Analytica Chimica Acta 648 (2009) 77–84

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration

Hongdong Li [a], Yizeng Liang [a,*], Qingsong Xu [b], Dongsheng Cao [a]

[a] Research Center of Modernization of Traditional Chinese Medicines, College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, PR China
[b] School of Mathematic Sciences, Central South University, Changsha 410083, PR China

ARTICLE INFO

Article history:
Received 12 February 2009
Received in revised form 18 June 2009
Accepted 18 June 2009
Available online 24 June 2009

Keywords:
Wavelength selection
Monte Carlo
Adaptive reweighted sampling
Model sampling
Near infrared
Multivariate calibration

ABSTRACT

By employing the simple but effective principle 'survival of the fittest' on which Darwin's Evolution Theory is based, a novel strategy for selecting an optimal combination of key wavelengths of multi-component spectral data, named competitive adaptive reweighted sampling (CARS), is developed. Key wavelengths are defined as the wavelengths with large absolute coefficients in a multivariate linear regression model, such as partial least squares (PLS). In the present work, the absolute values of regression coefficients of PLS model are used as an index for evaluating the importance of each wavelength. Then, based on the importance level of each wavelength, CARS sequentially selects *N* subsets of wavelengths from N Monte Carlo (MC) sampling runs in an iterative and competitive manner. In each sampling run, a fixed ratio (*e.g.* 80%) of samples is first randomly selected to establish a calibration model. Next, based on the regression coefficients, a two-step procedure including exponentially decreasing function (EDF) based enforced wavelength selection and adaptive reweighted sampling (ARS) based competitive wavelength selection is adopted to select the key wavelengths. Finally, cross validation (CV) is applied to choose the subset with the lowest root mean square error of CV (RMSECV). The performance of the proposed procedure is evaluated using one simulated dataset together with one near infrared dataset of two properties. The results reveal an outstanding characteristic of CARS that it can usually locate an optimal combination of some key wavelengths which are interpretable to the chemical property of interest. Additionally, our study shows that better prediction is obtained by CARS when compared to full spectrum PLS modeling, Monte Carlo uninformative variable elimination (MC-UVE) and moving window partial least squares regression (MWPLSR).

1. Introduction

Multivariate calibration models have been gaining extensive applications in the analysis of multi-component spectroscopic data due to their potential to extract chemically meaningful information, *e.g.* structure-related wavelengths, from the over-determined systems. But the measured spectral data on the modern spectroscopic instrument, such as ultraviolet or near infrared instruments, are usually of high colinearity, which is the commonplace faced by analytical chemists. To address this problem, a variety of techniques based on latent variables (LVs) have been proposed, such as principal component regression (PCR) [1,2] and partial least squares (PLS) [3,4]. Typically, the establishment of a calibration model usually includes all the measured wavelengths. It is obvious that such a full spectrum model is sure to contain much redundant information, which will of course have negative influence on the prediction ability of the developed model. In addition, from the point of view of model interpretation, it is really difficult for analytical chemists and/or chemometrists to determine which wavelengths or combinations are responsible for the property of interest. It has been demonstrated that, both experimentally and theoretically, improvement of the performance of the calibration model can be achieved by using the selected informative wavelengths not the full spectrum.

Generally, the selection criteria for wavelength can be categorized into two groups [5]. One is based on information content of the wavelength, such as signal-to-noise ratio. The other is based on the statistics related to the model's performance, *e.g.* RMSECV. Gemperline reviewed the work in the area of wavelength selection [6]. From an optimization perspective, the wavelength selection can be viewed as an optimizing process which maximizes the prediction performance of the calibration model. Thus, it is natural to employ the optimization algorithm, which tries to seek a good combination of wavelengths, to implement wavelength selection using the criteria mentioned above as the objection function. Genetic algorithm (GA) [5,7–15], simplex optimization [16], branch and bound

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

0003-2670/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2009.06.046

combination optimization [17,18], simulated annealing (SA) [16,19], and ant colony optimization (ACO) [20] have been applied to select the optimal subset of wavelengths. All these studies suggest that better prediction can be obtained using the selected wavelengths rather than the full spectrum, which is an indication of the importance of wavelength selection. But one should know that this kind of methods based on optimization methods is usually computationally intensive and sensible to the initialized solution.

Besides, a series of more direct methods have been proposed to conduct wavelength selection, such as iterative partial least squares (iPLS) [21], uninformative variable elimination (UVE) [22], Monte Carlo based UVE (MC-UVE) [23,24], moving window partial least squares (MWPLS) [25], successive projection [26,27], Bayesian linear regression (BLR) [28] and so on.

In essence, the developed wavelength-reduced model by wavelength selection is much more interpretable for the sake of some scientific insight into the relationship between digitalized spectra and the property to be investigated, *e.g.* concentration. The underlying assumption behind wavelength selection may be that the regression model will be biased from the 'true' one due to the distortion caused by the wavelengths which are irrelevant with respect to the property under investigation. Based on the reports [5–25,28–33], one can conclude that wavelength selection is a key factor for constructing a reliable and interpretable calibration model with good prediction accuracy.

In this study, we present a new strategy, termed competitive adaptive reweighted sampling (CARS), which has the potential to select an optimal combination of the wavelengths existing in the full spectrum coupled with partial least squares regression by using the simple but effective principle 'survival of the fittest' on which Darwin's Evolution Theory is based. With applications to one simulated dataset and one real NIR spectral dataset of two properties, CARS proves to be a promising procedure to conduct wavelength selection for building a high performance calibration model. Additionally, it should be pointed out that CARS is not designed for spectral data only. It is a general strategy and thus can be used for variable selection of other kinds of data, such as genomic, proteomic and metabolomic data. Moreover, it can also be coupled with discriminant analysis for biomarker discovery.

2. Theory and algorithms

2.1. Notation

The data matrix **X** contains m samples in rows and p variables in columns. Vector **y** with order $m \times 1$ denotes the measured property of interest. The superscript T denotes vector or matrix transpose. When modeling, both **X** and **y** are mean-centered.

Suppose the number of MC sampling runs of CARS is set to N. With this setting, CARS will sequentially select N subsets of wavelengths. Briefly speaking, in each sampling run, CARS works in four successive steps: (1) Monte Carlo for model sampling. (2) Employ EDF to perform enforced wavelength selection. (3) Adopt ARS to realize a competitive selection of wavelengths and (4) cross validation [34–37] is utilized to evaluate the subset. CARS will be discussed in great detail in the following sections.

2.2. Monte Carlo for model sampling

Like uninformative variable elimination [22,23], in each sampling run of CARS, a PLS model is built using the randomly selected samples (usually 80–90% of the calibration set) not all the samples in the calibration set. From the point of view of sampling, this process can be regarded as sampling in the model space combined with Monte Carlo strategy. We are intended to select the variables which are of high adaptability regardless of the variation of training samples.

2.3. PLS and weights of variables

PLS is a widely used procedure for modeling the linear relationship between **X** and **y** based on latent variables (LVs). Suppose that the scores matrix is denoted by **T**, which is a linear combination of **X** with **W** as combination coefficients [38], and **c** is the regression coefficient vector of **y** against **T** by least squares. Thus we have the following formula:

$$\boldsymbol{T} = \boldsymbol{XW} \tag{1}$$

$$\boldsymbol{y} = \boldsymbol{Tc} + \boldsymbol{e} = \boldsymbol{XWc} + \boldsymbol{e} = \boldsymbol{Xb} + \boldsymbol{e} \tag{2}$$

where **e** is the prediction error and $\mathbf{b} = \mathbf{Wc} = [b_1, b_2, \ldots b_p]^{\mathrm{T}}$ is the p-dimensional coefficient vector. The absolute value of the ith element in **b**, denoted $|b_i|$ $(1 \le i \le p)$ reflects the ith wavelength's contribution to **y**. Thus, it is natural to say that the larger $|b_i|$ is, the more important the ith variable is. For evaluating the importance of each wavelength, we define a normalized weight as:

$$\boldsymbol{w}_i = \frac{|b_i|}{\sum_{i=1}^{p} |b_i|}, i = 1,\ 2,\ 3, \ldots,\ p \tag{3}$$

Additional attention should be paid to that the weights of the eliminated wavelengths by CARS are set to zero manually so that the weight vector **w** is always p-dimensional.

2.4. Exponentially decreasing function

Suppose the full spectrum contains p wavelengths and N sampling runs are performed in CARS. As mentioned before, the wavelength selection in CARS consists of two steps. In the first step, EDF is utilized to remove the wavelengths which are of relatively small absolute regression coefficients by force. In the ith sampling run, the ratio of wavelengths to be kept is computed using an EDF defined as:

$$r_i = ae^{-ki} \tag{4}$$

where a and k are two constants determined by the following two conditions: (I) in the first sampling run, all the p wavelengths are taken for modeling which means that $r_1 = 1$, (II) in the Nth sampling run, only two wavelengths are reserved such that we have $r_N = 2/p$. With the two conditions, a and k can be calculated as:

$$a = \left(\frac{p}{2}\right)^{1/(N-1)} \tag{5}$$

$$k = \frac{\ln(p/2)}{N-1} \tag{6}$$

where ln denotes the natural logarithm.

Fig. 1 illustrates an example of EDF. As can be seen clearly, the process of wavelength reduction can be roughly divided into two stages. In the first stage, wavelengths are eliminated rapidly which performs a 'fast selection', whereas in the second stage, wavelengths are reduced in a very gentle manner, which is instead called a 'refined selection' stage in our study. Therefore, wavelengths of little or no information in a full spectrum can be removed in a stepwise and efficient way because of the advantage of EDF. That is the reason why we choose EDF. Its advantage will be demonstrated by our experiments in the following sections.

2.5. Adaptive reweighted sampling

Following EDF-based enforced wavelength reduction, adaptive reweighted sampling (ARS) is employed in CARS to further eliminate wavelengths in a competitive way. This step mimics the

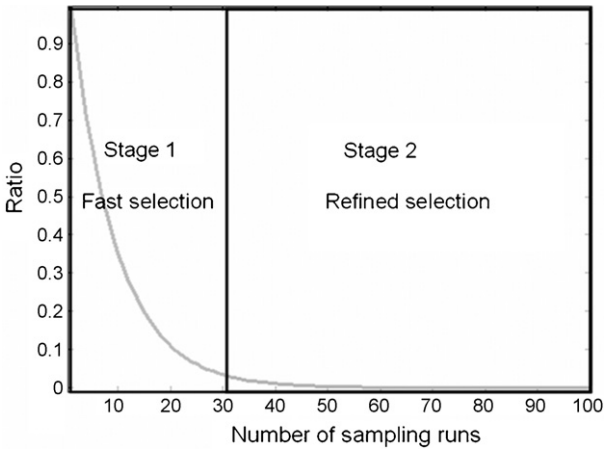

Fig. 1. Graphical illustration of the exponentially decreasing function. In the first stage, the number of the wavelengths is reduced fast while in the second stage, it decreases very slowly which realizes a refined selection.

'survival of the fittest' principle which is the basis of Darwin's Evolution Theory. Fig. 2 illustrates the meaning of adaptive reweighted sampling. Assume that we have five weighted variables which will be subjected to five random weighted sampling experiments with replacement. In Case 1, each variable has an equal weight 0.20 indicating that they can be sampled with an equal probability. The ideal result is that each variable is sampled one time. Case 2 shows variables 1 and 2 have the largest weight 0.30 while variables 4 and 5 are of the smallest weights 0.10. Thus, variables 1 and 2 are sampled twice, while variable 3 once. Variables 4 and 5 are not sampled by ARS and hence eliminated. Similar to Case 2, Case 3 demonstrates that only variables 1 and 3 are sampled in the five weighted sampling experiments due to their dominant weights, while variables 2, 4 and 5 are much less competitive and hence out of play because of their relatively weak weights.

2.6. General description of CARS

Fig. 3 shows the scheme of CARS algorithm. It is outlined clearly in Fig. 3 that CARS selects *N* subsets of variables by *N* sampling runs in an iterative manner and finally chooses the subset with the lowest RMSECV value as the optimal subset. In each sampling run, CARS works in four successive steps including Monte Carlo model sampling, enforced wavelength reduction by EDF, competitive wavelength reduction by ARS and RMSECV calculation for each subset. Of these, EDF-based wavelength reduction in combination with competitive wavelength reduction by ARS is a two-step procedure for wavelength selection. In summary, CARS employs a simple but effective principle 'survival of the fittest' and realizes to some extent the selection of an optimal subset of wavelength. In the following sections, the characteristics and

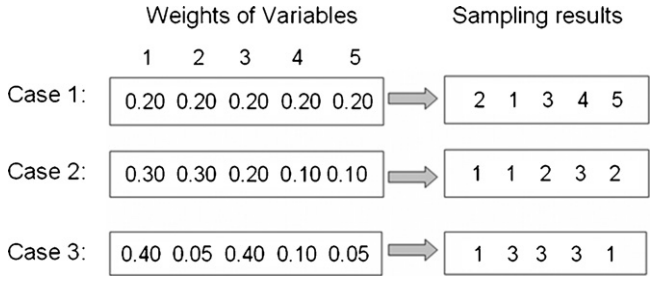

Fig. 2. Illustration of adaptive reweighted sampling technique using five variables in three cases as an example. The variables with larger weights will be selected with higher frequency.

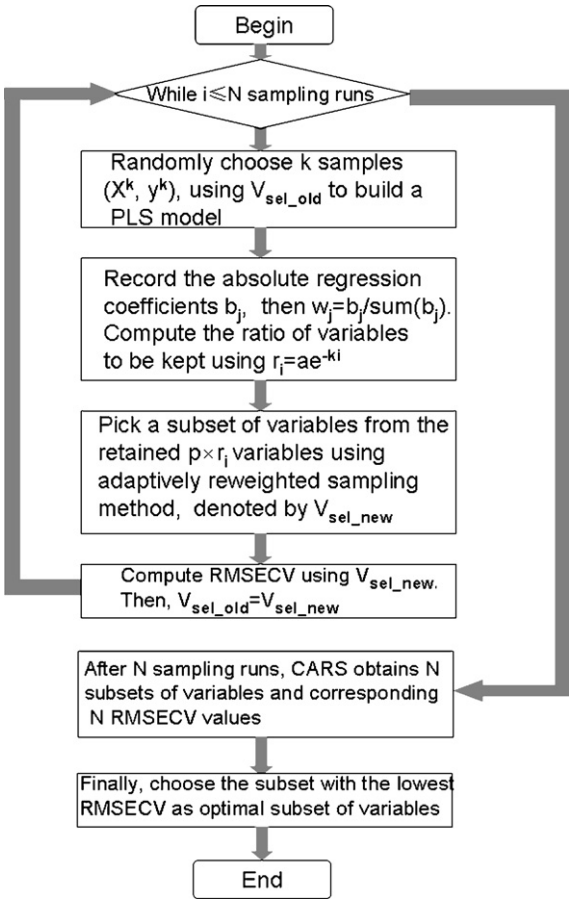

Fig. 3. Flow chart of CARS algorithm. When $i = 1$, all the variables are included to build a calibration model. Thus in this step, V_{sel_old} contains all the original variables. After *N* sampling runs, CARS obtains *N* subsets of variables and finally choose the subset with the lowest RMSECV value as the optimal one.

behaviors of CARS will be discussed in detail using one simulated dataset and one real world benchmark NIR dataset with two properties.

3. Data description

3.1. Simulated data set

This dataset, called SIMUIN, is simulated in the same way as in Ref. [22] which contains exactly five latent variables. The yielded relative eigenvalues by principal component analysis on the centered data are (%) 25.34, 23.02, 22.59, 21.49 and 7.57. SIMUIN consists of 25 samples in rows and 200 wavelengths in columns. The first 100 wavelengths are linearly related with **y** but the last 100 columns contain random numbers from 0 to 1, standing for uninformative wavelengths. The added noises are normally distributed in the range from 0 to 0.005.

3.2. Corn data set

This benchmark data set [39] consists of NIR spectra of 80 corn samples, measured on different types of NIR spectrometer. Each spectrum contains 700 data points measured in the wavelength range 2498–1100 nm at 2 nm intervals. In the present study, two sub-datasets are employed to investigate the performance of CARS. The first dataset uses the NIR spectra of 80 corn samples measured on *m5* instrument as **X** and the moisture value as dependent variable **y**. For the second dataset, we use the NIR spectra of 80 corn

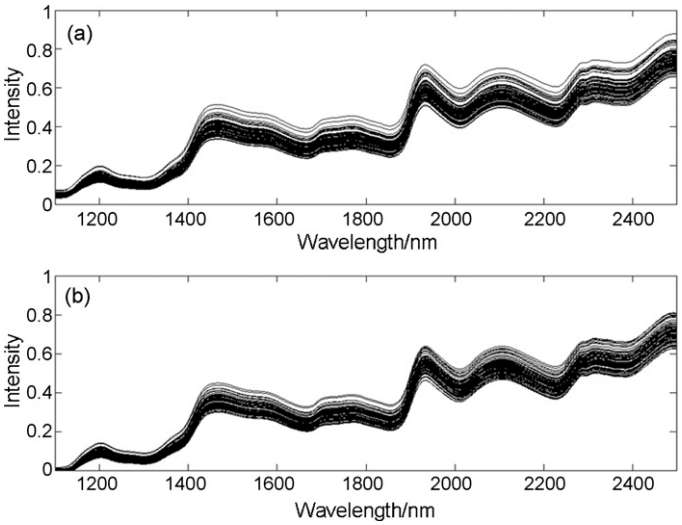

Fig. 4. The original NIR spectra of corn moisture (plot a) and corn protein data (plot b).

samples collected on *mp5* instrument as **X** and the protein content as the response variable **y**. The original spectra of the two data are shown in plots a and b of Fig. 4, respectively.

4. Results and discussion

4.1. Influence of number of MC sampling runs

In order to investigate the influence of the number of Monte Carlo sampling runs on CARS' performance, we have considered the following four cases: the number is set to 50, 100, 200 and 500. For each case and each of the three datasets, 50 replicate running of CARS is executed and RMSECV values are recorded. The resulted statistical box-plots are shown in Fig. 5. It can be found that the number of Monte Carlo sampling runs does not have significant influence on the performance of CARS. In the following sections, it is set to 100 as default.

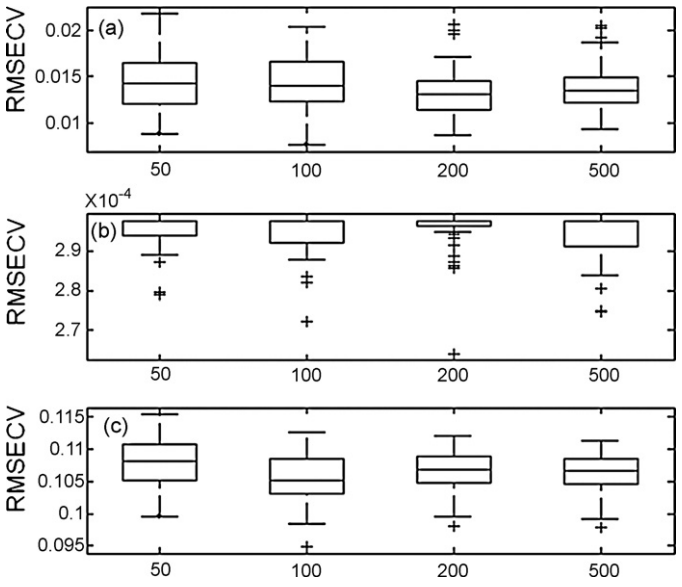

Fig. 5. The box-plots for each dataset with the number of Monte Carlo sampling runs of CARS set to 50, 100, 200 and 500, respectively. (a) Simulated dataset. (b) Corn moisture data. (c) Corn protein data.

Table 1
The results on the simulated dataset.

Methods	RMSECV	nLVs[a]	nVAR[a]	nUNV[a]
PLS[b]	1.101	7	200	–
PLS[c]	0.0200	5	100	–
MC-UVE-PLS	0.0209 ± 0.0006[d]	6 ± 1[d]	46 ± 20[d]	1 (235)
CARS-PLS	0.0139 ± 0.0023[d]	6 ± 1[d]	16 ± 4[d]	1 (1)

[a] nUNV stands for the number of selected different uninformative variables. The number in the bracket denotes the total times. nLVs and nVAR denotes the number of latent variables and selected variables, respectively.
[b] Results using full spectrum with 200 variables by PLS.
[c] Results using only the 100 simulated informative variables by PLS.
[d] Statistical results with the form mean value ± standard deviation from 500 replicate simulations.

4.2. Simulated data

This dataset is intended for investigating the ability for CARS to select key variables by eliminating the artificial noisy variables. 10-fold cross validation is used in this study to explore its predictive performance. Also, we compared CARS to MC-UVE, aiming only at demonstrating that CARS is indeed an alternative and efficient procedure for uninformative variable elimination not that which method is better.

This data is first autoscaled for each variable to have zero mean and unit variance before modeling. By 10-fold cross validation, the optimal number of latent variables of PLS model is 7. For MC-UVE, the number of Monte Carlo iterations is set to 500, and in each iteration 80% samples from this data are randomly chosen to build a PLS calibration model using seven latent variables. The regression coefficients for each variable are recorded in a vector. After 500 iterations, a coefficient matrix is obtained based on which a reliability index can be calculated for each variable. Then, all the variables are ranked in accordance with their reliability index. As known, cross validation is an effective and widely used technique for model/variable selection. Thus in our study, the number of variables to be selected is determined by 10-fold cross validation technique not by setting a cut-off value as done in Refs. [22,23]. Also the maximal number of selected variables is set to 100. With these settings, we run MC-UVE to eliminate the uninformative variables while simultaneously estimate its predictive performance. Further, it is noteworthy that only one running of MC-UVE is not sufficient due to the variation caused by Monte Carlo strategy. One remedy for this problem is to repeat it for many times. Therefore, MC-UVE is repeated 500 times in this case, which can help to get a deeper understanding of its behavior. For CARS, the number of MC sampling runs is set to 100. CARS is also rerun for 500 times and the results are recorded for further analysis.

Table 1 shows the results of MC-UVE and CARS on SIMUIN data, together with the results based on the full spectrum and only the informative variables. The RMSECV value using all the 200 hundred variables is 1.1010. By contrast, not only the RMSECV (=0.0200) but also the number of latent variables is reduced significantly when the model only includes the subset of the 100 informative variables. This phenomenon experimentally proves the necessity to perform variable selection or removing the uninformative variables before building a calibration model.

MC-UVE and CARS are applied in order to demonstrate whether better prediction can be obtained by selecting the reliable variable (MC-UVE) or key variables (CARS). From Table 1, one can find that CARS got much better prediction results, *i.e.* 0.0139 compared to 0.0209, but with a larger standard deviation (0.0023 compared to 0.0006), which indicates that the stability of CARS still needs improving although it can pick out variables leading to a model with good generalization performance. Interestingly, the number of the selected variables by CARS is relatively small (16 ± 4), which is one

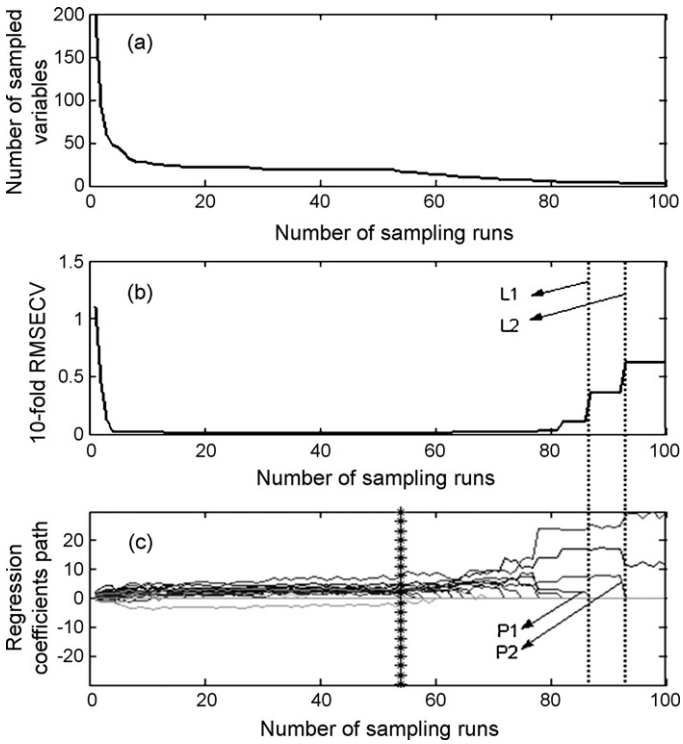

Fig. 6. The changing trend of the number of sampled variables (plot a), 10-fold RMSECV values (plot b) and regression coefficients of each variables (plot c) with the increasing of sampling runs. The line (marked by asterisk) denotes the optimal point where 10-fold RMSECV values achieve the lowest.

reason why we call them key variables. Moreover, only one uninformative variable is selected one time by CARS, which proves that it has the potential to eliminate uninformative variables as MC-UVE does.

Fig. 6 shows the changing trend of the number of sampled variables (plot a), 10-fold RMSECV values (plot b) and the regression coefficient path of each variable (plot c) with the increasing of sampling runs from one CARS running. As expected, the number of sampled variables decreases fast at the first stage of EDF and then very slowly at the second stage of EDF, which demonstrated that the proposed two phase selection, *i.e.* fast selection and refined selection, are indeed realized in CARS. The RMSECV values first descend quickly from sampling runs 1–10 which should be ascribed to the elimination of uninformative variables, then changes in a gentle way from sampling runs 20–60 corresponding to the phase that the sampled variables do not change obviously, and finally increase fast because of the loss of information caused by eliminating some key variables from the optimal subset (denoted by asterisk).

Also noteworthy is the coefficient path of each variable shown in plot c. Each line in plot c records the coefficients at different sampling runs for each variable. Thus, a subset of variables together with the regression coefficients can be extracted from each sampling run. The best subset with the lowest RMSECV value is marked by the vertical line denoted by asterisk. More interestingly, the RMSECV value jumps up to a higher stage at the sampling point (denoted dot line: L1), because the coefficient of one variable (denoted by P1) drops to zero just at the same time. The dot line marked by L2 is also the case when the coefficient of another variable denoted by P2 drops to zero. Such observations demonstrate the existence of key variables without which the model's performance would be reduced dramatically. That is why they are called key variables.

In general, this simulation study indicates that CARS is a promising method for variable selection. Wavelength selection for NIR spectral data will be discussed with great detail in the following.

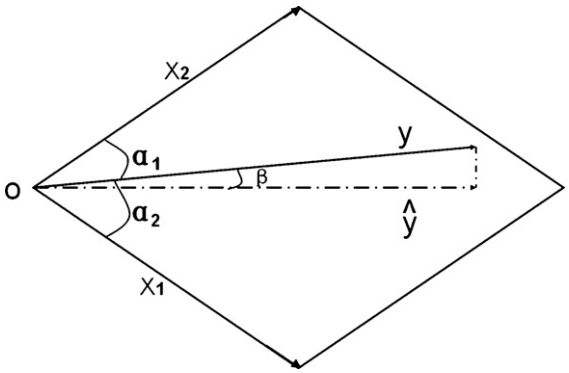

Fig. 7. As illustrated, α_1 denotes the angle between $\mathbf{X}_1$ and $\mathbf{y}$. α_2 denotes the angle between $\mathbf{X}_2$ and $\mathbf{y}$. β denotes the angle between $\mathbf{y}$ and its projection on the space spanned by $\mathbf{X}_1$ and $\mathbf{X}_2$. β is very small. The condition $\alpha_2 \gg \beta$ and $\alpha_2 \gg \beta$ holds in this case.

4.3. Corn moisture data

This NIR data is employed to specially address the situation that much better prediction results can only be obtained by combination of some variables, although each single variable is of relatively low correlation coefficient with $\mathbf{y}$. To search such a combination is an N-P hard problem and thus computationally infeasible.

Fig. 7 shows such a case. Both $\mathbf{x}_1$ and $\mathbf{x}_2$ are lowly correlated with $\mathbf{y}$. But $\mathbf{y}$ is so close to the subspace spanned by $\mathbf{x}_1$ and $\mathbf{x}_2$. The variable selection methods proposed by statisticians, such as forward stagewise selection [40], Lasso [41,42] and least angle regression [43], pick the most correlated variable with $\mathbf{y}$ at the first step in a greedy manner, which may cause the problem that some good combination might be missed. But in spectral data analysis, what interests analytical chemists is not the most correlated wavelengths but the chemically meaningful band or combinations of several bands.

In addition, a brief introduction of MWPLS is given for further proceeding. MWPLS is a wavelength interval selection procedure for multi-component spectral analysis. It establishes a PLS calibration model for each window (a continuous wavelength band) with a given number of latent variables. Then by moving the window on the whole measured wavelength region and changing the number of latent variables, a series of PLS models together with sums of squared residues (SSR) are calculated. Finally, the SSR is plotted versus the position of the moving window. Based on the obtained SSR plot, the wavelength interval with small SSR and fewer LVs are selected to build the final calibration model.

Fig. 8 depicts the wavelengths selected by UVE (plot a), MWPLS (plot b) and CARS (plot c), respectively. From plot a, one can see that two chemically meaningful wavelength bands 1894–1922 nm (Band 1) and 2098–2122 nm (Band 2), which are corresponding to the water absorption [12] and the combination of O–H bond [25], are selected by UVE. By contrast, the region around 1410 nm due to the first tone of O–H stretching mode leads to the minimal root mean squared errors of calibration (RMSEC) by MWPLS, while Band 1 and Band 2 are missed. When CARS is applied, only two wavelengths, *i.e.* 1908 and 2108 nm are picked out. It is noteworthy that the wavelength 1908 nm just belongs to Band 1 while 2108 nm to Band 2.

Table 2 shows the results of different methods or wavelength regions. The RMSECV values using Band 1 and Band 2 are 0.2394 (Q^2 = 0.5988, four latent variables) and 0.2747 (Q^2 = 0.4719, four latent variables), respectively. But it dramatically decreases to 0.0058 (Q^2 = 0.9998, four latent variables) when modeling by PLS using the combination of Band 1 and Band 2. This is one typical real world case as illustrated in Fig. 7. This phenomenon is an indication that combination of 1908 and 2108 nm has the most interpretability

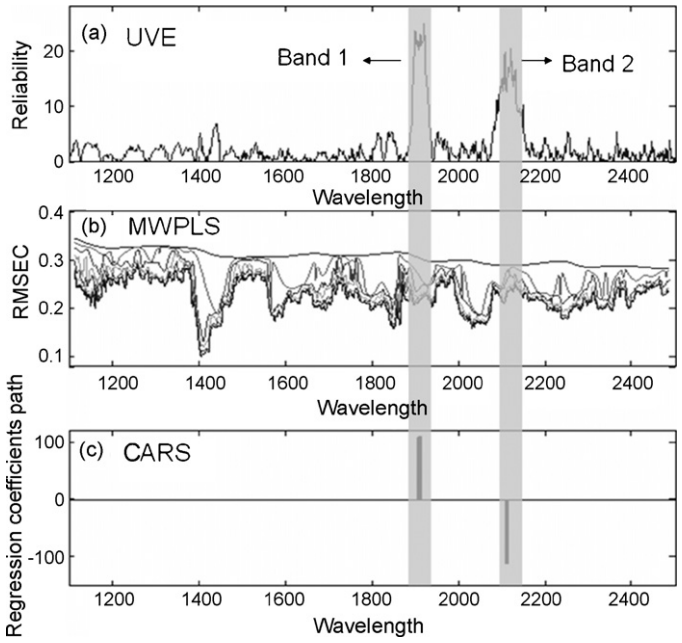

Fig. 8. Comparison of selected wavelengths by MC-UVE, MWPLS and CARS. The window size of MWPLS is fixed at 15. The iteration number of MC-UVE is 500 and the number of sampling runs of CARS is 500.

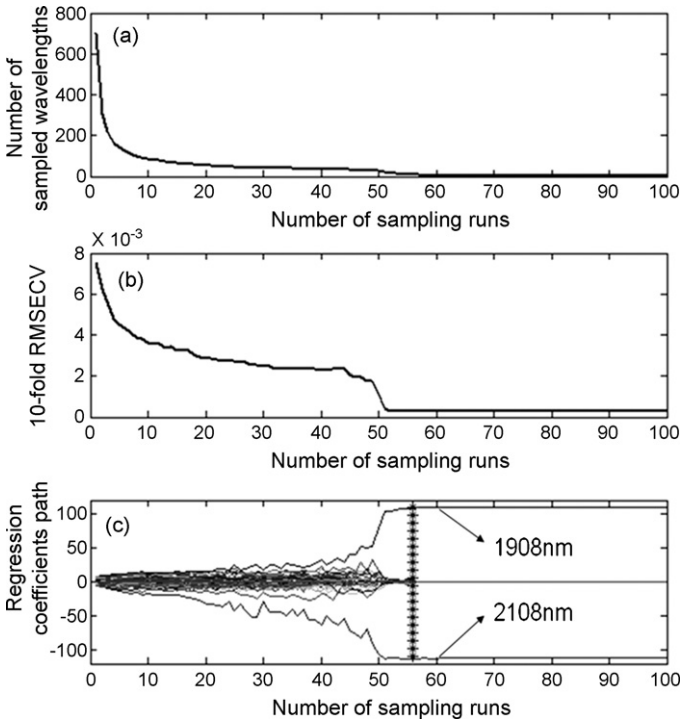

Fig. 9. Plots a and b show the changing of the number of sampled wavelengths and 10-fold RMSECV values. Plot c records the regression coefficient path of each wavelength. The vertical asterisk line denotes the optimal point where 10-fold CV values achieve the lowest.

for water content, from the point of view of either RMSECV or band assignment to chemical bond. As known, MWPLS is a procedure which takes a series of size-changing moving windows to identify and select a local wavelength band or several separate local bands in terms of the residuals and the number of latent variables. Thus it can only work well if the meaningful wavelength band exists in a narrow region. But for this case, Band 1 and Band 2 are so far away from each other that their combination cannot be detected by MWPLS. The results prove that MWPLS cannot deal with this situation well.

As mentioned before, both MC-UVE and CARS adopt Monte Carlo strategy to perform wavelength selection. Therefore, it is necessary to run the programmes many times to obtain statistically stable results. In our study, we run MC-UVE and CARS programmes 500 times, respectively. Both the mean and standard deviation are given in Table 2. The results demonstrate that better prediction is obtained by CARS combined with PLS. Moreover, the number of both latent variables and the selected wavelengths are significantly lower, which may be seen as a proof for Occam Razor Theory [44,45]. The reason why better prediction can be achieved using fewer wavelengths may be that wavelengths are heavily collinear and hence the model's variance can be reduced with fewer wavelengths. More interestingly, for each run of CARS, both the wavelength 1908 and 2108 nm are selected. Therefore, for this data, one can treat 1908 and 2108 nm, of very large absolute regression coefficients in calibration model, as the key wavelengths in terms of the selection of CARS.

Table 2
The results on corn moisture data.

Methods	RMSECV	nLVs	nVAR
PLS[a]	0.0229	10	700
PLS[b]	0.2394	4	15
PLS[c]	0.2747	4	13
PLS[d]	0.0058	4	28
MC-UVE-PLS	0.0032 ± 0.0004	10 ± 0	55 ± 6
MWPLS[e]	0.0383	10	119
CARS-PLS	0.0006 ± 0.0008	3 ± 2	3 ± 3

[a] Results using full spectrum in the range 1100–2498 nm.
[b] Results using the range 1894–1922 nm (Band 1, in Fig. 7).
[c] Results using the range 2098–2122 nm (Band 2, in Fig. 7).
[d] Results using the combination of 1894–1922 and 2098–2122 nm (Band 1 + Band 2, in Fig. 7).
[e] Results from the combination of four regions 1378–1438, 1558–1598, 1828–1868 and 1988–2078 nm.

Fig. 9c shows the regression coefficient path of each wavelength from one execution of CARS with the number of sampling runs set to 100. It can be seen in the first sampling run, that the absolute value of regression coefficient of each wavelength is very small. But with the number of sampling runs increased, the coefficients of some wavelengths get larger and larger while others become smaller and smaller. Specially, the coefficients even drop to zero if the corresponding wavelengths are eliminated by CARS because of their incompetence. Thus, the larger the absolute coefficient is, the more probable the corresponding wavelength can survive. This selection mechanism in CARS is somewhat like 'survival of the fittest' in Darwin's Evolution Theory. Each wavelength can be treated as an individual, and all the other wavelengths are naturally seen as its 'environment'. Based on this, CARS algorithm realizes the process of selecting the fittest individual by utilizing adaptive reweighted sampling technique. As Fig. 9c shows, the coefficients of wavelength 1908 and 2108 nm grow up first slowly, then quickly and finally reach the maximal absolute values above 100 (multiple runs of CARS lead to similar results, data not shown). These two wavelengths thus can be considered to be key wavelengths for this data. The optimal subset chosen by CARS can be extracted from the position denoted by the vertical asterisk line corresponding to the minimal 10-fold RMSECV value.

Further, we also statistically compute the selected frequency of each wavelength by running CARS 500 times. The result is shown in Fig. 10a. From Fig. 10a, one can find that the wavelengths 1908 and 2108 nm are not selected by chance because the frequencies of both are selected 500, which further prove that these two wavelengths are key to the calibration model. Generally, CARS can select an opti-

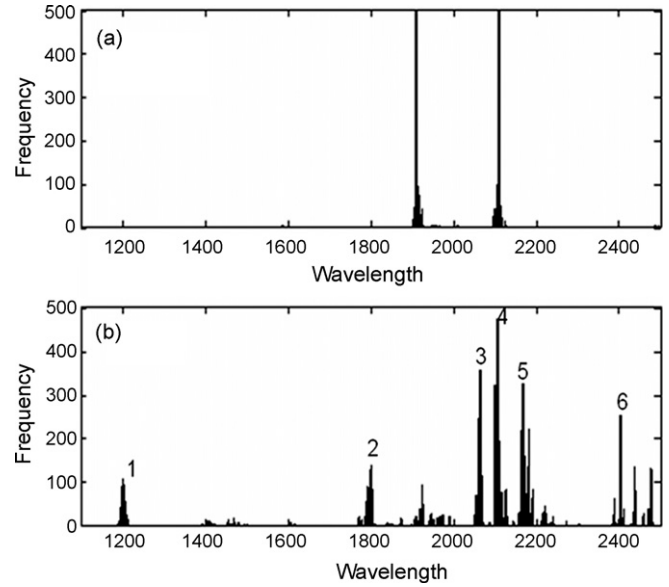

Fig. 10. The selected frequency of each wavelength by running CARS 500 times of corn moisture data (plot a) and corn protein data (plot b).

mal combination of chemically meaningful wavelengths that can lead to calibration model with better prediction ability.

4.4. Corn protein data

Fig. 11 shows the wavelength selection results obtained by MC-UVE, MWPLS and CARS. There exist common wavelength band by MC-UVE and MWPLS, such as the regions around 1202, 1760, 1974 and 2180 nm. Also great difference exists between these two methods, *e.g.* the bands around 1800, 1910, 2200 and 2400 nm. The fact that selected informative bands are distributed in a wide range

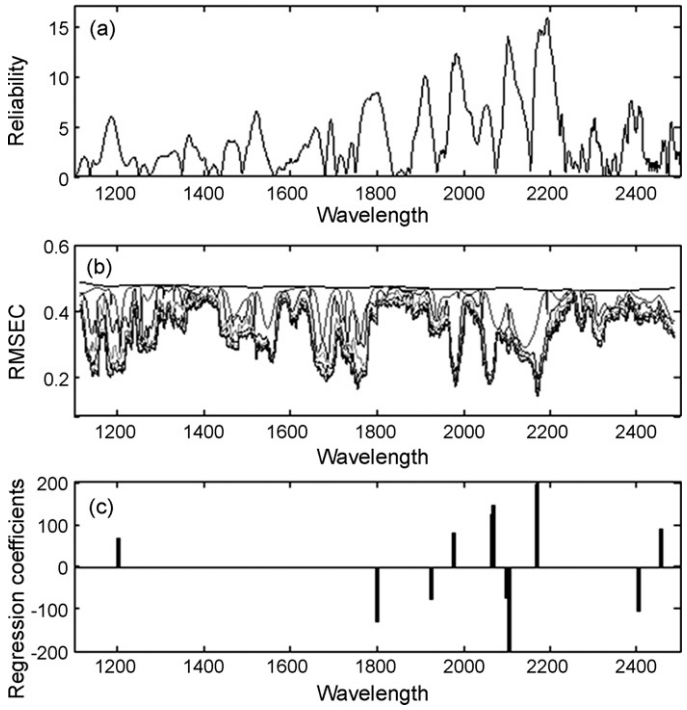

Fig. 11. Comparison of selected wavelengths by MC-UVE, MWPLS and CARS. The window size of MWPLS is set 15. The iteration number of MC-UVE and the number of sampling runs of CARS are both set to 500.

Table 3
The results on corn protein data.

Methods	RMSECV	nLVs	nVAR
PLS	0.1500	10	700
MC-UVE-PLS[a]	0.1214 ± 0.0005	8 ± 1	175 ± 12
MWPLS[b]	0.1325	9	106
CARS-PLS[a]	0.1067 ± 0.0033	8 ± 1	19 ± 5

[a] The mean and standard deviation are calculated from the results of 500 runs of MC-UVE and CARS, respectively.

[b] The chosen wavelength bands by MWPLS here are the combination of 1178–1208, 1658–1698, 1718–1778, 1968–1998, 2048–2068 and 2158–2178 nm.

is implicitly in agreement with the complex structure characteristics of protein, such as different vibration modes (stretching or bending) of C–H, O–H and N–H bond, the complicated microenvironment of C–H, O–H and N–H bond, and the interaction of them. Interestingly, some of the selected wavelengths by CARS are consistent with those by MC-UVE (1202, 1920, 1974 nm, etc.), others are consistent with those by MWPLS (1202, 1974, 2062, 2168 nm, etc.). Besides, the selected wavelength 2454 nm is unique to CARS. So, the performances of the three methods are sure to be different due to the difference of selected wavelengths. Table 3 presents the results of them together with that of full spectrum PLS. It is obvious that the best prediction in terms of RMSECV, are obtained by CARS. By comparison, CARS has a larger standard deviation than MC-UVE (0.0033 versus 0.0005), which means that the stability of CARS needs improving. One significant advantage is that the mean number of selected wavelengths by CARS is 19 with a standard deviation 5, which is much smaller than those of other methods. This phenomenon conveys that better prediction ability can be achieved with fewer wavelengths. Thus one can conclude that it is very necessary to first perform wavelength selection before building a calibration model. Moreover, it is also feasible to choose only the key wavelengths not a local continuous band or combination of several continuous bands for modeling because the severe collinear wavelengths can reduce the stability of calibration models. Occam's Razor Theory may account for this [44,45].

In order to investigate the stability behavior of CARS, we statistically calculate the frequency of each wavelength by running CARS 500 times. The result is shown in Fig. 10b. It can be found that only a small part of the wavelengths can be selected by CARS and the selected wavelengths are mainly distributed in six regions denoted by 1, 2, 2, 4, 5 and 6, respectively. This observation may be an indication that the wavelength in these six regions should be jointly meaningful to correlate protein content with the NIR spectra. Although it is hard to accurately assign the selected band to the chemical bond, the wide range covered by the selected wavelengths, can be a proof of the highly complexity of protein structure. Additionally, one should pay attention to the wavelengths with extremely high frequency, such as 2062, 2104, 2166, 2400 nm, etc. These wavelengths can be naturally considered to be key wavelengths. Moreover, one run of CARS can usually select a subset containing the wavelengths from the six regions. This may be a potential advantage of CARS.

It is also interesting to analyze the regression coefficient path of each wavelength as shown in Fig. 12c. As mentioned before, each line reflects the changing of coefficient of one wavelength. During CARS, some important wavelengths are retained while other incompetent ones are eliminated. The critical point denoted by asterisk line indicates the optimal subset with the lowest RMSECV. After this point, RMSECV values begin to increase because of the removing of some key wavelengths. For instance, RMSECV value rises up to a much higher level at the time denoted by dot line L1 because one wavelength (denoted by P1) is eliminated. The removal of another key wavelength (denoted by P2) also results in the sharp rising of RMSECV value (L2).

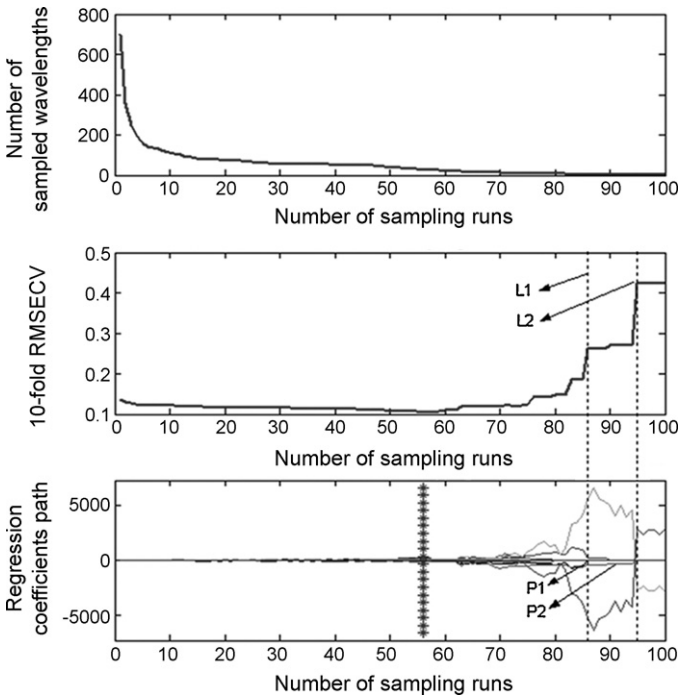

Fig. 12. Plots a–c, respectively, depict the changing of the number of sampled wavelengths, 10-fold RMSECV values and the regression coefficient path of each wavelength. The vertical asterisk line denotes the optimal point where 10-fold RMSECV values achieve the lowest.

5. Conclusions

This paper presents a new method for key wavelength selection using competitive adaptive reweighted sampling technique coupled with PLS. Based on the importance level of each wavelength, CARS sequentially selects *N* subsets of wavelengths from *N* sampling run. In each sampling run, the number of wavelengths to be selected by CARS is controlled by the proposed exponentially decreasing function and further by adaptive reweighted sampling. This sampling process is somewhat similar to the 'survival of the fittest' principle in Darwin's Evolution Theory. In an efficient and competitive way, CARS finally selects a combination of key wavelengths which is of great competence. With applications to one simulated dataset and one real NIR spectral dataset of two properties, it is demonstrated that CARS is a promising procedure to eliminate the uninformative variables and/or conduct wavelength selection for building a high performance calibration model. Our results indicate that wavelength selection is really necessary and better prediction can be obtained using a few chemically meaningful key wavelengths not a continuous band or combination of several continuous bands because the high collinear wavelengths may reduce the stability of the calibration model.

Although wavelength selection is performed by CARS coupled with PLS in this work, it should be pointed out that it can also be extended to be in combination with other modeling methods in either regression or pattern recognition. Our future work will focus on investigating the minute behavior of CARS and the application of CARS in other fields, such as biomarker discovery using genomic, proteomic and metabolomic data.

Acknowledgements

This work is financially supported by the National Nature Foundation Committee of P.R. China (Grants Nos. 20875104 and 10771217), the international cooperation project on traditional Chinese medicines of ministry of science and technology of China (Grant Nos. 2006DFA41090 and 2007DFA40680). The studies meet with the approval of the university's review board.

References

[1] P.J. Gemperline, A. Salt, J. Chemometr. 3 (1989) 343.
[2] M.K. Hartnett, G. Lightbody, G.W. Irwin, Chemometr. Intell. Lab. 40 (1998) 215.
[3] M. Sjostrom, S. Wold, W. Lindberg, J.-A. Persson, H. Martens, Anal. Chim. Acta 150 (1983) 61.
[4] P. Geladi, B.R. Kowalski, Anal. Chim. Acta 185 (1986) 1.
[5] A.S. Bangalore, R.E. Shaffer, G.W. Small, M.A. Arnold, Anal. Chem. 68 (1996) 4200.
[6] P.J. Gemperline, J. Chemometr. 3 (1989) 549.
[7] C.B. Lucasius, G. Kateman, TrAC 10 (1991) 254.
[8] C.B. Lucasius, M.L.M. Beckers, G. Kateman, Anal. Chim. Acta 286 (1994) 135.
[9] B. Hemmateenejad, M. Akhond, R. Miri, M. Shamsipur, J. Chem. Inf. Comp. Sci. 43 (2003) 1328.
[10] R.E. Shaffer, G.W. Small, M.A. Arnold, Anal. Chem. 68 (1996) 2663.
[11] Q. Ding, G.W. Small, M.A. Arnold, Anal. Chem. 70 (1998) 4472.
[12] D. Jouan-Rimbaud, D.-L. Massart, R. Leardi, O.E. De Noord, Anal. Chem. 67 (1995) 4295.
[13] T.-H. Li, C.B. Lucasius, G. Kateman, Anal. Chim. Acta 268 (1992) 123.
[14] A. Niazi, A. Soufi, M. Mobarakabadi, Anal. Lett. 39 (2006) 2359.
[15] H. Khajehsharifi, E. Pourbasheer, J. Chin. Chem. Soc. 55 (2008) 163.
[16] J.H. Kalivas, N. Roberts, J.M. Sutter, Anal. Chem. 61 (1989) 2024.
[17] K. Sasaki, S. Kawata, S. Minami, Appl. Spectrosc. 40 (1986) 185.
[18] Y.-Z. Liang, Y.-L. Xie, R.-Q. Yu, Anal. Chim. Acta 222 (1989) 347.
[19] U. Horchner, J.H. Kalivas, Anal. Chim. Acta 311 (1995) 1.
[20] M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad, M. Akhond, J. Chemometr. 20 (2006) 146.
[21] S.D. Osborne, R. Künnemeyer, R.B. Jordan, Analyst 122 (1997) 1531.
[22] V. Centner, D.-L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, C. Sterna, Anal. Chem. 68 (1996) 3851.
[23] W. Cai, Y. Li, X. Shao, Chemometr. Intell. Lab. 90 (2008) 188.
[24] Q.-J. Han, H.-L. Wu, C.-B. Cai, L. Xu, R.-Q. Yu, Anal. Chim. Acta 612 (2008) 121.
[25] J.-H. Jiang, R.J. Berry, H.W. Siesler, Y. Ozaki, Anal. Chem. 74 (2002) 3555.
[26] M.C. Ugulino Araújo, T.C.B. Saldanha, R.K.H. Galvão, T. Yoneyama, H.C. Chame, V. Visani, Chemometr. Intell. Lab. 57 (2001) 65.
[27] S. Ye, D. Wang, S. Min, Chemometr. Intell. Lab. 91 (2008) 194.
[28] T. Chen, E. Martin, Anal. Chim. Acta 631 (2009) 13.
[29] X.B. Zou, Y.X. Li, J.W. Zhao, J. Near Infrared Spectrosc. 15 (2007) 153.
[30] B. Cheng, X.H. Wu, D.Z. Chen, Spectrosc. Spect. Anal. 26 (2006) 1923.
[31] H.C. Goicoechea, A.C. Olivieri, J. Chem. Inform. Comp. Sci. 42 (2002) 1146.
[32] I.S. Liang Xu, Anal. Chem. 68 (1996) 2392.
[33] J.B. Philip, J. Chemometr. 6 (1992) 151.
[34] D.M. Allen, Technometrics 16 (1974) 125.
[35] M. Stone, J. R. Stat. Soc. B 36 (1974) 111.
[36] S. Wold, Technometrics 20 (1978) 397.
[37] Q.-S. Xu, Y.-Z. Liang, Chemometr. Intell. Lab. 56 (2001) 1.
[38] Q.-S. Xu, Y.-Z. Liang, H.-L. Shen, J. Chemometr 15 (2001) 135.
[39] http://software.eigenvector.com/Data/index.html.
[40] T. Hastie, J. Taylor, R. Tibshirani, G. Walther, Electron. J. Stat. 1 (2007) 1.
[41] R. Tibshirani, J. R. Stat. Soc. 58 (1996) 267.
[42] D. Ghosh, A.M. Chinnaiyan, J. Biomed. Biotechnol. 2 (2005) 147.
[43] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Ann. Stat. 32 (2004) 407.
[44] W.B. Roantree, Lancet 276 (1960) 600.
[45] A. Blumer, A. Ehrenfeucht, D. Haussler, M.K. Warmuth, Inform. Process. Lett. 24 (1987) 377.