Generalized Linear Regression on Sampled Signals and Curves: A *P*-Spline Approach

Brian D. MARX

Department of Experimental Statistics
Louisiana State University
Baton Rouge, LA 70803
(brian@stat.lsu.edu)

Paul H. C. EILERS

DCMR Rijnmond Environmental Agency
3119XT Schiedam
The Netherlands
(paul@dcmr.nl)

We consider generalized linear regression with many highly correlated regressors—for instance, digitized points of a curve on a spatial or temporal domain. We refer to this setting as *signal regression,* which requires severe regularization because the number of regressors is large, often exceeding the number of observations. We solve collinearity by forcing the coefficient vector to be smooth on the same domain. Dimension reduction is achieved by projecting the signal coefficient vector onto a moderate number of B splines. A difference penalty between the B-spline coefficients further increases smoothness—the P-spline framework of Eilers and Marx. The procedure is regulated by a penalty parameter chosen using information criteria or cross-validation.

KEY WORDS: B splines; Compression; Difference penalty; Multivariate calibration; Partial least squares; Penalized likelihood.

During the last five years, there has been a surge of renewed interest in techniques that once were specific to the chemometric community. Refer to Frank and Friedman (1993) for an excellent summary of chemometric estimation options, including partial least squares (PLS) and principal-component regression (PCR), among others. Simply put, modern technology is routinely generating experimental data that severely challenge standard regression models, and some chemometric approaches are providing reasonable data-analytic tools for the statistician. We revisit the multivariate calibration (MVC) problem, which relates signal-regressor information to a response variable of interest. From this point forward, we will refer to the MVC problem as *signal regression.* Some of the challenges posed in this framework are nontrivial and include the following: (a) The response variable of interest may be discrete or have a nonsymmetric distribution. (b) The number of regressors (p) may greatly exceed training observations (m); that is, $p \gg m$. (c) The regressors are highly correlated so that even if $m > p$ we have a very ill-conditioned problem. (d) Spatial information on the regressors often exists.

Similar challenges arise naturally in a variety of other applications—for example, log-spectra of digitized sequences of spoken syllables to predict phoneme classification, gray-scale pixel values from images used to model character recognition, light-scattering profiles used to understand features of wafer-etching experiments, and relating DNA histogram information to the presence of ovarian cancer, among others. All of these applications have severely ill-conditioned regressor information that can wreak havoc on classical (generalized linear) regression techniques. Some researchers have attacked the problem with selection methods, searching for a relatively small set of optimal regressors (e.g., selecting 20 among hundreds of wavelengths in a spectra). Despite success with variable selection for some signal-regressor information, often this approach can be unsatisfactory and impractical. For example, with smooth spectra, it is difficult to find a clear optimal subset of regressors among the often hundreds of candidates. More fundamentally, it is difficult to accept that any suggested optimum will be sharply defined in a smooth spectra. Any "optimal" regressor will differ very little from its neighbor at the next higher or lower wavelength.

In this article, we propose to integrate ideas from Hastie and Mallows (1993) (H&M) and Eilers and Marx (1996) (E&M) but to extend estimation and prediction in the framework of generalized linear regression (GLR). We aim to present an extremely practical solution for the signal-regression problem by forcing the regression coefficients to be smooth. We specifically consider the model $g(\mu) = \alpha_0 + X_{m\times p}\alpha_{p\times 1}$, where often $p \gg m$ and X is the severely ill-conditioned matrix of signal regressors. Denote g as the monotone, twice-differentiable link function, μ as the expected value of a random variable Y from exponential family of distributions, and α_0 as the intercept. H&M used the phrase *contrast template* for the vector α, because it contrasts important signal information useful for predicting the response; we will choose the term *signal coefficient vector.* Some other details of the generalized linear model will follow in Section 3; however, the unfamiliar reader could reference McCullagh and Nelder (1989) or Fahrmeir and Tutz (1994).

Like H&M, we attack the dimensionality of the signal coefficient vector α by projecting it onto a basis of smooth functions, B: $\alpha_{p\times 1} = B_{p\times n}\beta_{n\times 1}$, where $n < \min(m, p)$ and β is the vector of basis coefficients. The linear nature of the B-spline smoother makes it an attractive candidate for B, but the determination of the optimal number and positions

of the knots can be a complex task. We propose to use P splines (E&M), which is the combination of (a) projecting α onto a moderate number of equally spaced B splines and (b) further increasing smoothness by imposing a difference penalty on adjacent coefficients in the β vector. We will refer to our method as *P-spline signal regression* or PSR. The penalty is subtracted in the log-likelihood and consequently increases smoothness of α. With the penalty, the effective dimension of estimation of the regression model is reduced from p to less than n. We briefly discuss more about B splines in Sections 2 and 4.1. Some researchers, such as Alsberg (1993) and Denham and Brown (1993), preferred to smooth the signal directly (data compression) rather than the signal coefficient vector; we will draw connections to these two approaches in Section 3.3. The implementation of PSR with suggestions to optimize the penalty meta-parameter follow in Sections 3 and 4, respectively. Illustrative examples are found in Section 5, and a brief survey of other existing approaches can be found in Section 6. First, we revisit the standard signal-regression problem.

1. COMPARING THE PROPOSED PSR TECHNIQUE TO PLS AND PCR

One of the goals for this article is to make the threshold between our work and the current signal-regression literature as small as possible. To help strive toward this, we now present a standard linear signal problem in the chemometric context that compares PSR to PLS regression and to PCR. In Section 5 we will present more exotic examples that use both the generalized linear model (GLM) framework and extremely nonsmooth signals. As we will point out in Section 3.3, our PSR approach does not require that the spectra themselves be smooth; smoothness is only required in the associated regression vector. We first revisit the well-known chemometric example from Osborne, Fearn, Miller, and Douglas (1984) that was also used by Stone and Brooks (1990, ex. 5). This example related spectral information of biscuit dough to percentage of various constituents—% fat, % sucrose, % flour (dry weight basis), and % water. Consider *near infrared reflectance spectroscopic* (NIR) information (consisting of hundreds of digitizations of one signal) that can be used to predict a chemical response variable of interest. Refer to Figure 1 (top); there are $m = 39$ curves and each curve represents $p = 601$ regressors (1,200 nm to 2,400 nm, in steps of 2 nm). Denote $X_{39\times601}$ as a discrete representation of the observed signals. As we see from Figure 1 (top), the spectra have clearly shifted (due to unequal particle sizes). Differencing the columns of X effectively removes constants and sudden shifts that are not important to the regression. Figure 1 (bottom) displays the regressor information when the columns have been first-differenced. It is important to note that NIR photometric measurements can be faster and less costly than a chemical analysis. Parameter estimation associated with photometric information can provide predictive equations for chemical constituents and thus be of economic interest.

Initially there were $m = 40$ samples; however, on the advice of the authors, observation number 23 was discarded as an outlier, leaving $m = 39$. Moreover, the original spectra had a wider range of 1,100 nm to 2,498 nm (in steps of 2 nm). The channels at the ends, however, were known to be less reliable for instrumental reasons; hence, we only use 601 (600) original (differenced) wavelengths (1,200 nm to 2,400 nm). In an effort to access the predictive ability among the competitor techniques (PSR, PLS, PCR), we have randomly selected 15 of the 39 observations for a validation set (observation numbers 1, 2, 3, 7, 15, 17, 19, 20, 22, 29, 32, 33, 35, 39, 40). The remaining $m = 24$ observations were used to train the models. We wish to relate $Y = \%$ fat to $X_{24\times600}$, the (training) differenced spectral regressor information. Consider the model $E(Y) = \mu = \alpha_0 + X\alpha = \alpha_0 + XB\beta$, assuming an (approximately) Normal error structure. Clearly standard multiple regression does not work because the regressors are severely ill-conditioned. Ridge regression is often not a practical alternative in this setting because it can require enormous amounts of memory and unwieldy (iterative) matrix inversions (600×600 in the following example).

The predominant approaches to model signal regressors has been the use of PLS or PCR. PLS was born as a practical and ill-understood method of estimation (Wold 1975), but it has gradually obtained a firmer theoretical basis (Helland 1988), leading to much statistical and mathematical discussion. PCR uses the singular value decomposition (SVD). PLS strongly resembles the SVD and is related to the conjugate gradient algorithm, modified for optimal predictive power with respect to the response variable. Martens and Næs (1989) and Frank and Friedman (1993) both provided an excellent overview of PLS and the algorithmic details.

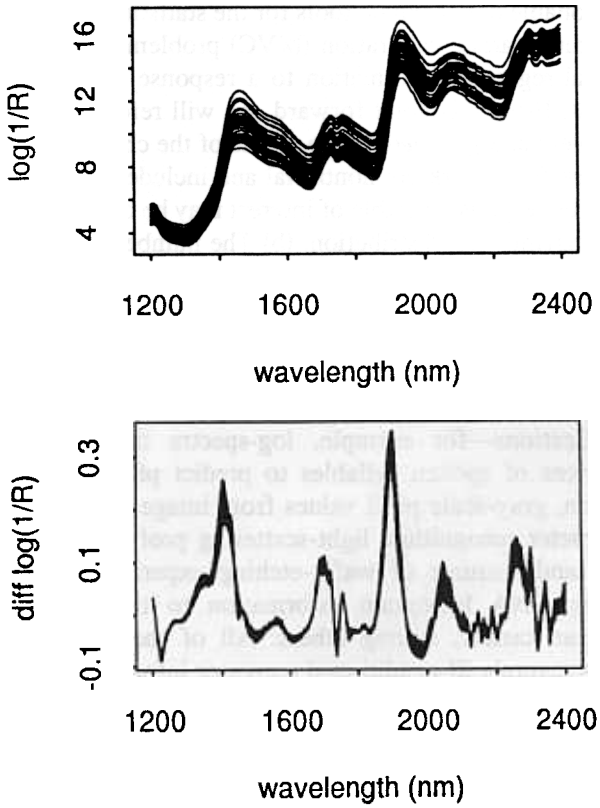

Figure 1. Biscuit: Top, Each Curve (m = 39) Represents p = 601 NIR Wavelengths; Bottom, the First-differenced NIR Spectra.

Alternatively, PCR uses a subset of orthogonally rotated regressors. PLS and PCR both produce a sequence of rank 1 orthogonal constructed variables that are linear combinations of the (autoscaled signal) regressors. A key difference between these two methods is the mechanism for choosing the loadings associated with their respective linear combinations. PLS chooses loadings that are based on the strength of simple linear correlation of the response with each regressor (wavelength), whereas PCR uses eigenvector loadings of the information matrix that are independent of the response.

PLS can be distinguished even further from PCR: Once PLS constructs a variable, then it is immediately related to the response. The second PLS-constructed variable has loadings based on the strength of correlation between the residual response and residual regressors, both orthogonal to the first constructed variable. This process continues for as many components as are desired. The optimal number of components can be determined by cross-validation, discussed in Section 4. PLS sometimes is termed *criss-cross* regression because it sequentially regresses constructed variables out of the (residual) response, then out of the (residual) regressors. Mainstream PLS or PCR considers least squares estimation but can be modified for non-Normal data and likelihood estimation (Marx and Smith 1990; Marx 1996). A revealing fact about PLS or PCR is that the order of the regressors is immaterial; that is, if the wavelengths are permuted arbitrarily, the PLS or PCR vectors will be permuted the same way. Our proposed PSR method uses additional structure, accounting for the indexing information along the signal, hence the smooth estimate of the coefficient vector, α.

For our PSR approach, we will not explain the choice of certain design parameters here; Section 4.1 contains some guidelines for the number and position of knots, degree of the B spline, order of the penalty, and so forth. PSR simplifies nicely in the standard setting to (penalized) least squares. To reduce dimensionality and regularize estimation, we project $\alpha_{600\times 1}$ onto a cubic B-spline matrix B (here with 23 equally spaced knots). Thus, α is compactly summarized by a 23-dimensional coefficient vector β vector. To further increase smoothness of the estimated α, a third-order difference penalty was attached to B-spline coefficients, β. The optimal penalty or tuning parameter $\lambda = 5e-7$ was found by cross-validation $(i, -i)$. We point out that essentially the same optimal dimension and standard error of prediction were achieved for the validation set when using either 23, 28, or 33 equally spaced knots.

Figure 2 (top) provides the estimated P-spline smooth $\hat{\alpha}_{600\times 1}$ (with 20.21 effective degrees of freedom) and its associated twice-standard-error bands. Apart from the intercept, a predicted value of % fat can be determined by the inner product of a given differenced spectrum and the vector represented by a solid curve in Figure 2. Based on the twice-standard error bands, we find some evidence that the first 100 coefficients (NIR range of 1,200 to 1,400 nm) are particularly important elements in the signal coefficient vector useful to predict % fat. Incidentally, using all 39 observations with the same P-spline design parameters yields a very similar result with an optimum dimension of 19.42.

We now attempt to approximate the "optimal" dimension for each of the three methods. To judge performance of a model as a predictor of new observations, we use "leave-one-out" cross-validation, which provides a vector of m predictions $\hat{y}_{i,-i}$. The evaluation criterion is the cross-validated $(i, -i)$ standard error of prediction (CVSEP):

$$\text{CVSEP}(i,-i) \quad \left(\sum_{i=1}^{m}(y_i \quad \hat{y}_{i,-i})^2/m\right)^{1/2}$$

It is worth mentioning that the CVSEP$(i, -i)$ for PSR can be constructed extremely swiftly because "hat" diagonal information can be used. For PLS or PCR, exact CVSEP needs to be computed by brute force, and it is considerably more taxing in computational time (i.e., by fitting m models at each deletion $-i$, then recentering, rescaling, and predicting at each respective i). Our experience shows that, for $m = 100$ observations, CVSEP is approximately 20 times faster for PSR than for PLS. Approximate CVSEP methods do exist for PLS that are efficient through the eigenvector decomposition of X to obtain latent variables that can use standard updating techniques.

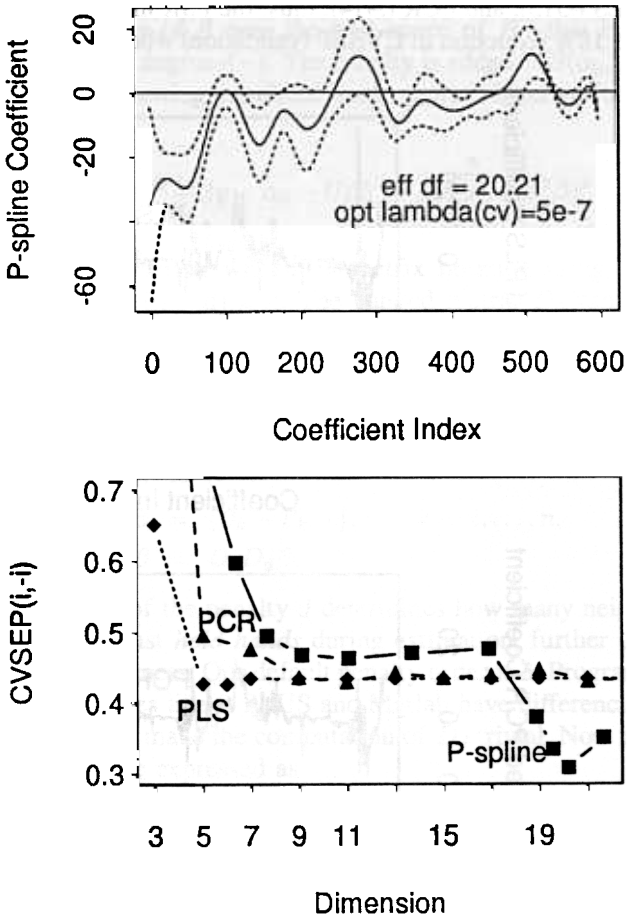

Figure 2. Biscuit: Top, Estimate (with twice standard error bands) of 600 Dimensional Signal Coefficient Vectors Using First-differenced Spectra; Bottom, a Comparison of Standard Error of Prediction (i; −i) by Methods and by (effective) Dimension.

Figure 2 (bottom) displays the CVSEP for % fat (in units of %) as a function of (effective) dimension and estimation technique. The x axis of Figure 2 represents effective dimensions for PSR and the number of factors used for PLS or PCR—this allows the presentation of all three methods on one figure. We caution that an effective dimension for PSR should not be interchanged with the number of factors. Effective dimension is explained in more detail in Section 4. Notice that PLS is outperforming (or nearly equal to) PCR for dimension 3 to 21. We expect such results because PLS (PCR) is (not) using the response information as it carves out orthogonal constructed variables. We see from the PLS line in Figure 2 that CVSEP$(i,-i)$ is minimized for dimension 6 (CVSEP = .427—99.99% of the variation in the information) and remains near this minimum up to dimension 22. The optimal dimension for PCR is 11 (CVSEP = .429) and also remains very near this value up to a dimension of 22. We find that the proposed PSR approach has an optimal dimension of near 20; the minimum CVSEP = .307 corresponds to an effective dimension of 20.21. Incidentally this is over a 28% CVSEP$(i,-i)$ reduction beyond that of either PLS or PCR, but we refrain from making this comparison and rather compare the three "optimal" competitor methods using the 15-observation validation set.

Using the "optimal" dimension for each method, the standard error of prediction for the validation set is PSR (.417), PLS (.515), and PCR (.514); thus, the PSR method has over an 18% reduction in CVSEP (validation) when compared to PLS or PCR. One might ask what would happen if we use 20 dimensions for PLS and PCR to put dimensionality on equal footing for the three methods. For a dimension of 20, we now have PLS (.475) and PCR (.468); PSR still maintains over an 11% reduction in CVSEP (validation) over these two methods.

Figure 3 displays the corresponding (unscaled) coefficients for PLS 20 (top) and PCR 20 (bottom). The graphs on the right side represent the coefficient difference between 20 and 6 components. The general feature of these graphs is that the PLS or PCR coefficients are extremely erratic along the indexing domain. This, of course, is in part because PLS and PCR do not use any indexing information that PSR does. Remedies to smooth PLS estimates were proposed by Goutis and Fearn (1996) and are briefly discussed in Section 6. Notice also that fewer differences exist between the PLS 20 and PLS 6 when compared to the differences found between PCR 20 and PCR 6. This feature is consistent with the plot of CVSEP found on the bottom of Figure 2. We analyze the biscuit data again in Section 5.2.

In summary, we see for this example that the proposed PSR method has some promising features: (a) It is easy to interpret and compute. (b) It has strong connections to classical regression. (c) It automatically builds in smooth structure associated with the coefficient index. (d) It has a CVSEP that can be computed swiftly. We will see that the PSR compactly summarizes results, can be easily extended to non-Normal responses, and can be used with data having

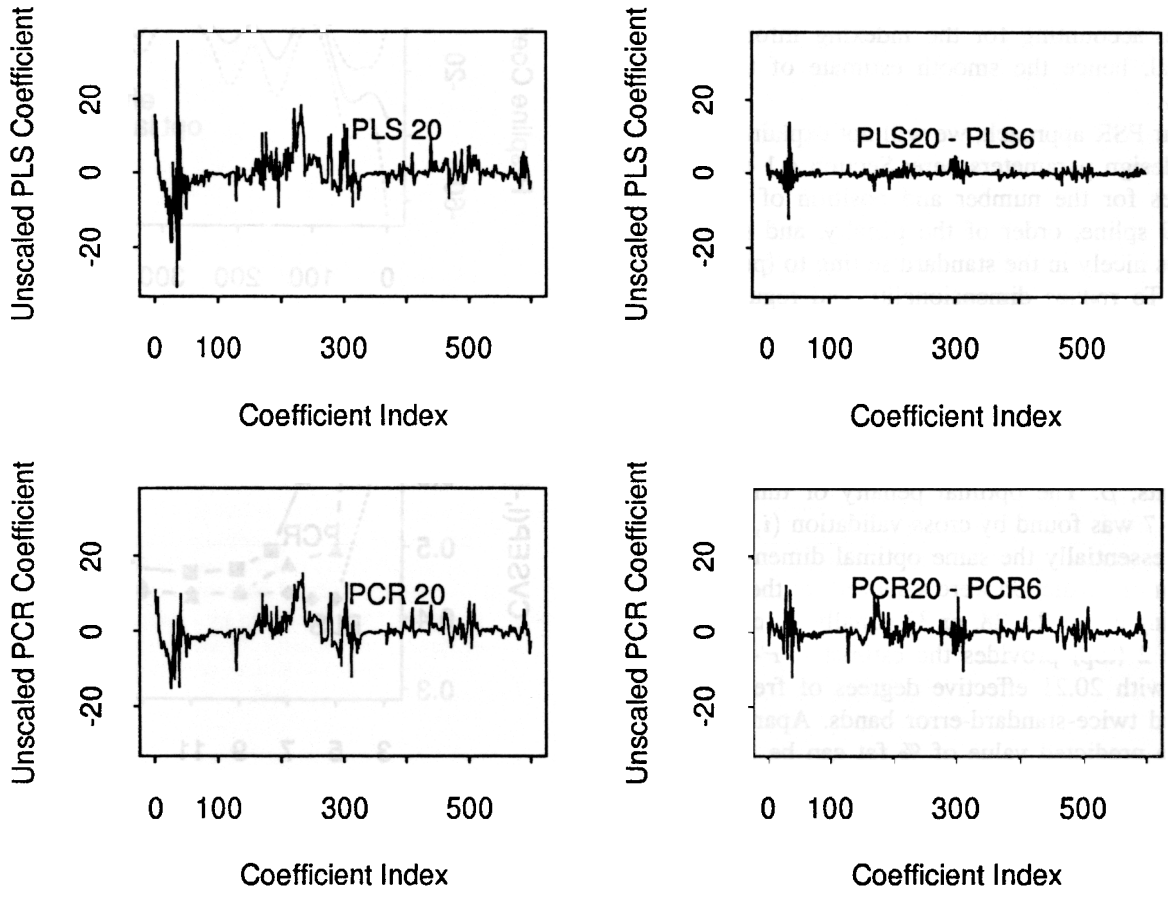

Figure 3. Biscuit: Top, Estimated PLS 20-Coefficient Vector (left) and Difference Between PLS 20 and PLS 6 Coefficients (right); Bottom, Estimated PCR 20-Coefficient Vector (left) and Difference Between PCR 20 and PCR 6 Coefficients (right).

nonsmooth signals. In the next sections, we will show how to combine B splines, discrete penalties, and ideas from GLM's to construct a relatively all-purpose, fast, and compact method for signal regression.

2. SMOOTHING THE SIGNAL COEFFICIENT VECTOR

Section 1 showed the basic idea of this article, forcing the signal coefficient vector to be smooth, and it worked well on a classic example from the multivariate calibration literature. In this section we first introduce regression on B splines to get smooth coefficients, then introduce the difference penalty for further smoothing. The idea of smoothness is not new: It was introduced by Hastie and Mallows (1993) in their discussion of Frank and Friedman (1993). H&M used the equivalent of smoothing splines, leading to large systems of equations with a size equal to the number of regressors. See Section 6 for some details. Additionally, H&M suggested that a projection of α onto a lower-dimensional basis of smooth functions might reduce the equations to a more manageable size, but they did not give details.

We use B splines to construct a smooth low-dimensional basis. They are easily computed and have excellent numerical properties. Section 4.1 contains references to the literature as well as practical guidelines for the design parameters necessary to construct B splines. Traditionally there has been one obstacle to the use of B splines as a general tool for smoothing—the choice of the number and placement of the "knots"—that is, the places where the smooth polynomial segments of the B splines join, as well as specify their limited support. Too many (few) knots will overfit (underfit) α. Optimization of the knots is a complicated nonlinear problem, leading to rather involved algorithms; see, for example, Kooperberg, Stone, and Truong (1995) or Friedman and Silverman (1989). O'Sullivan (1986) eliminated the knot-placement problem by combining the penalty for smoothing splines (Reinsch 1967) with a relatively large number of B splines based on equally spaced knots. Eilers and Marx (1996) simplified and generalized this idea by introducing a penalty on differences of the B-spline coefficients. E&M's approach, called P splines, allows an arbitrary order of the penalty with only minor changes to existing procedures for regression on B splines.

The signal-regression problem is to find a practical solution to the minimization of $S(\alpha_0, \alpha) = \|y - \alpha_0 - X\alpha\|^2$, where $y_{m\times 1}$ is the response vector and $X_{m\times p}$ is the signal-regressor matrix. Because the rows of X contain, for example, spectra, time series, spatial series, or histograms, generally p is much larger than m and the problem is ill-posed. We can only hope to get a sensible result by constraining α in some way. Our way is to require it to be smooth. By virtue that linear combinations of smooth B splines produce smooth curves, $\alpha_{m\times 1} = B_{m\times n}\beta_{n\times 1}$ can be a reasonable means to meet our requirement. We get

$$S(\alpha_0, \beta) = \|y - \alpha_0 - XB\beta\|^2 = \|y - \alpha_0 - U\beta\|^2,$$

where $U_{m\times n} = XB$. It is important to keep in mind that the matrix $U = XB$ has full column rank ($< m$). It is useful to view dimension reduction through U as the new known matrix of regressor variables used to estimate the unknown β. Of course we do not discard X, but given U, we do not need X in the estimating equations. This is advantageous because X can have hundreds of columns, whereas in our experience we have never found the need to include more than 40 columns in U. If n is chosen appreciably lower than m, the problem becomes a well-posed linear regression problem.

Varying the discrete number of knots to influence the extent of smoothing can create extra work because it is necessary to recompute the basis in B and hence $U = XB$. We avoid such schemes for knot selection. Relying on guidelines given by E&M, we choose a relatively generous number of equally spaced knots (once) that will make α more flexible than needed. Further smoothing is achieved through a difference penalty on the B-spline coefficient vector β:

$$P = \lambda \sum_{k=d+1}^{n} (\Delta_k^d \beta)^2,$$

where Δ_k^d indicates the kth difference operator of order d. The influence of the penalty is based on the magnitude of the nonnegative regularization parameter: $\lambda = 0$ provides unpenalized estimates, whereas large values of λ (e.g., 10^4) yield estimates of β near the null space of P—that is, a polynomial of degree $d-1$. The penalty is added to $S(\alpha_0, \beta)$, giving the following penalized least squares objective function:

$$S^* = S + P = \|y - \alpha_0 - U\beta\|^2 + \lambda \sum_{k=d+1}^{n} (\Delta^k \beta)^2.$$

The penalty can be written in matrix notation using D_d of dimension $(n - d) \times n$. The banded matrix D_d can be computed recursively, where D_1 has dimension $(n-1) \times n$, with $d_{ii} = -1, d_{i,i+1} = 1$, and all other elements are 0. We express an $n - d$ vector of differences as $D_d\beta$. We have

$$\begin{aligned} D_0\beta &= \beta \\ D_1\beta &= \{\beta_k - \beta_{k-1}\}, \qquad k = 2, \ldots, n, \\ D_{d+1}\beta &= D_1 D_d \beta. \end{aligned}$$

The order of the penalty d determines how many neighboring β's must *hold hands* during estimation, further ensuring smoothness. Our default penalty order is 3. Programming languages like S-PLUS and Matlab have differencing functions that make the computation of D_d trivial. Now the penalty can be expressed as

$$P = \lambda \beta^T D_d^T D_d \beta.$$

The minimization of S^* leads to the following system of equations for β:

$$(U^T U + \lambda D_d^T D_d)\beta = U^T y.$$

Note when $\lambda = 0$ that these are just the normal equations for linear regression. The penalty introduces very little

computational effort, and it is not necessary to recompute U^TU and U^Ty when λ is changed. Increasing λ makes β smoother. An optimum is searched by varying λ systematically and monitoring the prediction error, as measured by cross-validation. Details are given in Section 4.

3. THE GLR PENALIZED LIKELIHOOD FUNCTION AND *P* SPLINES

3.1 A Few Details and Notation of the GLR Log-likelihood

Consider broadening the framework to *generalized* linear signal regression,

$$g(\mu) = \alpha_0 + X\alpha = \alpha_0 + XB\beta = \alpha_0 + U\beta = \eta, \quad (1)$$

where the GLR notation has been defined in the introduction. Again (1) clearly displays H&M's mechanism to smoothly reduce dimensionality of the signal coefficient vector. We suggest that the unfamiliar reader refer to Dobson (1990), who provided a nice introductory presentation of how many statistical methods involving a linear predictor can be united through GLM's. GLR's can accommodate an entire family of response distributions. Common choices for the link function are the logarithm (for the Poisson distribution) and the logit (for the binomial distribution). The parameter estimates (for β associated with U) in most cases now must be iterated using an algorithm that resembles (iteratively) weighted least squares. In most applications, rarely does one have to derive the GLR details because tables exist (e.g., Fahrmeir and Tutz 1994, table 2.1) that specify the components and link function for common exponential family members.

Some specifics now follow for the interested reader that will lead up to the important method of scoring algorithm in Equation (3). The GLR requires that the response vector, $y_{m\times 1}$, have independent entries from a distribution in the exponential family: $f(y;\theta,\phi) = \exp[\{y\theta + c(\theta)\}/\phi + d(y)]$, where c and d are known functions and ϕ is a scale parameter. The parameter $\theta(\mu) = g(\mu)$ is the natural parameter or the canonical link function. Using (1), the dimensionality of estimation is reduced from m to $n+1$ by substituting θ with $\alpha_0 + U\beta = \eta$.

GLR estimation maximizes the log-likelihood of β instead of minimizing a sum of squares for the standard signal-regression problem presented in Section 2 (both are equivalent for Normal responses). The log-likelihood equation (here $\phi = 1$ without loss of generality) can be expressed through η as

$$l(\beta;U,y) = \sum_{i=1}^{N} \{[y_i\eta_i + c(\eta_i)] + d(y_i)\}. \quad (2)$$

Maximizing (2) yields the method of scoring iterative equations that simplifies to

$$\hat{\beta}_t = (U^T\hat{V}_{t-1}U)^{-1}U^T\hat{V}_{t-1}\hat{z}_{t-1}, \quad (3)$$

where $\hat{V} = \text{diag}(\hat{v}_{ii}) = \text{diag}\{[h'(\hat{\eta}_i)]^2/\text{var}(Y_i)\}$ and h' is the derivative of the inverse link function. The working vector has entries $\hat{z}_i = (y_i - \hat{\mu}_i)/h'(\hat{\eta}_i) + \hat{\eta}_i$. Convergence of (3) provides the (unpenalized) maximum likelihood (ML) parameter estimates. We see that estimation of β is no more difficult than (generalized) multiple linear regression. Existence of the ML solution is virtually guaranteed now because U has full-column rank and the dimension of unknown parameters has been reduced from p to n.

3.2 Penalizing the Log-likelihood and *P* Splines

The PSR approach attempts to maximize l in (2), but subject to the requirement that the estimates of adjacent β's do not differ much from each other. The modified log-likelihood now maximizes

$$l^* = l(\beta;U,y) - \tfrac{1}{2}\lambda\beta^T D_d^T D_d\beta, \quad (4)$$

where the subtrahend consists of the difference penalty ($d = 0,1,2,\ldots$) and the regularization penalty $\lambda \geq 0$. The factor $\frac{1}{2}$ is a small trick to get rid of a factor 2 that appears when differentiating the penalty. A P-spline approach transfers the decision associated with the number and position of B-spline knots to optimization of a continuous smoothing parameter.

Maximization of the penalized log-likelihood in (4) leads to small modification of the familiar scoring algorithm in (3),

$$\hat{\beta}_{\lambda,t} = (U^T\hat{V}_{t-1}U + \lambda D_d^T D_d)^{-1}U^T\hat{V}_{t-1}\hat{z}_{t-1} \quad (5)$$

It is useful to view (5) as a penalized form of an iterative weighted regression of the working vector on U, where $\hat{V}$ and $\hat{z}$ depend on the choice of λ. On convergence with fixed λ, we obtain the estimated smooth coefficient vector, $\hat{\alpha}_\lambda = B\hat{\beta}_\lambda$. Twice-standard-error bands can be constructed for $\hat{\alpha}_\lambda$ by noticing that

$$\begin{aligned} \text{var}(\hat{\beta}_\lambda) &= (U^T\hat{V}U + \lambda D_d^T D_d)^{-1} \\ &\quad \times U^T\hat{V}U(U^T\hat{V}U + \lambda D_d^T D_d)^{-1} \\ \text{var}(\hat{\alpha}_\lambda) &= B\ \text{var}(\hat{\beta}_\lambda)B^T. \end{aligned} \quad (6)$$

The preceding variance formulas are only asymptotically correct if λ is chosen a priori.

3.3 To Smooth the Coefficient Vector or to Smooth the Signal?

First off, we would like to emphasize that our proposed PSR approach does not require smooth signal regressors. Furthermore, despite some of the equivalencies (stated later) between smoothing the signal regressors and smoothing the signal coefficients, there is a difference in philosophy on what determines the optimal amount of smoothing. We do see from the term $XB\beta$ in (1) that one could conclude that the signal regressors are being smoothed. We would, however, like to stress that this does not imply that the signal regressors must be smooth but rather that the signal regressors might be smoothed first without doing much harm. We caution that the signal may be rougher than the coefficients and directly smoothing the signal regressors (optimally, based on, say, cross-validation) may require more smoothing than is necessary for an effective GLR. For example, the regressors may require 50 (unpenalized) B-spline

knots, whereas 20 knots could be sufficient to smooth the coefficient vector. Smooth signal regressors and a smooth coefficient vector are not the same. We will see our PSR approach applied to rather wild log-periodogram signal regressors in Section 5, where all that matters is that the coefficient vector is smooth. To summarize, (a) we seek a smooth coefficient vector, (b) the signal regressors do not have to be smooth, and (c) if the signal regressors are smooth, then we expect the coefficient vector to be smooth.

We now look at some details of smoothing signal regressors. Consider representing the signal-regressor matrix X by way of a coefficient matrix C (Alsberg 1993). This view is most useful when it is reasonable to assume that each row of X is of the same continuous functional type but with varying coefficients. In this way it is possible to represent or compress each discrete representation of the signal by a vector of coefficients. A guideline consists of the following: (a) For compatibility, each row of X should have the same set of basis functions over the same domain. (b) Each row of X is represented by a set of coefficients C (the compression). (c) C should be used instead of X in the GLR.

We treat each row of X as if we want to estimate it smoothly, minimizing $\|X^T - BC^T\|^2$. Here the columns of C^T are the coefficients associated with each row of X, B is a B-spline basis, and

$$\hat{C}^T = (B^TB)^{-1}B^TX^T. \tag{7}$$

Estimation is straightforward with $\hat{X}^T = B\hat{C}^T$. It is interesting to view the GLR problem in this light; we now have

$$g(\mu) = \hat{C}\beta^* = XB(B^TB)^{-1}\beta^* = \eta^*. \tag{8}$$

Equation (8) suggests that we must replace $U = XB$ with $U^* = U(B^TB)^{-1}$ in the likelihood function provided in (2). For unpenalized B-spline coefficients, this results in the iterative equation

$$\hat{\beta}_t^* = B^TB\hat{\beta}_t = B^TB(U^T\hat{V}_{t-1}U)^{-1}U^T\hat{V}_{t-1}\hat{z}_{t-1}. \tag{9}$$

Refer to Section 3.1 for notation details. Although the converged estimate of $\hat{\beta}^*$ differs from $\hat{\beta}$ by a multiple of B^TB, the estimates of $\hat{\mu}$ are identical to the ones found through (3) because the span$\{B\}$ = span$\{B(B^TB)^{-1}\}$. Equivalent results can be obtained by either projecting the matrix of signal regressors or the signal coefficient vector onto the matrix B, but strategies for choosing the optimal smoothing must be clearly defined.

4. OPTIMIZATION OF THE PENALTY

For linear models, we choose to optimize the value of the meta-parameter λ using cross-validation (CV) techniques. The idea is to leave out one observation, say with index i, fit the model with the remaining $m - 1$ observations, and then predict at the regressor variable location for the left-out observation. Cycling through all the observations, we arrive at

$$\text{CV} = \sum_{i=1}^{m}(y_i - \hat{y}_{i,-i})^2 = \sum_{i=1}^{m}\{(y_i - \hat{y}_i)/(1 - h_{ii})\}^2, \tag{10}$$

where h_{ii} is the ith diagonal element of the hat matrix $H(\lambda) = U(U^TU + \lambda D_d^TD_d)^{-1}U^T$. CV is computed almost as a by-product of the model fit; a good reference is Myers (1990). We should point out that recomputing CV for a variety of λ is actually less work than running a modest-sized multiple regression because U^TU and U^Ty stay the same at each run.

For other exponential family models, we propose to minimize the information criterion (IC):

$$\text{IC}(\lambda) = \text{deviance}(y; \hat{\beta}_\lambda) + \delta\dim(\hat{\beta}_\lambda). \tag{11}$$

When $\delta = 2$ and $\delta = \log(m)$, we have the Akaike information criterion (AIC) and the Bayesian information criterion, respectively. IC can be viewed as a compromise between goodness of fit and complexity of the model. Here $\dim(\hat{\beta}_\lambda)$ is the effective dimension of $\hat{\beta}_\lambda$; in general it will be less than n because the penalty constrains the freedom of $\hat{\beta}$. Following Hastie and Tibshirani (1990), we use the trace of the "hat" or smoother matrix that follows from convergence of (3). Denote

$$\dim(\hat{\beta}_\lambda) = \text{trace}\{\hat{H}(\lambda)\}, \tag{12}$$

where $\hat{H}(\lambda) = U(U^T\hat{V}U + \lambda D_d^TD_d)^{-1}U^T\hat{V}$ provides

$$\hat{\eta}_\lambda = \hat{H}(\lambda)\hat{z}. \tag{13}$$

The underlying interpretation of $\hat{H}(\lambda)$ is that it transforms the rough working vector $\hat{z}$ into the smooth $\hat{\eta}_\lambda$. In practice, especially with many observations, it is advantageous to compute the trace from the cyclical permutation,

$$\text{trace}\{H(\lambda)\} = \text{trace}\{(U^T\hat{V}U + \lambda D_d^TD_d)^{-1}U^T\hat{V}U\}, \tag{14}$$

because of the smaller size of the matrices involved. With a binomial or Poisson response, the computation of the deviance is straightforward. Thus, we see that IC can be computed swiftly. With a Normal response, an estimate of the variance of Y is needed; it is often computed using the residuals with a correction for effective degrees of freedom as

$$\widehat{\text{var}}(Y) = \sum_{i=1}^{m}(y_i - \hat{\mu}_{i\lambda})^2/\{m - \text{trace}[H(\lambda)]\} \tag{15}$$

(Hastie and Tibshirani 1990). For most practical purposes, it is sufficient to solve the problem for several values of λ and search for the minimum value of CV or IC.

General Recipe: B Splines (Knots, Degree), Penalty (Order, Optimization)

Not all readers will be familiar with B splines; the basic reference is de Boor (1978), and we find Dierckx (1993) particularly lucid. Algorithms for construction of B splines are routine in some statistical packages. De Boor (1978) provided an algorithm to compute B splines for a general placement of knots. For the interested reader, E&M presented a section "B-splines in a Nutshell" and also presented how the B-spline algorithm simplifies nicely for

equally spaced knots. To compute B splines, three things are needed—(1) the degree of the spline q, (2) the knots, and (3) the abscissa where the values of the splines are needed, indicated by u. Each column in the matrix of signal regressors $X_{m\times p}$ corresponds to a physical quantity, like wavelength in an NIR spectrogram or a frequency in a power spectrum. It is possible to use the original scale of the instrument (wavelength, frequency), but this is not necessary. Shifting and scaling of both u and the knots by the same amount leaves the B splines unchanged. So we recommend in practice to work with an easy nominal scale for u, such as $u_j = j - .5$ for $j = 1, \ldots, p$. For n' equally spaced B-spline (degree q) intervals, then the $(n = n' + q)$ knots are chosen as $t_s = sp/n'$ for $s = -q, \ldots, p + q$. To put u in the context of the original scale (defined as v), $u_j = (v_j - v_1)/(v_2 - v_1) + .5$.

1. Choose enough equally spaced B-spline knots ($n < m$ so U is full rank) to modestly exceed the dimension of the signal coefficient vector; that is, allow the smoother to be more flexible than it needs to be. If nothing is known about this dimension, then start with $n = 40$ (provided that $m > 40$).
2. We use cubic B splines (third degree) as a default. We find this is suitable for a variety of applications in practice.
3. Our default order for the penalty is 3 and in our experience rarely needs to exceed 4.
4. Perform a logarithmic grid search on the nonnegative penalty parameter λ and compute CV (for Normal responses) or IC (for non-Normal responses, such as binomial or Poisson).
5. The (effective) dimension of the regression is a function of λ and is computed through the trace of the (effective) *hat* matrix.
6. Plot CV or IC as a function of λ or (effective) dimension. Choose an optimal λ at the minimum, if it exists.
7. If the optimum effective dimension is considerably less than 40, then if desired the number of B-spline knots in the first step can be decreased. The value of the optimal penalty parameter will differ, but the associated deviance and effective degrees of freedom will be in the same neighborhood depending on the intricacy of the grid search.

5. ILLUSTRATIVE EXAMPLES

5.1 GLR Examples

Our first GLR example revisits data from Le Cessie and van Houwelingen (1992) (C&H) that explored the relationship between DNA content in ovarian cancer cells and the probability of surviving 24 months. There were 81 patients, but 11 patients' survival information was missing completely. Thus the analysis was restricted to those 70 patients whose survival ($Y = 0$; 28 patients) or death ($Y = 1$; 42 patients) status was available after two years. For each patient, the regressor information was the amount of DNA (in about 50 to 250) cancer cells, summarized by a (37-class) relative-frequency histogram. The readings ranged from 0C to 8C, with 1C corresponding to the amount of DNA in a haploid cell. For healthy persons we expect a large peak at 2C and a small peak at 4C. For patients with advanced ovarian cancer, a considerable fraction of cells can be at levels other than 2C or 4C. Figure 4 displays a randomly selected patient's histogram for both $Y = 0$ and $Y = 1$.

As pointed out by C&H, these data wreak havoc on standard binary regression approaches because neighboring heights of the histogram are highly correlated, and, furthermore, we have many regressors ($p = 37$) relative to the number of patients ($m = 70$). Le Cessie and van Houwelingen proposed a clever solution to this ill-conditioned problem that transformed the regressors so that a ridge solution penalizes first-differences in the coefficients. We consider the logit model

$$\log \frac{p_i}{1 - p_i} = \sum_{j=1}^{37} x_{ij}\alpha_j,$$

where p_i is the probability of 24-month survival and x_{ij} is the relative-frequency histogram for subject $i = 1, \ldots, 70$. The intercept term α_0 is not needed because rows of X sum to unity. We essentially reproduced the results of C&H with the special case of a first-order penalty and a penalized cubic B-spline basis. We choose, however, to illustrate our method with cubic B-spline basis (our default). With these data, ample flexibility for the signal coefficient vector is achieved with 10 knots. We should point out that, when using 20 knots, the deviance and effective degrees of freedom were nearly reproduced based on this optimal AIC. A third-order difference penalty (our default) on adjacent coefficients is used. The optimal $\lambda = .001$ based on AIC. Parameter estimation converged in three iterations. The resulting fit is also displayed in Figure 4 with twice-standard-error bands. There are 4.76 effective degrees of freedom with a deviance of 80.36 (on 65.24 residual degrees of freedom), and a 70.0% correct classification. Because healthy persons have mostly 2C and 4C DNA, the negative coefficients near these values are expected. Figure 4 additionally displays the PSR fit for an increased $\lambda = .1$, which has increased smoothness resembling a quadratic fit. This illustrates that, as λ increases, then P splines approach the null space of the difference penalty, a polynomial of order $d - 1$.

Our next example is one of phoneme classification described by Hastie, Buja, and Tibshirani (1995) and illustrates that the signal regressors do not have to be smooth to use our method. The data were log-periodograms of 32 ms time series of continuous speech. The database contains two speech frames of each phoneme from each speaker. The speech frames were represented by 512 samples at a 16-kHz sampling rate. Land and Friedman (1996) selected 160 speakers randomly from the 437 male subjects and only took the first 150 frequencies from each subject. The response variable is the phoneme *ao* (as in *water*, $Y = 1$) and *aa* (as in *dark*, $Y = 0$). The model of interest is

$$\log \frac{p_i}{1 - p_i} = \alpha_0 + \sum_{j=1}^{150} x_{ij}\alpha_j,$$

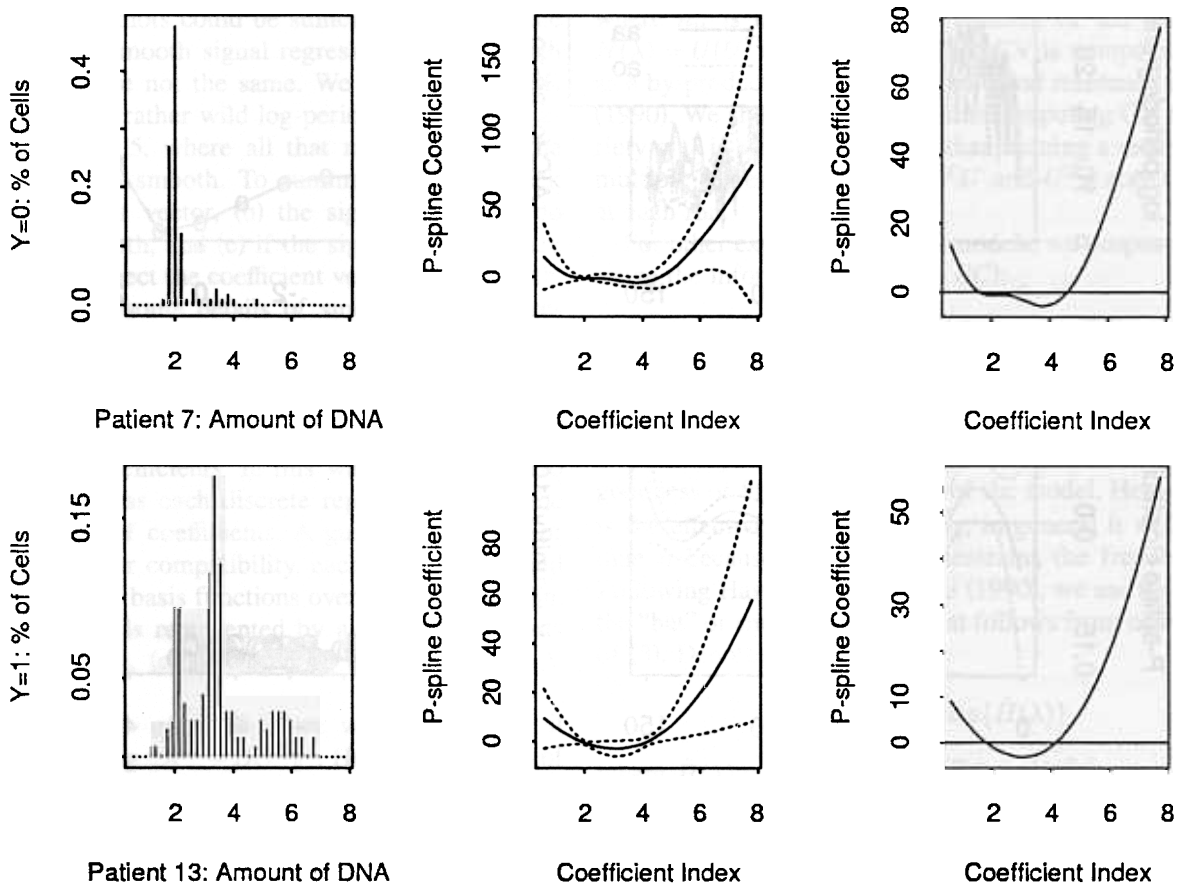

Figure 4. Typical DNA Regressor Information for Y = 0, 1 (top and bottom, left); PSR Coefficient Vector (with and without twice-standard-error bands) for Optimal λ = .001 (top, center and right); Oversmoothed PSR Coefficient Vector (with and without twice-standard-error bands) Using λ = .10 (bottom, center and right).

where p_i is the probability of *aa* and x_{ij} is the log-periodogram for subject $i = 1, \ldots, 160$. They investigated a variety of estimation techniques—cubic spline, variable fusion, ridge, and PLS; we consider PSR. Data also exist for distinguishing between *aa* and *iy* (as in *she*), but these are fairly easy to distinguish. We use L&F's *aa*–*ao* data. Figure 5 provides typical and extremely nonsmooth log-periodograms of the subjects.

Using a cubic (our default) P spline (13 knots were sufficient) and a third-order penalty (our default), the optimal value of $\lambda = 10$. See Figure 5 for the AIC plot. The deviance is 105.8 (on 151.1 residual degrees of freedom); the percent correct classification is 84.37%. There are 8.9 effective degrees of freedom. Figure 5 also displays the estimated coefficient vector with twice-standard-error bands. Notice that these bands suggest that there appears to be little information beyond (approximately) the 75th frequency. On the bottom right, we provide a plot of the *aa*–*ao* response versus $\hat{\eta}$ to illustrate the separation provided by the PSR coefficient vector. We should point out that the *iteratively reweighted partial least squares* method proposed by Marx (1996) does not converge for these data.

5.2 Further Comparisons of PSR to PLS in the Standard Setting

In Section 1 we illustrated PSR on a classic signal-regression example, and our method was shown to perform well compared to PCR and PLS. The latter is well established in chemometric theory and practice, so if we wish to challenge it, then one dataset is not enough. In addition to the % fat response, we also consider modeling the other biscuit ingredients (% flour, % sucrose, % water) using NIR spectra information. Moreover, we explore two other standard linear signal datasets (gasoline and wheat). Philip Hopke maintains an FTP site where he collects interesting data for chemometric applications. Both the gas and wheat datasets are publicly available from anonymous ftp://sun.mcs.clarkson.edu/pub/hopkepk/data/kalivas. First we consider relating NIR spectral information of $m = 60$ gasoline samples to their $Y =$ *octane number*. Figure 6 provides the gasoline NIR spectra (original and first-differenced), which range from 900 nm to 1,700 nm (in 2 nm intervals). Thus the discrete representation of the observed signals is in $X_{60\times 401}$. Notice that, unlike in the biscuit example, this spectra is not entirely smooth and has somewhat sharp spikes near 1,200 nm and also at 1,400 nm, especially when differenced. We use the differenced spectra. Recall (Sec. 3.3) that the PSR approach does not require that the spectra itself be smooth; smoothness is only required in the associated regression vector.

The second additional standard example uses NIR spectra of $m = 100$ wheat samples measured from 1,100 nm to 2,500 nm (in intervals of 2 nm) to predict responses of *protein content* and *moisture content*. The wheat spectra are

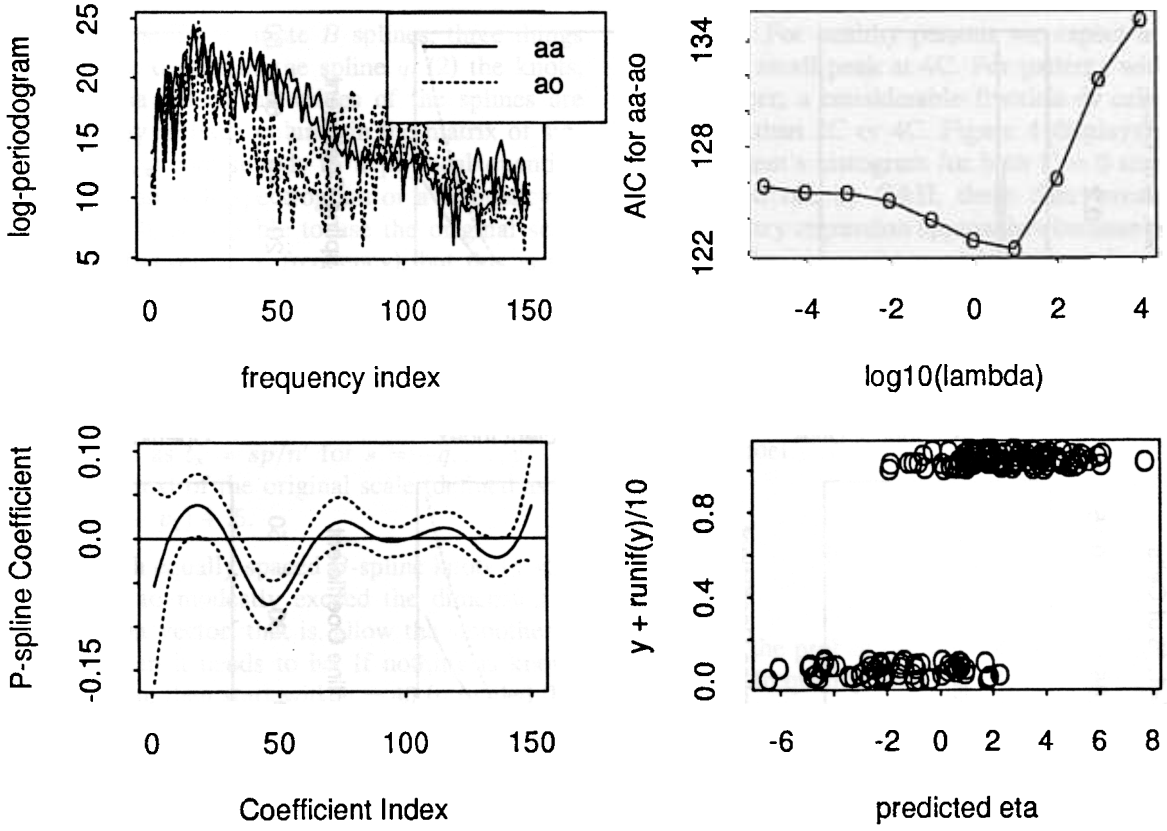

Figure 5. Typical Log-periodograms for aa–ao (top, left); AIC Plot as a Function of λ (top, right); PSR Coefficient Vector for Optimal λ = 10 With Twice Standard Error Bands (bottom, left); aa–ao Response as a Function of $\hat{\eta}$ (bottom, right).

smooth as seen in Figure 7; again we use first-differenced spectra.

In an effort to again compare our PSR to other methods, we decided to randomly split the observations into two groups: Two-thirds of the observations are used for training and the other one-third as a validation set. For ease of reproducibility, we chose every third observation (i.e., numbers 3, 6, 9, ...) as the validation set. We found that PLS was the only serious competitor to PSR; furthermore, PCR becomes impractical as the number of channels becomes large. Table 1 summarizes the optimal values of cross-validated standard errors (CVSEP) in three settings—(1) with all data, (2) with the two-thirds training data, (3) predicting the one-third validation set with the optimally trained model.

Notice that PSR is performing better than PLS in terms of CVSEP in all cases when using all the data. Based on these prediction results (cols. 2 and 3 of Table 1), however, there is no reason to prefer PSR over PLS or vice versa. We do find that PSR is a strong competitor to PLS when validating trained (optimal) models (outperforming for biscuit–fat, gasoline–octane, and wheat–protein). We note that we also repeated this exercise using the original (undifferenced) spectra. Based on CVSEP (one-third validation set), PSR was still a competitor to PLS, again outperforming for three of the seven responses (in this case for biscuit–water, gasoline–octane, and wheat–moisture).

6. A BRIEF SURVEY OF OTHER CURRENT APPROACHES

Beyond PLS and PCR, there have been several successful attempts to introduce smoothness ideas into the signal-regression problem. None of them have directly addressed the GLM framework to our knowledge. Here we give a brief survey, using the following notation: X is an $m \times p$ matrix of observed signals, one for each dependent-variable observation y_i (where m can be greater than p). Our presentation of these other methods is ordered such that we are moving from ideas of penalized regression (most similar to P splines) to smoothing the signal regressors (more dissimilar to P splines). Regularization is needed in any case to remove the singularity of the problem.

Le Cessie and van Houwelingen (1992) (C&H) estimated probabilities of binary outcomes; the rows of X are moderately sized histograms (37 cells) of DNA fragments. If $p_i = \Pr(y_i = 1)$, they used the GLM provided in (16). When estimating α by the standard GLM iterative algorithm, ill-conditioned systems of equations are found. To remedy this, the following penalized log-likelihood function was used with success:

$$l_{\mathrm{CH}}(\alpha; y, X) = l(\alpha; y, X) - \frac{1}{2}\lambda \sum_{j=2}^{p} (\alpha_j - \alpha_{j-1})^2. \quad (18)$$

The optimal value of λ was found by cross-validation. It is interesting to note that our P-spline approach is exactly equivalent to the C&H approach if we use zero-degree B splines with 37 equally spaced knots and a first-order penalty.

Hastie and Mallows (1993) proposed a smooth regression vector for the least squares problem. They interpreted the coefficients as samples from a continuous function, $\alpha_j =$

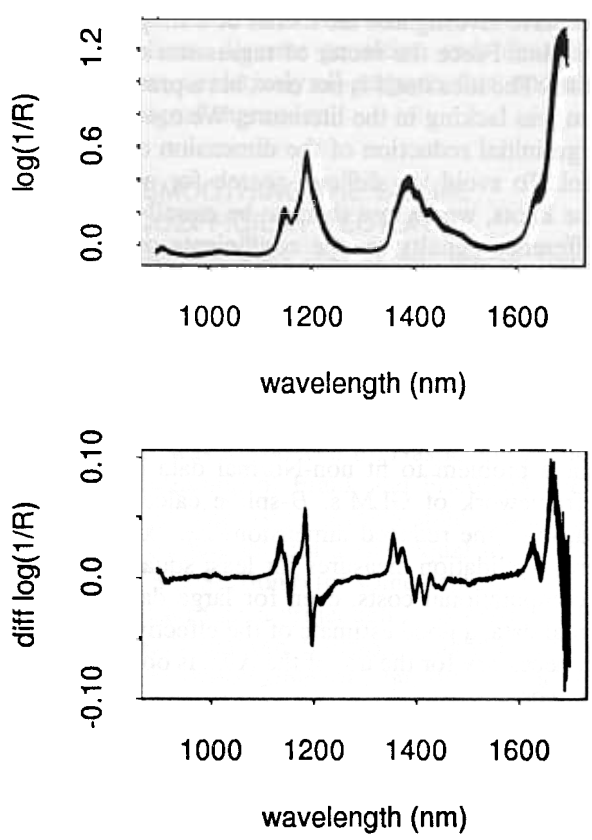

Figure 6. Gasoline: Each Curve (m = 60) Represents p = 401 NIR Wavelengths (top, original spectra; bottom, first-differenced spectra).

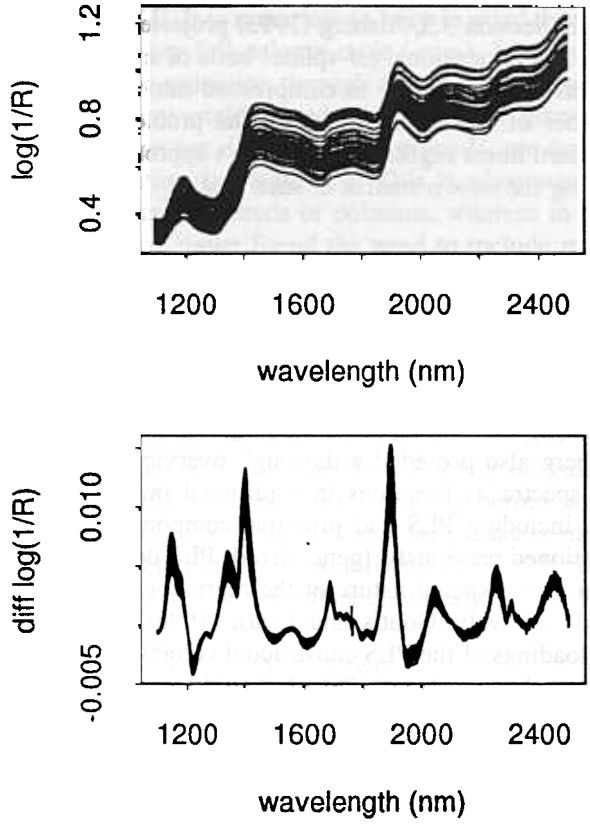

Figure 7. Wheat: Each Curve (m = 100) Represents p = 701 NIR Wavelengths (top, original spectra; bottom, first-differenced spectra).

$\alpha(t_j)$, and minimized

$$L_{\text{cubic}}(\alpha) = \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{p} x_i(t_j)\alpha(t_j) \right)^2 + \lambda \int (\alpha''(t))^2 \, dt. \quad (19)$$

The penalty is borrowed from the seminal work of Reinsch (1967). The equations that resulted were of order $p \times p$. As we have stressed, H&M recognized this and proposed to reduce the dimensionality of the signal coefficient vector using an (unspecified) basis B of smooth functions with dimension $p \times n$, such that $\alpha = B\beta$, where $n < \min(m, p)$. Thus, $E(Y|X) = XB\beta$, and solutions for α can be obtained by regressing Y on $U = XB$. Our P-spline approach provides a simple basis, combined with a continuously variable penalty.

Land and Friedman (1996) have also done some interesting work in the area of regression, complementing Frank and Friedman (1993). They proposed a *variable fusion* method that also can be viewed as a penalized least squares problem. L&F considered minimization of

$$L_{\text{fusion}}(\alpha) = \sum_{i=1}^{m} \left(y_i - \alpha_0 - \sum_{j=1}^{p} x_i(t_j)\alpha(t_j) \right)^2 + \lambda P(\{|\alpha_j - \alpha_{j+1}|, j = 1, \ldots, p-1\}), \quad (20)$$

where α_0 is the intercept term. The addendum in (20) ensures that neighboring coefficients do not differ too much from each other. Two penalties P were considered—(1) zero-order fusion of estimated coefficients resulting in a piecewise constant solution, $\lim_{\gamma \to \infty} \sum_{j=1}^{p-1} |\alpha_j - \alpha_{j+1}|^\gamma$, and (2) first-order fusion penalizing the sum of absolute-values first-differences of the coefficients, $\sum_{j=1}^{p-1} |\alpha_j - \alpha_{j+1}|$. They have extended these approaches to accommodate binary data.

The previous approaches explicitly demanded smoothness of the vector α. An alternative approach (Brown and Mäkeläinen 1992) exploited ideas from Bayesian estimation with a special prior for the spectra. They imposed an autoregressive structure to model their smoothness. This led to a system of augmented normal equations of size $p \times p$. As we

Table 1. Results of the Comparison PLS and PSR, for Cross-validation With all Data, for Cross-validation on a Training Set of Two-thirds of the Data, and for Predicting the Other Third of the Data With the Trained Model

	All data		Training		Prediction	
Data	PLS	PSR	PLS	PSR	PLS	PSR
Biscuit–fat	.421	**.325**	.573	**.475**	.296	**.293**
Biscuit–flour	.984	**.928**	.670	.673	1.323	1.440
Biscuit–sucrose	1.074	**.941**	1.034	1.065	1.283	1.380
Biscuit–water	.281	**.251**	.329	**.291**	.305	.458
Gasoline–octane	.224	**.223**	.283	**.252**	.264	**.191**
Wheat–moisture	.216	**.221**	.220	.235	.227	.246
Wheat–protein	.429	**.424**	.457	.528	.550	**.515**

NOTE: The numbers are RMS values of the differences between predicted and actual values (CVSEP). Differences of the spectra were used.

saw in Section 3.3, Alsberg (1993) projected the rows of X on a low-dimensional (B-spline) basis of smooth functions. In this way each row is compressed into a much smaller number of coefficients, making the problem amenable to standard linear regression. Alsberg's approach amounted to finding the $m \times n$ matrix C such that

$$L_A = \|y - C\beta\|^2, \tag{21}$$

where β has only n elements. No penalty was used. A similar approach was used by Denham and Brown (1993), although they modeled spectra from known chemical compositions instead of the reverse problem presented here. Alsberg also provided a thorough overview of manipulating spectra as functions in traditional multivariate methods, including PLS and principal-component analysis. As mentioned previously, (generalized) PLS does not take into account the spatial nature of the regressor index.

Alternatively, Goutis and Fearn (1996) recognized that the loadings of the PLS constructed variables (when plotted against the wavelength index) should resemble the smooth wavelength regressor information. They suggested using a Reinsch roughness penalty on the loadings while still preserving orthogonality of PLS components. This approach does achieve smoothness but adds another layer of work to an already highly nonlinear algorithm. Moreover, the smoothness penalty parameters must be optimized for each component.

Each of the approaches described previously goes a long way in solving the problem of signal regression, but each also has its shortcomings. Le Cessie and van Houwelingen's (1992) method is very general, but it leads to a large system of equations. Hastie and Mallows (1993) showed how to reduce dimensionality in principle but did not give details. Land and Friedman (1996) also had to confront large systems and further had the computational problem associated with the L_1 norm in the penalty. Alsberg (1993) as well as Denham and Brown (1993), reduced the dimension of the problem right from the start by exploiting the smoothness of the spectra but did not specify a smooth regression curve. Goutis and Fearn (1996) modified the PLS algorithm to ensure smoothness of each orthogonal component. Most authors gave little advice on choosing the dimension of the projection basis or optimizing the penalty.

7. DISCUSSION AND FUTURE RESEARCH

We believe that the results presented go far beyond the presentation of a new estimator in the standard setting. We present a rather simple model that easily accommodates the GLM framework with severely ill-conditioned regressors. The dimension of the model is dramatically and intuitively reduced, and estimation is not only fast but so is cross-validation. Our examples (not contrived) demonstrate competition to PLS/PCR with cross-validation. In many examples, smooth regression vectors make more sense (at least to us) than the extremely erratic coefficients associated with PLS/PCR. All in all, we think that our method stands the test quite gracefully.

We have investigated the merits of a simple idea in signal regression: Force the vector of regression coefficients to be smooth. The idea itself is not new, but a practical implementation was lacking in the literature. We use B splines to get a large initial reduction of the dimension of the regression model. To avoid the difficult search for optimal positions of the knots, we choose them to be equally spaced but use a difference penalty on the coefficients of the B splines. This is the P-spline signal-regression approach, blending the ease of B splines with continuous control over smoothness.

Because we stay very near to classical regression modeling, the arsenal of well-established outlier-detection methods and influence diagnostics is accessible. Furthermore, it is not a problem to fit non-Normal data by transplanting the framework of GLM's. B-spline calculations are fast. Because of the reduced dimensionality, exact computation of cross-validation measures for least squares adds little to the computational costs, even for large datasets. For non-Normal data, a good estimate of the effective model dimension, necessary for the use of the AIC, is obtained with little extra work.

At the heart of the signal-regression problem is ill-conditioned data, often with more regressors than observations. Constraining of the coefficients can help. The constraint for PCR is the projection of the regressors onto a lower-dimensional subspace, as it is for PLS, which additionally takes into account the response vector. Simply put, in any case you cannot expect to get hundreds and hundreds of meaningful estimated coefficients from dozens of observations. In PCR and PLS, the estimated coefficients suggest much relevant detail, but they are combinations of a low number of rough basis vectors. If one were to interpret the meaning of these details, then it would be in part dubious. For PLS and PCR, rough spectra lead to rough coefficients, having peaks or troughs where the spectra show peaks or troughs, suggesting meaningful detail where it is only spurious.

Alternatively, the P-spline signal-regression approach finds a smooth vector that does not allow any more detail than the data permit. PSR takes another route by constraining the regression vector itself: It is forced to be smooth. We wish to stress again that nonsmooth spectra do not imply nonsmooth regression coefficients. We do not claim that our smoothness penalty is the final answer, but it is sensible. It is only one type of penalty.

Of course, one can imagine that, for some datasets, it would be best if the regression coefficient vector had some nonsmooth behavior (kinks, jumps, or narrow peaks). Our present experience has shown no indication for need of such modifications. We expect that it would be rather difficult to detect the need for local nonsmoothness because of the underdetermined estimation problem. As noted, constraints of some type are needed to find any solution at all, and technically almost any constraint can produce a model with some predictive ability. The data by itself has little power to steer results.

In Section 6, we compared our approach to several applications of smooth regression in the literature. Two dif-

ferent interpretations are in use: One (H&M), like ours, imposes smoothness on the regression coefficients, but the other (Alsberg) emphasizes prior smoothing of signals (or orthogonal components derived from them). We think that the former interpretation is the more fruitful one: The signals do not have to be smooth for smooth regression coefficients. When signals are being smoothed, the temptation exists to replace the raw signals with the fit of smoothed signals as a figure of merit. This leads us in the wrong direction: It is the predictive value for the dependent variable that is often of most interest. There is some similarity with the basics of PCR and PLS: The former uses only X to construct regressors, but the latter takes both X and y into account. If prior smoothing is occupied with only X, the results will be less than optimal.

Many interesting subjects call for further research in this area. Obviously many more datasets have to be analyzed to compare performance of PSR to PLS and PCR. It remains to be seen whether or not the smooth coefficients of PSR methods have advantages over PLS in calibration transfer problems—for example, robustness of prediction against shifting, scaling, and warping of the spectra. Future work can consider settings with responses having multiple signals or other smooth additive components or perhaps varying coefficients, combining this research with that of Marx and Eilers (1998). A further step would be to introduce ideas from robust regression, like bounded influence functions or L_1 estimation.

Software for signal regression with P splines can be obtained from the authors.

ACKNOWLEDGMENTS

We thank Fearn, Hastie, Hopke, Land, le Cessie, and van Houwelingen for the use of their data. We also are indebted to editor Max D. Morris, the anonymous associate editor, and three anonymous referees for their thorough and constructive comments leading to a significantly improved article.

[Received May 1997. Revised August 1998.]

REFERENCES

Alsberg, B. K. (1993), "Representation of Spectra by Continuous Functions," *Journal of Chemometrics*, 7, 177–193.

Brown, P. J., and Mäkeläinen, T. (1992), "Regression, Sequenced Measurements and Coherent Calibration," in *Bayesian Statistics 4*, eds. J. M. Bernardo et al., Oxford, U.K.: Clarendon Press, pp. 97–108.

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

Denham, M. C., and Brown, P. J. (1993), "Calibration With Many Variables," *Applied Statistics*, 42, 515–528.

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

Dobson, A. J. (1990), *An Introduction to Generalized Linear Models*, London: Chapman and Hall.

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.

Fahrmeir, L., and Tutz, G. (1994), *Multivariate Statistical Modeling Based on Generalized Linear Models*, Berlin: Springer-Verlag.

Frank, I. E., and Friedman, J. H. (1993), "A Statistical View of Some Chemometric Regression Tools," *Technometrics*, 35, 109–148.

Friedman, J. H., and Silverman, B. W. (1989), "Flexible Parsimonious Smoothing and Additive Modeling" (with discussion), *Technometrics*, 31, 3–39.

Goutis, C., and Fearn, T. (1996), "Partial Least Squares on Smooth Factors," *Journal of the American Statistical Association*, 91, 627–632.

Hastie, T., Buja, A., and Tibshirani, R. (1995), "Penalized Discriminant Analysis," *The Annals of Statistics*, 23, 73–102.

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.

Helland, I. S. (1988), "On the Structure of Partial Least Squares Regression," *Communications in Statistics, Part B—Simulation and Computation*, 17, 581–607.

Kooperberg, C., Stone, C. J., and Truong, Y. K. (1995), "Hazard Regression," *Journal of the American Statistical Association*, 90, 78–94.

Land, S. R., and Friedman, J. H. (1996), "Variable Fusion: A New Method of Adaptive Signal Regression," Technical Report 114, Stanford University, Dept. of Statistics.

le Cessie, S., and van Houwelingen, J. C. (1992), "Ridge Estimators in Logistic Regression," *Applied Statistics*, 41, 191–201.

Martens, H., and Næs, T. (1989), *Multivariate Calibration*, New York: Wiley.

Marx, B. D. (1996), "Iteratively Reweighted Partial Least Squares Estimation for Generalized Linear Regression," *Technometrics*, 38, 374–381.

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

Marx, B. D., and Smith, E. P. (1990), "Principal Component Estimation for Generalized Linear Regression," *Biometrika*, 77, 23–31.

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: PWS-Kent.

Osborne, B. G., Fearn, T., Miller, A. R., and Douglas, S. (1984), "Application of Near Infrared Reflectance Spectroscopy to the Compositional Analysis of Biscuits and Biscuit Dough," *Journal of Scientific Food Agriculture*, 35, 99–105.

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

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

Stone, M., and Brooks, R. J. (1990), "Continuum Regression: Cross-Validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression," *Journal of the Royal Statistical Society*, Ser. B, 52, 237–269.

Wold, H. (1975), "Soft Modeling by Latent Variables: The Nonlinear Iterative Partial Least Squares Approach," in *Perspectives in Probability and Statistics, Papers in Honour of M.S. Bartlett*, ed. J. Gani, London: Academic Press, pp. 117–142.