Available online at www.sciencedirect.com
ScienceDirect
Chemometrics and Intelligent Laboratory Systems 90 (2008) 188–194
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra

Wensheng Cai, Yankun Li, Xueguang Shao *

Research Center for Analytical Sciences, College of Chemistry, Nankai University, Tianjin, 300071, PR China

Received 3 July 2007; received in revised form 28 September 2007; accepted 2 October 2007
Available online 11 October 2007

Abstract

Variable (or wavelength) selection plays an important role in the quantitative analysis of near-infrared (NIR) spectra. A modified method of uninformative variable elimination (UVE) was proposed for variable selection in NIR spectral modeling based on the principle of Monte Carlo (MC) and UVE. The method builds a large number of models with randomly selected calibration samples at first, and then each variable is evaluated with a stability of the corresponding coefficients in these models. Variables with poor stability are known as uninformative variable and eliminated. The performance of the proposed method is compared with UVE-PLS and conventional PLS for modeling the NIR data sets of tobacco samples. Results show that the proposed method is able to select important wavelengths from the NIR spectra, and makes the prediction more robust and accurate in quantitative analysis. Furthermore, if wavelet compression is combined with the method, more parsimonious and efficient model can be obtained.

Keywords: Near-infrared spectroscopy; Multivariate calibration; Monte Carlo (MC); Uninformative variable elimination (UVE)

1. Introduction

Chemometrical methods have become hot points and been widely applied in analytical chemistry in recent years. Especially, multivariate calibration methods have been playing indispensable roles in near-infrared (NIR) spectral quantitative analysis. Many multivariate calibration methods, such as principal component regression (PCR) [1], partial least squares (PLS) [2,3], artificial neural network (ANN) [4] and support vector regression (SVR), [5,6] are widely used to build the quantitative model in NIR spectral analysis.

The quality of a multivariate calibration model depends, among others, on the quality of both objects and variables. NIR spectra are typically consisted of broad, weak, non-specific and overlapped bands [7]. Moreover, NIR data sets may have thousands of wavelengths and hundreds or thousands of samples. Therefore, there may be some irrelevant variables for multivariate calibration. Elimination of uninformative variables can predigest calibration modeling and improve prediction results in terms of accuracy and robustness. Better quantitative calibration models may be obtained by selecting characteristic wavelengths including sample-specific or component-specific information instead of the full-spectrum. For this aim, several methods have been developed, such as the correlation coefficient method [8], interval PLS (iPLS) [9,10], stepwise regression analysis (SRA) [11], uninformative variable elimination (UVE) [12,13] and genetic algorithms (GA) [14,15].

Uninformative variable elimination by PLS (UVE-PLS) is a method for variable selection based on an analysis of regression coefficients of PLS [12]. The method consists of evaluating the reliability of each variable in the model through a variable selection criterion, i.e., the stability of each variable, and eliminating the uninformative variables. UVE-PLS method has been widely applied in analytical chemistry, and satisfactory prediction results are obtained comparing with many other methods of wavelengths selection [12,16–18]. However, in the course of acquiring stability values and the criterion, a leave-one-out jackknifing is adopted and extra random variables with the equal size of the spectra are needed. The procedure is time-consuming when it meets a large data set.

* Corresponding author. Tel.: +86 22 23503430; fax: +86 22 23502458.
E-mail address: xshao@nankai.edu.cn (X. Shao).

0169-7439/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2007.10.001

In this work, in virtue of the Monte Carlo (MC) technique and the stability criterion in UVE method, a modified method of UVE for variable selection is proposed and named as MC-UVE. The method builds a large number of models with randomly selected calibration samples at first, then by using the coefficients of these models, each variable is evaluated with a stability of the corresponding coefficient. Multiple models with different calibration subsets produced by the MC technique may effectively identify and encode more aspects of the relationship between independent and dependent variables than will a single model. Therefore, it can be expected to decrease the risk of over-fitting [19,20], and accordingly, evaluate the reliability of each variable more reasonably. Moreover, the MC-UVE is faster in computation than UVE for large data sets because extra random variables are not used. Calibration of NIR spectra and the routine ingredients (sugar compounds and nicotine) of tobacco samples were investigated with the proposed method. It is found that the accuracy of the predicted results obtained with the selected informative variables by MC-UVE is equivalent as or slightly better than that of the conventional PLS method obtained with full-spectrum data and that of the UVE-PLS method. Furthermore, Wavelet transform (WT) has been found to be a very efficient tool in processing analytical signals [21,22]. With WT technique, NIR spectra can be represented by only small amount of wavelet coefficients [23–26]. Therefore, selection of the informative wavelet coefficients with the MC-UVE method, named as WT-MC-UVE, was also investigated, and it was found that better results and more timesaving procedure can be obtained by fewer variables compared with the MC-UVE method.

2. Theory and algorithm

2.1. Uninformative variable elimination by PLS (UVE-PLS)

In linear least squares models, the predictions ŷ are computed by the equation:

$$\hat{y} = \mathbf{X}\boldsymbol{\beta} + b_0 \tag{1}$$

Where $\boldsymbol{X}$ is an $n \times p$ matrix containing p spectral responses of n samples, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients and b_0 is the model offset.

UVE-PLS method is presented in reference [12]. A regression coefficient matrix $\boldsymbol{\beta} = [\beta_1,\ldots,\beta_p]$ is calculated through a leave-one-out validation [12,13]. Because each coefficient β_j represents the contribution of the corresponding variable to the established model, the reliability of each variable j can be quantitatively measured by the stability defined as:

$$s_j = \text{mean}(\beta_j) / \text{std}(\beta_j) \quad j = 1 \ldots p \tag{2}$$

where mean(β_j) and std(β_j) are the mean and standard deviation of the regression coefficients of variable j. It is clear that, when the mean value of β_j is large and the standard deviation of β_j is small, the stability value is large. Therefore, the larger the stability, the more important the corresponding variable is. The variables whose stability is less than a threshold should be treated as uninformative and be eliminated.

In order to estimate a suitable *cutoff* threshold, an artificial random variable matrix $\boldsymbol{N}$ ($n \times p$) with very small amplitude (e.g. 10^{-11}) is added to the original data to compute their stability [13]. It is obvious that any variable whose stability is less than that of random variables should be known as uninformative and be eliminated. In practice, the *cutoff* threshold is generally defined by:

$$cutoff = k \times \max(\text{abs}(s_{\text{noise}})) \tag{3}$$

where k is an arbitrary value, e.g. 0.7 or 0.9.

2.2. Monte Carlo method

The Monte Carlo method, or called as the random imitative method, is a powerful and widely used technique for analyzing complex (multi-variable) problems. It is a stochastic technique, which is based on the use of random numbers and probability statistics to investigate problems. It has been applied in many fields such as statistical tests, optimization procedures, system analysis, and signal detection, etc. [27–29] In multiple regression analysis, Monte Carlo cross-validation (MCCV) is one of the most useful methods for modeling and prediction problems, which was first proposed by Picard and Cook [19] and its ability has been reported in several studies [30,31].

In this work, MC method is used in the procedure of acquiring stability of each variable. A large number of PLS models with different calibration samples selected by the MC technique are produced, then by using the regression coefficients of these models, the stability of the corresponding coefficient is calculated. The procedure has advantages of reducing dependence on single model and evaluating the reliability of each variable credibly to judge the remaining or rejection of them.

2.3. Monte Carlo combined with UVE (MC-UVE)

On the basis of the UVE and MC method, a combination of MC and UVE (MC-UVE) is used for variable selection in NIR spectral modeling. It uses the stability defined in UVE method to evaluate the reliability of each variable, but the stability values are obtained through the Monte Carlo method replacing the leave-one-out procedure in UVE. Moreover, instead of adding random noise variables to the original data matrix as in UVE method to estimate the *cutoff* threshold, the wavelengths to be selected are determined directly by their stability, which is more convenient. Then build PLS model by using the retained variables to predict unknown samples. The detailed procedures can be described as follows:

(1) All spectra (samples) are randomly divided into a training set, an assessing set and a prediction set. To ensure the concentration of training set covers all prediction samples, three samples of the highest and the lowest concentration are put into the training set manually.
(2) By the Monte Carlo technique, randomly select a certain amount (N_t) of samples from the training set as the training subset for constructing a PLS sub-model, and the

procedure is repeated M times. Then, a matrix of the PLS regression coefficients $\boldsymbol{\beta}$ ($M \times p$) are calculated [13], with which the stability of each variable s ($1 \times p$), is calculated by using the Eq. (2).

Following the principle of Monte Carlo technique, a large portion (e.g. 40–60%) of the training data should be set aside as a validation data set [20, 32]. In this work, N_t is set as a number of 50% the whole training set to construct PLS sub-models. M is set with 100, which is proved to be enough to assure a precise estimate of the stability.

(3) With the stability obtained above, a number (N_j) of the informative (stable) variables is selected for building the final PLS model, i.e., to rank the stability of all the variables from the highest to the lowest, and set the stability of the N_jth as the *cutoff* value. The variables whose stability is less than the *cutoff* are eliminated. The optimal value of N_j is discussed in the following section by using the prediction results of the assessing set.

(4) With the selected variables, build a PLS model by using the whole training set and predict the samples in the prediction set.

2.4. MC-UVE with wavelet transform (WT-MC-UVE)

WT has been found to be a very efficient tool in processing analytical signals, especially in compression of spectral data [25,33]. In this study, as a preprocessing tool, discrete wavelet transform (DWT) is used, which is a special case of WT that provides a compact representation of a signal in time and frequency. If the MC-UVE is applied to the wavelet coefficients, a more parsimonious PLS model can be obtained. Therefore, in the WT-MC-UVE method, wavelet coefficients are used to replace the raw spectra for variable selection.

3. Experimental and calculations

Two data sets were prepared for this work. One (data set 1) consists of 373 samples and nicotine content was measured for modeling. The other one (data set 2) consists of 288 samples and total sugar content was measured for modeling.

NIR spectra of tobacco lamina samples were measured on a Vector 22/N FT-NIR System (Bruker, Germany). Each NIR spectrum was recorded in the wavenumber range 4000–9000cm^{-1} (2500nm–1100nm) with the digitization interval ca. 4cm^{-1}. Each spectrum is composed of 1296 data points. Fig. 1 shows an example of the NIR spectra. The concentration of nicotine and sugar content in tobacco samples were measured on an Auto Analyzer III (Bran+Luebbe, Germany) following the procedures of the standard method.

For data set 1251 samples were used as training data set, 61 samples were used as assessing data set, and the other 61 samples were used as prediction data set. In the calculation of data set 2194 spectra were used as training data set, 47 samples were used as assessing data set, and the other 47 samples were used as prediction data set. In the comparison of MC-UVE-PLS, WT-MC-UVE-PLS, UVE-PLS and common PLS methods, the same training set and prediction set were used.

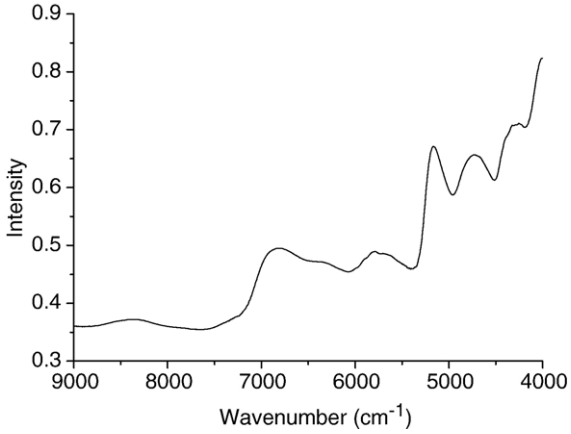

Fig. 1. An example of the measured NIR spectra of tobacco lamina samples.

On the other hand, in the calculation of wavelet transform, several wavelet filters, such as Daubechies, Symmlet, Coiflet, etc. and different decomposition scales are investigated. It is found that there is no significant difference between these filters and scales. Daubechies with vanishing moment 10, i.e., "db10" wavelet filter and scale = 9 are adopted in this work. In the calculations of WT-MC-UVE, the wavelet coefficients are taken as the input as described above.

4. Results and discussions

4.1. Determination of the principal factor number (n_f) for PLS modeling

The number of principal factor (n_f) of PLS is an important parameter in the modeling. Therefore, in this work, The parameter is determined with the root mean squared error of prediction (RMSEP) of the assessing set and the RMSEP of the calibration set in cross-validation (denoted by RMSECV). RMSEP is defined as

$$\mathrm{RMSEP} = \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right]^{1/2} \quad (4)$$

where $\hat{y}_i$ and y_i are the predicted and measured concentration of the ith sample, and n is the size of the assessing set. In the calculation of RMSECV, $\hat{y}_i$ is the predicted value in cross-validation, and n is the size of calibration set. Fig. 2(a) and (b) show the variation of RMSEP and RMSECV with the principal factor number of the three methods, i.e., PLS, UVE-PLS and MC-UVE-PLS method, for the two data sets, respectively. From the figures, it was clear that both RMSEP and RMSECV have a descending trend with the increase of the principal factor number, but the trend slowed down after $n_f > 10$. Therefore, Monte Carlo cross-validation with F test was used for confirming the suitable principal factor number, and the results show that 10–13 can be used. In order to make the model as less as complex and use an identical parameter in the three models, $n_f = 10$ was used further calculations.

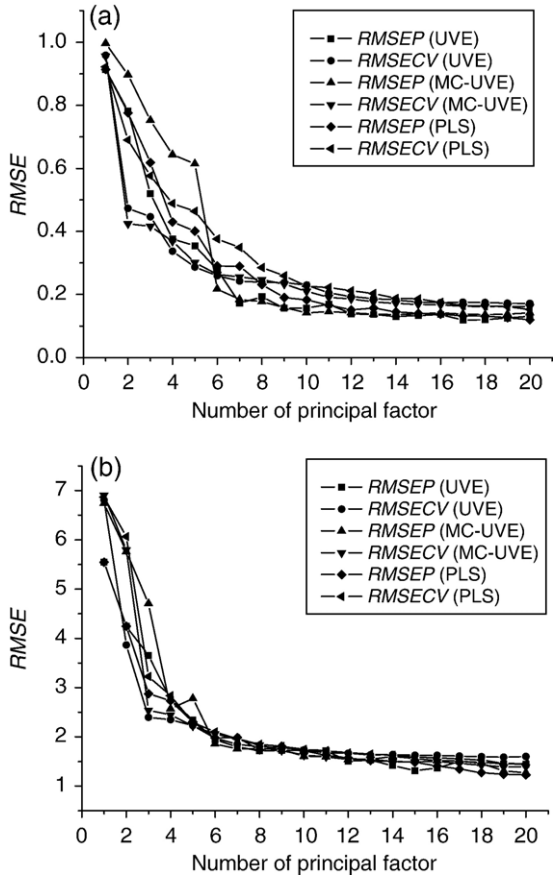

Fig. 2. Variation of RMSEP and RMSECV with the number of factors by UVE, MC-UVE and PLS methods for data set 1 (a) and data set 2 (b).

4.2. Variables selection for data set 1

Fig. 3(a) and (b) show the stability of each variable in the wavenumber 4000–9000cm^{-1} for the nicotine data set by UVE and MC-UVE method, respectively. In the figures, the dot lines show the cutoff, which is determined by $N_j = 100$ for the MC-UVE and by $k = 0.88$ for UVE method (it is an equivalent to $N_j = 109$). The vertical bar in Fig. 3(a) indicates the stability range of the added random noise. Variables whose stability lies within the dot lines will be eliminated, and the variables whose stability lies out of the dot lines are used for PLS calculation. With a comparison of the two figures in Fig. 3, it can be seen that the two curves are similar to each other. However, the positive peak around 5980cm^{-1} is enlarged by MC-UVE in Fig. 3(b). Therefore, the variables selected by UVE method are concentrated on two broad regions around 4450 and 6500cm^{-1}, and three narrow regions in the range of 4600– 4900cm^{-1}. However, more variables around 5980cm^{-1} and less variables in the range of 4600– 4900cm^{-1} are selected by MC-UVE.

To determine the number of retained variables (N_j) is the main aim of this study, which decides the stability and accuracy of the model. When the number of retained variables is too small, the robustness and accuracy of the model may be affected due to the loss of informative variables. On the contrary, if the number of retained variables is too large, uninformative variables may be contained in the model and make its performance poor. Therefore, the variation of the RMSEP of the assessing set with N_j is investigated. Fig. 4 shows the RMSEP obtained with N_j from 20 to 200 and a step of 20. For each N_j, a PLS model is developed and the model is then used to predict the assessing set. The mean value and the standard error (σ) of RMSEP through 30 repeated runs are shown in Fig. 4. It can be seen that, at the beginning, both the mean value and the standard error are large, then with the increase of N_j, both decrease sharply. Clearly, when N_j is 100, the lowest mean value of RMSEP is obtained. Then, when N_j is bigger than 100, with the increase of N_j, the mean value of RMSEP increases gradually with a little fluctuation. This indicates that with lesser variables, useful wavelengths cannot be completely included, so the quality of the model is bad. On the other hand, when more variables are used, irrelevant variables also affect the prediction results. Therefore, $N_j = 100$ is used for further study.

4.3. Variables selection for data set 2

The stabilities of each variable at wavenumber range 4000–9000cm^{-1} for sugar data set by UVE and MC-UVE methods are shown in Fig. 5(a) and (b), respectively. In the figures, the dot lines

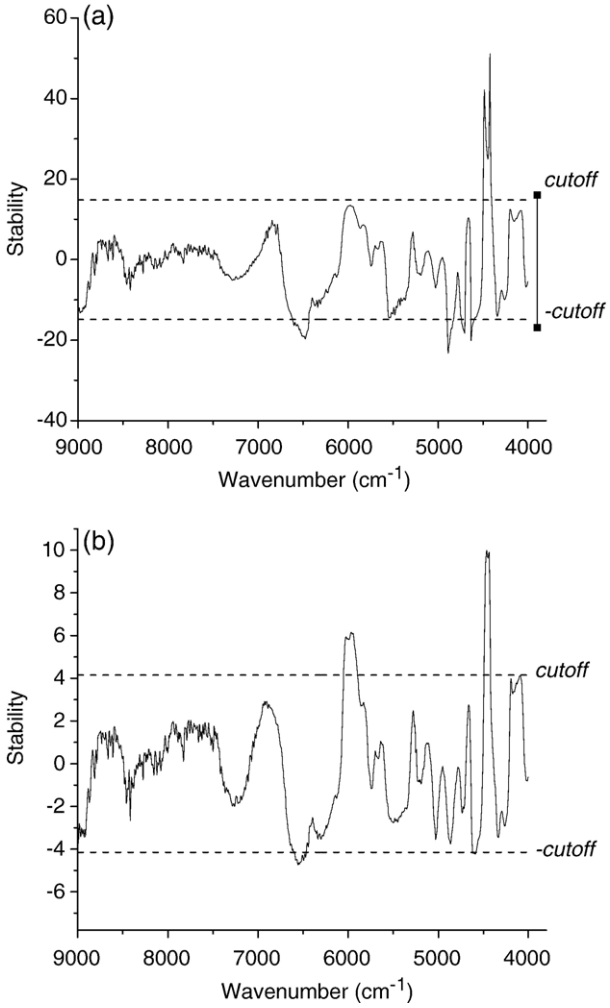

Fig. 3. The stability distribution of each variable for prediction of the nicotine by the UVE (a) and MC-UVE (b) method. The two dot lines indicate the lower and upper threshold. The vertical bar with "■" indicates the stability range of the added random noise.

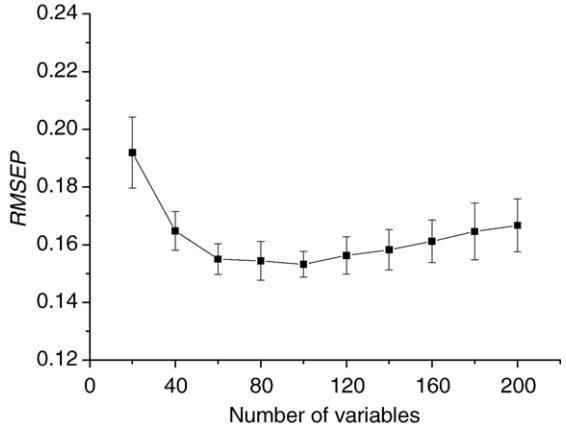

Fig. 4. Variation of RMSEPs with the number of selected wavelengths for data set 1. Standard deviation of 30 runs results is plotted as an error bar crossing the mean value.

are determined by $N_j = 200$ for MC-UVE and by $k = 0.65$ for UVE method (it is an equivalent to $N_j = 219$).

At first, the results for data set 2 are not like the situation for data set 1. Although the two curves are similar to each other, some difference can be seen from Fig. 5 that the distribution of the stability by MC-UVE is slightly more dispersed in full-spectrum than that by UVE method. This may be explained by the fact that the sugar content is a total amount of different sugars, such as glucose, levulose, sucrose and maltose etc. In Fig. 5(a), the selected variables are concentrated on three broad band around 4258, 5620, and 6000cm^{-1}, four narrow wavelength intervals around 4030, 5088, 5220, and 5324cm^{-1}, and several other wavenumbers around 4300, 4370, 5027 and 7100cm^{-1}. However, in Fig. 5(b), more wavenumbers in the range of 4100–4400cm^{-1} are selected, and the narrow peak around 5088cm^{-1} in Fig. 5(a) was depressed in Fig. 5(b).

With the same way as do for data set 1, Fig. 6 can be obtained, showing the variation of the RMSEP of the assessing set with the number of retained variables. It can be seen that, when N_j is 200, the lowest mean value of RMSEP is obtained. So $N_j = 200$ is used for further study. That the N_j for data set 2 is much bigger than that for data set 1 may also be explained by the complex of sugar compounds.

4.4. Comparison of the predicted results by MC-UVE-PLS, UVE-PLS and PLS methods

With the parameters discussed above, MC-UVE-PLS model was developed to predict the nicotine content of the 61 samples of data set 1 and the sugar content of the 47 samples of data set 2. The calculation, including the steps (2), (3), and (4) described above, was repeated 30 times. The mean RMSEP with their standard deviation (σ) were summarized in Table 1. As comparison, the RMSEPs of UVE-PLS and common PLS model with the same data sets were also listed in the table. With a comparison of the results in Table 1, it is clear that, for both of the two data sets, the three methods produced similar prediction with the MC-UVE-PLS being slightly better. However, fewer variables, 100 and 200 for the two data sets, respectively, are used in the MC-UVE-PLS model.

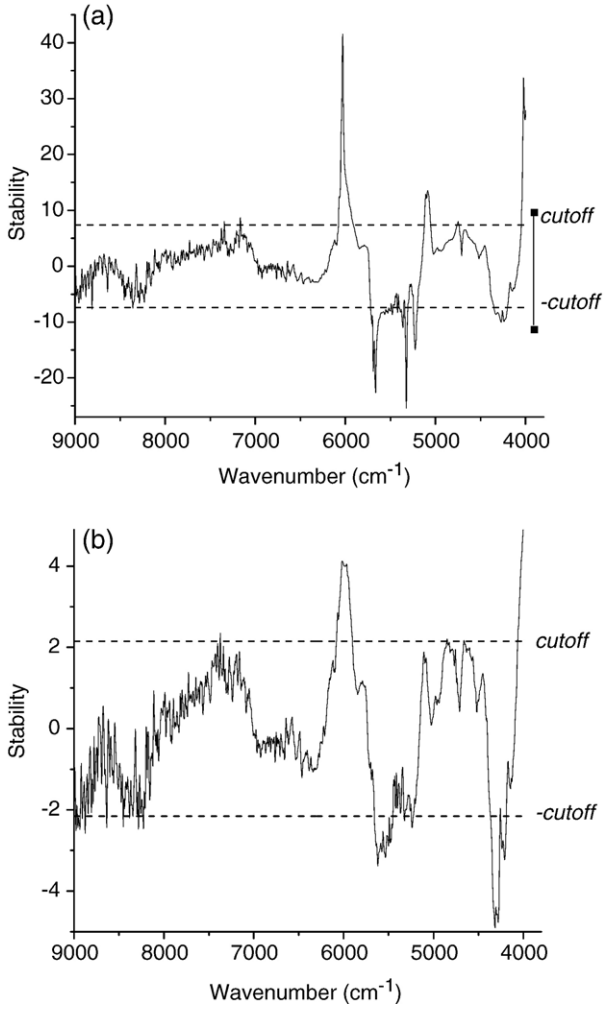

Fig. 5. The stability distribution of each variable for prediction of the total sugar by the UVE (a) and MC-UVE (b) method. The two dot lines indicate the lower and upper threshold. The line between two "■" indicates the stability range of the added random noise.

For a more fair comparison of MC-UVE-PLS and UVE-PLS, the optimal threshold, i.e., the value of k, for UVE-PLS is investigated. It is found that when $k = 0.5$ and 0.3, for the two

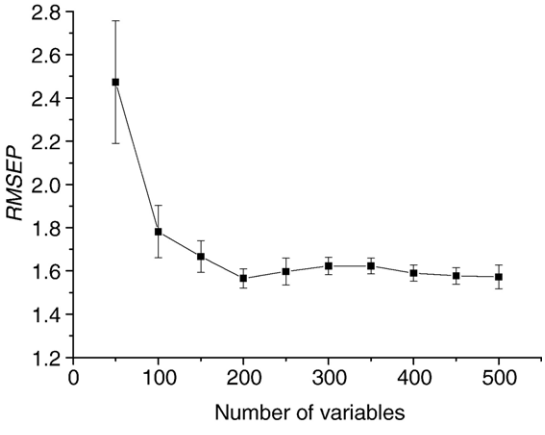

Fig. 6. Variation of RMSEPs with the number of selected variables for data set 2. Standard deviation of 30 runs results is plotted as an error bar crossing the mean value.

Table 1
A comparison of the results obtained by PLS, UVE-PLS, and MC-UVE-PLS

Model	Number of variables	RMSEP(σ)[a]
Data set 1		
PLS	1296	0.17
UVE-PLS	109 (k=0.88)	0.14
MC-UVE-PLS	100	0.14 (0.009)
UVE-PLS-2	466 (k=0.50)	0.14
Data set 2		
PLS	1296	1.71
UVE-PLS	219 (k=0.65)	1.68
MC-UVE-PLS	200	1.57 (0.049)
UVE-PLS-2	700 (k=0.30)	1.58

[a] The RMSEP for MC-UVE-PLS is the mean of 30 repeated runs. σ is the standard deviation of the 30 results.

data sets, respectively, the RMSEP of the assessing data set reaches a minimum, which corresponds the number of retained variables 466 and 700, respectively. With the parameters, RMSEP of the prediction set and the correlation coefficient are listed in Table 1 (in the line of UVE-PLS-2). It can be seen that more variables are retained by the optimal threshold of UVE, but the prediction results remain almost the same.

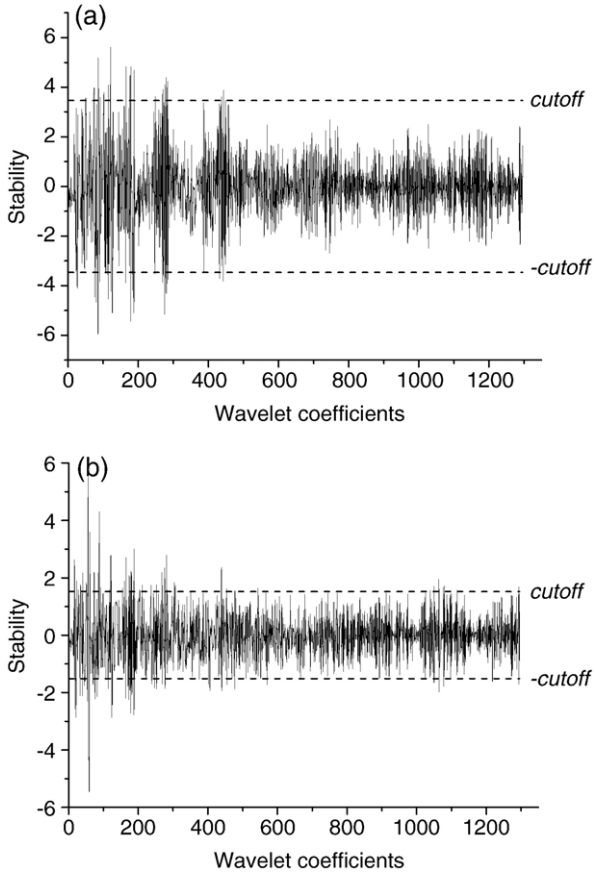

Fig. 7. The stability distribution of wavelet coefficients of data set 1 (a) and data set 2 (b) by MC-UVE. The dot lines indicate the cutoff.

4.5. WT-MC-UVE-PLS method

Due to the characteristic of WT, if wavelet coefficients are used in the MC-UVE method, more parsimonious and efficient model should be obtained. Fig. 7(a) and (b) show the stabilities of the wavelet coefficients of data set 1 and 2, respectively. The cutoff (dot line) is determined by N_j = 50 and 100, respectively. It can be seen that the stability distribution of wavelet coefficients is much different from that of the spectra as shown in Figs. 3 and 5. The stabilities of the largest scale coefficients (1–42) and smaller scales coefficients (after 300) are poorer than that of middle scales coefficients. This is because the large scale coefficients represent the information of background and the small scale coefficients represent the information of noise. With the stability of Fig. 7, the variation of the RMSEPs of the assessing set was investigated. Fig. 8 shows the results of mean RMSEP with their standard deviation (σ) of 30 repeated runs at different number of retained coefficients. It is clear that, for data set 1, with only 30 coefficients the prediction result can be as good as that of the MC-UVE-PLS by using 100 variables. For the data set 2, when 100 coefficients are used, similar results as that by MC-UVE-PLS using 200 variables can be obtained. Therefore, with wavelet compression, the MC-UVE can be significantly improved to produce more parsimonious and efficient model for NIR analysis.

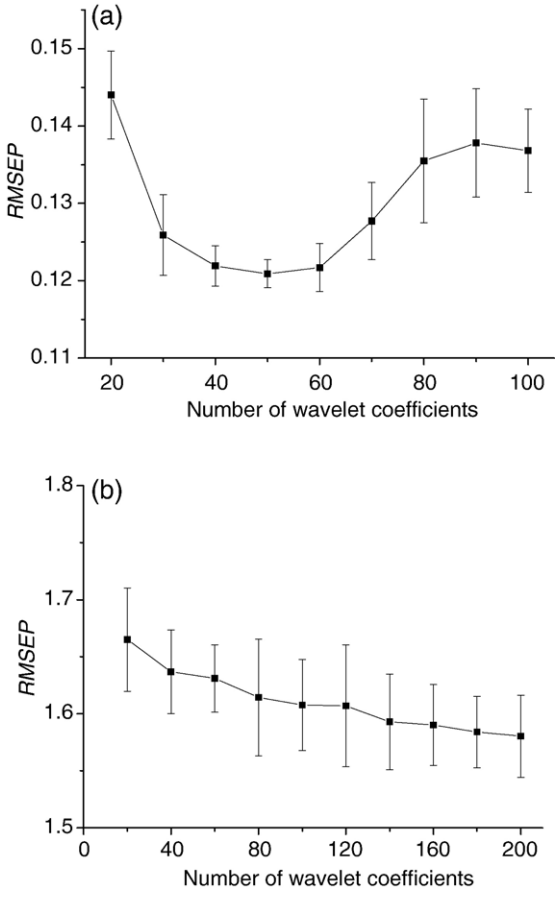

Fig. 8. Variation of the RMSEPs with the number of selected wavelet coefficients for data set 1 (a) and data set 2 (b). Standard deviation of 30 runs results is plotted as an error bar crossing the mean value.

Table 2
Results obtained by WT-MC-UVE-PLS with different number of coefficients

Data set	Number of coefficients	RMSEP(σ)[a]
Data set 1	30	0.13 (0.005)
	40	0.13 (0.003)
	50	0.13 (0.001)
Data set 2	60	1.64 (0.054)
	80	1.63 (0.048)
	100	1.60 (0.060)

[a] The RMSEP for MC-UVE-PLS is the mean of 30 repeated runs. σ is the standard deviation of the 30 results.

For further investigation of the WT-MC-UVE method, the prediction sets of the two data sets were predicted with the WT-MC-UVE-PLS model. Table 2 lists the results. Compared with the results in Table 1, it can be found that similar results are obtained even if less variables are used.

5. Conclusions

A modification of the UVE method, named as the MC-UVE method, for variable selection in NIR analysis was proposed based on an integration of the Monte Carlo technique and UVE method. The method calculate the stability of variables with a large number of PLS coefficients obtained by different training subset determined with MC technique, and then perform the variables selection according to the stability. With applications of the method for analysis of nicotine and sugar contents in tobacco samples, it was proved that the proposed method is an efficient tool. Equivalent or slightly better results can be obtained compared with full-spectral PLS and UVE methods. Furthermore, when it is combined with wavelet transform, the method can produce more parsimonious and efficient model.

Acknowledgements

This study is supported by the National Natural Science Foundation (Nos. 20575031 and 20775036), the Ph.D. Programs Foundation of Ministry of Education (MOE) of China (No. 20050055001).

References

[1] E.V. Thomas, D.M. Haaland, Anal. Chem. 62 (1990) 1091–1099.
[2] P. Geladi, B.R. Kowalski, Anal. Chim. Acta 185 (1986) 1–17.
[3] D. Chen, B. Hu, X.G. Shao, Q.D. Su, Anal. Bioanal. Chem. 381 (2005) 795–805.
[4] C. Borggard, H. Thodberg, Anal. Chem. 64 (1992) 545–551.
[5] A.I. Belousov, S.A. Verzakov, J. von Frese, J. Chemometr. 16 (2002) 482–489.
[6] Y.K. Li, X.G. Shao, W.S. Cai, Talanta 72 (2007) 217–222.
[7] M. Blanco, J. Coello, H. Iturriaga, S. Maspoch, J. Pages, Chemometr. Intell. Lab. Syst. 50 (2000) 75–82.
[8] W. Wu, B. Walczak, D.L. Massart, K.A. Prebble, I.R. Last, Anal. Chim. Acta 315 (1995) 243–255.
[9] L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Appl. Spectrosc. 54 (2000) 413–419.
[10] R. Leardi, L. Norgaard, J. Chemometr. 18 (2004) 486–497.
[11] R.F. Kokaly, R.N. Clark, Remote Sens. Environ. 67 (1999) 267–287.
[12] V. Centner, D.L. Massart, Anal. Chem. 68 (1996) 3851–3858.
[13] X.G. Shao, F. Wang, D. Chen, Q.D. Su, Anal. Bioanal. Chem. 378 (2004) 1382–1387.
[14] R. Leardi, A.L. Gonzalez, Chemometr. Intell. Lab. Syst. 41 (1998) 195–207.
[15] H.C. Goicoechea, A.C. Olivieri, J. Chem. Inf. Comput. Sci. 42 (2002) 1146–1153.
[16] D. Chen, X.G. Shao, B. Hu, Q.D. Su, Anal. Chim. Acta. 511 (2004) 37–45.
[17] R. Put, M. Daszykowski, Baczek, Y. Vander Heyden, J. Proteome Res. 5 (2006) 1618–1625.
[18] J. Polanski, R. Gieleciak, J. Chem. Inf. Comput. Sci. 43 (2003) 656–666.
[19] R.R. Picard, R.D. Cook, J. Am. Stat. Assoc. 79 (1984) 575–583.
[20] K. Baumann, N. Stiefl, J. Comput. Aid. Mol. Des. 18 (2004) 549–562.
[21] X.G. Shao, W.S. Cai, Rev. Anal. Chem. 17 (1998) 235–285.
[22] X.G. Shao, A.K.M. Leung, F.T. Chau, Accounts Chem. Res. 36 (2003) 276–283.
[23] J. Koshoubu, T. Iwata, S. Minami, Appl. Spectrosc. 54 (2000) 148–152.
[24] D. Jouan-Rimbaud, B. Walczak, R.J. Poppi, Anal. Chem. 69 (1997) 4317–4323.
[25] H.W. Tan, S.D. Brown, J. Chemometr. 16 (2002) 228–240.
[26] X.G. Shao, Y.D. Zhuang, Anal. Sci. 20 (2004) 451–454.
[27] F.M. Liang, C.H. Liu, R.J. Carroll, J. Am. Stat. Assoc. 102 (2007) 305–320.
[28] K. Hongo, Y. Kawazoe, H. Yasuhara, Int.J. Quantum. Chem. 107 (2007) 1459–1467.
[29] M. Marseguerra, A. Zoia, Ann. Nucl. Energy. 33 (2006) 1396–1407.
[30] Q.S. Xu, Y.Z. Liang, Y.P. Du, J. Chemometr. 18 (2004) 112–120.
[31] S. Gourvenec, J.A.F. Pierna, D.L. Massart, D.N. Rutledge, Chemom. Intell. Lab. Syst. 68 (2003) 41–51.
[32] Q.S. Xu, Y.Z. Liang, Chemometr. Intell. Lab. Syst. 56 (2001) 1–11.
[33] L.M. Shao, X.Q. Lin, X.G. Shao, Appl. Spectrosc. Rev. 37 (2002) 429–450.