Journal of Molecular Structure 1019 (2012) 97-102

Contents lists available at SciVerse ScienceDirect

Journal of Molecular Structure

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

Absorption characteristics and quantitative contribution of overtones and combination of NIR: Method development and validation

Zhisheng Wu, Min Du, Bing Xu, Zhaozhou Lin, Xinyuan Shi*, Yanjiang Qiao*

Beijing University of Chinese Medicine, Beijing 100102, China

The Key Laboratory of TCM-Information Engineering of State Administration of Traditional Chinese Medicine, Beijing 100102, China

ARTICLE INFO

Article history: Received 28 February 2012 Accepted 26 March 2012 Available online 3 April 2012

Keywords: Near-infrared spectroscopy Accuracy profile Frequency range Partial least squares

ABSTRACT

The present study focuses on the interpretation of the relationship between absorption characteristics and quantitative contribution in each near-infrared (NIR) frequency range. NIR spectra were measured for citric acid in each mixture solvent of water and ethanol. Accuracy profile theory was successfully utilized to validate the performance of partial least squares (PLS) models in each frequency range. It was found that saturated absorption characteristic (i.e. intensity) occurred in combination region (CR) and proportion of first combination-overtone (FCOT) region. The result of chemometric indicators, i.e. correlation coefficient (r), the root mean square errors of calibration and prediction (RMSEC and RMSEP). revealed the importance of second combination-overtone (SCOT) region in the calibration model performance. Simultaneously, it demonstrated that PLS model in FCOT or CR region could provide the lowest RMSEC and RMSEP, only when FCOT or CR region has rich transmission information. According to validation methodology parameters, i.e. relative bias, repeatability, intermediate precision, risk, lower limits of quantification (LLOO), these results also indicated that FCOT or CR region was preferred as variables selection region, but if saturated absorption FCOT or CR region occurred, SCOT region was a better choice. In addition, the validated results testified that absorption intensities of solvent seriously interfered with model performance, especially for FCOT or CR region. Therefore, this revelation work provided a guideline for variables selection in each NIR frequency range.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Near-infrared (NIR) is a promising instrument used for process analysis. The wavenumber of NIR is from 4000 cm⁻¹ to 12,500 cm⁻¹, which mainly covers overtones and combinations of molecular vibrations. These signals decrease significantly in absorption cross section, compared with the fundamental vibrational bands from mid-infrared. The functional groups, C–H, N–H, O–H, S–H [1–4] are almost exclusively involving the hydrogen atom. According to molecular vibrations, the NIR frequency range can be divided into four ranges: combination region (CR, 4000–4900 cm⁻¹), first combination-overtone (FCOT, 4900–7100 cm⁻¹), second combination-overtone (SCOT, 7100–10,000 cm⁻¹), and third overtone (TO, 10,000–12,500 cm⁻¹) [5].

There are two major absorption bands at FCOT frequency region for water molecule, which affected the accuracy and precision of test materials at aquatic solution [2,4–6]. Additionally, the NIR overall absorption intensities at CR region are typically 2–100 times that of the overtones between FCOT and SCOT regions. Therefore, absorption characteristics and quantitative contribution of each frequency region should be notable. In recent years, NIR has been more and more widely applied with success in the research of traditional Chinese medicine (TCM), and the main subjects are content measurement and geographical identification [7–11].

Water, ethanol, mixture of ethanol and water have been used as extraction solvent for TCM. In every Chinese patent prescription, extraction process employs ethanol or water as good solvent. However, NIR absorption characteristic of these extraction solvents in each frequency region was not illustrated. In previous way, NIR absorption could be characterized by various variables selection methods, which are accurate but belonging to the object itself and unable to satisfy the whole common needs [12,13]. To resolve this issue, we observed the absorption characteristics from SCOT range to CR range, i.e. saturated absorption. Citric acid was used as solute, and water, ethanol and mixture of ethanol and water were used as solvent respectively. Quantitative performance was investigated using partial least squares (PLS) models in each absorption characteristics of spectrum and each frequency region.

PLS model fitting result can generally be evaluated according to the following chemometric indicators: low root mean square error in cross-validation (RMSECV), low root mean square prediction error (RMSEP), high correlation coefficient (*r*), and low bias, etc.





^{*} Corresponding authors at: Beijing University of Chinese Medicine, Beijing 100102, China. Tel.: +86 10 84738621; fax: +86 10 84738661.

E-mail addresses: shixinyuan01@163.com (X. Shi), yjqiao@263.net (Y. Qiao).

^{0022-2860/\$ -} see front matter @ 2012 Elsevier B.V. All rights reserved. http://dx.doi.org/10.1016/j.molstruc.2012.03.058

[14–19]. Once a calibration is developed and favorable predictions are expected, they must be validated to be accepted for routine use. For external validation, an independent set of samples is needed. The external validation set may include only the samples that are similar to those that are intended to be analyzed. There are several validation parameters that must be determined in order to be consistent with the recommendations of International Conference of Harmonization (ICH) and other guidelines: accuracy, precision (repeatability and intermediate precision), specificity, linearity and range of application. Based on "β-expectation tolerance intervals", the accuracy profile gives PLS model a reliable validation and actual future performance [20–25]. It fully complies with the ICH Q2 (R1) regulatory documents as it integrates all the useful required methodology parameters, i.e. accuracy, precision, lower limit of quantification (LLOQ), Upper limit of quantization (ULOQ) range and linearity [21-23.25.26]. Therefore, a novel approach based on accuracy profile was followed to validate the quantitative results to obtain the quantitative contribution of each frequency region.

2. Accuracy profile theory

The basic idea of accuracy profile is the acceptability limit criterion, noted λ . It is assumed that end-users actually expect from an analytical procedure to return a result \hat{Z} which differs of the unknown target value Z of less than λ . This requirement can be shown by following equation:

$$|Z - \hat{Z}| < \lambda \tag{1}$$

A procedure can be validated if it is very likely that the requirement given by Eq. (1) is fulfilled, i.e.:

$$P(|Z - \hat{Z}| < \lambda) \ge \beta \tag{2}$$

 β is the probability indicating the possibility for a future determination to fall within the acceptability limits. It is possible to compute the so-called " β -expectation tolerance interval" (β -ETI). The β ETI is given by

$$\delta \pm \mathbf{Q}_t \mathbf{k}_s \mathbf{S}_R \tag{3}$$

 Q_t is the β quantile of the Student's *t*-distribution, δ is the bias and k_s is the expansion factor:

$$k_{\rm S} = \sqrt{1 + \frac{1}{pnB^2}} \tag{4}$$

With

$$B = \sqrt{\frac{R+1}{nR+1}}; \quad R = \frac{S_B^2}{S_r^2} \tag{5}$$

 S_B^2 and S_r^2 are the estimates of the between conditions variance and the within conditions variance (repeatability). The intermediate precision S_R^2 is obtained as:

$$S_R^2 = S_r^2 + S_B^2$$
(6)

It can be easily demonstrated that

$$k_s^2 S_R^2 = S_R^2 + \frac{n S_B^2 + S_r^2}{np}$$
(7)

The right term of the second member of Eq. (7) is the estimation of the variance of the overall mean that can be assimilated to the bias uncertainty for a nested design with *p* conditions for experiments and *n* replications within each condition. Thus

$$k_s^2 S_R^2 = S_R^2 + S_\delta^2 = u^2(Z)$$
(8)

And βETI can be given now by

$$\delta \pm Q_t u(Z) \tag{9}$$

Therefore, based on β -expectation tolerance intervals, the accuracy profile brought PLS model a visual representation in each absorption characteristics of NIR spectrum.

3. Materials and methods

3.1. Materials

Citric acid (Beijing Chemical Works) was purchased from the local manufacturer. Ethanol (Concord Technology Co., Ltd. Tianjin, China) was used according to analytical grade. Deionized water was purified by Milli-Q water system (Millipore Corp., Bedford, MA, USA).

3.2. NIR equipment and software

The NIR spectra were collected with the transmission mode using the Fourier Transform NIR with Thermonic resulting software (Thermo Nicolet Corporation). Each spectrum was scanned for 32 times with 16 cm⁻¹ resolution. The range of spectra was from 4000 cm⁻¹ to 10,000 cm⁻¹. Each sample was analyzed three times and the average result of three spectra was used for future analysis. The sample was hold in a circular liquid cuvette with plastic cap (Optical path is 8 mm).

Data analysis was performed with home-made routines programmed in MATLAB code (MATLAB v 7.0, The MatWorks, Massachussetts). The calculation of the accuracy profile based on the external validation set results were given by e. noval V3.0 (Arlenda, Liège, Belgium).

3.3. Experimental design

Ethanol/water solvent is divided into seven groups with different proportions (100: 0, 90:10, 70:30, 50:50, 30:70, 10:90, 0:100, v/ v), every group was prepared to construct different absorption characteristics. Citric acid was added into 500 mL volumetric -flask separately, and diluted with ethanol/water solvent as described above (containing 20 mg of citric acid per mL).

Each stock solution with different proportions were transferred to 25 mL volumetric flasks separately, and diluted with each solvent to volume for establishment of the calibration set. Similarly, for the validation set establishment, different volumes of stock solution were diluted in 25 ml-flasks. As seen in Table 1, different variability sources were separately created for the calibration and validation steps in order to obtain robust model.

3.4. Chemometric indicators

To display the result of NIR-calibration models, the optimum preprocessing method was established based on the lowest prediction residual error-sum squares (PRESS). Cross validation was used with a segment size of four, and produced PRESS plot. The first minimum point on the PRESS plot is usually used to determine

Table 1

Variability sources included in the calibration and validation sets.

	Calibration set Validation set		
Variability sources	Amount of variability		
Solute	Citric acid		
Solvents	Mixture of water and ethanol		
Concentration levels	9 (0.5, 1.5, 2.5, 3.5, 4.5, 6,	9 (1, 2, 3, 4, 5, 6.5, 8,	
(mg/mL)	7.5, 8.5,10)	9, 9.5)	
Operators	2	2	
Days	2	2	
NIR acquisition	3	3	

the optimum number of factors. In addition, PLS model could generally be evaluated according to the chemometric indicators, i.e. lowest root mean square error of prediction (RMSEC) and RMSEP. The related indicators were calculated as following:

$$PRESS = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$
(10)

$$RMSEC = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{N}}$$
(11)

$$RMSEP = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}}$$
(12)

4. Results and discussion

4.1. Absorption characteristics in each frequency region

Fig. 1 shows FT-NIR transmission spectra of citric acid in ethanol/water solvent in different proportions. More water proportion leads to less transmission information. In detail, the transmission intensities varied obviously in SCOT region, which depended on the water amount, while FCOT region shows little difference except in ethanol solution and 10% water (90% ethanol) solution. It emphasized that the transmittance value at CR region (4900–4000 cm⁻¹) and part of FCOT region (7000–6500 cm⁻¹) was close to zero, indicating saturated absorption in NIR system.

4.2. Model performance in each frequency region

The use of each NIR frequency region involves the development of calibration models to relate the absorption characteristics. The type of the data pre-processing technique prior to the development of calibration may greatly influence model performance. This work analyses the effect of some commonly used pre-processing techniques prior to PLS model, i.e. 1st order derivative (1st), 2nd order derivative (2nd), Savitzky-Golay smooth (s-g, 9, 2), 1st order derivative + Savitzky-Golay smooth (1st + s-g), 2nd order derivative + Savitzky-Golay smooth (2nd + s-g). The optimum number of latent factors and pre-processing method are determined by the lowest PRESS value. Fig. 2 shows the raw spectra provided the lowest PRESS value among all preprocessing methods.

According to the result presented in Fig. 2, the PLS models were developed based on the different NIR frequency regions. Fig. 3 shows the different prediction results in each frequency region (Latent factors for PLS models were shown in Supporting information). Only when FCOT region has rich transmission information such as ethanol solvent and 10% water (90% ethanol) solvent, the results clearly indicated that PLS model in FCOT region could

provide the lowest RMSEC and RMSEP in contrast to models in other frequency regions, i.e. ethanol solvent, the RMSEC, RMSEP and *r* were 0.132 mg/mL, 0.155 mg/mL and 0.9989.

For further comparison, the model performance decreased with the drastic interference of increase proportion of water. The prediction performance in FCOT or CR region was more susceptible than SCOT or ALL regions. When water percentage attained to 30%, the relative prediction errors of PLS model in FCOT or CR region did not satisfy the analysis needs. In addition, when FCOT or CR region was saturated absorption, the model performance using SCOT region was better than the one using FCOT or CR region in all solvents. The similar prediction performance between SCOT region and ALL region presented in Fig. 3 reinforced the importance of the SCOT region in the performance of the calibration model.

4.3. Validation of model performance in each frequency region in one solvent

Validation is a crucial and mandatory step to obtain a robust NIR method. Therefore, accuracy profile theory was used to evaluate PLS models in each frequency region. Fig. 4 displays the accuracy profile result obtained by PLS model in ethanol solvent. The acceptance limits were set at $\pm 10\%$ while the maximum risk to obtain results outside these acceptance limits was set at 5%. Holistic methodology parameters i.e. linearity, relative bias, uncertainty, repeatability, intermediate precision, the lower limit of quantification (LLOQ) and risk were calculated, as seen in Fig. 4.

The relative bias of PLS model in ethanol solution was assessed in each concentration level (Fig. 4a). It gives information on systematic error. The highest value at 1 mg/mL was that for CR region (50%). The relative bias at low concentration level decreased using SCOT or FCOT region compared with ALL region. But if FCOT region was saturated absorption i.e. pure water, PLS model using FOT region which provided poor validation results due to little transmission information (data are not shown).

Precision is the closeness of agreement among measurements from multiple sampling of a homogeneous sample under the recommended conditions. It gives some information on random errors and it can be evaluated at two levels: repeatability and intermediate precision. As shown in Fig. 4b and c, precision is defined in terms of relative standard deviation (RSD%) values for repeatability and intermediate precision. The highest RSD% was the CR region in each concentration level. PLS model using SCOT or FCOT region provided lower RSD% than ALL region.

Fig. 4d shows the risk parameter in each frequency region. The risk of having measurements outside the acceptance limits is directly derived from the above tolerance interval, using the same estimates and t distribution in a different manner, i.e. by computing the probability to be above the upper acceptance limit plus the probability of being below the lower acceptance limit instead of



Fig. 1. Raw spectra of citric acid in ethanol/water solvent with different proportions of water.



Fig. 2. Effect of number of PLS latent factors on PRESS values in each solvent.



Fig. 3. The results of calibration and external validation of PLS model in each region.

computing the interval where it is expected to observe β % of the future measurements. The same result indicated that PLS model using SCOT or FCOT region provided lower risk value than ALL region.

It is notable that the above same results appeared in each solvent. These results provided a guideline for variables selection region based on the quantification characteristic of each frequency region. In general, FCOT or CR region was still preferred as variables selection region in most time, but if saturated absorption occurs in FCOT or CR region, the validation methodology parameters shows SCOT region is a better choice because of the good transmission information.

4.4. Validation of model performance in SCOT or ALL region in various solvents

Fig. 5 shows the validation result of SCOT or ALL region in various solvents. The absorption intensity is known to increase due to the increase of water proportion in the solvent system. The result pointed out that the validation methodology parameters value increase occurs, i.e. LLOQ, repeatability, intermediate precision and risk. Taking 1 mg/mL concentration level of ALL region in various solvents as an example, the repeatability differed from 11.4% to 67.11%, respectively, the intermediate precision differed from 11.4% to 71.88%, and the risk differed from 44.86% to 90.46%. In addition, it should be noted that saturated absorption of FCOT or CR region resulted in poor quantitative results, as shown in Table 2. Even a linear regression model is not fitted for FCOT or CR region.

Table 2 presents the LOQ results of each frequency region in various solvents. The limits of quantization are obtained by calculating the smallest and highest concentration beyond which the accuracy limits or β -expectation limits go beyond the acceptance limits. The result illustrated that the stronger the absorption intensity was, the larger the LLOQ was. It was observed that LLOQ was the worst in CR region. The value exceeded the experimental



Fig. 4. Accuracy profile of PLS model in ethanol solution and in each region.



Fig. 5. ICH Q2(R1) validation criteria of PLS calibration model in SCOT and ALL region.

Table 2 Limit of quantization parameter in each solvent of each frequency region.

Solvents	LLOQ-ULOQ (mg/mL)				
	ALL	SCOT	FCOT	CR	
Ethanol	4.578-9.680	2.178-9.680	1.758-9.680	8.898-9.680	
90% Ethanol	5.860-9.680	1.592-9.680	2.294-9.680	ND-ND	
70% Ethanol	7.541-9.680	5.761-9.680	ND-ND	ND-ND	
50% Ethanol	7.451-9.680	7.926-8.766	ND-ND	ND-ND	
30% Ethanol	6.768-7.310	8.309-8.405	ND-ND	ND-ND	
10% Ethanol	ND-ND	7.238-9.680	ND-ND	ND-ND	
Water	8.782-8.968	ND-ND	ND-ND	ND-ND	

ND: not detected.

concentration design so that it was not obtained in Table 2. In addition, the LLOQ result still demonstrated in SCOT region was the best choice for NIR model. Certainly, while FCOT region has good transmission information, i.e. ethanol, the LLOQ was 1.758 mg/ mL and lower than the model using SCOT region.

Some explanations have been given for this phenomenon between the absorption intensity and the model performance. From the perspective of mathematical matrix, absorption intensity of various solvents was the major interference due to the difficulty of information extraction. On the other hand, as viewed from physical chemistry, the increased proportion of water resulted in the formation of hydrogen bonding.

5. Conclusion

The relationship between absorption characteristics and quantitative contribution in each near-infrared (NIR) frequency range has been revealed.

An experimental protocol was performed using citric acid in each mixture of water and ethanol solvent. PLS model was carried out and the quantification contribution to each frequency range was clarified. The absorption characteristics indicated that saturated absorption existed in CR range and FCOT range. Chemometric indicators, i.e. *r*, RMSEC and RMSEP, have revealed that SCOT region plays an important role in the entire NIR spectral region. In addition, FCOT or CR region could provide the lowest RMSEC and RMSEP, only when FCOT or CR region has rich transmission information.

Furthermore, each model was further validated based on the fitness-for-purpose approach, called the accuracy profile. According to the methodology parameters, i.e. relative bias, repeatability, and intermediate precision, the result showed that FCOT or CR region was preferred as a variables selection region, but if saturated absorption occurred in FCOT or CR region, SCOT region was a better choice. The validation result indicated that the increase of water proportion in the solvent system seriously interfered with the model accuracy and precision. Finally, it was shown that this work can provide a guideline for variables selection in each NIR frequency range. Additional research is necessary to thoroughly evaluate each frequency region, such as different optical paths condition and extraction solution of natural product, in order to identify all its strengths and limitations.

Acknowledgments

Financial support of this work from the Ministry of Science and Technology of China Major Special Project "Significant Creation of New Drugs" (No: 2010ZX09502-002) and Beijing municipal government for university affiliated with Party Central Committee "Research of the TCM pilot production scaling up technology" were gratefully acknowledged. Authors also thank Moonen Francois for his assistance in software usage.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.molstruc.2012.03. 058.

References

- [1] M.T.J. Benito, C.B. Ojeda, F.S. Rojas, Appl. Spectrosc. Rev. 43 (2008) 452-484.
- [2] E. Rasanen, N. Sandler, J. Pharm. Pharmacol. 59 (2007) 147–159.
- [3] Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond, N. Jent, J. Pharm. Biomed. Anal. 44 (2007) 683–700.
- [4] J. Luypaert, D.L. Massart, Y.V. Heyden, Talanta 72 (2007) 865–883.
- [5] G. Reich, Adv. Drug Delivery Rev. 57 (2005) 1109-1143.
- [6] K.A. Bakeev, E.W. Ciurczak, Spectroscopy 18 (2003) 32-35
- [7] Z. Wu, B. Xu, M. Du, C. Sui, X. Shi, Y. Qiao, J. Pharm. Biomed. Anal. 62 (2012) 1.
- [8] B. Xu, Z. Wu, Z. Lin, Anal. Chim. Acta 720 (2012) 22.
- [9] Y.H. Lai, Y.N. Ni, S. Kokot, Vib. Spectrosc. 56 (2011) 154-160.
- [10] W.L. Li, L.H. Xing, Y. Cai, H.B. Qu, Vib. Spectrosc. 55 (2011) 58-64.
- [11] Y.H. Wu, Y. Zheng, Q.Q. Li, J. Lqbal, L.Y. Zhang, W.B. Zhang, Y.P. Du, Vib.
- Spectrosc. 55 (2011) 201–206. [12] R.M. Balabin, S.V. Smirnov, Anal. Chim. Acta 692 (2011) 63–72.
- [13] X.B. Zou, J.W. Zhao, M.J.W. Povey, M. Holmes, H.P. Mao, Anal. Chim. Acta 667 (2010) 14–32.
- [14] B. Zai, X. Shi, Y. Qiao, J. Chin. Mater. Med. 35 (2010) 699-702.
- [15] W.L. Li, H.B. Qu, J. Pharm. Biomed. Anal. 52 (2010) 425-431.
- [16] W.L. Li, L.H. Xing, L.M. Fang, J. Wang, H.B. Qu, J. Pharm. Biomed. Anal. 53 (2010) 350–358.
- [17] Y. Mohri, Y. Sakata, M. Otsuka, J. Near Infrared Spectrosc. 17 (2009) 89-100.
- [18] T. Lillhonga, J. Grunwald, P. Geladi, J. Near Infrared Spectrosc. 17 (2009) 275-287.
- [19] C.C. Lau, C.O. Chan, F.T. Chau, D.K.W. Mok, J. Chromatogra. A 1216 (2009) 2130–2135.
- [20] A. Boubetra, F. Le Nestour, C. Allaert, M. Feinberg, Appl. Environ. Microbiol. 77 (2011) 3360–3367.
- [21] E. Ziemons, J. Mantanus, P. Lebrun, E. Rozet, B. Evrard, P. Hubert, J. Pharm. Biomed. Anal. 53 (2010) 510–516.
- [22] J. Mantanus, E. Ziemons, P. Lebrun, E. Rozet, R. Klinkenberg, B. Streel, B. Evrard, P. Hubert, Talanta 80 (2010) 1750–1757.
- [23] J. Mantanus, E. Ziemons, P. Lebrun, E. Rozet, R. Klinkenberg, B. Streel, B. Evrard, P. Hubert, Anal. Chim. Acta 642 (2009) 186–192.
- [24] A. Peinado, J. Hammond, A. Scott, J. Pharm. Biomed. Anal. 54 (2011) 13-20.
- [25] C. Bodson, E. Rozet, E. Ziemons, B. Evrard, P. Hubert, L. Delattre, J. Pharm. Biomed. Anal. 45 (2007) 356–361.
- [26] A.G. Gonzalez, M.A. Herrador, Anal. Chem. 26 (2007) 227-238.