Bangladesh J. Bot. **45**(3): 659-667, 2016 (September)

MULTIVARIATE PATTERN RECOGNITION OF DIFFERENT VARIETIES' LAVENDER BY HEAD SPACE-SOLID PHASE MICROEXTRACTION WITH GAS CHROMATOGRAPHY

JUN TANG*, JIHONG FU, HONG TONG AND XIANG LIAO

Center for Physical and Chemical Analysis, Xinjiang University, Urumqi, P. R. China

Key words: Lavender, Head space-solid phase, Microextraction, Principal component analysis, Partial least squares-discriminate analysis

Abstract

Head space-solid phase microextraction (HS-SPME) coupled with gas chromatography (GC) has been applied for the identification of the characteristic volatile profile of lavender with the aim to study its varieties from Yili Xinjiang areas. Quantitative determinations of lavender samples' common peaks were carried out by GC with flame ionization detection (GC-FID) and qualitative analysis by GC with mass spectrometry (GC-MS). Principal components analysis (PCA) showed group clustering of three types of lavender varieties using GC-FID data. Partial least squares-discriminant analysis (PLS-DA) models had excellent classification sensitivity both for the calibration set and for the test set samples. The PLS-DA method was evaluated by the statistical indices of the correct recognition rate. Results showed that PLS-DA model were much better than PCA in classificatory abilities. It could successfully identify the complex nonlinearity and correlations among input variables and minimize them. The proposed chemometric methods illustrating the very plummy multivariate classification models, it proved to be an effective strategy for identifying the varieties of lavender, especially give a feasible method in the lavender quality control for use in medicine.

Introduction

Lavender (*Lavandula angustifolia* Mill.) is one of the most widely cultivated species in Europe, the Middle East, Asia and Northern Africa (Zheljazkov *et al.* 2012). The traditional uses of lavender range from use as an aromatic toan antimicrobial agent in perfume, food and pharmaceutical industry. For example, lavender is applied commercially in the fragrance industry, including soaps, perfumes, skin lotions and other cosmetics. It is also employed in food manufacturing, such as flavoring beverages, ice cream, candy, chewing gum and so on (Zuzarte *et al.* 2010). It is well-known that lavender could provide people with sedative and anti-depressive effects, such as relaxing the tight muscles, or releasing the pain of burns and insect bites.It has become one of the most important traditional herbal medicine and has significant economic value. These days, people found out that the effective ingredients are those volatile aromatic compounds with various functional groups abundantly present in lavender (Hancianu *et al.* 2013).

Lavender and its oil have been reported to contain large quantities of terpenes and aromatic compounds, with linalool, linalyl acetate, lavandulyl acetate, camphor and 1,8-cineole being the predominant (Danh *et al.* 2012). Specifically, linalool and linalylacetate, key active components of lavender, are reported to have an effect on glutamatergic system, enhances the release of dopamine from rat brain striatum slices. These organic compounds occur in varying amounts depending on harvest seasons, the extraction methods and species (Fismer *et al.* 2012, Chioca *et al.* 2013). In addition, lavender's chemical compositions vary significantly in terms of varieties. Therefore, analytical techniques and new methods are needed to develop an evaluation system to assess the systematic of the natural plant.

*Author for correspondence: <tangjunwq@163.com>. Center for Physical and Chemical Analysis, Xinjiang University, Urumqi, P. R. China.

Solid phase microextraction (SPME) has been suggested as an alternative technique for the extraction of chemical composition. It permits both the qualitative and quantitative analysis of sampling simultaneously (Popiel *et al*. 2011). Depending on the specific objectives, the large amount of data obtained by HP-SPME-GC-FID makes it necessary to use multivariate methods. In this work, derived parameters from the chromatogram such as peak areas were used. When faced with these high-dimensional data, PCA is often used to reduce the dimension by creating a new set of uncorrelated variables to form a super matrix, which corresponds to Eigen vectors of the sample covariance matrix (Dupuy *et al.* 2010). As compared to multivariate calibration methods based on latent variables, PLS-DA has advantages in terms of simplicity and ease of interpretation. The aim of this work was to attempt HP-SPME-GC as one of the key analytical tools, combined with chemometrics for identifying the species to meet the quality control requirements for use in herbal drugs. Multivariate data analysis was employed for identification and classification of lavender varieties using the GC-FID peak areas.

Materials and Methods

Lavender flower samples (n = 59) were harvested during the growing season in 2012. Lavender cultivars, C-197(2) lavender (n = 15), French blue lavender (n = 29) and H-701 lavender (n = 15) were purchased from the farming herd round field in Yili Xinjiang region. Doctor Junwei Yang identified it as lavender genus. The lavender samples were shattered before used and packed in special plastic packet to avoid the loss of aroma. Samples stored at 4ºC until the analysis.

The volatile fraction of lavender was monitored by using the HS-SPME technique. SPME parameters were adjusted through a series of optimization studies, including SPME fiber type, length of exposure, and volume of head space. The volatile organic components were extracted from the head space of the vials using the best conditions at last. The fibre was purchased from SUPELCO(USA). It was coated with 50/30 μm of divinylbenzene/carboxen/polydimethylsiloxane (DVB/CAR/PDMS) and previously conditioned before inserting it into the GC-FID injector at 250ºC for 30 min.

For each SPME analysis, 0.16 g lavender sample was placed in a 30 ml glass vial and put in a warm water bath (69ºC). The fibre was inserted in the head space and maintaining for 92.12 min and then removed from the vial. For the desorption of compounds the fibre inserted into the GC-FID injector at 250ºC for 5 min.

The GC-FID (Agilent Technologies, 5890 GC-FID) measurements were performed by using a PET-5 capillary column (30 m, 0.25 mm I.D., 0.25 μm film thickness). Oven temperature was kept at 50ºC held steady for 5 min, programmed to 246ºC at a rate of 3ºC/min, and then held steady for 5 min. Injector and detector temperatures were kept at 250 and 280ºC, respectively. The nitrogen (99.99%) was used as carrier gas at a constant flow rate of 1.0 ml/min. Flow rates of hydrogen and air were kept at 30 and 300 ml/min, respectively.

GC-MS (Shimadzu Corporation, GC/MS-QP 2010) analyses were performed using the same column and chromatographic conditions. The helium (99.99%) was used as carrier gas with a total flow of 1.16 ml/min. The injector temperature was 250ºC and the injection mode was split less. The detector temperature was 280ºC and the acquisition mode was full scan (from 30 to 500 m/z). The target compounds were identified based on GC-FID retention times using the NIST147 and WILEY7 standard mass spectra library.

Only one side of the original data is represented in a low-dimensional sub space, useful information may be lost (Gennebäck *et al.* 2013). Multivariate analysis is a tool used to bring out pertinent information from a matrix of either physical or chemical measurements. The methods allow valuable information to be extracted from multivariate data arrays. PCA, often used as pre-processor, the most important application was the reduction of the number of variables (scores) and the representation of a multivariate data table in a low dimensional space (Ciosek *et al.* 2006, Felipe-Sotelo *et al.* 2008). Principal components (PCs) are linear combinations of the original ones and can be interpreted like spectra, the portion of data for classification does not coincide with that of maximum correlation, or in PCA language that carrying the maximum portion of variance.

Supervised and unsupervised methods are two types of dimension reduction algorithms. In contrast to unsupervised PCA, supervised PLS-DA was based on the relation between spectral intensity and sample characteristics, which should have an initial knowledge of the classes of the sample set (Worley *et al.* 2013, Ivorra *et al.* 2013). The classes are defined based on a priori information of the system or by an exploratory analysis, for example, using PCA (Almeida *et al.* 2005). For PLS-DA classification, the class (one of M) of each sample was coded with a (binary) vector of length M with 0 and 1. The M predicted scores by the ordinary PLS-DA method are used to predict the class of each sample (Balabin *et al.* 2011). PLS-DA does not allow attributing a sample to other groups than the ones first defined. As a consequence, all measured variables play the same role with respect to the class assignment (Galtier *et al.* 2011, Guzmán *et al.* 2013). PLS-DA yields graphical displays and offers interpretation tools that enable investigating the structure of the sensory data. Within the modeling, a strong emphasis was placed on the plotting of model parameters scores, weights, etc. In this work, loading plot and regression model diagram were employed to show the importance of variables while building the models and the separation of different kinds of lavender samples, respectively.

PLS-DA is often compared to PCA in terms of its ability to classify data or to discriminate between different groups (Hobro *et al.* 2010). The X matrix is the PCA scores; the Y matrix is constructed on columns of binary numbers where 1 represents the sample member of a class and 0 if it is not a member. The sample membership of a class is then modeled and predicted as a 0 or 1 within a threshold limit of usually 0.5. Diagnostic plots of different thresholds will reveal the stability of the classifications. Due to the high number of correlated variables in a PLS model, it is essential to determine the correct number of components to reduce the risk for over fitting (Fongaro *et al.* 2013). Therefore, it is necessary to test the model performance of each PLS-DA component and select the optimal number of components. The parameters of PLS-DA model were optimized using cross-validation conditioned on the training data set (Mannina *et al.* 2010). It is of vital importance to monitor the root mean square error (RMSE) of the predicted values when chose the correct the number of factors.

Fifty nine lavender samples' GC-FID common peak areas were normalized to the internal standard areas before the statistical analysis. The relative abundance of each sample was found by dividing the peak area between the total areas of all the components. The determined original chromatographic profiles were organized into a matrix as input variables, employing PCA, PLS-DA procedures to differentiate these samples. PCA and PLS-DA data analysis was carried out by using Simca-P software (version 13.0, 2011; Umetrics, Umeå, Sweden).

Results and Discussion

GC-FID was used to reveal the relationship among the different lavender samples based on compositions of head space compounds. By comparing retention times, totally 15 common peaks were discovered (Table 1). Each compound was further confirmed by GC-MS. The optimized HP-SPME-GC/MS method was applied in order to characterize the volatile profile of lavender samples. From the data, 15 common volatile organic compounds were extracted from the head space of lavender samples. Fig. 1 is the total ion chromatogram of C-197(2) lavender sample. In order to obtain clear differences between the samples, the determined compounds' relative peak areas were used as initial variables. Therefore, the signal of GC-FID is a data matrix (15 variables × 59 lavender samples = 885 data) that is obtained. Then it was analyzed using multivariate pattern recognition methods.

Table 1. The common peaks information of 59 lavender samples.

No.	Chemical name	No.	Chemical name	No.	Chemical name
1	2-methylbutyl p-decyloxybenzylidene p-aminocinnamate	6	linalyl acetate	11	trans-caryophyllene
2	linalool oxide	7	neryl acetate	12	3,16-diacetyl pseudo solasodine
3	trans-linalool oxide	8	8-acetoxy linalool	13	beta-farnesene
4	linalool	9	limonene oxide	14	hotrienyl acetate
5	butyric acid-3-hexenyl ester	10	farnesal	15	andcaryophyllene oxide

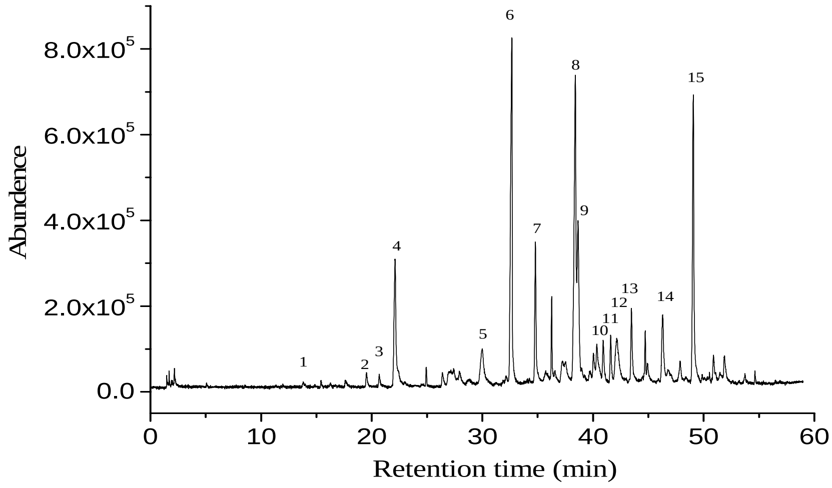

Fig. 1. Total ion chromatogram of C-197(2) lavender sample.

PCA only captures linear relations and does not pick the nonlinear relations. The extracted features are orthogonal and variant under transformation. A non-surveillance method was applied to visualize the data trends and to provide a first evaluation of the discriminant efficiency of the variables.

Applying PCA on the composed of data matrix, four components contributing the most to their formation were extracted, representing jointly 85.72% of the whole system variance. Thus, the first 4PCs reduced the 15-dimensional data set to a four-dimensional data set, with 14.28% loss of detail. The samples have been well separated in the score plot of the first two principal components, the first and second principal components describing 44.10% and 17.25% of the variability in the original observations, respectively, while both principal components account for 61.35% of the total variance. This analysis, therefore makes it possible to identify the measured variables that best explain the architectural variability observed.

Fig. 2 is the score plot of PC1vs PC2 for PCA-X of lavender samples. It was apparent that all samples were mixed scattered in score plot. Most of the cultivars could be generally separated from each other. It sorted into a clear separation of three differentiated groups (A, B and C) except for one of C-197(2) lavender sample (A7). PC1 positive value area differentiates the French blue cultivars from the C-197(2) lavender and H-701 lavender. Whereas PC1 negative value and PC2 positive value differentiate the H-701 lavender cultivars from the other two. Both PC1 and PC2 negative values differentiate the C-197(2) lavender cultivars from the other two, only A15, B14, C15 samples were wrongly classified to some extent. It was proved that these three samples were grouped in the cluster that did not belong to its classification.

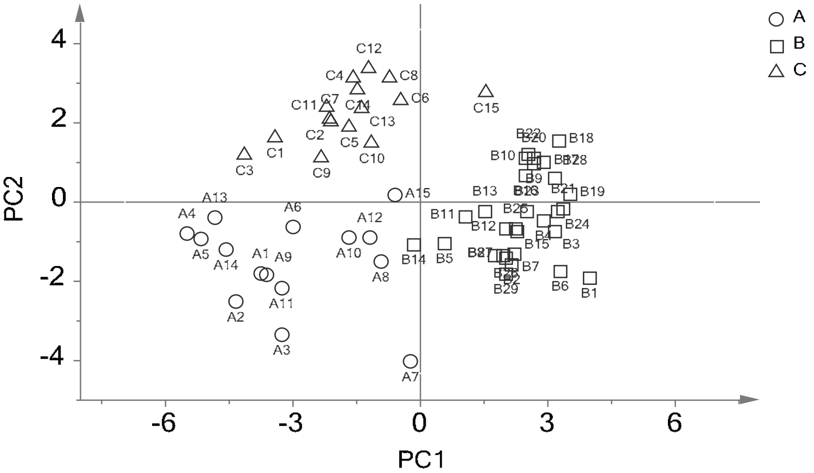

Fig. 2. The score plot for PCA-X of lavender samples (A: C-197(2) lavender, B: French blue lavender, C: H-701 lavender), the number is the same as the cultivars in supplement data.

As is well known, the PCA score plots provide useful information on the clustering of samples. The new variables were summarizing the X-variables that represent the PC1, PC2, etc. The first component explains the largest variation of the X space, followed by PC2, etc. Hence, the scatter plot of PC1 vs. PC2 was a window in the X space, displaying how the X observations are situated with respect to each other. This plot shows the possible presence of outliers, groups, similarities and other patterns in the data. The score plot was a map of the observations showed a clear histological classification that a two-dimensional window indicates grouping lavender samples into three clusters, a clear separation between C-197(2), French blue, H-701lavender categories, respectively.

The PLS-DA loading weights scatter plot displays the relation between the X and the Y-variables. To facilitate interpretation this plot by default symbol coded according to the model terms. X-variables situated in the vicinity of the dummy Y-variables have the highest discriminatory power between the classes. The plot is for the first and second components of the PLS-DA model. As shown in the graphic (Fig. 3A), there were 15 common compounds. Some compounds showed discriminating capability between the three kinds of lavender samples, the variables that mainly affect the separation between the different varieties of lavender. It was evident in the loading plot that C-197(2) variety was discriminated from the other two varieties of lavender by having positive values on PC1 and negative values on PC2, which exclusively contains farnesal, 3,16-diacetyl pseudo solasodine. French blue, was clustered by a negative values on PC1 due to high content of 2-methylbutyl p-decyloxybenzylidene, p-aminocinnamate, trans-linalool oxide, butyric acid-3-hexenyl ester, beta-farnesene, hotrienyl acetate, caryophyllene oxide to separate it from the other two. H-701 was clustered by having positive values on PC1 and PC2 due to high content of linalool, trans-caryophyllene to separate it from the other two. From the result of loading weights scatter plot, we can learn that the compounds correlated with different varieties of lavender.

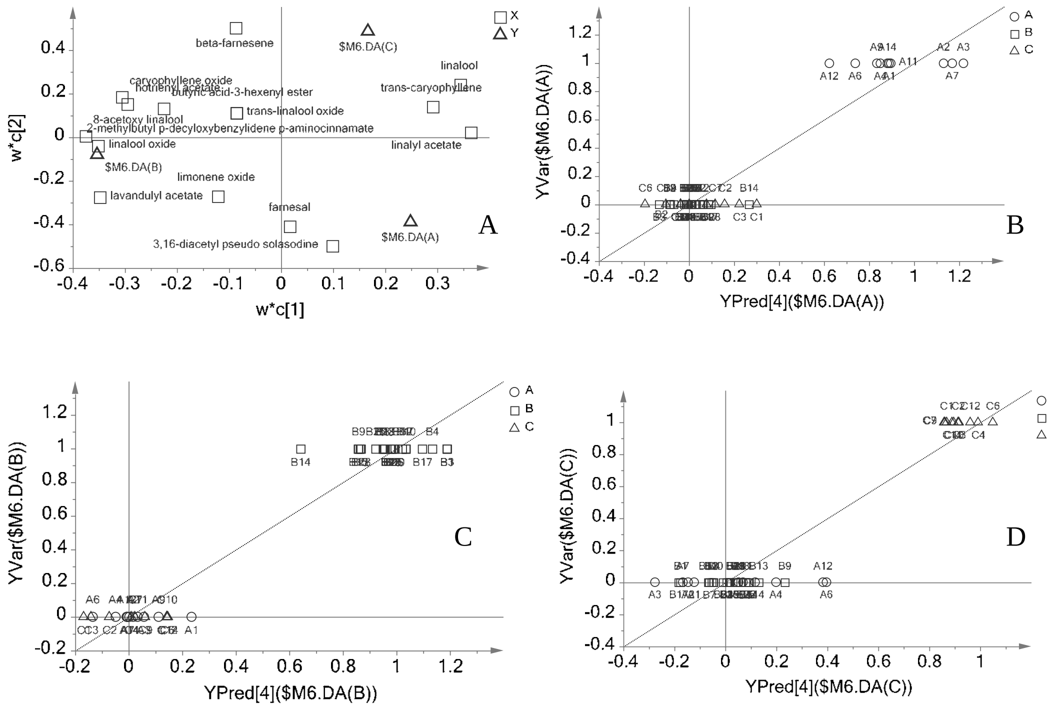

Fig. 3. PLS-DA loading weights scatter plot of the variables in the first two principal components and calibration set of regression model diagram (A: C-197(2), B: French blue, C: H-701), (A) The PLS-DA loading weights scatter plot displays the relation between the X and the Y-variables. The above w*c plot is a superimposition of the w* plot and the c plot, for the first and second components of the PLS-DA model. (B) Regression model diagram for discriminating C-197(2) lavender from French blue and H-701. (C) Regression model diagram for discriminating French blue lavender from C-197(2) and H-701. (D) Regression model diagram for discriminating H-701 lavender from French blue and C-197(2).

PLS-DA has one dummy Y variable per class with values of 1 or 0 depending on whether the observation belongs to the class or not. An ordinary PLS-DA analysis is made with this Y matrix with the X matrix containing the data characterizing the observations and classes. It is robust to colinearity issues and the noise and specific features of the training data not relevant to the problem under concern by tuning the number of latent variables used by the model. With less good models, the points are scattered around the regression line as in the plot below. PLS-DA played a significant role in discriminating samples that requires the data of interest to be splitted into two data sets, a calibration set and a validation set. When investigating lavender species, the calibration set comprised 42 samples and the validation set comprised of 17 samples.

The graphic (Fig. 3B, C, D) also displays the observed versus predicted values of the selected Y-variable. With a good model, all the points will fall close to this 45 degree line. The plot shows an example of a good model. C-197, French blue and H-701 species changed as Y-variable, all the calibration set samples ranged at 0.5 - 1.5 in M length, that indicating all the samples were classified correctly. Good results hold also when using the three different validation sets, the different cultivars were clearly separated from each other. It indicated that the success of the statistical tool for the recognition of the cultivars under study. Thus, PLS-DA model seems as well promising in view of discriminating the species.

Table 2 shows the classification parameters (RMSEE, RMSEP, R^2) and the percentage of the discrimination rate and prediction rate. As shown in Table1, all the calibration set objects have been correctly classified, leading to a correct discrimination among the different groups. The root mean square error of estimation (RMSEE) in the footer indicates the fit of the observations to the model. The R^2 value of the regression line indicates the goodness of fit. These results showed that the generated models have leading to a correct discrimination among the different groups.

Table 2. Identification efficiency of lavender samples with PLS-DA model, including calibration and validation sets.

Sample set		C-197(2)	French blue	H-701
Calibration set	R	0.8831	0.9477	0.8763
	RMSEE	0.1531	0.1201	0.1575
	Discriminant rate		100% (42/42)	
Validation set	R	0.8798	0.9372	0.8120
	RMSEP	0.1956	0.1364	0.1993
	Predict rate		94.12% (16/17)	

Cross validation was performed using random subsets meaning that each test set is formed from a random selection of objects such that a single object is only included in one test set. The number of latent variables selected for each model was determined by which giving the smallest root mean squared error of prediction (RMSEP). Three components were obtained by 20 times cross-validation. Fig. 4A, B, C) was permutation validation of the PLS-DA model. R^2, a measure of fit, i.e. how well the model fits the data, is the percent of variation of the training set explained by the model. Q^2 is the percent of variation of the training set predicted by the model according to cross validation. Q^2 indicates how well the model predicts new data. The idea of this validation is to compare the goodness of fit (R^2 and Q^2) of the original model with the goodness of fit of several models based on data where the order of the Y-observations has been randomly permuted, while the X-matrix has been kept intact.

The plot shows, for a selected Y-variable, on the vertical axis the values of R^2 and Q^2 for the original model (far to the right) and of the Y-permuted models further to the left. The horizontal axis shows the correlation between the permuted Y-vectors and the original Y-vector for the selected Y. The original Y has the correlation 1.0 with itself, defining the high point on the horizontal axis.

The plot above strongly indicates that the original model was valid. The criteria for validity: all Q^2-values to the left are lower than the original points to the right, the regression line of the Q^2-points intersects the vertical axis (on the left) at, or below zero. Note that the R^2-values always show some degree of optimism. However, when all R^2-values to the left are lower than the original point to the right, this was also an indication for the validity of the original model. From Figs 4A, B, C), all $Q^2$and R^2 values were higher in the permutation test than in the real model, revealing great predictability and goodness of fit.

The test set indicating a good predictive ability of the derived model. From results reported in Table 2, the selection of specific calibration subsets improved the prediction performance of the PLS-DA models in lavender samples as reflected by the RMSEP values. Although some samples were overlapped in the PCA (Fig. 2), all samples were clearly separated in the PLS-DA model (Fig. 3B, C, D). It indicated that non-correlated variation in X variables (the analyzed compounds) to Y variables (lavender variety) resulting in maximum separation in the PLS-DA model. It was not surprising that PLS-DA behaves better than PCA, e.g. sample A15, B14 and C15 were wrongly classified in PCA.

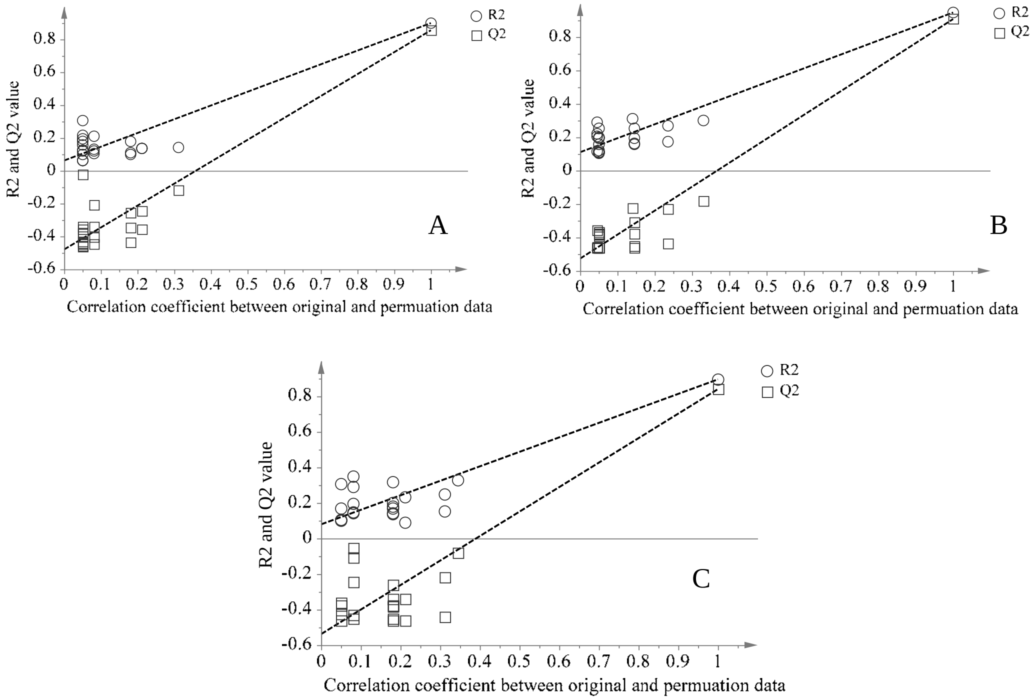

Fig.4. Permutation validation of the PLS-DA model, indicating great predictability (Q2) and goodness of fit (R2) of models leading to a correct discrimination. (A) The result of C-197(2) lavender. (B) The result of French blue lavender. (C) The result of H-701 lavender.

In summary, the results demonstrated that the specific HS-SPME-GC method combined with PCA and PLS-DA can effectively classify and identify the lavender samples by using the composition of volatile compounds. Moreover, PLS-DA was more efficient for the study of these types of data than unfolding PCA, as it offers an easier visualization of the data structure, in particular capturing the trends that are common to the different lavender in the samples mode plot. The results obtained can provide a useful comprehensive evaluation for lavender quality and give a reference for medicinal herb quality control. Through a standardized control, the variety of lavender will be improved.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 21265021). The authors are grateful to Mr. Qiang Wang (Center for Physical and Chemical Analysis of Xinjiang University, China) for providing the instruments and Mr. Junwei Yang (Xinjiang laboratory for the control of drugs, China) for identification of lavender genus.

References

Zheljazkov VD, Astatkie T and Hristov AN 2012. Lavender and hyssop productivity, oil content, and bioactivity as a function of harvest time and drying. Ind. Crops Prod. **36**: 222-228.

Zuzarte MR, Dinis AM, Cavaleiro C, Salgueiro LR and Canhoto JM 2010. Trichomes, essential oils *and in vitro* propagation of *Lavandula pedunculata* (Lamiaceae). Ind. Crops Prod. **32**: 580-587.

Hancianu M, Cioanca O, Mihasan M and Hritcu L 2013. Neuroprotective effects of inhaled lavender oil on scopolamine-induced dementia via anti-oxidative activities in rats. Phytomedicine **20**: 446-452.

Danh LT, Triet NDA, Han LTN, Zhao J, Mammucari R and Foster N 2012. Antioxidant activity, yield and chemical composition of lavender essential oil extracted by supercritical CO_2. J. Supercrit. Fluids **70**: 27-34.

Fismer KL and Pilkington K 2012. Lavender and sleep: A systematic review of the evidence. Integr. Med. **4**: e436-e447.

Chioca LR, Ferro MM, Baretta IP, Oliveira SM, Silva CR, Ferreira J, Losso EM and Andreatini R 2013. Anxiolytic-like effect of lavender essential oil inhalation in mice: Participation of serotonergic but not $GABA_A$/benzodiazepine neurotransmission. J. Ethnopharmacol. **147**: 412-418.

Popiel S and Sankowska M 2011. Determination of chemical warfare agents and related compounds in environmental samples by solid-phase microextraction with gas chromatography. J. Chromatogr. A. **1218**: 8457-8479.

Dupuy N, Galtier O, Ollivier D, Vanloot P and Artaud J 2010. Comparison between NIR, MIR, concatenated NIR and MIR analysis and hierarchical PLS model. Application to virgin olive oil analysis. Anal. Chim. Acta **666**: 23-31.

Gennebäck N, Malm L, Hellman U, Waldenström A and Mörner S 2013. Using OPLS-DA tofi and new hypotheses in vast amounts of gene expression data - Studying the progression of cardiac hypertrophy in the heart of aorta ligated rat. Gene **522**: 27-36.

Ciosek P, Brz´ozka WZ, Wr´oblewski, Martinelli E, Natale CD and D’Amico A 2006. Direct and two-stage data analysis procedures based on PCA, PLS-DA and ANN for ISE-based electronic tongue - Effect of supervised feature extraction. Talanta **67**: 590-596.

Felipe-Sotelo M, Tauler R, Vives I and Grimalt JO 2008. Assessment of the environmental and physiological processes determining the accumulation of organochlorine compounds in European mountain lake fish through multivariate analysis (PCA and PLS). Sci. Total Environ. **404**: 148-161.

Worley B, Halouska S and Powers R 2013. Utilities for quantifying separation in PCA/PLS-DA scores plots. Anal. Biochem. **433**: 102-104.

Ivorra E, Girón J, Sánchez AJ, Verdú S, Barat JM and Grau R 2013. Detection of expired vacuum-packed smoked salmon based on PLS-DA method using hyperspectral images. J. Food Eng. **117**: 342-349.

Almeida MRD, Correa DN, Rocha WFC, fi FJOS and Poppi RJ 2013. Discrimination between authentic and counterfeit banknotes using Raman spectroscopy and PLS-DA with uncertainty estimation. Microchem. J. **109**: 170-177.

Balabin RM and Safieva RZ 2011. Biodiesel classification by base stock type (vegetable oil) using near infrared spectroscopy data. Anal. Chim. Acta **689**: 190-197.

Galtier O, Abbas O, Dréau YL, Rebufa C, Kister J, Artaud J and Dupuy N 2011. Comparison of PLS1-DA, PLS2-DA and SIMCA for classification by origin of crude petroleum oils by MIR and virgin olive oils by NIR for different spectral regions. Vib. Spectrosc. **55**: 132-140.

Guzmán E, Baeten V, Pierna JAF and García-Mesa JA 2013. A portable Raman sensor for the rapid discrimination of olives according to fruit quality. Talanta **93**: 94-98.

Hobro AJ, Ski JK, Döll Mu and Lendl Bh 2010. Differentiation of walnut wood species and steam treatment using ATR-FTIR and partial least squares discriminant analysis (PLS-DA). Anal. Bioanal. Chem. **398**: 2713-2722.

Fongaro L and Kvaal K 2013. Surface texture characterization of an Italian pasta by means of univariate and multivariate feature extraction from their texture images. Food Res. Int. **51**: 693-705.

Mannina L, Marini F, Gobbino M, Sobolev AP and Capitani D 2010. NMR and chemometrics in tracing European olive oils: The case study of Ligurian samples. Talanta **80**: 2141-2148.

(Manuscript received on 7 May, 2016; revised on 13 June, 2016)