Chemometrics and Intelligent Laboratory Systems xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Short communication

A two-level partial least squares system for non-invasive blood glucose concentration prediction

Zheng-Ming Chuah [a], Raveendran Paramesran [a,*], Kavintheran Thambiratnam [a], Sin-Chew Poh [b]

[a] Dept. of Electrical Engineering, University of Malaya, Malaysia
[b] Dept. of Mechanical Engineering, University of Malaya, Malaysia

ARTICLE INFO

Article history:
Received 1 October 2009
Received in revised form 15 July 2010
Accepted 31 August 2010
Available online xxxx

Keywords:
Partial least squares
Non-invasive blood glucose prediction
Two-level partial least squares
OGTT
NIR

ABSTRACT

In this study, we propose and demonstrate a novel two-Level Partial Least Squares (2L-PLS) architecture for non-invasive blood glucose concentration measurement. A total of 290 Near-Infrared (NIR) spectroscopy readings from six laser diodes with discrete wavelengths of between 1500 nm and 1800 nm are obtained together with blood glucose concentration readings collected via Oral Glucose Tolerance Test (OGTT) experiments from a healthy volunteer over 4 days. While the conventional approach to predicting the blood glucose concentrations is to use a single Partial Least Squares (PLS) or non-linear PLS model, these systems do not achieve a high level of accuracy. As such, a 2L-PLS system consisting of one PLS model at the first level and three at the second level is proposed to enhance the prediction accuracy. A non-linear 2L-PLS system based on the same structure is also investigated in this study. The proposed 2L-PLS systems show improvements of 10 to 12% in the number of predictions that fall below a 5% error margin as compared to single-level PLS systems.

1. Introduction

The regulation of the body's blood glucose concentration is essential in maintaining a person's health. This is because glucose is the primary energy source for the body, and unregulated blood glucose concentrations can lead to numerous medical complications such as hyperglycemia and hypoglycemia. Current blood glucose monitoring systems are able to provide highly accurate readings, but suffer from a significant drawback in that they require a blood sample. The constant drawing of blood can induce substantial pain to the patient and increase the risk of infection due to the penetration of the skin, as well as incurring significant cost due to the number of test strips required. As such, a non-invasive blood glucose monitoring system that can overcome these limitations is highly desired and has now become the primary focus of research.

Most non-invasive systems are based on optical methods such as polarimetry [1], Raman, [2], Mid-Infrared (MIR) and Near-Infrared (NIR) spectroscopy [3–7], with MIR and NIR spectroscopy demonstrating the highest potential for blood glucose monitoring. Initial limitations such as low SNR levels and high frequency noise [8–12] have been addressed to increase the accuracy of the system, and more recently researchers have looked towards new data analysis methods to further enhance the accuracy of the systems. The Partial Least Squares (PLS) regression method is among the most successful in modelling the relationship between the predictors, **X**, and the responses, **Y**, and is able to provide good prediction results for co-linear data. However, due to the linear framework of the PLS regression method, it does not perform well when provided with non-linear data. To overcome this shortcoming, and to also obtain an inner non-linear model between the principal components of dependent variables and independent variables, S. Wold developed the non-linear PLS model [13,14], which makes full use of the PLS regression method's ability to project to latent projections for non-linear relations and thus allowing for more accurate predictions.

In this paper, we propose a novel method for analyzing NIR spectral data for blood glucose concentrations, using a linear and non-linear two-level partial least squares (2L-PLS) model. A total of 290 NIR spectroscopy readings are obtained via non-invasive blood glucose assays from six laser diodes with discrete wavelengths, together with actual blood glucose concentration measurements from a volunteer over 4 days. The proposed 2L-PLS system consists of 4 PLS models divided into two levels. The first level consists of one PLS model, whilst the second level consists of the three remaining PLS models. Both linear and non-linear 2L-PLS are developed, where the linear model uses the input data as predictors, while the non-linear model expands the predictors with their powers and cross-products. The results obtained from both 2L-PLS models are compared to determine their accuracy and performance.

2. Experimental data

The experimental data for this research is obtained via NIR based non-invasive blood glucose assays as shown in Fig. 1. The NIR light

* Corresponding author.
E-mail addresses: zm_8429@yahoo.com (Z.-M. Chuah), ravee@um.edu.my (R. Paramesran), kavintheran@gmail.com (K. Thambiratnam), pohsc@um.edu.my (S.-C. Poh).

0169-7439/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2010.08.015

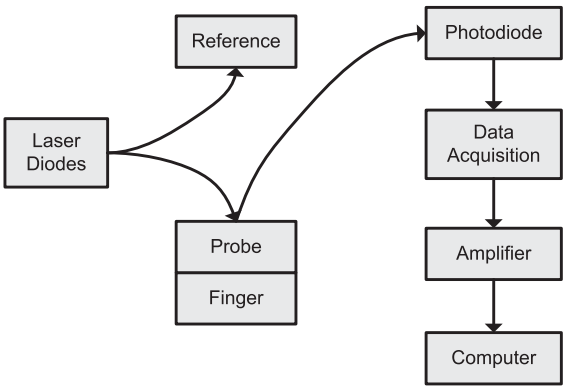

Fig. 1. Schematic diagram of the non-invasive blood glucose monitoring system.

source consists of a series of six laser diodes with discrete wavelengths of between 1550 nm and 1800 nm operating at output powers of 6 mW. Two assumptions are made; first, the chosen wavelengths exhibit minimal water attenuation (this is an important consideration as more than 95% of human blood constitutes water, which has very strong absorption in the NIR region) and second, the reflected signals from these selected wavelengths contain sufficient information on the blood glucose concentration [15]. The six laser diodes are triggered sequentially using a controlled stepping motor scanning system [4]. A portion of the NIR signal is collected as a reference while the rest is transmitted to the measurement site via an optical fiber located at the center of the measurement probe. The measurement site is the finger nail bed of the subject's index finger and is chosen as its properties are relatively constant between subjects (as opposed to the skin as it would be affected by factors such as sweat, pigmentation, etc.). In order to minimize errors, the positions of the test bed and fiber optic probe throughout this work are fixed. The temperature of the laser diodes are also measured and used together with the reference light measurements to compensate for variations in the NIR spectra readings arising from temperature fluctuations. The NIR light penetrates the skin to the blood vessel where it is absorbed by the glucose in the blood and the diffused light from the nail bed is collected by optical fibers in a circular distribution surrounding the central transmission fibers and guided to the photodiode, where the optical signal is converted into an electrical signal. The signal is then sent to a computer via a Universal Serial Bus (USB) connection for further analysis.

The *In vivo* blood glucose assay was carried out with the participation of a healthy volunteer subject fasting for at least 8 h before submitting to the test. The subject consumed a 35 g glucose solution approximately 10 min after the start of the experiment, and upon the commencement of the experiment the NIR spectra were collected at intervals of every minute from the measurement site. At the same time, a blood sample was taken from the index finger of the subject's other hand at intervals of approximately 10 min and measured using a traditional blood glucose meter. The experiment was repeated for 4 days with the same subject to obtain a total of 290 NIR data points and the measured blood glucose concentration levels were extrapolated before the calibration process. The collected blood glucose concentration range was between 4 *mmol/L* and 10 *mmol/L*. For the calibration, 60% of the collected data was used while the remaining 40% was used for validation. The NIR spectral readings used in the calibration were randomly selected.

3. Two-level partial least squares (2L-PLS) system

The PLS method is a popular multivariate data analytical method designed to handle intercorrelated regressors [16–18]. Originally proposed by Herman Wold as an econometric technique, it became popular in the field of chemometrics [19] and has been successfully used to regress NIR data [19,20]. However, the PLS model still produces high-error rates when used to regress non-linear data. In this aspect, a 2L-PLS system is introduced to overcome the high-error rates of the traditional single-level PLS model. Unlike the single PLS system, a 2L-PLS system consists of two levels, with a single PLS model at the first level and three PLS models at the second level. The main concept of the 2L-PLS system is to subdivide the overall blood glucose concentration into several smaller ranges, hence calibrating each PLS model in the second level with a smaller blood glucose concentration range as compared to the single PLS system. The PLS calibration model in the first level however is required to calibrate the full range of blood glucose concentrations from 4 mmol/L to 10 mmol/L. The overall structure of the 2L-PLS system in this study is shown in Fig. 2.

The NIR spectral data form the input to the PLS model in the first level and its output selects either one (or two) of the PLS models in the second level. Subsequently, the corresponding NIR spectra from the first level are used as the input of the chosen PLS model(s) in the second level of the system for the actual prediction. Both the proposed linear and non-linear 2L-PLS models operate in the same manner, with the exception that the input of the non-linear 2L-PLS model is expanded according to their powers and also by a sinusodial function, specifically to the powers of $-2, -1, 2, 3$ and also by the exponentials of $\exp(x)$ and $\exp(-x)$. This strategy makes the non-linear models simple to use but at the same time robust [21].

3.1. Performance evaluation

The performance of the PLS system in predicting the blood glucose concentration level is based on the calibration and validation data sets and can be evaluated by the Root Mean Square Error of Calibration

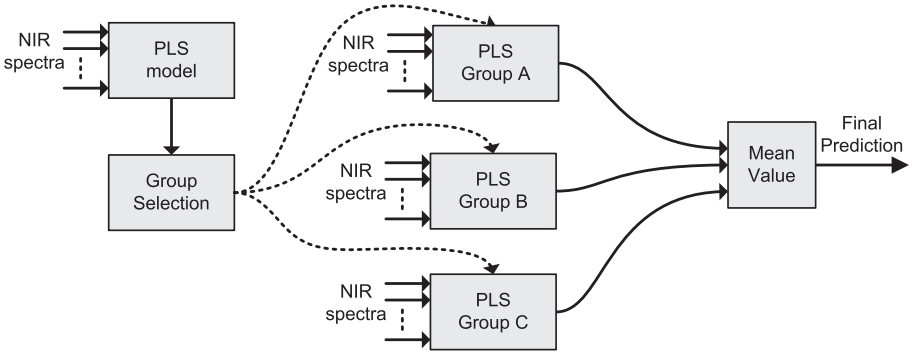

Fig. 2. Overall structure of the two-level partial least squares system.

Please cite this article as: Z.-M. Chuah, et al., A two-level partial least squares system for non-invasive blood glucose concentration prediction, Chemometr. Intell. Lab. Syst. (2010), doi:10.1016/j.chemolab.2010.08.015

(RMSEC) and the Root Mean Square Error of Prediction (RMSEP) respectively as per the following equations [6]:

$$\mathrm{RMSEC} = \left\{ \sum_{m=1}^{M} (y_m - \hat{y}_m)^2 / (M-A-1) \right\}^{1/2} \quad (1)$$

and

$$\mathrm{RMSEP} = \left\{ \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 / N \right\}^{1/2} \quad (2)$$

where y_m and y_n are the measured blood glucose concentrations, $\hat{y}_m$ and $\hat{y}_n$ are the predicted blood glucose concentrations, M and N are the number of observations for the calibration and validation data sets respectively and A is the number of PLS score factors used in the PLS calibration system.

As the score factors for the PLS models in the second level differ from each other, Eq. (1) must be slightly modified so that A will be the mean value of the score factors of the selected PLS models. The prediction performance of the validation data set however can be evaluated using Eq. (2) as it is. In this study, a score factor of 6 is used in all four PLS models of the linear 2L-PLS system. The non-linear system uses a score factor of 29 for the PLS model in first level and score factors of 13, 10 and 6 for the group A, group B and group C PLS models respectively, based on the Wold's R criterion [19].

3.2. Considerations for the 2L-PLS system

There are several considerations made in the design of the 2L-PLS system. Firstly, it is necessary for each PLS model at the second level to receive an adequate number of inputs to accurately model the relation between the predictors and responses matrices. Secondly, errors produced in the first level of the system may result in the selection of the wrong second level PLS model during the calibration phase. For example, if the measured blood glucose concentration is 7.2 *mmol/L* whilst the output from the first level is 6.8 *mmol/L* and the boundary value between the two groups is 7 *mmol/L*, then the corresponding NIR spectra will be categorized into an undesired group. In order to reduce the error-rate arising from erroneous groupings, the 2L-PLS system is designed such that the input ranges for the second level PLS models overlap with each other as shown in Fig. 3, where the values of a and b are the boundary values among these 3 groups before being overlapped. The percentage of overlapped data is given as h%, and subsequently the overlapping boundary values of a_1 and b_1 can be determined from the formulae of $a_1 = a - a^*h/100$ and $b_1 = b - b^*h/100$, respectively, with the size of the overlapping regions determined by the value of h. The grey areas in Fig. 3 indicate the overlapping regions.

In this study, the values of a and b for the linear 2L-PLS system are 6.6 *mmol/L* and 8.63 *mmol/L*, respectively and 7.3 *mmol/L* and 8.63 *mmol/L*, respectively for the non-linear 2L-PLS system. These values are chosen based upon the aforementioned consideration.

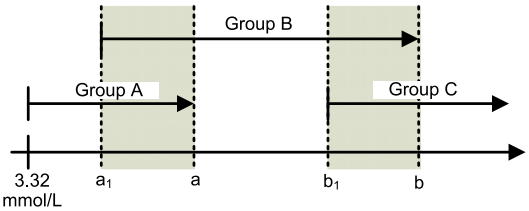

Fig. 3. The blood glucose concentration ranges for the PLS groups in the second level of the 2L-PLS system.

4. Results and discussions

The results obtained from the linear and non-linear 2L-PLS systems are reported together with the corresponding results from the single-level linear and non-linear PLS system in this section. The performance of the proposed systems is given along with its clinical accuracy using the Error Grid Analysis (EGA) method. Subsequently, the overall prediction performance for the calibration and validation data sets are presented and followed by the prediction performance of individual observation.

4.1. Performance of proposed system with different overlapping regions

Figs. 4 and 5 show the performance of the linear 2L-PLS system and non-linear 2L-PLS system, respectively, with overlapping percentages from 1% to 50%, where the performance is evaluated in terms of the RMSEC and RMSEP. The dotted lines in Figs. 4 and 5 indicate the overlapping percentage that chosen for the linear 2L-PLS system (27%) and the non-linear 2L-PLS system (22%). Hence, a_1 and b_1 for linear 2L-PLS system corresponds to 4.818 *mmol/L* and 6.2999 *mmol/L* respectively, whilst a_1 and b_1 for non-linear 2L-PLS system corresponds to 5.694 *mmol/L* and 6.7314 *mmol/L* respectively.

4.2. Error grid analysis

The Clarke EGA has been used to evaluate the clinical accuracy of blood glucose measurements made by a meter as compared to a standard reference value [22]. In the Clarke EGA, the grid is subdivided into 5 regions, designated A, B, C, D and E. Values that fall into Regions A and B are clinically acceptable, whereas values that fall into Regions C, D, and E are potentially dangerous and constitute clinically significant errors.

The EGA results for the PLS and the 2L-PLS systems for the calibration data set and testing data show that for the prediction data set, 92 observations (79.31%) fall into region A for the linear PLS, increasing to 98 observations (84.48%) from the linear 2L-PLS system. Similarly, 97 observations (83.62%) fall into region A for the non-linear PLS system and increases to 103 observations (88.79%) for the non-linear 2L-PLS system. Figs. 6 and 7 show the EGA plot of the validation data set for the proposed linear and non-linear systems, respectively. The '+' plots in Figs. 6 and 7 represent the EGA plot of the PLS system, whilst the '*' plots represent the EGA plot of the 2L-PLS system. It is observed that more predictions from the 2L-PLS system

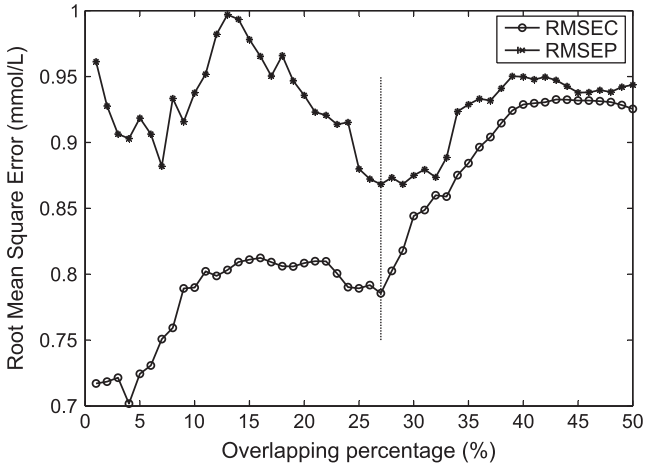

Fig. 4. The performance of the linear 2L-PLS system for overlapping percentages of 1% to 50%.

Please cite this article as: Z.-M. Chuah, et al., A two-level partial least squares system for non-invasive blood glucose concentration prediction, Chemometr. Intell. Lab. Syst. (2010), doi:10.1016/j.chemolab.2010.08.015

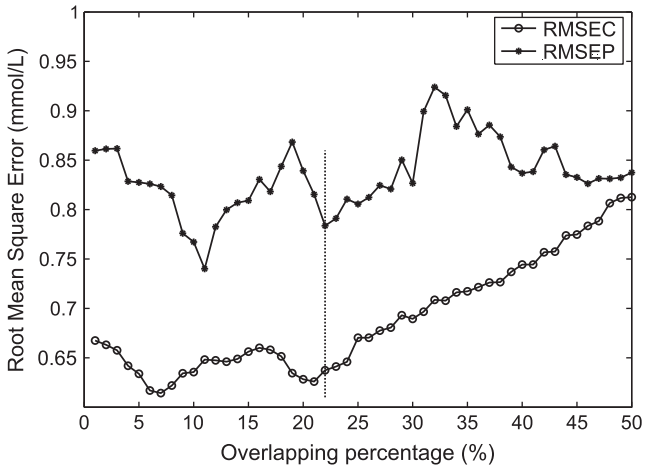

Fig. 5. The performance of the non-linear 2L-PLS system for overlapping percentages of 1% to 50%.

lie closer to the solid diagonal line of the EGA graph as compared to the single-level PLS system, thus demonstrating the higher accuracy of the 2L-PLS system. Additionally, all predictions from both the single-level PLS and the 2L-PLS systems fall within region A and region B of Figs. 6 and 7, indicating that all predicted results are within the clinically acceptable range.

4.3. Overall performance

Table 1 shows the overall prediction results for the linear and non-linear PLS and linear and non-linear 2L-PLS systems. The overall performance of these systems is evaluated in terms of the RMSEC and RMSEP for the calibration data set and validation data set respectively. An RMSEC of 0.7856 *mmol/L* (14.14 *mg/dL*) is obtained in the 2L-PLS system and is an improvement of 20.11% over the single-level PLS system as shown in Table 1 for the calibration data set. Similarly, the RMSEP of 0.8682 *mmol/L* (15.63 *mg/dL*) is obtained in the 2L-PLS system, which shows an improvement of 7.88% over the single-level PLS system for the validation data set. A similar observation is seen in the case of the non-linear PLS and 2L-PLS systems.

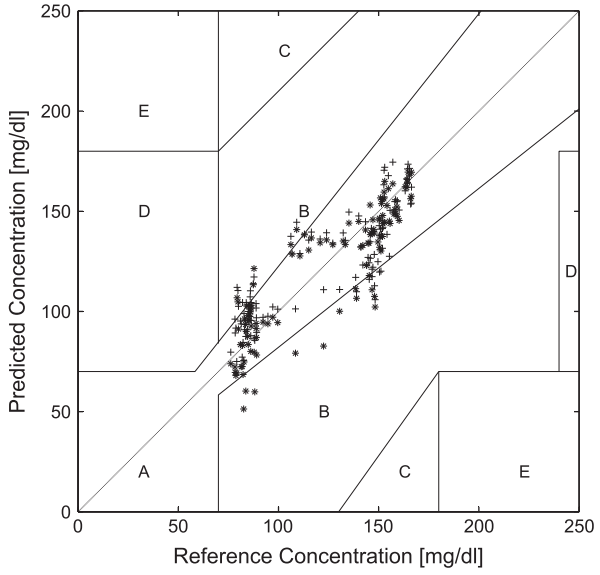

Fig. 6. The EGA plot of the testing data set for PLS system and 2L-PLS system.

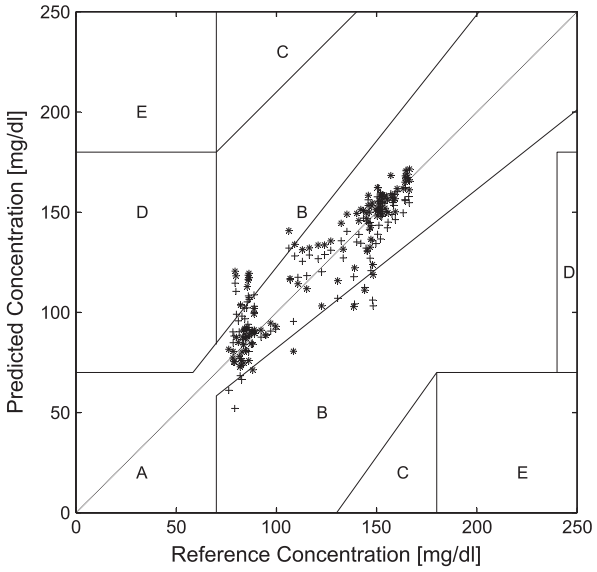

Fig. 7. The EGA plot of the testing data set for non-linear PLS system and non-linear 2L-PLS system.

The correlation coefficient between the predicted BGL and measured BGL for the calibration and validation data sets are represented in terms of *rC* and *rP* respectively. For the single-level PLS system, a correlation coefficient of 0.8476 is obtained, whilst for the 2L-PLS system, a better correlation coefficient of 0.8785 is obtained for the validation data set. A similar improvement was also observed for the case of the non-linear PLS models, where a correlation coefficient of 0.8981 is obtained from the non-linear single-level PLS system and improves to 0.9152 for the non-linear 2L-PLS system.

4.4. Performance of individual observations

The prediction results for each observation for calibration data set and validation data set are shown in Tables 2 and 3 respectively. The error percentage values constitute the distribution of samples in the calibration and validation data sets in accordance to the individual sample percentage error. Table 3 shows that for the validation data set, which determines the actual performance of the proposed system, 18 predictions from the linear 2L-PLS system have an error margin of less than 2.5%, whilst only 11 predictions from the linear single-level PLS system have a similir error margin. Table 3 also shows that 62 predictions from the non-linear 2L-PLS system have an error margin of less than 5% as compared to only 48 predictions from the non-linear single-level PLS system that have a similar error margin. This shows both the linear and non-linear 2L-PLS systems perform significantly better than their single-level counterparts, and show improvements

Table 1
Overall prediction results for the PLS, 2L-PLS, non-linear PLS and non-linear 2L-PLS systems.

		PLS	2L-PLS	Non-linear PLS	Non-linear 2L-PLS
Calibration data set	RMSEC (mmol/L)	0.9834	0.7856	0.7597	0.6477
	rC	0.8385	0.9003	0.9204	0.9353
Validation data set	RMSEP (mmol/L)	0.9425	0.8682	0.7778	0.7370
	rP	0.8476	0.8785	0.8981	0.9152

Please cite this article as: Z.-M. Chuah, et al., A two-level partial least squares system for non-invasive blood glucose concentration prediction, Chemometr. Intell. Lab. Syst. (2010), doi:10.1016/j.chemolab.2010.08.015

Table 2
Individual error percentage results for calibration data set.

	Number of observations			
	PLS	2L-PLS	Non-linear PLS	Non-linear 2L-PLS
>10%	101	73	53	46
5%–10%	36	41	48	46
2.5%–5%	16	22	41	39
1%–2.5%	14	22	20	22
0%–1%	7	16	21	21

Table 3
Individual error percentage results for validation data set.

	Number of observations			
	PLS	2L-PLS	Non-linear PLS	Non-linear 2L-PLS
>10%	62	46	40	30
5%–10%	28	32	28	24
2.5%–5%	15	20	27	31
1%–2.5%	8	11	12	16
0%–1%	3	7	9	15

of 10% and 12% respectively for errors that fall below the 5% error margin for the validation data set as compared to the linear and non-linear single-level PLS systems.

5. Conclusion

In this paper, a comparative analysis on the performance of using a single-level linear and non-linear PLS systems and a linear and non-linear 2L-PLS systems to predict the blood glucose concentration was studied. The PLS models use only six laser diodes operating at discrete wavelengths between 1500 nm and 1800 nm from an NIR light source as predictor variables. A total of 290 pairs of NIR spectra and measured blood glucose concentration readings are collected from a healthy volunteer subject over 4 days. The proposed 2L-PLS system consists of two levels, with a single PLS model in the first level and three PLS models in the second level. Both the linear and non-linear 2L-PLS systems show an improvement of 10% and 12% compared to their respective single-level PLS systems for errors that fall below the 5% error margin for the validation data set.

References

[1] B.D. Cameron, H.W. Gorde, B. Satheesan, G.L. Coté, The use of polarized laser light through the eye for noninvasive glucose monitoring, Diabetes Technol. Ther. 1 (1999) 135–143.
[2] M.S. Borchert, M.C. Storrie-Lombardi, J.L. Lambert, A noninvasive glucose monitor: preliminary results in rabbits, Diabetes Technol. Ther. 1 (1999) 145–151.
[3] K. Maruo, M. Tsurugi, J. Chin, T. Ota, H. Arimoto, Y. Yamada, M. Tamura, M. Ishii, Y. Ozaki, Noninvasive blood glucose assay using a newly developed near-infrared system, IEEE J. Sel. Top. Quantum Electron. 9 (2) (2003) 322–330.
[4] E.T. Ooi, X.Q. Zhang, J.H. Chen, P.H. Soh, K. Ng, J.H. Yeo, Noninvasive blood glucose measurement using multiple laser diodes, SPIE Conference Series, vol. 6445, 2007.
[5] H.M. Heise, A. Bittner, Multivariate calibration for near-infrared spectroscopic assays of blood substrates in human plasma based on variable selection using pls-regression vector choices, Fresenius J. Anal. Chem. 362 (1) (1998) 141–147.
[6] F.M. Ham, I.N. Kostanic, G.M. Cohen, B.R. Gooch, Determination of glucose concentrations in an aqueous matrix from nir spectra using optimal time-domain filtering and partial least-squares regression, IEEE Trans. Biomed. Eng. 44 (6) (1997) 475–485.
[7] Y. Yamakoshi, M. Ogawa, T. Yamakoshi, M. Satoh, M. Nogawa, S. Tanaka, T. Tamura, P. Rolfe, K. Yamakoshi, A new non-invasive method for measuring blood glucose using instantaneous differential near infrared spectrophotometry, EMBS 2007. 29th A Int. Conference of the IEEE, , 2007, pp. 2964–2967.
[8] J.J. Burmeister, M.A. Arnold, Evaluation of measurement sites for noninvasive blood glucose sensing with near-infrared transmission spectroscopy, Clin. Chem. 45 (9) (1999) 1621–1627.
[9] L. Yun-Han, H. Fu-Rong, L. Shi-Ping, C. Zhe, Detection limit of glucose concentration with near-infrared absorption and scattering spectroscopy, Chin. Phys. Lett. 25 (3) (2008) 1117–1119.
[10] N.G. Yee, G.G. Coghill, Factor selection strategies for orthogonal signal correction applied to calibration of near-infrared spectra, Chemom. Intell. Lab. Syst. 67 (2) (2003) 145–156.
[11] B. Lindholm-Sethson, S. Han, S. Ollmar, I. Nicander, G. Jonsson, F. Lithner, U. Bertheim, P. Geladi, Multivariate analysis of skin impedance data in long-term type 1 diabetic patients, Chemom. Intell. Lab. Syst. 44 (1–2) (1998) 381–394.
[12] W. Cai, Y. Li, X. Shao, A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra, Chemom. Intell. Lab. Syst. 90 (2) (2008) 188–194.
[13] B.S.S. Wold, N. Kettaneh-Wold, Nonlinear pls modeling, Chemom. Intell. Lab. Syst. 7 (1989) 53–65.
[14] S. Wold, Nonlinear partial least squares modelling. ii. Spline inner relation, Chemom. Intell. Lab. Syst. 14 (1992) 1–3.
[15] X. Zhang, J. Chen, E.T. Ooi, J.H. Yeo, Noninvasive blood glucose monitoring with laser diode, in: G.L. Coté, A.V. Priezzhev (Eds.), SPIE Conference Series, vol. 6094, 2006, pp. 95–102.
[16] K. Bennett, M. Embrechts, Advances in learning theory: methods, models and applications, Ch. An Optimization Perspective on Kernel Partial Least Squares Regression, IOS Press, The Netherlands, 2003, pp. 227–249.
[17] B. Li, P.A. Hassel, A.J. Morris, E.B. Martin, A non-linear nested partial least-squares algorithm, Comput. Stat. Data Anal. 48 (2005) 87–101.
[18] Y. Du, Y. Liang, J. Jiang, R. Berry, Y. Ozaki, Spectral regions selection to improve prediction ability of pls models by changeable size moving window partial least squares and searching combination moving window partial least squares, Anal. Chim. Acta 501 (2) (2004) 183–191.
[19] S. Wold, Cross-validatory estimation of the number of components in factor and principal components models, Technometrics 20 (4) (1978) 397–405.
[20] H. Martens, T. Naes, Multivariate Calibration, John Wiley & Sons Inc, 1989.
[21] A.B.S. Wold, J. Trygg, H. Antti, Some recent developments in pls modelling, Chemom. Intell. Lab. Syst. 58 (2001) 131–150.
[22] W.L. Clarke, The original clarke error grid analysis (ega), Diabetes Technol. Ther. 7 (5) (2005) 776–779.

Please cite this article as: Z.-M. Chuah, et al., A two-level partial least squares system for non-invasive blood glucose concentration prediction, Chemometr. Intell. Lab. Syst. (2010), doi:10.1016/j.chemolab.2010.08.015