Statistical Science
1996, Vol. 11, No. 2, 89–121

Flexible Smoothing with *B*-splines and Penalties

Paul H. C. Eilers and Brian D. Marx

Abstract. *B*-splines are attractive for nonparametric modelling, but choosing the optimal number and positions of knots is a complex task. Equidistant knots can be used, but their small and discrete number allows only limited control over smoothness and fit. We propose to use a relatively large number of knots and a difference penalty on coefficients of adjacent *B*-splines. We show connections to the familiar spline penalty on the integral of the squared second derivative. A short overview of *B*-splines, of their construction and of penalized likelihood is presented. We discuss properties of penalized *B*-splines and propose various criteria for the choice of an optimal penalty parameter. Nonparametric logistic regression, density estimation and scatterplot smoothing are used as examples. Some details of the computations are presented.

Key words and phrases: Generalized linear models, smoothing, nonparametric models, splines, density estimation.

1. INTRODUCTION

There can be little doubt that smoothing has a respectable place in statistics today. Many papers and a number of books have appeared (Silverman, 1986; Eubank, 1988; Hastie and Tibshirani, 1990; Härdle, 1990; Wahba, 1990; Wand and Jones, 1993; Green and Silverman, 1994). There are several reasons for this popularity: many data sets are too "rich" to be fully modeled with parametric models; graphical presentation has become increasingly more important and easier to use; and exploratory analysis of data has become more common.

Actually, the name nonparametric is not always well chosen. It might apply to kernel smoothers and running statistics, but spline smoothers are described by parameters, although their number can be large. It might be better to talk about "overparametric" techniques or "anonymous" models; the parameters have no scientific interpretation.

There exist several refinements of running statistics, like kernel smoothers (Silverman, 1986; Härdle, 1990) and LOWESS (Cleveland, 1979). Splines come in several varieties: smoothing splines, regression splines (Eubank, 1988) and *B*-splines (de Boor, 1978; Dierckx, 1993). With so many techniques available, why should we propose a new one? We believe that a combination of *B*-splines and difference penalties (on the estimated coefficients), which we call *P*-splines, has very attractive properties. *P*-splines have no boundary effects, they are a straightforward extension of (generalized) linear regression models, conserve moments (means, variances) of the data and have polynomial curve fits as limits. The computations, including those for cross-validation, are relatively inexpensive and easily incorporated into standard software.

B-splines are constructed from polynomial pieces, joined at certain values of x, the knots. Once the knots are given, it is easy to compute the *B*-splines recursively, for any desired degree of the polynomial; see de Boor (1977, 1978), Cox (1981) or Dierckx (1993). The choice of knots has been a subject of much research: too many knots lead to overfitting of the data, too few knots lead to underfitting. Some authors have proposed automatic schemes for optimizing the number and the positions of the knots (Friedman and Silverman, 1989; Kooperberg and Stone, 1991, 1992). This is a diffi-

Paul H. C. Eilers is Department Head in the computing section of DCMR Milieudienst Rijnmond,'s-Gravelandseweg 565, 3119XT Schiedam, The Netherlands (e-mail: paul@dcmr.nl). Brian D. Marx is Associate Professor, Department of Experimental Statistics, Louisiana State University, Baton Rouge, LA 70803-5606 (e-mail: brian@stat.lsu.edu).

cult numerical problem and, to our knowledge, no attractive all-purpose scheme exists.

A different track was chosen by O'Sullivan (1986, 1988). He proposed to use a relatively large number of knots. To prevent overfitting, a penalty on the second derivative restricts the flexibility of the fitted curve, similar to the penalty pioneered for smoothing splines by Reinsch (1967) and that has become the standard in much of the spline literature; see, for example, Eubank (1988), Wahba (1990) and Green and Silverman (1994). In this paper we simplify and generalize the approach of O'Sullivan, in such a way that it can be applied in any context where regression on B-splines is useful. Only small modifications of the regression equations are necessary.

The basic idea is not to use the integral of a squared higher derivative of the fitted curve in the penalty, but instead to use a simple difference penalty on the coefficients themselves of adjacent B-splines. We show that both approaches are very similar for second-order differences. In some applications, however, it can be useful to use differences of a smaller or higher order in the penalty. With our approach it is simple to incorporate a penalty of any order in the (generalized) regression equations.

A major problem of any smoothing technique is the choice of the optimal amount of smoothing, in our case the optimal weight of the penalty. We use cross-validation and the Akaike information criterion (AIC). In the latter the effective dimension, that is, the effective number of parameters, of a model plays a crucial role. We follow Hastie and Tibshirani (1990) in using the trace of the smoother matrix as the effective dimension. Because we use standard regression techniques, this quantity can be computed easily. We find the trace very useful to compare the effective amount of smoothing for different numbers of knots, different degrees of the B-splines and different orders of penalties.

We investigate the conservation of moments of different order, in relation to the degree of the B-splines and the order of the differences in the penalty. To illustrate the use of P-splines, we present the following as applications: smoothing of scatterplots; modeling of dose–response curves; and density estimation.

2. *B*-SPLINES IN A NUTSHELL

Not all readers will be familiar with B-splines. Basic references are de Boor (1978) and Dierckx (1993), but, to illustrate the basic simplicity of the ideas, we explain some essential background here. A B-spline consists of polynomial pieces, connected in a special way. A very simple example is shown at the left of Figure 1(a): one B-spline of degree 1. It consists of two linear pieces; one piece from x_1 to x_2, the other from x_2 to x_3. The knots are x_1, x_2 and x_3. To the left of x_1 and to the right of x_3 this B-spline is zero. In the right part of Figure 1(a), three more B-splines of degree 1 are shown: each one based on three knots. Of course, we can construct as large a set of B-splines as we like, by introducing more knots.

In the left part of Figure 1(b), a B-spline of degree 2 is shown. It consists of three quadratic pieces, joined at two knots. At the joining points not only the ordinates of the polynomial pieces match, but also their first derivatives are equal (but not their second derivatives). The B-spline is based on four adjacent knots: $x_1, \dots, x_4$. In the right part Figure 1(b), three more B-splines of degree 2 are shown.

Note that the B-splines overlap each other. First-degree B-splines overlap with two neighbors, second-degree B-splines with four neighbors and so on. Of course, the leftmost and rightmost splines have less overlap. At a given x, two first-degree (or three second-degree) B-splines are nonzero.

These examples illustrate the general properties of a B-spline of degree q:

- it consists of $q + 1$ polynomial pieces, each of degree q;
- the polynomial pieces join at q inner knots;
- at the joining points, derivatives up to order $q - 1$ are continuous;
- the B-spline is positive on a domain spanned by $q + 2$ knots; everywhere else it is zero;
- except at the boundaries, it overlaps with $2q$ polynomial pieces of its neighbors;
- at a given x, $q + 1$ B-splines are nonzero.

Let the domain from $x_{\min}$ to $x_{\max}$ be divided into n' equal intervals by $n' + 1$ knots. Each interval will be covered by $q + 1$ B-splines of degree q. The total number of knots for construction of the B-splines will be $n' + 2q + 1$. The number of B-splines in the regression is $n = n' + q$. This is easily verified by constructing graphs like those in Figure 1.

B-splines are very attractive as base functions for ("nonparametric") univariate regression. A linear combination of (say) third-degree B-splines gives a smooth curve. Once one can compute the B-splines themselves, their application is no more difficult than polynomial regression.

De Boor (1978) gave an algorithm to compute B-splines of any degree from B-splines of lower degree. Because a zero-degree B-spline is just a constant on one interval between two knots, it is simple to com-

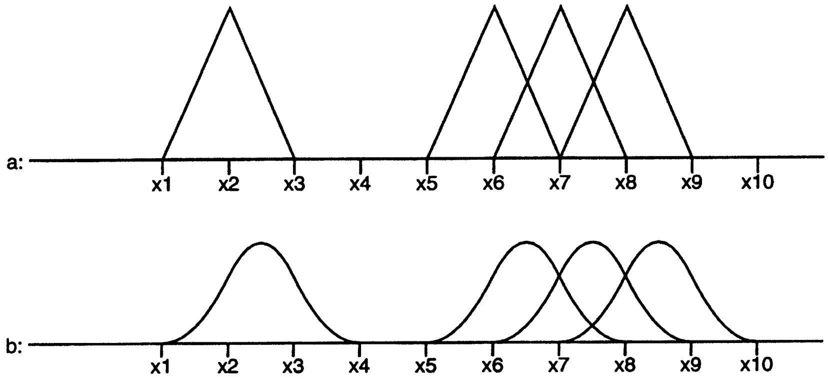

FIG. 1. *Illustrations of one isolated B-spline and several overlapping ones* (a) *degree* 1; (b) *degree* 2.

pute B-splines of any degree. In this paper we use only equidistant knots, but de Boor's algorithm also works for any placement of knots. For equidistant knots, the algorithm can be further simplified, as is illustrated by a small MATLAB function in the Appendix.

Let $B_j(x;\, q)$ denote the value at x of the jth B-spline of degree q for a given equidistant grid of knots. A fitted curve $\hat{y}$ to data (x_i, y_i) is the linear combination $\hat{y}(x) = \sum_{j=1}^{n} \hat{a}_j B_j(x;\, q)$. When the degree of the B-splines is clear from the context, or immaterial, we use $B_j(x)$ instead of $B_j(x;\, q)$.

The indexing of B-splines needs some care, especially when we are going to use derivatives. The indexing connects a B-spline to a knot; that is, it gives the index of the knot that characterizes the position of the B-spline. Our choice is to take the leftmost knot, the knot at which the B-spline starts to become nonzero. In Figure 1(a), x_1 is the positioning knot for the first B-spline. This choice of indexing demands that we introduce q knots to the left of the domain of x. In the formulas that follow for derivatives, the exact bounds of the index in the sums are immaterial, so we have left them out.

De Boor (1978) gives a simple formula for derivatives of B-splines:

$$\begin{aligned} h\sum_j a_j B'_j(x;\, q) &= \sum_j a_j B_j(x;\, q-1) \\ &\quad - \sum_j a_{j+1} B_{j+1}(x;\, q-1) \\ &= -\sum_j \Delta a_{j+1} B_j(x;\, q-1), \end{aligned} \tag{1}$$

where h is the distance between knots and $\Delta a_j = a_j - a_{j-1}$.

By induction we find the following for the second derivative:

$$h^2 \sum_j a_j B''_j(x;\, q) = \sum_j \Delta^2 a_j B_j(x;\, q-2), \tag{2}$$

where $\Delta^2 a_j = \Delta\Delta a_j = a_j - 2a_{j-1} + a_{j-2}$. This fact will prove very useful when we compare continuous and discrete roughness penalties in the next section.

3. PENALTIES

Consider the regression of m data points (x_i, y_i) on a set of n B-splines $B_j(\cdot)$. The least squares objective function to minimize is

$$S = \sum_{i=1}^{m} \left\{ y_i - \sum_{j=1}^{n} a_j B_j(x_i) \right\}^2. \tag{3}$$

Let the number of knots be relatively large, such that the fitted curve will show more variation than is justified by the data. To make the result less flexible, O'Sullivan (1986, 1988) introduced a penalty on the second derivative of the fitted curve and so formed the objective function

$$\begin{aligned} S &= \sum_{i=1}^{m} \left\{ y_i - \sum_{j=1}^{n} a_j B_j(x_i) \right\}^2 \\ &\quad + \lambda \int_{x_{\min}}^{x_{\max}} \left\{ \sum_{j=1}^{n} a_j B''_j(x) \right\}^2 dx. \end{aligned} \tag{4}$$

The integral of the square of the second derivative of a fitted function has become common as a smoothness penalty, since the seminal work on smoothing splines by Reinsch (1967). There is nothing special about the second derivative; in fact, lower or higher orders might be used as well. In the context of smoothing splines, the first derivative leads to simple equations, and a piecewise linear fit, while higher derivatives lead to rather complex mathematics, systems of equations with a high bandwidth, and a very smooth fit.

We propose to base the penalty on (higher-order) finite differences of the coefficients of adjacent B-splines:

$$S = \sum_{i=1}^{m} \left\{ y_i - \sum_{j=1}^{n} a_j B_j(x_i) \right\}^2 + \lambda \sum_{j=k+1}^{n} (\Delta^k a_j)^2. \tag{5}$$

This approach reduces the dimensionality of the problem to n, the number of B-splines, instead of m, the number of observations, with smoothing splines. We still have a parameter λ for continuous control over smoothness of the fit. The difference penalty is a good discrete approximation to the integrated square of the kth derivative. What is more important: with this penalty moments of the data are conserved and polynomial regression models occur as limits for large values of λ. See Section 5 for details.

We will show below that there is a very strong connection between a penalty on second-order differences of the B-spline coefficients and O'Sullivan's choice of a penalty on the second derivative of the fitted function. However, our penalty can be handled mechanically for any order of the differences (see the implementation in the Appendix).

Difference penalties have a long history that goes back at least to Whittaker (1923); recent applications have been described by Green and Yandell (1985) and Eilers (1989, 1991a, b, 1995).

The difference penalty is easily introduced into the regression equations. That makes it possible to experiment with different orders of the differences. In some cases it is useful to work with even the fourth or higher order. This stems from the fact that for high values of λ the fitted curve approaches a parametric (polynomial) model, as will be shown below.

O'Sullivan (1986, 1988) used third-degree B-splines and the following penalty:

$$h^2 P = \lambda \int_{x_{\min}}^{x_{\max}} \left\{ \sum_j a_j B_j''(x; 3) \right\}^2 dx. \tag{6}$$

From the derivative properties of B-splines it follows that

$$h^2 P = \lambda \int_{x_{\min}}^{x_{\max}} \left\{ \sum_j \Delta^2 a_j B_j(x; 1) \right\}^2 dx. \tag{7}$$

This can be written as

$$\begin{aligned} h^2 P = \lambda \int_{x_{\min}}^{x_{\max}} & \sum_j \sum_k \Delta^2 a_j \, \Delta^2 a_k \\ & \cdot B_j(x; 1) B_k(x; 1) \, dx. \end{aligned} \tag{8}$$

Most of the cross products of $B_j(x; 1)$ and $B_k(x; 1)$ disappear, because B-splines of degree 1 only overlap when j is $k-1$, k or $k+1$. We thus have that

$$\begin{aligned} h^2 P = \lambda \int_{x_{\min}}^{x_{\max}} \Bigg[& \left\{ \sum_j \Delta^2 a_j B_j(x; 1) \right\}^2 \\ & + 2 \sum_j \Delta^2 a_j \, \Delta^2 a_{j-1} \\ & \cdot B_j(x; 1) B_{j-1}(x; 1) \Bigg] dx, \end{aligned} \tag{9}$$

or

$$\begin{aligned} h^2 P = {} & \lambda \sum_j (\Delta^2 a_j)^2 \int_{x_{\min}}^{x_{\max}} B_j^2(x; 1) \, dx \\ & + 2\lambda \sum_j \Delta^2 a_j \, \Delta^2 a_{j-1} \\ & \cdot \int_{x_{\min}}^{x_{\max}} B_j(x; 1) B_{j-1}(x; 1) \, dx, \end{aligned} \tag{10}$$

which can be written as

$$h^2 P = \lambda \left\{ c_1 \sum_j (\Delta^2 a_j)^2 + c_2 \sum_j \Delta^2 a_j \, \Delta^2 a_{j-1} \right\}, \tag{11}$$

where c_1 and c_2 are constants for given (equidistant) knots:

$$\begin{aligned} c_1 &= \int_{x_{\min}}^{x_{\max}} B_j^2(x; 1) \, dx; \\ c_2 &= \int_{x_{\min}}^{x_{\max}} B_j(x; 1) B_{j-1}(x; 1) \, dx. \end{aligned} \tag{12}$$

The first term in (11) is equivalent to our second-order difference penalty, the second term contains cross products of neighboring second differences. This leads to more complex equations when minimizing the penalized likelihood (equations in which seven adjacent a_j's occur, compared to five if only squares of second differences occur in the penalty). The higher complexity of the penalty equations stems from the overlapping of B-splines. With higher order differences and/or higher degrees of the B-splines, the complications grow rapidly and make it rather difficult to construct an automatic procedure for incorporating the penalty in the likelihood equations. With the use of a difference penalty on the coefficients of the B-splines this problem disappears.

4. PENALIZED LIKELIHOOD

For least squares smoothing we have to minimize S in (5). The system of equations that follows from the minimization of S can be written as:

$$B^T y = (B^T B + \lambda D_k^T D_k) a, \tag{13}$$

where D_k is the matrix representation of the difference operator Δ^k, and the elements of B are $b_{ij} = B_j(x_i)$. When $\lambda = 0$, we have the standard normal equations of linear regression with a B-spline basis. With $k = 0$ we have a special case of ridge regression. When $\lambda > 0$, the penalty only influences the main diagonal and k subdiagonals (on both sides of the main diagonal) of the system of equations. This system has a banded structure because of the limited overlap of the B-splines. It is seldom worth the trouble to exploit this special structure, as the number of equations is equal to the number of splines, which is generally moderate (10–20).

In a generalized linear model (GLM), we introduce a linear predictor $\eta_i = \sum_{j=1}^{n} b_{ij} a_j$ and a (canonical) link function $\eta_i = g(\mu_i)$, where μ_i is the expectation of y_i. The penalty now is subtracted from the log-likelihood $l(y; a)$ to form the penalized likelihood function

$$L = l(y; a) - \frac{\lambda}{2} \sum_{j=k+1}^{n} (\Delta^k a_j)^2. \tag{14}$$

The optimization of L leads to the following system of equations:

$$B^T(y - \mu) = \lambda D_k^T D_k a. \tag{15}$$

These are solved as usual with iterative weighted linear regressions with the system

$$\begin{aligned} &B^T \tilde{W}(y - \tilde{\mu}) + B^T \tilde{W} B \tilde{a} \\ &\quad = (B^T \tilde{W} B + \lambda D_k^T D_k) a, \end{aligned} \tag{16}$$

where $\tilde{a}$ and $\tilde{\mu}$ are current approximations to the solution and $\tilde{W}$ is a diagonal matrix of weights

$$w_{ii} = \frac{1}{v_i} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2, \tag{17}$$

where v_i is the variance of y_i, given μ_i. The only difference with the standard procedure for fitting of GLM's (McCullagh and Nelder, 1989), with B-splines as regressors, is the modification of $B^T \tilde{W} B$ by $\lambda D_k^T D_k$ (which itself is constant for fixed λ) at each iteration.

5. PROPERTIES OF *P*-SPLINES

P-splines have a number of useful properties, partially inherited from B-splines. We give a short overview, with somewhat informal proofs.

In the first place: P-splines show no boundary effects, as many types of kernel smoothers do. By this we mean the spreading of a fitted curve or density outside of the (physical) domain of the data, generally accompanied by bending toward zero. In Section 8 this aspect is considered in some detail, in the context of density smoothing.

P-splines can fit polynomial data exactly. Let data (x_i, y_i) be given. If the y_i are a polynomial in x of degree k, then B-splines of degree k or higher will exactly fit the data (de Boor, 1977). The same is true for P-splines, if the order of the penalty is $k + 1$ or higher, whatever the value of λ. To see that this is true, take the case of a first-order penalty and the fit to data y that are constant (a polynomial of degree 0). Because $\sum_{j=1}^{n} \hat{a}_j B_j(x) = c$, we have that $\sum_{j=1}^{n} \hat{a}_j B'_j(x_i) = 0$, for all x. Then it follows from the relationship between differences and derivatives in (1) that all Δa_j are zero, and thus that $\sum_{j=2}^{n} \Delta a_j = 0$. Consequently, the penalty has no effect and the fit is the same as for unpenalized B-splines. This reasoning can easily be extended by induction to data with a linear relationship between x and y, and a second order difference penalty.

P-splines can conserve moments of the data. For a linear model with P-splines of degree $k + 1$ and a penalty of order $k + 1$, or higher, it holds that

$$\sum_{i=1}^{m} x^k y_i = \sum_{i=1}^{m} x^k \hat{y}_i, \tag{18}$$

for all values of λ, where $\hat{y}_i = \sum_{j=1}^{n} b_{ij} \hat{a}_j$ are the fitted values. For GLM's with canonical links it holds that

$$\sum_{i=1}^{m} x^k y_i = \sum_{i=1}^{m} x^k \hat{\mu}_i. \tag{19}$$

This property is especially useful in the context of density smoothing: the mean and variance of the estimated density will be equal to mean and variance of the data, for any amount of smoothing. This is an advantage compared to kernel smoothers: these inflate the variance increasingly with stronger smoothing.

The limit of a P-splines fit with strong smoothing is a polynomial. For large values of λ and a penalty of order k, the fitted series will approach a polynomial of degree $k - 1$, if the degree of the B-splines is equal to, or higher than, k. Once again, the relationships between derivatives of a B-spline fit and differences of coefficients, as in (1) and (2), are the key. Take the example of a second-order difference penalty: when λ is large, $\sum_{j=3}^{n} (\Delta^2 a_j)^2$ has to be very near zero. Thus each of the second differences has to be near zero, and thus the second derivative of the fit has to be near zero everywhere. In view of these very useful results, it seems that B-splines and difference penalties are the ideal marriage.

It is important to focus on the linearized smoothing problem that is solved at each iteration, because we will make use of properties of the smoothing matrix. From (16) follows for the hat matrix H:

$$H = B(B^T \tilde{W} B + \lambda D_k^T D_k)^{-1} B^T \tilde{W}. \tag{20}$$

The trace of H will approach k as λ increases. A proof goes as follows. Let

$$(21)\qquad Q_B = B^T \tilde{W} B \quad \text{and} \quad Q_\lambda = \lambda D^T D.$$

Write $\operatorname{tr}(H)$ as

$$(22)\qquad \begin{aligned} \operatorname{tr}[H] &= \operatorname{tr}\{(Q_B + Q_\lambda)^{-1} Q_B\} \\ &= \operatorname{tr}\{Q_B^{1/2}(Q_B + Q_\lambda)^{-1} Q_B^{1/2}\} \\ &= \operatorname{tr}\{(I + Q_B^{-1/2} Q_\lambda Q_B^{-1/2})^{-1}\}. \end{aligned}$$

This can be written as

$$(23)\qquad \operatorname{tr}(H) = \operatorname{tr}\{(I + \lambda L)^{-1}\} = \sum_{j=1}^{n} \frac{1}{1 + \lambda \gamma_j},$$

where

$$(24)\qquad L = Q_B^{-1/2} Q_\lambda Q_B^{-1/2}$$

and γ_j, for $j = 1, \ldots, n$, are the eigenvalues of L. Because k eigenvalues of Q_λ are zero, L has k zero eigenvalues. When λ is large, only the (k) terms with $\gamma_j = 0$ contribute to the leftmost term, and thus to the trace of H. Hence $\operatorname{tr}(H)$ approaches k for large λ.

6. OPTIMAL SMOOTHING, AIC AND CROSS-VALIDATION

Now that we can easily influence the smoothness of a fitted curve with λ, we need some way to choose an "optimal" value for it. We propose to use the Akaike information criterion (AIC).

The basic idea of AIC is to correct the log-likelihood of a fitted model for the effective number of parameters. An extensive discussion and applications can be found in Sakamoto, Ishiguro and Kitagawa (1986). Instead of the log-likelihood, the deviance is easier to use. The definition of AIC is equivalent to

$$(25)\qquad \mathrm{AIC}(\lambda) = \operatorname{dev}(y; a, \lambda) + 2 * \dim(a, \lambda),$$

where $\dim(a, \lambda)$ is the (effective) dimension of the vector of parameters, a, and $\operatorname{dev}(y; a, \lambda)$ is the deviance.

Computation of the deviance is straightforward, but how shall we determine the effective dimension of our P-spline fit? We find a solution in Hastie and Tibshirani (1990). They discuss the effective dimensions of linear smoothers and propose to use the trace of the smoother matrix as an approximation. In our case that means $\dim(a) = \operatorname{tr}(H)$. Note that $\operatorname{tr}(H) = n$ when $\lambda = 0$, as in (nonsingular) standard linear regression.

As $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ (for conformable matrices), it is computationally advantageous to use

$$(26)\qquad \begin{aligned} \operatorname{tr}(H) &= \operatorname{tr}\{B(B^T W B + \lambda D_k^T D_k)^{-1} B^T W\} \\ &= \operatorname{tr}\{(B^T W B + \lambda D_k^T D_k)^{-1} B^T W B\}. \end{aligned}$$

The latter expression involves only n-by-n matrices, whereas H is an m-by-m matrix.

In some GLM's, the scale of the data is known, as for counts with a Poisson distribution and for binomial data; then the deviance can be computed directly. For linear data, an estimate of the variance is needed. One approach is to take the variance of the residuals from the $\hat{y}_i$ that are computed when $\lambda = 0$, say, $\hat{\sigma}_0^2$:

$$(27)\qquad \begin{aligned} \mathrm{AIC} = \sum_{i=1}^{m} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\sigma}_0^2} &+ 2 \operatorname{tr}(H) \\ -2m \ln \hat{\sigma}_0 &- m \ln 2\pi. \end{aligned}$$

This choice for the variance is rather arbitrary, as it depends on the numer of knots. Alternatives can be based on (generalized) cross-validation. For ordinary cross-validation we compute

$$(28)\qquad \mathrm{CV}(\lambda) = \sum_{i=1}^{m} \left(\frac{y_i - \hat{y}_i}{1 - h_{ii}} \right)^2,$$

where the h_{ii} are the diagonal elements of the hat matrix H. For generalized cross-validation (Wahba, 1990), we compute

$$(29)\qquad \mathrm{GCV}(\lambda) = \sum_{i=1}^{m} \frac{(y_i - \hat{y}_i)^2}{(m - \sum_{i=1}^{m} h_{ii})^2}.$$

The difference between both quantities is generally small. The best λ is the value that minimizes $\mathrm{CV}(\lambda)$ or $\mathrm{GCV}(\lambda)$. The variance of the residuals at the optimal λ is a natural choice to use as an estimate of σ_0^2 for the computation of $\mathrm{AIC}(\lambda)$. It is practical to work with modified versions of $\mathrm{CV}(\lambda)$ and $\mathrm{GCV}(\lambda)$, with values that can be interpreted as estimates of the cross-validation standard deviation:

$$(30)\qquad \begin{aligned} \overline{\mathrm{CV}(\lambda)} &= \sqrt{\mathrm{CV}(\lambda)/m}; \\ \overline{\mathrm{GCV}(\lambda)} &= \sqrt{m\, \mathrm{GCV}(\lambda)}. \end{aligned}$$

The two terms in $\mathrm{AIC}(\lambda)$ represent the deviance and the trace of the smoother matrix. The latter term, say $T(\lambda) = \operatorname{tr}\{H(\lambda)\}$, is of interest on its own, because it can be interpreted as the effective dimension of the fitted curve.

$T(\lambda)$ is useful to compare fits for different numbers of knots and orders of penalties, whereas λ can vary over a large range of values and has no clear intuitive appeal. We will show in an example below

TABLE 1
Values of several diagnostics for the motorcycle impact data, for several values of λ

λ	0.001	0.01	0.1	0.2	0.5	1	2	5	10
$\overline{\mathrm{CV}}$	24.77	24.02	23.52	23.37	23.26	23.38	23.90	25.50	27.49
$\overline{\mathrm{GCV}}$	25.32	24.93	24.17	23.94	23.74	23.81	24.28	25.87	27.85
AIC	159.6	156.2	149.0	146.7	144.7	145.4	150.6	169.1	194.3
tr(H)	21.2	19.4	15.13	13.6	11.7	10.4	9.2	7.7	6.8

that a plot of AIC against T is a useful diagnostic tool.

In the case of P-splines, the maximum value that $T(\lambda)$ can attain is equal to the number of B-splines (when $\lambda = 0$). The actual maximum depends on the number and the distributions of the data points. The minimum value of $T(\lambda)$ occurs when λ goes to infinity; it is equal to the order of the difference penalty. This agrees with the fact that for high values of λ the fit of P-splines approaches a polynomial of degree $k-1$.

7. APPLICATIONS TO GENERALIZED LINEAR MODELLING

In this section we apply P-splines to a number of nonparametric modelling situations, with normal as well as nonnormal data.

First we look at a problem with additive errors. Silverman (1985) used motorcycle crash helmet impact data to illustrate smoothing of a scatterplot with splines; the data can be found in Härdle (1990) and (also on diskette) in Hand et al. (1994). The data give head acceleration in units of g, at different times after impact in simulated accidents. We smooth with B-splines of degree 3 and a second-order penalty. The chosen knots divide the domain of x (0–60) into 20 intervals of equal width. When we vary λ on an approximately geometric grid, we get the results in Table 1, where $\hat{\sigma}_0$ is computed from GCV(λ) at the optimal value of λ. At the optimal value of λ as determined by $\overline{\mathrm{GCV}}$, we get the results as plotted in Figure 2.

It is interesting to note that the amount of work to investigate several values of λ is largely independent of the number of data points when using $\overline{\mathrm{GCV}}$. The system to be solved is

$$(31)\qquad (B^TB + \lambda D_k^T D_k)a = B^T y.$$

The sum of squares is

$$(32)\quad S = |y - Ba|^2 = y^T y - 2a^T B^T y + a^T B^T Ba.$$

So B^TB and B^Ty have to be computed only once. The hat matrix H is m by m, but for its trace we found an expression in (26) that involves only B^TB and $D_k^T D_k$. So we do not need the original data for cross-validation at any value of λ.

Our second example concerns logistic regression. The model is

$$(33)\qquad \ln\left(\frac{p_i}{1-p_i}\right) = \eta_i = \sum_{j=1}^{n} a_j B_j(x_i).$$

The observations are triples (x_i, t_i, y_i), where t_i is the number of individuals under study at dose x_i, and y_i is the number of "successes." We assume that y_i has a binomial distribution with probability p_i and t_i trials. The expected value of y_i is $t_i p_i$ and the variance is $t_i p_i(1-p_i)$.

Figure 3 shows data from Ashford and Walker (1972) on the numbers of Trypanosome organisms killed at different doses of a certain poison. The data points and two fitted curves are shown. For the thick line curve $\lambda = 1$ and AIC $= 13.4$; this value of λ is optimal for the chosen B-splines of degree 3 and a penalty of order 2. The thin line curve shows the fit for $\lambda = 10^8$ (AIC $= 27.8$). With a second-order penalty, this essentially a logistic fit.

Figure 4 shows curves of AIC(λ) against $T(\lambda)$ at different values of k, the order of the penalty. We find that $k = 3$ can give a lower value of AIC (for $\lambda = 5$, AIC $= 11.8$). For $k = 4$ we find that a very high value of λ is allowed; then AIC $= 11.4$, hardly different from the lowest possible value (11.1). A large value of λ with a fourth-order penalty means that effectively the fitted curve for η is a third-order polynomial. The limit of the fit with P-splines thus indicates a cubic logistic fit as a good parametric model. Here we have seen an application where a fourth-order penalty is useful.

Our third example is a time series of counts y_i, which we will model with a Poisson distribution with smoothly changing expectation:

$$(34)\qquad \ln \mu_i = \eta_i = \sum_{j=1}^{n} a_j B_j(x_i).$$

In this special case the x_i are equidistant, but this is immaterial. Figure 5 shows the numbers of disasters in British coal mines for the years 1850–1962, as presented in (Diggle and Marron, 1988). The counts are drawn as narrow vertical bars, the line is the fitted trend. The number of intervals is 20, the B-splines have degree 3 and the order of the penalty is 2. An optimal value of λ was searched on the approximately geometric grid 1, 2, 5, 10 and

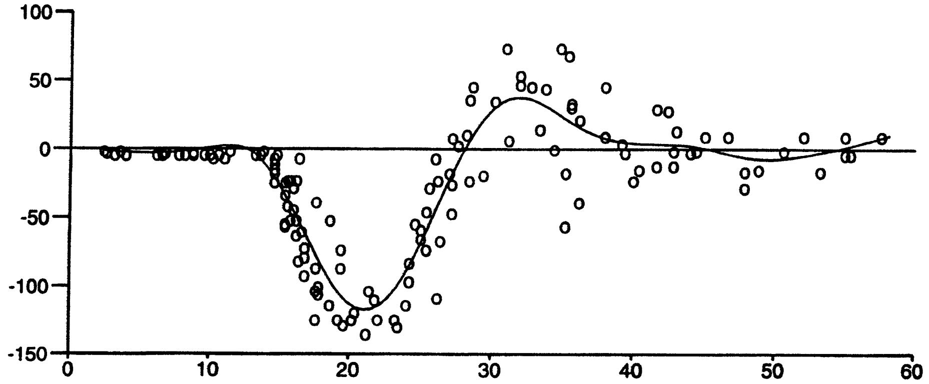

FIG. 2. *Motorcycle crash helmet impact data: optimal fit with B-splines of third degree, a second-order penalty and* $\lambda = 0.5$.

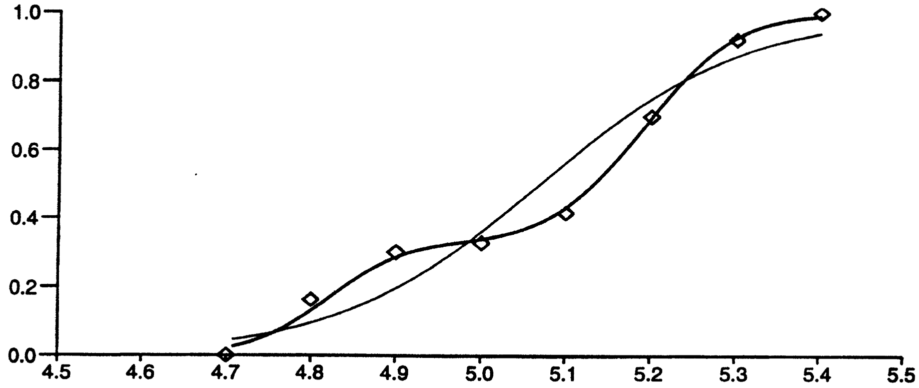

FIG. 3. *Nonparametric logistic regression of Trypanosome data: P-splines of order* 3 *with* 13 *knots, difference penalty of order* 2, $\lambda = 1$ *and* AIC = 13.4 (*thick line*); *the thin line is effectively the logistic fit* ($\lambda = 10^8$ *and* AIC = 27.8).

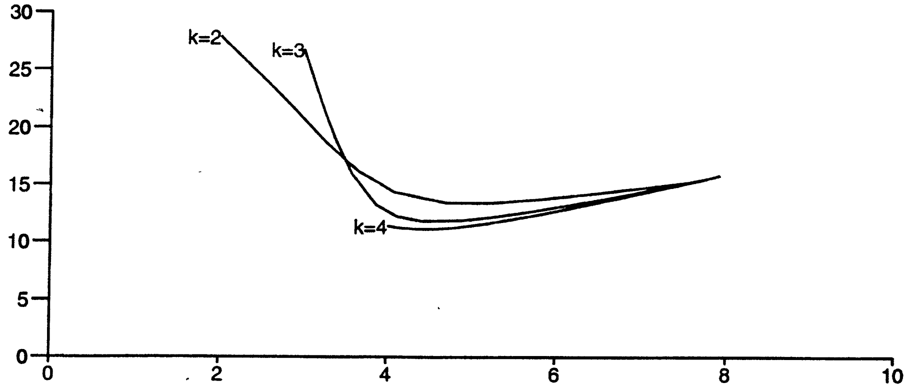

FIG. 4. AIC(λ) *versus* $T(\lambda)$, *the effective dimension, for several orders of the penalty* (k).

so on. The minimum of AIC (126.0) was found for $\lambda = 1{,}000$.

The raw data of the coal mining accidents presumably were the dates on which they occurred. So the data we use herę are in fact a histogram with one-year-wide bins. With events on a time scale it seems natural to smooth counts over intervals, but the same idea applies to any form of histogram (bin counts) or density smoothing. This was already noted by Diggle and Marron (1988). In the next section we take a detailed look at density smoothing with P-splines.

8. DENSITY SMOOTHING

In the preceding section we noted that a time series of counts is just a histogram on the time axis. Any other histogram might be smoothed in the same

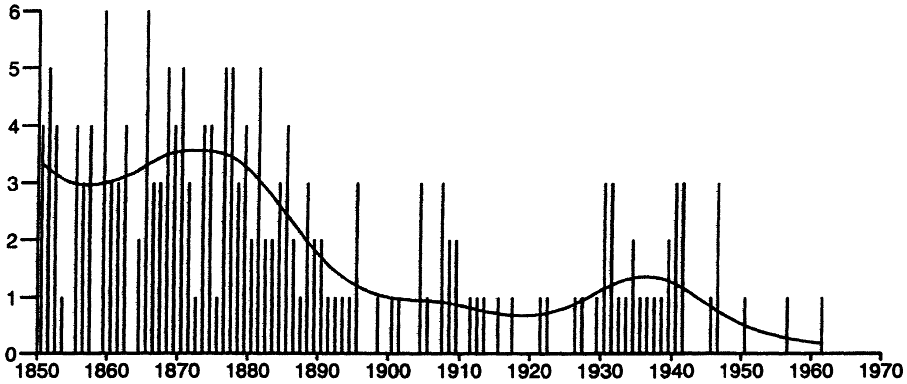

FIG. 5. *Numbers of severe accidents in British coal mines: number per year shown as vertical lines; fitted trend of the expectation of the Poisson distribution; B-splines of degree* 3, *penalty of order* 3, 20 *intervals between* 1850 *and* 1970, $\lambda = 1{,}000$ *and* AIC = 126.0.

way. However, it is our experience that this idea is hard to swallow for many colleagues. They see the construction of a frequency histogram as an unallowable discretization of the data and as a prelude to disaster. Perhaps this feeling stems from the well-known fact that maximum likelihood estimation of histograms leads to pathological results, namely, delta functions at the observations (Scott, 1992). However, if we optimize a penalized likelihood, we arrive at stable and very useful results, as we will show below.

Let y_i, $i = 1, \dots, m$, be a histogram. Let the origin of x be chosen in such a way that the midpoints of the bins are $x_i = ih$; thus y_i is the number of raw observations with $x_i - h/2 \le x < x_i + h/2$. If p_i is the probability of finding a raw observation in cell i, then the likelihood of the given histogram is proportional to the multinomial likelihood $\prod_{i=1}^m p_i^{y_i}$. Equivalently (see Bishop, Fienberg and Holland, 1975, Chapter 13), one can work with the likelihood of m Poisson distributions with expectations $\mu_i = p_i y_+$, where $y_+ = \sum_{i=1}^m y_i$.

To smooth the histogram, we again use a generalized linear model with the canonical log link (which guarantees positive μ):

$$\ln \mu_i = \eta_i = \sum_{j=1}^{n} a_j B_j(x_i) \tag{35}$$

and construct the penalized log likelihood

$$L = \sum_{i=1}^{m} y_i \ln \mu_i - \sum_{i=1}^{m} \mu_i - \lambda \sum_{j=k+1}^{n} \frac{(\Delta^k a_j)^2}{2}, \tag{36}$$

with n a suitable (i.e., relatively large) number of knots for the B-splines. The penalized likelihood equations follow from the minimization of L:

$$\sum_{i=1}^{m} (y_i - \mu_i) B_j(x_i) = \lambda \sum_{l=k+1}^{n} d_{jl} a_l. \tag{37}$$

These equations are solved with iteratively reweighted regression, as described in Section 4.

Now we let h, the width of the cells of the histogram, shrink to a very small value. If the raw data are given to infinite precision, we will eventually arrive at a situation in which each cell of the histogram has at most one observation. In other words, we have a very large number (m) of cells, of which y_+ are 1 and all others 0. Let I be the set of indices of cells for which $y_i = 1$. Then

$$\sum_{i=1}^{m} y_i B_j(x_i) = \sum_{i \in I} B_j(x_i). \tag{38}$$

If the raw observations are u_t for $t = 1, \dots, r$, with $r = y_+$, then we can write

$$\sum_{i \in I} B_j(x_i) = \sum_{t=1}^{r} B_j(u_t) = B_j^+, \tag{39}$$

and the penalized likelihood equations in (37) change to

$$B_j^+ - \sum_{i=1}^{m} \mu_i B_j(x_i) = \lambda \sum_{l=k+1}^{n} d_{jl} a_l. \tag{40}$$

For any j, the first term on the left-hand side of (40) can be interpreted as the "empirical sum" of B-spline j, while the second term on the left can be interpreted as the "expected sum" of that B-spline for the fitted density. When $\lambda = 0$, these terms have to be equal to each other for each j.

Note that the second term on the left-hand side of (40) is in fact a numerical approximation of an integral:

$$\begin{aligned} &\sum_{i=1}^{m} \mu_i B_j(x_i)/y_+ \\ &\quad \approx \int_{x_{\min}}^{x_{\max}} B_j(x) \exp\left\{ \sum_{l=1}^{n} a_l B_l(x) \right\} dx. \end{aligned} \tag{41}$$

TABLE 2
The value of AIC at several values of lambda for the Old Faithful density estimate

λ	0.001	0.01	0.02	0.05	0.1	0.2	0.5	1	10
AIC	50.79	48.21	47.67	47.37	47.70	48.61	50.59	52.81	65.66

The smaller h (the larger m), the better the approximation. In other words: the discretization is only needed to solve an integral numerically for which, as far as we know, no closed form solution exists. For practical purposes the simple sum is sufficient, but a more sophisticated integration scheme is possible. Note that the sums to calculate B_j^+ involve all raw observations, but in fact at each of these only $q+1$ terms $B_j(u_t)$ add to their corresponding B_j^+.

The necessary computations can be done in terms of the sufficient statistics B_j^+: we have seen their role in the penalized likelihood equations above. But also the deviance and thus AIC can be computed directly:

$$\begin{aligned} \operatorname{dev}(y; a) &= 2\sum_{i=1}^{m} y_i \ln(y_i/\mu_i) \\ &= 2\sum_{i=1}^{m} y_i \ln y_i - 2\sum_{i=1}^{m} y_i \sum_{j=1}^{n} a_j B_j(x_i) \qquad (42) \\ &= 2\sum_{i=1}^{m} y_i \ln y_i - 2\sum_{j=1}^{n} a_j B_j^+. \end{aligned}$$

In the extreme case, when the y_i are either 0 or 1, the term $\sum y_i \ln y_i$ vanishes. In any case it is independent of the fitted density.

The density smoother with P-splines is very attractive: the estimated density is positive and continuous, it can be described relatively parsimoniously in terms of the coefficients of the B-splines, and it is a proper density. Moments are conserved, as follows from (19). This implies that with third-degree B-splines and a third-order penalty, mean and variance of the estimated distribution are equal to those of the raw data, whatever the amount of smoothing; the limit for high λ is a normal distribution.

The P-spline density smoother is not troubled by boundary effects, as for instance kernel smoothers are. Marron and Ruppert (1994) give examples and a rather complicated remedy, based on transformations. With P-splines no special precautions are necessary, but it is important to specify the domain of the data correctly. We will present an example below.

We now take as a first example a data set from (Silverman, 1986). The data are durations of 107 eruptions of the Old Faithful geyser. Third-degree B-splines were used, with a third-order penalty. The domain from 0 to 6 was divided into 20 intervals to determine the knots. In the figure two fits are shown, for $\lambda = 0.001$ and for $\lambda = 0.05$. The latter value gives the minimum of AIC, as Table 2 shows. We see that of the two clearly separated humps, the right one seems to be a mixture of two peaks.

The second example also comes from (Silverman, 1986). The data are lengths of spells of psychiatric treatments in a suicide study. Figure 7 shows the raw data and the estimated density when the domain is chosen from 0 to 1,000. Third-degree B-splines were used, with a second-order penalty. A fairly large amount of smoothing ($\lambda = 100$) is indicated by AIC; the fitted density is nearly exponential. In fact, if one considers only the domain from 0 to 500, then λ can become arbitrarily large and a pure exponential density results. However, if we choose the domain from -200 to 800 we get a quite different fit, as Figure 8 shows. By extending the domain we force the estimated density also to cover negative values of x, where there are no data (which means zero counts). Consequently, it has to drop toward zero, missing the peak for small positive values. The optimal value of λ now is 0.01 and a much more wiggly fit results, with an appreciably higher value of AIC. This nicely illustrates how, with a proper choice of the domain, the P-spline density smoother can be free from the boundary effects that give so much trouble with kernel smoothers.

9. DISCUSSION

We believe that P-splines come near to being the ideal smoother. With their grounding in classic regression methods and generalized linear models, their properties are easy to verify and understand. Moments of the data are conserved and the limiting behavior with a strong penalty is well defined and gives a connection to polynomial models. Boundary effects do not occur if the domain of the data is properly specified.

The necessary computations, including cross-validation, are comparable in size to those for a medium sized regression problem. The regression context makes it natural to extend P-splines to semiparametric models, in which additional explanatory variables occur. The computed fit is described compactly by the coefficients of the B-splines.

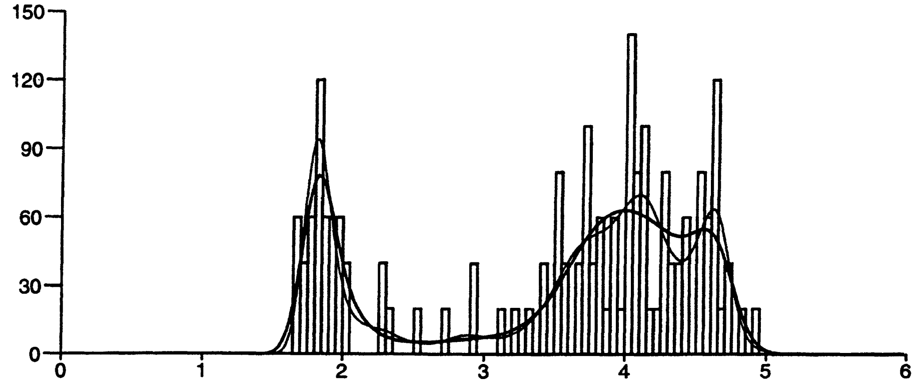

FIG. 6. *Density smoothing of durations of Old Faithful geyser eruptions: density histogram and fitted densities; thin line, third-order penalty with* $\lambda = 0.001$(AIC = 84.05); *thick line, optimal* $\lambda = 0.05$, *with* AIC = 80.17; *B-splines of degree* 3 *with* 20 *intervals on the domain from* 1 *to* 6.

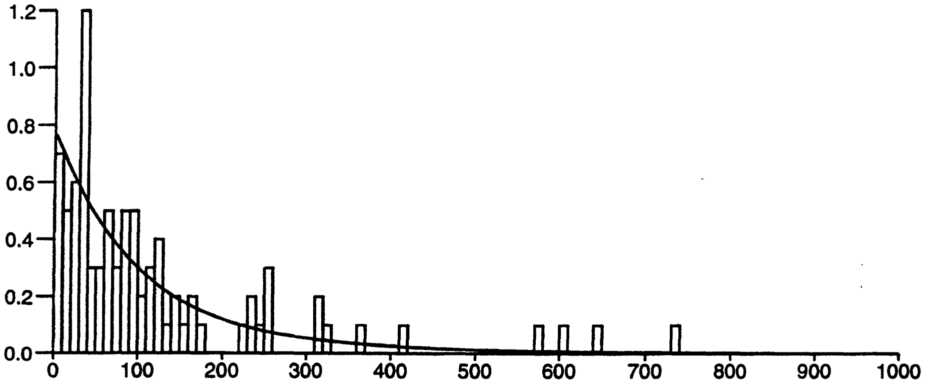

FIG. 7. *Density smoothing of suicide data: positive domain* (0–1,000); *B-splines of degree* 3, *penalty of order* 2, 20 *intervals,* $\lambda =$ 100, AIC = 69.9.

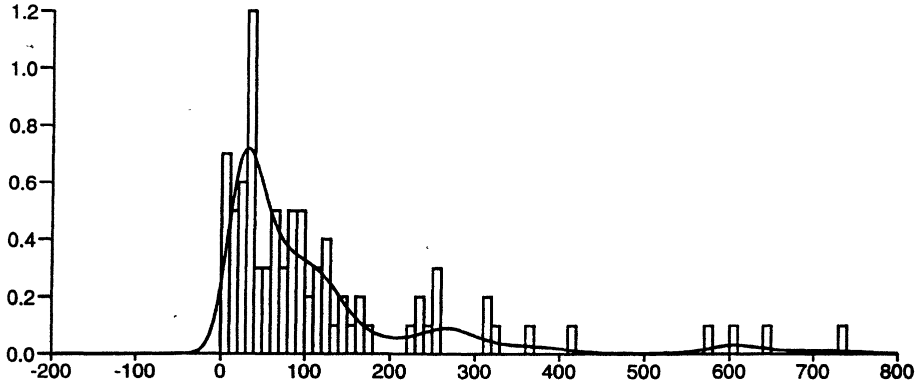

FIG. 8. *Density smoothing of suicide data: the domain includes negative values* (−200–800); *B-splines of degree* 3, *penalty of order* 2, 20 *intervals,* $\lambda = 0.01$, AIC = 83.6.

P-splines can be very useful in (generalized) additive models. For each dimension a B-spline basis and a penalty are introduced. With n knots in each base and d dimensions, a system of nd-by-nd (weighted) regression equations results. Backfitting, the iterative smoothing for each separate dimension, is eliminated. We have reported on this application elsewhere (Marx and Eilers, 1994, 1996).

Penalized likelihood is a subject with a growing popularity. We already mentioned the work of

O'Sullivan. In the book by Green and Silverman (1994), many applications and references can be found. Almost exclusively, penalties are defined in terms of the square of the second derivative of the fitted curve. Generalizations to penalties on higher derivatives have been mentioned in the literature, but to our knowledge, practical applications are very rare. The shift from the continuous penalty to the discrete penalty in terms of the coefficents of the B-splines is not spectacular in itself. But we have seen that it leads to very useful results, while giving a mechanical way to work with higher-order penalties. The modelling of binomial dose–response in Section 7 showed the usefulness of higher-order penalties.

A remarkable property of AIC is that it is easier to compute it for certain nonnormal distributions, like the Poisson and binomial, than for normal distributions. This is so because for these distributions the relationship between mean and variance is known. We should warn the reader that AIC may lead to undersmoothing when the data are overdispersed, since the assumed variance of the data may then be too low. We are presently investigating smoothing with P-splines and overdispersed distributions like the negative binomial and the beta-binomial. Also ideas of quasilikelihood will be incorporated.

We have paid extra attention to density smoothing, because we feel that in this area the advantages of P-splines really shine. Traditionally, kernel smoothers have been popular in this field, but they inflate the variance and have troubles with boundaries of data domains; their computation is expensive, cross-validation even more so, and one cannot report an estimated density in a compact way.

Possibly kernel smoothers still have advantages in two or more dimensions, but it seems that P-splines can also be used for two-dimensional smoothing with Kronecker products of B-splines. With a grid of, say, 10 by 10 knots and a third-order penalty, a system of 130 equations results, with half bandwidth of approximately 30. This can easily be handled on a personal computer. The automatic construction of the equations will be more difficult than in one dimension. First experiments with this approach look promising; we will report on them in due time.

We have not touched on many obvious and interesting extensions to P-splines. Robustness can be obtained with any nonlinear reweighting scheme that can be used with regression models. Circular domains can be handled by wrapping the B-splines and the penalty around the origin. The penalty can be extended with weights, to give a fit with nonconstant stiffness. It this way it will be easy to specify a varying stiffness, but it is quite another matter to estimate the weights from the data.

Finally, we like to remark that P-splines form a bridge between the purely discrete smoothing problem, as set forth originally by Whittaker (1923) and continuous smoothing. B-splines of degree zero are constant on an interval between two knots, and zero elsewhere; they have no overlap. Thus the fitted function gives for each interval the value of the coefficient of the corresponding B-spline.

APPENDIX: COMPUTATIONAL DETAILS

Here we look at the computation of B-splines and derivatives of the penalty. We use S-PLUS and MATLAB as example languages because of their widespread use. Also we give some impressions of the speed of the computations.

In the linear case we have to solve the system of equations

$$(B^T B + \lambda D_k^T D_k)\hat{a} = B^T y \tag{43}$$

and to compute $|y - B\hat{a}|^2$ and $\mathrm{tr}\{(B^T B + \lambda D^T D)^{-1} \cdot B^T B\}$. We need a function to compute B, the B-spline base matrix. In S-PLUS, this is a simple matter, as there is a built-in function `spline.des()` that computes (derivatives) of B-splines. We only have to construct the sequence of knots. Let us assume that `xl` is the left of the x-domain, `xr` the right, and that there are `ndx` intervals on that domain. To compute B for a given vector `x`, based on B-splines of degree `bdeg`, we can use the following function:

```
bspline <- function(x, xl, xr, ndx, bdeg) {
 dx <- (xr - xl) / ndx
 knots <- seq(xl - bdeg * dx, xr + bdeg * dx, by = dx)
 B <- spline.des(knots, x, bdeg + 1, 0 * x)$design
 B
}
```

Note that S-PLUS works with the order of B-splines, following the original definition of de Boor (1977): the order is the degree plus 1.

The matrix D_k can also be computed easily. The identity matrix of size `n` by `n` is constructed by `diag(n)` and there is a built-in function `diff()` to difference it. With a short loop we arrive at D_k. The computations thus are given as (with `pord` the order of the penalty) follows:

```
B <- bspline(x, xl, xr, ndx, bdeg)
D <- diag(ncol(B))
for (k in 1:pord) D <- diff(D)
a <- solve(t(B) %*% B + lambda * t(D) %*% D,
       t(B) %*% y)
```

```
  yhat <- B %*% a
  s <- sum((y - yhat)^2)
  Q <- solve(t(B) %*% B + lambda * t(D) %*% D)
        # matrix inversion
  t <- sum(diag(Q %*% (t(B) %*% B)))
  gcv <- s / (nrow(B) - t)^2
```

There is room to optimize the computations above by storing and reusing intermediate results.

MATLAB has no built-in function to compute B-splines, so we have to program the recursions ourself. We start with the recurrence relation that is given in de Boor (1978, Chapter 10):

$$\frac{B_{j,k}(x)}{t_{j+k}-t_j} = \frac{x-t_j}{t_{j+k-1}-t_j}\,\frac{B_{j,k-1}(x)}{t_{j+k-1}-t_j} + \frac{t_{j_k}-x}{t_{j+k}-t_j}\,\frac{B_{j+1,k-1}(x)}{t_{j+k}-t_{j+1}}, \tag{44}$$

where $B_{j,k}(x)$ in de Boor's notation is our $B_j(x;\, k-1)$ (de Boor uses order 1 for the constant B-splines, whereas we use degree 0). The use of a uniform grid of knots at distances $dx = (x_{\max} - x_{\min})/n'$ greatly simplifies the formulas. If we define $p = (x - x_{\min})/dx$, we arrive at the following recurrence formula:

$$B_j(x;\, k) = \frac{k+p-j+1}{k} B_{j-1}(x;\, k-1) + \frac{j-p}{k} B_j(x;\, k-1). \tag{45}$$

The recursion can be started with $k = 0$, because $B_j(x;\, 0) = 1$ when $(j-1)dx < x - x_{\min} \le jd$, and zero for all other j. Also, $B_j(x;\, k) = 0$ for $j < 0$ and $j > n$. This leads to the following function:

```
function B = bspline(x, xl, xr, ndx, bdeg)
 dx = (xr - xl) / ndx;
 t = xl + dx * [-bdeg:ndx-1];
 T = (0 * x + 1) * t;
 X = x * (0 * t + 1);
 P = (X - T) / dx;
 B = (T <= X) & (X < (T + dx));
 r = [2:length(t) 1];
 for k = 1:bdeg
  B = (P .* B + (k + 1 - P) .* B(:, r)) / k;
 end;
end;
```

The computation of D_k is a little simpler, because there is the built-in function `diff()` that accepts a parameter for the order of the difference. Consequently, in MATLAB the computations look like the following:

```
B = bspline(x, xl, xr, ndx, bdeg);
[m n] = size(B);
D = diff(eye(n), pord);
a = (B' * B + lambda * D' * D) \ (B' * y);
yhat = B * a;
Q = inv(B' * B + lambda * D' * D);
s = sum((y - yhat) .^ 2)
t = sum(diag(Q * (B' * B)));
gcv = s / (m - t)^2;
```

The formulas for the penalized likelihood equations describe how to incorporate the penalty when one has access to all the individual steps of the regression computations. If this is not the case, data augmentation can help. Instead of working with the matrices B of B-splines regressors and D_k of the penalty separately, and combining their inner products, augmented data can be constructed as follows:

$$\begin{bmatrix} y \\ 0 \end{bmatrix} \approx \begin{bmatrix} B \\ \sqrt{\lambda} D_k \end{bmatrix}, \tag{46}$$

where $\approx$ indicates regression of the left-hand vector on the right-hand matrix. For linear problems, it is enough to do this only one time. In generalized linear models, data augmentation has to be done anew in each of the iterations with weighted linear regressions.

We tested the above program fragments on a PC with 75-MHz Pentium processor, with S-PLUS 3.3 and MATLAB 4.2, both operating under Windows for Workgroups. The data were those from the motorcycle helmet experiment, as presented in Figure 2. There are 133 data points and we used 20 intervals on the x-domain. S-PLUS took about 0.9 second, Matlab about 0.2 second (for one value of λ). These times can be reduced to 0.6 second and 0.1 second, respectively, by storing and reusing some intermediate results ($B^T B$ and the inverse of $B^T B + \lambda D_k^T D_k$).

Functions for generalized linear estimation can be obtained from the first author. We are preparing a submission to Statlib.

ACKNOWLEDGMENTS

Our initial research, as presented in Eilers and Marx (1992), did not point out that O'Sullivan's work (O'Sullivan, 1986, 1988) implicitly used a modified second-order difference penalty. We are grateful to Professor Wahba for drawing attention to this connection and our oversight. We also thank the anonymous referee for many suggestions to improve our presentation.

REFERENCES

ASHFORD, R. and WALKER, P. J. (1972). Quantal response analysis for a mixture of populations. *Biometrics* **28** 981–988.

BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* MIT Press.

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

COX, M. G. (1981). Practical spline approximation. In *Topics in Numerical Analysis* (P. R. Turner, ed.). Springer, Berlin.

DE BOOR, C. (1977). Package for calculating with *B*-splines. *SIAM J. Numer. Anal.* **14** 441–472.

DE BOOR, C. (1978). *A Practical Guide to Splines.* Springer, Berlin.

DIERCKX, P. (1993). *Curve and Surface Fitting with Splines.* Clarendon, Oxford.

DIGGLE P. and MARRON J. S. (1988). Equivalence of smoothing parameter selectors in density and intensity estimation. *J. Amer. Statist. Assoc.* **83** 793–800.

EILERS, P. H. C. (1990). Smoothing and interpolation with generalized linear models. *Quaderni di Statistica e Matematica Applicata alle Scienze Economico-Sociali* **12** 21–32.

EILERS, P. H. C. (1991a). Penalized regression in action: estimating pollution roses from daily averages. *Environmetrics* **2** 25–48.

EILERS, P. H. C. (1991b). Nonparametric density estimation with grouped observations. *Statist. Neerlandica* **45** 255–270.

EILERS, P. H. C. (1995). Indirect observations, composite link models and penalized likelihood. In *Statistical Modelling* (G. U. H. Seeber et al., eds.). Springer, New York.

EILERS, P. H. C. and MARX, B. D. (1992). Generalized linear models with *P*-splines. In *Advances in GLIM and Statistical Modelling* (L. Fahrmeir et al., eds.). Springer, New York.

EUBANK, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* Dekker, New York.

FRIEDMAN, J. and SILVERMAN, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics* **31** 3–39.

GREEN, P. J. and SILVERMAN, B. W. (1994). *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, London.

GREEN, P. J. and YANDELL, B. S. (1985). Semi-parametric generalized linear models. In *Generalized Linear Models* (B. Gilchrist et al., eds.). Springer, New York.

HAND, D. J., DALY, F., LUNN, A. D., MCCONWAY, K. J. and OSTROWSKI, E. (1994). *A Handbook of Small Data Sets.* Chapman and Hall, London.

HÄRDLE, W. (1990). *Applied Nonparametric Regression.* Cambridge Univ. Press.

HASTIE, T. and TIBSHIRANI, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

KOOPERBERG, C. and STONE, C. J. (1991). A study of logspline density estimation. *Comput. Statist. Data Anal.* **12** 327–347.

KOOPERBERG, C. and STONE, C. J. (1992). Logspline density estimation for censored data. *J. Comput. Graph. Statist.* **1** 301–328.

MARRON, J. S. and RUPPERT, D. (1994). Transformations to reduce boundary bias in kernel density estimation. *J. Roy. Statist. Soc. Ser. B* **56** 653–671.

MARX, B. D. and EILERS, P. H. C. (1994). Direct generalized additive modelling with penalized likelihood. Paper presented at the 9th Workshop on Statistical Modelling, Exeter, 1994.

MARX, B. D. and EILERS, P. H. C. (1996). Direct generalized additive modelling with penalized likelihood. Unpublished manuscript.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

O'SULLIVAN, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statist. Sci.* **1** 505–527.

O'SULLIVAN, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM J. Sci. Statist. Comput.* **9** 363–379.

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

SAKAMOTO, Y., ISHIGURO, M. and KITAGAWA, G. (1986). *Akaike Information Criterion Statistics.* Reidel, Dordrecht.

SCOTT, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley, New York.

SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *J. Roy. Statist. Soc. Ser. B* **47** 1–52.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

WAHBA, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.

WAND, M. P. and JONES, M. C. (1993). *Kernel Smoothing.* Chapman and Hall, London.

WHITTAKER, E. T. (1923). On a new method of graduation. *Proc. Edinburgh Math. Soc.* **41** 63–75.

Comment

S-T. Chiu

Authors Paul Eilers and Brian Marx provide a very interesting approach to nonparametric curve fitting. They give a brief but very concise review of *B*-splines. I also enjoyed reading the part where the authors applied their procedure to some examples. As shown in the paper, the approach has several merits which deserve to be studied in more detail.

Similar to any nonparametric smoother, the proposed procedure needs a smoothing parameter λ to control the smoothness of the fitting curve. My com-

S-T. Chiu is with the Department of Statistics, Colorado State University, Fort Collins, Colorado 80523-0001.

ments mainly concern the selection of the smoothing parameter.

It is well known that the classical selectors such as AIC, GCV, Mallows's C_p and so on do not give a satisfactory result. For the regression case, more details about the defects can be found in Rice (1984) and Chiu (1991a). Scott and Terrell (1987) and Chiu (1991b) discuss the case of density estimation. The classical selectors have a large sample variation and a tendency to select a small smoothing parameter, thus producing a very rough curve estimate. It is natural to expect that they have a similar problem when applied to selecting the smoothing parameter for P-splines.

Several procedures have been suggested to remedy the defects of the classical procedures. Chiu (1996) provides a survey of some of these newer selectors for density estimation. For the regression case, some procedures are suggested in Chiu (1991a), Hall and Johnstone (1992) and Hall, Marron and Park (1992).

In the following, I provide a brief review to explain the defects and some remedy to the classical selectors for kernel regression estimate. Let us assume the simplest model of a circular design with equally spaced design points. $y_t = \mu(x_t) + \varepsilon_t$, where ε_t are i.i.d. noise. For the kernel estimate $\hat{\mu}_\beta$ with a bandwidth β, we often use the mean of sum of squared errors

$$R(\beta) = E\Big[\sum\{\hat{\mu}_\beta(x_t) - \mu(x_j)\}^2\Big] \tag{1}$$

to measure the closeness between $\hat{\mu}(x)$ and $\mu(x)$.

The goal of bandwidth selection is to select the optimal bandwidth which minimizes $R(\beta)$. Since in practice μ is unknown, we have to estimate $R(\beta)$ and use the minimizer of the estimated $R(\lambda)$ as an estimate of the optimal bandwidth. For example, Mallows's C_p has the form

$$\hat{R}(\beta) = \mathrm{RSS}(\beta) - T\sigma^2 + 2\sigma^2 w(0)/\beta. \tag{2}$$

Here $w(x)$ is the kernel and σ^2 is the error variance. Other classical procedures such as AIC and GCV have a similar form and were shown to be asymptotically equivalent in Rice (1984). All of these procedures rely on the residual sum of squares $\mathrm{RSS}(\beta)$.

Mallows (1973) proposed the procedure based on the observation that

$$R(\beta) = E\{\mathrm{RSS}(\beta)\} - T\sigma^2 + 2\sigma^2 w(0)/\beta.$$

As we will explain later, the main problem here is that $\mathrm{RSS}(\beta)$ is not a good estimate of its expected value.

By using the Fourier transform, (1) and (2) could be written, respectively, as

$$\begin{aligned} R(\beta) = 4\pi \sum_{j=1}^{N} I_S(\lambda_j)\{1 - W_\beta(\lambda_j)\}^2 \\ + \sigma^2 \sum_{j=1}^{N} W_\beta(\lambda)^2 + \sigma^2 \end{aligned} \tag{3}$$

and

$$\begin{aligned} \hat{R}(\beta) = 4\pi \sum_{j=1}^{N} \left\{ I_Y(\lambda_j) - \frac{\sigma^2}{2\pi} \right\} \{1 - W_\beta(\lambda_j)\}^2 \\ + \sigma^2 \sum_{j=1}^{N} W_\beta(\lambda)^2 + \sigma^2, \end{aligned} \tag{4}$$

where I_Y and I_S are the periodograms of Y_t and the signal $S_t = \mu(t/T)$, respectively, and $\lambda_j = 2\pi j/T$, $j = 1, \dots, N = [T/2]$. Also, $W_\beta(\lambda)$ is the transfer function of $w\{t/(\beta T)\}/(\beta T)$.

Comparing (3) and (4), we see that $\hat{R}$ attempts to use $I_Y(\lambda) - \sigma^2/(2\pi)$ to estimate $I_S(\lambda)$. The difficulty is that at high frequency, I_Y is dominated by the noise and thus does not give a good estimate of I_S.

Chiu (1991a) suggested truncating the high-frequency portion when we estimate $R(\beta)$,

$$\begin{aligned} \tilde{R}(\beta) = 4\pi \sum_{j=1}^{J} \left\{ I_Y(\lambda_j) - \frac{\sigma^2}{2\pi} \right\} \{1 - W_\beta(\lambda_j)\}^2 \\ + \sigma^2 \sum_{j=1}^{N} W_\beta(\lambda)^2 + \sigma^2. \end{aligned} \tag{5}$$

Here J is selected in such a way that there is no significant I_S beyond frequency λ_J. The selector $\tilde{R}(\beta)$ has a much better performance than the classical ones. Hall, Marron and Park (1992) proposed another procedure which downweights the contribution from the high-frequency part.

It is clear that the bases of the kernel regression are the sinusoid waves. The primary reason of success of criterion (5) is that most information about μ concentrates at low frequency. In other words, we just need quite a few bases to approximate the true curve well.

However, since each basis of the B-spline is very local to a certain interval, we cannot use just a few bases to approximate the curve over the whole region. In my opinion, this could be a big obstacle to the understanding and improvement of the classical smoothing parameter selectors.

REFERENCES

CHIU, S.-T. (1991a). Some stabilized bandwidth selectors for nonparametric regression. *Ann. Statist.* **19** 1528–1546.

CHIU, S.-T. (1991b). Bandwidth selection for kernel density estimation. *Ann. Statist.* **19** 1883–1905.

CHIU, S.-T. (1996). A comparative review of bandwidth selection for kernel density estimation. *Statist. Sinica* **6** 129–145.

HALL, P. and JOHNSTONE, I. (1992). Empirical functionals and efficient smoothing parameter selection. *J. Roy. Statist. Soc. Ser. B* **54** 519–521.

HALL, P., MARRON, J. S. and PARK, B. U. (1992). Smoothed cross-validation. *Probab. Theory Related Fields* **92** 1–20.

MALLOWS, C. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

RICE, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12** 1215–1230.

SCOTT, D. W. and TERRELL, G. R. (1987). Biased and unbiased cross-validation in density estimation. *J. Amer. Statist. Assoc.* **82** 1131–1146.

Comment

Douglas Nychka and David Cummins

One strength of the authors's presentation is the simple ridge regression formulas that result for the estimator. We would like to point out a decomposition using a different set of basis functions that helps to interpret this smoother. This alternative basis, derived from B-splines, facilitates the computation of the GCV function and confidence bands for the estimated curve.

To simplify this discussion assume that $W = I$ so that the hat matrix is

$$H = B(B^T B + \lambda D^T D)^{-1} B^T = G(I + \lambda \Gamma)^{-1} G^T,$$

$G = BQ_2^{-1/2}U$, $Q_2 = B^T B\ U$, $\Gamma = \mathrm{diag}(\gamma)$ and U is an orthogonal matrix such that $Q_2^{-1/2} D^T D Q_2^{-1/2} = U \Gamma U^T$. The columns of G can be identified with a new set of functions known as the Demmler–Reinsch (DR) basis. Specifically these are piecewise polynomial functions, $\{\psi_\nu\}$ so that the elements of G satisfy $\psi_\nu(x_i) = G_{i\nu}$. Besides having useful orthogonality properties the DR basis can be ordered by frequency and larger values of γ_ν will exhibit more oscillations (in fact $\nu - 1$ zero crossings). Figure 1(a) plots several of the basis functions for $m = 133$ equally spaced x's and 20 equally spaced interior knots. Figure 1(b) illustrates the expected polynomial increase in the size of γ_ν as a function of ν.

The Demmler–Reinsch basis provides an informative interpretation of the spline estimate. Let $\hat{f}$ denote the P-spline and let $\alpha = G^T y$ denote the least squares coefficients from regressing y on the DR basis functions:

$$\begin{aligned}\hat{f}(x_i) &= [Hy]_i = [G(I + \lambda\Gamma)^{-1} G^T y]_i \\ &= \sum_{\nu=1}^{m} \psi_\nu(x_i) \frac{\alpha_\nu}{1 + \lambda \gamma_\nu}.\end{aligned}$$

Note that the smoother is just a linear combination of the DR basis functions using coefficients that are downweighted (or tapered) by the factor $1/(1 + \lambda\gamma_\nu)$ from the least squares estimates. Because of the relationship between γ_ν and ψ_ν (see Figure 1), the basis functions that represent higher-frequency structure will have coefficients that are more severely downweighted. In this way the smoother is a low-pass filter, tending to preserve low-frequency structure and downweighting higher-frequency terms. The residual sum of squares and the trace of H can be computed rapidly (order n) using the DR representation. Thus the GCV function can also be evaluated in order n operations for a given value of λ.

Another application of the DR form is in computing a confidence band. Consider a set of candidate functions that contain the true function with the correct level of confidence. The confidence band is then the envelope implied by considering all functions in this set. For example, let $\hat{f}$ denote the function estimate and for $C_1, C_2 > 0$ let

$$\mathscr{B} = \left\{ h\text{: } h \text{ is a } B\text{-spline with coefficients } b, \right. \\ \left. \sum_{i=1}^{n} (\hat{f}(x_i) - h(x_i))^2 \le C_1 \text{ and } b^T D^T D\, b \le C_2 \right\}$$

Douglas Nychka is Professor of Statistics and David Cummins is with the Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203.

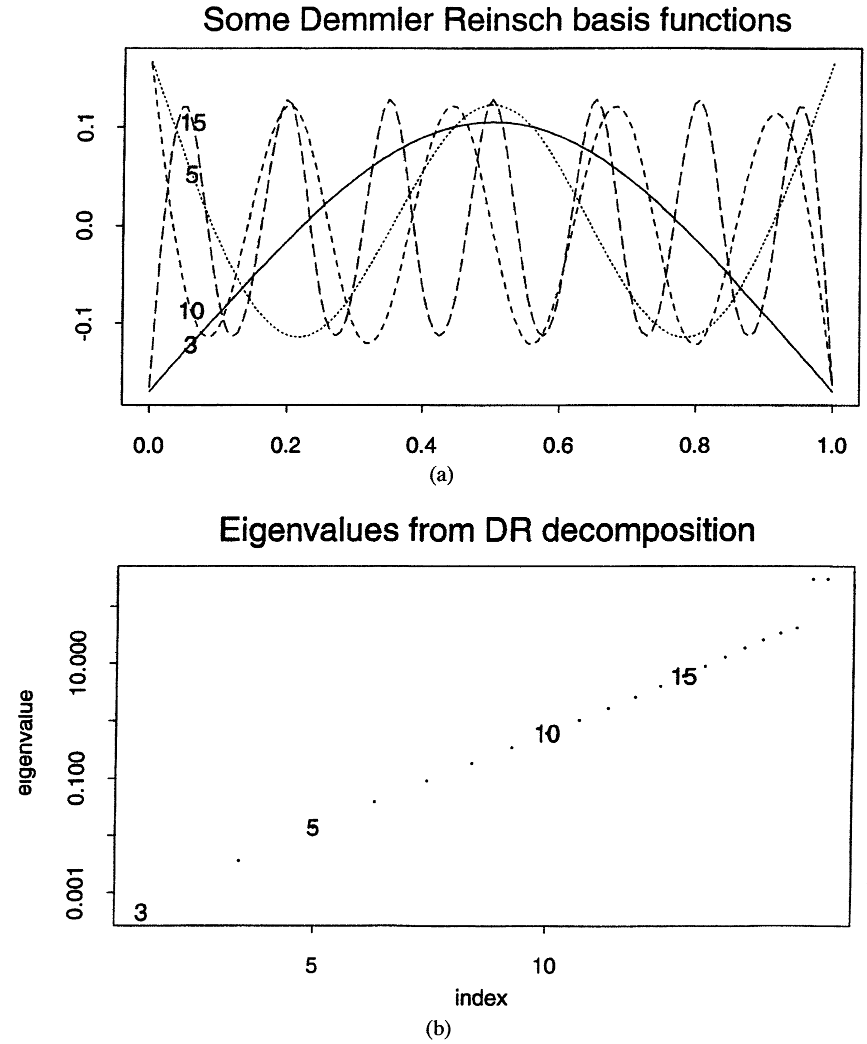

FIG. 1. *Illustration of several Demmler–Reinsch basis functions and the associated eigenvalues for* 20 *equally spaced knots,* 133 *equally spaced observations and second divided differences* ($k = 2$): *the upper plot* (a) *is* $\{\psi_\nu\}$ *for* $\nu = (3, 5, 10, 15)$; *the numerals identify the order of these basis functions and in the second plot* (b) *identify the eigenvalues for these functions.*

The constants C_1 and C_2 are determined so that $P\{f \in \mathscr{B}\}$ equals the desired confidence level. The upper and lower boundaries of the confidence band are then

$$U(x) = \max\{h(x)\colon h \in \mathscr{B}\}$$

and

$$L(x) = \min\{h(x)\colon h \in \mathscr{B}\}$$

In practice we work with the coefficients and thus the computation of U and L at each x is a minimization problem with two quadratic constraints. Using the DR basis reduces both constraints to quadratic forms with diagonal matrices and thus both are computable in order n operations. Moreover this strategy does not depend on the roughness penalty being divided differences but will work for any nonnegative matrix used as a penalty (e.g., thin plate splines). Currently we are investigating the choice of C_1 and C_2 based on the GCV estimate of f.

ACKNOWLEDGMENT

This work was supported by NSF Grant DMS-92-17866.

Comment

Chong Gu

I would like to begin by congratulating the authors Eilers and Marx for a clear exposition of an interesting variant of penalized regression splines. My comments center around three questions: Are P-splines really better? What does optimal smoothing stand for? And what does the future hold for nonparametric function estimation?

ARE P-SPLINES REALLY BETTER?

P-splines can certainly be as useful as other variants of penalized regression splines, but I am not sure that they are really advantageous over the others. It is true that with huge sample sizes, one may choose n much smaller than m to save on computation without sacrificing performance, but other variants of regression splines also share the same advantage. The mechanical handling of the difference penalty is certainly very interesting computationally, but as far as the end users are concerned, I do not see why the discrete penalties are necessarily advantageous over the continuous ones. Higher-order derivative penalties are certainly as feasible as discrete penalties computationally, albeit more difficult to implement, but the difference is irrelevant to the end users whose main interest is the interface.

The users may be more interested in what the program computes rather than how it computes, however, and in this respect, I only see P-splines lose out to penalized regression splines with the usual derivative penalties that everyone can understand. Being told that B-splines provide a good basis for function approximation, the users may simply ignore whatever other properties B-splines have and still have a clear picture about what they are getting from derivative penalties or, for that matter, from Whittaker's discrete penalties which use the differences of adjacent function values. With the P-splines, however, the intuition is unfortunately taken away from the users, and even with a thorough knowledge of all the properties of B-splines, I am not sure one can easily perceive what the penalty is really doing, other than that it is reducing the effective dimension in some not so easily comprehensible way.

Penalized smoothers with quadratic penalties are known to be equivalent to Bayes estimates with Gaussian priors. When $Q = D_k^T D_k$ is of full rank, the corresponding prior for the B-spline coefficients a has mean 0 and covariance proportional to Q^{-1}. When Q is rank-deficient, the prior has a "fixed effect" component diffuse in the null space of Q and a "random effect" component with mean 0 and covariance proportional to Q^{+}, the Moore–Penrose inverse of Q. From this perspective, P-splines differ from other variants of penalized regression splines only in the specification of Q.

WHAT DOES OPTIMAL SMOOTHING STAND FOR?

One probably can never overstate the importance of smoothing parameter selection for any successful practical application of any smoothing method. AIC and cross-validation are among the most accepted (and successful) working criteria for model selection, yet their optimalities are established, theoretically or empirically, only for specific problem settings under appropriate conditions. Naive adaptations of these criteria in new problem settings do *not* necessarily deliver fits that are nearly optimal.

Specifically, I am somewhat worried about the "optimality" of the naive adaptations of these criteria proclaimed in Section 6. First, it is not clear in what sense these criteria are "optimal" in the problem settings to which they are applied; second, there is no empirical (or theoretical) evidence illustrating the presumed "optimality." AIC or cross-validation *may* deliver nearly optimal fits, but they surely do not by themselves define the notion of optimality.

My worries stem from previous empirical experiments with smoothing parameter selection by myself and by others, especially in non-Gaussian regression problems (commonly referred to as generalized linear models). Using Kullback–Leibler discrepancy or its symmetrized version to define optimality, it has been found that a naive adaptation of GCV in non-Gaussian regression, which appears similar to what the authors suggest in Section 7, may return anything but nearly optimal fits. See, for example, Cox and Chang (1990), Gu (1992)

Chong Gu is Assistant Professor, Department of Statistics, Purdue University, West Lafayette, Indiana 47907.

and Xiang and Wahba (1996). For the density estimation problem in Section 8, I could not find the definition of the H matrix to understand the AIC proposed, but whatever it is, it should be subject to the same scrutiny before being recommended as "optimal."

In ordinary Gaussian regression, the optimality of GCV is well established in the literature. For the AIC score presented in (27), however, I would like some empirical evidence to be convinced of its optimality. The skepticism is partly due to some empirical evidence suggesting that *the trace of H may not be a consistent characterization of the effective dimension of the model*. Such evidence can be found in Gu (1996), available online at `http://www.stat.lsa.umich.edu/~chong/ps/modl.ps`.

WHAT DOES THE FUTURE HOLD FOR FUNCTION ESTIMATION?

In response to *Statistical Science*'s desideration for speculations regarding future research directions, I would like to take this opportunity to offer some of my thoughts.

It has long been said that all smoothing methods perform similarly in one dimension, provided that the smoothing parameter selection is done properly, yet time and again new and not so new methods keep being invented. The real challenge, however, seems to lie in *multivariate problems*. Amid the curse of dimensionality and potential structures associated with multivariate problems, the choice of methods can make a real difference in multidimension, in the ease of computation and smoothing parameter selection, in the convenience of incorporation of structures, and so on. Among methods with the most potential are the adaptive regression splines developed by Friedman, Stone and coworkers, and the smoothing splines developed by the Wisconsin spline school lead by Wahba. The penalized regression spline approach, however, seems somewhat handicapped by the lack of effective basis, say in dimensions beyond two or three.

More challenging still, an important line of research that has been largely neglected is *inference*. What one usually gets from the function estimation literature are point estimates possibly with asymptotic convergence rates, and intuitive smoothing parameter selectors not always accompanied by justifications. Besides a few entries based on the Bayes model of smoothing splines by Wahba (1983), Cox, Koh, Wahba and Yandell (1988), Barry (1993) and some follow-ups, practical procedures that offer *interval estimates*, *test of hypothesis*, and so on, are largely missing in the literature. To guard against the danger of overinterpreting data by the use of nonparametric methods, such inferential tools should be a top priority in future research. Under a Bayes model where the target function is treated as a realization of a stochastic process, the development may proceed within the conventional inferential framework. Under the traditional setting where the target function is considered fixed, however, one may have to turn his back on the conventional Neyman–Pearson thinking before he can call any useful inferential tools non-ad-hoc.

REFERENCES

BARRY, D. (1993). Testing for additivity of a regression function. *Ann. Statist.* **21** 235–254.

COX, D. D. and CHANG, Y.-F. (1990). Iterated state space algorithms and cross validation for generalized smoothing splines. Technical Report 49, Dept. Statistics, Univ. Illinois.

COX, D. D., KOH, E., WAHBA, G. and YANDELL, B. S. (1988). Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. *Ann. Statist.* **16** 113–119.

GU, C. (1992). Cross validating non Gaussian data. *Journal of Computational and Graphical Statistics* **1** 169–179.

GU, C. (1996). Model indexing and smoothing parameter selection in nonparametric function estimation. Technical Report 93-55 (rev.), Dept. Statistics, Purdue Univ.

WAHBA, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B* **45** 133–150.

XIANG, D. and WAHBA, G. (1996). A generalized approximate cross validation for smoothing splines with non-Gaussian date. *Statist. Sinica*. To appear.

Comment

M. C. Jones

Eilers and Marx present a clear and interesting account of their P-spline smoothing methodology. Clearly, P-splines constitute another respectable approach to smoothing. However, their good properties appear to be, broadly, on a par with those of various other approaches; the method is no nearer to, or further from, "being the ideal smoother" than others.

"P-splines have no boundary effects, they are a straightforward extension of (generalized) linear regression models, conserve moments (means, variances) of the data, and have polynomial curve fits as limits." Except for the third point, the same claims can be made of spline smoothing (Green and Silverman, 1994) or local polynomial fitting (Fan and Gijbels, 1996).

Conservation of moments seems unimportant. In regression, I do not see the desirability. In density estimation, simple corrections of kernel density estimates for variance inflation exist, but make little difference away from the normal density (Jones, 1991). Indeed, getting means and variances right is a normality-based concept, so corrected kernel estimators act in a normal-driven semiparametric manner. Efron and Tibshirani (1996) propose more sophisticated moment conservation, but initial indications are that this is no better nor worse than alternative semiparametric density estimators (Hjort, 1996).

"The computations, including those for cross-validation, are relatively inexpensive and easily incorporated into standard software." Again, proponents of the two competing methods I have mentioned would claim the same for the first half of this and advocates of regression splines would claim the lot.

The authors make no particularly novel contribution to automatic bandwidth selection. Cross-validation and AIC are in a class of methods (e.g., Härdle, 1990, pages 166–167) which, while not being downright bad, allow scope for improvement. Calculating thesebandwidth selectors quickly is less important than developing better selectors. For local polynomials, improvements are offered (for normal errors) by Fan and Gijbels (1995) and Ruppert, Sheather and Wand (1995) and unpublished work extends these to more general situations.

The comparison of (5) with (11) focusses on the small extra complexity of the latter. But which is more interpretable: a roughness penalty on a curve or on a series of coefficients? Changing the penalty in a smoothing spline setup allows different parametric limits (e.g., Ansley, Kohn and Wong, 1993); how can P-splines cope with this?

An exasperating aspect of spline-based approaches is the lack of straightforward (asymptotic) mean squared error–type results to indicate theoretical performance relative to kernel/local polynomial approaches for which such results are simply obtained and, within limitations, informative. I doubt whether P-splines can facilitate such developments (reason given below).

It seems that P-splines have no particular attractiveness for multivariate applications. The examples are noteworthy only for looking like results obtainable by other methods too.

The idea behind density estimation P-splines is to treat a fine binning as Poisson regression data. OK, but again equally applicable to other approaches and already investigated for local polynomial smoothing. Simonoff (1996, Section 6.4) and Jones (1996) explain how such regression approaches to density estimation are discretized versions of certain "direct" local likelihood density estimation methods (Hjort and Jones, 1996; Loader, 1996). Binning is the major computational device of all kernel-type estimators (Fan and Marron, 1994). The local likelihood approach is already deeply understood theoretically.

Comparison of P-splines's reasonable boundary performance with local polynomials's reasonable boundary performance is not yet available through theory or simulations.

An interesting point mentioned in the paper is the apparent continuum between few-parameter parametric fits at one end and fully "nonparametric" techniques at the other, with many-parameter para-

M.C. Jones is Reader in Statistical Science, Department of Statistics, The Open University, Walton Hall, Milton Keynes, MK7 6AA, United Kingdom.

metric models and semiparametric approaches in between: a dichotomy into parametric and nonparametric is inappropriate, and there is a huge grey area of overlap. The equivalent degrees-of-freedom ideas of Hastie and Tibshirani (1990) provide a fine (but possibly improveable?) attempt to give this continuum a scale. Theoretical development might be made more difficult by P-splines for reasons associated with quantifying the "nonparametricness" of intermediate methods.

Finally, we come back to my main point. In an admirable "personal view of smoothing and statistics," Marron (1996) gives a list of smoothing methods and another of factors (to which I might add others) involved in the choice between methods. Marron says "All of the methods ... listed ... have differing strengths and weaknesses in ... divergent senses. None of these methods dominates any other in all of the senses. ... Since these factors are so different, almost any method can be 'best', simply by an appropriate personal weighting of the various factors involved." P-splines are a reasonable addition to Marron's first list, but have no special status with respect to his second.

REFERENCES

ANSLEY, C. F., KOHN, R. and WONG, C. M. (1993). Nonparametric spline regression with prior information. *Biometrika* **80** 75–88.

EFRON, B. and TIBSHIRANI, R. (1996). Using specially designed exponential families for density estimation. *Ann. Statist.* **24** 000–000.

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.

FAN, J. and MARRON, J. S. (1994). Fast implementations of nonparametric curve estimators. *J. Comput. Graph. Statist.* **3** 35–56.

HJORT, N. L. (1996). Performance of Efron and Tibshirani's semiparametric denisty estimator. Unpublished manuscript.

HJORT, N. L. and JONES, M. C. (1996). Locally parametric nonparametric density estimation. *Ann. Statist.* **24** 1619–1647.

JONES, M. C. (1991). On correcting for variance inflation in kernel density estimation. *Comput. Statist. Data Anal.* **11** 3–15.

JONES, M. C. (1996). On close relations of local likelihood density estimation. Unpublished manuscript.

LOADER, C. R. (1996). Local likelihood density estimation. *Ann. Statist.* **24** 1602–1618

MARRON, J. S. (1996). A personal view of smoothing and statistics (with discussion). *Comput. Statist.* To appear.

RUPPERT, D., SHEATHER, S. J. and WAND, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **90** 1257–1270.

SIMONOFF, J. S. (1996). *Smoothing Methods in Statistics*. Springer, New York.

Comment

Joachim Engel and Alois Kneip

Paul Eilers and Brian Marx have provided us with a nice and flexible addition to the smoother's toolkit. Their proposed P-spline estimator can be considered as some compromise between the usual B-spline estimation and the smoothing spline approach. Different from many papers on B-splines, however, they do not consider the delicate problem of optimal knot selection. Instead, they propose to use a large number of equidistant knots. Smoothing is introduced by a roughness penalty on the difference of spline coefficients.

P-spline estimation is equivalent to smoothing splines when choosing as many knots as there are observations ($n = m$) with a knot placed at each data point. However, this is not the situation the authors have in mind. They propose to choose a large number n of knots, but $n < m$. Such an approach is of considerable interest. We know from personal experience that nonparametric regression fits based on B-splines are often visually more appealing than, for example, kernel estimates. The same seems to be true for P-splines if a moderate number of knots is used. Furthermore, as the authors indicate, P-splines together with the difference penalty enjoy many important practical advantages and are flexible enough to be applied in different modelling situations, for example, in additive models or self-modelling regression where the backfitting algorithm is used.

Nevertheless, we do not yet see much evidence for the authors's claim that P-splines "come near being the ideal smoother." For example, local polynomial regression is known to exhibit no boundary

Joachim Engel is with Wirtschaftstheorie II, Universität Bonn, and Department of Mathematics, PH Ludwigsburg, Germany. Alois Kneip is with Institut de Statistique, Université Catholique de Louvain, Belgium.

problems (in first order) and to possess certain optimality and minimax properties (Fan, 1993). For density estimation.Engel and Gasser (1995) show a minimax property of the fixed bandwith kernel method within a large class of estimators containing penalized likelihood estimators. The presented paper does not provide any argument, neither theoretical nor by simulations, supporting any superiority of P-splines over their many competitors.

In the regression case, the theoretical properties of P-splines might be evaluated by combining arguments of de Boor (1978) on the asymptotic bias and variance of B-splines in (dependence on m, the spline order k and the smoothness of the underlying function) with the well-known results on smoothing splines.

The authors propose to use AIC or cross-validation to select the smoothing parameter λ. However, a careful look at their method reveals that there are in fact two free parameters: λ and the number n of knots. If $n \approx m$, then we essentially obtain a smoothing spline fit, while results might be very different if $n \ll m$. Indeed, the estimate might crucially depend on n. Therefore, why not determine λ *and* n by cross-validation or a related method? The following theoretical arguments may suggest that such a procedure will work. Note that AIC and cross-validation are very close to unbiased risk estimation which consists of estimating the optimal values of λ and n by minimizing

$$\sum_{i=1}^{m}(y_i - \hat{\mu}_i)^2 + 2\sigma^2 \operatorname{tr}(H_{\lambda,n}),$$

where $H \equiv H_{\lambda,n}$ is the corresponding smoother matrix. Let $\text{ASE}(\lambda, n)$ denote the average squared error of the fit obtained by using some parameters λ and n. Under some technical conditions, it then follows from results of Kneip (1994) that, as $m \to \infty$,

$$\text{ASE}(\hat{\lambda}, \hat{m})/\text{ASE}(\lambda_{\text{opt}}, m_{\text{opt}}) \to_P 1.$$

Here $\hat{\lambda}$ and $\hat{m}$ are the parameters estimated by unbiased risk estimation, while λ_{opt} and m_{opt} represent the optimal choice of the parameters minimizing ASE.

Comment

Charles Kooperberg

Eilers and Marx present an interesting approach to spline modeling. While function estimation based on smoothing splines often yields reasonable results, the computational burden can be very large. If the number of basis functions is limited, however, the computations become much easier, and when the knots are equally spaced, the solution indeed becomes rather elegant. To increase the credibility of the claim that P-splines are close to the "ideal smoother," several issues need to be addressed:

1. In density estimation, when the range of the data is $\mathbb{R}$ ($\mathbb{R}^+$), it is useful that a density estimate be positive on $\mathbb{R}$ ($\mathbb{R}^+$), for example, for resampling. Some methods can estimate densities on bounded or unbounded intervals. P-splines do not seem to have this property: lower and upper bounds have to be specified and there seems to be no natural way to extrapolate beyond these bounds. Is there any way around that? Can infinity be a bound?

   How would one specify the bounds? From the suicide example it appears that this may influence the results considerably.
2. To use P-splines, additional choices need to be made. How many knots should one use? Is the procedure insensitive to the number of knots provided that there are enough of them? If so, how many is enough? How does the computational burden depend on the number of knots?

   What order of penalty should be used? Do you advocate examining several possible penalties, as in the logistic regression example, or do you have another recommendation, such as using $k = 3$ for density estimation so that the limit of your estimate as $\lambda \to \infty$ is a normal density? Since many smoothing and density estimation procedures are used as EDA tools, good defaults are very worthwhile.
3. It would be interesting to see an application of the P-spline methodology to more challenging data, such as the income data described below,

Charles Kooperberg is Assistant Professor, Department of Statistics, University of Washington, Seattle, Washington 98195-0001.

which involves thousands of cases, a narrow peak and a severe outlier.

How would the P-spline algorithm, where knots are positioned equidistantly, behave when there are severe outliers, which would dominate the positioning of the knots? Is it possible to position knots nonequidistantly, for example, based on order statistics?

4. Are there theoretical results about the large sample behavior of P-splines?

POLYNOMIAL SPLINES AND LOGSPLINE DENSITY ESTIMATION

Besides the penalized likelihood approach, there is an entirely different approach to function estimation based on splines. Whereas for P-splines both the number and the locations of the knots are fixed in advance and the smoothness is governed by a smoothing parameter, in the polynomial spline framework the number and location of the knots are determined adaptively using a stepwise algorithm and no smoothing parameter is needed. Such polynomial spline methods have been used for regression (Friedman, 1991), density estimation (Kooperberg and Stone, 1992), polychotomous (multiple logistic) regression (Kooperberg, Bose and Stone, 1997), survival analysis (Kooperberg, Stone and Truong, 1995a) and spectral density estimation (Kooperberg, Stone and Truong, 1995b).

In univariate polynomial spline methodologies the algorithm starts with a fairly small number of knots. It then adds knots in those regions where an added knot would have the most influence, using Rao (score) statistics to decide on the best location; after a prespecified maximum number of knots is reached, knots are deleted one at a time, using Wald statistics to decide which knot to remove. Out of the sequence of fitted models, the one having the smallest value for the BIC criterion is selected.

Polynomial spline algorithms for multivariate function estimation are similar, except that at each addition step the algorithm adds either a knot in one variable or a tensor product of two or more univariate basis functions. We have successfully applied such methodologies to data sets as small as 50 for one-dimensional density estimation and as large as 112,000 for a 63-dimensional polychotomous regression problem with 46 classes. For nonadaptive polynomial spline methodologies theoretical results regarding the L_2-rate of convergence are established. Stone, Hansen, Kooperberg and Truong (1996) provide an overview of polynomial splines and their applications.

Logspline density estimation, in which a (univariate) log-density is modeled by a cubic spline, is discussed in Kooperberg and Stone (1992) and Stone et al. (1996). Software for the 1992 version, written in C and interfaced to S-PLUS, is publically available from Statlib. (The 1992 version of LOGSPLINE employs only knot deletion; here, however, we focus on the 1996 version, which uses both knot addition and knot deletion.) LOGSPLINE can provide estimates on both finite and infinite intervals, and it can handle censored data.

The results of LOGSPLINE on the Old Faithful data and the suicide data are very similar to the corresponding results of P-splines [the suicide data is an example in Kooperberg and Stone (1992)]. Here we consider a much more challenging data set. The solid line in Figure 1 shows the logspline density estimate based on a random sample of 7,125 annual net incomes in the United Kingdom [Family Expenditure Survey (1968–1983)]. (The data have been rescaled to have mean 1.) The nine knots that were selected by LOGSPLINE are indicated. Note that four of these knots are extremely close to the peak near 0.24. This peak is due to the UK old age pension, which caused many people to have nearly identical incomes. In Kooperberg and Stone (1992) we concluded that the height and location of this peak are accurately estimated by LOGSPLINE. There are several reasons why this data is more challenging than the Old Faithful and suicide data: the data set is much larger, so that it is more of a challenge to computing resources (the LOGSPLINE estimate took 9 seconds on a Sparc 10 workstation); the width of the peak is about 0.02, compared to the range 11.5 of the data; there is a severe outlier (the largest observation is 11.5, the second largest is 7.8); and the rise of the density to the left of the peak is very steep.

To get an impression of what the P-splines procedure would yield for this data, I first removed the largest observation so that there would not be any long gaps in the data, reducing the maximum observation to 7.8. The dashed line in Figure 1 is the LOGSPLINE estimate to the data with fixed knots at $(i/20) \times 7.8$, for $i = 0, 1, \dots, 20$ (using 20 intervals, as in most P-spline examples.) The resulting fit should be similar to a P-spline fit with $\lambda = 0$. In this estimate it appears that the narrow peak is completely missed and that, because of the steep rise of the density to the left of the peak and the lack of sufficiently many knots near the peak, two modes are estimated where only one mode exists.

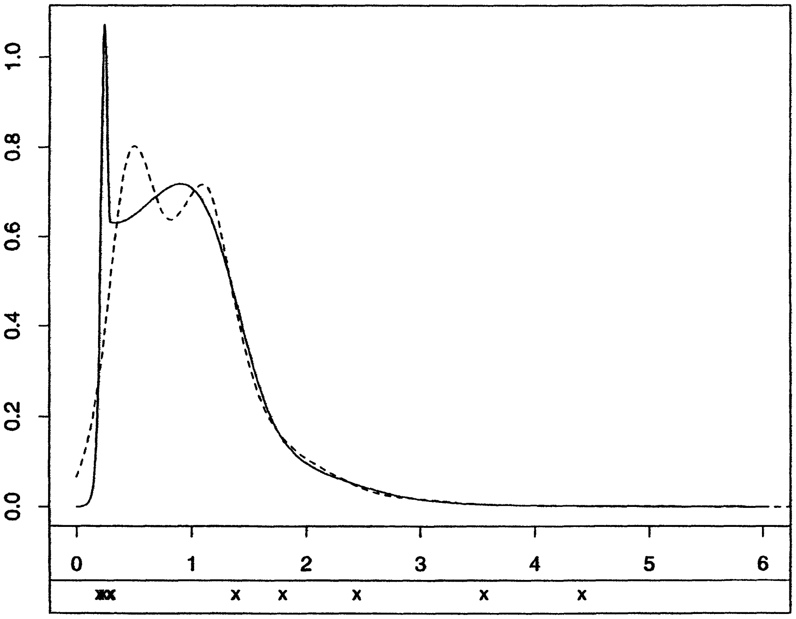

FIG. 1. *Logspline density estimate for the income data (solid line); the* x *indicate the locations of the knots; logspline approximation of the P-spline estimate with penalty parameter* 0 *(dashed line).*

It would be very much of interest to see how the P-spline methodology behaves on this data, and in particular whether it can accurately represent the sharp peak near 0.24.

ACKNOWLEDGMENT

Research supported in part by NSF Grant DMS-94-03371.

Comment

Dennis D. Cox

The main new idea in this paper is a roughness penalty based on the B-spline coefficients. There will be critics—I give some criticisms below—but there is considerable appeal in the simplicity of the idea. If I had to develop the software ab initio, it is clear that the roughness penalties proposed here would require less effort to implement than the standard ones based on L_2-norm of a second derivative.

There is a precedent for the use of the B-spline coefficients in such a direct way, from computer graphics (CG) and computer aided design (CAD). The "control point" typically used in parametric B-spline representations of curves and surfaces basically consists of the B-spline coefficients. See Foley and van Dam (1995, Section 11.2.3). This is demonstrated in Figure 1, where the control points for the solid curve are just random uniform added to a linear trend, and the same points are shrunk toward 0.5 before adding the trend to obtain the control points for the dashed curve. The ordinate of each control point is the cubic cardinal B-spline coefficient and the abscissa is the midpoint of support. In CG/CAD applications, the control points are manipulated to obtain a curve or surface with desirable shape or smoothness. The CG/CAD practitioners become familiar with these control points and develop

Dennis D. Cox is with Department of Statistics, Rice University, P.O. Box 1892, Houston, Texas 77251.

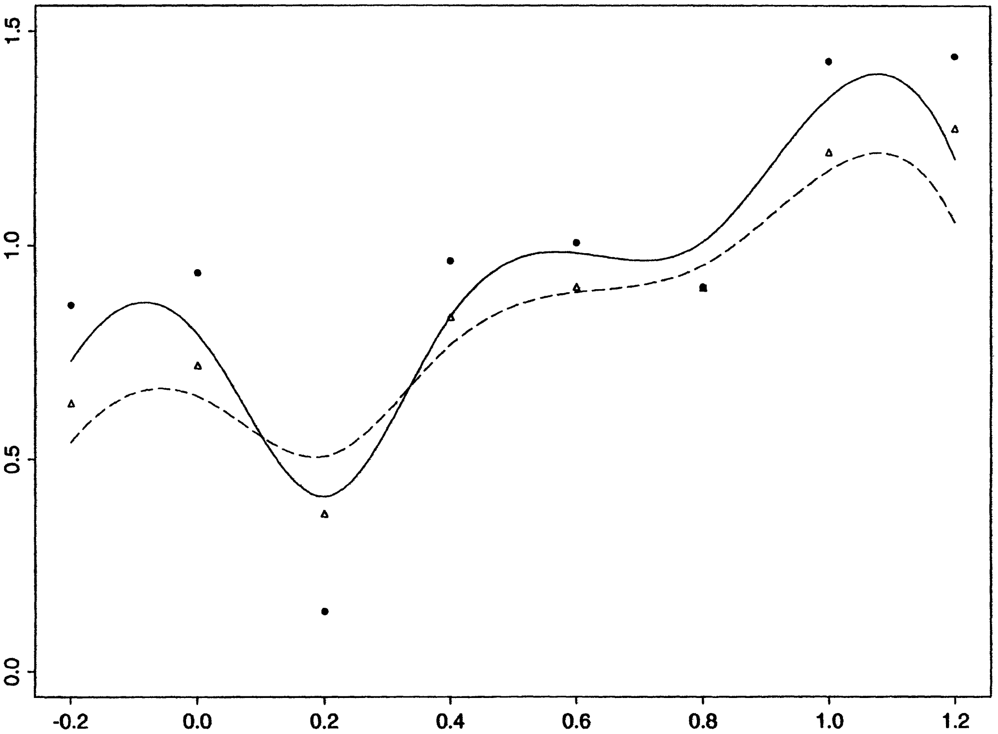

FIG. 1. *Example of control points: the solid curve derives from the solid control points, and the dashed curve from the triangular control points.*

a feel for their influence on the curve or surface. Similarly, statisticians may find after some effort that B-spline coefficients are very natural.

If I had equally easy to use software for smoothing splines or P-splines, I would prefer the former, partially from Bayesian considerations. The Bayesian interpretation of P-splines (i.e., the differenced B-spline coefficients are a Gaussian white noise under the prior) is more artificial than the usual priors as in Wahba (1978). In particular, the usual priors are specified independently of sample size, whereas one would want to use more B-splines with a larger sample. Furthermore, the integral of the second derivative squared is easier to interpret from a non-Bayesian perspective than the sum of squares of second differences of B-spline coefficients.

I take issue with the authors's claim that their method does not have boundary problems. P-splines are approximately equivalent to smoothing splines which do have boundary effects (Speckman, 1983). To explain, consider minimizing from equation (5),

$$S(\mathbf{a}) = \sum_{i=1}^{m}\left\{y_i - \sum_{j=1}^{n} a_j B_j(x_i)\right\}^2 + \lambda \sum_{j=3}^{n}(\Delta^2 a_j)^2.$$

A discrete form of the variational derivation in Speckman (1983) leads to the system

$$\begin{aligned}
&\lambda\Delta^2 a_3 + \sum_i B_1(x_i)\sum_j a_j B_j(x_i)\\
&\quad = \sum_i y_i B_1(x_i),\\
&\lambda\Delta^3 a_4 - \lambda\Delta^2 a_3 + \sum_i B_2(x_i)\sum_j a_j B_j(x_i)\\
&\quad = \sum_i y_i B_2(x_i),\\
&\lambda\Delta^4 a_k + \sum_i B_k(x_i)\sum_j a_j B_j(x_i)\\
&\quad = \sum_i y_i B_k(x_i), \quad 3 \le k \le n-2,\\
&-\lambda\Delta^3 a_n - \lambda\Delta^2 a_n + \sum_i B_{n-1}(x_i)\sum_j a_j B_j(x_i)\\
&\quad = \sum_i y_i B_{n-1}(x_i),\\
&\lambda\Delta^2 a_n + \sum_i B_n(x_i)\sum_j a_j B_j(x_i)\\
&\quad = \sum_i y_i B_n(x_i).
\end{aligned}$$

Notice that the equations for coefficients near the end involve lower-order differencing so there is less smoothness imposed.

ACKNOWLEDGMENT

Research supported by NSF Grant DMS-90-01726.

Comment

Stephan R. Sain and David W. Scott

We have been interested in formulations of the smoothing problem that are simultaneously global in nature with locally adaptive behavior. Roughness penalties based on functionals such as the integral of squared second derivatives of the fitted curve have enjoyed much popularity. The solution to such optimization problems is often a spline. The authors are to be congratulated for introducing the idea of penalizing on the smoothness of the spline coefficients, which reduces the dimensionality of the problem as well as reducing the complexity of the calculations. There is much to say for this approach.

It is generally of interest to try to work out the equivalent kernel formulation of all smoothing methods. This was done for Nadarya–Watson regression smoothing by Silverman (1984), who demonstrated the asymptotic manner in which the estimator adapted locally.

In the density estimation setting, we have been investigating the nature of the best locally adaptive density estimator along the lines of the Breiman–Meisel–Purcell estimator (Breiman, Meisel and Purcell, 1977)

$$(1)\quad \hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_i} K\left(\frac{x - x_i}{h_i}\right) = \frac{1}{n}\sum_{i=1}^{n} K_{h_i}(x - x_i).$$

The goal is to find an optimal set of bandwidths $\hat{h}_i, i = 1, \ldots, n$, without restrictions on the functional form. Sain and Scott (1996) explore an approach using a binned version of (1) where the bandwidths were found numerically by optimizing over a variation of the least-squares or unbiased cross-validation (UCV) criterion.

The surprising finding of our research is that the optimal estimator contains distinctly nonlocal as well as local adaptive features. That is, the bandwidths for some data points, particularly in the tails, are very large. This was rather unexpected since Terrell and Scott (1992) discussed the negative consequences of such large bandwidths when $h_i = h/\sqrt{f(x_i)}$, an idea suggested by Abramson (1982) and studied extensively in the literature. Furthermore, Sain and Scott (1996) showed that this "square-root law," in practice, lacks flexibility due to the dependence solely on the level of the underlying density. We refer the interested reader to those articles.

In Figure 1 we show three densities of the geyser data: (1) our optimal locally adaptive estimate; (2) a fixed kernel estimate (bandwidth also chosen by UCV); and (3) the authors's P-spline. The fixed bandwidth approach cannot find a single bandwidth to smooth both modes appropriately, leaving the right mode undersmoothed. The more flexible adaptive estimator recognizes the local structure of the underlying density and gives a clear representation of the two modes in the data (rejecting the possibility of a third mode) without excessive noise. The P-spline estimator yields an estimate lying somewhere between the two approaches. It is

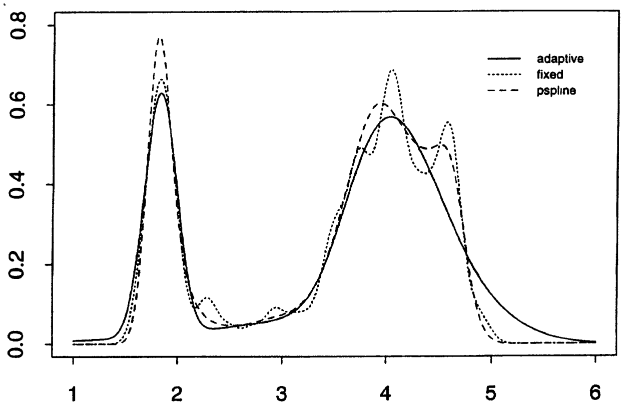

FIG. 1. *Estimated densities using the Old Faithful geyser data.*

Stephan R. Sain is with Southern Methodist University. David W. Scott is Professor of Statistics, Rice University, Houston, Texas 77251-1892.

interesting to note that the adaptive approach is not oversmoothed as the heights of the two modes are not affected.

The P-spline (and other similar estimators) allow some local adaptiveness through the penalty mechanism and the restrictions on the roughness of the fitted curve. However, the choice of knots (in this case an arbitrarily large number of equally spaced knots) can also affect the local nature of the P-spline estimator. It would be interesting to see more on how to choose the knots, including an "adaptive" approach that might lead to a more parsimonious model as well as better local behavior. Finally, one must wonder how the choice of the bin width for the initial histogram affects the resulting P-spline estimate. In our experience, reliance on such pilot estimators can lead to poor results as well as difficulties in automatic implementation.

We would be very interested to see a more thorough study of how P-splines behave vis-à-vis some notion of optimal local adaptiveness, and how the penalty and AIC criterion, as well as other parameters, can be set to achieve such behavior.

ACKNOWLEDGMENT

This research (second author) was supported in part by Grant DMS-96-26187.

Rejoinder

Paul H. C. Eilers and Brian D. Marx

INTRODUCTION

Do P-splines deserve a place in the spotlight? We claimed so, generating a lot of discussion. We are grateful for the many careful, positive and detailed comments. For our rejoinder, we group them as follows:

- extensions and clarifications, especially concerning optimal smoothing;
- challenges to the performance of P-splines;
- doubts about our claim that P-splines come near to the ideal smoother.

We will react in the same order, first commenting on the extensions. Then we will show analyses and examples to show that we can meet all challenges with standard P-splines, except adaptive flexibility (but the need for that is less than one might think). After that we present a kind of "consumer test," with a scoring table, to compare P-splines to the competition. Finally, we will conclude that P-splines meet nearly all challenges and summarize why they are attractive to use.

OPTIMAL SMOOTHING

We did not mean to imply that AIC and cross-validation are the final word on optimal smoothing. We advocated their use because they can be computed easily and fast, and because we have had good experiences in many real-life applications. But the search for optimal criteria has to continue, and there is no obstacle in P-splines to prevent the use of more sophisticated methods.

Chiu's filtering approach is interesting. However, it seems limited to equidistantly sampled data, because a (fast) Fourier transform is needed.

Nychka and Cummins introduce an interesting interpretation of P-spline smoothing as a projection on the Demmler–Reinsch basis. They use an equispaced x-vector implictly. For sparse designs, some of the columns of B may become empty, making $B^T B$ singular. We suspect that additional restrictions (like a small ridge penalty) then will be needed to make the construction of G possible. The advantages of the Demmler–Reinsch basis are mainly conceptual: the computation the trace of the hat matrix and GCV can already be done efficiently with P-splines.

While we are on this subject we would like to add that we do not understand the widespread preoccupation with the sum of squares functional $\int(f - \widehat{f})^2\,dx$ as a measure of performance in density estimation, and the detailed analyses that have been made. One would expect some deviance-like functional, or Kullback–Leibler distance, such as Gu is using. After all, no one fits a density to a histogram with least squares. It seems that mathematical tractability is the driving force behind it, reminding us of the drunkard searching under a street lantern for the keys he lost elsewhere in the dark. To help him, bright mathematicians hook up a metal detector to the lantern.

We note that for so-called second-generation criteria (Jones, Marron and Sheather, 1996), P-splines can be very useful when one has to estimate (integrals of squared) third or higher derivatives, because of the ease of generating high-degree B-splines. But we are very content about AIC. Figure 1 shows data from Cook and Weisberg (1994), giving the lean body mass of Australian athletes. These data were used as a test bed by Jones, Marron and Sheather (1996). We use AIC and get essentially the same amount of smoothing, "a second generation result at a first generation price." Again, we do not wish to imply that AIC is the final answer, but show that it is more useful than sometimes suggested.

THE KNOTS

In our paper we were rather conservative in the number of knots we used and advised to use. Yet many variations are possible. Below we will see examples with very many knots, even more than there are data points, giving a counterexample to Engel and Kneip's assumption. They are right that, with as many knots as there are data points, we come very near to the smoothing spline, if the x's are equidistant. If this is not the case, we need knots on a non-equidistant grid. But then the penalty has to change too: divided differences, like $(a_j - a_{j-1})/(t_j - t_{j-1})$ in the case of a first-order penalty, have to be used. We have not yet fully analyzed this situation, but we suspect some interesting results, because B-splines on an arbitrary grid are computed with a divided difference scheme.

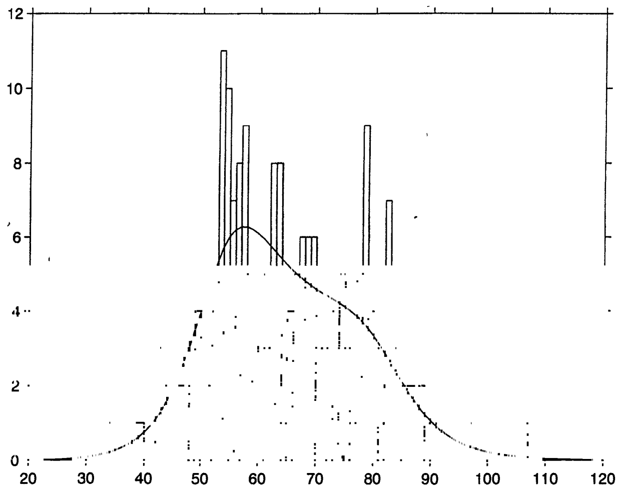

FIG. 1. *Histogram and density estimate of lean body mass of Australian athletes (male and female);* 20 *P-splines of degree* 3, *penalty of order* 3, $\lambda = 10$.

The number of knots is largely immaterial, as long as it is large enough. Yet the sugestion of Engel and Kneip to optimize both λ and n can have value when striving for parsimony. With $\lambda = 0$, we can change n until we find that for optimal smoothing a number between a certain $n' - 1$ and n' appears needed. Take $n = n'$ and increase λ for the last part of the road to optimality. However, problems may arise with sparse designs, in which case nonidentifiablity may occur without a penalty.

BOUNDARIES

Regarding Kooperberg's concern in choosing boundaries, we must discern "physical" from "technical" boundaries. A physical boundary is determined by the nature of the variable under study. In the suicide example, zero is a physical left boundary, because time intervals cannot have negative length, and so there can be no density below zero. If we choose as a technical boundary a certain negative value, we say that there can be negative data, that we did not observe them, but that it is reasonable to estimate a density in that region. In the same example the upper boundary is technical, because we know of no upper limit (well, maybe 80 years or so). It does not matter much which value we take, a practical choice is 2 times the highest observed value.

Kooperberg asks how we extrapolate. We do not do that: we choose the boundaries (when physically meaningful) wide enough to include the domain where "extrapolation" is wanted.

Cox shows that there is indeed a boundary effect, but of a more subtle kind than we were considering in our paper, where we meant the unpleasant property of kernels to distribute probability mass outside the (physical) boundaries in density estimation, or tending toward zero in kernel regression.

DEFAULT PARAMETER VALUES

As Kooperberg rightly remarks, we were vague in giving default values for some parameters. Rules of thumb might be the following: take the interval between knots as half the width of the narrowest peak that should be visible in a fitted curve; use B-splines of degree 3 and a penalty of order 3. We recommend plots of AIC or GCV against effective dimension, for the orders 1, 2 and 3 of the difference penalty. For density estimation by histogram smoothing we recommend 100 bins or more. It can do no harm to take a large number of B-splines, such as 50 or 100, because the penalty takes care of any overfitting. Of course, for faster computation it is best to have a small number of B-splines.

In some cases B-splines of degree zero, which are just constant between two knots, and zero elswhere, are sufficient. If we take the number of knots equal to the number of (equidistantly sampled) data, B is the identity matrix. We cannot resist the temptation to show how simple smoothing becomes that way. Let the vector y be the data and w be a vector of 0–1 weights to indicate missing data. Then the three MATLAB lines

```
I = eye(length(y));
D = diff(I, 3);
mu = (diag(w) + lambda * D' * D) \ (diag(w) * y);
```

do the trick. In fact this is just Whittaker's (1923) "graduation" algorithm.

ASYMPTOTICS

We did not yet have any asymptotic results on rates of convergence. However, borrowing from the asymptotic GLM theory, upon convergence with fixed λ, the asymptotic variance–covariance matrix of the P-splines coefficients is $\tilde{\Phi} = (Q_B + Q_\lambda)^{-1} Q_B (Q_B + Q_\lambda)^{-1}$. This result is particularly useful for straightforward construction of twice standard error bands for $g(\tilde{\mu})$, that is, $\mathrm{var}\{g(\tilde{\mu})\} = B\tilde{\Phi}B^T$. Other asymptotic theory follows regarding variance and bias of the P-spline coefficients. Of course, these have to be translated to propeties of the estimated curve, because the coeffients themselves have a limited interpretation.

ADAPTIVE SMOOTHING

P-splines have constant flexibility. Kooperberg, Sain and Scott discuss adaptive estimation, in which non-constant flexibility is needed. First we will analyze the income data, then Old Faithful.

We have to admit right from the start that P-splines in their present form cannot challenge the extreme control on flexibility that LOGSPLINE offers. Yet we can come a long way with constant flexibility. Figure 2 shows a histogram of the data with bin width 0.1 (because of the large number of bins, the counts are drawn as vertical lines in the midpoint of each bin). The same interval is used for the knots of the B-splines of degree 3, giving 153 of them. You cannot have such a large number without a penalty, because of severe identifiability problems. We see that the optimal fit is very near to the histogram itself, giving too wiggly a right tail. Without the outlier, we get nearly the same result, which is to be expected with a small amount of smoothing. Figure 3 is based on a part of the data, using smaller bins. The left peak is recovered rather well, but again the right part seems too wiggly. We conclude that the high number of observations and the outlier do not wreak havoc on P-splines, but that their fixed flexibility leads to a small amount of smoothing.

Incomes are positive and show a large ratio between maximum and minimum. Data of that type should always be studied also on a logarithmic scale. This is done in Figure 4. The estimated density looks very reasonable. An inflection between the two peaks is indicated that LOGSPLINE does not pick up.

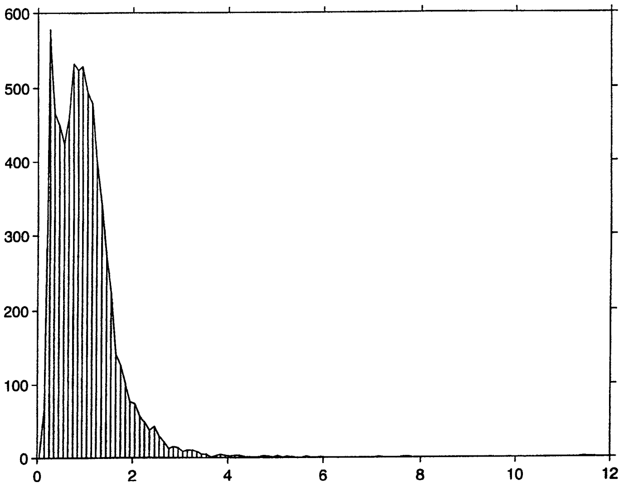

FIG. 2. *Histogram (bin width* 0.1) *and minimum AIC fit of* 153 *P-splines to the income data*; $\lambda = 10^{-6}$.

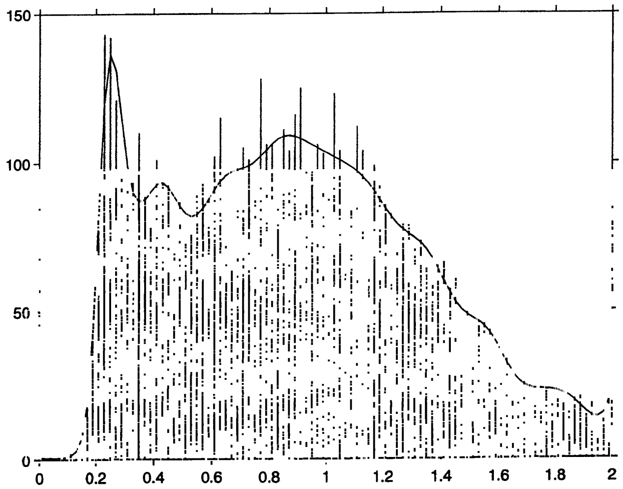

FIG. 3. *Histogram (bin width* 0.1) *of income data with* $x < 2$ *and minimum AIC fit of* 53 *P-splines to the income data*; $\lambda = 10^{-6}$.

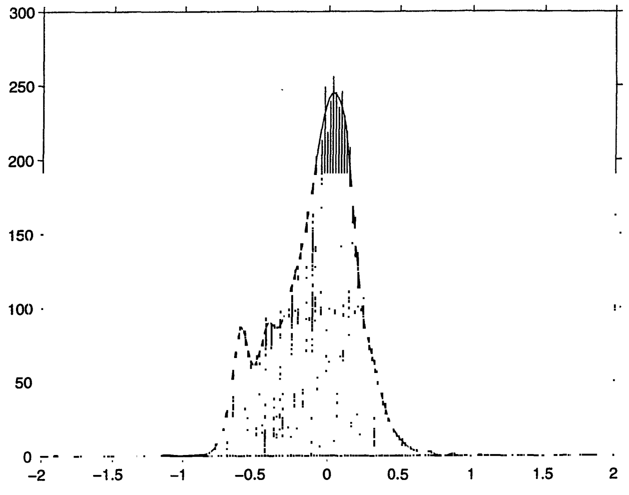

FIG. 4. *Histogram (bin width* 0.1) *and minimum AIC fit of* 53 *P-splines to the logarithm of the income data*; $\lambda = 10^{-6}$.

Logarithms also work well for the Old Faithful data, as Figure 5 shows. We estimated a density for the logarithms of the data and transformed that back to the linear scale. Note that the tail at the right side is much shorter than the one Sain and Scott present in their figure; is it variance inflation by the kernel smoother? Sain and Scott mention that they found unexpectedly larger kernel band widths in the tails. The left tail of that curve is rather strange, being appreciably larger than zero over a long stretch.

It is interesting to note that the effective dimension we computed is 10.8 for the linear data and 8.2 for the logarithmic data, indicating that on the latter scale appreciably stronger smoothing is allowed.

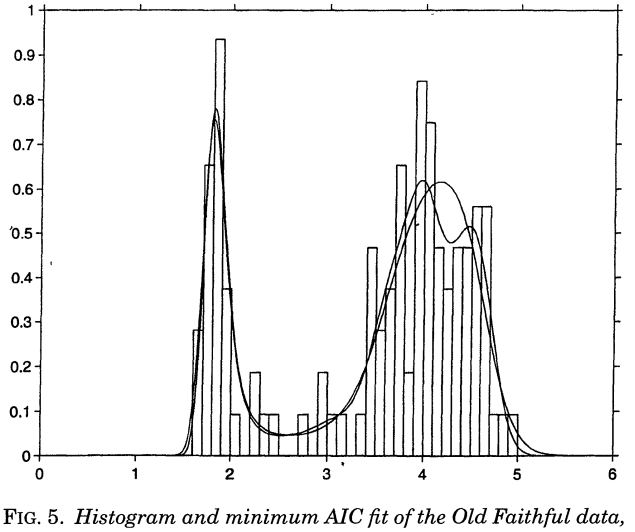

FIG. 5. *Histogram and minimum AIC fit of the Old Faithful data, on both a linear and a (back-transformed) logarithmic scale; the curve with two modes on the right is based on the linear scale.*

Of course, there will be situations in which a simple transformation will not work. A nonconstant flexibility might be realized by using suitable weights in the penalty, like $\sum v_j(\Delta^d a_j)^2$. A simple way of adaptive estimation might be borrowed from Fan and Gijbels (1995): divide the support in a number of (overlapping) intervals, do the smoothing for each of these separately, giving a number of optimal λ's. These can be interpolated to give a smooth curve of λ. The v's in the weighted penalty can be set proportional to the values of that curve at the knots.

Fan et al. (1996) studied kernel smoothing with continuously varying band width $b(x)$ by smooth interpolation of a low-dimensional set of points (x_j, b_j), and optimizing the b_j's. One can imagine optimizing the elements of v in a similar way.

DERIVATIVE OR DIFFERENCES?

Several discussants suggest that a penalty on the second derivative is more clear than one on the coefficients of B-splines. We think not. Those coefficients are the heights of the B-splines that build the fitted curve and so they have a direct intuitive interpretation. Smoothness demands that the heights of neighboring B-splines may not differ to much. The penalty lets the B-splines "hold hands" to withstand erratic fluctuations in the data.

Figure 6 shows simulated data and a B-spline fit without a penalty, while in Figure 7 a strong penalty ($\lambda = 2$, with second-order differences) is used. Note the smooth envelope that is suggested by the tops of the B-splines.

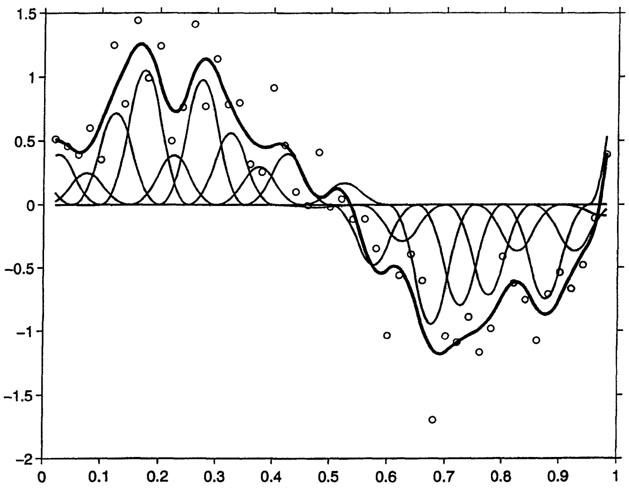

FIG. 6. *Simulated data, individual B-splines and the fitted curve (thick line) without a penalty.*

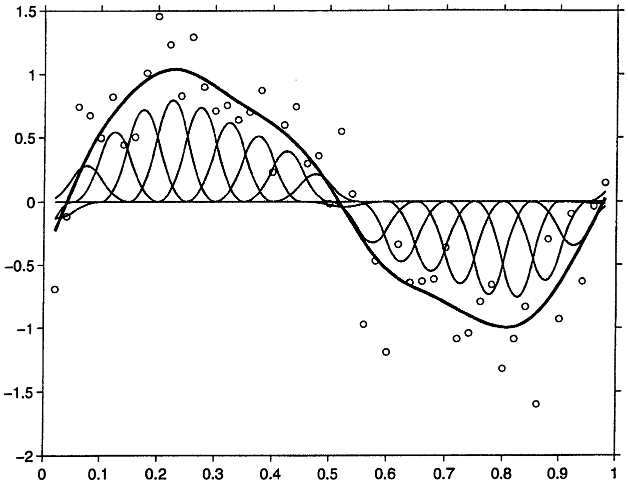

FIG. 7. *Simulated data, individual B-splines and the fitted curve (thick line) with a second order penalty and* $\lambda = 2$.

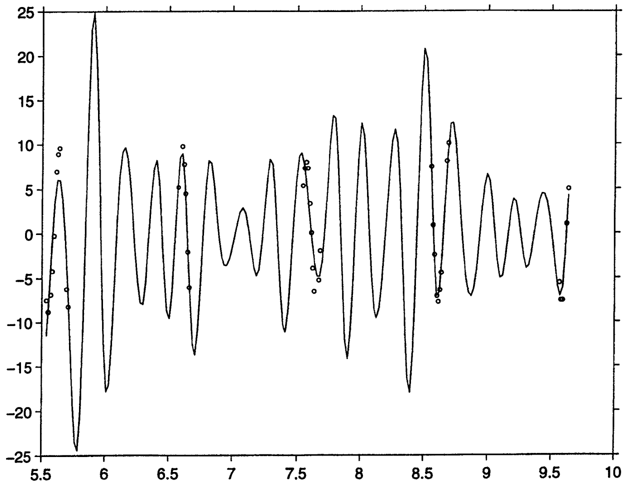

FIG. 8. *Velocities of a variable star (dots) and fit of* 50 *P-splines with a specialized penalty that forces the fit toward a sine curve.*

The parametric limit can be illustrated in the same way: when the tops of the B-splines are on a straight line (a parabola), the fitted curve is linear (quadratic).

As Cox rightly judges from CAD experience, with some practice one can develop a good intuitive understanding of B-splines.

We agree with Gu that the (very liberally stated) "the penalty is the prior." We also must admit that there the connection to derivatives is much clearer than one to differences of B-spline heights. At present we can only point to the near equivalence of both criteria, as shown in the paper, but we will search for a more intuitive understanding.

SPECIALIZED PENALTIES

We now come to penalties for special parametric limits and use an example to make our point. Suppose that we change the penalty $\lambda \sum (a_j - 2a_{j-1} + a_{j-2})^2$ to $\lambda \sum (a_j - 2ca_{j-1} + a_{j-2})^2$, with $c = \cos(2\pi t/p)$, and t the distance between knots, then for high λ the series a tends to a sine function with period p: $a_j = a_0 \cos(2\pi jt/p + \phi)$, with a_0 and ϕ determined by the data. This forces the fitted series toward a sine signal, interpolated by B-splines. With t small compared to p, this will effectively be a sine curve. Figure 8 shows a part of a series of measurements of velocities of a variable star, centered to have zero mean; the data were provided by Conny Aerts of Leuven University. The assumed value for the period p is 0.161 day.

The figure also shows that extreme holes in the data can be handled with P-splines. With the difference penalty this would work too, but there would be very large and very smooth swings up and down in regions without data, because there it is only smoothness that counts. This can be remedied by adding a small ridge penalty, another example of specializing the penalty to the problem.

For completeness we note that Eilers (1991a) used mixed penalties in a regression problem with ordered regressors. Eilers (1988) proposed to use penalized least squares to estimate autoregressive models for noisy signals with possibly missing data.

CONSERVATION OF MOMENTS

Jones does not see the need for conservation of moments in regression. We think it is important: if they were not conserved, the parametric model that is approached with strong smoothing would be different from the one obtained with polynomial regression. In density estimation, variance inflation is undesirable. The work of several first-class statisticians, including Jones himself, testifies to this.

MULTIVARIATE APPLICATIONS

As for multivariate applications, we have efficient MATLAB functions for two-dimensional P-splines, based on tensor products of one-dimenional B-splines. In two dimensions the probability of sparse data is high. To prevent identifiability problems, a penalty is nearly mandatory. We are not optimistic about generalizations to higher dimensions. In theory it is not so difficult, but the organization of the computations is difficult. Also the number of

TABLE 1

Consumer test of smoothing methods; the abbreviations stand for the following: KS, kernel smoother; KSB, kernel smoother with binning; LR, local regression; LRB, local regression with binning; SS, smoothing splines; SSB, smoothing splines with band solver; RSF, regression splines with fixed knots; RSA, regression splines with adaptive knots; PS, P-splines. The row "Adaptive flexibility available" means that a software implementation is readily available

Aspect	**KS**	**KSB**	**LR**	**LRB**	**SS**	**SSB**	**RSF**	**RSA**	**PS**
Speed of fitting	—	+	—	+	—	+	+	+	+
Speed of optimization	—	+	—	+	—	+	—	—	+
Boundary effects	—	—	+	+	+	+	+	+	+
Sparse designs	—	—	—	—	+	+	—	+	+
Semi parametric models	—	—	—	—	+	—	+	+	+
Non-normal data	+	+	+	+	+	+	+	+	+
Easy implementation	+	—	+	—	+	—	+	—	+
Parametric limit	—	—	+	+	+	+	+	+	+
Specialized limits	—	—	—	—	+	+	—	—	+
Variance inflation	—	—	+	+	+	+	+	+	+
Adaptive flexibility possible	+	+	+	+	+	+	—	+	+
Adaptive flexibility available	—	—	—	—	—	—	—	+	—
Compact result	—	—	—	—	—	—	+	+	+
Conservation of moments	—	—	+	+	+	+	+	+	+
Easy standard errors	—	—	+	+	—	+	+	+	+

the basis functions may easily become larger than the number of observations.

BIN WIDTH OF HISTOGRAMS

Sain and Scott like to see an investigation of how much the bin width of a histogram influences the *P*-spline density estimate. We do that empirically with the Old Faithful (that name gets a new meaning here) data. In Figure 9 we plot the estimated density for five values of the bin width: 0.2, 0.1, 0.05, 0.02 and 0.01. The optimal value λ_{opt} was found by trying a decreasing series of integer powers of 10, starting at 10^4, stopping when AIC started to rise. To the last three pairs $(\log \lambda, \text{AIC})$, a parabola was fitted; the location of its minimum gave $\log \lambda_{\text{opt}}$. It appears that bins of 0.2 are too coarse, but for the other values the curves are practically the same. As a rule of thumb we might derive that a histogram with 100 or more bins is a good choice.

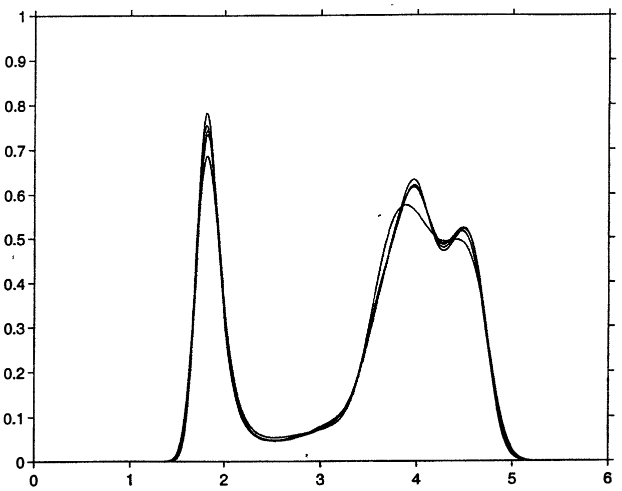

FIG. 9. *Five estimated densities for the Old Faithful data, based on histograms with bin widths* 0.2, 0.1, 0.05, 0.02, 0.01*; the curve with the lowest peaks is for bin width* 0.2.

A CONSUMER TEST

Several discussants doubt whether *P*-splines come as near to the ideal smoother as we claim. Every good property we mentioned can also be found in one or more other methods. In Table 1 we have constructed a "consumer test" of smoothers to make comparison easier. We have neglected most ad-hoc changes that have been published to remedy troubles like boundary effects and sparse designs, because they have not yet appeared in readily available software. The meaning of "specialized limits" will be explained below. Of course one can argue about some of the pluses or minuses of the competing methods, but the advantages of *P*-splines have a firm basis. Anyone wishing to use them can be on track in a few hours, in any language that supports matrix operations and/or regression, starting from the algorithms in our Appendix.

A HAPPY ENDING

Yes, we think that *P*-splines deserve a place in the spotlight. They are easy to use, easy to program and easy to understand. They respect boundaries, have no problems with sparse designs and give compact results. Polynomial and exponential (sinusoidal) limits can be forced with almost trivial changes to the difference operator in the penalty.

Yet there remains a lot to be done, especially on optimizing the weight of the penalty and on adaptive flexibility. A better understanding of the Bayesian interpretation of the penalty is needed. We will continue our research in these areas. We hope to meet many others there who also have recognized the charm of P-splines.

ACKNOWLEDGMENTS

Once again we thank the discussants for their inspiring comments. We also acknowledge the advice of (anonymous) referees and the support of Associate Editor David Scott (with extra thanks for organizing the discussion) and Editor Paul Switzer on our long trail to a paper in print.

ADDITIONAL REFERENCES

COOK, R. D. and WEISBERG, S. (1994). *Regression Graphics*. Wiley, New York.

EILERS, P. H. C. (1988). Autoregressive models with latent variables. In *COMPSTAT 1988 Proceedings* (D. Edwards and N. E. Raun, eds.). Physica-Verlag.

ENGEL, J. and GASSER, T. (1995). A minimax result for a class of nonparametric density estimators. *Nonparametric Statistics* **4** 327–334.

FAMILY EXPENDITURE SURVEY (1968–1983). Annual base tapes and reports (1968–1983). Dept. Employment, Statistics Division, Her Majesty's Stationary Office, London.

FAN, J. (1993). Local linear regression smoothers and their minimax efficiency. *Ann. Statist.* **21** 196–216.

FAN, J. and GIJBELS, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Roy. Statist. Soc. Ser. B* **57** 371–394.

FAN, J., HALL, P., MARTIN, M. A. and PATIL, P. (1996). On local smoothing of nonparametric curve estimators. *J. Amer. Statist. Assoc.* **91** 258–266.

FOLEY, J. D., VAN DAM, A., FEINER, S. K. and HUGHES, J. F. (1996). *Computer Graphics: Principles and Practice*. Addison-Wesley, Reading, MA.

FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19** 1–141.

JONES, M. C., MARRON, J. S. and SHEATHER, S. J. (1996). A brief survey of bandwidth selection for density estimation. *J. Amer. Statist. Assoc.* To appear.

KNEIP, A. (1994). Ordered linear smoothers. *Ann. Statist.* **22** 835–866.

KOOPERBERG, C., BOSE, S. and STONE, C. J. (1997). Polychotomous regression. *J. Amer. Statist. Assoc.* To appear.

KOOPERBERG, C., STONE, C. J. and TRUONG, Y. K. (1995a). Hazard regression. *J. Amer. Statist. Assoc.* **90** 78–94.

KOOPERBERG, C., STONE, C. J. and TRUONG, Y. K. (1995b). Logspline estimation of a possibly mixed spectral distribution. *J. Time Ser. Anal.* **16** 359–388.

SPECKMAN, P. L. (1983). Spline smoothing and optimal rates of convergence in nonparametric regression models. *Ann. Statist.* **13** 970–983.

STONE, C. J., HANSEN, M., KOOPERBERG, C. and TRUONG, Y. K. (1996). Polynomial splines and their tensor products in extended linear modeling. *Ann. Statist.* To appear.

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