Food Chemistry 115 (2009) 1430–1436

Contents lists available at ScienceDirect

Food Chemistry

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

Application of successive projections algorithm for variable selection to determine organic acids of plum vinegar

Fei Liu, Yong He *

College of Biosystems Engineering and Food Science, Zhejiang University, 268 Kaixuan Road, Hangzhou, Zhejiang 310029, China

ARTICLE INFO

Article history:
Received 13 September 2008
Received in revised form 4 December 2008
Accepted 25 January 2009

Keywords:
Visible and near infrared spectroscopy
Successive projections algorithm
Variable selection
Least squares-support vector machine
Plum vinegar
Organic acids

ABSTRACT

Visible and near infrared (Vis/NIR) spectroscopy was investigated to determine the acetic, tartaric and lactic acids of plum vinegar based on a newly proposed combination of successive projections algorithm-least squares-support vector machine (SPA-LS-SVM). SPA, compared with regression coefficients (RC), was applied to select effective wavelengths (EWs) with least collinearity and redundancies. Five concentration levels (100%, 80%, 60%, 40% and 20%) of plum vinegar were studied. Multiple linear regression (MLR) and partial least squares (PLS) models were developed for comparison. The results indicated that SPA-LS-SVM achieved the optimal performance for three acids comparing with full-spectrum PLS, SPA-MLR, SPA-PLS, RC-PLS and RC-LS-SVM. The root mean square error of prediction (RMSEP) was 0.3581, 0.0714 and 0.0201 for acetic, tartaric and lactic acids, respectively. The overall results indicated that Vis/NIR spectroscopy incorporated to SPA-LS-SVM could be applied as an alternative fast and accurate method for the determination of organic acids of plum vinegars.

1. Introduction

Nowadays, visible and near infrared (Vis/NIR) spectroscopy is widely employed as alternatives to wet chemistry procedures for qualitative and quantitative analysis in many fields, such as agriculture, pharmaceuticals, food, textiles, cosmetics, and polymer production industry (Yan, Zhao, Han, & Yang, 2005). The NIR spectroscopic analysis is largely depend on chemometric methods for the quantitative analysis of multicomponent systems or mixtures because side bands occur as a result of overtones and combination bands of fundamental vibrations (Bokobza, 1998). Among these methods, the most used are multiple linear regression (MLR), principle component regression (PCR) and partial least squares (PLS) (Martens & Naes, 1993). These methods can only handle the linear relationship between spectral data and chemical components. Whereas it is known that some latent nonlinear information is existed in the spectral data. In order to make advantage of the nonlinear information as well as the most linear information, some chemometrics, such as artificial neural network (ANN) (Despagne & Massart, 1998) and least squares-support vector machine (LS-SVM) (Suykens, Van Gestel, De Brabanter, De Moor, & Vandewalle, 2002; Suykens & Vandewalle, 1999) are proposed to solve these problems. Since spectral matrices have large amount of data, they are too complicated to be trained directly in the ANN or SVM models. This procedure is time-consuming and not convenient to fulfill the high speed features of spectroscopic techniques. Moreover, the full spectral regions may include wavelengths or wavelength bands which contribute more collinearity, redundancies and noise than relevant information to models (Ye, Wang, & Min, 2008). Hence, some variable selection methods have been proposed for the development of a parsimonious model for the quantitative and qualitative analysis.

The commonly used variable selection methods are as follows: the least condition number of the calibration matrix (Otto & Wegscheider, 1985), generalized simulated annealing (Kalivas, Roberts, & Sutter, 1989), genetic algorithm (Jouan-Rimbaud, Massart, Leardi, & De Noord, 1995), correlation coefficients and B-matrix coefficients (Min & Lee, 2005), x-loading weights (Esbensen, 2002; Liu, He, Wang, & Pan, 2007), wavelet transforms (Alsberg, Woodward, Winson, Rowl, & Kell, 1998), regression coefficients (Liu, He, & Wang, 2008a,b; Liu et al., 2007), independent component analysis (Hyvärinen, Karhunen, & Oja, 2001; Liu et al., 2008b), modeling power (Liu et al., 2008b; Sagrado & Cronin, 2008) and uninformative variable elimination (Centner et al., 1996). The successive projections algorithm (SPA) is a newly developed variable selection strategy for MLR and PLS calibration procedures (Araújo et al., 2001; Galvão et al., 2008). Herein, a new combination of SPA-LS-SVM is proposed for the determination of organic acids of plum vinegar using Vis/NIR spectra. SPA-LS-SVM was thought to be a powerful calibration method using the selected relevant variables as well as the linear and nonlinear spectral information. Recently, NIR spectroscopy has been applied for the discrimination of aging of vinegar during storage (Casale,

* Corresponding author. Tel.: +86 571 86971143; fax: +86 571 86971143.
E-mail address: yhe@zju.edu.cn (Y. He).

0308-8146/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.foodchem.2009.01.073

Abajo, Sáiz, Pizarro, & Forina, 2006), and prediction of chemical constituents such as organic acids during storage and aging (Sáiz-Abajo, González-Sáiz, & Pizarro, 2006), reducing sugars (Fu, Yan, Chen, & Li, 2005), total procyanidins (García-Parrilla, Heredia, Troncoso, & González, 1997), soluble solids content and pH of rice vinegar (Liu et al., 2008a) and discrimination of fruit vinegar varieties (Liu et al., 2008b). However, up to our knowledge, there were few reports on the determination of acetic, tartaric and lactic acids of plum vinegar using Vis/NIR spectroscopy and few studies focused on the incorporated SPA-LS-SVM method.

The objective of this paper is (1) to study the feasibility of using Vis/NIR spectroscopy to determine acetic, tartaric and lactic acids of plum vinegar; (2) to compare and evaluate the effective wavelengths (EWs) selected by successive projections algorithm (SPA) and regression coefficients (RC) and (3) to evaluate the newly proposed combination of SPA-LS-SVM method compared with MLR and PLS.

2. Materials and methods

2.1. Sample preparation

Plum vinegars with different batches of producing times were obtained in local market and they were all fermented vinegars. All plum vinegar samples were stored in the laboratory at a constant temperature of 25 ± 1 °C for more than 48 h to equalize the temperature. The original vinegar was diluted by distilled water. Five concentration levels (100%, 80%, 60%, 40% and 20%, v/v%) of original plum vinegar were prepared for the experiment. The dilution operation was based on the following four reasons. Firstly, the original plum vinegar had a high concentration of acids and it was not suitable for direct drink. The instructions of drinking plum vinegar also suggested using dilution operation for three to five times of original vinegar. This could meet the tastes of different people. Secondly, the dilution could expand the ranges of acid concentration. The wide range could make the samples in calibration set more general and robust. The samples of different batches were thought to be independent samples. Hence, the development model could be more stable and robust. Thirdly, the predictive ability of calibration model would be more powerful with the wide ranges of acid concentration. The prediction precision and generalization was higher because the calibration model covered a large range of acid concentration level. Fourthly, the develop model would be more suitable for *in situ* fermentation monitoring because the concentration of acid is not one fixed value and the concentration is kept changing during the fermentation stage. Sixty samples for each level and a total of 300 samples were prepared for spectral and chemical analysis. All vinegar samples were randomly divided into three data sets. The calibration set consisted of 150 samples with 30 for each level, the validation set consisted of 75 samples with 15 for each level, and the remaining samples were separated for prediction set. No single sample was used in calibration, validation and prediction sets at the same time. The calibration and validation sets were used for model-building, and the prediction set was applied for performance evaluation purpose.

2.2. Spectral collection and preprocessing

A handheld FieldSpec Pro FR (325–1075 nm)/A110070 spectroradiometer with Trademarks of Analytical Spectral Devices, Inc. (Analytical Spectral Devices, Boulder, USA) was applied for the spectral scanning. The field-of-view (FOV) of the spectroradiometer is 25°. The light source consists of a Lowell pro-lam interior light source assemble/128930 with Lowell pro-lam 14.5 V Bulb/128690 tungsten halogen bulb that could be used both in visible and near infrared region (325–1075 nm). The energy of light source could be adjusted according to the standard curve of spectroradiometer. The transmission mode was applied in this experiment. Fruit vinegar sample was placed in a cuvette with a 2 mm light path length. The transmission spectra were measured from 325 to 1075 nm with an average reading of 30 scans for each spectrum. For each sample, three replicate spectra were collected and the averaged spectrum of these three replicates was used as the data of this sample. All spectral data were stored in a personal computer and processed using the RS^3 software for Windows (Analytical Spectral Devices, Boulder, USA) designed with a Graphical User Interface.

Before the calibration stage, the spectral data should be preprocessed for an optimal performance. The transmission spectra were transformed into absorbance (absorbance = log(1/T)). The pretreatments were implemented by "The Unscrambler® 9.6" (CAMO AS, Oslo, Norway). The influence of the following data preprocessing methods had been compared including Savitzky–Golay smoothing (SG), multiplicative scatter correction (MSC), standard normal variate (SNV), and first and second derivative (1-derivative and 2-derivative).

2.3. Organic acid analysis

The reference value of organic acids were determined by high performance liquid chromatography (HPLC) with Trademarks of Waters 2695 (Waters, Milford, MA, USA). All plum vinegar samples were centrifuged at 10,000 rpm, and then were filtered through a 0.22 μm membrane prior to HPLC analysis. The instrument was equipped with a 717+ automatic injector with 20 μL once, and a 2996 Photodiode Array detector (PDA) UV at 210 nm. Separation was achieved using an Agilent Zorbax SB-C18 column (5 μm, 4.6 × 250 mm) at 25 °C. The mobile phase was 5/95 (v/v) methanol/water with flow rate of 1.0 mL/min. 0.01 mol/L potassium dihydrogen phosphate solution was used with pH of 2.6. The running time was set as 8 min. The settings were based on several trials and previous similar studies (Gao, Liao, Wang, & Hu, 2004). All organic acids were recorded on a computer-based data system.

2.4. Successive projections algorithm

SPA is a forward variable selection algorithm for multivariate calibration (Araújo et al., 2001; Galvão et al., 2008). SPA performs simple projection operations in a vector space to obtain subsets of useful variables with small collinearity. The main points are summarized here. Firstly, set the maximum number of variables *N* to be selected before a start vector is chosen in a space of *n*-dimensions (where *n* is the number of original variables). Subsequently, in an orthogonal sub-space, the vector of higher projection is selected and becoming the new starting vector. The choice of the orthogonal sub-space at each iteration is made in order to select only the non-collinear variables. The optimal initial variable and number of variables can be determined on the basis of the smallest root mean square error of validation (RMSEV) in validation set of MLR calibration. The details of SPA could be found in the literatures (Araújo et al., 2001; Galvão et al., 2008). Moreover, the selected variables, named EWs, could be used as the inputs of MLR, PLS and LS-SVM models.

2.5. PLS and RC analysis

Partial least squares (PLS) analysis is a widely utilized multianalysis and regression method for the spectroscopy analysis (Geladi & Kowalski, 1986; Martens & Naes, 1993). PLS analysis can be applied to develop a calibration model to progress the relationship between the spectral data and organic acids of plum vinegar. PLS analysis was also used as a way to select the effective wavelengths (EWs) by regression coefficient (RC) analysis. The regression coefficients calculated from the spectral data could calculate the response value *Y*-variables from the *X*-variables. The size of the coefficients gave an indication of which variables had the important impact on the response variables (*Y*). Its task was to find which variables were important for predicting *Y*-variable. Large absolute values indicated the importance and the significance of the effect on the prediction of *Y*-variable preference. The regression coefficients were calculated by the software "The Unscrambler® 9.6". Hence, these selected EWs could be employed as the input data matrix of PLS and LS-SVM models.

2.6. LS-SVM

Least squares-support vector machine (LS-SVM) is a state-of-the-art learning algorithm with a good theoretical foundation in statistical learning method. LS-SVM is capable of dealing with linear and nonlinear multivariate analysis and resolving these problems in a relatively fast way (Suykens & Vandewalle, 1999; Suykens et al., 2002). It employs a set of linear equations instead of quadratic programming (QP) problems to obtain the support vectors (SVs). SVM embodies the structural risk minimization (SRM) principle instead of traditional empirical risk minimization (ERM) principle to avoid overfitting problems. The details of LS-SVM algorithm could be found in the literatures (Guo, Liu, & Wang, 2006; Suykens et al., 2002). Before the application of LS-SVM, three crucial problems were required to solve, including the optimal input data set, proper kernel function and the optimal LS-SVM parameters. The optimal inputs had been settled by using the aforementioned EWs. Radial basis function (RBF) kernel as a nonlinear function was a more compacted supported kernel and able to reduce the computational complexity of the training procedure. Simultaneously, RBF kernel could handle the nonlinear relationships between the spectra and target attributes and give a good performance under general smoothness assumptions. Thus, RBF kernel was recommended as the kernel function of LS-SVM in this paper. There were two significant parameters to be decided in the LS-SVM model. The regularization parameter gam (γ) determined the tradeoff between minimizing the training error and minimizing model complexity. The parameter *sig*^{2} (σ^2) of RBF kernel function was the bandwidth and implicitly defines the nonlinear mapping from input space to some high dimensional feature space. In order to obtain the optimal combination of (γ, σ^2), a two-step grid search technique was employed with leave-one-out cross validation to avoid overfitting problems. The ranges of γ and σ^2 within (10^{-2}–10^5) were set based on experience and previous researches (Liu et al., 2008a,b). Grid search tries values of each parameter across the specified search range using geometric steps. The first step grid search was for a crude search with a large step size and the second step for the specified search with a small step size. After the process of grid search, the optimal combination of (γ, σ^2) would be achieved for the LS-SVM models. All the calculations were performed using MATLAB® 7.0 (The Math Works, Natick, USA). The free LS-SVM toolbox (LS-SVM v 1.5, Suykens, Leuven, Belgium) was applied with MATLAB 7.0 to develop the calibration models.

The evaluation indices of predictive capability for all developed models were correlation coefficients (*r*) and root mean square error of validation (RMSEV) and prediction (RMSEP), as in previous papers (Araújo et al., 2001; Galvão et al., 2008). Generally, a good model should have higher correlation coefficients value, lower RMSEV and RMSEP values.

3. Results and discussion

3.1. Spectral features

The raw absorbance spectra of plum vinegar are shown in Fig. 1. As can be seen, the trends of the spectra with different concentration levels were similar, and there were no obvious peaks and valleys in the absorbance spectra within the regions of 400–1000 nm. Moreover, there were some noise at the beginning and end parts of raw spectra. Hence, some pretreatments as stated above were compared for optimal prediction performance. The ranges of acetic acid, tartaric acid and lactic acid were 5.048–25.385, 1.188–6.112 and 0.091–0.539 g L^{-1}, respectively for all five concentration levels of plum vinegar. The reference values of organic acids covered a large scope which was good to develop a stable and robust calibration model.

3.2. Selection of EWs

As stated above, SPA was used for the selection of EWs for the determination of three organic acids of plum vinegar. It was worth noting that the validation set was applied for the guidance of selection of candidate subsets of variables. The prediction set was utilized in the final performance evaluation of the resulting models. It was not applied in any step of the calibration and validation procedures. A SPA-MLR procedure was applied for the calculation of a sequence of root mean square error of validation (RMSEV) values using the selected variable subsets. This process confirmed the achievement of the optimal number of selected EWs with an optimal RMSEV value, and this RMSEV value was not significantly larger than the minimum RMSEV value. The maximum number of selected EWs was set as 30. The influences of different preprocessing methods were compared including SG smoothing, SNV, first and second derivative. The results indicated that the raw spectra without preprocessing obtained the optimal validation results. Herein, only the results using raw spectra were shown for comparison with other calibration methods.

The scree plots for the selected number of EWs of each organic acid by applying SPA are shown in Fig. 2a–c. As can be seen, a sharp fall is shown in the starting part of the RMSEV curve as the number of selected EWs is increased from one to three. This indicated that the number of selected EWs should be three for the resolution of spectral overlapping features of the analytes. The trends of RMSEV curve are still descent after that point but the improvement becomes marginal with further increasing number of selected EWs, and thus the curve tends to level off after the determination of selected EWs by SPA cutoff procedure by *F*-test criterion with $\alpha = 0.25$ (Galvão et al., 2008). The numbers of selected EWs for acetic, tartaric and lactic acids were 12, 15 and 19, respectively. The selected EWs (circle markers) corresponding to raw spectra are shown in Table 1 and Fig. 2d–f for acetic, tartaric and lactic acids, respectively. As can be seen in Table 1, the selected EWs by SPA are sequenced in the order of relevance. This indicated that wavelength at 944 nm was the most relevant variable of 12 selected EWs for the prediction of acetic acid of plum vinegar. Comparing

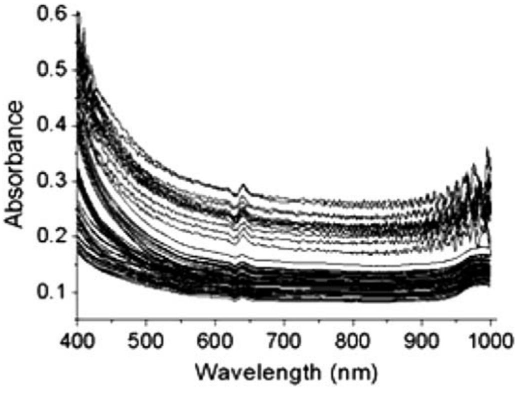

Fig. 1. The absorbance spectra of plum vinegars with five concentration levels.

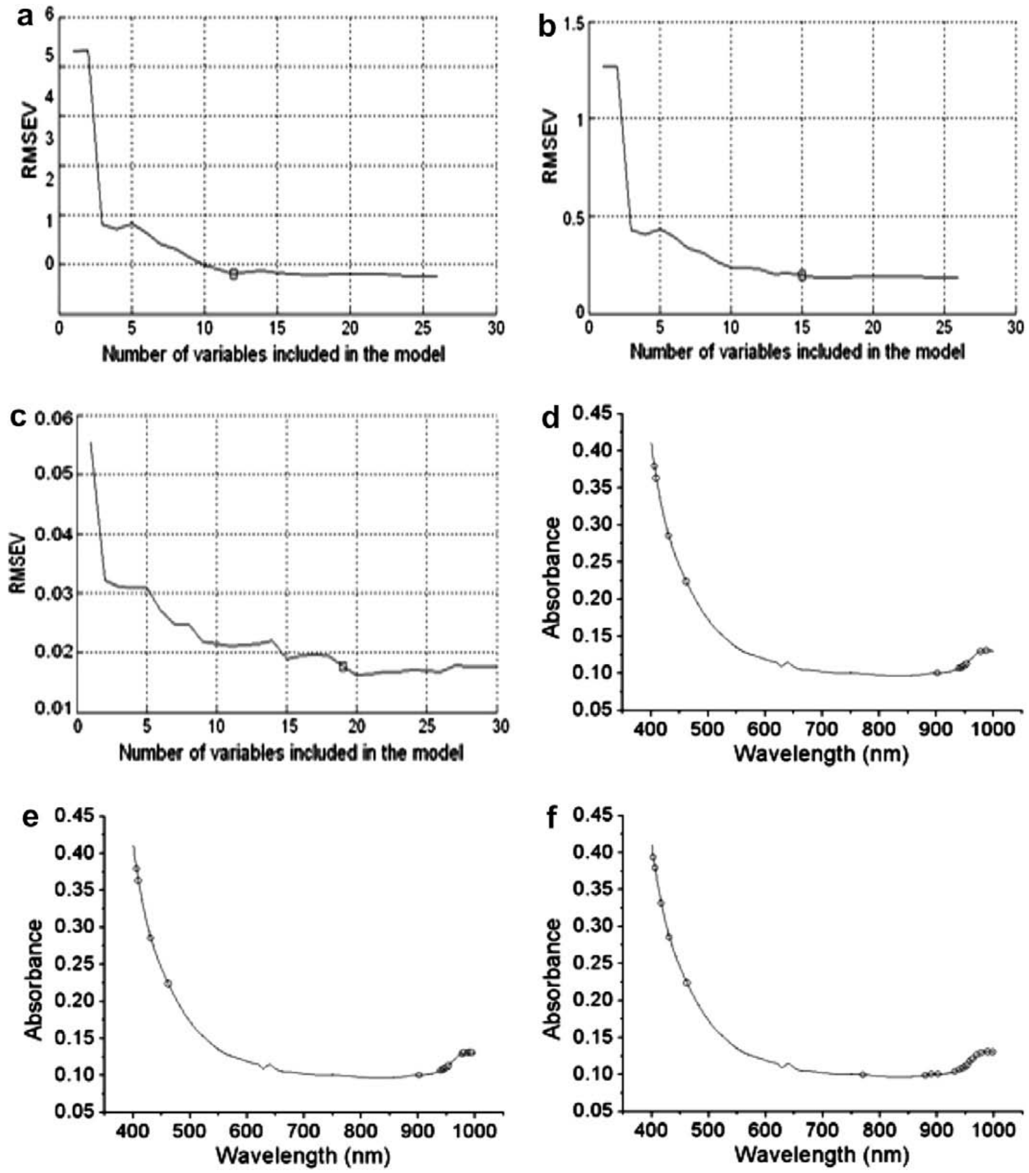

Fig. 2. RMSEV plots obtained for acetic acid (a), tartaric acid (b) and lactic acid (c). The selected EWs (shown in circle markers) for acetic, tartaric and lactic acids are presented in (d), (e), and (f), respectively.

Table 1
The selected EWs for three organic acids by SPA and RC.

Organic acid	Methods	No.	Selected EWs (nm)
Acetic acid	SPA	12	944, 951, 462, 902, 431, 978, 988, 409, 406, 944, 947, 940
	RC	8	406, 417, 918, 925, 944, 979, 993, 998
Tartaric acid	SPA	15	944, 951, 462, 902, 431, 978, 988, 409, 406, 944, 981, 947, 940, 995, 992
	RC	8	406, 417, 918, 925, 944, 979, 993, 996
Lactic acid	SPA	19	431, 771, 963, 959, 940, 970, 880, 891, 951, 932, 954, 403, 902, 998, 989, 417, 978, 406, 946
	RC	6	406, 417, 918, 944, 979, 993

the selected EWs for all three acids, the EWs for acetic acid and tartaric acid were almost the same, except that three more EWs (981, 995 and 992 nm) were chosen for tartaric acid. Moreover, the relevant sequence was same for the first ten EWs (seen in Table 1). The 19 EWs for lactic acid were quite different and only 7 EWs were identical with acetic and tartaric acids. The difference of selected EWs indicated that three different organic acids had different latent spectral features. Similar results could also be discovered with an inspection of RMSEV scree plots from Fig. 2a–c. The curves of Fig. 2a and Fig. 2b were quite similar from the first to the fifth variable (indicated by number 1–5 in Fig. 2a and Fig. 2b). The trend was flat for the first to the second variable, then a sharp fall to the third variable, and a marginal improvement to the fifth variable. From the fifth variable, a gradual descent was appeared to the selected number of variables. However, the RMSEV plots for lactic acid (Fig. 2c) was different from Fig. 2a and

Fig. 2b with a initial sharp fall for the first to the second variable, then a step descent to the selected number of variable.

For comparison, another variable selection method by regression coefficients was applied in this study. PLS analysis was utilized to develop a model with calibration and validation sets. The validation set was used to evaluate the calibration performance, and to achieve the optimal calibration model. The influences of the aforementioned preprocessing were compared, and the results indicated that pretreatment of SG and SNV was necessary for a better performance. Hence, the EWs were selected by RC after PLS analysis. The cutoff criterion was set as ±1.0, ±0.25 and ±0.02 for acetic, tartaric and lactic acids, respectively. This criterion was settled based on experience and previous studies (Liu et al., 2008b). The plots of regression coefficients are shown in Fig. 3a–c for acetic, tartaric and lactic acids, respectively. The dot lines indicated the upper and lower cutoff threshold values. The trends of these three curves were quite similar, and similar EWs were also selected as relevant variables. The selected EWs (circle markers) corresponding to the preprocessed spectrum by SG and SNV are shown in Table 1 and Fig. 3d–f for acetic, tartaric and lactic acids, respectively. The selected EWs for acetic and tartaric acid were almost the same except one EW as 998 nm for acetic acid, whereas 996 for tartaric acid. The selected 6 EWs for lactic acid were included in the EWs for acetic and tartaric acids. Compared with SPA, some EWs were both selected by SPA and RC, and they were 406 and 944 nm for both acetic and tartaric acids, whereas 406 and 417 nm for lactic acid. However, most of the EWs selected were not the same by SPA and RC. This might indicated the different mechanism of SPA and RC for variable selection.

3.3. Calibration models and prediction performance

Firstly, the full-spectrum PLS models were developed without variable elimination for the prediction of three organic acids. The performance was validated by the samples in prediction set. Different latent variables (LVs) were used in PLS models for prediction of different chemical components. The performance of different preprocessing methods was compared, and the results are shown in

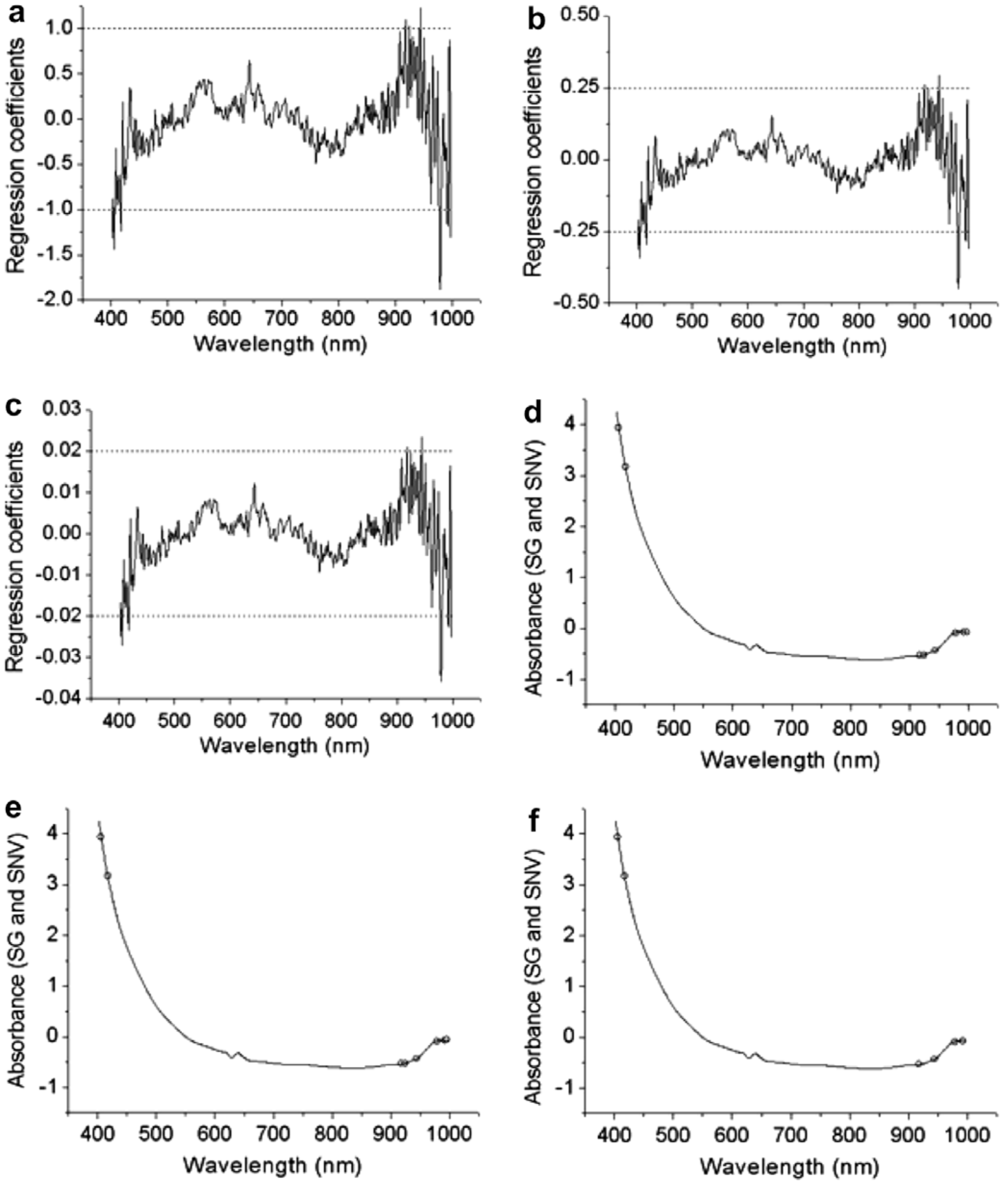

Fig. 3. The regression coefficients plots obtained for acetic acid (a), tartaric acid (b) and lactic acid (c). The two dot lines indicate the lower and upper cutoff threshold. The selected EWs (shown in circle markers) for acetic, tartaric and lactic acids are presented in (d), (e), and (f), respectively.

Table 2 for acetic, tartaric and lactic acids, respectively. As can be seen, the preprocessing of SG and SNV obtained the optimal performance for all three organic acids with highest correlation coefficients (r) values and least RMSEV/RMSEP values. Then the raw spectra performed better than the first and second derivative spectra. The reason might be that the treatment of first and second derivative brought in some noise which impaired the prediction ability of developed models.

Secondly, SPA-MLR models were developed using the selected EWs by SPA for the prediction of three organic acids of plum vinegar. The performance was also evaluated by 75 samples in prediction set. The prediction results are shown in Table 2. Herein, only the optimal calibration was shown applying raw spectra. As can be seen, in the validation set, the SPA-MLR model was slightly better than PLS models with SG and SNV spectra for all three organic acids. However, for the prediction set, full-spectrum PLS (SG and SNV) models performed slightly better than SPA-MLR models. It is worth mentioning that SPA-MLR models also performed an acceptable result considering the largely reduced number of variables.

Thirdly, PLS models were also developed using the EWs selected by SPA and RC. The SPA-PLS models were developed based on raw spectra, whereas RC-PLS models were developed using the preprocessed spectra by SG and SNV. The prediction results are shown in Table 2. As can be seen, all SPA-PLS models outperformed RC-PLS models for three organic acids. This indicated that the EWs selected by SPA were more powerful than those by RC. All RC-PLS models were not as good as full-spectrum PLS (SG and SNV) models for these three acids. However, SPA-PLS model yielded slightly better results than PLS (SG and SNV) model in validation set for the prediction of tartaric acid (seen in Table 2). On the overall, SPA-PLS models slightly outperformed SPA-MLR models in the prediction set for three acids, except that the correlation coefficients (r) values were a little lower for acetic ($0.9777 < 0.9799$) and lactic ($0.9760 < 0.9783$) acids.

Finally, the LS-SVM models were developed using the selected EWs by SPA and RC for the determination of organic acids of plum vinegar. The EWs were used as the inputs of LS-SVM models to develop SPA-LS-SVM and RC-LS-SVM models in order to reduce the training time. It was important to notice that the training time using LS-SVM increased with the square of the number of training samples and linearly with the number of variables (dimension of spectra) (Chauchard, Cogdill, Roussel, Roger, & Bellon-Maurel, 2004). Hence, the application of EWs would be helpful for the reduction of computation time. RBF kernel was recommended as the kernel function as analyzed above. The optimal model parameters (γ, σ^2) were determined by the two-step grid search technique with leave-one-out validation procedure. The searching ranges were set within (10^{-2}–10^5), and the optimal values were achieved with (3.1×10^4, 9.8), (2.5×10^3, 10.0) and (5.0×10^3, 7.7) by SPA-LS-SVM model, whereas (30.9, 4.9), (39.2, 9.5) and (14.9, 2.1) by RC-LS-SVM model for acetic, tartaric and lactic acids, respectively. The performance was validated by prediction set. The results are shown in Table 2. As can be seen, all SPA-LS-SVM models for three acids yielded better results than RC-LS-SVM models for both validation and prediction sets. This revealed that SPA was more powerful than RC for the selection of EWs as in the PLS models. However, all RC-LS-SVM models yielded better results than RC-PLS models. The reason might be that LS-SVM could make advantage of the latent nonlinear information of the spectral data which contributed a better prediction performance. On the overall, SPA-LS-SVM models achieved the optimal results among all developed models for the prediction of all three organic acids. The RMSEP values were 0.3581, 0.0714 and 0.0201 for acetic, tartaric

Table 2
The prediction results of organic acids of plum vinegar by different models.

Models	Preprocessing	EWs/LVs/(γ, σ^2)	Validation set		Prediction set	
			R[a]	RMSEV	r[b]	RMSEP
Acetic acid						
PLS	Raw	601/2/–	0.9769	1.5609	0.9794	1.5181
	SG + SNV	601/4/–	0.9878	1.1176	0.9885	1.1633
	1-derivative	601/2/–	0.8952	3.2220	0.8522	4.2076
	2-derivative	601/3/–	0.8566	3.7549	0.7253	5.4954
SPA-MLR	Raw	12/–/–	0.9937	0.8064	0.9799	1.6438
SPA-PLS	Raw	12/3/–	0.9843	1.2801	0.9777	1.5904
RC-PLS	SG + SNV	8/3/–	0.9627	2.0124	0.9654	1.8812
SPA-LS-SVM	Raw	12/–/(3.1×10^4, 9.8)	0.9990	0.2851	0.9985	0.3581
RC-LS-SVM	SG + SNV	8/–/(30.9, 4.9)	0.9915	0.9497	0.9798	1.6536
Tartaric acid						
PLS	Raw	601/2/–	0.9769	0.3719	0.9792	0.3641
	SG + SNV	601/4/–	0.9878	0.2662	0.9883	0.2801
	1-derivative	601/2/–	0.8951	0.7865	0.8500	1.0078
	2-derivative	601/3/–	0.8563	0.8963	0.7239	1.3123
SPA-MLR	Raw	15/–/–	0.9935	0.1963	0.9830	0.3944
SPA-PLS	Raw	15/4/–	0.9920	0.2173	0.9838	0.3513
RC-PLS	SG + SNV	8/2/–	0.9592	0.4827	0.9622	0.4719
SPA-LS-SVM	Raw	15/–/(2.5×10^3, 10.0)	0.9995	0.0475	0.9990	0.0714
RC-LS-SVM	SG + SNV	8/–/(39.2, 9.5)	0.9854	0.2900	0.9823	0.4152
Lactic acid						
PLS	Raw	601/2/–	0.9748	0.0308	0.9758	0.0310
	SG + SNV	601/4/–	0.9876	0.0213	0.9842	0.0254
	1-derivative	601/2/–	0.8945	0.0612	0.8474	0.0810
	2-derivative	601/3/–	0.8575	0.0709	0.7240	0.1048
SPA-MLR	Raw	19/–/–	0.9915	0.0176	0.9783	0.0330
SPA-PLS	Raw	19/2/–	0.9763	0.0298	0.9760	0.0312
RC-PLS	SG + SNV	6/2/–	0.9531	0.0411	0.9498	0.0432
SPA-LS-SVM	Raw	19/–/(5.0×10^3, 7.7)	0.9970	0.0104	0.9889	0.0201
RC-LS-SVM	SG + SNV	6/–/(14.9, 2.1)	0.9920	0.0182	0.9854	0.0289

[a] Correlation coefficients of validation set.
[b] Correlation coefficients of prediction set.

and lactic acids, respectively. This indicated that the new combination of SPA-LS-SVM was powerful and successful for the acid prediction in this specific study. SPA selected the most informative variables, and LS-SVM made good use of these EWs for calibration. The combination of SPA-LS-SVM offered the optimal prediction performance for the determination of organic acids of plum vinegar.

The overall results indicated that Vis/NIR spectroscopy combined with SPA-LS-SVM was successfully applied for the determination of acetic, tartaric and lactic acids of plum vinegar. Moreover, the results demonstrated that SPA was a powerful way for the selection of EWs which were relevant for spectroscopic analysis. Some further improvements could be made for the new SPA-LS-SVM method. Firstly, SPA was mainly focused on the selection of effective wavelengths with least collinearity and redundancies. Some attention should be paid on the relationship between the selected wavelengths and the molecular bands of acetic, tartaric and lactic acids. This would be helpful for us to understand the correlation between the wavelength and organic acids of plum vinegar. Moreover, the RBF kernel function could be replaced by other kernels such as linear kernel, polynomial kernel, and multilayer perceptron (MLP) in other applications when using SPA-LS-SVM method. Then the optimal kernel could be settled for the specific case. As is well known, the model developed by NIR spectroscopy has some limitations for its generalization. The model developed in this paper was most suitable for the case of determination of acids of plum vinegar. The proposed new approach of SPA-LS-SVM would be helpful for other applications. The model would also be applicable for other vinegars which had similar acids, such as apple cider vinegar and other fruit vinegars. If a more precise model needed for other vinegar, the samples of the vinegar should be added to the calibration set, and an expanded model could be developed using SPA-LS-SVM method. Then the prediction precision of the expanded model would be strengthened, and the generalization and robustness could be improved for more varieties of vinegars. Further studies would be focused on the further optimization of selected EWs by SPA, the relationship between the selected EWs and corresponding chemical components, and the generalization of an expanded model with other similar vinegars.

4. Conclusion

Vis/NIR spectroscopy combined with SPA-LS-SVM regression method was successfully utilized for the determination of acetic, tartaric and lactic acids of plum vinegar. SPA was proposed as a new powerful way for the selection of EWs, and the new developed combination of SPA-LS-SVM achieved the optimal prediction performance for all three acids comparing with full-spectrum PLS, SPA-MLR, SPA-PLS, RC-PLS and RC-LS-SVM models. The best prediction results by SPA-LS-SVM were that the RMSEP values for validation set were 0.3581, 0.0714 and 0.0201 for acetic, tartaric and lactic acids, respectively. The overall results indicated that Vis/NIR spectroscopy incorporated to SPA-LS-SVM could be applied as an alternative fast and accurate method for the determination of organic acids of plum vinegars. These results might be useful for the process and *in situ* monitoring of vinegar fermentation.

Acknowledgements

This study was supported by National Science and Technology Support Program (2006BAD10A09), the Teaching and Research Award Program for Outstanding Young Teachers in Higher Education Institutions of MOE, P.R.C., Natural Science Foundation of China (Project No. 30671213), 863 National High-Tech Research and Development plan (2007AA10Z210), Natural Science Foundation of Zhejiang (Y307158), Science and Technology Department of Zhejiang Province (Project No. 2005C12029).

References

Alsberg, B. K., Woodward, A. M., Winson, M. K., Rowland, J. J., & Kell, D. B. (1998). Variable selection in wavelet regression models. *Analytica Chimica Acta, 368*, 29–44.

Araújo, M. C. U., Saldanha, T. C. B., Galvão, R. K. H., Yoneyama, T., Chame, H. C., & Visani, V. (2001). The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems, 57*, 65–73.

Bokobza, L. (1998). Near infrared spectroscopy. *Journal of Near Infrared Spectroscopy, 6*, 3–17.

Casale, M., Abajo, M. J. S., Sáiz, J. M. G., Pizarro, C., & Forina, M. (2006). Study of the aging and oxidation processes of vinegar samples from different origins during storage by near-infrared spectroscopy. *Analytica Chimica Acta, 557*, 360–366.

Centner, V., Massart, D. L., De Noord, O. E., De Jong, S., Vandeginste, B. M., & Sterna, C. (1996). Elimination of uninformative variables for multivariate calibration. *Analytical Chemistry, 68*, 3851–3858.

Chauchard, F., Cogdill, R., Roussel, S., Roger, J. M., & Bellon-Maurel, V. (2004). Application of LS-SVM to non-linear phenomena in NIR spectroscopy: development of a robust and portable sensor for acidity prediction in grapes. *Chemometrics and Intelligent Laboratory Systems, 71*, 141–150.

Despagne, F., & Massart, D. L. (1998). Variable selection for neural networks in multivariate calibration. *Chemometrics and Intelligent Laboratory Systems, 40*, 145–163.

Esbensen, K. H. (2002). *Multivariate data analysis in practice* (5th ed.). Oslo: CAMO Process As.

Fu, X. G., Yan, G. Z., Chen, B., & Li, H. B. (2005). Application of wavelet transforms to improve prediction precision of near infrared spectra. *Journal of Food Engineering, 69*, 461–466.

Galvão, R. K. H., Araújo, M. C. U., Fragoso, W. D., Silva, E. C., José, G. E., Soares, S. F. C., et al. (2008). A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm. *Chemometrics and Intelligent Laboratory Systems, 92*, 83–91.

Gao, H. Y., Liao, X. J., Wang, S. G., & Hu, X. S. (2004). Simultaneous determination of eleven organic acids in fruit juice by reversed phase high performance liquid chromatography. *Chinese Journal of Analytical Chemistry, 32*, 1645–1648.

García-Parrilla, M. C., Heredia, F. J., Troncoso, A. M., & González, A. G. (1997). Spectrophotometric determination of total procyanidins in wine vinegars. *Talanta, 44*, 119–123.

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta, 185*, 1–17.

Guo, H., Liu, H. P., & Wang, L. (2006). Method for selecting parameters of least squares support vector machines and application. *Journal of System Simulation, 18*, 2033–2036.

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. New York: John Wiley and Sons.

Jouan-Rimbaud, D., Massart, D. L., Leardi, R., & De Noord, O. E. (1995). Genetic algorithms as a tool for wavelength selection in multivariate calibration. *Analytical Chemistry, 67*, 4295–4301.

Kalivas, J. H., Roberts, N., & Sutter, J. M. (1989). Global optimization by simulated annealing with wavelength selection for ultraviolet–visible spectrophotometry. *Analytical Chemistry, 67*, 2024–2030.

Liu, F., He, Y., & Wang, L. (2008a). Comparison of calibrations for the determination of soluble solids content and pH of rice vinegars using visible and short-wave near infrared spectroscopy. *Analytica Chimica Acta, 610*, 196–204.

Liu, F., He, Y., & Wang, L. (2008b). Determination of effective wavelengths for discrimination of fruit vinegars using near infrared spectroscopy and multivariate analysis. *Analytica Chimica Acta, 615*, 10–17.

Liu, F., He, Y., Wang, L., & Pan, H. M. (2007). Feasibility of the use of visible and near infrared spectroscopy to assess soluble solids content and pH of rice wines. *Journal of Food Engineering, 83*, 430–435.

Martens, H., & Naes, T. (1993). *Multivariate calibration*. London: Wiley.

Min, M., & Lee, W. S. (2005). Determination of significant wavelengths and prediction of nitrogen content fro citrus. *Transactions of the ASAE, 48*, 455–461.

Otto, M., & Wegscheider, W. (1985). Spectrophotometric multicomponent analysis applied to trace metal determinations. *Analytical Chemistry, 57*, 63–69.

Sagrado, S., & Cronin, M. T. D. (2008). Application of the modelling power approach to variable subset selection for GA-PLS QSAR models. *Analytica Chimica Acta, 609*, 169–174.

Sáiz-Abajo, M. J., González-Sáiz, J. M., & Pizarro, C. (2006). Prediction of organic acids and other quality parameters of wine vinegar by near-infrared spectroscopy. A feasibility study. *Food Chemistry, 99*, 615–621.

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. (2002). *Least squares support vector machines*. Singapore: World Scientific.

Suykens, J. A. K., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters, 9*, 293–300.

Yan, Y. L., Zhao, L. L., Han, D. H., & Yang, S. M. (2005). *The foundation and application of near infrared spectroscopy analysis*. Beijing: China Light Industry Press.

Ye, S. F., Wang, D., & Min, S. G. (2008). Successive projections algorithm combined with uninformative variable elimination for spectral variable selection. *Chemometrics and Intelligent Laboratory Systems, 91*, 194–199.