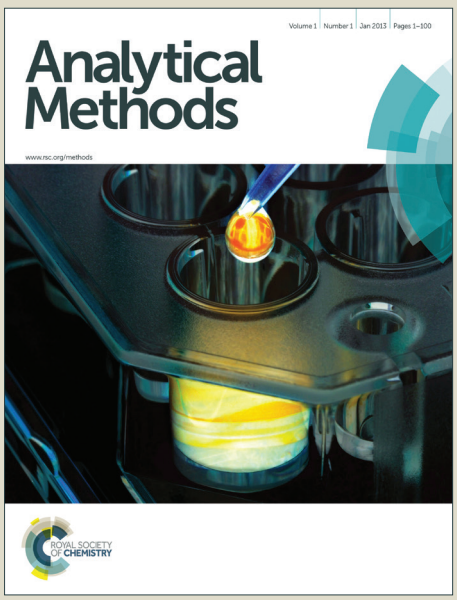
Analytical Methods

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

Analysis of different Flos Chrysanthemum tea samples with the use of two-dimensional chromatographic fingerprints, which were interpreted by different multivariate methods

Xiaoxiao Ding, [a,b] Yongnian Ni, [a,b*] and Serge Kokot [c]

[a] *State Key Laboratory of Food Science and Technology, Nanchang University, Nanchang 330047, China*

[b] *College of Chemistry, Nanchang University, Nanchang 330031, China*

[c] *School of Chemistry, Physics and Mechanical Engineering, Queensland University of Technology, Brisbane 4001, Australia*

* Corresponding author address: Department of Chemistry, Nanchang University, Nanchang 330031, China. E-mail: ynni@ncu.edu.cn; s.kokot@qut.edu.au

Abstract: Flos Chrysanthemum tea contains flavonoids, essential oils and caffeoylquinic acids. These substances are pharmacologically active but this activity is cultivar dependent. Seventy-six Flos Chrysanthemum samples collected from four cultivars (Hangju, Taiju, Gongju and Boju) were discriminated with the use of results from high performance liquid chromatography (HPLC) and gas chromatography - mass spectrometry (GC-MS). A two-dimensional chromatographic fingerprint data set of the four kinds of Flos Chrysanthemum cultivar was built from the combined GC/HPLC profiles and thirty variables were selected. Principal component analysis (PCA) and kernel – PCA (KPCA) were used for feature extraction. The score mapping graph indicated that these two PCA methods effectively extracted most information from the samples, and the four Flos Chrysanthemum cultivars were qualitatively differentiated. Furthermore, four supervised pattern recognition techniques, Radial basis function-neural network analysis (RBF-NN), Least squares support vector machines (LS-SVM), Linear discriminant analysis (LDA) and K-nearest neighbors (KNN), successfully predicted the validation samples.

Key words: Agricultural samples; Flos Chrysanthemum tea; HPLC; GC-MS; Multivariate analysis

1. Introduction

Flos Chrysanthemum, also, known as the 'white chrysanthemum', is the dried capitulum of Chrysanthemum morifolium Ramat (*C. morifolium* R.), which belongs to the Chrysanthemum genus in the Asteraceae family.[1] Generally, it is native to Asia but more particularly to Korea and northern Japan.[2,3] In China, the flowers of this plant are commonly used to make a pleasant and refreshing tea drink.[4] It has also been reported that Flos Chrysanthemum has anti-bacterial, -inflammatory, -oxidant and -mutagenic properties,[5,6] because it is rich in bioactive constituents, such as flavonoids, essential oils and caffeoylquinic acids.[7,8] Forty six flavonoids and seventeen caffeic acid derivatives have been identified in an aqueous-methanol extract of Flos Chrysanthemum with the use of high performance liquid chromatography (HPLC) equipped with a diode array detector, and electrospray ionization/mass spectrometry (ESI/MS).[9] Clifford et al.[10] used LC-MS^5 to identify chlorogenic acids and some caffeic acid derivatives of Chrysanthemum samples. Chang and Kim[11] detected forty-five volatile compounds in Flos Chrysanthemum samples extracted with the use of steam distillation.

In China, there are more than 20 cultivars of Flos Chrysanthemum grown in different geographical regions; these plants are often differently processed.[12] In general, it may be reasonably expected that the different cultivars will have different pharmacological effects, but interestingly the 'Chinese Pharmacopoeia', records all of them as 'Flos Chrysanthemum' with similar morphological characteristics.[1] The most common cultivar varieties are: Boju (Bozhou, Anhui), Chuju (Chuzhou, Anhui), Gongju (Huangshan, Anhui), and Hangju (Tongxiang, Zhejiang). Importantly, market prices of these cultivars are quite different, e.g. in China, the Gongju and Hangju prices are often as much as fifteen times higher than Boju. It is, therefore, important to have analytical techniques to distinguish and identify the cultivars so as to discourage adulteration practices, which could impact on the sale prices as well as, arguably, on the health of the consumers. The World Health Organization (WHO), the US Food and Drug Administration (FDA) and the European Medicines Agency (EMA)[13] have recommended the fingerprint approach for quality control of the Flos Chrysanthemum cultivars, and this was adopted in this study.

Different techniques such as HPLC, gas chromatography (GC), nuclear magnetic resonance (NMR) and near-Infrared reflectance spectroscopy (NIR) have been used to obtain the fingerprints of food or plant samples.[14-17] The chromatographic fingerprinting approach is, arguably, the most widely used in industry and

scientific work. It involves the building of a large data base of chromatographic profiles from compounds found in the relevant plants or foods, such that this data can provide comprehensive information on the plant or food samples of interest. The main characteristics of chromatographic fingerprinting include data integrity, fuzziness, similarities and differences.[18] Thus, given that chromatographic fingerprint matrices are often quite large, multivariate data analysis techniques, often referred to as chemometrics methods, become useful and necessary for data interpretation.[19]

There are several research reports concerning the quality of Chrysanthemum plants with the use of different methods. Liu et al.[4] characterized various Chrysanthemums of different species and provenance (White, Boju, and wild Chrysanthemums amongst others) with the use of HPLC. Xie et al.[20] analysed phenolic compounds with the use of HPLC-DAD-ESI/MS, in 12 samples of five different cultivars of Chrysanthemum morifolium flowers grown in China. The results of their anti-allergic assays were investigated as well. Zhong et al.[21] used GC/MS to analyze the volatile components o*f Chrysanthemum indicun* L. from eight populations in China, and 169 compounds were identified. Wang et al. analysed different types of white Chrysanthemum sample, and the plants of different varieties were discriminated. Interestingly, adulterated samples were separated from the non- adulterated ones with the use of a novel technique – the electronic tongue method, which mimicked human taste.[22] However, only a few studies have been carried out with Chrysanthemums, which involved the simultaneously discrimination of their molecular constituents, which belonged to several different categories or groupings.

In the present work, two-dimensional chromatographic fingerprints were collected and combined with the use of GC and HPLC methods. Four kinds of Flos Chrysanthemum cultivars, Hangju, Taiju, Gongju and Boju, were qualitatively discriminated with the use of two kinds of unsupervised pattern recognition approaches, principal component analysis (PCA) and kernel principal component analysis (KPCA). The same data matrices were analysed with the use of four supervised pattern recognition approaches: radial basis function-neural network analysis (RBF-NN), least square support vector machines (LS-SVM), linear discriminant analysis (LDA) and K-nearest neighbors (KNN). Calibration sets were developed for prediction of unknown samples.

2. Materials and methods

2.1. Chemicals and plants

Formic acid (Analytical Grade) was purchased from Red Star Chemical Factory (Beijing, China), Methanol (LC grade) was obtained from Damao Chemical Reagent Factory (Tianjin, China), and freshly twice-distilled water was used throughout the experiment.

Seventy-six Flos Chrysanthemum samples of four cultivars were collected from different pharmaceutical and tea stores in China. Samples 1 - 15 and 16 - 30 were the Hangju and Taiju kind, respectively, and they were cultivated in Tongxian, Zhejiang province. Taiju is the flower bud of the Hangju plant, whereas Hangju is the flower of the Hangju variety. Samples 31 - 55 and 56 - 76 were the Gongju and Boju varieties, respectively, which are grown in Huangshan mountain and Bozhou county, respectively, in the Anhui province. All the Flos Chrysanthemum samples were crushed into powder with a high-speed pulverizer (QE-100, Yili Instrument Co., Wuyi, China), and then passed through a 60 mesh sieve.

2.2. Sample preparation and analysis with the RP-HPLC-DAD technique

An accurately weighed powder sample (2.00 g) was reflux extracted in 45 mL 80% methanol for 2 hours on a 60 °C water bath. The extract was filtered through a filter paper into 50 mL volumetric flask and then diluted to the mark with 80% methanol. The final extract was filtered through a 0.45 μm organic membrane and used for HPLC analysis.

The assay was performed on an Agilent 1100 series (Agilent Technologies, Palo Alto, CA, USA) equipped with a G1315B diode array detector, a G1379A online vacuum degasser, a quaternary pump solvent management system, a G1311A autosampler, and an injector with a 100 mL loop. The chromatographic separation was carried out on an Agilent Zorbax Eclipse XDB-C18 column (4.6 mm × 250 mm, 5 μm) and an Agilent Zorbax high-pressure reliance cartridge guard column (C18, 4.6 × 12.5 mm, 5 μm). The mobile phase was methanol (A) and 0.1% formic acid aqueous solution (B), which was ultrasonically degassed and filtered through a 0.22 μm nylon membrane. The elution gradient program was as follows: 95-75% B (0-10min), 75-70% B (10-15 min), 70-65% B (15-18 min), 65-60% B (18-25 min), 60% B (25-35 min), 60-30% B (35-45 min), 30-10% B (45-50 min), 10-0% B (50-60 min) and the last elution was held constant for 5 min. The sample was then allowed to equilibrate at the initial conditions (5% A and 95% B) before the next sample was injected. The flow rate was set as 1.0 mL min^{-1}, the injection volume was 20 μL, the

detection wavelength was set at 254 nm for all compounds, and the column temperature was maintained at (25 ± 0.5) °C.

2.3. Headspace extraction of volatile compounds and GC-MS analysis

Each powder sample (1.00 g) was accurately weighed and transferred into the 20 mL headspace vials capped by polytetrafluoroethylene (PTFE)/silicone septa, and then such a sample was equilibrated for 45 min at 80 °C in a water bath. After this equilibration, 200 μL extract of the volatile compounds were absorbed with a CTC Analytics Headspace Syringes (Switzerland) and injected into the GC-MS instrument (Shimadzu GCMS-QP2010, Kyoto, Japan) equipped with an Agilent DB-5 MS column (30 m length × 0.25 mm inner diameter × 0.25 μm film thickness, Agilent Technologies, Palo Alto, CA, USA). The helium gas flow rate through the column was 1.00 mL min^{-1}, and the total flow rate was 9.00 mL min^{-1}, and the linear velocity was 36.5 cm s^{-1}. The inlet temperature was 140 °C, and 200 μL of the extract was injected in the split (5:1) mode. The initial oven temperature was 60 °C and was held at this point for 1 min, and then increased to 130°C at a rate of 4 °C min^{-1}, held for 1 min, and finally ramped up to 195 °C at 6 °C min^{-1}, and held for 1 min. The interface and ion source temperatures were 250 and 200 °C, respectively. The mass detector was set on 'Full scan' monitoring mode with a mass scanning range of m/z 50-500 and the detector voltage of 1.06 kV.

2.4. Statistical analysis

Multivariate analysis methods play an important role in differentiating different kinds of Flos Chrysanthemum cultivars. In this work, the characteristic GC and HPLC common peaks of all samples were matched by the Computer Aided Similarity Evaluation System (CASES, Chinese Pharmacopoeia Commission, Version 2004A). The above experimental data was processed in the same way but with the use of similarity analysis. Several, well-known multivariate, statistical, analytical methods were applied for the above work, and their principles are summarized below.

2.4.1. Unsupervised pattern recognition:

Principal components analysis (PCA): PCA is an unsupervised pattern recognition method of data analysis,

which is also used for data compression and feature extraction.[23,24] As it is a linear unsupervised method, no data grouping information has to be known before the analysis. PCA transforms the original variables into new ones, i.e., the principal components (PCs). These new variables are linearly combined with the original ones and are orthogonal to each other. In general, the number of PCs extracted is the same as the number of the original variables but usually only a few PC variables account for most of the data variance, i.e., the dimensionality of the original data matrix is reduced and the redundant information is eliminated. The first PC accounts for most of the data variance, PC2 the next largest amount, and so on. Generally, PCA decomposes the data set into score and loadings matrices. When the scores' matrix is visualized, hidden associations between the samples are often revealed, e.g., samples from the same origin will group together although no grouping information is apparent prior to PCA analysis. Additionally, when the loadings' matrix is visualized, then, the loadings will be displayed as vectors, the longest of which will represent the most influential or discriminating variables. If the score plots and loading plots are mathematically combined in a biplot, then the relationships between scores' groups and variables will become readily apparent, i.e. in the case of peak variables, their significance with respect to the objects becomes readily apparent – the closer and longer loadings vectors are to a group of objects the more influence they (and the associated original peak variables) will have on the group. Finally, the eigenvalues of the data matrix can also be obtained after PCA decomposition, and from this it is possible to calculate the variance of the total data variance every PC explains.

Kernel principal component analysis (KPCA): The PCA modeling discussed above refers to the resolution of linear data but when non- linear data is involved, the KPCA method becomes much more appropriate for the extraction of PCs. This method KPCA enables PCA calculations using nonlinear modelling. It was developed by generalizing the kernel method into PCA.[25] A kernel method, such as the support vector machine model, was originally used for solving complex nonlinear classification and regression problems,[26] and eventually, Scholkopf et al.[27] generalized this method into a classical PCA, which became known as the KPCA method. Unlike PCA, this method can only obtain the score matrix and the eigenvalue vectors. However, it can also reduce the number of dimensions by considering just the first few eigenvectors and mapping the input data onto a high dimensional feature space.

2.4.2. Supervised pattern recognition

Radial basis function-neural network analysis (RBF-NN): The non-linear neural network has been used for pattern recognition and classification of different foods.[28] This method involved information transmition between object units, which were treated as if they had the structure and function of the human brain nerve network, i.e., learning, memorizing, summarizing and extraction data processes were facilitated. The training of the learning process begins with the transmition of the signal from an intermediate layer to the output layer. Following this training process, the network is able to recognize and memorize the data. RBF-NN is a classical topology type of neural network because of its high non-linear mapping ability. It creates a two layer network: a hidden radial basis layer, and an output linear layer. The hidden layer has neurons with non-linear functions, e.g., the radial basis functions. In this context, the Gaussian function is commonly incorporated in the RBF-NN model.[29] The output layer is made of linear units, containing Purelin neurons, and its weighted input is calculated. The following steps are repeated until the network's mean square error falls below the maximum number of neurons to be used: (1) the network is simulated; (2) the input vector with the greatest error is found; (3) a neuron is added with weights equal to that vector; (4) the Purlin layer weights are redesigned to minimized error.[30]

Linear discriminant analysis (LDA): Linear discriminant analysis (LDA) is the most frequently used supervised pattern recognition, linear and parametric method. LDA is used in applications where classification of certain criteria is needed. Fisher criterion is a well known classification criterion.[31] Fisher's linear discriminant analysis aims to find the optimum linear boundaries among different data classes which maximizes the ratio between class variance and minimizes the ratio within class variance.[32] The latent variables obtained in the LDA are the linear combination of the original variables. From *k* classes, *k*-1 latent variable can be determined. LDA can also be used to reduce the dimensionality of data similar to PCA.[33]

Least square support vector machines (LS-SVM): SVM is a methodology based on statistical learning theory in the field of non-linear modelling. Suykens and his colleagues proposed a modified version of SVM, which is the least-squares SVM (LS-SVM).[34] This method does not produce the support vectors (SVs) with the use of quadratic methods rather linear equations suffice. In this method, the input data is transferred to high dimensional space, and appropriate kernel functions and optimum kernel parameters are required to complete the computations. Also, the regularization parameter (γ) and the radial basis function (RBF) parameter (σ) are obtained by a pattern search algorithm followed by the leave one out cross-validation (LOOCV)

method.[35]

K-nearest neighbors (KNN): K-nearest neighbors (KNN) method is a linear, non-parametric, supervised pattern recognition method, which is an uncomplicated, reliable classification method for unknown samples. These are divided into a training and an unknown sample sets prior to KNN analysis. The class of each of the *k* neighbours is known in the training set, and thus, the distance between an unknown sample and each of the training samples can be determined. Subsequently, each unknown sample will be classified into a class of k nearest neighbors, which belong to that class. The Euclidean distance and the correlation coefficient are the two criteria used for calculating the distances between samples, and the Euclidean distance is often adopted because variables are not strongly correlated in most cases.[36] In this algorithm, *k*, is an odd number (e.g. 1, 3, 5, 7 and 9), and influences the prediction rate. In general, an optimum *k* value is chosen by calculating the prediction and recognition rates with several *k* values.[37]

3. Results and discussion

3.1. Analysis of Flos Chrysanthemum samples with the use of GC data

A set of GC data was collected from the GC-MS real time analysis chemstation in QGD data format. The data were transformed into text format after the total ion chromatograms (TIC) of all groups were integrated. The TICs of the four different Chrysanthemum cultivars (Fig. 1A) are very similar, although it is quite evident that the Taiju samples contain more volatile constituents than the other three cultivars. The basis for this observation may be that less essential oils volatilize from the Taiju flower buds during blossoming. Also, TIC data suggested that Taiju and Hangju results were similar; this probably occurs because the two cultivars belong to the same species. Thirty-five strong GC peaks of volatile compounds were identified by comparing their mass spectral patterns with those from the NIST (http://srdata.nist.gov/chemistry) online databases (Table 1). The main volatile compounds in the four Flos Chrysanthemum cultivars were: 7-(1-Methylethylidene)bicyclo[4.1.0]heptane (**6**), 2H-Benzocyclohepten-2-one,3,4,4a,5,6,7,8,9-octahydro- **(10)**, Eucalyptol **(14)**, 3,5-Heptadienal,2-ethylidene-6-methyl- **(18)**, 2(1H)-Naphthalenone,octahydro-,trans (**20**), Isoborneol (**22**), Bicyclo[3.1.1]hept-2-en-4-ol,2,6,6-trimethyl-,acetate (**25**), 4,6,6-Trimethylbicyclo [3.1.1]hept-3-en-2-yl,acetate (**26**), Acetic acid,1,7,7-trimethyl-bicyclo[2.2.1]hept-2-yl,ester (**27**), Chrysanthenone (**30**), Benzene,1-(1,5-dimethyl-4-hexenyl)-4-methyl- (**31**), beta.-Sesquiphellandrene (**33),**

Bis(1-methyl-4-pentenyl)phthalate (**34**).

Seventeen common peaks were found (numbered peaks, Fig. 1A), and their related compounds (%) were listed (Table 2). These results showed that the Gongju samples have a very high level of G9 (**20**, 20.6%) and a relatively high level of G3 (**6**, 10.2%) compared with other Flos Chrysanthemum cultivars. Boju samples contain relatively low levels of G2 (**5**, 1.8%) and G5 (**10**, 1.4%) compared with other Flos Chrysanthemum cultivars. The G7 (**14**) compound makes a similar contribution to that found in the Gongju and Boju varieties, i.e., there are about 7.7% and 7.8% of the main volatile components, respectively - a little higher than those of Hangju and Taiju.

The retention time shifts of the GC profiles were corrected and the text formated GC data were processed with the use of CASES (version 2004A) software. This program is applied for extracting common peaks from all samples. The similarities (parameters: similarity range, mean and relative standard deviation) between samples from the same class or samples from different classes (Table 3) indicated that the former have a higher mean similarity than the samples from different classes. However, the relative standard deviations of the sample profiles were somewhat higher irrespective whether the samples came from the same or different class.

3.2. Interpretation of the reverse-phase HPLC results from Flos Chrysanthemum

A data matrix containing the results collected from the Flos Chrysanthemum samples with the use of reversed-phase HPLC, was exported from the chemstation as an analytical instrument association (AIA) file; then, these data were imported into CASES in order to match any common peaks from all the samples and thus, investigate their similarities. The parameter values (similarity range, mean and relative standard deviation) of the similarities between the samples (Table 3) indicated that the sample profiles from the same class have a higher mean similarity than that of the samples from different classes. However, there were still several samples from different classes, which were quite similar. Thus, the conclusion is that it is difficult to discriminate the Flos Chrysanthemum samples according to their similarity values. Also, the relative standard deviations of the similarity values derived from the liquid chromatographic profiles of the samples, were lower than those from the GC profiles. This occurred irrespectively weather these

profiles were derived from the same or different classes. These results indicated that the LC profiles of the Flos Chrysanthemum samples were more stable. The HPLC fingerprints of the four kinds of different Chrysanthemum cultivar were compared (Fig. 1B), and it was evident that the chromatograms from Hangju, Taiju, Gongju and Boju samples were reasonably similar, but of the four sets of these well matching profiles, those from Taiju and Hangju samples were particularly alike, (Fig. 1B), e.g. chromatographic profiles from Hangju and Taiju had thirteen matched common peaks (numerically annotated).

3.3. Extraction of variables from the two-dimensional fingerprints

It was previously demonstrated that data, collected from different analytical instruments, is combined then classification of objects is improved.[38] In this work, seventeen common GC peaks (76 × 17) and thirteen common LC peaks (76 × 13) were obtained after being processed by the CASES method. These data were normalized and combined to form a two dimensional data matrix. It has been suggested,[39] that the selection of characteristic variables should follow the requirement that n (number of sample) / f (number of the characteristic variable) should be > 5. On this basis, reliable statistical results of feature extraction could be obtained. Thus, the combined data matrix was submitted for PCA and a loadings matrix was obtained. The resulting loadings biplot (Fig. 2) has circles to indicate the thirty variables, the position of which may be thought of in terms of vectors originating at the origin. Conventionally, the longest loadings vectors indicate the most important variables. In this work, such vectors are represented by the 14 outermost circles. These 14 variables have significant roles in the discrimination of the four types of object, i.e. the four Chrysanthemum cultivars. The variables included six LC and eight GC characteristic peaks, and these 14 variables were utilized in the formation of a smaller data matrix (76 × 14), which was used for subsequent data analysis.

3.4. Data analysis with the use of methods involving principal components

PCA is a well-known qualitative method of data analysis including data from chromatographic fingerprints. Such analysis provides information regarding any relationships between objects, variables, and objects and variables. PCA results from the smaller data matrix (76 × 14) obtained from the four different Chrysanthemum cultivars, are displayed in Fig. 3. The first three PCs extracted from this matrix accounted for 80.36% of data variance: PC1 - 40.17%, PC2 - 29.13% and PC3 - 11.06% .The PC1–PC2 scores and

loadings biplot (Fig. 3A) accounted for 69.30% of data variance. When the objects with positive scores were projected onto PC1, the Gongju objects overlapped with the Boju ones, and **L7**, L10, L11, **L12**, L13, **G7**, **G9**, G15 and G16 vectors were associated with these objects, particularly the highlighted ones. The Hangju and Taiju samples spread out along PC1 with negative scores. Again, there is no clear separation of these two object groups, and the associated variables include: L9, G2, G5, G14 and G17, and only the G14 vector is rather short and weak. Thus, the Hangju and Taiju groups of samples are separated from the Gongju and the Boju ones on the basis of the variables noted above. Interestingly, when the objects were projected onto PC2, a better but still not quite complete separation of the four groups of cultivars was observed. Thus, the Gongju sample group with the highest positive PC2 scores, was followed by the Hangju and Taiju ones; these were compressed together around the origin, and these objects were followed by the Bosu group, which was well spread out with negative scores on this PC. However, the Hangju and Taiju samples were successfully separated along the PC3 axis of the PC1 vs. PC3 scores and loadings biplot (Fig. 3B). However, these two groups overlapped each other with negative scores on PC1. The other two groups, Gongju and Bosu, were also tightly grouped with positive scores on PC1. The same loadings vectors are responsible for these separations as in Fig. 3A, but they are differently positioned and their lengths are different as well. The three dimensional PCA scores plot - PC1 vs. PC2 vs. PC3, of the Flos Chrysanthemum samples (Fig. 3C) suggests that the four cultivar groups may be separated in this space. Thus, PCA provides useful, qualitative information regarding the objects, loadings and their inter-relationships. Also, a data matrix built from the PC1-PC3 scores would be sufficient to build quantitative models.

Flos Chrysanthemum samples were also investigated with the use of the KPCA classification method (Fig. 4). The top three PCs explained 64.22% of the variance with PC1, PC2 and PC3 contributing 31.94%, 22.21% and 10.07%, respectively. When the three score plot diagrams (Fig. 4A, 4B and 4C) were compared with their counterparts in Fig. 3A, 3B and 3C, then on the whole, the distribution of the groups is more or less similar with one important difference - the groups appear to be somewhat better discriminated in the respective KPCA plane. While that is an important aspect of the KPCA method, the important advantage of the PCA model is that it allows for the exploration of the relationship between the sample groups and the variables. Consequently, the PCA method would be generally preferred, with the KPCA method being used to investigate the discrimination of very similar samples.

3.5. Pattern recognition

Prediction modelling requires information from the analytes to build a calibration, validation and prediction data sets. In this work, the 76 Flos Chrysanthemum samples were divided into 51 randomly selected samples for calibration (10 Hangju, 10 Taiju, 17 Gongju and 14 Boju) and the rest 25 samples were used for prediction. Prediction models were built from the four pattern recognition methods, RBF-PLS, LS-SVM, LDA and KNN were used with the unknown Flos Chrysanthemum samples. Both, the PCA data matrix formed from the scores of the first three PCs, and the KPCA data matrix, formed from the scores of the first three KPCs, were used for four kinds of supervised pattern recognition model noted above. All of the four supervised pattern recognition models based on the two scores data matrices achieved recognition rates of 100% on the 51 calibration samples. For the prediction work, the Hangju, Taiju, Gongju and Boju data entries were labelled 1, 2, 3 and 4, respectively, and their prediction results were reported as integers (Table 4). However, the results for the RBF-PLS model were reported as non-integers. For these results, if the predicted value of a sample was in one of the following ranges: 1±0.3, 2 ±0.3, 3±0.3 or 4 ±0.3, then, the sample could be placed into one of the following classes: Hangju, Taiju, Gongju or Boju, respectively; otherwise, the sample had to be considered as an outlier. The results (Table 4) demonstrated that RBF-PLS, LS-SVM, LDA and KNN methods performed well, i.e., based on the data for the four types of the Flos Chrysanthemum , the individual samples in the prediction group were correctly placed. This implies that the four cultivars concerned were well discriminated. Aside from this overall success, sample #66 (Boju) was detected as an outlier; this prediction was achieved with the use of the PCA-RBF-PLS and KPCA- RBF-PLS models. The fact this outlier was detected by these two models indicated that both the PCA and KPCA methods were capable of detecting unsuitable or outlier samples. Apart from that important feature selection, the results show that all four models, RBF-PLS, LS-SVM, LDA and KNN, are suitable for predicting unknown, different kinds of Flos Chrysanthemum cultivar sample.

4. Conclusions

The work described, has demonstrated that the combined use of GC and HPLC fingerprints, provide a means for detecting the many characteristic, chemical components that can be found in complex, important

substances such as the four kinds of Flos Chrysanthemum cultivar. The two-dimensional matrix, formed from the collected data, can be resolved by multivariate methods of data analysis. Thus, the qualitative methods such as PCA and KPCA, enabled the discrimination of the objects and apportionment of the chemical components present in the Flos Chrysanthemum cultivars from Hangju, Taiju, Gongju and Boju provinces. The PCA method was shown to be somewhat more effective than the KPCA one; it was able to model the objects and variables. Importantly, PCA indicated the characteristic, important variables for different sample groups. The quantitative RBF-PLS, LS-SVM, LDA and KNN prediction models, all performed well for placing the cultivar objects into their correct groups of origin. Additionally, the prediction results from the four above methods, indicated that both PCA and KPCA were useful methods for data pretreatment prior to quantitative data analysis.

Acknowledgments

The authors greatly appreciate the financial support from the Natural Science Foundation of China (NSFC-21065007) and the State Key Laboratory of Food Science and Technology of Nanchang University (SKLF-ZZA-201302 and SKLF-ZZB-201303).

References

1 National Commission of Chinese Pharmacopoeir, Chinese Pharmacopeia, 1st Ed., Chemical Industry Press, Beijing, 2010.

2 M. Ukiya, T. Akihisa, K. Yasukawa, Y. Kasahara, Y. Kimura, K. Koike, T. Nikaido and M. Takido, J. *Agric. Food Chem*., 2001, **49**, 3187-3197.

3 P. Wang, X. A. Pan, G. M. Chen, J. Li, L. Liu, X. D. Liu, S. Jin, L. Xie and G. J. Wang, *J. Food Sci.*, 2012, **77**(6), 121-127.

4 F. Liu, E. S. Ong and S. F. Li, *Food Chem*., 2013, **141**, 1807-1813.

5 D. X. He, X. C. Ru, L. Wen, Y. C. Wen, H. D. Jiang, I. C. Bruce, J. Jin, X. Ma and Q. Xia, *J. Ethnopharmacol.*, 2012, **139**, 68-73.

6 M. Miyazawa and M. Hisama, *Biosci. Biotechnol. Biochem.,* 2003, **67**(10), 2091-2099.

7 N. T. Luyen, L. H. Tram, T. T. H. Hanh, P. T. Binh, N. H. Dang, C. V. Minh and N. T. Dat, *Phytochem. Lett.,* 2013, **6**, 322-325.

8 Y. R. Niu, L. L. Yin, S. H. Luo, J. Dong, H. Wang, Y. Hashi and S. Z. Chen, *Phytochem. Anal.*, 2013, **24**, 59-68.

9 L. Lin and J. M. Harnly, *Food Chem.,* 2010, **120**, 319-326.

10 M. N. Clifford, W. Wu, J. Kirkpatrick and N. Kuhnert, *J. Agric. Food Chem.*, 2007, **55**, 929-936.

11 K. Chang and G. Kim, *Food Sci. Biotechnol*., 2013, **22**(3), 659-663.

12 Y. J. Wang, X. W. Yang, Q. S. Guo and H. T. Liu, *Chromatographia*, 2009, **70**, 109-116.

13 J. R. Lucio-Gutierrez, A. Garza-Juarez, J. Coello, S. Maspoch, M. L. Salazar-Cavazos, R. Salazar-Aranda and N. W. D. Torres, *J. Chromatogr. A*, 2012, **1235**, 68-76.

14 L. Zhang, X. Y. Wang, J. Z. Guo, Q. L. Xia, G. Zhao, H. N. Zhou and F. W. Xie, *J. Agric. Food Chem.*, 2013, **61**, 2597-2605.

15 L. Calani, D. Beghe, P. Mena, D. D. Rio, R. Bruni, A. Fabbri, C. Dall'Asta and G. Galaverna, *J. Agric. Food Chem.*, 2013, **61**, 5600-5609.

16 G. Fan, M. Y. Zhang, X. D. Zhou, X. R. Lai, Q. H. Yue, C. Tang, W. Z. Luo and Y. Zhang, *Anal. Chim. Acta*, 2012, **747**, 76-83.

17 W. J. Dong, Y. N. Ni and S. Kokot, *J. Agric. Food Chem.*, 2013, **61**, 540-546.

18 Y. Z. Liang, P. S. Xie and K. Chan, *J. Chromatogr. B,* 2004, **812**, 53-70.

19 M. Casale, C. Casolino, P. Oliveri and M. Forina, *Food Chem.*, 2010, **118**, 163-170.

20 Y. Y. Xie, J. L. Qu, Q. L. Wang, Y. Wang, M. Yoshikawa and D. Yuan, *J. Agric. Food Chem.*, 2012, **60**, 12574-12583.

21 C. Zhang, M. J. Qin, P. Shu, J. L. Hong, L. Lu and D. X. He, *Chem. Biodiversity*, 2010, **7**, 2951-2962.

22 J. Wang and H. Xiao, *J. Food Sci. Technol.*, 2013, **50**(5), 986-992.

23 N. A. Sinkov and J. J. Harynuk, *Talanta*, 2013, **103**, 252-259.

24 M. Tufariello, S. Capone and P. Siciliano, *Food Chem.*, 2012, **132**, 2155-2164.

25 L.J. Cao, K. S. Chua, W. K. Chong, H. P. Lee and Q. M. Gu, Neurocomputing, 2003, **55**, 321-336.

26 M. B. Richman and I. Adrianto, *Phys. Chem. Earth,* 2010, **35**, 316-32*8.*

27 B. Scholkopf, A. Smola and K. B. Muller, *Neural Comput.*, 1998, **10**, 1299-1319.

28 F. Marini, *Anal. Chim. Acta*, 2009, **635**, 121-131.

29 S. Hashemi and M. R. Aghamohammadi, *Int. J. Electr. Power Energy Syst.,* 2013, **49**, 86-94.

30 S. A. Etemad, and A. Arya, *Neurocomputing*, 2014, **129**, 585-595.

31 H. M. Monavar, N. K. Afseth, J. Lozano, R. Alimardani, M. Omid and J. P. Wold, *Talanta*, 2013, **111**, 98-104.

32 E. Teye, X.Y. Huang, H. Dai and Q. S. Chen, *Spectrochim. Acta A*, 2013, **114**, 183-189.

33 R. J. Martis, U. R. Acharya, and L. C. Min, *Biomed. Signal Process. Control*, 2013, **8**, 437-448.

34 J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor and J. Vandawalle, Least Squares Support Vector Machines, World Scientific publishing, Singapore, 2002.

35 D. Wu, Y. He, S.J. Feng, and D.W. Sun, *J. Food Eng*., 2008, **84**, 124-131.

36 M. J. Anzanello, R. S. Ortiz, R. P. Limbergerb and P. Mayorga, *J. Pharm. Biomed. Anal.*, 2013, **83**, 209-214.

37 Q. S. Chen, J. W. Zhao, Z. Chen, H. Lin and D. A. Zhao, *Sens. Actuators B*, 2011, **159**, 294-300.

38 Y. N. Ni, B. H. Li and S. Kokot, *Anal. Method*, 2012, **4**, 4326-4333.

39 Y. N. Ni, M. H. Mei and S. Kokot, *J. Liq. Chromatogr. Related Technol.*, 2011, **34**, 1952-1964.

Table 1. Volatile components identified by GC-MS from Flos Chrysanthemum samples

Number	R_t (min)	Volatile components	Formula	Mol. weight	Similarity
1	3.18	Hexanal	$C_6H_{12}O$	100	88
2	3.67	1,6-Dimethylhepta-1,3,5-triene	C_9H_{14}	122	80
3	3.88	5-tert-Butyl-1,3-cyclopentadiene	C_9H_{14}	122	90
4	5.47	alpha.-Phellandrene	$C_{10}H_{16}$	136	92
5	5.69	(3E)-2,7-Dimethyl-3-octen-5-yne	$C_{10}H_{16}$	136	89
6	6.13	7-(1-Methylethylidene) bicyclo[4.1.0] heptane	$C_{10}H_{16}$	136	92
7	6.75	Bicyclo[2.2.1]heptane, 7,7-dimethyl-2-methylene	$C_{10}H_{16}$	136	88
8	6.85	Linalyl n-propionate	$C_{13}H_{22}O_2$	210	83
9	6.97	6-Methyl-5-heptene-2-one	$C_8H_{14}O$	126	89
10	7.19	2H-Benzocyclohepten-2-one, 3,4,4a,5,6,7,8,9-octahydro-	$C_{11}H_{16}O$	164	78
11	7.29	Psi-Cumene	C_9H_{12}	120	85
12	7.97	Terpinolen	$C_{10}H_{16}$	136	93
13	8.19	Benzene, 1-methyl-4-(1-methylethyl)-	$C_{10}H_{14}$	134	90
14	8.47	Eucalyptol	$C_{10}H_{18}O$	154	92
15	9.28	alpha.- Terpinolen	$C_{10}H_{16}$	136	91
16	9.68	cis-.beta.-Terpineol	$C_{10}H_{18}O$	154	90
17	10.15	4,7,7-Trimethylbicyclo [4.1.0]hept-2-ene	$C_{10}H_{16}$	136	88
18	11.42	3,5-Heptadienal, 2-ethylidene-6-methyl-	$C_{10}H_{14}O$	150	85
19	11.93	1,3,3-Trimethylcyclohex-1-ene-4-carboxaldehyde,(+,-)-	$C_{10}H_{16}O$	152	88
20	12.32	2(1H)-Naphthalenone, octahydro-, trans	$C_{10}H_{16}O$	152	87
21	12.86	2(10)-pinen-3-one	$C_{10}H_{14}O$	150	82
22	13.21	Isoborneol	$C_{10}H_{18}O$	154	88
23	13.49	3-Cyclohexen-1-ol, 4-methyl-1-(1-methylethyl)-	$C_{10}H_{18}O$	154	82
24	14.01	Terpineol	$C_{10}H_{18}O$	154	86

25	15.19	Bicyclo[3.1.1]hept-2-en-4-ol, 2,6,6-trimethyl-,acetate	$C_{12}H_{18}O_2$	194	90
26	16.15	4,6,6-Trimethylbicyclo [3.1.1]hept-3-en-2-yl acetate	$C_{12}H_{18}O_2$	194	88
27	17.13	Acetic acid, 1,7,7-trimethyl-bicyclo[2.2.1]hept-2-yl ester	$C_{12}H_{20}O_2$	196	93
28	18.07	1,5-Hexadiene, 2,5-dimethyl-3-methylene-	C_9H_{14}	122	84
29	18.28	1,6-Dimethylhepta-1,3,5-triene	C_9H_{14}	122	81
30	20.44	Chrysanthenone	$C_{10}H_{14}O$	150	82
31	23.73	Benzene, 1-(1,5-dimethyl-4-hexenyl)-4-methyl-	$C_{15}H_{22}$	202	93
32	24.50	Cyclohexene	$C_{15}H_{24}$	204	92
33	24.96	beta.-Sesquiphellandrene	$C_{15}H_{24}$	204	86
34	26.58	Bis(1-methyl-4-pentenyl) phthalate	$C_{20}H_{26}O_4$	330	77
35	28.48	Juniper camphor	$C_{15}H_{26}O$	222	83

Table 2. Common, characteristic, volatile compounds in the four Flos Chrysanthemum cultivars (%)

No.	R_t (min)	Identified compounds	Hangju (%)	Taiju (%)	Gongju (%)	Boju (%)
G1	3.18	1[a]	3.5[b] (0.9-5.3[c], 1.3[d])	2.7 (0.9-5.1, 1.4)	5.6 (3.0-7.6, 1.2)	2.8 (1.1-6.2, 1.4)
G2	5.69	5	4.9 (0.4-14.0, 3.2)	10.3 (6.8-15.6, 2.6)	9.7 (7.1-14.4, 2.1)	1.8 (0.2-9.9, 2.0)
G3	6.13	6	1.6 (0.3-6.5, 1.5)	2.0 (0.3-4.4, 1.0)	10.2 (5.9-17.5, 3.2)	3.3 (0.9-8.0, 2.3)
G4	6.75	7	1.0 (0.3-3.5, 0.8)	2.9 (1.1-5.3, 1.5)	1.0 (0.6-1.4, 0.2)	0.4 (0.1-1.4, 0.3)
G5	7.19	10	9.6 (3.5-14.7, 3.3)	7.1 (4.1-11.6, 1.9)	3.9 (1.5-5.2, 1.0)	1.4 (0.3-3.0, 0.9)
G6	7.97	12	1.6 (0.7-4.7, 1.0)	1.3 (0.8-2.2, 0.4)	1.6 (0.9-2.6, 0.4)	0.5 (0.2-1.6, 0.4)
G7	8.47	14	1.2 (0.2-4.7, 1.1)	2.4 (1.1-6.2, 1.2)	7.7 (5.1-11.3, 1.6)	7.8 (4.1-11.0, 2.1)
G8	9.28	15	1.7 (0.8-5.0, 1.0)	1.6 (1.0-2.6, 0.5)	1.6 (0.9-2.8, 0.4)	0.3 (0.1-0.6, 0.1)
G9	12.32	20	5.3 (1.1-14.0, 3.6)	2.4 (1.0-8.0, 1.9)	20.6 (15.0-26.3, 3.2)	7.5 (2.6-18.0, 4.6)
G10	13.21	22	2.3 (1.2-4.5, 0.9)	5.2 (2.3-7.6, 1.2)	3.4 (1.9-5.3, 0.8)	1.1 (0.5-1.7, 0.3)
G11	15.19	25	2.1 (0.4-8.9, 2.3)	0.6 (0.2-1.0, 0.2)	1.2 (0.2-3.8, 0.9)	0.9 (0.4-1.7, 0.3)
G12	17.13	27	1.4 (0.5-2.9, 0.7)	2.6 (1.3-3.7, 0.8)	1.6 (0.7-2.2, 0.4)	0.5 (0.2-1.4, 0.3)
G13	20.44	30	1.0 (0.4-2.4, 0.6)	0.6 (0.1-1.2, 0.3)	0.9 (0.2-2.2, 0.6)	0.5 (0.1-1.1, 0.2)
G14	23.73	31	3.9 (1.9-7.4, 1.3)	5.4 (3.1-7.6, 1.6)	2.5 (1.5-4.2, 0.7)	5.4 (1.2-9.6, 2.7)
G15	24.50	32	0.6 (0.2-1.1, 0.3)	0.8 (0.5-1.2, 0.2)	0.6 (0.2-1.1, 0.2)	1.7 (0.3-3.9, 1.0)
G16	24.96	33	1.4 (0.7-2.3, 0.4)	2.0 (0.9-3.3, 0.6)	1.3 (0.3-2.2, 0.5)	4.6 (1.2-8.8, 2.4)
G17	26.58	34	16.9 (12.2-25.8, 4.1)	8.0 (4.1-11.9, 2.8)	10.7 (4.5-16.0, 2.9)	8.2 (4.2-14.7, 3.2)

a. Numbers identifying the volatile components in the *Flos Chrysanthemum* samples in Table 1.

b, c and d. Average content (%), content range and standard deviation of each volatile component in all the *Flos Chrysanthemum* samples belonging to the same cultivar, respectively.

Table 3. Similarity of gas chromatograms and HPLC chromatograms of all Flos Chrysanthemum samples

Flos Chrysanthemum	Hangju	Taiju	Gongju	Boju
Gas chromatograms				
Hangju	0.510-0.948[a]	0.362-0.896	0.318-0.834	0.252-0.760
	(0.772[b], 0.137[c])	(0.659, 0.181)	(0.592, 0.181)	(0.492, 0.211)
Taiju		0.577-0.973	0.430-0.739	0.272-0.756
		(0.821, 0.129)	(0.572, 0.116)	(0.482, 0.212)
Gongju			0.817-0.992	0.350-0.808
			(0.953, 0.032)	(0.576, 0.196)
Boju				0.380-0.865
				(0.602, 0.176)
HPLC chromatograms				
Hangju	0.694-0.994[a]	0.643-0.978	0.433-0.857	0.594-0.871
	(0.856[b], 0.090[c])	(0.824, 0.068)	(0.759, 0.104)	(0.758, 0.076)
Taiju		0.723-0.987	0.514-0.845	0.562-0.838
		(0.832, 0.082)	(0.752, 0.076)	(0.753, 0.076)
Gongju			0.769-0.999	0.496-0.876
			(0.922, 0.049)	(0.755, 0.085)
Boju				0.832-0.992
				(0.920, 0.033)

a. Similarities between the two samples.

b. Mean of all similarities of the samples.

c. Relative standard deviation of all the similarities among the samples.

Table 4. Prediction results of 25 samples. PCA and KPCA scores data matrices related to the first three principal components, and then processed separately by four supervised pattern recognition models

Sample	Class[a]	PCA[b]				KPCA[c]			
		RBF-NN	LS-SVM	LDA	KNN	RBF-NN	LSSVM	LDA	KNN
4	1	0.971	1	1	1	1.048	1	1	1
8	1	1.063	1	1	1	1.123	1	1	1
12	1	0.952	1	1	1	0.936	1	1	1
13	1	1.116	1	1	1	1.190	1	1	1
15	1	0.975	1	1	1	0.916	1	1	1
20	2	2.217	2	2	2	2.260	2	2	2
25	2	2.101	2	2	2	1.770	2	2	2
27	2	2.065	2	2	2	2.053	2	2	2
28	2	2.264	2	2	2	2.263	2	2	2
29	2	1.884	2	2	2	1.830	2	2	2
33	3	3.001	3	3	3	3.000	3	3	3
34	3	3.001	3	3	3	3.000	3	3	3
39	3	3.005	3	3	3	3.000	3	3	3
41	3	3.007	3	3	3	3.000	3	3	3
42	3	2.997	3	3	3	3.000	3	3	3
44	3	3.004	3	3	3	3.000	3	3	3
48	3	3.002	3	3	3	3.000	3	3	3
51	3	3.010	3	3	3	3.004	3	3	3
62	4	3.955	4	4	4	4.107	4	4	4
63	4	4.039	4	4	4	4.030	4	4	4
66	4	**4.348**	4	4	4	**4.856**	4	4	4
67	4	4.009	4	4	4	3.988	4	4	4
68	4	4.041	4	4	4	4.020	4	4	4
70	4	4.032	4	4	4	4.026	4	4	4
75	4	3.936	4	4	4	3.903	4	4	4
Prediction Rate		96%	100%	100%	100%	96%	100%	100%	100%

a. The number of 1, 2, 3 and 4 represented Hangju, Taiju, Gongju and Boju, respectively.

b. Parameters of models based on PCA score data matrix. **RBF-NN model parameters**: mean squared error goal = 10^{-5}, spread of radial basis functions = 1, maximum number of neurons = 100 and number of neurons

to add between display = 1; **Optimal parameter for LS-SVM model**: γ = 4.4897 and σ^2 = 1.0089; Number of nearest neighbor for **KNN** was 1.

c. Parameters of models based on KPCA score data matrix. **RBF-NN model parameters**: mean square error goal = 10^{-7}, spread of radial basis functions = 1, maximum number of neurons = 50 and number of neurons to add between displays = 1; **Optimal parameter for LS-SVM model**: γ=0.472 and σ^2=0.5155; Number of nearest neighbor for **KNN** was 1.

Captions

Fig 1. Representative (**A**) GC-MS total ion chromatograms with seventeen common characteristic peaks, and (**B**) HPLC chromatograms with thirteen common characteristic peaks of four kinds of Flos Chrysanthemum cultivars, Hangju, Taiju, Gongju and Boju.

Fig. 2. Loadings plot of PCA projection analysis of peak areas for thirty common characteristic compounds present in 76 Flos Chrysanthemum samples. Circles: thirty characteristic constituents (variables); Outermost circles: the chosen characteristic variables (total 14).

Fig. 3. PCA plot of **(A)** PC1 - PC2, (B) PC1 - PC3 and (C) PC1 - PC2 - PC3 of compressed data matrix (76 samples × 14 variables).

Fig 4. Score plot of KPCA projection analysis of **(A)** KPC1-KPC2, **(B)** KPC1-KPC3 and **(C)** KPC1-KPC2-KPC3 of the compressed data matrix (76 samples × 14 variables).

Figure 1.

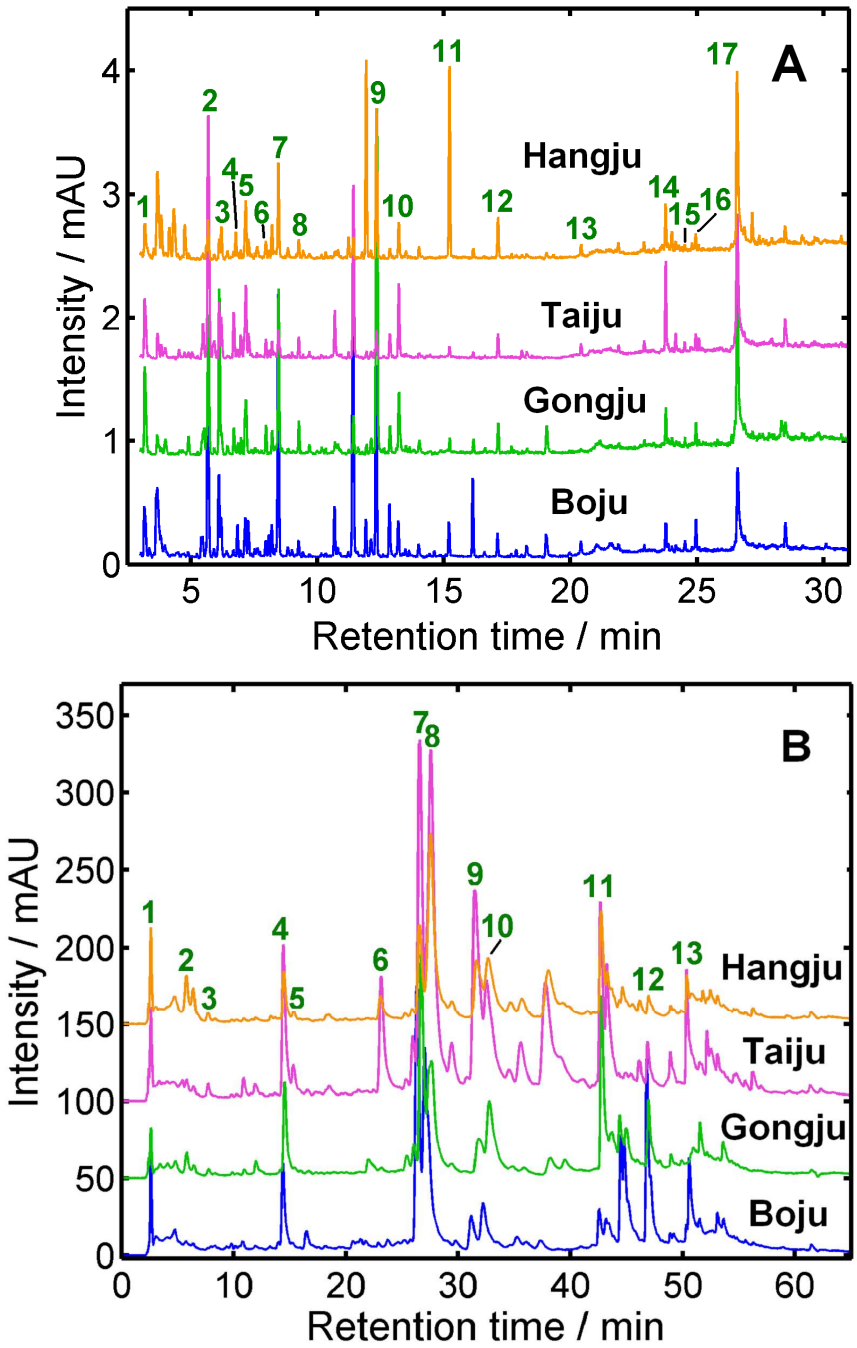


Figure 2

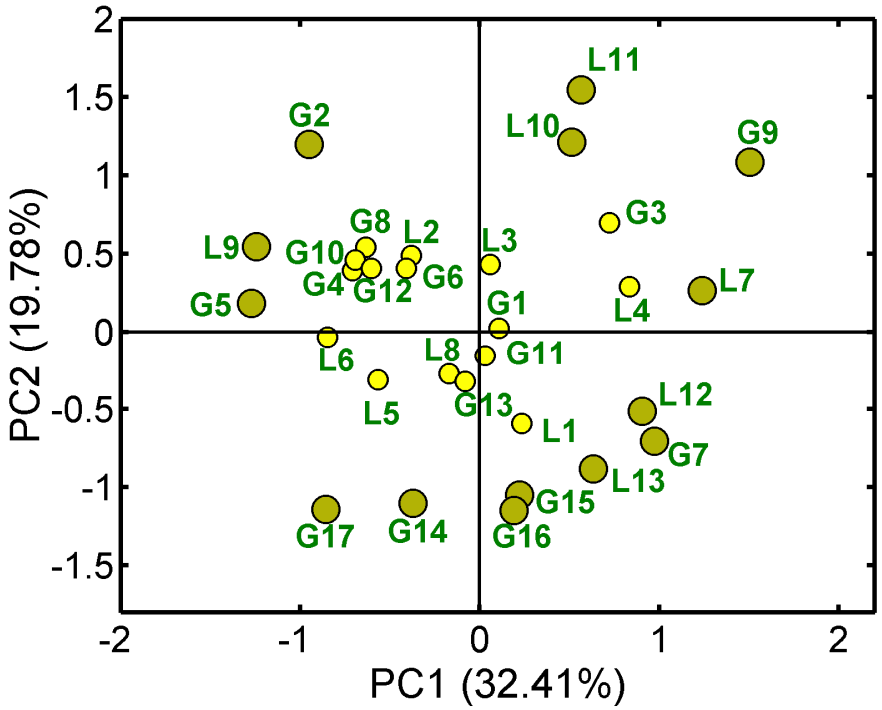


Figure 3

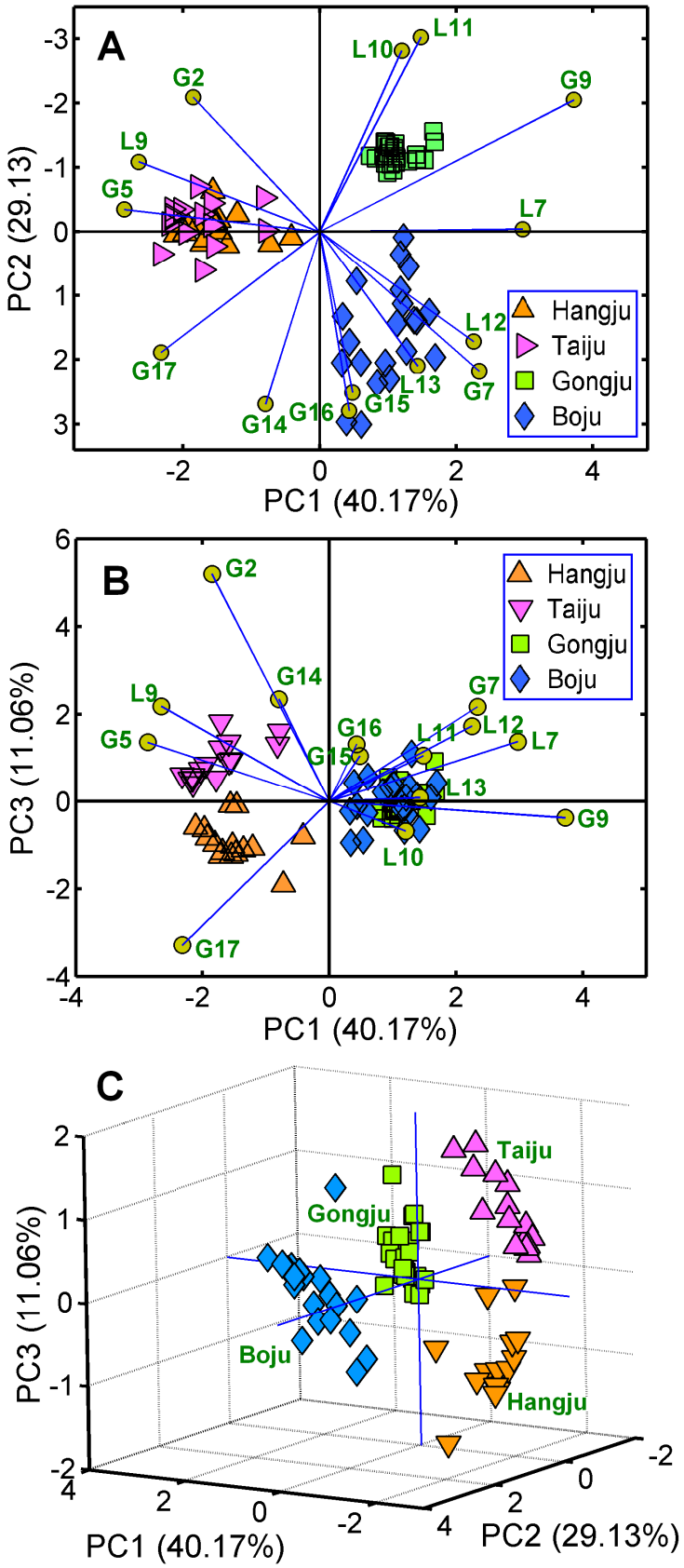


Figure 4