Statistical Modelling 2008; **8**(4): 367–383

Sharpening P-spline signal regression

Bin Li and Brian D. Marx
Department of Experimental Statistics, Louisiana State University, USA

Abstract: We propose two variations of P-spline signal regression: space-varying penalization signal regression (SPSR) and additive polynomial signal regression (APSR). SPSR uses space-varying roughness penalty according to the estimated coefficients from the partial least-squares (PLS) regression, while APSR expands the linear basis to polynomial bases. SPSR and APSR are motivated in the following two scenarios, respectively: (i) some region(s) of the regressor channels contain more useful information for prediction than others and (ii) the relationship between the response and regressor channels is highly nonlinear. We also extend the methods to the generalized linear regression setting. As illustration, we apply the methods to two published data sets showing highly competitive performance.

Key words: multivariate calibration; P-splines; partial least squares

Received June 2007; **revised** September 2007; **accepted** October 2007

1 Introduction

Nowadays, a wide range of scientific studies often provide data that are obtained in a form of curves and the observed data consist of sets of curves sampled on a fine grid. Ramsay and Silverman (2005) view "functional data analysis" (FDA) as an inclusive term for the analysis of data for which the ideal units are curves, and provide a clear overview on the foundation and applications of FDA. Although functional data share many common principles with multivariate data, they are different in the "atom" of a statistical analysis. From the perspective of FDA, each observed instance is represented as a function in an infinite-dimensional space, while each observation is only a realization of the function at many (equally spaced) digitized points. One common resolution is a "filtering approach", which projects the functional data onto a finite-dimensional space, i.e. each instance is approximated by a linear combination of a finite number of basis functions. The data are then represented by the resulting basis coefficients. The type of problem we consider in this paper is "functional linear models for scalar responses", in which one wants to model and predict the

Address for correspondence: Bin Li, Assistant Professor in Department of Experimental Statistics, Louisiana State University. E-mail: bli@lsu.edu. Telephone: (225) 578-1343. Address: Room 61 Agricultural Administration Building, Louisiana State University, Baton Rouge, LA 70803-5606 USA.

 10.1177/1471082X0800800403

scalar response from a sample curve via its functional form. For various issues and references for the recent development in this area, we refer the readers to Chapter 15 of Ramsay and Silverman (2005). In chemistry and engineering, this area is particularly motivated through multivariate calibration (MVC) and signal regression problems.

MVC is a signal regression problem with a rich set of ordered regressors, often equally spaced digitizations of a curve. In the chemometric community, the regressor "signals" are usually optical spectra, and the response is a concentration of a chemical analyte. The objective is one of reliable (external) prediction, that is train the model for an "optimal" vector of regressor coefficients, so that given a future optical spectra, the unknown chemical concentration, Y, can be predicted well. Depending on precision, the number of regressors (p) can be in the hundreds or even thousands, whereas in comparison the number of training observations (m) is relatively less, e.g. in the dozens. Hence from a statistical perspective, the problem is ill-posed. To give a quick idea of the ill-conditioning, Figure 1 (top panel) is a visual representation of a 3×512 regressor matrix, i.e. the X matrix for three observations and 512 regressors. Section 2 provides more data structure details. Martens and Næs (1989) give an excellent review on MVC. To reach out to all readers, we refer to the MVC problem as signal regression from now on.

As outlined in Eilers and Marx (2003), two general approaches have been used to make the signal regression problem well-posed: (a) reduction of the regression bases and (b) penalized estimation. The first approach can use, for example, principal component regression (PCR), partial least-squares (PLS) regression, or projection onto B-splines. The second approach of penalized regression comes in many forms, two of which are (i) ridge regression, which shrinks the regression coefficients towards zero, and (ii) penalized signal regression (PSR), which forces the vector of coefficients to vary smoothly with the signal's ordered regressors. Frank and Friedman (1993) provided an excellent overview of chemometric modeling tools. We will see (in Section 6) that even more recently a great deal of attention has been given to variants of support vector machine (SVM) approaches.

In this paper, we develop extensions to the P-spline signal regression approach (Marx and Eilers, 1999, 2005), which can have significant improvements in external prediction error for benchmark data sets in the literature. We consider a variety of modifications to PSR, including additive polynomial signals and variable weighting schemes that outperform PCR, PLS, and standard PSR regression. Further, in some cases, we find our proposed methodology to have prediction performance that is extremely competitive to that found in neural networks and SVMs, while at the same time being much easier to optimally tune and understand. As a special case, we additionally suggest a compromise approach between PLS and P-spline signal regression, using features of both to boost prediction performance.

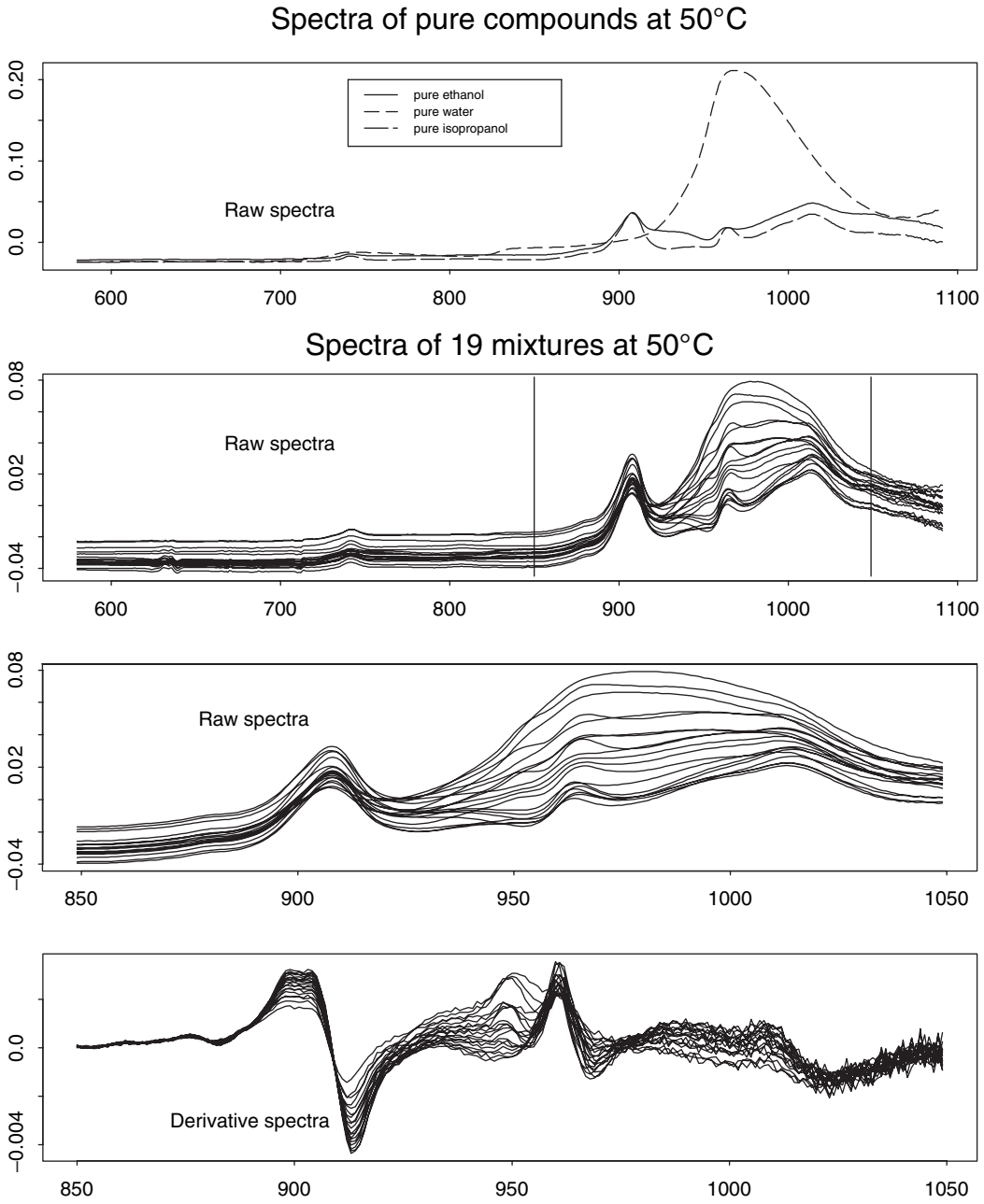

Figure 1 Measured spectra at 50 °C. The top panel provides the raw spectra for the pure components. The second and third panels displays the raw spectra for the 19 mixtures. The bottom panel gives derivative spectra.

2 The motivating example

Wülfert, *et al.* (1998) presented an experiment that involved mixtures of ethanol, water and isopropanol prepared according to the design given in Figure 2. Specific details can be found in their article, and the data are publicly available. Each of the 19 mixtures, as well as the three pure compounds had measured spectra under several temperature conditions: 30, 40, 50, 60, and 70 °C (±0.2 °C), which were short-wave near-infrared spectra ranging from 580 to 1091 nm, by 1 nm. There is a separate spectrum for each mixture and temperature combination (in total $(19+3) \times 5 = 110$). Wülfert *et al.* (2000) and Eilers and Marx (2002) took the temperature information into account in their models. Recently, Thissen *et al.* (2004) and Üstün *et al.* (2005) applied global (ignoring temperature information but using all the data) SVM approaches to tackle the data. Due to absorption and noise, all data analyses are performed only using an interior region of 200 wavelengths, i.e. 850 to 1049 nm. Figure 1 provides the spectra for the middle temperature 50 °C. The top panel displays raw spectra for pure components. The second panel displays the raw spectra for the 19 mixtures, where the vertical bars indicate the central 200 wavelength region, which is further displayed in the third panel. Finally, the bottom panel displays the spectra

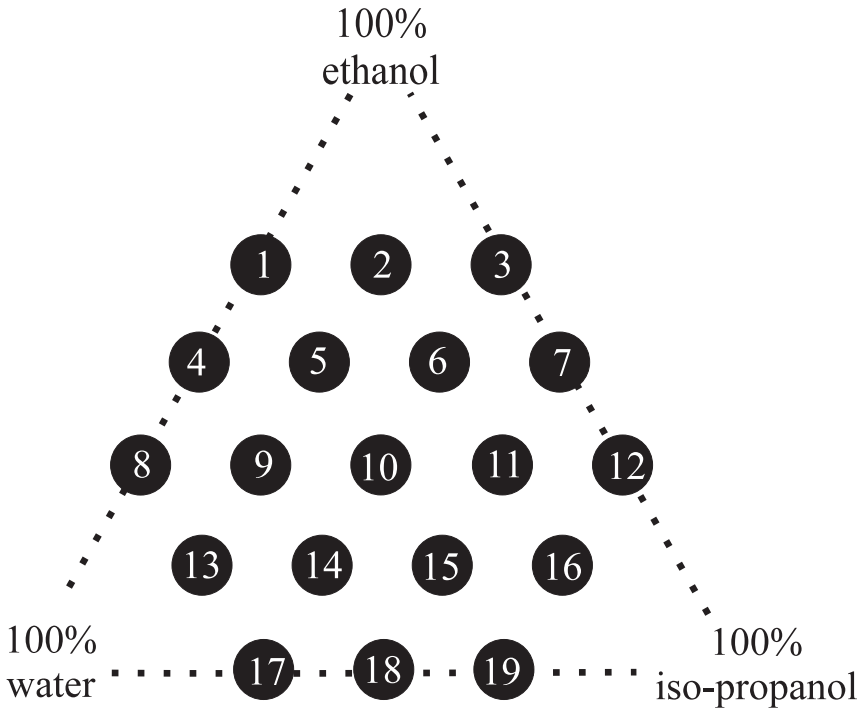

Figure 2 Mixture design for ethanol, water, and isopropanol mole fractions.

used in the analyses to follow: first-derivative spectrum (199 digitizations), which can effectively remove trends that may not be important to modeling.

3 P-spline signal regression in a nut shell

We assume familiarity with PCR and PLS techniques, but we remind the reader that these methods have in common that the order of the regressor channels is ignored; the same results will be obtained when the regressors are shuffled. From now on, bold face letters will be used for only augmented matrices and vectors.

On the other hand, PSR attempts to take advantage of the additional structure from the ordered regressors, while forcing the regression coefficients to be smooth. Consider the PSR model:

$$\mu = E(y) = \beta_0 + X_{m\times p}\beta_{p\times 1}, \tag{3.1}$$

where y is the realization of the response (e.g. ethanol, water, or isopropanol), X is the signal regressor matrix, β_0 is the unknown scalar intercept, and β is the unknown signal coefficient vector. Typically, the regression problem is inherently ill-posed as $p \gg m$.

The goal of PSR is smoothness in β, and this is achieved through dimension reduction by first projecting β onto a known (n-dimensional) basis of smooth functions, i.e. $\beta_{p\times 1} = B_{p\times n}\alpha_{n\times 1}$. We prefer a B-splines basis, in part because they are easy to compute and have excellent numerical properties (for details see de Boor, 1978 or Dierckx, 1995). The vector α is the unknown vector of basis coefficients often of modest dimension. The smoothed β coefficients are re-expressed linearly, in a multiple regression with regressors constrained in a smooth subspace.

Traditionally, the major obstacle with B-splines smooths is choice of the number and placement of knots—that is the places where the smooth polynomial segments of B-splines join, as well as specify their limited support. Too many (few) knots will lead to overfitting (underfitting), and optimization schemes to place the best number of knots are complicated nonlinear problems. P-splines (Eilers and Marx, 1996) circumvent this problem by (a) projecting β onto a rich B-spline basis using moderate number of equally spaced knots and (b) further increasing smoothness by imposing a difference penalty on adjacent B-spline coefficients in the α vector. Denote the number of B-splines as n and typically $n < p$ (but this is not essential due to the penalty). Notice that (3.1) can be rewritten as

$$\mu = \beta_0 + U_{m\times n}\alpha_{n\times 1}, \tag{3.2}$$

where $U = XB$. We now have a moderately sized regression problem, which is by design more flexible than needed. Additional smoothness and regularization come from a difference penalty on adjacent B-spline coefficients:

$$P = \lambda \sum_{k=d+1}^{n} \left(\Delta_k^d \alpha\right)^2, \tag{3.3}$$

where Δ_k^d indicates the kth difference operator of order d. In matrix terms, $P = \lambda\alpha' D_d' D_d \alpha$, where D_d is a $(n-d) \times n$ banded matrix of contrasts resulting from differencing adjacent rows of the identity matrix (I_n) d times. The order of the difference penalty can also moderate smoothing, i.e. increasing d translates into a wider footprint of the penalty affecting more neighbouring B-spline coefficients.

The PSR estimator is derived from minimizing

$$Q_P = ||y - XB\alpha||^2 + \lambda||D_d\alpha||^2, \tag{3.4}$$

with respect to α. The penalized least-squares solution simplifies as

$$\hat{\alpha}_\lambda = \left(U'U + \lambda D_d' D_d\right)^{-1} U'y. \tag{3.5}$$

Note when $\lambda = 0$ we have the unpenalized least-squares solution. The non-negative parameter λ regularizes and tunes the penalty and can be chosen by minimizing a cross-validation measure or information criterion through grid search.

A *general recipe* for PSR is given in Marx and Eilers (1999, Section 4). To give an idea of default design parameters, we typically use between 10 and 200 equally spaced cubic B-splines. However, we do vary the degree of the B-splines ($q = 3$ to 0), as well as the order of the difference penalty ($d = 3$ to 0). For fixed d, optimal λ is searched for systematically by monitoring, for instance, cross-validation prediction error. Results of these optima can be directly compared over the various $d = 0$–3. Given choice of d and λ, then the p-dimensional signal coefficient vector can be constructed, $\hat{\beta}_\lambda = B\hat{\alpha}_\lambda$. Since PSR is grounded in classical regression, diagnostics involving hat diagonal information (such as the PRESS statistics) are available, approximate (twice) standard error bands for $\hat{\beta}_\lambda$, as well as extensions into the generalized linear model.

4 Space-varying penalization signal regression (SPSR)

One can imagine that there may be some subsets or specific regions of the regressor channels that contain more useful information for prediction than others. Suppose that we have an n-vector of penalty weights w of the same dimension as α, then a modification of the PSR objective in (3.3) can be rewritten as $Q_P^\star = ||y - XB\alpha||^2 + \lambda||D_d^\star\alpha||^2$, where $D_d^\star = W^{-1/2}D_d$ and W is the diagonal matrix of weights. The reason for the inverse weight in the adjustment of the difference matrix is that channels with high weight should correspond to light smoothing locally. The λ now serves as a global tuning parameter, whereas the W allows for a local sharpening (dulling) when estimating the signal's (smooth) coefficient vector. One beauty of this approach is that existing (stable) code can be easily adapted by simply redefining the modified penalty matrix. In the next section, we will provide an approach to estimate the diagonal weight matrix W.

4.1 Selection of weights and a PLS – PSR compromise

The unknown diagonal weight matrix can be estimated using variety of existing methods; all that is necessary is a sensible weight for each region of the signal. An obvious choice is the (partial) correlation (in absolute value) or in other words the loadings for the first PLS component, i.e. (autoscaled) $U'y$, where $U = XB$ of dimension $m \times n$. Naturally, a higher rank PLS model can also provide the sensible weights, e.g. by summing the (absolute value) loading contributions. In a similar fashion, PCR can provide weights, but now with loadings corresponding to the orthogonal components that maximally explain variance in the transformed signal matrix U. In this paper, the weights in SPSR for the ordinary regression is calculated from the first-order PLS.

4.2 A few more words about choosing weights

It is perhaps worth of further investigation to sharpen the spectra directly. Suppose that we have a p-vector of weights that reflect some measure of "importance" associated with the regressor channels, forming a diagonal matrix Λ. We could view the "sharpened" regressors as $X^\star = X\Lambda$, rewrite (3.1) as $\mu^\star = X^\star\beta^\star$ (suppressing the intercept term for clarity). Given Λ, the (unpenalized) signal regression problem could now minimize $Q^\star = ||y - X^\star\beta^\star||^2$ as opposed to $Q = ||y - X\beta||^2$. Unfortunately, there is no benefit from "importance" weighting in this setting since $X\hat{\beta} = X^\star\hat{\beta}^\star$ or $\hat{\mu} = \hat{\mu}^\star$. Given this finding, one might altogether dismiss the idea of weighting the regressor channels. Yet, a potential benefit of weighting can exist when using the P-spline signal regression approach, i.e. $\hat{\mu} \neq \hat{\mu}^\star = X^\star(B'X^{\star'}X^\star B + \lambda D'D)^{-1}B'X^{\star'}y$, unless V is (proportional to) the identity matrix. The inequality actually is a consequence of either one of the two steps required in the P-spline approach: projection onto B-splines (when $\lambda = 0$) or through direct penalization without projection onto B-splines, e.g. ridge regression for the signal coefficient vector β. Implementation of this idea is trivial, as all that one must do is pass $X^\star$ into the function instead of X.

5 Additive polynomial signal regression (APSR)

One problem with PSR is that prediction quality is limited to estimated coefficients that are linear in the signal regressors, and this may be one explanation as to why PSR sometimes has difficulties competing with neural network approaches that use nonlinear features of the signals. We take a novel approach which combines ideas of the generalized additive model (Hastie and Tibshirani, 1990) with PSR. Define $X^k = \{x_{ij}^k\}$, where the superscript denotes the kth power. Note that X^k, a polynomial

transformed signal of order k, remains of dimension $m \times p$. Consider the following APSR(k) model:

$$\mu = \beta_0 + \sum_{k=1}^{K} X^k \beta_k \tag{5.1}$$

with the corresponding $p \times 1$ coefficient vector β_k, $k = 1, 2, \ldots, K$. Actually, (5.1) is a special case of a generalized linear additive smooth structure (GLASS) model (Eilers and Marx, 2002), where all of the model's "building blocks" are signal components. By projecting each β_k onto a common B-spline basis B and augmenting matrices and vectors, we can re-express (5.1) as

$$\mu = \beta_0 + \sum_{k=1}^{K} U_k \alpha_k = (1|U_1| \ldots |U_k)(\beta_0, \alpha_1, \ldots, \alpha_k)' = \boldsymbol{U}\boldsymbol{\alpha}, \tag{5.2}$$

where $U_k = X^k B$ $(m \times n)$ and α_k is the B-spline coefficient vector of length n. The penalized objective function is similar to (3.4), except using

$$\boldsymbol{P} = \boldsymbol{\alpha}' \quad \text{block diag}\left\{0, \lambda_1 D_d' D_d, \ldots, \lambda_k D_d' D_d\right\} \boldsymbol{\alpha},$$

and minimizing

$$Q_A = ||y - \boldsymbol{U}\boldsymbol{\alpha}||^2 + \boldsymbol{P},$$

with respect to $\boldsymbol{\alpha}$. The zero in the first position of the penalty ensures an unpenalized intercept. In practice, we often use a common order d in $\boldsymbol{P}$ and $k = 1$–3. In most cases, APSR requires a multi-dimensional search for the "optimal" λ vector and is discussed in the following section.

6 Results for the mixture experiment

The mixture data are split into the training and test sets. The pure compounds were not used. The training set consists of 13 mixtures (at each of 30, 40, 50, 60 and 70 °C): the outer ring + center point (mixtures 1, 2, 3, 4, 7, 8, 10, 12, 13, 16, 17, 18, 19). The independent external test set contains of the remaining six mixtures (per temperatures) in the inner ring (mixtures 5, 6, 9, 11, 14, 15). We separately model each compound (ethanol, water, and isopropanol) using the percent of mixtures as the response y and the derivative spectra (199 channels) as the regressors X. In all cases, we set the degree of B-splines to its default value $q = 3$ (cubic) and use 150 knots. The optimal values for parameters, d and λ, are found by minimizing the leave-one-out cross-validation (LOOCV) error on the training set.

LOOCV involves leaving a single observation from the training set, fitting the model using the remaining observations, and using the only omitted observation

to compute whatever the loss criterion we used. This is repeated such that each observation in the sample is omitted once. Although LOOCV is often computationally intensive, if the squared-error loss is used, LOOCV error can be computed exactly and efficiently for PSR using

$$LOOCV(\lambda) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\frac{y_i - \hat{\mu}_i}{1 - h_{ii}}\right)^2}, \tag{6.1}$$

where the h_{ii} are the diagonal elements of the "hat" matrix H, defined as

$$H = U\left(U'U + \lambda D_d' D_d\right)^{-1} U'. \tag{6.2}$$

LOOCV error can also be found for SPSR and APSR, by computing H using $D_d^{\star}$ and $(\boldsymbol{U}, \boldsymbol{P})$, respectively. The "effective dimension" of $\hat{\beta}_\lambda$ is commonly approximated using $ED(\hat{\beta}_\lambda) = \text{trace}(H)$. Effective dimension is also computed efficiently by taking advantage of invariance of trace under cyclical permutations, and this result will be useful in extensions to the generalized linear model below.

In practice, we did a linear grid search on $\log(\lambda)$ for the optimal value of λ by using the efficient gradient cleversearch(.) function developed by Susanne Heim for R/S-PLUS. To evaluate the predictive performance, we calculate the root-mean-square error of prediction (RMSEP) on the independent external test set. Table 1 presents the results and corresponding parameter values for PSR, SPSR, and APSR in the mixture experiment. For presentation in this paper, the SPSR method chooses sharpening weights using PLS(1) and APSR consists of two additive polynomial components ($k = 2$).

Figure 3 displays and compares the prediction performance of different approaches from the literature together with the newly proposed SPSR and APSR approaches. The first two methods we compared with are the global and local

Table 1 Results for the mixture experiment

Mixture/Methods	d	λ	LOOCV error	RMSEP
Ethanol/PSR	3	4.75×10^{-8}	0.01053	0.01288
Ethanol/SPSR	2	4.75×10^{-8}	0.00924	0.00956
Ethanol/APSR	3	5.99×10^{-6} and 1.0×10^{-8}	0.00781	0.00710
Water/PSR	3	3.94×10^{-7}	0.00630	0.00570
Water/SPSR	2	2.15×10^{-7}	0.00595	0.00527
Water/APSR	3	2.15×10^{-6} and 1.0×10^{-8}	0.00488	0.00311
Isopropanol/PSR	3	4.75×10^{-8}	0.00930	0.01117
Isopropanol/SPSR	2	1.59×10^{-7}	0.00952	0.01112
Isopropanol/APSR	3	7.74×10^{-7} and 3.59×10^{-5}	0.00923	0.00593

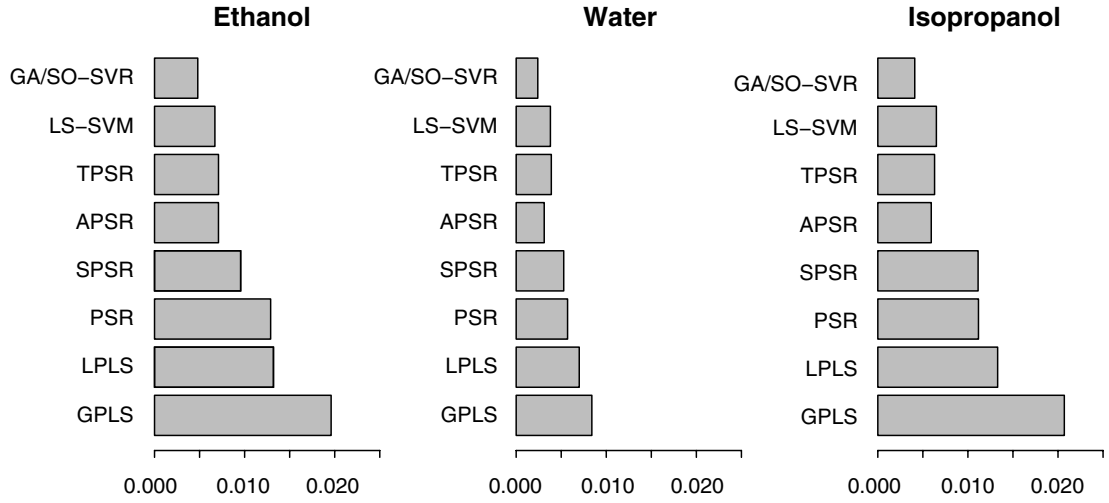

Figure 3 Prediction performances of different approaches from the literature together with the newly presented global model based on SPSR and APSR.

PLS approaches used in Wülfert *et al.* (1998). The former was set up for all temperatures at once, while the latter was set up for each temperature separately. We additionally compared our method with relatively newly proposed methods: two-dimensional tensor product PSR (TPSR, Eilers and Marx, 2003); least-squares support vector machine (LS-SVM, Thissen *et al.*, 2004); and generic algorithm/simplex optimization support vector regression (GA/SO-SVR, Üstün *et al.*, 2005). Note that TPSR approach uses a joined wavelength–temperature domain to determine the regression coefficients for an arbitrary temperature, whereas LS-SVM is a variant of nonlinear SVM using the least-squares loss function (instead of the ϵ-insensitive loss). Üstün *et al.* (2005) proposed using a genetic algorithm (GA), as well as a simplex optimization method to select the various optimal tuning parameters in GA/SO-SVR.

Based on Figure 3, we have the following remarks: (1) using the adaptive weighting scheme, the predictive performance of SPSR beats PSR in all three cases; (2) adding the quadratic terms into standard PSR substantially improves the predictive performance as shown in APSR, which outperform TPSR and LS-SVM in two of the three components; and (3) GA/SO-SVR outperforms our methods for all three mixture components. In Section 6.1, we specifically compare our methods with GA/SO-SVR in detail.

Figure 4 displays the estimated P-spline coefficient curves in PSR and SPSR methods. We see that the estimated coefficient curve in SPSR (grey) is less smooth than the estimated curve in PSR (black solid), i.e. the coefficient curve is sharpened in the SPSR approach. Figure 4 also shows the specifically scaled (in order to

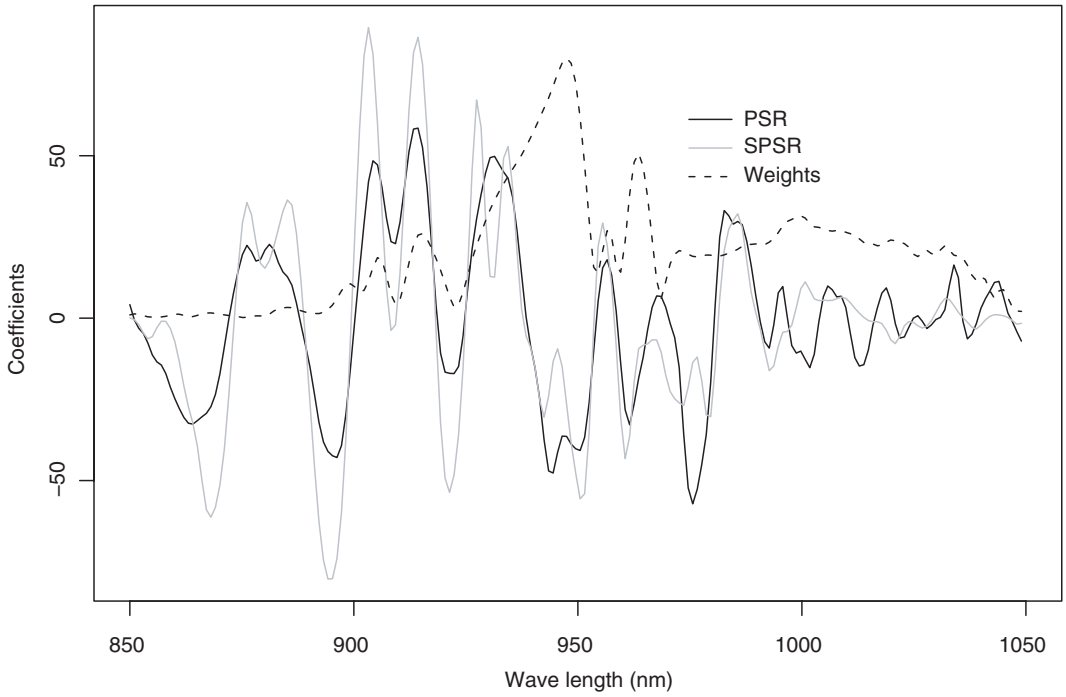

Figure 4 Estimated spectra coefficient curves in PSR (black, solid), SPSR (grey) and scaled weight vector (dash).

fit into the figure properly) weight vector $B\mathbf{w}$, where $\mathbf{w}$ is the diagonal of the $n \times n$ weight matrix W and B is the $p \times n$ B-splines basis matrix. The weight vector $B\mathbf{w}$ can be viewed as a smooth way to "connect the weight dots $\mathbf{w}$". In general, the estimated coefficients from SPSR corresponding to the regions with larger weights tend to have less local smoothness and more peaks and troughs (such as the middle part of the spectra). Note that in PSR, we put the penalty on the difference of the adjacent channel coefficients rather than the size of the coefficient itself. Thus, in SPSR, the regions with small weights do not necessarily to have small estimated coefficient values (in terms of absolute value).

6.1 SPSR/APSR vs. GA/SO-SVR

In all fairness, GA/SO-SVR has the strongest prediction performance among all presented competitors that were applied to the above mixture data. Certainly, part of the reason for the success of GA/SO-SVR is the extreme flexibility and inherent nonlinear features within the SVR methodology. Such features do come at some

price: GA/SO-SVR has three tuning parameters, and in order to select optimal values using the GA, the user needs to"expertly" specify six additional parameters. The CV optimization used by Üstün *et al.* is "leave-15-out" rather than "leave-1-out", which we believe is due to expensive computation cost in SVR. In comparison, our APSR approach, for example, has two tuning parameters, as well as a few miscellaneous (low integer) design settings: B-spline degree (3), penalty order (3), and sharpening components (1).

In any case, we believe that our methods remain highly competitive and deserving of a place in the researcher's toolbox. Unlike GA/SO-SVR, our approaches have the following features: (1) SPSR/APSR is grounded in (classical/penalized/generalized) linear regression methods; (2) they are easy to be optimally tuned, e.g. using LOOCV; (3) in practice, investigators can see the estimated regression coefficients, with twice-standard error bands (see Marx and Eilers, 1999) to help visually identify potentially important regions of the signal; (4) diagnostics, e.g. residual analysis and standard "delete-one", are easily accessible; and (5) our methods extend naturally to the generalized linear model and other additive structure models. The next section highlights this last point.

7 Extensions to the generalized linear model

In a generalized linear model (GLM) setting, PSR now has the form $g(\mu) = U\alpha = \eta$, where g is a monotone link function and $\mu = E(y)$, and y can be a non-normal (exponential family) response. Rather than minimizing (penalized) squared error loss, the PSR approach attempts to maximize the log-likelihood, $l = l(\alpha; y)$, but subject to the requirement that the estimates of adjacent α's do not differ much from each other. The modified log-likelihood now maximizes

$$l_P = l(\alpha; U, y) - \frac{1}{2}\lambda\alpha' D_d' D_d \alpha, \tag{7.1}$$

where the second term consists of the difference penalty, ($d = 0, 1, 2, ...$) 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. We find that the P-spline approach again 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 (7.1) leads to small modification of the familiar scoring algorithm:

$$\hat{\alpha}_{\lambda,t} = \left(U'\hat{V}_{t-1}U + \lambda D_d' D_d\right)^{-1} U'\hat{V}_{t-1}\hat{z}_{t-1}, \tag{7.2}$$

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 with respect to η. The working vector has entries $\hat{z}_i = (y_i - \hat{\mu}_i)/h'(\hat{\eta}_i) + \hat{\eta}_i$. The index t denotes the current iterate. It is useful to view (7.2) 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.

Upon convergence with fixed λ, we obtain the estimated smooth coefficient vector, $\hat{\beta}_\lambda = B\hat{\alpha}_\lambda$. The "hat" matrix becomes $H = U(U'\hat{V}U + \lambda D_d' D_d)^{-1}U'\hat{V}$ and effective dimension of $\hat{\beta}_\lambda$ is approximated as $ED(\hat{\beta}_\lambda) = \text{trace}\{U'\hat{V}U(U'\hat{V}U + \lambda D_d' D_d)^{-1}\}$.

Generalized SPSR and APSR estimations only require slight modifications of (7.1) to produce corresponding objectives $l_P^\star = l(\alpha; U, y) - \frac{1}{2}\lambda\alpha' D_d^{\star'} D_d^\star \alpha$ and $l_A^\star = l(\boldsymbol{\alpha}; \boldsymbol{U}, y) - \frac{1}{2}\boldsymbol{P}$, respectively. SPSR and APSR modifications to the iterative solutions in (7.2) and $ED(\lambda)$ also follow naturally. Notice that we choose the weights in SPSR (for the generalized linear model) through the iteratively reweighted PLS method proposed by Marx (1996), but other sensible weights can also be applied.

For training in the GLM setting, we optimize the penalty tuning parameter(s) using an information criterion. Although several variants of information criteria are available, we use a modified version of Akaike's Information Criterion (AIC):

$$\text{AIC}^*(\lambda) = \text{deviance}(y, \hat{\beta}_\lambda) + 2\ ED(\hat{\beta}_\lambda). \tag{7.3}$$

Note that the deviance in (7.3) is based on the penalized likelihood defined in (7.1). In addition, we use *ED*, the trace of the "hat" matrix, as an approximated effective dimension (in the same spirit as the degrees of freedom for a linear smoother proposed in Hastie and Tibshirani, 1990). Other information criteria can also be applied.

7.1 Phoneme classification example

We use the data for phoneme classification described by Hastie, *et al.* (1995), which also illustrates that the signal regressors do not have to be smooth to use our method. The data are 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 are 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

$$\log \frac{p_i}{1 - p_i} = f(x_{i1}, \ldots, x_{i150}), \tag{7.4}$$

where p_i is the probability of *aa* and x_{ij} is the log-periodogram for subject $i = 1, \ldots, 160$. Specifically, for both PSR and SPSR model, the right-hand side of

(7.4) is $\alpha_0 + \sum_{j=1}^{150} x_{ij}\alpha_j$. For the APSR model there is the additional quadratic term in the linear predictor: $\alpha_0 + \sum_{j=1}^{150}(x_{ij}\alpha_{j1} + x_{ij}^2\alpha_{j2})$.

Note that data also exist for distinguishing between *aa* and *iy* (as in *she*), but these are fairly easy to distinguish. The data are randomly split into the training and test sets. The training set consists of 128 subjects (80% of the observations in the data set), where the independent external test set contains of the rest of the 32 speakers. The left panel of Figure 5 provides typical and extremely non-smooth log-periodograms of the subjects.

We use cubic (our default value for q) P-spline and third-order penalty (our default for d) on 13 equally spaced knots, providing sufficient flexibility. The optimal values for λ is found by minimizing the AIC on the training set. To evaluate the predictive performance, we calculate the classification error rate (CER) on the independent external test set, defined as

$$CER = \frac{1}{l}\sum_{i=1}^{l} \mathbf{1}(y_i \neq \hat{y}_i), \tag{7.5}$$

where $\mathbf{1}(A)$ is an indicator function, equal to one when A holds, otherwise zero, and $\hat{y}_i = \mathbf{1}\ (p_i \geq 0.5)$. The size the external test set is generally denoted as l (32 for this example). In SPSR, we set the weights proportional to the absolute values of

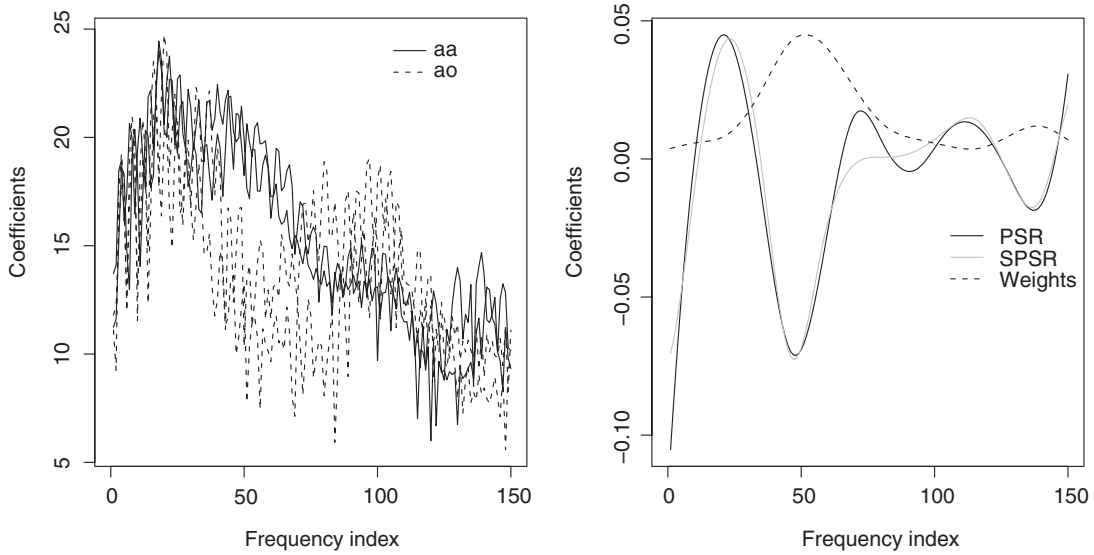

Figure 5 (Left) Typical and extremely non-smooth log-periodograms for two subjects in each class. (Right) Estimated spectra coefficient curves in PSR (black, solid), SPSR (grey), and scaled weight vector (dash).

Table 2 Results for the phoneme example

Methods	λ	AIC	CER
PLS	–	–	0.125
SVM	–	–	0.09375
PSR	26.8	106.22	0.09375
SPSR	37.3	104.59	0.0625
APSR	17.6 and 1×10^5	108.62	0.0625

the estimated coefficients of the iteratively reweighted PLS method with only the first component.

Table 2 presents the results and corresponding parameter values for PSR, SPSR, and APSR in the phoneme example. In addition, we applied the nonlinear SVM using Gaussian kernel on the data. The CER based on the test set for the SVM is 0.09375. Furthermore, we applied the generalized PLS method (Marx, 1996) on the data and explored the model with up to five components. The test errors (from one to five components) are 0.125, 0.125, 0.21875, 0.28125, and 0.28125, respectively. PSR has an error rate of 0.0925; however, SPSR and APSR have an error rate of 0.0625, outperforming generalized PLS. The right panel of Figure 5 shows the estimated P-spline coefficient curves in PSR (black solid) and SPSR (grey) methods. Although the estimated coefficient curve in SPSR is very close to the estimated curve in PSR at two ends, it de-emphasizes the region between the 70th and 110th frequency, where the corresponding weights (dash) are small.

8 Discussion

In this paper, we proposed two variations on the P-spline signal-regression approach. SPSR can be viewed as PSR with an adaptively chosen roughness penalty matrix D_d (for the basis coefficients). Specifically, SPSR uses the adaptive penalty weights $W^{-1/2}D_d$, where $W^{-1/2}$ is a diagonal matrix (e.g. using PLS loading information). When the estimated coefficients from PLS are rough (having peaks or troughs), the channels with large absolute values of estimated coefficients tend to be more relevant to the response variable, and thus these channels are locally penalized less than those channels with "smaller" PLS loadings. Standard PSR is achieved in the special case when W is proportional to the identity matrix. Certainly, other weighting schemes using other schemes (or prior knowledge) can be used and should be explored, e.g. using the variable importance summaries from the boosting method (Friedman, 2001). In cases where the underlying relationship between the response and regressors is nonlinear, the performance of PSR and SPSR can degrade seriously. To overcome this problem, APSR expands the linear basis (the regressors themselves)

to a polynomial basis. As shown in Section 6, APSR can substantially improve predictive performance.

For computational purposes, it is worth noting that the explicit solution for SPSR can be found efficiently through data augmentation tricks on PSR, i.e. replace D_d by D_d^*. For APSR, since it is a special case of GLASS model, APSR can be implemented through GLASS with slight modification. As a result, both SPSR and APSR inherit the computational benefits from PSR: (i) the necessary computations, including cross-validation, are comparable in size to those for a medium sized regression problem; (ii) the computed fits are described compactly by the coefficients of the B-splines; and (iii) unlike many nonlinear methods which need specific solvers (such as SVM needs quadratic programming solvers) and sophisticated tuning process (such as training a neural network with large amount of parameters), the PSR-based methods have simple structures and need essentially a least-square regression solver, which is widely available in many programming languages, to get their solutions.

We did explore the idea of combining SPSR and APSR. The penalized objective function is similar to (3.4), except using

$$\boldsymbol{P} = \boldsymbol{\alpha}' \text{ block diag}\left\{0, \lambda_1 D_{dk}^{\star\prime} D_{dk}^{\star}, \ldots, \lambda_k D_{dk}^{\star\prime} D_{dk}^{\star}\right\} \boldsymbol{\alpha},$$

where $D_{dk}^{\star} = W_k^{-1/2} D_d$ and W_k is the diagonal matrix of weights for the polynomial order k. Like SPSR, we chose the weights (for the polynomial order k) proportional to the loadings (in absolute value) for the first PLS component, i.e. autoscaled $(X^k B)' y$. For the mixture and phoneme examples, we did not achieve performance better than APSR. Optimization of the tuning parameters led to very small values for the λ's in both examples which led us to believe that this approach tended to overfit the data.

For further confirmation, we also applied our methods to other data sets, such as the highly cited Tecator near-infrared absorbance spectra data, which is available at http://lib.stat.cmu.edu/datasets/tecator. By using the APSR (on the second-derivative spectra) with the polynomial order at three (including the linear, quadratic, and cubic basis), the RMSEP (on the independent external test set) reduces to 0.566, a 79% reduction from PSR. APSR again outperforms neural network, leading to an RMSEP that is a factor 1.15 lower (see Borggaard and Thodberg, 1992).

In signal regression, we may often find it useful to remove trends by using first-order derivatives. Naturally, higher order derivative spectra can also be obtained. This not only exploits the intrinsic smoothness in the signaling process, but also may get closer to the underlying driving forces at work. Thus, we may include the order of derivative as one of the design parameter and choose its optimal value adaptively. For example, in the Tecator data, we found using the second-derivative spectra achieves better performance than using the others.

References

Borggaard C and Thodberg HH (1992) Optimal minimal neural interpretation of spectra. *Analytical Chemistry*, **64**, 545–51.

de Boor C (1978) *A practical guide to splines*. New York: Springer-Verlag.

Dierckx P (1995) *Curve and surface fitting with splines*. Oxford: Clarendon Press.

Eilers PHC and Marx BD (1996) Flexible smoothing with B-splines and penalties (with comments and rejoinder). *Statistical Science*, **11**, 89–121.

Eilers PHC and Marx BD (2002) Generalized linear additive smooth structures. *Journal of Computational and Graphical Statistics*, **11**(4), 758–83.

Eilers PHC and Marx BD (2003) Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems*, **66**, 159–74.

Frank IE and Friedman JH (1993) A statistical view of some chemometric regression tools (with Discussion). *Technometrics*, **35**, 109–48.

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, **29**, 1189–232.

Hastie T and Tibshirani R (1990) *Generalized additive models*. London: Chapman & Hall.

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

Land SR and Friedman JH (1996) *Variable fusion: a new method of adaptive signal regression*. Technical Report 114, Department of Statistics, Stanford University.

Martens H and Næs T (1989) *Multivariate calibration*. New York: John Wiley and Sons.

Marx BD (1996) Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics*, **38**, 374–81.

Marx BD and Eilers PHC (1999) Generalized linear regression on sampled signals and curves: a P-spline approach. *Technometrics*, **41**, 1–13.

Marx BD and Eilers PHC (2005) Multidimensional penalized signal regression. *Technometrics*, **47**, 13–22.

Ramsay JO and Silverman BW (2005) *Functional data analysis* (2nd ed.). Springer.

Thissen U, Üstün B, Melssen WJ and Buydens LMC (2004) Multivariate calibration with least-squares support vector machines. *Analytical Chemistry*, **76**, 3099–105.

Üstün B, Melssen WJ, Oudenhuijzen M and Buydens LMC (2005) Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization. *Analytica Chimica Acta*, **544**, 292–305.

Wülfert F, Kok W and Smilde A (1998) Influence of temperature on vibrational spectra and consequences for the predictive ability of multivariate models. *Analytical Chemistry*, **70**, 1761–67.

Wülfert F, Kok W, Noord O and Smilde A (2000) Linear techniques to correct for temperature induced spectra variation in multivariate calibration. *Chemometrics and Intelligent Laboratory Systems*, **51**, 189–200.