Uninformative variable elimination for improvement of successive projections algorithm on spectral multivariable selection with different calibration algorithms for the rapid and n...

ARTICLE *in* ANALYTICAL METHODS · AUGUST 2011

Impact Factor: 1.82 · DOI: 10.1039/c1ay05075c

5 AUTHORS, INCLUDING:

Di Wu
Zhejiang University
96 PUBLICATIONS **1,391** CITATIONS
SEE PROFILE

Xiaojing Chen
Wenzhou University
22 PUBLICATIONS **221** CITATIONS
SEE PROFILE

Analytical Methods

Dynamic Article Links

Cite this: *Anal. Methods*, 2011, **3**, 1790

www.rsc.org/methods

PAPER

Uninformative variable elimination for improvement of successive projections algorithm on spectral multivariable selection with different calibration algorithms for the rapid and non-destructive determination of protein content in dried laver

Di Wu,[a] Xiaojing Chen,*[b] Xiangou Zhu,[b] Xiaochun Guan[b] and Guichu Wu[b]

Received 13th February 2011, Accepted 18th May 2011
DOI: 10.1039/c1ay05075c

The potential of using partial least square based uninformative variable elimination algorithm (UVE_{PLS}) on successive projections algorithm (SPA) for spectral multivariable selection was evaluated. A case study was done on the visible and shortwave-near infrared (Vis-SNIR) spectroscopy for the rapid and non-destructive determination of protein content in dried laver. Three calibration algorithms, namely multiple linear regression (MLR), partial least square regression (PLS) and least-square support vector machine (LS-SVM), were used for the model establishment based on the selected variables of SPA, UVE_{PLS} and UVE_{PLS}-SPA, respectively. A total of 175 samples were prepared for the calibration ($n = 117$) and prediction ($n = 58$) sets. The performances of different pretreatments were compared. Both linear calibration algorithms of MLR and PLS and non-linear calibration algorithms of LS-SVM with linear kernel and RBF kernel obtained similar results based on certain variable selection strategies of SPA, UVE_{PLS} and UVE_{PLS}-SPA. The average improvement percentage of RPD values of four calibration algorithms was 38.66% by calculating SPA on UVE_{PLS} processed variables. Therefore there was much improvement of using UVE_{PLS} on SPA spectral multivariable selection with both linear and nonlinear calibration algorithms in this case. Moreover, the RPD values of both linear and non-linear models based on the thirteen selected variables of UVE_{PLS}-SPA show that coarse quantitative predictions of the protein determination in dried laver is possible based on Vis-SNIR spectra. We hope that the results obtained in this study will help both further chemometric (multivariate selection and calibration analysis) investigations and investigations in the sphere of applied vibrational (Near infrared, Mid-infrared and Raman) spectroscopy of sophisticated multicomponent systems.

1 Introduction

Visible-near infrared (Vis-NIR) spectroscopy has been widely adopted for low-cost, nondestructive analysis of fruits, vegetables, and grains.[1] It has the advantage of being rapid, low-cost, and nondestructive. Vis-NIR spectroscopy has the ability to predict multiple constituents and quality traits simultaneously. It permits an on-line monitoring of the food and production process and allows a fast intervention into the process when some deviation in the product standard is observed. The spectral data sets from the modern Vis-NIR spectroscopy instrumentations with a high resolution, often contain hundreds or thousands of variables. Those mass variables can cause the spectral data to be too complicated to be calibrated directly and make the calibration process time-consuming and inconvenient to fulfill the high speed feature of spectroscopy. The elimination of irrelevant variables can predigest calibration modeling and improve the results in terms of accuracy and robustness. Therefore model calibration process involves a selection on which wavelengths should be used to establish an optimal model. Recently both theoretical[2] and experimental evidence[3,4] have proved that characteristic wavelengths instead of full spectra can improve quantitative results.[5,6] Specific regions can generate more stable models with good interpretability.[7] More stable models with superior interpretability can be generated and this can produce the lower prediction error. Therefore it is important to select specific variables which contain useful information.

Successive projection algorithm (SPA) is a novel variable selection algorithm to solve the collinearity problems. It employs a simple projection operation in a vector space to select subsets of

[a]*College of Biosystems Engineering and Food Science, Zhejiang University, 866 Yuhangtang Road, Hangzhou, 310058, China*
[b]*College of Physics and Electronic Engineering Information, Wenzhou University, Chashan University Town Wenzhou, Zhejiang Province, 325035, China. E-mail: chenxj10@yahoo.cn; Fax: +86 577 86689027; Tel: +86 577 86689027*

variables with minimum collinearity.[8] Therefore its selected variables are minimally redundant. SPA can provide more reproducible results than genetic algorithm.[9] However the SPA operation is time consuming when the whole range spectra (WRS) which usually have thousands of variables were considered. Moreover, its selected variables from WRS may contain low signal noise ratio (S/N) which can affect the models performance.[10] Therefore the SPA calculation on the whole range spectra sometimes might not obtain a good result. It might be possible to improve the calibration model when SPA is followed to select variables with minimum redundant information from the informative variables with high S/N.

Partial least square regression (PLS) is an often-used calibration algorithm. Some researchers used loading weights and regression coefficients of PLS to implement variable selections.[7] However, these selections were manual and could weaken the performance of the calibration model without prior experienced knowledge about the spectra. Other researchers constructed PLS to implement variable selections, like interval PLS (iPLS), backward iPLS (biPLS) and synergy interval PLS (siPLS).[11,12] These methods split the WRS into several equidistant regions, and then establish PLS regression models for one or several sub-intervals. The best sub-interval or the optimal sub-interval combination is chosen based on its root mean square error of cross-validation (RMSECV). However, these algorithms can only consider one or a few wavelength intervals and the size of the sub-interval can affect the result of variable selection. When the sub-interval has fewer wavelength variables, its established model may miss some useful information. When more wavelength variables are considered in the sub-interval, they may include some useless data and cause a longer calculation time. Uninformation variable elimination is another variable selection algorithm based on the stability analysis of PLS regression coefficient (UVE_{PLS}).[13] In UVE_{PLS} process, the spectral data are added with the artificial random variables as a reference. So those spectral variables which have less important in the model than the random variables are eliminated. UVE_{PLS} can eliminate the variables which have no more informative variables for modeling than noise. The selected variables of UVE_{PLS} can avoid the model over-fitting and usually improve its predictive ability. UVE_{PLS} shows a better ability of variable selection than other PLS based algorithms, such as iPLS, biPLS and siPLS.[14]

The performance of UVE_{PLS} combined with SPA with different linear and nonlinear calibration algorithms was evidenced by a case study to determine the protein content of dried laver. Laver is a kind of edible seaweed with high proportions of protein and mineral salt. Laver's nutrition is influenced by environmental factors but also has a strong genetic component. To make enormous profits, dried lavers with low protein content are sold at a high price. These behaviors badly infringe on the rights and interests of consumers. Therefore the protein content determination in dried laver is critically important to both consumers and industries in public-health and economic terms. Usually protein is determined by the Micro-Kjeldahl method,[15] which is time-consuming and costly, and requires professional operation. Efforts to monitor the protein content in dried laver would be aided with nondestructive technologies that allow consumers to select protein content at the same level of price. Longwave near infrared (LNIR) spectroscopy has been successfully used for the protein measurement.[16–18] However, compared to the LNIR spectroscopy (1100–2500 nm), visible-shortwave near infrared (Vis-SNIR) spectroscopy (325–1100 nm) can obtain single-beam data reliably which can reduce the measurement time[19] and has fewer effects from the water vibration.[20] Moreover, Vis-SNIR spectroscopy instruments are usually much cheaper than those of LNIR spectroscopy. To the authors' knowledge, the current study was the first to evaluate the performance of UVE_{PLS}-SPA for multivariable selection based on different calibration algorithms and to use Vis-SNIR spectroscopy to characterize laver and specifically for the protein determination in laver.

The objective of this study was to evaluate the improvement ability of using UVE_{PLS} on SPA for the spectral multivariable selection with different linear and nonlinear calibration algorithms. Different spectral pretreatment algorithms, variable selection algorithms and calibration algorithms were compared for the protein determination in dried laver.

2 Materials and methods

2.1 Sample preparation and Vis-SNIR spectral measurement

In order to sample a broad range of protein content in laver, five brands of dried laver, from Quanzhou (L1, Fujian Province), Dongtou (L2, Zhejiang Province), Zhoushan (L3, Zhejiang Province), Cangnan (L4, Zhejiang Province), Lijian (L5, Fujian Province) respectively, were prepared on the basis of diverse geographic locations of original collection. All of these brands are popular in Chinese markets. Samples of one brand are from five batches with three production times. The protein content was determined by Kjeldahl method and the factor 6.38 was used to convert the nitrogen values to protein. The descriptive statistics for the protein contents are presented in Table 1.

A Vis-NIR spectrometer (USB4000 Miniature Fiber Optic Spectrometer, the Ocean Optics, lnc., USA) was used to measure Vis-SNIR reflectance spectra of dried laver samples. The spectral measurement was made at ambient temperature of 18–20 °C. Dried laver from each variety was fragmented and spread on a paper. An iron plate was used to make fragment's surface close to smooth. The probe of the spectrometer was placed above the surface of laver about 4 mm. The spectrum of each sample was the average of 30 successive scans. Finally 175 samples of laver samples were obtained. There were 37 samples for L1, 39 samples for L2, 35 samples for L3, 29 samples for L4 and 35 samples for L5. In order to obtain a 2 : 1 division of calibration/prediction spectra, the four samples of every six samples were selected into the calibration set. The calibration set contains 117 spectra and another 58 spectra constitute the prediction set. A low signal to noise ratio was in the spectra between 346 and 464 nm, and between 1017 and 1050 nm. Therefore the spectra containing

Table 1 Statistics of protein contents in dried laver

Data sets	Sample number	Maximum	Minimum	Mean	Standard deviation
Calibration	117	34.20	27.60	31.76	1.62
Prediction	58	33.85	28.19	31.78	1.58
All	175	34.20	27.60	31.77	1.60

useful information was determined between 464 and 1017 nm (2900 variables).

2.2 Spectra pretreatment

Five spectral pretreatment algorithms, including Savitzky-Golay (SG) smoothing,[21] standard normal variate (SNV),[22] multiplicative scatter correction (MSC),[23] 1st and 2nd derivatives (1-Der and 2-Der)), were implemented using "The Unscrambler V9.7" (CAMO PROCESS AS, Oslo, Norway). Savitzky-Golay smoothing is an averaging algorithm that fits a polynomial to the data points. SNV is a row-oriented transformation which centers and scales individual spectra. MSC is a transformation algorithm used to compensate for additive and/or multiplicative effects in spectral data. Derivative attempts to correct baseline drift in spectra. The performances of these pretreatment algorithms were compared based on PLS calibration. The spectral pretreatment calculation was implemented by "The Unscrambler®9.7" (CAMO AS, Oslo, Norway).

2.3 Variable selection algorithms

In this paper, two variable selection algorithms were investigated, namely partial least square regression based uninformation variable elimination (UVE_{PLS}) and successive projections algorithm (SPA).

2.3.1 Methodology of partial least square regression based uninformation variable elimination. UVE_{PLS} is based on the stability analysis of PLS regression coefficient.[13,24] The objective of UVE_{PLS} is to eliminate the variables which have no more information for modeling than noise. In the UVE algorithm, a PLS regression coefficient matrix $b = [b_1,\ldots b_p]$ is calculated through a leave-one-out validation; then the reliability of each variable can be quantitatively measured according to its stability. The stability of variable j can be calculated as:

$$s_j = mean(\beta_j)/std(\beta_j) \quad (1)$$

where $mean(\beta_j)$ and $std(\beta_j)$ are the mean and standard deviation of the regression coefficients of variable j. To estimate the uninformative wavelength variables, an artificial random variable matrix, with a range of approximately 10^{-5}, is established and appended to the spectral matrix. Then their S_j values, which are stability of random variable matrix, are computed. If a variable's absolute value of S_j value is smaller than the maximum absolute value of S_j values of random variable matrix, this variable is considered to be an uninformative variable. The process of UVE_{PLS} was executed in MATLAB 7.6 (The Math Works, Natick, USA).

2.3.2 Methodology of successive projections algorithm. Successive projections algorithm (SPA) is a promising variable selection algorithm. It can select variables with minimally redundant to solve the collinearity problems. In SPA process, a projection operation in a vector space is applied to select subsets of variables with a minimum collinearity.[8] In detail, the spectral data are disposed in a matrix $\boldsymbol{X}$ ($N \times K$) that the kth wavelength variable x_k corresponds to the kth column vector $\mathbf{x}_k \in \Re^N$. Let $M = \min(N - 1, K)$ be the maximum number of selected variables. The first step consists of projections carried on the $\boldsymbol{X}$ matrix, which generate k chains of M variables. The second step consists of evaluating candidate subsets of variables selected in the first step. A total of $M \times K$ subsets of variables are tested, and the best variable subset is selected. For this purpose root mean square error (RMSE) is adopted. The main feature of the algorithm can be found in literature.[25] The process of SPA was operated in MATLAB 7.6 (The Math Works, Natick, USA).

2.4 Chemometric calibration algorithms

In this paper, three calibration algorithms were investigated, namely partial least square regression, least-square support vector machine and multiple linear regression. In order to show how the chemometrics work, a flow program is shown in Fig. 1.

2.4.1 Methodology of partial least squares regression. Partial least squares regression (PLS) analysis[26] is widely used for calibration in present chemometric analysis. It can establish a regression model and perform the prediction of physiological concentrations. PLS finds the fundamental relations between the variable matrix Y (the properties of interest) and the variable matrix X (the spectra). PLS is particularly suited when variables are more than samples, and when there is multicollinearity among X values. The calculation of PLS was implemented by "The Unscrambler®9.7" (CAMO AS, Oslo, Norway).

2.4.2 Methodology of least-square support vector machine (LS-SVM). LS-SVM is an evolution of the standard support vector machine.[27] With the capability for both linear and nonlinear multivariate calibration, the LS-SVM can solve the multivariate calibration problems in a relatively fast way. To obtain the support vectors, a linear set of equations is used instead of a quadratic programming problem.[28] To evaluate the performances of different kernel functions, linear kernel and RBF kernel were compared, respectively. The linear kernel type is the simplest and most efficient kernel to perform similarity calculation. RBF kernel is a nonlinear function and a more compact supported kernel. It can reduce the computational complexity of the training procedure while giving good performance under general smoothness assumptions. Two important kernel parameters of LS-SVM need to be considered. $gam(c)$ is a regularization parameter for both linear kernel and RBF kernel. $sig^2(r^2)$ represents the bandwidth in the case of the RBF kernel.[29] We employed grid-search technique and leave-one-out cross validation to find out the optimal parameter values. In grid-search process, RMSECV was calculated and the optimum

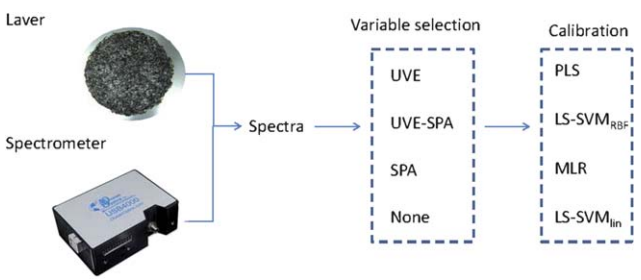

Fig. 1 Flow program of chemometric calculation process.

values of two parameters were selected when they produced smaller RMSECV. Details of the LS-SVM algorithm can be found in the literature.[18] LS-SVM toolbox (LS-SVM v 1.5, Suykens, Leuven, Belgium) was applied with MATLAB to derive all of the LS-SVM models.

2.4.3 Methodology of multiple linear regression (MLR). MLR is a commonly used calibration algorithm with features of being simple and easy to be interpreted. However it fails when the variable number is more than sample number and can be easily affected by the collinearity problems.[30] In this study the number of whole Vis-SNIR spectral wavelength variables is larger than sample number. Therefore it is not possible to run MLR directly on the WRS variables. Therefore the effective variable selection is necessary before MLR model establishment. Moreover, the selected variables with less collinearity would be helpful to improve the MLR model. The calculation of MLR was implemented by "The Unscrambler®9.7" (CAMO AS, Oslo, Norway).

2.5 Model evaluation standard

In this study, the performances of all established spectral models were evaluated in terms of the root mean square error of calibration (RMSEC) for the calibration set and root mean square error of prediction (RMSEP) and residual predictive deviation (RPD) for the prediction set. The large difference between RMSEC and RMSEP means that the model is overfitting. The coefficients of determination (r^2) were used for the evaluation of both calibration (r_{cal}^2) and prediction (r_{pre}^2) process. RPD is the standard deviation of reference data for the prediction samples divided by the standard error of prediction (SEP) and provides a standardization of the SEP.[31] Generally, a good model should have higher r_{cal}^2, r_{pre}^2 and RPD value, and lower RMSEC and RMSEP values.

3 Results and discussion

3.1 Choosing the best pretreatment algorithm

WRS-PLS models were established based on different pretreatment algorithms (Table 2). The SG smoothing, 1st D and 2nd D with the segment sizes of 35, 45 and 55, were calculated. The best results were obtained when the segment sizes were 35, 35 and 45 for SG smoothing, 1st D and 2nd D respectively (only the best results are shown in Table 2). After the SG smoothing, the result was not improved compared to that of WRS-PLS models with original spectra. Therefore SG smoothing was not considered to be combined with other pretreatment algorithms. This is due to the original spectra having less noise, which proved that the USB4000 Miniature Fiber Optic Spectrometer can obtain the spectra of dried laver with high quality. The best result was obtained based on the pretreatment of SNV. Its r_{pre}^2 was 0.8020, RMSEP was 0.7316, and RPD was 2.1507. The results of WRS-SNV-PLS model, MSC-WRS-PLS model and 1st D-WRS-PLS model were similar. The RPD values of MSC-WRS-PLS model and 1st D-WRS-PLS model were 98.25% and 97.21% that of the WRS-SNV-PLS model. The result of the 2nd D model was worse than that of other pretreatments. This is because the 2nd D enlarged some hidden feature peaks but also induced much noise into of the spectra of dried laver. Therefore, further analysis was done based on the spectra pretreated by the best pretreatment algorithm of SNV.

LS-SVM is a powerful spectral nonlinear calibration algorithm. To compare the result obtained by PLS, we calculated the SNV pretreated spectra based on LS-SVM model with linear

Table 2 Prediction results of protein content in dried laver using vis-SNIR vpectroscopy with chemometrics (available range of protein: 27.60–34.20%)

Pretreatment	Variable selection algorithm		Variable number	Calibration	Latent variables	Calibration		Prediction		
						r^2_{cal}	RMSEC	r^2_{pre}	RMSEP	RPD
/	/	/	2900	PLS	7	0.7486	0.8093	0.7227	0.8438	1.8690
SG smoothing	/	/	2900	PLS	7	0.7488	0.8089	0.7233	0.8426	1.8728
SNV	/	/	2900	PLS	8	0.8413	0.6430	0.8020	0.7316	2.1507
MSC	/	/	2900	PLS	8	0.8411	0.6434	0.7971	0.7445	2.1130
1st D	/	/	2900	PLS	5	0.8148	0.6945	0.7934	0.7489	2.0908
2nd D	/	/	2900	PLS	2	0.7246	0.8470	0.6334	0.9624	1.6314
SNV	/	/	2900	LS-SVM_{RBF}	/	0.8993	0.5237	0.8163	0.6819	2.3161
SNV	/	/	2900	LS-SVM_{lin}	/	0.9857	0.1986	0.7820	0.8620	1.8205
SNV	/	SPA	9	PLS	7	0.7812	0.7550	0.7139	0.9728	1.6894
SNV	/	SPA	9	LS-SVM_{RBF}	/	0.8071	0.7097	0.7305	0.9223	1.7753
SNV	/	SPA	9	LS-SVM_{lin}	/	0.7824	0.7531	0.7219	0.9479	1.7304
SNV	/	SPA	9	MLR	/	0.7831	0.7517	0.7183	0.9710	1.6883
SNV	UVE_{PLS}	/	217	PLS	7	0.8619	0.5999	0.8344	0.6536	2.4036
SNV	UVE_{PLS}	/	217	LS-SVM_{RBF}	/	0.9039	0.5032	0.8523	0.6178	2.5467
SNV	UVE_{PLS}	/	217	LS-SVM_{lin}	/	0.8937	0.5278	0.8466	0.6345	2.4813
SNV	UVE_{PLS}	SPA	13	PLS	8	0.8718	0.5778	0.8321	0.6589	2.3958
SNV	UVE_{PLS}	SPA	13	LS-SVM_{RBF}	/	0.8781	0.5649	0.8410	0.6385	2.4585
SNV	UVE_{PLS}	SPA	13	LS-SVM_{lin}	/	0.8752	0.5706	0.8300	0.6667	2.3568
SNV	UVE_{PLS}	SPA	13	MLR	/	0.8761	0.5680	0.8292	0.6756	2.3320

kernel (LS-SVM_{lin}) and LS-SVM model with RBF kernel (LS-SVM_{RBF}). The result is shown in Table 2. Compared to the result of SNV-WRS-PLS model, SNV-WRS-LS-SVM_{RBF} model's RMSEC decreased 18.55%, RMSEP decreased 6.79%, while RPD increased by 7.69%, showing that LS-SVM_{RBF} model was better than PLS model. However, when linear kernel was used, LS-SVM model's result became poorer than LS-SVM_{RBF} model and PLS model. Compared to the result of SNV-WRS-PLS model, SNV-WRS-LS-SVM_{lin} model's RMSEP increased 17.82% while RPD decreased by 15.35%. When the absolute difference values between RMSEC and RMSEP were calculated, the values of 0.1582 and 0.6635 were obtained for LS-SVM_{RBF} model and LS-SVM_{lin} model, respectively. However the value of PLS model was 0.0886, only 56.01% and 13.35% compared to those of LS-SVM_{RBF} model and LS-SVM_{lin} model. The result shows that LS-SVM_{RBF} model and LS-SVM_{lin} were more overfitting than PLS model. Therefore when WRS were considered, LS-SVM_{RBF} might be able to obtain better result but PLS model was more robust in this study.

3.2 SPA calculation based on the WRS

SPA was carried out for selecting effective wavelength variables from the WRS. Fig. 2 shows the RMSE scree plot obtained by SPA. Fig. 2 was used for the explanation of the selection procedure by SPA, and the distribution of selected variables in the spectral curve plot. As can be seen, the trends of RMSE curves become marginal in the starting part as the numbers of selected variables were from 1 to 6. Then a sharp fall is shown when the numbers of selected variables were from 6 to 7. The curve tends to level off after the determination of selected variables by F-test criterion with $\alpha = 0.25$.[9] The solid circle shows the selected variable number of nine. Therefore nine variables (RMSE = 0.83614) variables were selected. The selected wavelength variables were set as the input variables of MLR, PLS, LS-SVM_{RBF} and LS-SVM_{lin} models, respectively. The results are shown in Table 2. SNV-SPA-LS-SVM_{RBF} model obtained the best result. Its r_{pre}^2 was 0.7305, RMSEP was 0.9223, and RPD was 1.7753. However, the other three algorithms obtained similar results too. The RPD values of SNV-SPA- LS-SVM_{lin} model, SNV-SPA-PLS model and SNV-SPA-MLR model were 97.47%, 95.16% and 95.10% of that of SNV-SPA-LS-SVM_{RBF} model.

After the variable selection using SPA, although the variable numbers were much reduced (9 *vs.* 2900), the performances of PLS, LS-SVM_{RBF} and LS-SVM_{lin} calibration models became worse (MLR cannot be established based on the WRS). SNV-SPA-LS-SVM_{RBF} model had a RPD value of 1.7753, 23.35% decrease compared to that of SNV-WRS-LS-SVM model. The SNV-SPA-LS-SVM_{lin} model had a RPD value of 1.7304, a 4.95% decrease compared to that of SNV-WRS-LS-SVM_{lin} model. SNV-SPA-PLS model had a RPD value of 1.6894, 21.45% decrease compared to that of SNV-WRS-PLS model. It might be because SPA was operated on the whole spectra which have the low S/N.[10] Moreover, the SPA operation based on the whole spectra with thousands of variables is time-consuming. Thus it might be possible to improve the SPA performance and reduce the calculation time by eliminating uninformation variables before SPA.

3.3 UVE_{PLS} calculation based on the WRS

Fig. 3 shows the stability of each wavelength variable based on UVE_{PLS} with 10 LVs. Wavelength variables are on the left of the vertical line, while random variables are on the right side. The two horizontal lines are the lower and upper cutoffs. The variable whose stability is within the cutoff lines is treated as uninformative and be eliminated. Finally 217 wavelength variables were selected from 2900 WRS variables. That means 92.52% of the WRS variables were eliminated. The selected wavelength variables were used to establish the PLS, LS-SVM_{RBF} and LS-SVM_{lin} models, respectively. The result of SNV-UVE_{PLS}-LS-SVM_{RBF} model was the best. Its r_{pre}^2 was 0.8523, RMSEP was 0.6178 and RPD was 2.5467. The other two algorithms obtained similar results. The RPD values of SNV-UVE_{PLS}-LS-SVM_{lin} and SNV-UVE_{PLS}-PLS models were 97.43% and 94.38% of that of SNV-UVE_{PLS}-LS-SVM_{RBF} model.

UVE_{PLS} has improved the prediction result compared to that based on WRS. When PLS was used as the calibration algorithm, the RPD value of SNV-UVE_{PLS}-PLS model was 2.4036, 11.76% higher than that of the SNV-WRS-PLS model. When LS-SVM_{RBF} was used as the calibration algorithm, the RPD value of SNV-UVE_{PLS}-LS-SVM_{RBF} model was 2.5467, 9.96% higher than that of SNV-WRS-LS-SVM model. When LS-SVM_{lin} was used as the calibration algorithm, the RPD value of SNV-UVE_{PLS}-LS-SVM_{lin} model was 2.4813, 36.30% higher than that

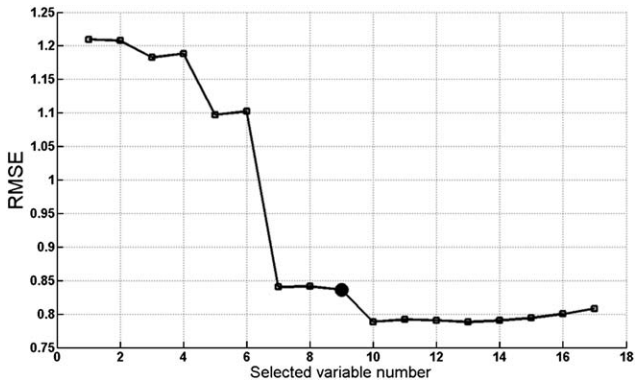

Fig. 2 RMSE scree plot of SPA operated based on the whole range spectra.

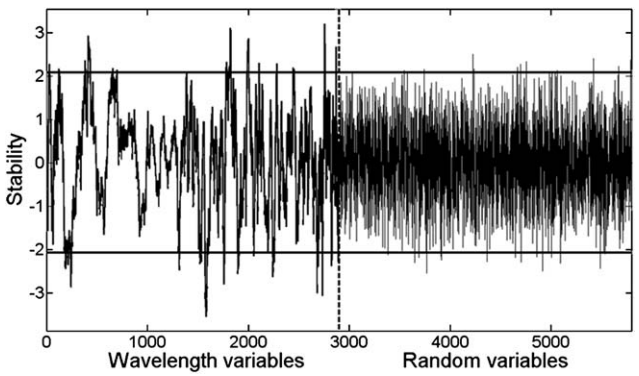

Fig. 3 Stability of each variable by UVE_{PLS} with 10 LVs. Two horizontal lines indicate the lower and upper cutoff.

of SNV-WRS-LS-SVM model. The average improvement percentage was 19.34%. Moreover, 92.52% of the WRS variables were eliminated after the process of UVE_{PLS} algorithm. Therefore, variables with no more information for modeling than noise were eliminated after UVE_{PLS} analysis.

3.4 SPA calculation based on UVE_{PLS} selected spectra

Although over ninety percent of WRS variables were eliminated after UVE_{PLS} process, there were still more than one hundred variables remaining. In order to obtain fewer variables and to make the model more simple and interpretable, SPA was used to further select the effective variables based on the selected variables of UVE_{PLS}. Based on the same F-test criterion, thirteen variables were selected. The selected thirteen wavelength variables were set as the inputs of the MLR, PLS, LS-SVM_{RBF} and LS-SVM_{lin} models, spectively. The results are shown in Table 2. SNV-UVEPLS-SPA-LS-SVM_{RBF} model obtained the best result. Its r_{pre}^2 was 0.8410, RMSEP was 0.6385, and RPD was 2.4585. The other three calibration algorithms obtained similar results. The RPD values of SNV-UVE_{PLS}-LS-SVM_{lin}, SNV-UVE_{PLS}-SPA-PLS model and SNV-UVE_{PLS}-SPA-MLR model were 95.86%, 97.45% and 94.85% of that of SNV-UVE_{PLS}-SPA-LS-SVM model.

The uninformation variable elimination by UVE_{PLS} was useful to improve the results of SPA (Table 2). When MLR was used as the calibration algorithm, the RPD value of SNV-UVE_{PLS}-SPA-MLR model was 2.3320, 38.13% higher than that of SNV-SPA-MLR model. When PLS was used as the calibration algorithm, the RPD value of SNV-UVE_{PLS}-SPA-PLS model was 2.3958, 41.81% higher than that of SNV-SPA-PLS model. When LS-SVM_{RBF} was used as the calibration algorithm, the RPD value of SNV-UVE_{PLS}-SPA-LS-SVM_{RBF} model was 2.4585, 38.48% higher than that of SNV-SPA-LS-SVM_{RBF} model. When LS-SVM_{lin} was used as the calibration algorithm, the RPD value of SNV-UVE_{PLS}-SPA-LS-SVM_{lin} model was 2.3568, 36.20% higher than that of SNV-SPA-LS-SVM_{lin} model.

3.5 Discussion

Our results show that UVE_{PLS} can improve the results of SPA based on all four calibration algorithms in this case (Fig. 4).

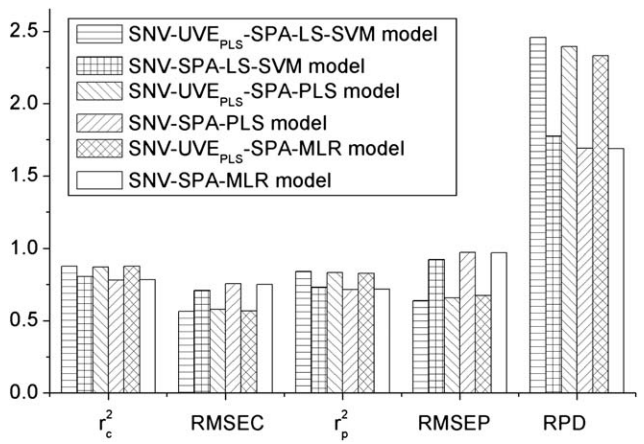

Fig. 4 Comparison of the evaluation parameters of different models based on UVE_{PLS}-SPA and SPA respectively.

When SPA was calculated directly on the WRS, the average RPD value of the models based on three calibration algorithms was 1.7208. That means the model can only discriminate between low and high values of the response variable.[32] However when UVE_{PLS} was used to eliminate uninformation variables and the retained variables were selected by SPA, the average RPD value was 2.3858, and shows that coarse quantitative prediction is possible.[32] It can be seen that the average improvement percentage of RPD was 38.66% by calculating SPA on UVE_{PLS} processed variables. Also the average increased rate of r_{cal}^2 and r_{pre}^2 of SNV-UVE_{PLS}-SPA-MLR model was 13.66%, and the average decreased rate of RMSEC and RMSEP was 27.43%, compared to those of SNV-SPA-MLR model. The average increased rate of r_{cal}^2 and r_{pre}^2 of SNV-UVE_{PLS}-SPA-PLS model was 14.08%, and the average decreased rate of RMSEC and RMSEP was 27.87%, compared to those of SNV-SPA-PLS model. The average increased rate of r_{cal}^2 and r_{pre}^2 of SNV-UVE_{PLS}-SPA-LS-SVM_{RBF} model was 11.96%, and the average decreased rate of RMSEC and RMSEP was 25.59%, compared to those of SNV-SPA-LS-SVM_{RBF} model. The average increased rate of r_{cal}^2 and r_{pre}^2 of SNV-UVE_{PLS}-SPA-LS-SVM_{lin} model was 13.42%, and the average decreased rate of RMSEC and RMSEP was 26.95%, compared to those of SNV-SPA-LS-SVM_{lin} model. Therefore it was proved that there was much improvement of using UVE_{PLS} on SPA spectral multivariable selection with both linear and nonlinear calibration algorithms in this case.

Both linear calibration algorithms of MLR and PLS and non-linear calibration algorithms of LS-SVM_{RBF} and LS-SVM_{lin} obtained similar results based on all variable selection strategies of SPA, UVE_{PLS} and UVE_{PLS}-SPA. LS-SVM did not obtain better results compared to other linear calibration algorithms in this study, although other papers show that LS-SVM usually can obtain better results than PLS and MLR.[4,18,33] MLR and PLS are simpler for use in practice than LS-SVM. They were suggested to be used for the spectral calibration in this study. MLR would be more preferred when SPA and UVE_{PLS}-SPA are used for the variables selection which can make the number of variables less than the number of samples, and can make MLR available for the calibration. Therefore, for different applications, both linear and non-linear calibration algorithms should be analyzed to choose the best one.

4 Conclusions

The improvement ability of using UVE_{PLS} on SPA for the spectral multivariable selection was evaluated for the rapid and non-destructive determination of protein content in dried laver. Different spectral pretreatment algorithms (SG smoothing, SNV, MSC, 1-Der and 2-Der), variable selection algorithms (UVE_{PLS}, SPA and UVE_{PLS}-SPA) and calibration algorithms (MLR, PLS, LS-SVM_{RBF} and LS-SVM_{lin}) were analyzed. The results proved that it was necessary to operate UVE_{PLS} before SPA, which could both reduce the calculation time and improve the model's performance. Moreover, Vis-SNIR spectroscopy was successfully utilized for the protein determination in dried laver. The RPD values of both linear and non-linear models based on the thirteen selected variables of UVE_{PLS}-SPA showed that coarse quantitative predictions were possible. We hope that

the results obtained by us will help both further chemometric investigations (multivariate selection and calibration analysis) and investigations in the sphere of vibrational (Near infrared, Mid-infrared and Raman) spectroscopy of multicomponent systems.

Acknowledgements

This study was supported by Scientific Research Fund of Zhejiang Provincial (ZJNSFC, No. Y3110289) and Research Fund of Wenzhou Technology Projects (No. G20100078).

References

1 H. Y. Cen and Y. He, *Trends Food Sci. Technol.*, 2007, **18**, 72–83.
2 C. Abrahamsson, J. Johansson, A. Sparen and F. Lindgren, *Chemom. Intell. Lab. Syst.*, 2003, **69**, 3–12.
3 D. Wu, X. Chen, P. Shi, S. Wang, F. Feng and Y. He, *Anal. Chim. Acta*, 2009, **634**, 166–171.
4 D. Wu, Y. He, J. Shi and S. Feng, *J. Agric. Food Chem.*, 2009, **57**, 1697–1704.
5 X. Chen and X. Lei, *J. Agric. Food Chem.*, 2009, **57**, 334–340.
6 X. Chen, D. Wu, Y. He and S. Liu, *Anal. Chim. Acta*, 2009, **638**, 16–22.
7 D. Wu, Y. He and S. Feng, *Anal. Chim. Acta*, 2008, **610**, 232–242.
8 M. C. U. Araujo, T. C. B. Saldanha, R. K. H. Galvao, T. Yoneyama, H. C. Chame and V. Visani, *Chemom. Intell. Lab. Syst.*, 2001, **57**, 65–73.
9 M. C. Breitkreitz, I. M. Raimundo, J. J. R. Rohwedder, C. Pasquini, H. A. Dantas, G. E. Jose and M. C. U. Araujo, *Analyst*, 2003, **128**, 1204–1207.
10 S. F. Ye, D. Wang and S. G. Min, *Chemom. Intell. Lab. Syst.*, 2008, **91**, 194–199.
11 L. Norgaard, A. Saudland, J. Wagner, J. P. Nielsen, L. Munck and S. B. Engelsen, *Appl. Spectrosc.*, 2000, **54**, 413–419.
12 R. Leardi and L. Norgaard, *J. Chemom.*, 2004, **18**, 486–497.
13 V. Centner, D. L. Massart, O. E. deNoord, S. deJong, B. M. Vandeginste and C. Sterna, *Anal. Chem.*, 1996, **68**, 3851–3858.
14 D. Wu, Y. He, P. C. Nie, F. Cao and Y. D. Bao, *Anal. Chim. Acta*, 2010, **659**, 229–237.
15 K. H. Ogbonda, R. E. Aminigo and G. O. Abu, *Bioresour. Technol.*, 2007, **98**, 2207–2211.
16 L. Qin, X. J. Shen, J. H. Chen and S. J. Zhu, *Spectrosc Spect Anal*, 2010, **30**, 635–639.
17 Q. Sun, J. H. Wang and D. H. Han, *Spectrosc Spect Anal*, 2009, **29**, 1818–1821.
18 D. Wu, Y. He, S. J. Feng and D. W. Sun, *J. Food Eng.*, 2008, **84**, 124–131.
19 A. Bittner, R. Marbach and H. M. Heise, *J. Mol. Struct.*, 1995, **349**, 341–344.
20 J. Reeves, III, *J. Near Infrared Spectrosc.*, 1994, **2**, 199–212.
21 A. Savitzky and M. J. E. Golay, *Anal. Chem.*, 1964, **36**, 1627.
22 R. J. Barnes, M. S. Dhanoa and S. J. Lister, *Appl. Spectrosc.*, 1989, **43**, 772–777.
23 I. S. Helland, T. Naes and T. Isaksson, *Chemom. Intell. Lab. Syst.*, 1995, **29**, 233–241.
24 X. Chen, H. Li, D. Wu, X. Lei, X. Zhu and A. Zhang, *Eur. Food Res. Technol.*, 2010, **230**, 981–988.
25 R. K. H. Galvao, M. C. U. Araujo, W. D. Fragoso, E. C. Silva, G. E. Jose, S. F. C. Soares and H. M. Paiva, *Chemom. Intell. Lab. Syst.*, 2008, **92**, 83–91.
26 R. W. Gerlach, B. R. Kowalski and H. O. A. Wold, *Anal. Chim. Acta*, 1979, **3**, 417–421.
27 J. A. K. Suykens, T. van Gestel, J. de Brabanter, B. de Moor and J. Vandewalle, *Least-Squares Support Vector Machines*. World Scientific, Singapore, 2002.
28 J. Z. Li, H. X. Liu, X. J. Yao, M. C. Liu, Z. D. Hu and B. T. Fan, *Anal. Chim. Acta*, 2007, **581**, 333–342.
29 D. Wu, H. Yang, X. Chen, Y. He and X. Li, *J. Food Eng.*, 2008, **88**, 474–483.
30 T. Naes and B. H. Mevik, *J. Chemom.*, 2001, **15**, 413–426.
31 P. C. Williams, *Near-Infrared Technology in the Agricultural and Food Industries*. American Association of Cereal Chemists, Saint Paul, MN, USA, 2001.
32 B. M. Nicolai, K. Beullens, E. Bobelyn, A. Peirs, W. Saeys, K. I. Theron and J. Lammertyn, *Postharvest Biol. Technol.*, 2007, **46**, 99–118.
33 D. Wu, P. C. Nie, J. Cuello, Y. He, Z. P. Wang and H. X. Wu, *J. Food Eng.*, 2011, **102**, 278–286.

