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.

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).

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-

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: $x_1, \ldots, x_4$. 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-

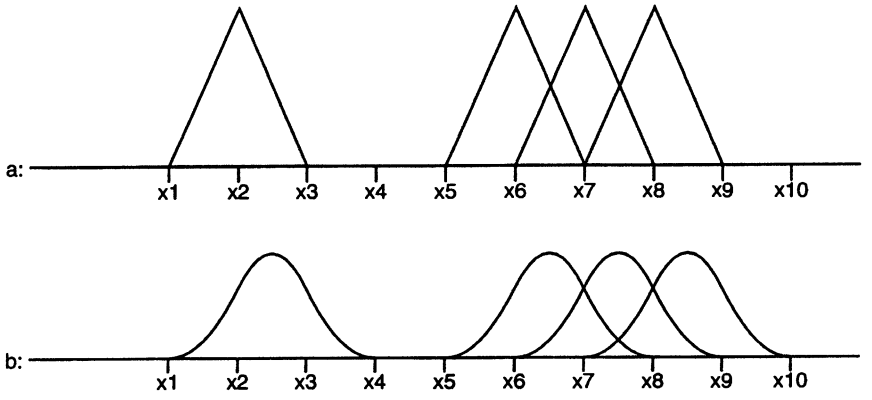

FIG. 1. *Illustrations of one isolated B-spline and several overlapping ones* (a) *degree* 1; (b) *degree* 2.

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

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

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

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

or

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

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:

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

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

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

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

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

This can be written as

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

where

$$L = Q_B^{-1/2} Q_\lambda Q_B^{-1/2} \tag{24}$$

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 $\text{tr}(H)$ 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

$$\text{AIC}(\lambda) = \text{dev}(y; a, \lambda) + 2 * \dim(a, \lambda), \tag{25}$$

where $\dim(a, \lambda)$ is the (effective) dimension of the vector of parameters, a, and $\text{dev}(y; a, \lambda)$ 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 $\dim(a) = \text{tr}(H)$. Note that $\text{tr}(H) = n$ when $\lambda = 0$, as in (nonsingular) standard linear regression.

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

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

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, $\hat{\sigma}_0^2$:

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

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

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

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

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

The difference between both quantities is generally small. The best λ is the value that minimizes $\text{CV}(\lambda)$ or $\text{GCV}(\lambda)$. 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 $\text{AIC}(\lambda)$. It is practical to work with modified versions of $\text{CV}(\lambda)$ and $\text{GCV}(\lambda)$, with values that can be interpreted as estimates of the cross-validation standard deviation:

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

The two terms in $\text{AIC}(\lambda)$ represent the deviance and the trace of the smoother matrix. The latter term, say $T(\lambda) = \text{tr}\{H(\lambda)\}$, 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
$\overline{\text{CV}}$	24.77	24.02	23.52	23.37	23.26	23.38	23.90	25.50	27.49
$\overline{\text{GCV}}$	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 $\overline{\text{GCV}}$, 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 $\overline{\text{GCV}}$. The system to be solved is

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

The sum of squares is

$$S = |y - Ba|^2 = y^T y - 2a^T B^T y + a^T B^T Ba. \tag{32}$$

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 $D_k^TD_k$. 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

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

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:

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

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

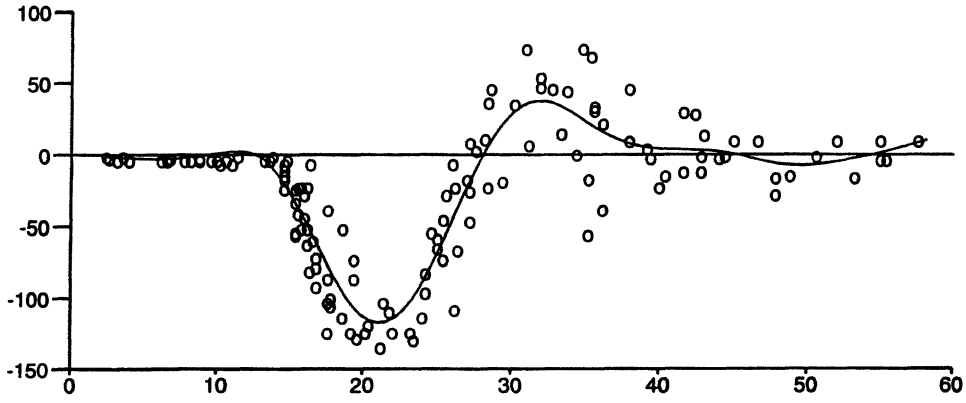

FIG. 2. *Motorcycle crash helmet impact data: optimal fit with B-splines of third degree, a second-order penalty and* $\lambda = 0.5$.

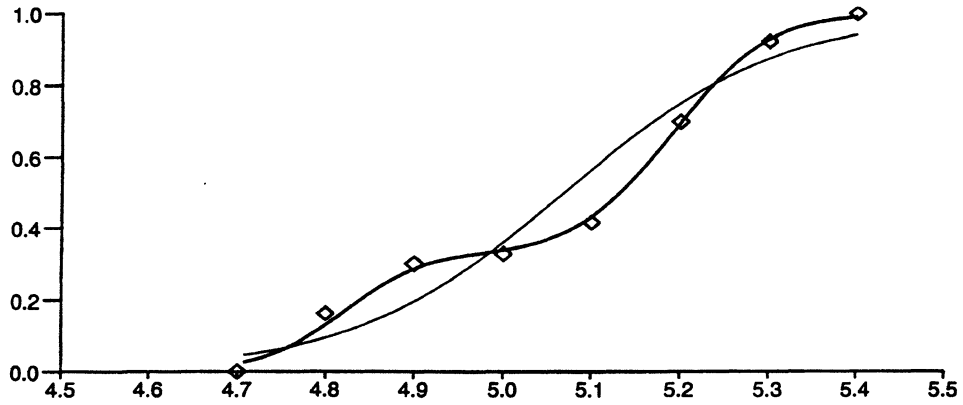

FIG. 3. *Nonparametric logistic regression of Trypanosome data: P-splines of order* 3 *with* 13 *knots, difference penalty of order* 2, $\lambda = 1$ *and* AIC = 13.4 *(thick line); the thin line is effectively the logistic fit* ($\lambda = 10^8$ *and* AIC = 27.8).

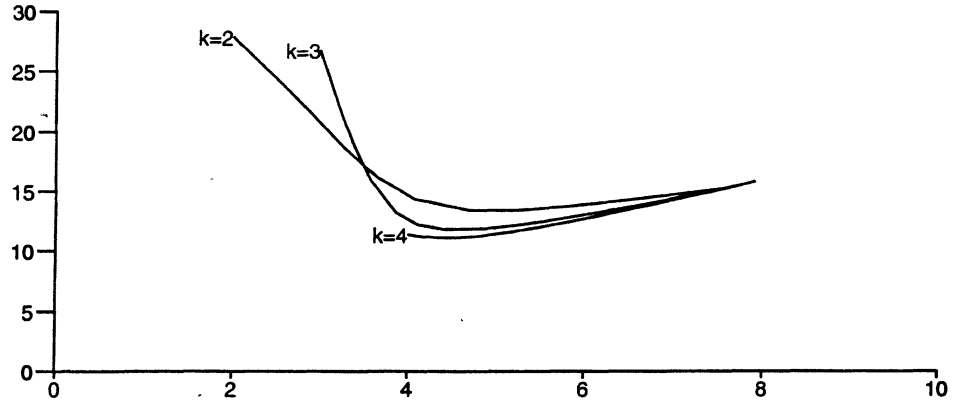

FIG. 4. AIC(λ) *versus* $T(\lambda)$, *the effective dimension, for several orders of the penalty* (k).

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 here 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

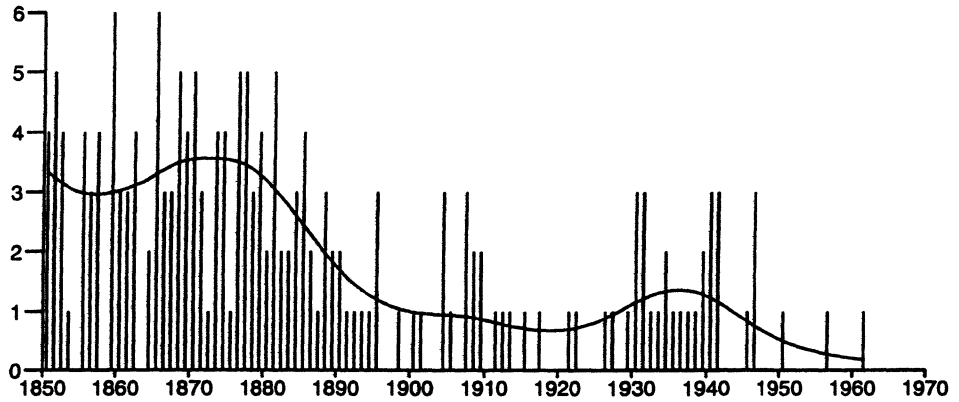

FIG. 5. *Numbers of severe accidents in British coal mines: number per year shown as vertical lines; fitted trend of the expectation of the Poisson distribution; B-splines of degree* 3, *penalty of order* 3, 20 *intervals between* 1850 *and* 1970, $\lambda = 1{,}000$ *and* AIC = 126.0.

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, $i = 1, \ldots, m$, 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 $t = 1, \ldots, r$, 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:

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

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 $B_j(u_i)$ 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:

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

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.

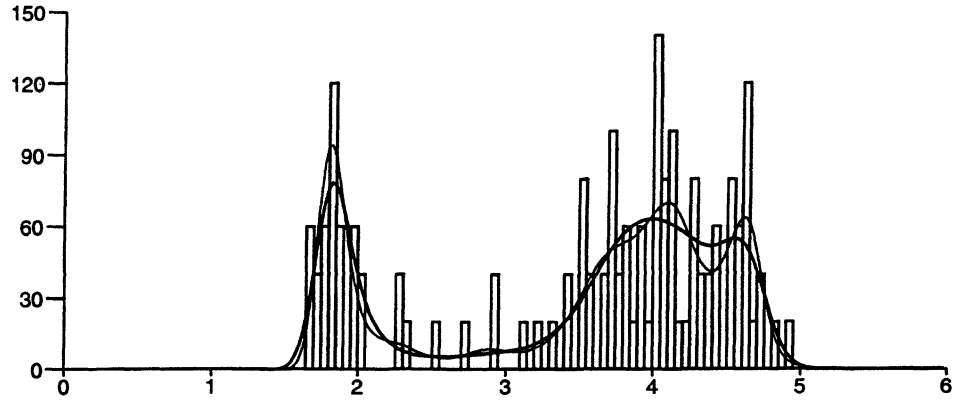

FIG. 6. *Density smoothing of durations of Old Faithful geyser eruptions: density histogram and fitted densities; thin line, third-order penalty with* $\lambda = 0.001$(AIC = 84.05); *thick line, optimal* $\lambda = 0.05$, *with* AIC = 80.17; *B-splines of degree* 3 *with* 20 *intervals on the domain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

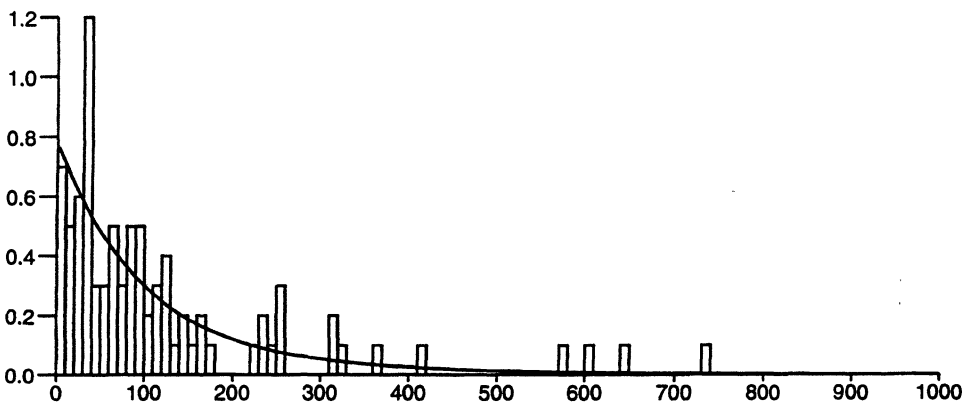

FIG. 7. *Density smoothing of suicide data: positive domain* (0–1,000); *B-splines of degree* 3, *penalty of order* 2, 20 *intervals,* $\lambda = 100$, AIC = 69.9.

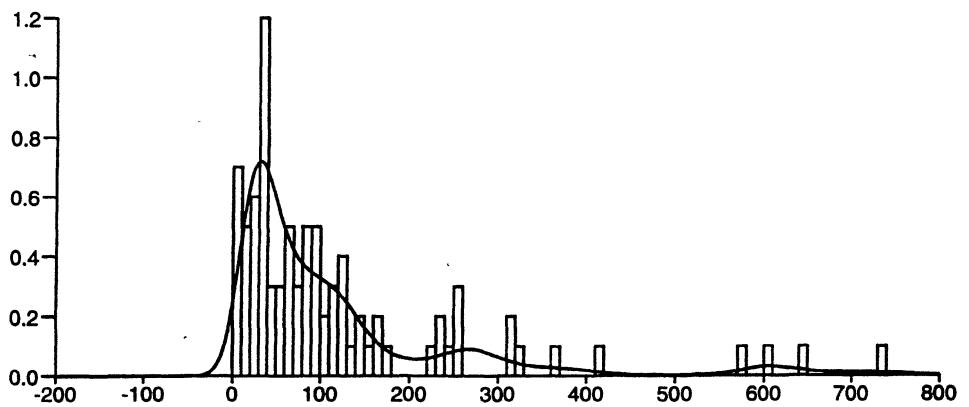

FIG. 8. *Density smoothing of suicide data: the domain includes negative values* (−200–800); *B-splines of degree* 3, *penalty of order* 2, *20 intervals,* $\lambda = 0.01$, AIC = 83.6.

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

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

and to compute $|y - B\hat{a}|^2$ and $\mathrm{tr}\{(B^TB + \lambda D^T D)^{-1} \cdot B^TB\}$. 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^TB and the inverse of $B^TB + \lambda D_k^T D_k$).

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.