Infrared Physics & Technology 53 (2010) 410–417

Contents lists available at ScienceDirect

Infrared Physics & Technology

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

A nonlinear model for calibration of blood glucose noninvasive measurement using near infrared spectroscopy

Qing-Bo Li, Li-Na Li, Guang-Jun Zhang *

Precision Opto-Mechatronics Technology, Key Laboratory of Education Ministry, Beihang University, B528 Xin-Zhu-Lou, 37 Xue Yuan Road, Hai Dian District, Beijing 100191, China

ARTICLE INFO

Article history:
Received 31 March 2010
Available online 5 August 2010

Keywords:
Blood glucose noninvasive measurement
Near infrared spectroscopy
Nonlinear
Kernel partial least squares

ABSTRACT

In order to improve prediction accuracy of calibration in human blood glucose noninvasive measurement using near infrared (NIR) spectroscopy, a modified uninformative variable elimination (mUVE) method combined with kernel partial least squares (KPLS), named as mUVE–KPLS, is proposed as an alternative nonlinear modeling strategy. Under the mUVE method, high-frequency noise and matrix background can be eliminated simultaneously, which provide a optimized data for calibration in sequence; under the kernel trick, a nonlinear relationship of response variable and predictor variables is constructed, which is different with PLS that is a complex model and inappropriate to describe the underlying data structure with significant nonlinear characteristics. Two NIR spectra data of basic research experiments (simulated physiological solution samples experiment in vitro and human noninvasive measurement experiment in vivo) are introduced to evaluate the performance of the proposed method. The results indicate that, after elimination high-frequency noise and matrix background from optical absorption of water in NIR region, a high-quality spectra data is employed in calibration; and under the selection of kernel function and kernel parameter, the best prediction accuracy can be got by KPLS with Gaussian kernel compared with Spline-PLS and PLS. It is encouraging that mUVE–KPLS is a promising nonlinear calibration strategy with higher prediction accuracy for blood glucose noninvasive measurement using NIR spectroscopy.

1. Introduction

Nowadays, diabetes is a modern disease in worldwide. There is no effective cure strategy in clinic so far. The recommendatory method proposed by World Health Organization (WHO) is that, monitoring blood glucose concentration frequently in home, and controlling blood glucose level by injecting corresponding dosage of insulin. Although most blood glucose monitor used in home is minimally invasive, pricking the finger several times a day makes the patient suffer from inevitable pain and infection. Thus, non invasive blood glucose measurement has been an intriguing concept and it has challenged researchers for a number of years [1–3]. It is proved that near infrared (NIR) spectroscopy combined with chemometrics is one of the most promising techniques for developing a noninvasive blood glucose monitoring system for diabetic patients [4–6]. However, there has no successful clinical application report about this technology so far because its prediction accuracy is not satisfied in clinic.

The main holdback to the prediction accuracy is that, the relatively low absorbance of glucose and the relatively high overlapping absorption of interferences in NIR region [7]. The optimal prediction performance cannot be got when calibration model is constructed by employing the NIR spectra data containing uninformative signal [8]. So the choice of proper data analysis strategies (i.e. pretreatment and calibration) is very important to extract weak glucose information from the overlapped spectra, which is critical to the successful implementation of an analysis using NIR spectroscopy. Thus, the purpose of this paper is mainly focus on basic research for improving prediction accuracy by extracting weak glucose information from NIR spectra using chemometrics.

It is well known that combination of appropriate pretreatment and high-accuracy quantitative calibration is benefit to improve prediction accuracy in chemometrics [9]. Currently, data preprocess is often utilized to eliminate noises or baseline. However, the research about elimination of matrix background interference in the field of blood glucose noninvasive measurement using NIR spectroscopy is very few [10]. In practice, the matrix background interference such as water is pivotal in NIR region for blood glucose noninvasive measurement. There two reasons: (i) water is the main components in human blood or tissue, and normally the glucose content is 0.1% of water in human blood and tissue; (ii) the optical absorption of water is distinct in NIR region. On the other hand, quantitative calibration is another key point to extract glucose information by building a relationship of response variable (concentration) and predictor variables (wavelength). Linear model such as partial least squares (PLS) regression is often used in some researches. But it is known that nonlinearity is an inherent trait for

* Corresponding author.
E-mail addresses: lina_buaa@126.com (L.-N. Li), LLN604@sina.com (G.-J. Zhang).

1350-4495/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.infrared.2010.07.012

systems, and linear model is inappropriate to describe the underlying data structure with significant nonlinear characteristics. Especially, for the noninvasive measurement of human blood glucose with NIR spectroscopy, the linear relationship based on Lambert–Beer Law is not tenable because of many factors existed, such as the complexity of blood components, the interaction between components, the distribution irregularity of the blood components because of the macromolecule existed (protein, fats, etc.), the affection of colored noise, baseline drift and so on. However, few studies are focus on nonlinear modeling especially in the field of blood glucose noninvasive measurement using NIR spectroscopy [11].

The works of this paper are mainly on the two aspects: (i) optimization of modeling data by pretreatment, and it is aiming to eliminate high-frequency noise and matrix background from optical absorbance of water in NIR region; (ii) construction of nonlinear model, and it is main to improve prediction accuracy by building a nonlinear relationship of glucose concentration and wavelength variables. So a modified uninformative variable elimination (mUVE) method combined with kernel partial least squares (KPLS) regression, named as mUVE–KPLS, is proposed in this paper. mUVE is used to eliminate high-frequency noise and matrix background simultaneously by wavelet multi-resolution technology. KPLS is followed to build a nonlinear relationship by kernel function transformation. Two NIR spectra data from basic research experiments (simulated physiological solution experiment in vitro and human noninvasive measurement experiment in vivo) are introduced to evaluate the performance of the proposed strategy. Under the same KPLS model, the performances of different pretreatment methods are discussed in this paper. And under the mUVE method, statistical evaluation and adaptability of KPLS with Covariance kernel (a linear approach same to the ordinary PLS), KPLS with Gaussian kernel, KPLS with Polynomial kernel and Spline-PLS (SPLS) (a nonlinear method) are compared in this study. In this research, a feasible nonlinear model with optimal parameter is selected for human blood glucose noninvasive measurement by NIR spectroscopy.

2. Experimental

2.1. Simulated physiological solution samples experiment in vitro

The simulated physiological solution samples experiment data is a NIR spectra data of aqueous solution samples with four components (water, glucose, cattle hemoglobin, and albumin), which is a simple simulation of human blood by adding main blood components in water. There are 50 samples. The range of glucose concentration is uniform distributed from 100 to 5000 mg dL^{-1}, the range of cattle hemoglobin concentration is random distributed from 0 to 1000 mg dL^{-1}, and the range of albumin concentration is random distributed from 0 to 400 mg dL^{-1}.

The preparation process for simulated physiological solution samples is described here. Firstly, the mother liquids are made up for every solute. Then, proper quantities of mother liquid of a solute are extracted by pipet and put it in a 100 mL solution flask. Afterwards, adding de-ionized water in the solution flask until 100 mL. Thus, samples with different concentrations are prepared. According to the two weights (before dilution and after dilution), the concentrations of water are calculated. According to the concentration of mother liquid and sampling cubic content, the concentrations of all solutes are calculated.

The NIR spectra data is got by Spectrum GX FITR instrument (Perkin–Elmer, America) with InSb detector cooled by liquid nitrogen. One millimeter quartz cell, peristaltic pump automatic sampling system and thermostatic apparatus are used for spectra measurement. At first, the spectrum of empty cell is measured as a background. Then, the spectra of simulated physiological solution samples are measured. The spectra scan scope is 4000–10,000 cm^{-1} (1000–2500 nm), the resolution is 4 cm^{-1}. The total number of the wavelength variables is 1501. The 50 samples are scanned randomly, that means it is not scanned according to the order of glucose concentration value, which is mainly for avoiding chance correlation in calibration.

In this experiment, there are six samples are picked out as outliers, thus 44 samples are retained in this study. At mean while, considering that there are two saturation regions and noises are very obviously after wavelength region of 1900 nm, so the sensitive glucose absorption region of 1000–1900 nm is employed in this study. There are 1182 wavelength variables in each sample. There are 30 samples selected as calibration set, and the rest 14 samples are selected as independent external prediction set.

2.2. Human noninvasive measurement experiment in vivo

In order to get a calibration model covering wide variety scope of human blood glucose concentration, the Oral Glucose Tolerance Test (OGTT) is introduced in this experiment. Originally, OGTT is a kind of glucose burden adjustability test for diagnosis of diabetes in clinic. For healthy person, under this test, the glucose concentration value will be increased from normal value to peak value, and then decreased from peak value to normal value. So it is possible to get a varying scope of glucose concentration within several hours. It can be used as a special experiment method for getting a calibration samples with a certain concentration variety, which gives an interesting way for the basic research of constructing calibration model in blood glucose noninvasive measurement using NIR spectroscopy.

The NIR spectra of human noninvasive measurement samples are got by NIR Quest 256-2.5 Spectrometer (Ocean Optics, America) with Hamamatsu G9208-256 InGaAs linear array detector. Detector range is 900–2500 nm. Operating temperature is −15 °C. Integration time is 28 ms. Lamp-house (HL-2000-HP) and reflection fiber probe (R400-7-VIS-NIR of type Y) are used in this experiment. The One Touch® Ultra®2 Blood Glucose Meter (LifeScan, Inc., America) is used for getting reference values of human blood glucose concentration.

Experiment procedure is described here. A healthy volunteer had been fasted for 8 h before this experiment began. Then he drunk 100 mL water with 75 g glucose within 5 min, in succession the NIR diffuse reflection spectra were collected from the finger pulp by reflection fiber. 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 got by the One Touch® Ultra®2 Blood Glucose Meter. In this experiment, the blood glucose concentration value of the volunteer increased gradually until arriving at the peak value 176.4 mg dL^{-1}, and then decreased. The experiment was over when the concentration value was down to the normal level 90 mg dL^{-1}. This experiment was done within 3 h.

There are 31 samples got in this experiment. Each spectrum has 256 variables. The blood glucose concentration scope is 90.0–176.4 mg dL^{-1}. At each point of blood glucose concentration measurement, NIR spectra are also got. There are 21 samples selected as calibration set for modeling, and the rest 10 samples are selected as independent external prediction set for evaluating the model.

3. Theory and algorithm

3.1. Modified uninformative variable elimination

The mUVE method is initially proposed by Shao et al. [12]. It is extended from the method called "uninformative variable elimination (UVE)" [13,14]. Instead of selecting the variables for multivariate calibration as UVE do, mUVE is proposed to investigate the corresponding wavelet component's contribution to the model. It is often utilized as a criterion to distinguish and eliminate background and noise by wavelet multi-resolution technology. It is well known that background and noise are always found in high-scale approximation component and low-scale detail coefficients. Using wavelet transformation, approximation component and detail coefficients at every scale can be got according to linear superposition theory. Based on mUVE criterion, the scales expressing information of background and noises can be determined. After removing corresponding components in the scales determined, new spectra with minimum unrelated information can be reconstructed. More details of mUVE see Refs. [12,15].

3.2. mUVE–KPLS model

KPLS is a novel nonlinear model developed by Rosipal et al. [16,17]. The theory and algorithm is given in an Appendix A. Here only the strategy of mUVE–KPLS is described.

In order to construct a parsimonious model with high-accuracy for human blood glucose noninvasive measurement using NIR spectroscopy, a hybrid nonlinear modeling strategy mUVE–KPLS is proposed in this research. The idea of the proposed approach is that: at first, mUVE is used to eliminate high-frequency noise and matrix background simultaneously; then, KPLS is followed to construct a nonlinear relationship of response variable $\boldsymbol{y}$ and treated predictor variables $\boldsymbol{x}$ for extracting glucose concentration information sufficiently.

In theory, there are four advantages when applying this hybrid strategy to the human blood glucose noninvasive measurement based on NIR spectroscopy: (i) a more practical calibration model can be constructed because that, a nonlinear model is built after elimination of unrelated information by mUVE; (ii) under the kernel trick of KPLS, a more parsimonious model can be got, only a n-dimensional square matrix is employed in PLS; (iii) the prediction accuracy can be improved greatly because that, nonlinear model is introduced instead of linear model, and pretreatment method based on mUVE is benefit to provide high-quality spectra for calibration by eliminating matrix background and high-frequency noises; (iv) the nonlinear mapping effect can be completed only by dot product, so the hybrid strategy is easy to be performed.

3.3. Evaluation of model

In order to estimate the number of factors and determine parameters of mUVE–KPLS model, leave-one-out (LOO) cross-validation (CV) procedure is introduced in this paper. The performance of calibration model is evaluated according to root mean squares error (RMSE) of calibration set (RMSEC) and prediction set (RMSEP). In order to evaluate the quality of obtained results from different models, relative error (RE) of predicted concentration is introduced here. It is calculated as:

$$\mathrm{RE} = 100\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} y_i^2}} \tag{1}$$

where $\hat{y}_i$ are predicted values, y_i are reference values, n is the number of samples.

In addition, p-value is introduced in this paper for evaluating validity of calibration model. p-Value is a measure of probability, the tail (part) probability of the distribution of a test statistic. t-Test with significance level $\alpha = 0.05$ is utilized in this study. The null hypothesis is that, there is no obviously difference between NIR spectroscopy method based on mUVE–KPLS and criterion method based on chemistry for human blood glucose measurement. The smaller of p-value the more likely you should reject the null hypothesis. Because, smaller p-value represents a smaller chance that the null hypothesis is going to be true.

4. Results and discussion

4.1. Model for the simulated physiological solution samples

In order to improve signal–noise-ratio (SNR) and have a treated spectra representing glucose information more closely, mUVE is employed first. For the NIR spectra of simulated physiological solution samples, each spectra signal is decomposed by wavelet transform, wavelet function is db3, Mallat decomposition level is 10. According to mUVE criterion, the components of level 1, 2 and 7 are removed as high-frequency noises and low-frequency background respectively. Then every spectra signal is reconstructed. The original spectra, reconstructed spectra and eliminated information are given in Fig. 1a–c. It is clear that the signal at the absorption area around 1450 nm and 1900 nm are removed distinctly (see Fig. 1c), which around the main absorption peak of water in NIR region. It is encouraging that the main matrix background interference from optical absorption of water in NIR region is eliminated.

Then, KPLS is employed to construct a nonlinear model for the reconstructed spectra. Gaussian kernel transformation and mean-center are used here. The Fig. 2a depicts mean-centered Gaussian kernel Gram matrix of calibration data for NIR data of simulated physiological solution samples. Fig. 2b depicts ordinary mean-centered matrix usually employed in PLS. It is clear that the original matrix with 30 × 1182 dimension is decreased to 30 × 30 dimension by kernel trick. It is clear that KPLS model is more parsimonious and not complex like as PLS.

In order to select Gaussian kernel parameter σ, the experiential values are employed in this paper, and the smallest root mean squares error of cross-validation (RMSECV) is introduced as target function for parameter selection. Under different Gaussian kernel parameters, the RMSECV curve for the NIR spectra of simulated solution samples is given in Fig. 3. According to the RMSECV curve, Gaussian kernel parameter $\sigma = 3$ is selected as an optimal value for this case.

4.2. Model for the human noninvasive measurement samples

For the NIR spectra of human noninvasive measurement samples, mUVE is utilized first to eliminate unrelated information

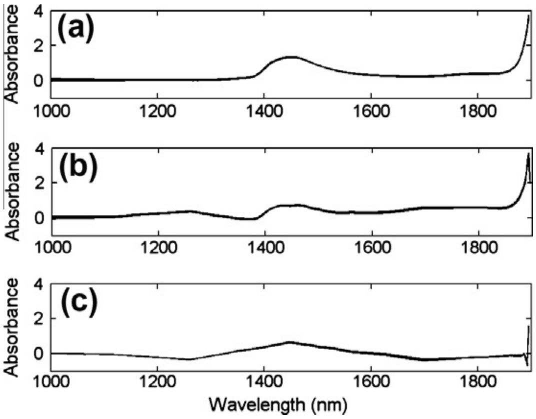

Fig. 1. Pretreatment results by mUVE for the NIR data of simulated physiological solution samples: (a) original spectra; (b) reconstructed spectra; (c) eliminated information.

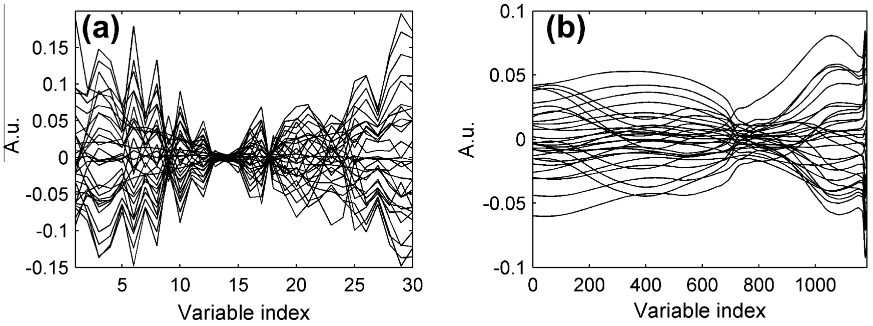

Fig. 2. Centralization results of calibration set for simulated physiological solution samples experiment: (a) the mean-centered kernel Gram matrix; (b) the ordinary mean-centered matrix.

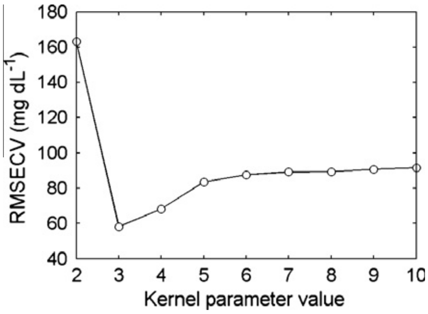

Fig. 3. The RMSECV by using different kernel parameters for calibration set of simulated physiological solution samples.

and improve spectra quality. Each spectra signal is also decomposed with the wavelet function db3, Mallat decomposition level is 8. According to mUVE criterion, the components of level 1 and 7 are removed as high-frequency noises and low-frequency background. Then spectra signal is reconstructed. The original spectra, reconstructed spectra and eliminated information are given in Fig. 4a–c. The situation is similar to simulated physiological solution samples experiment. It is clear that the information at the absorption area around 1450 nm and 1900 nm are also removed distinctly (see Fig. 4c). It is implied that the main matrix background interference (the optical absorption of water in NIR region) is almost eliminated by mUVE.

Then, the reconstructed spectra are employed to construct a nonlinear KPLS model. Gaussian kernel transformation and mean-center are also performed here. For comparison, the mean-centered Gaussian kernel Gram matrix is shown in Fig. 5a, the ordinary mean-centered matrix is shown in Fig. 5b. Similar to simulated physiological solution samples experiment, the original matrix with 21 × 256 dimension (always employed in PLS) is decreased to 21 × 21 dimension by kernel trick. Figs. 2 and 5 also indicate that, under the kernel transformation, solution using feature space is avoided; only dot product is used for completing the nonlinear transformation. It is implied that the original spectra model can be simplified by kernel trick; at meanwhile the nonlinear property of spectra data has been contained.

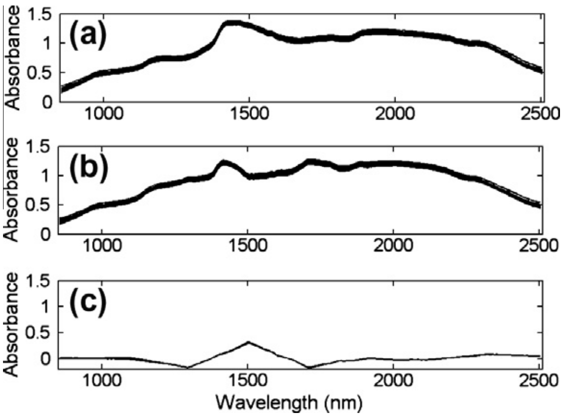

Fig. 4. Pretreatment results by mUVE for the NIR data of human noninvasive measurement samples: (a) original spectra; (b) reconstructed spectra; (c) eliminated information.

In order to select Gaussian kernel parameter σ for this experimental data, the experiential values are also employed here. Under different experiential parameters selected, the RMSECV curve for human noninvasive measurement samples is given in Fig. 6. It is clear that Gaussian kernel parameter $\sigma = 2$ is the best choice for this NIR experiment.

4.3. Performance comparison of pretreatment methods

In order to estimate effectiveness of pretreatment method for blood glucose noninvasive measurement using NIR spectroscopy, the usually used pretreatment methods, first-derivative spectrum and multiplicative scatter correction (MSC) are introduced in this research for comparing with mUVE.

First- and second-derivative spectra are commonly used for minimizing the problems due to overlap and baselines. For the NIR spectra of simulated physiological solution samples, first-derivative spectrum is formed after Savitzky–Golay algorithm with three points smooth. For the NIR spectra of human noninvasive measurement samples, first-derivative spectrum is formed after Savitzky–Golay algorithm with nine points smooth. Then KPLS is constructed by using the calibration samples. External independent prediction sets are introduced to evaluate prediction performance.

MSC is often used for minimizing light source variations. A number of effects can be successfully treated with MSC, such as path length problems, offset shifts, and interference. In this research, MSC is also used before calibration using KPLS for the two experimental NIR data.

Under different pretreatment methods described above, the prediction parameters are given in Table 1. The results indicate that the best prediction accuracy (RMSEP is 58.1 mg dL^{-1} and 9.4 mg dL^{-1} for the two NIR data respectively) can be got by using mUVE method. Although derivative spectra are commonly used to eliminate baseline and background, it is not useful to improve SNR. MSC is a signal treatment operation to spectrum only, concentration information is not considered in MSC algorithm. So some useful chemistry information is ignored. In addition, noises cannot be eliminated at all, which affect spectra quality and it is not always useful to calibration for blood glucose noninvasive measurement

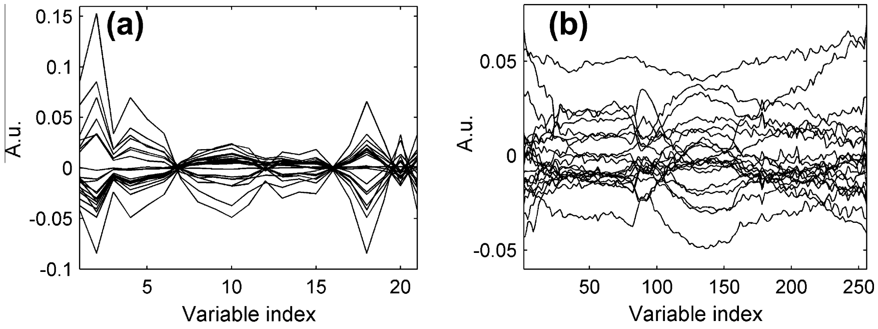

Fig. 5. Centralization results of calibration set for human noninvasive measurement samples experiment: (a) the mean-centered kernel Gram matrix; (b) the ordinary mean-centered matrix.

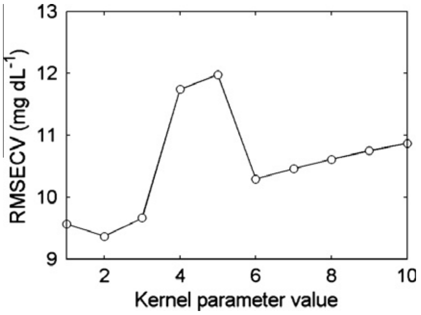

Fig. 6. The RMSECV by using different kernel parameters for calibration set of human noninvasive measurement samples.

using NIR spectroscopy. However, mUVE can be used to eliminate matrix background and noise simultaneously by wavelet multi-resolution technology, and the better spectra quality can be got. Especially, it is noticed that, under the mUVE method, one of the main matrix background interference, the optical absorption of water in NIR region, is almost eliminated. It is well known that water is the main matrix background interference in NIR region, which is not useful to calibration. And eliminating optical absorption of water is important to improve spectra quality and benefit to calibration for blood glucose noninvasive measurement using NIR spectroscopy.

4.4. Statistical evaluation and adaptability discussion

In order to estimate effectiveness of quantitative calibration models for glucose measurement using NIR spectroscopy, different strategies, mUVE–KPLS with Covariance kernel, mUVE–KPLS with Polynomial kernel and mUVE–SPLS are introduced to compare with mUVE–KPLS using Gaussian kernel. The calibration sets are employed to construct model; RMSEC is utilized to evaluate fitting performance. The independent external prediction sets are used to validate the constructed model; RMSEP and RE (%) are introduced to evaluate prediction performance and adaptability. In addition, *p*-value is employed to evaluate validity of calibration model. Prediction parameters based on different calibration models for the two NIR experimental data are given in Tables 2 and 3. Under different calibration strategies, the reference values and the predicted values of glucose concentration are also shown in Figs. 7 and 8 for the two experimental data respectively. The results indicate that:

(i) Under the pretreatment method of mUVE, according to the results of RMSEC given in Tables 2 and 3, the nonlinear model has better fitting performance than linear model (KPLS with Covariance kernel) for the two NIR experimental data.

(ii) For the independent external prediction set of simulated physiological solution samples, the best prediction accuracy is got by using SPLS and the RMSEP is 55.5 mg dL^{-1}, which is increased by 42.5% than linear model. Although the

Table 1
Prediction parameters of KPLS with different pretreatment methods for the two NIR data.

Pretreatment	Simulated physiological solution samples			Human noninvasive measurement samples		
	Factor	RMSEP (mg dL^{-1})	Correlation coefficient	Factor	RMSEP (mg dL^{-1})	Correlation coefficient
None	13	98.9	0.981	15	11.0	0.880
Derivative	15	81.9	0.992	12	13.8	0.805
MSC	11	71.0	0.990	12	15.1	0.784
mUVE	13	58.1	0.997	13	9.4	0.918

RMSEP, root mean squares error of prediction; MSC, multiplicative scatter correction; mUVE, modified uninformative variable elimination.

Table 2
Prediction parameters based on different calibration models for the NIR data of simulated physiological solution samples.

Calibration model	Factor	RMSEC (mg dL^{-1})	RMSEP (mg dL^{-1})	RE (%)	*p*-Value
mUVE–KPLS with Covariance kernel	3	67.4	96.6	2.3	0.42
mUVE–KPLS with Polynomial kernel	9	26.9	61.0	1.5	0.18
mUVE–KPLS with Gaussian kernel	13	24.6	58.1	1.4	0.21
mUVE–SPLS	12	13.4	55.5	1.3	0.50

RMSEC, root mean squares error of calibration; RE, relative error in the predicted concentration, in percentage; KPLS, kernel partial least squares; SPLS, Spline partial least squares; see Table 1 for the meaning of RMSEP, mUVE.

Table 3
Prediction parameters based on different calibration models for the NIR data of human noninvasive measurement samples.

Calibration method	Factor	RMSEC (mg dL^{-1})	RMSEP (mg dL^{-1})	RE (%)	*p*-Value
mUVE–KPLS with Covariance kernel	8	14.6	11.7	8.6	0.27
mUVE–KPLS with Polynomial kernel	14	0.3	10.6	7.8	0.50
mUVE–KPLS with Gaussian kernel	13	0.4	9.4	6.9	0.35
mUVE–SPLS	8	0.7	11.4	8.4	0.50

See Tables 1 and 2 for the meaning of RMSEC, RMSEP, RE, mUVE, KPLS, SPLS.

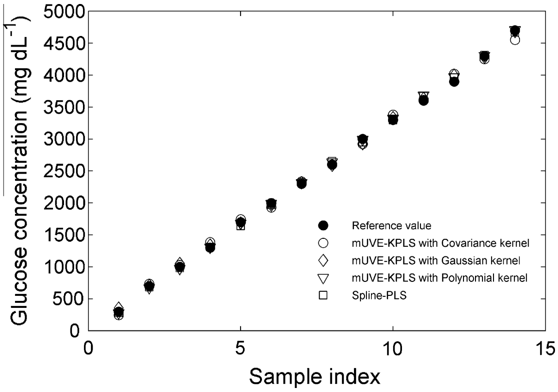

Fig. 7. Reference values and predicted values by using different modeling strategies for NIR data of simulated physiological solution samples.

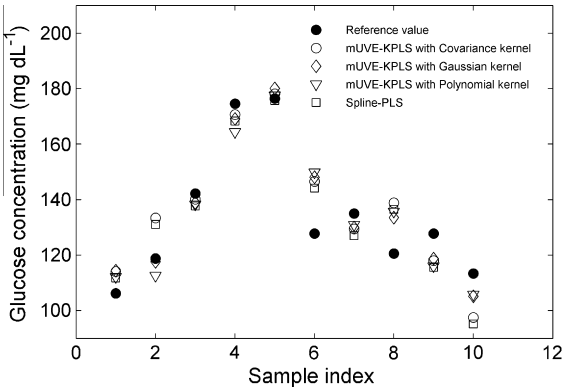

Fig. 8. Reference values and predicted values by using different modeling strategies for NIR data of human noninvasive measurement samples.

nonlinear KPLS model based on Gaussian kernel is not the best one among the calibration strategies considered, it still has good prediction performance, and the RMSEP is 58.1 mg dL^{-1}, which is increased by 39.9% than linear model. It is implied that, for the samples with simple components, the prediction performance of KPLS model based on Gaussian kernel is close to the SPLS.

(iii) According to the results shown in Tables 2 and 3, it is indicated that, even though SPLS has the best prediction performance to the NIR experiment data of simulated physiological solution samples in vitro, it seems that SPLS is not always has good performance, especially for the complex NIR data of human noninvasive measurement experiment in vivo. Under the hybrid nonlinear modeling strategy of mUVE–KPLS with Gaussian kernel, the best prediction accuracy is got and the RMSEP is 9.4 mg dL^{-1}, which is increased by 19.7% compared with linear model. However, under the modeling strategy of mUVE–SPLS, the RMSEP is 11.4 mg dL^{-1}, which is increased slightly by 2.6% compared with linear model. From this point of view, it is implied that mUVE–KPLS can has better robustness than mUVE–SPLS, especially when employed to the complex NIR data.

(iv) For the two experiment data, all the *p*-values are more than 0.05, which means that the four calibration models employed here are valid.

In addition, for the NIR data of human noninvasive measurement samples, the situation is more complex than simulated physiological solution samples, the complexity is mainly from two aspects: on the one hand, inherent physiological complexity of human body results in the complexity of the spectra, which including some uncertain information, such as temperature, pressure, psychological condition and so on; on the other hand, because of NIR spectra express the combination and overtone molecular vibrations associated with C–H and O–H bands of the glucose molecular, the information is overlapped. Even though, the best prediction accuracy is got by using mUVE–KPLS with Gaussian kernel in this research, and the RE (%) of independent external prediction set is 6.9 mg dL^{-1}. Although mUVE–SPLS method has good performance to the NIR data with certain element like as simulated solution samples experiment in vitro, it is not always done well to complex spectra data. The results also implied that, mUVE–KPLS based on Gaussian kernel has better adaptability to human blood glucose noninvasive measurement using NIR spectroscopy.

5. Conclusions

In this paper we commend a hybrid calibration strategy, mUVE–KPLS based on Gaussian kernel, for basic research of improving prediction accuracy in blood glucose noninvasive measurement using NIR spectroscopy. Two NIR data (simulated physiological solution samples experiment in vitro and human noninvasive measurement experiment in vivo) are introduced to evaluate prediction accuracy and adaptability of the proposed method. At meanwhile, by comparing with other quantitative linear and nonlinear calibration models, it is concluded that:

(i) Under the pretreatment method of mUVE, the noises and the great interference of matrix background from optical absorption of water in NIR region can be removed distinctly, which is useful to improve spectra quality and benefit to calibration.

(ii) KPLS is a simple-to-use model for the nonlinear mapping can be completed only by dot product; and a parsimonious calibration model with higher prediction can be constructed by kernel trick, that is a *n*-dimensional square matrix is employed in PLS, which is benefit to improve modeling efficiency and reduce the level of complexity of PLS.

(iii) For the NIR experimental samples with simple components, the prediction performance of KPLS model based on Gaussian kernel is close to SPLS.

(iv) For the more complex NIR experimental samples of human noninvasive measurement, KPLS with Gaussian kernel is more applicable by selection of appropriate kernel parameters.

This research indicates that the hybrid nonlinear modeling strategy of mUVE–KPLS can give us an alternative quantitative calibration approach for human blood glucose noninvasive measurement using NIR spectroscopy. It also can be utilized to other human blood components noninvasive measurement using NIR spectroscopy when appropriate kernel function and kernel parameters are determined.

Acknowledgements

This work is supported by the program for Changjiang Scholars and Innovative Research Team in University (IRT 0705), the National High Technology Research and Development Program of China, and the Innovation Foundation of BUAA for Ph.D. Graduate. The NIR data of simulated physiological solution samples is supported by the Biological Optics and Spectroscopy Application Laboratory of Tianjing University.

Appendix A. Kernel partial least squares

A.1. Kernel trick

Under the KPLS method, the limitation of PLS that it only can deal with linear system can be breached. KPLS differs from the previously mentioned nonlinear PLS algorithms [18–20], in that the original input data in space $\boldsymbol{R}$ are nonlinearly transformed into a feature space $\boldsymbol{F}$ of arbitrary dimensionality via nonlinear mapping $\boldsymbol{\Phi}(x)$, then a linear PLS model is created in feature space. The nonlinear transformation effect can be completed only by dot product as described in Eq. (A.1) [21,22].

$$k(x_i, x_j) = \boldsymbol{\Phi}(x_i)^T \boldsymbol{\Phi}(x_j) \tag{A.1}$$

where $k(x_i, x_j)$ denotes kernel function, which satisfies Mercer's theorem [16,17].

There are several kernel functions in common use. In this research we introduce the following kernel functions:

Covariance kernel: $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T \boldsymbol{x}_j$. It is defined as inner product of two spectra. KPLS based on Covariance kernel is equivalent to the ordinary PLS.

Polynomial kernel: $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^T \boldsymbol{x}_j + p)^q$, where q is the order of Polynomial, p is a tuning parameter and $p \geqslant 0$. It is clear that, when $p = 0$ and $q = 1$, the Polynomial kernel is also a Covariance kernel.

Gaussian kernel: $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-||\boldsymbol{x}_i - \boldsymbol{x}_j||^2/(2 \times \sigma^2))$, where σ is a turning parameter. Gaussian kernels are a special case of Radial Basis Function [23,24].

For spectra data, the form of kernel matrix (or Gram matrix) using kernel function can be described as follows:

$$\boldsymbol{K}_{\text{train}} = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \ldots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \ldots & k(x_2, x_n) \\ \ldots & \ldots & \ldots & \ldots \\ k(x_n, x_1) & k(x_n, x_2) & \ldots & k(x_n, x_n) \end{bmatrix} \tag{A.2}$$

$$\boldsymbol{K}_{\text{test}} = \begin{bmatrix} k(xt_1, x_1) & k(xt_1, x_2) & \ldots & k(xt_1, x_n) \\ k(xt_2, x_1) & k(xt_2, x_2) & \ldots & k(xt_2, x_n) \\ \ldots & \ldots & \ldots & \ldots \\ k(xt_l, x_1) & k(xt_l, x_2) & \ldots & k(xt_l, x_n) \end{bmatrix} \tag{A.3}$$

where $x_{tr} \in \boldsymbol{X}_{n \times m}$ $(0 < tr < n$, n: the number of training samples; m: the number of wavelength variables) denotes training set, $xt_{te} \in \boldsymbol{X}_{l \times m}$ $(0 < te < l$, l: the number of testing samples) denotes testing set. The kernel matrix of training set is a n-dimensional square matrix, in which each element is obtained by computing kernel function between original training samples; the kernel matrix of testing set is a $(l \times n)$-dimensional matrix, in which each element is obtained by computing kernel function between testing samples and training samples. It is clear that, when there are a number of wavelength variables, after kernel trick, only n-dimension data is employed to PLS, which decreases the complexity of PLS greatly.

A.2. KPLS algorithm

KPLS algorithm is based on kernel trick and PLS. Moreover, KPLS model is a more parsimonious model than PLS, and it can tackle nonlinear problems by selecting appropriate kernel function and parameters. The algorithm of KPLS can be summarized as follows:

Step 1. Initially, kernel function and kernel parameters are determined. Then, for $i = 1$ to A (A is the number of principal components), or until the stop criterion is met, the following steps will be repeated;
Step 2. Randomly initialize $\boldsymbol{u}$, $\boldsymbol{u} = \boldsymbol{y}$, where $\boldsymbol{y}$ is response variable, that is glucose concentration value in this study;
Step 3. $\boldsymbol{t} = \boldsymbol{K}_{\text{train}}^T \boldsymbol{u}$, $\boldsymbol{t} = \boldsymbol{t}/||\boldsymbol{t}||$;
Step 4. $\boldsymbol{c} = \boldsymbol{y}|^T \boldsymbol{t}$;
Step 5. $\boldsymbol{u} = \boldsymbol{y}\boldsymbol{c}$, $\boldsymbol{u} = \boldsymbol{u}/||\boldsymbol{u}||$;
Step 6. Repeat Steps 3–6 until the convergence;
Step 7. $\boldsymbol{K}_{\text{train}} \leftarrow (\boldsymbol{I} - \boldsymbol{t}\boldsymbol{t}^T)\boldsymbol{K}_{\text{train}}(\boldsymbol{I} - \boldsymbol{t}\boldsymbol{t}^T)$;
Step 8. $\boldsymbol{y} \leftarrow (\boldsymbol{I} - \boldsymbol{t}\boldsymbol{t}^T)\boldsymbol{y}$;
Step 9. End.

Note that $\boldsymbol{K}$ is kernel Gram matrix, which is computed and centered by using Eqs. ((A.2)–(A.5)). That is, the mean centering in the high dimensional space can be performed by substituting kernel matrices $\boldsymbol{K}_{\text{train}}$ and $\boldsymbol{K}_{\text{test}}$ with $\tilde{\boldsymbol{K}}_{\text{train}}$ and $\tilde{\boldsymbol{K}}_{\text{test}}$[16,17]:

$$\widetilde{\boldsymbol{K}}_{\text{train}} = \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}_n \boldsymbol{1}_n^T\right)\boldsymbol{K}_{\text{train}}\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}_n \boldsymbol{1}_n^T\right) \tag{A.4}$$

$$\widetilde{\boldsymbol{K}}_{\text{test}} = \left(\boldsymbol{K}_{\text{test}} - \frac{1}{n}\boldsymbol{1}_l \boldsymbol{1}_n^T \boldsymbol{K}_{\text{train}}\right)\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}_n \boldsymbol{1}_n^T\right) \tag{A.5}$$

where $\boldsymbol{I}$ is a n-dimensional identity matrix, $\boldsymbol{1}_n$, $\boldsymbol{1}_l$ represent vectors whose elements are ones, and its length is n and l respectively.

The calibration and validation data are evaluated by using Eqs. (A.6 and A.7):

$$\tilde{\boldsymbol{y}} = \boldsymbol{K}_{\text{train}}\boldsymbol{u}\left(\boldsymbol{T}^T \boldsymbol{K}_{\text{train}}\boldsymbol{u}\right)^{-1}\boldsymbol{T}^T\boldsymbol{y} \tag{A.6}$$

$$\tilde{\boldsymbol{y}}_t = \boldsymbol{K}_{\text{test}}\boldsymbol{u}\left(\boldsymbol{T}^T \boldsymbol{K}_{\text{train}}\boldsymbol{u}\right)^{-1}\boldsymbol{T}^T\boldsymbol{y} \tag{A.7}$$

where $\boldsymbol{T}$ is $(n \times A)$-dimension matrix, which is formed by the columns of latent vector $\boldsymbol{t}$.

References

[1] Carlos Eduardo Ferrantedo Amaral, Benhard Wolf, Current development in non-invasive glucose monitoring, Med. Eng. Phys. 30 (5) (2008) 541–549.

[2] Vanesa Sanz, Susana de Marcos, Javier Galbán, A blood-assisted optical biosensor for automatic glucose determination, Talanta 78 (3) (2009) 846–851.

[3] Qing-Bo Li, Guang-Jun Zhang, Ke-Xin Xu, Yan Wang, Application of digital Fourier filtering pretreatment method to improving robustness of multivariate calibration model in near infrared spectroscopy, Spectrosc. Spectr. Anal. 27 (8) (2007) 1484–1488.

[4] Mark A. Arnold, Gary W. Small, Noninvasive glucose sensing, Anal. Chem. 77 (17) (2005) 5429–5439.

[5] Kirsten E. Kramer, Gary W. Small, Robust absorbance computations in the analysis of glucose by near-infrared spectroscopy, Vib. Spectrosc. 43 (2) (2007) 440–446.

[6] Mukire J. Wabomba, Gary W. Small, Mark A. Arnold, Evaluation of selectivity and robustness of near-infrared glucose measurements based on short-scan Fourier transform infrared interferograms, Anal. Chim. Acta 490 (1–2) (2003) 325–340.

[7] Li-Na Li, Qing-Bo Li, Guang-Jun Zhang, A weak signal extraction method for human blood glucose noninvasive measurement using near infrared spectroscopy, J. Infrared Millim. Terahertz Waves 30 (11) (2009) 1191–1204.

[8] Ke-Xin Xu, Feng Gao, Hui-Juan Zhao, Biomedical Photonics, Science Publishing Company, Beijing, China, 2007.

[9] G. Carlomagno, L. Capozzo, G. Attolico, A. Distante, Non-destructive grading of peaches by near-infrared spectrometry, Infrared Phys. Technol. 46 (2004) 23–29.

[10] Li-Na Li, Guang-Jun Zhang, Qing-Bo Li, Pretreatment method research of near infrared spectra in blood component non-invasive measurement, Mod. Phys. Lett. B 23 (7) (2009) 925–937.

[11] Qing Ding, Gary W. Small, Mark 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.

[12] Da Chen, Xueguang Shao, Bin Hu, Qingde Su, A background and noise elimination method for quantitative calibration of near infrared spectra, Anal. Chim. Acta 511 (1) (2004) 37–45.

[13] Vitézslav Centner, Désiré-Luc Massart, Onno E. de Noord, Sijmen de Jong, Bernard M. Vandeginste, Cécile Sterna, Elimination of uninformative variables for multivariate calibration, Anal. Chem. 68 (21) (1996) 3851–3858.

[14] Hu-Wei Tan, Steven D. Brown, Wavelet analysis applied to removing non-constant, varying spectroscopic background in multivariate calibration, J. Chemomet. 16 (5) (2002) 228–240.

[15] Guang-Jun Zhang, Li-Na Li, Qing-Bo Li, Yu-Po Xu, Application of denoising and background elimination based on wavelet transform to blood glucose noninvasive measurement of near infrared spectroscopy, J. Infrared Millim. Waves 28 (2) (2009) 107–110.

[16] R. Rosipal, L.J. Trejo, Kernel partial least squares regression in reproducing kernel Hilbert space, J. Mach. Learn. Res. 2 (2001) 97–123.

[17] R. Rosipal, Kernel partial least squares for nonlinear regression and discrimination, Neural Network World 13 (3) (2003) 291–300.

[18] Lubomir Hadjiiski, Paul Geladi, Philip Hopke, A comparison of modeling nonlinear systems with artificial neural networks and partial least squares, Chemometr. Intell. Lab. Syst. 49 (1) (1999) 91–103.

[19] D. Pérez-Marín, A. Garrido-Varo, J.E. Guerrero, Non-linear regression methods in NIRS quantitative analysis, Talanta 72 (1) (2007) 28–42.

[20] Bahram Hemmateenejad, Mohammad A. Safarpour, A. Mohammad Mehranpour, Net analyte signal-artificial neural network (NAS–ANN) model for efficient nonlinear multivariate calibration, Anal. Chim. Acta 535 (1–2) (2005) 275–285.

[21] Kyungpil Kim, Jong-Min Lee, In-Beum Lee, A novel multivariate regression approach based on kernel partial least squares with orthogonal signal correction, Chemometr. Intell. Lab. Syst. 79 (1–2) (2005) 22–30.

[22] Bart M. Nicolaï, Karen I. Theron, Jeroen Lammertyn, Kernel PLS regression on wavelet transformed NIR spectra for prediction of sugar content of apple, Chemometr. Intell. Lab. Syst. 85 (2) (2007) 243–252.

[23] B. Walczak, D.L. Massart, The radial basis functions – Partial least squares approach as a flexible non-linear regression technique, Anal. Chim. Acta 331 (3) (1996) 177–185.

[24] B. Walczak, D.L. Massart, Application of radial basis functions – partial least squares to non-linear pattern recognition problems: diagnosis of process faults, Anal. Chim. Acta 331 (3) (1996) 187–193.