Research article

JOURNAL OF CHEMOMETRICS

Received: 26 March 2014, Revised: 31 July 2014, Accepted: 4 August 2014, Published online in Wiley Online Library: 17 September 2014

(wileyonlinelibrary.com) DOI: 10.1002/cem.2667

Evaluating the reliability of spectral variables selected by subsampling methods

Zhaozhou Lin[a], Xiaoning Pan[a], Bing Xu[a], Jiayu Zhang[a], Xinyuan Shi[a,b]* and Yanjiang Qiao[a,b]*

It is imperfect to evaluate a subsampling variable selection method using only its prediction performance. To further assess the reliability of subsampling variable selection methods, dummy noise variables of different amplitudes were augmented to the original spectral data, and the false variable selection number was recorded. The reliabilities of three subsampling variable selection methods including Monte Carlo uninformative variable elimination (MC-UVE), competitive adaptive reweighted sampling (CARS), and stability CARS (SCARS) were evaluated using this dummy noise strategy. The evaluation results indicated that both CARS and SCARS produced more parsimonious variable sets, but the reliabilities of their final variable sets were weaker than those of MC-UVE. On the contrary, only marginal improvement on the prediction performance was obtained using MC-UVE. Further experiments showed that removing white noise-like variables beforehand would improve the reliability of variables extracted by CARS and SCARS. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: noise variable; reliability; subsampling; Monte Carlo sampling; competitive adaptive reweighted sampling (CARS)

1. INTRODUCTION

Near-infrared (NIR) spectroscopy is a powerful and rapid analytical technique and has become a widespread tool for the analysis of agricultural, petroleum, chemical, and pharmaceutical samples [1–4]. During these analyses, one of the most crucial tasks is to construct a reliable model to handle the collinearity of the NIR spectra. Here, partial least squares (PLS) regression is the most effective and commonly used method. Generally, the established calibration model includes all measured wavelengths. From a statistical or data analysis perspective, it is quite difficult for even the experienced spectroscopists to determine the wavelengths that should be retained in calibration models. Variable selection methods that were originally designed to extract the most pertinent wavelengths from the full spectrum have drawn considerable attention in recent quantitative analyses. Both experimental and theoretical applications have demonstrated that the prediction and interpretation performance of the calibration model can be improved through variable selection [5–11].

In chemometrics, there are several methods to extract pertinent wavelengths [12,13]. However, when the calibration samples change, the selected wavelengths can hardly be consistent. Variable selection methods using re-sampling techniques can slightly reduce the variation in the variable set caused by changes in the calibration set. One such method is the Monte Carlo uninformative variable elimination (MC-UVE) [14]. Rather than adding random noise variables to estimate the cutoff value, MC-UVE determines the threshold directly by using the stability calculated with the Monte Carlo sampling (MCS) strategy. Competitive adaptive reweighted sampling (CARS) reduces the variation caused by changes in the calibration set by implementing an adaptive reweighted sampling [15]. Stability CARS (SCARS) [16] modifies the raw CARS to create a more parsimonious and reasonable model.

Generally, variable selection methods are evaluated using prediction accuracy, but studies on the reliability of subsampling variable selection methods are rare. In addition, the random errors in routine NIR analysis can generally be reduced but not eliminated. The suspicion on the reliability of subsampling methods cannot be eliminated if noise variables exist in the final variable set. Thus, white noise variables that mimicked the behavior of spectral variance caused by random error were used to evaluate the effectiveness and reliability of the different subsampling variable selection methods. Each of the three Monte Carlo based subsampling variable selection methods was repeated 500 times to give a stable result because repeated cross-validation in small-sample settings is less affected by the *error-counting* problem [17]. The proposed approach was tested on three datasets. It is clear that our approach reveals the illusive effect of noise-like variables on the reliability of subsampling variables selection methods. Therefore, it is suggested to remove noise-like variables beforehand to enhance the reliability of variable selected and thus the final analytical determination.

* *Correspondence to: X. Shi and Y. Qiao, Key Laboratory of TCM-Information Engineer of State Administration of TCM, Beijing University of Chinese Medicine, Beijing 100102, China*
E-mail: xyshi@126.com; yjqiao@263.net

a *Z. Lin, X. Pan, B. Xu, J. Zhang, X. Shi, Y. Qiao*
Beijing University of Chinese Medicine, Beijing, 100102, China

b *X. Shi, Y. Qiao*
Key Laboratory of TCM-Information Engineer of State Administration of TCM, Beijing 100102, China

2. METHODS

2.1. Monte Carlo uninformative variable elimination

In linear regression, a calibration model is expressed as follows:

$$y = X\beta + e \tag{1}$$

here, **X** is a column-centered $n \times p$ matrix containing p spectral responses of n samples. Both $\boldsymbol{y}$ and $\boldsymbol{e}$ are $n \times 1$ vectors, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of the regression coefficients.

For spectral data, the regression coefficients estimated by using the PLS model are preferable. Thus, only the PLS model is considered. Typically, the original PLS model is constructed by using all measured spectral variables. However, the noise variables or the other variables containing irrelevant information may deteriorate the accuracy of the PLS model. Centner *et al.* proposed the UVE-PLS approach to eliminate the negative effect of the uninformative variables. Similar to UVE-PLS, MC-UVE [9,14] calculates the reliability of each variable to sieve out the uninformative variables, but the regression vectors are estimated with the calibration subset sampled by N MCS runs. This forms the reliability criterion c defined as follows:

$$c_j = \frac{\beta_j}{s\left(\beta_j\right)}, j = 1,\ 2, \ldots, p \tag{2}$$

with

$$s\left(\beta_j\right) = \left(\sum_{i=1}^{N} \frac{\left(\beta_{ij} - \beta_j\right)^2}{N-1} \right)^{1/2}$$

where c_j is the reliability (i.e., stability) of the jth wavelength. The term β_{ij} denotes the regression coefficient of variable j in the PLS model of the ith MCS; $\boldsymbol{\beta}_j$ and s(β_j) are the mean and the standard deviation of all β_{ij} for the jth wavelength, respectively.

With this sorted stability, certain informative variables are selected to construct the final PLS model. The number of informative variables can be optimized by changing the number of variables used.

2.2. Competitive adaptive reweighted sampling

CARS considers the variability of the regression coefficients caused by the variation in the calibration set via a Monte Carlo strategy. In each run of CARS, a certain number of samples are selected to form the current calibration set. By adopting dual elimination procedures, that is, enforced wavelength reduction and adaptive reweighted sampling (ARS), redundant variables can be repeatedly eliminated. The entire algorithm is briefly outlined here. For details, please refer to [15].

At the very beginning of the CARS algorithm, a subset of n samples is selected by MCS to estimate the regression coefficient $\boldsymbol{\beta}$. In order to evaluate the importance of the ith variable, a normalized weight is defined as follows:

$$w_i = \frac{|\beta_i|}{\sum_{i=1}^{p} |\beta_i|}, i = 1,\ 2,\ 3, \ldots, p \tag{3}$$

During the enforced wavelength reduction step, the ratio of the variables remaining in the jth sampling run is updated using the following function:

$$r_j = ae^{-kj} \tag{4}$$

with constants defined by the following two equations:

$$a = \left(\frac{p}{2}\right)^{1/(N-1)}$$

and

$$k = \frac{\ln(p/2)}{N-1}$$

The ratio filters the variables that are less important. Informative variables are retained in the final variable subset, although the definition of the ratio function is not directly related to the importance of each variable. Based on the sorted variables retained, the ARS procedure further condenses the variables subset.

The aforementioned steps are sequentially repeated N times. In each run, the root mean square error (RMSE) of the cross-validation (RMSECV) is calculated for the current variable subsets. Finally, the subset with the lowest RMSECV is selected as the optimal variable subset. The RMSE calculation is provided as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2} \tag{5}$$

where for cross-validation, y_i is the reference property value for the ith sample of the calibration set, $\hat{y}_i$ is the predicted property value of the ith sample in the calibration set, and n is the number of samples. For the root mean square of prediction (RMSEP), y_i is the reference value for the ith sample in the prediction set, and $\hat{y}_i$ is the predicted value of the ith sample in the prediction set.

2.3. Stability competitive adaptive reweighted sampling

The overall framework of SCARS [16] is similar to that of CARS, except that important variables are defined as the variables with large stability. SCARS selects N subsets of informa-

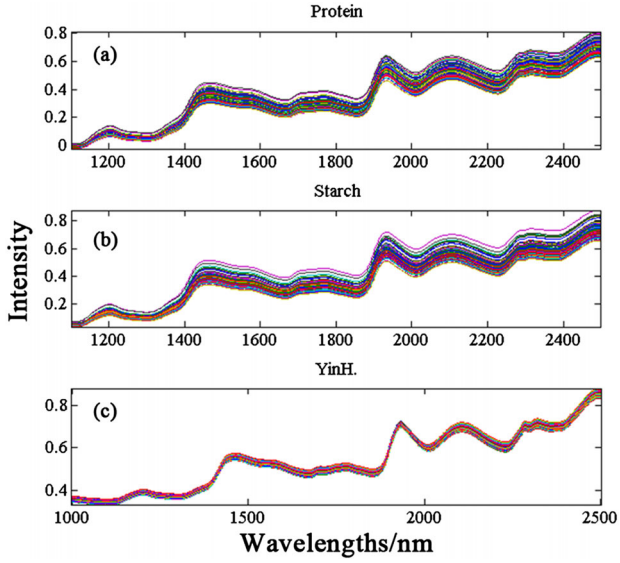

Figure 1. The raw spectra of corn dataset for protein (a) and starch (b) together with *Yin H.* (c).

Journal of CHEMOMETRICS

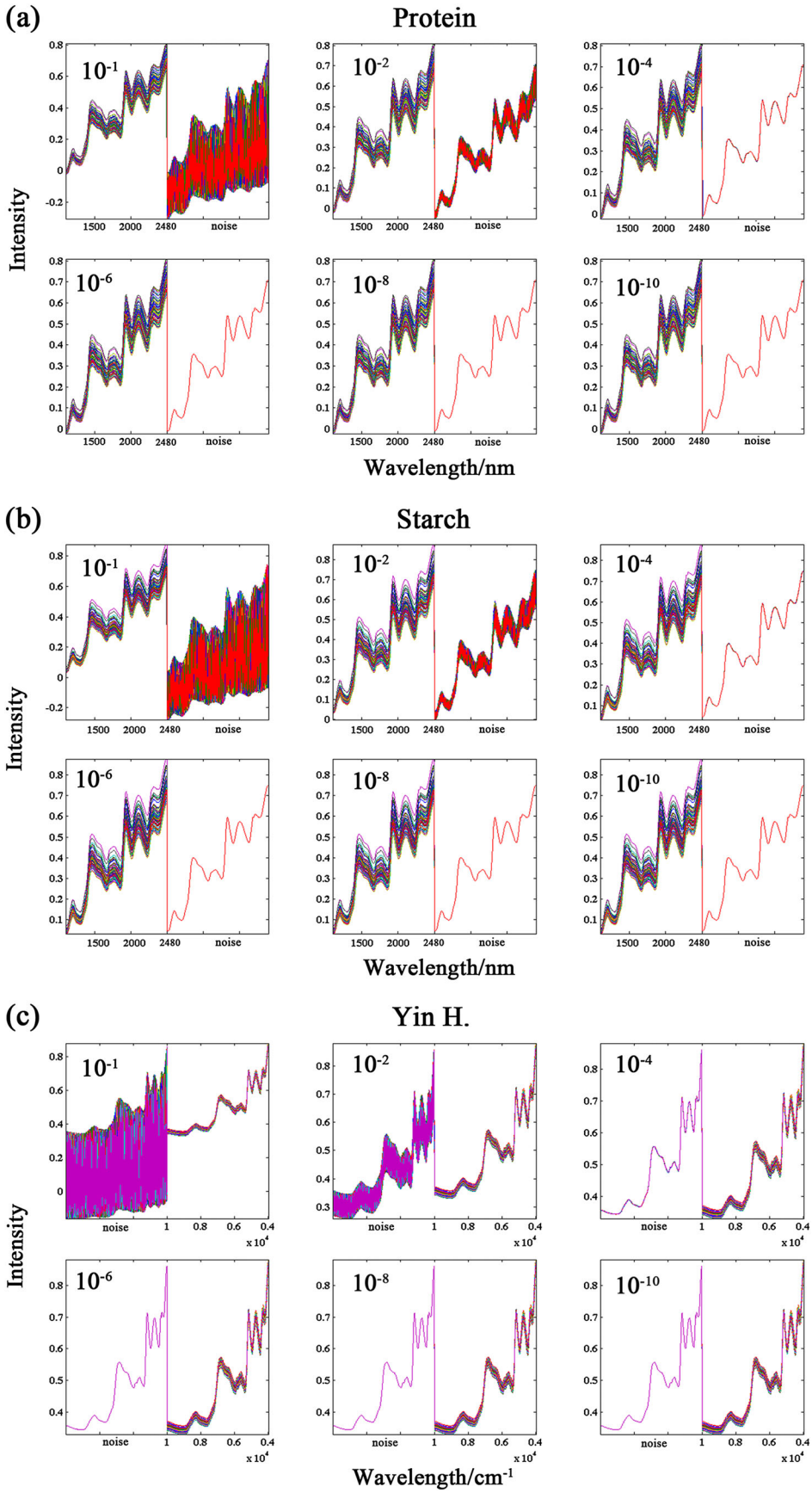

Figure 2. The augmented spectra data for protein, starch, and *Yin H.* samples in noise scale varying among 10^{-1}, 10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}, and 10^{-10}.

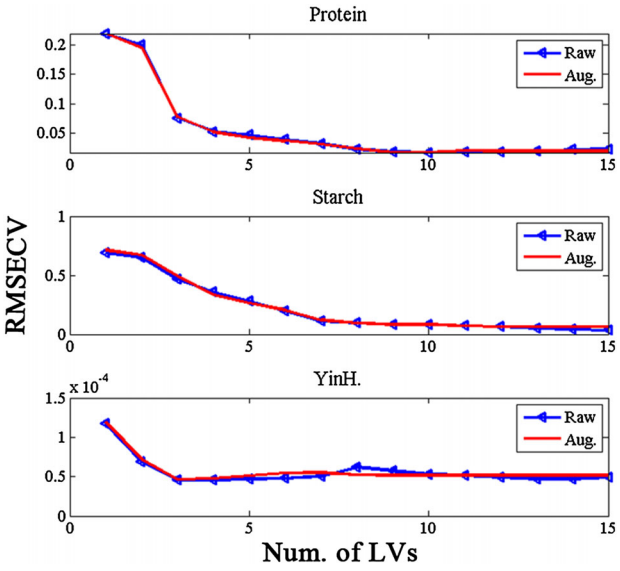

Figure 3. A comparison between the root mean square error of the cross-validation (RMSECV) curve of raw spectra data and that of the augmented (Aug.) data. LVs, latent variables.

tive variables in a stepwise manner by using N iterative loops. Initially, all variables are embedded in the survival variable subset. During each loop, SCARS randomly selects n_{sam} samples M times. The stability of each variable is computed with Equation (2). Then, the enforced wavelength selection and ARS procedure remove uninformative variables. A PLS model is built, and RMSECV is calculated for each variable subset. The variable subset with smallest RMSECV is selected as the final subset.

Although the stability criterion adopted in SCARS is the same as that for MC-UVE, the prediction performance of SCARS improves markedly versus MC-UVE.

3. EXPERIMENTAL

3.1. Datasets

Three datasets were used to investigate the reliabilities of the aforementioned subsampling methods. Two can be downloaded from http://www.eigenvector.com/data/Corn/index.html. Each downloaded dataset contains 80 NIR spectra for 80 corn samples. Every spectrum covered 1100–2498 nm at 2-nm intervals. The spectra measured on *m*5 were modeled to predict the starch content of corn samples, while the prediction accuracy of the protein content of the corn samples was investigated using spectra collected on *mp*5. All spectral data were mean centered, and no extra preprocessing was performed. An overlay plot of the original spectra was shown in Figure 1(a, b).

The other dataset [18] contains 68 NIR spectra from *Yin Huang Granule* (*Yin H.*) samples, which were manufactured by JXJM Co., Ltd. (Jiang Xi, China). The NIR spectra were collected at 8 cm^{-1} interval over the spectral range from 10,000 to 4000 cm^{-1} using Antaris FT-NIR System (Thermo Scientific, Madison, WI, USA) equipped with an integrating sphere system. Each sample was analyzed in triplicate, with spectra obtained by averaging 32 scans. Assay values were determined by high-performance liquid chromatography. The raw NIR spectra of *Yin H.* samples were shown in Figure 1(c).

The dummy noise matrix [19,20] was created as follows: First, the means of the calibration absorbance spectra were digitally duplicated n times (n denotes the number of calibration samples) and converted to reflectance mode. White noise at different levels was then added to the reflection units. Finally, the contaminated spectra in reflectance mode were converted back to absorbance mode. The augmented spectra at different noise levels were presented in Figure 2.

3.2. Software

All calculations were performed on a PC equipped with an i7-processor using MATLAB (MathWorks, Natick, MA, USA)

Table I. The prediction performance of PLS, CARS, SCARS, and MC-UVE on three augmented datasets in terms of RMSECV, RMSEP, TNVS, and NDNV

Data	Method	RMSECV	RMSEP	NDNV	TNVS
Protein	PLS	0.1248	0.1713	700	400
	CARS	0.0517 (0.0064)	0.1708 (0.0176)	32 (11)	76 (25)
	SCARS	0.0687 (0.0091)	0.1624 (0.0173)	12 (6)	58 (42)
	MC-UVE	0.1237 (0.0082)	0.1572 (0.0041)	0	80 (30)
Starch	PLS	0.2982	0.1983	700	400
	CARS	0.1239 (0.0132)	0.1954 (0.0323)	9 (6)	45 (18)
	SCARS	0.1622 (0.0232)	0.1548 (0.0541)	0	2 (16)
	MC-UVE	0.3318 (0.1384)	0.3043 (0.0455)	0	70 (50)
Yin H.	PLS	0.0067	0.0069	1557	3114
	CARS	0.0061 (6.1733[a])	0.0066 (8.3944[a])	0	6 (4)
	SCARS	0.0061 (2.0309[a])	0.0068 (5.7971[a])	0	2 (17)
	MC-UVE	0.0073 (7.5698[b])	0.0069 (2.0593[b])	0	50 (150)

In the parenthesis is the interquartile range of 500 repeated runs.
PLS, partial least squares; CARS, competitive adaptive reweighted sampling; SCARS, stability CARS; MC-UVE, Monte Carlo uninformative variable elimination; RMSECV, root mean square error of the cross-validation; RMSEP, root mean square of prediction; TNVS, total number of variables selected; NDNV, number of dummy noise variable.
[a]Indicates the order of magnitude is 10^{-5}.
[b]Indicates the order of magnitude is 10^{-4}.

running Windows 7 Professional operating system. The MC-UVE, CARS, and SCARS functions were obtained from or were modifications of functions in the toolbox downloaded from http://code.google.com/p/carspls/. The Kennard–Stone function employed was written with MATLAB.

4. RESULTS AND DISCUSSION

4.1. Data augmented with noise variables

Both the corn spectra and the *Yin H.* spectra were augmented with dummy noise spectra. The deviation of the dummy spectral variables was fixed at 10^{-4} based on the variance of repeatedly measured spectra. Each augmented dataset was split into two independent datasets by the Kennard–Stone algorithm [21]. Specifically, the corn data were split into 60 against 20, that is, 60 for the calibration set and 20 for the test set. The *Yin H.* samples were split into 45 calibration samples and 23 testing samples. The effectiveness and reliability of the three Monte Carlo subsampling variable selection methods were investigated using these datasets.

As shown in Figure 3, the RMSECV curve decreased gradually with increasing latent variables (LVs) until it plateaued near nine LVs. The RMSECV curves of the augmented spectra data nicely approach those of the raw data. In other words, there was no significant difference between the RMSECV curves of the augmented and the raw datasets. This means that the noise variables have a limited effect on the prediction ability of the augmented PLS model. Therefore, the maximum number of LVs in the (S)CARS algorithm and MC-UVE algorithm were set at nine for the corn datasets. Similarly, the maximum number of LVs allowed in the (S)CARS algorithm and MC-UVE algorithm were set at three for the *Yin H.* samples. The corresponding regression vectors were then used to predict the assay values of samples in the test sets.

During CARS run, a 10-fold cross-validation and 100 times MCS were executed using parameters from the literatures [15,16]. The entire CARS procedure was repeated 500 times. Meanwhile, most of the parameters adopted in the SCARS

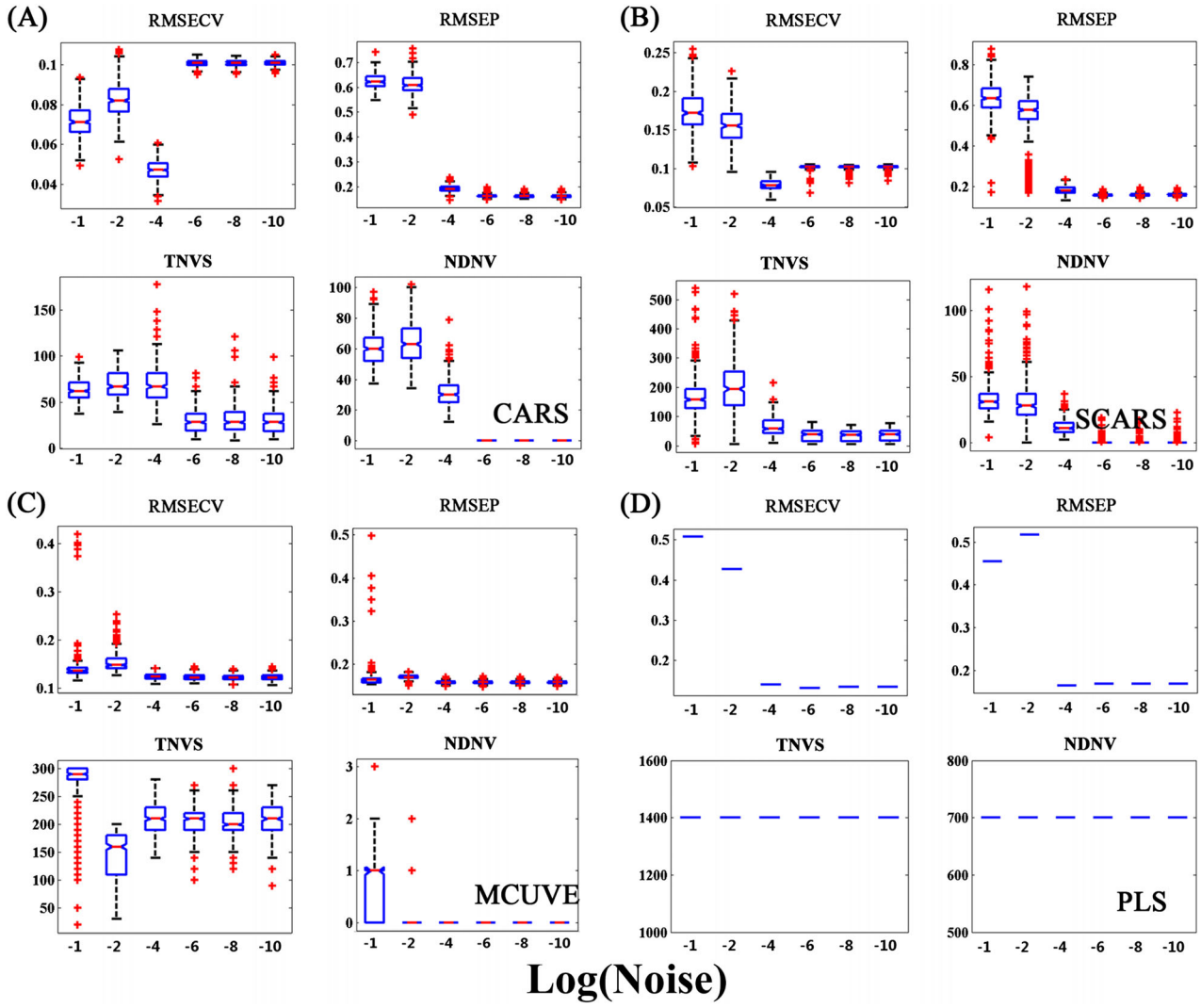

Figure 4. The boxplots of four investigated methods on protein dataset with noise in different scales. The root mean square error of the cross-validation (RMSECV), root mean square of prediction (RMSEP), total number of variables selected (TNVS), and number of dummy noise variable (NDNV) metrics were repeatedly calculated 500 times. CARS, competitive adaptive reweighted sampling; SCARS, stability CARS; MC-UVE, Monte Carlo uninformative variable elimination; PLS, partial least squares.

algorithm were determined similar to that of CARS. Additionally, 36 calibration samples were randomly selected to estimate the stability of the SCARS algorithm, and the number of MCS per loop was set to 100. For the MC-UVE algorithm, 45 calibration samples were randomly selected per sampling run. To optimize the number of variables selected in the final model, a serial number of variables ranging from 20 to 300 at increments of 10 were investigated. In addition, the number of variables remaining in the final model and the number of the pure noise part (false variable selection number) were recorded for each run in the three subsampling method. The median, rather than mean, of each indicator was presented in Table I because the median is more robust than the mean.

The number of variables remaining in the final model of the MC-UVE had a median value of 180 for protein data (Table I). Among them, there was no dummy noise variable, and its prediction performance improved to some extent versus the plain PLS model in terms of RMSEP. The RMSECV median value of SCARS method decreased markedly from 0.1248 to 0.0517. Unfortunately, a certain number of selected variables were dummy noise. Therefore, it was difficult to be confident in the variables selected using SCARS. For CARS, both the median values of RMSECV and RMSEP decreased markedly versus the raw PLS model. However, dummy noise variables still existed in the final variable set. These results led us to conclude that MC-UVE was more acceptable than the CARS and SCARS method in terms of the reliability of the final variable set. Moreover, the difference between RMSECVs and RMSEPs in CARS and SCARS was large. Therefore, even RMSECV can be an unbiased estimate of the prediction ability of the calibration model. It may be more sensitive to the noise-like variables.

From the results presented in Table I for the starch data, it can be observed that CARS performed no worse than the plain PLS algorithm. SCARS outperformed the other two Monte Carlo-based subsampling methods because there was no dummy noise in the final variable set. The prediction performance of SCARS improved markedly versus the plain PLS model. Although no improvement was obtained with the MC-UVE method, there

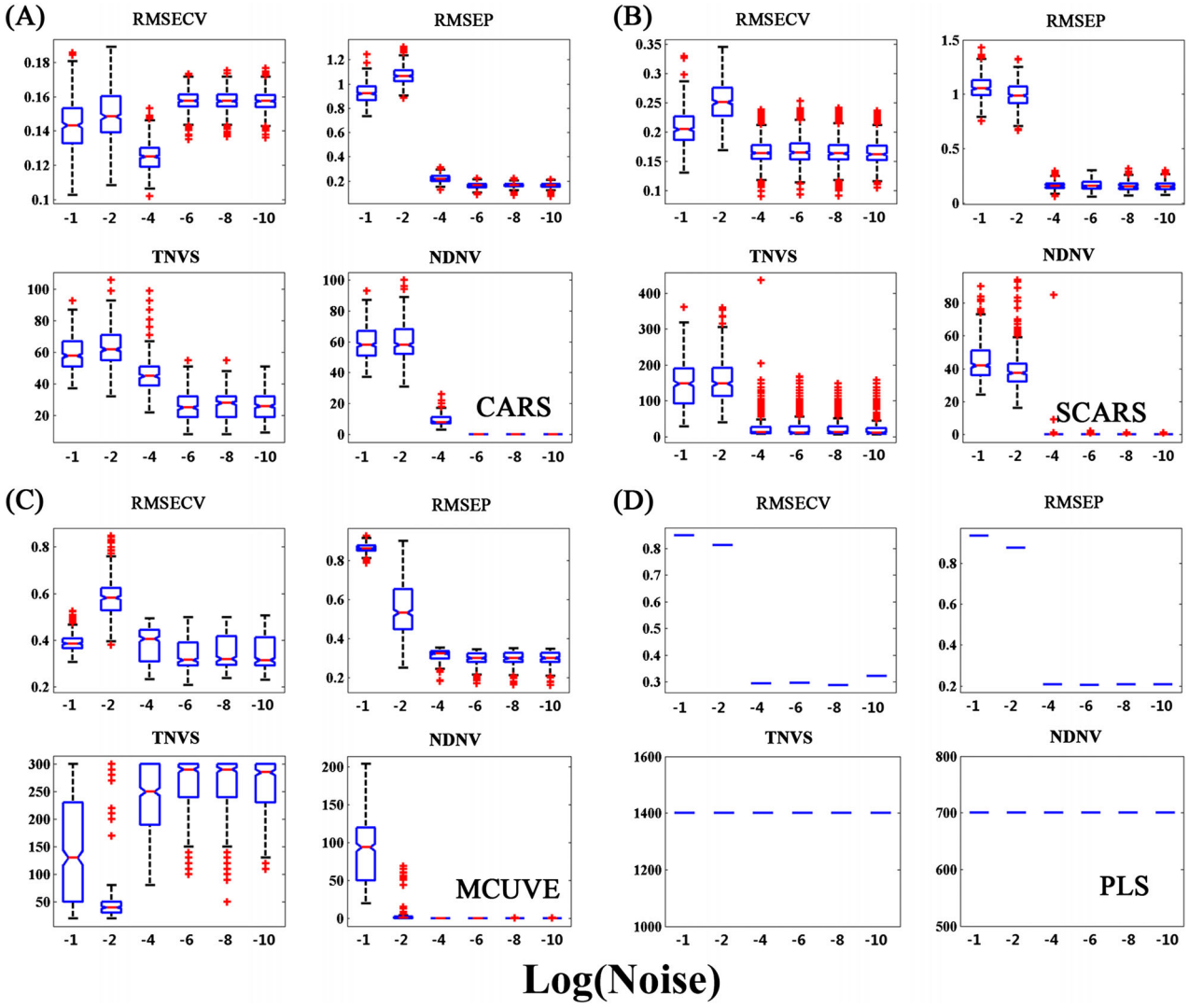

Figure 5. A comparison among the performance of three Monte Carlo-based methods on the starch data augmented with noise varying from 10^{-1} to 10^{-10}. Subplots (A), (B), (C), and (D) correspond to the results of competitive adaptive reweighted sampling (CARS), stability CARS (SCARS), Monte Carlo uninformative variable elimination (MC-UVE), and partial least squares (PLS), respectively. RMSECV, root mean square error of the cross-validation; RMSEP, root mean square of prediction; TNVS, total number of variables selected; NDNV, number of dummy noise variable.

was no dummy noise variable in the final variable set. These results supported our previous conclusion that the variables selected by MC-UVE were reliable but that its efficiency should be improved. Furthermore, noise variability may be one of the elements that contributed most to the imbalance between RMSECV and RMSEP.

The results obtained on *Yin H.* data indicated that the RMSECV median is comparable to RMSEP when there was no dummy noise-like variable in the final variable set. Together with the results of protein and starch data, it can be concluded that modeling strategies matter more than the ranking function.

4.2. Data augmented with noise in different scales

Dummy noise matrixes in different amplitudes were augmented to the original dataset to consider the scale effect. The MC-UVE, CARS, and SCARS methods were adopted to select informative variables in this noisy context. The parameters, especially the number (or maximum number) of LVs, were the same as those in Section 4.1 because it was assumed that the noise variables mainly contribute to the last LVs and the contribution of the last LVs was limited (Figure 3).

As shown in Figure 4, the results of the four investigated methods using data augmented with noise drawn randomly from the open interval (0, 1) were far less than satisfactory. The variables selected by the CARS method were nearly all noise variables. Worse still, noise variables appeared in the variable set selected by MC-UVE (Figures 4(C) and 5(C)). Fortunately, random variation in this scale was rare.

When the noise amplitude dropped to 10^{-4}, the median RMSEP values of CARS and SCARS decreased drastically to 0.1911 and 0.1823, respectively (Figure 4). Meanwhile, the median RMSECV values for CARS and SCARS reached their minimum, which seemed more satisfactory than that of the plain PLS. However, a significant part of the selected variables was the dummy noise variable. This suggested that the corresponding model has a poor reliability.

The dummy noise variable could be excluded completely if the noise amplitude continued to drop. Moreover, both the median RMSECV and RMSEP values became stable. Therefore, noise with relative small amplitude did not alter the final results significantly. But, caution must be taken when the RMSECV metrics was used to evaluate the prediction performance. Although the variable selected by MC-UVE was reliable, limited improvement in prediction performance was seen. The boxplots also

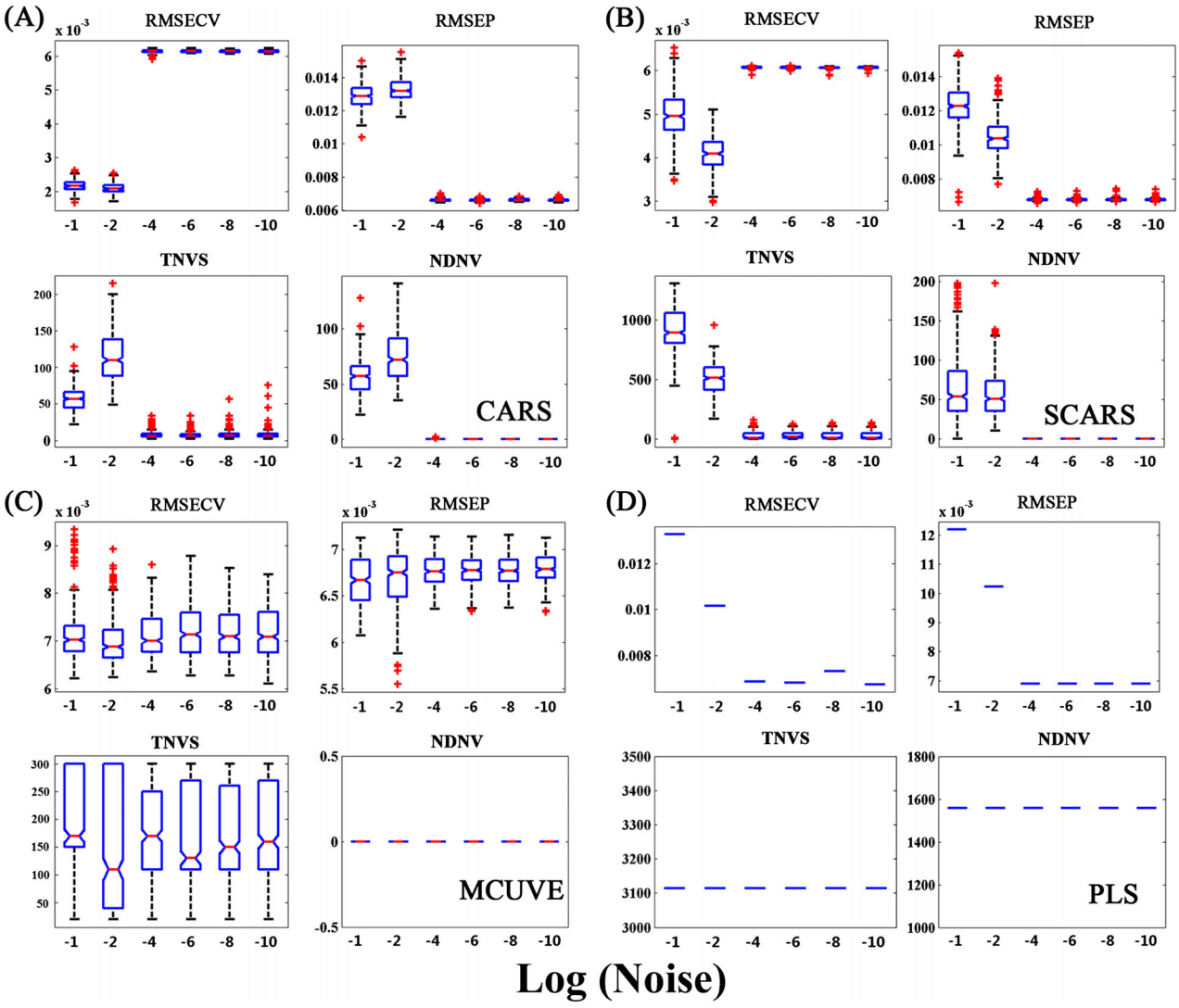

Figure 6. A summary of the results obtained from four investigated methods on *Yin H.* data. For details, refer to Figure 5.

Table II. The performance of PLS, CARS, and SCARS on the datasets reduced by using MC-UVE

Data	Method	RMSECV	RMSEP	NDNV	TNVS
Protein	PLS	0.1391	0.1819	0	400
	CARS	0.1046 (0.0015)	0.1559 (0.0059)	0	20 (8)
	SCARS	0.1052 (0.0023)	0.1571 (0.0095)	0	16 (7)
Starch	PLS	0.3397	0.2608	0	400
	CARS	0.1697 (0.0094)	0.1948 (0.0217)	0	17 (4)
	SCARS	0.2324 (0.0193)	0.2585 (0.0399)	0	10 (8)
Yin H.	PLS	0.0068	0.0069	0	1000
	CARS	0.0062 (4.8705[a])	0.0066 (7.7733[a])	0	6 (4)
	SCARS	0.0061 (2.1731[a])	0.0068 (6.4518[a])	0	12 (31)

In the parenthesis is the interquartile range of 500 repeated runs.
PLS, partial least squares; CARS, competitive adaptive reweighted sampling; SCARS, stability CARS; MC-UVE, Monte Carlo uninformative variable elimination; RMSECV, root mean square error of the cross-validation; RMSEP, root mean square of prediction; TNVS, total number of variables selected; NDNV, number of dummy noise variable.
[a]Indicates the order of magnitude is 10^{-5}.

show that the RMSECV median is comparable to that of RMSEP when there was no dummy noise-like variable in the final variable set. Similar conclusions could be drawn from the metrics calculated from the starch data (Figure 5).

Figure 6 is a graphical comparison of the PLS results on variables selected by MC-UVE, CARS, and SCARS. No dummy noise variable appeared in the variable set selected by MCUVE (Figure 6(C)), which adds confidence to this approach. With a small number of variables selected with CARS and SCARS, the PLS model behaved as well as the full spectrum PLS model. Furthermore, both the CARS and SCARS methods resisted all the noise variables when the amplitude was below 10^{-2}.

These results support previous assumptions that the RMSECV metric is more sensitive to the noise-like variables. Furthermore, unavoidable noise variation can make the RMSECV more satisfied than expected. SCARS is an embedding backward elimination strategy that performs better than MC-UVE in terms of prediction performance. But the reliability of the variables selected by SCARS is weaker than that of MC-UVE. For reliability, CARS performed no better than SCARS, although the absolute value of regression coefficients was used to rank features in CARS. Because the variance of repeatedly measured spectra is about 10^{-4}, it is better to remove potentially false informative wavelengths before the CARS or SCARS approach is used.

4.3. Data preselected by MC-UVE

Although the variables selected by CARS and SCARS are more parsimonious and predictable than those from MC-UVE, their reliability must be improved. Thus, all the three datasets were pretreated by MC-UVE. In this section, each spectral dataset was augmented with a noise matrix (Section 4.1). The number of variables remaining in the final set of the MC-UVE was directly set at 400 for the corn data and 1000 for *Yin H.* data because MC-UVE was used as a rough filter. The RMSECV curves of the PLS model constructed with the three reduced datasets plateaued near eight, seven, and three LVs, respectively. Therefore, the maximum number of LVs was fixed at eight, seven, and three for both CARS and SCARS. The other parameters remained the same as in Section 4.1.

There was no dummy noise variable in the variable sets selected by CARS and SCARS for all three datasets (Table II). The difference between RMSECV and RMSEP median values for the reduced protein data was decreased. This means that with the dummy noise-like variables removed by MC-UVE, the reliability of the final models improved. For the starch data, however, the models constructed by applying SCARS on the reduced data were obviously worse than those of the augmented data (Table I). A more predictable variable selection method should probably be integrated. These observations led us to the conclusion that the reliability of variables selected by CARS or SCARS can be improved when they are coupled with MC-UVE.

5. CONCLUSIONS

The dummy noise variable was used as an indicator to evaluate the reliability of the variables selected by a subsampling variable selection method. A comparison study of the reliabilities of the three Monte Carlo-based subsampling methods illustrated that the variables selected by MC-UVE were more reliable than those selected by CARS and SCARS. However, applying SCARS or CARS to spectral data produced more parsimonious and predictable variable sets. The results of adding different levels of normally distributed noise to the mean of the spectra clearly illustrated that the dummy noise variable nicely satisfies the RMSECV metrics than expected. Moreover, the reliability of the variables selected by CARS or SCARS could be improved when they were coupled with MC-UVE. In other words, removing noise-like variables beforehand will be beneficial for improving the reliability of variables extracted separately by CARS and SCARS. Random error in routine NIR analysis can generally be reduced but not eliminated. Thus, removing noise-like variable beforehand will be also beneficial for reducing the potential side effects caused by random error.

Acknowledgements

The authors would like to thank the anonymous reviewers for their kind and insightful comments. Financial supports from the Joint Development Program Supported by Beijing Municipal Education

Commission—Key Laboratory Construction Project and the graduate research projects of Beijing University of Chinese Medicine (no. 2013-JXBZZ-XS-112) are gratefully acknowledged. The computation was partially supported by CHEMCLOUDECOMPUTING (Beijing University of Chemical Technology, Beijing, China).

REFERENCES

1. Lee M-J, Seo D-Y, Lee H-E, Wang I-C, Kim W-S, Jeong M-Y Choi GJ. In line NIR quantification of film thickness on pharmaceutical pellets during a fluid bed coating process. *Int. J. Pharm.* 2011; **403**: 66–72.
2. Kohonen J, Reinikainen S-P Höskuldsson A. Block-based approach to modelling of granulated fertilizers' quality. *Chemom. Intell. Lab. Syst.* 2009; **97**: 18–24.
3. Pomerantsev AL, Rodionova OY, Melichar M, Wigmore AJ Bogomolov A. In-line prediction of drug release profiles for pH-sensitive coated pellets. *Analyst* 2011; **136**: 4830–4838.
4. Ricci C, Eliasson C, Macleod N, Newton P, Matousek P Kazarian S. Characterization of genuine and fake artesunate anti-malarial tablets using Fourier transform infrared imaging and spatially offset Raman spectroscopy through blister packs. *Anal. Bioanal. Chem.* 2007; **389**: 1525–1532.
5. Smit S, van Breemen MJ, Hoefsloot HCJ, Smilde AK, Aerts JMFG de Koster CG. Assessing the statistical validity of proteomics based biomarkers. *Anal. Chim. Acta* 2007; **592**: 210–217.
6. Roger JM, Palagos B, Bertrand D Fernandez-Ahumada E. CovSel: variable selection for highly multivariate and multi-response calibration: application to IR spectroscopy. *Chemom. Intell. Lab. Syst.* 2011; **106**: 216–223.
7. Liu F, He Y Wang L. Determination of effective wavelengths for discrimination of fruit vinegars using near infrared spectroscopy and multivariate analysis. *Anal. Chim. Acta* 2008; **615**: 10–17.
8. Liebmann B, Friedl A Varmuza K. Determination of glucose and ethanol in bioethanol production by near infrared spectroscopy and chemometrics. *Anal. Chim. Acta* 2009; **642**: 171–178.
9. Han Q-J, Wu H-L, Cai C-B, Xu L Yu R-Q. An ensemble of Monte Carlo uninformative variable elimination for wavelength selection. *Anal. Chim. Acta* 2008; **612**: 121–125.
10. Rossi F, Francois D, Wertz V, Meurens M Verleysen M. Fast selection of spectral variables with B-spline compression. *Chemom. Intell. Lab. Syst.* 2007; **86**: 208–218.
11. Brás LP, Lopes M, Ferreira AP Menezes JC. A bootstrap-based strategy for spectral interval selection in PLS regression. *J Chemom.* 2008; **22**: 695–700.
12. Balabin RM Smirnov SV. Variable selection in near-infrared spectroscopy: benchmarking of feature selection methods on biodiesel data. *Anal. Chim. Acta* 2011; **692**: 63–72.
13. Xiaobo Z, Jiewen Z, Povey MJW, Holmes M Hanpin M. Variables selection methods in near-infrared spectroscopy. *Anal. Chim. Acta* 2010; **667**: 14–32.
14. Cai W, Li Y Shao X. A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. *Chemom. Intell. Lab. Syst.* 2008; **90**: 188–194.
15. Li H, Liang Y, Xu Q Cao D. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Anal. Chim. Acta* 2009; **648**: 77–84.
16. Zheng K, Li Q, Wang J, Geng J, Cao P, Sui T, Wang X Du Y. Stability competitive adaptive reweighted sampling (SCARS) and its applications to multivariate calibration of NIR spectra. *Chemom. Intell. Lab. Syst.* 2012; **112**: 48–54.
17. Braga-Neto UM Dougherty ER. Is cross-validation valid for small-sample microarray classification? *Bioinformatics* 2004; **20**: 374–380.
18. Wu Z. The basic theories and methods research of NIR technology on process analysis of Chinese Medicine [Doctor]: Beijing University of Chinese Medline; 2012.
19. Haaland DM Easterling RG. Application of new least-squares methods for the quantitative infrared analysis of multicomponent samples. *Appl. Spectrosc.* 1982; **36**: 665–673.
20. Sáiz-Abajo MJ, Mevik BH, Segtnan VH Næs T. Ensemble methods and data augmentation by noise addition applied to the analysis of spectroscopic data. *Anal. Chim. Acta* 2005; **533**: 147–159.
21. Kennard RW Stone LA. Computer aided design of experiments. *Technometrics* 1969; **11**: 137–148.