Equidistant combination moving window MLR method for wavenumbers selection of NIR analysis

T. Pan,* J. Xie, H. Chen, H. Yin and X. Chen

Key Laboratory of Optoelectronic Information and Sensing Technologies of Guangdong Higher Educational Institutes, Department of Optoelectronic Engineering, Jinan University, Guangzhou 510632, P.R. China.
**E-mail: tpan@jnu.edu.cn*

Keywords: near infrared spectral analysis, serum, glucose, MWPLS, ECMWMLR

Introduction

Wave-number selection of near-infrared spectral (NIRS) analysis is very important for improving model prediction effectiveness, in reducing model complexity, and in designing special NIR spectroscopy instruments with high SNR. The multiple linear regression (MLR) method has the advantages of simple modeling, but cannot overcome spectral collinearity. Combining the advantages of MLR and principal component analysis (PCA), partial least squares (PLS) was developed. PLS could overcome the collinearity and became widely used in NIRS analysis. The moving window partial least squares (MWPLS) method was developed and effectively used to select the optimal spectral regions.[1–3]

In this paper, an equidistant combination moving window multiple linear regression (ECMWMLR) method for wave-number combination selection for NIRS analysis is presented. The ECMWMLR method extracts equidistant data from the continuous spectral range, to establish a MLR model, retains the simplicity of MLR, and also can overcome spectral collinearity, when the data gap selection is appropriate. The selected wave-number combination is the mode of "quasi continuous", so that it can be easily preprocessed by methods such as Savitzky-Golay (SG) smoothing,[4–8] and the prediction effectiveness can be further improved.

Using the NIRS analysis of serum glucose as an example, the optimal wave-number combination was selected by ECMWMLR. As a comparison, the results were compared to results obtained by MWPLS.

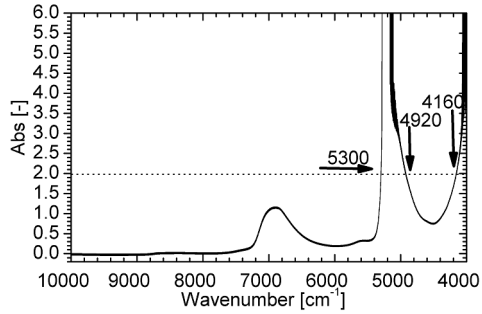

Figure 1. Near-infrared spectra of 191 human serum samples.

Experiment and methods

A total of 191 human serum samples were collected. Glucose concentration of the samples was measured by a routine chemical method, and ranged from 3.53 to 6.15 mmol L^{-1}. The mean and standard deviation are 4.90 and 0.59 mmol L^{-1}. The NIR spectra were measured from 10000 to 4000 cm^{-1} by Nicolet 5700 FT-NIR spectrometer using a 2 mm transmission accessory, and 64 scans at 4 cm^{-1} resolution were co-added for each spectrum.

The average NIR spectra of 191 samples are shown in Figure 1.

Since the spectra near 5200 cm^{-1} and 4000 cm^{-1} are shown with very strong absorption, low spectral energy, low information quality and much noise, bands where the absorbance was higher than 2 were eliminated, and the combination bands of 10000–5301 and 4918–4160 cm^{-1} were selected as whole wave-number regions for modeling. Based on the requirement for prediction of the effectiveness of the optimal single wave-number model, all samples were divided into a calibration set (131 samples) and a prediction set (60 samples).

The ECMWMLR method extracted equidistant data in the continuous spectral range, compared to the established MLR model. The moving window method was applied to the selection of equidistant wave-numbers. For ECMWMLR, numbers of adopted wave-numbers (N_E), gaps of adopted wave-numbers (G) and beginning wave-numbers (B), were set from 2 to 100, from 1 to

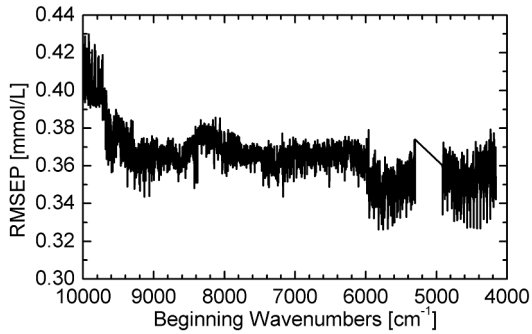

Figure 2. *RMSEP* of optimal model corresponding to beginning wavenumbers by ECMWMLR.

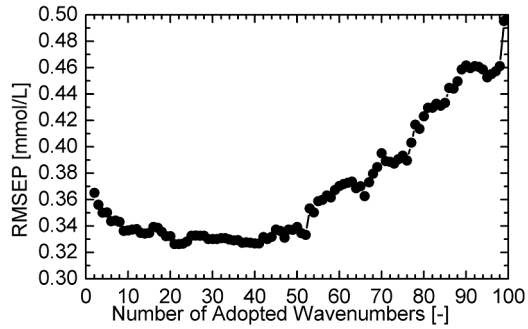

Figure 3. *RMSEP* of optimal model corresponding to number of adopted wavenumbers by ECMWMLR.

250 and from 4160 to 10000 cm^{-1}, respectively. The MLR model corresponding to each parameter combination (N_E, G, B) was established and the optimal wave-number combination was selected, according to the prediction results.

For MWPLS, numbers of adopted wave-numbers (N_M), beginning wave-numbers (B) and PLS factors (F) were set from 2 to 2830, 10000 to 4160 cm^{-1} and from 1 to 30, respectively. The PLS model corresponding to each parameter combination (N_M, B, F) was established, and the optimal spectral region was selected by the prediction of the prediction set.

The selected wave-number combinations and spectral bands were preprocessed by a SG smoothing method, and the models were re-established. The SG smoothing parameters included the order of derivative (OD), degree of polynomial (DP) and number of smoothing points (NSP).

The model evaluation indicators include root mean squared error of prediction (*RMSEP*), correlation coefficient of prediction (R_P) and the relative root mean squared error of prediction to the mean of chemical values of all samples in the prediction set (*RRMSEP%*). The *RMSEP* was used as the goal of model optimisation and parameter design.

Table 1. Selected wavenumbers combinations, spectral regions and their prediction effects.

	MWPLS		ECMWMLR		
Wavenumbers (cm^{-1})	6705–5415	5961–5519	Begin	7130	6000
			End	5753	5591
N_M	**670**	230	N_E	22	**31**
			G	34	**7**
OD	**0**	4	—	0	**0**
DP	**6**	6	—	6	**2**
NSP	**81**	49	—	13	**7**
F	**16**	12	—	—	—
RMSEP (mmol L^{-1})	**0.342**	0.351	—	0.326	**0.316**

Results and discussion

By using ECMWMLR and MWPLS methods respectively, the wave-number combinations and spectral bands of the NIR analysis for serum glucose were selected. The selected wave-number combinations and spectral regions were preprocessed by SG smoothing, and further treatment, to establish MLR and PLS models respectively.

RMSEP of the optimal model corresponding to each beginning wave-numbers, and each number of adopted wave-numbers by ECMWMLR are shown in Figure 2 and Figure 3, respectively.

Selected wave-number combinations, spectral ranges and their prediction effects are summarised in Table 1.

The results show that the optimal prediction effects obtained by ECMWMLR were slightly better than those obtained by the MWPLS method.

Conclusion

Optimal models for NIRS analysis of serum glucose were selected by MWPLS and ECMWMLR. For the wavelength combination of the optimal model by ECMWMLR, N_E was 31, G was 7, and B was 6000 cm^{-1}. The smoothing mode was the original spectral smoothing, 2, 3 degree polynomial, 7 smoothing points. The prediction correlation coefficient, R_P, was 0.851, *RMSEP* was 0.316 mmol L^{-1} and the *RRMSEP* reached 6.5%. For the spectral range of the optimal model by MWPLS, the optimal wavelength range was 6705–5415 cm^{-1}, N_M was 670, smoothing mode was the original spectral smoothing, 6 degree polynomial, 81 smoothing points, F was 16, the prediction correlation coefficient R_P was 0.823, the *RMSEP* was 0.342 mmol L^{-1} and the *RRMSEP* reached 7.0 %. The results show that the optimal prediction effects obtained by ECMWMLR were slightly better than those obtained by MWPLS. ECMWMLR can be effectively applied to the wave-number selection for NIR analysis.

Acknowledgement

This work was supported by the NSF of China (10771087), the NSF of Guangdong (7005948), the Science and Technology Project of Guangdong (2007B030501008, 2009B030801239), the Science and Technology Project of Guangzhou (2007Z3-E0281).

References

1. J. Jiang, R. James Berry, H.W. Siesler and Y. Ozaki, *Anal. Chem*. **74,** 3555 (2002).
2. Y. Du, Y. Liang, J. Jiang, R.J. Berry and Y. Ozaki, *Analy. Chim. Acta* **501,** 183 (2004).
3. S. Kasemsumran, Y. Du, K. Maruo and Y. Ozaki, *Chemometr. Intell. Lab. Syst*. **82,** 97 (2006).
4. A. Savitzky and M.J.E. Golay, *Anal. Chem*. **36,** 1627 (1964).
5. K. Nakanishi, A. Hashimoto, T. Pan, *et al*., *Appl. Spectrosc*. **57,** 1510 (2003).
6. T. Pan, A. Hashimoto, M. Kanou, *et al*., *Bioprocess and Biosystems Engineering* **26,** 133 (2003).
7. J. Chen, T. Pan and X. Chen, *Optics and Precision Engineering* **14,** 1 (2006).
8. P. Cao, T. Pan and X. Chen, *Optics and Precision Engineering* **15,** 1952 (2007).