Chemometrics and Intelligent Laboratory Systems 163 (2017) 7–15

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

A novel ensemble L1 regularization based variable selection framework with an application in near infrared spectroscopy

Zhang Rui[a,b,c], Chen Yuanyuan[a,b,c,*], Wang Zhibin[a,b,c], Li Kewu[a,b,c]

[a] State Key Laboratory for Electronic Measurement Technology, North University of China, Taiyuan 030051, China
[b] Key Lab of Instrumentation Science & Dynamic Measurement, North University of China, Ministry of Education, Taiyuan 030051, China
[c] Engineering Technology Research Center of Shanxi Province for Opto-Electronic Information and Instrument, North University of China, Taiyuan 030051, China

ARTICLE INFO

Keywords:
variable selection
Uninformative variable elimination
Ensemble L1 regularization framework
Near infrared spectroscopy

ABSTRACT

Variable selection is an essential part during the whole process of qualitative and quantitative analysis of spectroscopy. Traditional methods like interval partial least square (iPLS), uninformative variable elimination (UVE), successive projections algorithm (SPA) etc. often have some disadvantages such as many parameters need to be tuned, weak robustness and so on. To solve these problems, this paper proposed a novel variable selection framework which combines UVE algorithm and ensemble L1 regularization framework together. The whole process of proposed method includes two phases: rough and fine selection. Firstly, UVE algorithm is used to eliminate the uninformative variables (rough selection). Secondly, the variable selection problem is mapped into a L1 regularization optimization problem with constraint (fine selection). To improve the stability and robustness of proposed method, an ensemble variable selection framework is designed which ensemble the results of many L1 regularization selectors. To validate the performance of proposed method, the following two public near infrared spectral datasets were tested: (1) Spectral (range from 900 nm to 1700 nm) and octane data of gasoline; (2) Spectral (range from 1100 nm to 2498 nm) and moisture data of corn. The experimental results showed that the proposed method can not only select the most featured wavelengths, but also can improve the stability and robustness of variable selection results.

1. Introduction

In recent years, infrared (IR) spectroscopy has gained wide acceptance in different fields by virtue of its advantages over other analytical techniques [1]. Variable or feature selection, also called “frequency” or “wavelength” selection when applied to spectroscopic data, is a critical step in data analysis, as it allows interactive improvement of the quality of data during the calibration procedure. Generally, more features mean more information, so an ideal model should perform better. However, the use of the full spectrum does not always yield optimal results because it usually includes regions without relevant information [2]. Indeed, there exists a trade-off between the number of selected variables and the generalization error. The main reasons of needing for variable selection before modeling are as follows. Firstly, the number of available samples is usually small with respect to the data dimensionality. Secondly, during the on-line monitoring, the hardware limitations lead to the fact that it is not possible to acquire more than a certain number of spectra in a given period of time. Thirdly, correct selection of variables in order to gather a small subset with a decreased sensitivity to non-linearity or discard those wavelengths most markedly contributing to it suffices in some cases. Additionally, variable selection may also help experts to understand which features are relevant in particular applications [3].

Variable selection algorithms can be classified into filters and wrappers. Filter methods select subset of features as a preprocessing step, independently of the inducing (learning) algorithm. Wrappers utilize the classifier (learning machine) performance to evaluate the goodness of feature subsets. Table 1 summarizes both the advantages and disadvantages of some classical previous variable selection methods/algorithms. The detailed introduction of these methods can be obtained from the corresponding reference papers.

From Table 1, it can be seen that those traditional variable selection algorithms either have many parameters need to be tuned (ANN, GA, interval selection method etc.), or the number of selected variable is still large (UVE, knowledge based selection etc.). More importantly, the robustness of above mentioned algorithms is often weak because **their performance often depends** on those parameters. Hence, to reduce the number of adjustable parameters, improve the robustness and stability of variable selection results, this paper proposed a novel variable selection **method** based on UVE algorithm and ensemble L1

* Corresponding author at: State Key Laboratory for Electronic Measurement Technology, North University of China, Taiyuan 030051, China.
E-mail address: chenyy@nuc.edu.cn (C. Yuanyuan).

http://dx.doi.org/10.1016/j.chemolab.2017.01.020
Received 18 September 2016; Received in revised form 25 January 2017; Accepted 27 January 2017
Available online 13 February 2017
0169-7439/ © 2017 Elsevier B.V. All rights reserved.

Table 1
Comparison of some previous variable selection methods/algorithms.

Algorithms	Advantages	Disadvantages	Reference
Knowledge based selection	Simple and easy use by the spectroscopist	Need for experience and good understanding	[1]
Successive Projections algorithm (SPA)	Minimize collinearity problems in multiple linear regression (MLR)	Selected variables have low signal-to-noise ratio (SNR)	[1,11–14]
Uninformative Variable Elimination (UVE)	Remove non-informative variables and can avoid model over-fitting	Large number of selected variables	[1,2,15–18]
Simulated Annealing (SA)	Able to traverse local minima and achieve the global minimum	Impossible to select two of the same variables in the same numerical string	[1,19–21]
Artificial Neural Network (ANN)	Allow more complex relationships between inputs and outputs	Being a "black box" heuristic tool and difficult to interpret	[1,22,23]
Genetic Algorithm (GA)	Probabilistic and non-local search process	Tend to be extremely slow and too many adjustable factors need to be configured	[1,24–26,30–33]
Interval selection method	Graphic output giving an overview of the interesting spectral areas	Starting wavelength, ending wavelength and wavelength interval etc. need to be optimized	[1,27–29,34,35]

regularization **framework**. Firstly, UVE algorithm was used to eliminate those uninformative variables. Secondly, considering the number of selected variables after UVE is still large, L1 regularization method was adopted to further select the most featured variables. Meanwhile, to improve the robustness and stability of variable selection results, the idea of ensemble learning which consists of many variable selectors was applied to solve this problem.

The following of this paper is organized as follows. Section 2 gives an introduction of the principle of UVE algorithm and ensemble L1 regularization based variable selection framework in detail. Section 3 assesses the performance of proposed algorithm based on two public near infrared spectral dataset (Corn dataset and Gasoline dataset). The discussion and conclusions are drawn in Section 4.

2. Algorithm Description

As shown in Fig. 1, the proposed ensemble L1 regularization based variable selection framework mainly includes the following four steps:

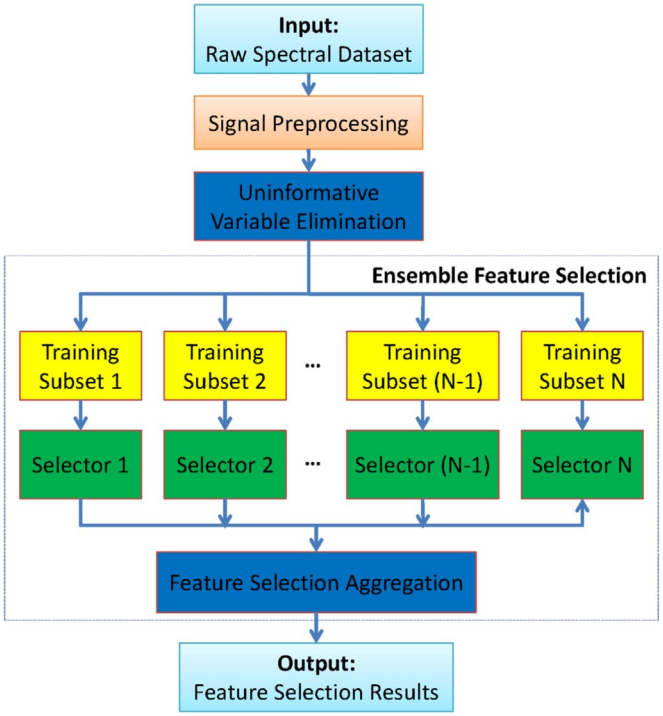

Fig. 1. Flowchart of ensemble L1 regularization based variable selection framework. (The whole procedure of proposed method can be divided into four steps: (1) Signal Preprocessing; (2) Rough selection with UVE algorithm; (3) Fine selection with ensemble L1 regularization method; (4) feature selection aggregation).

(1) **Signal preprocessing**. The raw near infrared spectral signals normally have some random noises and **exist** baseline drift phenomenon which will affect the qualitative and quantitative analysis. To remove this unrelated information, signal preprocessing methods such as smoothing, first and second order derivation and so on are often used.

(2) **Uninformative variable elimination**. According to the principle of near-infrared spectroscopy, there are only few featured wavelengths are high correlated with the target component while the contribution of other wavelengths is similar to random noises. Hence, based on this assumption, UVE algorithm was proposed to eliminate those uninformative variables.

(3) **Ensemble L1 regularization based variable selection. On one hand**, because the number of selected variables with UVE algorithm is often still large, **it is necessary to** further select the most featured variables. **On the other hand, considering the fact that the performance of traditional variable selection methods (such as iPLS, genetic algorithm, particle swarm optimization etc.) are often depend on the related parameters, which cause the above-mentioned methods are not robust enough. Hence, to find the most featured variables and** improve the robustness of variable selection results, an ensemble variable selection framework which includes many L1 regularization selectors is proposed. The selectors are independent of each other and for each selector, the variable selection problem is mapped into a sparse optimization problem and L1 regularization method is applied to find the best solution.

(4) **Variable selection aggregation**. As mentioned in step (3), each selector will give one variable selection result (the combination of most featured variables). Considering the fact that there are many parameters have strong impact on the variable selection results, which may result in weak stability and robustness. Hence, to improve the stability and robustness, some aggregation strategies such as voting, **weighted average** etc. are often used to formulate the final variable selection results.

Next, the principle of UVE algorithm and L1 regularization based variable selection method will be introduced in detail.

The basic idea of UVE is that the variables can be classified as informative or uninformative according to a stability parameter s_j. Variables with s_j values below a defined threshold level *cutoff* are considered as uninformative and removed from the original data set [2]. Artificial random variables are added to the data as a reference so that those variables which play a less important role in the model than the random variables are eliminated. The whole algorithm is described in Appendix A.

Without loss of generality, considering the general regression problem with J factors:

$$Y = \sum_{j=1}^{J} X_j \beta_j + \varepsilon \tag{1}$$

where Y is an $n \times 1$ vector, $\varepsilon \sim N_n(0, \sigma^2 I)$, X_j is an $n \times p_j$ matrix corresponding to the jth factor and β_j is a coefficient vector of size p_j, $j = 1, \ldots, J$. Denoting $X = (X_1, X_2, \cdots, X_J)$ and $\beta = (\beta_1', \cdots, \beta_J')'$, Eq. (1) can be written as

$$Y = X\beta + \varepsilon \tag{2}$$

The goal **of variable selection** is to select **the most** important factors for accurate estimation in Eq. (2). **In other words, those variables have large coefficient values generally indicate that they contain more information about the output variable while those variables with small coefficient values contain less information and can be ignored. Hence,** this amounts to deciding whether to set the vector β_j to zero vectors for each j. A commonly considered special case of Eq. (1) is when $p_1 = \cdots = p_J = 1$. This is the most studied model selection problem. **In the theory of L-norm regularization, L1 often indicates sparse representation or optimization.** Tibshirani [4] proposed the popular lasso, which is defined as

$$\hat{\beta}(\lambda) = \arg\min_{\beta}(\|Y - X\beta\|^2 + \lambda\|\beta\|_1) \tag{3}$$

where λ is a tuning parameter which controls the sparsity of β and $\|\bullet\|_1$ stands for the vector L1-norm. The L1-norm penalty induces sparsity in the solution.

To calculate the optimized solution, the group shooting lasso algorithm proposed by Yuan and Fu et al. [5,6] was adopted in this paper. The whole procedure of proposed framework is described in Appendix B.

3. Experimental results and discussion

To validate the performance of proposed ensemble L1 regularization variable selection framework, in this paper, we applied this algorithm to two public near infrared spectroscopy datasets: corn dataset and gasoline dataset [7]. (The dataset files in MATLAB.mat format are listed in Additional file 1 and 2).

3.1. Datasets description

3.1.1. Corn dataset

This dataset consists of 80 samples of corn measured on 3 different NIR spectrometers. The wavelength range is 1100–2498 nm at 2 nm intervals (700 channels). The moisture, oil, protein and starch values for each of the samples are also included.

3.1.2. Gasoline dataset

This dataset consists of 60 samples of gasoline measured by NIR spectrometers. The wavelength range is 900–1700 nm at 2 nm intervals (400 channels). The gasoline value for each of the samples is also included.

The raw NIR spectrums of Corn and Gasoline datasets were shown in Fig. 2 (The corresponding MATLAB code is listed in Additional file 3). It is easy to find that even the whole scanned wavelength range is large, only few featured wavelengths have strong correlation with the target (output) component. Hence, it is very necessary to select the most important wavelengths before the qualitative and quantitative analysis.

3.1.3. Experimental results

As illustrated above, the whole process of proposed variable selection framework can be divided into two phases: rough and fine selection. During the rough selection period, the UVE algorithm was adopted to eliminate the uninformative variables; during the fine selection period, the ensemble L1 regularization variable selection framework was applied to **find** the best combination of featured variables. In the following section, we will analyze the performance of proposed method during each phase, individually.

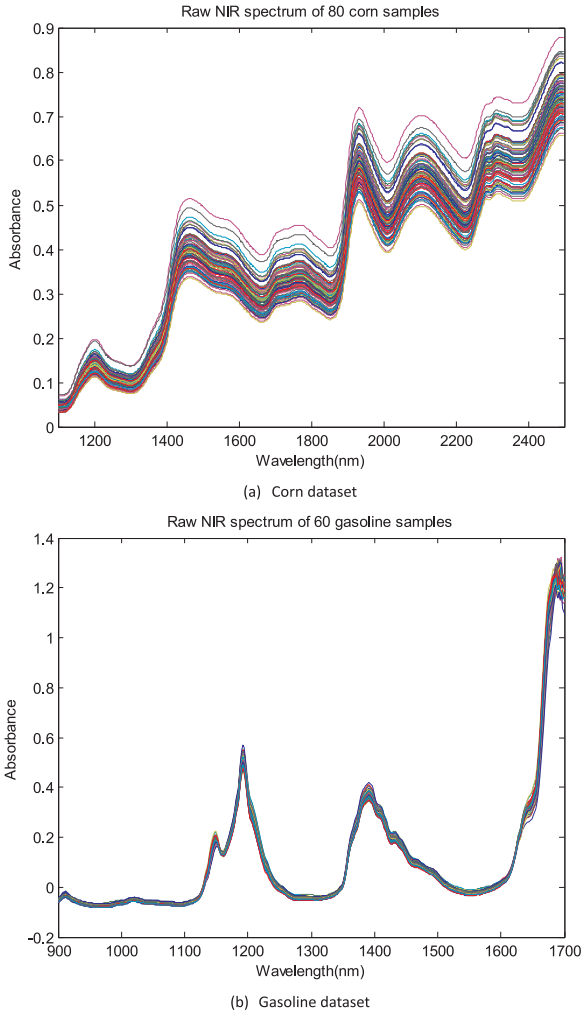

Fig. 2. Raw near infrared spectrum plot.

In this paper, partial least squares (PLS) method was took to establish the regression model and generate the regression coefficient for each wavelength. Fig. 3 (The corresponding MATLAB codes are listed in Additional file 4 and 5, in which UVE_PLS function contains the core code of UVE algorithm) describes the wavelength selection results of Corn and Gasoline dataset with UVE algorithm. The top panel shows the value of stability index s of each real variable (wavelength) and random variables. It is clearly that among the whole wavelength range, there exist many wavelengths whose corresponding s absolute values are smaller than the maximum s of random variables. The bottom panel shows the wavelength selection results by eliminate those uninformative variables. From Fig. 3**(a),** it can be seen that the selected wavelengths of Corn dataset mainly locate at the following sub ranges: [1364,1368], [1394,1456], [1502,1684], [1766,1814], [1880,1920], [1936,1998], [2020,2048], [2310,2320], [2342,2354], [2422,2446], [2474,2480] nm; from Fig. 3**(b),** it can be seen that the selected wavelengths of Gasoline dataset mainly locate at the following sub ranges: [1146,1158], [1194,1198], [1204,1250], [1350,1392], [1396,1406], [1520,1524] nm. The selected sub ranges are identical to the featured absorbance peaks, which indicate that the UVE algorithm can effectively eliminate those uninformative variables and suitable for rough selection. However, similar to the previous works, we can find that after rough uninformative variable elimination, 251 out of 700 wavelengths of Corn dataset and 80 out of 400 wavelength of

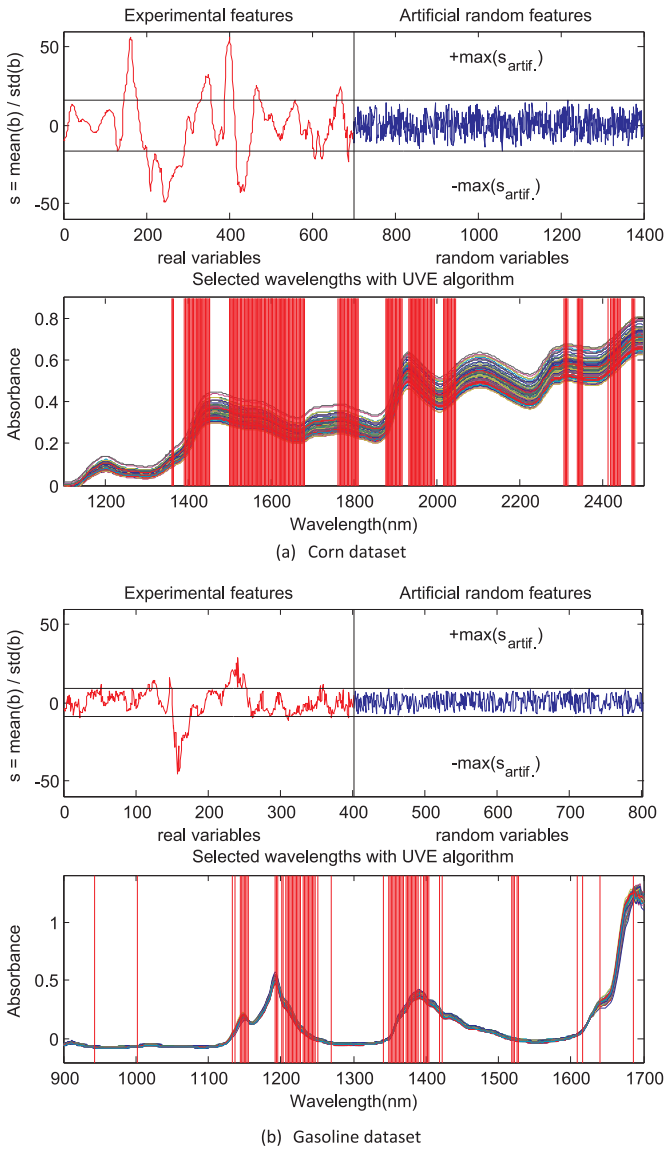

Fig. 3. variable selection results with UVE algorithm. (**Top**: the stability index s of real variables (red line) and random variables (blue line); **bottom**: selected wavelengths combination). (a) Corn dataset. (b) Gasoline dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

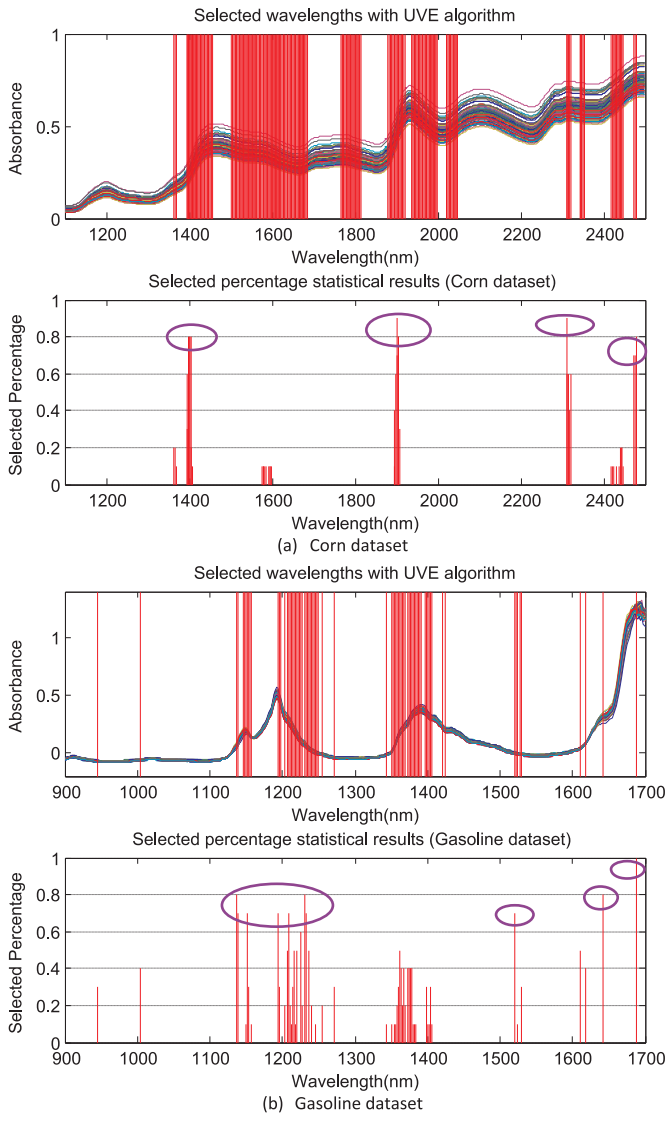

Fig. 4. Wavelengths selection results with ensemble L1 regularization method. (**Top:** selected wavelengths combination (same as the bottom panel in Fig. 3); bottom: selected percentage statistical results (blue eclipse denote the selected percentage above 60%)). (a) Corn dataset. (b) Gasoline dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Gasoline dataset were kept, respectively. The number of selected variables by using UVE algorithm is often still so large that it is not suitable for the further qualitative and quantitative analysis.

Further, during the fine variable selection phase, ensemble L1 regularization method was applied to find the best combination of featured wavelengths. There are many selectors in the ensemble framework according to Fig. 1, here we set the population size of selectors to 10. For Corn dataset, each selector was built based on randomly generated 60 out of 80 samples; for Gasoline dataset, each selector was built based on randomly generated 40 out of 60 samples. According to the sampling theory, we can guarantee that the selectors are independent between each other due to the random mechanism.

The wavelength selection results were illustrated in Fig. 4 (The corresponding MATLAB codes are listed in Additional file 6, 7 and 8, where vanilla_glasso.zip is a group lasso toolbox). In order to help the readers' observation and understanding, we redraw the bottom panel of Fig. 3 as the top panel of Fig. 4. The bottom panel of Fig. 4 describes the selected percentage statistical results of each wavelength. The blue eclipses denote those wavelengths whose selected percentage is above 60%, in other words, among overall 10 selectors, more than 6 select those wavelengths. By selecting the wavelengths with high percentage, the stability and robustness can be guaranteed.

The comparison of selected wavelengths after UVE algorithm and ensemble L1 variable selection framework is listed in Table 2. It can be found that after ensemble L1 regularization variable selection, the number of selected wavelengths reduced a lot. For both Corn and Gasoline dataset, only about 10 most featured wavelengths were selected. It indicates that the proposed variable selection framework which contains two phases (rough and fine selection) is effective for the high dimensional variable selection problem.

3.2. Discussion

3.2.1. Effects of sparsity on variable selection

From Eq. (5), we can conclude that the tuning parameter λ controls the sparsity of the final variable selection results β, with λ increase, β becomes sparser. However, if the value of λ is too large, β may become too sparse to keep the most featured variables. Hence, here we try to investigate how λ affect the wavelength selection results for both Corn and Gasoline dataset.

Table 2
The comparison of selected wavelengths for both Corn and Gasoline dataset after UVE algorithm and ensemble L1 regularization variable selection framework.

Phase	Selected wavelengths (nm)		Number of selected wavelengths	
	Corn	Gasoline	Corn	Gasoline
Rough selection (After UVE algorithm)	[1364,1368], [1394,1456], [1502,1684], [1766,1814], [1880,1920], [1936,1998], [2020,2048], [2310,2320], [2342,2354], 2418, [2422,2446], [2474,2480]	946, 1004, 1136, 1138, [1146,1158], [1194,1198], [1204,1250], 1254, 1272, 1344, [1350,1392], [1396,1406], 1420, 1424, [1520,1524], 1528, 1530, 1610, 1618, 1642, 1688	251	80
Fine selection (After ensemble L1 regularization variable selection framework)	[1398,1404], [1900,1906], 2310, 2474, 2476, 2480	1136, 1138, 1152, 1194, 1208, 1230, 1232, 1520, 1642, 1688	12	10

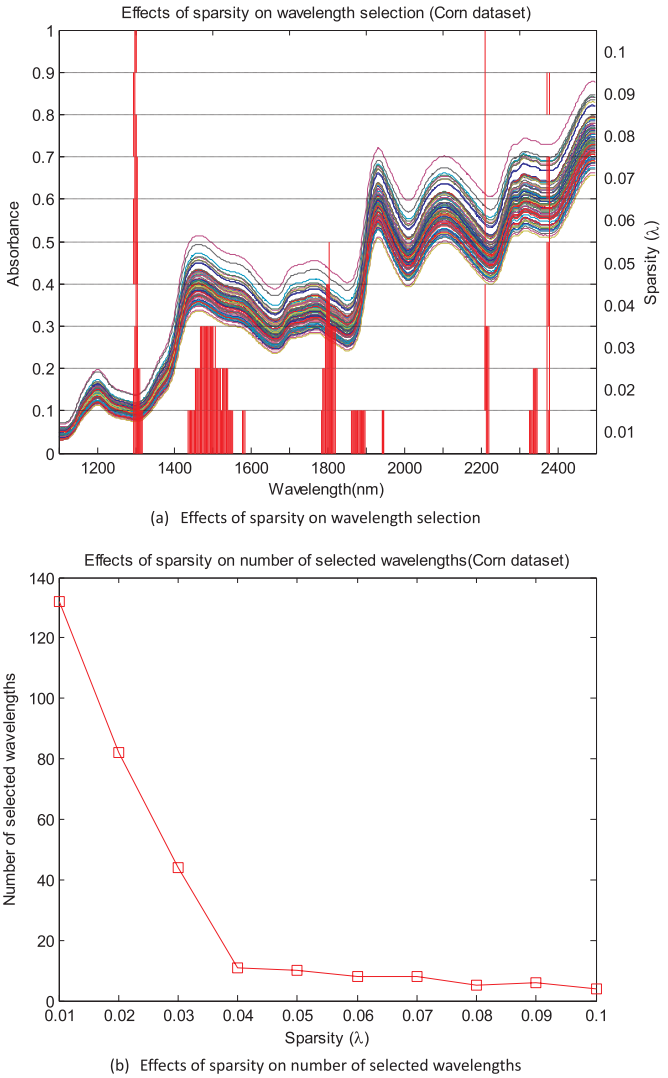

Fig. 5. Effects of sparsity on variable selection (Corn dataset). (a) Effects of sparsity on wavelength selection. (b) Effects of sparsity on number of selected wavelengths.

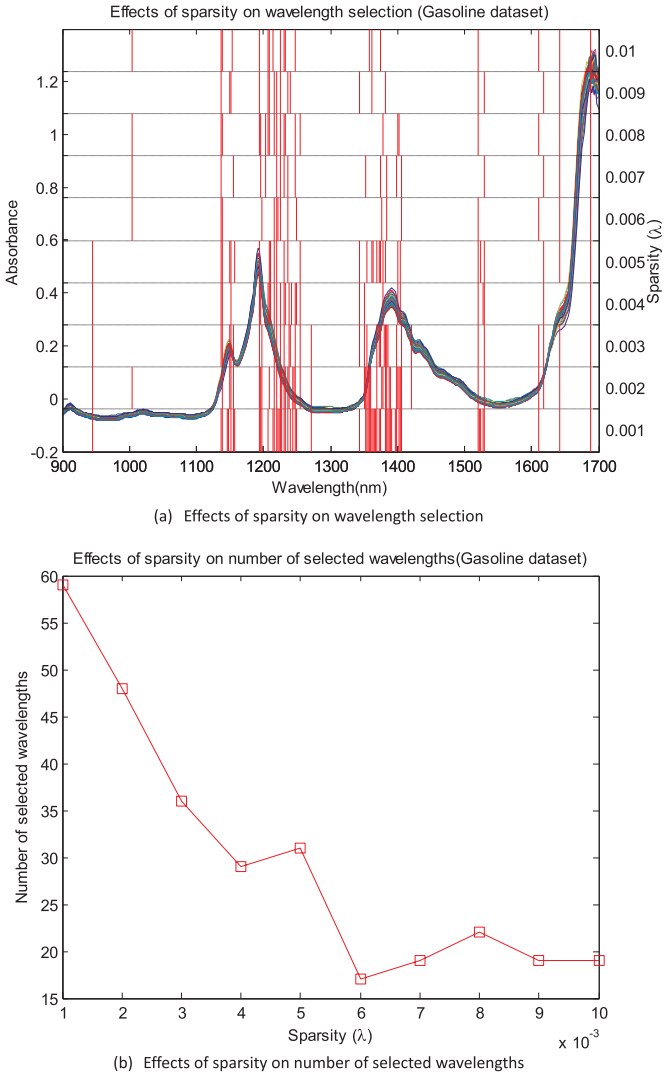

Fig. 6. Effects of sparsity on variable selection (Gasoline dataset). (a) Effects of sparsity on wavelength selection. (b) Effects of sparsity on number of selected wavelengths.

The effects of sparsity tuning parameter λ on wavelength selection results of Corn dataset are shown in Fig. 5 (The corresponding MATLAB codes are listed in Additional file 9 and 10). From Fig. 5 **(b)**, it can be seen that with the change of λ from 0.01 to 0.1, the number of selected wavelengths decrease from about 130 to 10. Meanwhile, from Fig. 5 **(a)**, we can find that while $\lambda > 0.03$, the wavelength selection results will ignore some featured wavelengths locate at 1500, 1800 nm nearby.

Similarly, the effects of sparsity tuning parameter λ on wavelength selection results of Gasoline dataset are shown in Fig. 6. From Fig. 6 **(b)** we will see the same trend as Corn dataset, which is with the change of λ from 0.001 to 0.01, the number of selected wavelengths decrease from about 60 to 20. Meanwhile, compared with Corn dataset, the wavelength selection results of Gasoline dataset is better through the whole range of λ without lose any featured wavelength.

Consequently, in other real applications we should set the sparsity parameter λ to a suitable value so that on one hand the number of selected variables is not too large, on the other hand the selected variables will not ignore any most featured variables.

3.2.2. *Effects of variable selection on quantitative analysis*

Another question we need **to** concern **is that** how the variable selection results will affect the further qualitative and quantitative analysis. In this paper, we compared three quantitative regression model which were established based on the whole range spectrum,

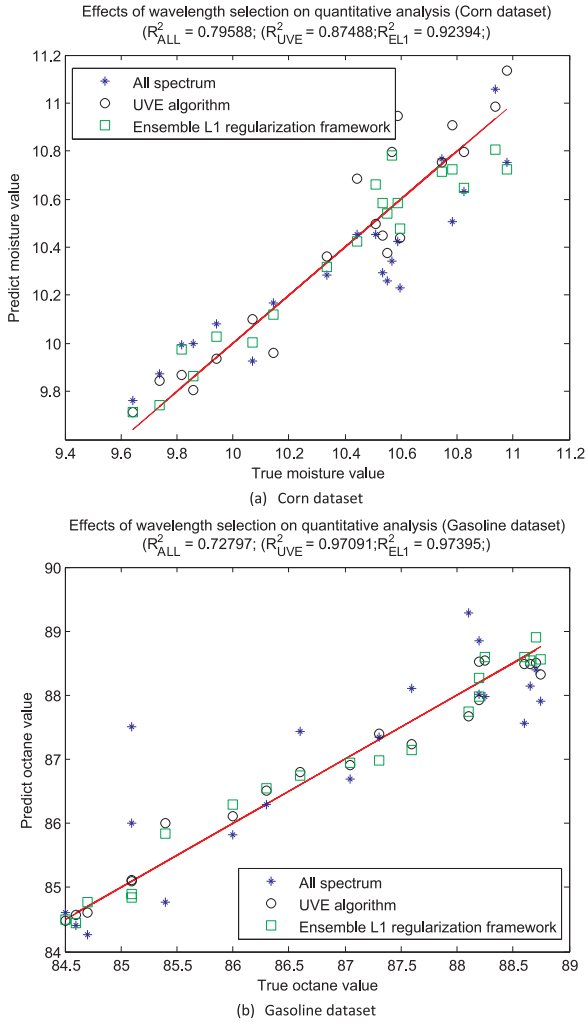

Fig. 7. Effects of variable selection on quantitative analysis. (a) Corn dataset. (b) Gasoline dataset.

selected wavelengths with UVE algorithm and proposed ensemble L1 regularization variable selection framework, respectively. The moisture and octane value were chosen as the output target for Corn and Gasoline dataset, respectively. The regression models were built with extreme learning machine (ELM) algorithm which was proposed by Huang et al. [8,9]. Compared with traditional backpropagation neural networks, ELM does not have to iterative adjusting the connecting weights and bias, it maps the training process to a problem of solving a group of linear equations. Besides, Huang et al. have proved that if the activation function g: $R \rightarrow R$ is infinitely differentiable in any interval, the connected weight between neurons of input and hidden layer and bias of neuron in hidden layer can be chosen randomly.

In this paper, to improve the generalized performance of the quantitative regression models, the 10-fold cross validation method was implemented to find the best number of hidden neurons. Fig. 7 **(a)** and **(b)** (The corresponding MATLAB codes are listed in Additional file 11, 12 and 13, where elm.zip is ELM algorithm toolbox) illustrate the effects of variable selection on quantitative analysis of Corn and Gasoline dataset, respectively. It is clearly to find that the generalized performance of models established with selected wavelengths is better than with whole range spectrum (R^2 denotes for determined coefficient, which is more close to 1 means the corresponding model has better generalized performance). Additionally, although the number of selected wavelengths by proposed method is much smaller than UVE algorithm, the generalized performance of corresponding models is closely. In other words, by using the proposed method, the complexity of further qualitative and quantitative model can reduce a lot while the generalized performance is not influenced.

3.3. Comparisons of variable selection stability between different methods

As mentioned above, the purpose of this paper is not only to overcome the disadvantage of traditional UVE algorithm (the number of selected variables is often still large), but also to improve the stability and robustness of variable selection results. Hence, we want to check whether the proposed method has better stability than traditional methods. Based on the previous works, the Jaccard coefficient [10] was selected to evaluate the stability and robustness, which is defined as:

$$S_{total} = \frac{2\sum_{i=1}^{k}\sum_{j=i+1}^{k} S(FS_i, FS_j)}{k(k-1)} \tag{4}$$

$$S(FS_i, FS_j) = \frac{|FS_i \cap FS_j|}{|FS_i \cup FS_j|} = \frac{\sum_l I(FS_i^l = FS_j^l = 1)}{\sum_l I(FS_i^l + FS_{jf}^l > 0)} \tag{5}$$

where FS_i is the variable selection results of the ith trial, which is a binary vector ("1" or "0" denotes the corresponding variable is selected or not); k denotes the number of trials; the value of $S(FS_i, FS_j)$ is between 0 and 1, $S(FS_i, FS_j) = 0$ and $S(FS_i, FS_j) = 1$ mean that FS_i and FS_j are totally different and same, respectively. Hence, if the value of S_{total} is more closely to 1, the corresponding method is more stable and robust.

Here, we compared the proposed method with genetic algorithm, UVE algorithm, particle swarm optimization algorithm and binary bat algorithm. Each method was running for 50 times, the comparison results are listed in Table 3. It is clearly that among the three methods, the proposed method has the highest value of S_{total} while genetic algorithm has the lowest value of S_{total}, which indicate that the proposed method has better stability and robustness than traditional variable selection methods. To find whether there **exists** statistical difference between the proposed method and other four methods, here the Student's test approach was used. The p-value was 0.0382, 0.0316, 0.0264 and 0.0225, respectively. It means that there is a statistical significant difference between the proposed method and genetic algorithm and UVE algorithm.

4. Conclusions

This paper proposed a novel variable selection framework which combines traditional UVE algorithm and L1-**norm** regularization method together. The experimental results of two public near infrared spectrum dataset (Corn and Gasoline) indicated that by using the proposed method, on one hand the most featured wavelengths can be effectively selected, on the other hand the stability and robustness of variable selection results can be improved. Meanwhile, the effects of variable selection on further quantitative analysis proved that although the generalized performance has no obvious difference between the proposed method and UVE algorithm, the complexity of regression model is much simpler because UVE algorithm is a rough selection

Table 3
Comparisons of variable selection stability between different methods.

Method	Jaccard coefficient (S_{total})	
	Corn Dataset	Gasoline Dataset
Proposed method	0.946	0.925
Genetic algorithm	0.748	0.727
UVE algorithm	0.835	0.862
PSO algorithm	0.814	0.836
Binary bat algorithm	0.868	0.882

which only eliminate those uninformative variables while the proposed method is two phases (rough and fine) selection which can find the most featured variables. Additionally, the effects of sparsity on variable selection showed that in the real applications, it is necessary to set the sparsity tuning parameter to a suitable value so that it will not only reduce the number of selected variables, but also will not ignore any featured variable. In the future, we will take more efforts to find more commonly guidance for the adaptive setting of sparsity tuning parameter based on more real application datasets.

Conflict of interests

The authors declare that they have no proprietary, financial, professional, or other personal competing interests of any nature or kind.

Acknowledgements

The authors would like to thank the National Natural Science Foundation of China (NSFC, Grant No. 61605176) and the Key Program for International S & T Cooperation Projects of China (Grant No. 2012DFA10680 & 2013DFR10150) for the financial support. And Prof. Chen Yuanyuan thanks to the many researchers whom have offered the previous remarkable works in this field.

Appendix A

In linear models, the prediction $\hat{y}$ is computed with Eq. (A-1).

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_px_p + e \tag{A-1}$$

where each coefficient b_j represents the contribution of the corresponding variable to the established model, the stability of each variable j can be quantitatively measured by the stability defined as:

$$s_j = \frac{mean\,(b_j)}{std\,(b_j)}, j = 1, \ldots, p \tag{A-2}$$

where $mean\,(b_j)$ and $std\,(b_j)$ are the mean and standard deviation of the regression coefficients of variable j. It is clear that, when the mean value of b_j is large and the standard deviation of b_j is small, the stability value is large, which means the corresponding variable is important. The variables whose stability is less than a threshold should be treated as uninformative and be eliminated.

In order to estimate a suitable threshold level *cutoff*, an artificial random variable matrix $N(n \times p)$ with very small amplitude (e.g., 10^{-11}) is added to the original data to compute their stability.

There are several options to establish a limit value for the stability of a variable to be considered as uninformative. The cutoff threshold value has a strong impact on the wavelength selection results. With respect to Corn and Gasoline dataset used in this paper, the effects of cutoff threshold value on wavelength selection results are illustrated in Figs. A1–A3, respectively.

It is clearly to find that, if the cutoff threshold value is set too high, the number of selected wavelength with UVE algorithm is reduced so rapidly that some important wavelength features is also eliminated. Hence in this paper, considering the fact that after the rough selection, we will have a fine selection procedure, hence here the maximum and minimum s_j value of the added random variables is used as the threshold level.

Appendix B

Step 1: centering the input X and output Y through Eq. (B-1):

$$\begin{cases} \mathbf{X} \leftarrow \mathbf{X} - \overline{\mathbf{X}} \\ \mathbf{Y} \leftarrow \mathbf{Y} - \overline{\mathbf{Y}} \end{cases} \tag{B-1}$$

where $\overline{\mathbf{X}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$, $\overline{\mathbf{Y}} = \frac{1}{N}\sum_{i=1}^{N}y_i$.

Step 2: Gram-Schmidt orthonormalization of input X through Eq. (B-2):

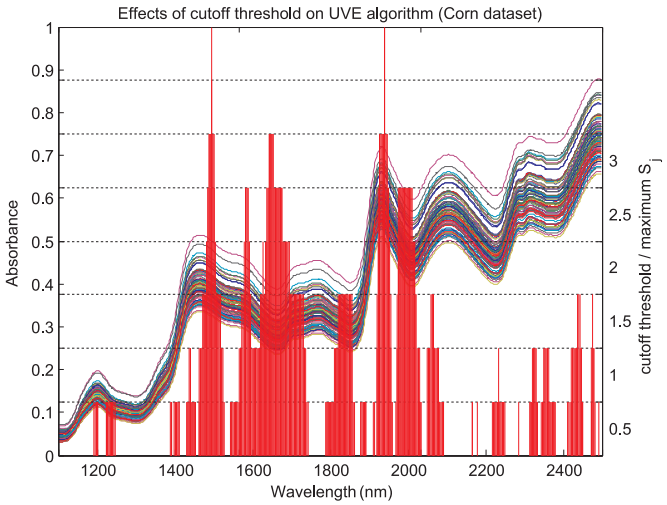

Fig. A1. Effects of cutoff threshold on UVE algorithm (Corn dataset).

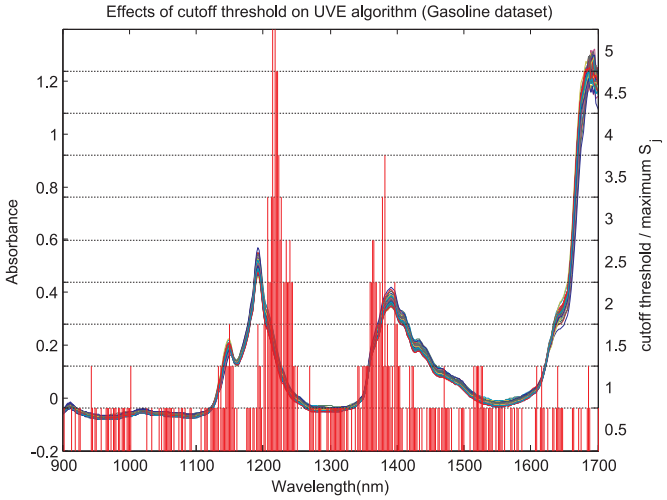

Fig. A2. Effects of cutoff threshold on UVE algorithm (Gasoline dataset).

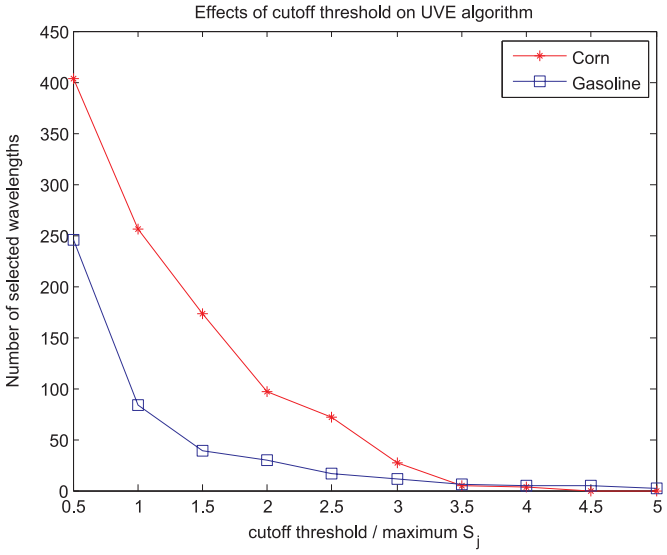

Fig. A3. Effects of cutoff threshold on the number of selected wavelengths.

$$\begin{cases} \tilde{\mathbf{x}}_1 = \mathbf{x}_1, & \boldsymbol{\eta}_1 = \frac{\tilde{\mathbf{x}}_1}{\|\tilde{\mathbf{x}}_1\|} \\ \tilde{\mathbf{x}}_j = \mathbf{x}_j - \sum_{i=1}^{j-1} \langle \mathbf{x}_j, \boldsymbol{\eta}_i \rangle \boldsymbol{\eta}_i, & \boldsymbol{\eta}_j = \frac{\tilde{\mathbf{x}}_j}{\|\tilde{\mathbf{x}}_j\|}, \ j = 2, \cdots, P \end{cases} \tag{B-2}$$

Step 3: solve Eq. (3) with ridge regression algorithm, and set the solution as the initial value of β as shown in Eq. (B-3):

$$\boldsymbol{\beta}_0 = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} \tag{B-3}$$

Step 4: for each block (in this paper, block size is equal to 1), modify the value of β through Eq. (B-4):

$$\boldsymbol{\beta}_j = \begin{cases} \left(1 - \frac{\lambda\sqrt{P_j}}{\|\mathbf{S}_j\|}\right)\mathbf{S}_j, & \left(1 - \frac{\lambda\sqrt{P_j}}{\|\mathbf{S}_j\|}\right) \geq 0 \quad j = 1, 2, \cdots, J \\ \mathbf{0}, & \text{others} \end{cases} \tag{B-4}$$

Step 5: iterative modify the value of β through **Step 4** until the stop condition (such as reach the maximum iterations, meet the iterative error limit etc.) is satisfied.

Step 6: give out the final sparse solution $\hat{\beta}$.

Appendix C. Supplementary material

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.chemolab.2017.01.020.

References

[1] Z. Xiaobo, Z. Jiewen, M.J. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy, Anal. Chim. Acta 667 (2010) 14–32.

[2] J.K. Javier Moros, Guillermo Quintas, Salvador Garrigues, New cut-off criterion for uninformative variable elimination in multivariate calibration of near-infrared spectra for the determination of heroin in illicit street drugs, Anal. Chim. Acta 6 (2008) 150–160.

[3] Mv.H. Frenay Benoit, Yoan Miche, Michel Verleysen, Amaury Lendasse Feature selection for nonlinear models with extreme learning machines, Neurocomputing 102 (2013) 111–124.

[4] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B (Methodol.) (1996) 267–288.

[5] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J. R. Stat. Soc.: Ser. B (Stat. Methodol.) 68 (2006) 49–67.

[6] W.J. Fu, Penalized regressions: the bridge versus the lasso, J. Comput. Graph. Stat. 7 (1998) 397–416.

[7] J.H. Kalivas, Two data sets of near infrared spectra, Chemom. Intell. Lab. 37 (1997) 255–259.

[8] G.-B. Huang, Q.-Y. Zhu, C.-K. Siew, Extreme learning machine: theory and applications, Neurocomputing 70 (2006) 489–501.

[9] G.-B. Huang, D.H. Wang, Y. Lan, Extreme learning machines: a survey, Int. J. Mach. Learn. Cybern. 2 (2011) 107–122.

[10] Y. Saeys, T. Abeel, Y. Van de Peer, Robust Feature Selection Using Ensemble Feature Selection Techniques, in: W. Daelemans, B. Goethals, K. Morik (Eds.) Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part II, Springer Berlin Heidelberg, Berlin, Heidelbergpp2008, pp. 313–325.

[11] M.C.U. Araujo, T.C.B. Saldanha, R.K.H. Galvao, T. Yoneyama, H.C. Chame, V. Visani, The successive projections algorithm for variable selection in spectroscopic multicomponent analysis, Chemom. Intell. Lab. 57 (2001) 65–73.

[12] F. Liu, Y. Jiang, Y. He, Variable selection in visible/near infrared spectra for linear and nonlinear calibrations: a case study to determine soluble solids content of beer, Anal. Chim. Acta 635 (2009) 45–52.

[13] M.J.C. Pontes, J. Cortez, R.K.H. Galvao, C. Pasquini, M.C.U. Araujo, R.M. Coelho, M.K. Chiba, M.F. de Abreu, B.E. Madari, Classification of Brazilian soils by using LIBS and variable selection in the wavelet domain, Anal. Chim. Acta 642 (2009) 12–18.

[14] A.G. Ouyang, J. Liu, Classification and determination of alcohol in gasoline using NIR spectroscopy and the successive projections algorithm for variable selection, Meas. Sci. Technol. 24 (2013).

[15] V. Centner, D.L. Massart, O.E. deNoord, S. deJong, B.M. Vandeginste, C. Sterna, Elimination of uninformative variables for multivariate calibration, Anal. Chem. 68 (1996) 3851–3858.

[16] C. Abrahamsson, J. Johansson, A. Sparen, F. Lindgren, Comparison of different variable selection methods conducted on NIR transmission measurements on intact tablets, Chemom. Intell. Lab. 69 (2003) 3–12.

[17] W.S. Cai, Y.K. Li, X.G. Shao, A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra, Chemom. Intell. Lab. 90 (2008) 188–194.

[18] S.F. Ye, D. Wang, S.G. Min, Successive projections algorithm combined with uninformative variable elimination for spectral variable selection, Chemom. Intell. Lab. 91 (2008) 194–199.

[19] J.H. Kalivas, N. Roberts, J.M. Sutter, Global Optimization by Simulated Annealing with Wavelength Selection for Ultraviolet Visible Spectrophotometry, Anal. Chem. 61 (1989) 2024–2030.

[20] H. Swierenga, P.J. de Groot, A.P. de Weijer, M.W.J. Derksen, L.M.C. Buydens, Improvement of PLS model transferability by robust wavelength selection, Chemom. Intell. Lab. 41 (1998) 237–248.

[21] H. Swierenga, F. Wulfert, O.E. de Noord, A.P. de Weijer, A.K. Smilde, L.M.C. Buydens, Development of robust calibration models in near infra-red spectrometric applications, Anal. Chim. Acta 411 (2000) 121–135.

[22] R. Todeschini, D. Galvagni, J.L. Vilchez, M. del Olmo, N. Navas, Kohonen artificial neural networks as a tool for wavelength selection in multicomponent spectrofluorimetric PLS modelling: application to phenol, o-cresol, m-cresol and p-cresol mixtures, Trac-Trend Anal. Chem. 18 (1999) 93–98.

[23] Z. Boger, Selection of quasi-optimal inputs in chemometrics modeling by artificial neural network analysis, Anal. Chim. Acta 490 (2003) 31–40.

[24] M.H. Givianrad, M. Saber-Tehrani, S. Zarin, Genetic algorithm-based wavelength selection in multicomponent spectrophotometric determinations by partial least square regression: application to a sulfamethoxazole and trimethoprim mixture in bovine milk, J Serb. Chem. Soc. 78 (2013) 555–564.

[25] R. Leardi, M.B. Seasholtz, R.J. Pell, Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data, Anal. Chim. Acta 461 (2002) 189–200.

[26] A. Durand, O. Devos, C. Ruckebusch, J.P. Huvenne, Genetic algorithm optimisation combined with partial least squares regression and mutual information variable selection procedures in near-infrared quantitative analysis of cotton-viscose textiles, Anal. Chim. Acta 595 (2007) 72–79.

[27] L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy, Appl. Spectrosc. 54 (2000) 413–419.

[28] B. Hemmateenejad, M. Akhond, F. Samari, A comparative study between PCR and PLS in simultaneous spectrophotometric determination of diphenylamine, aniline, and phenol: effect of wavelength selection, Spectrochim. Acta A 67 (2007) 958–965.

[29] A.F.C. Pereira, M.J.C. Pontes, F.F. Gambarra, S.R.B. Santos, R.K.H. Galvao, M.C.U. Araujo, NIR spectrometric determination of quality parameters in vegetable oils using iPLS and variable selection, Food Res. Int. 41 (2008) 341–348.

[30] J.H. Cheng, D.W. Sun, H. Pu, Combining the genetic algorithm and successive projection algorithm for the selection of feature wavelengths to evaluate exudative characteristics in frozen–thawed fish muscle, Food Chem. 197 (2016) 855–863.

[31] X.D. Sun, M.X. Zhou, Y.Z. Sun, Variables selection for quantitative determination of cotton content in textile blends by near infrared spectroscopy, Infrared Phys. Technol. 77 (2016) 65–72.

[32] K.A. Attia, M.W. Nassar, M.B. El-Zeiny, A. Serag, Firefly algorithm versus genetic algorithm as powerful variable selection tools and their effect on different multivariate calibration models in spectroscopy: a comparative study, Spectrochim. Acta Part A: Mol. Biomol. Spectrosc. 170 (2017) 117–123.

[33] Z. Li, Wavelength Selection for Quantitative Analysis in Terahertz Spectroscopy Using a Genetic Algorithm, IEEE Trans. Terahertz Sci. Technol. 6 (2016) 658–663.

[34] X. Li, C. Sun, L. Luo, Y. He, Determination of tea polyphenols content by infrared spectroscopy coupled with iPLS and random frog techniques, Comput. Electron. Agric. 112 (2015) 28–35.

[35] H. Cao, X. Yan, S.S. Ge, H. Ren, Variable selection based on information tree for spectroscopy quantitative analysis, Anal. Methods 7 (2015) 6612–6618.