Chemometrics and Intelligent Laboratory Systems 96 (2009) 196–202

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Multivariate calibration with single-index signal regression

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

[a] Methodology and Statistics, Faculty of Social and Behavioral Sci, Utrecht University, 3508 TC, Utrecht, The Netherlands
[b] Department of Experimental Statistics, Louisiana State University, Baton Rouge, LA 70803, United States

ARTICLE INFO

Article history:
Received 16 November 2008
Received in revised form 20 January 2009
Accepted 3 February 2009
Available online 12 February 2009

Keywords:
Multivariate calibration
P-splines
Projection pursuit regression

ABSTRACT

In general, linearity is assumed to hold in multivariate calibration, but this may not be true. Penalized signal regression can be extended with an explicit link function between linear prediction and response, in the spirit of single-index models. Like the vector of calibration coefficients, the unknown link function is being estimated by P-splines. Application to simulations and three data sets shows that if a non-linearity is present, it will be picked up by the model and prediction will be improved.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

Multivariate calibration (MVC) has seen many new developments in recent years. The beaten path of ridge regression, PLS (partial least squares), and PCR (principal components regression) has been left by adventurous souls in search of new ways of approaching an old problem. Two of us (BM and PE) have pioneered the use of smoothness penalties in one and two dimensions [13,14,7], which was recently improved by PLS-derived variable weighting [12]. Others have successfully explored support vector machines (SVM) [15] and genetic algorithms (GA) [17]. In the race to improve MVC, some data sets have become the test bench for comparing methods, e.g. the Tecator data (with responses water, fat, and protein) and the Cargill corn study (with responses moisture, oil, protein, and starch), among others. The data sets usually have near-infrared (NIR) spectra for regressors and are discussed in more detail in the coming sections.

From statistical perspective, 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 often concentration of a chemical analyte. In this paper, we refer to such as the MVC problem.

The central idea of advanced MVCs is the use of the so-called kernel. The kernel replaces the familiar measures of distance and correlation, which stem from linear algebra, by non-linear functions. The idea is that a non-linear kernel in the linear space of the observations corresponds to a linear regression or discrimination function in a higher dimensional space. The higher-dimensional space is never explicitly constructed: it is implied by the kernel function.

We conjectured that the surprising power of the SVM in MVC should be explained by the fact that it implicitly deals with non-linearities. Unfortunately, it does not tell us anything about the character of the non-linearities. A method that explicitly models non-linear behavior would provide more insight. In fact, one could view MVC as using an identity link function, where such a model may be (slightly) misspecified. In effect, there may exist a true, but "missing link" function (that is nonlinear and monotone) [4]. In this paper we propose such a model that can estimate the "missing link" and improve external prediction.

Our starting point is penalized signal regression (PSR), where the prediction $\hat{y}_i$ of a concentration is expressed through $\mathrm{E}(\hat{y}_i)=\mu_i=\sum_j x_{ij}\beta_j$, where the vector x_i is the measured spectrum, and the vector of coefficients β is forced to be smooth [13]. We introduce a small modification: $\mu_i=f(\sum_j x_{ij}\beta_j)$, where $f(\cdot)$ is a smooth (and possibly monotone) function, to be estimated from the data. This model is known as "projection pursuit" [10], generally without smoothness demands on β. James and Silverman [11] generalized projection pursuit though functional adaptive model estimation. Bai et al. [1] have explored penalized single-index models with longitudinal data. To our knowledge, projection pursuit has never been applied to MVC. We show that it gives excellent predictions, highly competitive to, and sometimes better than, the SVM. Moreover, the function $f(\cdot)$ explicitly shows the type and the amount of the non-linearity.

2. Benchmark datasets

2.1. Tecator data

The data consists of 215 near-infrared absorbance spectra of meat samples, recorded on a Tecator Infratec Food Analyzer. Each observation consists of a 100-channel absorbance spectrum in the wave length range 850–1050 nm, contents of water, fat and protein. One

---

* Corresponding author.
E-mail addresses: p.eilers@erasmumc.nl (P.H.C. Eilers), bli@lsu.edu (B. Li), bmarx@lsu.edu (B.D. Marx).

0169-7439/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2009.02.001

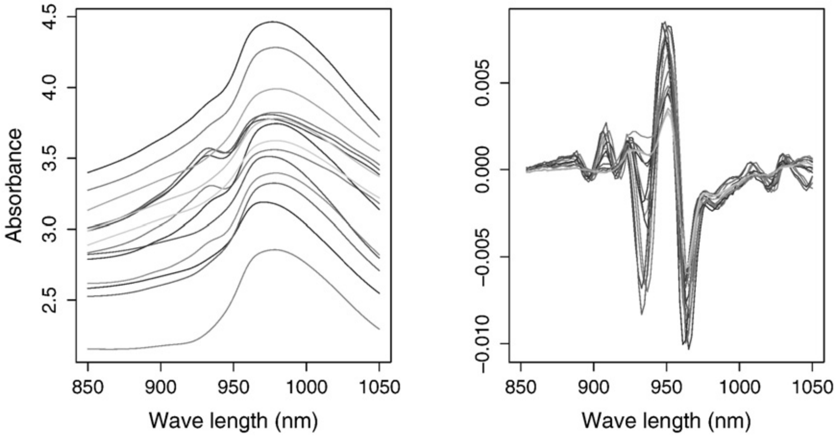

Fig. 1. Fifteen sample curves for Tecator data (left), and the corresponding sample curves after second order differencing (right). The level of greyness for each curve is proportional to its corresponding fat content.

goal is to predict the fat content of a meat sample on the basis of its near infrared (NIR) absorbance spectrum. The dataset are split in a training/monitoring/testing set of 129/43/43 samples. Tecator data, which was originally used by Borggaardand Thodberg [2], is available at http://lib.stat.cmu.edu/datasets/tecator. Fig. 1 shows 15 examples in Tecator data. The level of greyness for each sample is proportional to its fat content.

2.2. Cargill corn study

The second data set was taken from the Cargill study and contains 80 NIR spectra of corn samples measured on three different NIR spectrometers (referred to as: m5, mp5 and mp6) for the prediction of moisture, oil, protein and starch content, respectively. The spectra were measured from 1100 to 2498 nm at a spectral resolution of 2 nm. The data can be obtained at http://software.eigenvector.com/Data/Corn/index.html. Fig. 2 shows 30 examples in corn data and corresponding moisture, oil, protein and starch content. The corn data are split in a training/testing set, each contains 39 samples.

3. Recap: P-spline signal regression

Penalized signal regression (PSR) directly uses the ordinal structure of the regressors (e.g. along wavelength) and intentionally forces the regression coefficients to be smooth along the order index. A standard regression approach is

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

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. In chemometric applications, this regression problem is inherently ill-posed since $p \gg m$.

In short, PSR achieves smoothness in β through dimension reduction by first projecting β onto a known (n dimensional) basis of smooth functions, i.e. $\beta_{p \times 1} = B_{p \times n}\alpha_{n \times 1}$. A B-splines basis is used because it is easy to compute and has excellent numerical properties [3,5]. The basis B uses equally-spaced knots and is rich enough to provide more flexibility than needed. The vector α is the unknown vector of basis coefficients. We stress that PSR does not smooth the spectra, but rather assumes that smoothing the vector of coefficients does not adversely affect external prediction. In fact, PSR allows rough spectra (such as with differenced spectra). With penalization, optimal smoothing of spectra is not equivalent to optimal smoothing of the coefficients. Details are discussed in Marx and Eilers [13].

The P-spline approach [6] avoids the difficulties of the optimal placement and number of knots by: (a) projecting β onto a B-spline basis using moderate number of equally-spaced knots, and (b) further

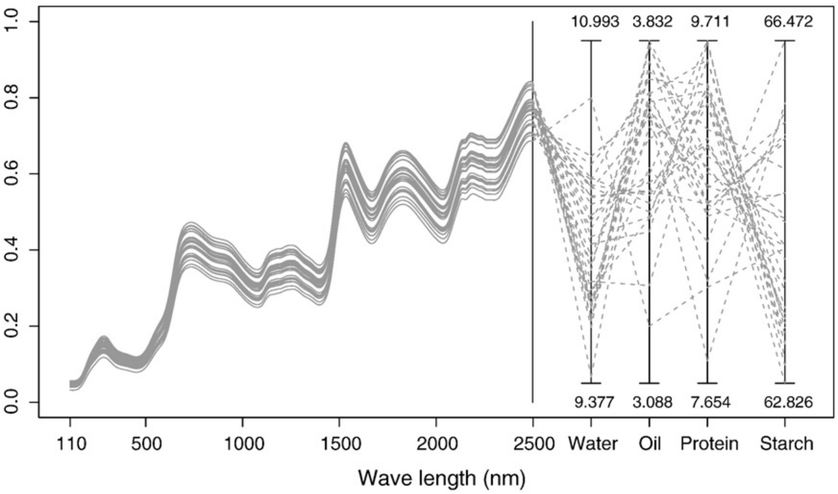

Fig. 2. Thirty sample curves for corn data with corresponding moisture, oil, protein and starch content.

increasing smoothness by imposing a difference penalty on adjacent B-spline coefficients in the α vector. Denote the number of B-splines as n (typically $n<p$). We re-express Eq. (1) as

$$\mu = \beta_0 + U_{m \times n}\alpha_{n \times 1}, \tag{2}$$

where $U=XB$. We now have a moderately sized regression problem, using regressors U. As stated, smoothness and regularization comes from a difference penalty on adjacent B-spline coefficients:

$$P = \lambda \sum_{k=d+1}^{n} \left(\Delta^d \alpha_k\right)^2, \tag{3}$$

where Δ^d indicate the 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 neighboring B-spline coefficients.

Only to simplify presentation, we set $\beta_0=0$. An intercept is used in all of the analyses to follow. The PSR estimator is derived from minimizing

$$Q_P = ||\,y - XB\alpha\,||^2 + \lambda\,||\,D_d\alpha\,||^2, \tag{4}$$

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

$$PSR(U, y, \lambda, D_d, n) = \hat{\alpha}_\lambda = \left(U'U + \lambda D_d' D_d\right)^{-1} U'y. \tag{5}$$

Eq. (5) can be easily adjusted for an intercept term by substituting $U_1=(1,U)$ and difference matrix becomes $D_{d1}=(0,D_d)$, and this is exactly what we do for all of our data analyses. The non-negative parameter λ tunes the penalty and can be chosen by minimizing a cross-validation measure through grid search. Leave-one-out cross-validation (LOOCV) involves leaving a single observation out 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

$$LOOCV(\lambda) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\frac{y_i - \hat{y}_i}{1 - h_{ii}}\right)^2}, \tag{6}$$

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

$$H = U\left(U'U + \lambda D_d' D_d\right)^{-1} U'. \tag{7}$$

The term “hat” matrix is used because H turns y into predicted values $\hat{y}=Hy$. In practice, we perform a linear grid search on $\log(\lambda)$ for the optimal value of λ using the efficient gradient *cleversearch*$(\cdot)$ function developed by Susanne Heim for R/S-PLUS.

A *general recipe* for PSR is given in (Marx and Eilers [13], Section 4). In SISR we have two hyper-parameters (one each for f and β) and a few (low integer) design parameters (the degree of the B-spline, the order of the penalty, and the number of equally-spaced knots). To give an idea of default design parameters, we typically use between 10 and 200 equally spaced cubic B-splines. We can vary the degree of the B-splines ($q=3$ to $q=0$), but usually default with cubic. The order of the difference penalty can vary ($d=3$ to $d=0$). For fixed d, optimal λ is searched for systematically by monitoring, for instance, cross-validation prediction error. Results of these optimum can be directly compared over the various $d=0,1,2,3$. Given choice of d and λ, then the p-dimensional signal coefficient vector can be constructed, $\hat{\beta}_\lambda = B\hat{\alpha}_\lambda$.

4. Methodology

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 machine learning approaches, e.g. support vector machine and neural networks, that use nonlinear features of the signals. We take a novel approach which combines ideas of the projection pursuit regression (see e.g. [10] with PSR). Our model has the form $\mu=f(U\alpha)$, where the function f is unspecified and estimated along with the coefficient vector α using some flexible nonparametric smoothing method. This model is known as the *single-index model*, which take only the first step of projection pursuit regression. The basic appeal of single-index model is its simplicity. Consequently, a modification of the PSR objective in Eq. (4) can be rewritten as

$$Q_P^\star = ||\,y - f(U\alpha)\,||^2 + \lambda\,||\,D_d\alpha\,||^2. \tag{8}$$

We refer to this method as Single-Index Signal Regression (SISR), where the penalty is implied as this is a generalization of PSR.

4.1. The model fitting algorithm

Given the B-spline coefficient vector α, the estimation of function f becomes a one-dimensional smoothing problem, and we can apply any scatter-plot smoother to obtain its estimate. In this paper, we estimate f using a (cubic) P-spline scatter smoother [6] for the following reasons: 1. P-splines smoothers are easy to use and optimize. 2) Heavy smoothing (with a second order penalty) leads to approximately monotone linear f, which is expected by the thermodynamic properties. 3) The first derivative of f (denoted as f'), which is needed in our algorithm, can be easily computed. Derivatives of smoothers with equally spaced B-splines have the pleasant property that they are equivalent to $B_{(q-1)}(\Delta\alpha)/b$, where q is the degree, Δ denotes the first difference operator, and b is the step length on the equally spaced knots. For simplicity reason, denote $S(V,W,\lambda,d,n)$ as the operation of fitting a cubic P-spline scatter smoother on V (the input variable) and W (the response) using the penalty tuning parameter λ and d difference order on n equally spaced knots.

Once given an estimate of f, the coefficient vector α can be estimated using a (first-order) Taylor series approximation of the function f (about the current estimate, α_0). Specifically, if α_0 is the current estimate for α, then the current estimate of $\mu=f(U\alpha)$ can be approximated by

$$f(U\alpha) \approx f(U\alpha_0) + f'(U\alpha_0)U(\alpha - \alpha_0). \tag{9}$$

Using Eq. (9), we have an approximation of $Q_P^\star$

$$\begin{aligned} Q_P^\star &\approx ||y - f(U\alpha_0) - f'(U\alpha_0)U(\alpha - \alpha_0)||^2 + \lambda||D_d\alpha||^2 \\ &= ||[y - f(U\alpha_0) + f'(U\alpha_0)U\alpha_0] - f'(U\alpha_0)U\alpha||^2 + \lambda||D_d\alpha||^2 \\ &= ||y^\star - U^\star\alpha||^2 + \lambda||D_d\alpha||^2, \end{aligned} \tag{10}$$

where $y^\star = y - f(U\alpha_0) + f'(U\alpha_0)U\alpha_0$ and $U^\star = \mathrm{diag}\{f'(U\alpha_0)\}U$. Note that Eq. (10) implies that given f, the optimal α that minimizes the right-hand side of (10) can be obtained through a $PSR(U^\star, y^\star, \lambda, D_d, n)$.

Hence, in our algorithm, we first carry out a PSR with target y on U (Step 1). Then, given α, an estimate of f is obtained (Step 2). The two steps, estimation of f and α are iterated until convergence of $\hat{\alpha}$. We set the B-spline basis degree $q=3$ (cubic splines) as default value for

Table 1
Summary results for the Tecator experiment.

Methods	λ	d	n	Validation error	Test error
PLS	–	–	–	2.83	2.86
PCR	–	–	–	2.82	2.92
PSR	2.64×10^{-8}	3	100	2.74	1.85
SISR	$(3.36\times10^{-6}, 0.161)$	(3,2)	(100,20)	1.73	1.39

both steps. Again only for simplicity of presentation, the intercept term is suppressed ($\beta_0=0$) in the algorithm.

```
Algorithm SISR
1. Initializations:
    • Choose the tuning parameter values $(\lambda_1,\lambda_2)$ for Step 1 and 2
    • Choose number of knots $(n_1,n_2)$
    • Choose penalty order $(d_1,d_2)$
    • Choose the tuning parameter $\lambda_0$ for the initial Step 1 (default value is $10^6$)
    • Create $U=XB$
    • Calculate $\hat{\alpha}=PSR(U,y,\lambda_0,d_1,n_1)$
2. Cycle until convergence of $\hat{\alpha}$'s
    • Estimate $\hat{f}$ and its derivative $\hat{f}\,'$ from $S(U\hat{\alpha},y,\lambda_2,d_2,n_2)$
    • Obtain $y^\star$ and $U^\star$
    • Update $\hat{\alpha}=PSR(U^\star,y^\star,\lambda_1,d_1,n_1)$
3. Prediction: $\hat{y}^{new}=\hat{f}(x^{new}B\hat{\alpha})$
end algorithm
```

Remark 1. Using the large value of λ_0 for the initial Step 1 provides more chance for the algorithm to find the optimal α and f rather than falls into a local minima when a small value of λ_0 is used. An alternative way is to try several different initial values of λ_0 to avoid the solution falls into a local minima.

Remark 2. Denote $||\alpha||^2=\sum_{k=1}^{n}\alpha_k^2$, α^{cur} (α^{pre}) is the α vector for the current (previous) iteration. The algorithm terminates when

$$\frac{(\alpha^{cur}/\,||\,\alpha^{cur}\,||\,)-(\alpha^{pre}/\,||\,\alpha^{pre}\,||\,)}{\alpha^{cur}/\,||\alpha^{cur}||}<\epsilon,$$

where ϵ is a prespecified convergence tolerance (default value is 10^{-3}). Note that $\alpha^{cur}/||\alpha^{cur}||$ scaled α vector to have unit L_2 norm.

4.2. Model selection

In SISR, there are two penalty tuning parameters $\lambda=(\lambda_1,\lambda_2)$. There are two ways to find the optimal values for λ. (1) Given an independent validation set, the optimal values for λ can be found by minimizing the error on the validation set (sometimes referred to as a monitoring set). (2) Otherwise, the optimal values for the tuning parameters can be found by minimizing the cross-validation error. Note that one advantage of using PSR is that the computation of LOOCV error is inexpensive, i.e. no need to refit the model after omitting one observation. However, in SISR, the cross-validation error can only be calculated *externally* through refitting the model omitting the validation set. To evaluate the external predictive performance, we calculate the root-mean-square error of prediction (RMSEP) on an independent test set:

$$RMSEP=\sqrt{\frac{1}{m^{test}}\sum_{i=1}^{m^{test}}(y_i-\hat{y}_i)^2} \tag{11}$$

where m^{test} is the number of observations on the test set and $\hat{y}_i$ is the predicted response for the ith subject in the external test set.

5. Examples and simulation

5.1. Results for Tecator data

The model is fitted on the 129 training samples, while the optimal value(s) for the tuning parameters are found by minimizing the validation errors on the monitoring set. Hence, we bypass the computationally expensive computing for the external LOOCV errors in the Tecator study. The second order derivative spectra (98 channels) is used as the regressor X. Table 1 presents and compares the prediction performance of different approaches from the literature together with the newly proposed SISR approach and the corresponding parameters for the PSR and SISR methods. We see that the test error for SISR is about 25% lower than the one from PSR. Note that only five cycles were needed until convergence in SISR. Fig. 3 displays the estimated function f and the coefficient curves $\hat{\beta}=\beta_{p\times n}\hat{\alpha}_{n\times1}$ in SISR (grey solid) and PSR (black dash). Note that (1) the estimated $\hat{f}$ function is a monotonically increasing and nonlinear function; (2) the coefficient curve in SISR is smoother than the one in PSR.

5.2. Results for the corn study

Recently, Feudale et al. [9] have investigated the effects of various orthogonal signal correction (OSC) algorithms on the modelling power of PLS. In order to investigate the influence of OSC, Feudale et al. [9] applied PLS to non-processed spectra, OSC-corrected spectra according to the algorithms described by Wise and Gallagher (which is available online) and Fearn [8], respectively, and on piecewise orthogonal signal correction (POSC) corrected spectra as well Feudale et al. [9].

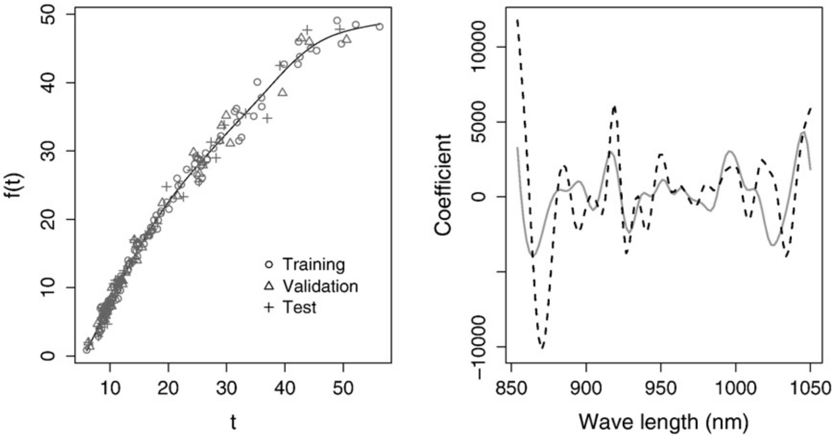

Fig. 3. Estimated function $\hat{f}$ (left) and estimated spectra coefficient curves $\hat{\beta}$(right) for SISR (grey solid) and PSR (black dash).

Table 2
Results for SISR in the corn study.

Methods	λ	Number of cycles	LOOCV error	Test error
Water	1.0×10^{-7}, 1.0×10^{2}	3	0.0081	0.0098
Oil	1.0×10^{-5}, 1.0×10^{5}	2	0.0469	0.0791
Protein	3.2×10^{-5}, 3.2×10^{2}	2	0.1135	0.1227
Starch	1.0×10^{-7}, 3.2×10^{4}	2	0.1142	0.1714

Table 3
Comparison of the prediction errors of various approaches for the corn study.

Methods	Water	Oil	Protein	Starch
PLS	0.0190	0.0742	0.1732	0.2946
PLS Wise OSC	0.0189	0.0741	0.1733	0.2946
PLS Fearn OSC	0.0190	0.0742	0.1732	0.2946
PLS/POSC	0.0433	0.0555	0.1211	0.2759
GA/Sim-SVM	0.0100	0.0654	0.1190	0.1806
PSR	0.0098	0.0719	0.1226	0.1581
SISR	0.0098	0.0791	0.1227	0.1714
SISR (differenced)	0.0130	0.0770	0.1071	0.1877

Table 4
Comparison of the prediction errors of various approaches for the mixture data.

Methods	Ethanol	Water	Isopropanol
PSR	0.0129	0.0057	0.0112
SISR	0.0165	0.0049	0.0137
VPSR	0.0096	0.0053	0.0111
Polynomial	0.0089	0.0031	0.0059
TRSP	0.0071	0.0039	0.0063
LS-SVM	0.0067	0.0038	0.0065
GA/Sim-SVM	0.0048	0.0024	0.0041

In addition, Üstün et al. [17] applied GA/SO-SVR on the corn data. Here, we employed the same data pre-treatment and training and test set selection has been performed as described in Feudale et al. [9]. Üstün et al. [17] proposed a sophisticated machine learning method using support vector regression (SVR), which usually includes four tuning parameters. In order to find these optimal tuning parameters, they suggested to use genetic algorithm (GA) combined with simplex search algorithm. Note that the genetic algorithm needs six additional tuning parameters, which need “expert's knowledge” to choose a reasonable setting. Similarly, it is well-known that it is very difficult to fine tune the parameters and model structures (such as number of hidden neurons and layers) in the neural network.

Since no validation set is assigned in corn study, we chose the optimal tuning parameters in SISR (λ_1,λ_2) based on minimizing the LOOCV errors. Table 2 presents the prediction results and corresponding parameters for SISR on the corn data set. Table 3 displays and compares the prediction performance of different approaches from the literature together with our newly proposed SISR approach. The last row of Table 3 uses SISR with first differenced spectra. We find mixed gains and losses when using the differenced spectra instead of the raw spectra. Since the results from the previous literature use non-differenced spectra, we suggest using the raw spectra SISR results to ensure a fairer comparison.

Fig. 4 displays the estimated function $\hat{f}$ in SISR, and the estimated coefficient curves for water in SISR (grey solid) and PSR (black dash). Unlike the Tecator example, here the estimated $\hat{f}$ is essentially a 45 degree line. This can be seen by the different size of the λ_2 in two examples. The λ_2 in Tecator study is 0.616, while in corn example λ_2 is 100 for water. Recall the $d_2 = 2$, and large λ_2 produces linear $\hat{f}$. Hence, in this case, the SISR defaults back to the simpler, yet extremely competitive, PSR method of Marx and Eilers [13]. As expected, the coefficient curve $\hat{\beta}$ for SISR is very close to the estimates from PSR.

5.3. Results of the mixture experiment

Another well-known test bench data set for MVC is the three-component (ethanol, water, iso-propanol) mixture experiment [18,19]. The mixture data contains 19 mixtures, and each mixture has measured spectra under five temperature conditions: 30, 40, 50, 60, and 70 °C (±0.2 °C). Unlike the previous the Tecator and corn examples, LOOCV is used for optimization of tuning parameters, rather than using an independent validation data set. We also did apply SISR on these mixture data, however the prediction error is more or less par with standard PSR. Thissen et al. [15], Üstün et al. [17], and Li and Marx [12] show summaries of the best prediction performance of all available algorithms, as well as present the details as to how the data are split for training models and external prediction. Table 4 displays the external prediction errors (RMSEP) of the mixture data on various methods, now including our SISR approach. The P-spline based approaches (PSR, SISR, VPSR, and polynomial) each used 150 equally-spaced and a third order difference penalty for the spectra coefficient vector. To estimate f in SISR, 20 equally-spaced knots were used with a second order difference penalty. For reproducibility, the optimal (λ_1,λ_2,LOOCV) were {(10^{-7}, 1000, 0.01038), (10^{-7}, 100, 0.00497), (10^{-7}, 0.1, 0.01201)} for ethanol, water, and iso-propanol, respectively.

We believe that the relatively uncompetitive results of SISR are partially due to the nature of the experiment design: For each component there is “effectively” only five unique “levels” (0, 16, 33, 50, 66%). As such, the estimated function f for each of the three components looked very much like smooth version of a step function. Poor prediction of f that is based on only five unique levels of y translates

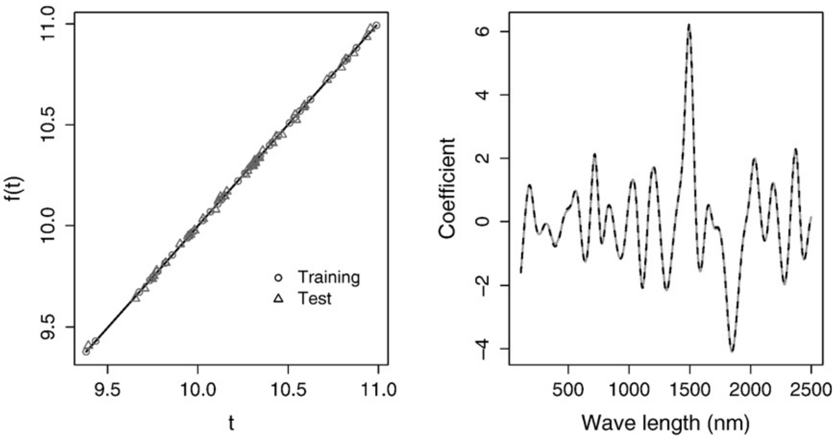

Fig. 4. Estimated function $\hat{f}$ (left) and estimated spectra coefficient curves $\hat{\beta}$(right) for SISR (grey solid) and PSR (black dash).

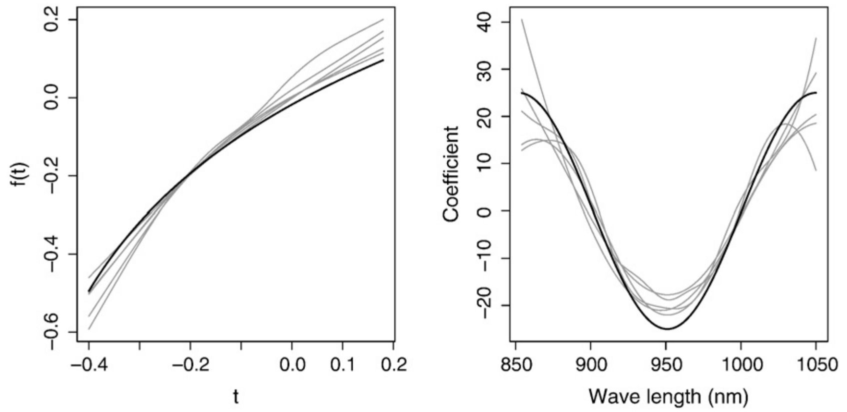

Fig. 5. Estimated functions $\hat{f}$ (left) and estimated spectra coefficient curves $\hat{\beta}$ (right) based on five random replications of the simulation.

into poor external prediction, especially when the estimated *f* needs to be interpolated or extrapolated. We do not recommend SISR when the response only has a few numbers of fixed levels as in this mixture experiment. We suggest that the researcher first check the number of levels for the response *y*, and if there are only a small number of different values, SISR should not be applied.

5.4. Simulation

In order to gain insight into the nature of our method, we applied SISR on a simulated regression example based on Tecator data with known underlying generating mechanism. The original Tecator spectra (100 channels) are used as the regressor *X*. We simulate data from the true model

$$y_i = \log_{10}\left(x_i'\beta\right) + \epsilon_i, \quad \epsilon_i \sim N\left(0,\sigma^2\right), \qquad (12)$$

where the coefficients β follows the cosine curve and σ is set at 0.03. We use the same data splitting scheme from Section 1 and ran five random replications of the simulation. We would expect the estimated $\hat{f}$ from SISR to be near the logarithm function. The left panel in Fig. 5 shows estimated functions $\hat{f}$ (grey) and the shifted logarithm function (black), which is the true underlying *f*. The right panel shows the estimated coefficient curves (grey) and the scaled underlying cosine function (black). We see a snapshot of evidence that SISR can successfully identify the underlying *f* and β.

6. Discussion

We have shown how to estimate non-linear relationships in multivariate calibration, by combining the single index model with penalized signal regression. In simulations, single-index signal regression (SISR), as we call it, recovered a curved relationship reliably. In real data sets the results were mixed. In the Tecator data set a non-linear effect, pointing towards saturation, was discovered and prediction was improved. In the corn data set no curvature and no improvement were found, and SISR defaulted to PSR. In the mixture data the model suffered from the relative sparsity of the design points, and we were not able to reliably estimate a non-linearity.

It is clear that our proposal is not the last word on non-linearities in multivariate calibration. Compared to simplicity of kernel methods, inspired by support vector machines, where one plugs in common alternatives to inner products of vectors, SISR takes more work and is less robust. But when it works and a true non-linearity is found and explicitly modeled, and it can give insights into the physical and chemical process underlying the measurements. In contrast kernel methods are just black boxes.

Our model is related to the problem of estimating an unknown link function in generalized linear models [4]. In the present case the response is assumed to have a normal distribution. Penalized signal regression has also been used for binary classification [13]. The response then is assumed to follow a Bernoulli distribution with probability π_i and the linear predictor is $\eta = \log(\pi/(1-\pi)) = X\beta$. To use SISR in this application, the model would have to be changed to $\log(\pi/(1-\pi)) = f(X\beta) = \eta$, with $f(\cdot)$ the smooth and monotone "missing link" that is to estimated. Tibshirani et al. [16] have worked on a related problem. For other distributions, like Poisson or Gamma, the linear predictor would have to be transformed in a similar way.

Much interesting and useful research is waiting. Prediction intervals and uncertainty could be explored using bootstrap approaches. An obvious generalization is SISR on "images", two-dimensional spectra or other data matrices [14]. In principle it also look possible to put the nonlinearity in another place: instead of $\eta_i = f\left(\sum_j x_{ij}\beta_j\right)$, a non-linear response in each spectroscopic channel could be assumed: $\eta_i = \sum_j f(x_{ij})\beta_j$). These are challenges we are studying now and we hope to report on them in due times.

Although kernel methods are black boxes, they seem to perform quite well. It seems possible to combine the single-index idea with kernels. Related to this the question is whether or not certain non-linearities imply certain types of kernels.

References

[1] Y. Bai, W.K. Wing, Z.Y. Zhu, Penalized quadratic inference functions for single-index models with longitudinal data, Journal of Multivariate Analysis 100 (2009) 152–161.
[2] C. Borggaard, H.H. Thodberg, Optimal minimal neural interpretation of spectra, Analytical Chemistry 64 (1992) 545–551.
[3] C. de Boor, A Practical Guide to Splines, Springer-Verlag, New York, 1978.
[4] C. Cox, Generalized linear models — the missing link, Journal of the Royal Statistical Society, Series C 33 (1984) 18–24.
[5] P. Dierckx, Curve and surface fitting with splines, Clarendon Press, Oxford, 1995.
[6] P.H.C. Eilers, B.D. Marx, Flexible smoothing with B-splines and penalties (with comments and rejoinder), Statistical Science 11 (1996) 89–121.
[7] P.H.C. Eilers, B.D. Marx, Multivariate calibration with temperature interaction using two-dimensional penalized signal regression, Chemometrics and Intelligent Laboratory 66 (2003) 159–174.
[8] T. Fearn, On orthogonal signal correction, Chemometrics and Intelligent Laboratory Systems 50 (2000) 42–52.
[9] R.N. Feudale, H. Tan, S.D. Brown, Piecewise orthogonal signal correction, Chemometrics and Intelligent Laboratory Systems 63 (2002) 129–138.
[10] J.H. Friedman, W. Stuetzle, Projection pursuit regression, Journal of the American Statistical Association 76 (1981) 817–823.
[11] G. James, B. Silverman, Functional adaptive model estimation, Journal of the American Statistical Association 100 (2005) 565–576.
[12] B. Li, B.D. Marx, Sharpening P-spline signal regression, Statistical Modelling, An International Journal 8 (4) (2008) 367–383.
[13] B.D. Marx, P.H.C. Eilers, Generalized linear regression on sampled signals and curves: a P-spline approach, Technometrics 41 (1999) 1–13.
[14] B.D. Marx, P.H.C. Eilers, Multidimensional penalized signal regression, Technometrics 47 (2005) 13–22.

[15] U. Thissen, B. Üstün, W.J. Melssen, L.M.C. Buydens, Multivariate calibration with least-squares support vector machines, Analytical Chemistry 76 (2004) 3099–3105.

[16] R. Tibshirani, M. Saunders, S. Rosset, et al., Sparsity and smoothness via the fused lasso, Journal of the Royal Statistical Society. Series B 67 (2005) 91–108.

[17] B. Üstün, W.J. Melssen, M. Oudenhuijzen, L.M.C. Buydens, Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization, Analytica Chimica Acta 544 (2005) 292–305.

[18] F. Wülfert, W. Kok, A. Smilde, Influence of temperature on vibrational spectra and consequences for the predictive ability of multivariate models, Analytical Chemistry 70 (1998) 1761–1767.

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