Generalized Linear Additive Smooth Structures

Paul H. C. Eilers and Brian D. Marx

This article proposes a practical modeling approach that can accommodate a rich variety of predictors, united in a generalized linear model (GLM) setting. In addition to the usual ANOVA-type or covariate *linear* (L) predictors, we consider modeling any combination of *smooth additive* (G) components, *varying coefficient* (V) components, and (discrete representations of) *signal* (S) components. We assume that G is, and the coefficients of V and S are, inherently smooth—projecting each of these onto B-spline bases using a modest number of equally spaced knots. Enough knots are used to ensure more flexibility than needed; further smoothness is achieved through a difference penalty on adjacent B-spline coefficients (P-splines). This linear re-expression allows all of the parameters associated with these components to be estimated simultaneously in one large GLM through penalized likelihood. Thus, we have the advantage of avoiding both the backfitting algorithm and complex knot selection schemes. We regulate the flexibility of each component through a separate penalty parameter that is optimally chosen based on cross-validation or an information criterion.

Key Words: Generalized additive models; Multivariate calibration; P-splines; Signal regression; Varying-coefficient models.

1. INTRODUCTION

Linear structures occur at many places in statistical models. We encounter them in the classical, as well as in the generalized linear model (GLM, Nelder and Wedderburn 1972). In the last decade or so, we have seen many extensions: the generalized additive model (GAM, Hastie and Tibshirani 1990), the varying-coefficient model (VCM, Hastie and Tibshirani 1993) and penalized regression on signals (Hastie and Mallows 1993; Marx and Eilers 1999). This article provides a practical mechanism to simultaneously combine all of these models into one generalized additive structure, avoiding backfitting and with excellent

Paul H. C. Eilers is Assistant Professor, Department of Medical Statistics, Leiden University Medical Center, 2300 RC Leiden, The Netherlands (E-mail: p.eilers@lumc.nl). Brian D. Marx is Professor, Department of Experimental Statistics, Louisiana State University, Baton Rouge, LA 70803 (E-mail: bmarx@lsu.edu).

Journal of Computational and Graphical Statistics, Volume 11, Number 4, Pages 758–783
DOI: 10.1198/106186002844

control over smoothness. Our approach is based on penalized B-splines or P-splines (Eilers and Marx 1996). Smooth components are estimated with B-splines (purposely overfitting using a modest number of equally spaced knots) combined with a difference penalty on their coefficients, to reduce unnecessary flexibility. A nice feature of this approach is that smooth components are coerced into linear structures. We invented the acronym GLASS: generalized linear additive smooth structures.

Section 2 discusses the components and illustrates the richness of GLASS. We then give an overview of P-splines in Section 3. Using P-splines, each model component can be described by penalized regression on a moderately sized B-spline basis. These components can be mixed and matched like *LEGO*$^{\text{TM}}$ construction blocks; details are provided in Section 4. The building blocks are the regressors in a large penalized GLM; details of the estimation procedure are given in Section 5. Choosing the right amount of smoothing (optimal penalty parameters), using an information criterion or cross-validation, is discussed in Section 6. Section 7 contains several illustrative examples, while Section 8 discusses computational details. We close with a brief discussion.

2. COMPONENTS OF GLASS

This section gives an informal description of the model components. We start with the familiar generalized linear regression model (of which linear regression is a special case). Let the data be (y, X), where y $(m \times 1)$ is the response variable with independent entries from a particular member of the exponential family of distributions, and X $(m \times p)$ contains explanatory variables, which can contain a constant vector to account for a possible intercept. The model is:

$$E(Y) = \mu = h(\eta), \tag{2.1}$$

where μ is the expected value of the response, and $h(\cdot)$ is an inverse (monotone and twice differentiable) link function. For the standard GLM, we have a linear predictor in the form:

$$\eta_L = X_L \alpha_L. \tag{2.2}$$

We add the subscript L here to discern this X from other matrices that will be introduced below. The unfamiliar reader can reference McCullagh and Nelder (1989), who also presented efficient algorithms for estimating α. GLMs are standard in most modern statistical software.

Whereas the GLM constructs the linear predictor as a linear combination of the columns of X, a generalized additive model (GAM) assumes smooth nonlinear functions (Hastie and Tibshirani 1990):

$$\eta_G = f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p), \tag{2.3}$$

where x_j indicates column j of X, $j = 1, \ldots, p$. A GAM can be estimated by backfitting: updating an approximate solution by repeated weighted smoothing of the scatterplot formed by x_j and the partial residuals $y - \sum_k \tilde{f}_k(x_j) + \tilde{f}_j(x_j)$, where the tilde indicates an approximate solution. In principle any type of smoother can be used. Marx and Eilers (1998) showed how to write a GAM as a large penalized GLM, using P-splines, eliminating backfitting. GAMs are standard in some statistical software, perhaps most notably in S-Plus.

In contrast to the GLM, where the regression coefficients in α are assumed to be constant, the varying coefficient model (VCM) was proposed by Hastie and Tibshirani (1993). This model accommodates situations in which one or more of the coefficients are allowed to vary smoothly (interact) over an additional m-vector t, for example, time or space. The analog for (2.2) now is

$$\eta_V = f_1(t)x_1(t) + f_2(t)x_2(t) + \cdots + f_p(t)x_p(t), \tag{2.4}$$

where we explicitly indicate the dependence of each column of X on one specific t. In principle there can be a unique t for each column of X, yielding $T_V = (t_1, \ldots, t_p)$. Hastie and Tibshirani (1993) described an algorithm for estimating the smooth curves that make up the VCM, through backfitting and minimization of a criterion that penalizes smoothness much like Reinsch (1967). Their S-Plus software is available at StatLib (http://lib.stat.cmu.edu).

Finally, we have signal regression. Now the rows of $X_{(m\times n)}$ are (equidistantly sampled) signals: for example, time series, discrete (optical) spectra or histograms. The number of columns (n) of X generally is much larger (often 10 or 100 times) than the number of observations (m). One column of X represents the same point in time (for time series), the same wavelength (for spectra) or the same histogram bin. An n-vector t indexes the particular signal associated with the ordered columns. Usually the regression problem is singular; in the chemometric literature it is known as multivariate calibration. One possible solution was presented by Marx and Eilers (1999), extending a proposal by Hastie and Mallows (1993). The assumption is that detailed knowledge of a large vector of regression coefficients is impossible to obtain, given the relatively small number of observations. As the columns of X are ordered, it may be reasonable to assume that neighboring coefficients have similar values, that is, it is a smooth vector. The η in (2.1) now can be viewed as

$$\eta_S = X_S f_S, \tag{2.5}$$

where we assume f_S to be a smooth function of t. Marx and Eilers (1999) presented a solution, based on P-splines, called penalized signal regression (PSR). Software is available on www.stat.lsu.edu/bmarx.

To keep the notation simple, we introduced only one signal regressor. But it will become clear below (Section 5) that multiple signals, each of possibly differing dimension and domain, can be accommodated as well.

We are interested in a structure where any or all of the four components presented above can be combined at will. The regressor information can be summarized as follows:

$$X = [X_L|X_G|X_V|X_S], \tag{2.6}$$

where

- the $m \times p_L$ matrix X_L contains ANOVA-type or covariate regressors that are to be treated in the standard GLM way;
- the columns of the $m \times p_G$ matrix X_G are each candidates for smooth (GAM) additive modeling;
- the columns of the $m \times p_V$ matrix X_V are regressors for which we want to find coefficients along corresponding indexing column vectors $t \in T_V$;
- the rows of the $m \times n$ matrix X_S contain sampled signals, indexed by a vector t.

We assume that X_L at least contains a vector of ones when an intercept is to be included. Although presented in a general form, a composite additive predictor can be constructed as follows

$$\eta = \eta_L + \eta_G + \eta_V + \eta_S. \tag{2.7}$$

We will show, in the next two sections, how using P-splines can coerce the above η into a *linear* predictor and further how to construct building blocks that correspond to three of the four models (GAM, VCM, PSR) using P-splines (the GLM needs no special treatment). As a service to the reader, who might not be familiar with P-splines, we first present an overview in the next section. A full account can be found in the article by Eilers and Marx (1996) and its accompanying discussion.

3. A CRASH COURSE ON P-SPLINES

A B-spline is a bell-shaped curves resembling a Gaussian density. In contrast to the latter, a B-spline has only local support, as it is constructed from smoothly joining polynomial segments, as shown in Figure 1. The positions on the horizontal axis where the segments come together are called the knots. We will use only equidistant knots, but B-splines can be defined for an arbitrary grid of knots. Many details and algorithms can be found in the books by de Boor (1978) and Dierckx (1993). Usually a B-spline is said to have order $q + 1$ if the polynomial segments have degree q. In contrast to this custom, we will call it a B-spline of degree q; the reason being that we will also introduce difference penalties of certain orders and wish to avoid confusion. For example, a cubic (quadratic) B-spline has degree 3 (2) and consists of cubic (quadratic) polynomial segments.

When computing a set of B-splines, each shifted by one knot distance, we get a basis of local functions that is well suited for smoothing of a scatterplot of points (x_i, y_i), $i = 1, \ldots, m$. If $b_{ij} = B_j(x_i)$, $j = 1, \ldots, K$ $(< m)$ indicates the value of the jth B-spline at x_i, and $B = [b_{ij}]$, we minimize

$$S = |y - B\alpha|^2, \tag{3.1}$$

with the explicit solution

$$\hat{\alpha} = (B'B)^{-1}B'y. \tag{3.2}$$

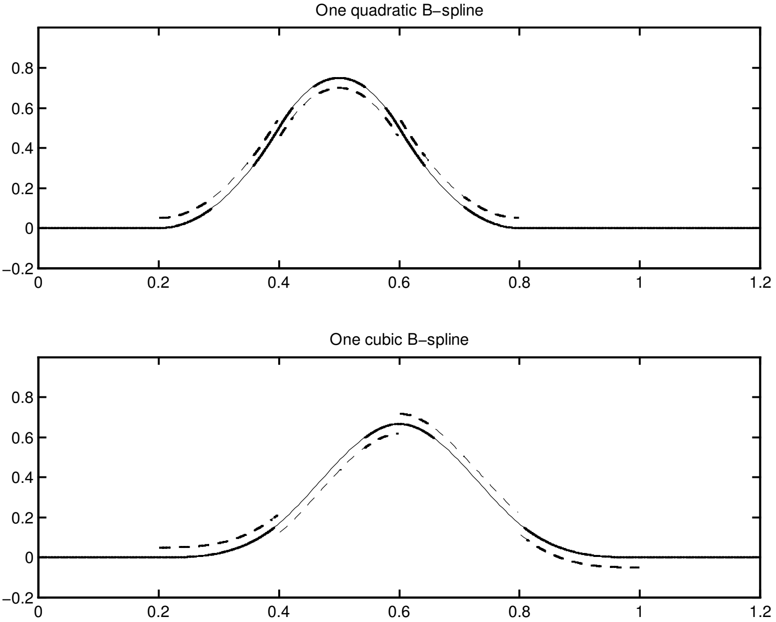

Figure 1. One quadratic (top panel) and one cubic (bottom panel) B-spline based on knots at 0.2, 0.4, 0.6, 0.8, and 1. To emphasize the construction, the polynomial pieces have also been drawn as broken lines, alternately shifted vertically by 0.05 and −0.05.

Given $\hat{\alpha}$, the estimated point on the curve at any (new) x is $\sum_j B_j(x)\hat{\alpha}_j$. This way smoothing is reduced to linear regression. The number of B-splines is the degree plus the number of segments between the left-most and right-most knot.

We emphasize that we consider only equidistant spacing of knots. Some details are now clarified. If x is the vector that gives the values for which the B-splines are to be computed, (rounded values of) $\min(x)$ and $\max(x)$ are used as the boundaries of the domain of x. The domain is divided in equal divisions. The number of B-splines is the number of divisions plus the degree of the splines. For example, with two divisions and cubic B-splines, there are three "visible" knots, two at the boundaries and one in the middle of the domain. Three "invisible" knots are constructed by the recursive algorithm that compute the splines.

The amount of smoothing is determined by the size of the B-spline basis and thus implicitly by the number of knots. The smaller the number of knots, the smoother the curve. This is shown in Figure 2 for a well-known dataset (Härdle 1990). Using leave-one out cross-validation we find that 17 (equally spaced) B-splines are optimal. Fits with more and less B-splines are shown for comparison. Note that cross-validation can be done very quickly, since

$$y_i - \hat{y}_{-i} = (y_i - \hat{y}_i)/(1 - h_{ii}), \tag{3.3}$$

where h_{ii} is the ith element on the diagonal on the "hat" matrix H, $\hat{y} = B(B'B)^{-1}B'y = Hy$, and $\hat{y}_{-i}$ is the fitted value for y_i that would be obtained if the model were estimated with y_i left out. It follows that $h_{ii} = b_i'(B'B)^{-1}b_i$, where b_i' indicates the ith row of B.

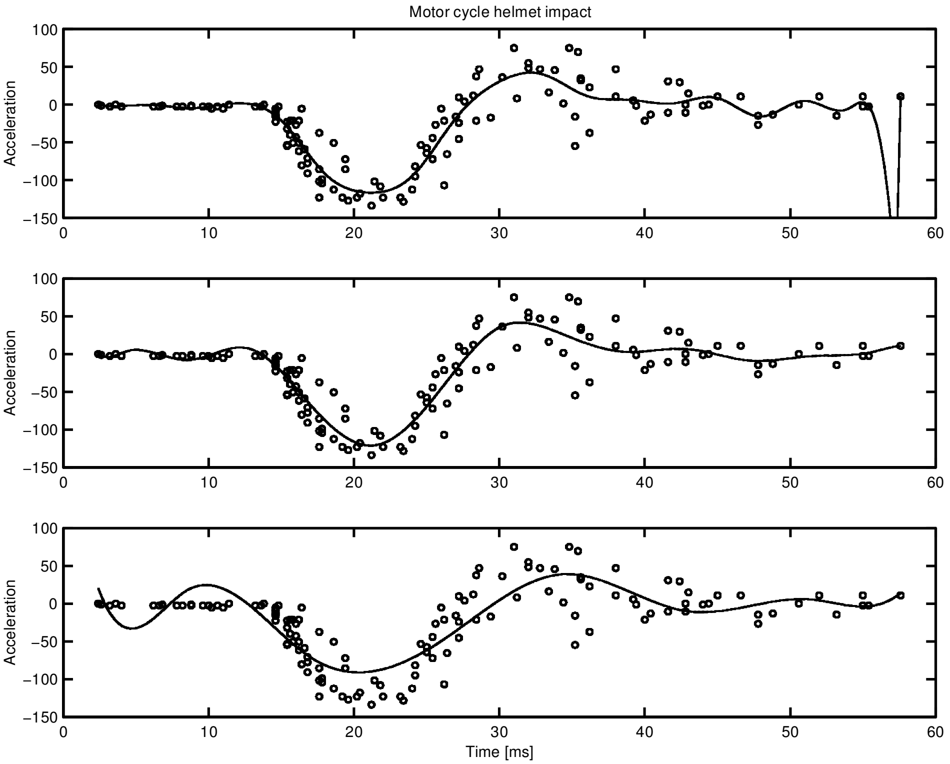

Figure 2. Fitting a smooth curve to a scatterplot using equally spaced cubic B-splines. In the middle panel the number of B-splines is optimal according to least-squares cross-validation. The upper (lower) panel illustrates overfitting (underfitting).

Hence the diagonal elements of H and the cross-validation residuals can be computed with little additional work. More details on optimal smoothing can be found in Section 6.

The upper panel of Figure 3 shows the central part of the dataset and a purposely overfitted curve, based on too many B-splines. We see that the heights of neighboring B-splines—which are proportional to the corresponding coefficients—differ strongly. We would get a smoother result if, in some way, we could force the coefficients to vary more smoothly, as illustrated in the lower panel of Figure 3. This is exactly the purpose of an additional penalty, weighted by a positive regularization parameter λ, that we attach to (3.1):

$$S^* = |y - B\alpha|^2 + \lambda |D_d \alpha|^2. \tag{3.4}$$

The matrix D constructs dth order differences of α:

$$D_d \alpha = \Delta^d \alpha. \tag{3.5}$$

The first difference of α, $\Delta^1 \alpha$ is the vector with elements $\alpha_{j+1} - \alpha_j$, for $j = 1 \ldots K - 1$. By repeating this computation on $\Delta\alpha$, we arrive at higher differences like $\Delta^2\alpha$ and $\Delta^3\alpha$. The $(n-1) \times n$ matrix D_1 is sparse, with $d_{j,j} = -1$ and $d_{j,j+1} = 1$ and all other elements

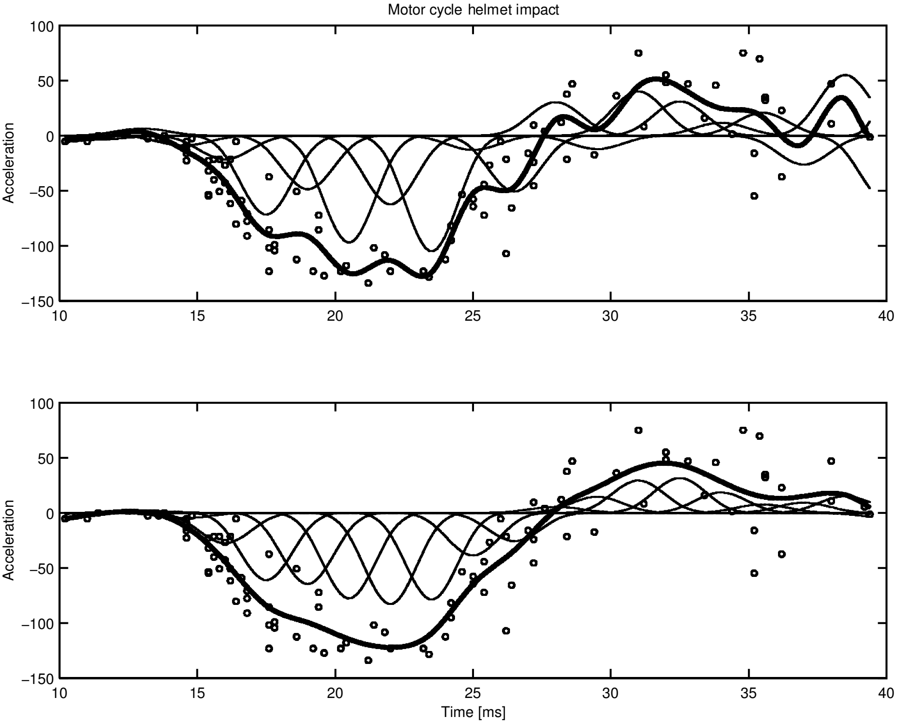

Figure 3. The P-spline principle illustrated using central portion of the motorcycle data. Scatterplot of data with individual B-splines (thin lines) and the corresponding fitted curve (thick line). Top panel: due to overfitting the heights of neighboring B-splines differ strongly. Bottom panel: a difference penalty forces the heights to change more gradually, yielding a smoother curve.

zero. Examples of D_1 and D_2 of small dimension look like

$$D_1 = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}; \quad D_2 = \begin{bmatrix} 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \end{bmatrix}.$$

Actually, the number of equally spaced knots does not matter much provided that enough are chosen to ensure more flexibility than needed: the penalty gives continuous control for further smoothing. The solution of (3.4) is

$$\hat{\alpha} = (B'B + \lambda D_d' D_d)^{-1} B'y, \tag{3.6}$$

and the hat matrix is now given by

$$H = B(B'B + \lambda D_d' D_d)^{-1} B'. \tag{3.7}$$

Cross-validation to find an optimal value of λ is just as fast as in the nonpenalized case.

Since P-splines are based on (penalized) linear regression, it is quite natural to transplant the methodology to the generalized linear model. Thus non-normal responses, such as

Poisson counts or binary data, can be smoothly modeled with $E(Y) = \mu = h(B\alpha)$. Now we must maximize a penalized log-likelihood function,

$$l^* = l(\alpha; B, y) - \frac{1}{2}\lambda|D_d\alpha|^2, \tag{3.8}$$

which is equivalent to minimizing (3.4) for normal responses. The $\frac{1}{2}$ is a trick to eliminate the 2 after differentiating. The solution for (3.8) can be achieved through iterative weighted regression:

$$\hat{\alpha}_{t+1} = (B'\hat{W}_tB + \lambda D_d'D_d)^{-1}B'\hat{W}_t\hat{z}_t,$$

where $\hat{z} = (y - \hat{\mu})/h'(\hat{\eta}) + B\hat{\alpha}$ is the working dependent variable and the weight matrix is $\hat{W} = \text{diag}[\{h'(\hat{\eta})\}^2/\text{var}(Y)]$.

To illustrate P-splines with a binary response and a logit link function, we use the *kyphosis* dataset in S-Plus. These data were also modeled at some length in a case study presented by Hastie and Tibshirani (1990, sec. 10.2). The response is the binary outcome of *presence* (1) or *absence* (0) of postoperative spinal deformity in children. The regressor used here is *age* of patient (months). There are $m = 81$ observations; 17 ones and 64 zeros. Figure 4 displays the fitted smooth probability of *kyphosis* as a function of *age*, for varying λ, using 13 equally spaced knots, cubic B-splines and a second-order difference penalty. The twice standard error bands for the predicted probability are also displayed; construction of these bands is discussed in further detail in Section 5. We see some evidence of increased probability of kyphosis near 100 months age. For certain GLMs, it is easier to monitor and minimize an information criterion, like AIC. Figure 4 suggests $\lambda \approx 10$, which yields a smooth fit having effective df of about 2 and a deviance of approximately 74 on 79 residual df. Information criteria and effective df are discussed in greater detail in Section 6.

As also seen in Figure 4, larger λ lead to smoother results, even with many B-splines in the basis. One can show that the fitted curve approaches a polynomial of degree $d - 1$ as λ gets large. P-splines can be interpreted as a projection onto a relatively smooth subspace with additional smoothing caused by the penalty, as with a smoothing spline. O'Sullivan (1986) did this literally, using the integral of the square of the second (or a higher) derivative of the fitted curve as the penalty. In practice, our penalty and O'Sullivan's penalty give essentially the same results, but for the latter it takes careful programming to construct the equivalent of $D_d'D_d$, especially for $d > 2$; yet for the difference penalty approach, higher orders are a simple mechanical procedure.

The penalty also solves problems with ill-conditioning that may occur with non-penalized B-splines. In some datasets one encounters a rather uneven distribution of data points on the x-axis. Some of the B-spline coefficients then will have large coefficients and will be very sensitive to small changes in the data. This can especially occur at the boundaries of the domain of x. It is visible in the rightmost portion of the top panel of Figure 2. In extreme cases some coefficients may even be inestimable, because of missing data on the support of some of the B-splines. One may also encounter rather erratic behavior of leave-one-out cross-validation.

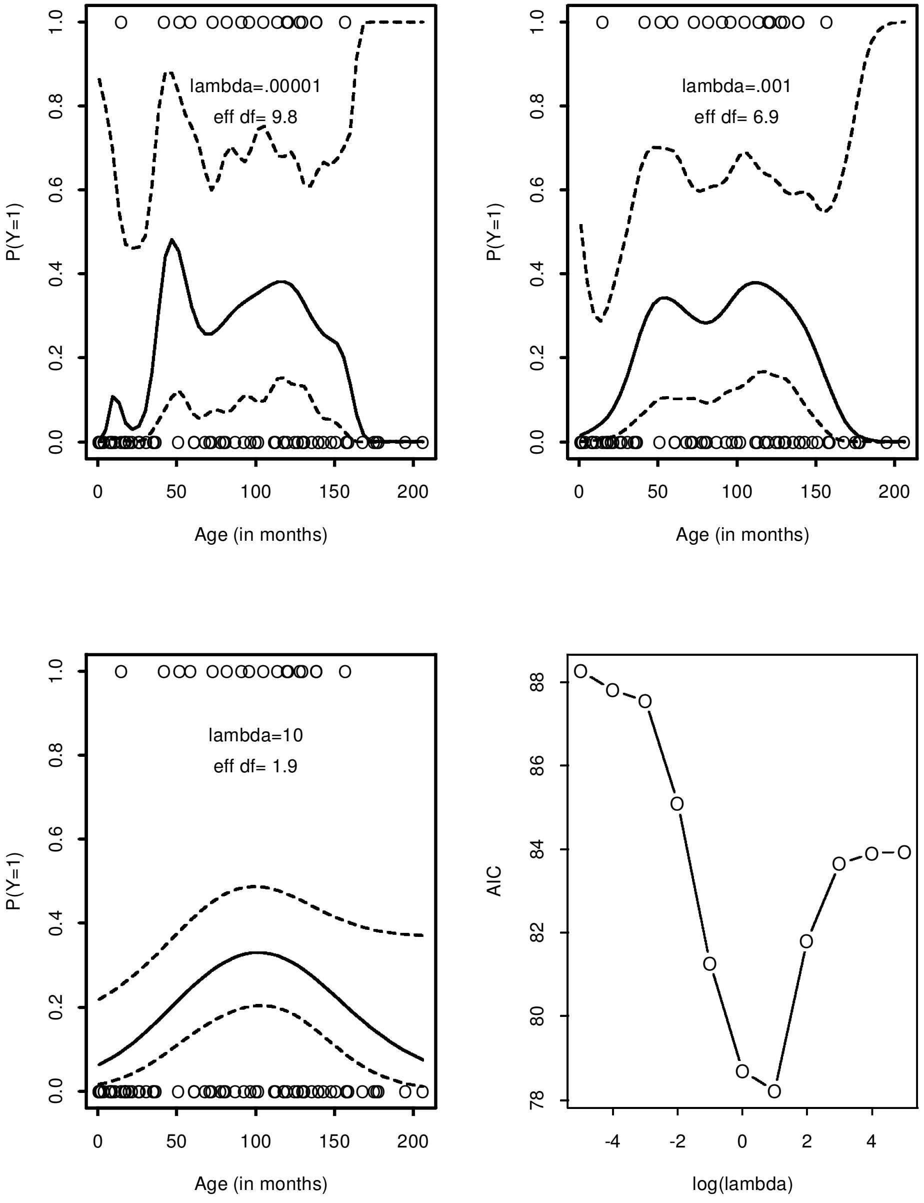

Figure 4. Smooth estimated probability of kyphosis as a function of age (with twice standard error bands). Three panels illustrate varying amount of smoothness using a second-order difference penalty. The bottom-right panel displays AIC with increasing log(λ).

These problems have led to the practical advice to use unequally spaced knots, determined by quantiles of x. Unequally spaced knots would ruin the simplicity of P-splines and, because of the penalty, there is no need for them. Generally we also see no need for choosing (new) unequally spaced knots after finding a curve. Of course, for some datasets, variable smoothness might be strongly indicated. Ruppert and Carroll (2000) gave an example, and a solution: a spatially varying penalty implemented as a diagonal weighting matrix W, such that $D_d'D_d$ is replaced by $D_d'WD_d$. In this article we will not make use of this extension.

To summarize, our practical recipe is:

- use a quadratic or cubic B-splines basis of moderate size (10 to 50), but large enough to overfit the data;
- use a second- or third-order difference penalty;
- optimize λ by minimizing a cross-validation measure or an information criterion, for example, AIC; and
- report the results as the vector of coefficients (and the parameters of the B-spline basis), the fitted curve can be reconstructed from this information.

B-splines are easily computed. S-Plus has a function to compute B-splines for arbitrary knots. Eilers and Marx (1996) presented a short Matlab function to compute B-splines with equally spaced knots. In S-Plus or Matlab, one can just apply the function `diff()` d times to the identity matrix to get D_d.

In the rejoinder to the discussion of Eilers and Marx (1996), there is a "consumer score card" that compares P-splines to many other types of smoothers. Their good properties make them very suitable to construct the building blocks of GLASS, which we describe in the next section.

4. B-SPLINE BUILDING BLOCKS

In Section 2, we introduced the models we consider as components of GLASS, summarized in (2.6) and the description following it. In this section we show how to translate each component into building blocks, using B-splines. The building blocks are combined to form one large GLM regressor matrix. In Section 5 we will add difference penalties in P-spline style.

One main point that we emphasize here is that smooth functions are constructed in a simple way: $f = B\alpha$, whether its for a GAM component or smooth coefficient vectors in varying coefficient regression or signal regression components. In fact, we will see below that it is convenient to view the building blocks in a more modular fashion by further generalizing the notation to $Mf = MB\alpha = U\alpha$. The B-splines are constructed on a domain, generally denoted as t, and M is a modifying matrix, each depending on the type of component. In this way, each component has a pair (t, M). Thus (2.7) can be viewed in a compact linear fashion through

$$\eta = \eta_L + \eta_G + \eta_V + \eta_S = \eta_L + U^\star\alpha^\star, \tag{4.1}$$

where $U^\star = [U_1|, \ldots, |U_p]$ and $\alpha^\star = [\alpha_1'|, \ldots, |\alpha_p']'$ for p mixed and matched GAM, varying and signal components.

4.1 The Linear Portion

The standard generalized linear model portion is denoted: $\eta_L = X_L \alpha_L$, where X_L is a standard $(m \times p_L)$ matrix of ANOVA-type or covariate regressors and α_L is the associated coefficient vector. The pair (t, M) is null.

4.2 The Smooth Additive Portion

Now we consider the jth column $(j = 1, \ldots, p_G)$ of the component X_G in (2.6) and thus the corresponding portion of (4.1). Recall that $f_j(\cdot)$ is modeled by K_j equally spaced B-splines with coefficients $\alpha_j = [a_{jk}]$. More explicitly, we have:

$$f_j(x_{ij}) = \sum_{k=1}^{K_j} B_k^j(x_{ij}) a_{jk} = \sum_{k=1}^{K_j} b_{ik}^j a_{jk}. \tag{4.2}$$

Here the pair is $(t = x_j, M = I)$. In general, the B-spline bases can be different in their support, as well as the degree of the B-splines. Thus each GAM term in (4.1) is simply $U\alpha = B\alpha$.

4.3 The Varying Coefficient Portion

Consider the jth column, x_j of X_V (the $m \times p_V$ matrix of candidates for varying coefficient modeling) and the corresponding indexing m-vector t. Hence we have $x_j(t)$, and want to find a smooth coefficient vector $f_j(t)$. Thus, we re-express each varying coefficient term in (4.1) with $U\alpha = \text{diag}\{x_j\}B\alpha$, where the pair is $(t, M = \text{diag}\{x_j\})$.

4.4 The Signal Portion

In (2.5) we defined $\eta_S = X_S f(t)$, where X_S is a $m \times n_S$ matrix of signal regressors, $f(t)$ is the smooth signal coefficient vector, and t is the corresponding row index of length n_S. We reduce the signal component in (4.1) to $U\alpha = X_S B\alpha$, where the pair is $(t_{n\times 1}, M = X_S)$. As with the varying coefficient components, this step can dramatically reduce the dimensionality of the problem. For example, in a typical chemometric application (see Section 7) optical spectra are given for nearly 500 wavelengths, while a B-spline basis of size 20 or less is often sufficiently flexible. The U associated with the signal has dimension $m \times K_S$, is full column rank and can be computed in advance. Of course X_S is not discarded, but just is not needed during parameter estimation. Multiple signals would lead to multiple U matrices; Marx, Eilers, and Auer (2002) modeled multiple signals for a large scale medical study on tumors.

5. GLASS REGRESSORS AND PROPOSED ESTIMATION

Despite the formidable initial appearance of the GLASS, the use of B-splines as described in the previous section makes it easy to model smooth curves by regression while curtailing potentially enormous dimension. In each case, B-spline bases are constructed and used to (linearly) reduce the problem to a moderately large sized standard GLM.

One can imagine arbitrary combinations of component submatrices, $X = [X_L|X_G|X_V|X_S]$. From X, we see that the GLM model matrix can be constructed:

$$R = [X_L|U^\star] \tag{5.1}$$

of dimension $m \times c$ with $c = p_L + \sum_{j=1}^{p_G} K_{Gj} + \sum_{j=1}^{p_V} K_{Vj} + K_S$. For the generalized linear model setting, we have

$$\mu = h(\eta) = h(R\theta) \tag{5.2}$$

with the GLASS parameter vector of $\theta = (\alpha_L', \alpha^{\star'})'$. If the GAM portion is non-null, note that we do not need to include the intercept term in X_L since the column of ones is already in the span of R. We now explain how to estimate these components simultaneously through a modified scoring algorithm using P-spline methodology.

The penalized log-likelihood function that was presented in Section 3 must now be further modified such that a separate difference penalty is attached to the log-likelihood function for each of α in $\alpha^\star$. This results in maximizing the penalized version of the log-likelihood function

$$l^*(\theta; y, R) = l(\theta; y, R) - \frac{1}{2} \sum_{\alpha_j \in \alpha^\star} \lambda_j |D_{dj}\alpha_j|^2, \tag{5.3}$$

where the p nonnegative λ's are the regularization parameters. Note that $l(\theta; y, R)$ is the usual log-likelihood of the standard GLM, and we see in (5.3) that further smoothing is achieved through difference penalties as outlined in Section 3.

The Fisher scoring algorithm can be modified to the following iterative estimation technique

$$\hat{\theta}_{t+1} = (R'\hat{W}_t R + P)^{-1} R'\hat{W}_t \hat{z}_t, \tag{5.4}$$

where again $\hat{W}$ are the GLM weights and $\hat{z}$ is the working dependent variable, evaluated at iteration t. The matrix P is a block-diagonal matrix of proper dimension ($c \times c$), which has zeros in the block for the X_L terms, followed sequentially by diagonal blocks of appropriately chosen $\lambda D_d' D_d$ for each contribution in $\alpha^\star$, respectively. With the P-spline approach, note that the number or positions of knots do not have to be changed to vary smoothness, rather the various λ's can regulate such control continuously. This is efficient since R only has to be computed once in the estimating equations above.

We would like to point out, with the sum of (many) linear combinations in a model, we cannot avoid aliasing: the columns of R will in general be linearly dependent, leading to underdetermined systems of equations. The smoothness penalty will not remedy this. Our way out is to add a small “ridge penalty" (with a weight of, say, 10^{-6}) on all coefficients for B-splines. Specifically we replace P with $P + 10^{-6}I$, where I is the identity matrix of proper dimension. This way the length of these vectors of regression coefficients are pushed gently toward a minimum, by centering them around zero.

Standard error bands can be constructed for $\hat{\eta}$ using the diagonal elements of the estimated covariance matrix $\hat{C}$:

$$\hat{C} \approx R(R'\hat{W}R + P)^{-1}R'\hat{W}R(R'\hat{W}R + P)^{-1}R'. \tag{5.5}$$

The preceding covariance formula is only asymptotically correct if the λs are chosen a priori. Additionally, GLM deletion diagnostics to help identify influential observations can be constructed using the converged effective hat matrix (Pregibon 1979)

$$H(\lambda) = R(R'\hat{W}R + P)^{-1}R'\hat{W}. \tag{5.6}$$

With Normal errors and the identity link, W is the identity matrix and standard regression diagnostics can be used routinely, for example to identify outliers or particular influence on θ parameter estimates. Myers (1990) provided a nice coverage of such diagnostics for standard multiple regression.

6. OPTIMAL SMOOTHING

The GLASS model involves one or more penalty parameters, and in any application one has to optimally and objectively choose values for them. The ingredients of leave one out cross-validation were given in Section 3 (for standard regression), which leads to the cross-validation statistic

$$\mathrm{CV} = \left(\frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_{-i})^2 \right)^{\frac{1}{2}}.$$

Again $y_i - \hat{y}_{-i} = (y_i - \hat{y}_i)/(1 - h_{ii})$, and now h_{ii} are the diagonal elements of the hat matrix H from (5.6). We can write

$$\hat{\theta} = (R'R + P)^{-1}R'y,$$

and

$$\hat{y} = R\hat{\theta} = R(R'R + P)^{-1}R'y = Hy,$$

since W is proportional to the identity. Thus, the model is fitted only once and the diagonal of H is computed. Because of the moderate number of columns of R, computation is not expensive. One searches for a minimum of CV by varying the smoothing parameters in a systematic way, or by using an optimization routine. We mention other options in Section 9.

For generalized linear applications there is no simple expression for cross-validation, because the link function generally is nonlinear and the deviance is nonquadratic. We propose to simply monitor and minimize an information criterion, such as

$$\text{AIC} = D(y, \hat{\mu}) + 2f(T),$$

where $D(y, \hat{\mu})$ is the model deviance. Using cyclical permutations for trace computations,

$$T = \text{trace}(H) = \text{trace}(R'WR(R'WR + P)^{-1})$$

upon convergence. We interpret T as the effective dimension of the model. Hurvich, Simonoff, and Tsai (1997) noted that the original definition of AIC, with $f(T) = T$ has a tendency to undersmooth. They showed that $f(T) = 1 + 2(T + 1)/(m - T - 2)$ is an improvement, especially in applications with many parameters. Certainly other information criteria, such as Bayesian or Schwartz's can also be used. We must add that the information criterion generally assumes that the observations obey, say, a binomial or Poisson distribution. Frequently one encounters overdispersion, in which case the appropriateness of the criterion is doubtful.

7. APPLICATIONS

We illustrate the use of GLASS with several experimental datasets. The emphasis is on its practical use: we will give little attention to the interpretation of the models themselves and the estimated results. The first example was chosen because it was presented in the literature to illustrate slow convergence of the backfitting algorithm for variable-coefficient models. The second and third examples consider modeling of time series with trends and seasonal components of varying strength, for normal and Poisson responses, respectively. Binomial responses are modeled with a factor (L), a smooth (G), and signals (S) regressors in the last example.

7.1 EXAMPLE 1

We revisit the data from Daniel and Wood (1980, p. 142) which presented a marketing price-volume study with 124 observations taken over a six-month period in 1970 (holidays omitted). The response variable is the (log 10) volume of gasoline sold (*log(volume)*). The explanatory variables are *date* index (ranging from 48 to 228), *price*, and *differentialprice* to competition. Daniel and Wood considered standard multiple regression approaches, adding indicator variables for *weekday* and for *month* (discrete trend).

As with using indicators for *month*, we note that zero degree B-splines (constant curves, each covering just one interval between knots) are also capable of producing nearly identical rough stepwise trends. However, one can imagine the usefulness of modeling the time or *date* trend smoothly, rather than through a rough estimate of stepwise trend (see Figure 5,

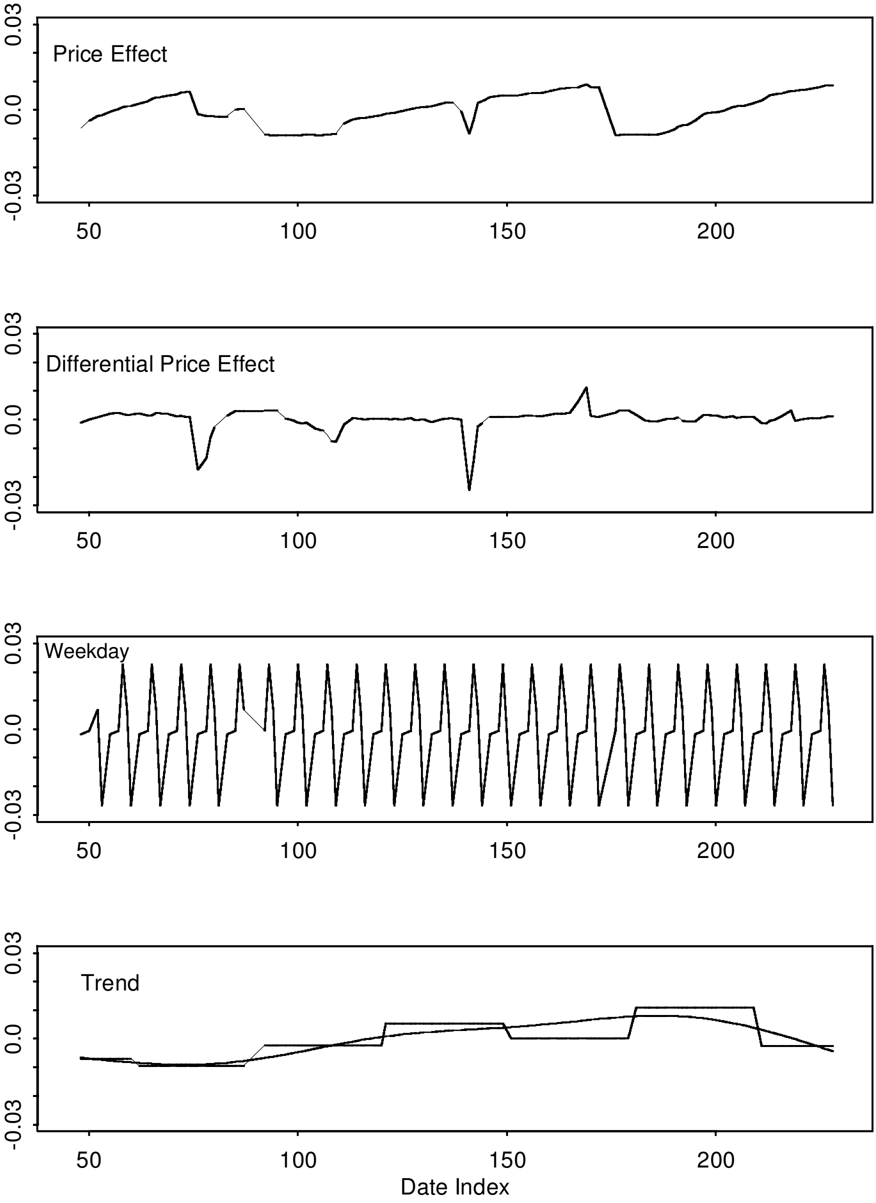

Figure 5. Daniel and Wood (1980). From top to bottom: L effects for price, differential price, and weekday; Smooth trend for date index (or varying intercept). For comparison, Daniel and Wood's stepwise month trend is overlaid the bottom figure.

bottom). In fact, Green (1993) tried to estimate such a smooth nonparametric trend, but found convergence of backfitting to be extremely slow. He additionally proposed a noniterative penalized least squares algorithm that leads to a system of equations with as many unknowns as the number of observations. In this setting, the backfitting algorithm alternates between regressing the residual of the smooth on X_L and smoothing the residual of X_L. We confirmed that it took several hundreds of iterations for a ten-fold reduction of size of the error, and the number of iterations increased with the number of B-splines used. Despite this lack of success, we would like to offer for the record that we did find considerable relief in convergence rates by simply subtracting the mean from each column of X_L: two to five iterations gave a ten-fold reduction of the error.

Nevertheless, our GLASS solution is more attractive than the above scheme because it is faster and it allows economic cross-validation. We recognize that a time-varying intercept term (over the *date* index) is equivalent to a smooth trend component for the *date* index. Such a GLASS model now can be coerced into the framework of direct GAMs using penalized likelihood (Marx and Eilers 1998) with a solution found in one step. The model can be written as

$$\hat{y} = X_L\hat{\alpha}_L + \hat{f}(t) = [X_L|B_G]\hat{\theta} = X_L\hat{\alpha}_L + B_G\hat{\alpha}_G,$$

where $\hat{f}(t)$ is the smooth time trend and X_L contains the fixed regressors. Figure 5 displays the estimated effects using constant coefficients, or the linear portion, for *price*, *differential price* and *weekday* indicators (6 linear df), while smoothly estimating the time or *date* trend (X_G). The matrices X_V and X_S are empty. There are no longer any convergence problems since backfitting is avoided. The smooth in Figure 5 (bottom) was fit with cubic B-splines (eight equally spaced knots) and a second-order difference penalty on the B-spline coefficients. The optimal $\lambda = 0.1$ (effective df of 4.56) minimized cross-validated standard error of prediction, CV $=$ 0.0085. Figure 6 presents the response (*log(volume)*) as a function of *date*, as well as the above model's predicted values and residuals.

7.2 Example 2

Our next example models seasonal time series of monthly (January 1963 through December 1986) concentrations of sulphur dioxide, SO_2, air pollution in the Netherlands. The data were collected by the Rijnmond Environmental Agency in units of μg/m^3, originally hourly values at approximately 30 monitoring stations. These measurements were then averaged over all stations and all hours in each month, producing the $m = 288$ monthly composite responses. Figure 7 (top) presents the SO_2 concentrations (in log 10 unit) as a function of time. We find a downward trend. Additionally we see a strong seasonal component with peaks in the winter and troughs in the summer, however with decreasing magnitude in time. We choose to GLASS model log(SO_2) with one smooth component (X_G) in time and two entries for the varying coefficients (X_V), seasonal sine and cosine waves. We propose

$$E\{\log(\mathrm{SO}_2)_i\} = \alpha_0 + f(i) + f_1(i)\sin(\omega i) + f_2(i)\cos(\omega i), \qquad (7.1)$$

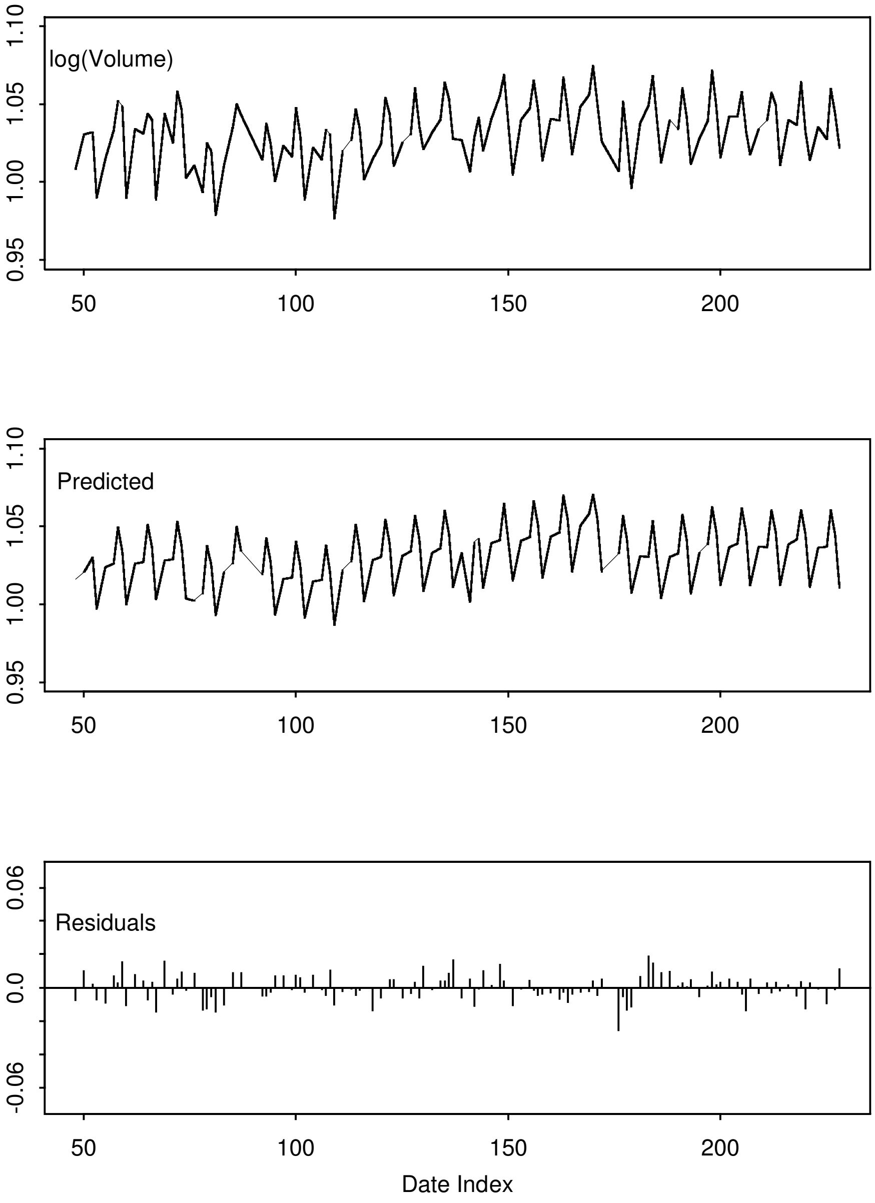

Figure 6. Daniel and Wood (1980). From top to bottom: The log(volume), predicted log(volume), and residuals plotted by date index.

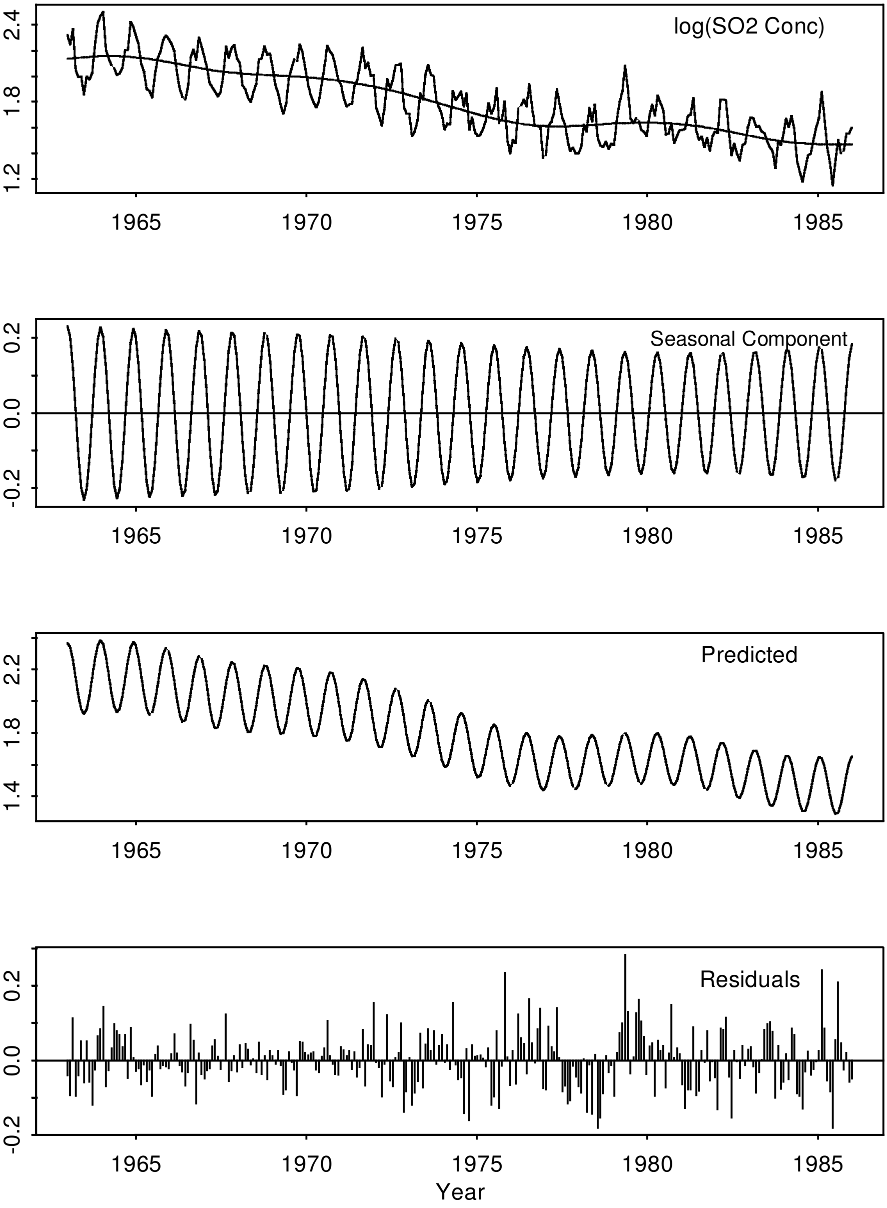

Figure 7. log(Sulphur Dioxide) air pollution. From top to bottom: Monthly (log) averages over time with smooth trend; varying seasonal components; predicted values based on optimal smoothing; and residuals.

where $i = 1, \ldots, 288$ indexes months. An approximately Normal error distribution is assumed. The seasonal component uses the circular frequency $\omega = 2\pi/12$ with coefficients $(\alpha_{V1i}, \alpha_{V2i})$ indexed by i to modulate slowly varying signals. For both G and V, cubic B-spline bases with 13 (10) equally spaced knots (interval segments) were used with a second-order difference penalty on the B-spline coefficients. A grid search for the optimal penalty parameter produced $\lambda_G = 0.10$ (effective dimension 8.94) and (common) $\lambda_V = 10$ (effective dimension 3.54 for each of the sine and cosine portion). These optima were based on minimizing CV $= 0.0798$.

Figure 7 (top) also displays the 8.94 df estimated (downward) smooth time trend. Figure 7 (second from top) displays the sum of the seasonal sine and cosine components (multiplied by their respective varying coefficients). The plots of the fitted values and the residuals are shown in the bottom two panels. The latter shows that this simple model does not catch all the structure in the data: the residuals show strong serial correlation and their variance seems to vary. Also the fit is quite bad around 1979, which was an unusually severe winter in the Netherlands.

7.3 Example 3

The above SO_2 model assumes a sinusoidal seasonal component, a rather strong assumption. If this is not satisfactory, then more harmonics can be added to allow more complicated seasonal patterns. Consider the discrete count time series taken from Zeger (1988). Figure 8 (top) displays the observed (and fitted) counts of monthly polio incidences in the United States (reported to the U.S. Center for Disease Control) during the years 1970 through 1987. For these data we consider extending the X_V matrix with pairs of columns, giving the sines and cosines at double frequency. Using the Poisson random component and the log-link function, we fit the GLASS

$$\log(\mu_i) = \alpha_0 + f(i) + \sum_{k=1}^{2} \{f_{1k}(i) \sin(k\omega i) + f_{2k}(i) \cos(k\omega i)\}.$$

The above model allows varying coefficients for the sine and cosine frequencies of $2\pi i/12$ and $2\pi i/6$, where $i = 1, \ldots, 216$ is the month index.

Figure 8 (second, third, and fourth panel) also provides the multiplicative factors for the smooth trend, annual seasonal component, and semi-annual seasonal component, respectively. Multiplying these portions produces the fitted curve in the top panel. The deviance residuals are also displayed (bottom). We further address their autoregressive tendency in Section 9. Some details of this Poisson model include: convergence in 4 scoring iterations, the trend is fit with 8.3 df, each seasonal term is fit with (a common) 7.6 df, and cubic B-splines with 13 knots and a second-order penalty for each term. Optimal smoothing was determined by an AIC grid search on the smooth ($\lambda = 0.1$) and a common seasonal smoothing parameter ($\lambda = 0.1$). The residual deviance is approximately 215 on 176 df. This deviance is not a reliable measure of goodness-of-fit since the large sample chi-square theory is violated in two ways: we have many small or zero counts and the number of

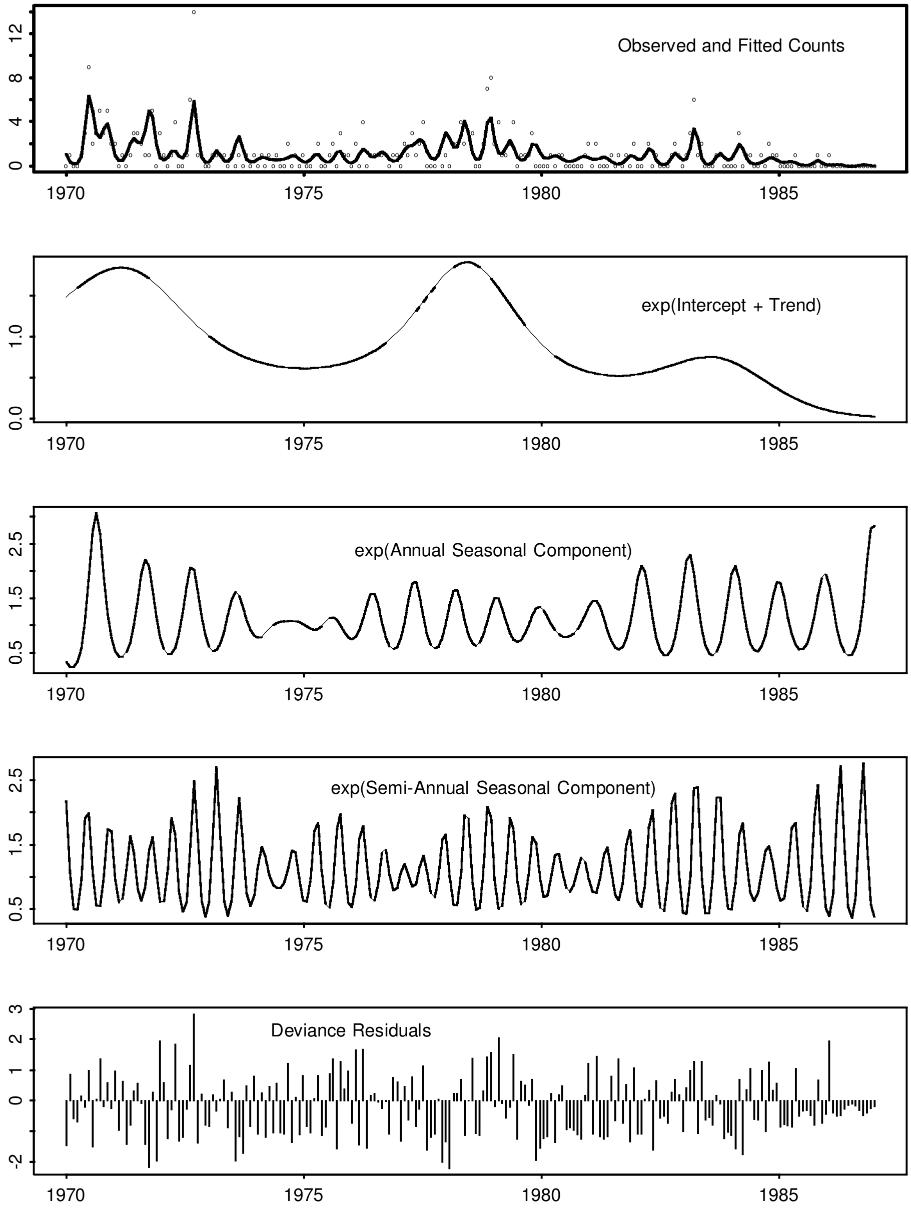

Figure 8. USA polio incidence (Poisson, log link). From top to bottom: Monthly observed and fitted counts (based on optimal AIC smoothing); multiplicative factor for smooth time trend (including intercept); multiplicative factors for the varying annual and semi-annual seasonal components; and the deviance residuals.

settings for the regressors (time) is not fixed as $m \to \infty$.

It is possible to fit constrained forms of the model, in which the relative strengths of the sines and cosines do not vary with time. In such a model the shape of each period is the same, but the size varies smoothly. These are interesting extensions, but they lead us into bilinear models, which are outside the scope of this article.

7.4 EXAMPLE 4

The last example offers elements of both signal regression and smooth additive modeling to predict the probability that a freesia's bud will flower. The signal was a near-infrared reflectance (NIR) spectroscopy taken on $m = 100$ branch bunches. Within a branch bunch, several branches (in most cases 20) were collected and then split into two groups: The first group (usually consisting of 14 branches) was put into a vase and monitored for successful budding, and the other branches (usually 6) was used for NIR spectroscopy. For a given source, the total combined buds (N_i) in the vase were counted, $i = 1, \ldots, 100$. We are interested in how the NIR and other factors are related to the number of buds that produce flowers in vase i, $y_i \in \{0, \ldots, N_i\}$. For the moment, we assume $y_i \sim$ binomial(N_i, p_i), with p_i unknown.

Some details on the explanatory information follow. The NIR spectra consisted of 476 log reflectance (1/R) readings ranging from 600 nm to 2,500 nm in equal steps of 4 nm. See Figure 9 (top, left) which displays the NIR spectra for the $m = 100$ freesia sources. There also existed a variety of other lab information (*storage time*, *drying time*, etc.); we are particularly interested in the two *cultivar* levels. To predict the probability of a *successful flowering* (1) or *unsuccessful flowering* (0) bud, we entertain the logit model

$$\log \frac{p_i}{1 - p_i} = X_L \alpha_L + f(N) + U_S f_S = X_L \alpha_L + B_G \alpha_G + X_S B_S \alpha_S,$$

where X_L contains the intercept term and a 0/1 dummy for cultivar, N is the number of buds found on a branch in a source, and X_S is the NIR spectra matrix (100 × 476).

Both the smooth additive (9 knots) and the signal (11 knots) terms were fit with cubic P-splines. Third-order difference penalties were used. Convergence was achieved in three iterations of the GLM scoring algorithm. Optimal smoothing (based on a AIC grid search) showed that a good deal of flexibility was needed for each of these terms ($\lambda_G = 0.01$, effective df $= 6.9$ and $\lambda_S = 0.001$, effective df $= 8.9$). The overfitting of the smooth and signal terms may be a result of the overall fit which is further discussed below. A ridge penalty of 10^{-4} was also used to remedy aliasing. Figure 9 (upper, right) displays the smooth additive portion for N, along with the partial residuals. This smooth is significant (Δ deviance $=$ 255.7 on 6.9 df) and generally suggests that a decrease in odds of budding with increasing N. Note that in the residuals associated with small N all correspond to *cultivar* 1 and have positive residuals, indicating a warning of a pure-region effect. For moderate and large N, we also find several distinct negative residuals associated with *cultivar* 2. The *cultivar* effect is significant (Δ deviance $= 19.9$ on 1 df). Figure 9 (lower, left) provides the

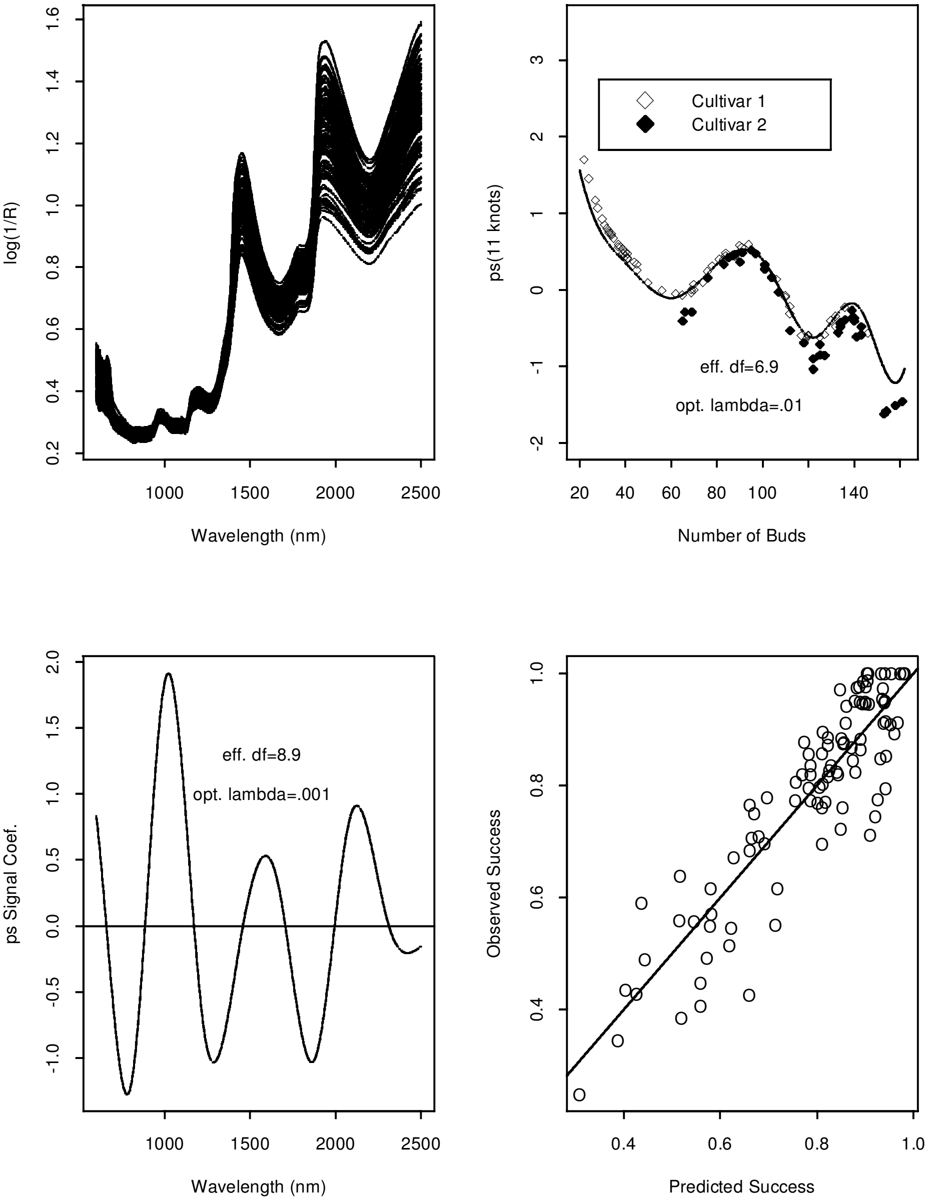

Figure 9. Freesia data. NIR spectra readings of $m = 100$ Freesia sources (upper left); estimated spectra (476) coefficient vector (lower, left); P-spline smooth of N number of buds with partial residuals (upper, right); observed versus predicted probability of success (lower right).

significant signal coefficient vector (on $t = (600, 2500)$) for the X_S matrix; Δ deviance = 182.2 on 8.9 df.

Technically, the model is a success (it quickly gives straightforward estimates of all parameters without any signs of numerical problems), and the fit to the data is improved by adding the various L, G, and S portions. Figure 9 (lower, right) displays the observed versus fitted predicted probabilities of *successful flowering* which at first glance appears somewhat promising. However, the residual deviance is 378.3 on 82.2 df, indicating that a binomial model may not be appropriate or that other explanatory information may be needed. For reasons described in the previous example, this deviance is not a reliable measure of goodness-of-fit. Although we do not pursue outliers here, we find three deviance residuals less than -2.3. Lastly, overdispersion is likely present, and caution should be taken when using AIC for optimal smoothing.

8. COMPUTATIONAL DETAILS

Using the details presented in Marx and Eilers (1998), we have broadened the S-Plus P-spline `ps()` function to accommodate not only smooth G terms, but also varying V and signal S terms. Like `ps()`, the new `glass()` functions work directly with the existing `gam()` function. Thus, the variety of arguments (such as `link`, `family`, etc.) are all accessible. The `glass()` function parallels the `bs()` (B-spline) function, except equidistant knots (interval segments) are used rather than ones on the quantiles. However, `glass()` has its own arguments: the `degree=3` of B-splines, number of `ps.intervals` (knots−`degree`), `order=3` of penalty, and regularization parameter `lambda=0`. Additionally for varying terms, `varying.index` must be specified (NULL is default); this is the t_c vector. When signal S entries are used, then `signal.index` (i.e., t_r, NULL default) and `x.signal` (the signal matrix, NULL default) must be specified. Based on the specific arguments provided in `glass()`, the D_d matrix is constructed and either B_G, U_V or U_S is passed to `glass.wam()` which orchestrates the (iterative) penalized method of scoring (avoiding the call to backfitting). The penalization is constructed through data augmentation techniques. An example function might look like:

```
glass1 <- gam(Y ~ glass(X1, ps.int=10, degree=2, order=1,
lambda=.01)+ X2 + glass(X3, ps.int=8, varying.index=t.c, degree=3,
order=3, lambda=10) + glass(1:length(Y), ps.int=20,
spectra.index=t.r, x.signal=S1, lambda=1) + factor(block),
family=binomial(link=logit), na.action=na.omit).
```

The `glass1` object would fit a logistic regression on binary Y using a P-spline smooth for $X1$, a linear term for $X2$, a varying coefficient term (on $t.c$) for $X3$, a signal term for $S1$ (on $t.r$), and a block. The usual `gam` list would be available with some additional entries, such as: linear and nonlinear df's for terms. The function `plot.glass()` provides plots of smoothes, varying coefficients, and signal coefficients with twice pointwise standard error bands if desired. The S-Plus code is available at www.stat.lsu.edu/bmarx and Matlab code (for a variety of common applications) is available from the first author.

9. DISCUSSION

We have presented a model based on generalized additive structures and penalties and illustrated its use with various applications. We believe that we have a sensible solution to a complex problem. Although some of the ideas presented have been addressed in the existing literature, a practical approach was lacking, especially when addressing all of these structures simultaneously. We hope that this work will highlight just how simple P-splines are to use and how they lead to such a manageable system of equations. GLASS offers a unified modeling framework that eliminates the need for backfitting and knot selection schemes, and further allows easy computation of diagnostics, compact results useful for prediction, computation of standard errors and fast cross-validation.

In our opinion, P-splines are the only viable smoother for the GLASS framework. With backfitting, kernel smoothers and local likelihood methods can be used for the GAM and VCM components. The computation of standard errors and diagnostics is complicated and cross-validation is expensive. For signal regression these smoothers are useless. Smoothing splines can be used only in principle: they lead to (very) large systems of equations. Although these can possibly be sparse for the GAM and VCM components, this is not the case for signal regression. Regression splines with fixed knots can encounter problems with unevenly distributed data. A solution is to locate the knots at evenly spaced percentiles of the data, but this gives only coarse control over smoothness, as the number of knots is an integer. Knot optimization is a nonlinear problem; we also have doubts of its applicability to signal regression. The closest to P-splines come penalized truncated polynomials, as presented by Ruppert and Carroll (2000). Implicitly they only used a second-order difference penalty. Elsewhere, in the context of GAM (Marx and Eilers 1998), we have shown that it may be useful to try several orders of the difference penalty to detect potential polynomial fits to the data. Another aspect of truncated polynomial bases, that up to now seems to have largely gone unnoticed, is their extremely bad numerical condition.

In larger GLASS models several smoothing parameters occur, each of which has to be optimized. Our present approach is to simply do a grid search; as cross-validation or computation of AIC is fast, this is feasible. More elegant approaches can possibly derived from mixed model theory. Verbyla, Cullis, and Kenward (1999) show how smoothing splines can be written as a mixed model. In the discussion to that article Eilers (1999a) showed how the same can be done for P-splines. Algorithms and software for the estimation of variance components in mixed models are rather well-developed (see Coull, Ruppert, and Wand 2001).

Our development of GLASS did assume independent observations. In the applications to time series data we noted apparent autocorrelation of residuals. With a normal response (and identity link) a solution is to introduce a covariance matrix, say Σ, derived from an autoregressive or moving average model and minimize a sum of squares that is weighted by

Σ^{-1}. Currie and Durbán (2001) reported good results on the use of P-splines with correlated responses in a mixed model framework. Further conditional models can be explored that add lag(Y) in the X_L or X_V portion. When the response is non-normal, no such simple solution exists. A first step in the right direction would be to adopt GEE (generalized estimating equations) (Zeger 1988).

As we showed, each of the GAM, VCM and PSR components constructs a linear predictor as $MB\alpha$, where B is the B-spline matrix, α the coefficient vector, and M a component-specific matrix: the identity matrix for GAM, a diagonal matrix for VCM, and the signals matrix for PSR. This suggests a search for additional components. An example is regression on time series at several delays, like distributed lags of order l (Malinvaud 1970): $\hat{y}_i = \sum_{j=0}^{l} x_{i-j}\alpha_j$, with x given and smooth α to be estimated. The rows of M would be shifted copies of x.

In this article we consider only a univariate response. One area of future work could be to extend GLASS to a multivariate response, like the multivariate GAM of Yee and Wild (1996). GLASS building blocks can also be used for modeling sets of crossed or nested curves. This would offer large computational advantages compared to the smoothing splines used by Brumback and Rice (1998) and Verbyla, Cullis, and Kenward (1999), as indicated by Eilers (1999b), and additionally open the possibility to use VCM and PSR components.

ACKNOWLEDGMENTS

We thank N. Marissen of the Research Station for Floriculture and Glasshouse Vegetables (Aalsmeer, the Netherlands) and R. Frankhuizen of DLO-State Institute for Quality Control of Agricultural Products (Wageningen, the Netherlands) for permission to use the freesia data. We thank Hilko van der Voet of the Agricultural Mathematics Group (GLW-DLO, Wageningen, the Netherlands) for his support in obtaining the data.

[Received June 2001. Revised July 2001.]

REFERENCES

Brumback, B. A., and Rice, J. A. (1998), "Smoothing Spline Models for the Analysis of Nested and Crossed Samples of Curves" (with discussion), *Journal of the American Statistical Association*, 93, 961–994.

Coull, B. A., Ruppert, D., and Wand M. P. (2001), "Simple Incorporation of Interactions into Additive Models," *Biometrics*, 57, 539–545.

Currie, I., and Durbán (2001), "Curve and Spline Analysis of Variance," in *Statistical Modeling. Proceedings of the 16th International Workshop on Statistical Modelling, Odense.*, eds. B. Klein and L. Korsholm, L. pp. 127–134.

Daniel, C., and Wood, F. S. (1980), *Fitting Equations to Data*, New York: Wiley.

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

Dierckx, P. (1993), *Curve and Surface Fitting with Splines*, Oxford: Clarendon Press.

Eilers, P. H. C. (1999a), Discussion of "The Analysis of Designed Experiments and Longitudinal Data by Using Smoothing Splines" by Verbyla, Cullis, and Kenward, *Applied Statistics*, 48, 269–300.

——— (1999b), "Curve and Spline Analysis of Variance," in *Statistical Modeling. Proceedings of the 14th International Workshop on Statistical Modelling, Graz.*, eds. H. Friedl and G. Kauermann, pp. 173–180.

Eilers, P. H. C., and Marx, B.D. (1996), "Flexible Smoothing Using B-Splines and Penalized Likelihood" (with comments and rejoinder), *Statistical Science*, 11, 89–121.

Green, P. (1993), Discussion of "Varying-Coefficient Models" by Hastie and Tibshirani, *Journal of the Royal Statistical Society*, Ser. B, 55, 757–796.

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge: University Press.

Hastie, T., and Mallows, C. (1993), Discussion of "A Statistical View of Some Chemometrics Regression Tools" by I. E. Frank and J. H. Friedman, *Technometrics*, 35, 140–143.

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, London: Chapman and Hall.

——— (1993), "Varying-Coefficient Models" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 55, 757–796.

Hurvich, C. M., Simonoff, J. S., and Tsai, C.-L. (1997), "Smoothing Parameter Selection in Nonparametric Regression Using an Improved AIC Criterion," *Journal of the Royal Statistical Society*, Ser. B, 60, 271–293.

Malinvaud, E. (1970), *Statistical Methods of Econometrics*, Amsterdam: North-Holland.

Marx, B. D., and Eilers, P. H. C. (1998), "Direct Generalized Additive Modeling with Penalized Likelihood," *Computational Statistics and Data Analysis*, 28, 193–209.

——— (1999), "Generalized Linear Regression on Sampled Signals and Curves: A P-Spline Approach," *Technometrics*, 41, 1–13.

Marx, B. D., Eilers, P. H. C., and Auer, D. P. (2002), "Generalized Additive Models with Signal Components from Medical Instruments," Technical Report RR-02-38, Department of Experimental Statistics, Louisiana State University.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

Myers, R. H. (1990), *Classical and Modern Regression with Applications* (2nd ed.), Boston: Duxbury Press.

Nelder, J. A., and Wedderburn, R. W. M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society*, Ser. A, 135, 370–384.

O'Sullivan, F. (1986), "A Statistical Perspective on Ill-Posed Inverse Problems" (with discussion), *Statistical Science*, 1, 505–527.

Pregibon, D. (1979), "Data Analytic Methods for Generalized Linear Models," Ph.D. Dissertation, University of Toronto.

Reinsch, C. (1967), "Smoothing by Spline Functions," *Numerische Mathematik*, 10, 177–183.

Ruppert, D., and Carroll, R. J. (2000), "Spatially-Adaptive Penalties for Spline Fitting," *Australian and New Zealand Journal of Statistics*, 42, 205–223.

Verbyla, A. P., Cullis, B. R., and Kenward, M. G. (1999), "The Analysis of Designed Experiments and Longitudinal Data by Using Smoothing Splines," *Applied Statistics*, 48, 269–300.

Yee, T., and Wild, C. J. (1996), "Vector Generalized Additive Models," *Journal of the Royal Statistical Society*, Ser. B, 58, 481–493.

Zeger, S. L. (1988), "A Regression Model for Time Series of Counts," *Biometrika*, 75, 621–629.