Improving the Creditability and Reproducibility of Variables Selected From Near Infrared Spectra

Zhaozhou Lin, Yanling Pei, Zhao Chen, Xinyuan Shi, Yanjiang Qiao
College of Chinese Medicine
Beijing University of Chinese Medicine
Beijing, China

Xinyuan Shi, Yanjiang Qiao
Research Center of TCM-information Engineering
State Administration of Traditional Chinese Medicine of the People's Republic of China
Beijing, China

Abstract—**A method based on an assembly of two metrics, including the variable importance in projection (VIP) and the PLS regression coefficients B, was developed for wavelength selection in multivariate calibration of spectral data. The proposed algorithm termed VIP-CARS combined the two metrics in a sequential and iterative manner, rather than directly introducing VIP into CARS-PLS. This approach is particularly attractive for quantification due to its relatively higher reproducibility and robustness compared to the CARS procedure. The method was tested on datasets taken from the corn and Rukuaixiao Tablets. It was shown that a small number of well-defined relevant spectral variables were identified with the proposed approach, providing easy spectral interpretation and high creditability. Moreover, with the implementation of the VIP-CARS algorithm, the prediction performance of the final model and the reproducibility of the selected wavelengths were also improved.**

Keywords—Variable Selection; Competitive Adaptive Reweighted Sampling (CARS); Reproducibility; Creditability; Near-Infrared Spectroscopy (NIR)

I. INTRODUCTION

Vibrational spectra, which consisted of overlapping absorption bands, interference from diffuse light scatter, and instrumental noise are usually of high co-linearity [1, 2]. Examples of methodologies giving such complex spectra are near-infrared (NIR) spectroscopy, Raman spectroscopy and Nuclear Magnetic Resonance (NMR) spectroscopy. Typically, the established multivariate calibration model includes all the measured wavelengths. Viewed from a statistical or data analysis perspective, it is really difficult for analytical chemists, even experienced spectroscopists, to determine which wavelengths or combinations should be kept in calibration models [3].Therefore, variable selection methods designed to decrease the aforementioned confusions have drawn considerable attention in quantitative analysis. In the last decade, dozens of inspired techniques appeared on the subject. With well selected variables, more efficient quantitative models can be built, and that may result in significantly reduced computation time, enhanced interpretability and increased robustness.

Methods for variable/wavelength selection can be categorized into two distinctive groups [3, 4]: 1) one is designed to identify the most contributive individual wavelengths or their combinations, and 2) the other is aimed at selecting the most informative spectral intervals or their assemblages. The variable selection criteria are based on the statistics related to the model's performance, e.g. **RMSECV** [5-7], **RPD** [8, 9], or a particular predefined function. The individual wavelength methods rank the contribution of the individual wavelengths according to one or several metrics, directly or indirectly evaluating the prediction performance of the calibration model, and then setting a cutoff criterion to segment informative/uninformative variables. The metrics used for ranking variables are the PLS weight vector **W** [10] , the absolute value of the partial least square (PLS) regression coefficients **B** [1, 11, 12], the posterior probability of the Bayesian method, the variable importance on the projection **VIP** [1, 12], or a combination of these metrics [13]. The cutoff criteria used are determined either based on prior knowledge, or through statistical analysis of uncertainties in the parameters using the Monte Carlo, the Bootstrap, or the Jackknife re-sampling methods. The best combination of these spectral wavelengths was found by Hongdong Li *et al*. [4] using competitive adaptive reweighted sampling (CARS) method. Heuristic search methods, such as the Genetic algorithm (GA) [14, 15], Generalized Simulated Annealing [16, 17] (GSA), ant colony optimization (ACO) [18] were applied to identify the wavelengths most relevant to one or several analytes in the samples.

The interval methods, on the other hand, aim at selecting the most informative spectral bands/intervals. Essentially, intervals are consecutive wavelengths obtained by splitting the spectra into a certain number of units or through iterative construction of these units using, for example, moving windows. The most informative intervals are also assessed by using various metrics and some cutoff strategies. Intervals Partial Least Square (iPLS), one of such interval approaches, was proposed by Nørgaard *et al*. [19]. The spectra was first divided into a certain number of equal width intervals, and then a local PLS model was built on each interval. The model showing the best prediction performance was chosen as the final PLS model. Synergy interval PLS [9] was investigated for finding the combination of spectral intervals which leads to the best PLS model. Bootstrap-VIP [20], on the other hand, was proposed as a method for searching the lower boundary on the number of wavelength intervals. Furthermore, the best

Funding for this project was available through the National Science and Technology Major Project of the Ministry of Science and Technology of China (No. 2010ZX09502-002).

978-1-4673-4713-6/10/$26.00 ©2013 IEEE

combination of spectral intervals was found by Xu *et al*. [21], using Monte Carlo Cross Validation (MCCV) stacked regression.

Generally, most of these methods are computationally less intensive, since they are not formulated as exhaustive search approaches. However, it is argued that the interpretation is still limited, since the variables selected by certain methods are not consistent or sensitive to noise variables [3, 22].

In this paper, a variable selection method, termed VIP-CARS, was proposed to improve the reproducibility of selected wavelengths for PLS regression models. It consists of selecting the variables whose VIP score is significantly greater than a predefined cutoff value and identifying the most relevant wavelengths existing in the reduced data. Removing irrelevant wavelengths prior to modeling is not only interesting from a predictive point of view, but may also help in accelerating the computing speed. Furthermore, the randomly selected wavelengths are reduced compared with those obtained by CARS. This phenomenon was shortly discussed in section 4.2. After the proposed approach was applied to two datasets, corn and Rukuaixiao Tablet, more consistent and efficient calibration models were obtained.

II. METHODS

2.1 Notations

Matrices are represented by bold capital letters, vectors by bold lowercase letters, and scalars by italic characters. The superscript t denotes matrix and vector transpose. The matrix of instrumental responses is denoted by $\mathbf{X}$ *(n×p)*, where *n* and *p* indicate the number of samples and variables/wavelengths, respectively. The measured property is denoted by $\boldsymbol{y}$ *(n×1)*.

2.2 PLS and variable importance on the projection (VIP)

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

Considering the case of one single response $\boldsymbol{y}$ and *p* variables, the structure of the PLS calibration model with *h* latent variables can be expressed as follows:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{t} + \mathbf{E} \tag{1}$$

$$\boldsymbol{y} = \mathbf{T}\boldsymbol{b} + \boldsymbol{f} \tag{2}$$

$$\mathbf{T} = \mathbf{X}\mathbf{W}^{*} \tag{3}$$

$$\mathbf{W}^{*} = \mathbf{W}(\mathbf{P}^{t}\mathbf{W})^{-1} \tag{4}$$

Equation. (1-4), $\mathbf{X}(n\times p)$, $\mathbf{T}(n\times h)$, $\mathbf{P}(p\times h)$, $\boldsymbol{y}(n\times 1)$ and $\boldsymbol{b}(h\times 1)$ respectively represent the predictor matrix(e.g. the instrumental responses), $\mathbf{X}$ scores, $\mathbf{X}$ loadings, the measured responses, and regression coefficients of $\mathbf{T}$. The *k*th element of column vector of $\boldsymbol{b}$ explains the relationship between response $\boldsymbol{y}$ and $\boldsymbol{t}_k$, the *k*th column vector of $\mathbf{T}$. Meanwhile, $\mathbf{E}$ $(n\times p)$ and $\boldsymbol{f}$ $(n\times 1)$ respectively stand for random errors of X and y. PLS-weights $\mathbf{W}$ $(p\times h)$ are obtained to make $\|\boldsymbol{f}\|$ (Euclidian norm) as small as possible [1]. Cross-validation using ten random subsets was generally used to select the number of PLS component (*h*).

The VIP score of a prediction model is a summary of the importance of the projections to h latent variables. The score for the kth variable/wavelengths can be calculated by (5). On the other hand, the average of squared VIP scores equals 1, the "greater than one rule" is generally used as a criterion for variable selection [1, 23].

$$\mathrm{VIP}_j = \sqrt{p\sum_{k=1}^{h}\left(SS(b_k t_k)\left(\frac{w_{jk}}{\|w_k\|}\right)^2\right) / \sum_{k=1}^{h} SS(b_k t_k)} \tag{5}$$

$$SS(b_k t_k) = b_k^2 t_k^t t_k \tag{6}$$

where k=1,2,…,h,; p is the number of columns of $\mathbf{X}$; w_{jk} is the loading weight of the *j*th variable in the *k*th component; b_k, t_k, and w_k are the *k*th elements or vectors of $\boldsymbol{b}$, $\mathbf{T}$, $\mathbf{W}$ respectively.

2.3 Bootstrap-VIP approach

Bootstrap-VIP was first designed to but not limited to select wavelength intervals in spectral imaging applications. The Bootstrap algorithm was used to assess the importance of each wavelength on the predictions of sample quality. Similarly to Monte Carlo simulations, the dataset is randomly re-sampled *N* times with replacement. During each loop, the PLS-Bootstrap algorithm was used to select wavelength based on the variable importance of the projection metric (VIP). In order to ensure that stable bootstrap uncertainty intervals were obtained, re-sampling of 500 [20] times was used. The typical greater-than-one rule was adopted to filter uninformative wavelengths. Given the uncertainty in the VIP metrics, a wavelength was considered relevant when its average VIP value, along with its one standard deviation error bar (obtained from the bootstrap),was above 1.0 .

2.4 Competitive adaptive reweighted sampling (CARS)

CARS was proposed to select the most relevant combination of variables (or wavelengths) during a successive selecting procedure. Based on the regression coefficients obtained by the PLS model, CARS iteratively selects N subsets of variables from N Monte Carlo (MC) sampling processes. During each process, fixed ratios of samples are randomly selected to establish a calibration model. Next, with the regression coefficients obtained, a two-step variable selection

procedure is adopted to select the relevant wavelengths. Finally, cross validation is used to choose the subset (the most relevant combination of wavelengths) showing the lowest root mean square error [4] .The method proceeds as follows:

Step 1: MC sampling

Randomly select k samples ($\boldsymbol{X}_i$,$\boldsymbol{y}_i$), i stands for the ith loop.

Build a PLS model based on the dominating variables *Vsel_old* , then, record the regression coefficients ***beta***:

$$\boldsymbol{beta} = \mathbf{W}^{*}\boldsymbol{b} \tag{7}$$

Step 2: Sort the variables in a descending order according to the absolute value of their regression coefficients. Update the ratio of variables to be kept.

$$\boldsymbol{r}_i = \boldsymbol{a}\boldsymbol{e}^{-ki} \tag{8}$$

where,

$$a = (\frac{p}{2})^{1/(N-1)} \tag{9}$$

$$k = \frac{\ln(p/2)}{N-1} \tag{10}$$

ln denotes the natural logarithm, N represents the Nth sampling process.

The exponent function's trace in Step 2 decreases rapidly in the first stage, whereas in the second stage, the trace progresses gently. This will facilitate the selection process [4].

Step 3: Condense current dataset to have $p\times r_i$ variables. Then draw a subset of variables from the retained $p\times r_i$ variables using adaptively reweighted sampling method, according to a normalized weight w_i.

$$w_i = \frac{|beta_i|}{\sum_{i=1}^{p}|beta_i|}\ ,\ i = 1, 2, 3..., p \tag{11}$$

Essentially, adaptively reweighted sampling method in Step 3 is a weighted sampling algorithm. The variables with larger weights will be selected with higher frequency, and this will accelerate the selection process.

Step 4: Compute RMSECV using *Vsel_new* .Then *Vsel_old* = *Vsel_new*

Step 5: Let i=i+1. If i<N return to step 1, else continue.

Step 6: Choose the subset with the minimum RMSECV as optimal combination of variables/wavelengths and build the final calibration model.

2.5 The VIP-CARS method

The VIP-CARS approach consists of removing irrelevant wavelengths using Bootstrap-VIP first and then identifying the best combination of wavelengths by CARS. Details of the VIP-CARS procedure are provided below.

Firstly, the VIP scores of each wavelength are calculated on the PLS model constructed using the dataset re-sampled by Bootstrap algorithm. Then, the wavelengths with VIP scores lower than a predefined threshold are removed. Next, the regression coefficients beta (7) are estimated with k samples ($\mathbf{X}_i$,$\boldsymbol{y}_i$) randomly selected from the reduced dataset. According to the updated ratio r_i and weight $\boldsymbol{w}_i$ (8-11), a subset of variables can then be drawn from the retained $p\times r_i$ variables using adaptively reweighted sampling method. The aforementioned sophisticated selection procedures, *viz.* step 1 to 4 of CARS in section 2.4, are repeated for N times. Ultimately, the calibration model is built on the wavelengths whose combination shows the minimum RMSECV.

Since the VIP threshold would naturally affect the final selected wavelengths, and it was indicated that the cutoff criterion should be a function of the data structure [1], therefore 10 cutoff points linearly spaced between and including 0.83 and 1.21 were investigated in this study.

III. EXPERIMENTAL

Two datasets were employed in this study to investigate the performance of the proposed method. One experiment was the protein content in corn and the other was the thickness of coating layer for Rukuaixiao Tablet. Without extra description, each dataset was mean-centered prior to further investigation.

3.1 Corn dataset

The data set [24] consisted of 80 NIR spectra, each corresponding to an independent corn sample, covering the spectral range 1100-2498 nm at 2 nm intervals. The spectral data set measured on m5 instrument and the protein content was used to assess the performance of the previous mentioned methods. The original spectra of the corn data is shown in Fig. 1 (a). The Bootstrap-VIP, VIP-CARS and CARS algorithms as well as the traditional PLS methods were performed on this data for comparison.

3.2 Rukuaixiao Tablet data set

The data set contained a total of 104 NIR spectra of 104 coated tablets from 3 production batches at seven sampling points (3 tablets per point) and a batch of authorized tablets, which were kindly supplied by Pharmaceutical factory of Beijing University of Chinese Medicine. The NIR spectra of the samples were collected at 4 cm^{-1} interval over the spectral range 4000-10,000 cm^{-1} with an Antaris FT-NIR System (Thermo scientific, Madison, USA) equipped with an integrating sphere system. Each sample was analyzed in sextuplet, with spectra obtained by averaging 32 scans and equilibrated at 25 ^{0}C for 10 min before scanning. The thickness of the coating layer was measured by a caliper. The raw spectra of the Rukuaixiao Tablet data are plotted in Fig. 1(b). The 104 × 1557 data matrix was adopted to compare the

prediction power of the aforementioned methods. Furthermore, the Rukuaixiao Tablet dataset was also used to test the robustness and the reproducibility of both CARS and VIP-CARS algorithms.

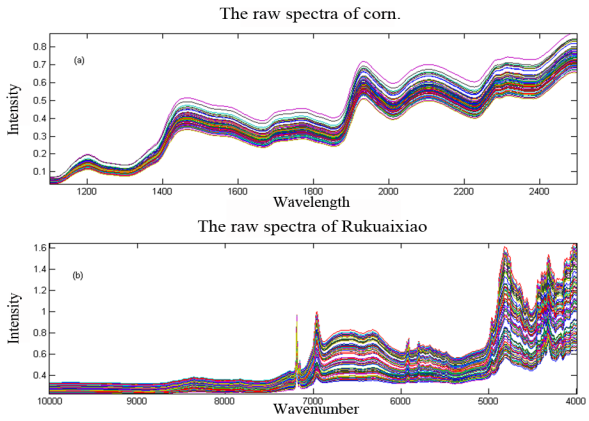

Figure 1. The raw NIR spectra of corn dataset (a) and Rukuaixiao Rukuaixiao Tablet dataset (b).

All calculations were performed on a personal computer i7 880 processor with 6GB RAM under the Win7 Professional operating system using Matlab 7.9 (Mathworks, Inc., Natick, MA). The Bootstrap-VIP routines were implementations of well-established algorithms. The CARS and VIP-CARS algorithms were obtained from or modifications of functions in the CARS toolbox (http://code.google.com/p/carspls/).

IV. RESULTS AND DISCUSSIONS

4.1 Parameter setup

There are certain numbers of critical parameters that needed to be optimized, including the re-sampling times, cutoff value, etc. The RMSECV curves of the PLS calibration models for the Corn protein and Rukuaixiao Tablet datasets were found to level off around 10 and 5 LVs, respectively. Therefore, in order to compare with the full spectrum PLS, the upper limit of the number of latent variables was set to 10 on the corn protein dataset and 5 on the Rukuaixiao Tablet dataset, for both CARS and VIP-CARS approaches. In the VIP-CARS procedure, Bootstrap re-sampling 50 times was found to yield very consistent results. The optimizations of the other parameters are discussed at large in the following sections.

4.1.1 Cutoff values of the VIP-CARS method

The cutoff value of zero to be used is not absolute to assess the significance of regression coefficient. In fact, it has been suggested that a proper cutoff value may vary between 0.8 and 1.2 [1]. Cutoff values, linearly spaced between and including 0.83 and 1.21, were considered in this section to explore the impact of the VIP thresholds on the proposed VIP-CARS approach. For each threshold, 100 times Monte Carlo re-sampling and 500 re-duplicate running of VIP-CARS were executed on each dataset and the RMSECV values were recorded. Such a large number of subsets and repeated runs were selected to ensure that a stable result can be obtained on each dataset. A comparison is provided in Fig. 2 between the VIP-CARS selection with a threshold of 1.0 and the other cutoff values. Even if the parameters of CARS procedure were fixed, minimizing RMSECV is still a 2-dimensional optimization problem which requires selecting both the number of latent variables of the reduced model and the optimal VIP cutoff. To investigate the impact of VIP thresholds on selected wavelengths, the number of latent variables for the pre-selection step (Bootstrap-VIP) was kept the same.

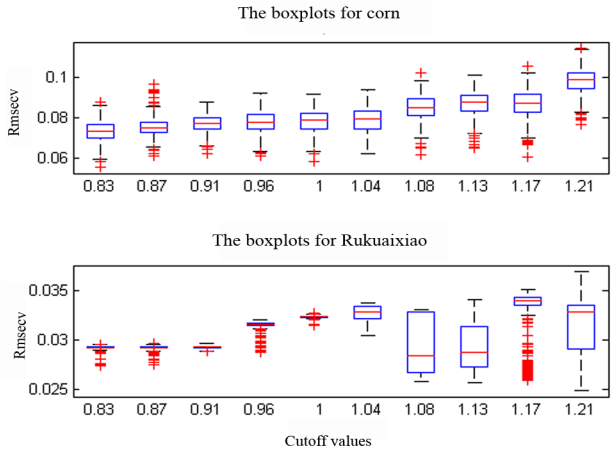

Figure 2. The box-plots of VIP-CARS for each dataset with the cutoff value 0.8300, 0.8722, 0.9144, 0.9567, 0.9989, 1.0411, 1.0833, 1.1256, 1.1678, 1.21, respectively. The selection process was executed for 500 times on each cutoff value.

The results presented in Fig. 2 show that the optimal VIP cutoff values vary between 0.8 and 1.2, providing marginal improvements on RMSECV values compared to the 1.0 threshold. These observations indicate that a VIP cutoff value of 1.0 is a good starting point, but can later be fine-tuned if necessary. The cutoff value in this study was set to 0.83 on corn and 0.91 on Rukuaixiao Tablet, based on the RMSECV and its variance at different cutoff values.

The dependence of the VIP cutoff on the scaling of spectral data needs to be further investigated. The greater-than-one rule has a particular meaning for autoscaled data, which may not hold for other scaling techniques [20]. However, after preparing the analysis of datasets using mean-centering only, it was found that the VIP =1.0 threshold was still a good initial value (RMSECV = 0.0703 on corn). Similar results were obtained in terms of both the predictive power and the selected spectral ranges.

4.1.2 The size of calibration subset and the number of Monte Carlo re-sampling runs

There are two other parameters required to be investigated, i.e. the randomly selected k samples, and the N Monte Carlo

sampling runs. With the purpose of finding out the adapted model for the data set, the VIP-CARS was used in different cases as N = 50, 100, 200, 500 and k = 0.5n, 0.6n, 0.7n, 0.8n, 0.9n, n (n, the size of the dataset). In order to gain a statistical perspective of the proposed method, the VIP-CARS procedure was repeated for 500 times.

The root-mean-square errors of 10 fold cross-validation (RMSECVs) for VIP-CARS in all cases, together with their variances are shown in Table 1. It can be seen from the table that both the value of N and k have certain influence on the RMSECVs calculated by the VIP-CARS procedure. The RMSECVs calculated on corn data set decrease gradually as the number of re-sampling runs increase from 50 to 500. But as for the Rukuaixiao Tablet data set, this does not present much variation except for the expected standard deviation of RMSECV. It should further be noted that the re-sampling operation of CARS is different from that of Bootstrap-VIP since larger values of Bootstrap re-sampling times efficiently treat unbalanced and non-smooth subsets at the expense of increased computation time [20]. But that is not the case with CARS. Larger values of Monte Carlo re-sampling times affect the exponent function mostly in Step 2 of the CARS algorithm.

TABLE I. THE MEAN AND STANDARD DEVIATION OF RMSECV OBTAINED BY **VIP-CARS** WITH DIFFERENT N AND K FOR DATA SETS CORN PROTEIN AND RUKUAIXIAO TABLET.

	k=0.5n	**k=0.6n**	**k=0.7n**	**k=0.8n**	**k=0.9n**	**k=n**
Corn						
N=50	0.0720±0.0069	0.0728±0.0058	0.0728±0.0051	0.0720±0.0053	0.0717±0.0050	0.0701±0.0043
N=100	0.0732±0.0058	0.0720±0.0049	0.0712±0.0057	0.0700±0.0049	0.0703±0.0045	0.0698±0.0047
N=200	0.0726±0.0064	0.0705±0.0063	0.0702±0.0055	0.0699±0.0052	0.0695±0.0050	0.0703±0.0043
N=500	0.0725±0.0062	0.0711±0.0059	0.0701±0.0058	0.0692±0.0058	0.0692±0.0049	0.0701±0.0050
Rukuaixiao						
N=50	0.0293±0.0003	0.0292±0.0002	0.0292±0.0002	0.0292±0.0002	0.0292±0.0002	0.0293±0.0002
N=100	0.0292±0.0002	0.0292±0.0002	0.0292±0.0002	0.0293±0.0001	0.0293±0.0001	0.0293±0.0001
N=200	0.0292±0.0002	0.0292±0.0002	0.0293±0.0002	0.0293±0.0001	0.0293±0.0001	0.0293±0.0001
N=500	0.0292±0.0002	0.0292±0.0002	0.0293±0.0002	0.0293±0.0001	0.0293±0.0001	0.0293±0.0001

As the size of calibration subsets vary from 0.5n to n, the decreasing trend of RMSECV becomes clear. The RMSECV reaches its minimum at k=0.8n and 0.9n. Therefore, a size of 0.9n, which was adopted in the original CARS methods, is a good initial point for both data. In the following study, the number of MC sampling runs was set to 500

4.2 Comparision of the variables selected by CARS and VIP-CARS

One objective of this study was to further indicate that stability of the proposed methods would lead to the identification of reproducible, rather than occasional, relevant variables. Given the reproducibility of the wavelengths selected, they can not only be used by spectroscopists, but additionally, their interpretation becomes easier compared with the result obtained by the full spectrum PLS. In order to inspect in detail the uncertainty caused by the Monte Carlo sampling, the methods were repeated for 500 times on the two datasets, and the resulting wavelengths were recorded. The results obtained on the two distinct datasets can be used to illustrate whether the results obtained with the proposed approaches, and the comparison studies with other methods, yield consistent conclusions across different properties and data structures.

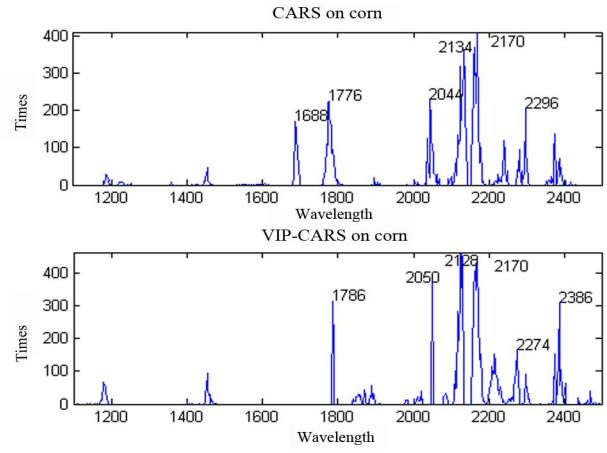

Figure 3. The frequency of each wavelength selected by running CARS and VIP-CARS 500 times for corn data set.

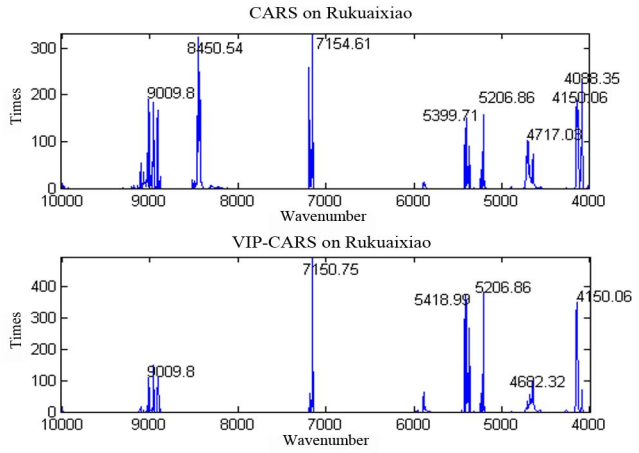

Figure 4. The frequency of selected wavenumbers by appling CARS and VIP-CARS 500 times on Rukuaixiao data set.

The specific wavelengths selected by CARS and VIP-CARS are shown in Fig. 3 and Fig. 4 for all of the two datasets. As is illustrated in Fig. 3, performance of each method is different across the data sets. The wavelengths selected by VIP-CARS spread across the full spectrum; however, only a limited part of wavelengths were selected. Besides, the variable-wise frequency peak of the dominant wavelengths may be an indication of the high complexity of NIR spectra. The performance of CARS approach is similar to VIP-CARS except that the dominant wavelengths are at around 1688nm. However, the results shown in Fig. 3 and Fig. 5 illustrate that a certain number of dominant wavelength peaks are of higher amplitude compared with the results obtained by CARS, including the peaks around 1786, 2050, 2134, and 2170 nm. Moreover, these wavelengths selected on the corn data can be attributed to the $CONH_2$ and $CONH_2(R)$ chemical groups, which are critical features of protein molecules. It means that the reproducibility of the relevant dominant peaks selected using VIP-CARS improved significantly. The variables selected by CARS in Fig. 4 are mainly distributed in six regions. The wave numbers selected by both approaches, at around 7150 cm^{-1}, are of high frequencies. Fig. 6 explicitly demonstrates the improvement in the reproducibility of the dominant peaks. The results presented above indicate that, compared with the results obtained by CARS, the reproducibility of wavelengths selected by VIP-CARS is significantly improved.

The number of latent variables and wavelengths selected by the investigated methods, together with its RMSECV, are shown in Table 2. It is clear that VIP-CARS can reduce the number of selected wavelengths and leads to apparently improved prediction power compared with full spectrum PLS regression. Moreover, the prediction performance of VIP-CARS on corn protein data set is far better than that of CARS. In addition, it should be noted that the RMSECV values of Bootstrap-VIP decrease significantly compared with the full spectrum PLS regression. That means Bootstrap-VIP is not only a good selection method for the interpretation phase, but also a reliable method to improve the predictive power.

In general, compared to just CARS, VIP-CARS produces more reproducible and credible wavelengths.

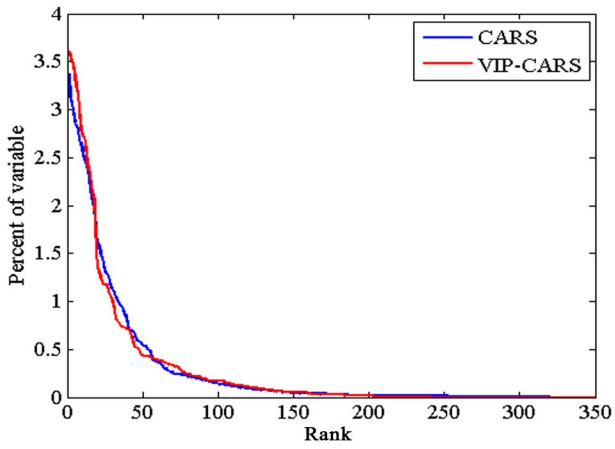

Figure 5. Summary of the percentage of each variable ranked descend for the Corn dataset.

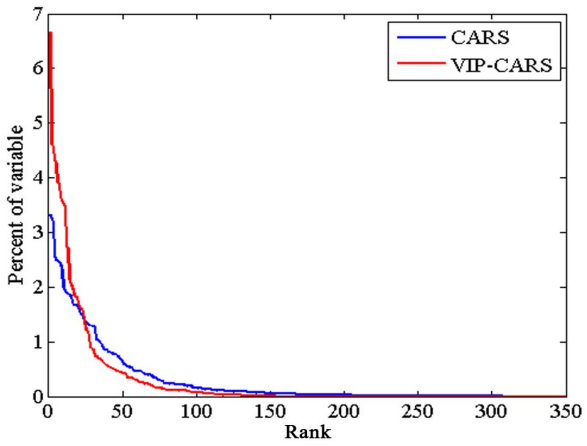

Figure 6. Summary of the percentage of selected wavenumbers for the Rukuaixiao dataset.

V. CONCLUSIONS

In this publication, VIP-CARS was proposed as a variable selection method. The performance of the proposed approach was evaluated by comparing the prediction ability of the resulting PLS model, the reproducibility of the selected wavelengths and the robustness to noise variable, to those obtained by the plain CARS. The results of application to two datasets indicate that more robust informative wavelengths can be obtained by VIP-CARS method. The impact of Monte

Carlo sampling on the selected wavelengths was reduced to some content. This means that although the VIP-CARS is essentially a modification of CARS, it leads to significantly improved wavelength reproducibility, creditability and robustness. Furthermore, with the implementation of the VIP-CARS algorithm, improvement was also observed in the prediction performance of the final model.

Analytical chemists will very likely benefit from the VIP-CARS method, especially when hundreds of variables are involved.

TABLE II. THE RMSECV AND NUMBER OF RETAINED VARIABLES OBTAINED BY **PLS**, **BOOTSTRAP-VIP**, **CARS**, **VIP-CARS** ON BOTH CORN PROTEIN DATASET AND RUKUAIXIAO TABLET DATASET.

METHOD	CORN[A]			RUKUAIXIAO[B]		
	LVS	NVAR	RMSECV	LVS	NVAR	RMSECV
PLS	10	700	0.1212	5	1557	0.0361
BOOTSTRAP-VIP	10	274	0.0912	5	431	0.0331
CARS	9±1	24±11	0.0723*	4±1	20±17	0.0286*
VIP-CARS	9±1	27±8	0.0687*	4±1	18±7	0.0288*

* The median of RMSECVs are present here to provide a robust estimate on the behavior of investigated methods.

[A] The maximum number of latent variables was set to 10.

[B] The maximum number of latent variables was set to 5.

REFERENCES

[1] I.-G. Chong and C.-H. Jun, "Performance of some variable selection methods when multicollinearity is present," Chemometrics and Intelligent Laboratory Systems, vol. 78, pp. 103-112, 2005.

[2] A. L. Xia, H.-L. Wu, Y. Zhang, S.-H. Zhu, Q.-J. Han, and R.-Q. Yu, "A novel efficient way to estimate the chemical rank of high-way data arrays," Analytica Chimica Acta, vol. 598, pp. 1-11, 2007.

[3] L. P. Brás, M. Lopes, A. P. Ferreira, and J. C. Menezes, "A bootstrap-based strategy for spectral interval selection in PLS regression," Journal of Chemometrics, vol. 22, pp. 695-700, 2008.

[4] H. Li, Y. Liang, Q. Xu, and D. Cao, "Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration," Analytica Chimica Acta, vol. 648, pp. 77-84, 2009.

[5] Y. P. Du, S. Kasemsumran, K. Maruo, T. Nakagawa, and Y. Ozaki, "Ascertainment of the number of samples in the validation set in Monte Carlo cross validation and the selection of model dimension with Monte Carlo cross validation," Chemometrics and Intelligent Laboratory Systems, vol. 82, pp. 83-89, 2006.

[6] S. Kasemsumran, Y. P. Du, K. Maruo, and Y. Ozaki, "Selective removal of interference signals for near-infrared spectra of biomedical samples by using region orthogonal signal correction," Analytica Chimica Acta, vol. 526, pp. 193-202, 2004.

[7] A. U. Vanarase, M. Alcalà, J. I. Jerez Rozo, F. J. Muzzio, and R. J. Romañach, "Real-time monitoring of drug concentration in a continuous powder mixing process using NIR spectroscopy," Chemical Engineering Science, vol. 65, pp. 5728-5733, 2010.

[8] B. Igne, J.-M. Roger, S. Roussel, V. Bellon-Maurel, and C. R. Hurburgh, "Improving the transfer of near infrared prediction models by orthogonal methods," Chemometrics and Intelligent Laboratory Systems, vol. 99, pp. 57-65, 2009.

[9] D. Wu, Y. He, P. Nie, F. Cao, and Y. Bao, "Hybrid variable selection in visible and near-infrared spectral analysis for non-invasive quality determination of grape juice," Analytica Chimica Acta, vol. 659, pp. 229-237, 2010.

[10] F. Lindgren, P. Geladi, S. Rännar, and S. Wold, "Interactive variable selection (IVS) for PLS. Part 1: Theory and algorithms," Journal of Chemometrics, vol. 8, pp. 349-363, 1994.

[11] C. Tistaert, B. Dejaegher, N. Nguyen Hoai, G. Chataigné, C. Rivière, V. Nguyen Thi Hong, et al., "Potential antioxidant compounds in Mallotus species fingerprints. Part I: Indication, using linear multivariate calibration techniques," Analytica Chimica Acta, vol. 649, pp. 24-32, 2009.

[12] N. Sorol, E. Arancibia, S. A. Bortolato, and A. C. Olivieri, "Visible/near infrared-partial least-squares analysis of Brix in sugar cane juice: A test field for variable selection methods," Chemometrics and Intelligent Laboratory Systems, vol. 102, pp. 100-109, 2010.

[13] J. P. A. M. R.F. Teófilo, M.M.C. Ferreira, "Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression," Journal of Chemometrics vol. 23, pp. 32-48, 2009.

[14] T. Li, H. Mei, and P. Cong, "Combining nonlinear PLS with the numeric genetic algorithm for QSAR," Chemometrics and Intelligent Laboratory Systems, vol. 45, pp. 177-184, 1999.

[15] "Genetic Algorithm Applied to Selection of Wavelength in Partial Least Squares for Simultaneous Spectrophotometric Determination of Nitrophenol Isomers," Analytical Letters, vol. 39, pp. 2359-2372, 2006.

[16] E. Llobet, O. Gualdrón, M. Vinaixa, N. El-Barbri, J. Brezmes, X. Vilanova, et al., "Efficient feature selection for mass spectrometry based electronic nose applications," Chemometrics and Intelligent Laboratory Systems, vol. 85, pp. 253-261, 2007.

[17] Z. Xiaobo, Z. Jiewen, M. J. W. Povey, M. Holmes, and M. Hanpin, "Variables selection methods in near-infrared spectroscopy," Analytica Chimica Acta, vol. 667, pp. 14-32, 2010.

[18] M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad, and M. Akhond, "Ant colony optimisation: a powerful tool for wavelength selection," Journal of Chemometrics, vol. 20, pp. 146-157, 2006.

[19] L. Norgaard, A. Saudland, J. Wagner, J. P. Nielsen, L. Munck, and S. B. Engelsen, "Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy," Appl. Spectrosc., vol. 54, pp. 413-419, 2000.

[20] R. Gosselin, D. Rodrigue, and C. Duchesne, "A Bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications," Chemometrics and Intelligent Laboratory Systems, vol. 100, pp. 12-21, 2010.

[21] L. Xu, J.-H. Jiang, Y.-P. Zhou, H.-L. Wu, G.-L. Shen, and R.-Q. Yu, "MCCV stacked regression for model combination and fast spectral interval selection in multivariate calibration," Chemometrics and Intelligent Laboratory Systems, vol. 87, pp. 226-230, 2007.

[22] R. Singh, K. V. Gernaey, and R. Gani, "Model-based computer-aided framework for design of process monitoring and analysis systems," Computers & Chemical Engineering, vol. 33, pp. 22-42, 2009.

[23] S. Gaudet, K. A. Janes, J. G. Albeck, E. A. Pace, D. A. Lauffenburger, and P. K. Sorger, "A Compendium of Signals and Responses Triggered by Prodeath and Prosurvival Cytokines," Molecular & Cellular Proteomics, vol. 4, pp. 1569-1590, October 1, 2005 2005.

[24] http://www.eigenvector.com/data/Corn/index.htm

[25] A. Gustavo González and M. Ángeles Herrador, "A practical guide to analytical method validation, including measurement uncertainty and accuracy profiles," TrAC Trends in Analytical Chemistry, vol. 26, pp. 227-238, 2007.