ELSEVIER Computational Statistics & Data Analysis 28 (1998) 193–209

COMPUTATIONAL STATISTICS & DATA ANALYSIS

Direct generalized additive modeling with penalized likelihood

Brian D. Marx [a,*], Paul H.C. Eilers [b]

[a] *Department of Experimental Statistics, Louisiana State University, Baton Rouge, LA 70803, USA*

[b] *DCMR Milieudienst Rijnmond, Netherlands*

Received September 1995; received in revised form December 1997

Abstract

Generalized additive models (GAMs) have become an elegant and practical option in model building. Estimation of a smooth GAM component traditionally requires an algorithm that cycles through and updates each smooth, while holding other components at their current estimated fit, until specified convergence. We aim to fit all the smooth components simultaneously. This can be achieved using penalized B-spline or P-spline smoothers for every smooth component, thus transforming GAMs into the generalized linear model framework. Using a large number of equally spaced knots, P-splines purposely overfit each B-spline component. To reduce flexibility, a difference penalty on adjacent B-spline coefficients is incorporated into a penalized version of the Fisher scoring algorithm. Each component has a separate smoothing parameter, and the penalty is optimally regulated through extensions of cross validation or information criterion. An example using logistic additive models provides illustrations of the developments.

Keywords: B-splines; Difference penalty; P-splines; Smoothing

1. Introduction

We revisit the generalized *additive* model (GAM) which fits a response variable Y by a sum of smooth functions of the explanatory variables, X_j for $j=1,\ldots,p$ (Hastie and Tibshirani, 1986, 1990). For a Normal response, the model is

$$\mu = E(Y) = \alpha + \sum_{j=1}^{p} f_j(X_j) = \eta_A, \tag{1}$$

* Corresponding author.

0167-9473/98/$19.00
PII S0167-9473(98)00033-4

where the $f_j(\cdot)$ are smooth functions. For non-normal responses, the approach of the generalized *linear* model (GLM) is extended and adapted (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989). For instance, an appropriate distribution is chosen for Y (from the exponential family of distributions) and a link function $g(\cdot)$ (monotone, twice differentiable) is introduced. Recall that the linear predictor of the GLM or additive linear effects of explanatory information is

$$g(\mu) = \alpha + \sum_j X_j \beta_j = \eta_L, \tag{2}$$

where the β_j are the unknown parameters. The GAM further generalizes the GLM by replacing the above linear predictor with $g(\mu) = \eta_A$. GAMs can provide useful approximations to the regression surface, upholding the richness of GLMs, but relaxing the linear (polynomial) structure of the additive effects. This additional flexibility relieves the difficulty of searching for the perfect linear relationship between each explanatory variable and the response, yet explains the variability of the response in an additive manner. If each component or a subset of components of a GAM is assumed to be relatively smooth, then these alterations permit users to entertain the vast parametric-free modeling literature as candidates for the additive structure models. Hastie and Tibshirani (1990) (H&T) suggested widespread applications.

Several options are available to reach the objective of fitting the smooth functions of a GAM. Perhaps most commonly *backfitting* is used which is a procedure that resembles the Gauss–Seidel algorithm for an iterative solution of systems of linear equations. To give the unfamiliar reader some intuition and insight, we find that backfitting is most easily explained with a simple example containing two explanatory variables with a normal response. Suppose that the mean of Y, $\mu = E(Y) = \alpha$, and for the moment that $f_2(\cdot)$ is known. The smooth $f_1(\cdot)$ can be found by smoothing the residual $y - \alpha - f_2(X_2)$ by any convenient smoother. Now if we do not know $f_2(\cdot)$, but rather an approximation, then we can improve the estimate of $f_2(\cdot)$ by smoothing $y - \alpha - f_1(X_1)$. This process can be cycled until the smooth fits settle down. Generally, convergence is obtained, even with very rough initial estimates like $f_1(\cdot) = f_2(\cdot) \equiv 0$. The backfitting algorithm is easily generalized to more than two explanatory variables: improve $f_k(\cdot)$ by smoothing of $y - \alpha - \sum_{j \neq k} f_j(X_j)$, cycling over all j, repeating the process until convergence. With non-normal data, generalized residual and smoothing weights using *working* response vectors can be defined to get the desired results, but for non-normal data these must be iterated through, for example, the method of scoring or Newton–Raphson algorithms. Details of the *local scoring algorithm* can be found in H&T.

Several authors have noted that GAMs can be reduced to large GLMs and fundamental work on modelling additive splines can be revisited in Stone and Koo (1985), Stone (1986) and Buja et al. (1989). Although backfitting or local scoring works well in practice, a more direct procedure is desirable – ideally, one that stays near to the regression model of a GLM. In such a case, the derivation of the standard errors, regression diagnostics and convergence properties would be easier and clearer. Several authors have studied this problem, leading to two different approaches: (i) using smoothing splines and another (ii) using regression splines.

First, smoothing splines are discussed in detail in Green and Silverman (1994) and applications to the GAM can be found in H&T. The use of smoothing splines for each component of a GAM does reduce the problem to the (penalized) GLM framework with a smoothness regularizing parameter. Despite connections to the GLM, a drawback to this approach is the fact that a potentially enormous system of equations results: the number of equations in this system is equal to the number of observations (N) where the number of variables is equal to the product of the number of observations times the number of smoothes ($N \times p$). The penalty prevents singularity. Because of the arbitrary combinations of the explanatory variables, there is no chance to exploit a banded structure of the equations that is done for the one-dimensional smoothing splines. The curse of dimensionality is at work here.

Secondly, Hastie and Tibshirani (1990, Ch. 9) studied the use of regression splines. A smooth function is now modeled as the sum of B-splines. This technique is revisited in the next section and smoothing boils down to (generalized) linear regression. Each smooth function is projected to a much lower dimensional sub-space of say $n_j \ll N$. This way, a GAM is reduced to a GLM. The beauty of this method is that B-splines are easy to construct. H&T discuss these advantages, but also discuss the awkward problem of optimizing the position and number of knots that are necessary to define the B-splines. Stone (1994) considers multidimensional splines where the knot placement problem becomes even larger.

We propose to use the strengths of both of these methods. In a previous publication we have shown that knot optimization can be avoided by combining B-splines with a difference penalty (Eilers and Marx, 1996). This approach leads to a one-dimensional smoother with many attractive properties, which we termed P-splines. In this article we aim to show that P-splines are very attractive for GAMs as well, extending our previous work into a p-dimensional additive case. Some of the advantages include:

- GAM estimation is reduced to (generalized) linear regression with a tractable penalty;
- the system of equations is of low dimension and is easy to solve;
- all of the smooths are estimated simultaneously;
- the resulting GAM fit is compactly summarized by a relatively few number of parameters that facilitate future prediction;
- standard errors and regression diagnostics can be computed with relative ease.

These claims are detailed in the following sections, and GAMs with P-splines (termed P-GAMs) will be applied to data in a case study example. First, we provide some background on the GLM and GAM, as well as on B-splines and P-splines.

2. Background and notation for the GLM and GAM

2.1. The generalized linear model

Before delving into P-splines and P-GAMs, we provide some details of the GLM. We suggest that the unfamiliar reader to refer to Dobson (1990) who provided a

nice introductory presentation of how many statistical methods involving a linear predictor can be united through generalized linear models. The GLM can accommodate an entire family of response distributions, thus is responsible for broadening the domain of the standard linear model which often needed mathematically contrived transformations to coerce normality. Thus, statisticians are now faced with a rich modelling mechanism to directly fit the response at hand using a variety of discrete or continuous response distributions and any monotone transformation (link function). For a thorough overview and standard theory of the GLM refer to McCullagh and Nelder (1989) or Fahrmeir and Tutz (1994). The parameter estimates (the β's in η_L) in most cases now must be iterated using an algorithm described below that resembles (iteratively) weighted least squares. We attempt to constructively use the details of the GLM to serve as an impetus for our notation. In many applications, rarely does one have to derive the details given below since tables exist (e.g. Fahrmeir and Tutz, 1994, Table 2.1) that specifies, e.g. the c,d components as well as the details of the canonical link function for common exponential family members (Normal, Bernoulli, Poisson, Gamma, etc.).

Some specifics now follow for the interested reader that will lead up to the important method of scoring algorithm provided below in (4). The GLM requires that the random response vector, $Y_{N\times 1}$, has independent entries Y_i following a distribution in the (canonical form) exponential family and is expressed as, $f(y;\theta,\phi)=\exp[\{y\theta+c(\theta)\}/\phi+d(y)]$, where c,d are known functions. Further, $\theta(\mu)$ is the canonical link function, also referred to the natural parameter of the distribution. It can be shown that $E(Y)=\mu=c'(\theta)$, providing the crucial connection between θ and μ, i.e. $\theta=(c')^{-1}(\mu)=g(\mu)$. In constructing the joint distribution for the Y_i, we find as many θ_i to estimate as there are observations (for $i=1,\ldots,N$). The dimensionality of estimation is reduced from N to p by substituting θ_i with the linear predictor η_L.

Given the set of p explanatory variables, GLMs use the relationship $g(\mu_i)=\eta_L$, where $\mu_i=E(Y_i)$ and g is a monotone, twice differentiable link function with a unique inverse, $h:=g^{-1}$. The estimation of β does not depend on having knowledge of ϕ (which is assumed constant over the observations). The loglikelihood equation (here $\phi=1$, without loss of generality) can be expressed (since $g(\mu_i)=\theta_i=\eta_i$) as

$$l(\beta;X,y)=\sum_{i=1}^{N}\{[y_i\eta_i+c(\eta_i)]+d(y_i)\}. \tag{3}$$

The maximum likelihood estimation of the parameters is typically based on maximizing (3) through the method of scoring iterative equations which simplifies to

$$\hat{\eta}_t=\hat{\alpha}_{t-1}\mathbf{1}+X(X^{\mathrm{T}}\hat{V}_{t-1}X)^{-1}X^{\mathrm{T}}\hat{V}_{t-1}\hat{z}_{t-1}, \tag{4}$$

where, if convergence is attained, the estimated information matrix $\hat{\Phi}=X^{\mathrm{T}}\hat{V}X$, $\hat{V}=\mathrm{diag}(\hat{v}_{ii})=\mathrm{diag}[\{h'(\hat{\eta}_i)\}^2/\mathrm{var}(Y_i)]$, $\hat{z}_i=\hat{\eta}_i+\hat{e}_i/h'(\hat{\eta}_i)$, and $\hat{e}_i=y_i-\hat{\mu}_i$. Here the estimates of V and $\hat{z}$ must be updated at each iteration step until convergence, because they are a function of the iterated $\hat{\eta}_{t-1}$.

2.2. The generalized additive model

We now provide the local scoring algorithm, before providing our own more direct method of estimation of smooth components. For generalized additive models, the components of $g(\mu_i)=\eta_A$ are conventionally fitted using a blend of backfitting and the iterative method of scoring algorithm in Eq. (4). Since each smooth may be estimated in a nonlinear fashion, there exists an additional backfitting cycle within each scoring iterate, specifically:

1. initialize η_A with $g(\bar{Y})$; $f_1=\cdots=f_p=0$;
2. construct and update adjusted dependent variable and weights as above, but using $\hat{\eta}=\alpha+\sum_{j=1}^{p}\hat{f}_j$ and $\hat{\mu}=h(\hat{\eta})$;
3. Cycle $j=1,\ldots,p$: $\hat{f}_j=S_j(z-\sum_{j\neq j}f_j\,|\,x_j,\hat{W})$, where S is the smoothing operator;
4. iterate and cycle until specified convergence is met.

The algorithm cycles through each smooth, trying to improve the fit for the one dimension while holding the other $p-1$ dimensions at their current estimates. Hence, the local scoring algorithm is essentially backfitting with an adjusted dependent variable and weights.

3. Revisiting the B-spline smoother

Among the several popular smoothing techniques, e.g. cubic splines, loess and kernel smoothers, the user is often provided graphical summaries of nonparametric fits. Although nonparametric modeling provides rich exploratory flexibility, it is often not easily used for future prediction. Recall that an objective of our proposed approach is to provide methods useful, in practice, through a compact representation of GAM components useful for prediction. We will see that penalized B-splines are a natural choice. First, some background is provided since not all readers will be familiar with B-splines.

The basic B-spline reference is de Boor (1978), and we find Dierckx (1993) to have a nice presentation. Perhaps the best way to introduce B-splines is by means of an example. The top graph in Fig. 1 shows linear (degree=1) B-splines. On the left are three points, x_1,x_2,x_3. These horizontal positions are called knots, and they are equally spaced in this illustration. At these knots polynomial pieces – in this case linear-join together. At the left of Fig. 1 (top) there is one B-spline of degree 1. This consists of two linear pieces; one from x_1 to x_2, and the other from x_2 to x_3. They are fused together at x_2. The knots are x_1, x_2 and x_3. This B-spline is zero to the left of x_1 and to the right of x_3. To the right of Fig. 1 (top), three more (overlapping) B-splines of degree 1 are provided, each based on 3 knots. Similarly, in Fig. 1 (bottom), B-splines of degree 2 are shown; they each consist of three quadratic pieces joined at two knots. B-splines do overlap to construct the basis; at a given x, two first degree or three second degree B-splines are non-zero.

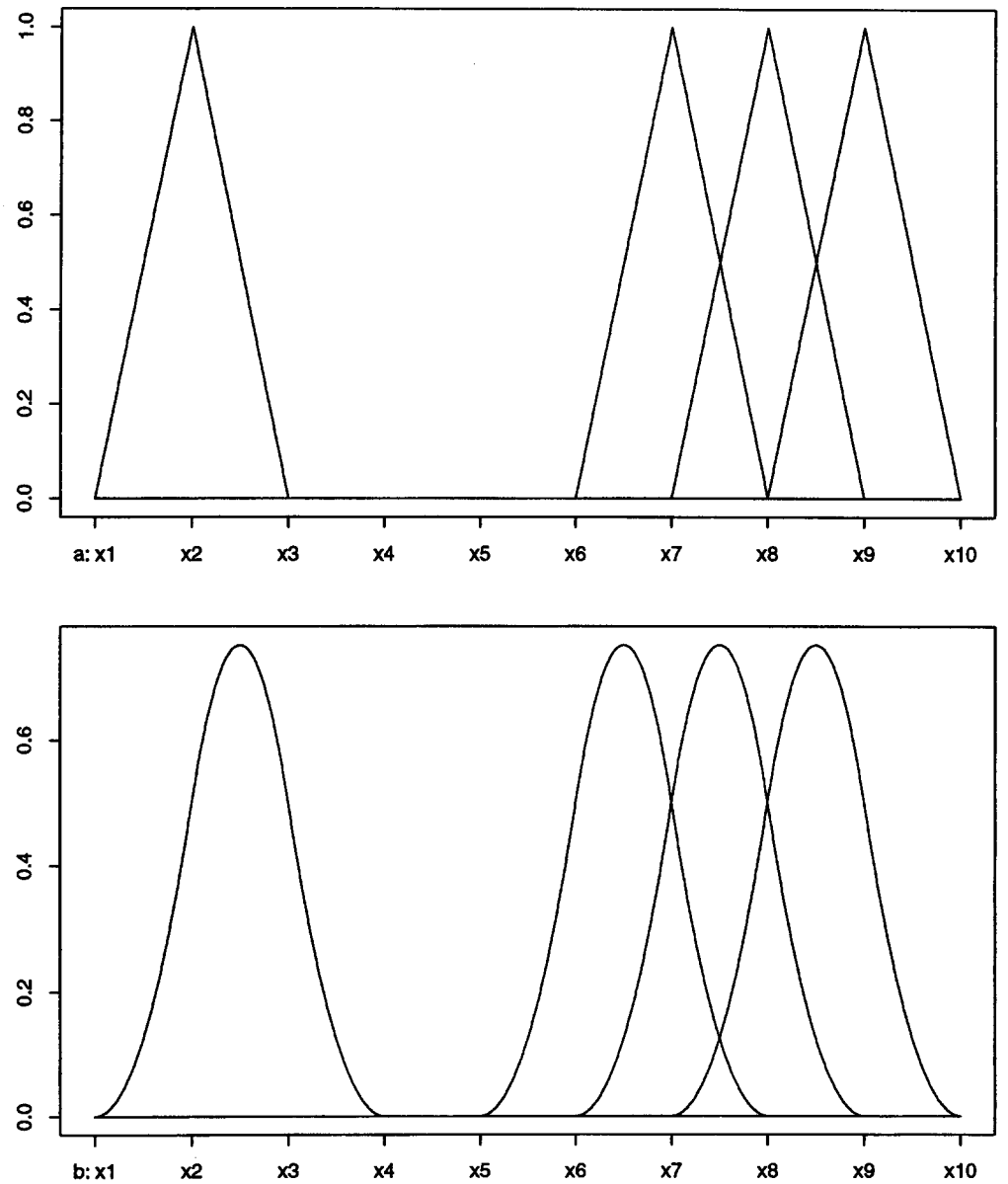

Fig. 1. Illustration of B-spline bases: degree=1 (*top*) and degree=2 (*bottom*).

Additionally, a B-splines smooth basis is independent of the response variable, and only depends on a few pieces of information: (i) the range of the explanatory variable; (ii) the number and position of the knots (we will choose a modest number that are equally spaced); and (iii) the degree of the B-spline (commonly cubic or third degree). More generally, a B-spline of degree q: consists of (i) $q+1$ polynomial pieces of degree q; (ii) these polynomial pieces join at q inner knots; (iii) the points of fusion of the polynomial pieces have continuous derivatives up to

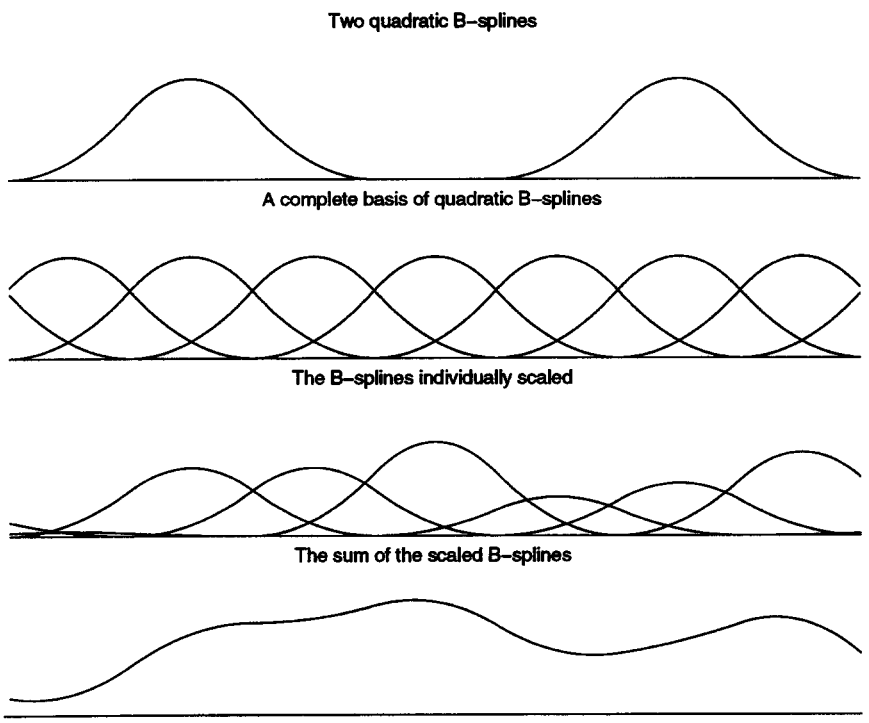

Fig. 2. Illustration of B-spline constructed smooth.

order $q-1$; (iv) a B-spline is positive on a domain spanned by $q+2$ knots (zero elsewhere); (v) except at the boundaries, a B-spline overlaps with $2q$ neighboring polynomial pieces; (vi) at a given x, $q+1$ B-splines are non-zero. Of practical value is a recursive algorithm to compute B-splines of degree q from B-splines of degree $q-1$.

Let the jth GAM component utilizes the B-spline smoother, then $f_j = B_j \boldsymbol{a}_j$. Note that f_j is reduced to a linear sum ($j=1,\ldots,p$), where B_j is the known B-spline matrix (with n_j knots) of dimension $N \times n_j$. Refer to the unknown $\boldsymbol{a}_j$ as the vector of coefficient associated with the B-spline bases. One can also view the $\boldsymbol{a}_j$ as the *amplitudes* of the B-splines. Refer to Fig. 2 which shows how B-splines actually produce smooth curves. The top two portions of Fig. 2 again show the building blocks of quadratic B-spline bases. In the third portion of Fig. 2, each B-spline is multiplied by its corresponding amplitude in $\boldsymbol{a}_j$. The bottom portion of Fig. 2 shows the resulting smooth which is simply the sum of the scaled B-splines. Linear combinations of smooth B-spline bases produce smooth (univariate) curves.

Fig. 2 shows how to construct a smooth curve from B-splines with given coefficients. Most of the time we are confronted with just the reverse problem: find the coefficients that give a smooth fit to the data of interest (scatterplot, time series, histogram). Because smoothness is an inherent property of B-splines, we only have to consider the fit to the data. Let the data be y_i and x_i, with $i=1,\ldots,N$. Also

let $b_{it} = B_t(x_i)$ be the value of the B-spline t at x_i. Further, let $\sum_{t=1}^{n} b_{it}a_t$ be the sum of B-splines. The fit to the data can now be expressed by the sum of squared differences

$$S = \sum_{i=1}^{N} \left(y_i - \sum_{t=1}^{n} b_{it}a_t \right)^2 .$$

The solution for the vector a is given by regression of y on the matrix B. For non-normal data, we apply the framework of the GLM: the linear predictor η is modelled as a sum of B-splines and the iterative method of scoring is used. It is evident that the smoothness of the curve will depend on the number of B-splines used to construct it. This does give one way to regulate smoothness, however, a course way since the number of B-splines is a relatively small integer. A more flexible way would be desirable.

Not only the number of B-splines influences the smoothness of the curve, also the values of the coefficients or amplitudes are important. If these are all nearly equal, then the curve will be very flat. If the amplitudes vary wildly, the curve will show many wiggles. This observation is key to a new way to regulate smoothness: constraining the amplitudes by means of a penalty (E&M).

4. Direct P-GAM likelihood and estimation

Clearly, if the unknown $\boldsymbol{a}_j$ are estimated by standard maximum likelihood, then there is no spatial restriction on the estimates of adjacent amplitudes. Hence, if neighboring estimated coefficients are erratic, then undesirable anomalies can be produced in the fitted smooth. We propose a smoothness requirement of the B-spline parameters or amplitudes. Additionally, one obvious drawback to the B-spline smoother is that the user has to optimize the number and position of knots. To both regularize smoothness and avoid knot selection schemes, P-splines (E&M) recommend using a large number of equally spaced knots (say between 10 & 30), but prevented overfitting by attaching a difference penalty on adjacent B-spline coefficients, ensuring smoothness. Continuous positive smoothing parameters regulate the penalty opposed to the discrete flexibility provided by knot selection.

We consider a P-GAM in the form of $g(\mu) = \boldsymbol{B}\boldsymbol{a} = \eta$, where $\boldsymbol{B} = (\mathbf{1} \| B_1 \| \ldots \| B_p)$ is the $N \times (1 + \sum_{j=1}^{p} n_j)$ regressor matrix, and $\boldsymbol{a} = (\alpha, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_p)'$. P-splines directly fit GAMs through a slightly modified method of scoring algorithm and avoid the call to backfitting. Our proposed P-GAM technique essentially eliminates step 3 (in Section 2.2) of the local scoring algorithm. Overfit B-splines for each GAM component, while penalizing estimation of each vector $\boldsymbol{a}_j$, $j = 1, \ldots, p$ (based on finite differences of adjacent B-splines coefficients), results in maximizing the penalized version of the log-likelihood

$$l^* = l(\boldsymbol{y}; \boldsymbol{a}) - \frac{1}{2} \sum_{j=1}^{p} \lambda_j \boldsymbol{a}_j' P_j \boldsymbol{a}_j, \tag{5}$$

where $\lambda_j \geq 0$, for all $j>0$ are the smoothing parameters, and (for the canonical link function)

$$l(\boldsymbol{y};\boldsymbol{a}) = \sum_{i=1}^{N}\{[y_i\eta_i + c(\eta_i)] + d(y_i)\}.$$

The term η_i is the ith element of $\boldsymbol{Ba}=\eta$.

We now take a closer look at the structure of the penalty given in the subtractend of (5). Define $P_j = (D_j^d)'D_j^d$, where $d = 0,1,2,\ldots$. The matrix D_j^d, of dimension $(n_j - d)\times n_j$ is the building block of the penalty with its (banded) rows consisting of dth-order polynomial contrasts. For fixed component j, this banded matrix corresponds to the matrix representation of the difference operator of order d. For the jth component, we express a $n_j - d$ vector of differences as $D_j^d\boldsymbol{a}_j$, where

$$D_j^0\boldsymbol{a}_j = \boldsymbol{a}_j \quad \text{and} \quad D_j^{d+1}\boldsymbol{a}_j = D_j^1 D_j^d \boldsymbol{a}_j,$$
$$D_j^1\boldsymbol{a}_j = \{a_{jk} - a_{j,k-1}\} \quad k = 2,\ldots,n_j.$$

When $d=0$, we have $P_j = I_{n_j\times n_j}$ which reduces to ridge regression with B-splines. For $d>0$, P_j has a banded structure. In principle, each P_j can have differing orders, and in practice, an order between one and three is usually adequate; $d>4$ is rarely needed. It should be noted that what we denote as *degree* is one less than what de Boor refers to as *order* of the B-spline. Our notation is consistent with S-plus algorithms.

Since P-GAMs only utilize penalized GLM Fisher scoring (and not the backfitting algorithm), maximization of l^* leads to an iterative estimation technique for $\boldsymbol{a}$ given by

$$\hat{\boldsymbol{a}}_{t+1} = (\boldsymbol{B}'\hat{\boldsymbol{W}}_t\boldsymbol{B} + \boldsymbol{P})^{-1}\boldsymbol{B}'\hat{\boldsymbol{W}}_t\hat{\boldsymbol{z}}_t, \tag{6}$$

until specified convergence. We denote $\hat{W} = \text{diag}\{[h'(\hat{\eta}_i)]^2/\text{var}(Y_i)\}$ and $\hat{z}_i = \hat{\eta}_i + (y-\hat{\mu}_i)/h'(\hat{\eta}_i)$ construct the Fisher scoring weight matrix and adjusted dependent vector used in GLM estimation, respectively. The matrix $\boldsymbol{P} = \text{blockdiag}(0,\lambda_1 P_1,\ldots,\lambda_p P_p)$. The zero in the (1,1) position of $\boldsymbol{P}$ corresponds to the intercept term. Note that (6) estimates all p components of the GAM simultaneously.

Upon convergence, $\hat{\mu} = g^{-1}(\boldsymbol{B}\hat{\boldsymbol{a}})$ and $\hat{f}_j = B_j\hat{\boldsymbol{a}}_j$. When the $\lambda_j = 0$, for all $j>0$, the iterative process reduces to GLM estimation with a B-spline basis. It is tempting to try to borrow the asymptotic normal theory of the standard GLM, i.e. if $\lambda_j = 0$, for all $j>0$, $\hat{\text{cov}}(\hat{f}_j) \approx B_j(\boldsymbol{B}'\hat{\boldsymbol{W}}\boldsymbol{B})^{-1}B_j'$. When $\lambda_j>0$, for at least one j, it is natural to try to generalize this result to

$$\hat{\text{cov}}(\hat{f}_j) \approx B_j(\boldsymbol{B}'\hat{\boldsymbol{W}}_t\boldsymbol{B} + \boldsymbol{P})^{-1}\boldsymbol{B}'\hat{\boldsymbol{W}}_t\boldsymbol{B}(\boldsymbol{B}'\hat{\boldsymbol{W}}_t\boldsymbol{B} + \boldsymbol{P})^{-1}B_j'. \tag{7}$$

The diagonal elements of (7) are useful to construct twice standard error bands for P-GAM smooth in the example provided in Section 6. All smoothers have some form of regularization whether it is the number and positions of knots, the degree of the local polynomial, or fraction of points used in the smoothing window. For P-GAMs, these decisions are transferred into the parameter vector λ. One caveat

with any smoothers: the asymptotic distribution of its estimator is not necessarily Normal, or for that matter even assured to be symmetric. We do not pursue this issue further here.

Due to the fact that the columns of each B_j sum to one, $\text{rank}(\boldsymbol{B}) = 1 - p + \sum_{j=1}^{p} n_j$ and $\boldsymbol{B}'\hat{\boldsymbol{W}}_t\boldsymbol{B}$ is inherently singular. We find that using a very small (10^{-6}) ridge penalty on the entire system of equations works well for all the examples that we have encountered. Thus the length of $\boldsymbol{a}$ is gently pushed toward zero, stabilizing estimation. It should be pointed out that deficiency in rank with B-splines can occur outside the P-GAM setting, and under the simplest applications. It is possible for $\text{rank}(B_j) < n_j$. This will occur when the kth column of B_j is the vector $\mathbf{0}$, i.e. when zero observations fall under B_{jk}. This less than full rank (LTFR) condition presents itself if n_j is large relative to N, if the explanatory variable X_j is unevenly dispersed in its own range, or both. Another feature of the P-GAM approach is that LTFR caused by gaps in the domain greater than one B-spline *footprint* will be automatically taken care of by penalized B-splines. We have another argument in favor of using a difference penalty: B-splines presenting the LTFR condition are interpolated by adjacent B-splines.

As pointed out in Section 1, H&T (1990, p. 150) have in fact proposed a penalized modification of the Fisher scoring algorithm. Their idea instead is maximize the log-likelihood, much like the work of O'Sullivan (1986), they consider the penalized likelihood,

$$j(f_1, \ldots, f_p) = l(\eta; \boldsymbol{y}) - \frac{1}{2} \sum_{j=1}^{p} \lambda_j \int \{f_j''(x)\}^2 \, \mathrm{d}x. \tag{8}$$

The expression in (8) $\sum_{j=1}^{p} \lambda_j \int \{f_j''(x)\}^2 \, \mathrm{d}x$ reduces to the form $\sum_{j=1}^{p} \lambda_j \boldsymbol{f}_j' K_j \boldsymbol{f}_j$ if each coordinate function is a cubic spline. The matrices K_j are certain quadratic penalty matrices (see H&T, Section 2.10 or Green and Silverman, 1994, Ch. 1). Note that each K_j is of order N, resulting in potentially an enormous system of equations, and can be numerically complex. Although there is nothing particularly special about the second derivative of $\boldsymbol{f}$, higher-order derivatives are rarely used in the above penalty. Perhaps the reason for the rarity of quintic splines is due to the practical problems of constructing higher-order equations. In using the difference penalty for each component, in a block diagonal fashion, P-GAMs provide a straightforward discrete approximation to quintic splines and higher-order derivative penalties. The connection of the above penalty to the work suggested by O'Sullivan (1986, 1988) is provided in Eilers and Marx (1996).

5. Smoothing parameters, Hat matrix and diagnostics

We choose the vector λ to both balance goodness of fit and complexity of the model. In principle, each component can be optimally regulated with its own penalty parameter, thus preserving the independent nature of the additive components. The choice of the vector of positive penalty parameters λ_j, $j > 0$, is not entirely trivial,

requiring a grid search of some type. For response distributions that have known scale parameter, such as the binomial and Poisson, information criterion (IC) can be easily extended to P-GAMs. Define

$$\mathrm{IC}(\lambda) = \mathrm{dev}(\boldsymbol{y}; \boldsymbol{a}, \lambda) + \delta\,\mathrm{trace}(\hat{\boldsymbol{H}}), \tag{9}$$

where $\hat{\boldsymbol{H}} = \boldsymbol{B}(\boldsymbol{B}'\hat{\boldsymbol{W}}\boldsymbol{B} + \boldsymbol{P})^{-1}\boldsymbol{B}'\hat{\boldsymbol{W}}$ is the converged smoother matrix. When $\delta = 2$ and $\delta = \log(N)$, we have the Akaike (AIC) and the Bayesian (BIC) information criteria, respectively. The trace of $\hat{\boldsymbol{H}}$ is the estimated effective dimension of the entire P-GAM fit, and is more efficiently computed using $\mathrm{trace}\{\hat{\boldsymbol{H}}\} = \mathrm{trace}\{\boldsymbol{B}'\hat{\boldsymbol{W}}\boldsymbol{B}(\boldsymbol{B}'\hat{\boldsymbol{W}}\boldsymbol{B} + \boldsymbol{P})^{-1}\}$. IC is appealing since it only requires a single application of Penalized Fisher scoring for each vector λ in the grid search. However, an estimate variance is needed to compute the deviance with normal, Gamma, or negative binomial response distributions. Because of the regression context with B-splines, linearized likelihood (generalized) cross-validation (GCV) can be done inexpensively and is particularly useful for this latter case. A desirable choice of the vector λ is one which minimizes (G)CV or IC. These details can be found in Hastie and Tibshirani (1990).

5.1. *The effective dimension of the fit*

The second term in (9) is of interest because $T(\lambda) = \mathrm{trace}(\hat{\boldsymbol{H}})$ can be interpreted as the effective dimension of the fitted GAM, with or without a penalty. When each λ_j is set to zero, we have the well-known result that the trace of the hat matrix is equal to the number of regressors. Here the maximum value of $T(\lambda)$ is achieved, which depends on the number and distribution of data points. With at least one nonzero λ_j, $T(\lambda)$ is the complexity of the GAM model. The minimum value of $T(\lambda)$ is met when all $\lambda_j \to \infty$. Due to the fact that the columns of each P-GAM component basis sums to unity, $\min\{T(\lambda)\} = 1 - p + \sum_{j=1}^{p} d_j$ (Eilers and Marx, 1996).

5.2. *Regression diagnostics*

The (linear approximation) regression diagnostics for the logistic model (Pregibon, 1981; Hosmer and Lemeshow, 1989), and their extensions to the GLM can be routinely implemented with the P-GAM approach. Computations are not taxing since we can use the diagonal elements of the effective hat matrix $\hat{\boldsymbol{H}}$ to generate $(i, -i)$ deletion diagnostics. Thus, for example, the effect of each covariate pattern on the fit of the model due to deletion can be assessed through change in Pearson chi-square or deviance, as well as the effect on the change of B-spline amplitudes $\boldsymbol{a}_j$.

5.3. *Connecting to parametric models*

An advantage of using P-splines in GAMs (P-GAMs) over using cubic splines (or other smoothers) as GAM components is that the null space of the difference penalty are polynomials of degree $d_j - 1$. This is one selling point of P-GAMs. In addition to offering the opportunity for local polynomial fitting, the difference penalty of order

d_j has the built-in property that as $\lambda_j \to \infty$, then the jth fitted GAM component will approach a global polynomial of degree $d-1$, provided the degree of the B-spline $q_j \geq d_j$. This result is very useful, in practice, for the following reason: if the measured IC or CV is maintained while imposing a large penalty parameter, say $\lambda_j = 10^4$ on the jth smooth GAM component, then the jth GAM component may be polynomial of degree $d_j - 1$. Thus, by weaking a penalty parameter, a mechanism will be available to provide a continuum between smooth GAM components and parametric polynomial functions. An underlying beauty of the proposed penalized approach is that if a latent polynomial trend exists within a GAM component, then it can be revealed by imposing a large penalty coefficient. We see this feature of P-GAMs in the example to follow.

6. Illustrative example

We revisit the Kyphosis case study presented by Hastie and Tibshirani (1990, Section 10.2) as an application of generalized additive modeling. The response is the binary outcome of the `presence` of `(1)` or the `absence` of `(0)` of postoperative spinal deformity in children. Regressors used in the GAM are: `Age` of patient (months); `Number` of vertebrae levels involved; and the `Start` position of the level of vertebrae involved in surgery. There are $N = 81$ observations; 17 ones and 64 zeros (after omitting the two observations, one corresponding to `Age` $= 243$ and the other to `Number` $= 14$). Hastie and Tibshirani considered several logistic additive models, typically fitting each term with a smoothing spline with a nominal $\mathrm{d}f_j = 3$. See Fig. 3 (top) which displays both the GAM fit using smoothing splines $\mathrm{d}f_j = 3$.

The estimated components are displayed with their associated partial residuals and twice standard error bands. We see from Fig. 3 that the odds of the `presence` of Kyphosis is highest in children about months of `Age` and also when the `Start` position of surgery involves the lower vertebrae. Notice the groupings of negative residuals in the partial residual plots for `Age` and `Start` reminding us of Hastie and Tibshirani's warning of a pure-region effect. The largest positive partial residual in each of the three figures is a potentially valuable observation since it corresponds to a `presence (1)` in a region with very low predicted probability. H&T also provided an analysis of deviance for subset smooth models, subset semi-parametric models, and parametric (polynomial and piecewise linear) models. We do not intend to compare P-GAM methodology in such an exhaustive fashion here.

We fitted P-GAMs by varying the orders of the penalty and considered: $d = 2, 3, 4$. A grid search was performed to obtain optimal (based on AIC in this case) λ(`Age`), λ(`Number`), λ(`Start`) on powers of 10^{γ}, where $\gamma = -4, -3, -2, \ldots, 4$. Thus, the grid search required 3×9^3 GAM fits. The top 10 performers, by order of penalty, are provided in Table 1.

For the purpose of this illustration, we overfit with cubic ($q_j = 3$, for all j) B-splines by partitioning the domain of each explanatory variable into 11 equidistant knots each for `Age` and `Start`, and nine knots for `Number`. The knots include boundary regions. Through penalization, we achieve a parsimonious representation

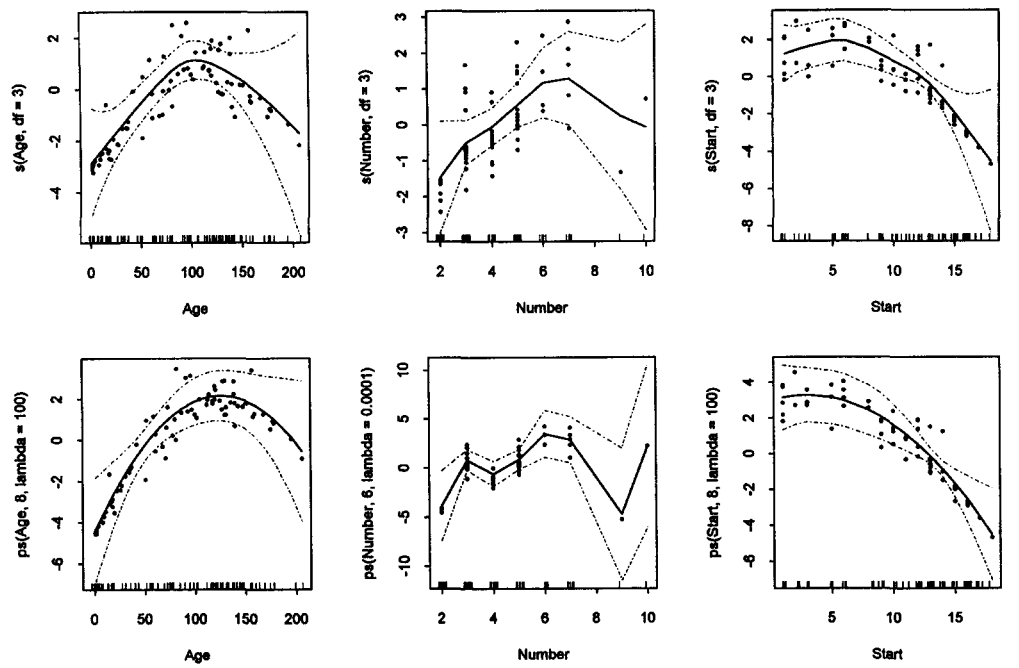

Fig. 3. Estimated additive fit, twice standard error bands & partial residuals: (*top*) Smoothing splines with $df = 3$ for each component, dev = 45.14 on $df_{err} = 71.09$; (*bottom*) cubic P-splines, third-order penalty & optimal λ chosen by AIC, dev = 37.17 on $df_{err} = 66.74$.

Table 1
Top 10 performers based on AIC, by degree of penalty (d), for the mode P-GAM ($X_1 =$ Age, $X_2 =$ Number, $X_3 =$ Start). Note that $\log_{10}(\lambda_j) = \gamma_j$

	$d=2$				$d=3$				$d=4$		
AIC	γ_1	γ_2	γ_3	AIC	γ_1	γ_2	γ_3	AIC	γ_1	γ_2	γ_3
59.249	−1	−4	1	58.173	4	−4	4	60.733	4	−4	−2
59.273	−1	−4	0	58.173	3	−4	4	60.733	3	−4	−2
59.324	−1	−4	2	58.173	4	−4	3	60.733	2	−4	−2
59.334	−1	−4	3	58.174	3	−4	3	60.734	1	−4	−2
59.335	−1	−4	4	58.175	2	−4	4	60.735	0	−4	−2
59.492	0	−4	0	58.176	2	−4	3	60.846	−1	−4	−2
59.493	0	−4	1	58.176	4	−4	2	61.145	4	−4	−3
59.590	0	−4	2	58.176	3	−4	2	61.145	3	−4	−3
59.603	0	−4	3	58.178	2	−4	2	61.145	2	−4	−3
59.604	0	−4	4	58.197	1	−4	4	61.146	1	−4	−3

of the GAM components as desired. Notice that sorting the grid search in Table 1 for the vector λ by AIC is strikingly fruitful. The γ_js surface in meaningful clusters. In the case when $d = 2$, farming out top competitors results in obvious choices for $\lambda(\texttt{Age}) = 10^{-1}, \lambda(\texttt{Number}) = 10^{-4}$ while $\lambda(\texttt{Start})$ can range freely and as high as 10^4 (suggesting linearity). H&T (Section 4.2) suggested that regressor `Start` may be

well fitted using a linear parametric component. This is confirmed in Fig. 4 (top). The effective df for Age, Number, and Start are 3.68, 6.39, and 1.45 respectively for $d=2$.

Similar patterns exist when d is increased to 3, in Fig. 3 (bottom). Note that now AIC is minimized for choices of $\lambda(\texttt{Number})=10^{-4}$, while $\lambda(\texttt{Age})=\lambda(\texttt{Start})=10^4$, indicating that the variables Age and Start may be inherently quadratic in nature. Recall that large penalties yield quadratic fits with a third-order penalty. These values of $\lambda=10^4$ are really overkill, and essentially the same fit is achieved using $\lambda(\texttt{Age})=\lambda(\texttt{Start})=100$ (the ninth rank for $d=3$), with effective df of 2.01, 6.23, and 2.01 respectively. From the competing models with $d=3$, we observed a dramatic switch from $\lambda(\texttt{Number})=10^{-4}$–$10^4$, when AIC reached 61.831. Remember, the nature of AIC is to compromise between goodness of fit and complexity of the model. A (nonsignificant) linear trend will be fitted for Number with $\lambda(\texttt{Number})=10^4$, increasing deviance considerably. However, AIC is compensated for by a significant reduction in the effective degrees of freedom for the Number component; hence another competitive model. Similar explorations with $d=4$ are provided in Fig. 4 (bottom) where again a quadratic trend (overfit cubic due to large $\lambda(\texttt{Age})$). Notice the increased slope for Start after vertebrae number 12; this is where the vertebrae change from thoracic to lumbar (effective d$f=4.63$). Again to avoid such enormous penalties, we present the fourth ranked which reduces $\lambda(\texttt{Age})$ from 10^4 to 10^1 with essentially the same fit (effective d$f=2.97$).

In all cases presented in Table 1, Number chooses $\lambda=10^{-4}$ suggesting either the absence of a polynomial trend, lack of significance, or a degree larger than cubic. All of our P-GAM fits converged in 6–8 iterations of the scoring algorithm (bypassing the call to backfitting). For the sake of interest, we fitted a parametric polynomial model (quadratic in Age & Start, linear in Number) yielding an AIC = 61.455. This illustration endorses that P-GAMs can be an effective approach to find parametric components of a GAM. As mentioned, Hastie and Tibshirani have in fact proposed optional parametric models for this data. They provided a quadratic approximation for Age, as P-GAMs suggest. After further investigation H&T decided to explain the quadratic feature in Start by a piecewise linear fit with a switch point at Start = 12, where the vertebrae change from thoracic to lumbar, which is consistent with our exploratory results using $d=4$.

7. Computational details

We have written the *S-plus* functions necessary to construct P-GAMs, namely, `ps()` and `ps.wam()`. These functions work directly with the existing `gam()` function and are available upon request from the authors. The `ps()` function parallels the `bs()` or B-spline function, except that equidistant knots are constructed, not on the quantiles of the regressor. The `ps()` function has arguments to specify the number of `ps.intervals`, the `degree` of the B-spline (default = 3), `order` of the penalty (default = 3) and the regularization parameter `lambda` (default = 0). *Attributes* are constructed from `ps()`, such as the D^d matrix. This D^d matrix is easily constructed

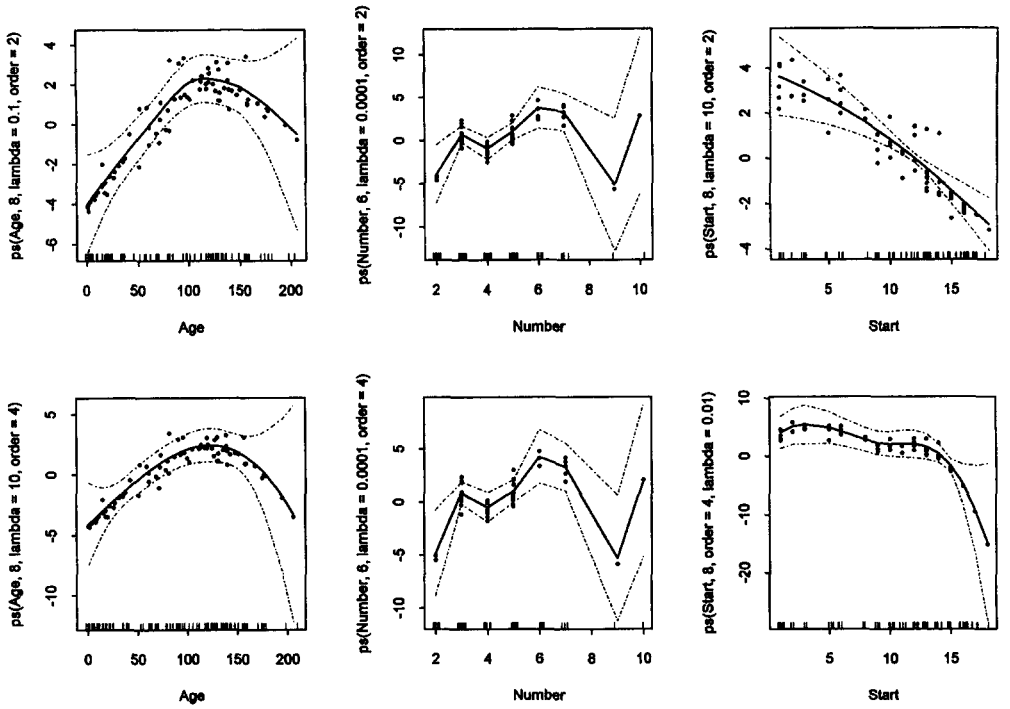

Fig. 4. Estimated additive fit, twice standard error bands and partial residuals: (*top*) P-splines using 2nd order penalty with dev = 37.39 on $df_{\text{err}} = 65.52$; (*bottom*) fourth-order penalty with dev = 34.03 on $df_{\text{err}} = 63.43$. Optimal λ chosen by AIC.

in *S-plus* by d repeated applications of the `difference()` function, e.g. D^1 is simply `difference(diag(n))`. Data augmentation is a useful tool to achieve penalization. The function `ps.wam()` orchestrates the iterative penalized method of scoring. Until specified convergence, a generalized weighted linear model is fitted on the following augmented response, regressors, and weights:

$$\boldsymbol{y}^* = \begin{pmatrix} \boldsymbol{y} \\ \mathbf{0} \end{pmatrix}, \qquad \boldsymbol{B}^* = \begin{pmatrix} \boldsymbol{B} \\ \boldsymbol{D}^* \end{pmatrix}, \qquad \boldsymbol{w}^* = \begin{pmatrix} \boldsymbol{w} \\ \mathbf{1} \end{pmatrix}, \tag{10}$$

where $\mathbf{1}$, $\mathbf{0}$ are vectors of dimension $\sum_{j=1}^{p}(n_j - d_j)$. We define the matrix,

$$\boldsymbol{D}^* = \text{blockdiag}(0, \sqrt{\lambda_1} D_1^d, \ldots, \sqrt{\lambda_p} D_p^d).$$

Constructing P-GAMs in *S-plus* has the same syntax as GAMs, e.g.

```
pgam1 <- gam(Kyphosis ~ ps(Age, ps.intervals=10, degree=3, order=3,
lambda=1)  + Number + ps(Start, ps.intervals=10, lambda=10), data=
kyphosis, family=binomial)
```

which fits P-splines to `Age` and `Start` and a linear fit to `Number`. Analysis of deviance is straight forward; the degrees of freedom are approximated by the trace of the smoothing matrix. Other design or class variable can be routinely handled.

8. Conclusions

We hope that this paper can be used as a platform to provide researchers with a practical application for model building. P-spline smoothers are originally presented in Eilers and Marx (1996). These smoothers have a wide range of thought provoking properties and contemporary applications, including: scatterplot smoothing, generalized linear models and density estimation. P-GAMs extend the work of P-splines into the p-dimensional additive model. As we mentioned in previous work, penalized likelihood is a subject of growing popularity. However, penalties are routinely defined in terms of the square of the second derivative of the fitted curve. For example, refer to the book by Green and Silverman (1994). Applications to higher-order derivatives, e.g. quintic, are rare. Corresponding algorithms are rarer. P-splines provide a mechanism to approximate higher-order penalties. With little effort, B-splines generalize to any degree and penalties to any order. P-splines provide a continuum to higher-order parametric polynomial models. If n knots are used for each dimension of the P-GAM, then $n \times p$ parameters summarize the fit, compared to nonlinear counterparts which can have a summary as large as $N \times p$ fitted values and corresponding second derivatives. Lastly, it should be evident that iterating with P-splines is just Gauss–Seidel iteration. If convergence is met, then it gives the same solution as solving for the *large* GLM model directly; thus there is no need to simulate.

Acknowledgements

The authors are grateful to Trevor Hastie for sharing his lucid understanding of the subject and guidance to link our *S-plus* code to the existing `gam()` function. We also extend our thanks to the Editor Stanley Azen, an anonymous associate editor, and two anonymous referees for their thorough and constructive review.

References

de Boor, C., 1978. A Practical Guide to Splines. Springer, Berlin.

Buja, A., Hastie, T., Tibshirani, R., 1989. Linear smoothers and additive models. Ann Statist. 17, 453–555.

Dierckx, P., 1993. Curve and Surface Fitting with Splines. Clarendon Press, Oxford.

Dobson, A.J., 1990. An Introduction to Generalized Linear Models, Chapman & Hall, London.

Eilers, P.H.C., Marx, B.D., 1996. Flexible smoothing using B-splines and penalized likelihood (with comments and rejoinder). Statist. Sci. 11 (2), 89–121.

Fahrmeir, L., Tutz, G., 1994. Multivariate Statistical Modelling based on Generalized Linear Models. Springer, Berlin.

Green, P.J., Silverman, B.W., 1994. Nonparametric Regression and Generalized Linear Models. Chapman & Hall, London.

Hastie, T., Tibshirani, R., 1986. Generalized additive models. Statist. Sci., 1, 297–318.

Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman & Hall, New York.

McCullagh, P., Nelder, J.A., 1989. Generalized linear models, 2nd ed. Chapman & Hall, London.

Hosmer, D.W., Lemeshow, S., 1989. Applied Logistic Regression. Wiley, New York.

Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized linear models. J. Roy. Statist. Soc. A 135, 370–384.

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.

Pregibon, D., 1981. Logistic regression diagnostics. Ann. Statist. 9, 705–724.

Stone, C.J., 1986. The dimensionality reduction principle for generalized additive models. Ann. Statist. 14, 590–610.

Stone, C.J., 1994. The use of polynomial splines and their tensor products in multivariate function estimation. Ann. Statist. 22, 118–184.

Stone, C.J., Koo, K.Y., 1985. Additive splines in statistics. Proc. Statistical Computing Section of the American Statistical Association, pp. 45–47.