Y}^{\mathrm{T}}\lambda^* = Y^{\mathrm{T}}(C + \alpha I)^{-1}\, c(t),$$

which is *exactly* the same equation as for cubic spline interpolation.

We are loath to interpret the cubic spline as a correlation function of stationary data with zero mean, but prefer to keep the relationship formal. It is merely expressing the type of non-uniqueness described at the beginning of this contribution. In two dimensions, the link is obtained through covariance functions like $h^2 \log h$, rather than h^3; see Dubrule (1983).

Having said all this, does Dr Silverman think it is possible to develop a Bayesian interpretation for kriging, by adopting his finite dimensional approach for splines?

Dr J. Cuzick (Imperial Cancer Research Fund Labs, London): I enjoyed very much Dr Silverman's lucid exposition of the theory of spline smoothing. However I do not find the material in Section 7 on the estimation of functionals entirely satisfactory, be they linear or non-linear. My point is made most transparent by consideration of an estimate for the fourth derivative at some fixed point (a linear functional), where $\hat{g}^{(4)} = 0$ except at knots where it is undefined. This extreme example highlights the need for a smoother estimate of g for approximating derivatives and suggests that efficient estimates should be based on a penalty function with terms containing higher order derivatives when g is known to possess higher order smoothness.

Professor J. Durbin (London School of Economics): In my review paper on time series analysis for the Society's 150th Anniversary Conference (Durbin, 1984) I referred briefly to the problem of seasonal adjustment and expressed the view that while existing seasonal-adjustment techniques require modernisation, methods based on the Kalman filter are ultimately more likely to be found preferable for the purpose than techniques based on ARIMA modelling. One of my reasons for this belief arises from the prevalence of changes in behaviour over time of economic and social time series. However, I have been greatly impressed by the suitability of the techniques discussed by Dr Silverman for the study of phenomena displaying slow structural change. Consequently I believe it would be of great interest to see the development of methods for the estimation of seasonal components in the presence of change over time analogous to those used in the paper for the estimation of trend. I would be interested to hear whether Dr Silverman would regard his approach as suitable for the seasonal adjustment problem.

Dr G. K. Eagleson (CSIRO, Australia): It has been a pleasure to read the clearly presented and timely paper of Dr Silverman. While he has noted three main purposes for which regression may be used, he does not mention what I believe to be another important use: the comparison of two data sets. Can the techniques of non-parametric regression outlined in his paper be used to that end? In particular, can the estimates of local variance be used to assess whether two fitted regression lines could be assumed to be the same?

Dr R. L. Eubank (Southern Methodist University, Texas, USA): Let me begin by thanking Dr Silverman for an interesting and thought-provoking paper. His use of regression type diagnostics for smoothing splines coincides with many of my own thoughts on this subject (Eubank, 1984a, b). The close connection between smoothing splines and polynomial regression leads one to believe that diagnostics appropriate for use with smoothing splines should resemble those currently in use by regression analysts. In this regard, it is well known that diagnostic procedures should include information about the design as well as the fit. Design diagnostics for smoothing splines are provided by the leverage values, $A_{ii}(\alpha)$. It can be shown that $0 \leqslant A_{ii}(\alpha) \leqslant 1$ and that a leverage value too near one indicates a sensitive point in the design where an observation will tend to dominate its own fit. A diagnostic which encompasses both information about an observation's leverage as well as its fit is (in Silverman's notation)

$$\text{DFITS}_i = | A_{ii}(\alpha)/(1 - A_{ii}(\alpha)) |^{\frac{1}{2}} | r_i |, \quad i = 1, \ldots, n.$$

This particular diagnostic indicator can be motivated from analogous quantities used in regression analysis and can provide valuable information over that available from measures focusing on residuals alone. Many other regression type diagnostics can also be suggested.

Concerning interval estimation, there are several alternatives to Silverman's method based on sample estimates of the influence curve for smoothing splines. One of these (cf. Wold, 1971) can be described as follows. For simplicity assume that $w_i = 1$, $i = 1, \ldots, n$, and let $\hat{g}^{[i]}$ and $\hat{\gamma}^{[i]}$ denote the smoothing spline estimate and vector of coefficient estimates, under the B-spline

basis, when (t_i, Y_i) has been deleted from the data. It can be shown that

$$\hat{\nu}^{[i]} = [\alpha\Omega + B^{\mathrm{T}}B]^{-1}B^{\mathrm{T}}\left[Y - e_i\,\frac{(Y_i - \hat{g}(t_i))}{1 - A_{ii}(\alpha)}\right],$$

where e_i is the ith column of the $n \times n$ identity matrix, which gives $\hat{g}^{[i]}(t) = \Sigma_{j=1}^{n}\hat{\nu}_j^{[i]}\beta_j(t)$. Given a functional Ψ we then define pseudovalues

$$P_i(\Psi) = \Psi(\hat{g}) - (n-1)\,(\Psi(\hat{g}) - \Psi(\hat{g}^{[i]})), \quad i = 1, \ldots, n,$$

and obtain the jackknife variance estimate (Efron, 1982)

$$S_{\Psi}^2 = [n(n-1)]^{-1}\,\Sigma_1^n\,\{P_i(\Psi) - \bar{P}(\Psi)\}^2,$$

where $\bar{P}(\Psi) = n^{-1}\,\Sigma_1^n\,P_i(\Psi)$. An approximate 95 per cent confidence interval for $\Psi(g)$ is provided by $\Psi(\hat{g}) \pm 2S_{\Psi}$. The computation of S_{Ψ}^2 simplifies considerably when Ψ is linear. Dr Silverman's approximations for the $A_{ii}(\alpha)$ have some obvious applications to jackknife interval estimation.

Jackknife confidence intervals do not require the assumption of normal errors and are more computationally expedient than the Bayesian approach when Ψ is nonlinear. It should also be noted (see e.g. Hinkley, 1977) that jackknife methods might be expected to be robust against nonhomogeneous error variances. It would be interesting to compare jackknife methods to Silverman's approach with estimated weights in this setting.

Drs W. Greblicki and M. Pawlak (Institute of Engineering Cybernetics, Wroclaw, Poland): We would like to point out one difference between various nonparametric estimates and we shall do it in the context of density estimation. For an unknown density f having p derivatives, Wahba (1975) suggested using a kernel K such that

$$\int y^i K(y)\,dy = 0, \quad i = 1, \ldots, p-1 \text{ and } \int y^p\,|\,K(y)\,|\,dy < \infty.$$

Then, for the kernel estimate $\hat{f}(x)$ of $f(x)$

$$E\{\hat{f}(x) - f(x)\}^2 \leqslant c_1 h^{2p-1} + c_2/nh, \tag{1}$$

c_1 and c_2 are positive and h is the smoothing parameter. Hence, for $h(n) \sim n^{-1/2p}$

$$E\{\hat{f}(x) - f(x)\}^2 = O(n^{-(2p-1)/2p}). \tag{2}$$

In turn, for the estimate $\bar{f}$ using the cosine orthonormal series

$$E\{\bar{f}(x) - f(x)\}^2 \leqslant c_3 N^{-(2p-1)} + c_4 N/n, \tag{3}$$

where N is the number of orthonormal functions in the estimate—see also Wahba (1975). For $N(n) \sim n^{1/2p}$, the rate of the cosine series estimate equals that in (2). Thus, generally speaking, the rates achieved for the kernel and the cosine series are the same. The same conclusion is true for mean integrated square error—see Rosenblatt (1971) for the kernel estimate and Greblicki and Pawlak (1984) for the orthogonal series.

Now let the density be analytic, i.e., let it have all derivatives (which is often the case) and let the kernel be selected in the way suggested by Wahba (1975) (now p is a number arbitrarily chosen by a statistician). Then, the right side in (1) and consequently the rate in (2) remain unchanged. For the cosine series estimate, however,

$$E\{\bar{f}(x) - f(x)\}^2 \leqslant c_3 N^{-1/\delta} + c_4 N/n \tag{4}$$

$\delta > 0$. Now, taking e.g. $N(n) \sim n^{\epsilon}$, $\epsilon > 0$, we get

$$E\{\bar{f}(x) - f(x)\}^2 = O(n^{-1+\epsilon}).$$

By selecting ϵ sufficiently small, one can obtain a rate better than that achieved for the kernel estimate. A similar property can be also observed for estimates employing other orthogonal series (see e.g. Greblicki and Pawlak, 1984).

In view of this it seems that, for analytic densities, orthogonal series estimates behave better than the kernel estimate. This difference between these two estimates is caused by the fact that the kernel is selected according to p, i.e. the number of existing derivatives of the unknown density, while in the orthogonal series estimate the kernel is uniquely determined by the

Christoffel-Darboux formula and gives a good fit of the estimate to smooth densities. Similar properties of both estimates can also be observed while estimating regression functions according to the model examined by Silverman in his paper. One problem, however, arises for analytic regressions (and densities): do spline estimates behave as the kernel or the orthogonal series estimate?

Dr Wolfgang Härdle (Johann Wolfgang Goethe–Universität, Germany): Dr Silverman's article on the spline smoothing approach to curve fitting is an excellent contribution to the understanding of data smoothing. He points out the various attractive features and shows in a variety of examples the wide applicability of spline smoothing. I found the clear and elegant discussion of Section 3, pointing out the relationships among spline smoothing and kernel regression, very stimulating.

My comments will address (a) the generalized cross-validation method (Section 4) and (b) the proposal of an automatic choice of the smoothing parameter in the case of robust spline smoothing (Section 8.1).

The generalized cross-validation method can be considered as a member of the smoothing parameter selection procedures:

"Choose α to minimize the score

$$S_n(\Xi; \alpha) = \mathrm{RSS}(\alpha)\, \Xi(n^{-1} \operatorname{tr} A(\alpha))."$$

Here Ξ denotes a "selection penalty" with expansion $\Xi(u) = 1 + 2u + \Xi''(\xi)u^2$. The generalized cross-validation score GXVSC(α) has penalty $\Xi(u) = (1-u)^{-2}$. A FPE-Type penalty $\Xi(u) = (1+u)/(1-u)$ (Akaike, 1970) or Shibata's (1981) $\Xi(u) = (1+2u)$ are also possible.

Note that

$$E\, S_n(\Xi; \alpha) = E \left\{ n^{-1} \sum_{i=1}^{n} \epsilon_i^2 + n^{-1} \sum_{i=1}^{n} (\hat{g}(t_i) - g(t_i))^2 + 2n^{-1} \sum_{i=1}^{n} \epsilon_i (\hat{g}(t_i) - g(t_i)) \right\}$$

$$\times \left[1 + 2n^{-1} \operatorname{tr} A(\alpha) + O((n^{-1} \operatorname{tr} A(\alpha))^2) \right]$$

$$= \sigma^2 + E\, n^{-1} \sum_{i=1}^{n} (\hat{g}(t_i) - g(t_i))^2 - 2n^{-1} \operatorname{tr} A(\alpha)\, \sigma^2$$

$$+ 2n^{-1} \operatorname{tr} A(\alpha)\, \sigma^2 + O((n^{-1} \operatorname{tr} A(\alpha))^2).$$

So, asymptotically minimizing $S_n(\Xi; \alpha)$ is the same as minimizing $n^{-1} \sum_{i=1}^{n} (\hat{g}(t_i) - g(t_i))^2$. This expansion also suggests that all possible selectors $S_n(\Xi; \alpha)$ are asymptotically equivalent. However, a Ξ with a large second derivative in a neighbourhood of zero could be preferred in order to penalize more for undersmoothing.

Which of the possible penalties Ξ should be applied in practice?

Denote the robust spline by $\hat{g}_R$. An automatic choice of the smoothing parameter by means of (8.3) is a natural extension of the cross-validation score XVSC(α).

Let $\rho(s; t) = \rho(s - t)$, $\psi = \rho'$ and let $V_n(\psi)$ be a consistent estimate of $E\psi^2 / E\psi'$. I propose the following score

$$W_n(\alpha) = 2n^{-1} \sum_{i=1}^{n} \rho(Y_i - \hat{g}_R(t_i)) + 2n^{-1} \operatorname{tr} A(\alpha)\, V_n(\psi)$$

as a smoothing parameter selector. The idea behind $W_n(\alpha)$ is that, by Taylor expansion,

$$E\, W_n(\alpha) = 2E\rho(\epsilon) + 2 \left\{ n^{-1} \sum_{i=1}^{n} E[\psi(\epsilon_i)\, (g(t_i) - \tilde{g}(t_i))] \right.$$

$$\left. + n^{-1} \operatorname{tr} A(\alpha)\, E V_n(\psi) \right\} + n^{-1} \sum_{i=1}^{n} E[\psi'(\epsilon_i)\, (g(t_i) - \tilde{g}(t_i))^2],$$

where $\tilde{g}(s)$ is the linear approximation to $\hat{g}_R$, as given by Cox (1983). The first term on the right hand side is independent of α, the second vanishes and the third term is the quantity of interest.

How is $W_n(\alpha)$ related to XVSC?

Professor D. V. Lindley (Somerset): This admirable paper makes the non-parametric fitting of curves approachable both by the mathematician, because of the elegance and simplicity of its argument, and by the practitioner, because of available numerical techniques.

The precise interpretation of the roughness penalty eludes me. The way it is introduced in Section 2, and the language used to describe it, suggests it is a loss function. The approach in Section 6 gives it a probabilistic description. Mathematically this confusion does not matter because only the product of loss and probability matters and therefore the product can be interpreted either way. But when we go beyond the technique to the standpoint (to use de Finetti's phrase) we must ask what it means in its probabilistic interpretation. First, it has probability one on cubic splines with knots at the t_i. In other words, change a t_i and the opinion changes drastically. This is very unrealistic but viewed as an approximation to an underlying distribution may not matter. Second, what does it say about these splines: what does it mean to say the γ_i are multivariate normal with zero means and precision matrix Ω? For example, what is the correlation between $g(s)$ and $g(t)$? Does it depend on how many knots there are between s and t or is it, as one would have in most applications, purely a function of $t - s$? Third, what does it say for extrapolation? It apparently implies linearity outside (t_1, t_n). If $\hat{g}''(s) > 1$ for all s in (t_1, t_n), would this be a reasonable inference? Statisticians should be encouraged to think about their model and I find it hard to be sure what the techniques so well presented here imply.

It would be interesting to compare tonight's methods with the non-parametric methods of Young (1977) using polynomials instead of splines. Notice that Young, working within the Bayesian framework, evaluated his equivalent of the smoothing parameter α, his λ, within that framework, though he had to use a modal value. Perhaps cross-validation could be replaced by a coherent method.

Dr J. S. Marron (University of North Carolina, Chapel Hill, USA): I very much enjoyed reading Dr Silverman's excellent presentation of recent advances in the spline smoothing approach to nonparametric regression estimation.

The asymptotic representation (3.2) of the spline estimator in terms of a kernel estimator allows nearly immediate application of the results of some unpublished papers by Härdle and Marron. In particular it may be shown, under reasonable assumptions, that if ISE denotes the Integrated Squared Error:

$$\text{ISE} = \int \{\hat{g}(s) - g(s)\}^2 \, ds,$$

and if $\hat{\alpha}$ denotes the minimizer of the cross-validation score XVSC, then the estimator g with the smoothing parameter taken to be $\hat{\alpha}$ has the compelling optimality property:

$$\text{ISE}(\hat{\alpha})/\min_{\alpha} \text{ISE}(\alpha) \to 1,$$

almost surely.

When the cross-validation scores GXVSC or AGXV are used to find $\hat{\alpha}$ more work is required to establish results like that above, but the difficulties should not be too great in view of the approximations presented in Section 4.

The fact that the local bandwidth given in (3.4) is a multiple of $n^{-1/4}$ is disturbing at first glance because traditional asymptotic theory calls for a multiple of $n^{-1/9}$, for this kernel. However, the above result says that this is not a problem if the bandwidth is chosen by cross-validation since the optimal selection of the scale factor α will automatically make the necessary adjustment. As Dr Silverman points out, it is the factor of $f(x)^{-1/4}$ in (3.4) which makes the spline estimator seem superior to a fixed bandwidth kernel estimator.

In Section 5, the problem of nonconstant variance is addressed by attaching weights to the sum of squares part of (2.1). Continuing along these lines one may wonder: does it seem worthwhile to also introduce a weighting scheme to the integral part of (2.1)? For example, in the case of the motorcycle impact data, it seems the estimator might be improved by having more smoothing for

times less than 15, and perhaps much less smoothing near what appears to be a "sharp corner" at about time 15.

A natural question regarding the material of Section 5.4 is: would it be profitable to choose the number of nearest neighbours, k, in (5.8) by a cross-validatory procedure? The dependence of the estimator on k certainly seems more tenuous than the dependence on α, yet a cross-validatory procedure seems more in the spirit of letting the data speak for themselves.

Drs H-G. Müller and U. Stadtmüller (University of Marburg; University of Ulm, Germany): There is a wealth of theoretically interesting and practically relevant results in the paper of Dr Silverman. We want to focus on one of his main results which shows that smoothing splines correspond essentially to a variable bandwidth kernel estimator. For kernel estimators, it is well known that under certain regularity conditions, using a kernel of order k (i.e. with $(k-1)$ vanishing moments) the optimal local bandwidth at a point t depends on t via the expression

$$\{\sigma^2(t)/g^{(k)}(t)^2 f(t)\}^{1/(2k+1)}, \qquad (*)$$

where $\sigma^2(t)$ denotes the local residual variance, g is the curve to be estimated and f the design density. Dr Silverman shows that cubic smoothing splines correspond to a kernel with $k = 4$ and adapt to $f(t)^{-1/4}$.

It is easy to devise variable bandwidth kernel estimators where the bandwidth varies according to $f(t)^{-\alpha}$, $\alpha \in [0, 1]$, and which should correspond to cubic smoothing splines if $k = 4$, $\alpha = 1/4$, whereas the asymptotically optimal value for α would be $\alpha = 1/9$. Preliminary simulation results show that the IMSE (Integrated Mean Squared Error) of such kernel estimators with bandwidth varying according to $f(t)^{-\alpha}$ is nearly the same for all $\alpha \in [0, 0.5]$. It is not clear whether the smoothing spline could be outdone by a kernel estimator adapting with an optimal α. At least in equidistant designs kernel estimators are according to simulation results not worse than smoothing splines.

For kernel estimators, it is possible to adapt also to $g^{(k)}(t)$ according to (*) by pilot estimators and this seems to be much more relevant with respect to IMSE than adaption to $f(t)$. It can be shown that by such an adaption to local curvature of g asymptotic IMSE is decreased and this could also be demonstrated for small sample sizes by simulations (Müller and Stadtmüller, 1984). Therefore we would rather prefer to use locally adaptive kernel estimators than smoothing splines, because they allow for much more flexibility in the manner and extent of adaption.

Further advantages of kernel estimates compared to smoothing splines are according to our experience: boundary effects can be taken care of (practical and theoretical relevance for global measures of deviation); multivariate extensions can easily be implemented, especially for a rectangular design; choice of different kernels allows to derive curve estimates of different degrees of smoothness, different rates of convergence and of derivatives; simplicity of the kernel method allows qualitative assessment of finite local bias and variance.

Dr F. O'Sullivan (University of California at Berkeley): I am glad to have the opportunity to contribute to the discussion of Dr Silverman's fine paper.

Generalized cross validation treats all points in the design space equally and so is likely to be unduly sensitive by very high leverage points. Maybe one interpretation of the asymptotic generalized cross validation score is that the knot density estimate implicitly produces a smoothing of leverage values in order to stabilize the original score? By directly considering leverage values, that is, diagonal elements of the hat matrix, perhaps we should be motivated to experiment with still other cross validation methods, especially in situations where there are extreme points in the design space or one is interested in criteria other than the predictive mean square error.

An alternative cross validation score for the situation in Section 8.1 could be motivated directly from the pseudo-process. One might consider a score of the form

$$n^{-1} \sum_{i=1}^{n} [\eta_i - \hat{g}(t_i)]^2 / [1 - n^{-1} \operatorname{tr} A(\alpha)]^2$$

where η_i are the pseudo-observations and $A(\alpha)$ is the hat-matrix for the "equivalent least squares problem" i.e. corresponding to the pseudo-observations and pseudo-weights. Huber (1979) suggests this score in the context of robust smoothing. An investigation of the score applied to

generalized linear models is given in O'Sullivan, Yandell and Raynor (1984). The basic finding is that the score performs well from the point of view of a weighted mean square error criterion, where the weights are related to the expected Fisher information obtainable at the design points. A simple algorithm for computing the score may be worth describing. Keeping with the notation in the paper, the hat-matrix for the equivalent least squares problem is given by:

$$A(\alpha) = B[B^{\mathrm{T}} WB + \alpha\Omega]^{-1} B^{\mathrm{T}} W,$$

where the diagonal of W contains the vector of pseudo-weights. Thus the trace term can be written as

$$\mathrm{tr}\, A(\alpha) = \mathrm{tr}\,(B^{\mathrm{T}} WBS^{-1}),$$

where $S = [B' WB + \alpha\Omega]$. Now using the Cholesky decomposition of S, as given in equation (7.1), the trace simplifies to:

$$\mathrm{tr}\, A(\alpha) = \mathrm{tr}\,(B^{\mathrm{T}} WB(L^{\mathrm{T}})^{-1} L^{-1}) = \sum_i \{\sum_k d_{ik} < c_k, c_i >\}$$

where $d_{ik} = (B^{\mathrm{T}} WB)_{ik}$ and $< c_k, c_i >$ are inner products between columns of L^{-1}. A further simplification occurs because the matrix $B^{\mathrm{T}}WB$ is band limited so that the terms in brackets are few – at most seven with the roughness penalty in the paper. Now expanding upon a trick in Elden (1977), the column inner products can be found recursively, starting with the last column, from the relation:

$$L^{-1} L = I.$$

It follows that the trace, and consequently the cross validation score, is directly computable in roughly $40n$ multiplications/divisions. More importantly, a similar technique can be used to calculate directly individual leverages: The ith leverage value is the norm of the ith column of the matrix $L^{-1} W^{1/2} B$. However the expression for this norm involves only a very limited number (at most three for the situation of the paper) of inner products between neighbouring columns of L^{-1}. So it turns out that the entire set of leverage values can be obtained directly with the same computational effort as the trace calculation. Among other things, these leverages could be used to compute the exact confidence intervals given in equation (6.10).

With regard to multivariate smoothing, a less general model for the multivariate surface is defined by letting some transformation of the expected response be a linear combination of one-dimensional functions of the independent variables or variables derived from the independent variables: see Stone (1984). Along with being practically very useful, there are definite computational advantages associated with these models. Now, ideally one would like to have methods with which to evaluate and compare alternative specifications for such models, so it would be interesting to study how the empirical Bayesian analysis of the paper might apply.

Drs R. L. Parker and J. A. Rice (University of California, San Diego, USA): We wish to contrast the approach of smoothing splines so clearly presented by Dr Silverman, with another approach to smoothing with splines, which we will call least squares splines. In this approach, a small number of knots ($p < n$) are specified and the unknown function is approximated by least squares from the linear space spanned by B-splines with those knots (deBoor, 1978, Chapter XIV). Asymptotic properties of this scheme have been investigated by Agarwal and Studden (1980).

Perhaps the principal advantage of this scheme over smoothing splines is that the linear system to be solved is of order p rather than n. Since the estimate is linear in the data, it is of the form given by equation (3.1) of Dr Silverman's paper. Although we do not know the asymptotically equivalent kernel in this case, for any given design the weight function can be easily determined by feeding a vector Y consisting of one 1 and the rest zeros into the estimation program. An example is given in Fig. D4. The most striking aspect of the figure is the presence of very pronounced side-lobes, which can give rise to oscillations in the fit unless the knot placement is done with a combination of care and luck.

We have used both forms of spline smoothing as low-pass filters to remove slow secular trends from time series. In this situation, since n can be quite large, least squares splines might seem especially appealing. However, it is sometimes impossible to enforce a high enough degree of smoothness. Figure D5 shows a data series consisting of the sum of three high-frequency components, Gaussian noise, and a linear trend. Suppose that we wish to filter out the high-frequency

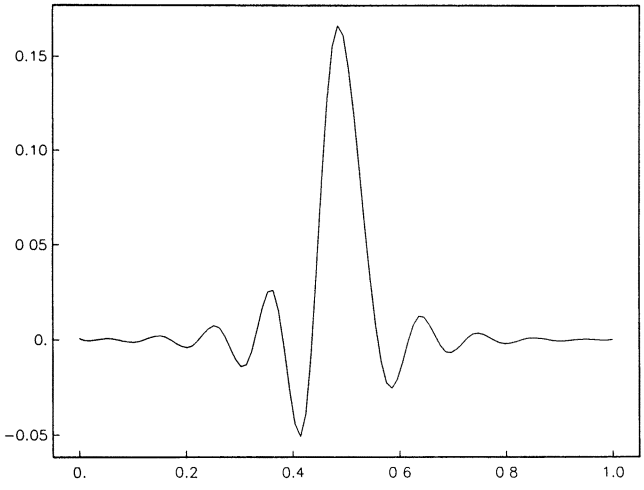

Fig. D4. The impulse response function corresponding to the 50th of 100 equispaced points in [0, 1] for a least squares spline with 20 equispaced knots.

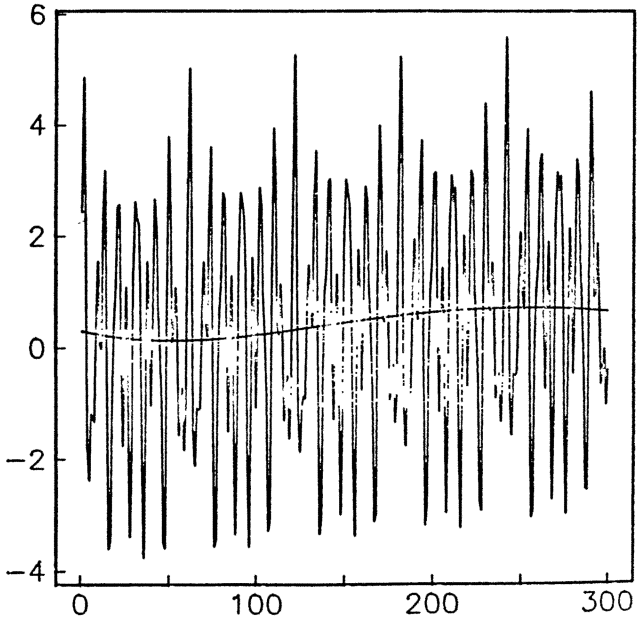

Fig. D5. The data series is the sum of a linear trend, three high frequency cosinusoids and Gaussian noise. The low pass filter is a least squares spline with a single interior knot.

components, leaving the linear trend intact. The illustration shows that with one interior knot, we were unable to recover the proper low-frequency behaviour. Another advantage of smoothing splines is that the degree of smoothness can be adjusted continuously.

This defect of least squares splines can be remedied by adding a penalty term involving the square of the second derivative to the objective function: then one is approximating the smoothing spline from a low-dimensional space of B-splines. It is permissible to use evenly spaced knots which considerably simplifies the programming. We have found this modification to be very satisfactory.

Dr D. A. Preece (Rothamsted Experimental Station): Much as I enjoyed this paper, I remain very uneasy about the analyses provided both by the author and by Cook and Weisberg (1982, pp. 40–42) for the geyser data. Faced with the two point-clusters, within the first of which there is very little visible indication of a relationship between the variates, I myself would not wish to consider any sort of regression relationship until I knew the time-sequence for the points. I would then also want to know how the values of the two given variates are related to those of at least a third, namely "time since *previous* eruption". Sadly, Cook and Weisberg give only the scatter plot of their own two variates, so other relationships cannot be deduced. The clustering seems likely to have a physical interpretation that might well become apparent from knowledge of the time sequence; that interpretation could then perhaps motivate a statistical analysis.

Dr A. E. Raftery (Trinity College, Dublin): One problem for which nonparametric spline regression may be useful, albeit with some modifications, is that of developing measures of intergenerational social mobility which are suitable for cross-national comparisons.

The results of social mobility surveys are usually reported in the form of two-way cross-classifications of respondent's occupation by father's occupation. The classification is often constructed by ranking occupations on an effectively continuous status scale and then discretizing. Different numbers and definitions of categories have been used in different surveys and different countries, making comparisons difficult. This suggests using as a mobility measure an estimate of a population parameter defined without reference to the way occupations are categorized, unlike most measures proposed to date, which do depend crucially on the categorization used.

Sociologists often wish to decompose observed mobility into "structural" mobility, forced by intergenerational changes in the occupational structure, and "relative" mobility. In order to assess the latter, we are led by the above considerations to define, for each individual, X to be the rank of his or her father in the population, and Y to be his or her own rank, each normalized so as to lie between 0 and 1. We then have the regression problem of estimating the conditional distribution of Y given X, or, equivalently, the joint distribution of X and Y (the equivalence being due to the fact that, assuming a continuous underlying scale, X is uniformly distributed between 0 and 1). Once this is done, measures of relative social mobility suitable for cross-national comparisons, such as $E(|X-Y|)$, are immediately available.

A simple approach to the estimation problem which seems to work well for the particular data sets analysed is outlined by Raftery (1983, 1985). However, the spline smoothing approach is much more elegant, but there are difficulties. Both X and Y are grouped, although, as pointed out by Good (1983), this should not present a major problem. Further, both X and Y are ranks rather than actual observations. Finally, a number of data sets are to be smoothed simultaneously in such a way as to preserve comparability. I should be grateful for any suggestions Dr Silverman may have as to how his results could be extended to deal with this case.

Professor P. M. Robinson (London School of Economics): The author assumes that the residuals in his model (1.1) are independent, which is not necessarily reasonable where data that are ordered over time or some other dimension are concerned. By analogy with known results for kernel and other regression estimators, the presence of serial correlation may not affect some of the asymptotic properties of the author's estimators. However, unless n is suitably large, serial correlation seems likely in practice to affect adversely the precision of curve fits based on the author's objective functions, to have implications for the choice of smoothing parameter, and to suggest the use of alternative curve-fitting formulae which, in a manner possibly analogous to that of generalized least squares, explicitly takes account of serial correlation. Irregular spacing of the t_i would cause complications here, but in the parametric setting methods are now available for efficient estimation in the presence of serial correlation, and for testing for serial correlation, in the presence of irregularly-spaced time series observations. The author refers briefly in Section 5.3 to an analysis of the time series nature of the motor-cycle impact data, and it would be interesting

to know whether he has carried out such an analysis on these data or whether, in fact, the observed residuals appeared to support his assumption of independence.

Dr A. H. Seheult (University of Durham): I have two comments. The first relates to choice of which derivative to penalise and connections with methods of analysing agricultural field experiments arranged in separate lines of equally-spaced contiguous narrow plots. It is sometimes appropriate to assume an additive decomposition for yields $\mathbf{y}$ of the form $\mathbf{y} = \mathbf{D}\tau + \xi + \eta$, where ξ represents a smooth fertility gradient, $\mathbf{D}$ is the design matrix for treatment effects τ and η represents independent homoscedastic errors: see Wilkinson *et al.* (1983). A current issue amounts to a choice between using first differences $\Delta_1 \mathbf{y}$ or second differences $\Delta_2 \mathbf{y}$ in some plausible weighted least squares analysis: see Green *et al.* (1983), Besag and Kempton (1984) and Wilkinson (1984). The choice of analysis is essentially equivalent to a choice between penalising first differences or second differences in an appropriate penalty function, similar in form to (2.1), for some interpretation and choice of the "tuning constant" α.

One resolution of the above issue would be to penalise both first and second differences in a penalty function of the form $\alpha_1 \xi^{\mathrm{T}} \Delta_1^{\mathrm{T}} \Delta_1 \xi + \alpha_1 \xi^{\mathrm{T}} \Delta_2^{\mathrm{T}} \Delta_2 \xi + \eta^{\mathrm{T}} \eta$. Presumably, cross-validation, or a similar method, could be used to provide an automatic choice of (α_1, α_2); but is it worth the extra computational effort? Perhaps a restricted optimization along the axes $\alpha_1 = 0$ and $\alpha_2 = 0$, corresponding to first and second differences, separately, would be more economical and provide a direct choice. More generally, does Dr Silverman know of other situations where it might be appropriate to penalise more than one derivative simultaneously?

My second comment concerns the relationship between generalized cross-validation and Tukey's rule; see Mosteller and Tukey (1977, p. 386) or Anscombe (1981, p. 358). The right hand side of (4.4) can be written as s^2/ν where $\nu = n - \mathrm{tr}\, A(\alpha)$ and $s^2 = \mathrm{RSS}(\alpha)/\nu$ is an approximate unbiased estimate of the error variance. Minimizing s^2/ν has an intuitive appeal when ν, which increases with α, is interpreted as a measure of complexity of the fitted curve: it offers a choice between the two conflicting aims of having simple structure (ν large) and small error. However, in line with Tukey's suggestion it is s^2/df which should be minimized, where df (the error degrees of freedom) is a measure of the statistical stability of s^2. In standard linear model theory, ν and df coincide but here, if we assume s^2 is approximately a chi-squared variate and the first four moments of the errors are proportional to those of the standard Gaussian distribution, we obtain $df = \nu^2/\mathrm{tr}\{I - A(\alpha)\}^2$. It should be noted that, as α tends to zero, ν tends to zero but df tends to a nonzero limit, as does s^2: this suggests that application of Tukey's rule would tend to lead to smaller values of α than generalized cross-validation, and this appears to be borne out in practice.

Mr R. Thompson (AFRC Animal Breeding Research Organisation): To paraphrase the title of Section 3, I would like to ask what is (generalized) cross-validation actually doing to the data?

For $E(y) = X\alpha$ and var $(y) = V = \sigma^2(I + ZZ^{\mathrm{T}}\gamma)$ then the weighted least squares estimate of $\alpha = (X^{\mathrm{T}} V^{-1} X)^{-1} X^{\mathrm{T}} V^{-1} y$ can be thought of as combination of estimates for (i) $(I - Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}})y$ and (ii) $Z^{\mathrm{T}}y$. The weight given to (i) and (ii) depends on the value of γ and so γ can be thought of as a "smoothing" parameter. If y is normally distributed and estimation of σ^2 and is based on residuals, $Sy = (I - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})y$, then estimates of σ^2 and γ satisfy

$$\Sigma[u_i^2 \lambda_1/(1 + \lambda_i\gamma)^2] = \sigma^2 \Sigma[\lambda_i/(1 + \lambda_i\gamma)] \quad \text{and} \tag{1}$$

$$\Sigma[u_i^2 (1 + \lambda_i\gamma)/(1 + \lambda_i\gamma)^2] = \sigma^2 \Sigma(1 + \lambda_i\gamma)/(1 + \lambda_i\gamma)], \tag{2}$$

where $u_i = P_i^{\mathrm{T}} Sy$ where P is orthogonal to S and SVS and $P_i^{\mathrm{T}} ZZ^{\mathrm{T}} P_i = \lambda_i$ (Patterson and Thompson, 1971). This can be interpreted as fitting a linear regression $\sigma^2(1 + \lambda_i\gamma)$ to squares of residuals. u_i^2. As u_i^2 is a chi-squared variables the weight given to u_i^2 is inversely proportional to $(1 + \lambda_i\gamma)^2$.

By contrast the GXV choice of parameters satisfies

$$\Sigma[u_i^2 \lambda_i/(1 + \lambda_i\gamma)^3] = \sigma^2 \Sigma[\lambda_i/(1 + \lambda_i\gamma)^2] \quad \text{and}$$

$$\Sigma[u_i^2 (1 + \lambda_i\gamma)/(1 + \lambda_i\gamma)^3 = \sigma^2 \Sigma[1 + \lambda_i\gamma)/(1 + \lambda_i\gamma)^2].$$

This is one of the same form as (1) and (2) except that the weight given to u_i^2 is inversely proportional to $(1 + \lambda_i\gamma)^3$. This is presumably an efficient scheme if u_i^2 has an inverse Gaussian distribution!

I would like to know if this formulation helps in understanding the choice of the spline smoothing parameter or if the computational burden is too great.

Professor Grace Wahba (University of Wisconsin–Madison): I would like to make a few clarifying remarks about the rather amazing properties of the confidence intervals obtained by estimating the smoothing parameter by generalized cross validation and then reverting to the Bayesian model to get confidence intervals. First, it is my belief that it is rather important to have a good value of the smoothing parameter (as obtained by GCV, for example) for the whole thing to work. Secondly, the frequentist property of these confidence intervals (very close to 95 per cent of the 95 per cent confidence intervals cover the true curves in Monte Carlo simulations) has to be interpreted as across the curve, and not necessarily for each point. To clarify this remark, consider the model of (1.1) with $n = 100$ and the variance of the e_is as constant. If one goes through the simulation exercise one will find that the confidence intervals cover say, 95 or 94 or maybe 96 of the points. Now take the same true curve and run the Monte Carlo again, that is, draw a new set of e_is. The same pleasing result is likely to obtain. However among the 4 or 5 or 6 points whose confidence intervals did not cover the truth, one may find that several of them were the same points whose confidence intervals failed to cover the truth in the first replication, and so forth. An extreme example of this may be found in Wahba (1983) Fig. 4 where the confidence interval at $t = 0.5$ fails to cover the point. In repeated replications with the same "true" curve the confidence interval at $t = 0.5$ repeatedly failed, although if all confidence intervals across the curve are taken together the total coverage is remarkably close to 95 per cent. (In that figure the first derivative had a jump at $t = 0.5$.) This phenomenon is most likely to appear where the bias is unusually large locally, and is likely to occur in the smoothing spline estimates where there is an unusually large local curvature, or worse, a kink in the curve. The cubic smoothing spline tends to smooth over extremely high curvature regions. A high curvature at the boundary where there is less neighbouring information causes more difficulty than if it occurred in the interior. In summary, these confidence intervals get the square bias right on the average, but not necessarily pointwise.

Nevertheless, if interpreted the right way, these confidence intervals appear to be amazingly reliable and generally tend to reflect what is going on, e.g. they are broader near the boundary and in regions of sparse data, etc. I would like to see the heuristic mathematics supporting them which appears at the end of Wahba (1983) put on a more rigorous footing. I would conjecture that there are at least rough mathematical justifications for Silverman's procedure for obtaining confidence intervals for some non-linear functionals, primarily those involving most or all of the whole curve, as opposed to those which are strictly local. As rather extreme examples, one would expect the results to be more trustworthy when estimating, say, the L_2 norm of a function than when estimating, say, the absolute first derivative at a specified point which happens to be near a "kinky" point. In any case, with high speed computing at most peoples fingertips, an experimenter can get a general feel concerning how reliable the entire procedure for a particular non-linear functional will be for his or her experiment by starting out with a reasonably representative (for their experiment) collection of possible "truths", simulating several data sets for each "truth" considered, going through the procedure of Section 7.3 for each data set, for the functional of interest, and comparing the results with the "truth". Possible surprises due to nonlinearity as noted by Silverman would become evident and, if the experimenter is concerned about unduly concentrated bias, its possible effects can be studied with suitable examples. Like Dr Silverman, I believe that it is perfectly respectable to use the computer as a tool whenever that will provide insight (provided, of course one knows enough about what the results ought to look like to get the programme debugged!).

Drs M. A. Tanner and W. H. Wong (University of Wisconsin–Madison; University of Chicago): We will comment on a few points raised by this important paper.

1. How do the results of Section 3 depend on the form of the penalty? For example, if higher derivatives are used in the penalty, e.g., $\int\{g^{(4)}(x)\}^2\,dx$, how would the dependency of the bandwidth on the density of the design points change? Does this consideration provide an approach for choosing the form of the penalty (or at least the degree of the derivative)?

2. We agree that the local dependency of the bandwidth on the density of the design points is a desirable property. However, this does not completely address the problem. There is another

aspect of local bandwidth that is important, even when the design points are uniformly spaced. The bias of the convolution approximation is small when the local curvature is small, in which case one can afford to use a wider bandwidth. Thus, the size of the bandwidth should depend on the local curvature of the (unknown) function. In this way, adaptation using a preliminary estimate of curvature may be a good idea. How can this be done in the spline context?

3. In Section 5, the local sum of squares of generalized residuals is used to estimate the local variation. However, this raw sum of squares decomposes into the local squared bias plus the variance. It seems preferable to estimate the local bias by the local mean and the local variance by the variance of these quantities. In this way, oversmoothing will not be confounded with non-homogeneity of variance.

4. The Bayesian development in Section 6 is very nice. But the intervals which are developed in this section do not seem to be real confidence intervals. To draw an analogy with kernel regression, the standard error is being used as if the bandwidth is chosen deterministically. Thus, in repeated experiments, the correct coverage probability will not be achieved. Of course, one can argue that only the distribution conditional on the selected bandwidth is relevant. This issue may seem similar to that in the Hinkley–Runger (Box–Cox?) versus Bickel–Doksum controversy. However, it seems that the argument for conditioning on the bandwidth is much weaker than conditioning on the scale in the Box–Cox transformation.

5. Finally, in order to obtain a clearer perspective of the usefulness of spline smoothing methodology, it may be desirable to include a comparison with competitive methodologies such as local linear fit (Cleveland (1979)) or nonlinear data smoothers (Velleman (1980)). A direct comparison in specific examples may be quite illuminating.

Acknowledgements

Dr Wong was supported by Research Grant MCS-8301459 of the National Science Foundation. Dr Tanner was supported by Research Grant MCS-8300977 of the National Science Foundation.

Professor Sidney Yakowitz (University of Arizona): I am grateful for Dr Silverman's masterful guidance through the hinterlands of spline regression. The connections given in Section 3 with kernel methods, with which I am more familiar, were particularly illuminating, and as a computational specialist, I cannot help but admire the developments in Section 4 which reduce the number of operations in evaluating a cross-validation score from $O(n^2)$ to $O(n)$.

The matter closest to my heart and realm of experience is the theory of diagnostics and error variance. I applaud Dr Silverman's accomplishments in this direction. For some while, "geostatisticians" and hydrologists have been painfully aware of the need for such a theory, and in response to this need, an activity commonly known as "kriging" (e.g., Journel and Huijbregts (1978), or in a more scholarly vein, Chapter 4 of Ripley (1981)) has flourished. In essence, kriging is a second-order linear filter predictor, where the covariance function must usually be inferred from the same data on which the prediction is based. Since it is a second-order theory, in principle squared-error estimates are immediately available. The drawback, as I see it, is that since the design points come from a bounded region, there is no reason to think that the covariance estimate is consistent.

Recently, Yakowitz and Szidarovszky (1985) have devised a data-based estimator $\hat{\sigma}_n^2(t)$ for the variance $\sigma_n^2(t) = \operatorname{var}\{g(t) - g_n(t)\}$ of the kernel nonparametric regression estimate $g_n(t)$. We established that, as $n \to \infty$,

$$\hat{\sigma}_n^2(t)/\sigma_n^2(t) \to 1, \text{ in probability.}$$

By making the kernel bandwidth decrease at the "right" rate, it is readily seen that the bias component of the estimate $g_n(t)$ becomes negligible and that the convergence rate, under second-order differentiability assumptions, becomes arbitrarily close to the Stone (1980) optimal rate.

The author replied later, in writing, as follows:

I am only sorry that limitations of time and space prevent me from answering every single point made in the discussion and from referring to every contributor by name. Some of the most interesting and valuable contributions stand alone without the need for further comment from me. I confess that in places I shall adopt the view of Good and Gaskins (1980) that "as is customary . . . most of our reply will be concerned with elaborations and mild dis-

agreements . . . because this is more useful than agreeing".

I am most grateful to Professor Whittle for supplying further historical information. In the circumstances, perhaps I should have delayed reading this paper until 1986, the centenary of Mr King's remarks! It was with trepidation that I presented to the Society a paper with the word Bayesian in the Summary—if not in the title—so Professor Whittle's kind remarks on the Bayesian aspects were most welcome. I believe that the degeneracy that Professor Whittle's likelihood displays as $N \to \infty$ is another unpleasant consequence of the peculiar behaviour of the full infinite-dimensional model discussed briefly in Section 6.1 of the paper.

However, an approach via maximum likelihood is possible, though some fudging is necessary to get over the partially improper nature of the prior distribution. Use the suggestion of Dr Green and separate out the constant and linear components of the fit as coefficients β_0 and β_1. The distribution of the curve and the data then depends on the three parameters λ, β_0 and β_1, and the predictive (or marginal) distribution of the data can be found; see Wahba (1983b) for a closely related discussion. The maximum of this predictive likelihood will correspond to the maximum of Professor Whittle's likelihood with N replaced by $n-2$ and the roughness penalty restored to its original form. Unfortunately, some heuristic calculations show that the minimizer of this likelihood will give a value of λ that leads to substantial undersmoothing, but further work on this topic is clearly necessary. Wahba (1983b) investigates a different criterion, motivated by similar considerations; she also shows, by theoretical arguments and by simulation, that her criterion will lead to undersmoothing, and that cross-validation is preferable.

Professor Whittle and several other discussants consider the multivariate case. Professor Whittle's elegant Fourier transform argument gives a special case of Theorem 1 of Meinguet (1979), which implies that, for the roughness penalty approach to work in a space of dimension m, it is necessary for the degree of the derivatives in the roughness penalty to be strictly greater than $\frac{1}{2}m$. I am not sure whether I would agree that increasing the degree of the differentials is relaxing the statistical assumptions on g; thinking in terms of the Fourier series for g, a high degree roughness penalty is placing very strong decay conditions on the sequence of high frequency coefficients, though it is also allowing a larger class of functions to have zero roughness.

Of course the multivariate case would have taken several papers all on its own. I believe that the numerical difficulties of the roughness penalty approach alluded to by Dr Scott are caused not by the dimensionality being greater than 3 but by the number of data points being too large. However the dominant role of boundary effects in higher dimensions rings a warning bell for kernel regression and for convoluted smoothing as well as for roughness penalty methods. I agree that dimension reducing ideas like projection pursuit, or something along the lines of Dr Marriott's suggestion (ii), are more likely to be useful.

Several contributors mention kriging. The relations between kriging and other smoothing methods are not very widely understood and I am grateful to Professor Cressie for discussing them. As far as I can see, the basic assumption of kriging, that the unknown function is a realization of a certain stochastic process, is nothing other than a Bayesian prior model, interpreted, as Professor Whittle puts it so succinctly, in a frequentist non-personal sense. The method for updating the prior to take account of observed data does not appear, on the face of it, to be in accordance with Bayes' Theorem. A particular problem with kriging is that geological surfaces are known not to look anything like realizations of isotropic Gaussian processes—because of phenomena like faults—and therefore it is to be hoped that work will be done on more careful modelling of surfaces. I am sure that both the theoretical understanding and the practical methodology of kriging would be enhanced by a Bayesian view of the procedure. However I would be (pleasantly) surprised if the finite-dimensional interpretation requested by Professor Cressie would be available for kriging as it stands.

Professor Titterington discusses connections with ridge regression. It is important to consider the conceptual differences between spline smoothing and ridge regression; I prefer to view the comparison in the Bayesian setting, but a non-Bayesian interpretation is equally possible. The context in which ridge regression was introduced (see Hoerl and Kennard, 1970) was to cope with problems, in the usual multiple regression setting, where the matrix $X^{\mathrm{T}}X$ is not very well conditioned; the solution is then to put a spherical normal prior on the parameter vector. In spline smoothing, on the other hand, the vector of parameters $g(t_i)$ is given a highly non-spherical prior distribution but the matrix $X^{\mathrm{T}}X$ is just the identity, if the $g(t_i)$ are considered to be the parameters. Spline smoothing can be forced into the ridge regression framework but only at the

expense of introducing a parametrization of curves g in which squared distance between two curves becomes, effectively, $\int(g_1'' - g_2'')^2$. This is a rather unnatural thing to do, especially if assessments of estimation accuracy are then made in terms of prediction errors.

Turning to Professor Titterington's specific remarks, I am glad that he agrees with me that subjective choice of smoothing will often suffice; I hope that Professor Barnard would regard a subjectively chosen α as satisfactory so long as its value were clearly reported as an "assumption" of the technique. Enormous image-enhancement problems are precisely an example where the techniques of my Section 4 should come into their own; ordinary cross-validation will be impossible, but approximations like AGXV will be extremely feasible, since they cost essentially nothing to compute.

Professor Titterington's equation (*) could be tautological or impossible to satisfy, depending on how σ^2 is to be estimated. Indeed cross-validation can severely undersmooth occasionally; the simulations, and the theoretical discussion, of Silverman (1984b) show that, at least for non-uniform design points, AGXV is much less prone to this difficulty. I agree that the curves of Figs 3 and 6 are fairly similar; however there is no guarantee that this should be so, and furthermore the posterior probability regions are very different from the unweighted data. The mechanism for providing "simultaneous inference" regions is provided in Section 7 of the paper; what is required is the posterior distribution of, say, $\sup |g - \hat{g}|$. The upper 95 per cent point of this distribution can be found by simulation from the posterior and this will give the half-width of the required region. Similarly a lower probability bound for the number of inflexions can be obtained.

I agree that the probability intervals are pointwise intervals; however in my finite-dimensional formulation they are certainly meaningful at all points, not just at the knots. I have discussed, very heuristically, some connections between Section 2 and density estimation in Section 7 of Silverman (1984a). I do not know whether there are consequences for the choice of smoothing parameter, but the argument given there does indicate that a roughness penalty approach may adapt well when estimating long-tailed densities. Finally I strongly support Professor Titterington's remarks about the need for future practical investigations. Of course, what we need to avoid is theoretical work of the wrong kind, not *all* theoretical work!

A very fair assessment of Dr Scott's suggested parametric model, least squares splines, is given by Drs Parker and Rice. As they point out, possibly the best way of using this approach is to develop a hybrid method, minimizing the penalized residual sum of squares over the space of splines on a fixed grid of knots. Incidentally, I doubt that the method is very much simpler computationally than non-parametric spline smoothing – it all depends which package you happen to have on your computer.

Dr Kent's spherical spline proposal sounds most interesting as an example of an extension of the technique. I am not aware of any specific "errors-in-variables" formulation of the spline smoothing approach, and this would be a very useful topic for further work.

Dr Robinson's question (see p. 50), also alluded to by Dr Marron, on processes whose degree of smoothness may not be stationary, is related to my misgivings about the stationarity of models for kriging. In the one-dimensional case, this problem can essentially be overcome, as far as the estimation of g itself is concerned, by using non-uniform patterns of weights and a constant smoothing parameter; however the penalized likelihood interpretation of the method will no longer hold. It is much more difficult to allow the smoothing parameter itself to vary, and change-point inference for α is something that would require a great deal of ingenuity to devise.

The finite-dimensional Bayesian model has aspects which worry several of the contributors. Mr Upsdell and Dr Ansley might be amused by Professor Lindley's personal comment to me that "it is a joy not to have to go into Hilbert space". I take the point that the model seems somewhat unsatisfactory because of its dependence on the position of the t_i. Although I have no hard evidence to offer, my experience is that moving t_i a little does not alter the inferences very much and certainly does not have the drastic consequences feared by Professor Lindley, no doubt, as he suggests, because of an approximation property of some kind. In particular, at the data points themselves, the posterior variance/covariance structure agrees with that obtained by Wahba (1978) for an essentially stationary infinite-dimensional prior. Although I find it much harder to contemplate personal beliefs about curves and surfaces in terms of covariances than Professor Lindley and Dr O'Hagan appear to, I hope that this helps to reconcile them to the finite dimensional model.

Is Professor Smith's consideration of the limiting case with zero errors really very interesting? It certainly is pushing the Bayesian aspects of the method beyond the extremes I would envisage

for it. What worries me about using the Wiener process formulation in a Bayesian context is that I know that the sample path behaviour of Brownian motion is very peculiar indeed and I am therefore hesitant about saying that my prior "knowledge" is represented by a Wiener process model (which includes a coded version of the works of Shakespeare in every realization). Incidentally, it is instructive to sketch a graph of your perceived idea of a Wiener process and to compare the result with Fig. 9.1 of Grimmett and Stirzaker (1982).

There are, of course, connections between my Bayesian formulation and Dr O'Hagan's much more general approach, and I apologise to him for not referring to them explicitly in the paper. Dr Ansley also points out connections with other very interesting related work with Bayesian connotations. To put the record straight, it must be pointed out that the formulation of Wahba (1983) is not dependent on the observed design points in the way Dr Ansley suggests. I entirely agree that more information about the effect of restricting the prior is needed. A possible personal-Bayesian approach—entirely feasible using the simulation methodology of Section 7—is, having factored out the constant and linear components of the curve, to simulate from the prior distribution of curves for various values of α and hence to choose a value of α producing curves that accord with prior ideas. This kind of idea is advocated by Stewart (1979). It is difficult to illustrate the results in published form, but on an interactive graphics terminal it is easy to get an idea of what the prior looks like. In addition the prior distribution of functionals of the curve can be obtained.

I accept completely that my approach is inappropriate for extrapolation. Perhaps, in providing clearly preposterous extrapolations beyond short ranges it is doing a service in making it clear that such extrapolation is (nearly) always preposterous.

In answer to Mr Upsdell's last point, the difference between the smoothing problem and the rational-real problem is that, in smoothing, the quantity $\int g''^2$ is of the essence, in that it is the quantity that measures how smooth we think the curve is, and hence on which our prior beliefs depend. If I were speculating whether a number were rational I would not use the normal distribution to quantify my beliefs.

Several contributors mention the problem of possible correlation of errors, and the connections with time series analysis. I am not sure whether it is appropriate to do as Dr Diggle has and merely modify equation (4.4), since, as Professor Robinson points out, an autoregressive error assumption will have consequences right through the development of the method. Nevertheless I certainly agree with his general conclusions. Mr Bates's suggestion of using the method of Cochrane and Orcutt (1949) is an obvious thing to try once we have decided how (in his notation) to select the smoothing parameter α for a given autoregressive coefficient ρ. Though others will disagree with me, I do not think there is very much real philosophical distinction between α and ρ.

May I reassure Dr Jennison a little? The spline method, as shown in Section 2, will also smooth over very local behaviour. It would be quite straightforward to use his very sensible definition of maximum gradient in the simulations to get a corresponding estimate and posterior standard deviation. Does he not feel that the broad intervals in Fig. 9 *do* illustrate the lack of reliability of gradient estimates?

I suspect that Professor Dawid's first suggestion might possibly make more difference than his second; my practical experience, and the conclusion (ii) of Section 3, indicate that the factor $n/(n-1)$ will hardly affect the curve at all. The only immediate difficulty concerning his first suggestion is that the determination of the curve for a given β (in his notation) requires a search over curves found for various values of α. This will increase the computational burden considerably, particularly for the cross-validation score.

Cleveland's (1979) approach, as mentioned by several contributors, is a robust refinement of local weighted moving averages. It is worth pointing out that it is essentially only in the "robustification" step that it matters whether local linear or local constant fit is used; a fair comparison would therefore be with the robust spline approach based on (8.1). A careful comparison, including the effect if the data are non-uniformly spaced, would be well worth making. In its local weighted nature Cleveland's approach has connections with kriging as well. Another candidate for careful comparison would be the estimator suggested by Drs Müller and Stadtmüller.

Professor Atkinson's proposal that deleted (or externally studentized) residuals should be used is interesting and deserves further investigation. The details are likely to be somewhat different

from the ordinary regression case because the hat matrix is no longer a projection matrix. As in ordinary regression, the effect of external Studentization is likely to be to increase residuals that are already large.

I doubt whether Mr Besag issues his students with pencils and rulers instead of teaching the reproducible and (in Professor Barnard's sense) objective method of linear regression. My "cause" was not to claim that spline smoothing was the best method in all cases but that it is a good method in many cases. Professor Atkinson is right about the motor-cycle data in that they are the superposition of measurements made by several instruments but the data I have presented are precisely in the form they were made available to me. Mr Besag would fit a logit curve in Fig. 8, but such a parametric logit model leaves no room for departures from the model and when used blindly (as very frequently in practice) will often suppress important features of the data, such as multiple growth spurts or negative growth in places. Mr Besag's general point about examples versus applications is well taken; I was well aware that alternative methods of analysis are always suggested when examples are presented. I feel that the flexibility and broad applicability of the spline smoothing approach is already demonstrated by the examples in the paper and I would encourage readers to try out the method for themselves. Only in this way will a realistic assessment of the scientific value of nonparametric smoothing emerge.

A natural approach to Mr Bates's first question, alluded to by Professor Wahba, would be to fit a multiple linear regression first, and then to model a smooth dependence of the residuals on, say, one of the explanatory variables. Another avenue of possible investigation, mentioned by Dr O'Sullivan, is to fit a model that is a sum of functions each of whcih is a smooth function of one of the coordinates only. Thus instead of fitting $z = g(x, y) +$ error we would fit $z = g_1(x) + g_2(y) +$ error, and penalize for the roughness of g_1 and g_2 separately. Developments of this kind are obviously likely to be valuable for Mr Bates's second problem as well.

In answer to Professor Collomb, and to Professors Greblicki and Pawlak, the asymptotic rate of convergence of the spline method is the same, as they will have guessed from Section 3, as the kernel method using the kernel κ. Loosely speaking, the method gives the optimal rate of mean square error convergence, $n^{-8/9}$, for curves g with four derivatives, but does not give any improvement for smoother curves. For a careful discussion and further references, see Rice and Rosenblatt (1981, 1983). It should be said that the practical impact of improvements from $n^{-8/9}$ to $n^{-1+\epsilon}$ is not likely to be very great. The rate can be improved further, subject to more stringent conditions, by using higher derivatives in the roughness penalty, as suggested by Drs Tanner and Wong and also by Dr Cuzick. Details of results for such penalties corresponding to Section 3 are given in Silverman (1984a). The power of the dependence of $h(t)$ on $f(t)$ is nearer to zero, but the sensitivity of the asymptotic arguments is probably too great for the results to be taken too literally in practice. I must admit to a similar suspicion about the asymptotics discussed by Drs Müller and Stadtmüller, and also about some of the results referred to by Professor Collomb.

Dr Cuzick makes a good point in demonstrating that sensible results will not be obtained for functionals that depend on high derivatives of g. As he says, higher order roughness penalties will then be appropriate. I should, of course, have qualified my remarks about the generality of the approach of Section 7.2 to exclude such functionals. A natural approach, which deserves further investigation, is to use simulation from the prior to get the prior distribution of a particular functional and to use this as a method of choosing the smoothing parameter. Careful work also needs to be done on relevant methods of automatic choice for estimation of functionals, especially when derivatives are involved.

Carefully chosen roughness penalties may well be the way forward in the seasonal adjustment problem mentioned by Professor Durbin. The key property is the "null family" (see Silverman, 1982b) of curves that have zero roughness. For the penalty $\int g''^2$ discussed in the paper this is precisely the family of linear trends. A possible roughness penalty for studying slowly changing seasonal effects would be $\int (g''' + \omega^2 g')^2$, which would regard pure seasonal curves of period $2\pi/\omega$ as perfectly smooth. This is an obvious topic for important future work, and answers a question raised by Dr Seheult.

Dr Eubank makes several excellent points about diagnostics and other matters. The generalized cross-validation approach of course neglects differences between the leverage values $A_{ii}(\alpha)$; this may be an advantage, because it reduces the influence of high leverage data points in the choice of smoothing parameter. The same appears to be true of the other methods mentioned in the extremely illuminating discussion of Dr Härdle. I hope that Dr Marron's remarks about cross-

validation will reassure sceptics. In answer to his last question, and to the third point of Drs Wong and Tanner, refinements to the local weight estimation procedure are obviously worth considering though I would frankly be surprised if they made more than a fairly slight practical difference.

Dr O'Sullivan makes several fascinating points. In particular I am very interested in his suggestion that AGXV goes further than GXV in reducing the influence of high leverage points. Taking into account the burden of finding the residual sum of squares for the given α, his computational remarks appear to reduce the computer time required for exact GXV to about twice that required for my procedure. It must be admitted that this will not be a vast overhead for moderate problems, but the compact nature of the AGXV calculations may be an advantage on microcomputers.

Drs Eagleson and Raftery both ask about simultaneous smoothing of two or more data sets. Obviously it is sensible to use the same smoothing parameter for both curves, and the natural approach is to minimize the sum of the two cross-validation scores. Dr Eagleson's question could perhaps then be approached by examining the posterior distribution of the difference between the two curves. I think that Dr Raftery's problem requires a bivariate penalized likelihood smoothing procedure for its satisfactory solution and that the present paper does very little more than provide a philosophical framework for this.

In conclusion, my warmest thanks are due to *all* the discussants for providing such a broad perspective on the material of the paper, for pointing out valuable additional references, for frank and open criticism (in the best traditions of the Society) and for giving me, and I hope others, a great deal of food for further thought and suggestions for further research.

Dr A. Robinson (University of Bath): My comments concern the dynamic calculation of the spline smoother, discussed by Wecker and Ansley (1983) who apply Kalman filtering to Wahba's (1978) model. Mr R. A. Moyeed and myself have used the finite-dimensional formulation of Section 6.1 to produce a more accessible approach to dynamic calculation in certain cases, e.g. when precise signal extraction is not the goal. For simplicity, assume no weighting and the smoothing parameter α is fixed. If n denotes values after n observations γ_n has posterior mean $\sigma^{-2} S_n^{-1} B_n^{\mathrm{T}} Y_n$ and variance $S_n^{-1} = (\alpha\Omega_n + \sigma^2 B_n^{\mathrm{T}} B_n)^{-1}$. The formulation of 6.1 may be associated with the dynamic linear model with observation equation, $y_\nu = b_\nu \gamma_{n+1} + \epsilon_\nu$, where y_ν is the $(n+1)$st observation at point t_ν $(t_l < t_\nu < t_{l+1})$; $\epsilon_\nu \sim N(0, \sigma^2)$ is the observation error; b_ν is the νth row of B_{n+1} and γ_{n+1} are the coefficients of these B-splines based on knots

$$\{t_1 < \ldots < t_l < t_\nu < t_{l+1} < \ldots < t_n\}.$$

The system equation is $\gamma_{n+1} = G_{n+1}\, \gamma_n + w_{n+1}$ where $G_{n+1} = B_{n+1}^{-1} B_n$ and $B_n^{(\nu)}$ is B_n with $(0, \ldots, 0, B_{n,\,l-1}(t_\nu), B_{n,\,l}(t_\nu), B_{n,\,l+1}(t_\nu), B_{n,\,l+2}(t_\nu), 0, \ldots, 0)$ inserted between rows l and $l+1$. The system error is $w_{n+1} \sim N(0, W_{n+1})$ where

$$W_{n+1} = (S_{n+1} - \sigma^{-2} b_\nu^{\mathrm{T}} b_\nu)^{-1} - G_{n+1}\, S_n^{-1}\, G_{n+1}^{\mathrm{T}}.$$

Computationally the method is straightforward, i.e. the mean is updated as follows:

$$\gamma_{n+1}^{\mathrm{post}} = \gamma_{n+1}^{\mathrm{prior}} + c_\nu(y_\nu - y_\nu^{\mathrm{prior}}) = G_{n+1}\, \gamma_n^{\mathrm{post}} + c_\nu(y_\nu - y_\nu^{\mathrm{prior}})$$

where c_ν is the νth column of $S_{n+1}^{-1} B_{n+1}^{\mathrm{T}}$ and y_ν^{prior} is the prior mean of y_ν obtained from $\{y_1, \ldots, y_n\}$. The last equation may also provide a frame for dynamic diagnostics and choice of subsequent design points.

Finally I would like to ask Professor Silverman if he considers the global choice of α by cross-validation restricts the adaptability of this method of fitting to a process whose "degree of smoothness" may not be stationary? Where desirable, would some sort of change point analysis for α be feasible?

REFERENCES IN THE DISCUSSION

Abramson, I. S. (1982) On bandwidth variation in kernel estimates—a square root law. *Ann. Statist.*, **10**, 1217–1223.

Agarwal, G. and Studden, W. (1980) Asymptotic integrated mean square error using least squares and bias minimizing splines. *Ann. Statist.*, 8, 1307–1325.

Akaike, H. (1970) Statistical predictor identification. *Ann. Inst. Statist. Maths*, **22**, 203–217.

Anscombe, F. (1981) *Computing in Statistical Science through APL*. New York: Springer-Verlag.

Ansley, C. F. and Kohn, R. (1984) On the equivalence of two stochastic approaches to spline smoothing. *J. Appl. Prob.*, to appear.

Ansley, C. F. and Wecker, W. E. (1983) Extensions and examples of the signal extraction approach to regression. In *Applied Time Series Analysis of Economic Data* (A. Zellner, ed.), pp. 181–198. Washington, D.C.: U.S. Bureau of the Census.

Atkinson, A. C. (1981) Two graphical displays for outlying and influential observations in regression. *Biometrika*, **68**, 13–20.

Belsey, D. A., Kuh, E. and Welsch, R. E. (1980) *Regression Diagnostics*. New York: Wiley.

Besag, J. E. and Kempton, R. (1984) Spatial methods in the analysis of agricultural field trials. *Proc. 12th International Biometrics Conference*, pp. 80–88.

Buse, A. and Lim, L. (1977) Cubic splines as a special case of restricted least squares. *J. Amer. Statist. Ass.*, **72**, 64–68.

Cleveland, W. S. (1979) Robust locally weighted regression and smoothing scatter plots. *J. Amer. Statist. Ass.*, **74**, 829–836.

Cochrane, D. and Orcutt, G. H. (1949) Application of least squares regression to relationships containing auto-correlated error terms. *J. Amer. Statist. Ass.*, **44**, 32–61.

Collomb, G. (1977a) Quelques propriétés de la méthode du noyau pour l'estimation nonparamétrique de la régression en un point fixé. *C. R. Acad. Sci. Paris*, A, **285**, 289–292.

——— (1977b) Estimation nonparamétrique de la régression par la méthode du noyau: propriétés de convergence asymptotiquement normale indépendante. *Ann. Sci. de l'Univ. de Clermont*, **15**, 24–46.

——— (1982) From data anlaysis to nonparametric statistics: recent developments and a computer realization for exploratory techniques in regression or prediction. In *Compstat 82*, (H. Caussinus *et al.*, eds), pp. 173–178. Vienna: Physica-Verlag.

——— (1983) Nonparametric time series analysis and prediction: uniform a.s. convergence of the window and K-NN autoregressive estimates. *Math. Operationsfors. und Statist., Series Statistics*, to appear.

——— (1984) Propriétés de convergence presque complète du prédicteur à noyau. *Zet. für Wahrsch. verw. Gebiete*, **66**, 441–460.

Collomb, G. and Härdle, W. (1984) Strong uniform convergence rates in robust nonparametric time series analysis and prediction: kernel regression estimation from dependent observations. Submitted for publication.

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

Cressie, N. (1985) Kriging non-stationary data. *J. Amer. Statist. Ass.*, submitted for publication.

Critchley, F. (1978) Multidimensional scaling: a short critique and a new method. In *Compstat 78* (L. C. A. Corsten and J. Hermans, eds), pp. 297–303. Vienna: Physica-Verlag.

De Leeuw, J. (1982) Nonlinear principal component analysis. In *Compstat 82* (H. Caussinus *et al.*, eds). Vienna: Physica-Verlag.

Draper, N. R. and van Nostrand, R. C. (1979) Ridge regression and James-Stein estimation: review and comments. *Technometrics*, **21**, 451–466.

Dubrule, O. (1983) Two methods with different objectives: splines and kriging. *J. Int. Ass. for Math. Geol.*, **15**, 245–257.

Durbin, J. (1984) Time series analysis. *J. R. Statist. Soc.* A, **147**, 161–173.

Efron, B. (1982) *The Jackknife, the Bootstrap and other Resampling Plans*. CBMS-NSF. SIAM Monograph No. 38.

Elden, L. (1977) Algorithms for regularization of ill-conditioned least square problems. BIT, **17**, 134–145.

Eubank, R. L. (1984a) Approximate regression models and splines. *Commun. in Statist.* A, **13**, 433–484.

——— (1984b) The hat matrix for smoothing splines. *Statist. and Prob. Letters*, 2, 9–14.

Friedman, J. H. and Stuetzle, W. (1981) Projection pursuit regression. *J. Amer. Statist. Ass.*, **76**, 817–823.

Gasser, T. and Muller, H. G. (1979) Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation* (T. Gasser and M. Rosenblatt, eds). pp. 23–68. Heidelberg: Springer-Verlag.

Gifi, A (1981) *Nonlinear Multivariate Analysis*. Leiden: University Press.

Good, I. J. (1983) Probability estimation by maximum penalized likelihood for large contingency tables and for other categorical data. *J. Statist. Comput. Simul.*, **17**, 66–67.

Good, I. J. and Gaskins, R. A. (1980) Density estimation and bump-hunting by the penalized maximum likelihood method exemplified by scattering and meteorite data. *J. Amer. Statist. Ass.*, **75**, 42–73.

Greblicki, W. and Pawlak, M. (1984) Hermite series estimates of a probability density and its derivatives. *J. Multivar. Anal.*, **15**, in press.

Green, P. J. (1985) Linear models for field trials, smoothing and cross-validation. *Biometrika*, to appear.

Green, P. J., Jennison, C. and Seheult, A. H. (1983) In discussion of Wilkinson *et al. J. R. Statist Soc*, B, **45**, 193–195.

Grimmett, G. R. and Stirzaker, D. R. (1982) *Probability and Random Processes*. Oxford: Clarendon.

Hinkley, D. V. (1977) Jackknifing in unbalanced situations. *Technometrics*, **19**, 285–292.

Hoerl, A. E. and Kennard, R. W. (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–82.

Journel, A. G. and Huijbregts, C. J. (1978) *Mining Geostatistics*. New York: Academic Press.
Kohn, R. and Ansley, C. F. (1983) On the smoothness properties of the best linear unbiased estimate of a stochastic process observed with noise. *Ann. Statist.*, **11**, 1011–1017.
Liero, H. (1982) On the maximal deviation of the kernel regression function estimate. *Math. Operationsfors. und Statist., Series Statistics*, **13** (2), 171–182.
Matheron, G. (1963) Principles of geostatistics. *Econ. Geol.*, **58**, 1246–1266.
Mosteller, F. and Tukey, J. W. (1977) *Data Analysis and Regression*. Reading, Mass.: Addison-Wesley.
Müller, H. G. and Stadtmüller, U. (1984) Variable bandwidth kernel estimators of regression curves. Unpublished manuscript, University of Ulm, Federal Republic of Germany.
O'Hagan, A. (1978) Curve fitting and optimal design for prediction (with Discussion). *J. R. Statist. Soc.* B, **40**, 1–42.
O'Sullivan, F., Yandell, B. and Raynor, W. (1984) Automatic smoothing of regression functions in generalized linear models. Technical Report No. 734, Statistics Dept, University of Wisconsin at Madison.
Patterson, H. D. and Thompson, R. (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.
Prakasa Rao, B. L. S. (1983) *Nonparametric Functional Estimation*. New York: Academic Press.
Phillips, D. L. (1962) A technique for the numerical solution of certain integral equations of the first kind. *J. Ass. Comp. Mach.*, **9**, 84–97.
Raftery, A. E. (1983) A non-parametric approach to measuring social mobility. *Int. Statist. Inst., 44th Session*, pp. 379–383.
——— (1985) Social mobility measures for cross-national comparisons. *Quality and Quantity*, to appear.
Ramsay, J. O. (1982) Some statistical approaches to multidimensional scaling (with Discussion). *J. R. Statist. Soc.* A, **145**, 285–312.
Rice, J. and Rosenblatt, M. (1981) Integrated mean square error of a smoothing spline. *J. Approx. Theo.*, **33**, 353–369.
Ripley, B. (1981) *Spatial Statistics*. New York: Wiley.
Rosenblatt, M. (1971) Curve estimates. *Ann. Math. Statist.*, **42**, 1815–1842.
Scott, E. M. (1984) A review of existing methods of non-parametric regression estimation. Submitted for publication.
Shibata, R. (1981) An optimal selection of regression variables. *Biometrika*, **68**, 45–54.
Smith, P. (1979) Splines as a useful and convenient statistical tool. *Amer. Statist'n*, **33**, 57–62.
Stone, C. J. (1980) Optimal rates of convergence for non-parametric estimators. *Ann. Statist.*, **8**, 1348–1360.
——— (1984) Additive regression and other nonparametric models. Technical Report No. 33, Statistics Dept, University of California at Berkeley.
Stone, M. (1983) In discussion of Copas. *J. R. Statist. Soc.* B, **45**, 336–338.
Van Rijckevorsel, J. (1982) Canonical analysis with *B*-splines. In *Compstat 82* (H. Caussinus *et al.*, eds). Vienna: Physica-Verlag.
Velleman, P. (1980) Definitio and comparison of robust non-linear data smoothing algorithms. *J. Amer. Statist. Ass.*, **75**, 609–615.
Wahba, G. (1975) Optimal convergence properties of variable knots, kernel and orthogonal series methods for density estimation. *Ann. Statist.*, **3**, 15–29.
——— (1981) Bayesian confidence intervals for the cross validated smoothing spline. Technical Report 645, Dept of Statistics, University of Wisconsin at Madison.
——— (1983b) A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. Technical Report 712, Dept of Statistics, University of Wisconsin in Madison.
——— (1984) Cross-validated spline methods for the estimation of multivariate functions from data on functionals. In *Statistics: an Appraisal* (H. A. David and H. T. David, eds), pp. 205–235. Iowa: State University press.
Walden, A. T. and Prescott, P. (1983) Identification of trends in annual maximum sea levels using robust locally weighted regression. *Esturine, Coastal and Shelf Sci.*, **16**, 17–26.
Watson, G. S. (1964) Smooth regression analysis. *Sankhya* A, **26**, 359–372.
——— (1972) Trend surface analysis and spatial correlation. *Geol. Soc. America, Special Paper*, **146**, 39–46.
——— (1984) Smoothing and interpolation by kriging and with splines. *J. Int. Ass. Math. Geol.*, **16**, 601–615.
Weinert, H. L., Byrd, R. H. and Sidhu, G. S. (1980) A stochastic framework for recursive computation of spline functions: Part II, Smoothing splines. *J. Optimiz. Theo. and Applic'ns*, **30**, 255–268.
Whittaker, E. T. and Robinson, C. (1924) *The Calculus of Observations*. Glasgow: Blackie & Son.
Wilkinson, G. N. (1984) Nearest neighbour methodology for design and analysis of field experiments. *Proc. XII Int. Biometrics Conference*, pp. 69–79.
Wilkinson, G. N., Eckert, S. R., Hancock, T. W. and Mayo, O. (1983) Nearest neighbour (NN) analysis of field experiments (with Discussion). *J. R. Statist. Soc.* B, **45**, 151–211.
Wold, S. (1974) Spline functions in data analysis. *Technometrics*, **16**, 1–11.
Yakowitz, S. and Szidarovszky, F. (1985) A comparison of kriging with nonparametric regression methods. *J. Multivar. Anal.*, **16**, 21–53.
Young, A. S. (1977) A Bayesian approach to prediction using polynomials. *Biometrika*, **64**, 309–317.