Research Article

Moving-Window-Improved Monte Carlo Uninformative Variable Elimination Combining Successive Projections Algorithm for Near-Infrared Spectroscopy (NIRS)

Weiwei Jiang,¹ Changhua Lu,^{1,2} Yujun Zhang,² Wei Ju,³ Jizhou Wang,⁴ Feng Hong,¹ Tao Wang,¹ and Chunsheng Ou¹

¹School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China

²Anhui Institute of Optics Fine Mechanics, Chinese Academy of Sciences, Hefei 230031, China

³School of Internet, Anhui University, Hefei 230039, China

⁴Department of Electronics, Hefei University, Hefei 230061, China

Correspondence should be addressed to Chunsheng Ou; ouchunsheng@hfut.edu.cn

Received 17 February 2020; Revised 26 June 2020; Accepted 11 July 2020; Published 3 August 2020

Academic Editor: Alessandra Durazzo

Copyright © 2020 Weiwei Jiang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The MC-UV-SPA method is commonly proposed as a variable selection approach for multivariate calibration. However, the SPA tends to select wavelength variables that are sparsely distributed over the wavelength ranges of the variables selected by the MC-UV-SPA algorithm, and the MC-UV-SPA cascade cannot improve the problem of wavelength point discontinuity. It is addressed in this paper by proposing a moving-window- (MW-) improved MC-UV-SPA wavelength selection algorithm. The proposed algorithm improves the continuity of the selected wavelength variables and thereby better exploits the advantages of the MC-UV-SPA algorithm and the SPA to obtain regression models with high prediction accuracy. The MC-UV-SPA, MC-UV-SPA, and MC-UV-SPA-MW algorithms are applied for conducting wavelength variable selection for the NIR spectral absorbance data of corn, diesel fuel, and ethylene. Here, partial least squares regression (PLSR) models reflecting the oil content of corn, the boiling point of diesel fuel, and the ethylene concentration are established after conducting wavelength selection using the MC-UV-SPA algorithm, and corresponding multiple linear regression (MLR) models are established after conducting wavelength selection using the MC-UV-SPA and MC-UV-SPA-MW algorithms. Experimental results demonstrate that the progressive elimination of uncorrelated and collinear variables generates increasingly simplified partial-spectrum models with greater prediction accuracy than the full-spectrum model. Among the three wavelength selection algorithms, the MC-UV-SPA selected the least number of wavelength variables, while the proposed MC-UV-SPA-MW algorithm provided models with the greatest prediction accuracy.

1. Introduction

With the characteristics of simple, rapid, noninvasive, and no sample pretreatment, near-infrared (NIR) spectroscopy [1] has been adopted as a popular analytical tool for both qualitative and quantitative analyses in various fields [2–5]. The quantitative analysis of NIR spectral data is generally conducted through the construction of regression models, such as those based on principle component analysis (PCA) [6], partial least squares (PLS) regression [7], and multiple linear regression (MLR) [8], which take the characteristic wavelengths of the spectral data as input variables. However,

the development of modern analytical instruments has led to the capability of acquiring NIR spectral data that can easily contain hundreds to tens of thousands of individual wavelengths [9]. Thus, the full-band spectral data were adopted for modeling, but the model contained a large amount of redundant information, which resulted in inefficiency [10]. In addition, spectral data usually contain noise, interference, and/or mixed spectral components that can often greatly detract from the prediction accuracy of full-spectrum models developed for spectral data analysis [11]. Yun et al. pointed out that there are three ways to address these problems, namely, regularization, dimension reduction,

and variable selection [12]. Among the above-discussed methods, variable selection has become the dominant method of interest in recent years for the development of NIR spectral analysis technology and chemometrics [11–14].

The goal of wavelength selection is to identify the most informative wavelengths for use as variables in partial-spectrum regression models. Here, uninformative wavelength variables have either no effect or a negative effect on the modeling performance. The wavelength selection process fulfills three purposes, including (1) providing models with greater predictive capability, (2) obtaining wavelength variables that provide greater modeling efficiency, and (3) providing simpler models with improved interpretability [9]. The most commonly employed wavelength selection algorithms developed thus far include uninformative variable elimination (UVE) and the successive projections algorithm (SPA).

The goal of UVE, first proposed by Centner et al. [15], is not to select variables directly, but to effectively eliminate uninformative variables in the spectral data, such that only informative wavelength variables remain. The SPA employs simple projection to select variables with a minimum of collinearity, but variables selected by SPA may make little contribution to multivariate calibration, which can affect model prediction [16]. A significant development in recent years has been the combined use of different algorithms through a cascade strategy, where the results of one wavelength selection algorithm are used as the inputs of the next selection algorithm in a stepwise manner. This can combine the advantages of various wavelength selection algorithms in a complementary way and thereby obtain better and more effective prediction results. The common variable selection method combined with SPA method can greatly simplify the model and improve the prediction accuracy. This strategy has been effectively used in many studies to address the problem associated with the application of the SPA to NIR spectral data by first reducing the dimension of the spectral data by applying some initial algorithm such as UVE, MC-UVE, particle swarm optimization (PSO), or genetic algorithm (GA) optimization [16–20]. Among them, UVE and MC-UVE are commonly used as the primary wavelength algorithms of SPA. For example, Ye et al. proposed the combination of UVE and SPA to integrate the bright side of each, successfully applied to the NIR spectroscopic analysis of nicotine in tobacco lamina and active pharmaceutical ingredients in intact tablets for variable selection, UVE was employed to select informative variables, and SPA was followed to select variables that have minimum redundant information from the informative variables [20]. Li et al. proposed a new combination of MC-UVE and SPA, MC-UVE was employed to select informative variables in the full spectrum, and SPA was also employed as a powerful method for further characteristic variable selection [18].

Nonetheless, most of the informative wavelengths in a molecular NIR spectrum typically exhibit some continuity, where wavelength points adjacent to an informative wavelength point also represent informative wavelengths [21]. However, the MC-UVE algorithm and the SPA are both wavelength selection algorithms based on optimal

wavelength points, which are most likely isolated points along the full NIR spectrum. The MC-UVE-SPA cascade cannot improve the problem of wavelength point discontinuity, which may result in the least number of selected wavelength variables, but the modeling effect is not the best. Fan et al. constructed a model for visible/NIR spectral data reflecting the lycopene content based on wavelength variable selection obtained using UVE, SPA, and CARS individually and in various two-stage cascaded combinations [22]. The UVE-SPA combination was found to retain the smallest number of wavelength variables of all the selection algorithms considered, but the prediction accuracy of the model constructed using this wavelength variable set was the worst of all models obtained using all other wavelength selection algorithms. Sun et al. showed that the prediction results of the model constructed by the cascaded wavelength selection algorithm were not always the most accurate, and the prediction results of the improved cascaded wavelength selection algorithm were better than those of the direct two-stage cascaded strategy [23].

Few studies have considered improving the continuity of the selected wavelength in the wavelength point selection algorithm. Therefore, this paper considers the continuity of the wavelength selected by the MC-UVE-SPA. In this study, this is employed as a moving-window-improved cascade strategy for wavelength selection that is herein denoted as the MC-UVE-SPA-MW algorithm. First, the uninformative variable is eliminated by MC-UVE, the collinear variable is eliminated by SPA, and then the wavelength variables are selected by extending outward from the optimal wavelength points by MC-UVE-SPA in conjunction with a moving window. This reduces the number of isolated wavelength variables, preserves the continuity between informative wavelength points in an NIR spectrum, and expects to improve the accuracy of the established prediction model.

2. Materials and Methods

2.1. Experiments and Data. Experiments based on the NIR spectral absorbance data of corn, diesel fuel, and ethylene were employed for verifying the wavelength variable selection performance of the proposed MC-UVE-SPA-MW algorithm and were conducted using the libPLS toolkit [24], while the remaining code was written and executed in the MATLAB R2017b environment.

2.1.1. Corn Spectral Data. The NIR spectral absorbance data for corn were provided by Eigenvector Research, Inc. (<http://www.eigenvector.com/data/Corn/index.html>). The m5 spectra of corn data set consist of 80 corn samples measured over a wavelength range of 1100–2498 nm in 2 nm intervals. Accordingly, the data set includes a total of 700 wavelength points. It also contains four component reference values of moisture, oil, protein, and starch contents determined by chemical methods for each sample. Table 1 shows the maximum, minimum, and average values of the relative concentrations of moisture, oil, protein, and starch in the 80 corn samples.

TABLE 1: Statistical analysis of the relative concentrations of corn.

Component to be tested	Max	Min	Ave
Moisture	10.9930	9.3770	10.2335
Oil	3.8320	3.0880	3.4984
Protein	9.7110	7.6540	8.6683
Starch	66.4720	62.8260	64.6956

2.1.2. Diesel Fuel Spectral Data. The NIR spectral absorbance data for diesel fuel were provided by the Southwest Research Institute (SWRI) (<http://www.eigenvector.com/data/SWRI/index.html>). The data set comprises unprocessed spectra derived from 784 diesel fuel samples measured over a wavelength range of 550~750 nm in 2 nm intervals. Accordingly, the data set includes a total of 401 wavelength points. The data set also contains various properties including the boiling point, cetane number, density, freezing point, total aromatic hydrocarbon content, and viscosity. Some of the parameter samples have missing values (NaN), which are eliminated during the experiment. Table 2 shows the maximum, minimum, and average values of the boiling point of diesel fuel.

2.1.3. Ethylene Gas Spectral Data. Ethylene gas samples were prepared within a closed cell filled with nitrogen gas at a pressure of 1 atm and a temperature of 296 K by distributing C_2H_4 gas into the cell to form samples with 72 known C_2H_4 concentrations ranging from 60.15 ppm to 200.5 ppm in 2.005 ppm intervals. The C_2H_4 gas distribution device adopted a gas distribution platform, shown in Figure 1, independently developed by the Hefei Material Science Research Institute of the Chinese Academy of Sciences. Through visual control software, set the gas distribution proportion according to the requirements, adjust the volume ratio of the auxiliary gas nitrogen and the gas to be distributed through the high-precision gas distribution platform, and configure the required concentration of standard gas according to the requirements. Fourier transform infrared (FTIR) spectroscopy was applied to capture the spectral absorbance intensity of the gas in a sealed sample cell. The optical path length of the cell was 10 m, and the range of the measured wavenumbers was 400~5000 cm^{-1} with a resolution of 1 cm^{-1} . The apodization function used a Hamming window, the number of scans was 16, and a total of 96 spectral data of different concentrations were collected.

Accordingly, the data set includes a total of 4601 wavelength points. The absorption spectrum of C_2H_4 gas obtained from the HITRAN database (<http://hitran.iao.ru/>) over a wavenumber range of 400~5000 cm^{-1} is shown in Figure 2. Figure 3 presents the background spectral intensity measured after the closed cell was filled with nitrogen gas at room temperature. Figure 4 presents the measured absorption spectral intensity of the cell after adding various concentrations of C_2H_4 gas. A comparison of Figures 3 and 4 indicates that the spectral intensities in the two regions of 794~1105 cm^{-1} and 2917~3242 cm^{-1} are drastically different due to the spectral absorption characteristics of the added C_2H_4 gas.

TABLE 2: Statistical analysis of the boiling point of diesel fuel.

Property to be tested	Effective sample size	Max	Min	Ave
Boiling point	395	297	182	258.37

2.2. Evaluation Indices. The NIR spectral absorbance data are first preprocessed to generate normalized data for facilitating consistent analyses. The normalized data are then divided into a calibration data set and a prediction data set, which are respectively applied for establishing the various regression models and for testing the established models, by adopting the Kennard–Stone method (3:1). The extent of information provided by the selected wavelength variables is generally difficult to directly evaluate. Therefore, indirect evaluation methods are usually adopted. Typically, the information value of wavelength variables is evaluated according to the prediction accuracy of the model constructed with the selected wavelengths. The indices for evaluating the prediction accuracy of regression models are the root mean square error of cross validation (RMSECV) for calibration set, the root mean square error of prediction, the correlation coefficient (r), and the relative percent deviation (RPD) for prediction set. These indices are defined as follows:

$$\begin{aligned}
 RMSECV &= \sqrt{\frac{\sum_{i=1}^n (y_k - y_k)^2}{n-1}}, \\
 RMSEP &= \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n-1}}, \\
 r &= \frac{\sum_{i=1}^n (y_i - y_{AVE})(\hat{y}_i - \hat{y}_{AVE})}{\sqrt{\sum_{i=1}^n (y_i - y_{AVE})^2 (\hat{y}_i - \hat{y}_{AVE})^2}}, \\
 RPD &= \sqrt{\frac{\sum_{i=1}^n (y_i - y_{AVE})^2}{\sum_{i=1}^n (\hat{y}_i - y_i)^2}}.
 \end{aligned} \tag{1}$$

Here, n is the number of samples in the calibration set or the prediction set, y_k is the measured value and $\hat{y}_k$ is the predicted value of sample i in calibration set, y_i is the measured value and $\hat{y}_i$ is the predicted value of sample i in prediction set, and y_{AVE} and $\hat{y}_{AVE}$ are the respective average measured value and the average predicted value of all samples in prediction set.

We note that the evaluated prediction performance increases with decreasing RMSE and increasing r and RPD. The RMSE is denoted as the RMSECV when referring to the value associated with the calibration data set and as the RMSEP when referring to the value associated with the prediction data set.

2.3. MC-UVE-SPA Method. The fundamental basis of UVE is to use the stability of the regression coefficient vector characteristic of a constructed PLS multiple regression model as a measure of the significance of a given wavelength. However, the UVE tends to suffer from model overfitting

FIGURE 1: Schematic diagram of the standard gas spectra acquisition system.

FIGURE 2: NIR absorbance spectrum for C₂H₄ gas from the HITRAN database.

[25]. This was addressed by the development of Monte Carlo (MC) UVE (MC-UVE), proposed by Cai et al. [26], which replaces the leave-one-out cross-validation (LOOCV) process calculating the regression coefficient matrix $\beta = [\beta_1, \beta_2, \dots, \beta_j]$ in conventional UVE with the MC cross-validation (MCCV) process. The reliability of each variable j can be quantitatively measured by

$$S_j = \frac{\text{mean}(\beta_j)}{\text{std}(\beta_j)}, \quad (2)$$

where $\text{mean}(\beta_j)$ and $\text{std}(\beta_j)$ are the mean and standard deviation of the regression coefficients of variable j . The greater the absolute value of stability, the more important the corresponding variable. The stability of uninformative variables should be less than a threshold.

The SPA, first proposed by Brègman [27], is a forward-cycling variable selection method. For spectral data analysis, each cycle of the process calculates the projection of a selected wavelength on an unselected wavelength and includes the unselected wavelength with the largest projection vector in the set of selected wavelengths [28]. This process is repeated for each selected wavelength as it is added to the set until the selected wavelength set includes a specified number of wavelengths [16]. More detailed information on the steps of SPA can be seen in literature [16, 29]. In selecting the next wavelength, each of the newly selected wavelengths has the lowest correlation with the previous one. Therefore, SPA can effectively eliminate collinear wavelength variables and reduce the number of dimensions of the sample spectrum, which accordingly reduces the calculation burden of the model.

FIGURE 3: Measured background NIR spectrum with the cell filled with nitrogen gas (N_2).

FIGURE 4: Measured C_2H_4 gas absorbance spectrum.

The MC-UVE-SPA method is a combination method of MC-UVE and SPA. Jiangbo Li et al. proved that the combination (MC-UVE-SPA) of both Monte Carlo uninformative variable elimination (MC-UVE) and successive projections algorithm (SPA) was more effective than MC-UVE or SPA alone [30]. Although the effect of UVE-SPA is better than that of using UVE or SPA alone, there is still something to be improved. In this paper, the UVE-SPA is improved by using the wavelength effective continuity and its effectiveness is verified by experiments.

2.4. Proposed MC-UVE-SPA-MW Wavelength Selection Algorithm. The proposed wavelength selection algorithm first applies MC-UVE to the calibration data set to construct

a PLS regression model. The threshold of the MC-UVE process is set to provide a number of wavelength variables that minimize the RMSECV of the constructed PLS regression model. The largest number of principal components (PCs) was set to 10, and the optimal number of PCs was determined based on the minimum RMSECV value. Subsequently, the wavelength variables retained by the MC-UVE algorithm are applied as the input of the SPA. Here, an MLR model is constructed based on the wavelength variables selected by the SPA for conducting cross-validation analysis, where the number of selected wavelength variables is determined according to the minimum of the RMSECV of the constructed MLR model. In order to reduce the number of isolated wavelength variables and maintain the continuity of adjacent information wavelength points of near-infrared

spectrum, it extends outward from the best wavelength point selected by UVE-SPA. In the original spectrum, the optimal wavelength point selected by the MC-UVE-SPA is used as the starting point of a moving window of width w . The wavelengths in the moving window are used as the selected wavelengths, and the number of wavelengths finally selected varies with the window width. Set the window width $w = 2$ (Left), 2 (Right), or 3. Here, 2 (Left) means to extend the selected wavelength point to the left to 2 wavelength points, 2 (Right) means to extend the selected wavelength point to the right to 2 wavelength points, and 3 means to extend the selected wavelength point to 3 wavelength points using the selected wavelength point as the center point. The optimal window width is determined by the minimum RMSECV of the MLR model. The processing flow of the proposed MC-UVE-SPA-MW algorithm and extending the wavelength point outward are illustrated in Figure 5.

3. Results and Discussion

3.1. Corn Spectral Data Experiments. The wavelength variable stability distribution map of the PLS regression model reflecting the oil concentration in corn constructed for calibration set using the MC-UVE algorithm is presented in Figure 6. Here, all wavelengths greater than the threshold value shown by the horizontal red line in the figure are selected for use in the model. This threshold was selected to provide the number of wavelength variables corresponding to the minimum RMSECV of the constructed PLS regression model. This is illustrated in Figure 7, where the RMSECV of the constructed PLS regression model is plotted with respect to the number of selected wavelength variables. It can be seen from Figure 7 that the RMSECV is relatively large when the number of wavelength variables is small, and the RMSECV drops sharply as the number of selected variables increases. This is because an overly small number of wavelength variables exclude useful information, and the prediction accuracy of the model is therefore improved as an increasing amount of useful information is incorporated into the model. A minimum value of RMSECV = 0.0289 is obtained when the number of selected wavelength variables is 106, and the RMSECV increases again when the number of variables exceeds 106. This increase results from the impact of selecting an increasing number of uninformative variables on the prediction accuracy of the model. We also note that the RMSECV changes very little when the number of wavelength variables exceeds 300. Thus, the MC-UVE algorithm eliminates a large number of wavelengths that are not related to the oil concentration of corn, where the final number of selected wavelength variables is just 15.1% of the full-spectrum value of 700.

The optimal number of 106 wavelength variables selected by MC-UVE is then used as the inputs of the SPA, which iteratively generates wavelength variable combinations using each wavelength as a starting point and applies them for constructing an MLR model. The wavelength combination corresponding to the minimum RMSECV of the MLR model is then taken as the optimal wavelength combination. The relationship between the number of selected wavelength

variables and the RMSECV of the MLR model constructed from variables selected by the MC-UVE-SPA is shown in Figure 8, where we note that the minimum RMSECV is obtained when the number of selected variables is 37. Thus, the SPA further reduces the number of informative wavelengths mainly by eliminating collinear variables in the MLR model, where the final number of selected wavelength variables is reduced to just 5.3% of the full-spectrum value of 700.

In the original spectrum, the optimal wavelength point selected by the MC-UVE-SPA is used as the starting point or center of a moving window of width $w = 2$ (Left), 2 (Right), or 3. The results of the PLS or MLR model constructed using the wavelength variables selected by different algorithms are shown in Figure 9, and the details are listed in Table 3 along with the results obtained for different models. In Table 3, the optimal number of PLS principal components was 10. As shown in Table 3, there were 37 characteristic wavelengths selected by the MC-UVE-SPA, accounting for only 5.3% of the total number of wavelengths, and the accuracy of the algorithm is better than that of MC-UVE algorithm, which is due to the elimination of wavelength collinearity. The wavelengths selected by the MC-UVE-SPA-MW are extended by the algorithm proposed in this paper. When the window width $w = 2$ (Left) and $w = 3$, the model accuracy of the MC-UVE-SPA-MW algorithm is higher than that of the MC-UVE-SPA. When $w = 2$ (Left), the MC-UVE-SPA-MW expands 37 wavelength variables selected by the MC-UVE-SPA to 64. At this point, RMSEP is 0.0381, r value is 0.9713, RPD value is 16.3666, and the model is optimal. Although the MC-UVE-SPA-MW provides improved continuity by increasing the number of wavelength variables from those obtained by the MC-UVE-SPA, the final number is still just 9.1% of the full-spectrum value of 700.

The wavelength variables selected from the NIR spectral absorbance data of a single sample by the MC-UVE, MC-UVE-SPA, and proposed MC-UVE-SPA-MW algorithms are compared in Figure 10. The results in Figure 10 are derived from the fact that oil is a complex organic molecule with infrared and NIR spectral absorption that occupies a wide wavenumber band ranging 3900~12000 cm^{-1} (833~2564 nm). This is mainly caused by the frequency doubling and frequency combinations of the stretching and vibrational energy level transitions of hydrogen-containing groups. From the results of Figure 10, we note that the wavelength variables selected by the MC-UVE, MC-UVE-SPA, and proposed MC-UVE-SPA-MW algorithms are mainly distributed between 1662~1790, 2222~2268, 2288~2316, 2390~2428, and 2476~2498 nm, which is exactly the range of the spectral absorption peaks generated by the first and second frequency doubling of the -C-H stretching vibrations of the -CH₂, -CH₃, and -CH-CH- functional groups of oil [31].

We note from Figure 10 that the moving window employed by the MC-UVE-SPA-MW algorithm expands the wavelength variables selected by the MC-UVE-SPA, resulting in a greater number of wavelength variables than that obtained by the MC-UVE-SPA, and the improved continuity of the wavelength variables selected by the MC-

FIGURE 5: Processing flow of the proposed moving-window-improved MC-UVE-SPA (MC-UVE-SPA-MW) wavelength selection algorithm based on the Monte Carlo unininformative variable elimination (MC-UVE) algorithm cascaded with the successive projections algorithm (SPA).

FIGURE 6: Wavelength variable stability distribution map of the partial least squares (PLS) regression model reflecting the oil concentration using the MC-UVE algorithm.

UVE-SPA-MW algorithm is very apparent in Figure 10 compared with the wavelength variables selected by the MC-UVE-SPA. We can also note from Table 3 that the full-spectrum model was relatively complicated, and its prediction accuracy was the worst of all models considered due to the impact of the large number of uninformative wavelength variables included within the model. In comparison, the models established with spectral data selected by the MC-UVE, MC-UVE-SPA, and MC-UVE-SPA-MW ($w=2L$, $2R$, 3) algorithms are all greatly simplified, and better model prediction accuracies are uniformly obtained. We also note from the table that, of the five wavelength selection

algorithms, the MC-UVE-SPA selected the least number of wavelengths and the MC-UVE-SPA-MW ($w=2L$) algorithm provided a model with the greatest prediction accuracy.

3.2. Diesel Spectral Data Experiments. The number of wavelength variables selected from the NIR spectral data of diesel fuel reflecting the boiling point by the MC-UVE, MC-UVE-SPA, and MC-UVE-SPA-MW ($w=3$, $2L$, $2R$) algorithms were, respectively, 262, 30, 83, 58, and 59, as shown in Table 4. These respectively represent 65.3%, 7.5%, 20.7%,

FIGURE 7: Relationship between the root mean square error (RMSE) of the corn oil concentration predicted by the PLS regression model and the number of selected wavelength variables.

FIGURE 8: Relationship between the RMSE of the corn oil concentration predicted by the multiple linear regression (MLR) model and the number of selected wavelength variables using the MC-UVE-SPA.

14.5%, and 14.7% of the 401 wavelength variables included in the full spectrum. The prediction results of the PLS or MLR models constructed from the selected wavelength variables are shown in Figure 11, and the details are listed in Table 4 along with the results obtained for a full-spectrum PLS model. We note from Table 4 and Figure 11 that the models established with spectral data selected by MC-UVE, MC-UVE-SPA, and MC-UVE-SPA-MW ($w = 2L, 2R, 3$) algorithms are greatly simplified compared with the full-spectrum model. MC-UVE retains 262 wavelength points, and the prediction accuracy is the worst of all the models considered, which may be due to the existence of wavelength collinearity. When SPA algorithm is used to further screen the wavelength points selected by MC-UVE, only 30 wavelength points are retained, while the prediction accuracy of the model is

greatly improved, RMSEP is reduced to 8.8676, r value is increased to 0.9341, and RPD value is increased to 2.4650. We note from Figure 12 that the moving window employed by the MC-UVE-SPA-MW expands the wavelength variables selected by the MC-UVE-SPA and improves the continuity of the wavelength variables selected by the MC-UVE-SPA-MW. When the window width $w = 2$ (Left), 2 (Right), and 3, the accuracy of the three models obtained by the MC-UVE-SPA-MW are all improved. When $w = 3$, the MC-UVE-SPA-MW expands 30 wavelength variables selected by the MC-UVE-SPA to 83. At this point, RMSEP is reduced to 5.9694, R value is increased to 0.9752, RPD value is increased to 3.9994, and the model is optimal. We can also note from Table 4 and Figure 11 that of the five wavelength selection algorithms, the MC-UVE-SPA selected the least number of wavelengths and the MC-UVE-SPA-MW ($w = 3$) algorithm provided a model with the greatest prediction accuracy.

3.3. Ethylene Gas Spectral Data Experiments. The number of wavelength variables selected from the spectral data reflecting the C_2H_4 concentration by the MVUVE, MC-UVE-SPA, and MC-UVE-SPA-MW ($w = 3, 2L, 2R$) algorithms were respectively 214, 17, 48, 34, and 34 as shown in Figure 13. These respectively represent 4.7%, 0.37%, 1.0%, 0.74%, and 0.74% of the 4601 wavelength variables included in the full spectrum. It can be determined from Figure 13 that greater than half of the selected wavelength variables fall within the strong absorption regions in the wavenumber ranges $794 \sim 1105 \text{ cm}^{-1}$ and $2917 \sim 3242 \text{ cm}^{-1}$. These results can be explained according to the description given on the HITRAN web page, which states that the absorption spectral band of C_2H_4 gas is in the range of $614 \sim 3242 \text{ cm}^{-1}$, and that the two isotopes $H_2^{12}C^{12}CH_2$ and $H_2^{12}C^{13}CH_2$ of C_2H_4 present strong absorption bands in the wavenumber ranges of $794 \sim 1105 \text{ cm}^{-1}$ and $2917 \sim 3242 \text{ cm}^{-1}$, respectively. From Figure 4, it can be seen that in some areas that are not C_2H_4 absorption bands, the spectral intensity has a significant linear relationship with C_2H_4 content, which may be due to the interference caused by the background spectrum with the change of C_2H_4 concentration, so in some areas that are not C_2H_4 absorption bands, the wavelength point is also selected.

The details regarding the prediction results of the PLS or MLR models constructed from the selected wavelength variables are listed in Table 5 along with the results obtained for a full-spectrum PLS model. We again note from the table that the full-spectrum model is more complicated, and its prediction accuracy was the worst of all models considered. In comparison, the models established with spectral data selected by the MC-UVE, MC-UVE-SPA, and MC-UVE-SPA-MW algorithms are all greatly simplified, and better model prediction accuracies are uniformly obtained. Of the five wavelength selection algorithms, we again note that the MC-UVE-SPA selected the least number of wavelengths and the MC-UVE-SPA-MW ($w = 3$) algorithm provided a model with the greatest prediction accuracy.

FIGURE 9: Comparison of corn oil concentration predictions from the prediction data set using the (a) MC-UVE-PLS, (b) MC-UVE-SPA, (c) MC-UVE-SPA-MW-MLR ($w = 2L$), (d) MC-UVE-SPA-MW-MLR ($w = 2R$), and (e) MC-UVE-SPA-MW-MLR ($w = 3$).

TABLE 3: Number of selected wavelengths and corn oil concentration prediction results.

Wavelength selection algorithm	Number of wavelengths	Model type	RMSECV	RMSEP	r	RPD
Full spectrum	700	PLS	0.1431	0.1345	0.9257	1.0550
MC-UVE	102	PLS	0.0289	0.1447	0.9024	0.7948
MC-UVE-SPA	37	MLR	0.0114	0.1225	0.9267	2.0225
MC-UVE-SPA-MW($w = 2L$)	64	MLR	0.0035	0.0381	0.9713	16.3666
MC-UVE-SPA-MW($w = 2R$)	64	MLR	0.0101	0.1009	0.8062	1.3703
MC-UVE-SPA-MW($w = 3$)	82	MLR	0.0089	0.0532	0.9335	6.1627

FIGURE 10: Wavelengths selected from the NIR spectral data reflecting corn oil concentration by the (a) MC-UVE, (b) MC-UVE-SPA, and (c) MC-UVE-SPA-MW algorithms.

3.4. Summary of the Experimental Results. It can be noted from the above experimental results that the prediction accuracy of the models established by the wavelength

selection algorithm are higher than that of the full-spectrum model. The MC-UVE-SPA selects the least characteristic wavelengths and eliminates the collinearity

TABLE 4: Number of selected wavelengths and boiling point prediction results.

Wavelength selection algorithm	Number of wavelengths	Model type	RMSECV	RMSEP	r	RPD
Full spectrum	401	PLS	4.6350	8.3841	0.9103	2.3704
MC-UVE	262	PLS	4.3184	10.0221	0.8311	1.7710
MC-UVE-SPA	30	MLR	3.5503	8.8676	0.9341	2.4650
MC-UVE-SPA-MW($w = 2L$)	58	MLR	3.9921	7.0442	0.9534	2.8212
MC-UVE-SPA-MW($w = 2R$)	59	MLR	4.0024	6.8573	0.9556	2.8982
MC-UVE-SPA-MW($w = 3$)	83	MLR	2.6478	5.9694	0.9752	3.9994

The maximum number of PCs was again set to 10 for both PLS models.

FIGURE 11: Comparison of diesel fuel boiling point predictions from the prediction data set using (a) the MC-UVE-PLS, (b) MC-UVE-SPA-MLR, (c) MC-UVE-SPA-MW-MLR ($w = 2L$), (d) MC-UVE-SPA-MW-MLR ($w = 2R$), and (e) MC-UVE-SPA-MW-MLR ($w = 3$).

FIGURE 12: Wavelengths selected from the NIR spectral data reflecting the boiling point of diesel fuel by the (a) MC-UVE, (b) MC-UVE-SPA, and (c) MC-UVE-SPA-MW algorithms.

between variables. The prediction accuracy of the model established by the MC-UVE-SPA is higher than that established by MC-UVE. The number of characteristic

wavelengths finally selected by the MC-UVE-SPA-MW is more than that of MC-UVE-SPA, but with better prediction accuracy.

FIGURE 13: Measured C_2H_4 gas absorbance spectra and wavelength variables selected by the MC-UVE, MC-UVE-SPA, and MC-UVE-SPA-MW algorithms.

TABLE 5: Number of selected wavelengths and C_2H_4 gas concentration prediction results.

Wavelength selection algorithm	Number of wavelengths	Model type	RMSECV	RMSEP	r
Full spectrum	4601	PLS	0.1431	0.1342	0.9801
MC-UVE	214	PLS	0.1009	0.1212	0.9887
MC-UVE-SPA	17	MLR	0.1375	0.0537	0.9913
MC-UVE-SPA-MW($w = 2L$)	34	MLR	0.0989	0.0495	0.9988
MC-UVE-SPA-MW($w = 2R$)	34	MLR	0.0988	0.0486	0.9990
MC-UVE-SPA-MW($w = 3$)	48	MLR	0.1005	0.0434	0.9999

The maximum number of PCs was again set to 10 for both PLS models.

4. Conclusions

The present study addressed the sparsity of wavelength variables selected by the cascaded MC-UVE-SPA through the application of a moving window, which improved the continuity of the selected wavelength variables, and thereby better exploited the advantages of the MC-UVE algorithm and the SPA to obtain regression models with high prediction accuracy. The advantages of the proposed MC-UVE-SPA-MW were demonstrated by applying the MC-UVE, MC-UVE-SPA, and MC-UVE-SPA-MW algorithms to the selection of wavelength variables from the NIR spectral absorbance data of corn, diesel fuel, and ethylene, and PLS and MLR models reflecting the oil content of corn, the boiling point of diesel fuel, and the ethylene concentration were thereby established and tested. The experimental results demonstrated that the progressive elimination of uncorrelated and collinear variables generated increasingly simplified partial-spectrum models with greater prediction accuracy than the full-spectrum model. Among the three wavelength selection algorithms, the MC-UVE-SPA selected the least number of wavelength variables and the proposed MC-UVE-SPA-MW algorithm provided models with the greatest prediction accuracy.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by grants from the Major National Science and Technology Special Project of China (JZ2015KJZZ0254) and the Key Projects of Natural Science Research in Universities in Anhui, China (KJ2018A0544).

References

- [1] C. Pasquini, "Near infrared spectroscopy: fundamentals, practical aspects and analytical applications," *Journal of the Brazilian Chemical Society*, vol. 14, no. 2, pp. 198–219, 2003.
- [2] X. Sun, M. Zhou, and Y. Sun, "Spectroscopy quantitative analysis cotton content of blend fabrics," *International Journal of Clothing Science and Technology*, vol. 28, no. 1, pp. 65–76, 2016.

- [3] M. Schwanninger, J. C. Rodrigues, and K. Fackler, "A review of band assignments in near infrared spectra of wood and wood Components," *Journal of Near Infrared Spectroscopy*, vol. 19, no. 5, pp. 287–308, 2011.
- [4] Y.-H. Yun, Y.-C. Wei, X.-B. Zhao, W.-J. Wu, Y.-Z. Liang, and H.-M. Lu, "A green method for the quantification of polysaccharides in *Dendrobium officinale*," *RSC Advances*, vol. 5, no. 127, pp. 105057–105065, 2015.
- [5] C. K. Vance, D. R. Tolleson, K. Kinoshita, J. Rodriguez, and W. J. Foley, "Near infrared spectroscopy in wildlife and biodiversity," *Journal of Near Infrared Spectroscopy*, vol. 24, no. 1, pp. 1–25, 2016.
- [6] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, 1933.
- [7] P. Geladi and B. R. Kowalski, "Partial least-squares regression: a tutorial," *Analytica Chimica Acta*, vol. 185, no. 1, pp. 1–17, 1986.
- [8] A. A. Kardamakis and N. Pasadakis, "Autoregressive modeling of near-IR spectra and MLR to predict RON values of gasolines," *Fuel*, vol. 89, no. 1, pp. 158–161, 2010.
- [9] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 3003.
- [10] B. Lu, N. Liu, H. Li et al., "Quantitative determination and characteristic wavelength selection of available nitrogen in coco-peat by NIR spectroscopy," *Soil and Tillage Research*, vol. 191, pp. 266–274, 2019.
- [11] M. J. Anzanello and F. S. Fogliatto, "A review of recent variable selection methods in industrial and chemometrics applications," *European Journal of Industrial Engineering*, vol. 8, no. 5, p. 619, 2014.
- [12] 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 in Analytical Chemistry*, vol. 113, pp. 102–115, 2019.
- [13] B. Nadler and R. R. Coifman, "The prediction error in CLS and PLS: the importance of feature selection prior to multivariate calibration," *Journal of Chemometrics*, vol. 19, no. 2, pp. 107–118, 2005.
- [14] T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø, "A review of variable selection methods in Partial Least Squares Regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 118, no. 8, pp. 62–69, 2012.
- [15] V. Centner, D.-L. Massart, O. E. de Noord, S. de Jong, B. M. Vandeginste, and C. Sterna, "Elimination of uninformative variables for multivariate calibration," *Analytical Chemistry*, vol. 68, no. 21, pp. 3851–3858, 1996.
- [16] G. Tang, Y. Huang, K. Tian et al., "A new spectral variable selection pattern using competitive adaptive reweighted sampling combined with successive projections algorithm," *The Analyst*, vol. 139, no. 19, pp. 4894–4902, 2014.
- [17] Y. Li, Y. Guo, C. Liu et al., "SPA combined with swarm intelligence optimization algorithms for wavelength variable selection to rapidly discriminate the adulteration of apple juice," *Food Analytical Methods*, vol. 10, no. 6, pp. 1965–1971, 2017.
- [18] J.-B. Li, C.-J. Zhao, W.-Q. Huang et al., "A combination algorithm for variable selection to determine soluble solid content and firmness of pears," *Analytical Methods*, vol. 6, no. 7, pp. 2170–2180, 2014.
- [19] Z. Xiaobo, Z. Jiewen, M. Hanpin, S. Jiyong, Y. Xiaopin, and L. Yanxiao, "Genetic algorithm interval partial least squares regression combined successive projections algorithm for variable selection in near-infrared quantitative analysis of pigment in cucumber leaves," *Applied Spectroscopy*, vol. 64, no. 7, pp. 786–794, 2010.
- [20] S. Ye, D. Wang, and S. Min, "Successive projections algorithm combined with uninformative variable elimination for spectral variable selection," *Chemometrics and Intelligent Laboratory Systems*, vol. 91, no. 2, pp. 194–199, 2008.
- [21] 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," *The Analyst*, vol. 140, no. 6, pp. 1876–1885, 2015.
- [22] W. Fan, Y.-Y. Li, Y.-K. Peng et al., "Nondestructive determination of lycopene content based on visible/near infrared transmission spectrum," *Chinese Journal of Analytical Chemistry*, vol. 46, no. 9, pp. 1424–1431, 2018.
- [23] Z. Sun, J. Fan, J. Wang et al., "Assessment of the human albumin in acid precipitation process using NIRS and multi-variable selection methods combined with SPA," *Journal of Molecular Structure*, vol. 1199, p. 126942, 2020.
- [24] H.-D. Li, Q.-S. Xu, and Y.-Z. Liang, "libPLS: an integrated library for partial least squares regression and linear discriminant analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 176, pp. 34–43, 2018.
- [25] R. Zhang, Y.-Y. Chen, Z.-B. Wang, and L. Kewu, "A novel ensemble L1 regularization based variable selection framework with an application in near infrared spectroscopy," *Chemometrics and Intelligent Laboratory Systems*, vol. 163, pp. 7–15, 2017.
- [26] W. Cai, Y. Li, and X. Shao, "A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra," *Chemometrics and Intelligent Laboratory Systems*, vol. 90, no. 2, pp. 188–194, 2008.
- [27] L. M. Brègman, "Finding the common point of convex sets by the method of successive projection," *Proceedings of the USSR Academy of Sciences*, vol. 162, no. 3, pp. 487–490, 1965.
- [28] X. Peng, T. Shi, A. Song, Y. Chen, and W. Gao, "Estimating soil organic carbon using VIS/NIR spectroscopy with SVMR and SPA methods," *Remote Sensing*, vol. 6, no. 4, pp. 2699–2717, 2014.
- [29] Y.-H. Liu, Q.-Q. Wang, X.-W. Gao, and A.-G. Xie, "Total phenolic content prediction in *Flos Lonicerae* using hyperspectral imaging combined with wavelengths selection methods," *Journal of Food Process Engineering*, vol. 42, no. 6, Article ID e13224, 2019.
- [30] J. Li, H. Zhang, B. Zhan, Y. Zhang, R. Li, and J. Li, "Non-destructive firmness measurement of the multiple cultivars of pears by Vis-NIR spectroscopy coupled with multivariate calibration analysis and MC-UVE-SPA method," *Infrared Physics & Technology*, vol. 104, Article ID 103154, 2020.
- [31] P. Hourant, V. Baeten, M. T. Morales, M. Meurens, and R. Aparicio, "Oil and fat classification by selected bands of near-infrared spectroscopy," *Applied Spectroscopy*, vol. 54, no. 8, pp. 1168–1174, 2000.