Wavelength selection and stability for near infrared spectroscopy analysis based on different divisions for calibration set and prediction set

Tao Pan*, Jun Xie, Huazhou Chen and Hao Yin

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

Introduction

Selecting the appropriate wavelength range for near infrared (NIR) spectroscopic analyses is important for improving the efficacy of model prediction, reducing model complexity, and for designing special NIR spectroscopy instruments, especially for NIR analysis of complex systems. In this paper, an equidistant combination moving window multiple linear regression (ECMWMLR) method[1-3] for wavelength selection is demonstrated. NIR spectroscopy analysis of human serum glucose was used as an example; the optimal wavenumber combinations based on ECMWMLR were selected. As a comparison, the results based on moving window partial least squares (MWPLS)[4] were also obtained.

To get stable results, all models were obtained using the average data from 50 different divisions of the calibration and prediction sets.

Materials and Methods

Experimental materials, instrument and measurement method

One hundred and ninety one serum samples with glucose concentrations ranging from 3.53 to 6.15 mmol.l^{-1} were collected. The mean value and the standard deviations were 4.90 and 0.59 mmol.l^{-1}, respectively. Spectra were collected using a Nicolet 5700 FT-NIR spectrometer with a 2 mm quartz cuvette. The spectral range was 10000–4000 cm^{-1} with 64 scans at 4 cm^{-1} resolution. Spectra near 5200 cm^{-1} and 4000 cm^{-1} were spurious (absorbance higher than 2; spectral noise) and were thus eliminated, and the combination of 10000–5301 and 4918–4160 cm^{-1} was used to represent the whole spectral region for wavenumber selection.

Dividing for calibration set and prediction set

All samples were divided into calibration and prediction sets at a ratio of about 2:1. Model parameters were changed as the calibration and prediction sets were subdivided, to avoid fluctuations of prediction effects. In order to establish objective models, a division method for the calibration and prediction sample sets based on the optimal single wavenumber prediction bias (OSWPB) was proposed.[5] A total of 50 different divisions of the calibration set and the prediction set were made. Calibration models were established for each division. Model predictions (e.g. root mean squared error of prediction) in 50 different divisions were averaged for each combination of model parameters. Based on the average data, the stable optimal model could be selected.

ECMWMLR method

Parameters of the ECMWMLR method include the beginning wavenumber (B), the number of adopted wavenumbers (N_E) and the gap of adopted wavenumbers (G). B was set from 4160 to 4920 and from 5300 to 10000 (cm^{-1}), N_E was set from 1 to 60, and G was set from 1 to 250. A multiple linear regression model for each parameter combination (B, N_E, G) was generated.

MWPLS method

Parameters of the MWPLS method include the beginning wavenumber (B), the number of adopted wavenumbers (N_M) and the PLS factor (F). B was set from 4160 to 4920 and from 5300 to 10000 (cm^{-1}), N_M was set from 1 to 2831, and F was set from 1 to 30. A PLS model for each parameter combination (B, N_M, F) was generated.

Model evaluation indicators

The model evaluation indicators include the root mean squared error of prediction (RMSEP) and correlation coefficient of predication (R_P). The parameters RMSEP and R_P were calculated for all 50 divisions, each

Reference paper as:
T. Pan, J. Xie, H. Chen and H. Yin (2012).Wavelength selection and the stability for near infrared spectroscopy analysis based on different divisions for calibration set and prediction set, in: Proceedings of the15th International Conference on Near Infrared Spectroscopy, Edited by M. Manley, C.M. McGoverin, D.B. Thomas and G. Downey, Cape Town, South Africa, pp. 276-278.

parameter was averaged, and the mean RMSEP value was used as the goal of model optimisation and parameter design.

Results and Discussion

The NIR spectra of 191 human serum samples

The NIR spectra of 191 human serum samples are shown in Figure 1. All samples were divided among the calibration (131 samples) and prediction (60 samples) sets using the prediction of the optimal single wavenumber (7232 cm^{-1}) model.

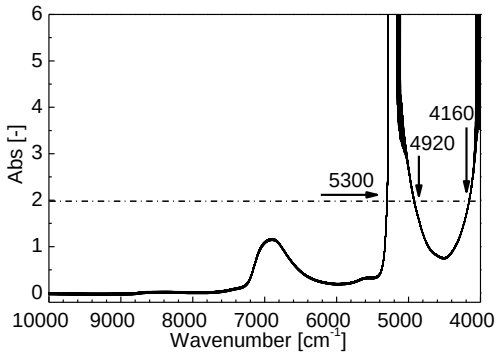

Figure 1. Near infrared spectra of 191 human serum samples.

RMSEP of the optimal models based on ECMWMLR

The optimal model for the fixed B, changed N_E and changed G was selected according to RMSEP; RMSEP of the optimal model corresponding to each B is shown in Figure 2. The optimal model for the fixed N_E, changed B and changed G was also selected according to RMSEP (Figure 3).

RMSEP of the optimal models based on MWPLS

The optimal models for the fixed B and changed N_M, and fixed N_M and changed B, were selected according to RMSEP values (Figures 4 and 5, respectively).

Selected wavenumber combination

The global optimal model based on ECMWMLR was selected, and the corresponding B, N_E and G were 4684 cm^{-1}, 11 and 74, respectively. The mean value and standard deviation of RMSEP were 0.389 and 0.028 $mmol.l^{-1}$, respectively. The global optimal model based on MWPLS was selected, and the corresponding B, N_M and F were 5602 cm^{-1}, 190 and 5, respectively. The selected wavelength range was 5967–5602 cm^{-1}, and the mean value and standard deviation of RMSEP were 0.406 and 0.025 $mmol.l^{-1}$, respectively.

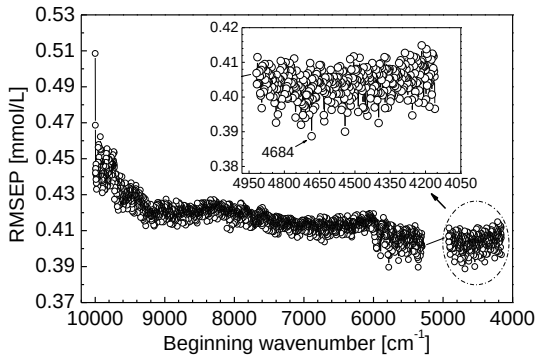

Figure 2. RMSEP of optimal model based on ECMWMLR corresponding to each beginning wavenumber.

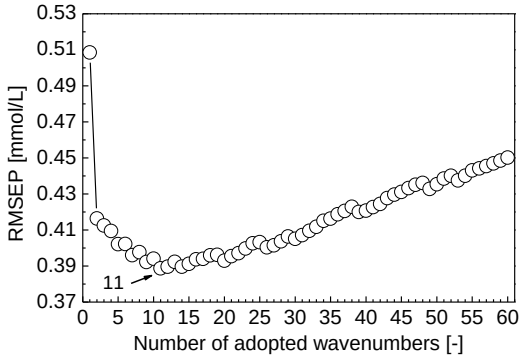

Figure 3. RMSEP of optimal model based on ECMWMLR corresponding to each number of adopted wavenumbers.

Reference paper as:
T. Pan, J. Xie, H. Chen and H. Yin (2012).Wavelength selection and the stability for near infrared spectroscopy analysis based on different divisions for calibration set and prediction set, in: Proceedings of the15th International Conference on Near Infrared Spectroscopy, Edited by M. Manley, C.M. McGoverin, D.B. Thomas and G. Downey, Cape Town, South Africa, pp. 276-278.

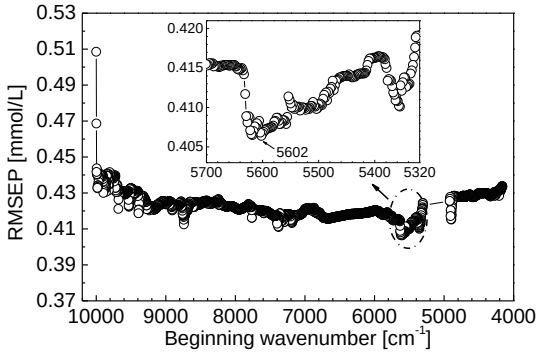

Figure 4. RMSEP of optimal model based on MWPLS corresponding to each beginning wavenumber.

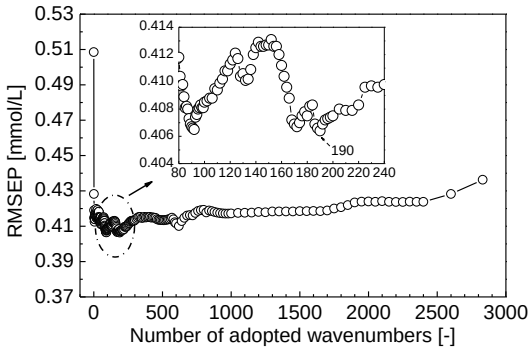

Figure 5. RMSEP of optimal model based on MWPLS corresponding to each number of adopted wavenumbers.

Conclusion

Results presented here were obtained using the average data of prediction from 50 different divisions of the calibration and prediction sets. The optimal ECMWMLR model only used 11 wavenumbers to provide a stable model with a slightly better prediction than an optimal MWPLS model. The equivalent ECMWMLR method extracted spectral data and overcame spectral collinearity, while retaining the simplicity of MLR. Unlike other discrete combination methods, ECMWMLR can be used in conjunction with spectral preprocessing (e.g. Savitzky-Golay smoothing) to further improve predictive ability. ECMWMLR is an effective method for selecting the appropriate NIR wavelengths for use in calibrations.

Acknowledgements

This work was supported by NSF of China (10771087, 61078040), the science and technology project of Guangdong province (2009B030801239).

References

1. T. Pan, J. Xie, H.Z. Chen, H. Yin and X.D. Chen, in *Near Infrared Spectroscopy: Proceedings of the 14th International Conference*, Ed by S. Saranwong, S. Kasemsumran, W. Thanapase and P. Williams, IM Publications, Chichester, UK, pp. 867-870 (2010).
2. T. Pan, H.Z. Chen, J. Xie, H. Yin and X.D. Chen, in *Near Infrared Spectroscopy: Proceedings of the 14th International Conference*, Ed by S. Saranwong, S. Kasemsumran, W. Thanapase and P. Williams, IM Publications, Chichester, UK, pp. 159-162 (2010)
3. T. Pan, J. Xie, Y. Shan, H.Z. Chen, *Adv. Mater. Res.* **181-182**, 647-650 (2011).
4. J.H. Jiang, R.J. Berry, H.W. Siesler and Y. Ozaki, *Anal. Chem.* **74**, 3555-3565 (2002).
5. J. Xie, T. Pan, J.M. Chen, H.Z. Chen and X.H. Ren, *Chinese J. Anal. Chem.* **38**, 342-346 (2010).

Reference paper as:
T. Pan, J. Xie, H. Chen and H. Yin (2012).Wavelength selection and the stability for near infrared spectroscopy analysis based on different divisions for calibration set and prediction set, in: Proceedings of the15th International Conference on Near Infrared Spectroscopy, Edited by M. Manley, C.M. McGoverin, D.B. Thomas and G. Downey, Cape Town, South Africa, pp. 276-278.