Clustering and Classification of Analytical Data

Barry K. Lavine
Clarkson University, Potsdam, USA

Clustering and classification are the major subdivisions of pattern recognition techniques. Using these techniques, samples can be classified according to a specific property by measurements indirectly related to the property of interest (such as the type of fuel responsible for an underground spill). An empirical relationship or classification rule can be developed from a set of samples for which the property of interest and the measurements are known. The classification rule can then be used to predict the property in samples that are not part of the original training set. The set of samples for which the property of interest and measurements is known is called the training set. The set of measurements that describe each sample in the data set is called a pattern. The determination of the property of interest by assigning a sample to its respective category is called recognition, hence the term pattern recognition.

For pattern recognition analysis, each sample is represented as a data vector $\boldsymbol{x} = (x_1, x_2, x_3, x_j, \ldots, x_n)$*, where component* x_j *is a measurement, e.g. the area* a *of the* j*th peak in a chromatogram. Thus, each sample is considered as a point in an* n*-dimensional measurement space. The dimensionality of the space corresponds to the number of measurements that are available for each sample. A basic assumption is that the distance between pairs of points in this measurement space is inversely related to the degree of similarity between the corresponding samples. Points representing samples from one class will cluster in a limited region of the measurement space distant from the points corresponding to the other class. Pattern recognition (i.e. clustering and classification) is a set of methods for investigating data represented in this manner, in order to assess its overall structure, which is defined as the overall relationship of each sample to every other in the data set.*

1 INTRODUCTION

Since the early 1980s, a major effort has been made to substantially improve the analytical methodology applied to the study of environmental samples. Instrumental techniques such as gas chromatography, high-performance liquid chromatography (HPLC) and X-ray fluorescence spectroscopy have dramatically increased the number of organic and inorganic compounds that can be identified and quantified, even at trace levels, in the environment. This capability, in turn, has allowed scientists to attack ever more complex problems, such as oil and fuel spill identification, but has also led to an information-handling problem.[1]

The reason for this problem is that in any monitoring effort it is necessary to analyze a large number of samples in order to assess the wide variation in composition that an environmental system may possess. The large number of samples that must be analyzed and the number of constituents that must be measured per sample give rise to data sets of enormous size and complexity. Often, important relationships in these data sets cannot be uncovered, when the data are examined one variable at a time, because of correlations between measurement variables, which tend to dominate the data and prevent information from being extracted.

Furthermore, the relationships sought in the data often cannot be expressed in quantitative terms, such as the source of a pollutant in the environment. These relationships are better expressed in terms of similarity or dissimilarity among groups of multivariate data. The task that confronts the scientist when investigating these sorts of relationships in multivariate data, is twofold:

- Can a useful structure based on distinct sample groups be discerned?

Encyclopedia of Analytical Chemistry
Edited by Robert A. Meyers. © John Wiley & Sons Ltd, Chichester. ISBN 0471 97670 9

- Can a sample be classified into one of these groups for the prediction of some property?

The first question is addressed using principal component analysis (PCA)[2] or cluster analysis,[3] whereas the second question is addressed using pattern recognition methods.[4]

PCA is the most widely used multivariate analysis technique in science and engineering.[5] It is a method for transforming the original measurement variables into new variables called principal components. Each principal component is a linear combination of the original measurement variables. Often, only two or three principal components are necessary to explain all of the information present in the data. By plotting the data in a coordinate system defined by the two or three largest principal components, it is possible to identify key relationships in the data, that is, find similarities and differences among objects (such as chromatograms or spectra) in a data set.

Cluster analysis[6] is the name given to a set of techniques that seek to determine the structural characteristics of a data set by dividing the data into groups, clusters, or hierarchies. Samples within the same group are more similar to each other than samples in different groups. Cluster analysis is an exploratory data analysis procedure. Hence, it is usually applied to data sets for which there is no a priori knowledge concerning the class membership of the samples.

Pattern recognition[7] is a name given to a set of techniques developed to solve the class-membership problem. In a typical pattern recognition study, samples are classified according to a specific property using measurements that are indirectly related to that property. An empirical relationship or classification rule is developed from a set of samples for which the property of interest and the measurements are known. The classification rule is then used to predict this property in samples that are not part of the original training set. The property in question may be the type of fuel responsible for a spill, and the measurements are the areas of selected gas chromatographic (GC) peaks. Classification is synonymous with pattern recognition, and scientists have turned to it and PCA and cluster analysis to analyze the large data sets typically generated in monitoring studies that employ computerized instrumentation.

This article explores the techniques of PCA, cluster analysis, and classification. The procedures that must be implemented to apply these techniques to real problems are also enumerated. Special emphasis is placed on the application of these techniques to problems in environmental analysis.

2 PRINCIPAL COMPONENT ANALYSIS

PCA is probably the oldest and best known of the techniques used for multivariate analysis. The overall goal of PCA is to reduce the dimensionality of a data set, while simultaneously retaining the information present in the data. Dimensionality reduction or data compression is possible with PCA because chemical data sets are often redundant. That is, chemical data sets are not information rich. Consider a gas chromatogram of a JP-4 fuel (Figure 1), which is a mixture of alkanes, alkenes, and aromatics. The gas chromatogram of a JP-4 fuel is characterized by a large number of early-eluting peaks, which are large in size. There are a few late-eluting peaks, but their size is small. Clearly, there is a strong negative correlation between the early- and late-eluting peaks of the JP-4 fuel. Furthermore, many of the alkane and alkene peaks are correlated, which should not come as a surprise as alkenes are not constituents of crude oil but instead are formed from alkanes during the refining process. In addition, the property of a fuel most likely to be reflected in a high resolution gas chromatogram is its distillation curve, which does not require all 85 peaks for characterization.

Redundancy in data is due to collinearity (i.e. correlations) among the measurement variables. Collinearity diminishes the information content of the data. Consider a set of samples characterized by two measurements, X_1 and X_2. Figure 2 shows a plot of these data in a two-dimensional measurement space, where the coordinate axes (or basis vectors) of this measurement space are the variables X_1 and X_2. There appears to be a relationship between these two measurement variables, which suggests that X_1 and X_2 are correlated, because fixing the value of X_1 limits the range of values possible for X_2. If the two measurement variables were uncorrelated, the enclosed rectangle in Figure 2 would be fully populated by the data points. Because information is defined as the scatter of points in a measurement space, it is evident that correlations between the measurement variables decrease the information content of this space. The data points, which are restricted to a small region of the measurement

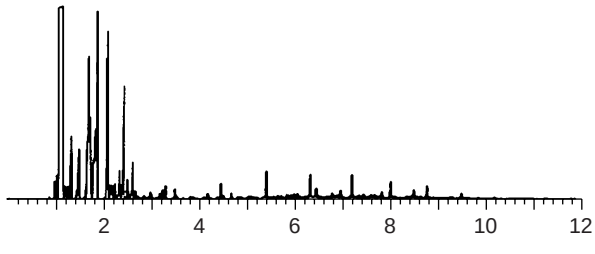

Figure 1 A high-resolution capillary column gas chromatogram of a JP-4 fuel.

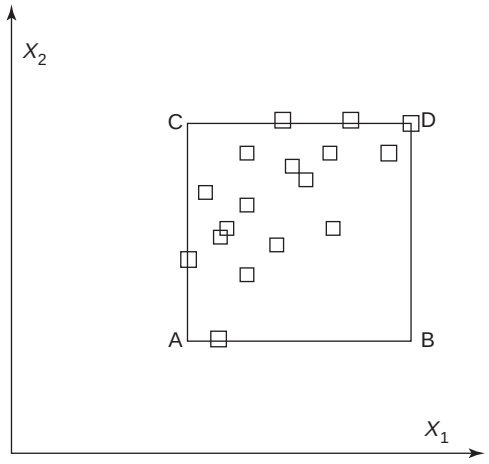

Figure 2 Seventeen hypothetical samples projected onto a two-dimensional measurement space defined by the measurement variables X_1 and X_2. The vertices, A, B, C, and D, of the rectangle represent the smallest and largest values of X_1 and X_2. (Adapted from Mandel.[8])

space due to correlations among the variables, could even reside in a subspace if the measurement variables are highly correlated. This is shown in Figure 3. Here X_3 is perfectly correlated with X_1 and X_2 because X_1 plus X_2 equals X_3. Hence, the seven sample points lie in a plane even though each data point has three measurements associated with it.

2.1 Variance-based Coordinate System

Variables that have a great deal of redundancy or are highly correlated are said to be collinear. High collinearity between variables is a strong indication that a new set of basis vectors can be found that will be better at conveying the information content present in data than axes defined by the original measurement variables. The new basis set that is linked to variation in the data can be used to develop a new coordinate system for displaying the data.

$$X = \begin{bmatrix} X_1 & X_2 & X_3 \\ 0.5 & 0.5 & 1.0 \\ 1.9 & 0.7 & 2.6 \\ 2.0 & 2.0 & 4.0 \\ 0.3 & 1.8 & 2.1 \\ 1.9 & 1.7 & 3.6 \\ 1.2 & 0.2 & 1.4 \\ 1.9 & 0.9 & 2.8 \end{bmatrix}$$

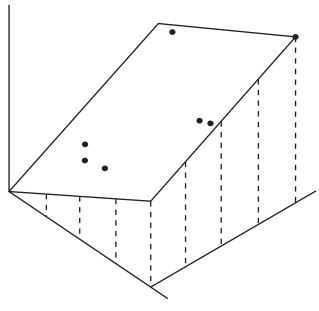

Figure 3 In the case of strongly correlated variables, the data points may reside in a subspace of the original measurement space. (Adapted from *Multivariate Pattern Recognition in Chemometrics*.[4] Copyright 1992, with permission from Elsevier Science.)

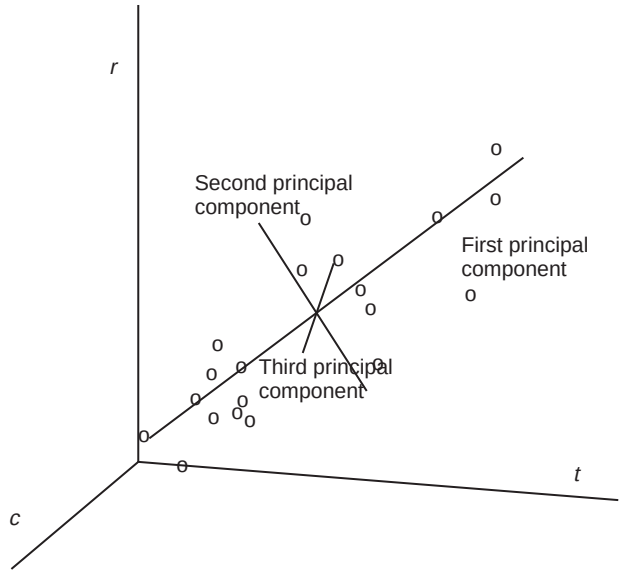

Figure 4 Principal component axes developed from the measurement variables *r*, *c*, and *t*. (Reproduced by permission from Brown[5] courtesy of Society of Applied Spectroscopy.)

The principal components of the data define the variance-based axes of this new coordinate system. The largest or first principal component is formed by determining the direction of largest variation in the original measurement space and modeling it with a line fitted by linear least squares (Figure 4) that passes through the center of the data. The second largest principal component lies in the direction of next largest variation – it passes through the center of the data and is orthogonal to the first principal component. The third largest principal component lies in the direction of next largest variation – it also passes through the center of the data, it is orthogonal to the first and second principal components, and so forth. Each principal component describes a different source of information because each defines a different direction of scatter or variance in the data. (The scatter of the data points in the measurement space is a direct measure of the data's variance.) Hence, the orthogonality constraint imposed by the mathematics of PCA ensures that each variance-based axis will be independent.

2.2 Information Content of Principal Components

One measure of the amount of information conveyed by each principal component is the variance of the data explained by the principal component. The variance explained by each principal component is expressed in terms of its eigenvalue. For this reason, principal components are usually arranged in order of decreasing eigenvalues or waning information content. The most informative principal component is the first and the least

informative is the last. The maximum number of principal components that can be extracted from the data is the smaller of either the number of samples or number of measurements in the data set, as this number defines the largest number of independent variables in the data.

If the data are collected with due care, one would expect that only the first few principal components would convey information about the signal, as most of the information in the data should be about the effect or property of interest being studied. However, the situation is not always this straightforward. Each principal component describes some amount of signal and some amount of noise in the data because of accidental correlation between signal and noise. The larger principal components primarily describe signal variation, whereas the smaller principal components essentially describe noise. When smaller principal components are deleted, noise is being discarded from the data, but so is a small amount of signal. However, the reduction in noise more than compensates for the biased representation of the signal that results from discarding principal components that contain a small amount of signal but a large amount of noise. Plotting the data in a coordinate system defined by the two or three largest principal components often provides more than enough information about the overall structure of the data. This approach to describing a data set in terms of important and unimportant variation is known as soft modeling in latent variables.

PCA takes advantage of the fact that a large amount of data is usually generated in monitoring studies when sophisticated chemical instrumentation, which is commonly under computer control, is used. The data have a great deal of redundancy and therefore a great deal of collinearity. Because the measurement variables are correlated, 85 peak gas chromatograms do not necessarily require 85 independent axes to define the position of the sample points. Utilizing PCA, the original measurement variables that constitute a correlated axis system can be converted into a system that removes correlation by forcing the new axes to be independent and orthogonal. This requirement greatly simplifies the data because the correlations present in the data often allow us to use fewer axes to describe the sample points. Hence, the gas chromatograms of a set of JP-4 and Jet-A fuel samples may reside in a subspace of the 85-dimensional measurement space. A plot of the two or three largest principal components of the data can help us to visualize the relative position of the Jet-A and JP-4 fuel samples in this subspace.

2.3 Case Studies

With PCA, we are able to plot the data in a new coordinate system based on variance. The origin of the new coordinate system is the center of the data, and the coordinate axes of the new system are the principal components of the data. Employing this new coordinate system, we can uncover relationships present in the data, that is, find distinct samples subgroups within the data. This section shows, by way of two published studies, how principal components can be used to discern similarities and differences among sample within a data set.

2.3.1 Troodos Data Set

In the first study, 143 rock samples collected in the Troodos region of Cyprus were analyzed by X-ray fluorescence spectroscopy for 10 metal oxides, which contained information about the formation of these rocks. If the formation of the entire Troodos region occurred at the same time, one would expect all of the rocks to be similar in composition. However, if there are distinct subgroups in the data, other conclusions may have to be drawn about the formation of the Troodos region. This study[9] was initiated to settle a controversy about the geological history of Cyprus.

Figure 5 shows a plot of the two largest principal components of the 143 rock samples. (The original Troodos data set was modified for the purpose of this principal component mapping exercise.) Samples 65 and 66 appear to be outliers in the plot as they are distant from the other samples. As a general rule, outliers should deleted because of the least-squares property of principal components. In other words, a sample that is distant from the other points in the measurement space can pull the principal components towards it and away from the direction of maximum variance. Figure 6 shows the results of a principal component mapping experiment

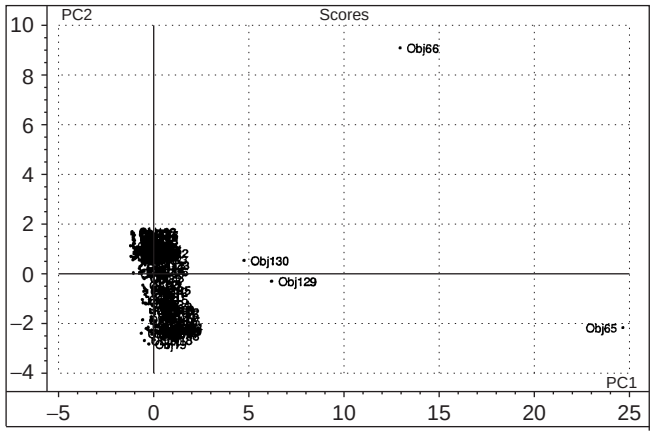

Figure 5 A principal component map of the 143 rock samples. Samples 65 and 66 are outliers. (The original data set was modified for this principal component mapping exercise.) The principal component map was generated using the program UNSCRAMBLER.

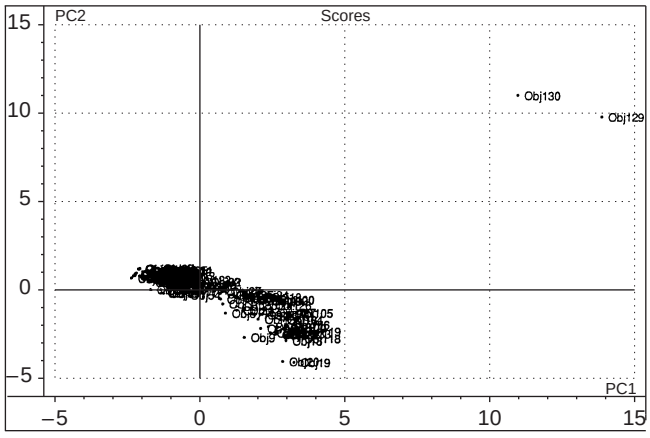

Figure 6 A principal component map of the Troodos rock samples with samples 65 and 66 removed. Samples 129 and 130 appear as outliers in the plot. (The original data set was modified for this principal component mapping exercise.) The principal component map was generated using the program UNSCRAMBLER.

with samples 65 and 66 removed from the data. It is evident from the plot that samples 129 and 130 are also outliers. Figure 7 summarizes the results of a principal component mapping experiment with samples 65, 66, 129, and 130 removed. Although samples 19 and 20 are probably outliers and are also candidates for removal, it is evident from the principal component plot that the rock samples can be divided into two groups, which would suggest that other conclusions should be drawn about the geological history of the Troodos region. The clustering of the rocks samples into two distinct groups was not apparent until the four outliers were removed from the data.

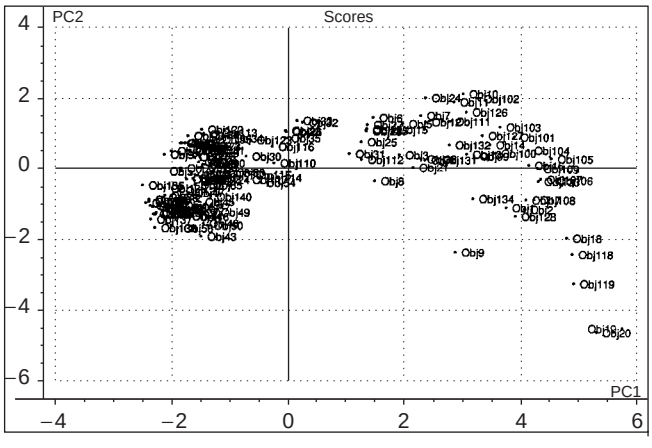

Figure 7 A principal component map of the Troodos rock samples with samples 65, 66, 129, and 130 removed. The principal component map was generated using the program UNSCRAMBLER.

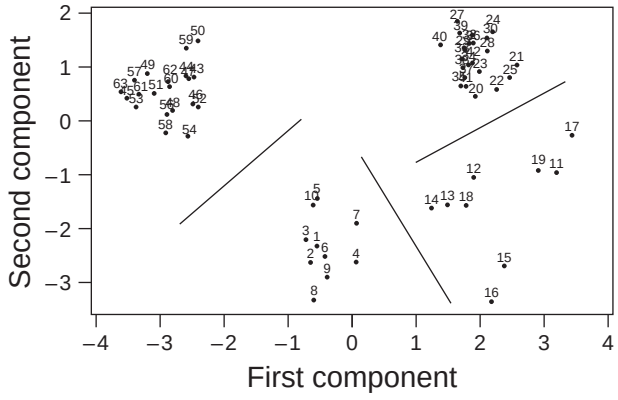

Figure 8 A plot of the two largest principal components for the 63 Indian artifacts developed from the concentration data of 10 metals. The principal component map was generated using the program SCAN.

2.3.2 Obsidian Data Set

The second study[10] also involves X-ray fluorescence data. Sixty-three Indian artifacts (such as jewelry, weapons, and tools) made from volcanic glass, were collected from four quarries in the San Francisco Bay area. (Samples 1–10 are from quarry 1, samples 11–19 are from quarry 2, samples 20–42 are from quarry 3, and samples 43–63 are from quarry 4.) Because the composition of volcanic glass is characteristic of the site and tends to be homogeneous, it is reasonable to assume that it should be possible to trace these artifacts to their original source material. In this study, the investigators attempted to do this by analyzing the 63 glass samples for 10 elements: Fe, Ti, Ba, Ca, K, Mn, Rb, Sr, Y, and Zn. Next, a PCA was performed on the data (63 artifacts with 10 features per artifact). The goal was to identify the overall trends present in the data. Figure 8 shows a plot of the two largest principal components of the data. From the principal component map, it is evident that the 63 Indian artifacts can be divided into four groups, which correspond to the quarry sites from which the artifacts were collected. Evidently, the artifacts in each quarry were made from the same source material. This result is significant because it provides the archaeologists with important information about the migration patterns and trading routes of the Indians in this region. Further details about the obsidian data can be found elsewhere.[10]

3 CLUSTER ANALYSIS

Cluster analysis is a popular technique whose basic objective is to discover sample groupings within data. The technique is encountered in many fields, such as biology, geology, and geochemistry, under such

names as unsupervised pattern recognition and numerical taxonomy. Clustering methods are divided into three categories, hierarchical, object-functional, and graph theoretical. The focus here is on hierarchical methods, as they are the most popular.

For cluster analysis, each sample is treated as a point in an n-dimensional measurement space. The coordinate axes of this space are defined by the measurements used to characterize the samples. Cluster analysis assesses the similarity between samples by measuring the distances between the points in the measurement space. Samples that are similar will lie close to one another, whereas dissimilar samples are distant from each other. The choice of the distance metric to express similarity between samples in a data set depends on the type of measurement variables used.

Typically, three types of variables – categorical, ordinal, and continuous – are used to characterize chemical samples. Categorical variables denote the assignment of a sample to a specific category. Each category is represented by a number, such as 1, 2, 3, etc. Ordinal variables are categorical variables, in which the categories follow a logical progression or order, such as 1, 2, and 3 denoting low, middle, and high, respectively. However, continuous variables are quantitative. The difference between two values for a continuous variable has a precise meaning. If a continuous variable assumes the values 1, 2, and 3, the difference between the values 3 and 2 will have the same meaning as the difference between the values 2 and 1, because they are equal.

Measurement variables are usually continuous. For continuous variables, the Euclidean distance is the best choice for the distance metric, because interpoint distances between the samples can be computed directly

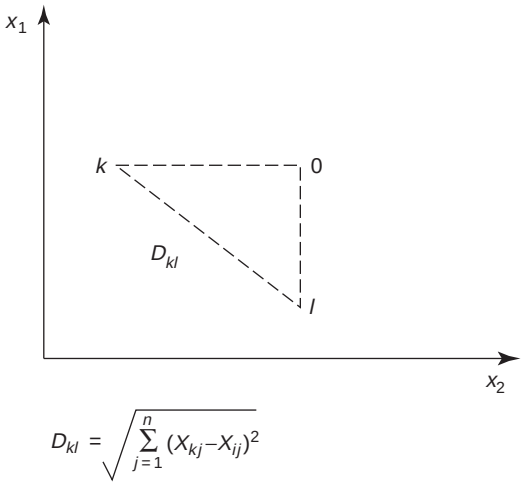

Figure 9 Euclidean distance between two data points in a two-dimensional measurement space defined by the measurement variables x_1 and x_2. (Reprinted from ref. 3.)

(Figure 9). However, there is a problem with using the Euclidean distance, which is the so-called scaling effect. It arises from inadvertent weighting of the variables in the analysis that can occur due to differences in magnitude among the measurement variables. For example, consider a data set where each sample is described by two variables: the concentration of Na and the concentration of K as measured by atomic flame emission spectroscopy. The concentration of Na varies from 50 to 500 ppm, whereas the concentration of K in the same samples varies from 5 to 50 ppm. A 10% change in the Na concentration will have a greater effect on Euclidean distance than a 10% change in K concentration. The influence of variable scaling on the Euclidean distance can be mitigated by autoscaling the data, which involves standardizing the measurement variables, so that each variable has a mean of zero and a standard deviation of 1 (Equation 1):

$$x_{i,\text{standardized}} = \frac{x_{i,\text{orig}} - m_{i,\text{orig}}}{s_{i,\text{orig}}} \tag{1}$$

where $x_{i,\text{orig}}$ is the original measurement variable i, $m_{i,\text{orig}}$ is the mean of the original measurement variable i, and $s_{i,\text{orig}}$ is the standard deviation of the original measurement variable i. Thus, a 10% change in K concentration has the same effect on the Euclidean distance as a 10% change in Na concentration when the data is autoscaled. Clearly, autoscaling ensures that each measurement variable has an equal weight in the analysis. For cluster analysis, it is best to autoscale the data, because similarity is directly determined by a majority vote of the measurement variables.

3.1 Hierarchical Clustering

Clustering methods attempt to find clusters of patterns (i.e. data points) in the measurement space, hence the term cluster analysis. Although several clustering algorithms exist, e.g. K-means, K-median, Patrick-Jarvis, FCV (fuzzy clustering varieties), hierarchical clustering is by far the most widely used clustering method. The starting point for a hierarchical clustering experiment is the similarity matrix which is formed by first computing the distances between all pairs of points in the data set. Each distance is then converted into a similarity value (Equation 2):

$$s_{ik} = 1 - \frac{d_{ik}}{d_{\max}} \tag{2}$$

where s_{ik} (which varies from 0 to 1) is the similarity between samples i and k, d_{ik} is the Euclidean distance between samples i and k, and $d_{\max}$ is the distance between the two most dissimilar samples (i.e. the largest distance) in the data set. The similarity values are organized in the form of a table or matrix. The similarity matrix is

then scanned for the largest value, which corresponds to the most similar point pair. The two samples constituting the point pair are combined to form a new point, which is located midway between the two original points. The rows and columns corresponding to the old data points are then removed from the matrix. The similarity matrix for the data set is then recomputed. In other words, the matrix is updated to include information about the similarity between the new point and every other point in the data set. The new nearest point pair is identified, and combined to form a single point. This process is repeated until all points have been linked.

There are a variety of ways to compute the distances between data points and clusters in hierarchical clustering (Figure 10). The single-linkage method assesses similarity between a point and a cluster of points by measuring the distance to the closest point in the cluster. The complete linkage method assesses similarity by measuring the distance to the farthest point in the cluster. Average linkage assesses the similarity by computing the distances between all point pairs where a member of each pair belongs to the cluster. The average of these distances is used to compute the similarity between the data point and the cluster.

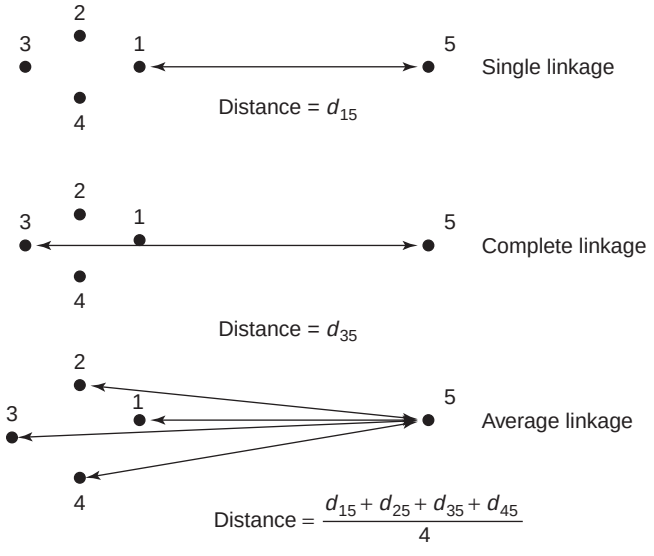

Figure 10 The distance between a data cluster and a point using single linkage, complete linkage, and average linkage. (Reproduced by permission from Lavine[20] courtesy of Marcel Dekker, Inc.)

Table 1 HPLC data set

Column	k_1	k_2	k_3
A	0.31	17.8	3
B	0.10	9.30	3
C	0.11	21.5	1
D	0.58	22.0	2
E	0.50	16.0	1

(Reprinted from *Multivariate Pattern Recognition in Chemometrics*.[4] Copyright 1992, by permission of Elsevier Science.)

To illustrate hierarchical clustering, consider the data shown in Table 1. (The example shown here is an adaptation of the exercise described in Chapter 6 of reference 4.) Five HPLC columns were characterized by the capacity factor values obtained from three substances, which served as retention probes. To perform single-linkage hierarchical clustering on this chromatographic data, it is necessary to first compute the similarity matrix for the data, given as Table 2.

The similarity matrix (Table 2) is then scanned for the largest value, which corresponds to the two HPLC columns that are most similar. An examination of the similarity matrix suggests that chromatographic columns A and B with a score of 0.79 are the most similar. Hence, chromatographic columns A and B should be combined to form a new point. The rows and columns corresponding to the two original points (A and B) are removed from the similarity matrix. The similarity matrix for the data set is then updated to include information about the similarity between the new point and every other point (C, D, and E) in the data set. In this study, the investigators chose to use the single-linkage criterion for assessing the similarity between a data point and a point cluster, see Table 3. (Using the single-linkage method, the similarity between point D and the cluster consisting of columns A and B is the larger of the values of 0.69 and 0.17, see Table 2. For complete linkage, the similarity between this cluster and point D is the smaller of the two values.)

The updated similarity matrix is then scanned for the largest value; the new nearest pair is combined to form a single point, which is points D and E. The rows and columns corresponding to points D and E are deleted

Table 2 Similarity matrix

Columns	A	B	C	D	E
A	1.00	0.79	0.58	0.69	0.61
B	0.79	1.00	0.36	0.17	0.34
C	0.58	0.36	1.00	0.51	0.72
D	0.69	0.17	0.51	1.00	0.75
E	0.61	0.34	0.72	0.75	1.00

Table 3 Updated similarity matrix

Columns	A, B	C	D	E
A, B	1.00	0.58	0.69	0.61
C	0.58	1.00	0.51	0.72
D	0.69	0.51	1.00	0.75
E	0.61	0.72	0.75	1.00

Table 4 Updated similarity matrix

Columns	A, B	C	D, E
A, B	1.00	0.58	0.61
C	0.58	1.00	0.72
D, E	0.61	0.72	1.00

Table 5 Updated similarity matrix

Columns	A, B	C, D, E
A, B	1.00	0.69
C, D, E	0.69	1.00

from the similarity matrix. The similarity matrix for the data set is then updated (Table 4) to include information about the similarity between the new point (D + E) and every other point in the data set. This process is repeated (Table 5) until all points are merged into a single cluster.

The results of a hierarchical clustering study are usually displayed as a dendogram, which is a tree-shaped map of the intersample distances in the data set. The dendogram shows the merging of samples into clusters at various stages of the analysis and the similarities at which the clusters merge, with the clustering displayed hierarchically. The dendogram for the single-linkage analysis of the HPLC data is shown in Figure 11. Interpretation of the results is intuitive, which is the major reason for the popularity of these methods.

3.2 Practical Considerations

A major problem in hierarchical clustering (or cluster analysis for that matter) is defining a cluster. Contrary to many published reports, there is no cluster validity measure that can serve as an indicator of the quality of a proposed partitioning of the data. Hence, clusters are defined intuitively, depending on the context of the problem, not mathematically, which limits the utility of this technique. Clearly, prior knowledge about the problem is essential when using these methods. The criterion for determining the threshold value for similarity is often subjective and depends to a large degree on the nature of the problem investigated – for example, the goals of the study, the number of clusters sought, previous experience, and common sense.

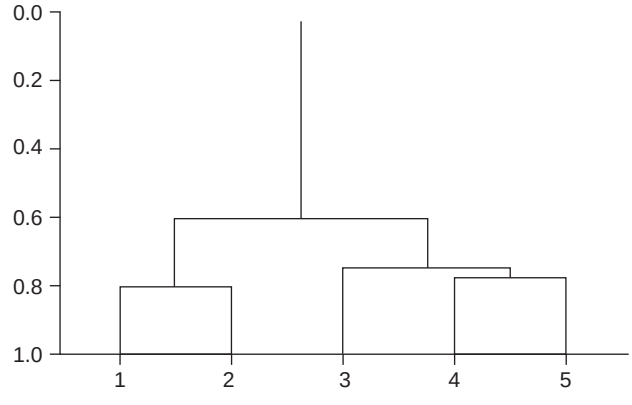

Figure 11 Single-linkage dendogram of the HPLC data set. There are two groups of HPLC columns: 1, 2 and 3, 4, 5. (Reprinted from *Multivariate Pattern Recognition in Chemometrics*.[4] Copyright 1992, with permission from Elsevier Science.)

All clustering procedures yield the same results for data sets with well-separated clusters. However, the results will differ when the clusters overlap. That is why it is a good idea to use at least two different clustering algorithms, such as single and complete linkage, when studying a data set. If the dendograms are in agreement, then a strong case can be made for partitioning the data into distinct groups as suggested by the dendograms. If the cluster memberships differ, the data should be further investigated using average linkage or PCA. The results from average linkage or PCA can be used to gauge whether the single or farthest linkage solution is the better one.

All hierarchical clustering techniques suffer from so-called space distorting effects. For example, single-linkage favors the formation of large linear clusters instead of the usual elliptical or spherical clusters. As a result, poorly separated clusters are often chained together. However, complete linkage favors the formation of small spherical clusters. Because of these space-distorting effects, hierarchical clustering methods should be used in tandem with PCA to detect clusters in multivariate data sets.

All hierarchical methods will always partition data, even randomly generated data, into distinct groups or clusters. Hence, it is important to ascertain the significance level of the similarity value selected by the user. For this task, a simple three-step procedure is proposed. First, a random data set is generated with the same correlation structure, the same number of samples, and the same number of measurements as the real data set that is currently being investigated. Second, the same clustering technique(s) is applied to the random data. Third, the similarity value, which generates the same number of clusters as identified in the real data set, is determined from the dendogram of the random data. If the similarity value is substantially larger for the real data set, the likelihood of having inadvertently exploited random variation in the data to achieve clustering is probably insignificant.

3.3 Case Studies

Hierarchical clustering methods attempt to uncover the intrinsic structure of a multivariate data set without making a priori assumptions about the data. This section,

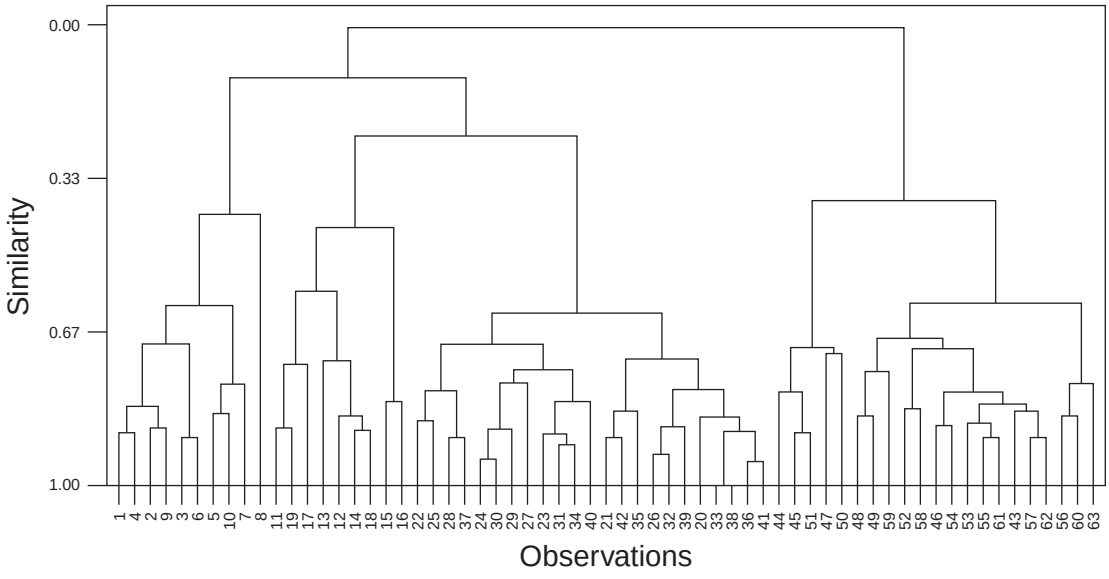

Figure 12 A single-linkage dendogram of the obsidian data set. The dendogram was generated using the program SCAN.

shows, by way of two published studies, how clustering methods can be used to find clusters of points in data.

3.3.1 Obsidian Data Set

The first study is the obsidian data set, discussed in section 2.3.2. A principal component map of the data (Figure 8) revealed four distinct clusters, which correspond to the sites from which these artifacts were obtained. To confirm the four-cluster hypothesis, the investigators also analyzed their data using single-linkage analysis. The resulting dendogram (Figure 12) indicated that it is reasonable to divide the glass samples into four categories based on the quarry sites from which these artifacts were obtained. (At similarity 0.40, samples 1–10 form a cluster, samples 11–19 form another cluster, samples 20–42 form a third cluster, and samples 43–63 form the fourth cluster.) Because the dendogram and principal component map of the data are in strong agreement, the investigators decided to partition their data into four groups based on the quarry labels of the samples.

3.3.2 Artificial Nose Data Set

The second study involves data from an artificial nose.[11] A salesman claims that an electronic nose can successfully sniff odor. The salesman obtained data from the literature to support his claim. Of the 41 compounds in the data set, compounds 1–21 are etheral, compounds 22–32 are pungent, and compounds 33–41 are minty. Each compound was characterized by six electronic measurements. Using a back-propagation neural network algorithm, the salesman was able to correctly classify all of the compounds in the data set. Should we then accept the salesman's claim that an electronic nose can sniff odor?

In order to validate the salesman's claim, the odor data was analyzed using single and complete linkage hierarchical clustering. Figures 13 and 14 show dendograms of the odor data. It is evident from the dendograms that dividing the 41 compounds into three categories based on odor type cannot be justified. (Although the data can be divided into three clusters by complete linkage at a 0.56 similarity, the cluster memberships cannot be correlated to compound odor – why? Cluster 1 consists of samples 41, 21, 19, 16, 13, 39, 31, 36, 29, 8, 34, and 32. Cluster 2 consists of samples 38, 5, 26, 14, 24, 15, 37, 6, 33, 35, 20, 30, and 7. Cluster 3 consists of samples 40, 17, 4, 22, 2, 25, 10, 28, 18, 27, 23, 12, 11, 1, 9, and 3.) Evidently, the six electronic measurements from the nose do not contain sufficient discriminatory information to force the compounds to cluster on the basis of odor type. Therefore, the salesman's claim about the efficacy of the proposed artificial nose should not be accepted at face value.

4 PATTERN RECOGNITION

So far, only exploratory data analysis techniques, i.e. cluster analysis and PCA, have been discussed. These techniques attempt to analyze data without directly using information about the class assignment of the samples. Although cluster analysis and PCA are powerful methods for uncovering relationships in large multivariate data sets, they are not sufficient for developing a classification rule that can accurately predict the class-membership of

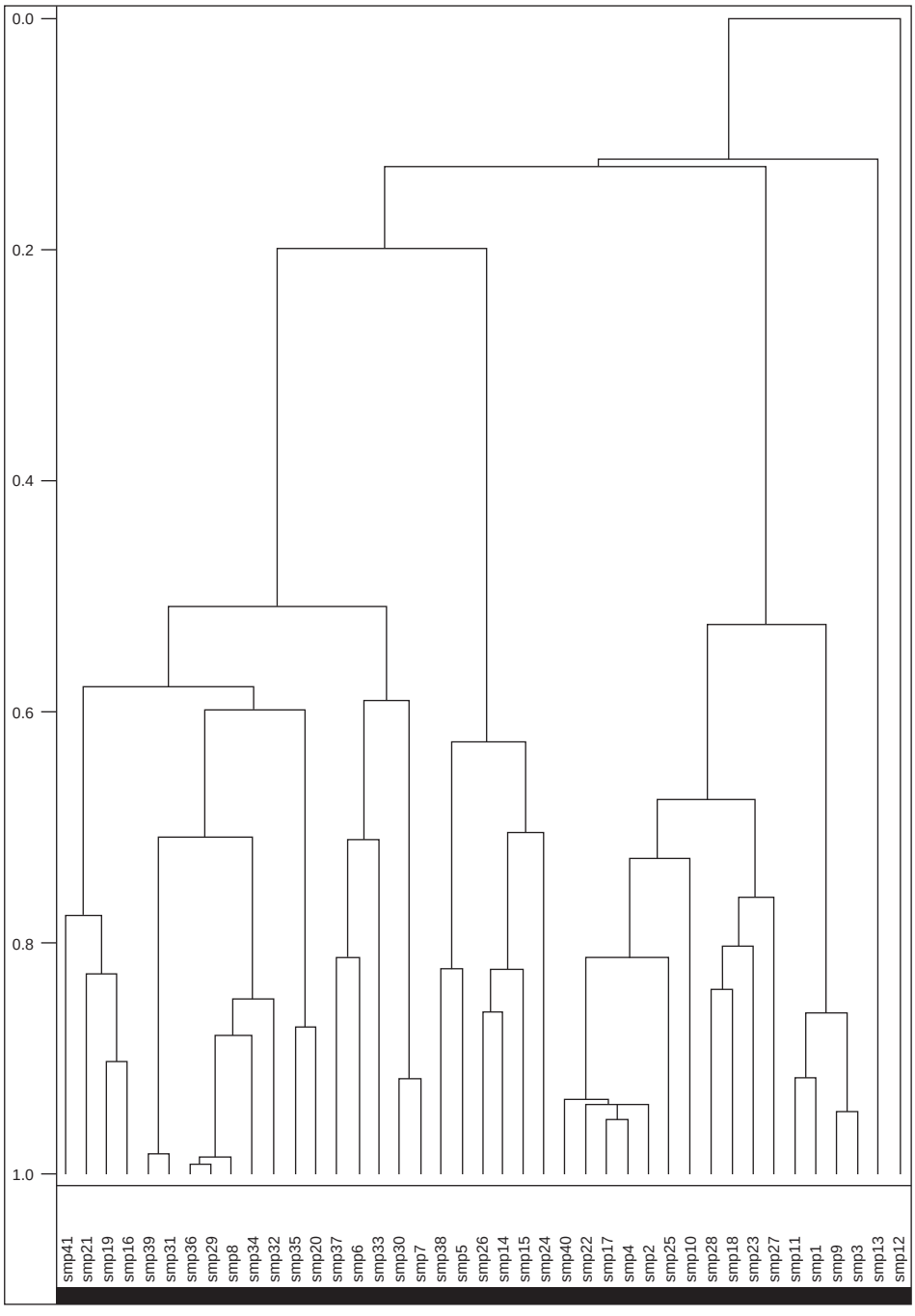

Figure 13 A single-linkage dendogram of the nose data set. The dendogram was generated using the program Pirouette.

an unknown sample. In this section, pattern recognition techniques will be discussed. These techniques were originally developed to categorize a sample on the basis of regularities in observed data. The first applications of pattern recognition to chemistry were studies involving low-resolution mass spectrometry.[12] Since then, pattern recognition techniques have been applied to a wide variety of chemical problems, such as chromatographic fingerprinting,[13–15] spectroscopic imaging,[16–18] and data interpretation.[19–21]

Pattern recognition techniques fall into one of two categories: nonparametric discriminants, and similarity-based classifiers. Nonparametric discriminants,[22–24] such as neural networks, attempt to divide a data space into different regions. In the simplest case, that of a binary classifier, the data space is divided into two regions.

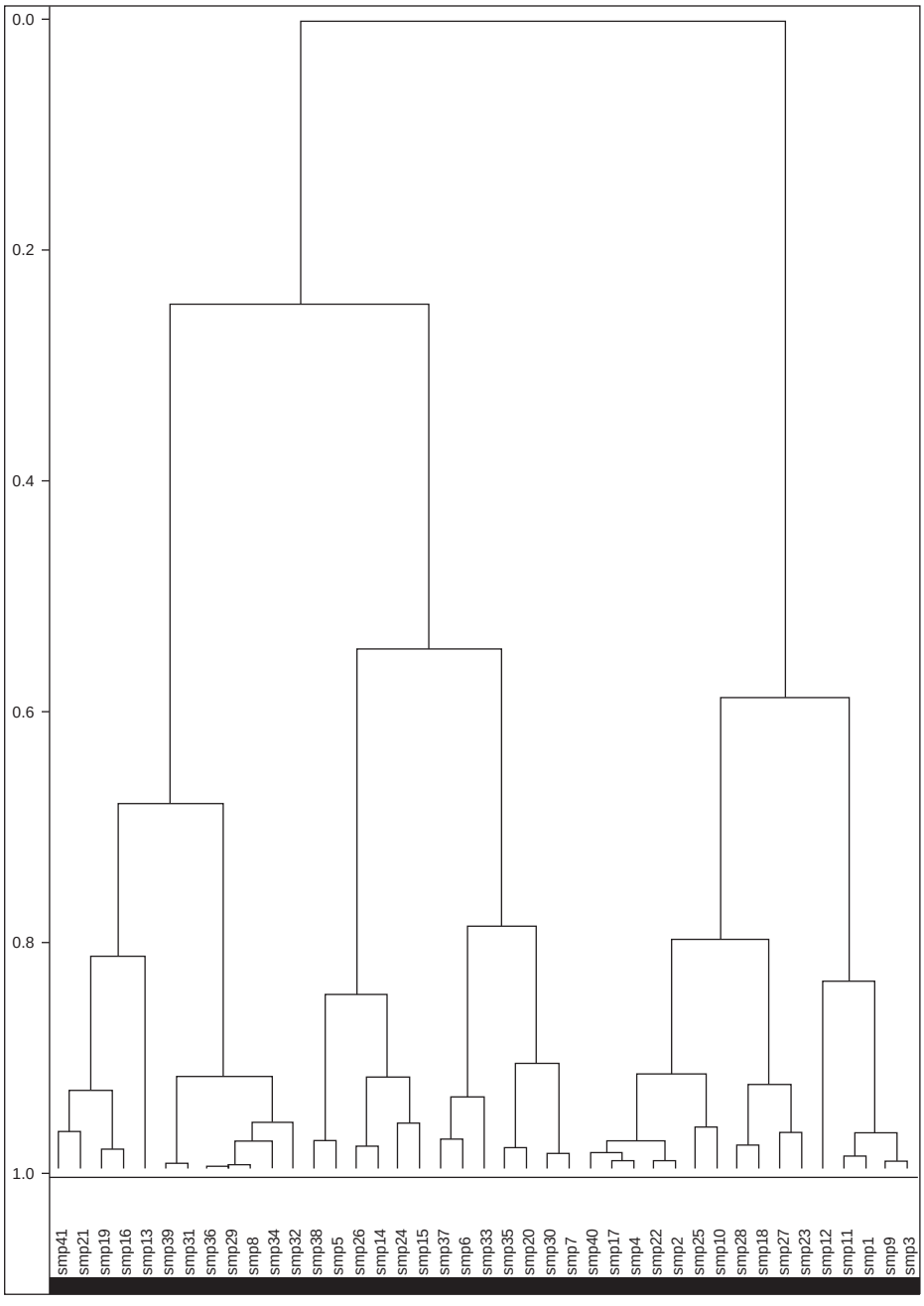

Figure 14 A complete-linkage dendogram of the nose data set. The dendogram was generated using the program Pirouette.

Samples that share a common property (such as fuel type) will be found on one side of the decision surface, whereas those samples comprising the other category will be found on the other side. Nonparametric discriminants have provided insight into relationships contained within sets of chemical measurements. However, classification based on random or chance separation[24] can be a serious problem if the data set is not sample rich. Because chemical data sets usually contain more variables than samples, similarity-based classifiers are generally preferred.

Similarity-based classifiers, e.g. *k*-nearest neighbor (KNN)[25] and soft independent modeling by class analogy (SIMCA),[26–30] treat each chromatogram or spectrum as a data vector $\mathbf{x} = (x_1, x_2, x_3, \ldots, x_j, \ldots, x_p)$ where component x_j is the area of the jth peak or the absorbance value of the jth wavelength. Such a vector can also be viewed as a point in a high-dimensional measurement space. A basic assumption is that distances between points in the measurement space will be inversely related to their degree of similarity.

Using a similarity-based classifier we can determine the class-membership of a sample by examining the class label of the data point closest to it or from the principal component model of the class, which lies closest to the sample in the measurement space. In chemistry, similarity-based classification rules are implemented using either KNN or SIMCA.

4.1 *k*-Nearest Neighbor

For its simplicity, KNN is a powerful classification technique. A sample is classified according to the majority vote of its KNNs, where k is an odd integer (one, three, or five). For a given sample, Euclidean distances are first computed from the sample to every other point in the data set. These distances arranged from smallest to the largest are used to define the sample's KNNs. A poll is then taken by examining the class identities among the point's KNNs. Based on the class identity of the majority of its KNNs, the sample is assigned to a class in the data set. If the assigned class and the actual class of the sample match, the test is considered a success. The overall classification success rate, calculated over the entire set of points, is a measure of the degree of clustering in the set of data. Clearly, a majority vote of the KNNs can only occur if the majority of the measurement variables concur, because the data is usually autoscaled.

KNN cannot furnish a statement about the reliability of a classification. However, its classification risk is bounded. In other words, the Bayes classifier will generate the optimal classification rule for the data, and 1-nearest neighbor has an error rate which is twice as large as the Bayes classifier. (To implement the Bayes classifier, one must have knowledge about all the statistics of the data set including the underlying probability distribution function for each class. Usually, this knowledge is not available.) Hence, any other classification method, no matter how sophisticated, can at best only improve on the performance of KNN by a factor of two.

4.2 Soft Independent Modeling by Class Analogy

In recent years, modeling approaches have become popular in analytical chemistry for developing classification rules because of the problems with nonparametric discriminants. Although there are a number of approaches to modeling classes, the SIMCA method, based on PCA, has been developed by Wold for isolating groups of multivariate data or classes in a data set. In SIMCA, a PCA is performed on each class in the data set, and a sufficient number of principal components are retained to account for most of the variation within each class. Hence, a principal component model is used to represent each class in the data set. The number of principal components retained for each class is usually different. Deciding on the number of principal components that should be retained for each class is important, as retention of too few components can distort the signal or information content contained in the model about the class, whereas retention of too many principal components diminishes the signal-to-noise. A procedure called cross-validation[31] ensures that the model size can be determined directly from the data. To perform cross-validation, segments of the data are omitted during the PCA. Using one, two, three, etc., principal components, omitted data are predicted and compared to the actual values. This procedure is repeated until every data element has been kept out once. The principal component model that yields the minimum prediction error for the omitted data is retained. Hence, cross-validation can be used to find the number of principal components necessary to describe the signal in the data while ensuring high signal-to-noise by not including the so-called secondary or noise-laden principal components in the class model.

The variance that is explained by the class model is called the modeled variance, which describes the signal, whereas the noise in the data is described by the residual variance or the variance not accounted for by the model. (The residual variance is explained by the secondary principal components, which have been truncated or omitted from the principal component model.) By comparing the residual variance of an unknown to the average residual variance of those samples that make up the class, it is possible to obtain a direct measure of the similarity of the unknown to the class. This comparison is also a measure of the goodness of fit of the sample to a particular principal component model. Often, the F-statistic is used to compare the residual variance of a sample with the mean residual variance of the class.[32] Employing the F-statistic, an upper limit for the residual variance can be calculated for those samples belonging to the class. The final result is a set of probabilities of class-membership for each sample.

An attractive feature of SIMCA is that a principal component mapping of the data has occurred. Hence, samples that may be described by spectra or chromatograms are mapped onto a much lower dimensional subspace for classification. If a sample is similar to the other samples in the class, it will lie near them in the principal component map defined by the samples representing that class. Another advantage of SIMCA is that an unknown is only assigned to the class for which it has a high probability. If the residual variance of a sample exceeds the upper limit for every modeled class in the data set, the sample would not be assigned to any of the classes because it is either an outlier or comes from a class that is not represented in the data set. Finally, SIMCA is sensitive to the quality of the

data used to generate the principal component models. As a result, there are diagnostics to assess the quality of the data, such as the modeling power[33] and the discriminatory power.[34] The modeling power describes how well a variable helps the principal components to model variation, and discriminatory power describes how well the variable helps the principal components to classify the samples in the data set. Variables with low modeling power and low discriminatory power are usually deleted from the data because they contribute only noise to the principal component models.

SIMCA can work with as few as 10 samples per class, and there is no restriction on the number of measurement variables, which is an important consideration, because the number of measurement variables often exceeds the number of samples in chemical studies. Most standard discrimination techniques would break down in these situations because of problems arising from collinearity and chance classification.[24]

4.3 Feature Selection

Feature selection is a crucial step in KNN or SIMCA, because it is important to delete features or measurements that contain information about experimental artifacts or other systematic variations in the data not related to legitimate chemical differences between classes in a data set. For profiling experiments of the type that are being considered (see section 4.4) it is inevitable that relationships may exist among sets of conditions used to generate the data and the patterns that result. One must realize this in advance when approaching the task of analyzing such data. Therefore, the problem is utilizing information contained in the data characteristic of the class without being swamped by the large amount of qualitative and quantitative information contained in the chromatograms or spectra about the experimental conditions used to generate the data. If the basis of classification for samples in the training set is other than desired group differences, unfavorable classification results for the prediction set will be obtained despite a linearly separable training set. The existence of these confounding relationships is an inherent part of profiling data. Hence, the goal of feature selection is to increase the signal-to-noise ratio of the data by discarding measurements on chemical components that are not characteristic of the source profile of the classes in the data set. Feature selection in the context of pattern recognition is described in greater detail in the next section by way of the two worked examples.

4.4 Case Studies

Pattern recognition is about reasoning, using the available information about the problem to uncover information contained within the data. Autoscaling, feature selection, and classification are an integral part of this reasoning process. Each plays a role in uncovering information contained within the data.

Pattern recognition analyses are usually implemented in four distinct steps: data preprocessing, feature selection, classification, and mapping and display. However, the process is iterative, with the results of a classification or display often determining a further preprocessing step and reanalysis of the data. Although the procedures selected for a given problem are highly dependent upon the nature of the problem, it is still possible to develop a general set of guidelines for applying pattern recognition techniques to real data sets. In this section, a framework for solving the class-membership problem is presented by way of two recently published studies on chromatographic fingerprinting of complex biological and environmental samples.

4.4.1 Fuel Spill Identification

The first study[35] involves the application of gas chromatographic and pattern recognition (GC/PR) methods to the problem of typing jet fuels, so that a spill sample in the environment can be traced to its source. The test data consisted of 228 gas chromatograms of neat jet fuel samples representing the major aviation fuels (JP-4, Jet-A, JP-7, JPTS, and JP-5) found in the USA. The neat jet fuel samples used in this study were obtained from Wright Patterson Air Force Base or Mulkiteo Energy Management Laboratory (Table 6). They were splits from regular quality control standards, which were purchased by the United States Air Force (USAF) to verify the authenticity of the manufacturer's claims.

The prediction set consisted of 25 gas chromatograms of weathered jet fuels (Table 7). Eleven of the 25 weathered fuels were collected from sampling wells as a neat oily phase found floating on top of the well water. Eleven of the 25 weathered fuel samples were extracted from the soil near various fuel spills. The other three fuel samples had been subjected to weathering in a laboratory.

The neat jet fuel samples were stored in sealed containers at $-20\,^{\circ}C$. Prior to chromatographic analysis, each fuel sample was diluted with methylene chloride

Table 6 Training set

Number of samples	Fuel type
54	JP-4 (fuel used by USAF fighters)
70	Jet-A (fuel used by civilian airliners)
32	JP-7 (fuel used by SR-71 reconnaissance plane)
29	JPTS (fuel used by TR-1 and U-2 aircraft)
43	JP-5 (fuel used by Navy jets)

Table 7 Prediction set

Sample	Identity	Source	Sample	Identity	Source
PF007	JP-4	A[a]	MIX1	JP-4	C[c]
PF008	JP-4	A[a]	MIX2	JP-4	C[c]
PF009	JP-4	A[a]	MIX3	JP-4	C[c]
PF010	JP-4	A[a]	MIX4	JP-4	C[c]
PF011	JP-4	A[a]	STALE-1	JP-4	D[d]
PF012	JP-4	A[a]	STALE-2	JP-4	D[d]
PF013	JP-4	A[a]	STALE-3	JP-4	D[d]
KSE1M2	JP-4	B[b]	PIT1UNK	JP-5	E[e]
KSE2M2	JP-4	B[b]	PIT1UNK	JP-5	E[e]
KSE3M2	JP-4	B[b]	PIT2UNK	JP-5	E[e]
KSE4M2	JP-4	B[b]	PIT2UNK	JP-5	E[e]
KSE5M2	JP-4	B[b]			
KSE6M2	JP-4	B[b]			
KSE7M2	JP-4	B[b]			

[a] Sampling well at Tyndall AFB: the sampling well was near a previously functioning storage depot. Each well sample was collected on a different day.
[b] Soil extract near a sampling well: dug with a hand auger at various depths. Distance between sampling well Tyndall and soil extract was approximately 80 yards.
[c] Weathered fuel added to sand.
[d] Old JP-4 fuel samples that had undergone weathering in a laboratory refrigerator.
[e] Sampling pit at Keywest Air Station: two pits were dug near a seawall to investigate a suspected JP-5 fuel leak.

and injected onto a fused silica capillary column (10 m × 0.10 mm) using a split injection technique. The fused silica capillary column was temperature programmed from 60 °C to 270 °C at 18 °C min^{-1}. High-speed gas chromatograms representative of the five fuel types (JP-4, Jet-A, JP-7, JPTS, and JP-5) are shown in Figure 15.

The gas chromatograms were peak matched using a computer program[36] that correctly assigned peaks by first computing the Kovats retention index (KI)[37] for compounds eluting off the GC column. Because the *n*-alkane peaks are the most prominent features present in the gas chromatograms of these fuels, it is a simple matter to compute KI values. The peak-matching program then developed a template of peaks by examining integration reports and adding peaks to the template that did not match the retention indices of previously observed peaks. By matching the retention indices of the peaks in each chromatogram with the retention indices of the features in the template, it was possible to produce a data vector for each gas chromatogram. Each feature in a peak-matched chromatogram was assigned a value corresponding to the normalized area of the peak in the chromatogram. (If the peak was present, its normalized area from the integration report was assigned to the corresponding element of the vector. If the peak was not present, the corresponding feature was assigned a value of zero.) The number of times a particular peak was found to have a nonzero value was also computed, and features below a user-specified number of nonzero occurrences (which was set equal to 5% of the total number of fuel samples in the training set) were deleted from the data set. This peak-matching procedure yielded a final cumulative reference file containing 85 features, though not all peaks were present in all chromatograms.

Because outliers have the potential to adversely influence the performance of pattern recognition methods, outlier analysis was performed on each fuel class in the training set prior to pattern recognition analysis. The generalized distance test[38] at the 0.01 significance level was implemented via SCOUT[39] to identify discordant observations in the data. Three Jet-A and four JP-7 fuel samples were found to be outliers and were subsequently removed from the data set. The training set, comprising 221 gas chromatograms of 85 peaks each, was analyzed using pattern recognition methods. Prior to pattern recognition analysis, the data were autoscaled so that each feature had a mean of zero and standard deviation of one within the set of 221 gas chromatograms. Hence, each gas chromatogram was initially represented as an 85-dimensional data vector, $\mathbf{x} = (x_1, x_2, x_3, \ldots, x_j, \ldots, x_{85})$, where x_j is the normalized area of the jth peak.

The first step in the study was to apply PCA to the training set data. Figure 16 shows a plot of the two largest principal components of the 85 GC peaks obtained from the 221 neat jet fuel samples. Each fuel sample or gas chromatogram is represented as a point in the principal component map of the data. The JP-7 and JPTS fuel samples are well separated from one another and from the gas chromatograms of the JP-4, Jet-A, and JP-5 fuel samples, suggesting that information about fuel type is

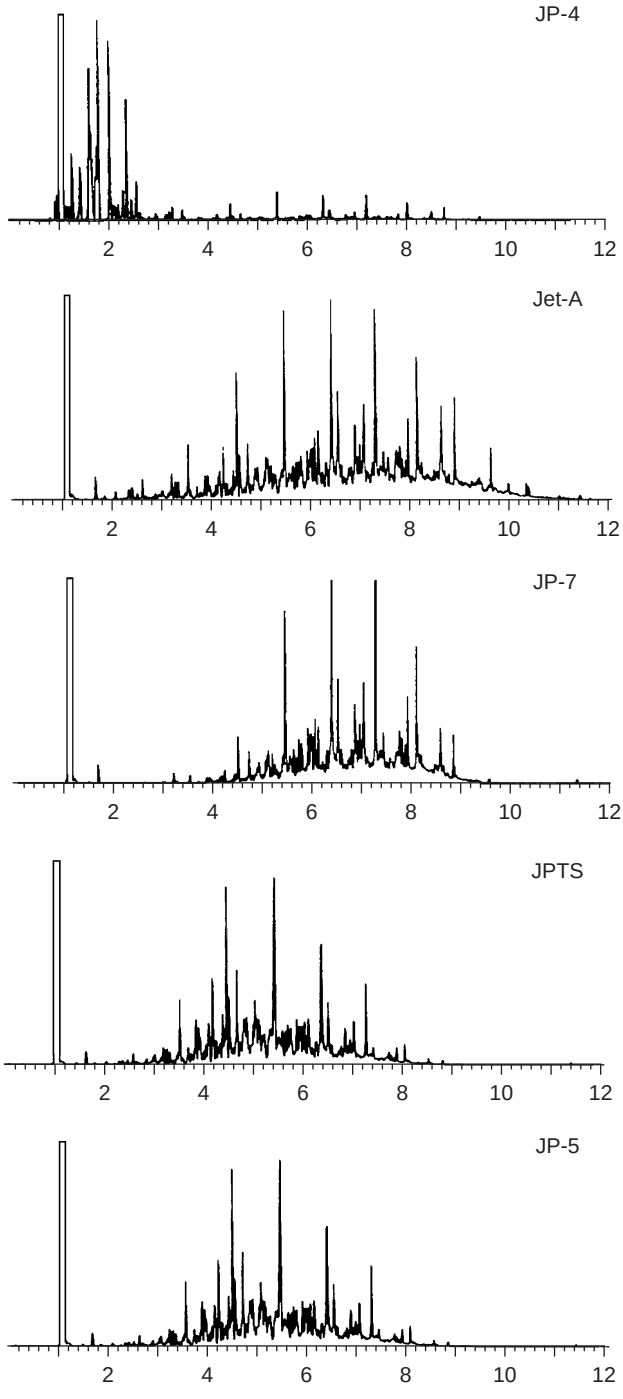

Figure 15 High speed gas chromatograms of JP-4, Jet-A, JP-7, JPTS, and JP-5. (Reproduced with permission from Lavine et al.[35] Copyright 1995 American Chemical Society.)

present in the high speed gas chromatograms of the neat jet fuels. However, the overlap of the JP-5 and Jet-A fuels in the principal component map suggests that gas chromatograms of these two fuels share a common set of attributes, which is not surprising in view of their

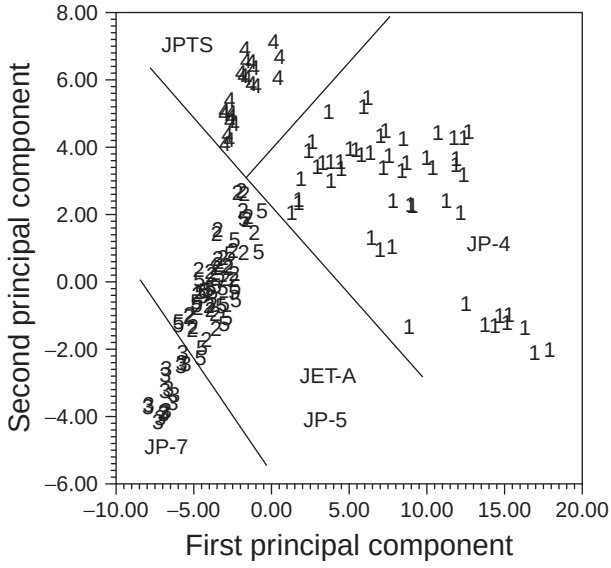

Figure 16 A plot of the two largest principal components of the 85 GC peaks for the 221 neat jet fuel samples. The map explains 72.3% of the total cumulative variance: 1 = JP-4, 2 = Jet-A, 3 = JP-7, 4 = JPTS, 5 = JP-5. (Reproduced with permission from Lavine et al.[35] Copyright 1995 American Chemical Society.)

similar physical and chemical properties.[40] Mayfield and Henley[41] have also reported that gas chromatograms of Jet-A and JP-5 fuels were more difficult to classify than gas chromatograms of other types of jet fuels. Nevertheless, they concluded that fingerprint patterns exist within GC profiles of Jet-A and JP-5 fuels characteristic of fuel type, which is consistent with a plot of the second and third largest principal components of the training set data. The plot in Figure 17 indicates that differences do indeed exist between the GC profiles of Jet-A and JP-5 fuels. However, the second and third largest principal components do not represent the direction of maximum variance in the data. (In fact, they only represent 23.1% of the total cumulative variance or information content of the data.) Hence, it must be concluded that the bulk of the information contained within the 85 GC peaks is not about differences between the GC profiles of Jet-A and JP-5 fuels.

To better understand the problems associated with classifying the gas chromatograms of Jet-A and JP-5 fuels, it was necessary to reexamine this classification problem in greater detail. Figure 18 shows a plot of the two largest principal components of the 85 GC peaks of the 110 Jet-A and JP-5 neat fuel samples. It is evident from an examination of the principal component map that Jet-A and JP-5 fuel samples lie in different regions, suggesting that Jet-A and JP-5 fuels can be differentiated from each other on the basis of their GC profiles. However, the points representing the JP-5 fuels form

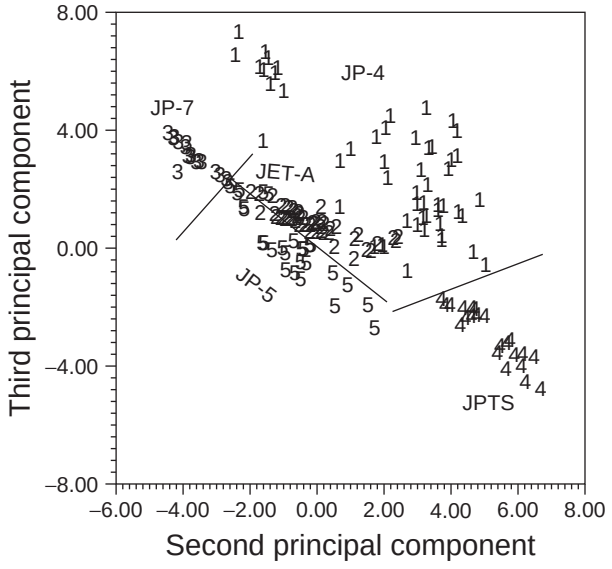

Figure 17 A plot of the second and third largest principal components of the 85 GC peaks for the 221 neat jet fuel samples. The map explains 23.1% of the total cumulative variance: 1 = JP-4, 2 = Jet-A, 3 = JP-7, 4 = JPTS, 5 = JP-5. (Reproduced with permission from Lavine et al.[35] Copyright 1995 American Chemical Society.)

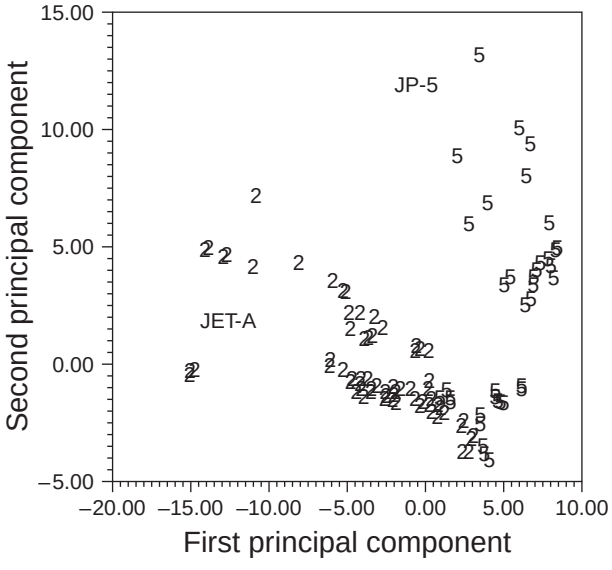

Figure 18 A principal component map of the 110 neat Jet-A and JP-5 fuel samples developed from the 85 GC peaks. The JP-5 fuel samples can be divided into two distinct groups: fuel samples that lie close to the Jet-A fuels and fuel samples distant from the Jet-A fuels. (Reproduced with permission from Lavine et al.[35] Copyright 1995 American Chemical Society.)

two distinct subgroups in the map, which could pose a problem as an important requirement for a successful pattern recognition study is that each class is represented by a collection of samples that are in some way similar. The subclustering suggests a lack of homogeneity among the samples representing the JP-5 fuels. Therefore, it is important to identify and delete the GC peaks responsible for the subclustering of the JP-5 fuels.

The following procedure was used to identify the GC peaks strongly correlated with the subclustering. First, the JP-5 fuel samples were divided into two categories on the basis of the observed subclustering. Next, the ability of each GC peak alone to discriminate between the gas chromatograms from the two JP-5 subclusters was assessed. The dichotomization power of each of the 85 GC peaks was also computed for the following category pairs: JP-5 versus JP-4, JP-5 versus Jet-A, JP-5 versus JP-7, and JP-5 versus JPTS. A GC peak was retained for further analysis only if its dichotomization power for the subclustering dichotomy was lower than for any of the other category pairs. Twenty-seven GC peaks that produced the best classification results when the chromatograms were classified as Jet-A, JPTS, JP-7, or JP-5 were retained for further study.

Figure 19 shows a plot of the two largest principal components of the 27 GC peaks obtained from the 221 neat jet fuel samples. It is evident from the principal component map of the 27 features that the five fuel classes are well separated. Furthermore, the principal component map of the data does not reveal subclustering within any class. This indicates that each fuel class is represented by a collection of samples that are in some way similar when the 27 GC peaks are used as features.

A five-way classification study involving JP-4, Jet-A, JP-7, JPTS, and JP-5 fuels was also undertaken using

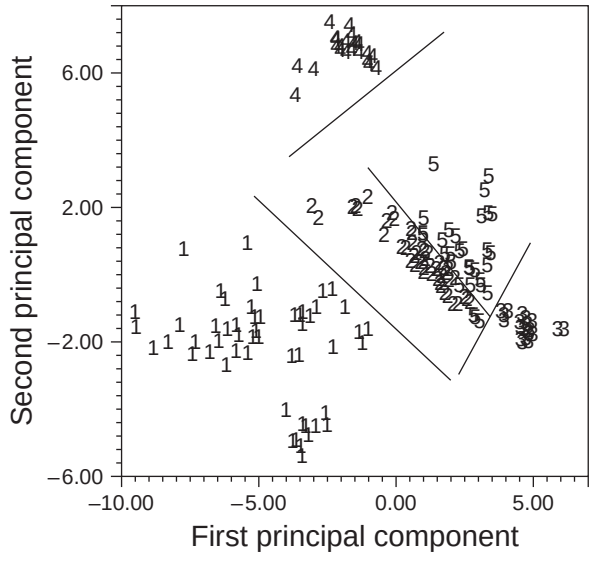

Figure 19 A principal component map of the 27 GC peaks: 1 = JP-4, 2 = Jet-A, 3 = JP-7, 4 = JPTS, 5 = JP-5. (Reproduced with permission from Lavine et al.[35] Copyright 1995 American Chemical Society.)

SIMCA. A principal component model for each fuel class in the training set was developed from the 27 GC peaks. The complexity of the principal component model was determined directly from the data using the technique of cross-validation. For each class, a single principal component was used to model the data. The gas chromatograms in the training set were then fitted to these models, and the residual – that is the sum of the squares difference between the original gas chromatogram and the chromatogram reproduced by the model – was computed for each gas chromatogram.

Each gas chromatogram in the training set was then classified on the basis of its goodness of fit. The probability of a gas chromatogram belonging to a particular fuel class was determined from its residual for the corresponding principal component model by way of an F-test. A gas chromatogram was assigned to the fuel class for which it had the lowest variance ratio. However, if the variance ratio exceeded the critical F-value for that class, then the sample would not be assigned to it. Results from the five-way classification study involving the training set samples are summarized in Table 8. The recognition rate for JP-4, Jet-A, JP-7, and JPTS fuels is very high. However, Jet-A is more difficult to recognize because of its similarity to JP-5, which is undoubtedly the reason for SIMCA classifying 16 Jet-A fuel samples as JP-5.

The ability of the principal component models to predict the class of an unknown fuel was first tested using a method called internal validation. The training set of 221 gas chromatograms was subdivided into 13 training set – prediction set pairs. Each training set had 204 gas chromatograms and each prediction set had 17 gas chromatograms. The members of the sets were chosen randomly. Furthermore, a particular chromatogram was present in only 1 of the 13 prediction sets generated. Principal component models were developed for each of the training sets and tested on the corresponding prediction set. The mean classification success rate for these so-called prediction sets was 90.5%.

To further test the predictive ability of the 27 GC peaks and the classification models associated with them, an external prediction set of 25 gas chromatograms was employed. The gas chromatograms in the prediction set were run a few months before the neat jet fuel gas chromatograms were run. The results of this study are shown in Table 9. All the weathered fuel samples were correctly classified. This is an important result, as the changes in composition that occur after a jet fuel is released into the environment constitute a major problem in fuel spill identification. These changes may arise from evaporation of lower-molecular-weight alkanes, microbial degradation, and the loss of water-soluble compounds due to dissolution.[42] Because the weathered

Table 8 SIMCA training set results

Class	F-criterion[a]			
	Principal components	Number in class	Correct	Percentage
JP-4	1	54	54	100
Jet-A[b]	1	67	51	76.1
JP-7[b]	1	28	27	96.4
JPTS	1	29	29	100
JP-5	1	43	43	100
Total		221	204	92.3

[a] Classifications were made on the basis of the variance ratio $F = [s_p/S_o]^2[N_q - NC - 1]$, where s_p^2 is the residual of sample p for class i, S_o^2 is the variance of class i, N_q is the number of samples in the class, and NC is the number of principal components used to model the class. A sample is assigned to the class for which it has the lowest variance ratio. However, if the sample's variance ratio exceeds the critical F-value for the class, then the sample cannot be assigned to the class. The critical F-value for each training set sample is $F_\alpha = 0.975$ $[(M - NC),(M - NC)(N_q - NC - 1)]$ where M is the number of measurement variables or GC peaks used to develop the principal component model.

[b] Misclassified Jet-A and JP-5 fuel samples were categorized as JP-5.

Table 9 Prediction set results

Samples	F-values[a]				
	JP-4	JET-A	JP-7	JPTS	JP-5
PF007	3.82	10.04	58.9	12.4	7.43
PF008	3.69	9.62	57.6	12.5	7.14
PF009	3.71	9.84	57.6	12.6	7.32
PF010	3.30	16.7	73.7	11.8	10.8
PF011	3.57	9.64	58.9	12.8	7.39
PF012	4.11	7.74	78.2	13.5	12.04
PF013	4.33	8.19	79.8	12.6	12.3
KSE1M2	2.83	24.4	63.9	30.4	11.21
KSE2M2	2.25	16.2	70.8	21.6	11.09
KSE3M2	2.51	9.41	71.0	17.3	10.2
KSE4M2	2.40	10.11	71.3	17.83	10.4
KSE5M2	2.33	7.76	56.4	17.9	7.61
KSE6M2	1.87	13.4	69.3	20.8	10.4
KSE7M2	2.21	9.85	67.3	18.3	9.78
MIX1	1.33	34.9	71.3	38.2	13.3
MIX2	1.33	11.93	53.3	20.9	7.37
MIX3	1.44	12.3	55.2	20.6	7.71
MIX4	1.59	9.51	48.6	19.9	6.27
STALE-1	1.72	73.7	151.9	54.7	31.5
STALE-2	0.58	28.7	123.8	30.9	22.6
STALE-3	0.541	28.7	127.3	29.9	22.6
PIT1UNK	6.62	1.19	6.106	33.02	0.504
PIT1UNK	6.57	1.15	6.03	32.9	0.496
PIT2UNK	6.51	1.14	6.14	32.8	0.479
PIT2UNK	6.51	1.14	6.27	32.7	0.471

[a] An object is assigned to the class for which it has the lowest variance ratio. However, if the variance ratio exceeds the critical F-value for that class, then the object cannot be assigned to it. Critical F-values of prediction set samples are obtained using one degree of freedom for the numerator and $N_q - NC - 1$ degrees of freedom for the denominator.[32] For JP-4, the critical F-value at $\alpha = 0.975$ is $F(1, 52) = 5.35$, and it is $F(1, 41) = 5.47$ for JP-5.

fuel samples used in this study were recovered from a subsurface environment, loss of lower alkanes due to evaporation will be severely retarded. Furthermore, dissolution of water-soluble components should not pose a serious problem as only a small fraction of the fuel's components is soluble in water.[43] Hence, the predominant weathering factor in subsurface fuel spills is probably biodegradation, which does not appear to have a pronounced effect on the overall GC profile of the fuels. Clearly, weathering of aviation turbine fuels in a subsurface environment will be greatly retarded compared to surface spills, thereby preserving the fuel's identity for a longer period of time.

4.4.2 Africanized Honeybees

GC/PR has also been used to develop a potential method for differentiating Africanized honeybees from European honeybees.[44–47] The test data consisted of 109 gas chromatograms (49 Africanized and 60 European) of cuticular hydrocarbons obtained from bodies of Africanized and European honeybees. Cuticular hydrocarbons were obtained by rinsing the dry or pinned bee specimens in hexane for approximately 15 min. The cuticular hydrocarbon fraction analyzed by gas chromatography was isolated from the concentrated washings by means of a silicic acid column. Hexane was used as the eluent. The extracted hydrocarbons (equivalent to 4% of a bee) were coinjected with authentic *n*-alkane standards. KIs were assigned to compounds eluting off the column. These indices were used for peak identification.

Each gas chromatogram contained 40 peaks corresponding to a set of standardized retention time windows. A typical GC trace of the cuticular hydrocarbons from an Africanized honeybee sample is shown in Figure 20. The GC column had about 5000 plates. The hydrocarbon extract was analyzed on a glass column (1.8 m × 2 mm) packed with 3% OV-17 on Chromosorb® WAW DMCS packing (120–140 mesh).

The gas chromatograms were translated into data vectors by measuring the area of the 40 GC peaks. However, only 10 of the GC peaks were considered for pattern recognition analysis. Compounds comprising these peaks were found in the wax produced by nest bees, and the concentration pattern of the wax constituents is believed to convey genetic information about the honeybee colony. Because the feature selection process was carried out on the basis of a priori considerations, the probability of inadvertently exploiting random variation in the data was minimized.

Each gas chromatogram was normalized to constant sum using the total integrated area of the 40 GC peaks. Also, the training set data were autoscaled to ensure

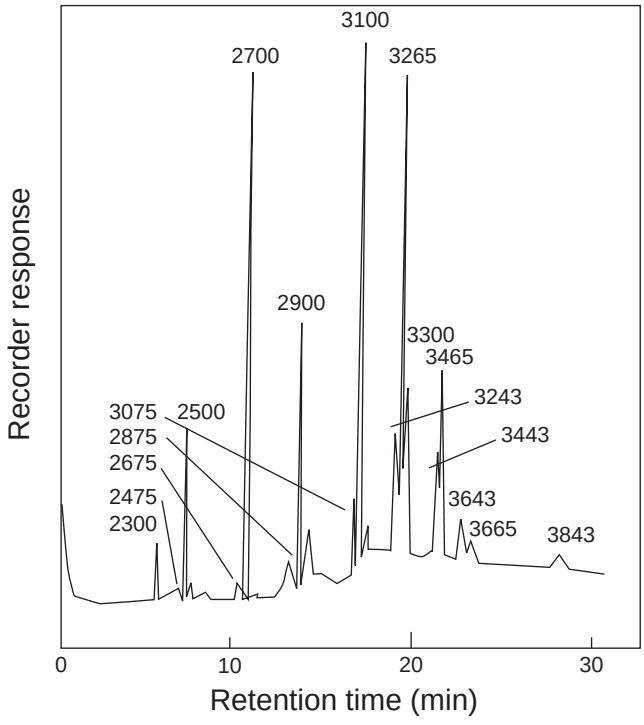

Figure 20 A gas chromatogram of the cuticular hydrocarbons from an Africanized honey bee. The KI values for the peaks used in the pattern recognition study were 2300, 2675, 2875, 3075, 3100, 3243, 3265, 3300, 3443, and 3465. (Reproduced with permission from Lavine and Carlson.[15] Copyright 1987 American Chemical Society.)

that each feature had equal weight in the analysis. The normalized and autoscaled data were then analyzed using KNN, which classifies the data vectors in the training set according to a majority vote of its KNNs. Hence, a sample will be classified as an Africanized or European bee only if the majority of its KNNs in the measurement space are Africanized bees. When the 1-nearest neighbor classification rule was applied to the 10 GC peaks, it could correctly classify every chromatogram in the training set. This result indicates that Africanized and European bee specimens are well separated from each other in the feature space defined by the 10 GC peaks.

To test the predictive ability of these descriptors and the classifier associated with them, a prediction set of 55 gas chromatograms (15 Africanized and 40 European) was employed. The distances between the prediction set samples and the samples in the training set were calculated, with class assignments computed in the same manner as in the training phase. Using the 1-nearest neighbor classification rule, a classification success rate of 100% was achieved for the gas chromatograms in the prediction set. This result is important because it demonstrates that information

derived solely from cuticular hydrocarbons can categorize bees as to subspecies. This suggests a direct relationship between the concentration pattern of these compounds and the identity of the subspecies (Africanized or European). Clearly, these results imply that GC/PR can be used to identify the presence of the African genotype in honeybees.

5 SOFTWARE

There are a number of Windows 95/98 software packages sold by commercial vendors that can be used for clustering and classification. UNSCRAMBLER (Camo A/S, Olav Tryggvasonsgt. 24, N-7011 Trondheim, Norway) offers data preprocessing, PCA, SIMCA classification, and graphics in a flexible package. Pirouette (Infometrix Inc., P.O. Box 1528, 17270 Woodinville-Redmond Road NE, Suite 777, Woodinville, WA 98072-1528) has a nice user interface, with good quality graphics. The package has data preprocessing, hierarchical clustering, PCA, KNN, and SIMCA classification. Pirouette, which has been validated according to the United States Food and Drug Administration Standard Operating Procedure, is a good introductory package because of its broad functionality.

SCAN (Minitab Inc., 3081 Enterprise Drive, State College, PA 16801-3008) has PCA, hierarchical clustering, KNN, SIMCA, and discriminant analysis (quadratic, linear, regularized, and DASCO, which is an advanced classification method in chemometrics). The user interface, which is similar to the popular Minitab statistics package, has many advanced editing features, such as brushing. The package is a good mix of statistical and pattern recognition methods. SIRIUS (Pattern Recognition Associates, P.O. Box 9280, The Woodlands, TX 77387-9280) is a graphics oriented package intended for modeling and exploratory data analysis, such as SIMCA and PCA. The PLS TOOLBOX (Eigenvector Technologies, P.O. Box 483, 196 Hyacinth, Manson, WA 89931) is for Matlab and contains routines for PCA, discriminant analysis, and cluster analysis.

6 CONCLUSION

In this article, a basic methodology for analyzing large multivariate chemical data sets is described. A chromatogram or spectrum is represented as a point in a high-dimensional measurement space. Exploratory data analysis techniques (PCA and hierarchical clustering) are then used to investigate the properties of this measurement space. These methods can provide information about trends present in the data. Classification methods can then be used to further quantify these relationships. The techniques, which have been found to be most useful, are nonparametric in nature. As such, they do not attempt to fit the data to an exact functional form; rather, they use the data to suggest an appropriate mathematical model, which can identify structure within the data. Hence, the approach described in this article relies heavily on graphics for the presentation of results, because clustering and classification methods should be used to extend the ability of human pattern recognition to uncover structure in multivariate data. Although the computer can assimilate more data at any given time than can the chemist, it is the chemist, in the end, who must make the necessary decisions and judgements about their data.

ABBREVIATIONS AND ACRONYMS

FCV	Fuzzy Clustering Varieties
GC	Gas Chromatographic
GC/PR	Gas Chromatographic and Pattern Recognition
HPLC	High-performance Liquid Chromatography
KI	Kovats Retention Index
KNN	*k*-Nearest Neighbor
PCA	Principal Component Analysis
SIMCA	Soft Independent Modeling by Class Analogy
USAF	United States Air Force

RELATED ARTICLES

Environment: Water and Waste **(Volume 4)**
Solid-phase Microextraction in Environmental Analysis • Underground Fuel Spills, Source Identification

Field-portable Instrumentation **(Volume 5)**
Solid-phase Microextraction in Analysis of Pollutants in the Field

Chemometrics **(Volume 11)**
Chemometrics

Gas Chromatography **(Volume 12)**
Data Reduction in Gas Chromatography

Infrared Spectroscopy **(Volume 12)**
Spectral Data, Modern Classification Methods for

REFERENCES

1. B.R. Kowalski, 'Analytical Chemistry as an Information Science', *TRACS*, **1**, 71–74 (1988).
2. I.T. Jolliffee, *Principal Component Analysis*, Springer-Verlag, New York, 1986.
3. D.L. Massart, L. Kaufman, *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*, John Wiley & Sons, New York, 1983.
4. R.G. Brereton (ed.), *Multivariate Pattern Recognition in Chemometrics*, Elsevier, Amsterdam, 1992.
5. S.D. Brown, 'Chemical Systems Under Indirect Observation: Latent Properties and Chemometrics', *Appl. Spectrosc.*, **49**(12), 14A–31A (1995).
6. L. Kaufman, P.J. Rousseeuw, *Finding Groups in Data*, John Wiley & Sons, New York, 1990.
7. G.L. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, John Wiley & Sons, New York, 1992.
8. J. Mandel, 'The Regression Analysis of Collinear Data', *J. Res. Natl. Bur. Stand.*, **90**(6), 465–476 (1985).
9. P. Thy, K. Esbensen, 'Seafloor Spreading and the Ophiolitic Sequence of the Troodos Complex: A Principal Component Analysis of Lava and Dike Compositions', *J. Geophysical Research*, **98**(B7), 11799–11805 (1993).
10. B.R. Kowalski, T.F. Schatzki, F.H. Stross, 'Classification of Archaeological Artifacts by Applying Pattern Recognition Techniques to Trace Element Data', *Anal. Chem.*, **44**, 2176–2180 (1972).
11. H. Abe, S. Kanaya, Y. Takahashi, S. Sasaki, 'Extended Studies of the Automated Odor-sensing Systems Based on Plural-semiconductor Gas Sensors with Computerized Pattern Recognition Techniques', *Anal. Chim. Acta*, **215**, 155–168 (1988).
12. P.C. Jurs, B.R. Kowalski, T.L. Isenhour, C.N. Reilley, 'Computerized Learning Machines Applied to Chemical Problems: Molecular Structural Parameters from Low Resolution Mass Spectrometry', *Anal. Chem.*, **42**(12), 1387–1394 (1970).
13. J.A. Pino, J.E. McMurry, P.C. Jurs, B.K. Lavine, A.M. Harper, 'Application of Pyrolysis/Gas Chromatography/Pattern Recognition to the Detection of Cystic Fibrosis Heterozygotes', *Anal. Chem.*, **57**(1), 295–302 (1985).
14. A.B. Smith, A.M. Belcher, G. Epple, P.C. Jurs, B.K. Lavine, 'Computerized Pattern Recognition: A New Technique for the Analysis of Chemical Communication', *Science*, **228**(4696), 175–177 (1985).
15. B.K. Lavine, D. Carlson, 'European Bee or Africanized Bee? Species Identification Through Chemical Analysis', *Anal. Chem.*, **59**(6), 468A–470A (1987).
16. P. Geladi, H. Grahn, *Multivariate Image Analysis*, John Wiley and Sons, New York, 1996.
17. W.H.A. van den Broek, D. Wienke, W.J. Melssen, R. Feldhoff, T. Huth-Fehre, T. Kantimm, L.M.C. Buydens, 'Application of a Spectroscopic Infrared Focal Plane Array Sensor for On-line Identification of Plastic Waste', *Appl. Spectrosc.*, **51**(6), 856–865 (1997).
18. P. Robert, D. Bertrand, M.F. Devaux, A. Sire, 'Identification of Chemical Constituents by Multivariate Near-infrared Spectral Imaging', *Anal. Chem.*, **64**, 664–667 (1992).
19. D.D. Coomans, D.I. Broeckaert, *Potential Pattern Recognition in Chemical and Medical Decision-making*, Research Studies Press Ltd, Letchworth, England, 1986.
20. B.K. Lavine, 'Signal Processing and Data Analysis', in *Practical Guide to Chemometrics*, ed. S.J. Haswell, Marcel Dekker, New York, 1992.
21. F.W. Pjipers, 'Failures and Successes with Pattern Recognition for Solving Problems in Analytical Chemistry', *Analyst*, **109**(3), 299–303 (1984).
22. P.D. Wasserman, *Neural Computing*, Van Nostrand Reinhold, New York, 1989.
23. J. Zupan, J. Gasteiger, *Neural Networks for Chemists*, VCH Publishers, New York, 1993.
24. B.K. Lavine, P.C. Jurs, D.R. Henry, 'Chance Classifications by Nonparametric Linear Discriminant Functions', *J. Chemom.*, **2**(1), 1–10 (1988).
25. B.R. Kowalski, C.F. Bender, 'Pattern Recognition. A Powerful Approach to Interpreting Chemical Data', *J. Am. Chem. Soc.*, **94**, 5632–5639 (1972).
26. B.R. Kowalski, S. Wold, 'Pattern Recognition in Chemistry', in *Classification, Pattern Recognition and Reduction of Dimensionality*, eds. P.R. Krishnaiah, L.N. Kanal, North Holland, Amsterdam, 1982.
27. M. Sjostrom, B.R. Kowalski, 'A Comparison of Five Pattern Recognition Methods based on the Classification Results from Six Real Data Bases', *Anal. Chim. Acta*, **112**, 11–30 (1979).
28. B. Soderstrom, S. Wold, G. Blomqvist, 'Pyrolysis Gas Chromatography Combined with SIMCA Pattern Recognition for Classification of Fruit-bodies of Some Ectomycorrhizal Suillus Species', *J. Gen. Microbiol.*, **128**, 1783–1794 (1982).
29. S. Wold, 'Pattern Recognition: Finding and Using Regularities in Multivariate Data', in *Food Research and Data Analysis*, eds. H. Martens, H. Russwurm, Applied Science, Essex, England, 1983.
30. S. Wold, 'Pattern Recognition by Means of Disjoint Principal Component Models', *Patt. Recog.*, **8**, 127–139 (1976).
31. S. Wold, 'Cross-validatory Estimation of the Number of Components in Factor and Principal Components Models', *Technometrics*, **20**, 397–406 (1978).
32. S. Wold, C. Albano, U. Edlund, K. Esbensen, S. Hellberg, E. Johansson, W. Lindberg, M. Sjostrom, 'Pattern Recognition by Means of Disjoint Principal Component Models (SIMCA). Philosophy and Methods', in *Proc. Symp. Applied Statistics, NEUCC*, eds. A. Hoskuldsson, K. Esbensen, RECAU and RECKU, Copenhagen, 183–218, 1981.

33. S. Wold, C. Albano, W.J. Dunn III, U. Edlund, K. Esbensen, S. Hellberg, E. Johansson, W. Lindberg, M. Sjostrom, 'Multivariate Data Analysis in Chemistry', in *Chemometrics, Mathematics and Statistics in Chemistry*, eds. B.R. Kowalski, D. Reidel Publishing Company, Dordrecht, 1984.
34. S. Wold, M. Sjostrom, 'SIMCA, a Method for Analyzing Chemical Data in Terms of Similarity and Analogy', in *Chemometrics, Theory and Application*, ed. B.R. Kowalski, American Chemical Society, Washington, DC, 1977.
35. B.K. Lavine, H. Mayfield, P.R. Kroman, A. Faruque, 'Source Identification of Underground Fuel Spills by Pattern Recognition Analysis of High-speed Gas Chromatograms', *Anal. Chem.*, **67**, 3846–3852 (1995).
36. H.T. Mayfield, W. Bertsch, 'An Algorithm for Rapidly Organizing Gas Chromatographic Data into Data Sets for Chemometric Analysis', *Comput. Appl. Lab.*, **1**, 13–137 (1983).
37. E. Kovats, in *Advances in Chromatography*, eds. J.C. Giddings, R.A. Keller, Marcel Dekker, New York, Vol. 1, 230, 1965.
38. R.A. Johnson, D.W. Wichern, 'Applied Multivariate Statistical Analysis', Prentice Hall, Englewood Cliffs, New Jersey, 1982.
39. M.A. Stapanian, F.C. Garner, K.E. Fitzgerald, G.T. Flatman, J.M. Nocerino, 'Finding Suspected Causes of Measurement Error in Multivariate Environmental Data', *J. Chemom.*, **7**, 165–176 (1993).
40. *Handbook of Aviation Fuel Properties*, Coordinating Research Council, Inc., Atlanta, GA, 1983.
41. H.T. Mayfield, M. Henley, in *Monitoring Water in the 1990's: Meeting New Challenges*, eds. J.R. Hall, G.D. Glayson, American Chemical Society for Testing of Materials, Philadelphia, PA, 1991.
42. J.C. Spain, C.C. Sommerville, L.C. Butler, T.J. Lee, A.W. Bourquin, 'Degradation of Jet Fuel Hydrocarbons in Aquatic Communities', USAF Report ESL-TR-83-26, AFESC, Tyndall AFB, 1983.
43. W.E. Coleman, J.W. Munch, R.P. Streicher, H.P. Ringhand, W.F. Kopfler, 'Optimization of Purging Efficiency and Quantification of Organic Contaminants from Water using a 1-L Closed Looped Stripping Apparatus and Computerized Capillary Columns', *Arch. Environ. Contam. Toxicol.*, **13**, 171–180 (1984).
44. B.K. Lavine, D.A. Carlson, D.R. Henry, P.C. Jurs, 'Taxonomy Based on Chemical Constitution: Differentiation of Africanized Honeybees from European Honeybees', *J. Chemomet.*, **2**(1), 29–38 (1988).
45. B.K. Lavine, A.J.I. Ward, R.K. Smith, O.R. Taylor, 'Application of Gas Chromatography/Pattern Recognition Techniques to the Problem of Identifying Africanized Honeybees', *Microchem. J.*, **39**, 308–316 (1989).
46. B.K. Lavine, D.A. Carlson, 'Chemical Fingerprinting of Africanized Honeybees by Gas Chromatography/Pattern Recognition Techniques', *Microchem. J.*, **42**, 121–125 (1990).
47. B.K. Lavine, A.J.I. Ward, J.H. Han, R.K. Smith, O.R. Taylor, 'Taxonomy Based on Chemical Constitution: Differentiation of Heavily Africanized Honeybees from Moderately Africanized Honeybees', *Chemolab.*, **8**, 239–243 (1990).