Non-Negative Matrix Factorization-Based SIMCA Method to Classify Traditional Chinese Medicine by HPLC Fingerprints

Bin Qu and **Yuzhu Hu***

Department of Analytical Chemistry, China Pharmaceutical University, Tongjiaxiang 24, Nanjing 210009, China

Abstract

Non-negative matrix factorization (NMF) is a novel technique that decomposes multivariate data into a smaller number of basis vectors and encodings under non-negative constraints. SIMCA (Soft independent modeling by class analogy) is a statistical method for supervised classification of data proposed by S. Wold. In this paper, each classification model is built separately using NMF algorithm rather than principal component analysis (PCA), and each group is described by NMF basis vectors. Then new observations are projected into each NMF model. The residual distances from the new objects to each NMF model are calculated and *F*-test is used to predict whether a new object belongs to some specific group. The capabilities of this novel SIMCA method, named as NMF-based SIMCA, for discriminating two origins of traditional Chinese medicine, *Xiangdan* injection, are studied and compared with classical PCA-based SIMCA method. The results show that NMF can more clearly separate samples than PCA. Better classification accuracy (over 90.0%) is achieved. Factors selection and robust evaluation of classification are studied. It is indicated that in some cases NMF-based SIMCA outperformed PCA-based SIMCA method. NMF and NMF-based SIMCA show promising features and can be practiced in pattern recognition field for discrimination purposes.

Introduction

Herbal medicines and traditional Chinese medicines (TCMs) have been widely used for preventive and therapeutic purposes for centuries, but have been paid particular attention to its efficacy, safety, and quality control during the past decade. Simple quantitative analysis of one or several main components or pharmaceutical active compounds in herbal medicines cannot represent its quality. The fingerprint technique is considered as an effective method to assess the quality of TCMs. The chromatographic fingerprint method has become one of the most frequently applied approaches which can provide the whole profile of not only the marker compounds but also unknown components. Since 2000, chromatographic fingerprints of botanical injections have been demanded by Chinese State Food and Drug Administration as standards for quality control. So far, many chromatographic fingerprints were reported, such as *Ginkgo biloba* (1), *Panax ginseng* (2), *Salvia miltiorrhiza* Bunge (3), *Radix Angelicae Sinensis* (4), *Ganoderma lucidum* (5), and *Qingkailing* injection (6), *Qingfu Guanjieshu* capsule (7), *Shuang-Huang-Lian* oral liquid (8), *Shenmai* injection (9), *Danshen* injection (10). Meanwhile, high-performance thin-layer chromatography (11), high-performance liquid chromatography (HPLC) (12), ultra-performance liquid chromatography (13), gas chromatography (GC) (14), capillary electrophoresis (15), high-speed counter-current chromatography (16), and related hyphenated LC–mass spectrometry (MS) (17), GC–MS (18) approaches have already been employed to develop fingerprints.

Although it is possible to visually differentiate the chromatograms, the process is subjective and not quantitative. Thus similarity evaluation or pattern recognition methods should be taken into consideration for reasonable definition of the class of the herbal medicines. Several chemometric pattern recognition techniques, including hierarchical clustering analysis (HCA) (19), principal component analysis (PCA) (20), and partial least squares-discriminant analysis (PLS-DA) (21) have been widely applied for authentication. HCA is a statistical method for finding relatively homogeneous clusters of cases based on measured characteristics. It starts with each case in a separate cluster and then combines the clusters sequentially, reducing the number of clusters at each step until only one cluster is left. PCA is a popular technique in pattern recognition. It is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. It can be used for dimensionality reduction in a data set by retaining those characteristics of the data set that contribute most to its variance, by keeping lower-order principal components and ignoring higher-order ones. Some low-order components often contain the "most important" aspects of the

*Author to whom correspondence should be addressed: email njhuyuzu@126.com.

data. However, PCA is theoretically the optimum transform for a given data set, although in specific applications the PCA vectors may not always provide the maximum separation of particular subsets of the original data. PLS-DA is performed in order to sharpen the separation between groups of observations by rotating PCA components so that a maximum separation among classes is obtained, and to understand which variables carry the class separating information. It is a regression extension of PCA that takes advantage of class information to attempt to maximize the separation between groups of observations.

In 1999, D.D. Lee and H.S. Seung introduced non-negative matrix factorization (NMF) (22) in its modern formulation as a method to decompose images using non-negative constraints. The main difference between NMF and other classical factorization methods relies on the non-negative constraints imposed on the model. It enforces the constraint that the factors must be non-negative (i.e., all elements must be equal to or greater than zero). These constraints tend to lead to a parts-based representation of the data because they allow only additive, not subtractive, combinations of data items. In this way, the factors produced by this method can be interpreted as parts of the data or as subsets of elements that tend to occur together in sub-portions of the dataset. D.D. Lee and H.S. Seung (22) applied this method to a set of face images and showed that the resulting basis functions represented localized features that corresponded with intuitive notions of the parts of faces (eyes, mouths, noses, etc.). By contrast, they noted that the application of PCA to image data yielded components with no obvious visual interpretation. As a result, NMF leads to decompositions with only non-negative values, unlike PCA with positive and negative values which are no visual interpretation. Therefore, in some cases, NMF may be more suitable than PCA, because it provides an alternative basis set. The comprehensible properties of the NMF method and the intuitivism of the results it provides, have attracted attention of many researchers in different fields of science, and the algorithm is actually applicable to a wide variety of problem domains, such as bioinformatics (23–26), artificial intelligence and pattern recognition (27–29), electrical engineering (30–32), etc. In some cases, NMF has been successfully applied for discrimination purpose. F. Guimet showed that NMF was a powerful technique for learning a meaningful parts-based representation of the fluorescence excitation-emission matrices of different sets of olive oils, and the capabilities of NMF together with Fisher's linear discriminant analysis (LDA) for discriminating between various types of olive oils from excitation-emission fluorescence spectra were studied (33).

Soft independent modeling by class analogy (SIMCA) (34,35) is a pattern recognition technique that builds a model for each class individually on the basis of a specific number of principal components (PCs) and a critical distance with probabilistic meaning. Objects are assigned to a class according to the distances from the class model. If a sample falls in a specific region of the hyperspace, it is assigned to the corresponding class. An *F*-statistic of the variance ratio between the tested object and the training set is used to accept or reject the object to a particular group. A non-significant *F* ratio indicates that the test sample belongs to the training set, though a significant *F* ratio indicates that the test sample is different from the training set. As a class-modeling tool, SIMCA builds a separate model for each category: samples fitting the model are accepted by that category, and samples falling outside the model are considered as outliers for the specific class. If more than one class is modeled, each sample can be assigned to a single category, to more than one category, or to no category at all. After advocated by S. Wold, SIMCA has been widely applied in many fields, such as food science (36–38), pharmacology (39), medicinal chemistry (40), metabonomics (41), and proteomics (42).

Xiangdan injection is a Chinese materia medica preparation produced by many Chinese pharmaceutical companies. It is an aseptic water solution containing the aqueous extracts of *Radix et Rhizoma Salviae Miltiorrhizae* (Danshen) and *Lignum Dalbergiae Odoriferae* (Jiangxiang), which can be used for improving microcirculation, causing coronary vasodilatation, suppressing the formation of thromboxane, inhibiting platelet adhension and aggregation, and protecting against myocardial ischemia (43). In the present study, we apply the NMF algorithm to build classification models, which we refer to as NMF-based SIMCA, to classify two origins of *Xiangdan* injection by HPLC fingerprints. For comparison, the traditional SIMCA advocated by S. Wold is referred to as PCA-based SIMCA. The results have been made comparison between these two SIMCA methods.

Theory and Algorithm

Non-negative matrix factorization (NMF)

The NMF algorithm is a method that compresses a set of n objects and m variables in an $n \times m$ matrix V into a smaller number of basis vectors and their encodings. The factorization is of the form:

$$V = WH + E \qquad \text{Eq. 1}$$

where the columns of the matrix W are called the basis vector and each column of the matrix H is called an encoding. E is the residual matrix. The encoding consists of the coefficients by which matrix V is represented with a linear combination of basis vectors. The dimensions of the matrix factors W and H are $n \times r$ and $r \times m$, respectively. r is the rank of the factorization, and is generally chosen so that $(n + m)\, r < nm$. The product WH can be regarded as a compressed form of the data V. The basis vectors are analogous to PCA scores because they contain information about the objects, whereas the encodings are analogous to PCA loadings, because they are related to variables.

The method starts by randomly initializing matrices W and H with positive values, which iteratively updated to minimize the objective function

$$F = \sum_{i=1}^{n} \sum_{\mu=1}^{m} \left[V_{i\mu} \log \frac{V_{i\mu}}{(WH)_{i\mu}} - V_{iu} + (WH)_{i\mu} \right] \qquad \text{Eq. 2}$$

subjected to the non-negative constraints. Different update rules can be applied for minimizing the objective functions (44). Here we apply the divergence-based update equations (44):

$$W_{ia} = W_{ia} \frac{\sum_{\mu} H_{a\mu} V_{i\mu} / (WH)_{i\mu}}{\sum_{\mu} H_{av}} \quad \text{Eq. 3}$$

$$H_{a\mu} = H_{a\mu} \frac{\sum_{i} W_{ia} V_{i\mu} / (WH)_{i\mu}}{\sum_{k} W_{ka}} \quad \text{Eq. 4}$$

NMF-based SIMCA

Like PCA-based SIMCA, each group is independently modeled using NMF algorithm, and it could be described by a different number of basis vectors of NMF. Class borders defining the quality of acceptable objects are then constructed around the NMF model. In PCA-based SIMCA method the classification results show that models established with different numbers of PCs lead to different success rates. The selection of the appropriate number of PCs, r, is a crucial point in SIMCA. Cross-validation and calculating eigenvalue have been proposed for determining how many PCs are most suited to describe a class (45). However, there is not a clear criterion for selecting the number of significant factors, r, in NMF algorithm and no general rules are given. In this paper, eigenvalues of PCA are used to estimate r of NMF. The cumulative percentage eigenvalue is used to approximately determine what proportion of the data has been modeled. The closer to 100%, the more faithful the model. A simple rule is to accept PCs whose cumulative eigenvalues account for more than a certain percentage (e.g., 95%) of the data. A high explained variance (expl. var.) means that the model describes most of the information contained in the original data.

Suppose the following r dimensional NMF model is obtained:

$$V_k = W_k H_k + E_k \quad \text{Eq. 5}$$

with V_k the matrix $(n \times m)$ for the training set K, W_k the basis vectors $(n \times r)$, H_k the encoding matrix $(r \times m)$ and E_k the matrix of the residuals $(n \times m)$.

The residuals of the training class K towards the model are assumed to follow a normal distribution with a residual standard deviation

$$s = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} e_{ij}^2 / [(n-r)(m-r)]} \quad \text{Eq. 6}$$

After obtaining the basis vectors W_k and the encodings H_k for the training set V_k, the test set v_{test} is projected into the model as following: fixing H_k and starting with positive random w, equation (3) is iterated until the objective function (2) is updated to the minimum. Thus, a representation of a new data vector w_{test} according to the encoding defined in H_k is obtained.

The residual vector e_{test} of test set v_{test} is calculated as

$$e_{test} = v_{test} - w_{test} H_k \quad \text{Eq. 7}$$

The Euclidean distance from the test set to the model is then obtained as

$$s_{test} = \sqrt{\sum_{i=1}^{n} e_{test}^2 / (m-r)} \quad \text{Eq. 8}$$

The dissimilarity between the NMF models and the training objects (with known prior classification) and the test objects (no prior classification) is measured by the goodness of fit according to an F-test. A confidence limit is obtained by defining a critical value of the Euclidean distance towards the model. This is given by

$$s_{crit} = \sqrt{F_{crit} S} \quad \text{Eq. 9}$$

F_{crit} is the tabulated one-sided value for $(m-r)$ and $(m-r)(n-r)$ degrees of freedom. In this paper the level of significant is set to 5 %. If $s_{test} < s_{crit}$, then the test set belongs to the training class K, otherwise it does not.

Experimental

Samples, apparatus, chemicals, and reagents

Samples collected

A total of 60 commercial *Xiangdan* injection samples from two different pharmaceutical manufactures in China (Guangdong Xianfeng Pharmaceutical Co., Ltd., Guangdong, China, and Shanghai Zhongxi Pharmaceutical Co., Ltd., Jiading, China, code named as G and H, respectively) were collected. Thirty samples were from company G and the other 30 samples were from company H. Each sample was produced at different batches, and they were all complied with the requirements of quality standard (No. WS3-B-3289-98) published by Chinese Pharmacopoeia Commission (CPC).

Chemicals and reagents

Protocatechuic aldehyde, tanshinol, salvianolic acid B, and salvianolic acid A (Figure 1) were a gift from Dr. Min Song of China Pharmaceutical University, Nanjing, China. All chemicals and solvents were of analytical grade.

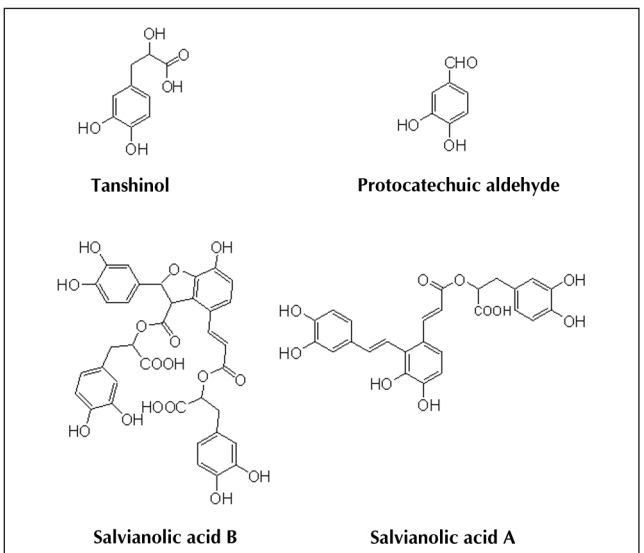

Figure 1. Chemical structures of protocatechuic aldehyde, tanshinol, salvianolic acid B and salvianolic acid A.

Apparatus

A Shimadzu HPLC system (Shimadzu, Kyoto, Japan) was composed of two LC-20AD pumps, SPD-20A ultraviolet detector, CTO-20A column oven, SIL-20A auto-sampler, and LCsolution workstation.

Analytical procedure

Preparation of sample solution

To prepare sample solution, 0.5 mL *Xiangdan* injection sample was diluted into 10 mL with double distilled water. Then the solution was filtered through a 0.45-µm Millipore membrane (mixed cellulose esters) and the filtrate was used as sample solution.

Preparation of reference solutions

The reference solutions for the Xiangdan injection analysis was prepared by dissolving 0.2 mg of protocatechuic aldehyde, tanshinol, salvianolic acid B, and salvianolic acid A reference substances in 1 mL of methanol, respectively.

Method of preparing HPLC fingerprint

The chromatographic conditions were: column: Alltima C_{18} (250 mm × 4.6 mm i.d., 5 µm) (Elite Inc., Dalian, China); injection volume, 20 µL; flow rate, 1.0 mL/min; column temperature, 30°C; detection wavelength, 280 nm. The mobile phase consisted of 1% aqueous acetic acid (A) and 1% acetic acid of methanol solution (B) using a linear gradient program of 10 –100% B for 0–55 min, and 100–10% B for 55–60 min.

Data analysis

Data arrangement

To accurately capture the information encoded in a chromatogram, a chromatographic fingerprint was mathematically represented by a vector of response value of signal, peak area or height of the peak, etc. According to the literature (46–48) in the present study, vector z was assumed to represent HPLC fingerprint

$$z = [z_1, z_2, \ldots, z_n]$$

where z_i denotes absolute area of the each peak of fingerprint and n is the number of the chosen peaks. The vectors were stacked in a matrix of size (samples × number of peaks), and the matrix was used for multivariate analysis.

Software

All the data processing (PCA, NMF, SIMCA) was performed with subroutines developed under MATLAB 7.0 software (The Mathworks, Natick, MA).

Results and Discussion

HPLC fingerprint of *Xiangdan* injection

A representative HPLC fingerprint of *Xiangdan* injection is shown in Figure 2. Compare the chromatograms of the sample solutions with the chromatograms of the reference solution. The 10.8 min, 15.6 min, 30.1 min, and 33.3 min of the reference substances were for tanshinol, protocatechuic aldehyde, salvianolic acid B, and salvianolic acid A.

The HPLC method validation of the fingerprint analysis was performed on the basis of the retention time and the peak area. Thirteen peaks were selected as characteristic peaks in the chromatogram of sample solutions, because the sum of these peak areas was over 95% of the total HPLC chromatogram and it could represent the whole fingerprint. The relative retention times (RRT) and the relative peak areas (RPA) of each characteristic peak related to the reference peak were calculated. The peak 3 (protocatechuic aldehyde) was chosen as reference peak as it has a high and stable content. All the peaks' RRT and RPA were obtained with reference to this substance. The precisions were represented by the relative standard deviation (RSD). The injection precision was determined by replicated injection of the same sample six times in a day. The RSD of RRT and RPA of all characteristic peaks were in the range of 0.2–1.1% and 0.7–3.1% (n = 6), respectively. The intra-day and inter-day precisions of the method were evaluated using multiple preparations of the same sample. Five replicate samples were prepared and analyzed in a single day and on three consecutive days. The intra-day precisions were 0.4–0.9% (n = 5) for RRT and 1.4–3.1% (n = 5) for

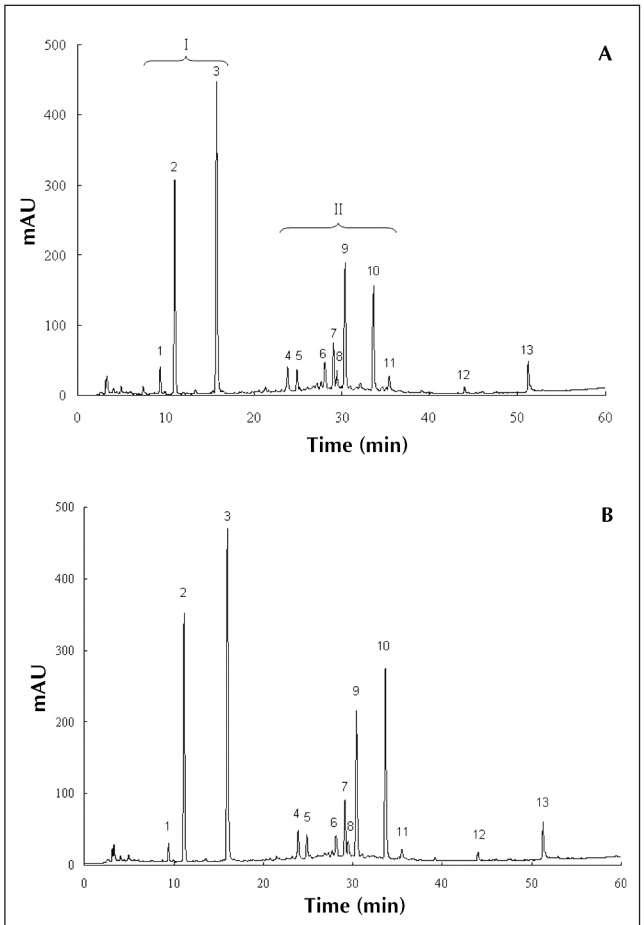

Figure 2. A representative HPLC fingerprint of Xiangdan injection of class G (A) and class H (B). The chromatogram was divided into two major sections: section I (t_R from 5 to 20 min), and section II (t_R from 22 to 35 min). The peaks are tanshinol (2), protocatechuic aldehyde (3), salvianolic acid B(9), and salvianolic acid A(10).

RPA, while the inter-day precisions were 0.6-1.2% (n = 5) for RRT and 2.1–3.8 % (n = 5) for RPA. The stability test was assessed by successive injection of the same sample in 0, 2, 4, 6, 12, and 24 h. The RSD of RRT and RPA were lower than 0.4% and 3.5, respectively. The results indicated that the proposed method is precise and repeatable. Sixty samples of *Xiangdan* injection (G1–G30 and H1–H30) were measured to obtain chromatographic fingerprints.

Visualization of the reduced data of total samples

All the 60 samples were constructed a data matrix with dimensions 60 × 13. Two dimensional reduction of the data set was considered in visualization. The data set was reduced by PCA and NMF independently. Plots of scores of PC2 versus PC1 for objects, G1–G30 and H1–H30, are displayed in Figure 3A. In the PC1–PC2 score plot, all of one class is separated from the other along PC2. Along PC1, very little separation was obtained. Sample G30, which belonged to class G, seemed to be more likely assigned to class H. Samples H11 and H28 seemed to neither belong to class G nor class H. The category threshold was not distinct. Compared with PCA scores plot, NMF basis vectors (BV) plot (Figure 3B) shows that the reduced data by NMF is more clearly separated than that by PCA. The samples were centralized

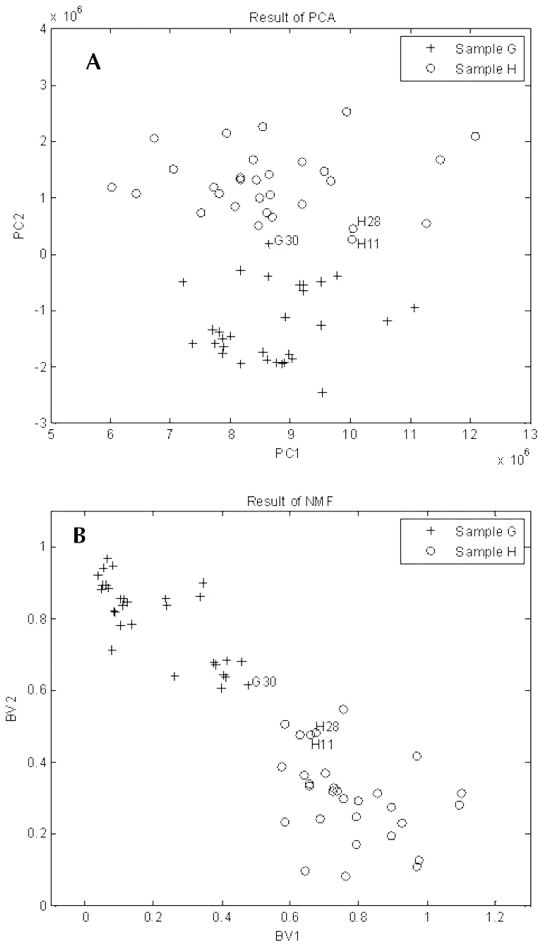

Figure 3. Results of PCA (A) vs NMF (B) for reducing the total 60 samples in visualization. 'PC' stands for principal component of PCA, and 'BV' means basis vector of NMF.

in two diagonal regions of the NMF BVs plot independently. Samples were divided into two well distinct classes, corresponding to class G and class H. Samples G30, H11, and H28 were all assigned to respective class correctly.

Generally, the angle between two vectors x and y, is given by

$$\theta = \arccos\left(\frac{x \cdot y}{|x||y|}\right) \quad \text{Eq. 10}$$

Here the dot product x · y is given by

$$x \cdot y = \sum_{i=1}^{n} x_i y_i \quad \text{Eq. 11}$$

where x_i and y_i are coordinates of vectors x and y, respectively, and the length |x| is given by

$$|x| = \sqrt{\sum_{i=1}^{n} x_i^2} \quad \text{Eq. 12}$$

PCA reduced dimensionality and generated new components PC1 and PC2 in the directions of maximal variance. The PC1 and PC2 must be orthogonal, so the angle between them was 90°. However, NMF involves factorization into matrices with non-negative entries. Because of the non-negativity requirements, NMF algorithm makes the basis vectors non-orthogonal. In this case, the angle between BV1 and BV2 is 59°. So they do not correspond to directions of maximal variance. This is the significant feature of NMF distinguished from PCA.

Selection of the training and test sets

The samples in the training set were chosen on the basis of the scores plot of PC2 versus PC1. The criterion was to cover the entire variability domain in order to assume that the data in the training set were sufficiently representative for the class structure. The rest of the samples were included in the test set. In this paper, each original data set was split into a training set of 15 samples and a test set of another 15 samples.

Classification analysis

PCA was used to estimate factors by means of calculating eigenvalues. Two-factor models of training set G with 99.2% expl. var. and training set H with 98.7% expl. var. were obtained. Two factors were considered firstly. NMF was applied to the matrix of the training set. The procedure was followed for all the NMF-based SIMCA models presented in this paper. Table I shows the percentage of correct classification for the training set and the test set. The results of correct classification for all the samples in the training set and the test set of class G and class H

Table I. Percentage of Correct Cassification for the Xiangdan Injection by SIMCA Method

		NMF-based SIMCA	PCA-based SIMCA (not mean-centered)	PCA-based SIMCA (mean-centered)
Training set	G	93.3	86.7	93.3
	H	93.3	93.3	93.3
Test set	G	86.7	66.7	66.7
	H	93.3	93.3	80.0
Total		91.7	85.0	83.3

were all over 85%. Regarding all the samples H, a 93.3% of correct classification was obtained for the training and the test set. For all the samples, 5 samples were misclassified and the total correct classification was achieved at 91.7%.

PCA-based SIMCA is the original SIMCA method. Therefore we applied PCA-based SIMCA to the same data sets and compared the results to those obtained by NMF-based SIMCA. Generally, the training set is mean-centered prior to building the model and the test set is also estimated on the centroid of the training set before residual analysis. However data must not be centered before applying NMF algorithm, because they must remain non-negative. So in this paper, not mean-centered PCA-based SIMCA was also performed and compared. The classification results are also shown in Table I. Only for the samples H of the training set, a 93.3% of correct classification was obtained, as in the case of NMF-based SIMCA. As can be seen, some results of classification were worse than that of NMF. For samples G of the test set, the correct classification rate is below 80%, both in the cases of not mean-centered and mean-centered PCA-based SIMCA models. The percentage of correct classification in the two models was lower than that in NMF for the training set and the test set.

As a class modeling tool, two additional figures of merit are usually computed to evaluate the performances of SIMCA, sensitivity and specificity. Sensitivity is the percentage of samples from the modeled class that are accepted by the class model, and specificity is the percentage of samples from other classes which are rejected by the class model. The results of NMF-based and PCA-based SIMCA modeling on the data set are reported in Table II. For both models, high specificity (100%) for the class G and class H were obtained. The discriminating performance relied on the sensitivity of the model. The NMF-based classification model showed an acceptable classification within the training set and a better predictive ability in the test set than PCA-based model.

Coomans plot is usually as a useful tool in the interpretation of SIMCA method. In a Coomans plot, the two axes represent the distance of each sample from a specific category, so that each class model is drawn as a rectangle corresponding to the critical distance (p = 0.05) from the class. Any sample having a distance to the corresponding centroid greater than the critical distance is considered as being outside the class model and, as a consequence, rejected as an outlier for the specific category (graphically, it is plotted outside the rectangle defining the class model). Moreover, the samples plotted onto the lower left square of the diagram are assigned to both classes. In this case, the Coomans plot of NMF-based SIMCA is shown in Figure 4. The distance from the model for class G is plotted against that from class H. On both axes, one indicates the critical distance. When inspecting the plot, it is possible to observe that the samples belonging to the class G and class H are almost mapped onto their individual category rectangle and all of them fall significantly apart from the lower left square of the figure. By plotting objects in this plot, their classification was immediately clear. It is easy to visualize the classification that samples G15, G24, G30, H14, and H24 are misclassified to the outlier zone. None of these objects was wrongly identified to overlap zone.

More details can be confirmed by HPLC fingerprint chromatograms. For convenience of evaluation, the chromatogram was divided into two major sections and characterized as followed: section I (t_R region from 5 to 20 min), and section II (t_R region from 22 to 35 min). These major peaks are of most significance to the overall characteristics of the fingerprint pattern. According to the reference solutions, the highest peak (3) in the fingerprint chromatogram presented (Figure 2A) is attributed to protocatechuic aldehyde, peak 2 to tanshinol, peak 9 to salvianolic acid B, and peak 10 to salvianolic acid A.

In the literature (2), the HPLC fingerprint of EGb761 can be seen as the standard pattern against which to compare other *Ginkgo biloba* preparations from different sources. However, there is still not a standard or reference for *Xiangdan* injection. The quality standard (No.WS3-B-3289-98) published by CPC only limits the concentration of protocatechuic aldehyde not less than 0.17 mg/mL. So in this study, the model was built independently by its own products. This model will be regarded as the reference of the products. Each original product will

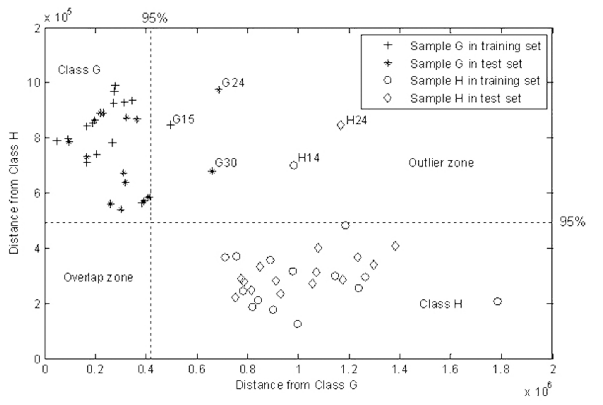

Figure 4. Coomans plots of NMF-based SIMCA model. Horizontal and vertical dotted lines represent the critical distance (p = 0.05) of the class G and H models, respectively.

Table II. SIMCA Sensitivity and Specificity

	NMF-based SIMCA				PCA-based SIMCA (not mean-centered)				PCA-based SIMCA (mean-centered)			
	Sensitivity		Specificity		Sensitivity		Specificity		Sensitivity		Specificity	
	G	H	G	H	G	H	G	H	G	H	G	H
Training set	93.3	93.3	100	100	86.7	93.3	100	100	93.3	93.3	100	100
Test set	86.7	93.3	100	100	66.7	93.3	100	100	66.7	80.0	100	100

have their individual model. The outliers can be viewed in the Coomans plot. It indicates that samples G15, G24, G30, H14, and H24 are misclassified to the outlier zone. In the fingerprint chromatogram of G15 (Figure 5A), peak 3 was uncharacteristically predominating. In the cases of samples G24 and G30, peaks 9 and 10 were uncharacteristically predominating (Figure 5B of G24). And for samples H14 and H24, peak 9 was weak and peak 10 was dramatically weak (Figure 5C of H14). Although these products were quantitative of protocatechuic aldehyde by convenient HPLC test, these outliers would not be evident. Therefore, simply determining the content of protocatechuic aldehyde alone is not sufficient for the quality assessment of *Xiangdan* injection products. Establishing HPLC fingerprint to construct an overall pattern seems specific for the identification and quality evaluation.

The NMF algorithm was started with randomly initializing matrices W and H with positive values, which are iteratively updated to minimize a divergence function. MATLAB command, rand(n,r) and rand(r,m), were run to generate random matrices W and H with uniformly distributed pseudo-random numbers on the unit interval. Due to this stochastic nature of the NMF algorithm (24), results might differ from one run to the other. One hundred repeated runs were carried out with random initial conditions, and the probability distribution of results is shown that one more sample was misclassified, so another correct classification rate of 90.0% took place, but the frequency was lower. Only two results (90.0% correct at the frequency of 37.0% and 91.7% correct at frequency of 63.0%) were produced, so the result of correct classification rate is stable and the algorithm of NMF-based SIMCA is robust.

F. Guimet chose the best factor with best classification of NMF model by selecting several different numbers of factors, such as factors (2–6), when NMF combined with Fisher's LDA was performed to discriminate various types of olive oils (33). In this study, factors (1–3) were also considered. The results of total correct classification rate and corresponding probability distribution were listed in Table III, after 100 repeated and random runs, respectively. Although the total correct classification rate was unique when factor was 1, the results were lower than that when the number of factors was 2. When it was 3, 93.3% of total correct classification rate was obtained, but the frequency was lower, and more than three results were generated at different low frequency. Therefore, 3 factors were not suitable to build NMF model. After comparison, the best classification could be obtained from the NMF-based SIMCA model using 2 factors, and it was feasible and suitable to select factors by means of eigenvalue.

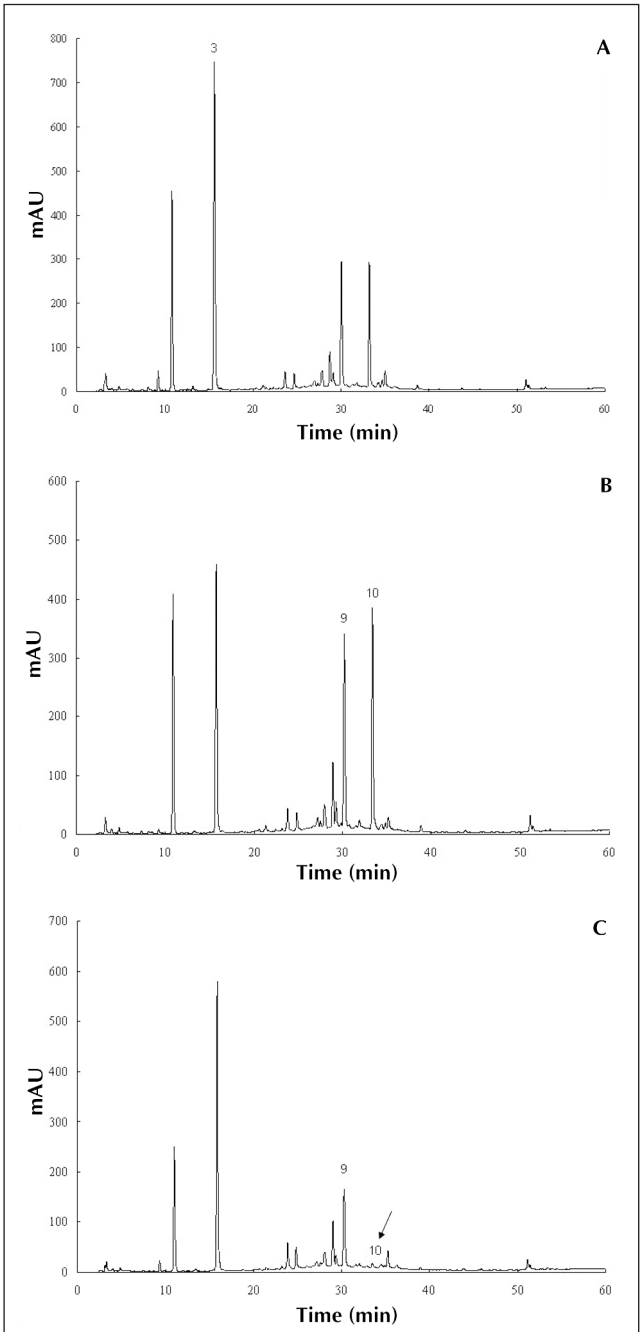

Figure 5. The HPLC fingerprint of sample G15 (A) showing peak 3 uncharacteristically predominating, sample G24 (B) showing peak 9 and 10 uncharacteristically predominating and sample H14 (C) showing peak 10 dramatically weak. Note the abnormally increased or decreased peaks dominating the fingerprint in comparison with the representative fingerprint in Figure 2.

Table III. Correct Classification Rate (CCR)*

	Factors										
	1	2		3							
CCR	81.7%	91.7%	90.0%	81.7%	83.3%	85.0%	86.7%	88.3%	90.0%	91.7%	93.3%
Frequency	100%	63%	37%	5%	8%	17%	29%	6%	13%	13%	9%

* When factors were 1–3. The frequencies of the corresponding CCR were statistics after 100 runs.

Fisher's LDA is another traditional discrimination method. F. Guimet firstly checked the capabilities of NMF together with Fisher's LDA for classification of different sets of olive oils (33). In this paper, NMF was applied to the matrix containing the chromatograms of the training set including the 15 samples of class G and 15 samples of class H. After obtaining the basis vectors and the encodings for the training set, the test set of another 15 samples of class G and 15 samples of class H was projected to the models and the basis vectors of the new samples were calculated. Afterwards, the NMF basis vectors of the training set were used to compute the Fisher's linear discriminant functions for discriminating between the two classes of *Xiangdan* injections. The NMF basis vectors of the test set were also used as test set for Fisher's LDA. The boundaries between classes were set taking the centroid of each class and drawing a line halfway between the pair of centroids. In this case, all the samples of training set and test set were classified in the correct class by NMF-Fisher's LDA. The classifications were 100% for class G and class H, respectively. By comparison of the results in the Table I, LDA showed a better performance. However, the different LDA performance with expect to SIMCA has to be expected to some extent. LDA is based on the whole data set of the training set which consists of all the classes to build model. It is often called forms of hard modeling. SIMCA, as it named soft independent modeling of class analogy, is regarded as a form of soft modeling used in pattern recognition. It independently builds a separate model for each category. Additional class can be added independently to the existing model without any changes. In contrast, when LDA is used, the entire modeling procedure must be repeated if extra numbers of groups are added, since the Fisher's score need be recalculated.

Cross-validation of NMF-based SIMCA

In NMF-based SIMCA, each class is modeled separated by NMF. The performance of the method depends not only on the difference between classes, but also strongly on the training set for each class. In this case, the class G and H contain some inhomogeneities. In general, some of dissimilarities observed in the fingerprints within one class can be related to factors such as origins, growth periodicity of herbs, and climate. The samples for each class were obtained from different batches, so some natural heterogeneity in the individual classes was expected. Although this natural variance was observed, each sample included in the database had passed all quality tests and was therefore released for use. As a result, all fingerprints are kept for the data analysis with SIMCA, since they represent a real life situation one encounters in the pharmaceutical industry and cannot be considered as analytical outliers. However, checking for atypical objects must be carried out carefully, since samples in the database define the quality of the classification models.

In SIMCA, a classification model is constructed for each class individually. The performance of the model is evaluated by doing leave-one-out cross validation (LOOCV) within the corresponding class. The confidence level was set at 95% which indicates that the theoretical expected amount of wrongly rejected samples is 5%. The number of BVs used for modeling is 2, and the correct classification rate obtained by LOOCV were 86.7% and 80.0% for the original data of class G and H, respectively.

As mentioned previously, some objects were misclassified. The first reason for that has to do with the natural heterogeneity and the dimension of the dataset. In LOOCV, the system was perturbed by leave out one object at the time to estimate the classification performance. As soon as an extreme sample is left out, the remaining objects will not span the same space anymore and as a result the object left out will not be classified in this class. This is a realistic way of evaluating the performance of the method, because in real life situation, one must expect that new samples with extreme characteristics will be submitted for prediction. The results show that NMF-based SIMCA is sensitive to dissimilarities between objects. Some authors propose to develop stable models by deleting all outliers and repeating the model for the remaining objects (49) or investigate principal component outlier detection methods (50) and the robustness of the SIMCA method (51). However, in this application no objects should be removed, because the data set represents real world variances. As more samples are included in the data set and made the database as representative as possible, better classification results will be obtained.

Conclusions

This work presents the first attempt to propose a chemometric methodology, NMF-based SIMCA, to evaluate the quality of one of the most important TCM preparations, *Xiangdan* injection. The HPLC fingerprint technique, which is not necessary to give information on the molecular nature of the compounds, gives a useful and information-rich fingerprint characteristic of each sample. This blind method, coupled with suitable chemometric tools, may represent a powerful tool for discrimination and classification of different batches of samples. NMF, as a new feature extraction method, produces positive basis vectors and encodings under non-negative constraints of the algorithm. Due to the difference between HPLC fingerprint and face images advocated by D.D. Lee and H.S. Seung (22), the basis vectors and encodings therefore cannot represent local fingerprint peaks of the chromatograms, which is not like intuitive notions of the faces' parts in the application of face images. However, NMF is still a powerful method of feature extraction to provide a new feature in non-negative space. The parts-based representative capability of NMF combined with classical SIMCA method, NMF-based SIMCA method, was proposed, and the discrimination between two different origins of *Xinagdan* injection was studied. NMF and PCA were compared for reducing fingerprint data matrix in visualization and discriminating analysis through SIMCA method. In some cases, correct classifications above 90.0% were achieved, and NMF-based SIMCA yielded higher classification accuracy than PCA-based SIMCA with the total samples of the training set and the test set. The results indicate the promising features of NMF and NMF-based SIMCA method and they are applicable techniques for discrimination purposes. Further research on the method of selecting initial matrix is needed in order to improve robustness of NMF.

References

1. Y.B. Ji, Q.S. Xu, Y.Z. Hu, and Y.V. Heyden. Development, optimization and validation of a fingerprint of Ginkgo biloba extracts by high-performance liquid chromatography. *J. Chromatogr. A* **1066:** 97–104 (2005).
2. P.S. Xie, S.B. Chen, Y.Z. Liang, X. Wang, R. Tian, and R. Upton. Chromatographic fingerprint analysis—a rational approach for quality assessment of traditional Chinese herbal medicine. *J. Chromatogr. A* **1112:** 171–180 (2006).
3. M. Gu, S.F. Zhang, Z.G. Su, and O.Y. Fan. Fingerprinting of Salvia miltiorrhiza Bunge by non-aqueous capillary electrophoresis compared with high-speed counter-current chromatography. *J. Chromatogr. A* **1057:** 133–140 (2004).
4. G.H. Lu, K. Chan, Y.Z. Liang, K. Leung, C.L. Chan, Z.H. Jiang, and Z.Z. Zhao. Development of high-performance liquid chromatographic fingerprints for distinguishing Chinese Angelica from related umbelliferae herbs. *J. Chromatogr. A* **1073:** 383–392 (2005).
5. Y. Chen, S.B. Zhu, M.Y. Xie, S.P. Nie, W. Liu, C. Li, X.F. Gong, and Y.X. Wang. Quality control and original discrimination of Ganoderma lucidum based on high-performance liquid chromatographic fingerprints and combined chemometrics methods. *Anal. Chim. Acta* **623:** 146–156 (2008).
6. S.K. Yan, W.F. Xin, G.A. Luo, Y.M. Wang, and Y.Y. Cheng. An approach to develop two-dimensional fingerprint for the quality control of Qingkailing injection by high-performance liquid chromatography with diode array detection. *J. Chromatogr. A* **1090:** 90–97 (2005).
7. Y. Xie, Z.H. Jiang, H. Zhou, X. Cai, Y.F. Wong, Z.Q. Liu, Z.X. Bian, H.X. Xu, and L. Liu. Combinative method using HPLC quantitative and qualitative analyses for quality consistency assessment of a herbal medicinal preparation. *J. Pharm. Biomed. Anal.* **43:** 204–212 (2007).
8. Y.H. Cao, L.C. Wang, X.J. Yu, and J. Ye. Development of the chromatographic fingerprint of herbal preparations Shuang-Huang-Lian oral liquid. *J. Pharm. Biomed. Anal.* **41:** 845–856 (2006).
9. X.H. Fan, Y. Wang, and Y.Y. Cheng. LC/MS fingerprinting of Shenmai injection: a novel approach to quality control of herbal medicines. *J. Pharm. Biomed. Anal.* **40:** 591–597 (2006).
10. J.L. Zhang, M. Cui, Y. He, H.L. Yu, and D.A. Guo. Chemical fingerprint and metabolic fingerprint analysis of Danshen injection by HPLC-UV and HPLC-MS methods. *J. Pharm. Biomed. Anal.* **36:** 1029–1035 (2005).
11. S.B. Chen, H.P. Liu, R.T. Tian, D.J. Yang, S.L. Chen, H.X. Xu, A.S. Chan, and P.S. Xie. High-performance thin-layer chromatographic fingerprints of isoflavonoids for distinguishing between Radix Puerariae Lobate and Radix Puerariae Thomsonii. *J. Chromatogr. A* **1121:** 114–119 (2006).
12. W. Jin, R.L. Ge, Q.J. Wei, T.Y. Bao, H.M. Shi, and P.F. Tu. Development of high-performance liquid chromatographic fingerprint for the quality control of Rheum tanguticum Maxim. ex Balf. *J. Chromatogr. A* **1132:** 320–324 (2006).
13. M. Liu, Y.G. Li, G.X. Chou, X. Cheng, M. Zhang, and Z. Wang. Extraction and ultra-performance liquid chromatography of hydrophilic and lipophilic bioactive components in a Chinese herb Radix Salviae Miltiorrhizae. *J. Chromatogr. A* **1157:** 51–55 (2007).
14. F.Q. Guo, T.Z. Liang, C.J. Xu, L.F. Huang, and X.N. Li. Comparison of the volatile constituents of Artemisia capillaris from different locations by gas chromatography-mass spectrometry and projection method. *J. Chromatogr. A* **1054:** 73–79 (2004).
15. Y.B. Ji, G. Alaerts, C.J. Xu, Y.Z. Hu, and Y.V. Heyden. Sequential uniform designs for fingerprints development of Ginkgo biloba extracts by capillary electrophoresis. *J. Chromatogr. A* **1128:** 273–281 (2006).
16. M. Gu, G.F. Zhang, Z.G. Su, and O.Y. Fan. Identification of major active constituents in the fingerprint of *Salvia miltiorrhiza* Bunge developed by high-speed counter-current chromatography. *J. Chromatogr. A* **1041:** 239–243 (2004).
17. J. Kang, L. Zhou, J.H. Sun, J. Han, and D.A. Guo. Chromatographic fingerprint analysis and characterization of furocoumarins in the roots of Angelica dahurica by HPLC/DAD/ESI-MSn technique. *J. Pharm. Biomed. Anal.* **47:** 778–785 (2008).
18. L.F. Hu, S.P. Li, H. Cao, J.J. Liu, J.L. Gao, F.Q. Yang, and Y.T. Wang. GC-MS fingerprint of Pogostemon cablin in China. *J. Pharm. Biomed. Anal.* **42:** 200–206 (2006).
19. P. Zou, Y. Hong, and H.L. Koh. Chemical fingerprinting of Isatis indigotica root by RP-HPLC and hierarchical clustering analysis. *J. Pharm. Biomed. Anal.* **38:** 514–520 (2005).
20. X.L. Piao, J.H. Park, J. Cui, D.H. Kim, and H.H. Yoo. Development of gas chromatographic/mass spectrometry-pattern recognition method for the quality control of Korean Angelica. *J. Pharm. Biomed. Anal.* **44:** 1163–1167 (2007).
21. L.Z. Yi, D.L. Yuan, Y.Z. Liang, P.S. Xie, and Y. Zhao. Quality control and discrimination of pericarpium citri reticulatae and pericarpium citri reticulatae viride based on high-performance liquid chromatographic fingerprints and multivariate statistical analysis. *Anal. Chim. Acta* **588:** 207–215 (2007).
22. D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature* **401:** 788–791 (1999).
23. Y. Gao and G. Church. Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics* **21:** 3970–3975 (2005).
24. J.P. Brunet, P. Tamayo, T.R. Golub, and J.P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U.S.A.* **101:** 4164–4169 (2004).
25. P. Carmona-Saez, R.D. Pascual-Marqui, F. Tirado, J.M. Carazo, and A. Pascual-Montano. Biclustering of gene expression data by non-smooth non-negative matrix factorization. *BMC Bioinformatics* **7:** 78–85 (2006).
26. P. Fogel, S.S. Young, D.M. Hawkins, and N. Ledirac. Inferential, robust non-negative matrix factorization analysis of microarray data. *Bioinformatics* **23:** 44–49 (2007).
27. S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas. Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification.-*IEEE Trans. Neural Netw.* **17:** 683–695 (2006).
28. D. Guillamet, J. Vitria, and B. Schiele. Introducing a weighted non-negative matrix factorization for image classification. *Pattern Recognit. Lett.* **24:** 2447–2454 (2003).
29. F. Shahnaz, M.W. Berry, V.P. Pauca, and R.J. Plemmons. Document clustering using nonnegative matrix factorization. *Inf. Process. Manage.* **42:** 373–386 (2006).
30. F.J. Theis and G.A. Garcia. On the use of sparse signal decomposition in the analysis of multi-channel surface electromyograms. *Signal Process.* **86:** 603–623 (2006).
31. P. Sajda, S.Y. Du, T.R. Brown, R. Stoyanova, D.C. Shungu, X.L. Mao, and L.C. Parra. Nonnegative matrix factorization for rapid recovery of constituent spectra in magnetic resonance chemical shift imaging of the brain. *IEEE Trans. Med. Imaging* **23:** 1453–1465 (2004).
32. R. Zdunek, and A. Cichocki. Nonnegative matrix factorization with constrained second-order optimization. *Signal Process.* **87:** 1904–1916 (2007).
33. F. Guimet, R. Boque, and J. Ferre. Application of non-negative matrix factorization combined with fisher's linear discriminant analysis for classification of olive oil excitation-emission fluorescence spectra. *Chemom. Intell. Lab. Syst.* **81:** 94–106 (2006).
34. B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. de Jong, P.J. Lewi, and J. Smeyers-Verbeke. *Handbook of Chemometrics and Qualimetrics, Part B*, Elsevier, Amsterdam, 1998, pp. 228
35. S. Wold. Pattern recognition by means of disjoint principal components models. *Pattern Recognition* **8:** 127–139 (1976).
36. D. Gonzalez-Arjona, G. Lopez-Perez, V. Gonzalez-Gallero, and A.G. Gonzalez. Supervised pattern recognition procedures for discrimination of whiskeys from gas chromatography/mass spectrometry congener analysis. *J. Agric. Food Chem.* **54:** 1982–1989 (2006).
37. F. Marini, A.L. Magri, R. Bucci, F. Balestrieri, and D. Marini. Class-modeling techniques in the authentication of Italian oils from Sicily with a Protected Denomination of Origin (PDO). *Chemometrics Intell. Lab. Syst.* **80:** 140–149 (2006).
38. M. Cocchi, C. Durante, A. Marchetti, C. Armanino, and M. Casale. Characterization and discrimination of different aged'Aceto Balsamico Tradizionale di Modena'products by head space mass spectrometry and chemometrics. *Anal. Chim. Acta* **589:** 96–104 (2007).
39. P.R.N. Wolohan and R.D. Clark. Predicting drug pharmacokinetic properties using molecular interaction fields and SIMCA. *J. Comput. Aid. Mol. Des.* **17:** 65–76 (2003).
40. J.J. Sutherland, and D.F. Weaver. Three-dimensional quantitative structure-activity and structure-selectivity relationships of dihydrofolate reductase inhibitors. *J. Comput.-Aided Mol. Des.* **18:** 309–331 (2004).
41. E.Holmes, A.W. Nicholls, J.C. Lindon, S.C. Connor, J.C. Connelly, J.N. Haselden, S.J. Damment, M. Spraul, P. Neidig, and J.K. Nicholson. Chemometric models for toxicity based on NMR spectra of biofluids. *Chem. Res. Toxicol.* **13:** 471–478 (2000).
42. E. Marengo, E. Robotti, M. Bobba, and P.G. Righetti. Evaluation of the variables characterized by significant discriminating power in the application of SIMCA classification method to proteomic studies. *J. Proteome Res.* **7:** 2789–2796 (2008).
43. T.O. Cheng. Danshen: A versatile Chinese herbal drug for the treatment of coronary heart disease. *Int. J. Cardiol.* **113:** 437-438 (2006).
44. D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. *Adv. Neural Inform. Process. Syst.* **13:** 556–562 (2001).
45. R.G. Brereton. *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*, John Wiley & Sons, Chichester, 2003, pp. 196
46. Y.P. Li, Z. Hu, and L.C. He. An approach to develop binary chromatographic fingerprints of the total alkaloids from Caulophyllum robustum by high performance liquid chromatography/diode array detector and gas chromatography/mass spectrometry. *J. Pharm. Biomed. Anal.* **43:** 1667–1672 (2007).
47. F. Gan and R.Y. Ye. New approach on similarity analysis of chromatographic fingerprint of herbal medicine. *J. Chromatogr. A* **1104:** 100–105 (2006).
48. X.H. Fan, Y.Y. Cheng, Z.L. Ye, R.C. Lin, and Z.Z. Qian. Multiple chromatographic fingerprinting and its application to the quality control of herbal medicines. *Anal. Chim. Acta* **555:** 217–224 (2006).
49. M.P. Derde, D. Coomans, and D.L. Massart. Effect of scaling on class modeling with SIMCA. *Anal. Chim. Acta* **141:** 187–192 (1982).
50. B. Mertens, M. Thompson, T. Fearn. Principal component outlier detection and SIMCA: a synthesis. *Analyst* **119:** 2777–2784 (1994)
51. K.V. Branden, and M. Hubert. Robust classification in high dimensions based on the SIMCA method. *Chemometrics Intell. Lab. Syst.* **79:** 10–21 (2005).

Manuscript received March 8, 2009;
revision received May 22, 2009.