Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 111 (2013) 31–36

Contents lists available at SciVerse ScienceDirect

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

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

An efficient method of wavelength interval selection based on random frog for multivariate spectral calibration

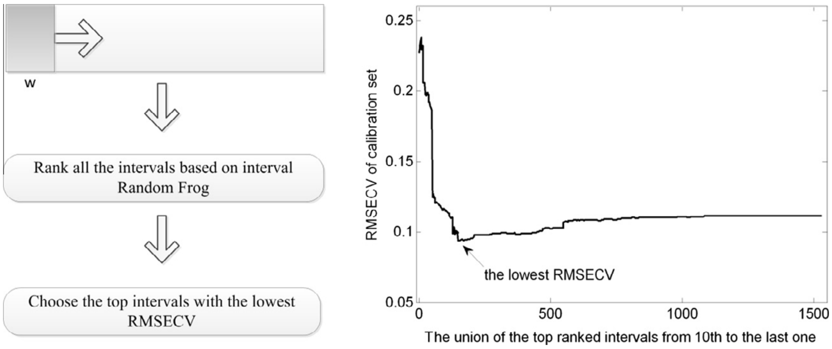
Yong-Huan Yun [a], Hong-Dong Li [a], Leslie R. E. Wood [a], Wei Fan [a], Jia-Jun Wang [b], Dong-Sheng Cao [a], Qing-Song Xu [c], Yi-Zeng Liang [a,*]

[a] College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, PR China
[b] Products Research Center, Honghe Cigarette Factory, Mile 652300, PR China
[c] School of Mathematics and Statistics, Central South University, Changsha 410083, PR China

HIGHLIGHTS

- Wavelength interval selection method: considers all the possible spectral intervals.
- A new idea: choose the best intervals from all the ranked overlapping intervals.
- More efficient compared with other wavelength interval selection methods.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 23 January 2013
Received in revised form 8 March 2013
Accepted 16 March 2013
Available online 28 March 2013

Keywords:
Random frog
Interval
Wavelength selection
Partial least squares
Near infrared spectra

ABSTRACT

Wavelength selection is a critical step for producing better prediction performance when applied to spectral data. Considering the fact that the vibrational and rotational spectra have continuous features of spectral bands, we propose a novel method of wavelength interval selection based on random frog, called interval random frog (iRF). To obtain all the possible continuous intervals, spectra are first divided into intervals by moving window of a fix width over the whole spectra. These overlapping intervals are ranked applying random frog coupled with PLS and the optimal ones are chosen. This method has been applied to two near-infrared spectral datasets displaying higher efficiency in wavelength interval selection than others. The source code of iRF can be freely downloaded for academy research at the website: http://code.google.com/p/multivariate-calibration/downloads/list.

Introduction

In recent years, multivariate calibration has been widely applied in vibrational and rotational spectral data such as infrared (IR), near infrared (NIR) and Raman spectroscopy [1,2]. The goal of multivariate calibration is to construct a predictive model, mostly linear calibration model, relating chemical measured variables like wavelengths to properties of interest like concentration values. With the advances in modern spectroscopic instrument, the expanded amounts of measured data are usually of high collinearity. To address this common problem, a variety of linear regression methods based on latent variables (LVs) have been developed, such as partial least squares (PLSs) [3] and principal component regression (PCR) [4]. Typically, these methods are usually used to carry out full-spectrum calibration due to a theoretical demonstration that the addition of spectral channels always improves the prediction performance under certain assumptions [5]. However,

[*] Corresponding author. Tel.: +86 731 8830824; fax: +86 731 8830831.
E-mail address: yizeng_liang@263.net (Y.-Z. Liang).

1386-1425/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.saa.2013.03.083

many papers have either theoretically or experimentally proved that it is very important and essential to conduct wavelength selection to gain better prediction performance [6–10]. The aim of wavelength selection is to select the informative wavelengths which are responsible for the property of interest. In other words, the removal of the uninformative and/or interfering variables contributes to construct a reliable and interpretable calibration model with good prediction accuracy. In addition, Zou et al. [11] summarized the importance of wavelength selection by means of chemical, physical and statistical basis.

So far, many methods of variable selection have been applied in multivariate calibration. These methods can be categorized into two classes: single wavelength selection and wavelength interval selection. During the past decades, a series of single wavelength selection methods have been proposed, such as uninformative variable elimination (UVE) [12,13], Monte Carlo based UVE (MC-UVE) [14], competitive adaptive reweighted sampling (CARS) [15,16], Latent projective graph (LPG) [17], influential variable (IV) [18], successive projection algorithm (SPA) [19], stepwise selection [20], genetic algorithm (GA) [21–26], simulated annealing (SA) [21], and PLS regression combined with sure independence screening (PLSSIS) [27]. The importance of individual wavelengths is calculated on the basis of the statistical features of the variables and regression model through some kind of criteria such as correlation coefficient, variable influence on projection, Akaike information criterion (AIC), and the mean squared error in prediction (MSEP). However, single wavelength selection methods are neither intuitive nor easy to interpret the selected variables corresponding to chemical property because they are selected independently. Also individual wavelengths are not robust to noise. Therefore, considering the fact that the vibrational and rotational spectra have continuous features of spectral bands, it is reasonable and interpretable to select spectral bands instead of scatter spectral points. For instance, the vibrational and rotational spectral band relating to chemical band generally has a width of 4–200 cm^{-1}. Many methods of wavelength interval selection have been developed following the idea of spectral band selection such as interval PLS (iPLS) [28], moving window PLS (MWPLS) [29], and as well as improvements made on them based on optimization algorithm [30–36]. The principle of iPLS consists of splitting the spectra into equal-width intervals, and developing sub-PLS models for each one. The sub-intervals with the lowest value of the root mean squared error of cross-validation (RMSECV) are chosen as the best. However, they are not the optimal ones. Many methods based on iPLS were developed to optimize the combination of the selected intervals, such as backward iPLS (biPLS) [33], and synergy iPLS (siPLS). The main advantage of this kind of method is that it uses a graphical display to focus on a choice of better sub-intervals and conduct comparison among the prediction performance of local models and the full-spectrum model. Instead of just testing a series of adjacent but nonoverlapping intervals, which would miss some more informative ones, MWPLS was proposed to overcome this drawback. It builds a series in a window that moves through the whole spectra and then chooses the informative intervals with low model complexity and low value of the sum of residuals. Because it considers all the possible continuous intervals, it can select all the possible informative intervals but not the optimized ones. Changeable size moving window partial least squares (CSMWPLSs) and searching combination moving window partial least squares (SCMWPLSs) [31] based on MWPLS were proposed to search for an optimized spectral interval and an optimized combination of spectral interval from informative intervals using a local optimized algorithm. Although the results have achieved some improvement, it is limited due to the use of local optimized algorithm not global optimized algorithm. In addition, when a high spectral resolution is used, e.g., 1 or 2 cm^{-1}, the many spectral points will make the calculation take a very long time. In addition, Balabin et al. [37] made a comprehensive comparison between different wavelength selection methods on biodiesel data, including the above two categories.

Considering the continuity of spectra and all the possible continuous spectral intervals, in this study, we propose a novel and efficient wavelength interval selection based on random frog [38], called interval random frog (iRF). Random frog is a reversible jump Markov Chain Monte Carlo (RJMCMC)-like algorithm that was originally proposed to apply into gene selection. It conducts a search in the model space through both fixed-dimensional and trans-dimensional moves between different models, and then a pseudo-MCMC chain is computed and used to calculate selection probability of each variable. Afterwards, variables can be selected in terms of the ranking of all variables.

Unlike the iPLS and iPLS-based methods, iRF considers all the possible continuous spectral intervals to find the possible informative ones, that is, spectra are first divided into sub-intervals by moving window of a fixed width over the whole spectra. These overlapping intervals are ranked applying random frog coupled with PLS, and the optimal ones are chosen. This approach is referred to a novel idea compared with other wavelength interval selection methods. The performance of iRF was tested on two near infrared spectra datasets. The results show that iRF is a more efficient wavelength interval selection method than other ones such as iPLS, biPLS, siPLS and MWPLS.

Theory and algorithms

Random frog coupled with PLS

Random frog is a mathematically simple and computationally efficient method that borrows the framework of reversible jump MCMC [39,40]. Random frog coupled with PLS means that PLS is used as a modeling method in random frog. Let $\mathbf{X}$, of size $n \times p$, denotes the spectral matrix consisting of n samples in rows and p variables (wavelengths or wavenumbers) in columns and $\mathbf{Y}$, of size $n \times 1$, denotes the property of interest like concentration values.

Before running the random frog method, five tuning parameters that are set to control its performance should be initialized.

(1) N : the number of iterations. It needs to be sufficiently large to achieve convergence depending on the number of variables.
(2) Q : the number of variables between 1 and p contained in the initialized variable set. Q had an impact on the iterative process only at the first time but had no significant effects on the overall performance of random frog.
(3) Θ : a factor controlling the variance of a normal distribution from which the number of variables is sampled to enter a candidate variable subset
(4) ω : the role of this parameter is explained in **Step 3** of the following section.
(5) η : a parameter, which is the upper bound of the probability for accepting a candidate variable subset $\mathbf{V}^*$ whose performance is not better than $\mathbf{V}_{0,}$ ranges from 0 to 1.

After initialization of parameters, random frog works in five steps:

Step 1: A variable subset $\mathbf{V}_0$ consisting of Q variables is initialized randomly.
Step 2: A random number Q^* is generated, which is the number of candidate variable subset $\mathbf{V}^*$, from a normal distribution Norm $(Q, \theta Q)$, where Q and θQ are the mean and standard deviation of this distribution, respectively.

Step 3: A candidate variable subset $\mathbf{V}^*$ is proposed based on Q^* variables. There are three possible situations: (1) If $Q^* = Q$, let $\mathbf{V}^* = \mathbf{V}_0$, (2) if $Q^* < Q$, a PLS model is first established using $\mathbf{V}_0$, and the regression coefficient of each variable in this PLS model is recorded and compared with each other. The $Q - Q^*$ variables related with the smallest absolute regression coefficients are deleted from $\mathbf{V}_0$. The rest Q^* variables constitute a candidate subset $\mathbf{V}^*$, (3) If $Q^* > Q$, a variable subset $\mathbf{S}$ with $\omega(Q^* - Q)$ variables randomly sampled from $\mathbf{V} - \mathbf{V}_0$ is produced. A PLS model is built using the combination of $\mathbf{V}_0$ and $\mathbf{S}$. Afterwards, the Q^* variables which have the largest absolute regression coefficients in this PLS model are retained and collected as a candidate subset $\mathbf{V}^*$.

Step 4: The acceptance of $\mathbf{V}^*$ is determined by computing the root mean squared error of cross-validation (RMSECV) using $\mathbf{V}_0$ and $\mathbf{V}^*$, respectively, obtaining RMSECV and RMSECV*. If RMSECV* $\leqslant$ RMSECV, accept $\mathbf{V}^*$ as $\mathbf{V}_1$. Otherwise accept $\mathbf{V}^*$ as $\mathbf{V}_1$ with probability ηRMSECV/RMSECV*. It is clear that RMSECV/RMSECV* < 1, so ηRMSECV/RMSECV* < η. Finally, $\mathbf{V}_0$ is updated using the variables in $\mathbf{V}_1$ to return to **Step 2**. This iteration is repeated until N loops have finished.

Step 5: A selection probability of each variable after N iterations is computed. The frequency of the jth variable, $j = 1,2,\ldots,p$, that has been selected in these N variable subsets is denoted as N_j. The selection probability of each variable can be calculated using Eq. (1). All the variables are ranked in terms of selection probability.

$$\text{Probability}_j = \frac{N_j}{N}, \quad j = 1,2,\ldots,p \tag{1}$$

Normal distribution is employed to control the number of variables implicitly and for the addition and deletion of variables, which provides the basis of model searching in a general model space. The criteria of absolute regression coefficient in the model used as a measure of variable importance guarantees that the more important a variable, the more likely it is to be selected.

Random frog is also viewed as an extension on model population analysis (MPA) [41–44] because it is based on analyzing a large amount of sub-models that are sampled from the model space.

Interval random frog (iRF)

The method of iRF proposed in this study is a wavelength interval selection method based on the framework of random frog PLS. Spectra are first divided into sub-intervals through moving window of a fixed width, denoted as w, over the whole spectra, which can obtain all the possible intervals. These overlapping intervals are regarded as 'variables' when applying random frog PLS. The union of members of different sub-intervals is used, when establishing a PLS model. Each candidate interval has w spectral points. Assess the intervals by the sum of the absolute regression coefficient of each spectral point. Except this, other approaches are the same as five steps in random frog. Finally, all intervals are ranked. In order to better understand this method, a graphic flowchart and simple example of iRF are briefly shown in Fig. 1.

Dataset and software

Tobacco dataset

The tobacco dataset [16] was obtained with the measurement of 300 samples by Nicolet Antaris FT-NIR spectrometer in transflective mode. Each spectrum consists of an average of 32 scans at intervals of 4 cm^{-1} within the wavenumbers range 10,000–4000 cm^{-1} (1557 spectral points). The total nicotine of the tobacco samples determined by continuous flow method was considered as the property of interest.

Milk dataset

The milk dataset consists of 67 milk samples, acquired directly from the local market in Changsha, China. The samples were measured using an Antaris II FT-NIR spectrometer (Thermo Fisher, USA) in transflective mode. Each spectrum consisted in an average of 32 scans at intervals of 4 cm^{-1} within the wavenumbers range 10,000–4000 cm^{-1} (1557 spectral points). The protein of milk was considered as the property of interest. Protein content was determined by Kjeldahl method as described by GB/T 5413.1-1997 (National Standards of PR China) and the factor 6.38 was used to convert the nitrogen values to protein. Rose-Gottlieb and Kjeldahl are the reference method for measuring the content of the property of interest. Of note, five outliers were detected and removed based on the Monte-Carlo outlier detection approach [45]. The dataset is available for readers on website: http://code.google.com/p/multivariate-calibration/downloads/list.

Software

A general-purpose computer with Inter Core i5 3.2 GHz CPU and 3 GB of RAM and Microsoft Windows XP operating system was used. All the calculations were performed by in MATLAB 2010b.

Results and discussion

The proposed iRF wavelength interval method was compared with others, including MWPLS, iPLS, biPLS and siPLS. Of note, siPLS is a greedy algorithm that investigates all the combination of predefined intervals to get the best intervals. In this study, three intervals were considered for siPLS. The datasets were first normalized to have zero mean. To better compare the prediction performances of these methods, an independent test set was used for our study. Furthermore, in order to ensure a uniform distribution of subsets, the dataset was divided into calibration set (80% of the dataset) and independent test set (20% of the dataset) on the basis of Kennard–Stone (KS) method [46]. The calibration set was used for wavelength selection and establishing the PLS model, while the independent test set was used for validation of the calibration model. For all of the calculations on these two datasets, the maximum number of latent variables was limited to 12. In model calibration, the optimal number of latent variables which was used for validation was determined by 10-fold cross-validation method with the lowest RMSECV. In addition, the performance of the model was assessed by the root mean square error of calibration (RMSEC) and the root mean square error of prediction (RMSEP)

Initialization of parameters in iRF

The method of iRF has six parameters to be initialized, including interval width w, N, Q, θ, ω and η. Because chemical bands generally have a width of 4–200 cm^{-1} in spectra and the intervals divided by iRF overlap greatly, an interval width w of 5–20 spectral points is should be set. As previously discussed, the larger the N is, the more likely iRF method is to select the best intervals but the higher the computational cost is. For these two datasets, according to our experience, N equal to 10,000 is sufficient. In regard to Q, it has an influence on the iterative process only at the first time but has no significant effects on the overall performance. Q was set to 50 so that the initialized set would contain more intervals. The other three parameters θ, ω and η, which do not have

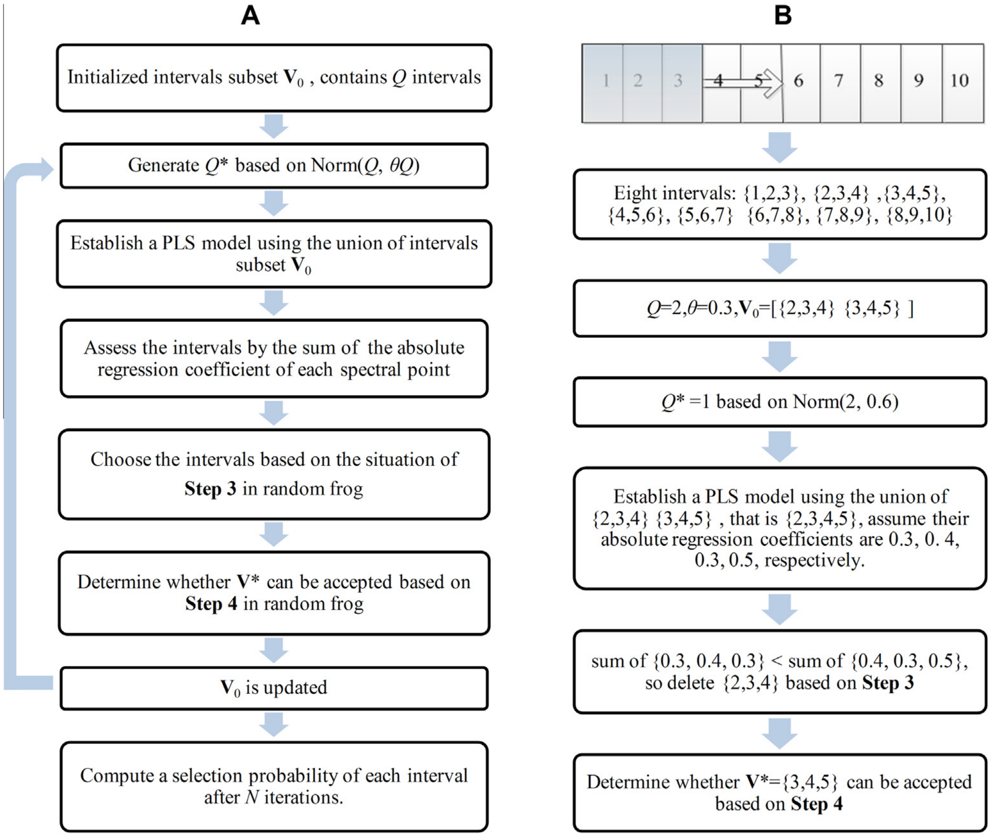

A
Initialized intervals subset V_0 , contains Q intervals
Generate Q* based on Norm(Q, θQ)
Establish a PLS model using the union of intervals subset V_0
Assess the intervals by the sum of the absolute regression coefficient of each spectral point
Choose the intervals based on the situation of Step 3 in random frog
Determine whether V* can be accepted based on Step 4 in random frog
V_0 is updated
Compute a selection probability of each interval after N iterations.
B
1 2 3 4 5 6 7 8 9 10
Eight intervals: {1,2,3}, {2,3,4} ,{3,4,5}, {4,5,6}, {5,6,7} {6,7,8}, {7,8,9}, {8,9,10}
Q=2,θ=0.3,V_0=[{2,3,4} {3,4,5}]
Q* =1 based on Norm(2, 0.6)
Establish a PLS model using the union of {2,3,4} {3,4,5} , that is {2,3,4,5}, assume their absolute regression coefficients are 0.3, 0. 4, 0.3, 0.5, respectively.
sum of {0.3, 0.4, 0.3} < sum of {0.4, 0.3, 0.5}, so delete {2,3,4} based on Step 3
Determine whether V*={3,4,5} can be accepted based on Step 4

Fig. 1. (A) a simple flowchart of iRF method; and (B) a simple example of iRF method.

significant effects on the results, were set to 0.3, 3 and 0.1, respectively.

Tobacco dataset

For this dataset, the *w* of iRF was set 20, and 1538 intervals that contained all possible intervals with 20 spectral points were obtained. In addition, the window size of MWPLS was set to 20, and the number of divisions was set to 40 for iPLS. What deserves special mention is that the intervals obtained by iRF and MWPLS were overlapping but the intervals divided by iPLS were nonoverlapping. Therefore, the conditions for iRF and MWPLS were similar to the ones for iPLS.

Regarding iRF, the selection probability of intervals cannot be reproduced due to the random sampling. To diminish the impact of this random factor, iRF was conducted 20 times and the average was used. Moreover, it is necessary to seek an optimal number of intervals from the ranked intervals. The top 20 ranked intervals are listed in Table 1. But the RMSECV of the union of the top 20 ranked intervals is not minimal. The RMSECV of the union of 10th to the last one (1538th) of the ranked intervals was computed (Fig. 2), so as to find the optimal number of intervals. And then the union of the optimal intervals was used for predicting the test set. We can see from Fig. 2 that the top 51 intervals are the optimal intervals with the lowest RMSECV on the calibration set.

The results of different methods are shown in Table 2. From Table 2, we can see that all the wavelength interval selection methods are superior to the full-spectrum PLS. Besides, iRF is significantly better than other wavelength interval selection methods with not only a lower RMSEC and RMSEP but also a smaller number of variables. All of the methods chose 12 as the optimal component of PLS based on 10-fold cross-validation. Four intervals were selected by iRF, which are located at 4000–4208 cm^{-1},

Table 1
The top 20 intervals (position in the dataset) for the tobacco dataset.

Rank	Intervals	Rank	Intervals	Rank	Intervals	Rank	Intervals
1	8–27	6	6–25	11	5–24	16	19–38
2	7–26	7	10–29	12	18–37	17	4–23
3	11–30	8	12–31	13	21–40	18	3–22
4	13–32	9	15–34	14	17–36	19	16–35
5	9–28	10	14–33	15	20–39	20	22–41

The union of the top 20 intervals is 3–41, that is 4007–4154 cm^{-1}.

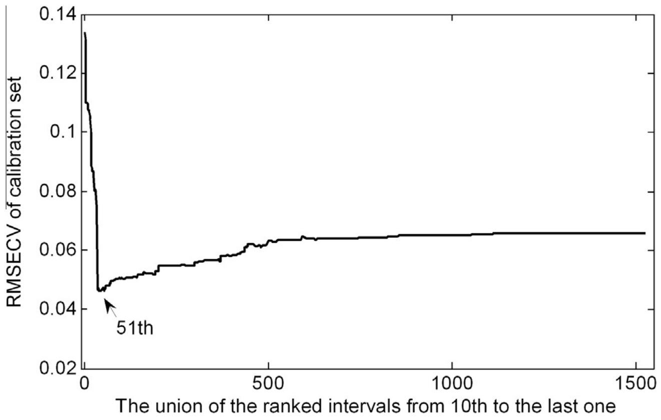


Fig. 2. The RMSECV of the union of the top ranked intervals from 10th to the last (1538th) on the tobacco dataset. The top 51 intervals are the optimal intervals with the lowest RMSECV on the calibration set.

Table 2
Results of different wavelength interval selection methods on the tobacco dataset.

Methods	Selected wavelength intervals (cm^{-1})	nVAR	nLVs	RMSEC	RMSEP
PLS	10,000–4000	1557	12	0.0595	0.0527
MWPLS	4000–4219, 4239–4655, 5612–5805, 5851–6067	275	12	0.0440	0.0473
iPLS	4000–4748, 5654–6102	312	12	0.0450	0.0485
biPLS	4000–4297, 4451–4597, 7455–7756	195	12	0.0409	0.0429
siPLS	4000–4297, 4451–4597	117	12	0.0400	0.0449
iRF	4000–4208, 4389–4474, 4586–4690, 5685–5770	129	12	0.0399	0.0419

4389–4474 cm^{-1}, 4586–4690 cm^{-1}, and 5685–5770 cm^{-1}. The selected intervals are shown in Fig. 3. The range of 4000–4208 cm^{-1} and 4389–4474 cm^{-1} are related to the combination of the fundamental stretching and bending vibrations of C—H/C—C, [47]. The interval of 4586–4690 cm^{-1} is ascribed to the second overtone of N—H bending, while the range of 5685–5770 cm^{-1} belong to the first overtone of C—H stretching of methyl. It should be pointed out that the intervals selected by MWPLS and iPLS contain most of the same ones selected by iRF, which indicates that the intervals selected by MWPLS and iPLS still have some redundant and uninformative variables. As for biPLS and siPLS, their results are not better than iRF due to the omission of the informative intervals approximately located at 5685–5770 cm^{-1}. Although biPLS and siPLS (consider all the possible combination of three intervals) are the improved version of iPLS to select the best combination of several intervals, they are still limited. Of course, the more intervals the spectra split, the better results can be obtained, but the longer the computational time is. For example, if siPLS considers the combination of five intervals with 40 divisions, it will lead to calculate 40*39*38*37*36/5! = 658, 008 times.

Milk dataset

Because the milk dataset has the same spectral points as the tobacco dataset, the procedure was conducted similarly. The *w* of iRF as well as the window size of MWPLS was set to 20, while the number of divisions was set to 40 for iPLS.

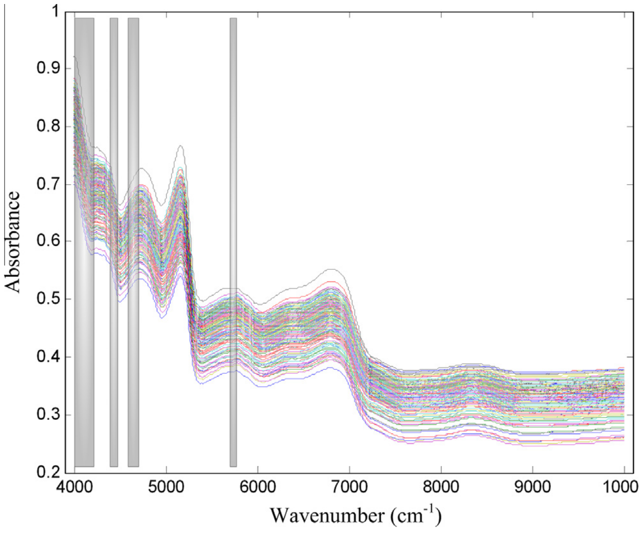


Fig. 3. The selected intervals on the tobacco dataset by iRF.

Table 3
Results of different wavelength interval selection methods on the milk dataset.

Methods	Selected wavelength intervals (cm^{-1})	nVAR	nLVs	RMSEC	RMSEP
PLS	10,000–4000	1557	12	0.0604	0.0642
MWPLS	4061–4370, 4451–4601, 4686–4864, 5488–5778	244	9	0.0546	0.0598
iPLS	4150–4898, 5504–5951	312	10	0.0664	0.0525
biPLS	4150–4748, 5203–5350, 7459–7756, 8512–8659, 8964–9110, 9264–9708	390	11	0.0438	0.0579
siPLS	4451–4597, 4752–4898, 6106–6252	117	6	0.0662	0.0551
iRF	4007–4239, 4258–4408, 4516–4682, 4725–4810, 5026–5315, 6025–6106, 8682–8763, 9260–9330, 9569–9673, 9928–10,000	357	10	0.0426	0.0555

Table 3 shows the comparison of different methods. Fig. 4 shows the RMSECV of the union of the top ranked intervals from 10th to the last (1538th). The top 165 intervals are the optimal intervals with the lowest RMSECV on the calibration set. From Table 3, one can see that all the wavelength interval selection methods are superior to the full-spectrum PLS on the basis of RMSEP. However, the RMSEC of iPLS and siPLS are even larger than the full-spectrum-PLS. As for the other three methods, both RMSEC and RMSEP of iRF are lower than MWPLS and biPLS, which indicates that iRF is indeed a more efficient wavelength interval method and the intervals selected by iRF are more informative. Ten intervals selected by iRF are 4007–4239 cm^{-1}, 4258–4408 cm^{-1}, 4516–4682 cm^{-1}, 4725–4810 cm^{-1}, 5026–5315 cm^{-1}, 6025–6106 cm^{-1}, 8682–8763 cm^{-1}, 9260–9330 cm^{-1}, 9569–9673 cm^{-1} and 9928–10,000 cm^{-1}, which are shown in Fig. 5. Furthermore, most of the selected intervals are associated with the chemical property. Intervals 4007–4239 cm^{-1} and 4258–4408 cm^{-1} are associated with the second overtone of secondary amine, while 4516–4682 cm^{-1} is related to the third overtone of C—H bending of —CH2 group. Interval 4725–4810 cm^{-1} is corresponding to C=O carbonyl stretch, second overtone of primary amide [48], and 6025–6106 cm^{-1} is corresponding to the first overtone of N—H stretch. Interval 9928–10,000 cm^{-1} is attributed to the second overtone of N—H stretch, but other methods did not select this spectral region. Noticeably, the intervals selected by biPLS contain most of the intervals selected by iRF except 9928–10,000 cm^{-1}. As

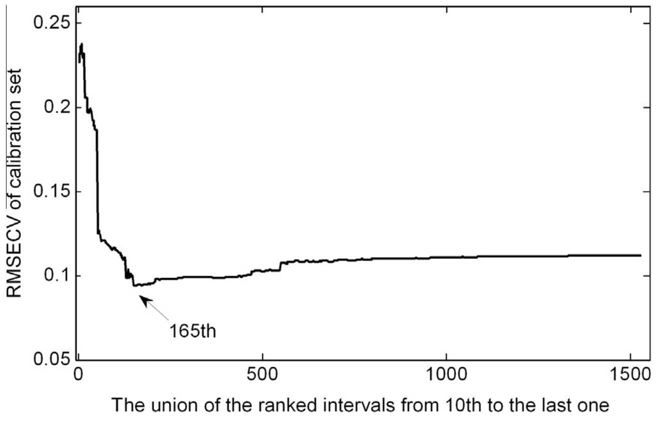


Fig. 4. The RMSECV of the union of the top ranked intervals from 10th to the last (1538th) on the milk dataset. The top 165 intervals are the optimal intervals with the lowest RMSECV on the calibration set.

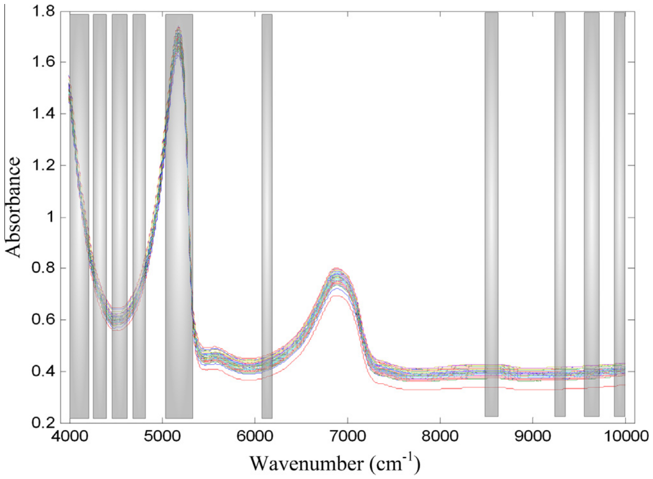


Fig. 5. The selected intervals on the milk dataset by iRF.

a consequence, the results from biPLS are not better than the ones from iRF. The region of 9928–10,000 cm^{-1} is also indicated as more informative. As for the left four unexplained intervals, we think that they were likely to be overlapped with chemical bands of other properties such as water and fat. For example, the range of 5026–5315 cm^{-1} is the absorption of water band. Therefore, interval of 5026–5315 cm^{-1} was most probably overlapped with water band completely.

Conclusions

A novel wavelength interval selection method based on the framework of random frog PLS, called interval random frog, was proposed and investigated using two NIR datasets. The results show that the proposed method selects more informative intervals and works better than other wavelength interval selection methods including MWPLS, iPLS, biPLS and siPLS. This method considers all possible spectral intervals and ranks all the intervals based on the absolute regression coefficient of PLS model. It can be said that iRF is an efficient method to be applied for spectral calibration. One of the drawbacks of iRF method is the low reproducibility as a result of random sampling. Hence, it is better to conduct many runs of iRF and use the average result.

The idea of considering all possible spectral intervals and using the ranked intervals is a new one and is useful for wavelength selection. For example, GA-PLS carries out variable selection based on the frequency of each variable. Therefore, we believe that it is worth applying it in the GA algorithm.

Acknowledgements

This work is financially supported by the National Nature Foundation Committee of PR China (Grant Nos. 21275164, 21075138 and 11271374). The studies meet with the approval of the university's review board.

References

[1] T. Hasegawa, in: J. Chalmers, P.R. Griffiths (Eds.), Handbook of Vibrational Spectroscopy, Wiley, Chichester, UK, 2001, p. 2293.
[2] B. Stuart, Infrared spectroscopy, in: Kirk–Othmer Encyclopedia of Chemical Technology, John Wiley & Sons Inc (2000).
[3] S. Wold, M. Sjöström, L. Eriksson, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.
[4] S. Wold, K. Esbensen, P. Geladi, Chemom. Intell. Lab. Syst 2 (1987) 37–52.
[5] A. Lorber, B.R. Kowalski, J. Chemom. 2 (1988) 67–79.
[6] J.H. Kalivas, N. Roberts, J.M. Sutter, Anal. Chem. 61 (1989) 2024–2030.
[7] C.H. Spiegelman, M.J. McShane, M.J. Goetz, M. Motamedi, Q.L. Yue, G.L. Coté, Anal. Chem 70 (1998) 35–44.
[8] L. Xu, I. Schechter, Anal. Chem 68 (1996) 2392–2400.
[9] D. Jouan-Rimbaud, B. Walczak, D.L. Massart, I.R. Last, K.A. Prebble, Anal. Chim. Acta 304 (1995) 285–295.
[10] J.H. Kalivas, Chemom. Intell. Lab. 37 (1997) 255–259.
[11] X. Zou, J. Zhao, M.J.W. Povey, M. Holmes, H. Mao, Anal. Chim. Acta 667 (2010) 14–32.
[12] V. Centner, D.-L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, C. Sterna, Anal. Chem 68 (1996) 3851–3858.
[13] W. Cai, Y. Li, X. Shao, Chemom. Intell. Lab 90 (2008) 188–194.
[14] Q.-J. Han, H.-L. Wu, C.-B. Cai, L. Xu, R.-Q. Yu, Anal. Chim. Acta 612 (2008) 121–125.
[15] H.-D. Li, Y.-Z. Liang, Q.-S. Xu, D.-S. Cao, Anal. Chim. Acta 648 (2009) 77–84.
[16] K. Zheng, Q. Li, J. Wang, J. Geng, P. Cao, T. Sui, X. Wang, Y. Du, Chemom. Intell. Lab 112 (2012) 48–54.
[17] X. Shao, G. Du, M. Jing, W. Cai, Chemom. Intell. Lab. Syst. 114 (2012) 44–49.
[18] X. Shao, M. Zhang, W. Cai, Anal. Methods 4 (2012) 467.
[19] M.C.U. Araújo, T.C.B. Saldanha, R.K.H. Galvão, T. Yoneyama, H.C. Chame, V. Visani, Chemom. Intell. Lab 57 (2001) 65–73.
[20] H. Martens, T. Naes, Multivariate Calibration, Wiley, New York, 1989.
[21] K. Sasaki, S. Kawata, S. Minami, Appl. Spectrosc. 40 (1986) 185–190.
[22] J. Yang, V. Honavar, IEEE Intell. Syst 13 (1998) 44–49.
[23] A.S. Bangalore, R.E. Shaffer, G.W. Small, M.A. Arnold, Anal. Chem. 68 (1996) 4200–4212.
[24] C.B. Lucasius, M.L.M. Beckers, G. Kateman, Anal. Chim. Acta 286 (1994) 135–153.
[25] R. Leardi, J. Chemom. 14 (2000) 643–655.
[26] J. Ghasemi, A. Niazi, R. Leardi, Talanta 59 (2003) 311–317.
[27] X. Huang, Q.-S. Xu, Y.-Z. Liang, Anal. Methods 4 (2012) 2815.
[28] L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Appl. Spectrosc. 54 (2000) 413–419.
[29] J.-H. Jiang, R.J. Berry, H.W. Siesler, Y. Ozaki, Anal. Chem. 74 (2002) 3555–3565.
[30] M. Arakawa, Y. Yamashita, K. Funatsu, J. Chemom. 25 (2011) 10–19.
[31] Y.P. Du, Y.Z. Liang, J.H. Jiang, R.J. Berry, Y. Ozaki, Anal. Chim. Acta 501 (2004) 183–191.
[32] S. Kasemsumran, Y.P. Du, K. Maruo, Y. Ozaki, Chemom. Intell. Lab. 82 (2006) 97–103.
[33] R. Leardi, L. Nørgaard, J. Chemom. 18 (2004) 486–497.
[34] A.G. Frenich, D. Jouan-Rimbaud, D.L. Massart, S. Kuttatharmmakul, M.M.n. Galera, J.L.M.n. Vidal, Analyst 120 (1995) 2787.
[35] Q. Chen, P. Jiang, J. Zhao, Spectrochim. Acta. A 76 (2010) 50–55.
[36] X. Zou, Z. Zhao, X. Huang, Y. Li, Chemom. Intell. Lab. 87 (2007) 43–51.
[37] R.M. Balabin, S.V. Smirnov, Anal. Chim. Acta 692 (2011) 63–72.
[38] H.-D. Li, Q.-S. Xu, Y.-Z. Liang, Anal. Chim. Acta 740 (2012) 20–26.
[39] P.J. Green, Biometrika 82 (1995) 711–732.
[40] H.F. Lopes, A Note on Reversible Jump Markov Chain Monte Carlo, Graduate School of Business, The University of Chicago, 2006.
[41] H.-D. Li, Y.-Z. Liang, D.-S. Cao, Q.-S. Xu, TRAC, Trends Anal. Chem. 38 (2012) 154–162.
[42] H.-D. Li, M.-M. Zeng, B.-B. Tan, Y.-Z. Liang, Q.-S. Xu, D.-S. Cao, Metabolomics 6 (2010) 353–361.
[43] H.-D. Li, Y.-Z. Liang, Q.-S. Xu, D.-S. Cao, J. Chemom. 24 (2010) 418–423.
[44] H.-D. Li, Y.-Z. Liang, Q.-S. Xu, D.-S. Cao, IEEE ACM. T. Comput. Bi. 8 (2011) 1633–1641.
[45] D.-S. Cao, Y.-Z. Liang, Q.-S. Xu, H.-D. Li, X. Chen, J. Comput. Chem. 31 (2010) 592–602.
[46] R.W. Kennard, L.A. Stone, Technometrics 11 (1969) 137–148.
[47] H. Xu, Z. Liu, W. Cai, X. Shao, Chemom. Intell. Lab. 97 (2009) 189–193.
[48] J.J. Workman, Appl. Spectrosc. Rev. 31 (1996) 251–320.