Infrared Physics & Technology 61 (2013) 176–183

Contents lists available at ScienceDirect

Infrared Physics & Technology

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

A strategy for multivariate calibration based on modified single-index signal regression: Capturing explicit non-linearity and improving prediction accuracy

Xiaoyu Zhang, Qingbo Li, Guangjun Zhang *

Precision Opto-mechatronics Technology, Key Laboratory of Education Ministry, School of Instrumentation Science and Opto-electronics Engineering, Beihang University, Beijing 100191, China

HIGHLIGHTS

- A new spectral multivariate calibration method is proposed for non-linearity.
- In quantitative analysis, the proposed method shows the best prediction performance.
- The proposed method explicitly exhibits the type and amount of the non-linearity.
- The proposed method has distinct adaptability for the complex spectra model.
- The good performance by the proposed method for biochemical analysis can be expanded.

ARTICLE INFO

Article history:
Received 12 November 2012
Available online 27 August 2013

Keywords:
Multivariate calibration
Modified single-index signal regression
Non-linearity
Spectrometric quantization

ABSTRACT

In this paper, a modified single-index signal regression (mSISR) method is proposed to construct a non-linear and practical model with high-accuracy. The mSISR method defines the optimal penalty tuning parameter in P-spline signal regression (PSR) as initial tuning parameter and chooses the number of cycles based on minimizing root mean squared error of cross-validation (RMSECV). mSISR is superior to single-index signal regression (SISR) in terms of accuracy, computation time and convergency. And it can provide the character of the non-linearity between spectra and responses in a more precise manner than SISR. Two spectra data sets from basic research experiments, including plant chlorophyll nondestructive measurement and human blood glucose noninvasive measurement, are employed to illustrate the advantages of mSISR. The results indicate that the mSISR method (i) obtains the smooth and helpful regression coefficient vector, (ii) explicitly exhibits the type and amount of the non-linearity, (iii) can take advantage of nonlinear features of the signals to improve prediction performance and (iv) has distinct adaptability for the complex spectra model by comparing with other calibration methods. It is validated that mSISR is a promising nonlinear modeling strategy for multivariate calibration.

1. Introduction

It is well known that high-accuracy quantitative calibration is beneficial to improve prediction accuracy in chemometrics [1]. Accordingly, quantitative calibration is a key point to extract analyte information by building a relationship of response variable (property) and predictor variables (wavelength). Linear model such as partial least squares (PLS) regression is often used in some researches. Whereas it is known that non-linearity is an inherent trait for systems, and linear model is inappropriate to describe the underlying data structure with significant nonlinear characteristics. For example, in the leaf chlorophyll nondestructive measurement using Vis–NIR spectroscopy, differences in species, healthy state, growing state and so on complicate the leaf spectrum and induce additional mendacious and nonlinear factors; in the noninvasive measurement of human blood glucose with NIR spectroscopy, the linear relationship based on Lambert–Beer Law is not tenable due to many factors, such as the complexity of blood components, the interaction between components, the distribution irregularity of blood components because of macromolecules (protein, fats, etc.), the effect of colored noise, baseline drift and so on [2–8].

Single-index signal regression (SISR) is a nonlinear method that combines ideas of projection pursuit regression [9] with P-spline signal regression (PSR) [10]. SISR is related to the problem of estimating 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 estimated

* Corresponding author. Tel./fax: +86 10 82338768.
E-mail address: xiaoyuzhangzi@126.com (G. Zhang).

1350-4495/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.infrared.2013.08.008

using P-splines [11]. To exploit and explore an explicit nonlinear effect is the added benefit provided using SISR. However, using a large value of initial tuning parameter in SISR falls into a local minimum, and prediction performance is poor. And it is very difficult to attain convergence for SISR. Additionally, SISR is time consuming.

Correspondingly, a modified single-index signal regression (mSISR) that defines the optimal penalty tuning parameter in the method of PSR as initial tuning parameter and chooses the number of cycles based on cross-validation is proposed in this paper. The approach outperforms SISR in terms of precision and computational speed, and it can avoid the problem of being hard to converge. Moreover, unlike "black box" approach, e.g. kernel partial least squares (KPLS) [12,13], the approach explicitly and accurately models the non-linearity, allowing us to learn something about its features, while enhancing insight into the measurement process. Two spectra data sets from basic research experiments, namely plant chlorophyll nondestructive measurement and human blood glucose noninvasive measurement, are adopted to evaluate the performance of the proposed strategy. The regression coefficients and prediction performance of the models constructed by different methods are analyzed in this paper. And the adaptability of the proposed method to the complex spectra model with many uncertain factors is discussed. In this research, a feasible nonlinear model with optimal parameters is selected for multivariate calibration.

2. Method

Since P-spline signal regression and single-index signal regression are the basis for the proposed method, they are described before the proposed method in order to make the reader understand the proposed method in depth.

2.1. P-spline signal regression (PSR)

Consider a standard regression approach:

$$E(y) = X_{m\times p}\beta_{p\times 1} \tag{1}$$

where y is the realization of the response, X is the spectra matrix and β is the unknown regression coefficient vector. Typically the number p of regressors far exceeds the number m of observations.

The goal of PSR is smoothness in β, and this is achieved through dimension reduction by first projecting β onto a rich B-spline basis using moderate number of equally spaced knots (n-dimensional, $n < p$), i.e. $\beta_{p\times 1} = B_{p\times n}\alpha_{n\times 1}$. Some specifics of the B-spline basis see Ref. [14]. The vector α is the unknown vector of basis coefficients of modest dimension. Notice that Eq. (1) can be rewritten as:

$$E(y) = U_{m\times n}\alpha_{n\times 1} \tag{2}$$

where $U = XB$. Then PSR further increases smoothness by imposing a difference penalty on adjacent B-spline coefficients in the α vector [10,15].

The penalized least-squares solution simplifies as [10,15]

$$\mathrm{PSR}(U, y, \lambda, d, n) = \hat{\alpha} = (U^T U + \lambda D_d^T D_d)^{-1} U^T y \tag{3}$$

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 d of the difference penalty can moderate smoothing. The non-negative tuning parameter λ regularizes the penalty and can be chosen through a logarithmic grid search.

PSR typically uses between 10 and 200 equally spaced cubic B-splines. The order of the difference penalty can vary ($d = 3$ to 0). For fixed d, increasing λ makes α smoother and optimal λ is searched for systematically by monitoring cross-validation prediction error. Results of these optima can be directly compared over the various $d = 0, 1, 2, 3$. Given choice of d and λ, then the p-dimensional regression coefficient vector can be constructed, $\hat{\beta} = B\hat{\alpha}$[11].

2.2. Single-index signal regression (SISR)

The SISR model has the form $E(y) = f(U\alpha)$, where the function $f(\cdot)$ is assumed to be smooth and is estimated from the data using P-splines, having its own additional tuning parameter. SISR is extremely flexible: even minor departures in f from the identity function can lead to relatively dramatic changes in the estimated coefficient vector, while significantly improving prediction. The model fitting algorithm [11] is described below.

Firstly, SISR carries out a PSR with the response y on U, and the basis coefficient estimates $\hat{\alpha}$ can be calculated through a $\mathrm{PSR}(U, y, \lambda_0, d_1, n_1)$, where λ_0 is the initial tuning parameter, d_1 is the penalty order, and n_1 is the number of B-splines.

Secondly, a cubic P-spline scatter smoother is employed to obtain the estimation of function f, which driven by $\hat{\alpha}$. The penalty on γ ensures a smooth f; recall that γ is the vector of B-spline coefficients with equally-spaced knots placed along estimated linear predictor $U\hat{\alpha}$. For simplicity in notation, denote $S(U\hat{\alpha}, y, \lambda_2, d_2, n_2)$ as the operation of fitting a cubic P-spline scatter smoother on $U\hat{\alpha}$ (the input variable) and y (the response) using the penalty tuning parameter λ_2 and difference order d_2 on the n_2 equally-spaced B-splines. Apparently, f and its derivative f' can be estimated from $S(U\hat{\alpha}, y, \lambda_2, d_2, n_2)$.

Thirdly, the basis coefficient estimates $\hat{\alpha}$ can be updated using a first-order Taylor series approximation of the function f (about the current estimate α_0 for α), i.e., with fixed f, the optimal value for α can be obtained through a $\mathrm{PSR}(U^*, y^*, \lambda_1, d_1, n_1)$, where λ_1 is also the penalty tuning parameter, $y^* = y – f(U\alpha_0) + diag\{f'(U\alpha_0)\}U\alpha_0$ and $U^* = diag\{f'(U\alpha_0)\}U$.

To simultaneously estimate the final coefficient vector and nonlinear relationship, the estimation between f and α is iterative, which is extremely tractable, essentially boiling down to repeated alternate applications of PSR and P-spline smoothing on "working" responses and predictors. Eilers et al. [11] proposed to cycle back and forth between PSR and P-spline smoothing until convergence of $\hat{\alpha}$.

2.3. Modified single-index signal regression (mSISR)

For SISR, using a large value of λ_0 falls into a local minimum, and prediction performance is poor. Additionally, it is computationally intensive to try several different initial values of λ_0 to avoid the solution falls into a local minimum. Therefore we propose to define the optimal λ in the method of PSR (see Section 2.1) as the initial tuning parameter λ_0 in mSISR, which can avoid falling into a local minimum, provide more chance for the algorithm to find the optimal α and f, and as a consequence improve prediction performance of quantitative calibration model.

On the other hand, it is very difficult to attain convergence of $\hat{\alpha}$ for SISR. Accordingly, the number of cycles in the mSISR method is chosen based on minimizing root mean squared error of cross-validation (RMSECV). This prevents underfitting and overfitting, so the coefficient vector and the nonlinear function of linear prediction can be accurately estimated and, furthermore, a reliable prediction can be obtained.

Besides, the mSISR model is also driven by the non-negative penalty regularization parameters $\lambda = (\lambda_1, \lambda_2)$, which drive the continuous control over smoothness. Generally, we perform a two dimensional linear grid search, where each element of (λ_1, λ_2) is varied on a logarithm scale. And the optimal values for (λ_1, λ_2) are determined by minimizing RMSECV. Since several model parameters need to be determined by cross-validation, which is

computationally intensive, for the sake of convenience, we use the optimal n and d in the PSR method as n_1 and d_1, respectively.

In theory, the proposed strategy has four advantages for multivariate calibration: (i) a more practical calibration model can be constructed, because the selection of model parameters requires less computational time and the problem that it is very difficult to attain convergence of $\hat{\alpha}$ will not occur; (ii) under the B-spline trick, a more parsimonious model can be got, because only a matrix U of size $m \times n_1$ is employed through dimension reduction; (iii) the prediction accuracy can be improved greatly, because an accurate nonlinear relationship of response variable and wavelength variables are constructed for extracting the useful information sufficiently; (iv) because f is estimated more precisely, the character of the non-linearity between spectra and responses can be exhibited in a more precise manner, which can give more insights into the physical and chemical process underlying the measurements.

3. Experimental

3.1. Vis–NIR experiment of leaf chlorophyll nondestructive measurement

Leaf biochemical parameter such as chlorophyll content can provide valuable insight into the physiological performance of plants. *Epipremnum aureum* was used as a representative plant because the thickness of different locations for the same leaf decreases gradually from leaf root to leaf apex. It is theorized that the chlorophyll content is symmetrically distributed for the same leaf. Six *E. aureum* leaves with different green and sapless levels were selected. All of them were healthy and homogeneous in color without anthocyanin pigmentation or visible symptoms of damage. Spectra of six different locations per sample were measured, and 36 sample spectra were obtained as predictor variables for the response variable, i.e., chlorophyll content. So in this study, only the nonlinear effect caused by thickness difference of the same species is taken into account to avoid alternating influence.

An Ocean Optics (Dunedin, Florida) spectrometer and diffuse reflectance sample accessories *Y* style fiber were used for spectra measurement. The light source was a white light. A white panel (Spectralon, Labsphere, North Sutton, New Hampshire) was used as a 100% reflectance standard for all measurements. The parameters of the spectrometer were as follows: spectrum scanning range, 350–1050 nm; number of pixels, 3648; integration time, 15 ms; average time, 20 ms; width of smooth window, 3. The data were stored in the form of reflectance. Due to the low spectral intensity of the halogen lamp used below 450 nm and the resulting noise in the measured spectra, only reflectance data above 450 nm were considered.

To obtain reference values of chlorophyll content, each leaf was cut into fragments and extracted with 80% aqueous solution with acetone and then centrifuged. The absorption spectra of the acetone extract were measured with the same spectrophotometer. The concentration of chlorophyll was calculated based on the absorbance measured at 646.6, 663.6, and 750 nm according to the Porra formula [16].

3.2. NIR experiment of human blood glucose noninvasive measurement

Body oral glucose tolerance test (OGTT) is a kind of glucose burden adjustability test for diagnosis of diabetes in clinic. For the healthy person, under this test, a varying scope of glucose concentration can be obtained in a short period of time [17,18]. It can be used as a special experiment method for getting a calibration model with a certain concentration variety, which gives a novel way for the research of constructing calibration model in blood glucose noninvasive measurement using NIR spectroscopy. Because of the human individual difference, it is impossible to build a universal prediction model by the current NIR spectroscopy technology for blood glucose noninvasive measurement. But the personal knowledge base is utilized usually. So in this paper, the experiment only included one volunteer. All the modeling is aiming to this one. Of course, the method can be repeated similarly for the other people.

NIR spectra were obtained by using Nicolet FTIR 6700 Spectrometer (Thermo scientific, America) with InGaAs 2.6 μm detector, CaF_2 beam splitter and white light source. An integrating sphere sampling accessory was used in this experiment. The One Touch® Ultra®2 Blood Glucose Meter (LifeScan, Inc., America) was used for getting reference values of blood glucose concentration.

Experiment procedure is described here. A healthy volunteer had been fasted for 8 h before the experiment began. Then he drank 100 mL solution with 75 g glucose within 5 min, and the NIR diffuse reflection spectra were collected from the finger pulp. At the time of sampling, the measurement position, measurement pressure as well as the psychology of the volunteer kept invariableness as far as possible. At meantime, the corresponding blood glucose reference values were obtained by using One Touch® Ultra®2 Blood Glucose Meter. In the course of this experiment, the human blood glucose concentration value increased gradually to the peak value 217.8 mg dL^{-1}, and then decreased. The experiment was over when the concentration value was down to the normal level 84.6 mg dL^{-1}. The whole time cost in this experiment was 3 h.

There are twenty-one samples got by this experiment. The blood glucose concentration scope is 84.6–217.8 mg dL^{-1}. The spectra scan scope is 1000–2500 nm with 3112 variables.

3.3. Calculation and software

Because of the limit of experiment condition, such as glucose concentration only changed in 3 h for OGTT experiment, pricking the finger several times is painful and the position that can be pricked is also a limited region, there are many variables and fewer observations for these two data sets. So the root mean squared error of prediction (RMSEP) of cross-validation is employed as an evaluation criterion for the predictive ability of calibration model. All further calculations are performed with Matlab 7.6.0 (The Mathworks, Inc., Natick, MA, USA). The convergence of $\hat{\alpha}$ in SISR cannot be obtained for these two data sets, so the SISR results and comparisons to SISR have not been displayed in this paper.

4. Results

4.1. Model for Vis–NIR data of plant leaf samples

Fig. 1 presents the raw spectra of plant leaf samples. mSISR is employed to construct a nonlinear model for the raw spectra which is used as the regressor. The nonlinear response is obtained by generating a linear response and a nonlinear function. The linear response is constructed using the raw spectra from the plant leaf experiment and a linearly estimated coefficient vector. From Section 2.3, this paper chooses the optimal model parameters (except λ_0, d_1 and n_1) in mSISR based on minimizing RMSECV. The coefficient vector is determined using standard PSR with 58 equally-spaced B-splines, a zero order difference penalty and a tuning parameter 10^5 (initial value 10^5). The nonlinear function is determined using 5 equally-spaced B-splines, a third order difference penalty and a tuning parameter 10^{-6}. Note that 26 cycles are needed in mSISR.

The unique contribution of mSISR, i.e. the explicit and accurate estimation of the nonlinear link function, is highlighted. Since all

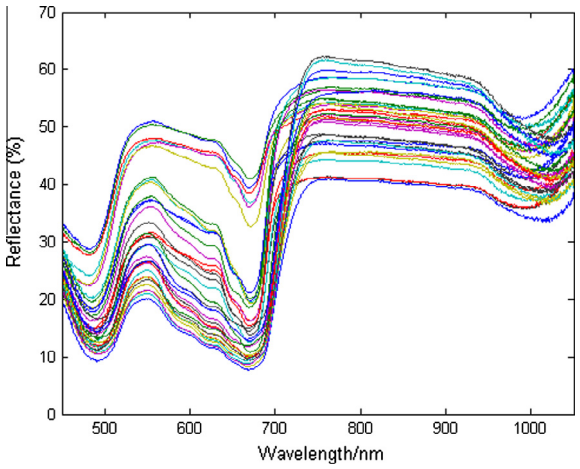

Fig. 1. Original spectra of plant leaf experiment.

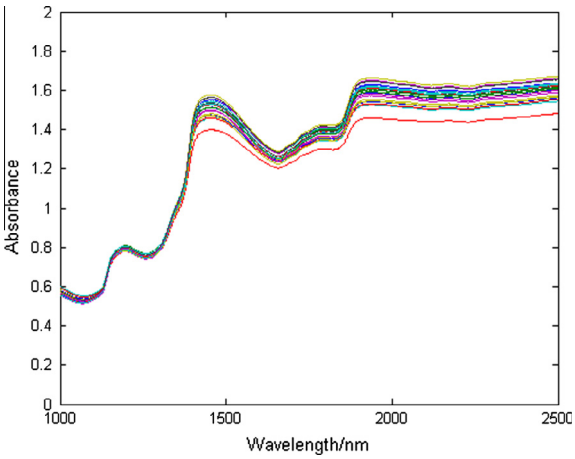

Fig. 3. Original spectra of OGTT experiment.

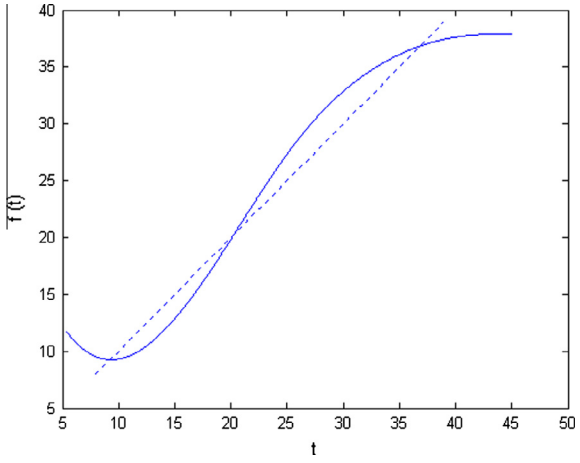

Fig. 2. Estimated nonlinear function (solid line) for plant leaf experiment.

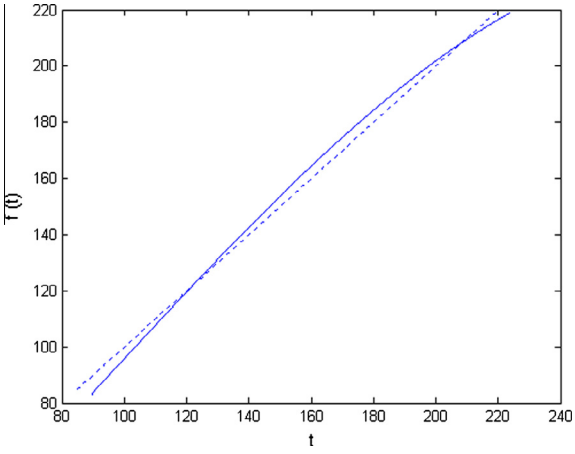

Fig. 4. Estimated nonlinear function (solid line) for OGTT experiment.

nonlinear functions generated during the cross-validation cannot be listed one by one due to limitations of space, without loss of generality, a nonlinear function f estimated using all leaf data and optimal model parameters is displayed. Fig. 2 shows f, pointing towards saturation, relative to the dashed identity line. The result suggests two quite sharp bends near ten and thirty in an otherwise quite smooth function for leaf chlorophyll. We expect that mSISR can capture explicit and accurate nonlinearity caused by thickness difference through the estimation of nonlinear function, while have a smaller RMSEP than other counterparts.

In order to estimate effectiveness of quantitative calibration models for leaf chlorophyll content nondestructive measurement using Vis–NIR spectroscopy, different strategies, PLS, KPLS using Gaussian kernel and PSR are introduced to compare with mSISR. The model parameters and prediction results are given in Table 1 for this experiment.

4.2. Model for NIR data of human OGTT samples

In this part, the raw spectra of human OGTT samples are presented in Fig. 3. The raw spectra are employed to construct a nonlinear mSISR model. Similarly, RMSECV is used for optimization of model parameters (except λ_0, d_1 and n_1) in mSISR. The approach takes 27 equally-spaced B-splines, a third order difference penalty and a tuning parameter 10^{-6} (initial value 0.01) for the spectra coefficient vector. To estimate the nonlinear function, 5 equally-spaced B-splines are used with a third order difference penalty and a tuning parameter 0.01. And only fourteen cycles are needed in mSISR. A nonlinear function f is constructed using all human OGTT data and optimal model parameters. The estimated link function f, as shown in Fig. 4, is clearly monotonically increasing and exhibits the recovery of some of the true underlying nonlinear response features. Similar to Vis–NIR experiment of plant leaf, Table 2 gives the model parameters and prediction results of different calibration methods for the OGTT data set.

Table 1
Comparison between different calibration methods for the Vis–NIR data of plant leaf samples.

Method	h	σ	λ_0	λ	d	n	Number of cycles	RMSEP (mg dL^{-1})	Correlation coefficient
PLS	5	–	–	–	–	–	–	3.8	0.948
KPLS	9	350	–	–	–	–	–	2.9	0.974
PSR	–	–	–	10^5	0	58	–	3.6	0.958
mSISR	–	–	10^5	(10^5, 10^{-6})	(0, 3)	(58, 5)	26	2.3	0.984

Table 2
Comparison between different calibration methods for the NIR data of human OGTT samples.

Method	h	σ	λ_0	λ	d	n	Number of cycles	RMSEP (mg dL^{-1})	Correlation coefficient
PLS	8	–	–	–	–	–	–	22.5	0.861
KPLS	11	4×10^5	–	–	–	–	–	21.9	0.864
PSR	–	–	–	0.01	3	27	–	14.6	0.941
mSISR	–	–	0.01	$(10^{-6}, 0.01)$	(3, 3)	(27, 5)	14	11.1	0.967

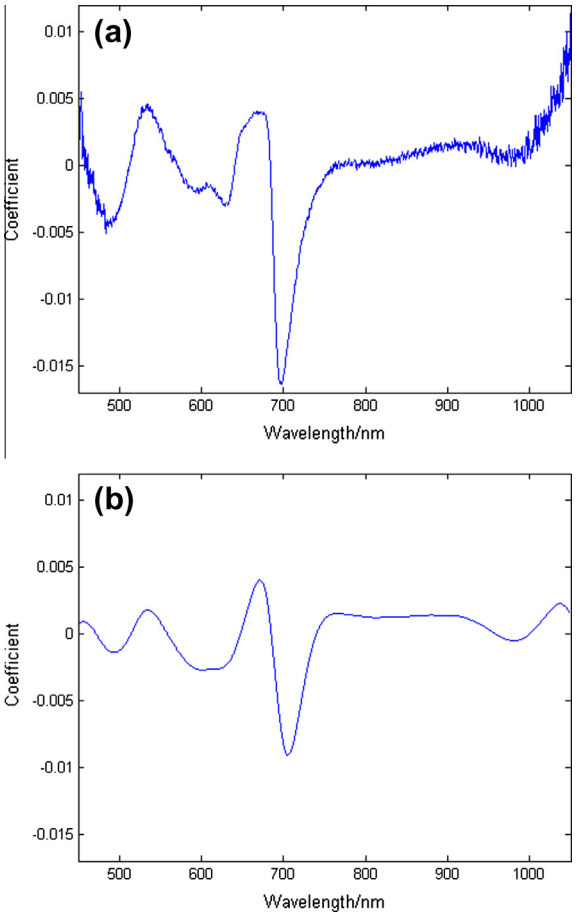

Fig. 5. Regression coefficient curves for the Vis–NIR data of leaf samples obtained by (a) PLS and (b) mSISR.

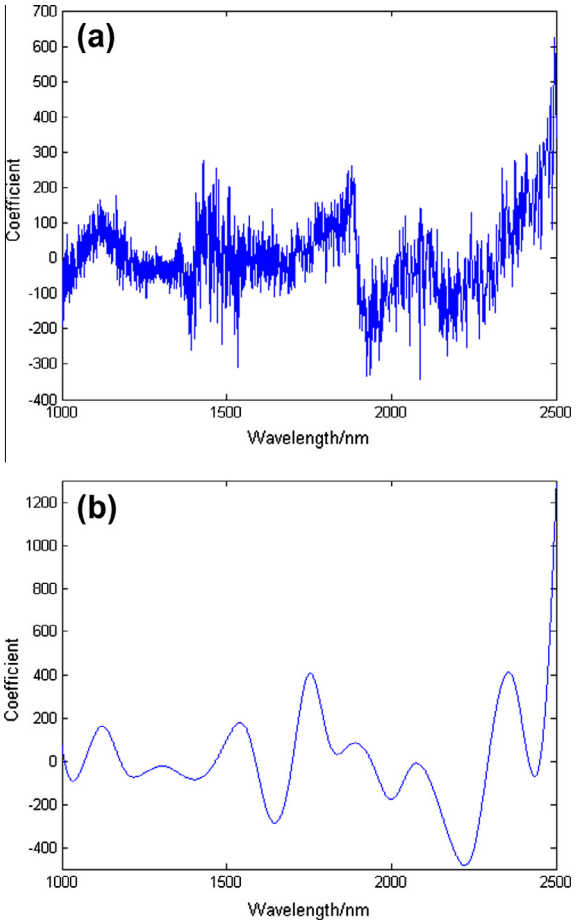

Fig. 6. Regression coefficient curves for the NIR data of human OGTT samples obtained by (a) PLS and (b) mSISR.

5. Discussion

5.1. Model regression coefficient comparison and analysis

The regression coefficient is an important parameter of any calibration model. It is generated in the calibration process and is used to predict the composition of the sample. A discussion about regression coefficient curve of the model is necessary. The regression coefficient curve, with weak noise and obvious valleys and peaks, is helpful not only for the interpretation of model, but also for the improvement of prediction accuracy, because the tested composition of sample suffers little interference. As a result, smooth regression coefficients make more sense and have advantages over extremely erratic coefficients. Constructing the model based on the smooth regression coefficient curve with weak noise and obvious valleys and peaks is very beneficial, and can improve the predictive accuracy of model; the extremely erratic regression coefficient curve with high noise and unconspicuous valleys and peaks is adverse to the construction of high-accuracy model.

Like f, regression coefficients are estimated using all experimental data and optimal model parameters. The regression coefficient curves of PLS and mSISR models for the leaf data are shown in Fig. 5. It can be seen that the regression coefficient curve of PLS has some noise. However, the regression coefficient curve of mSISR is smooth without noise and the valley and peak positions are obvious. This is mainly because mSISR automatically builds in smooth structure associated with the coefficient index, by virtue that a linear combination of smooth B-splines produces a smooth curve, and smoothness is further increased by using a difference penalty and a penalty tuning parameter; besides, the estimation of nonlinear function and the choice for initial tuning parameter and number of cycles in mSISR can lead to a more precise estimation of coefficient vector. The smooth coefficients of mSISR method have advantages over the PLS coefficients for Vis–NIR experiment of plant leaf, which is consistent with the prediction results found in Table 1.

Fig. 6 displays the regression coefficient curves of PLS and mSISR models for the OGTT data. From this figure, it is observed

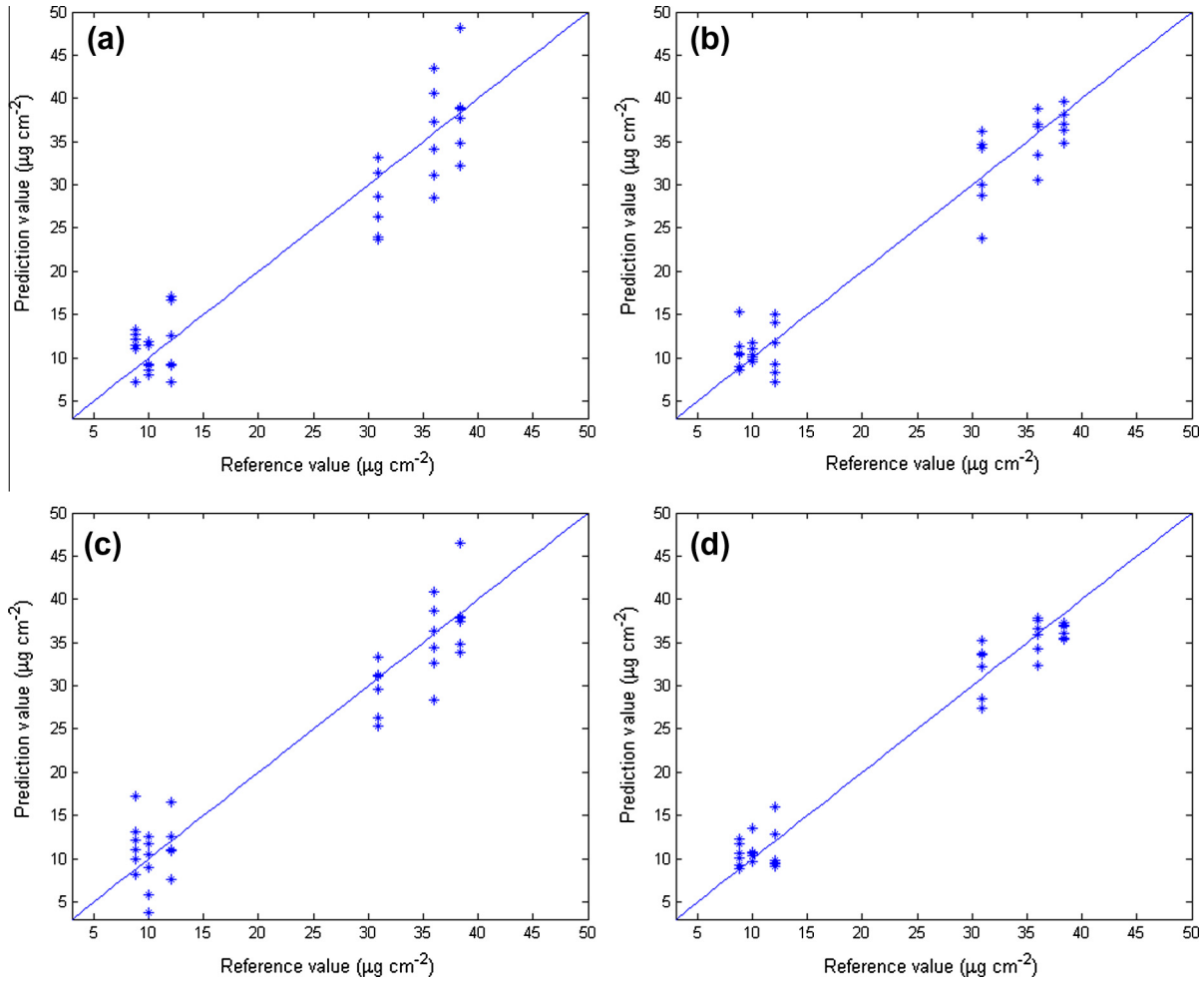

Fig. 7. Prediction vs. reference values for the Vis–NIR data of leaf samples obtained by (a) PLS, (b) KPLS, (c) PSR and (d) mSISR. The solid lines correspond to the ideal, unity correlation between prediction and reference contents.

that the regression coefficient curve of PLS has very high noise and some non-smooth behavior (kinks, jumps, or narrow peaks), furthermore, the peak and valley positions of the curve are concealed. The general feature of Fig. 6a is that the PLS coefficients are extremely erratic along the indexing domain. Calibration with mSISR has a smooth and noiseless regression coefficient curve, and the peak and valley regions could be divided obviously. Consequently, the predictive accuracy of human blood glucose noninvasive measurement using NIR spectroscopy can be further improved, which is also consistent with the corresponding increase in RMSEP of 22.5 (PLS) compared to 11.1 (mSISR).

5.2. Model prediction capability comparison and analysis

The main objective of this article is to investigate whether the model constructed using the mSISR method can be used in the actual measurement. We focus on a prediction performance study that directly compares the mSISR model to the linear PLS and PSR models and to the nonlinear KPLS model using Gaussian kernel.

A good calibration model needs a relatively bigger correlation coefficient and a smaller RMSEP. As shown in Table 1, calibration with mSISR for the leaf data set obtains the best results, i.e., the lowest RMSEP and the highest correlation. Under the nonlinear modeling strategy of mSISR, the RMSEP is 2.3 μg cm^{-2}, which is decreased 39% of one with PLS, 21% of one with KPLS using Gaussian kernel and 36% of one with PSR. The differences between the RMSEPs turn out to be pronounced. Fig. 7 shows prediction vs. reference values for all the samples in the Vis–NIR data set of plant leaf. It is clearly visible from Fig. 7 that predictions made with mSISR are more precise as with other methods. mSISR obtains an overall good agreement between estimated and true contents. The reasons are as follows: the choice for initial tuning parameter λ_0 avoids falling into a local minimum, so the better α and f can be obtained; the number of cycles and the optimal values for λ_1, λ_2, d_2, n_2 can be determined by RMSECV, which prevents overfitting or underfitting, and as a consequence helps in estimating α and f accurately; α is further regularized by accounting for the nonlinear effect caused by thickness difference, yielding stronger prediction results with reasonable parameters; the smooth regression coefficient vector β is meaningful and beneficial to the chlorophyll content information extraction, so the prediction performance of calibration model can be improved.

Table 2 illustrates that for the OGTT data set, the best prediction accuracy is still obtained by the nonlinear mSISR method, and the RMSEP is decreased 51% of one with PLS, 49% of one with KPLS using Gaussian kernel and 24% of one with PSR. The correlation coefficient between prediction and reference values is very distinct for different calibration methods, and the calibration with mSISR still has the best correlation. Under different calibration strategies, the reference values and the prediction values of glucose concentration are also shown in Fig. 8 for human OGTT experiment. From Fig. 8, it is observed that the prediction made with mSISR obviously outperforms the other methods. It can be concluded that when a

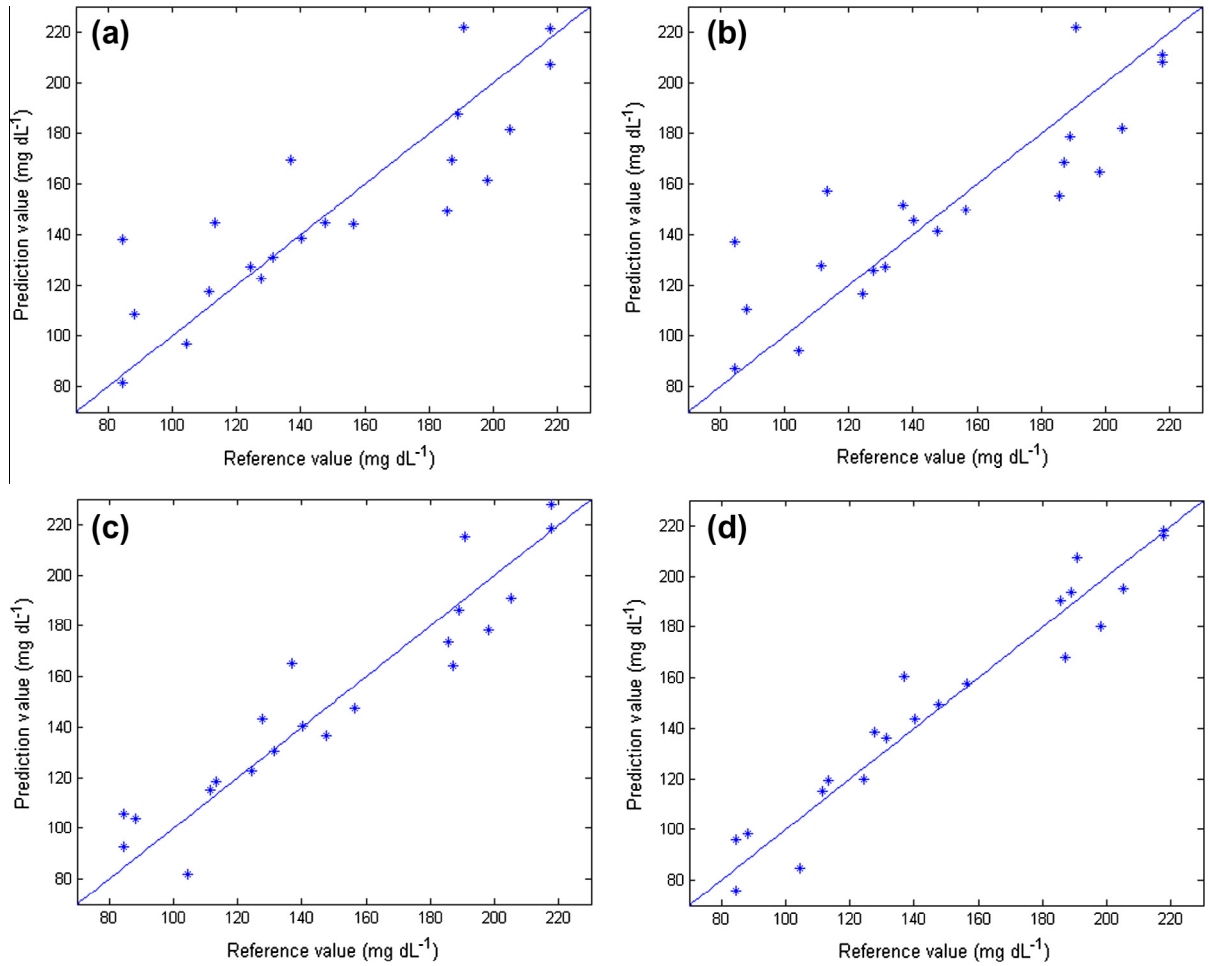

Fig. 8. Prediction vs. reference values for the NIR data of human OGTT samples obtained by (a) PLS, (b) KPLS, (c) PSR and (d) mSISR. The solid lines correspond to the ideal, unity correlation between prediction and reference concentrations.

non-linearity is present in the data, the mSISR algorithm is able to reliably extract the relevant information and successfully identify the underlying f and β, furthermore, excellent predictions of f and β translate into the excellent prediction of glucose concentration.

For Vis–NIR experiment of plant leaf, the RMSEP of PLS model amounts to a relative gain of about 24% in comparison with that of KPLS model based on Gaussian kernel; for the complex NIR data of human noninvasive measurement experiment, the RMSEP of KPLS model based on Gaussian kernel is decreased slightly by 3% compared with PLS. However, the RMSEP of mSISR is substantially lower than that of PLS obtained for these two data sets. From this point of view, it is implied that mSISR has better robustness than its nonlinear competitor KPLS, especially when employed to the complex spectra data. It is encouraged that mSISR can be used as a powerful tool for establishing a complex spectra model with good robustness.

5.3. Adaptability discussion for complex spectra model

The spectra model of human blood glucose noninvasive measurement is more complex than the spectra model of plant chlorophyll nondestructive measurement. The complexity is mainly from two aspects: on the one hand, the sample components in human OGTT experiment are more complex than those in plant leaf experiment; on the other hand, the spectra model obtained by plant leaf experiment is steady comparatively because components in the leaf sample are invariable, but the spectra model of human blood glucose measurement can be affected by many factors, such as measurement position, measurement time, measurement pressure and human physiological state, which are strongly uncertain and cannot be estimated [1]. The adaptability to complex spectra model is a very important part. It is an appraisement item for the multivariate calibration method. The prediction results of different calibration methods for the two spectra data indicate that:

(i) Under the method of mSISR, it can be seen that the RMSEP is decreased 39% from 3.8 to 2.3 $\mu g\ cm^{-2}$ for the experimental data of plant leaf; and the RMSEP is decreased 51% from 22.5 to 11.1 $mg\ dL^{-1}$ for the experimental data of OGTT. It is validated that this method is adapted to the complex OGTT spectra model, in which the improvement of prediction accuracy of human blood glucose noninvasive measurement is better than that of plant leaf experiment.

(ii) For the spectra data with certain element like as leaf samples experiment, the RMSEP of KPLS model based on Gaussian kernel is decreased compared with PLS. Whereas the prediction performance of KPLS model using Gaussian kernel is close to that of PLS model for the complex spectra of blood glucose noninvasive measurement. These results demonstrate that the prediction by using KPLS is worse for complex spectra data. Conversely, the best prediction accuracy is got by using mSISR for these two spectra data, and the predictions of mSISR are both by far superior to those of PLS. Accord-

ingly, comparing mSISR with its nonlinear competitor KPLS gives a clear outcome: mSISR also has better adaptability to complex spectra model. In this regard, mSISR is a better alternative for multivariate calibration.

(iii) Under different calibration strategies, the correlations for different spectra data sets are calculated respectively. It is clear that, for the spectra model of plant leaf, the correlation coefficients of different calibration methods are about 0.97, which means that different modeling strategies utilized for this experimental spectra model all have very similar correlation. However, for the more complex spectra model, which is obtained by OGTT experiment, the correlation is very different for different calibration methods, and calibration with the mSISR method has the best correlation. These phenomena provide convincing evidence that the mSISR method has distinct adaptability and robustness to the complex spectra model, which has a very important application meaning for multivariate calibration.

6. Conclusion

A modified single-index signal regression is proposed in this paper for nonlinear modeling. We have shown how to estimate nonlinear relationship in multivariate calibration with mSISR, by the appropriate selection of model parameters. The basic appeal of mSISR is its simplicity. mSISR is straight-forward to use: it uses the entire (“raw”) signal and works without any data preprocessing. It is superior to SISR in terms of accuracy, computation time and convergency. Besides, it picks up the type and amount of the non-linearity in a more precise manner compared with SISR. mSISR quantification is successfully applied to two experimental spectra data sets for analysis of biochemical parameter. We stress that our mSISR approach is not only a competitor, but has some clear advantages: since it can capture the smooth and accurate nonlinear function and regression coefficient information, for these two data sets it performs much better than PLS, KPLS using Gaussian kernel and PSR in terms of precision; the nonlinearity between spectra and responses is clearly estimated with a smooth function, and the explicit estimation of nonlinearity can provide some insights into the physical and chemical process underlying the measurements, which we view as a contribution over “black box” approaches; it has better adaptability for complex spectra model, which possesses potential capability to multivariate calibration. It is expected that the optimal prediction accuracy can be obtained when the most informative wavelength bands, the fitting pretreatment method and the mSISR calibration are all used in the study. Therefore mSISR is a promising method. The excellent performance by mSISR for biochemical parameter determination can be expanded and more stable for future practical applications.

Acknowledgements

This work is supported by Programs for Changjiang Scholars and Innovative Research Team (PCSIRT) in University of China (IRT0705) and National Natural Science Foundation (60708026).

References

[1] X.Y. Zhang, Q.B. Li, G.J. Zhang, Modified robust continuum regression by net analyte signal to improve prediction performance for data with outliers, Chemom. Intell. Lab. Syst. 107 (2) (2011) 333–342.
[2] C.E. Ferrante do Amaral, B. Wolf, Current development in non-invasive glucose monitoring, Med. Eng. Phys. 30 (5) (2008) 541–549.
[3] Q.B. Li, G.J. Zhang, K.X. Xu, Y. Wang, Application of digital Fourier filtering pretreatment method to improving robustness of multivariate calibration model in near infrared spectroscopy, Spectrosc. Spect. Anal. 27 (8) (2007) 1484–1488.
[4] Q.B. Li, K.X. Xu, Y. Wang, Primary discussion on prerequisites to noninvasive blood glucose, J. Tianjin Univ. 36 (2) (2003) 139–142.
[5] S. Arazuri, C. Jarén, J.I. Arana, Selection of the temperature in the sugar content determination of Kiwi fruit, Int. J. Infrared Millimeter Waves 26 (4) (2005) 606–607.
[6] C. Jarén, S. Arazuri, M.J. Garća, P. Arnal, J.I. Arana, White asparagus harvest data discrimination using NIRS technology, Int. J. Infrared Millimeter Waves 27 (3) (2006) 391–401.
[7] C. Jarén, J.C. Ortuño, S. Arazuri, J.I. Arana, M.C. Salvadores, Sugar determination in grapes using NIR technology, Int. J. Infrared Millimeter Waves 22 (10) (2001) 1521–1530.
[8] Q. Ding, G.W. Small, M.A. Arnold, Evaluation of nonlinear model building strategies for the determination of glucose in biological matrices by near-infrared spectroscopy, Anal. Chim. Acta 384 (3) (1999) 333–343.
[9] J.H. Friedman, W. Stuetzle, Projection pursuit regression, J. Am. Stat. Assoc. 76 (1981) 817–823.
[10] B.D. Marx, P.H.C. Eilers, Generalized linear regression on sampled signals and curves: a P-spline approach, Technometrics 41 (1999) 1–13.
[11] P.H.C. Eilers, B. Li, B.D. Marx, Multivariate calibration with single-index signal regression, Chemom. Intell. Lab. Syst. 96 (2009) 196–202.
[12] R. Rosipal, L.J. Trejo, Kernel partial least squares regression in reproducing kernel Hilbert space, J. Mach. Learn. Res. 2 (2001) 97–123.
[13] R. Rosipal, Kernel partial least squares for nonlinear regression and discrimination, Neural Netw. World 13 (3) (2003) 291–300.
[14] P.H.C. Eilers, B.D. Marx, Flexible smoothing with B-splines and penalties (with comments and rejoinder), Stat. Sci. 11 (1996) 89–121.
[15] X.Y. Zhang, Q.B. Li, G.J. Zhang, Multivariate calibration for spectral analysis based on P-spline signal regression with net analyte signal, Spectroscopy 28 (4) (2013) 40–47.
[16] R.J. Porra, W.A. Thompson, P.E. Kriedemann, Determination of accurate extinction coefficients and simultaneous equations for assaying chlorophylls a and b extracted with four different solvents: verification of the concentration of chlorophyll standards by atomic absorption spectroscopy, Biochim. Biophys. Acta 975 (1989) 384–394.
[17] S.F. Malin, T.L. Ruchti, T.B. Blank, S.N. Thennadil, S.L. Monfre, Noninvasive prediction of glucose by near-infrared diffuse reflectance spectroscopy, Clin. Chem. 45 (9) (1999) 1651–1658.
[18] K.X. Xu, Q.J. Qiu, J.Y. Jang, X.Y. Yang, Non-invasive glucose sensing with near-infrared spectroscopy enhanced by optical measurement conditions reproduction technique, Opt. Lasers Eng. 43 (10) (2005) 1096–1106.