Received October 29, 2019, accepted November 13, 2019, date of publication November 19, 2019,
date of current version December 2, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2954115

An Variable Selection Method of the Significance Multivariate Correlation Competitive Population Analysis for Near-Infrared Spectroscopy in Chemical Modeling

YUXI WANG[1], ZHENHONG JIA[1], AND JIE YANG[2]

[1]College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China
[2]Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China

Corresponding author: Zhenhong Jia (jzhh9009@sohu.com)

This work was supported in part by the National Science Foundation of China under Grant 61665012 and Grant U1803261, and in part by the International Science and Technology Cooperation Project of the Ministry of Education of the People's Republic of China under Grant DICE 2016–2196.

ABSTRACT The high dimensionality of spectral datasets makes it difficult to select the optimal subset of variables. This paper presents a new method for variable selection called the significant multivariate competitive population analysis (SMCPA), Which combines ideas of significant multivariate correlation (SMC) and model population analysis, and employs weighted bootstrap sampling (WBS) and exponential decline function (EDF) competition methods. In this study, the values of SMC distributions are used as an index for evaluating the importance of each wavelength. Then, based on the importance level of each wavelength. SMCPA sequentially selects N subsets of spectral wavelengths by N Monte Carlo sampling in an iterative and competitive procedure. In each sampling run, a fixed ratio of samples is used to build a calibrated partial least-squares model, and then SMC is performed to obtain the score and threshold values. Next, based on the significant multivariate correlation scores, the key variables are selected by two steps: the compulsory selection of exponential decline function and the competitive selection of adaptive weighted sampling. Finally, cross-validation(CV) is applied to select the optimal subset with the lowest root mean square error. This method is tested on three NIR spectral datasets and compared against three high-performance variable selection methods. The experimental results show that the proposed algorithm has the highest efficiency and the best selection effect, and can usually locate the optimal combination of key wavelength variables in a dataset. The evaluation result after PLS modeling is also the best.

INDEX TERMS Spectrochemical analysis, variable selection, the significant multivariate correlation, weighted bootstrap sampling, model population analysis, monte Carlo sampling, analytical techniques, partial least squares method.

I. INTRODUCTION

With the characteristics of simple, rapid, noninvasive, cost-effective and no sample pretreatment, near infrared spectroscopy has been widely adopted as a popular analytical tool for both qualitative and quantitative analysis in petrochemical, pharmaceutical, environmental, clinical, agricultural, food and biomedical fields [1]. Modern near infrared spectroscopy analysis involves the rapid acquisition of a large number of absorbance values for a selected spectral range, and then uses the information contained in the spectral curve to predict the chemical composition of the sample by extracting the appropriate variables of interest [2]. However, the typically present relatively weak, nonspecific, highly overlapping and discontinuity of the near-infrared spectral region it make it difficult to directly extract the information related to the content of components in near-infrared spectra of substances and provide reasonable spectral analysis [3], Moreover, the large number of spectral variables in most datasets encountered in spectral chemometrics often

The associate editor coordinating the review of this manuscript and approving it for publication was Stavros Souravlas.

complicates the prediction of a dependent variable, and the analysis of those high-dimensional spectral data faces many challenges,one of which is the so called non-deterministic polynomial-time(NP)hard problem, i.e., relatively large number of variables compared to the number of samples. With the large number of spectral variables, NIR spectra usually include some noise and interfering variables that render the predictive property of interest unreliable, and modeling of such datasets will has a high risk of over-fitting [4]. Therefore, it is very important to study the sensitive, fast and accurate method of extracting relevant variables to predict the chemical composition of samples in chemometrics, and to develop a dimension reduction method to address those problems [5]. Projection methods such as partial least-squares (PLS) [6] and principal component regression(PCR) [7] have been widely used to address these high-dimensional collinear data. By replacing the original variables with a few latent variables or principal components of larger variance, the impact of collinearity, band overlaps and redundant noise irrelevant to the property of interest can be reduced [8]. However, researchers have found that variable selection can further improve the prediction ability of PLS or PCR models, Variable selection improves the stability of the model and improves model interpretability as well, even if multi-collinearity exists in multivariate calibration [9]. Therefore, variable selection techniques have play a key role in the analysis of high-dimensional spectral datasets [10], [11]. Liang and colleagues confirmed the importance and necessity of variable selection to obtain a subset of spectral information in complex analysis systems [12], [13]. The purpose of variable selection can be summarized as follows [14]: (1) improve the model predictive ability; (2) provide faster and more cost-effective predictors by reducing the curse of dimensionality; (3) give a better understanding and interpretation of the underlying process that generated the data.

In view of the benefits of variable selection, variable selection methods based on different strategies have been proposed in large of numbers. These include the classical methods [15], [16]; the penalized methods [17], [18]; the intelligent learning algorithms [19], [20]; variable sorting strategies based on PLS model parameters [21]–[23] and variable selection methods based on the model population analysis (MPA)strategy [23]–[38] and so on. MPA as an open ensemble learning framework, in which different blocks can be filled and statistical tools to extract important information from the models. The concept of MPA was first proposed by Li et al. in the field of variable selection [39]. The key elements of MPA are random sampling and statistical analysis, The core idea of MPA is to statistically analyze(i.e. statistical test) the performance of a large population of sub-models generated from random sampling and to extract interesting information from outputs of the sub-models. An important feature of MPA is that it considers the output of interest not as a single value but a distribution [40], MPA is a very effective strategy for developing variable selection methods. There are many algorithms based on this framework, such as the competitive adaptive reweighted sampling method (CARS) [31], variable permutation population analysis(VPPA) [27], iteratively optimizing variable space using weighted binary matrix sampling for variable selection method(VISSA) [32], variable combination population analysis(VCPA) [24], bootstrapping soft shrinkage method (BOSS) [28], iteratively variable subset optimization (IVSO) [35], least absolute shrinkage and selection operator (SEPA-LASSO) [37], Fisher optimal subspace shrinkage (FOSS) [29] and model adaptive space shrinkage(MASS) [26] and so on.

In this study, we focus on developing more efficient variable selection methods. First, SMCPA attaches great importance to the stability and reliability of proposed method. Some newly suggested methods made the compared with several existing methods employing simple quantitative "visual" comparison and slightly smaller RMSEP, and concluded that the proposed method is better than others. However, a statistical analysis, which indicates the unreliability of the proposed methods, was not performed. It is necessary to validate the reliability of the algorithm through statistical analysis or statistical testing [41], [42]. SMCPA combines with the new criteria of variable importance (SMC), statistical test(F-test) and some mathematical ideas(such as MC-sampling, EDF and WBS, etc.), and Statistical methods to demonstrate the reliability of the proposed method (SMCPA). Stability could be used as a metric to evaluate the prediction performance of models, similar to Deng *et al.* [40]. SMCPA inherits the advantages of the MPA framework, statistically analyzing the interesting output of a large number of sub-models and extracting useful information. The ensemble strategy (MPA) is also a good method for addressing the instability problem as it can obtain more accurate, stable and robust prediction results by combining all the predictions of multiple sub-models built from different subsets. Second, SMCPA focused on multi-objective optimization, and the global optimization of the algorithm makes the variable selection method more meaningful [42], [44]. In the SMCPA algorithm, the RMSECV, RMSEP, number of variables(nVRA), number of latent variables(nLV), coefficient of determination of cross-validation(Q2_cv), coefficient of determination of test set(Q2_test), and computation time (T) are regarded as the objective functions. With more than one objective function optimized simultaneously, and thus ensuring that the PLS modeling the selected optimal subset of variables by SMCPA will have better prediction accuracy and simpler with greater interpretability. Third, The SMCPA can eliminate the negative impact of noise in the dataset,locate the information variables quickly and reduce the possibility of noise and irrelevant variables being selected as key variables. This study also considers the problem of an absolute regression coefficient as a key variable evaluation metric. Many of the variable selection methods are developed with absRC as an important variable's evaluation criterion [43], such as CARS, BOSS, IVOS and FOSS etc. However, absRC does not always reflect the real information of a variable's importance [34], which will have a negative impact on the variable selection method. Therefore, we chooses SMC as the criterion of variable importance. The key points in SMC are estimating for each variable the correct sources of variability resulting from the PLS regression (i.e., regression variance and residual variance), and using them for statistically determining a variable's importance with respect to the regression model, and SMC discards the orthogonal variance decomposition to prevent the influence of nonrelevant information contained in datasets [42]. Therefore, the SMCPA can also address the spectral datasets with noise. Fourth, the proposed variable selection methods are more to increase the computational cost to obtain smaller prediction error (RMSEP) values, such as VCPA, BOSS and VISSA algorithms, which all achieve better prediction results by increasing computational cost [43]. This study balances the relationship between computational cost and predictive ability, and considers the computation time as an objective function. Fifth, SMCPA preserves the synergy and combination effects among the variables and gradually eliminates irrelevant variables through the shrinkage strategy. Additionally, SMCPA always selects variables that are better correlated with analyte contents or properties of interest. Therefore, it is necessary to interpret the selected variables associated with the functional group (CH, OH, SH, and NH) of the analyte or property of interest. Therefore, the SMCPA algorithm has strong interpretability and avoids the blind selection of variables. Finally, although a large number of variable selection methods have been developed in NIR spectra over the last several decades, many problems still need to be addressed and solved. Prediction accuracy, model interpretability, and computational complexity are three important pillars of any variable selection procedures. It is a great challenge to develop a method that can strike a good balance between the three pillars [42]. Our research has a good balance of these three parts; SMCPA not only has the best prediction accuracy and interpretability but also the minimum computational cost. We evaluate an algorithm from additional perspectives. The performance of SMCPA was tested on three groups of NIR spectral datasets and three high performing variable selection methods.

II. THEORY AND METHOD

A. NOTATION

The data matrix X is assumed to contain p variables in rows and n samples in columns, and the vector y of size n × 1 denotes the measured property of interest. A superscript T denotes the vector or matrix transpose. When establishing the PLS model, both X and y are mean-centered.

In this study, we assume that the number of exponential decline function (EDF) iterations and the number of Monte Carlo sampling (MCS) runs is set to N. The ratio of samples in each random MCS run is R. Using the above settings, SMCPA can be divided into four steps in each iteration: (1) A subset of variables is randomly established using Monte Carlo sampling with a fixed selection ratio. (2) The distribution of SMC values of the output of the sub-models is statistically analyzed(F-test) and its values ranked, then EDF is used to force the elimination of uninformative or redundancy variables. (3) Normalized SMC scores as the weights of each wavelength, and the WBS method is used to further eliminate the weaker weight variables. Variables with larger weights have greater a probability of being retained, while variables with weaker weights are less competitive, and populations with variables are gradually eliminated. (4) The N variable subsets were cross validated to evaluate the subset. The minimum RMSECV subset is selected as the optimal subset.

B. MONTE CARLO SAMPLING OF DATA SUBSETS

MCS is an important statistical tool for analyzing complex (multivariate) problems [45]. It is a stochastic method based on the use of random numbers and probability statistics. The samples and variables are both randomly selected with a fixed number, respectively. MCS is implemented both in sample space and variable space of the calibration set to obtain sub-datasets, In each sampling run, the selected sub-dataset from the calibration set is considered as the training set, while the remaining part is regarded as the test set. A large number of PLS models are established on the sub-datasets generated by many MCS runs. We obtain prediction errors through model population analysis instead of depending on a single model. Statistical analysis can be used to analyze each sub-model output to evaluate the unknown parameters of interest in each sub-dataset. The randomly sampled sub-datasets are represented as $(X_{\mathrm{sub}}, y_{\mathrm{sub}})_i$, where $i = 1, 2, \ldots, N$.

C. IMPORTANCE VARIABLE DETERMINATION METHODS AND SIGNIFICANT MULTIVARIATE CORRELATION

The interpretation of variable importance carried out using parameters calculated from the PLS model is referred to as the popular filter methods [42], [43], [46], Variable Importance in the Projection (VIP) [5], Selectivity Ratio (SR) [47] and the PLS regression coefficients (RC). The Significant Multivariate Correlation (SMC) [48] methodology was introduced and showed favorable results when compared with the above important variable determination methods.

D. VARIABLE IMPORTANCE IN THE PROJECTION

In the VIP approach [49], the importance of the variables is established using the projection information of the independent variable X and the response y. The VIP score is used to evaluate the importance of each variable in the PLS model. For a PLS model with h latent variables, the VIP score of the jth variable is calculated as follows:

$$P = \sqrt{\frac{p\sum_{k=1}^{h}\left[SS\left(c_k t_k\right)\left({}^{w_{jk}}/_{\|w_k\|}\right)^2\right]}{\sum_{k=1}^{h} SS\left(c_t t_t\right)}} \tag{1}$$

The average VIP is equal to 1, because the SS of all VIP values is equal to the number of variables in X. This means that if all X-variables have the same contribution to the model, they will have a VIP value equal to 1. VIP values larger than 1 point to the most important variables, and generally VIP values below 0.5 are considered irrelevant variables.

E. SELECTIVITY RATIO

The **s** electivity ratio (SR) [47] takes the regression coefficient vector of PLS and calculates a target projection (TP) by projecting the rows of X onto the normalized regression coefficient vector and the load by projecting the columns of X onto the score vector:

$$\mathrm{t}_{TP} = Xb \big/ \|b\| = \hat{y} \Big/ \|b\| \tag{2}$$

$$\mathrm{p}_{TP} = X^T t_{TP} \Big/ \left(t_{TP}^T t_{TP} \right) \tag{3}$$

The explained and residual variance can be calculated from the variable matrix X, TP scores, and the loads:

$$X = t_{TP} p_{TP}^T + E_{TP} \tag{4}$$

$$\mathrm{V}_{i,\mathrm{exp}} = \left\| t_{TP_i} p_{TP_i}^T \right\|^2, \quad \mathrm{i} = 1, 2, \ldots, \mathrm{p} \tag{5}$$

$$\mathrm{V}_{i,res} = \left\| e_{TP_i} \right\|^2, \quad \mathrm{i} = 1, 2, \ldots, \mathrm{p} \tag{6}$$

Equations (5) and (6) determine SR as the ratio of the explained variance $\mathrm{V}_{\mathrm{i,exp}}$ and the residual variance $\mathrm{V}_{\mathrm{i,res}}$ for each variable:

$$\mathrm{SR}_\mathrm{i} = \mathrm{V}_{\mathrm{i,exp}} \big/ \mathrm{V}_{i,res}, \quad i = 1, 2, \ldots, p \tag{7}$$

$$\mathrm{SR}_\mathrm{i} > F_{crit} = F(\alpha, N-2, N-3) \tag{8}$$

SR was applied to re-quantify the X-variance to improve interpretation of variable importance by means of the target rotation or the orthogonal filtering strategy. The purpose is to allocate information proportional to the covariance between X and y and simultaneously isolate orthogonal irrelevant variation. A critical threshold value for determining variable importance was suggested in Ref [47]. where in the SR_i is assessed against the F-distribution with n − 2 and n − 3 degrees of freedom. In this work,an F-test (95%) was selected to select candidate targets.

F. REGRESSION COEFFICIENTS BETA

The interpretation of variable importance is through the vector of regression coefficient (RC), which is a single measure of association between each variable and the response. The variables with a small absolute value of this filter measure can be eliminated. Also in this case thresholding may be based on significance consideration from jackknifing or bootstrapping. Those resampling techniques, such as jack-knife and bootstrap, are often used to determine confidence intervals [22]. Various resampling techniques are available for PLS regression coefficients, but none offer a straightforward ranking of variable importance in the model. Usually, the absolute value of the regression coefficient is used as a guide, but this does not always reflect the true importance of the variable, and is affected by factors such as noise.

G. THE SIGNIFICANT MULTIVARIATE CORRELATION

The SMC is an important part of the variable selection method. The key points in SMC are to estimate for each variable the correct sources of variability resulting from the PLS regression (i.e., regression variance and residual variance), and use them to statistically determine a variable's importance with respect to the regression model. For variance evaluation, SMC uses the PLS regression coefficient vector, *b*, to define the covariance between the X-variable and the response variable *y* in the combination of the vector of predicted values, $\hat{y}$, as a new latent score vector of the PLS model in eq. (9) and the regression coefficient vector, see eq. (9). However, dissimilar to the target projection method, SMC discards the orthogonal variance decomposition in eq. (3) in order to prevent the influence of non-relevant information contained in X. Therefore, without this rotation, the normalized regression coefficient vector will be used as a load vector , p_{sMC}, and the reconstructed of X can be represented in eq. (10) [50].

$$\mathrm{t}_{sMC} = Xp_{sMC} = X\frac{b}{\|b\|} = \frac{\hat{y}}{\|b\|} \tag{9}$$

When predicted in the PLS model $\hat{y} = Xb$;

$$\mathrm{X} = t_{sMC} p_{sMC}^T + E_{sMC} = \frac{\left(\hat{y}\, b^T\right)}{\|b\|^2} + E_{sMC} \tag{10}$$

Lacking the actual regression step(orthogonal variance decomposition), SMC is not a complete basic rotation for the explained variance $||\mathrm{t}_{\mathrm{smc}}\mathrm{p}_{\mathrm{smc}}^\mathrm{T}||^2$(or regression variance) in eq. (11) may not be orthogonal to the estimated residual variance $||\mathrm{e}_{\mathrm{smc}}||^2$ in eq. (12). However, it reflects the relevant variation in the predicted response projected back onto the original X-variable space via the PLS regression vector.

$$\mathrm{V}_{\mathrm{i},reg} = \left\| t_{SMCp\,SMC_i^T} \right\|^2 = \left\| \frac{\hat{y}\, b_i^T}{\|b_i\|^2} \right\|^2 \tag{11}$$

$$\mathrm{V}_{\mathrm{i},res} = \left\| e_{SMC_i} \right\|^2 = \left\| x_i - \frac{\hat{y}\, b_i^T}{\|b_i\|^2} \right\|^2 \tag{12}$$

The variables whose F-values exceed the critical threshold of the F-test (determined by the selection of the significance level) are considered to be important variables. Those two variances(the explained variance and residual variance) are obtained in the form of individual regression of each X-variable to the common score vector, with the loadings as the regression coefficients in eq.(10) for SMC. The analysis of variance test(ANOVA) is the most appropriate for the regression significance. In the ANOVA ,the F-test is carried out using the mean-squared error, which is the raw sums of squares divided by the appropriate degrees of freedom. For the SMCi test values, we have an F distribution of 1 numerator and n-2 denominator degrees of freedom is used. F(1-α, 1, n-2), where α is the chosen significance level. In the experiments α is 0.01. The F-test is used to assess variables that are statistically significant with respect to their relationship to y. F_{smci} exceeding the F-test critical threshold value (as determined via the choice of significance level) are considered important variables. Moreover, the variables are ranked based on their respective F-values with a defined significant threshold value. Eq. (13)–Eq. (17) [48].

$$\mathrm{S}_{\mathrm{i,reg}} = V_{i,reg}\big/1 \tag{13}$$

$$\mathrm{S}_{\mathrm{i,res}} = V_{i,res}\big/(n-2) \tag{14}$$

$$\mathrm{F}_{\mathrm{smc}_i} = \frac{S_{i,reg}}{S_{i,res}} \tag{15}$$

$$F_{test\_threshold} = F\,(1-\alpha, 1, n-2) \tag{16}$$

$$F_{smc_i} > F_{test\_threshold} \tag{17}$$

Note that the impact on the threshold value for the SMC test is dependent on the actual rotation effect. Due to the lack of the regression step, the variance decomposition is biased in that the $\mathrm{S}_{\mathrm{i,res}} + \mathrm{S}_{\mathrm{i,reg}}$ may not always equal the sum of squares $||\mathrm{x_i}||^2$, and only $\mathrm{p_{smc}} \approx \mathrm{p_{TP}}$ the basic rotation effects can be ignored. When the basic rotation effect is negligible, the residual variance is orthogonal to the explained variance, and the false positive rate associated with a non-parametric distribution should be close to the theoretical null value of 0.05. Otherwise, the rotation effect cannot be neglected in the presence of a large irrelevant variation effect, which is an uninformative variable in the NIR dataset [50]. As the first property of the basic sequence, the TP loading is the rotation of the regression coefficients vector toward the dominant eigenvector of $\mathrm{X^TX}$, which may be independent of the response. For this reason, depending on the actual magnitude of the rotation, $\mathrm{P_{TP}}$ may be less proportional to the covariance of the X-variables and the response variable y. However, the rotation impact has been removed in SMC. The potential false positive rate of SMC is lower than the theoretical value.

In the Ref [48], it has been proven that SMC provides a better subset of variables than VIP, SR, and abs RC, and has the fewest false negative and false positive errors. Moreover, the prediction performance of the PLS model is improved by the reduced deviation of the variable selection method. This reduces the possibility of mistaking irrelevant and noisy information exclusive to important variables and reduces the likelihood of treating important variables as irrelevant or uninformative. Therefore, when using complex NIR data variables, SMC can reduce the false positive rate and improve the result of important variable selection. The basic concept of SMC allows for the maximum use of the information obtained in the basic sequence to identify important variables in the PLS model. The SMC method best highlights the minimum deviation and statistically significant variables in the model.

Next, we compare the importance variables defined by SMC, VIP, SR, and the abs RC using a beer spectral dataset. Given that MCS is used in this study, a fixed ratio of spectral samples (i.e., randomly selected ratios of 80–90%)creates a large number of subsets,produces calibrated PLS models, and obtains each of variable SMC score, VIP score, SR score and the abs RC. Thus, we conducted experiments with between 80% and 90% of the randomly selected samples, and repeated each experiment 60 times. Finally, 80% and 90% of the samples were selected for statistical averaging and then visualized to show the feasibility of using SMC as an important variable selection criterion.

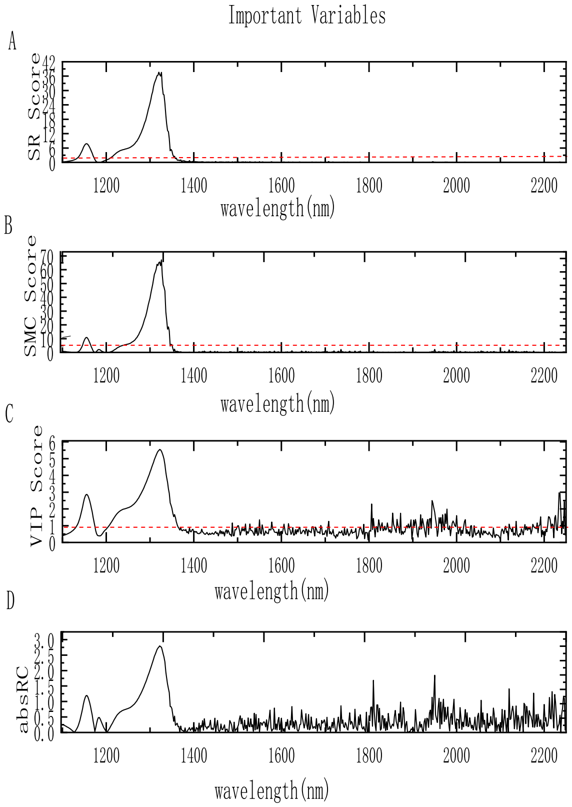


FIGURE 1. Importance of variables is illustrated by SR, SMC, VIP, and the abs RC of 80% samples randomly selected from a spectral dataset. The red dotted lines in the SR, SMC and VIP diagrams represent the threshold for selecting the importance of variables.

In the above figures, there is not much change between the graphs when 80% and 90% of samples are randomly selected. The important variables defined by the VIP scores and the abs RC also include a large number of irrelevant variables. A large number of interference variables in the curve will have a great impact on the variable selection algorithm, which regards these variables as the importance variables and will greatly increase the possibility of irrelevant and interference variables being selected into key variables, whereas the distribution of SMC or SR values of the output of the sub-models is statistically analyzed, and those curves have fewer interference variables. In fact, the important variable regions defined by SR and SMC coincide with the yeast substrate chemical properties in the beer dataset, and the 1150–1350 nm region in the beer dataset corresponds to the first overtone of the O-H stretch bond vibration and the second overtone of the

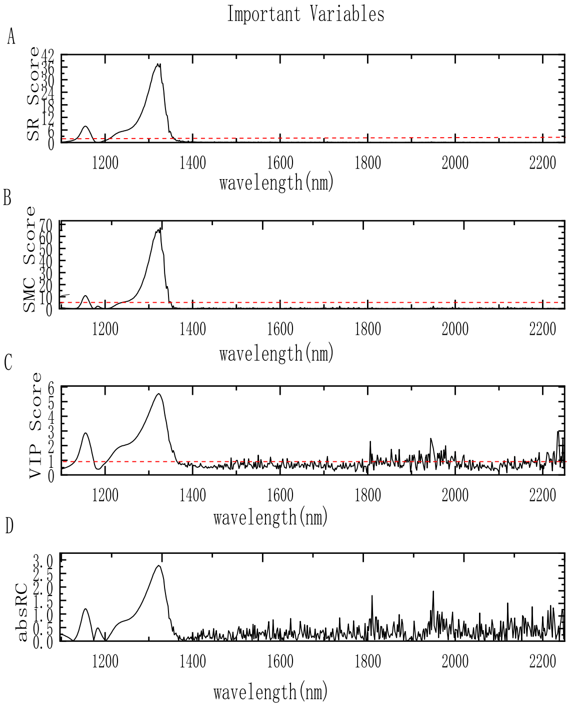


FIGURE 2. **Importance of variables is illustrated by SR, SMC, VIP, and the abs RC of 90% samples randomly selected from a spectral dataset. The red dotted lines in the SR, SMC and VIP diagrams represent the threshold for selecting the importance of variables.**

C-H stretch bond. This conforms with the chemical properties of interest to be studied in the beer dataset. Improved interpretability with simple models is one of the three purposes of variable selection, so it is necessary to make an interpret of selected variables associated with the functional group (CH, OH, SH, NH, etc.) of the analyte or property of interest. The SMC method can also ensure that variables with minimum deviation and the statistical significance are highlighted in the model. Through the above theoretical analysis and this example proved that SMC is more suitable than VIP, SR and Absolute RC as criteria for evaluating important variables in this study. As mentioned above, we choose the distribution value of SMC as the basis for evaluating the importance of each variable.

SMC = $[\mathrm{smc}_1, \mathrm{smc}_2, \ldots, \mathrm{scm}_\mathrm{p}]^\mathrm{T}$ is a P-dimensional SMC score Vector, in which score values are greater than threshold. The i-th element smc_i in the SMC score reflects the i-th wavelength contribution to y. Here, we evaluate the importance of each wavelength, i.e., rank the SMC scores, where higher-ranked variables are more important. Normalized weights are also defined for bootstrap sampling to compete for important variables:

$$\mathrm{w}_i = \frac{smc_i}{\sum_{i=1}^{p} smc_i}, \quad i = 1, 2, \ldots, p \tag{18}$$

Note that the weight of the eliminated wavelength is forcibly changed to zero and the weight vector is always p-dimensional.

H. EXPONENTIALLY DECREASING FUNCTION

In the proposed method, EDF is used to mimic the principle of "natural selection" or survival of the fittest." [28], [32] EDF can be divided into two stages, First, many unimportant variables are quickly eliminated. In this initial stage, the corresponding elimination ratio is relatively large, and the elimination of irrelevant information is relatively strong. This stage is called "Fast selection. In the second stage, following this reduction in the number of uninformative and unimportant variables, the elimination ratio of EDF becomes smaller and approaches zero, thus preventing errors in eliminating key variables. This second stage is called "Refined selection."

Let ri be the variable reserve rate of the first iteration, and a and k be determined by the following two conditions: (1) In the first iteration, all p wavelengths are used for modeling, which means that all variables are preserved, i.e., the variable retention ratio r1 = 1. (2) In the Nth iteration, only two wavelengths are retained, which means that the variable retention rate rN = 2/p. Mathematically, this process is formulated as follows:

$$r_i = ae^{-ki}, \quad i = 1, 2, \ldots, \mathrm{N} \tag{19}$$

$$a = \left(\frac{p}{2}\right)^{\frac{1}{N-1}} \tag{20}$$

$$\mathrm{k} = \frac{\ln\left({}^{p}\!/_{2}\right)}{N-1} \tag{21}$$

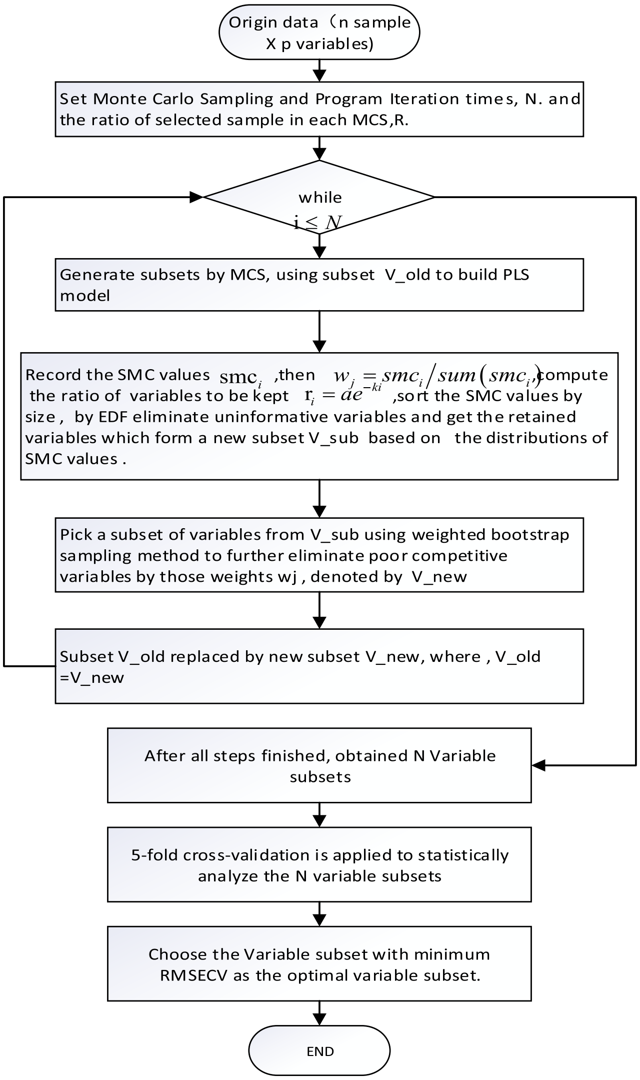
I. THE WEIGHTED BOOTSTRAP SAMPLING

The bootstrap method is a statistical technique for random sampling with replacement [51]. It was introduced by B. Efron in 1979 and was then successfully used in many fields such as model evaluation. Recently, it has also been used for wavelength selection [29]. Weighted bootstrap sampling (WBS) is an improved version of bootstrap sampling with different weights on sampling objects. It is should be noted that the variables selected may not be unique, and some variables may be selected repeatedly, while others will not be selected at all. The weighted bootstrap sampling (WBS) is an improved sampling technique for BSS that also allows replacement during sampling. In this study, after the wavelength variables have been reduced using EDF, the WBS algorithm is applied to further eliminate weaker weight variables, similar to the "survival of the fittest" principle. Variables with larger weights have a greater probability of being retained, whereas those weaker weight variables are less competitive, and gradually eliminated in the pool of variables.

J. GENERAL DESCRIPTION OF SMCPA

The SMCPA method chooses the optimal variable scheme. Through Monte Carlo N times sampling, N subsets of variables are selected iteratively. Finally, the subset with the minimum RMSECV is selected as the optimal subset. In each sampling run, SMCPA runs in four successive steps including Monte Carlo model sampling. The distribution of

SMC values of the output of the sub-models is statistically analyzed (F-test) and its values ranked, enforced wavelength reduction by EDF, further competition wavelength reduction by WBS and RMSECV calculation for each subset. The combination of EDF forced variables reduction and WBS competitive variables reduction is a two-step procedure for wavelength selection. The block diagram is as follows:


Origin data （n sample X p variables)
Set Monte Carlo Sampling and Program Iteration times, N. and the ratio of selected sample in each MCS,R.
while
$i \leq N$
Generate subsets by MCS, using subset V_old to build PLS model
Record the SMC values smc_i ,then $w_j = smc_i / sum(smc_i)$,compute the ratio of variables to be kept $\mathrm{r}_i = ae^{-ki}$,sort the SMC values by size , by EDF eliminate uninformative variables and get the retained variables which form a new subset V_sub based on the distributions of SMC values .
Pick a subset of variables from V_sub using weighted bootstrap sampling method to further eliminate poor competitive variables by those weights wj , denoted by V_new
Subset V_old replaced by new subset V_new, where , V_old =V_new
After all steps finished, obtained N Variable subsets
5-fold cross-validation is applied to statistically analyze the N variable subsets
Choose the Variable subset with minimum RMSECV as the optimal variable subset.
END

K. MODEL VALIDATION

The predictive ability of the models is assessed by 5-fold cross-validation and independent prediction set. Each dataset was divided into a calibration set and prediction set using the Kennard-Stone (KS) method, resulting the root mean squared error of cross-validation(RMSECV),the root mean squared error of prediction(RMSEP), the coefficient of determination of cross-validation(Q2_cv) and the coefficient of determination of the test set (Q2_tets), The calibration set was used for wavelength selection and building the model while the prediction set was used for validating the calibration model.

III. DATASETS AND SOFTWARE

We selected three public spectral datasets for our experiments, namely the corn dataset, wheat dataset, and beer dataset. The corn and wheat data sets are ideal spectral data sets for acquisition. whereas the beer dataset consists of 60 samples containing a rather large noisy part due to an absorbance that is too strong, in a region dominated by the water component. The proposed algorithm can obtain better variable screening results for both relatively ideal public spectral datasets and poor public spectral dataset with noise interference, and the evaluation result after PLS modeling was also the best.

A. WHEAT DATASET

This NIR dataset consists of 100 wheat samples in which the protein value is considered as the property of interest y. The spectrum was recorded from 1100–2500 nm with 701 spectral points at intervals of 2 nm. Because of the ‘large p, small n’ problem, the original spectrum is compressed into a maximum of 200 points by an appropriate window size, as performed by Riccardo Leardi [52]. Setting the window size to 4, this dataset decreased to 175 variables with an average of every four original variables. The dataset was divided into a calibration set of 80 samples and independent test set 20 samples on the basis of the Kennard–Stone (KS) method.

B. CORN DATASET

This dataset consists of 80 samples of corn measured by three different NIR spectrometers. Each spectrum contains 700 spectral points at intervals of 2 nm within the range 1100–2498 nm. In the present study, the NIR spectra of 80 corn samples measured by the m5 instrument were considered as X and the moisture value was considered as the property of interest y. In addition, the dataset was divided into a calibration set (80% of the dataset, 60 × 700) and an independent test set (20% of the dataset) on the basis of the KS method. The corn dataset is available from http://www.eigenvector.corn/data/corn/index.html.

C. BEER DATASET

This spectral dataset [53] was recorded with a 30-mm quartz cell directly on the undiluted degassed beer, and collected at intervals of 2 nm within the range 400–2250 nm. For this study, the NIR region 1100–2250 nm (576 data points) was chosen. The original extract concentration, which indicates the substrate potential for the yeast to ferment into alcohol, is considered as the property of interest. The dataset was divided into a calibration set (40 samples) and the independent test set (20 samples) on the basis of the Kennard–Stone (KS) method. The independent test set was constituted by selecting every third sample from the original dataset.

D. SOFTWARE AND SCRIPTS

All codes were written and performed in MATLAB(version 2016A,the MathWorks, Inc.) on a general-purpose Lenovo computer with an intel i5 3.2GHz CPU and 4GB RAM, with

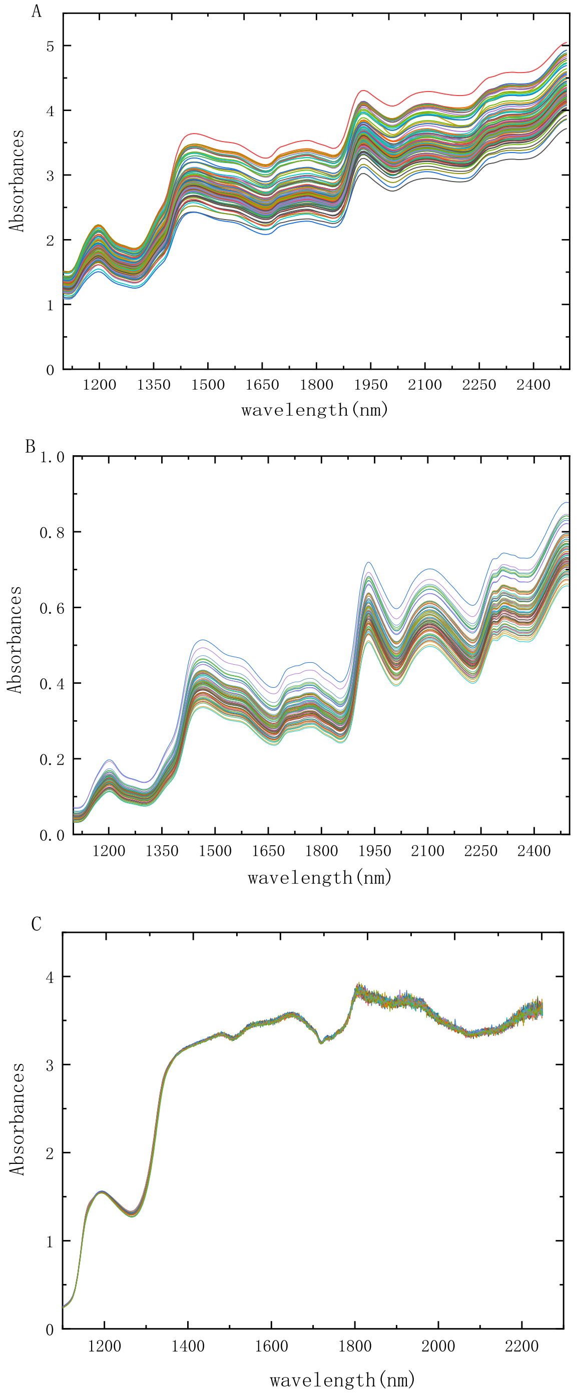


FIGURE 3. (A) Original NIR spectra of wheat dataset (B) original NIR spectra of corn dataset (C) original NIR spectra of beer dataset.

a Microsoft Windows 10 professional Version operating system. The MATLAB codes for BOSS are freely available from http://www.mathworks.com/matlabcentral/fileexchange/52770-boss. The MATLAB source code for VCPA is available for academic research from http://www.mathworks.com/matlabcentral/fileexchange/authors/498750. The MATLAB source code for CARS is available from http://code.google.com/p/carspls/.

IV. RESULTS AND DISCUSSION

In this study, all data sets were divided into calibration sets and independent test sets based on the Kennard-Stone (KS) method. The KS method aims to cover multidimensional space by maximizing the Euclidean distance between each pair of selected samples. The calibration set was used for variable selection and goodness of fit. and the independent test set was used to validate the calibration model for predictions, and the PLS model based on NIPALS-PLS in our paper. In addition, in order to evaluate the performance of the SMCPA, we compared with the excellent methods CARS, VCPA and BOSS methods. Through the reference [31] and experiments, the parameters of CARS were set as follow: the number of Monte Carlo sampling runs was set to 100, and in each MC sampling run, a fixed ratio (e.g. 80%) of samples was first randomly selected to establish a calibration model and the Exponentially decreasing function runs were 100. Through the reference [24] and experiments, the parameters of VCPA were set as follows: the binary matrix sampling runs were 1000,the Exponentially decreasing function runs were 100,the number of the left variables was 14 and the ratio of best models was 10%.Through the reference [28] and experiments for the BOSS parameters, the Bootstrap sampling (BSS) runs are 1000 and 1000 subsets were generated, the number of variables in the new subsets was about 0.632 times that of the previous subsets and the ratio of best models was 10%. For all methods, the PLS maximum latent variable is limited to 10 and the number of latent variables was determined by a 5-fold cross-validation. Each data set was mean-centered before modeling. Each method was executed 50 times to obtain statistical results and to ensure a fair comparison.

A. EFFECT OF NUMBER OF MC SAMPLING RUNS

To investigate the effects of the number of MCS runs and the fixed selection ratio of samples on SMCPA performance, we considered the following case: the monte carlo sampling number was set to 50, 100, 200, 300, and 500. For each case and each of the three datasets, 50 replicate runs of SMCPA were executed and RMSECV values were recorded. Experiments with different selection ratios,we found that the fixed selection ratio does not have a significant influence on the performance of SMCPA. Therefore, we do not discuss the effect of the fixed selection ratio on the SMCPA, and only consider the influence of the number of MCS runs, and the selection ratio was set to 0.8 in all experiments.

In the following sections, the wheat, corn, and beer datasets set the optimal numbers of MCS runs were 50, 100, and 100, respectively. The resulting statistical box-plots are shown in Figure 4.

B. WHEAT DATASET

The statistical results of variable selection methods,i.e., CARS, VCPA,BOSS and SMCPA, over 50 runs on the wheat dataset, are summarized in Table 1. It can be clearly seen

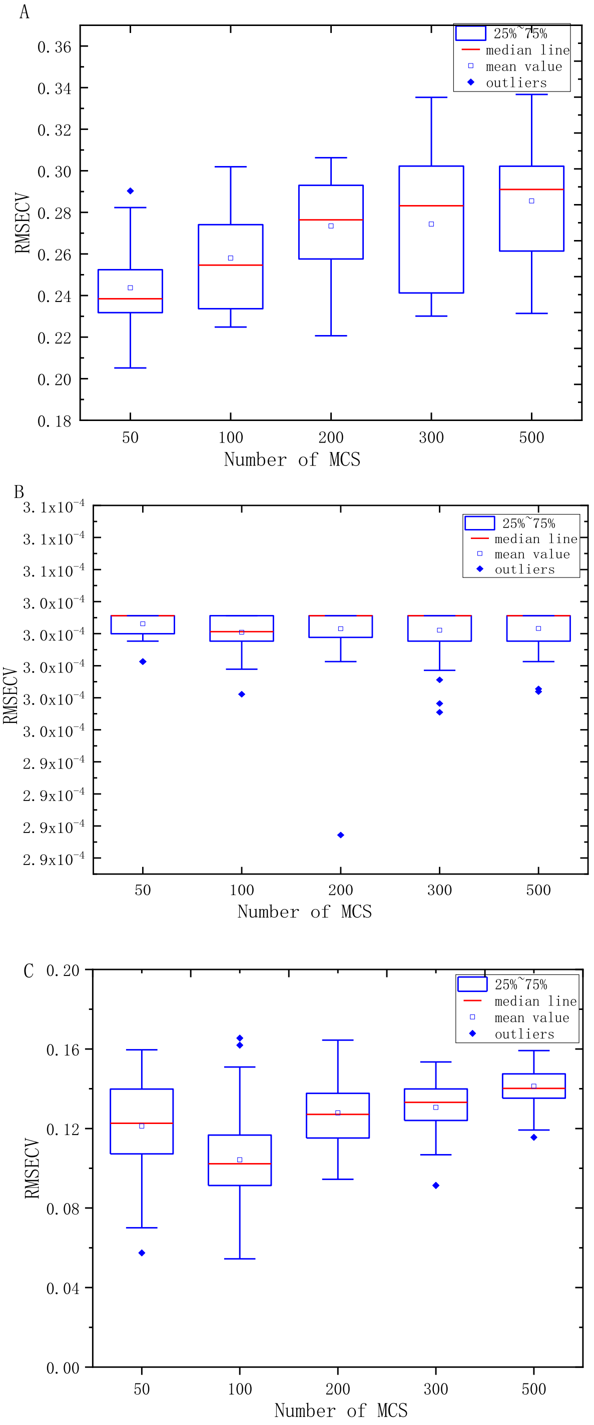


FIGURE 4. Dataset box plots for 50, 100, 200, 300, and 500 MCS iterations. (A) Wheat dataset. (B) Corn dataset. (C) Beer dataset.

that the prediction performances of the four variable selection methods are better than that of PLS full-spectrum,which means that variable selection is a necessary and important step for the NIR model. The four variable selection methods have improved prediction performance both on cross-validation and the test set predictions. The enhancement of prediction ability by the SMCPA algorithm is significant Compared to the PLS models of full spectral PLS models. The values of RMSECV and RMSEP for SMCPA decrease from 0.607 and 0.519 to 0.232 and 0.234, respectively, and the values of Q2_cv and Q2_test increase from 0.748 and 0.774 to 0.942 and 0.889, respectively. Compared to other selection methods, the SMCPA algorithm has better prediction performance in terms of RMSECV, RMSEP , Q^2_cv, and Q^2_test. Taking RMSEP as an example,RMSEP of SMCPA is 0.234,while the other selection methods, CARS,VCPA and BOSS are 0.418,0.289 and 0.305. The number of selected variables by SMCPA decreases from 175 to 6.7. SMCPA selected fewer variables with lower RMSECV and RMSEP, which means that it can achieve better prediction performance with fewer variables for wheat dataset. In terms of algorithm computation cost, the average computation times of SMCPA and CARS is about 0.8 s, which proves the efficiency of two methods. While VCPA and BOSS have much computation times of 142.09s and 36.34s, respectively, although they have a better performance models. The both of VCPA and BOSS increase computational cost to obtain better model prediction performance. It is obvious that SMCPA is the best in prediction performance of model and computational efficiency.

The variables selected by different selection methods over 50 runs on wheat datasets are displayed in FIGURE 5. The information variables in the 1150–1350 nm region were selected for all four algorithms. This region belongs to the second overtone of the C-H stretching bands and the first overtone of the O-H stretching bands. The spectral features of samples in the near-infrared(1000–2500nm) spectral region are associated with the vibrational modes of functional groups. The organic species present in the sample have a distinct spectral fingerprint in the NIR region, namely a relatively strong overtone absorption and mode combination relative to several functional groups(C-H,N-H,O-H,i.e.).

From Figure.5, we found that VCPA, BOSS, and SMCPA selected variables in similar regions that were roughly distributed in the range 1150–1350 nm. However, the CARS selected variables were distributed over the entire 1100–2500 nm spectral range, which means that selected variables by CARS include uninformative or interfering variables cause to the poor prediction performance. The SMCPA, BOSS and VCPA selected wavelength regions were consistent with the chemical properties of wheat dataset, and selected those relevant variables that had better and accurate performance of models.

C. CORN MOISTURE DATASET

The statistical results of variable selection methods, i.e., CARS, VCPA,BOSS and SMCPA,over 50 runs on the corn moisture datasets are summarized in Table 2. Compared to the PLS models of full spectral PLS models, the RMSECV and RMSEP for SMCPA decrease by 98.6% and 98.1%, respectively. The values of RMSECV and RMSEP decreased from 0.02111 and 0.01522 to 0.00030 and 0.0.00031, respectively, and the values of Q2_cv and Q2_test increased from 0.96710 and 0.93780 to 0.99999 and 0.99999, respectively.

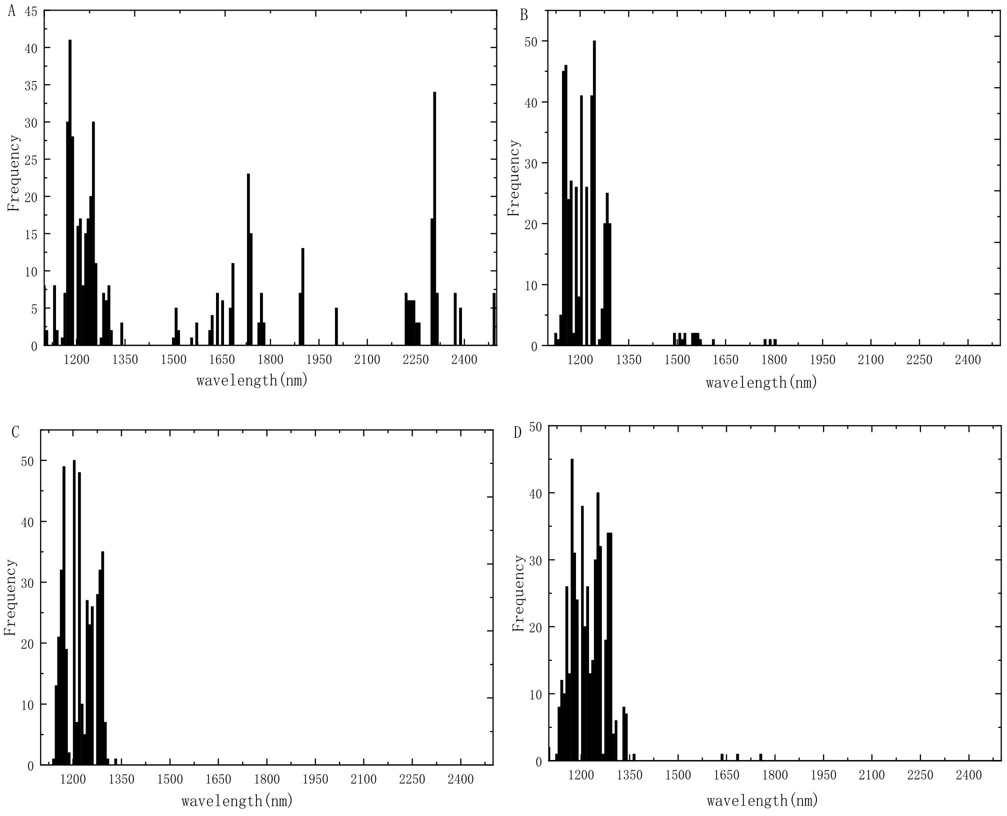


FIGURE 5. The frequency of selected variables within 50 times on the wheat dataset: (A) CARS, (B) VCPA, (C) BOSS, (D) SMCPA.

TABLE 1. Results for the wheat dataset. nVAR: number of variables; nLVs: number of latent variables; RMSECV: root-mean-square error of cross-validation; RMSEP: root-mean-square error of prediction; Q2_CV: coefficient of determination of cross-validation; Q2_test: coefficient of determination of test set; T/s: average computation time for 50 replicate runs; and statistical results with the form mean value ± standard deviation over 50 runs.

Characteristics	PLS	CARS	VCPA	BOSS	SMCPA
nVAR	175	12.9±4.4	8.9±1.0	6.6±0.7	6.7±0.7
nLVs	10	8.4±1.3	7.5±1.1	7.4±0.9	7.3±.0.5
RMSECV	0.607	0.257±0.034	0.249±0.019	0.235±0.031	0.232±0.032
RMSEP	0.519	0.418±0.019	0.289±0.005	0.305±0.004	0.234±0.005
Q2_CV	0.748	0.877±0.017	0.936±0.001	0.935±0.001	0.942±0.006
Q2_test	0.774	0.827±0.052	0.829±0.028	0.862±0.039	0.889±0.005
T/s	N/A	0.82	142.09	36.34	0.81

Compared to other selection methods,the SMCPA algorithm has better prediction performance in terms of the RMSECV,RMSEP, Q^2_cv, and Q^2_test, SMCPA showed the best results; RMSECV(0.00030), RMSEP(0.00031), Q^2_cv(0.99999), and Q^2_test(0.99999),were followed by CARS(0.00048, 0.000053, 0.99993 and 0.99992), VCPA (0.00039, 0.00045, 0.99995 and, 0.99996), BOSS(0.00036, 0.00039, 0.99997 and 0.99995).

Obviously, SMCPA shows the best predictive performance among all the four variable selection methods. From the 95% confidence interval, SMCPA yields the lowest standard deviation, indicating higher stability. In addition, SMCPA has selected fewer variables than other methods with lower RMSECV and RMSEP, which means that it can achieve better prediction performance with fewer variables for this dataset. In terms of algorithm computational cost, the average computational time of SMCPA was approximately 1.35 s, which demonstrates that the SMCPA comprehensively improves the efficiency of variable selection. Due to the high quality of spectral acquisition in the corn dataset, the other three algorithms also exhibit good variable selection results. However, the both of VCPA and BOSS are increase the computational cost to obtain better prediction performance of models. It is obvious that SMCPA is

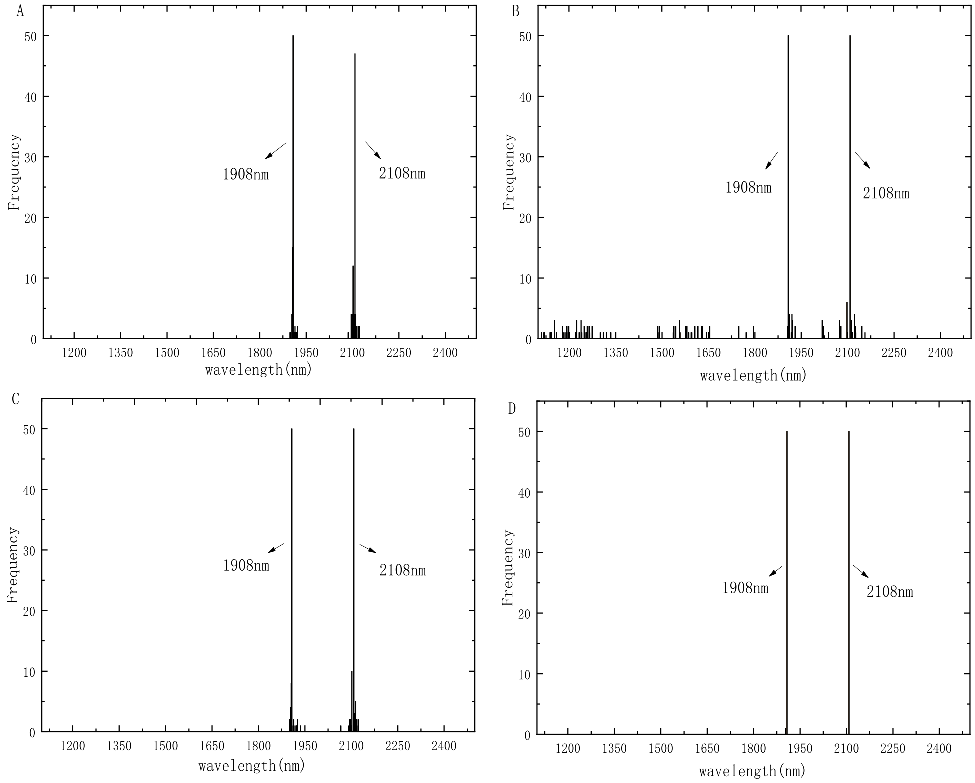


FIGURE 6. The frequency of selected variables within 50 times on the corn moisture dataset: (A) CARS, (B) VCPA, (C) BOSS, (D) SMCPA.

TABLE 2. Results for the corn dataset. nVAR: number of variables; nLVs: number of latent variables; RMSECV: root-mean-square error of cross-validation; RMSEP: root-mean-square error of prediction; Q2_CV: coefficient of determination of cross-validation; Q2_test: coefficient of determination of test set; T/s: average computation time for 50 replicate runs; statistical results have the form mean value ± standard deviation over 50 runs.

Characteristic s	PLS	CARS	VCPA	BOSS	SMCPA
nVAR	700	3.8±2.1	6.2±2.3	4.6±2.7	2±0.1
nLVs	10	3.7±1.4	5.7±2.1	4.6±2.6	2±0.2
RMSECV	0.02111	0.00048±0.00081	0.00039±0.00022	0.00036±0.00031	0.00030±0.00011
RMSEP	0.01522	0.00053±0.00082	0.00045±0.00053	0.00039±0.00023	0.00031±0.00001
Q2_CV	0.96710	0.99993±0.00033	0.99995±0.00021	0.99997±0.00012	0.99999±0.00000
Q2_test	0.93780	0.99992±0.00024	0.99996±0.00012	0.99995±0.00036	0.99999±0.00000
T/s	N/A	1.42	152.61	45.50	1.35

the best in model prediction performance and computational efficiency.

The variables selected by different methods over 50 runs on corn moisture datasets are displayed in Figure 6. The 1908nm and 2108 nm wavelengths, which correspond to the water absorption and the combination of O-H bond, were proved to be the key wavelengths in the reference [31], [35]. From the Figure.6 both of wavelengths were selected very frequently by all the methods. We found that CARS could not always select key variable 2108 nm in 50 replicate runs. For VCPA,

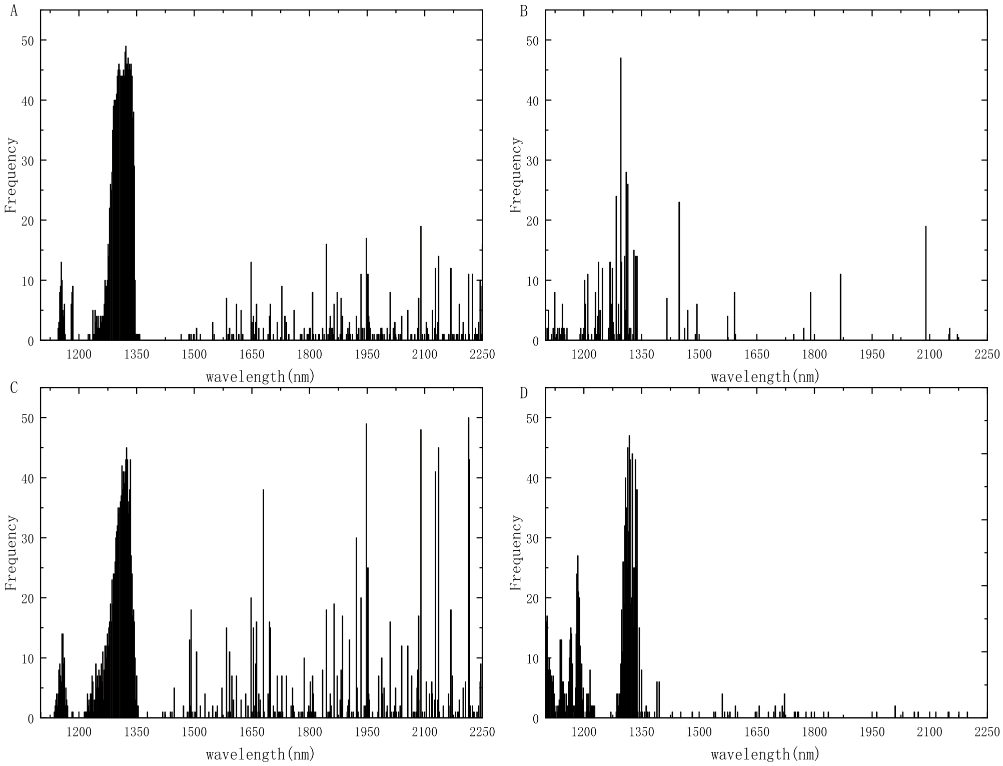


FIGURE 7. The frequency of selected variables within 50 times on the beer dataset: (A) CARS, (B) VCPA, (C) BOSS, (D) SMCPA.

in addition to choosing these two variables, it also choose many other wavelengths across the entire 1100–2500 nm NIR regions. The selected variables by VCPA included uninformative or interfering variables that lead to the lower prediction performance. SMCPA could always select two key variables in 50 replicate runs, 1908nm and 2108nm, which means that SMCPA is a very effective and stable variable selection algorithm.

D. BEER DATASET

The statistical results of variable selection methods, i.e., CARS, VCPA,BOSS and SMCPA, over 50 runs on the beer datasets are summarized in Table 3. From the table,it can be shown that variable selection methods have made much more predictive performance compared with the PLS full-spectrum. The RMSECV and RMSEP for SMCPA decreased by 86.5% and 83.8%, respectively, the values of RMSECV and RMSEP decreased from 0.622 and 0.822 to 0.084 and 0.133, respectively, and the values of Q2_cv and Q2_test increased from 0.940 and 0.852 to 0.998 and 0.995, respectively. Compared with CARS, VCPA, and BOSS, SMCPA has better prediction ability in cross-validation and the test results, the values of RMSECV and RMSEP were the lowest, and the values of Q2_CV and Q2_test were the highest. The enhancement of prediction ability by the SMCPA was significant. Although VCPA can also achieve very good predictive performance, it has some drawbacks. In the Ref [35], VCPA will eventually search for the optimal variable subset from the remaining 14 variables. Thus,VCPA is inclined to select fewer variables, which is unsuitable for many datasets. In the Ref [25], we found that VCPA is limited by the computer memory,When the number of residual variables is 16, it cannot be computed as a 65535(2^{16}-1) combination in a common computer due to an out-of-memory error. From the 95% confidence interval, SMCPA yields the lowest standard deviation, indicating higher stability. The beer dataset is an NIR spectral ensemble of 60 beer samples containing a rather large noisy part [53], so the selection variables methods and the PLS of models will be affected by the noise. However, the advantage of SMCPA is that it can eliminate the negative impact of noise and achieve the best predictive performance. In terms of algorithm computational cost, the average computational times of SMCPA and CARS were approximately 1 s, which proves the efficiency of the two methods. While VCPA and BOSS have higher computation times of 162.13s and 56.22s, respectively, although they have better performance models. Both VCPA and BOSS increase the computational cost to obtain better prediction performance of models. It is obvious that SMCPA is the best in predicting model performance and computational efficiency.

The variables selected by different methods over 50 runs on the beer datasets are displayed in Figure 7. CARS and BOSS tend to select a large number of variables, while the nVAR obtained by VCPA and SMCPA are lower. We found

TABLE 3. Results for the beer dataset. nVAR: number of variables; nLVs: number of latent variables; RMSECV: root-mean-square error of cross-validation; RMSEP: root-mean-square error of prediction; Q2_CV: coefficient of determination of cross-validation; Q2_test: coefficient of determination of test set; T/s: average computation time for 50 replicate runs; statistical results have the form mean value ± standard deviation over 50 runs.

Characteristics	PLS	CARS	VCPA	BOSS	SMCPA
nVAR	567	41.0±17.8	11.7±0.98	47.7±18.5	14.1±9.3
nLVs	10	5.7±1.1	9.5±0.7	7.9±1.9	3.5±1.3
RMSECV	0.622	0.168±0.031	0.095±0.006	0.100±0.006	0.084±0.001
RMSEP	0.822	0.502±0.009	0.172±0.306	0.591±0.049	0.133±0.029
Q2_CV	0.940	0.986±0.090	0.996±0.001	0.995±0.002	0.998±0.001
Q2_test	0.852	0.933±0.082	0.984±0.002	0.923±0.012	0.995±0.001
T/s	N/A	1.03	162.13	56.22	1.02

that VCPA selects the lowest variables of all methods (in Table 3),but SMCPA achieves better prediction performance than VCPA, we can infer the reason may be that VCPA misses some informative variables. In Figure 7, these four methods could succeeded in selecting informative regions of 1100nm–1350nm, which corresponds to the first overtone of the O-H stretch bond vibration. However, CARS, VCPA, and BOSS still select some redundant and uninformative variables outside of those informative regions(1100nm-1350nm). For example, variables selected by CARS were distributed in 1100-1350nm and 1600-2250nm, and variables selected by BOSS almost scattered across almost the entire NIR spectrum. The regions of 1350-2250nm were likely to have a large number of uninformative or interfering variables, leading to lower prediction performance of models. For the understanding and interpretation of the selected variables, the frequency of each variable within 50 replicate runs is shown in Figure.7. The SMCPA selection variables were mainly concentrated in the region 1100–1350 nm, which is consistent with the chemical properties of interest. This clearly demonstrates the excellent selectivity of SMCPA.

V. CONCLUSION

This paper has proposed a novel variable selection method, SMCPA. There are some obvious advantages to this approach. First, SMCPA theoretically make sure that iteratively and shrinks variable space to obtain the best variable combination, and the statistical analysis information (SMC) of the interesting output of a large number of sub-models is highlighted. Second, SMCPA combines new criteria of variable importance (SMC), statistical analysis(F-test) and some mathematical ideas (such as MC-sampling, EDF and WBS, etc.) in the algorithm, reducing the risk of eliminating informative variables and taking variable combination effects and interpretability into consideration. Finally, SMCPA has balance between computation ability and predictability.

With applications three real NIR spectral datasets, it was proved that SMCPA is a promising method for eliminating uninformative variables and constructing a high-performance calibration model. The results indicate that variable selection is really necessary and better prediction could be obtained using a few chemically meaningful key variables.

It should be noted that although SMCPA was employed on an NIR dataset, it is a general strategy and therefore could be coupled with other modeling methods and applied to other areas, such as genomics, proteomics, bioinformatics, metabolomics, quantitative structure–activity relationship (QSAR), etc. Our future work will focus on investigating the application of SMCPA in other fields.

REFERENCES

[1] R. Tauler, B. Walczak, and S. D. Brown, *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*. Amsterdam, The Netherlands: Elsevier, 2009.

[2] L.-L. Wang, Y.-W. Lin, X.-F. Wang, N. Xiao, Y.-D. Xu, H.-D. Li, and Q.-S. Xu, "A selective review and comparison for interval variable selection in spectroscopic modeling," *Chemometrics Intell. Lab. Syst.*, vol. 172, pp. 229–240, Jan. 2018.

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

[4] B.-C. Deng, Y.-H. Yun, Y.-Z. Liang, D.-S. Cao, Q.-S. Xu, L.-Z. Yi, and X. Huang, "A new strategy to prevent over-fitting in partial least squares models based on model population analysis," *Analytica Chim. Acta*, vol. 880, pp. 32–41, Jun. 2015.

[5] L. Shi, J. A. Westerhuis, J. Rosen, R. Landberg, and C. Brunius, "libPLS: An integrated library for partial least squares regression and linear discriminant analysis," *Bioinformatics*, vol. 35, pp. 972–980, Mar. 2019.

[6] O. M. Kvalheim, R. Arneberg, O. Bleie, T. Rajalahti, A. K. Smilde, and J. A. Westerhuis, "Variable importance in latent variable regression models," *J. Chemometrics*, vol. 28, no. 8, pp. 615–622, Aug. 2014.

[7] A. M. Aguilera, M. Escabias, C. Preda, and G. Saporta, "Using basis expansions for estimating functional PLS regression: Applications with chemometric data," *Chemometrics Intell. Lab. Syst.*, vol. 104, pp. 289–305, Dec. 2010.

[8] J. Fan and R. Li, "Statistical challenges with high dimensionality: Feature selection in knowledge discovery," in *Proc. Int. Congr. Math.*, 2006, pp. 595–622.

[9] L. Shi, J. A. Westerhuis, J. Rosén, R. Landberg, and C. Brunius, "Variable selection and validation in multivariate modelling," *Bioinformatics*, vol. 35, pp. 972–980, Mar. 2019.

[10] H. Liu, X. Li, J. Li, and S. Zhang, "Efficient outlier detection for high-dimensional data," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 48, no. 12, pp. 2451–2461, Dec. 2018.

[11] H. Chun and S. Keleş, "Sparse partial least squares regression for simultaneous dimension reduction and variable selection," *J. Roy. Stat. Soc. Stat. Methodol. B*, vol. 72, no. 1, pp. 3–25, 2010.

[12] H.-D. Li, Y.-Z. Liang, X.-X. Long, Y.-H. Yun, and Q.-S. Xu, "The continuity of sample complexity and its relationship to multivariate calibration: A general perspective on first-order calibration of spectral data in analytical chemistry," *Chemometrics Intell. Lab. Syst.*, vol. 122, pp. 23–30, Mar. 2013.

[13] Y.-H. Yun, Y.-Z. Liang, G.-X. Xie, H.-D. Li, D.-S. Caoa, and Q.-S. Xuc, "A perspective demonstration on the importance of variable selection in inverse calibration for complex analytical systems," *Analyst*, vol. 138, no. 21, pp. 6412–6421, 2013.

[14] X. Song, Y. Huang, H. Yan, Y. Xiong, and S. Min, "A novel algorithm for spectral interval combination optimization," *Analytica Chim. Acta*, vol. 948, pp. 19–29, Dec. 2016.

[15] F. G. Blanchet, P. Legendre, and D. Borcard, "Forward selection of explanatory variables," *Ecology*, vol. 89, pp. 2623–2632, Sep. 2008.

[16] T. Yamashita, K. Yamashita, and R. Kamimura, "A stepwise AIC method for variable selection in linear regression," *Commun. Statist., Theory Methods*, vol. 36, no. 13, pp. 2395–2403, 2007.

[17] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Statist. Softw.*, vol. 33, no. 1, pp. 1–22, 2010.

[18] Y.-W. Lin, N. Xiao, L.-L. Wang, C.-Q. Li, and Q.-S. Xu, "Ordered homogeneity pursuit lasso for group variable selection with applications to spectroscopic data," *Chemometrics Intell. Lab. Syst.*, vol. 168, pp. 62–71, Sep. 2017.

[19] T. Ge, B. Wei, D. Wu, F. Peng, J. Liu, Y. Tang, S. Xiong, and Z. Zhang, "The optimal wavelengths for light absorption spectroscopy measurements based on genetic algorithm–particle swarm optimization," *J. Appl. Spectrosc.*, vol. 85, pp. 109–118, Mar. 2018.

[20] Y. Lee and C.-H. Wei, "A computerized feature selection method using genetic algorithms to forecast freeway accident duration times," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 25, no. 2, pp. 132–148, 2010.

[21] M. Farrés, S. Platikanov, S. Tsakovski, and R. Tauler, "Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation," *J. Chemometrics*, vol. 29, pp. 528–536, Oct. 2015.

[22] N. L. Afanador, T. N. Tran, and L. M. C. Buydens, "Use of the bootstrap and permutation methods for a more robust variable importance in the projection metric for partial least squares regression," *Analytica Chim. Acta*, vol. 768, pp. 49–56, Mar. 2013.

[23] Y.-H. Yun, D.-M. Wu, G.-Y. Li, Q.-Y. Zhang, X. Yang, Q.-F. Li, D.-S. Cao, and Q.-S. Xu, "A strategy on the definition of applicability domain of model based on population analysis," *Chemometrics Intell. Lab. Syst.*, vol. 170, pp. 77–83, Nov. 2017.

[24] X.-X. Long, H.-D. Li, W. Fan, Q.-S. Xu, and Y.-Z. Liang, "A model population analysis method for variable selection based on mutual information," *Chemometrics Intell. Lab. Syst.*, vol. 121, pp. 75–81, Feb. 2013.

[25] Y.-H. Yun, W.-T. Wang, B.-C. Deng, G.-B. Lai, X. Liu, D.-B. Ren, Y.-Z. Liang, W. Fan, and Q.-S. Xu, "Using variable combination population analysis for variable selection in multivariate calibration," *Anal. Chim. Acta*, vol. 862, pp. 14–23, Mar. 2015.

[26] Y. Wang, F. Jiang, B. B. Gupta, S. Rho, Q. Liu, and H. Hou, "Variable selection and optimization in rapid detection of soybean straw biomass based on CARS," *IEEE Access*, vol. 6, pp. 5290–5299, 2018.

[27] M. Wen, B.-C. Deng, D.-S. Cao, Y.-H. Yun, R.-H. Yang, H.-M. Lu, and Y.-Z. Liang, "The model adaptive space shrinkage (MASS) approach: A new method for simultaneous variable selection and outlier detection based on model population analysis," *Analyst*, vol. 141, no. 19, pp. 5586–5597, 2016.

[28] J. Bin, F. Ai, W. Fan, J. Zhou, X. Li, W. Tang, and Y. Liang, "An efficient variable selection method based on variable permutation and model population analysis for multivariate calibration of NIR spectra," *Chemometrics Intell. Lab. Syst.*, vol. 158, pp. 1–13, Nov. 2016.

[29] B.-C. Deng, Y.-H. Yun, D.-S. Cao, Y.-L. Yin, W.-T. Wang, H.-M. Lu, Q.-Y. Luo, and Y.-L. Liang, "A bootstrapping soft shrinkage approach for variable selection in chemical modeling," *Analytica Chim. Acta*, vol. 908, pp. 63–74, Feb. 2016.

[30] Y.-W. Lin, B.-C. Deng, L.-L. Wang, Q.-S. Xu, L. Liu, and Y.-Z. Liang, "Fisher optimal subspace shrinkage for block variable selection with applications to NIR spectroscopic analysis," *Chemometrics Intell. Lab. Syst.*, vol. 159, pp. 196–204, Dec. 2016.

[31] B. Mahanty, "Alternate deflation and inflation of search space in reweighted sampling: An effective variable selection approach for PLS model," *Chemometrics Intell. Lab. Syst.*, vol. 174, pp. 45–55, Mar. 2018.

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

[33] B.-C. Deng, Y.-H. Yun, Y.-Z. Liang, and L.-Z. Yic, "A novel variable selection approach that iteratively optimizes variable space using weighted binary matrix sampling," *Analyst*, vol. 139, no. 19, pp. 4836–4845, 2014.

[34] K. Zheng, Q. Li, J. Wang, J. Geng, P. Cao, T. Sui, X. Wang, and Y. Dua, "Stability competitive adaptive reweighted sampling (SCARS) and its applications to multivariate calibration of NIR spectra," *Chemometrics Intell. Lab. Syst.*, vol. 112, pp. 48–54, Mar. 2012.

[35] Y.-H. Yun, J. Bin, D.-L. Liu, L. Xu, T.-L. Yan, D.-S. Cao, and Q.-S. Xu, "A hybrid variable selection strategy based on continuous shrinkage of variable space in multivariate calibration," *Analytica Chim. Acta*, vol. 1058, pp. 58–69, Jun. 2019.

[36] W. Wang, Y. Yun, B. Deng, W. Fan, and Y. Liang, "Iteratively variable subset optimization for multivariate calibration," *RSC Adv.*, vol. 5, no. 116, pp. 95771–95780, 2015.

[37] B.-C. Deng, Y.-H. Yun, P. Ma, C.-C. Lin, D.-B. Ren, and Y.-Z. Liang, "A new method for wavelength interval selection that intelligently optimizes the locations, widths and combinations of the intervals," *Analyst*, vol. 140, no. 6, pp. 1876–1885, 2015.

[38] R. Zhang, F. Zhang, W. Chen, H. Yao, J. Ge, S. Wu, T. Wu, and Y. Du, "A new strategy of least absolute shrinkage and selection operator coupled with sampling error profile analysis for wavelength selection," *Chemometrics Intell. Lab. Syst.*, vol. 175, pp. 47–54, Apr. 2018.

[39] H.-D. Li, Y.-Z. Liang, D.-S. Cao, and Q.-S. Xu, "Model-population analysis and its applications in chemical and biological modeling," *TrAC Trends Anal. Chem.*, vol. 38, pp. 154–162, Sep. 2012.

[40] B.-C. Deng, Y.-H. Yun, and Y.-Z. Liang, "Model population analysis in chemometrics," *Chemometrics Intell. Lab. Syst.*, vol. 149, pp. 166–176, Dec. 2015.

[41] J. Fan and J. Lv, "A selective overview of variable selection in high dimensional feature space," *Statistica Sinica*, vol. 20, no. 1, p. 101, 2010.

[42] Y.-H. Yun, H.-D. Li, B.-C. Deng, and D.-S. Cao, "An overview of variable selection methods in multivariate analysis of near-infrared spectra," *TrAC Trends Anal. Chem.*, vol. 113, pp. 102–115, Apr. 2019.

[43] C. Pasquini, "Near infrared spectroscopy: A mature analytical technique with new perspectives—A review," *Analytica Chim. Acta*, vol. 1026, pp. 8–36, Oct. 2018.

[44] K. Sindhya, K. Miettinen, and K. Deb, "A hybrid framework for evolutionary multi-objective optimization," *IEEE Trans. Evol. Comput.*, vol. 17, no. 4, pp. 495–511, Aug. 2013.

[45] A. Singhee and R. A. Rutenbar, "Why quasi-Monte Carlo is better than Monte Carlo or latin hypercube sampling for statistical circuit analysis," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 29, no. 11, pp. 1763–1776, Nov. 2010.

[46] T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø, "A review of variable selection methods in partial least squares regression," *Chemometrics Intell. Lab. Syst.*, vol. 118, pp. 62–69, Aug. 2012.

[47] O. M. Kvalheim, "Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots," *J. Chemometrics*, vol. 24, pp. 496–504, Jul./Aug. 2010.

[48] T. N. Tran, N. Lee Afanador, L. M. C. Buydens, and L. Blanchet, "Interpretation of variable importance in partial least squares with significance multivariate correlation (sMC)," *Chemometrics Intell. Lab. Syst.*, vol. 138, pp. 153–160, Nov. 2014.

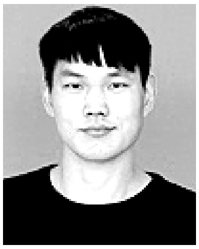
[49] I.-G. Chong and C.-H. Jun, "Performance of some variable selection methods when multicollinearity is present," *Chemometrics Intell. Lab. Syst.*, vol. 78, pp. 103–112, Jul. 2005.

[50] N. L. Afanador, T. N. Tran, L. Blanchet, and L. M. C. Buydens, "Variable importance in PLS in the presence of autocorrelated data—Case studies in manufacturing processes," *Chemometrics Intell. Lab. Syst.*, vol. 139, pp. 139–145, Dec. 2014.

[51] F. Meinecke, A. Ziehe, M. Kawanabe, and K.-R. Müller, "A resampling approach to estimate the stability of one-dimensional or multidimensional independent components," *IEEE Trans. Biomed. Eng.*, vol. 49, no. 12, pp. 1514–1525, Dec. 2002.

[52] R. Leardi, "Application of genetic algorithm–PLS for feature selection in spectral data sets," *J. Chemometrics*, vol. 14, nos. 5–6, pp. 643–655, 2010.

[53] L. Nørgaard, 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, no. 3, pp. 413–419, 2000.



YUXI WANG received the bachelor's degree in electronic information engineering from the School of Physical Science and Information Engineering, Liaocheng University, China, in 2016. He is currently pursuing the master's degree with the School of Information Science and Engineering, Xinjiang University, China. His research direction is spectral analytical chemistry and metrological analytical chemistry.

ZHENHONG JIA received the B.S. degree from Beijing Normal University, Beijing, China, in 1985, and the M.S. degree and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 1987 and 1995, respectively. He is currently a Professor with the Autonomous University Key Laboratory of Signal and Information Processing Laboratory, Xinjiang University, China. His research interests include digital image processing, and photoelectric information detection and sensor.

JIE YANG received the Ph.D. degree from the Department of Computer Science, Hamburg University, Germany, in 1994. He is currently a Professor with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China. His major research interests are object detection and recognition, data fusion and data mining, and medical image processing.