Partial Least Square with Savitzky Golay Derivative in Predicting Blood Hemoglobin Using Near Infrared Spectrum

Mohd Nazrul Effendy Mohd Idrus[1], and *Kim Seng* Chia[1,*]

[1]Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Malaysia

Abstract. Near infrared spectroscopy (NIRS) is a reliable technique that widely used in medical fields. Partial least square was developed to predict blood hemoglobin concentration using NIRS. The aims of this paper are (i) to develop predictive model for near infrared spectroscopic analysis in blood hemoglobin prediction, (ii) to establish relationship between blood hemoglobin and near infrared spectrum using a predictive model, (iii) to evaluate the predictive accuracy of a predictive model based on root mean squared error (RMSE) and coefficient of determination r_p^2. Partial least square with first order Savitzky Golay (SG) derivative preprocessing (PLS-SGd1) showed the higher performance of predictions with RMSE = 0.7965 and r_p^2 = 0.9206 in K-fold cross validation. Optimum number of latent variable (LV) and frame length (f) were 32 and 27 nm, respectively. These findings suggest that the relationship between blood hemoglobin and near infrared spectrum is strong, and the partial least square with first order SG derivative is able to predict the blood hemoglobin using near infrared spectral data.

1. Introduction

Hemoglobin concentration can be used to diagnose anemia [1]. Hemoglobin (Hb) is a protein molecule in red blood cells that contains an iron molecule to carries oxygen from the lungs to the rest of body [2]. Generally, people with hemoglobin level 12.0 g/dL or higher was defined as non-anemia, 10 to 11.9 g/dL was mild anemia, 9.9 to 7.0 g/dL was moderate anemia and lower than 7.0 g/dL was severe anemia [1]. Normally, blood hemoglobin was measured using Cyanmethemoglobin method with some blood drawn from patient to be mixed with reagent chemicals for analysis [3]. However, this method has limitations in term of time consuming, require reagent chemical in analysis and invasive method. In this way, NIRS method is a promising fast response, noninvasive and prominent technique to measure blood hemoglobin.

NIRS is a simple and reliable technique widely used in various field such as medical [4–10], food [11–15], agrochemical [12, 13, 15], and fuel [16]. NIRS measures overtones and combination tones of the fundamental molecular vibrations in especially the asymmetric vibrations and these properties make NIR useful for analyzing in biological system [17]. However, different reviews have demonstrated that fundamental reasons which limit the use of NIRS based on several factors; interference resulting in poor S/N ratio, calibration issues, baseline drift, thermal noise and proper selection of wave- length [18]. These problems of NIRS can degrade accuracy performance of the predictions. Due to these issues and nonlinearity of NIRS spectral data, multivariate calibration modelling method need to be developed for quantitative analysis of the target component in complex samples. Preprocessing, calibration and validations is a common process rely in developing multivariate calibration modelling [19].

Different types of multivariate calibration methods have been applied into NIRS spectral data to extract the relevant part of information for a large dataset to predict concentration from samples [16, 20]. The major concern for these multivariate calibration methods with spectral data in data nonlinearity [20, 21]. An appropriate calibration modelling need to be investigated to give an optimum performance predictions from the spectral sample data. Partial least squares (PLS) was famous predictive modelling used in NIRS due to its advantages of rapidity, simplicity and practicability [21, 22]. With a linear method combination of principal multi linear regression (MLR) and component analysis (PCA), PLS be able to handle data with strong co-linearity and noise, as well as in situations with the number of variables more than the number of samples [22]. PLS model show superior model in number of component in terms of effectiveness compared to multi linear regression (MLR,) principal component regression (PCR) [20]. Moreover, PLS has shown much better performance compared to artificial neural network (ANN) in term of RMSEP [24]. However, conventional PLS model need to use prior preprocessing step result to confront with the change in interferent structure in the test set and

* Corresponding author: kschia@uthm.edu.my

reducing the prediction error [8].

SG preprocessing method has been used successfully to remove unwanted signal from spectral data and overcome most common issues in raw spectral data from NIR [5, 10, 14, 15] . However, little studies have been conducted to investigate the effect of the preprocessing to the predictive accuracy of predictive models [25]. Thus, PLS combined with different SG preprocessing techniques (i.e. smoothing, first and second order derivative) is proposed to be investigated to predict the blood hemoglobin concentration in this study.

2. Material and methods

2.1 Research methodology

Figure 1 describe the general idea of the research methodology in this research. Raw spectral data from near infrared spectroscopy (NIRS) will be preprocessed with SG derivative. After that, PLS multivariate calibration will be used for modelling the spectral data and generate prediction value of blood hemoglobin.

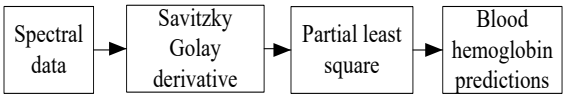

Fig. 1. The flow of blood hemoglobin predictions.

2.2 Spectral data

The spectral dataset were adopted from IDRC ShootOut 2010 that provided by Karl Norris [26]. Blood samples were analyzed during the period from 1990 to 1992 with an NIRSystems 6500 spectrometer with a transmission amplifier mounted in the sample transport. All spectra have 700 variables, from 1100 to 2498 nm, with a 2 nm interval as shown in **Fig. 2**.

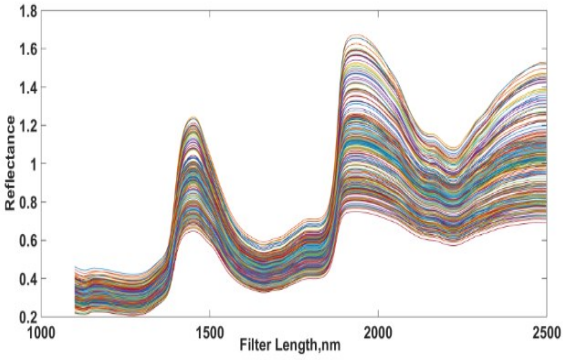

Fig. 2. Blood hemoglobin reflectance raw spectral data.

The data set contain n=231 sets for calibration and n=194 testing sets for blind test used to measure concentration of blood hemoglobin from blood constituent. The blood hemoglobin reference method was measured by Coulter STKS monitor, which is made by the Coulter Corporation of Hialeah, FL [27]. Descriptive statistics of the samples and reference blood hemoglobin showing, number of samples (n), minimum (Min), maximum (Max), mean and standard deviation (Std) as shown in Table 1.

Table 1. Descriptive statistics of the blood hemoglobin

	n	Min	Max	Mean	Std
Calibration	231	10.30	17.30	13.78	1.66
Testing	194	6.50	18.20	12.20	2.83

2.3 Savitzky Golay preprocessing

Preprocessing works to performed data loading, preprocessing zero order SG, first order SG derivative and second order SG derivatives to remove unwanted signal before the spectral data going to modelling process. Sets of 231 sample data $(X = 1:231)$ was processed with three level of SG preprocessing method produced smoothing SG (SG0), first order SG derivative (SG1) and second order SG derivative (SG2) data. The coefficients of SG (C0, C1 and C2) were generated by using built in matrix routine function from MATLAB simulation software (MATLAB® Version8.4 (R2014b)). Middle value from desired order derivative can be estimated by dot product of each value of C0, C1 and C2 represented coefficient differentiation filter with spectral data using following equation:

$$Y_j = (C \otimes y)_j = \sum_{i=-\frac{m-1}{2}}^{\frac{m-1}{2}} C_i y_{i+j} \quad (1)$$

Where C_i is set of SG coefficient; y_{i+j} is related set of data before treatment; Y_j is observed value after treatment. The range spectral data was treated between m set data $\frac{m-1}{2}$ and $n - \frac{m-1}{2}$. Where m and n is measured number and total number of frame length.

2.4 Partial least square

Partial least square was carried out using MATLAB simulation software (MATLAB® Version8.4 (R2014b)). General concept idea behind model of PLS modelling is to decompose both the design matrix predictor X and matrix of response Y as following equations:

$$X = TP^T \quad (2)$$

$$Y = UQ^T \quad (3)$$

Where X is an $n \times m$ matrix of predictors, Y is an $n \times p$ matrix of response. T and U is $n \times l$ matrix that are projections of X score and Y score respectively. P and Q are $m \times l$ and $p \times l$ orthogonal loading matrices respectively. The algorithm will yield the PLS regression estimates B_{ij} and B_o after estimating the factor and loading matrices T, U, P and Q for the linear regression as following equations:

$$Y_p = \sum X_{ij} B_{ij} + B_o \quad (4)$$

Where B_{ij} and B_o is PLS regression coefficient. Y_p is predicted value of blood hemoglobin. In this research, the coefficients of PLS regression were generated by using the MATLAB built-in matrix

routines function from MATLAB. PLS with SG preprocessing have four stage process start with preprocessing, training, validation and testing process.

2.5 Validations

K-fold cross validation has been used to evaluate performance of PLS model [27, 28]. There are three steps in K-fold cross validation. First, the data set was randomly divided into 5 disjoint folds with approximately equal size. Second, each fold turn to be test the model induced from the other k -1 folds with certain arrangement. After that, root mean square error of cross validation (RMSECV), root mean square error of prediction (RMSEP) and coefficient of determination of prediction r_p^2 can be determined to characterize prediction accuracy capacity of created model. The RMSECV, RMSEP and r_p^2 were calculated as follows:

$$\text{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}-y)^2}{n}} \quad (5)$$

where n is the total number of samples, $\hat{y}$ and y denote the predicted blood Hb and reference Blood Hb from calibration data set, respectively. While root mean squared error of prediction (RMSEP) is used to measures the accuracy of the predictions of the calibration model with new unseen of data set can be computed as

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_p-y_p)^2}{n}} \quad (6)$$

where n is the total number of samples, $\hat{y}$ and y denote the predicted blood Hb and reference Blood Hb from new unseen data set, respectively. The coefficient of determination of prediction r_p^2 used interpreted proportion of the variance in the predicted from reference value output of regression analysis is defined as

$$r_p^2 = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} \quad (7)$$

Where $\bar{y}$ is mean of reference data, $\hat{y}$ and y denote the predicted blood Hb and reference Blood Hb from new unseen data set, respectively. SSE and SST denoted as residual sum of squares and total sum of squares respectively.

3. Results and discussion

3.1 Savitzky Golay preprocessing

Fig. 3 shows the output of smoothing SG after raw spectra data from near infrared spectroscopy has been applied smoothing SG preprocessing. The result indicates that the spectra has been smooth and signal to noise ratio (SNR) has been increase without greatly distorting the original signal.

The raw spectra data has been treated with a set of 37nm frame length to produce convolution coefficients relatively. Thus, 38nm frame length from starting point (1100 to 1176nm) and end point (2422 to 2498nm) has been remove because convolution coefficients that has been applied to all data sub-sets to give estimates of the smoothed signal at the central point of each subset for 37nm frame length. It means 76 from 700 variable information of spectra data has been lost during preprocessing. Therefore, the use of number of frame length should be optimized to avoid more elimination information during preprocessing. Output spectra data after first order SG derivative preprocessing (SGd1) as shown at Fig. 4. The result indicate that baseline shift effect has been remove after the SGd1 process. The useful information of original spectra still available for modelling is between range (1306 to 1630 nm, 1824 to 2150 nm and 2232 to 2460).

In spite of that, all information spectral data from 1136 to 2460 nm can be used for modelling process. While Fig. 5 indicate more information from spectral data has been neglected after second order SG derivative preprocessing has been applied.

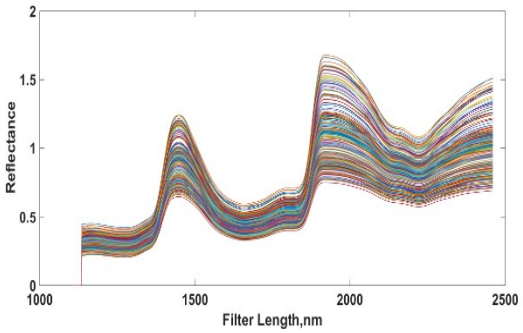

Fig. 3. The preprocessed spectral data after SG smoothing

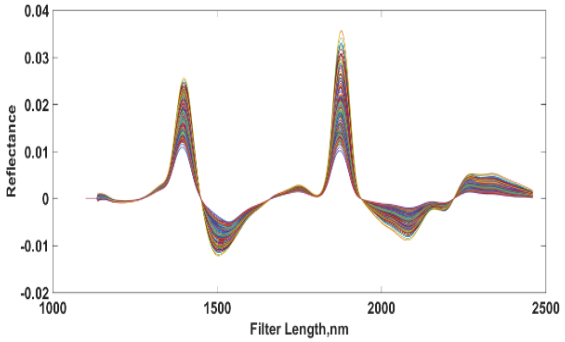

Fig. 4. The preprocessed spectral data after 1st order SG derivative

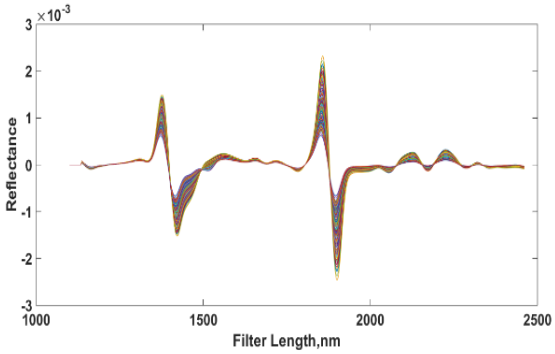

Fig. 5. The preprocessed spectral data after 2nd order SG derivative

3.2 Performance calibration of SG-PLS

Table 2 shows the RMSECV of PLS with smoothing, first order and second order SG derivative pre-processed spectral data with optimal filter and different LVs. As can be seen, best five number RMSECV from each type of preprocessing were presented. It indicated best of 25 models have been developed.

Table 2. RMSECV of PLS with Smoothing, 1st and 2nd Order SG Derivative pre-processed spectral data with optimal filter and different LVs.

Pre-processed method	*f* (nm)	LV	RMSECV (gd/L)
PLS Smoothing Savitzky-Golay (PLS-SGd0)	77	35	0.2194
	95	32	0.223
	97	34	0.2231
	93	35	0.2235
	95	33	0.2238
PLS First Order Savitzky-Golay Derivative (PLS-SGd1)	27	32	0.2178
	21	30	0.2182
	27	33	0.2184
	37	35	0.2185
	67	34	0.2201
PLS Second Order Savitzky-Golay Derivative (PLS-SGd2)	79	31	0.2163
	79	32	0.2185
	97	34	0.2189
	97	32	0.2190
	97	33	0.2195

Minimum value of RMSECV is 0.2163 gd/L follow by 0.2178 gd/L while highest is 0.2238 gd/L. Generally, value of RMSECV of 25 selected models is not much different with 2.18% different. Optimization on the use of an appropriate of frame length after PLS-SGd0 pretreatment resulted in high scale between 93nm to 99nm. This may happen because of the data still have baseline shift and slope effects after smoothing pretreatment process.

RMSECV value for PLS-SGd1 having a minimum number of frame length 37nm and range number of LVs is between 21 and 37. This slight improvement performance might have resulted from removing the slope effect. When higher filter length used to optimize performance of the model, the more original signal will be neglected because new value after preprocessing is a central value from dot product of convolution coefficient with spectral data. Optimal frame lengths were used for PLS-SGd2 are between 79nm to 99nm and LVs are between 31 to 33. It indicated that much information has been lost due the preprocessing process. Different order of SG preprocessing influencing number of frame length and latent variable used to perform optimum RMSECV.

3.3 Performance prediction of SG-PLS

192 new unseen samples from calibration has been used to measured performance prediction of PLS. From the result, indicates that optimization of frame length and number of LVs can improve performance of modelling for example PLS-SGd1 showed the higher performance of predictions (f = 27, LV = 32, r^2_p = 0.9206) followed by PLS-SGd0 (f = 77, LV = 35, r^2_p = 0.9190) as shown in Table 3. This could be due to the minimum number of frame length and lower number of LVs used. The results indicate that PLS-SGd1 have more robustness in testing new unseen test sets samples where the samples were frozen for storage and thawed before used for analysis.

Table 3. The Performance of PLS with different SG pre-processed spectral data with optimal filter and latent variable (LV) using Cross Validation

Pre-processed method	*f* (nm)	LV	RMSECV (gd/L)	RMSEP (gd/L)	r^2_p
PLS-SGd0	77	35	0.2194	0.8046	0.9190
PLS-SGd1	27	32	0.2178	0.7965	0.9206
PLS-SGd2	79	31	0.2163	0.8200	0.9159

4. Conclusions

As a conclusion, partial least square modelling is promising to predict the blood hemoglobin from near infrared spectral data. With the optimal frame length and latent variables, partial least square with first order Savitzky Golay derivative preprocessing (f = 27, LV = 32) showed the highest performance of prediction, i.e. RMSEP = 0.7965 gd/L and r^2_p = 0.9206 in K-fold cross validation. Thus, optimization in frame length and latent variable is crucial to improve the prediction performance. Next, findings also show that the smaller size the frame length, and the lower the number of LVs give a better prediction.

The author would like to acknowledge Research and Innovation Fund provided by the Office for Research, Innovation, Commercialization and Consultancy Management (ORICC), RMC, Universiti Tun Hussein Onn Malaysia (UTHM) for providing financial support, and Faculty of Electrical and Electronic Engineering, UTHM for providing facilities for this study.

5. References

1. M. Chan, "Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity," *Geneva, Switz. World Heal. Organ.*, pp. 1–6, 2011.
2. F. B. Jensen, A. Fago, and R. E. Weber, "Hemoglobin Structure and Function," *Fish Physiol.*, vol. 17, no. March 2016, pp. 1–40, 1998.

3. J. Rosenblit *et al.*, "Evaluation of three methods for hemoglobin measurement in a blood donor setting.," *Sao Paulo Med. J.*, vol. 117, no. 3, pp. 108–112, 1999.
4. P. Venkatesan, C. Dharuman, and and S. Gunasekaran, "squares s regression with A comparative study of principal component regression and partial least square application to FTIR diabetes data," *Indian J. Sci. Technol.*, vol. 4, no. 7, pp. 740–746, 2011.
5. N. N. M. Yatim *et al.*, "Noninvasive glucose level determination using diffuse reflectance near infrared spectroscopy and chemometrics analysis based on in vitro sample and human skin," *Syst. Process Control (ICSPC), 2014 IEEE Conf.*, no. March 2016, pp. 30–35, 2014.
6. S. Ramasahayam, K. S. Haindavi, B. Kavala, and S. R. Chowdhury, "Non invasive estimation of blood glucose using near infra red spectroscopy and double regression analysis," *Proc. Int. Conf. Sens. Technol. ICST*, pp. 627–631, 2013.
7. A. Al-Mbaideen and M. Benaissa, "Determination of glucose concentration from NIR spectra using independent component regression," *Chemom. Intell. Lab. Syst.*, vol. 105, no. 1, pp. 131–135, 2011.
8. M. Goodarzi, S. Sharma, H. Ramon, and W. Saeys, "Multivariate calibration of NIR spectroscopic sensors for continuous glucose monitoring," *TrAC - Trends Anal. Chem.*, vol. 67, pp. 147–158, 2015.
9. J. Hennrich, C. Herff, D. Heger, and T. Schultz, "Investigating deep learning for fNIRS based BCI," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, vol. 2015–Novem, pp. 2844–2847, 2015.
10. C. Wood *et al.*, "Near infra red spectroscopy as a multivariate process analytical tool for predicting pharmaceutical co-crystal concentration," *J. Pharm. Biomed. Anal.*, vol. 129, pp. 172–181, 2016.
11. K. S. Chia, H. A. Rahim, and R. A. Rahim, "Artificial Neural Network Coupled with Robust Principal Components in Near Infrared Spectroscopic Analysis," *IEEE 8th Int. Colloq. Signal Process. its Appl. Artif.*, no. 36, pp. 19–22, 2012.
12. S. Manager and T. Chemometric, "Artificial Neural Networks and Near Infrared Spectroscopy - A case study on protein content in whole wheat grain," no. 1, pp. 1–6, 2013.
13. P. Lin, Y. M. Chen, Y. He, G. W. Hu, X. L. Fu, and C. L. Gu, "Study on Nonlinear Multivariate Methods Combined with the Visible Near-Infrared Spectroscopy (Vis/NIRS) Technique for Detecting the Protein Content of Cheese," *Food Bioprocess Technol.*, vol. 7, no. 12, pp. 3359–3369, 2014.
14. N. C. T. Mariani, G. H. De Almeida Teixeira, K. M. G. De Lima, T. B. Morgenstern, V. Nardini, and L. C. C. Júnior, "NIRS and iSPA-PLS for predicting total anthocyanin content in jaboticaba fruit," *Food Chem.*, vol. 174, pp. 643–648, 2015.
15. T. R. Viegas, A. L. M. L. Mata, M. M. L. Duarte, and K. M. G. Lima, "Determination of quality attributes in wax jambu fruit using NIRS and PLS," *Food Chem.*, vol. 190, pp. 1–4, 2016.
16. H. Abdul Rahim, C. Kim Seng, and R. Abdul Rahim, "Prediction of soluble solid content in pineapple using adaptive linear neuron," *Sensors and Transducers*, vol. 168, no. 4, pp. 243–248, 2014.
17. L. Norgaard, B. Rasmus, and S. B. Engelsen, "Principal Component Analysis and Near Infrared Spectroscopy," *A white Pap. from FOSS*, pp. 2–7, 2014.
18. J. Yadav, A. Rani, V. Singh, and B. M. Murari, "Prospects and limitations of non-invasive blood glucose monitoring using near-infrared spectroscopy," *Biomed. Signal Process. Control*, vol. 18, pp. 214–227, 2015.
19. M. Blanco and I. Villarroya, "NIR spectroscopy: A rapid-response analytical tool," *TrAC - Trends Anal. Chem.*, vol. 21, no. 4, pp. 240–250, 2002.
20. R. M. Balabin, R. Z. Safieva, and E. I. Lomakina, "Comparison of linear and nonlinear calibration models based on near infrared (NIR) spectroscopy data for gasoline properties prediction," *Chemom. Intell. Lab. Syst.*, vol. 88, no. 2, pp. 183–188, 2007.
21. H. Yang, P. R. Griffiths, and J. D. Tate, "Comparison of partial least squares regression and multi-layer neural networks for quantification of nonlinear systems and application to gas phase Fourier transform infrared spectra," *Anal. Chim. Acta*, vol. 489, no. 2, pp. 125–136, 2003.
22. X. Shao, X. Bian, J. Liu, M. Zhang, and W. Cai, "Multivariate calibration methods in near infrared spectroscopic analysis," *Anal. Methods*, vol. 2, no. 11, p. 1662, 2010.
23. S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: A basic tool of chemometrics," *Chemom. Intell. Lab. Syst.*, vol. 58, no. 2, pp. 109–130, 2001.
24. X. Jintao, Y. Liming, L. Yufei, L. Chunyan, and C. Han, "Noninvasive and fast measurement of blood glucose in vivo by near infrared (NIR) spectroscopy," *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.*, vol. 179, pp. 250–254, 2017.
25. K. S. Chia, H. Abdul Rahim, and R. Abdul Rahim, "Evaluation of common pre-processing approaches for visible (VIS) and shortwave near infrared (SWNIR) spectroscopy in soluble solids content (SSC) assessment," *Biosyst. Eng.*, vol. 115, no. 1, pp. 82–88, 2013.
26. C. S. Rules, "Chemometrics ShootOut Rules," *Int. Diffus. Reflectance Conf. 2012*, no. June, 2012.

27. J. T. Kuenstner, K. H. Norris, and W. F. McCarthy, "Measurement of hemoglobin in unlysed blood by near-infrared spectroscopy," *Appl. Spectrosc.*, vol. 48, no. 4, pp. 484–488, 1994.
28. G. Jiang and W. Wang, "Error estimation based on variance analysis of k -fold cross-validation," *Pattern Recognit.*, vol. 69, pp. 94–106, 2017.
29. T.-T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," *Pattern Recognit.*, vol. 48, no. 9, pp. 2839–2846, 2015.