Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 210 (2019) 362–371

Contents lists available at ScienceDirect

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

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

Short Communication

A modification of the bootstrapping soft shrinkage approach for spectral variable selection in the issue of over-fitting, model accuracy and variable selection credibility

Hong Yan [a], Xiangzhong Song [a], Kuangda Tian [a], Jingxian Gao [a], Qianqian Li [b], Yanmei Xiong [a,*], Shungeng Min [a,*]

[a] College of Science, China Agricultural University, Beijing 100193, PR China
[b] School of Marine Science, China University of Geoscience, Beijing 100083, PR China

ARTICLE INFO

Article history:
Received 13 August 2018
Received in revised form 4 October 2018
Accepted 20 October 2018
Available online 24 October 2018

Keywords:
Variable selection
Stabilized bootstrapping soft shrinkage approach
Stability of RC
Credibility
Over-fitting

ABSTRACT

In this study, we proposed a new computational method stabilized bootstrapping soft shrinkage approach (SBOSS) for variable selection based on bootstrapping soft shrinkage approach (BOSS) which can enhance the analysis of chemical interest from the massive variables among the overlapped absorption bands. In SBOSS, variable is selected by the index of stability of regression coefficients instead of regression coefficients absolute value. In each loop, a weighted bootstrap sampling (WBS) is applied to generate sub-models, according to the weights update by conducting model population analysis (MPA) on the stability of regression coefficients (RC) of these sub-models. Finally, the subset with the lowest RMSECV is chosen to be the optimal variable set. The performance of the SBOSS was evaluated by one simulated dataset and three NIR datasets. The results show that SBOSS can select the fewer variables and supply the least RMSEP and latent variable number of the PLS model with the best stability comparing with methods of Monte Carlo uninformative variables elimination (MCUVE), genetic algorithm (GA), competitive reweighted sampling (CARS), stability of competitive adaptive reweighted sampling (SCARS) and BOSS.

1. Introduction

During the last few decades, variable selection, regarded as a statistic tool which plays an essential place in quantitative analysis of near infrared (NIR), mid infrared (MIR) and Raman spectroscopy, aims to give the minimum prediction errors of established models, meanwhile, the relations between variables and the properties of samples can be illuminated according to it [1].

In NIR, absorption bands of fundamental frequency vibrations and combination of vibrations make it possible for quantitative analysis. However, adverse issues are also inevitable, as absorption bands are usually overlapping. Therefore, suitable chemometrics algorithm is necessary to deal with NIR spectrum, with the purpose to eliminate the uninformative variables effectively by using variable selections.

Partial least squares regression (PLS) can process large numbers of noisy and unrelated variables, which has been commonly used in multivariate regression analysis. Meanwhile, the issue that elimination of uninformative variables can improve the performance of PLS models has been demonstrated with theoretical [2–5] and experimental evidences [6–13]. Recently, researchers began to employ different variable selection methods to assess their performance [12,14–22]. The results have shown that variable selection has great improvement in accuracy and robustness in quantitative model, and the relationship between variables and chemical information can be solidified more satisfactory [21].

A family of methods of variable selection grounded on PLS regression have been developed successfully [23]. Some of them focus on selecting combination of single variables with good performance, such as uninformative variables elimination (UVE) [2], variable importance projection (VIP) [24] etc., another of which are mostly based on model population analysis (MPA) [25], such as iteratively retains informative variables (IRIV) [26], variable iterative space shrinkage approach (VISSA) [22], variable combination population analysis (VCPA) [27], bootstrapping soft shrinkage approach (BOSS) [28] and margin influence analysis (MIA) [26]. Also, some of which consider statistical features of the variables, e.g. successive projection algorithm (SPA) [29], random frog [19], competitive reweighted sampling(CARS) [1], stability of competitive adaptive reweighted sampling (SCARS) [21] and Bayesian linear regression (BLR) [30], while others aim at seeking the best combination on spectral intervals, e.g. moving windows partial least square (MWPLS) [31], interval partial least square (iPLS) [32] and interval random frog (iRF) [33].

Recently, a new variable selection method named bootstrapping soft shrinkage approach (BOSS) has been developed by Deng et al. in 2016

* Corresponding authors.
E-mail addresses: xiongym@cau.edu.cn (Y. Xiong), minsg@cau.edu.cn (S. Min).

https://doi.org/10.1016/j.saa.2018.10.034
1386-1425/

[19]. The core parts of BOSS are weighted bootstrap sampling (WBS) and MPA. The absolute value of RC is criterion of variables.

This method has shown significant improvement of prediction ability compared with other high performing selection methods such as Monte Carlo uninformative variable elimination (MCUVE) [2], competitive adaptive reweighted sampling (CARS) [1] and genetic algorithm PLS (GA-PLS) [28], but the model credibility of BOSS was still in need of further improving. In our research, stability of RC should be taken into account as an index, such as MCUVE and SCARS. Therefore, we tried to modify BOSS with new index as the stability of RC because RC are diverse in different models. In all existing methods, loading weights, RC and variable importance in projection are common filter measures for the variable importance. By introducing this criterion, both the model accuracy and credibility were enhanced, also, the problem of over-fitting was fixed.

In practice, SBOSS confirmed the significance of each variable and then found out the optimal combination of variables. To evaluate SBOSS, we applied this data-driven approach on one simulated dataset and three NIR datasets, and for comparison, five variable methods, such as GA, MCUVE, CARS, SCARS and BOSS, were involved in.

In following parts, it would be demonstrated how the stability of RC was used as an index to extract the optimal variable combination, how this filter improved the accuracy and stability apparently in calibration modeling, how the selected variables corresponded to the chemical information with uncorrelated contents, and how the over-fitting in PLS model was avoided by the combination of MPA, CV and stability of RC.

2. Theory and Algorithm

2.1. Stability of RC

The stability of a variable was used to optimize BOSS. In PLS, spectral data matrix **X** contains p variables in columns and n samples in rows. Vector **y** with order $n \times 1$ denotes the measured property of interest. The equation of relationship of X and y is below:

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

The β includes p RC, E was error vector. After N sampling runs, a matrix B ($[\beta_1, \beta_2, \ldots \beta_N]$) containing N corresponding regression coefficient vectors can be obtained. The stability should be defined as:

$$c_k = \left|\overline{b_k}/s(b_k)\right| \tag{2}$$

The c_k is the stability of kth variable in N sampling runs. $\overline{b_k}$ is the average value, $s(b_k)$ is the standard deviation of kth variable in N sampling runs. c_k can indicate the importance of kth variable and its high ability in modeling.

2.2. Bootstrap Sampling (BSS) and Weighted Bootstrap Sampling (WBS)

BSS is a statistical technique for random sampling with replacement. All the objects have the same possibility to be selected in each test. Different with BSS, WBS hires different weights for objects. In SBOSS, both BSS and WBS are used. More details refer to related articles [28].

2.3. The Element of MPA (Model Population Analysis)

MPA could be considered as a general framework for developing new methods by statistically analyzing some interesting parameters (RC, prediction errors, etc.) of a number of sub-models. The generation of large population models is based on the sampling method. Monte Carlo sampling (MCS) is often used in the sample space and variable space. The procedure of MPA is below:

(1) Generate a sub dataset by MCS, bootstrap sampling and binary matrix sampling are both used as to draw datasets in the context of MPA. (2) Establishing a sub-model for each sub-dataset. (3) Analyze the outputs of all the sub-models to extract some information, this is the most important point of MPA.

RMSECV of 5-fold cross validation is regarded as the assessment of the model. The ratio of best models is defined as σ, the best sub-models with the best σ have the lowest RMSECV.

2.4. The Stabilized Bootstrapping Soft Shrinkage (SBOSS) Method

With the N sampling runs, SBOSS will recurrently select N subsets. Briefly, SBOSS operate in four steps in each sampling run: (1) BSS is used to generate subsets. (2) PLS sub-models are built and find the best models with MPA. (3) New weights for variables are obtained. (4) Due to the new weights, WBS is applied to produce new subsets.

The same with BOSS, the calculation of SBOSS also has many loops. WBS and MPA are applied remain informative variables. The details can be found in reference [28].

After BSS sampling, the variables were indexed by the stability of variables. PLS sub-models are built. RC are calculated for each extracted model. Change all the elements in regression vector to absolute value and normalize each regression vector to have unit length. Sum up the normalized regression vector to obtain new weights for variables. Sum up the normalized stability regression vector to generate new weights for variables.

$$w_i = \sum_{k=1}^{K} c_{i,k} \tag{3}$$

K is the number of sub-models, $c_{i,k}$ is the absolute value of normalized stability of regression coefficient for variable i in the kth sub-models.

After that, new subsets are created by WBS with new weights. And all the loops except BSS sampling are repeated until the number of variables in the new subsets equals to 1. In the end, the subset with the lowest RMSECV is chosen to be the optimal variable set. The flowchart is demonstrated in Fig. 1.

2.5. A Brief Introduction of the Compared Methods

2.5.1. MCUVE

The linear calibration model is expressed in the following:

$$c_j = \frac{\beta_j}{s(\beta_j)} j = 1, 2, \ldots . p \tag{4}$$

where β is the regression coefficient vector, X represents observation matrix, y the response vector, e symbolizes the random error vector and E (e) and Cov (σ) denote the expectation and covariance, respectively.

The UVE-PLS procedure is based on analyzing the β RC in Eq. (4), in which the original variables is added by an equal number of random variables with very small range (about 10^{-10}). The stability criterion c is defined by

$$c_j = \frac{\beta_j}{s(\beta_j)} j = 1, 2, \ldots . p \tag{5}$$

$$s(\beta_j) = \left(\sum_{i=1}^{n} \frac{(\beta_{ij} - \beta_j)^2}{n-1} \right)^{1/2} \tag{6}$$

where c_j is utilized on the conjunction of the addition of random variables and the original data, β_j stands for the RC of the jth wavelength variable when leaving out the ith calibration sample, and n is the

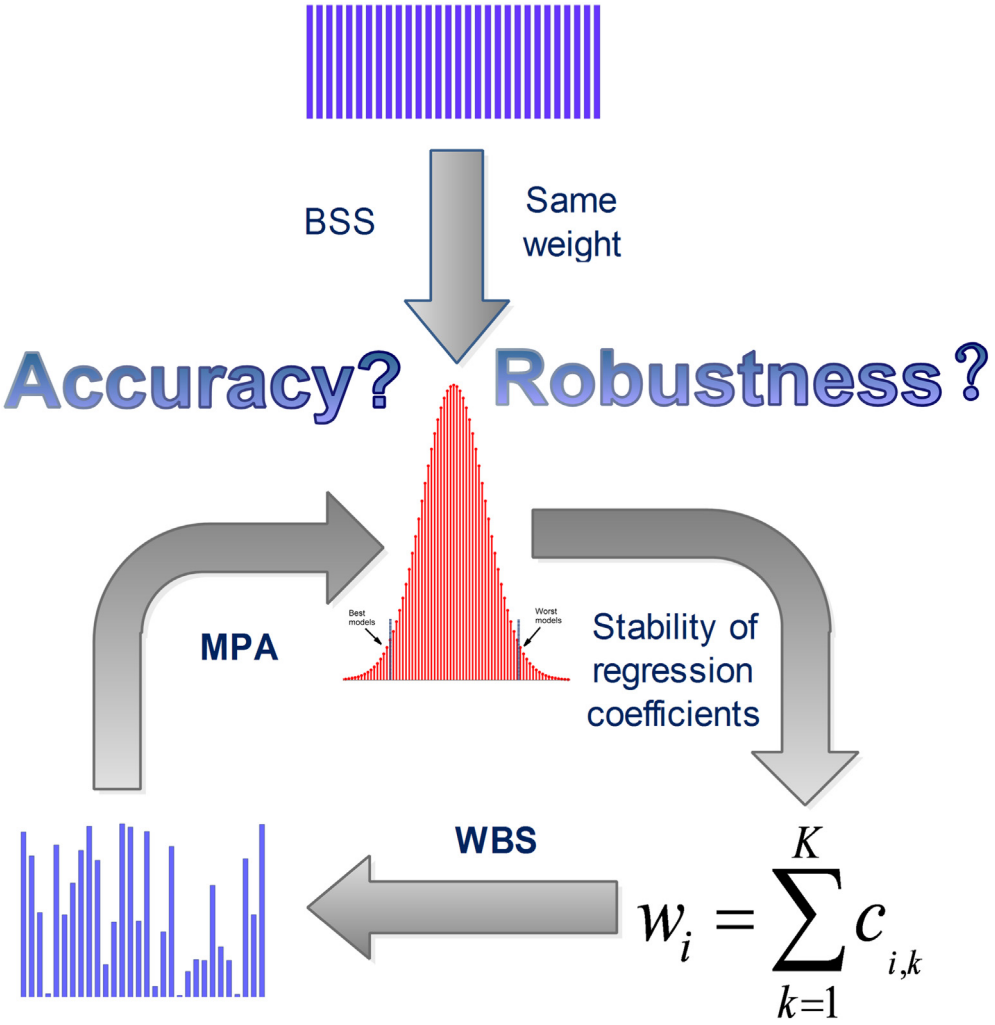

Fig. 1. The flowchart of SBOSS.

number of calibration samples. β_j and $s(\beta_j)$ denote respectively the mean and the standard deviation of all β_{ij} for the jth wavelength variable, and β_{ij} is achieved by the leave-one-out method.

The threshold of the elimination of uninformative variables is obtained by the following equation:

$$|(c_j)| < |\max(c_{artif})| \tag{7}$$

where (c_j) is the stability criterion for the jth wavelength in the original data; and $|\max(c_{artif})|$ is the absolute value of the maximum value for (c_j) from the added random variables. The modification of UVE, which gave an objective cut-off level for the stability criterion c in the form of jack-knife estimated standard error, has been proposed to enhance the interpretability of the results [34].

In the MCUVE, Monte Carlo strategy is introduced to UVE instead of leave-one-out strategy: random choosing M samples from all the calibration samples to build PLS models for calculating the regression coefficient β, then repeating the procedure for N times. So Eq. (6) becomes the following:

$$s(\beta_j) = \left(\sum_{i=1}^{N} \frac{(\beta_{ij} - \beta_j)^2}{N-1} \right)^{1/2} \tag{8}$$

Here, β_{ij} symbolizes the regression coefficient of the jth wavelength in PLS model, which is built by the ith M random chosen samples. In practice, by means of Monte Carlo method the amount of computation complexity can be reduced substantially.

2.5.2. GA-PLS

GA was proposed by Lucasius and Kateman according to the 'Darwin's classical rules' [35], which is mainly ruled by the 'struggle of life'.

(1) Define the parameters of the GA-PLS:The parameters of GA-PLS in this paper are listed in Table 1 (Supplementary material).
(2) Initiation of population: each chromosome in the population is row vector containing as many genes as there are variables, each gene being coded as 1 if the corresponding variable was selected and 0 if not. The structure of each chromosome is determined in a totally random way. Of note, each chromosome would be checked to avoid having the same structure in the population.
(3) Evaluation of the response: based on the variables selected by each chromosome, a number of subset data could be extracted from the full data. The larger the value of the cross-validated explained variance, the better the chromosome.
(4) Crossover and mutation: in order to generate two new chromosomes of the existing population is randomly selected to carry out crossover and mutation approach as well as to evaluate the cross-validated explained variance of the new offspring. At this step, two new chromosomes should be also checked to avoid containing the same variables.

(5) Update the population by comparing the CVEV of the two new chromosomes with the one of the existing chromosomes of the current population. The updating rule is that each chromosome of the new offspring would survive if it is better than the worst chromosome which would be discarded late.
(6) Go back to Step (4) when the amount of the evaluations does not satisfy the criterion of the entrance of backward selection. When it is satisfied, backward selection is conducted to choose the best subset of the population.
(7) If the criterion of the final termination is reached, the whole evolution process of GA has ended. If not, go back to Step (4). As we can see the parameters from the Table 1 (Supplementary material), the amount of evaluations is set to 200. Thus, only if the evaluations reach 200, the GA run is terminated.
(8) After processing the predefined runs, the selection frequency of each variable could be obtained. Rank the variables by the selection frequencies, and then choose the optimal subset with the maximum CVEV according to the ranking of variables.

2.5.3. CARS

CARS is based on absolute RC to evaluate the importance of variables. Monte Carlo is employed for sampling. The exponentially decreasing function (EDF) is then employed to enforce feature selection, removing variables with small absolute RC by force. In the ith sampling run, the ratio of wavelengths to be kept is computed using an EDF defined as:

$$r_i = ae^{-ki} \tag{9}$$

where a and k are two constants determined by the following two conditions: (I) int the first sampling run, all the p wavelength are taken for modeling which means that $r_1 = 1$, (II) in the Nth sampling run, only two wavelengths are reserved such that we have $r_N = 2/p$. With the two conditions, a and k can be calculated as:

$$a = (p/2)^{1/(N-1)} \tag{10}$$

$$k = \ln(p/2)/(N-1) \tag{11}$$

where ln denotes the natural logarithm.

Consecutively, adaptive reweighted sampling (ARS) is performed to realize a competitive feature selection based on the RC. This step mimics the 'survival of the fittest' principle which is the basis of Darwin's Evolution Theory.

Finally, cross validation is adopted to select the subsets according to the lowest RMSECV.

Both CARS and BOSS are ground on large absolute RC to evaluate the importance of each variable. The variables with larger absolute RC have great opportunities to be selected. In CARS, Monte Carlo strategy is adopted for sampling. The exponentially decreasing function (EDF) is then employed to enforce feature selection, and small absolute RC are removed. Continuously, adaptive reweighted sampling (ARS) is applied to select the key variables [36]. Eventually, the subset will be selected by cross validation with the lowest RMSECV.

SCARS is a method based on CARS, a more informative criterion, i.e. the variable stability was employed to select important variables. The definition of stability is the absolute value of regression coefficient divided by its standard deviation.

BOSS is a newly proposed method. The only difference between BOSS and SBOSS is that BOSS still considers the regression coefficient. The procedure of BOSS has been demonstrated in previous research [28].

2.6. Model Validation

Centering (pre-processing method) was applied in all the datasets, to evaluate the performance of four promising variable selection methods, namely CARS, SCARS, BOSS and SBOSS. Mean-centered were applied before modeling, and the optimal number of latent variables was determined by 5-fold cross validation. The root-mean-square error of calibration (RMSEC), root-mean-square error of the prediction of test set (RMSEP), Q^2_{cv} and Q^2_{test} were used to assess model performance. Moreover, the number of optimal latent variables (nLVs) and the number of variables selected (nVAR) were also recorded.

$$RMSEC = \sqrt{\sum_{i=1}^{Ncal}\left(y_i - \hat{y}_i\right)^2 / Ncal} \tag{12}$$

$$Q^2_{cv} = 1 - \sum_{i=1}^{Ncal}\left(y_i - \hat{y}_i\right)^2 / \sum_{i=1}^{Ncal}(y_i - \bar{y}_i)^2 \tag{13}$$

While y_i is the experimental of the predicted properties, and y_i^ and $\bar{y}_i$ represent predicted and average respectively. *Ncal* is the number of calibration samples of the training set. RMSEP and Q^2_{test} hold the equation following the same as RMSEC and Q^2_{cv}.

$$SD = \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 / (n-1)} \tag{14}$$

Each method was repeated 50 times to assess the stability and reproducibility. The standard derivation (SD) was employed to calculate stability with Eq. (14). Where X_i and $\bar{X}$ are predicted and average value, separately. n is the number of all samples. The smaller the value of the stability, the more stable is the method.

3. Experimental Section

3.1. Simulated Dataset [2]

SIMUIN data consists of 50 samples in rows and 200 wavelengths in columns. The first 100 wavelengths are linearly related with y, and the last 100 variables are random numbers from 0 to 1. In the second part, noises are in the range of 0 to 0.005. K-Stone sampling was applied to form calibration set (30 samples) and independent test set (20 samples). The calibration set is used for variable selection and modeling, and the validation set is used for prediction.

3.2. Corn Dataset

NIR datasets of corn were got from the website: http://www.eigenvector.com/data/Corn/index.html. The datasets contain 80 samples of corn. Each spectrum is in the range of 1100–2498 nm within 700 wavelengths at intervals of 2 nm. The properties of protein and starch are analytical targets. 32 samples were used to make up the calibration set, 32 samples were used as the validation set and 16 samples were employed as the test set according to the K-Stone sampling.

3.3. Iodine Value (IV) of Edible Oil Dataset

Soybean oil, olive oil, peanut oil and blend oil products were obtained from local supermarket. Iodine value were carried out by a standard titration method which is according to the official methods described in Method for animal and vegetable fats and oils-determination of iodine value (ISO3961:1996, MOD). 59 samples were prepared by mixing the four kinds of oil with the concentration of soybean oil, olive oil, peanut oil, blend oil from 0% to 85.46%,0% to 69.34%,0% to 88.35%,0% to 85.46%, respectively. NIR spectra were collected with

5 mm quartz cuvette by AutoCalib™ Demo (HAMAMATSU, Japan). The spectra were acquired over the range 1100–2100 nm (middle gain resolution, 2000 ms scans) at room temperature. Between each spectrum, the quartz cuvette was rinsed by the next sample. The dataset was split into Calibration set (36 samples) and independent test set (23 samples) by K-Stone sampling.

3.4. Beer Dataset

NIR datasets of corn were got from the website: http://www.mli.kvl.dk/foodtech/special/specials.htm. This spectra dataset was obtained with 30 mm quartz cell and collected at intervals of 2 nm within the wavenumbers range 400–2250 nm (926 points). The property of interest links to the original extract concentration. The dataset was divided into calibration set (36 samples) and independent test set (24 samples) by K-Stone sampling.

3.5. Reagents

Potassium iodide (AR, Sinopharm Chemical Reagent Co., Led. China), Sodium thiosulfate pentahydrate (AR, Sinopharm Chemical Reagent Co., Led. China), Cyclohexane (CP, Sinopharm Chemical Reagent Co., Led. China), glacial acetic acid (AR, Sinopharm Chemical Reagent Co., Led. China).

3.6. Software

All codes and datasets computations were written and applied in Matlab (V2014a, Mathworks, USA) on my personal computer (SSD) with an Intel Core i5-4210U 2.4 GHz CPU and 8 GB RAM for analysis.

4. Results and Discussion

4.1. Simulated Data

This dataset is designed to investigate general applicability of SBOSS. 5-fold cross validation is applied to explore its predict ability [37]. Meanwhile, MCUVE, GA, CARS, SCARS and BOSS are used for comparison, aiming at indicating that SBOSS is truly a promising method to eliminate uninformative variable, but not to make a decision which method is better.

Performance of different variable selection models are illustrated in Table 1. The results between two full-variables PLS models obviously show that uninformative variables have a great impact on the model efficiency. Compared to the full spectrum model, GA, CARS, BOSS, SCARS and SBOSS gave great improvements in RMSEP. The RMSEP decreased from 0.4947 to 0.0149, 0.0112, 0.0145, 0.0118 and 0.0111. It is worth noting that, the uninformative variables were well managed by CARS, BOSS, SCARS and SBOSS, but MCUVE led to poor competence because of the introduction of three uninformative variables. SBOSS showed the lowest RMSEP (0.0111) with the same nLVs with others. The standard deviation (SD) of six methods ranges from 0.0015 to 0.0022, which has little difference.

From Table 1, it can be observed that compared with other methods, SBOSS selected less variables and gave good RMSEP, together with SBOSS, also SCARS did not select uninformative variables. CARS and BOSS both perform well in RMSEP, but the two methods still select one or two uninformative variables. Above this, taking stability of regression coefficient into account is truly essential. We can ascertain that SBOSS is indeed an alternative method for variable selection. In conclusion, SIMUIN data has indicated that SBOSS is feasible to eliminate uninformative variables and improve predict ability. Three practical NIR datasets will be discussed in the next.

4.2. Corn Protein Data

The original NIR spectra of corn samples are presented in Fig. 1 (Supplementary material). In the corn dataset, on the base of 5-fold cross-validation on full spectra, the maximum number of LVs (latent variables) was set in 10. The results are demonstrated in Table 2, Fig. 2. In data analysis of spectrum, the most important part is not the most related wavelengths but the combinations of several bands which are chemical meaningful.

Common variables are existed among the four variable selection methods, including the regions around 1680, 1800 and 2180 nm. It can be noticed that selected variables cover a wide range linking to the complicated structure of protein, e.g. C—H, O—H and N—H bond with different vibration pattern, complex microenvironment of the three bonds, and the interaction of them.

As shown in Table 2, MCUVE gave a slighter better result than the full-spectrum PLS model, while other five methods showed obvious improvement in RMSEP. In total, 79 variables were retained by MCUVE, whereas fewer variables were selected by GA, CARS, SCARS, BOSS and SBOSS. GA, CARS and SCARS obtained comparable results, BOSS and SBOSS gave better results. The lowest RMSEP was acquired by SBOSS with fewer variables as SBOSS can avoid selecting uninformative variables, which have been investigated in simulated data. What's more, it should be noted that SBOSS also provide the lowest SD, even smaller than SCARS, which also hires stability of RC as index. The results on corn protein data demonstrated the efficiency of combination of MPA and stability of RC. Both the accuracy and robustness can be assured. The phenomenon expresses that with fewer variables, better prediction results can be acquired. As a result, it is essential to carry out variable selection before establishing calibration models. What's more, considering that collinear variables can reduce the stability of calibration models, choosing only the key variables is a practicable way for modeling.

The variables selected by CARS, BOSS SCARS and SBOSS of protein are displayed in Fig. 3. BOSS and SBOSS obviously have similar selected variables, compared with SCARS and CARS, SBOSS eliminates variables around 2210 nm and 1200 nm, and it retain the region around 1450 nm compared with BOSS, at the same time, MCUVE remove the region around 1450 nm, GA didn't retain the variables around 1450 nm

Table 1
The results on the SIMUIN dataset of different methods.

Characteristics	PLS[a]	PLS[b]	MCUVE-PLS		GA-PLS		CARS		BOSS		SCARS		SBOSS	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
nVAR	200(100)[c]	100(0)	78(3)	±4	60(2)	±18	35(1)	±7	53(1)	±8	46(0)	±8	31(0)	±9
nLVs	4	4	4	±0	3	±0	4	±0	4	±0	4	±0	5	±1
RMSEC	0.4673	0.0091	0.0396	±0.0257	0.0135	±0.0059	0.0106	±0.0015	0.0095	±0.0006	0.01	±0.0022	0.0075	±0.0007
RMSEP	0.4947	0.0112	0.0392	±0.0291	0.0149	±0.0055	0.0112	±0.0015	0.0145	±0.0015	0.0118	±0.0016	0.0111	±0.0013

nVAR: The number of selected variables. nLVs: The number of selected latent variables of PLS.
RMSEC: Root mean square error of calibration. RMSEP: Root mean square error of prediction.
SD: Standard deviation in 50 runs.
[a] Results using full spectrum with 200 variables by PLS.
[b] Results using only the 100 simulated informative variables by PLS.
[c] Number in the bracket denotes the number of uninformative variables used in the model.

Table 2
The results on the corn dataset of different methods.

Element	Characteristics	PLS	MCUVE-PLS		GA-PLS		CARS		BOSS		SCARS		SBOSS	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Protein	nVAR	700	79	±5	31	±9	49	±14	16	±6	16	±7	25	±7
	nLVs	10	8	±3	7	±1	10	±0	8	±2	9	±1	11	±2
	RMSEC	0.1387	0.0764	±0.0327	0.0516	±0.0140	0.0718	±0.0144	0.0366	±0.0172	0.0617	±0.0122	0.0183	±0.0049
	RMSEP	0.1275	0.1091	±0.0417	0.0683	±0.0170	0.0688	±0.0137	0.0411	±0.0172	0.0636	±0.0101	0.0275	±0.0048
Starch	nVAR	700	80	±14	33	±9	39	±7	13	±4	16	±8	13	±3
	nLVs	9	10	±1	9	±2	10	±0	9	±1	9	±1	9	±1
	RMSEC	0.3053	0.1638	±0.0107	0.0929	±0.0239	0.1451	±0.0273	0.0943	±0.0082	0.1181	±0.0163	0.0779	±0.0049
	RMSEP	0.2479	0.1810	±0.0120	0.1134	±0.0362	0.134	±0.026	0.0944	±0.0168	0.1243	±0.022	0.0859	±0.0121

and 1660 nm, which result in that SBOSS shows the best performance of RMSEP (0.0275).

Moreover, the variables in the absorption peak of the spectrum (2100 nm–2200 nm, 1440 nm,1600 nm–1800 nm) were possibly employed in the modeling. The variables in the range from 2200 nm to 2400 nm (with little information) that were selected by CARS were completed removed by MCUVE, GA, BOSS and CARS, while SCARS gave better results than CARS, because it retained fewer variables in this region than CARS.

4.3. Corn Starch Data

The results of six different methods in corn dataset are given in Table 2. It is obvious that all the variable selection methods gave better prediction results compared to the PLS full spectrum. SBOSS gave the lowest RMSEP (0.0859), followed by BOSS, GA, SCARS, CARS and MCUVE. it has an improvement above 48% to the PLS full spectrum model. Moreover, SBOSS yields the lowest SD (0.0121), which performs higher stability. As can be seen from Fig. 4, MCUVE and CARS didn't remove the variables between 1200 nm to 1400 nm which doesn't correspond to starch, therefore the worst results were obtained. BOSS and SBOSS have selected fewer variables than other four methods with lower RMSEP, which demonstrated that better prediction results can be achieved with fewer variables. All the methods have selected the region around 1748 nm and 1766 nm which correspond to the second overtone of C—H.

4.4. IV Dataset

Fig. 1 (Supplementary material) shows the NIR raw spectra of mixed edible samples over the spectral range of 1000–2200 nm. There are five absorption regions, which are similar with the literatures describing the location of NIR regions specific to edible oils. The two peaks centered around 1168 and 1210 nm correspond to the second overtone of CH stretching vibration. The combination of the C—H stretching and vibration with other vibration modes of the concerned molecule associated

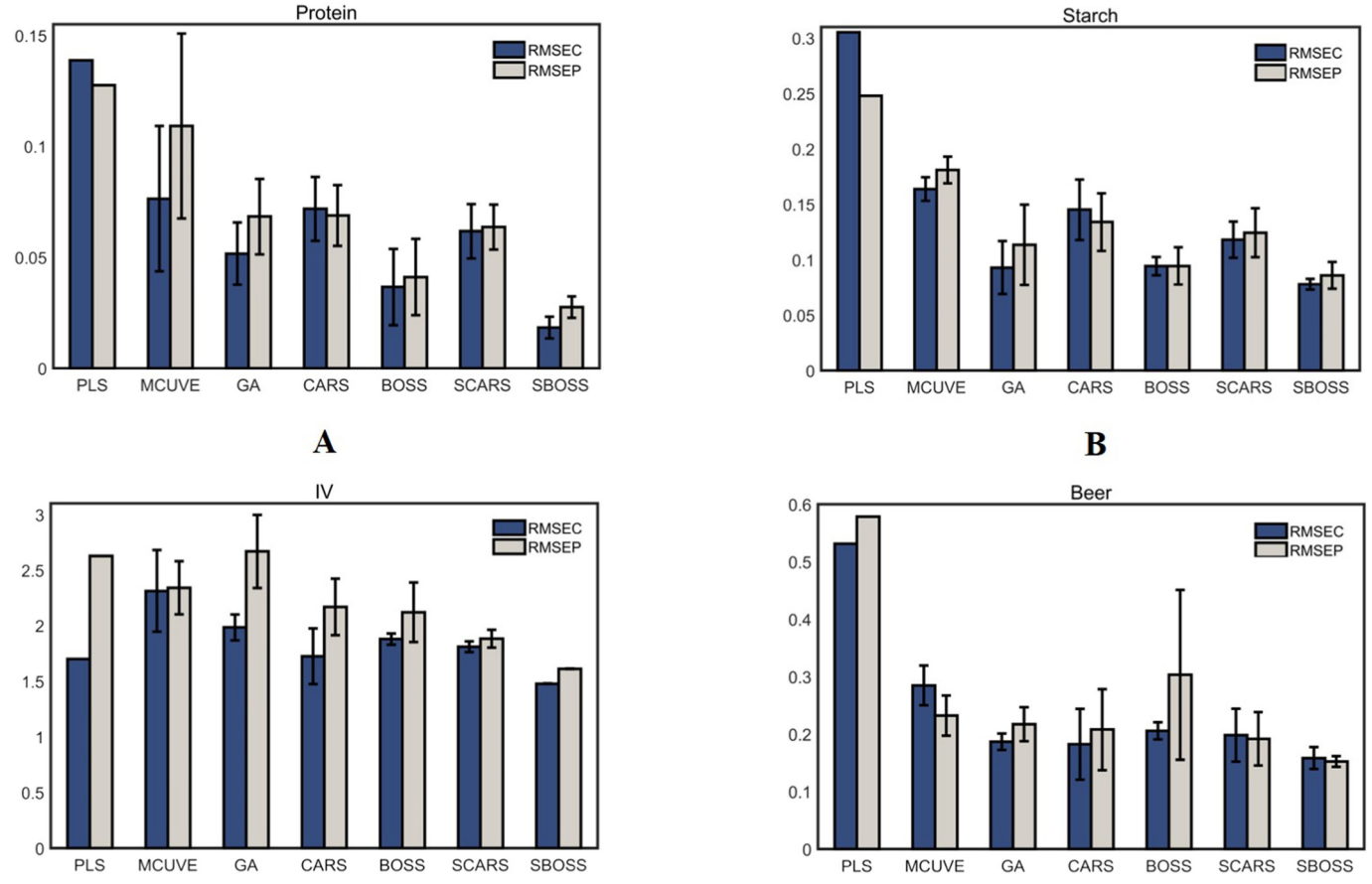

Fig. 2. RMSEC and RMSEP of PLS, MCUVE, GA, CARS, BOSS, SCARS and SBOSS on datasets (A) corn protein (B) corn starch (C) IV (D) beer.

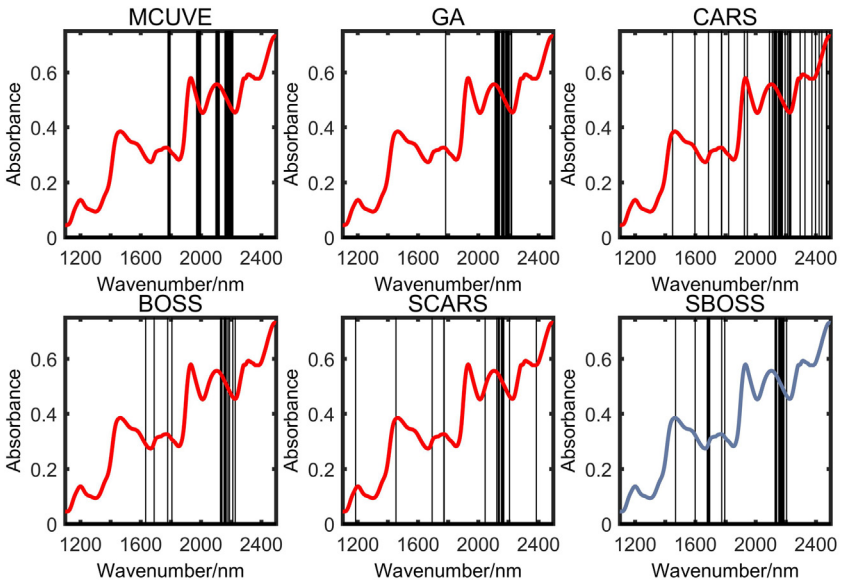

Fig. 3. The variables selected by MCUVE, GA, CARS, BOSS, SCARS and SBOSS on protein of corn datasets.

with the regions around 1392 and 1414 nm. And two peaks centered 1726 nm and 1761 nm linked to the first overtone of the CH stretching vibration.

Table 3 and Fig. 2 show the results of IV of edible oil. Compared to the full spectrum PLS model, the six variable selection methods didn't give great improvements. Still and all, SBOSS showed the largest improvements (20.2%) of RMSEP with the lowest SD (0.0022). The other five methods all have much larger SD than SBOSS. Combined with Fig. 5, it is obvious that GA obtained the least selected variables, but the worst prediction performance it got. It didn't select the regions around 1392 nm, 1414 nm which correspond to the combination of the C—H stretching and vibration with other vibration modes of the concerned molecule and the region 1761 nm linked to the first overtone of the C—H stretching vibration. Elimination of informative variables lead to the bad outcomes of GA. SCARS and BOSS have similar selected variables except the region around 1400 nm. Both of them selected all the informative variables which linked to oil, thus, they gave nearly the same RMSEP. In IV dataset, SBOSS still gave a good performance not only in prediction ability but also stability.

4.5. Beer Dataset

In beer spectrum, there is a large variation in the visual part of the spectra going from 400 nm to approximately 700 nm. This is due to variation in the visual appearance of the beers, which vary from very light beers to very dark beers. The area has a high variance, but it has little or no relationship to the chemical property to be predicted. In the high spectral range, dominated by absorbance of water, high absorbances lead to noisy measurements, which may also have a certain influence on the regression models but are not related to the parameter to be predicted. The remaining part of the spectrum is dominated by C—H and N—H stretching overtones except for the O—H second overtone of water at approximately 970 nm. This data set is interesting because it contains the two features that mostly lead to suboptimal models

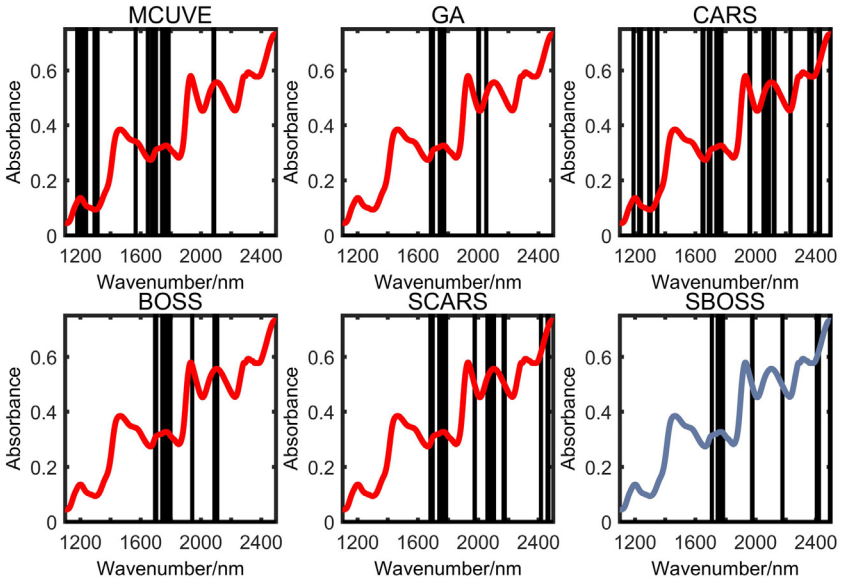

Fig. 4. The variables selected by MCUVE, GA, CARS, BOSS, SCARS and SBOSS on starch of corn datasets.

Table 3
The results on the IV dataset of different methods.

Characteristics	PLS	MCUVE-PLS		GA-PLS		CARS		BOSS		SCARS		SBOSS	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
nVAR	3108	809	±430	9	±7	144	±90	30	±12	135	±74	187	±9
nLVs	3	2	±1	2	±0	3	±1	4	±1	2	±0	3	±1
RMSEC	1.6999	2.3119	±0.3671	1.9835	±0.1172	1.724	±0.2489	1.8784	±0.0508	1.8101	±0.048	1.4785	±0.0017
RMSEP	2.6271	2.3403	±0.2382	2.6676	±0.3290	2.1683	±0.2548	2.1203	±0.2681	1.8819	±0.0803	1.6123	±0.0022

when using non-relevant variables. The left (low wavelength) part of the spectrum contains highly systematic but non-relevant variation. The right (high wavelength) part is mainly unsystematic noise and also irrelevant for predictions.

The noisy part is typically not too difficult to handle but it leads to spurious correlations. In this case, such spuriously selected variables are easily detected visually because of the spectral nature of the data; if one variable is selected and the neighboring variables are not, then it is an indication that the result is not to be trusted. Such visual aids are not feasible for all types of data, but we can use them here to see how different approaches handle noisy data. The systematic but irrelevant left part of the spectrum is more challenging and will typically lead to problems. The information in that area has no physical relation to the quality but because of the high variance and the limited number of samples, variables from that region are typically included by variable selection methods due to the high variance and accidental moderate correlation.

The purpose of acquiring the data is to predict the real extract concentration, which is a measure of the ability of the yeast to ferment alcohol. It is used as quality parameter in the beer production. Predicting the real extract from spectroscopic measurements can provide a fast quality measurement in the beer production.

The results of different methods are listed in Table 4 and Fig. 2. It can be seen that compared to the PLS full spectrum model, the prediction has showed great improvement with variable selection. CARS and BOSS performs worst. The reason why CARS and BOSS do not play well may be that the index based on RC is not an optimal choice. SBOSS gave the best RMSEP and SD. One detail we should notice is that the performance of BOSS on the dataset is abnormal. The stability of BOSS is the worst, and the abnormal results in 50 replicate runs are listed in Table 2 (Supplementary material). Over-fitting is serious.

To give a better understanding and explanation of the selected variables, the variables selected by MCUVE, GA, CARS, BOSS, SCARS and SBOSS are shown in Fig. 6. The variables selected by SCARS and SBOSS are more concentrated on the region between 1100 and 1500 nm, which is linked to the absorption of 1st overtone of O—H stretching bond vibration.

GA, SCARS and MCUVE have comparative performance, while variables retained by MCUVE were fifteen times of those by SCARS, GA and SBOSS. However, CARS and BOSS abnormal retain the region 1600-2200 nm which is mainly noise information. That's why the results CARS and BOSS are bad. Beer, as a mixture, consists thousands of chemicals, so it has more disturbance than simple system. Therefore, retaining uninformative variables usually bring in bad results when irrelevant information interferes. Compared with BOSS and CARS, SBOSS prevent the over-fitting issue efficiently. It not only gave the outstanding predict ability but also show the best stability. On top of that, the index of stability of RC may be a good choice. It can ensure the accuracy and stability of prediction.

Meanwhile, the RMSECV in the sub-models decrease during the iterations and reach the minima at iteration 12 (Fig. 2 in Supplementary material). This method takes into consideration of variable subsets at different levels of nVAR, which is reasonable since the optimal nVAR is unknown before and during variable selection. The weights of variables change during the iterations as it is shown in Fig. 3 in Supplementary material. The variables which have large weight at the beginning may not turn out to be unimportant for modeling with small weight in the later iterations. The optimal variable set is obtained in iteration 12. The most informative variables are thus obtained at around 1100 nm, 1200 nm and 1500 nm, which are reserved by all six variable selection methods.

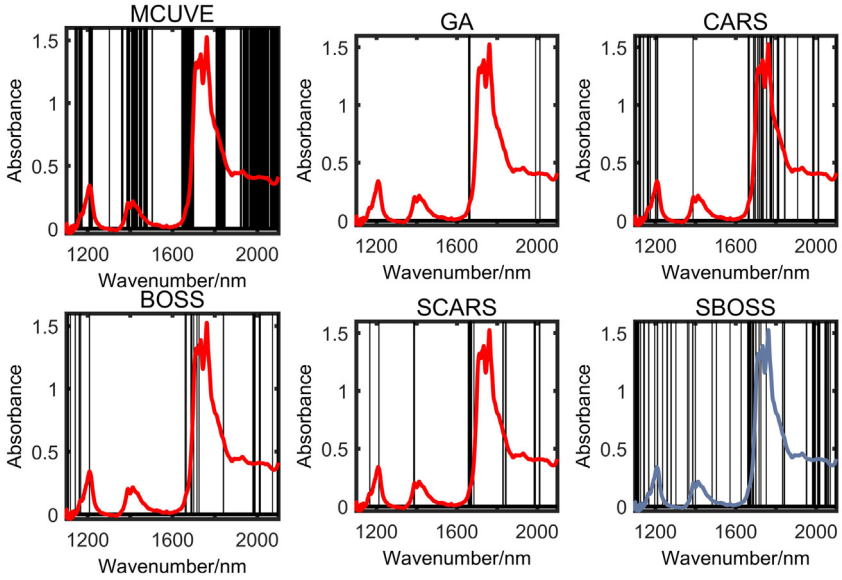

Fig. 5. The variables selected by MCUVE, GA, CARS, SCARS, BOSS and SBOSS on IV dataset.

Table 4
The results on the beer dataset of different methods.

Characteristics	PLS	MCUVE-PLS		GA-PLS		CARS		BOSS		SCARS		SBOSS	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
nVAR	926	132	±12	5	±2	104	±28	18	±18	20	±28	4	±1
nLVs	3	3	±0	3	±1	3	±0	3	±0	3	±1	3	±0
RMSEC	0.531	0.2847	±0.0346	0.1866	±0.0143	0.1822	±0.0615	0.2055	±0.0149	0.1979	±0.0459	0.1582	±0.0191
RMSEP	0.5786	0.2321	±0.0351	0.2169	±0.0295	0.2078	±0.0705	0.3032	±0.1477	0.1917	±0.0465	0.1524	±0.0093

4.6. Stability of RC and MPA

Stability of a variable considers the stability of regression coefficient, not only the RC. It can guarantee that the most relevant variables have larger probabilities to participate in the model by combining with MPA. In SBOSS, sub-models are computed due to the weights of variables, and the weights are achieved according to the stability of RC. It makes the results more reliable.

As a general framework for statistically extracting interesting information from a large population of sub-models towards better understanding of the chemical data, MPA is promising in developing new chemometrics algorithms, ranging from variable selection and model evaluation to outlier detection and applicability domain definition.

4.7. Soft Shrinkage Strategy

In soft shrinkage strategy, the variables which may be considered as less important are assigned smaller weights, thus, these variables still have the chance to participate the models. At the same time, they may prove to be informative. The risk of ignoring essential variables can be reduced by using soft shrinkage during the whole process. WBS also played a role in the good performance of BOSS, different weights assign to different objects, so the objects with larger weights have more chances to be selected. Applying WBS can avoid the influence of collinearity of RC, also, the combination of BSS and WBS make the variable ranking more reliable.

4.8. Over-Fitting and Cross Validation Issue

BOSS and SBOSS is based on 5-fold CV, and the basic problem in learning is to test all of the training data on some set of previously unseen CV data, and then to pick up the smallest RMSECV. Therefore, the stability of RMSECV should be assured.

Investigations have shown that variable selection has the ability to lower the risk of over-fitting by reducing the dimensionality of the variable space. However, methods based on RC still can't avoid over-fitting. In spectra, noise and uninformative variables are main factors that may have influence on the stability of computation. Related research has indicated that RC is susceptible to noise. Firstly, PLS can't extract effective variables on account of the existence of high level noise, which will bring in poor RC, such as the beer dataset. Meantime, the RC of uninformative variables are usually labile. Therefore, the RC can't represent the real model in effect the existence of two types of variables. Secondly, in practice, CV doesn't give warning of over-fitted models, and investigations have shown that model stability is a good indicator of over-fitting as well as under-fitting [38]. When a PLS model is built, the ultimate result is the regression vector which represents the model. The stability of the regression vector thus reflects the stability of the model. On top of that, in our study, combination of the stability of the RC and the prediction of CV is adopted as a criterion for model selection which is based on MPA can prevent over-fitting efficiently. In SBOSS loop, through multi-models, the model prediction ability and stability is acquired. It can select the optimal variable combination and avoid over-fitting in PLS models.

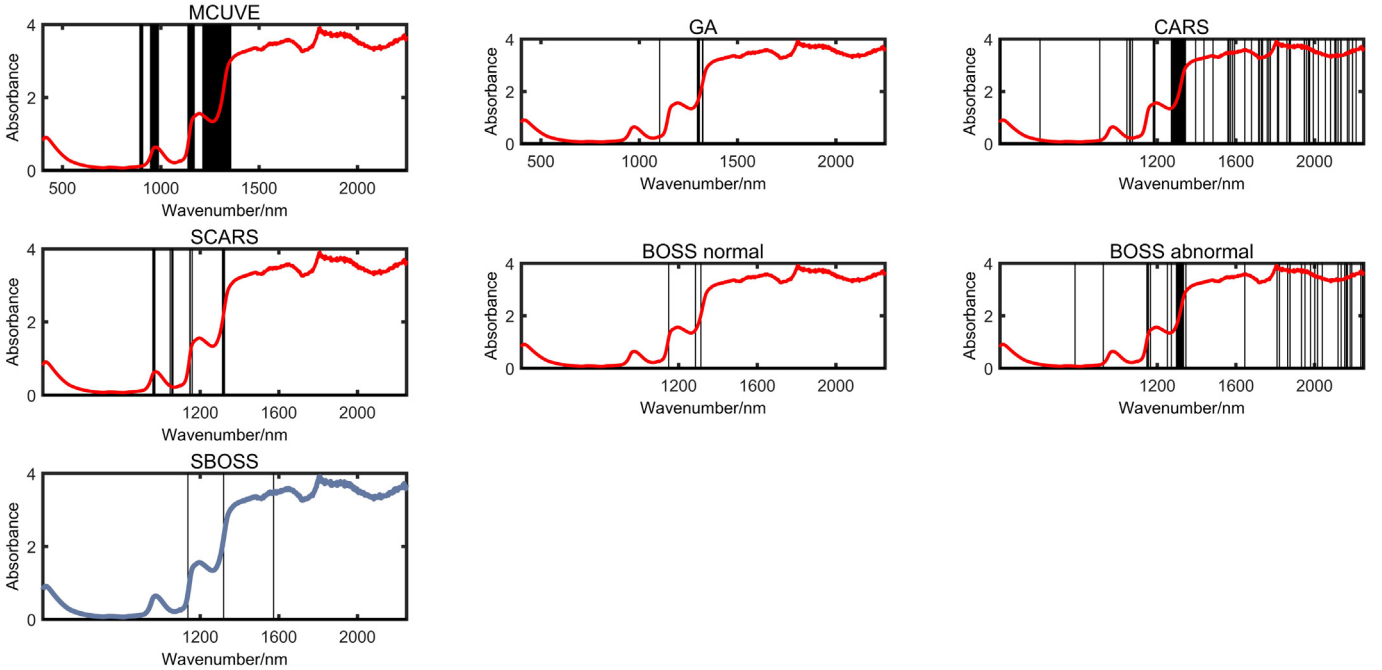

Fig. 6. The variables selected by MCUVE, CARS, BOSS and SBOSS on beer dataset.

5. Conclusion

In the framework of ensemble, we introduced an effective and promising variable selection approach that we term SOBSS as a technique to deal with overlapped spectral variables and uncorrelated variance of NIR spectra system for identification and quantitative analysis. The application based on three datasets demonstrated the improvements of prediction ability by adopting SBOSS to the shape of NIR vibration. The model selected by SBOSS showed both When compared with five outstanding variable selection methods, including GA-PLS, MC-UVE, CARS, SCARS and BOSS, SBOSS was demonstrated as a better strategy with the good accuracy, model stability and variable selection credibility, and the combination of MPA, model stability and CV avoided over-fitting efficiently. The outstanding performance of SBOSS indicates that it is a good alternative of variable selection in multivariate calibration.

In addition, SBOSS suggested a broader utility for variable selection with chemical interests, and the stability of RC can be considered the optimal search index in further investigations. Although variable selection was performed by SBOSS coupled with PLS in this study, it is a general strategy that can also be coupled with other regression and classification methods and applied into other fields, such as genomics, bioinformatics, metabolomics and quantitative structure–activity relationship (QSAR).

Acknowledgments

This work has been help with Xiangzhong Song, language revised by Kuangda Tian.

Disclosure of potential conflicts of interest

All authors declare that they have no conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.saa.2018.10.034.

References

[1] H. Li, Y. Liang, Q. Xu, D. Cao, Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration, Anal. Chim. Acta 648 (2009) 77–84.
[2] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, C. Sterna, Elimination of uninformative variables for multivariate calibration, Anal. Chem. 68 (1996) 3851–3858.
[3] D. Jouan-Rimbaud, D.L. Massart, R. Leardi, O.E. De Noord, Genetic algorithms as a tool for wavelength selection in multivariate calibration, Anal. Chem. 67 (1995) 4295–4301.
[4] U. Hörchner, J.H. Kalivas, Simulated-annealing-based optimization algorithms: fundamentals and wavelength selection applications, J. Chemom. 9 (1995) 283–308.
[5] C.H. Spiegelman, M.J. Mcshane, M.J. Goetz, M. Motamedi, Q.L. Yue, G.L. Coté, Theoretical justification of wavelength selection in PLS calibration: development of a new algorithm, Anal. Chem. 70 (1998) 35–44.
[6] C.H. Spiegelman, M.J. McShane, M.J. Goetz, M. Motamedi, Q.L. Yue, G.L. Coté, Theoretical justification of wavelength selection in PLS calibration: development of a new algorithm, Anal. Chem. 70 (1998) 35–44.
[7] H.C. Goicoechea, A.C. Olivieri, A new family of genetic algorithms for wavelength interval selection in multivariate analytical spectroscopy, J. Chemom. 17 (2003) 338–345.
[8] Y. Du, Y. Liang, J. Jiang, R. Berry, Y. Ozaki, Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares, Anal. Chim. Acta 501 (2004) 183–191.
[9] R. Leardi, L. Nørgaard, Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions, J. Chemom. 18 (2004) 486–497.
[10] X. Zou, J. Zhao, Y. Li, Selection of the efficient wavelength regions in FT-NIR spectroscopy for determination of SSC of 'Fuji' apple based on BiPLS and FiPLS models, Vib. Spectrosc. 44 (2007) 220–227.
[11] S.W. Lin, K.C. Ying, S.C. Chen, Z.J. Lee, Particle swarm optimization for parameter determination and feature selection of support vector machines, Expert Syst. Appl. 35 (2008) 1817–1824.
[12] S. Ye, D. Wang, S. Min, Successive projections algorithm combined with uninformative variable elimination for spectral variable selection, Chemom. Intell. Lab. Syst. 91 (2008) 194–199.
[13] H. Xu, Z. Liu, W. Cai, X. Shao, A wavelength selection method based on randomization test for near-infrared spectral analysis, Chemom. Intell. Lab. Syst. 97 (2009) 189–193.
[14] M. Shamsipur, V. Zare. Shahabadi, B. Hemmateenejad, M. Akhond, Combination of ant colony optimization with various local search strategies. A novel method for variable selection in multivariate calibration and QSPR study, QSAR Comb. Sci. 28 (2009) 1263–1275.
[15] Z. Xiaobo, Z. Jiewen, M.J. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy, Anal. Chim. Acta 667 (2010) 14–32.
[16] F. Allegrini, A.C. Olivieri, A new and efficient variable selection algorithm based on ant colony optimization. Applications to near infrared spectroscopy/partial least-squares analysis, Anal. Chim. Acta 699 (2011) 18–25.
[17] S.F.C. Soares, R.K.H. Galvão, M.C.U. Araújo, E.C. Da Silva, C.F. Pereira, S.I.E. De Andrade, F.C. Leite, A modification of the successive projections algorithm for spectral variable selection in the presence of unknown interferents, Anal. Chim. Acta 689 (2011) 22–28.
[18] H.D. Li, Y.Z. Liang, D.S. Cao, Q.S. Xu, Model-population analysis and its applications in chemical and biological modeling, TrAC Trends Anal. Chem. 38 (2012) 154–162.
[19] H.D. Li, Q.S. Xu, Y.Z. Liang, Random frog: an efficient reversible jump Markov chain Monte Carlo-like approach for variable selection with applications to gene selection and disease classification, Anal. Chim. Acta 740 (2012) 20–26.
[20] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, A review of variable selection methods in partial least squares regression, Chemom. Intell. Lab. Syst. 118 (2012) 62–69.
[21] K. Zheng, Q. Li, J. Wang, J. Geng, P. Cao, T. Sui, X. Wang, Y. Du, Stability competitive adaptive reweighted sampling (SCARS) and its applications to multivariate calibration of NIR spectra, Chemom. Intell. Lab. Syst. 112 (2012) 48–54.
[22] B.C. Deng, Y.H. Yun, Y.Z. Liang, L.Z. Yi, A novel variable selection approach that iteratively optimizes variable space using weighted binary matrix sampling, Analyst 139 (2014) 4836–4845.
[23] J.P. Andries, Y.V. Heyden, L.M. Buydens, Predictive-property-ranked variable reduction with final complexity adapted models in partial least squares modeling for multiple responses, Anal. Chem. 85 (2013) 5444–5453.
[24] M. Farrés, S. Platikanov, S. Tsakovski, R. Tauler, Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation, J. Chemom. 29 (2015) 528–536.
[25] H.D. Li, Y.Z. Liang, Q.S. Xu, D.S. Cao, Model population analysis for variable selection, J. Chemom. 24 (2010) 418–423.
[26] Y.H. Yun, W.T. Wang, M.L. Tan, Y.Z. Liang, H.D. Li, D.S. Cao, H.M. Lu, Q.S. Xu, A strategy that iteratively retains informative variables for selecting optimal variable subset in multivariate calibration, Anal. Chim. Acta 807 (2014) 36–43.
[27] Y.H. Yun, W.T. Wang, B.C. Deng, G.B. Lai, X.B. Liu, D.B. Ren, Y.Z. Liang, W. Fan, Q.S. Xu, Using variable combination population analysis for variable selection in multivariate calibration, Anal. Chim. Acta 862 (2015) 14–23.
[28] B.C. Deng, Y.H. Yun, D.S. Cao, Y.L. Yin, W.T. Wang, H.M. Lu, Q.Y. Luo, Y.Z. Liang, A bootstrapping soft shrinkage approach for variable selection in chemical modeling, Anal. Chim. Acta 908 (2016) 63–74.
[29] M.C.U. Araújo, T.C.B. Saldanha, R.K.H. Galvão, T. Yoneyama, H.C. Chame, V. Visani, The successive projections algorithm for variable selection in spectroscopic multicomponent analysis, Chemom. Intell. Lab. Syst. 57 (2001) 65–73.
[30] C. Tao, E. Martin, Bayesian linear regression and variable selection for spectroscopic calibration, Anal. Chim. Acta 631 (2009) 13–21.
[31] J.H. Jiang, R.J. Berry, H.W. Siesler, Y. Ozaki, Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data, Anal. Chem. 74 (2002) 3555–3565.
[32] L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S. Engelsen, Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy, Appl. Spectrosc. 54 (2000) 413–419.
[33] Y.H. Yun, H.D. Li, L.R. Wood, W. Fan, J.J. Wang, D.S. Cao, Q.S. Xu, Y.Z. Liang, An efficient method of wavelength interval selection based on random frog for multivariate spectral calibration, Spectrochim. Acta A Mol. Biomol. Spectrosc. 111 (2013) 31–36.
[34] Q.J. Han, H.L. Wu, C.B. Cai, L. Xu, R.Q. Yu, An ensemble of Monte Carlo uninformative variable elimination for wavelength selection, Anal. Chim. Acta 612 (2008) 121–125.
[35] M. Forina, S. Lanteri, M.C. Cerrato Oliveros, C. Pizzaro Millan, Selection of useful predictors in multivariate calibration, Anal. Bioanal. Chem. 380 (2004) 12.
[36] D. Vallenet, E. Belda, A. Calteau, S. Cruveiller, S. Engelen, A. Lajus, F. Le Fevre, C. Longin, D. Mornico, D. Roche, Z. Rouy, G. Salvignol, C. Scarpelli, A.A.T. Smith, M. Weiman, C. Medigue, MicroScope-an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data, Nucleic Acids Res. 41 (2013) E636–E647.
[37] G.H. Fu, Q.S. Xu, H.D. Li, D.S. Cao, Y.Z. Liang, Elastic net grouping variable selection combined with partial least squares regression (EN-PLSR) for the analysis of strongly multi-collinear spectroscopic data, Appl. Spectrosc. 65 (2011) 402–408.
[38] B.C. Deng, Y.H. Yun, Y.Z. Liang, D.S. Cao, Q.S. Xu, L.Z. Yi, X. Huang, A new strategy to prevent over-fitting in partial least squares models based on model population analysis, Anal. Chim. Acta 880 (2015) 32–41.