Asian Journal of Chemistry; Vol. 27, No. 1 (2015), 335-339

ASIAN JOURNAL OF CHEMISTRY

http://dx.doi.org/10.14233/ajchem.2015.17888

Determination of Total Ester Content in Chinese Liquor by Combining Near-Infrared Spectroscopy and Wavelet-Based Calibration

C. TAN[1,3,*], H. CHEN[2], T. WU[1], Z. LIN[1] and L. WANG[1]

[1]Department of Chemistry and Chemical Engineering and Key Lab of Process Analysis and Control of Sichuan Universities, Yibin University, Yibin 644007, P.R. China
[2]Hospital, Yibin University, Yibin 644007, P.R. China
[3]Computational Physics Key Laboratory of Sichuan Province, Yibin University, Yibin 644007, P.R. China

*Corresponding author: Tel/Fax: +86 831 3551080; E-mail: chaotan1112@163.com

Received: 25 April 2014; *Accepted*: 1 July 2014; *Published online*: 26 December 2014; AJC-16566

Chinese liquor is one of the famous distilled spirits in China. How to accurately quantify total ester content of liquor is a problem. The feasibility of combining near-infrared spectroscopy with a calibration model for total ester content quantization is investigated. One hundred and thirty-seven samples of commercial bottled liquors were used for experiment. A new calibration procedure called reconstructed partial least squares, which is a combination of partial least squares, wavelet transform and mutual information, was developed. It is actually partial least squares modeling in reconstructed original domain coupled with mutual information-induced variable selection in wavelet domain. Three kinds of calibration procedure were used for comparison. It concluded that, compared to the reference methods, reconstructed partial least squares can produce better models without increased complexity for an end-user. Even if the proposed reconstructed partial least squares is only be used to determine the total ester content, it can be a potential tool for near-infrared analysis of other complex samples.

Keywords: Ester, Liquor, Near-infrared, Wavelet, Mutual information.

INTRODUCTION

Chinese liquor is one of the famous distilled spirits in the world and has been consumed for centuries[1]. As a complex mixture, it consists of hundreds of flavor compounds present different concentrations. Among others, various kinds of esters are the most components and therefore the total ester content (TEC) is one of the most important physicochemical indexes of liquor as the price of liquor is closely related to its amount of esters[2,3]. Over the years, the test of Chinese liquor mainly includes sensory test and chromatographic analysis. The former relies on person's sense of taste and smell to evaluate integrally liquor flavor and is difficult to ensure scientific and objective results. Chromatographic analysis is often a time-consuming, laborious and inexpensive operation[4]. An ideal method for the determination of chemical composition in liquor should be non-invasive, non-destructive and rapid. The application of spectroscopic techniques in liquor analysis has developed considerably[4-9]. Specially, near-infrared (NIR) spectroscopy, *i.e.*, the electromagnetic spectrum between 750 and 2,500 nm, offers many advantages including its speed, the absence of or reduced need for sample pretreatment and the absence of the use of chemicals. Also, it presents smaller absorption bands in at least 1 order of magnitude for each successive overtone, allows the use of more concentrated samples and longer optical paths than those used in middle-infrared spectroscopy.

In the quantitative application of NIR technique, the importance of reliable calibration model is indisputable for prediction of composition contents or characterization of unknown samples and often determines its availability[10-13]. Thus, it is crucial to study the methods of model construction. In recent years, great effort has been made to improve the predictive model. These works include spectral pretreatment techniques, variable selection methods and different robust strategies. Even if many methods have been developed, the most commonly used multivariate calibration method is undoubtedly partial least squares (PLS)[14]. The partial least squares can handle datasets even when the number of variables is much larger than the number of samples. Also, partial least squares can model weak non-linearity by using a few extra latent variables. However, care must be taken because training a partial least squares model with too many factors will tend to over-fit untrained data. Furthermore, in some situations it can be an advantage to reduce the number of variables in order to, among others, obtain (a) improvement of the model predictions, (b) a better interpretation or (c) lower measurement costs. Mutual information (MI), a measure from information theory, has proved to be a powerful criterion measuring variable dependences and been successfully used in variable selection[15,16]. It has the unique advantage to be model-dependent and nonlinearity. Model-independent means that no assumption is made about the model that will be used while nonlinearity means that the mutual information measures the non-linear relationships between variables.

In recent years, to improve understanding and prediction of calibration models, the application of wavelet transform as a pre-processing step prior to any type of modeling has attracted widespread attention[17]. The basis for such approaches is the concept of multi-resolution, the ability to separate a signal according to frequency, is one of the important advantages offered by the wavelet transform[18]. Chemical spectra measured from spectroscopic instruments are in fact information that can considered as a signal composed of various frequency components, not obviously remarkable in the original domain. When one applies calibration methods to raw spectra in general, the final model is based on the highest resolution level only. This means that it is sometimes difficult to detect dependencies between the spectrum space and, *e.g.*, the concentration space of a compound which originate at different scales. By using wavelet and multi-resolution analysis, it is possible to capture the information at the different scales separately and to investigate the contribution of each scale to the final model. More specifically, wavelet can transform the original spectrum into the wavelet domain. The corresponding information can be represented by the wavelet coefficients. Due to the built-in effect of information concentration, there are many wavelet coefficients with very small amplitude, which can be regarded as uninformative. By certain operations, spectral information can be concentrated into a small number of variables. Furthermore, since wavelet transform decompose the spectra according to scale while retaining wavelength information, features such as overlapping bands, noise or a variable baseline are often separated into different variables in the wavelet domain. Thus, it becomes easier to identify relevant features. Several methods have already been proposed for selecting the relevant variables in wavelet domain for partial least squares (PLS). The wavelet transform itself does not produce a compressed version of the original. Thus, it is often used by coupling with a variable selection approach in order to eliminate the wavelet coefficients that do not hold valuable information.

In the present work, based on partial least squares, wavelet transform and mutual information-induced variable selection, a simple and effective calibration procedure is developed for analyzing the total ester content of liquor by NIR spectroscopy. It is named reconstructed partial least squares, *i.e.*, partial least squares modeling in reconstructed original domain coupled with mutual information-induced variable selection in wavelet domain. In reconstructed partial least squares, the original spectra of the training set are first decomposed into a set of wavelet representations at a depth of scale (level) by action of the wavelet prism transform. Then, the mutual information value between each wavelet coefficient variable and the response variable is calculated, resulting in a mutual information spectrum; by retaining a subset set of coefficients with higher mutual information, an update training set consisting of wavelet coefficients is obtained and can be reconstructed/converted back to the original domain. Based on this, a partial least square (PLS) model can be constructed and optimized. The optimal wavelet and decomposition level are determined by experiment. Three kinds of calibration procedure/methods, *i.e.*, conventional full-spectrum partial least squares in original domain (FPLS), partial least squares in original domain coupled with mutual information-induced variable selection (OPLS) and direct partial least squares in mutual information-based wavelet coefficients (WPLS), were used for comparison purpose.

EXPERIMENTAL

Samples collection and reference analysis: One hundred and thirty-seven samples of commercial bottled liquors belonging to seven brands were purchased in local stores of west China. Each sample was analyzed using reference methods for total ester content. Reference analyses were in accordance with the Official Methods of Analysis for Chinese liquor (GB/T 10245.5-1989). All analyses were done in duplicate.

Instrument and spectra collection: Liquor bottles were opened and subsamples were scanned on in transmission mode (4000-12000 cm^{-1}) using a near-infrared spectrometer coupled with a automated transmission module (Antaris II, Thermo fisher, USA). Spectral data collection was made using Vision software-TQ Analyst. Samples were scanned in a rectangular curette with a 1 mm path length and temperature equilibrated at 25 °C for 2 min in the instrument before scanning. Spectral data were stored as the logarithm of the reciprocal of transmittance [log (1/T)], at 4 nm intervals. The spectrum of each sample was the average of 32 successive scans, resulting in 2074 data points/variables for each spectrum. The absorbance in the region of 8000-12000 cm^{-1} was very weak.

Mutual information-based variable selection: Variable selection, also called "feature" or "wavelength" selection when applied to spectroscopic data. The goal of variable selection is to identify a subset of spectral variables that produce the smallest possible errors when used to perform operations such as making quantitative determinations or discriminating between dissimilar samples. An optimal way to do variable selection is to try all combinations of variables and select the best ones. This sounds simple, but is, in practice, impossible for a number of reasons. Furthermore, even if it is possible to test all combinations of variables, the risk of over-fitting would be detrimental unless the number of samples was much higher than the number of combinations of variables. Among others, for these reasons, a number of variable selection methods have been developed which try to find a good set of variables rather than the optimal set of variables.

The mutual information (MI) is a statistical measure of arbitrary dependencies between two variables. Mutual information is based on Shannon's information theory and is very useful in a prediction context. Unlike other parametric estimators, such as the correlation, the mutual information does not make any assumption about what type of relation could exist between the variables. It can thus be used in a wide range of contexts, including for the selection of variables. More specifically, mutual information enables to assess the amount of information contained in the candidate variable x that can be used to predict the response variable y. Mutual information is zero only if there is no dependence between the two variables. However, the exact measure of the mutual information is not possible in practice. Indeed, the exact measure of mutual information is possible only when the probability density functions (PDF) of x and y are known. Often, the PDF are not known (we only know a few samples, not the distribution of samples) and must be estimated. mutual information estimators have thus been developed to compute an approximation of the mutual information between x and y, *i.e.*, I(x,y), in the finite sample case. Histograms and kernels may be used for that purpose. More information about the mutual information concepts and calibrations are available[19,20].

Wavelet transformation: The wavelet transformation (WT) is a multi-resolutional signal processing tool that has found several applications in de-noising, feature extraction and signal compression. Wavelets, as often used as the discrete wavelet transform (DWT), have been shown to be highly effective in improving the performance of calibration type problems in many fields of NIR spectroscopy. The DWT of a spectrum can be obtained in a fast manner by using a filter bank. The basic structure of the filter bank consists of a pair of low-pass and high-pass filters, followed by a down-sampling operation, which discards one in every two points of the filtering outcome. The down-sampled output of the low-pass filter, termed "approximation", is a smoothed version of the original spectrum at a coarser resolution. The down-sampled output of the high-pass filter, termed "detail", mainly correspond to high frequency noise, as well as sharp features of the original spectrum, such as narrow peaks. Such an operation can be reapplied to the approximation coefficients up to the set decomposition levels. The result of the wavelet transform comprises the final approximation, as well as the detail obtained along the entire filter bank. With a slight abuse of language, this result will be hence forth termed "wavelet coefficients". Due to the finite length of the filters employed in the filter bank, approximation or detail coefficient in each level corresponds to a reduced range of wavelengths within the spectrum. Unlike the Fourier transform, the wavelet transform can use a variety of different basis functions with different properties. Once the spectrum is transformed into the wavelet coefficients, it is convenient to perform variable compression/selection. By utilizing inverse wavelet transform, the compressed/treated coefficients can be converted back to its original domain for constructing model. The theory of wavelets is well established in chemometrics[21,22].

Wavelet-mutual information-reconstructed partial least squares: Here, we use the newly developed calibration method, *i.e.*, reconstructed partial least squares (RPLS), which is based on the wavelet transform and mutual information-induced variable selection.

RESULTS AND DISCUSSION

All 137 samples were split into two subsets: The training (calibration) set and the test set. Considering the fact that the evaluation of a model is valid only if the test set contains information similar to the training set, a special scheme was used to avoid possible bias in subset selection. That is, an algorithm of representative sample selection named SPXY[23] (sample set partitioning based on joint x-y distances) for sorting all samples, followed by an alternative re-sampling with a ratio of 2/1. That is, one-third of samples were selected into the test set while the other 2/3 of the samples constituted the training set. As a result, the training set and the test set consist of 91 and 46 samples, respectively. Basically, the range of the response (total ester content) in the test set covers the range in the calibration set.

Based on the determination of total ester content, the performances of the proposed reconstructed partial least squares and three other reference methods (FPLS, OPLS and WPLS) are compared. The wavelet chosen for this study is 'db2' because it has a shape suitable for describing infrared signal. Also, its advantages have been confirmed by some researches. When using wavelet transform, which wavelet and how many coefficients (variables) should be retained is of great importance and also depends heavily on the task. As far as multivariate calibration is concerned, one model designer focuses on maintaining as much valuable information for predicting the response as possible. In order to give an insight of how many variables (wavelet coefficients) is enough, Fig. 1 gives the RMSECV (root-mean-squared error of cross validation) values associated with OPLS, reconstructed partial least squares and WPLS models as a function of the number of variables. Despite significant fluctuation, it can be seen in Fig. 1 that using 100 variables is enough and reasonable since using more variables is useless for reducing the RMSECV of each kind of model. So, the number of variables is fixed at 100 for quantitative analysis. The decomposition level is also crucial in these methods related to wavelet transform. The level is optimized by using the (RMSECV) as the measure. The RMSECV values of both reconstructed partial least squares and WPLS models *versus* decomposition level are shown in Fig. 2, implying that the optimal decomposition level is six for both methods.

In order to analyze the advantages from wavelet transform, as an example, Fig. 3 represents the original spectra and the corresponding reconstructed spectra after selecting 100

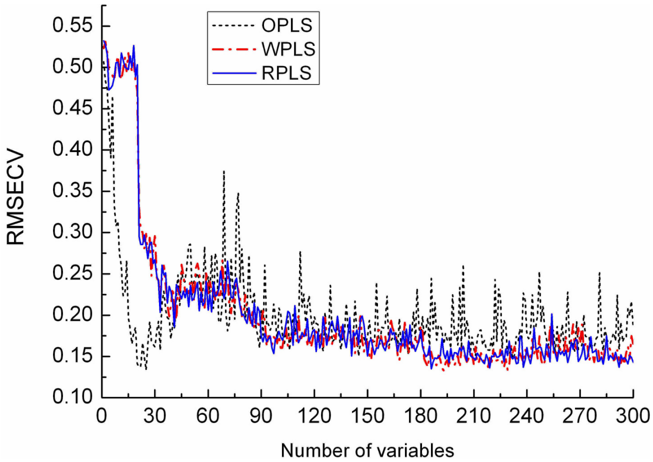

Fig. 1. RMSECV values associated with OPLS, RPLS and WPLS models as a function of the number of variables

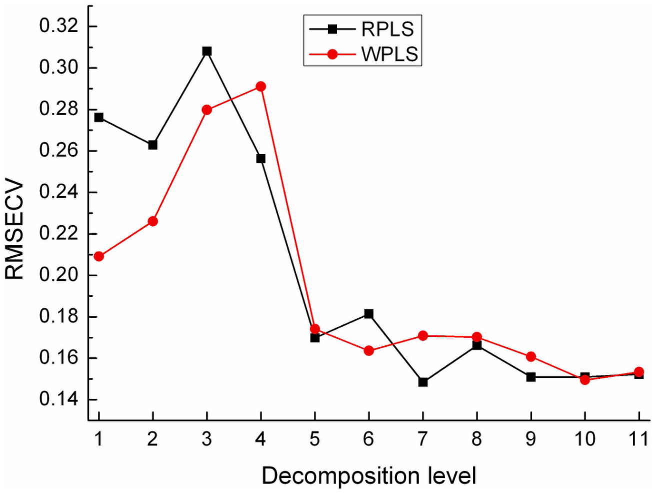

Fig. 2. RMSECV values of RPLS and WPLS models *versus* decomposition level

variables in wavelet domain. As seen in Fig. 3, the reconstructed spectra highlight some important features, while ignoring some uninformative parts. Specifically, only the variables in the region of 4000-5500 cm^{-1} exhibit significant absorption strength, which indicating that wavelet transform coupled with mutual information-induced variable selection have the role of information collection. The reconstructed partial least squares seems to be contrary to those cases, which intend to remove an entire scale for certain purpose and therefore has the risk of losing an informative component, resulting in information leakage. Table-1 summarizes the distribution of the selected 100 wavelet coefficients in various scales/levels. Clearly, the selected wavelet coefficients at the D1-D3 levels account for about 90 % of the total number and the approximation coefficients are never selected. Based on this, it is possible to prepare a so-called scalogram (not shown here), in which each tile represents the area covered by a wavelet basis function in time-frequency domain. Also, it is convenient to observe the positions of the selected wavelet coefficients or to locate the most important regions in the original spectra influenced by these coefficients.

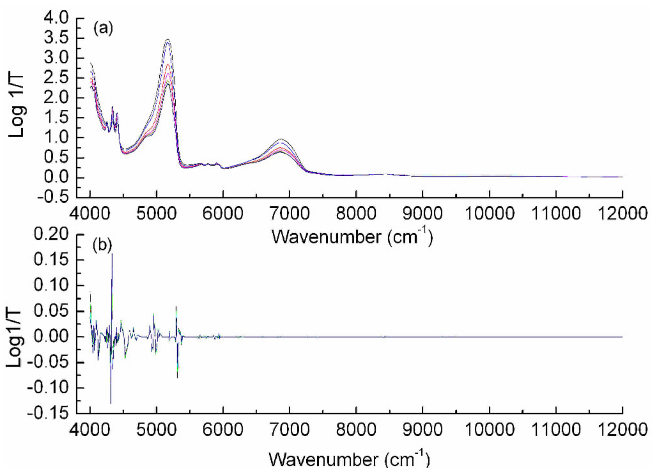

Fig. 3. Original spectra and the corresponding reconstructed spectra after selecting 100 variables in wavelet domain

The performances of the different models are evaluated in terms of root mean square error of calibration (RMSEC) and the root mean square error of prediction (RMSEP), corresponding to the training set and the test set. Table-2 summarizes the performance comparison of four kinds of optimal models and the mean of latent variables (Lvs) of 100 runs. Each time, the optimal Lvs is corresponding to the lowest RMSECV. It is clear in Table-2 that the OPLS model has a lower RMSEC/RMSECP compared to the FPLS model. Both reconstructed partial least squares and WPLS enable to produce better models than OPLS. Also, even if the WPLS model shows lower RMSEC values than the reconstructed partial least squares model, it has a higher RMSEP values, implying that a possible overfitting problem exists for the WPLS model. It seems that the reconstructed partial least squares model hold better generalization ability. For further comparison of the prediction results of OPLS, reconstructed partial least squares and WPLS, Fig. 4 illustrate the actual/measured *vs.* predicted points for the test set. In this plot, the data points will fall on the diagonal if the model fits the data perfectly. From deviation from the diagonal, it is able to visually analyze the performance of models. As seen in these plots, the points associated to either reconstructed partial least squares or WPLS models fall on the diagonal more compactly than OPLS, confirming their better predictive ability. It could be explained that partial least squares is actually a linear method, although it may handle mild non-linearity by including extra latent variables into the model. The number of esters in liquors, *i.e.*, total esters, is a mixture containing a variety of esters compositions. Furthermore, Chinese liquor is a solid fermented product containing many other substances such as fusel oil, acids and other organics. A NIR spectrum is composed of overtones and combinations of fundamental vibrations of corresponding organic groups from the mid-infrared. Therefore, the relationship between the NIR spectra and total ester content is maybe complicated which is more inclined to non-linear rather than linear. It seems that reconstructed partial least squares combine the advantages from wavelet multi-resolution analysis and mutual information-induced variable selection. So, reconstructed partial least squares can construct superior model to the reference methods.

TABLE-2
A COMPARISON OF THE MEAN RESULTS OF FOUR CALIBRATION METHODS FROM 100 RUNS

Index	FPLS	OPLS	WPLS	RPLS
RMSEC[a]	0.5105	0.3833	0.3045	0.3537
RMSEP[b]	0.5283	0.4711	0.4226	0.3659
LVs[c]	16	13	14	14

TABLE-1
DISTRIBUTION OF SELECTED 100 WAVELET COEFFICIENTS IN VARIOUS SCALES/LEVELS

Level	A	D6	D5	D4	D3	D2	D1
Selected coefficients	0	1	1	9	23	33	33
Total coefficients	35	35	67	132	261	520	1038
Percentage	0	1	1	9	23	33	33

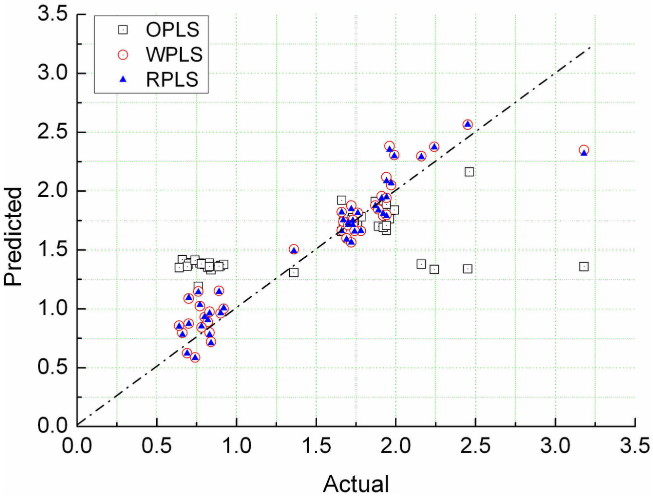

Fig. 4. Scatter plot of predicted *vs.* actual value of the test set for OPLS, RPLS and WPLS models

In fact, when a calibration/regression method is applied to raw spectra in the wavenumber domain, the final model is based on the highest resolution level only. This means that is it is difficult to detect dependencies between the spectrum spaces, *e.g.*, the concentration space of total ester content, which maybe originates from different scales associated to the peak shape, as introduced above. By using wavelet-based regression, it is possible to make full use of the information distributed in various frequency bands. It may be just because of this that reconstructed partial least squares performs well. In fact, the near-infrared spectra are inherently multiscale. For example, noise usually locates in high-frequency region while background and drifts often appear at the lowest frequency ranges, *i.e.*, approximation coefficients.

Compared to partial least squares, although the proposed reconstructed partial least squares method differs in the modeling process, the modeling results were identical. That is, a partial least squares-based model can actually represented as a b-coefficient vector with the same length as the number of spectral channels. What should be stressed is that, to predict the response (total ester content) of a new sample, the only requirement is to multiply its spectrum with b-coefficient vector. In other words, once the reconstructed partial least squares model is established, it can be performed for prediction within the same time. It is only the process of constructing a model that involves variable selection and wavelet transform. From the perspective of application, the reconstructed partial least squares does not increase any complexity of the calibration while enhancing its performance.

Conclusion

It is verified that NIR spectroscopy coupled with the developed methods, *i.e.*, reconstructed partial least squares, is a suitable tool for quantification of the most important liquor parameter, *i.e.*, total ester content. This method combines the advantage of wavelet multi-resolution with mutual information for capturing the non-linear relationship. So, it enables to produce better models without increased complexity for an end-user, compared to three reference methods. Even if the proposed reconstructed partial least squares is only be used to determine the total ester content, it can be a potential tool for near-infrared analysis of other complex samples. The idea of considering the inherent local features of spectral signals in both time and frequency domains, instead of in a single domain, may be of great potential for developing future calibration methods.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (21375118), the Applied Basic Research Programs of Science and Technology Department of Sichuan Province of China (2013JY0101), Scientific Research Foundation of Sichuan Provincial Education Department of China (12ZA201, 13ZB0300),Yibin Municipal Innovation Foundation (2013GY018) and Innovative Research and Teaching Team Program of Yibin University (Cx201104).

REFERENCES

1. C.W. Li, J.P. Wei, Q. Zhou and S.Q. Sun, *J. Mol. Struct.*, **883-884**, 99 (2008).
2. F. Shen, Y. Ying, B. Li, Y. Zheng and J. Hu, *Food Res. Int.*, **44**, 1521 (2011).
3. B.Z. Peng, M.H. Long, T.L. Yue and Y.H. Yuan, *Transactions of the CSAE*, **22**, 216 (2005).
4. D. Cozzolino, G. Cowey, K.A. Lattey, P. Godden, W.U. Cynkar, R.G. Dambergs, L. Janik and M. Gishen, *Anal. Bioanal. Chem.*, **391**, 975 (2008).
5. A. Alcázar, J.M. Jurado, A. Palacios-Morillo, F. de Pablos and M.J. Martín, *Food Contr.*, **23**, 258 (2012).
6. X.L. Mo, W.L. Fan and Y. Xu, *J. Inst. Brew.*, **115**, 300 (2009).
7. X.Y. Niu, F. Shen, Y.F. Yu, Z. Yan, K. Xu, H.Y. Yu and Y. Ying, *J. Agric. Food Chem.*, **56**, 7271 (2008).
8. F. Shen, F.Z. Li, D. Liu, H. Xu, Y. Ying and B. Li, *Food Contr.*, **25**, 458 (2012).
9. H.Y. Yu, X.Y. Niu, H.J. Lin, Y.B. Ying, B.B. Li and X.X. Pan, *Food Chem.*, **113**, 291 (2009).
10. F. Liu, X.J. Ye, Y. He and L. Wang, *J. Food Eng.*, **93**, 127 (2009).
11. Y. Yu, Y.B. Ying, X.P. Fu and H.S. Lu, *J. Food Qual.*, **29**, 339 (2006).
12. R.A. Viscarra Rossel, *J. Near Infrared Spectrosc.*, **15**, 39 (2007).
13. H. Shinzawa, J.H. Jiang, P. Ritthiruangdej and Y. Ozaki, *J. Chemometr.*, **20**, 436 (2006).
14. A. Höskuldsson, *Chemom. Intell. Lab. Syst.*, **55**, 23 (2001).
15. N. Benoudjit, D. Francois, M. Meurens and M. Verleysen, *Chemom. Intell. Lab. Syst.*, **74**, 243 (2004).
16. A. Dionisio, R. Menezes and D.A. Mendes, *Physica A*, **344**, 326 (2004).
17. L. Gao and S.X. Ren, *Spectrochim. Acta A*, **71**, 959 (2008).
18. H.W. Tan and S.D. Brown, *Anal. Chim. Acta*, **490**, 291 (2003).
19. F. Rossi, A. Lendasse, D. Francois, V. Wertz and M. Verleysen, *Chemom. Intell. Lab. Syst.*, **80**, 215 (2006).
20. C. Tan and M.L. Li, *Spectrochim. Acta A*, **71**, 1266 (2008).
21. F. Ehrentreich, *Anal. Bioanal. Chem.*, **372**, 115 (2002).
22. M. Jing, W.S. Cai and X.G. Shao, *Chemom. Intell. Lab. Syst.*, **100**, 22 (2010).
23. R.K.H. Galváo, M.C.U. Araújo, G.E. José, M.J.C. Pontes, E.C. Silva and T. Saldanha, *Talanta*, **67**, 736 (2005).