Analytica Chimica Acta 948 (2016) 19–29

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

A novel algorithm for spectral interval combination optimization

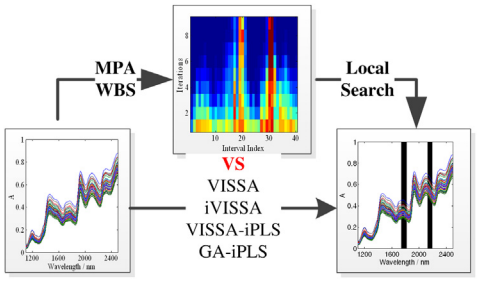
Xiangzhong Song [a], Yue Huang [a, b], Hong Yan [a], Yanmei Xiong [a, *], Shungeng Min [a, **]

[a] College of Science, China Agricultural University, Beijing, 100193, PR China
[b] Third Class Tobacco Supervision Station, Beijing, 101121, PR China

HIGHLIGHTS

- A new wavelength interval combination optimization algorithm was proposed based on model popular analysis strategy.
- The combination of spectral intervals can be optimized in a soft shrinkage manner.
- Its computational intensity is economic benefit from fewer tune parameters and faster convergence speed.
- WBS was proved to be a more efficient sampling method than WBMS especially for implementing MPA strategy.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 25 August 2016
Received in revised form
26 October 2016
Accepted 28 October 2016
Available online 2 November 2016

Keywords:
Wavelength selection
Interval combination optimization (ICO)
Model population analysis (MPA)
Weighted bootstrap sampling (WBS)
Weighted binary matrix sampling (WBMS)

ABSTRACT

In this study, a new wavelength interval selection algorithm named as interval combination optimization (ICO) was proposed under the framework of model population analysis (MPA). In this method, the full spectra are divided into a fixed number of equal-width intervals firstly. Then the optimal interval combination is searched iteratively under the guide of MPA in a soft shrinkage manner, among which weighted bootstrap sampling (WBS) is employed as random sampling method. Finally, local search is conducted to optimize the widths of selected intervals. Three NIR datasets were used to validate the performance of ICO algorithm. Results show that ICO can select fewer wavelengths with better prediction performance when compared with other four wavelength selection methods, including VISSA, VISSA-iPLS, iVISSA and GA-iPLS. In addition, the computational intensity of ICO is also economical, benefit from fewer tune parameters and faster convergence speed.

1. Introduction

Spectroscopic datasets collected by high throughput instruments are usually faced with the non-deterministic polynomial time (NP)-hard problem. This kind of datasets usually consists of large number of variables and relatively few samples due to the constraint of actual experimental conditions and costs. Multivariate calibration techniques such as principal component regression (PCR) and partial least squares regression (PLS) are usually employed to address this problem by extracting latent information from spectroscopic dataset. However, more and more researches have proved that variable selection is still beneficial for these multivariate calibration techniques from both experimental and theoretical aspects [1–5]. The benefits of variable selection can be

* Corresponding author.
** Corresponding author.
E-mail addresses: xiongym@cau.edu.cn (Y. Xiong), minsg@263.net (S. Min).

http://dx.doi.org/10.1016/j.aca.2016.10.041
0003-2670/

summarized in four main aspects: (1) the prediction ability of calibration model can usually be improved by eliminating uninformative or interfering variables; (2) new calibration model based on informative variables will be easier to interpret; (3) the computational speed of new model will be boosted; (4) low cost of dedicated online or inline analytical instrument with less spectral channels may be produced under the guide of variable selection [6].

In essence, variable selection is aimed to find an optimal combination of variables for the best prediction performance. However, as the number of variable combinations grows exponentially along with the increase of variables, the rough search is always impractical. Thus, a large number of variable selection methods have been proposed based on different strategies in the past decades, such as stepwise strategy, e.g. forward selection and backward elimination [7]; variable ranking strategy based on parameters of PLS model [8–10], e.g. loading weights [11,12], regression coefficients [13,14], variable in projection (VIP) [15], stability [16–19], and selective ratio [20]; optimization strategy based on artificial intelligent algorithms, e.g. genetic algorithm (GA) [21], simulated annealing (SA) [22,23], particle swarm optimization (PSO) [24] and ant colony optimization (ACO) [25]; projection strategy, e.g. successive projection algorithm (SPA) [26]. Besides, it is worth noting that model population analysis (MPA) strategy proposed by Liang's group can also be used for variable selection [27]. Based on this strategy, a series of variable selection methods has been proposed in recent years, such as iteratively retaining informative variables (IRIV) [28], variable combination population analysis (VCPA) [29], variable iterative space shrinkage approach (VISSA) [30,31], bootstrapping soft shrinkage (BOSS) [32].

As a general framework for designing new chemometrics or bioinformatics algorithms, MPA emphasizes that information should be extracted by analyzing a number of sub-models statistically, because the results or parameters of one single model are not always reliable. In detail, MPA usually contains three stages: (1) sub-datasets generation procedure, where random sampling method is applied to obtain a series of sub-datasets from variable or sample space, such as jackknife sampling [33], bootstrap sampling (BSS) [34], binary matrix sampling (BMS) [35]; (2) modeling procedure, where a series of sub-models are established based on sub-datasets generated in the previous step; (3) statistical analysis procedure, where interested outputs (e.g., RMSECV value) of all these sub-models are analyzed statistically.

Advantages of using MPA strategy to variable selection can be concluded in two aspects: (1) MPA extracts information from a large number of sub-models, which is beneficial for avoiding the uncertainty of one single model. (2) Synergistic or combination effects between different variables are more possible to be retained by MPA since random variable combinations are generated during the optimization process. Additionally, the strategy of soft shrinkage, which can avoid removing important variables by mistake, can also be regarded as an advantage of some new methods (e.g. VISSA and BOSS) developed from MPA. By this strategy, insignificant variables are not eliminated directly, but are assigned with a smaller sampling weight, ensuring that the process of optimization is implemented in the soft shrinkage way. Besides, weighted binary matrix sampling (WBMS) [30] and weighted bootstrap sampling (WBS) [36] are also two commonly used weighted random sampling methods. Up to now, there is no comparison of their performance yet.

Certainly, variable selection methods based on MPA have some drawbacks. First, their computational burden is much heavier than other methods, because they not only need to establish a large number of sub-models in each loop, but also require many loops to realize iteration convergence. Secondly, overfitting of these methods is at high risk due to the large number of variables combination [3]. Specially, WBMS generates sub-datasets too strictly depending on the sampling weights, even if the sampling weight of one variable becomes 1 by chance, it still has to be included in the future iterations.

Undoubtedly, for most kinds of spectral data, especially for near infrared spectroscopy, the selection of wavelength intervals seems more reasonable than single spectral points [3]. Because the informative variables within specific absorbing bands certainly contain similar information, which may lead some individual variable selections to chaos runs [37]. In contrast, interval selection methods can provide a more stable result. Chemical meaning can also be explained much easier. Furthermore, the selection of intervals can decrease the computational burden by reducing the number of possible combinations. It was more likely to avoid selecting single wavelengths in the noisy area which may have spurious correlations with the interested property [3]. Hence, there are a lot of spectral interval selection methods reported, such as interval partial least squares (iPLS) [38], moving windows PLS (MWPLS) [39] and many variants based on them [40–43]. Besides, some strategies commonly used for individual variable selection such as SPA [44], GA [45,46], ACO [47], etc. have also been modified for selecting informative intervals in recent years. However, MPA strategy and soft shrinkage strategy have rarely been applied to spectral interval selection.

New wavelength interval selection named as interval combination optimization (ICO) is proposed by coupling WBS with MPA, which can address drawbacks mentioned above together. In this study, three NIR datasets were applied to validate the performance of ICO. For comparison, four wavelength selection methods, including VISSA, interval VISSA (iVISSA), VISSA-iPLS and GA-iPLS, were also performed as references.

2. Theory and algorithm

2.1. Weighted binary matrix sampling (WBMS)

WBMS provides a random sampling strategy using a binary matrix [30]. In this K × P size binary matrix, K is the total sampling number and P is the number of objects. In each column of the binary matrix, “1” represents the object will be retained for modeling, while “0” represents the object will be excluded, and the ratio of “1” in each column will be updated according to the weight in each iteration. After the ranking order of each column is permutated, a new binary matrix is generated. In this new binary matrix, each row represents one random sampling procedure. Obviously, the greater the weight is, the greater the selected probability. And if the weight of one object is 1, it will be selected in every random sampling procedure, which means that it will have no possibility to be excluded. If the weight of one object is 0, it has no possibility to be retained by any random sampling procedure, which means that it will be eliminated.

2.2. Weighted bootstrap sampling (WBS)

WBS is a random sampling technique with replacement derived from BSS [36]. In WBS, one weight is allocated to one object firstly, which is between 0 and 1. Then WBS selects objects with a strategy like the roulette wheel. In this strategy, each object is corresponding to one slot on the roulette, and the size of which is proportional to the weight of the corresponding object. One object is selected in each run of this roulette. The theoretical selected probability of one object in each run can be calculated according to Equation (1). Therefore, even if the weight of one object reaches 1, it still has a chance to be excluded.

$$p_i = \frac{w_i}{\sum_1^n w_i} \quad (1)$$

where n is the number of objects. In one randomly sampling procedure of WBS, the roulette needs to repeat R times, which is determined by the average number of objects in the previous step (note that the initial R is equal to n), and the objects which are selected no less than once will be retained as one object subset. Then we can get a large number of new subsets by running WBS repeatedly, and the average number of selected objects in these new subsets will be about 0.632 times of the previous ones. Thus, the number of selected objects can be shrunk automatically by implementing WBS iteratively.

2.3. Interval combination optimization (ICO) algorithm

As a new interval selection method under the framework of MPA, the flowchart of ICO algorithm is illustrated in Fig. 1. The procedure of ICO can be summarized in the following steps:

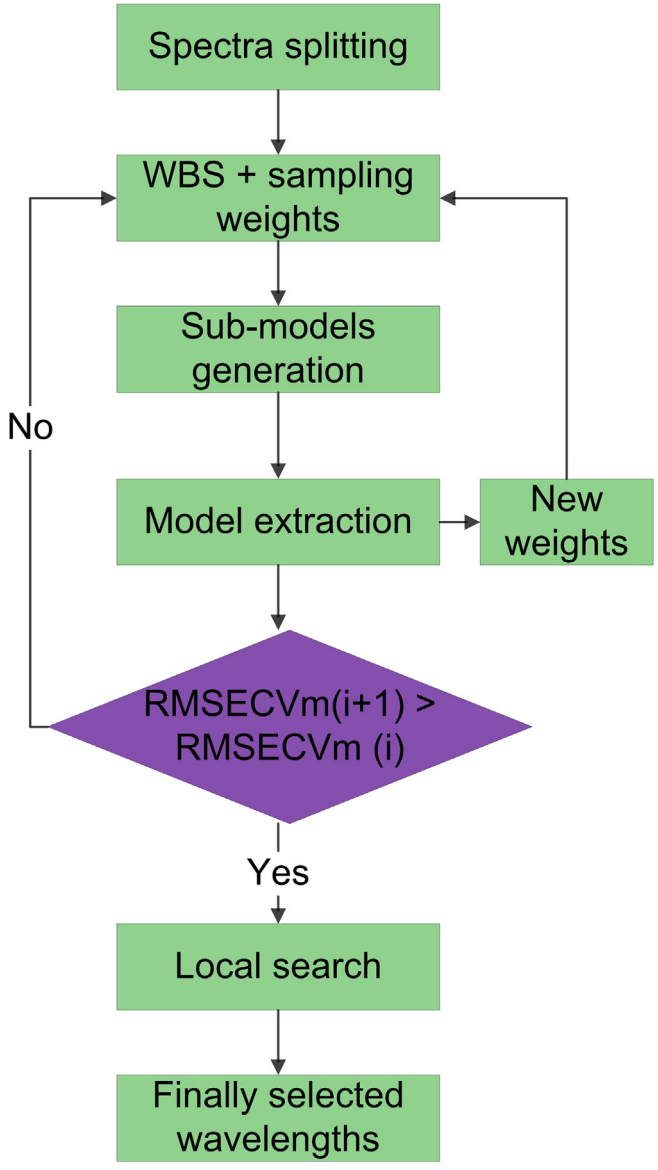


Fig. 1. Flowchart of ICO algorithm.

(1) The spectra are split into N (e.g., 40) equal-width intervals.
(2) WBS is used to generate M (e.g., 1000) random combinations of these intervals. Note that the initial weight of each interval is set to 1.
(3) RMSECV value of each interval combination is calculated by PLS algorithm and five-fold cross validation.
(4) Extract a ratio α (e.g., 10%) of best interval combinations with lowest RMSECV values, the mean value of these RMSECV values (denoted as $RMSECV_{mean}$) is recorded.
(5) Count the appearance frequency of each interval in the best interval combinations, and the new weight of interval i can be calculated according to Equation (2).

$$w_i = \frac{f_i}{k_{\text{best}}} \quad (2)$$

where f_i represents the appearance frequency of interval i in the best interval combinations, k_{best} is the number of the best interval combinations.

(6) If the $RMSECV_{mean}$ decreases, the loop goes on by repeating step (2)–(5). Otherwise, the loop terminates.
(7) Find the interval combination with the lowest RMSECV value in the final iteration, which will be considered as the best interval combination.
(8) Optimize the widths of the finally selected intervals. In this step, the variables near the boundaries of finally selected intervals will be evaluated by comparing the prediction errors (RMSECV) of the sub-models with and without them successively. If the RMSECV value of the sub-model without one variable is lower than that of with it, it will be removed. Otherwise, it will be retained.

2.4. Brief introduction of compared methods

2.4.1. Variable iterative space shrinkage approach (VISSA)

VISSA is an optimization algorithm based on MPA and WBMS [30], main procedures of which can be summarized into three main steps. Step 1: WBMS is used to generate a number (e.g., 5000) of sub-datasets, where the initial weight of each variable is set to 0.5. Step 2: a PLS model is built on each sub-dataset, prediction error (RMSECV) of which is recorded. Step 3: MPA is applied to extract information from the prospective of prediction errors of all these sub-models. The appearance probability of each variable in the 10% best models is calculated, which will be used as the new weight of each variable in step 1.

In general, the weight of each variable is updated continuously by repeating these three steps during the whole optimization procedure. From the introduction of WBMS in 2.1 we know that if the weight of one variable reaches 1, it will be included in the final set. On the contrary, if the weight of one variable reaches 0, it will be excluded from the final set. If the weight of one variable is between 0 and 1, it will be regarded as candidate variable need to be evaluated in the next iteration. Finally, VISSA is terminated automatically when the weights of all variables are constant.

2.4.2. iVISSA algorithm

Interval variable iterative space shrinkage approach (iVISSA) is an iterative method for wavelength interval selection proposed by Deng et al. [31]. In this method, the location and combination of informative individual wavelengths are searched by VISSA strategy in the global search procedure. The width of each interval is optimized in the local search procedure, which is implemented by the same way in step (8) of Section 2.3. Finally the locations, widths and

Table 1
Parameters of the GA optimization procedure in the GA-iPLS algorithm.

Population size: 30 chromosomes
On average, 5 variables per chromosome in the original population
Response: cross-validated explained variance % (the optimal number of latent variables of PLS is determined by 5 fold cross-validation)
Maximum number of variables selected in the same chromosome: 30
Probability of cross-over: 50%
Probability of mutation: 1%
Maximum number of latent variables of PLS: 10
Number of runs: 100
The amount of evaluations: 100

combinations of the informative wavelength intervals are intelligently optimized by implementing global search and local search procedure alternatively.

2.4.3. VISSA-iPLS algorithm

VISSA-iPLS algorithm is proposed in this study firstly. In this algorithm, the full spectra are divided into a number of equal-width intervals, and then the optimal interval combination is searched softly by VISSA strategy. Detail descriptions of VISSA strategy can be founded in Section 2.4.1. Because the searching targets are reduced significantly by the simple idea of iPLS, the computational burden of VISSA-iPLS will be much less than VISSA. What's more, local search is also conducted in VISSA-iPLS algorithm by the same way in step (8) of Section 2.3. It should be noted that the biggest difference between VISSA-iPLS and ICO is that they employed different randomly sampling methods.

2.4.4. GA-iPLS algorithm

GA-iPLS algorithm is a wavelength interval selection method based on GA strategy [45,48]. Similar to VISSA-iPLS algorithm, the full spectra are divided into a number of equal-width intervals firstly, and then the combinations of these intervals are searched by GA strategy. Detail descriptions of GA strategy can be founded in Ref. [49] and the parameters of GA used in this study are listed in Table 1. Because the selection results of GA strategy are always varied due to the random components in its search process, the selection frequency of each interval in 100 iterations is always calculated. Then PLS models are built on interval combinations consisted of 1, 2, 3, 4 … intervals with the highest selection frequencies respectively, and the interval combination with the lowest RMSECV value will be selected finally.

Table 2
Results of ICO with different parameters on corn dataset (the number of intervals was set to 40). M represents the number of random combinations generated in each iteration, α represents the extract ratio of best combinations, nVAR represents the number of selected variables; nLV represents number of latent variables, the number in parentheses is standard deviation of results in 20 repeated runs.

M	α	nVAR	nLV	RMSEP	Time/s
100	0.05	70.70 (5.39)	10.00 (0.00)	0.0119 (0.0045)	8.28 (0.76)
	0.10	69.30 (1.34)	10.00 (0.00)	0.0112 (0.0016)	9.43 (0.60)
	0.20	69.95 (4.25)	10.00 (0.00)	0.0110 (0.0021)	9.79 (0.75)
	0.30	71.20 (4.44)	10.00 (0.00)	0.0125 (0.0030)	10.57 (0.69)
	0.40	73.65 (9.20)	10.00 (0.00)	0.0204 (0.0105)	11.30 (0.78)
	0.50	85.30 (42.90)	9.95 (0.22)	0.0262 (0.0287)	12.06 (1.86)
500	0.05	69.00 (0.00)	10.00 (0.00)	0.0106 (0.0000)	43.25 (1.85)
	0.10	69.00 (0.00)	10.00 (0.00)	0.0106 (0.0000)	45.08 (1.06)
	0.20	69.00 (0.00)	10.00 (0.00)	0.0106 (0.0000)	48.33 (0.92)
	0.30	69.00 (0.00)	10.00 (0.00)	0.0106 (0.0000)	52.53 (1.58)
	0.40	69.65 (1.73)	10.00 (0.00)	0.0111 (0.0017)	55.98 (1.78)
	0.50	71.20 (4.88)	10.00 (0.00)	0.0115 (0.0017)	55.87 (2.32)
1000	0.05	69.00 (0.00)	10.00 (0.00)	0.0106 (0.0000)	89.11 (2.93)
	0.10	69.00 (0.00)	10.00 (0.00)	0.0106 (0.0000)	92.19 (2.17)
	0.20	69.00 (0.00)	10.00 (0.00)	0.0106 (0.0000)	96.33 (2.17)
	0.30	69.00 (0.00)	10.00 (0.00)	0.0106 (0.0000)	102.45 (2.40)
	0.40	69.90 (4.02)	10.00 (0.00)	0.0110 (0.0019)	106.90 (5.83)
	0.50	71.15 (5.98)	10.00 (0.00)	0.0118 (0.0037)	113.00 (4.98)

3. Experimental

3.1. NIR spectra of corn samples

This benchmark NIR datasets of corn samples are available at the website: http://www.eigenvector.com/data/Corn/index.html. NIR spectra measured by the m5 NIR spectrometer were used in this study. The spectra were acquisitioned within the range 1100–2498 nm at intervals of 2 nm. Thus, each spectrum consists of 700wavelength points. The protein content (%, w/w) was considered as property of interest. In addition, 80 samples were split into calibration set (60 samples) and validation set (20 samples) according to the SPXY algorithm [50].

3.2. Diffuse reflectance NIR spectra of soil samples

This dataset was downloaded at the website: http://www.models.life.ku.dk/NIRsoil. It consists of 108 samples generated at a long-term field experiment in Abisko, northern Sweden (68°21′N, 18°49′E) [51]. The spectra were recorded within the range of 400–2500 nm at 2 nm intervals. Soil organic matter (SOM) content (%, w/w) was considered as the interested property in this study, which was measured as loss on ignition at 550 °C. After the deletion of six outliers which were detected by the Monte-Carlo outlier detection approach [52], the remaining 102 samples were split into calibration set (62samples) and validation set (40 samples) according to the SPXY algorithm [50].

3.3. Transmittance NIR spectra of pharmaceutical tablets

655 transmittance spectra of pharmaceutical tablets were downloaded from the web of http://software.eigenvector.com/Data/tablets/index.html. The spectra measured on Instrument I (Foss NIR Systems 6500 spectrometer) were used in this study. The spectra were recorded within the range of 600–1898 nm at an interval of 2 nm. The active pharmaceutical ingredient (API) content (%, w/w) of each individual tablet was analyzed by HPLC method. This dataset has been split into three different files, thus 155 spectra in the calibration file were used as calibration set, 460

Table 3
Results of different methods on three datasets. nVAR represents the number of selected variables; nLV represents number of latent variables.

Datasets	Methods	nVAR	nLV	RMSEP	Time/s
Corn protein	PLS	700	10	0.1442	0.14
	VISSA	172	10	0.1102	3027.63
	iVISSA	241	10	0.1318	3946.88
	VISSA-iPLS	105	10	0.0196	107.33
	GA-iPLS	70	10	0.0118	118.83
	ICO	**69**	**10**	**0.0106**	**83.69**
Soil SOM	PLS	1050	10	1.76	0.17
	VISSA	268	10	1.26	5008.11
	iVISSA	501	10	1.41	6387.31
	VISSA-iPLS	229	10	1.18	223.83
	GA-iPLS	182	10	1.12	145.09
	ICO	**131**	**10**	**0.93**	**159.06**
Tablets API	PLS	650	5	0.75	0.20
	VISSA	182	5	0.62	3075.00
	iVISSA	275	5	0.63	3679.47
	VISSA-iPLS	139	5	0.47	197.03
	GA-iPLS	48	5	0.65	120.11
	ICO	**34**	**5**	**0.42**	**99.33**

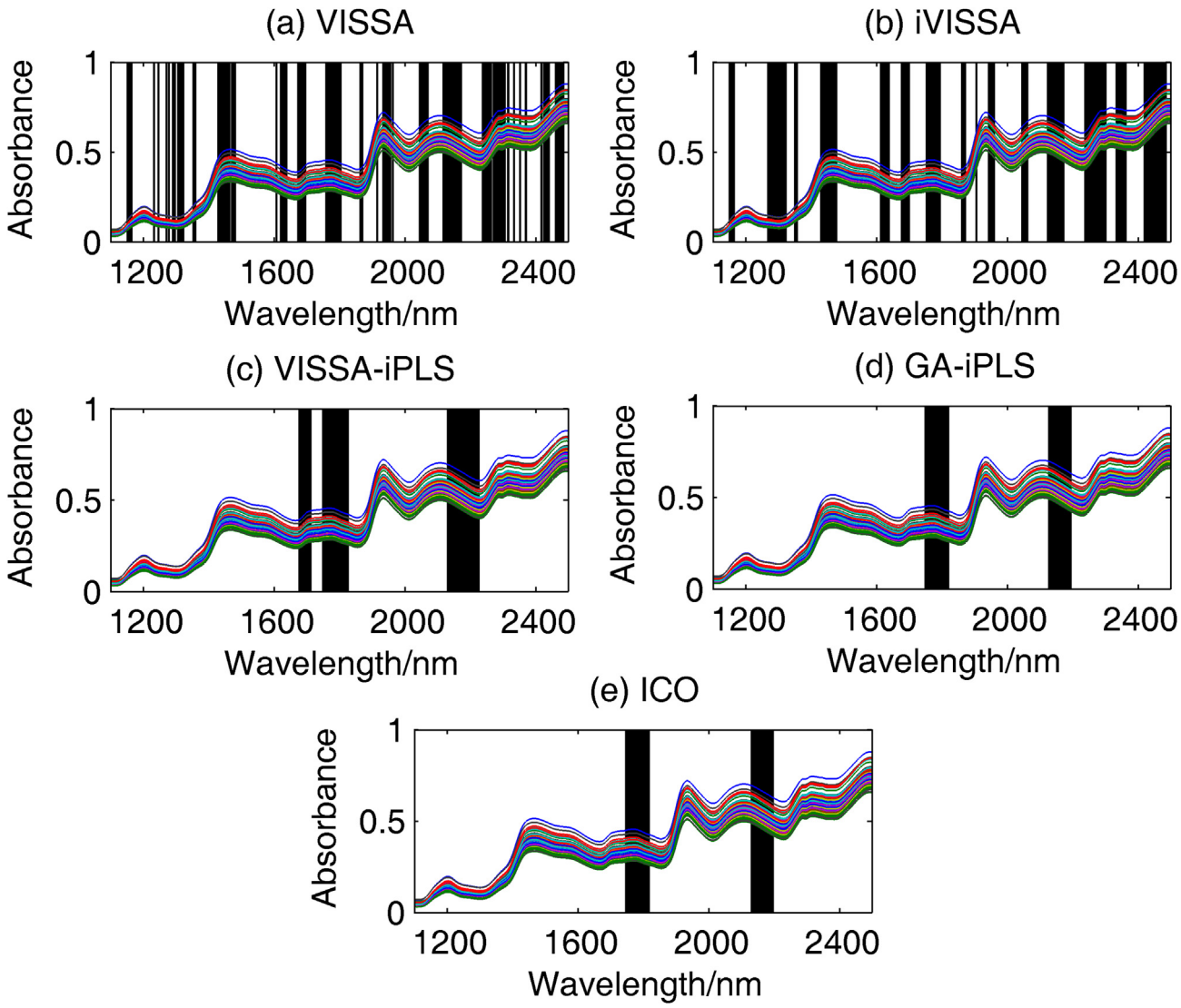


Fig. 2. NIR spectra of corn samples and wavelengths selected by different methods.

spectra in the validation file were used as validation set, and 40 spectra in the test file were used as test set.

3.4. Calculation and software

All computations were performed in MATLAB (Version 2011a, the MathWorks, Inc.) on a general personal computer (configured with Intel® Pentium® G630 CPU (2.7 GHz), 2 GB RAM, and Microsoft® Windows XP operating system). MATLAB codes of both VISSA and iVISSA were downloaded from the web of http://www.mathworks.com/matlabcentral/fileexchange on March 31, 2016. ICO, GA-iPLS and VISSA-iPLS were all realized with home-made codes which are available upon request.

3.5. Modeling strategy and model evaluation parameters

In this study, all the data was centered to have zero mean before modeling. Partial least squares regression was used for modeling, and the maximum number of latent variables was set to 10. The optimal number of latent variables was determined by 5-fold cross validation (ASTM E1655-05). The quality of the model is assessed by the root mean squared error of prediction set (RMSEP), which is calculated according to Equation (3).

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \quad (3)$$

where n is the number of samples, yi and ŷi represent the measured and predicted values of the i th sample, respectively.

4. Results and discussion

4.1. The influence of different parameters on ICO algorithm

It should be noted that there are only three parameters including N, M, α need to be optimized in ICO algorithm, which is relatively few when compared with some intelligent optimization algorithms such as GA, ACO, and PSO. Firstly, as pointed by Ref. [53], if N is too low, the interval will be too broad, which may lead to the neglect of some narrow peaks; if N is too large, the results will rely too much on a local scale, and computational burden of the searching procedure will be increased. In consideration of the fact that the actual width of most NIR spectral absorbance peaks is

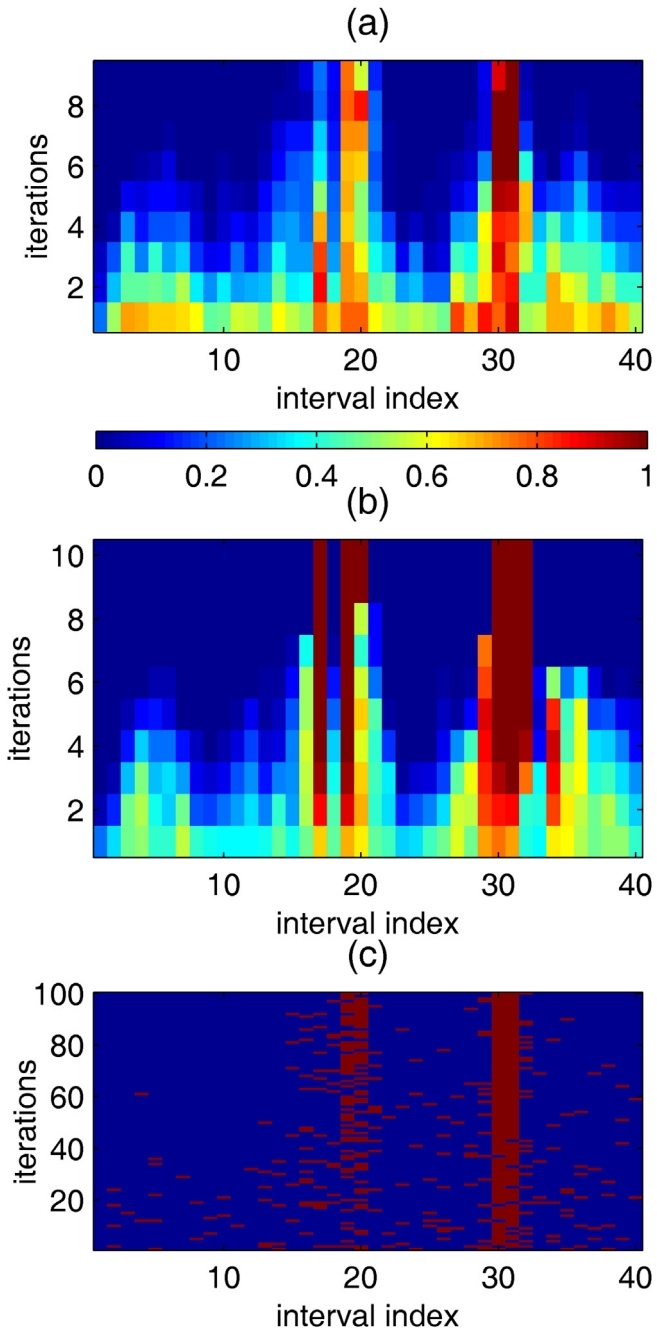


Fig. 3. Sampling weights of each interval in the optimization process of (a) ICO, (b) VISSA-iPLS and (c) GA-iPLS on corn dataset.

usually around 20–100 nm, 20 or 40 intervals are commonly recommended in most literature [42]. Thus, in order to test the ability of ICO for retaining the synergistic effect between different intervals more adequately, N was set to 40 for all datasets in this study. However, since the width of intervals is fixed, we should pay attention to two kinds of wavelengths, including informative wavelengths outside and noise or uninformative wavelengths inside of the finally selected intervals. Therefore, local search also should be conducted to optimize the widths of finally selected intervals respectively according the continuity of spectra.

The other two parameters M and α represent the number of random combinations need to be generated and the extracting ratio of better combinations respectively. Because they are main parameters for implementing MPA strategy, the influence of these two parameters on the final results of variable selection methods based on MPA has been discussed in previous literature [29,30,54]. From these literature, we can find that although all these methods are not sensitive to these two parameters, larger M and lower α still tend to generate more accurate and stable results. In addition, the same conclusion can also be obtained when we take corn dataset as example, and the results of ICO with different parameters on corn dataset are displayed in Table 2. As was shown in this table, if M is not large enough, e.g. 100, the results of ICO will be unstable no matter how many does α set to, which can be proved by the standard deviation value of number of variables selected by ICO. Whereas, if α is too large, e.g. 0.5, the results of ICO will also be unstable due to the inclusion of some bad information. Furthermore, it's also worth noting that the computational time of ICO is approximately proportional to M. Reasons may be that the computation time of ICO is determined by two factors including M and the number of iterations need to be conducted, whereas the latter one was almost fixed due to the implementation of WBS. Thus, in order to get stable results within appropriate computation time, M and α were set to 1000 and 0.05 respectively in ICO algorithm.

4.2. Analysis of protein content in corn samples

The results of different methods on corn dataset are displayed in Table 3. As was shown in this table, all variable selection methods show better predictive performance when compared with the full spectra PLS model, which demonstrates the necessity of conducting wavelength selection. From this table, we can also find that all three interval selection methods GA-iPLS, VISSA-iPLS and ICO can select fewer wavelengths with even better predictive performance in much less computation time when compared with VISSA. It indicates that the selection of spectral intervals rather than individual wavelengths can not only reduce the computational burden of one optimization strategy, but also help to improve the effectiveness of selected wavelengths. In detail, all three interval selection methods need to conduct optimization among only 40 intervals. Whereas, VISSA should conduct optimization among 700 individual wavelengths, which will lead to much higher risk of overfitting. Furthermore, we can also find that the performance of iVISSA is even worse than VISSA in terms of both number of selected wavelengths and RMSEP, which may also due to overfitting. As was introduced in section 2.4.3, local search and VISSA were implemented alternatively in iVISSA, which will increase its computational burden and risk of overfitting simultaneously.

Fig. 2 shows the NIR spectra of corn samples and wavelengths selected by five different methods. As can be observed from Fig. 2(a), wavelengths selected by VISSA are distributed across the whole spectral range, which is consistent with its performance in previous report [30]. It indicates that although VISSA can retain the synergistic effect between different wavelengths more adequately, it is still faced with high possibility of overfitting, which is similar with the fact that GA-PLS can hardly perform well when there are more than 200 variables need to be optimized [53]. The distribution of selected wavelengths of iVISSA is displayed in Fig. 2(b), which is quite similar with VISSA. We can also observe that some discrete individual wavelengths in 1250–1320 nm and 2320–2350 nm selected by VISSA have been expanded into continuous regions selected by iVISSA through implementing local search strategy. It demonstrates that although the finally selected wavelengths of iVISSA are some continuous spectral regions, it still relies on discrete wavelengths selected by VISSA to determine the location of finally selected spectral interval.

Fig. 2(c)–(e) display the spectral regions selected by VISSA-iPLS,

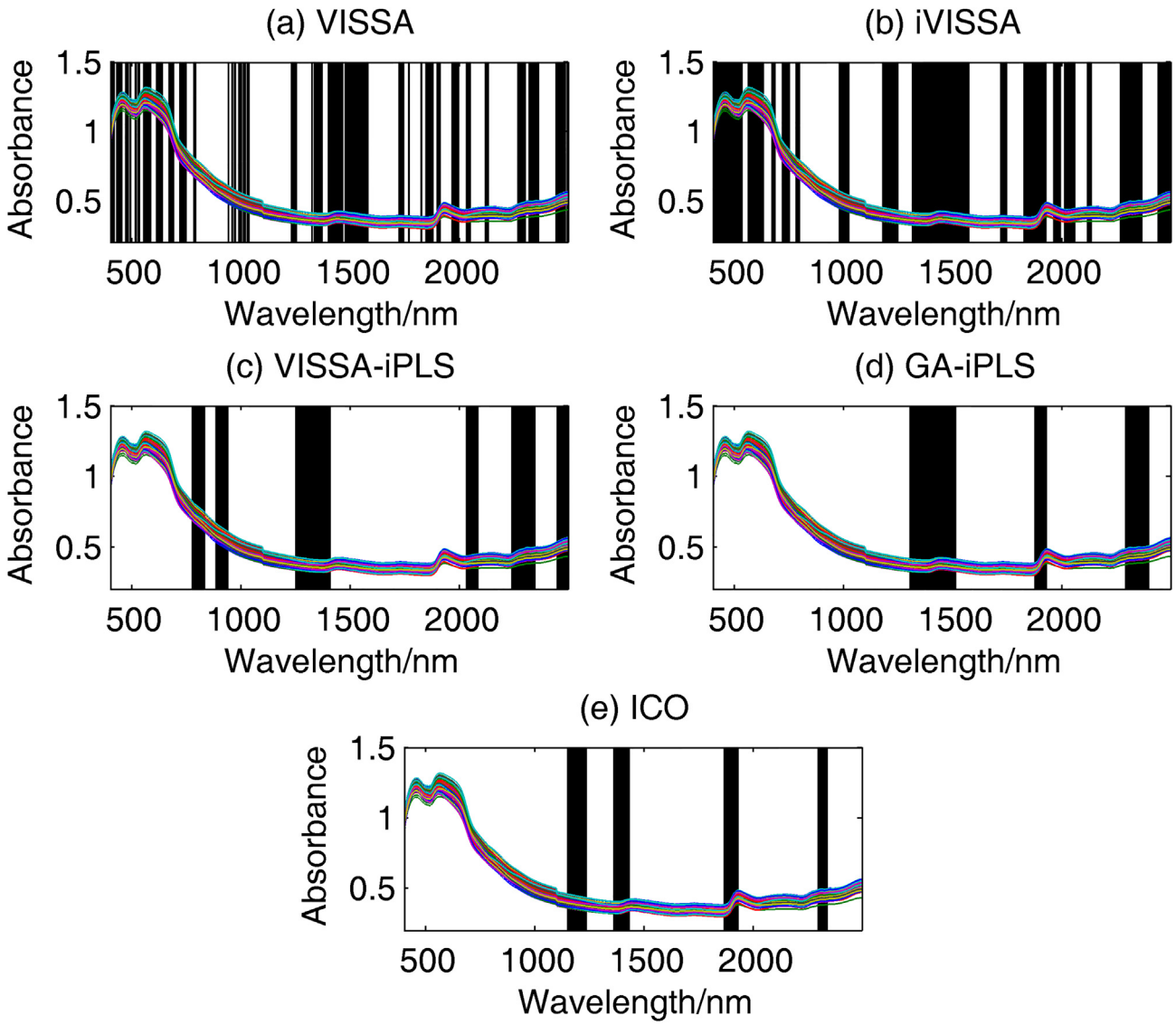


Fig. 4. NIR spectra of soil samples and wavelengths selected by different methods.

GA-iPLS and ICO respectively. It was shown that all these three methods can select wavelengths from the regions of 1740–1820 nm and 2130–2190 nm, which can be assigned to the first overtone of C–H stretching and varied vibration combinations of N–H in protein structure respectively. Furthermore, these two regions have been proved to be important for analysing protein content in Refs. [13,55], indicating that these methods are both efficient wavelength selection methods. However, VISSA-iPLS still selected some more wavelengths outside of these two regions, such as 1670–1710 nm and 2192–2224 nm, which may explain why VISSA-iPLS gave slightly worse predictive performance than ICO. In addition, we can find that there is only very small difference between the wavelengths selected by ICO and GA-iPLS, which is caused by local search in ICO.

Fig. 3(a) and (b) show the sampling weights of each interval in the iteration process of ICO and VISSA-iPLS respectively. In these figures, the sampling weight of each interval is illustrated by different colors. In detail, if one interval is in dark red, its sampling weight is 1; if one interval is in dark blue, its sampling weight is 0. And the sampling weight of one interval in other color is between 0 and 1. As was shown in these figures, the optimal interval combination is searched in a soft shrinkage manner in both of these two methods. In detail, uninformative intervals were not eliminated directly, but were assigned to smaller sampling weights, which can help to lower the risk of removing informative intervals by mistake. In Fig. 3(b), we can find that the sampling weight of one interval will always be 1 in the next iterations, as long as it has the chance to become 1. The reason for this phenomenon is that WBMS generates random combinations of different intervals according to sampling weights strictly, which has been described in Section 2.1. In contrast, even if the sampling weights of some informative intervals become 1 in the iteration process of ICO, they still have a chance to be excluded in the next iterations. For example, in Fig. 3(a), although the sampling weight of 30th interval has become 1 in the 8th iteration, its sampling weight can still be less than 1 in the 9th iteration. The reason is that WBS is a random sampling method with replacements. Thus, ICO can ensure that every informative interval can still have a chance to be evaluated in the next iteration, which is also beneficial for avoiding the problem of iteration termination with local minimum.

After a number of iterations, the sampling weights of most intervals will be decreased to 0 (dark blue) in the last iteration, which

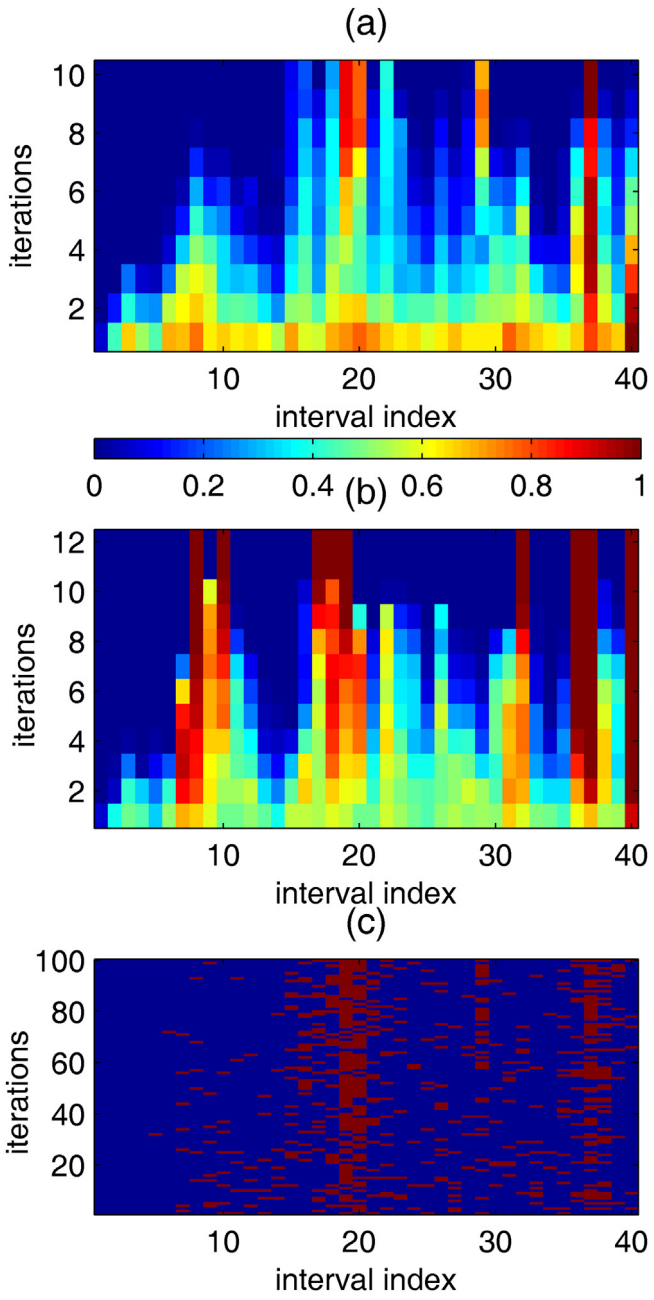


Fig. 5. Sampling weights of each interval in the optimization process of (a) ICO, (b) VISSA-iPLS and (c)GA-iPLS on soil dataset.

means they will be excluded. However, there is a great difference between the sampling weights of intervals retained in the last iteration between these two methods. The sampling weight of all the retained intervals in the last iteration of VISSA-iPLS can only be equal to either 1 or 0, whereas, the sampling weight of some retained intervals (e.g. 17, 18, 19, 20, 30th interval) in the last iteration of ICO can still be equal to non-integer that is between 0 and 1. It demonstrates that ICO can ensure that most retained intervals still have a chance to be excluded even in the last iteration.

Fig. 3(c) show selected intervals in the iteration process of GA-iPLS, and there are only two pure colors in this figure. Intervals in deep red were selected; intervals in deep blue were excluded. In this figure, we can observe that GA-iPLS always selects different intervals in different iteration due to the crossover and mutation step in the process of GA optimization. In order to find the optimal interval combination, GA-iPLS has to conduct stepwise selection according to the selected frequency of each interval. Finally, GA-iPLS selected four intervals, which are the same with intervals selected by ICO. It demonstrates that although the selection strategies of these two methods are different, they are both efficient for informative interval selection in corn dataset.

4.3. Analysis of soil organic matter (SOM) content in soil samples

Results of different methods on soil samples are also displayed in Table 2. It can be seen from this table that all variable selection methods have made some improvement in terms of RMSEP when compared with full spectrum PLS, which is similar with the results of corn samples. In addition, we can also find that all three interval selection methods (VISSA-iPLS, GA-iPLS and ICO) can select less wavelengths with better predictive performance in much less computation times when compared with iVISSA and VISSA, which proves once again the effectiveness of implementing wavelength selection on intervals rather than individual wavelengths. It also should be noted that although iVISSA selects wavelengths based on VISSA, its predictive performance cannot outperform VISSA. We believe that this is due to overfitting, because iVISSA has to conduct local search around too many individual wavelengths selected by VISSA. In addition, ICO can give the best performance in terms of both the number of selected wavelengths and RMSEP value among all three interval selection methods, which indicates that ICO is superior to VISSA-iPLS and GA-iPLS.

Fig. 4 shows the spectra of soil samples and wavelengths selected by different methods. From this figure, we can observe that the wavelengths selected by VISSA and iVISSA were also quite similar, which is consistent with the results of corn dataset. They all tend to select wavelengths from the whole spectral range, which can hardly avoid selecting some wavelengths from uninformative regions. On the contrary, ICO selected wavelengths from only four spectral regions with the best predictive performance. These four regions were around 1200 nm, 1430 nm, 1920 nm, 2300 nm respectively, which have been proved to be related with the soil organic matter in previous reports [56]. In detail, the wavelengths around 1200 nm and 2300 nm are related to the second overtone of C—H stretch and the combination absorbance of C—H vibration in various kinds of organic matters respectively; the wavelengths around 1430 nm can be attributed to OH groups in water or cellulose, or to CH2 groups in lignin; the wavelengths around 1920 nm can be assigned to OH groups in water or various functional groups present in cellulose, lignin, glucan, starch, pectin and humic acid.

Fig. 4(c) shows wavelengths selected by VISSA-iPLS. As was shown in this figure, we can find that VISSA-iPLS not only selected wavelengths from some informative regions mentioned above, but also selects wavelengths from some uninformative regions, such as regions around 800 nm and 900 nm, which may explain its relatively poor predictive performance when compared with ICO. In contrast, GA-iPLS selected wavelengths only from informative regions, which is shown in Fig. 4(d). However, GA-iPLS neglected one informative region around 1200 nm, thus it still cannot give comparable predictive performance when compared with ICO.

The sampling weights of each interval during the iteration process in ICO, VISSA-iPLS and GA-iPLS are displayed in Fig. 5. As was shown in Fig. 5(a) and (b), we can find that the sampling weights of most intervals during the iteration process in ICO are less than 1, which can ensure that all the retained intervals can have a chance to be excluded in the next iteration. In contrast, the sampling weights of some intervals finally retained by VISSA-iPLS have increased to 1 quickly in the first few times of iteration, which will have no chance to be excluded in the next iteration. This is also the

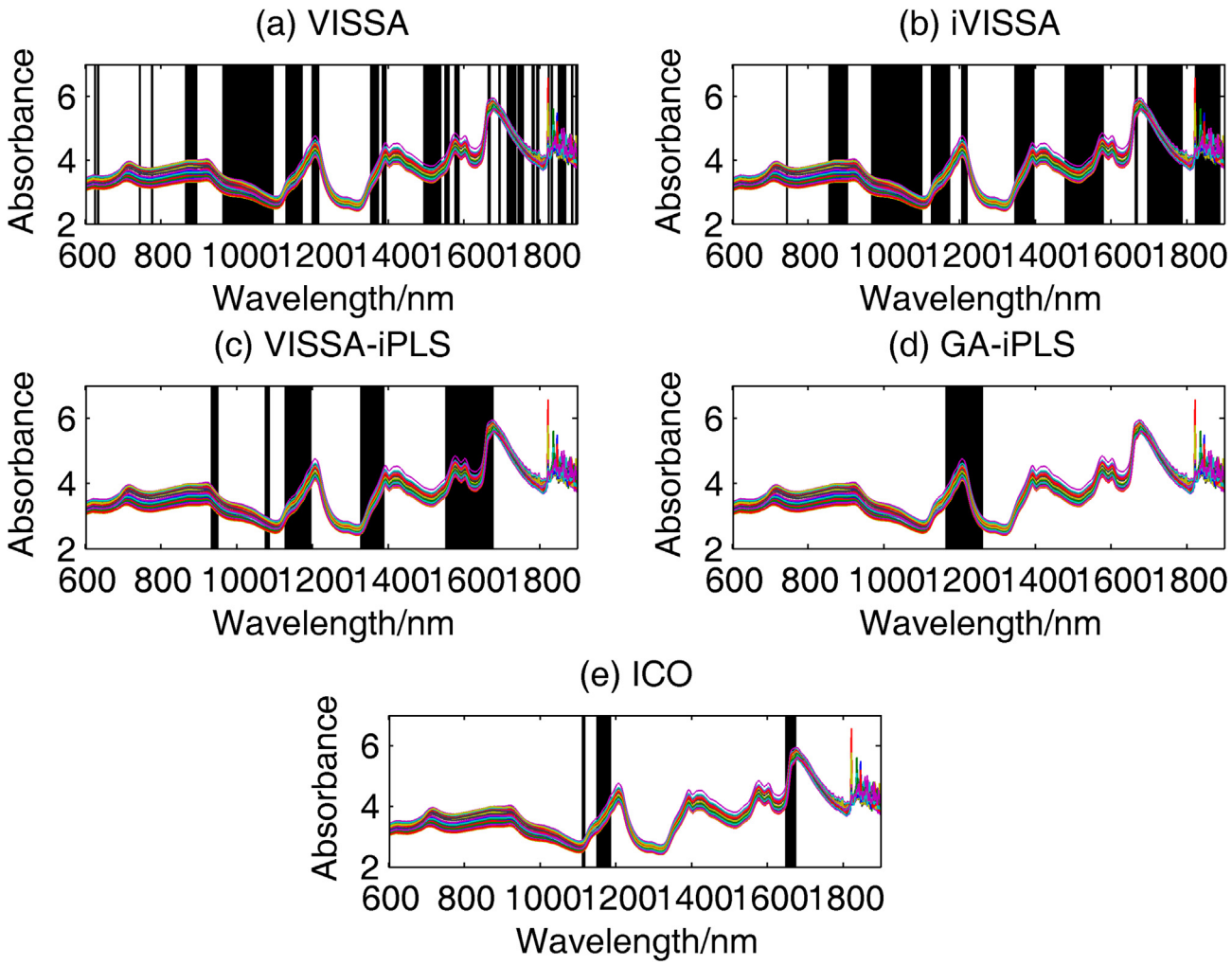


Fig. 6. NIR spectra of pharmaceutical tablets and wavelengths selected by different methods.

reason why VISSA-iPLS cannot select informative intervals as efficiently as ICO. The deep reason to this phenomenon is also that they employed different randomly sampling methods.

Fig. 5(c) shows intervals selected in the iteration process of GA-iPLS. In this figure, we can observe that although the selected frequencies of most informative intervals are relatively high, such as 19, 20, 29, 36, 37th interval, GA-iPLS still neglected the 16th interval due to low selected frequency, which may explain why it cannot give comparable predictive performance with ICO.

4.4. Analysis of active pharmaceutical ingredient (API) content in single tablet

The results of different methods on pharmaceutical tablets are also displayed in Table 3. In this dataset, the maximum number of latent variables was set to 5 for avoiding overfitting, which is different to the other datasets. This is because that the optimal number of latent variables of the full spectra model is equal to 5, which was much less than 10. From this table, we can find that although the number of wavelengths selected by ICO was much less than that of other four methods, it still gave the best predictive performance, indicating that ICO is a more efficient wavelength interval selection method. In addition, all three interval selection methods consumed much less computation time than VISSA and iVISSA, indicating that the selection of intervals rather than individual wavelengths can improve the computation speed of one selection strategy significantly.

Fig. 6 shows the spectra of pharmaceutical tablets and wavelengths selected by different methods. Different from the other two datasets, a small spectral region (1800–1898 nm) with apparent low S/N appears in the original spectra of pharmaceutical tablets. However, both VISSA and iVISSA still selected some wavelengths from this region. What's worse, iVISSA selected even more wavelengths from this region than VISSA, which may explain its poor predictive performance. In contrast, all three interval selection methods can avoid selecting wavelengths from this region, indicating that the selection of wavelength intervals instead of individual wavelengths can help to avoid selecting single wavelengths in the noisy area which may have spurious correlations with the responded property. Benefit from local search procedure in ICO and VISSA-iPLS, the initially selected equal width intervals have become spectral regions with different widths finally, which can be observed in Fig. 6(c) and (e). However, GA-iPLS selected wavelengths from only one broad region around 1200 nm, which are shown in Fig. 6(d). It may explain why it gave worse predictive performance than other wavelength selection methods.

Fig. 7 shows the sampling weights of each interval during the iteration process in ICO, VISSA-iPLS and GA-iPLS respectively. As was shown in this figure, we can observe that 39th and 40th interval in the low S/N regions were excluded through the whole

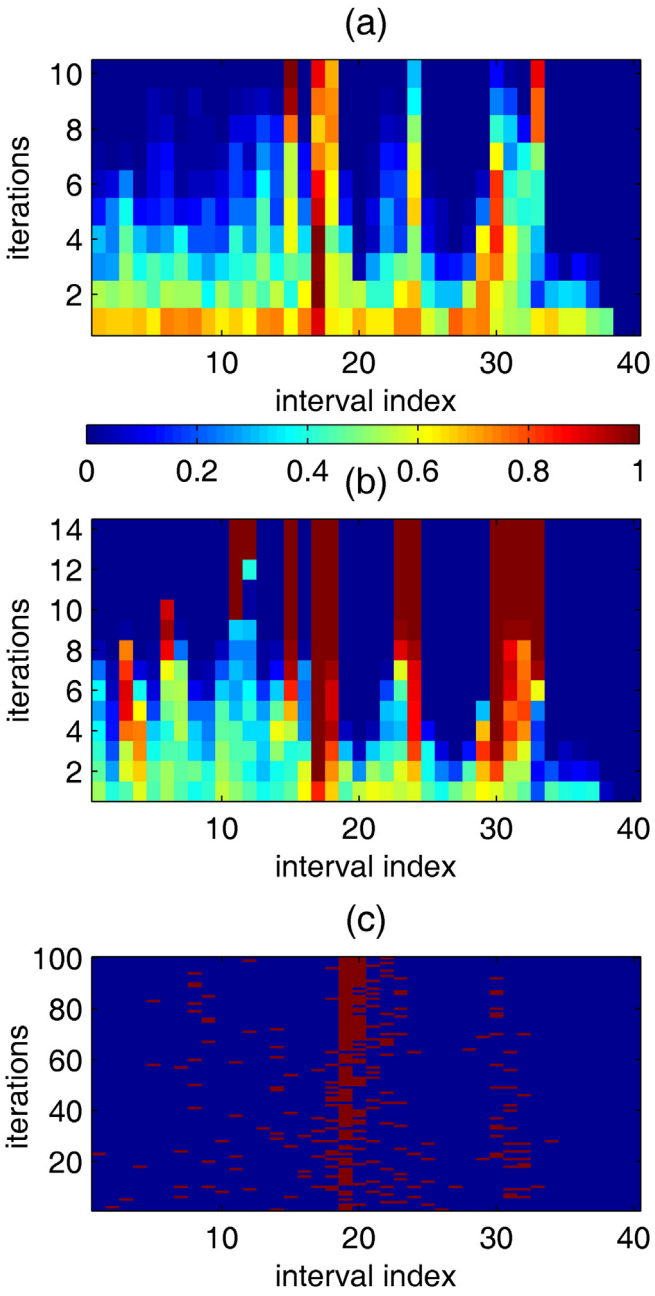


Fig. 7. Sampling weights of each interval in the optimization process of (a) ICO, (b) VISSA-iPLS and (c) GA-iPLS on pharmaceutical tablets dataset.

iteration process of all three interval selection methods, which demonstrates once again the advantage of conducting selection on wavelength interval. Furthermore, ICO and VISSA-iPLS conducted 10 and 14 iterations to reach iteration convergence respectively, indicating that the convergence speed of ICO is faster than VISSA-iPLS. In contrast, GA-iPLS didn't converge after 100 iterations, because there are too many random components in GA strategy. Hence, although GA also generated many different interval combinations during the process of optimization, some important intervals still have no chance to get selected frequencies high enough to be selected finally, because most selected intervals were mainly around the 19th interval (Fig. 7(c)). This phenomenon also indicates that the reason why GA-iPLS failed to find the optimal interval combination in this dataset.

5. Conclusion

A new wavelength interval selection method named as ICO was proposed by coupling MPA and WBS in this study. In this method, spectral interval was used instead of individual wavelengths. Then, the optimal combination of spectral intervals can be searched in a soft shrinkage mode. Moreover, the widths of finally selected intervals can also be optimized in ICO automatically. Three different NIR spectral datasets were applied to validate the performance of ICO. Results showed that, ICO can select the optimal wavelength interval combination with the best prediction performance effectively. It demonstrates that ICO not only inherits the advantages of MPA and soft shrinkage strategy, but also overcomes the disadvantages of WBMS by introducing appropriate scales of random components into the sampling step with WBS. Furthermore, it was also proved that the selection of interval rather than individual wavelengths can indeed reduce the risk of overfitting, as well as computational burden of MPA. Hence, ICO may be a good alternative wavelength selection method for spectroscopic data.

Acknowledgements

This research is financially supported by National Natural Science Foundation of China (Grant No. 31301685).

References

[1] C.H. Spiegelman, M.J. McShane, M.J. Goetz, M. Motamedi, Q.L. Yue, G.L. Coté, Theoretical justification of wavelength selection in PLS calibration: development of a new algorithm, Anal. Chem. 70 (1998) 35–44.
[2] Y.-H. Yun, Y.-Z. Liang, G.-X. Xie, H.-D. Li, D.-S. Cao, Q.-S. Xu, A perspective demonstration on the importance of variable selection in inverse calibration for complex analytical systems, Analyst 138 (2013) 6412–6421.
[3] C.M. Andersen, R. Bro, Variable selection in regression—a tutorial, J. Chemom. 24 (2010) 728–737.
[4] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, A review of variable selection methods in partial least squares regression, Chemom. Intell. Lab. 118 (2012) 62–69.
[5] 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.
[6] V. Giovenzana, R. Civelli, R. Beghi, R. Oberti, R. Guidetti, Testing of a simplified LED based vis/NIR system for rapid ripeness evaluation of white grape (Vitis vinifera L.) for Franciacorta wine, Talanta 144 (2015) 584–591.
[7] J.M. Sutter, J.H. Kalivas, Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection, Microchem. J. 47 (1993) 60–66.
[8] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira, Sorting variables by using informative vectors as a strategy for feature selection, J. Chemom. 23 (2009) 32–48.
[9] J.P. Andries, Y. Vander Heyden, L.M. Buydens, Predictive-property-ranked variable reduction in partial least squares modelling with final complexity adapted models: comparison of properties for ranking, Anal. Chim. Acta 760 (2013) 34–45.
[10] M. Farrés, S. Platikanov, S. Tsakovski, R. Tauler, Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation, J. Chemom. 29 (2015) 528–536.
[11] F. Lindgren, P. Geladi, S. Rännar, S. Wold, Interactive variable selection (IVS) for PLS. Part 1: theory and algorithms, J. Chemom. 8 (1994) 349–363.
[12] F. Lindgren, P. Geladi, A. Berglund, M. Sjöström, S. Wold, Interactive variable selection (IVS) for PLS. Part II: chemical applications, J. Chemom. 9 (1995) 331–342.
[13] H. Li, Y. Liang, Q. Xu, D. Cao, Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration, Anal. Chim. Acta 648 (2009) 77–84.
[14] J.P. Andries, Y. Vander Heyden, L.M. Buydens, Improved variable reduction in partial least squares modelling based on predictive-property-ranked variables and adaptation of partial least squares complexity, Anal. Chim. Acta 705 (2011) 292–305.
[15] O.M. Kvalheim, R. Arneberg, O. Bleie, T. Rajalahti, A.K. Smilde, J.A. Westerhuis, Variable importance in latent variable regression models, J. Chemom. 28 (2014) 615–622.
[16] V. Centner, D.-L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, C. Sterna, Elimination of uninformative variables for multivariate calibration, Anal. Chem. 68 (1996) 3851–3858.
[17] W. Cai, Y. Li, X. 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] Q.-J. Han, H.-L. Wu, C.-B. Cai, L. Xu, R.-Q. Yu, An ensemble of Monte Carlo uninformative variable elimination for wavelength selection, Anal. Chim. Acta 612 (2008) 121–125.
[19] K. Zheng, Q. Li, J. Wang, J. Geng, P. Cao, T. Sui, X. Wang, Y. Du, Stability competitive adaptive reweighted sampling (SCARS) and its applications to multivariate calibration of NIR spectra, Chemom. Intell. Lab. 112 (2012) 48–54.
[20] T. Rajalahti, R. Arneberg, A.C. Kroksveen, M. Berle, K.M. Myhr, O.M. Kvalheim, Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles, Anal. Chem. 81 (2009) 2581–2590.
[21] R. Leardi, A.L. Gonzalez, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemom. Intell. Lab. 41 (1998) 195–207.
[22] 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.
[23] U. Hörchner, J.H. Kalivas, Simulated-annealing-based optimization algorithms: fundamentals and wavelength selection applications, J. Chemom. 9 (1995) 283–308.
[24] X. Wang, J. Yang, X. Teng, W. Xia, R. Jensen, Feature selection based on rough sets and particle swarm optimization, Pattern Recognit. Lett. 28 (2007) 459–471.
[25] M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad, M. Akhond, Ant colony optimisation: a powerful tool for wavelength selection, J. Chemom. 20 (2006) 146–157.
[26] M.C.U. Araújo, T.C.B. Saldanha, R.K.H. Galvão, T. Yoneyama, H.C. Chame, V. Visani, The successive projections algorithm for variable selection in spectroscopic multicomponent analysis, Chemom. Intell. Lab. 57 (2001) 65–73.
[27] H.D. Li, Y.Z. Liang, Q.S. Xu, D.S. Cao, Model population analysis for variable selection, J. Chemom. 24 (2010) 418–423.
[28] Y.-H. Yun, W.-T. Wang, M.-L. Tan, Y.-Z. Liang, H.-D. Li, D.-S. Cao, H.-M. Lu, Q.-S. Xu, A strategy that iteratively retains informative variables for selecting optimal variable subset in multivariate calibration, Anal. Chim. Acta 807 (2014) 36–43.
[29] Y.-H. Yun, W.-T. Wang, B.-C. Deng, G.-B. Lai, X.-b. Liu, D.-B. Ren, Y.-Z. Liang, W. Fan, Q.-S. Xu, Using variable combination population analysis for variable selection in multivariate calibration, Anal. Chim. Acta 862 (2015) 14–23.
[30] B.-C. Deng, Y.-h. Yun, Y.-z. Liang, L.-z. Yi, A novel variable selection approach that iteratively optimizes variable space using weighted binary matrix sampling, Analyst 139 (2014) 4836–4845.
[31] B.-C. Deng, Y.-H. Yun, P. Ma, C.-C. Lin, D.-B. Ren, Y.-Z. Liang, A new method for wavelength interval selection that intelligently optimizes the locations, widths and combinations of the intervals, Analyst 140 (2015) 1876–1885.
[32] B.-C. Deng, Y.-H. Yun, D.-S. Cao, Y.-L. Yin, W.-T. Wang, H.-M. Lu, Q.-Y. Luo, Y.-Z. Liang, A bootstrapping soft shrinkage approach for variable selection in chemical modeling, Anal. Chim. Acta 908 (2016) 63–74.
[33] R.G. Miller, The jackknife -a review, Biometrika 61 (1974) 1–15.
[34] R.W. Johnson, An Introduction to the Bootstrap, Chapman & Hall.
[35] H. Zhang, H. Wang, Z. Dai, M.S. Chen, Z. Yuan, Improving accuracy for cancer classification with a new algorithm for genes selection, Bmc Bioinforma. 13 (2012) 1–20.
[36] P. Barbe, P. Bertail, The weighted bootstrap, Lect. Notes Stat. 98 (1995).
[37] G. Tang, Y. Huang, K. Tian, X. Song, H. Yan, J. Hu, Y. Xiong, S. Min, A new spectral variable selection pattern using competitive adaptive reweighted sampling combined with successive projections algorithm, Analyst 139 (2014) 4894–4902.
[38] L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S. Engelsen, Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy, Appl. Spectrosc. 54 (2000) 413–419.
[39] J.-H. Jiang, R.J. Berry, H.W. Siesler, Y. Ozaki, Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data, Anal. Chem. 74 (2002) 3555–3565.
[40] Y. Du, Y. Liang, J. Jiang, R. Berry, Y. Ozaki, Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares, Anal. Chim. Acta 501 (2004) 183–191.
[41] L. Munck, J.P. Nielsen, B. Møller, S. Jacobsen, I. Søndergaard, S.B. Engelsen, L. Nørgaard, R. Bro, Exploring the phenotypic expression of a regulatory proteome-altering gene by spectroscopy and chemometrics, Anal. Chim. Acta 446 (2001) 169–184.
[42] Y.-H. Yun, H.-D. Li, L.R. Wood, W. Fan, J.-J. Wang, D.-S. Cao, Q.-S. Xu, Y.-Z. Liang, An efficient method of wavelength interval selection based on random frog for multivariate spectral calibration, Spectrochim. Acta A 111 (2013) 31–36.
[43] L.P. Brás, M. Lopes, A.P. Ferreira, J.C. Menezes, A bootstrap-based strategy for spectral interval selection in PLS regression, J. Chemom. 22 (2008) 695–700.
[44] A. de Araújo Gomes, R.K.H. Galvão, M.C.U. de Araújo, G. Véras, E.C. da Silva, The successive projections algorithm for interval selection in PLS, Microchem. J. 110 (2013) 202–208.
[45] H.C. Goicoechea, A.C. Olivieri, A new family of genetic algorithms for wavelength interval selection in multivariate analytical spectroscopy, J. Chemom. 17 (2003) 338–345.
[46] M. Arakawa, Y. Yamashita, K. Funatsu, Genetic algorithm-based wavelength selection method for spectral calibration, J. Chemom. 25 (2011) 10–19.
[47] F. Allegrini, A.C. Olivieri, A new and efficient variable selection algorithm based on ant colony optimization. Applications to near infrared spectroscopy/partial least-squares analysis, Anal. Chim. Acta 699 (2011) 18–25.
[48] Z. Xiaobo, Z. Jiewen, H. Xingyi, L. Yanxiao, Use of FT-NIR spectrometry in non-invasive measurements of soluble solid contents (SSC) of 'Fuji'apple based on different PLS models, Chemom. Intell. Lab. 87 (2007) 43–51.
[49] Y.-H. Yun, D.-S. Cao, M.-L. Tan, J. Yan, D.-B. Ren, Q.-S. Xu, L. Yu, Y.-Z. Liang, A simple idea on applying large regression coefficient to improve the genetic algorithm-PLS for variable selection in multivariate calibration, Chemom. Intell. Lab. 130 (2014) 76–83.
[50] R.K.H. Galvão, M.C.U. Araujo, G.E. José, M.J.C. Pontes, E.C. Silva, T.C.B. Saldanha, A method for calibration and validation subset partitioning, Talanta 67 (2005) 736–740.
[51] R. Rinnan, Å. Rinnan, Application of near infrared reflectance (NIR) and fluorescence spectroscopy to analysis of microbiological and chemical properties of arctic soil, Soil Biol. Biochem. 39 (2007) 1664–1673.
[52] D.S. Cao, Y.Z. Liang, Q.S. Xu, H.D. Li, X. Chen, A new strategy of outlier detection for QSAR/QSPR, J. Comput. Chem. 31 (2010) 592–602.
[53] R. Leardi, L. Nørgaard, Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions, J. Chemom. 18 (2004) 486–497.
[54] W. Wang, Y. Yun, B. Deng, W. Fan, Y. Liang, Iteratively variable subset optimization for multivariate calibration, RSC Adv. 5 (2015) 95771–95780.
[55] G.-H. Fu, Q.-S. Xu, H.-D. Li, D.-S. Cao, Y.-Z. Liang, Elastic net grouping variable selection combined with partial least squares regression (EN-PLSR) for the analysis of strongly multi-collinear spectroscopic data, Appl. Spectrosc. 65 (2011) 402–408.
[56] E. Ben-Dor, Y. Inbar, Y. Chen, The reflectance spectra of organic matter in the visible near-infrared and short wave infrared region (400–2500 nm) during a controlled decomposition process ☆, Remote Sens. Environ. 61 (1997) 1–15.