A new spectral variable selection pattern using competitive adaptive reweighted sampling combined with successive projections algorithm

Article *in* The Analyst · June 2014

DOI: 10.1039/c4an00837e

Analyst

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/c4an00837e

A new spectral variable selection pattern using competitive adaptive reweighted sampling combined with successive projections algorithm

Guo Tang,[a] Yue Huang,*[ab] Kuangda Tian,[a] Xiangzhong Song,[a] Hong Yan,[a] Jing Hu,[a] Yanmei Xiong[a] and Shungeng Min*[a]

The competitive adaptive reweighted sampling-successive projections algorithm (CARS-SPA) method was proposed as a novel variable selection approach to process multivariate calibration. The CARS was first used to select informative variables, and then SPA to refine the variables with minimum redundant information. The proposed method was applied to near-infrared (NIR) reflectance data of nicotine in tobacco lamina and NIR transmission data of active ingredient in pesticide formulation. As a result, fewer but more informative variables were selected by CARS-SPA than by direct CARS. In the system of pesticide formulation, a multiple linear regression (MLR) model using variables selected by CARS-SPA provided a better prediction than the full-range partial least-squares (PLS) model, successive projections algorithm (SPA) model and uninformative variables elimination-successive projections algorithm (UVE-SPA) processed model. The variable subsets selected by CARS-SPA included the spectral ranges with sufficient chemical information, whereas the uninformative variables were hardly selected.

Received 9th May 2014
Accepted 30th June 2014

DOI: 10.1039/c4an00837e

www.rsc.org/analyst

Introduction

In recent years, near-infrared (NIR) spectroscopy has gained wide acceptance in different fields such as agriculture[1–4] and the petrochemical[5,6] and pharmaceutical industries[7,8] by virtue of its advantages in recording spectra for solid and liquid samples without any pretreatment. Generally, NIR spectroscopy is used in combination with multivariate techniques for qualitative or quantitative analysis. However, with the existence of uninformative or irrelevant variables in raw spectra, bad or inefficient prediction results are usually obtained. To solve this problem, suitable projection or selection techniques are usually used.[9–14] It is now widely accepted that a well-performed variable selection can make models have a better prediction.[15] Variable selection aims at obtaining a subset of spectral information that gives the smallest possible errors when used to make quantitative determinations or to discriminate between dissimilar samples.

The competitive adaptive reweighted sampling method (CARS) is a recently proposed variable selection method that has been proved very efficient when applied to NIR data.[16,17] The successive projections algorithm (SPA), proposed as a variable selection strategy by M. C. U. Araújo, *et al.*,[18] shows the advantage of acquiring a small representative subset of full-spectrum variables with minimum collinearity. SPA has been successfully applied to select variables in NIR spectroscopy,[19–22] as well as for coefficient selection in wavelet regression models.[23–25]

CARS can select the variables with large coefficients in a multivariate linear regression model, and employing the variables selected by CARS for modeling can avoid model over-fitting and usually improves its predictive ability. However, partial least squares (PLS) refinement is still required, as too many variables are still retained after CARS for simple multiple linear regression (MLR). On the other hand, SPA employs simple projection to select variables with a minimum of collinearity, but variables selected by SPA may make little contribution to multivariate calibration, which can affect model prediction. The combination of the two methods will integrate the bright side of each, and a similar method that adapts this idea, *i.e.* the uninformative variables elimination-successive projections algorithm (UVE-SPA), had been proposed by Ye *et al.*[26] Nevertheless, CARS has been proved more efficient than UVE in variable selection as many fewer variables were selected and comparative or even better results could be obtained.[16,17] Coupling CARS with SPA may achieve more satisfactory results than using UVE-SPA.

In this work, successive projection algorithms combined with the competitive adaptive reweighted sampling (CARS-SPA) method was proposed for spectral variable selection in which SPA was employed for variable selection after CARS discarded the unimportant variables by regression. The proposed method was applied to two systems of NIR data, namely, the nicotine in tobacco lamina and the active ingredient in pesticide formulation. The corresponding MLR models were established over the spectral variables selected by CARS-SPA. Moreover, UVE, SPA, UVE-SPA

[a]College of Science, China Agricultural University, Beijing 100193, P.R. China. E-mail: minsg@263.net; orange07@126.com; Fax: +86 10 62733091; Tel: +86 10 62733091
[b]Beijing Third-class Supervision Station of Tobacco, Beijing 101121, P.R. China

and CARS were also investigated the same sample sets, and their performances were compared with the proposed method.

Theory

CARS

The CARS is a strategy for variable selection by selecting the variables with large absolute coefficients in a multivariate linear regression model such as PLS. The details of CARS can be found in ref. 16, and the principles of CARS are summarized as follows:

(1) The absolute values of regression coefficients of the PLS model are calculated and used as an index for evaluating the importance of each variable.

(2) CARS sequentially selects N subsets of wavelengths from N Monte Carlo sampling runs in an iterative and competitive manner based on the importance level of each variable. In each sampling run, a fixed ratio of samples is first randomly selected to establish a calibration model.

(3) A two-step procedure, including exponentially decreasing function (EDF)-based enforced wavelength selection and adaptive reweighted sampling-based competitive wavelength selection, is then adopted to select the key variables based on the regression coefficients.

(4) Finally, cross validation is applied to choose the subset with the lowest root mean square error of cross validation.

UVE

UVE is a method of variable selection based on stability analysis of regression coefficients b. The details of UVE can be found in ref. 13. In the present study, the main steps of UVE are taken as follows:

(1) First, PLS regression is performed on instrumental response data X_{cal} and property values y of calibration set, and the optimal number of latent variables (LVs) is determined.

(2) Then, a noise matrix with the same size of the X_{cal} is generated, and the elements are multiplied with a small constant to make their impact on the model negligible. The noise matrix is appended to the original one to form an extended matrix with twice as many variables as the original.

(3) PLS models are made on the extended matrix and y in manner of leave-one-out cross validation. This leads to a matrix of b values with as many rows as samples and one column for each variable, both original and random.

(4) The c value of each variable is calculated as the average of the b values of each column divided by the standard deviation of that column.

(5) The cut-off value is set as the maximum of absolute value c among the random variables. Every original variable with equal or lower absolute value of c is assumed to contain nothing but noise and is eliminated.

SPA

In SPA, the selection of variables is cast in the form of a combinatorial optimization problem with constraints that are formed according to a sequence of projection operations. Moreover, the projection operations are used to choose subsets of variables with little collinearity to minimize redundancy. SPA is aimed at selecting variables for use in multiple linear regression (MLR) models. SPA employs simple projection operations in a vector space to obtain subsets of variables with small collinearity. The general procedure of SPA[18,27] is summarized as follows:

Step 1: before the first iteration ($n = 1$), let $x_j = j$th column of X_{cal}; $j = 1, \ldots, J$.

Where X_{cal} is the spectra matrix of calibration set.

Step 2: let S be the set of wavelengths which have not been selected yet, *i.e.*,

$$S = \{j \text{ such that } 1 \leq j \leq J \text{ and } j \notin \{k(0), \ldots, k(n-1)\}\}.$$

Step 3: calculate the projection of x_j on the subspace orthogonal to $x_{k(n-1)}$ as

$$\mathbf{P}_{x_j} = x_j - (x_j^T x_{k(n-1)}) x_{k(n-1)} (x_{k(n-1)}^T x_{k(n-1)})^{-1}$$

for all $j \in S$, where $\mathbf{P}$ is the projection operator.

Step 4: let $k(n) = \arg(\max \|\mathbf{P}_{x_j}\|, j \in S)$.

Step 5: let $x_j = \mathbf{P} x_j, j \in S$.

Step 6: let $n = n + 1$. If $n < N$, go back to step 2.

Step 7: the resultant wavelengths are $\{k(n); n = 0, \ldots, N-1\}$.

The optimal number of employed variables is determined by the performance in the following MLR model.

UVE-SPA and CARS-SPA methods

UVE-SPA and CARS-SPA are both combination methods. CARS or UVE is firstly employed to select the key variables, and then SPA is used to select variables from the key variables that have minimum collinearity. In terms of spectral variable selection, both methods have obvious advantages in two aspects: (1) make the association of variables and property closer; (2) the variables that SPA requires is significantly reduced. In particular, the advantage of CARS-SPA over UVE-SPA lies mainly in the efficiency differences between CARS and UVE. Because fewer variables are retained by CARS than by UVE, SPA can work more efficiently with the reserved information and greatly improve the modeling efficiency. Furthermore, as CARS selects the variables with large regression coefficient, whereas UVE eliminates the variables with low signal to noise ratio (S/N), the discrepancy between the two techniques would give different variable selection results. As is well known, MLR cannot handle the original variables for severe colinearity exists in raw spectral data. Since UVE-SPA and CARS-SPA not only eliminate the unimportant but also collinear variables, MLR can process the retained variables instead of seeking latent variables with PLS. In this study, MLR is subsequently conducted after UVE-SPA or CARS-SPA.

Experimental

Reagents

Deltamethrin emulsion (25 g L^{-1}, Bayer Crop Science, China); original deltamethrin (98.1%, Jiangsu Huangma Agrochemicals,

Table 1 Evaluation of the reference methods (unit: %)

	Number	1	2	3	4	5	6	Mean	RSD	$\overline{RSD}$
Tobacco	1	1.71	1.71	1.69	1.7	1.7	1.71	1.70	0.82	0.73
	2	2.48	2.47	2.48	2.47	2.49	2.48	2.48	0.75	
	3	3.5	3.51	3.49	3.50	3.50	3.50	3.50	0.63	
Pesticide	1	1.43	1.43	1.43	1.43	1.42	1.43	1.43	0.52	0.56
	2	2.53	2.54	2.53	2.54	2.54	2.54	2.54	0.55	
	3	4.35	4.36	4.37	4.36	4.36	4.36	4.36	0.63	

China); dimethylbenzene (A.R., Beijing Chemical Works, China) and carbon tetrachloride (A.R., Beijing Chemical Works, China). All the reagents were stored in the refrigerator at 4 ± 1 °C before the experiment.

Reference methods

As the models were established on the data obtained from the reference methods, the standard error of the reference methods plays an important role in chemometrics modeling. In this study, two actual data sets were used to evaluate the proposed variable selection technique:

(1) Nicotine in tobacco lamina: the concentration of nicotine in tobacco laminae was measured by continuous flow analysis method with an external standard method. In this data set, the nicotine content ranges from 1.06% to 4.37%, and the mean value is 2.42%.

(2) Deltamethrin concentration in pesticide formulation: the exact content of the deltamethrin in the formulation was measured by HPLC with an external standard method. In this data set, the concentration of deltamethrin ranges from 0.11% to 5.39% (w/w), and the mean value is 2.80%.

To evaluate the reference methods, three samples of each data set were measured for six times, and the results are shown in Table 1.

Diffuse reflectance spectra of tobacco samples

NIR diffuse reflectance spectra of 500 tobacco lamina samples were measured using a FT-NIR spectrometer (Spectrum ONE NTS, PerkinElmer, USA). The spectra were recorded over the wavenumber range of 10 000–4000 cm^{-1} at 8 cm^{-1} resolution. Each spectrum was the average of 64 scans. For grouping, 400 samples were randomly selected as the modeling set, and the other 100 samples were used as the prediction set. Among the modeling set, 250 samples were treated as calibration set by K-Stone sampling,[28] and the other 150 samples were considered as the validation set.

Transmittance spectra of pesticide formulation samples

Three batches of the commercial deltamethrin formulation were used to prepare the samples, and 80 samples were prepared in total. Among the samples, a prediction set that consisted of 20 samples was prepared independently to evaluate the model. The other 60 modeling samples were divided into calibration set (40 samples) and validation set (20 samples) by K-Stone sampling. Each sample was prepared with certain amount of formulation, dimethylbenzene and deltamethrin for spectra collection. In order to avoid collinearity, the three reagents were added randomly. The gross mass of each sample was around 15 g and the concentration ranged from 0.11% to 5.39% (w/w). The transmittance spectra of the prepared pesticide formulation were recorded by a FT-NIR spectrometer (Spectrum ONE NTS, PerkinElmer, USA). The spectra were recorded over the wavenumber range of 12 500–4000 cm^{-1} at 8 cm^{-1} resolution. Each spectrum was the average of 64 scans. The cuvette was rinsed with carbon tetrachloride between the samples.

Simulated data

A dataset, called SIMUIN, is simulated in the way as ref. 13, which contains exactly five latent variables. The yielded relative eigenvalues by principal component analysis on the centered data are (%) 23.40, 20.94, 19.26, 18.57 and 17.83. SIMUIN consists of 80 samples in rows and 200 wavelengths in columns. The first 100 wavelengths are linearly related with y, whereas the last 100 columns contain random numbers from 0 to 1, representing uninformative wavelengths. The added noises are normally distributed in the range 0 to 0.005. Simulated data was randomly grouped into a modeling set (60 samples) and a prediction set (20 samples). Among the modeling set, 40 samples were set as calibration and the other 20 samples as validated by K-Stone sampling.

Software and scripts

The spectra files were imported into Matlab (v7.11, MathWorks, USA) for data analysis. The scripts used in this study are based on ref. 13, 18 and 26. The CARS scripts are available at http://www.code.google.com/p/carspls/, and the other scripts are also available upon request.

Modeling strategy

First, from all the samples, a modeling set was randomly selected and the remainder was used as a prediction set (for prepared pesticide samples, the modeling set and prediction set were prepared separately). Among the modeling set, samples of calibration set were selected by K-stone sampling and the remaining samples were taken as a validation set to test the performance of the model. As is well known, cross validation is an effective and widely used technique for modeling and variable selection. Thus, in this study, the models were all

established over a 10-fold cross validation technique. Parameters of algorithms were optimized according to the root mean square error of cross-validation (RMSECV) generated from the calibration. After optimization of the preliminary modeling, a validation set was used to validate its performance (root mean square error of validation, RMSEV). Lastly, the model was evaluated by an independent sample sets to testify its prediction ability, *i.e.*, root mean error square error of prediction (RMSEP).

Results and discussion

Influence of number of Monte Carlo sampling runs

To investigate the influence of the number of Monte Carlo sampling runs on the performance of CARS, the following four cases were taken into consideration in which the number of sampling runs was individually set as 50, 100, 200 and 500 times. In each case for the three datasets, 50 replicates of CARS running were executed, and simultaneously, RMSECV values were recorded. Results of statistical box-plots are shown in Fig. 1. There is no obvious evidence that the number of Monte Carlo sampling runs has any significant influence on the performance of CARS. Therefore, the number of Monte Carlo sampling runs was set 100 times in this study.

Investigation of the simulated data

The data matrix SIMUIN was implemented to evaluate the efficiency of the proposed CARS-SPA method. It was compared with the other three approaches as UVE, CARS and UVE-SPA, aiming to ascertain that CARS-SPA is indeed a characteristic and alternative procedure for variable selection, but not to decide which method is observably the best.

This data was first auto-scaled at each variable to obtain zero mean and unit variance before modeling. The number of variables N selected by SPA was determined by RMSECV of corresponding MLR model. Results are illustrated in Fig. 2a. The optimal number of variables to be employed in SPA, UVE-SPA and CARS-SPA was 5 in each case. Performance of the models using different variable selection methods is listed in Table 2. Results between two full-variables PLS models clearly show that the uninformative variables have a great impact on the model efficiency. It is worth noting that, by using UVE and CARS, the uninformative variables were well managed; thus, the models' performance was equally comparable with the PLS model with only informative variables. Compared with UVE and CARS, which selected few variables and gave good RMSEP, SPA employed the same variable number, but the introduction of two uninformative variables led to poor competence. This indicated that SPA cannot deal with the uninformative variables efficiently. Interestingly, results of CARS-SPA and UVE-SPA were acceptable, but their results were both inferior to single CARS and UVE. The explanation can be that the first 100 variables were all informative; *i.e.* the fewer variables employed, the less chemical information was collected to build the regression. Therefore, the modeling performance descended along with the employed variables decreasing (PLS > UVE ≥ CARS > CARS-SPA ≈ UVE-SPA). It should be noted that the retained variables

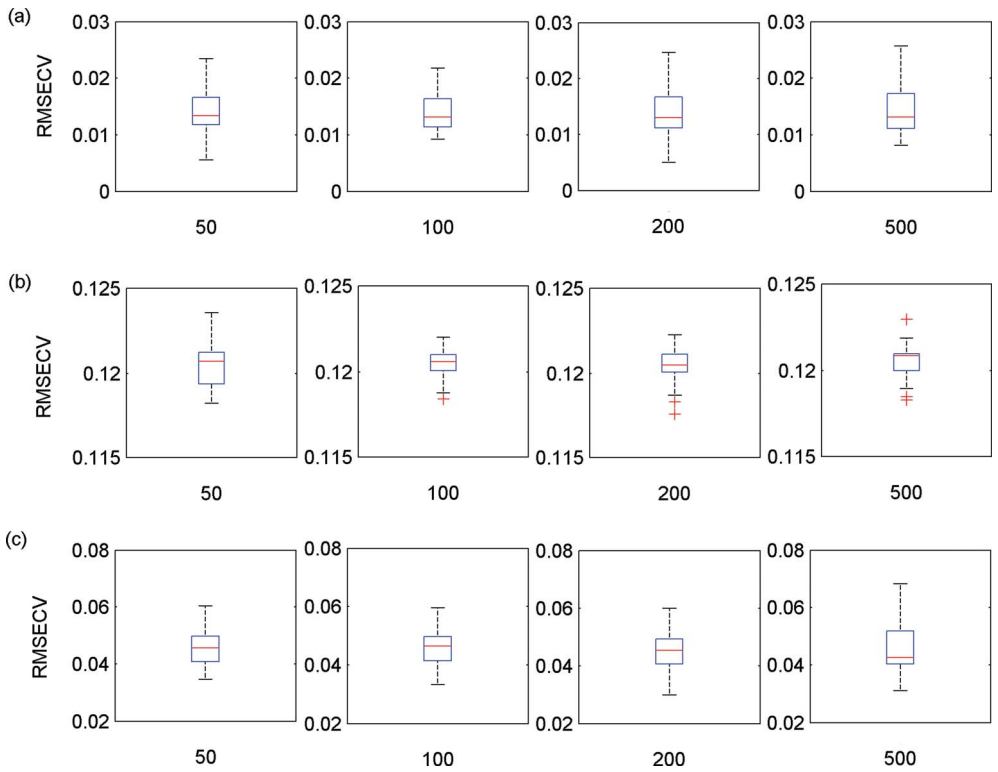

Fig. 1 Box-plots for each dataset with the number of Monte Carlo sampling runs of CARS set at 50, 100, 200 and 500. (a) Simulated dataset. (b) Tobacco nicotine data. (c) Pesticide formulation data.

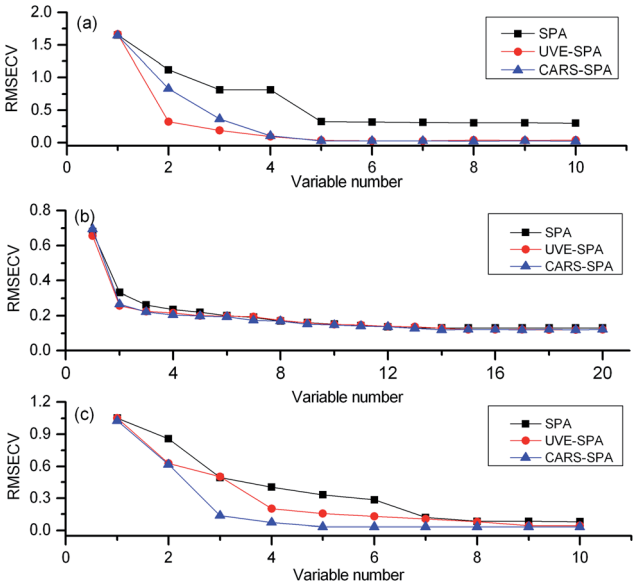

Fig. 2 Variable number *versus* RMSECV of different SPA models: simulated data (a), tobacco nicotine data (b), pesticide formulation data (c).

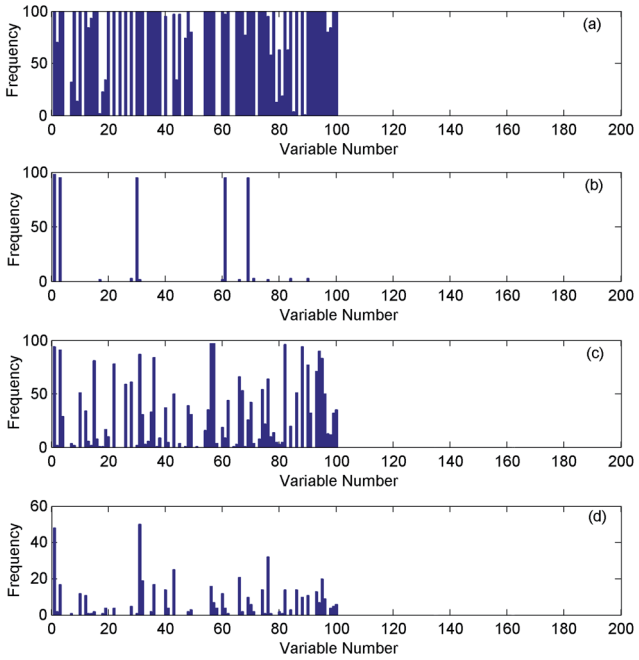

Fig. 3 Variable selection frequencies of UVE (a), UVE-SPA (b), CARS (c) and CARS-SPA (d) in 100 calculations of the simulated data.

Table 2 Results of different algorithms in modeling of simulated data

Methods	PLS factors	Variables	RMSECV	RMSEV	RMSEP
PLS[a]	6	200(100)[c]	0.0926	2.5181	2.6336
PLS[b]	4	100(0)	0.0090	0.0092	0.0115
UVE-PLS	4	65(0)	0.0097	0.0109	0.0123
CARS-PLS	4	21(0)	0.0118	0.0127	0.0141
SPA-MLR	—	5(2)	0.3230	0.4498	0.3552
UVE-SPA-MLR	—	5(0)	0.0310	0.0360	0.0332
CARS-SPA-MLR	—	5(0)	0.0261	0.0296	0.0349

[a] Results using full spectrum with 200 variables by PLS. [b] Results using only the 100 simulated informative variables by PLS. [c] Number in the bracket denotes the number of uninformative variables used in the model.

varied in different calculations. To obtain comprehensive observation of selection, all modeling calculations of each algorithm were repeated 100 times. Fig. 3 plots the variable frequencies in 100 calculations of UVE, CARS, UVE-SPA and CARS-SPA, respectively. As is shown, these four techniques can all eliminate the uninformative variables (latter 100 variables). Since UVE eliminates only a few informative variables (former 100 variables), the frequency varies highly in different selections. CARS behaved similarly to UVE, except that the selection frequency differs at some variables. Fig. 3b and d show that the frequency of UVE-SPA intensively distributes on several specific variables, while CARS-SPA moderately reduced the frequency of the variables selected by CARS, simultaneously ensuring the probability of selection distributes more widely. Theoretically, every variable in the first 100 variables of this system is informative as they are all contributing to the regression. In this sense, the selections of CARS and CARS-SPA are more reasonable.

Analysis of nicotine in tobacco lamina

The original NIR spectra of tobacco lamina are presented in Fig. 4a. First derivative, a widely used spectral-preprocessing method, can remove most of the influence of baseline variation. As the drift of baseline of the spectra always had a great impact on the model performance in solid samples, the first derivative spectra with 9 points smoothing by a Savitzky–Golay filter with a second-order polynomial were used. The number of variables N to be selected by SPA was determined by the RMSECV of the MLR model and the results are illustrated in Fig. 2b. As shown in Fig. 2b, the optimal variables to be employed should be 13, 15 and 14 for SPA, UVE-SPA and CARS-SPA, respectively. The performance of the PLS and MLR models with optimal parameters are listed in Table 3.

As shown in Table 3, UVE and CARS gave a slightly better result than the full-spectrum PLS model. In total, 257 variables were retained by UVE, whereas 52 variables were selected by CARS. Although the two models obtained comparable results, fewer PLS factors were used in the CARS-PLS model, as only one-fifth of the variables were employed than the fraction in the UVE-PLS model. Both UVE-SPA and CARS-SPA achieved performance comparable to that of UVE and CARS with fewer variables as SPA removes the colinearity between the original variables. However, it should be noted that the least variables used in the SPA-MLR model had no direct relation to a robust prediction (or to a satisfactory result). This is because SPA cannot completely avoid selecting uninformative variables, which have been investigated in simulated data (shown in Table 2). Fig. 5 shows the variables selected by SPA, UVE-SPA and CARS-SPA. The three algorithms all retained the variables in the absorption peaks (derivative spectra) that were located in the range of 4000–5200 cm^{-1} (the first combination of stretching

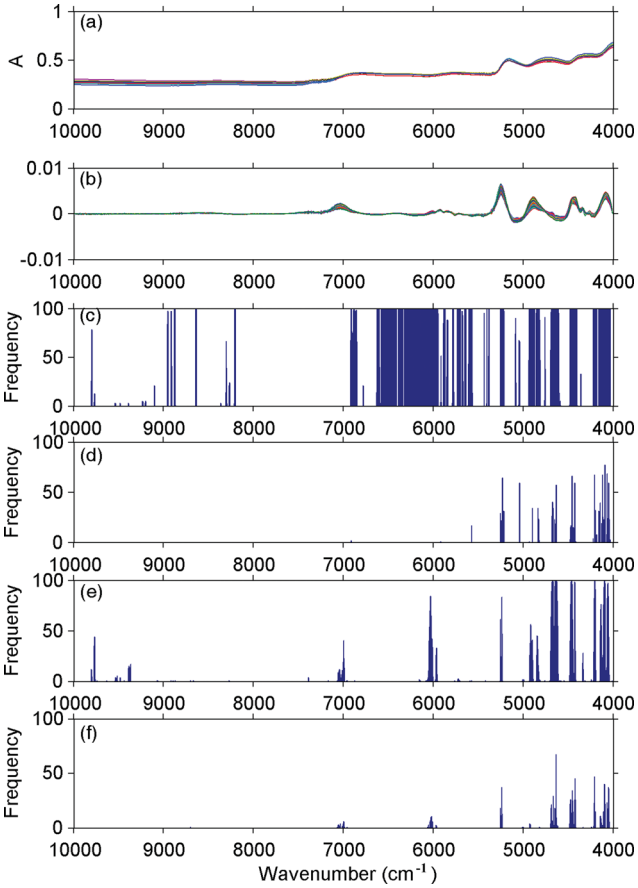

Fig. 4 Original spectra (a), derivative spectra of the tobacco samples (b) and variable selection frequencies of UVE (c), UVE-SPA (d), CARS (e) and CARS-SPA (f) in 100 calculations.

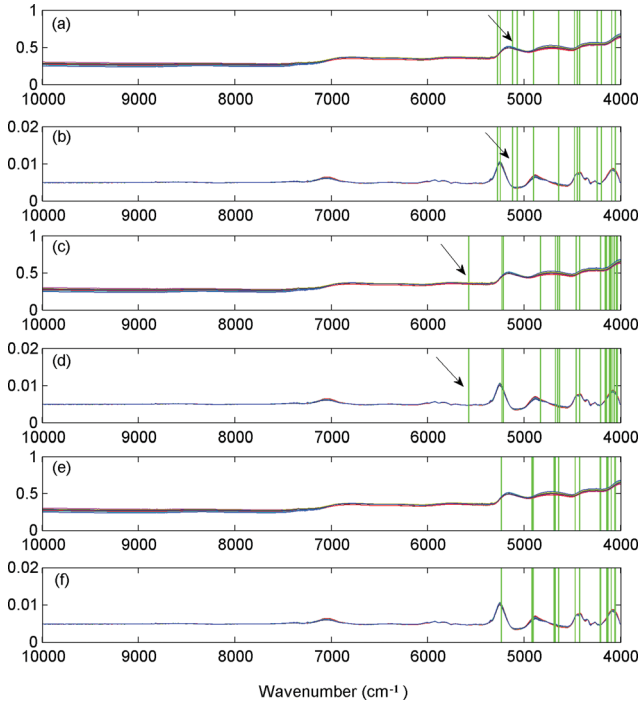

Fig. 5 Selected variables by SPA with raw spectra (a) and derivative spectra (b), UVE-SPA with raw spectra (c) and derivative spectra (d), CARS-SPA with raw spectra (e) and derivative spectra (f) in analysis of nicotine in tobacco.

vibration of X–H and second overtone of C=O).[29] The difference among the three procedures was that SPA and UVE-SPA tended to select more less-informative variables (annotated with arrows, *e.g.*, variables in 5050 cm^{-1} for SPA and those in 5700 cm^{-1} for UVE-SPA), whereas CARS-SPA ignored most of them. On the other hand, as the retained variables varied in different calculations; thus, further research should be performed with the two algorithms. Therefore, the overview of the selected variables by different calculations was performed. In Fig. 4c–f, the variable frequencies in 100 calculations of UVE, CARS, UVE-SPA and CARS-SPA are presented, respectively. The selection frequency of a variable indicated in some respects the importance of the variable in the calibration. Obviously, there was a considerable difference between UVE and CARS (Fig. 4c and e). UVE selected almost all variables with same high frequency, whereas the variables selected by CARS posed a more selective and more dispersive distribution of variable selection frequency. As is shown in Fig. 4d and f, variables in the range from 4000 cm^{-1} to 5200 cm^{-1} (the first combination of stretching vibration of X–H and second overtone of C=O[29]) were both accepted by UVE-SPA and CARS-SPA in the calibration because of their high spectral absorption (contain much information or variation). However, the variables in 5700 cm^{-1} (with a weaker signal, *e.g.*, less informative) were also included in the UVE-SPA calculation. The explanation can be that the coordination of baseline variables is required in UVE calculation (to obtain the criterion of elimination). In contrast, CARS-SPA employed the variables in the range of 6000 cm^{-1} (first overtone of C–H[29]) and 7000 cm^{-1} (the second combination of stretching

Table 3 Results of different algorithms in determination of nicotine in tobacco lamina

Methods	PLS factors	Variables	RMSECV	RMSEV	RMSEP
PLS	8	1555	0.1185	0.1288	0.1291
UVE-PLS	7	257	0.1197	0.1263	0.1259
CARS-PLS	6	52	0.1195	0.1261	0.1247
SPA-MLR	—	13	0.1298	0.1263	0.1293
UVE-SPA-MLR	—	15	0.1201	0.1194	0.1218
CARS-SPA-MLR	—	14	0.1192	0.1193	0.1204

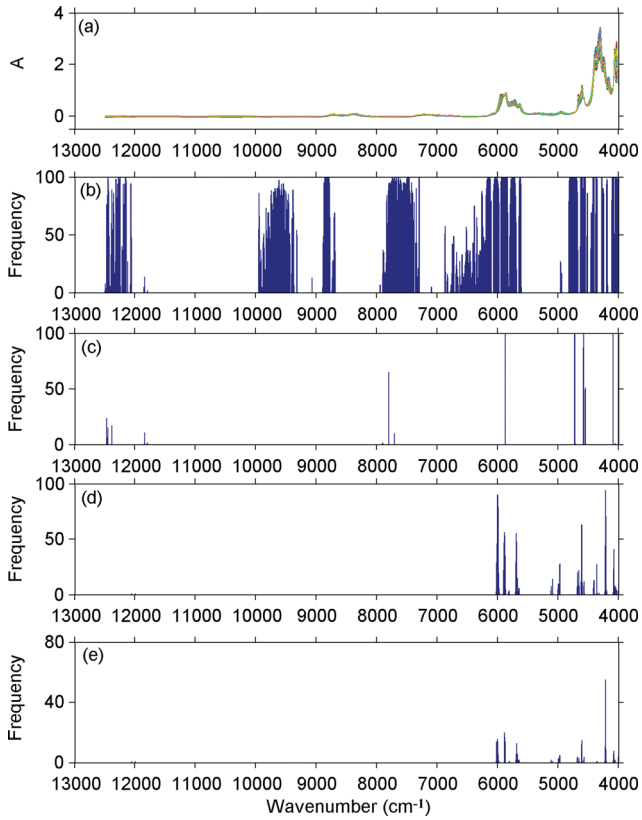

Fig. 6 Original spectra of the pesticide formulation samples (a) and the variable selection frequencies of UVE (b), UVE-SPA (c), CARS (d) and CARS-SPA (e) in 100 calculations.

vibration of C–H and first overtone of O–H[29]), which had more chemical information instead of those baseline variables. Furthermore, the frequency of CARS-SPA was more dispersive than that of CARS, indicating that the variables were equally important in the calibration. Moreover, the variables in the absorption peak of the derivative spectrum (4000–5200 cm^{-1}, 6000 cm^{-1} and 7000 cm^{-1}) were still possibly employed in the modeling. The variables in the range from 9000 cm^{-1} to 10 000 cm^{-1} (with little information) that were selected by CARS were completely removed by CARS-SPA. In addition, it should be noted that the variables selected by CARS were almost a subset of those by UVE, but more less-informative variables were reserved by UVE. Nevertheless, the retained variables may differ after SPA calculation. In CARS-SPA, few uninformative variables were introduced to the following SPA procedure; therefore, SPA can perform its search with less interference. Informative variables that made a connotative contribution to the regression were finally retained. However, case of UVE-SPA was different: many uninformative variables employed in the SPA calculation led to the entombment of the competitive but connotative variables to the regression.

Analysis of active ingredient in pesticide formulation

The obtained spectra of the prepared samples are shown in Fig. 6a. More sophisticated models using spectral pretreatments including derivative, smoothing and normalization, among others, were also attempted. However, they made no significant difference or they were even worse than the non-preprocessed model; therefore, no pretreatments were used in this analysis. The number of variables *N* to be selected by SPA was determined by the RMSECV of the MLR model and the results are presented in Fig. 2c. According to Fig. 2c, the optimal number of variable employed in the following MLR models should be 8, 9 and 5 for SPA, UVE-SPA and CARS-SPA, respectively. Table 4 outlines the results of PLS and MLR models. As shown in Table 4, unlike the tobacco system, CARS outperformed UVE, although both of them obtained a better result than the full-spectrum PLS model. Variables retained by UVE were thirty times of those selected by CARS. Unlike tobacco, which consists of thousands of chemicals, the pesticide formulation is a simple system composed of several components. In the simple system, full spectral data seems surplus for preprocessing, and usually brings in bad results when irrelevant information interferes. Therefore, much better results will be obtained when variable selection is used for fewer but essential variables. It should be noted that both UVE-SPA and CARS-SPA achieved more impressive performances than direct UVE and CARS, indicating the obvious advantage of the variable selection technique in the simple system. Furthermore, the CARS-SPA model selected only five variables, which was only the half of those selected by UVE-SPA and SPA. Fig. 7 shows the variables selected by SPA, UVE-SPA and CARS-SPA from the spectrum of deltamethrin (dissolved in carbon tetrachloride). As designated with arrows in Fig. 7, the variables in the range of 6500–12 500 cm^{-1}, which had little chemical information related to the deltamethrin, were completely discarded by CARS-SPA, but SPA and UVE-SPA still employed some of them. This further demonstrated that SPA and UVE-SPA cannot avoid choosing those uninformative variables that were used by UVE. All these three algorithms selected the variables in the two main absorptions of deltamethrin: 4000–5000 cm^{-1} (the first

Table 4 Results of different algorithms in determination of active ingredient in formulations

Methods	PLS factors	Variables	RMSECV	RMSEV	RMSEP
PLS	6	4251	0.1250	0.1320	0.1335
UVE-PLS	6	1178	0.0846	0.0883	0.0907
CARS-PLS	4	38	0.0425	0.0489	0.0494
SPA-MLR	—	8	0.0823	0.0584	0.0661
UVE-SPA-MLR	—	9	0.0439	0.0396	0.0432
CARS-SPA-MLR	—	5	0.0327	0.0300	0.0330

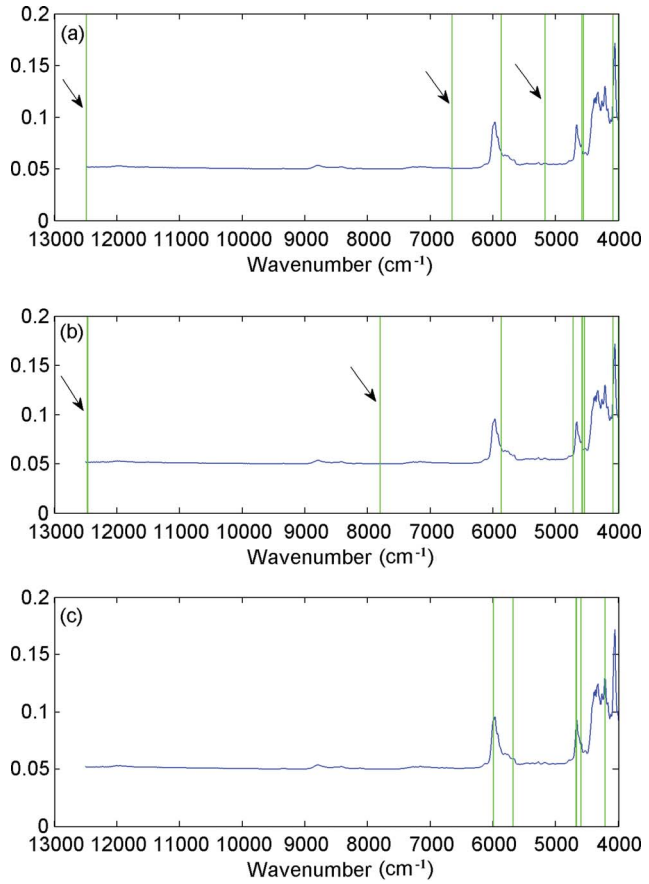

Fig. 7 Selected variables by SPA (a), UVE-SPA (b) and CARS-SPA (c) in analysis of deltamethrin in pesticide formulation.

combination of stretching vibration of C–H) and 5700–6000 cm^{-1} (the first overtone of C–H). Fig. 6b–e plot the spectra of the samples and the variable frequency in 100 calculations of UVE, CARS, UVE-SPA and CARS-SPA, respectively. In the analysis of deltamethrin, the variables selected by UVE were almost those with high frequency, whereas CARS selected fewer variables, and the selected frequencies of each variable were also more dispersive. As shown in Fig. 6c, the selection frequency of UVE-SPA was much more concentrated on several variables compared to that in the analysis of nicotine. In contrast, the selection frequency of CARS and CARS-SPA were much more dispersive than that in tobacco system. In CARS-SPA (Fig. 6e), the variables that were highly related to the deltamethrin absorption (4000–4800 cm^{-1} and 5700–6000 cm^{-1}) were reserved. However, UVE-SPA (Fig. 6c) still employed some 'baseline' variables in addition to the characteristic absorption. Moreover, the frequency of CARS-SPA was also more dispersive (both in selection range and frequency) than that of UVE-SPA, which again demonstrated that, compared with UVE-SPA, CARS-SPA can select the variable subsets not only with chemical information of the analyte but can also exclude the uninformative ones. Intuitively, variables located in absorption peak can possibly be employed, but variables located in the non-absorption waveband are usually discarded.

Conclusion

A new method designated as CARS-SPA was proposed for variable selection by combining SPA with CARS. Prior to SPA, which selected variables for multivariate calibration, CARS was performed to select the key variables with large regression coefficient, which made little contribution to calibration. The investigation of the simulated data indicated that the proposed method can avoid selecting the uninformative variables in modeling. Moreover, compared with the similar method UVE-SPA, CARS was more selective than UVE, fewer variables were employed and the search range of variables by SPA was further narrowed, making variable selection more efficient. The proposed method was successfully applied to NIR spectroscopic analysis of nicotine in tobacco lamina and the active ingredient in pesticide formulation for variable selection. As shown in the results, SPA or UVE-SPA cannot avoid selecting the uninformative variables, whereas CARS-SPA tends to retain the variables with certain chemical information in each absorption peak without involving the uninformative ones. Although CARS-SPA did not impress us with a considerably better result than UVE-SPA in a complex system such as tobacco, it still proved to be an effective and alternative technique for variable selection. The variables selected by CARS-SPA covered the full spectral range, and the uninformative variables are barely selected. This indicated that CARS-SPA can refine important variables that make positive contributions to the regression.

Acknowledgements

This research is financially supported by the National Natural Science Foundation of China (Grant no. 20575076) and the Chinese Universities Scientific Fund (no. 2012QJ028).

References

1 T. Azzouz and R. Tauler, *Talanta*, 2007, **74**, 1201–1210.
2 J. Duan, Y. Huang, Z. Li, B. Zheng, Q. Li, Y. Xiong, L. Wu and S. Min, *Ind. Crops Prod.*, 2012, **40**, 21–26.
3 S. Armenta, S. Garrigues and M. Guardia, *Vib. Spectrosc.*, 2007, **44**, 273–278.
4 B. Jamshidi, S. Minaei, E. Mohajerani and H. Ghassemian, *Comput. Electron. Agr.*, 2012, **85**, 64–69.
5 F. S. Falla, C. Larini, G. C. Le Roux, F. H. Quina, L. L. Moro and C. O. Nascimento, *J. Pet. Sci. Eng.*, 2006, **51**, 127–137.
6 R. M. Balabin, R. Z. Safieva and E. I. Lomakina, *Anal. Chim. Acta*, 2010, **671**, 27–35.
7 A. Porfire, L. Rus, A. L. Vonica and I. Tomuta, *J. Pharm. Biomed. Anal.*, 2012, **70**, 301–309.
8 C. H. Howland and S. W. Hoag, *Int. J. Pharm.*, 2013, **452**, 82–91.
9 F. Liu, Y. He and L. Wang, *Anal. Chim. Acta*, 2008, **615**, 10–17.
10 H. Xu, Z. Liu, W. Cai and X. Shao, *Chemom. Intell. Lab. Syst.*, 2009, **97**, 189–193.
11 S. Osborne, R. Künnemeyer and R. Jordan, *Analyst*, 1997, **122**, 1531–1537.

12 K. Hasegawa, Y. Miyashita and K. Funatsu, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 306–310.
13 V. Centner, D. Massart, O. de Noord, S. de Jong, B. Vandeginste and C. Sterna, *Anal. Chem.*, 1996, **68**, 3851–3858.
14 S. Sæbø, T. Almøy, J. Aarøe and A. H. Aastveit, *J. Chemom.*, 2007, **20**, 54–62.
15 H. Namkung, Y. Lee and H. Chung, *Anal. Chim. Acta*, 2008, **606**, 50–56.
16 H. Li, Y. Liang, Q. Xu and D. Cao, *Anal. Chim. Acta*, 2009, **648**, 77–84.
17 W. Fan, Y. Shan, G. Li, H. Lv, H. Li and Y. Liang, *Food Anal. Methods*, 2012, **5**, 585–590.
18 M. U. Araújo, T. B. Saldanha, R. H. Galvão, T. Yoneyama, H. Chame and V. Visani, *Chemom. Intell. Lab. Syst.*, 2001, **57**, 65–73.
19 M. C. Breitkreitz, I. M. Raimundo, J. R. Rohwedder, C. Pasquini, H. D. Filho, G. E. José and M. U. Araújo, *Analyst*, 2003, **128**, 1204–1207.
20 A. C. Pereira, M. C. Pontes, F. G. Neto, S. B. Santos, R. H. Galvão and M. U. Araújo, *Food Res. Int.*, 2008, **41**, 341–348.
21 A. S. Soares, R. H. Galvão, M. U. Araújo, S. C. Soares, L. A. Pinto and J. Braz, *Chem. Soc.*, 2010, **21**, 1626–1634.
22 F. Liu, Z. L. Jin, M. S. Naeem, T. Tian, F. Zhang, Y. He, H. Fang, Q. F. Ye and W. J. Zhou, *Food Bioprocess. Technol.*, 2011, **4**, 1314–1321.
23 C. J. Coelho, R. H. Galvão, M. U. Araújo, M. F. Pimentel and E. C. Silva, *Chemom. Intell. Lab. Syst.*, 2003, **66**, 205–217.
24 L. A. Pinto, R. H. Galvão and M. U. Araújo, *Anal. Chim. Acta*, 2010, **682**, 37–47.
25 M. J. C. Pontes, J. Cortez, R. K. H. Galvão, C. Pasquini, M. C. U. Araújo, R. M. Coelho, M. K. Chiba, M. F. Abreu and B. E. Madari, *Anal. Chim. Acta*, 2009, **642**, 12–18.
26 S. Ye, D. Wang and S. Min, *Chemom. Intell. Lab. Syst.*, 2008, **91**, 194–199.
27 S. F. C. Soares, A. A. Gomes and M. C. U. Araújo, *TrAC, Trends Anal. Chem.*, 2013, **42**, 84–98.
28 R. W. Kennard and L. A. Stone, *Technometrics*, 1969, **11**, 137–148.
29 J. J. Kelly and J. B. Gallis, *Anal. Chem.*, 1990, **62**, 1444–1451.