Analyzing Chemical Data in More Than Two Dimensions

A Tutorial on Factor and Cluster Analysis

Thomas P. E. Auf der Heyde[1]
University of the Western Cape, Bellville 7530, South Africa

The extraction of information from the results of a chemical experiment most often involves the analysis of a considerable number of variables. Not infrequently a small number of these may contain most of the chemical information, while the majority add little or nothing that is interpretable in chemical terms. The decision—as to which variables are important and which are not—is often made on the basis of *chemical intuition* or *experience*, i.e. *subjective* criteria, rather than on more objective ones. This approach, instead of resulting from ill will on behalf of the chemist, is generally dictated by practical considerations, since the analysis of multidimensional data via standard techniques—such as scatter plots or correlation coefficients—often represents considerable hurdles.

Consider, for example, a chemical experiment that aims to investigate the effect of variations in molecular geometry on solubilities across a series of solutes and solvents. Such an experiment would necessarily involve correlating many different variables with one another: interatomic distances and angles, polarities, dielectric constants, temperatures, solubilities, molecular weights. Any attempt to seek correlations between these variables by means of scatterplots, say, would have the effect of (subjectively) collapsing an n-dimensional problem (where n represents the number of variables) onto two dimensions, with the consequent loss of all information not associated with these two. In many cases this may be an appropriate simplification, although it clearly cannot be so in general.

The problem faced by the chemist is therefore twofold. First, how can we avoid subjectively carving up the data into two-dimensional subsets, i.e., how can all n variables be *analyzed simultaneously* in order to reveal correlations between them? Second, how can the dimensionality of the problem be *objectively* reduced in order to interpret and visualize these correlations?

Multivariate statistical techniques that are suitable for these purposes have been developed and extensively applied in the social sciences, but their application to chemical problems has been mainly limited to analytical chemistry. The two most commonly applied methods are those of factor and cluster analysis. Factor analysis essentially searches for correlations among all variables simultaneously, extracting linear combinations of highly correlated variables that describe, in turn, the greatest amount of sample variance, the second greatest, and so on. Its main use therefore lies in dimensionality reduction. Cluster analysis represents a complementary technique that groups together similar points in the multidimensional data space, thereby yielding clusters or clouds of data points that are often useful in classifying the data. Used in conjunction the two methods afford the chemist an excellent means of visually representing the main characteristics of the data distribution in an objective way.

A recent analysis of crystallographic data on the molecular geometries of 196 five-coordinate d^8 metal complexes (*1*) affords a good example where a multidimensional treatment has yielded more clarity than previous investigations did. Five-coordinate complexes typically adopt either trigonal bipyramidal (TBP) or square-based (or rectangular) pyramidal (SQP) conformation, whereby the "typical" SQP has defied definition.[2] Cluster analysis of the data revealed that the complexes clustered about three archetypal—or "average"—conformations: the (expected) TBP, and two kinds of SQP that differ from one another in the amount by which the metal atom is displaced from the basal plane. The geometries of these three "average" conformations could be determined, and it was shown that the more "flattened" SQP (fSQP)—with its metal atom close to the basal plane—was characterized by an angle of 171° between pairs of ligand atoms trans to one another in this plane, while in the more "elevated" SQP (eSQP) this angle had a value of 163°. The results of a factor analysis suggested that this difference may be significant, since it was revealed that only the eSQP—and not the fSQP—is capable of distorting toward a TBP via the well-known Berry mechanism (*2*). The multidimensional analysis therefore enabled a more complete classification of five-coordinate conformations than had previously been possible, and it moreover revealed chemically meaningful results that had not been accessible by other means.

This tutorial is intended to introduce factor and cluster analysis at a level that will afford a senior student—and the practicing chemist, for that matter—some insights into the workings of the computer packages employing these methods, without overwhelming him or her with mathematical detail in the process.[3] We shall attempt this by first introducing some basic statistical concepts and terms, thereafter touching on the philosophical basis of factor analysis. Its mathematical basis will be sketched in broad outlines only. Cluster analysis will be similarly presented with an emphasis on its philosophical and qualitative aspects, rather than its quantitative mathematical detail. Both methods are demonstrated by the analysis of a simple, hypothetical three-dimensional data set chosen so as to enable anybody with a rudimentary understanding of matrix algebra and a working knowledge of a hand calculator to follow the calculations. Finally, the results of the analysis will also be graphically interpreted in order to illustrate the graphical use to which factor and cluster analysis may be put.

[1] Present address: Princeton University, Princeton, NJ 08544.

[2] The typical TBP geometry is defined by its point group symmetry, D_{3h}, while the point group symmetry of the SQP (C_{4v}) nonetheless allows it a degree of freedom with respect to its pyramidality, i.e., the elevation of the metal atom out of the plane defined by the four basal ligands (the square base of the pyramid).

[3] The application of factor analysis to analytical chemistry has been outlined by Malinowski and Howery (*3*), while Massart and Kaufman have described the use of cluster analysis (*4*). Both techniques are now commonly available in most statistical computer packages such as BMDP, SPSS, and CLUSTAN (*5*), and their usage is adequately described in manuals to such packages.

Basic Concepts and Terms

The Data Matrix

To start with, the data consisting of n readings or measurements on m objects are considered to constitute an $m \times n$ data matrix **D**, the elements of which are d_{ij} (where d_{ij} is the value of the jth variable for the ith object) and the columns of which list the variables $(1, 2, \ldots, j, \ldots, n)$.

$$\mathbf{D} = \text{objects} \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ i \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & \ldots & j & \ldots & n \end{array} \\ \begin{pmatrix} d_{11} & d_{12} & \ldots & d_{1j} & \ldots & d_{1n} \\ d_{21} & & & & & \\ \vdots & & & \vdots & & \vdots \\ d_{i1} & \ldots & \ldots & d_{ij} & & \\ \vdots & & & \vdots & & \vdots \\ d_{m1} & \ldots & \ldots & \ldots & \ldots & d_{mn} \end{pmatrix} \end{array}$$

Hence object i can be thought of as being represented by a row vector $\mathbf{d}_i = (d_{i1}, d_{i2}, \ldots, d_{in})$ called the *pattern* or *response vector*, which essentially defines a representative point for object i in the n-dimensional space spanned by the n variables.

Standardization and Scaling

In chemical problems units of measurement are usually chosen on the basis of availability and convenience and to a large extent depend on the gradation of the instruments one is using. They seldom, however, bear much relation to the natural or inherent variation of the characteristic being measured. In the description of molecular geometry, for example, changes of the order of 1 Å in interatomic distances might be relatively more important than changes of the order of 10° in interatomic angles. Similarly, two angle variables might have identical (or nearly so) means of 120°, say, while their ranges vary from 90° to 150°, and from 110° to 130°, respectively. The purpose of standardization and scaling is to express each observation in terms of the inherent variations of the system.

Although there are a number of techniques for standardization and scaling, we will concentrate on the so-called z *transformation* (*4*). This transformation expresses an observation as the number of standard deviations from the mean and leads to a matrix **Z** consisting of z scores

$$z_{ij} = \frac{d_{ij} - \bar{d}_j}{s_j} \tag{1}$$

where

$$\bar{d}_j = \frac{1}{m}\sum_{i=1}^{m} d_{ij} \tag{2}$$

and

$$s_{2/j} = \frac{1}{m-1}\sum_{i=1}^{m} (d_{ij} - \bar{d}_j)^2 \tag{3}$$

Here $\bar{d}_j$ is the sample *mean* of the jth variable while s_j^2 is the sample *variance*, or the square of the *standard deviation* (s_j, or sometimes also σ_j) of the sample. An important property (which will be illustrated later) of the z_{ij} values is that their covariance matrix is the same as the correlation matrix of the d_{ij}'s.

Measures of Similarity

In order to group together observations (representative points) some criterion of "similarity" will obviously need to be developed. Each object in the n-dimensional space will need to be compared with every other object in order to group together into the same cluster those that are "similar," while assigning dissimilar ones to different clusters. Two such measures will be considered here.

Covariance and Correlation. If the $m \times n$ data matrix **D** is premultiplied by its $n \times m$ transpose $\mathbf{D}^T$, after subtracting the mean of each variable, an $n \times n$ square matrix is obtained. After dividing its elements by the number of objects minus 1 it is called the *covariance* matrix **C.**

An element of this matrix is given by

$$c_{kl} = \frac{1}{m-1}\sum_{i=1}^{m} (d_{ik} - \bar{d}_k)(d_{il} - \bar{d}_l) \tag{4}$$

where

$$\bar{d}_k = \frac{1}{m}\sum_{i=1}^{m} d_{ik}$$

The matrix can be written as

$$\mathbf{C} = \frac{1}{m-1}\left[\mathbf{D}^T\mathbf{D} - \frac{1}{m}\mathbf{D}^T\mathbf{i}\mathbf{i}^T\mathbf{D}\right]$$

where **i** is a vector whose components are all 1.

It should be noted, firstly, that the diagonal elements of this matrix are equal to the variances of the n variables and, secondly, that the matrix is symmetric about the diagonal. Moreover, the sum of the diagonal elements, or the *trace* of **C,** is equal to the total variance in the data set. c_{kl} is large and positive when for most objects the values of variables k and l deviate from the mean in the same direction. The covariance c_{kl} of the two variables is therefore a measure of their association. This covariance or correlation between the two variables is often also expressed by the *correlation coefficient* r_{kl} where

$$r_{kl} = \frac{c_{kl}}{s_k \cdot s_l} \tag{5}$$

s_k and s_l are the standard deviations of variables k and l, respectively, and r_{kl} is hence a standardized covariance that lies between -1 and $+1$. For each element c_{kl} of the covariance matrix a correlation coefficient can be derived, and the covariance matrix **C** may consequently be transformed into a correlation matrix **R.** The covariance and/or correlation matrices almost invariably represent the basis of departure for the subsequent factor and cluster analyses.

Distance Measurements. In some cases, especially in cluster analysis, it may prove convenient to express the similarity of two observations in terms of the *distance* between the two representative points in the n-dimensional parameter space. Thus, the *Euclidean distance* x_{kl} between two points k and l in n-dimensional space is given by

$$x_{kl}^2 = \sum_{j=1}^{n} (d_{kj} - d_{lj})^2 = (\mathbf{d}_k - \mathbf{d}_l) \cdot (\mathbf{d}_k - \mathbf{d}_l)^T \tag{6}$$

where $(\mathbf{d}_k - \mathbf{d}_l)$ is the difference vector between the pattern vectors $\mathbf{d}_k$ and $\mathbf{d}_l$ for objects k and l, respectively, while $(\mathbf{d}_k - \mathbf{d}_l)^T$ is its transpose.

Massart and Kaufman (*4*) have shown, however, that correlation between variables in the n-dimensional space results in a distortion of the relation between the representative points, so that the Euclidean distance is an insufficient measure of similarity. The *Mahalanobis* distance, on the other hand, tends to compensate for this effect of correlation by incorporating the inverse of the covariance matrix ($\mathbf{C}^{-1}$) into the distance equation

$$x_{kl}^2 = (\mathbf{d}_k - \mathbf{d}_l) \cdot \mathbf{C}^{-1} \cdot (\mathbf{d}_k - \mathbf{d}_l)^T \tag{7}$$

The Mahalanobis distance measures the distance between two objects in terms of the inherent variation and covariation of the characteristics used to describe the system under consideration. It might also be thought of as the multivariate generalization of the z transformation in the following manner: The Mahalanobis distance from the ith object to the centroid, $\bar{\mathbf{d}}$, of the data or observations is

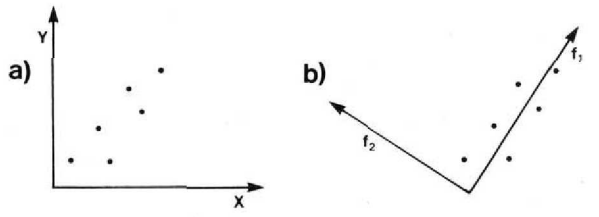

Figure 1. Diagram showing the relation between data distribution and (a) variable axes x and y and (b) factors $\mathbf{f}_1$ and $\mathbf{f}_2$. Note that the origin of the factor space actually lies at the midpoint of the data distribution; the factor axes in (b) are merely intended to emphasize the directions of variance.

$$x_i^2 = (\mathbf{d}_i - \bar{\mathbf{d}}) \cdot \mathbf{C}^{-1} \cdot (\mathbf{d}_i - \bar{\mathbf{d}})^{\mathrm{T}}$$

where the centroid $\bar{\mathbf{d}}$ is the vector of means taken over all objects. This distance is the number of "standardized units" that the ith object is from the centroid $\bar{\mathbf{d}}$. It should be noted, that, if the covariance matrix **C** (and hence also its inverse) happened to be the identity matrix **I**, then the Mahalanobis distance reduces to the Euclidean distance. This is seen immediately on substituting **I** in place of **C** in eq 7. Seen in this perspective, the Mahalanobis distance has similar appeal as does the z transformation while the Euclidean distance appears rather forced and unnatural. Most cluster analysis computer packages offer the option of using this distance measure rather than the straightforward Euclidean distance.

Factor Analysis[4]

The Philosophical Basis

Essentially factor analysis involves the transformation of the n orthogonal axes (representing the variables) that span the parameter space into n new axes (representing linear combinations of the variables), such that these new axes lie along the directions of maximum variance. This basic concept can easily be visualized with the help of a two-dimensional example. Consider the distribution depicted in Figure 1(a).

It is obvious from Figure 1(a) that the direction of maximum variance lies neither along the x axis nor along the y axis, but rather along some direction between them, i.e., along some combination of x and y. Similarly, the axis describing the direction of the second greatest amount of variation *away* from the principal direction of variance is coincident neither with x nor with y. Figure 1(b) depicts the identical distribution to that of Figure 1(a), but referred to a new set of axes $\mathbf{f}_1$ and $\mathbf{f}_2$, such that $\mathbf{f}_1$ represents the direction of greatest variance and $\mathbf{f}_2$ that of the greatest variance *orthogonal* to $\mathbf{f}_1$. Now, if the variation along $\mathbf{f}_2$ is minimal compared to that along $\mathbf{f}_1$, then it could justifiably be argued that the combination of x and y represented by $\mathbf{f}_1$ is adequate in describing the distribution of the data points in the two-dimensional space spanned by x and y. In other words, a reduction in the dimensionality of the data point distribution from two to one has been achieved.

In the case of an n-dimensional problem what factor analysis therefore yields are up to n orthogonal factors (linear combinations of the original variables) lying along, respectively, the axis of largest variance, the axis of second largest variance, of third largest variance, and so on. Often the number of factors needed to describe, say, 90% of the sample variance is less than n, so that factor analysis essentially affords one a technique whereby the dimensionality of the parameter space can be reduced, i.e., it is a dimension reduction method.

However, factor analysis offers a second important tool for multidimensional analysis that derives, in fact, from its original application in the social sciences and from which it took its name. Consider, for example, a hypothetical survey of lung cancer sufferers. These might be asked to complete questionnaires in which, among many other items, they are asked to indicate whether they are male or female, what the color of their hair is, how many cigarettes they smoke daily, what their incomes are, and so on. When the results of such a survey are subjected to factor analysis, what would very conceivably arise is a situation whereby one factor would be seen to account for most of the variance in the sample population, with other factors adding very little additional information. If this principal factor were examined for the components of the original variables present in it, it is very likely that the number of cigarettes smoked would feature as one of the components, while sex, for example, would not. The conclusion then would be that smoking is one of the "factors" that is correlated with lung cancer!

In other words, factor analysis can also reveal those underlying factors or combinations of the original variables that principally determine the structure of the data distribution and that not infrequently are related to some *real* influencing factor in the sample population. The task of the chemist, in our case, would then be to *interpret* in chemical terms those underlying factors extracted out of the data matrix by factor analysis.

The Mathematical Basis

The descriptive approach outlined by Murray–Rust in a series of papers on computer analysis of molecular geometry (*6*) will be used here. Essentially the mathematical basis of factor analysis rests on *eigenanalysis* of the covariance or correlation matrix (*3*, *4*). Eigenanalysis of a matrix **M** involves finding unique pairs of vectors $\mathbf{e}_i$ and scalars λ_i, called *eigenvectors* and *eigenvalues*, respectively, such that the following equation is satisfied

$$\mathbf{M} \cdot \mathbf{e}_i = \lambda_i \cdot \mathbf{I} \cdot \mathbf{e}_i$$

where **I** is the identity matrix.

Since the covariance matrix **C** is symmetrical about its diagonal, it will have real and nonnegative eigenvalues λ_i, and corresponding eigenvectors $\mathbf{e}_i$ can hence be obtained. Thus, eigenanalysis of an $n \times n$ covariance matrix, say, will yield n pairs of eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ and eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$. The n factors are then obtained from

$$\mathbf{F} = \mathbf{E} \cdot \mathbf{\Lambda}^{1/2} \qquad (8)$$

where **F** is the $n \times n$ matrix of the factors, **E** is the $n \times n$ matrix whose columns are the eigenvectors, and $\mathbf{\Lambda}^{1/2}$ is the $n \times n$ diagonal matrix composed of the square roots of the eigenvalues. Ordinarily the matrix **E** is composed of the normalized eigenvectors, i.e., eigenvectors whose lengths have all been normalized to unity, since, if this is the case, then

$$\mathbf{C} = \mathbf{F} \cdot \mathbf{F}^{\mathrm{T}}$$

i.e., there is a check offered of whether the factors extracted from **C** are the correct ones, in that multiplying the factor matrix **F** by its transpose $\mathbf{F}^{\mathrm{T}}$ ought to again yield the original covariance matrix **C**. (This point is explained in more detail in the Appendix).

The factors appear as linear combinations of the original variables in the form

$$\mathbf{f} = a_1 x_1 + a_2 x_2 + \ldots + a_i x_i$$

where x_i represents the original variables, while the coeffi-

[4] In this paper we follow the BMDP manual (see ref *5*) in the use of the term "factor analysis". It has been argued that the term "principal component analysis" might be more appropriate for the technique used here. See Chatfield, C., Collins, A. J. *Introduction to Multivariate Analysis*; Chapman and Hall: London, 1980.

cients $a_1, a_2, \ldots, a_i$ give an indication of the relative importance of the corresponding variable in the factor. These coefficients are often called *loadings*.

A further important point to note is that the λ's represent the proportion of the total sample variance which the corresponding eigenvectors (factors) account for. Thus the factor with the largest eigenvalue will be the most important, or principal factor, and will lie along the axis of maximum variance of the data.[5]

The factors $\mathbf{f}_n$ obtained from the columns of the factor matrix **F** are sometimes called *abstract* factors if they do not relate directly to any chemical information, but represent rather a composite mixture of the original variables. In order to obtain chemically meaningful factors, the abstract factor matrix **F** may be transformed or *rotated* into chemically meaningful data. This is accomplished using a $n \times n$ rotation matrix **A** such that the new factors g are obtained from

$$\mathbf{G} = \mathbf{F} \cdot \mathbf{A}$$

Essentially such a rotation corresponds to a rotation of the axes representing the abstract factors in the factor space until they become coincident with a set of chemically meaningful "chemical" factor axes in that space. There are two methods for doing this. The first, orthogonal rotation, preserves the orthogonal relation of the abstract factor axes on rotation, while the second, oblique rotation, does not. However, since rotation may lead to a subjective interpretation of the results of factor analysis, it needs to be treated cautiously, as has been pointed out repeatedly by Murray–Rust (*6c*, *6d*).

Finally, in order to analyze graphically the results of factor analysis it is necessary to convert the original data matrix **D** (or the matrix **Z** of z scores in the case of standardized data) into a matrix **S** of *factor scores*

$$\mathbf{S} = \mathbf{D} \cdot \mathbf{F} \qquad (9)$$

The factor scores for a given observation simply represent the coordinates of its representative point in the n-dimensional space spanned by the n factors, in much the same way as the values of the variables represent the coordinates for the data point in the original data space. The representative point in the original data space is therefore simply transformed into a new one as the original data space is transformed into the new factor space.

A Worked Example

In order to demonstrate some of the statistical techniques outlined above, a simple and hypothetical three-dimensional data set consisting of 12 observations will be subjected to a simple analysis. Starting with the data matrix **D** with the 12 cases of measurements of three variables x, y, and z.

	Case	x	y	z
	1	−5	−1	2
	2	−5	−4	−1
	3	−4	−2	3
	4	−3	−4	4
	5	−2	0	1
D =	6	−1	−2	0
	7	2	2	−2
	8	3	3	−1
	9	4	0	−3
	10	5	2	−1
	11	6	3	−5
	12	6	1	−7
$\bar{d}$ =		0.50	−0.17	−0.83
s^2 =		18.45	6.15	10.15
range =		11	7	11

where the means $\bar{d}$ and sample variances s^2 are obtained as shown in eqs 2 and 3.

The corresponding covariance matrix **C** as obtained from eq 4 is

$$\mathbf{C} = \frac{1}{11}\left[\mathbf{D}^{\mathrm{T}}\mathbf{D} - \frac{1}{12}\mathbf{D}^{\mathrm{T}}\mathbf{i}\mathbf{i}^{\mathrm{T}}\mathbf{D}\right] = \begin{pmatrix} 18.45 & & \\ 8.64 & 6.15 & \\ -11.09 & -4.79 & 10.15 \end{pmatrix}$$

and the corresponding correlation matrix **R** as obtained from eq 5 is

$$\mathbf{R} = \begin{pmatrix} 1.00 & & \\ 0.81 & 1.00 & \\ -0.81 & -0.61 & 1.00 \end{pmatrix}$$

The matrix **Z** of z scores obtained from the data matrix **D** according to eq 1 is

	−1.28	−0.33	0.89
	−1.28	−1.54	−0.05
	−1.05	−0.74	1.20
	−0.81	−1.54	1.52
	−0.58	0.07	0.57
Z =	−0.35	−0.74	0.26
	0.35	0.87	−0.37
	0.58	1.28	−0.05
	0.81	0.07	−0.68
	1.05	0.87	−0.05
	1.28	1.28	−1.31
	1.28	0.47	−1.94
$\bar{d}$ =	0.0	0.0	0.0
s^2 =	1.0	1.0	1.0
range =	2.56	2.82	3.46

The corresponding covariance matrix $\mathbf{C}_z$ is

$$\mathbf{C}_z = \frac{1}{11}\mathbf{Z}^{\mathrm{T}} \cdot \mathbf{Z} = \begin{pmatrix} 1.00 & & \\ 0.81 & 1.00 & \\ -0.81 & -0.61 & 1.00 \end{pmatrix}$$

which is identical to the correlation matrix **R** of the unstandardized data matrix **D**, as was pointed out above.

From the values of the correlation coefficients r in the matrix **R**, it becomes obvious that there is a high degree of correlation between x, y, and z and, furthermore, that any pair of variables can describe between 37% and 66% of the variance of the sample. This may be gleaned from the squares of the correlation coefficients which each represent the proportion of the variance that can be explained by the linear relatedness of the two parameters involved.

Since in this case both the scale and the range of the raw

---

[5] The philosophical relation between eigenanalysis and factor analysis is premised, in fact, on the definition of an eigenvector **e** of a matrix **M** as *a vector that is transformed into a multiple of itself by* **M**, i.e.

$$\mathbf{M} \cdot \mathbf{e} = \lambda \mathbf{I} \cdot \mathbf{e}$$

where λ is a scalar called the eigenvalue of **M** and **I** is the identity matrix. Suppose a covariance matrix **C** can be obtained from a given data set **D**, i.e., a matrix that describes the covariance between the variables describing **D**. Suppose further that some given linear combination **e** of these variables describes the axis of maximum variance in **D**. Now, if more data taken *from the same parent population* as **D** were added to **D**, then this should not seriously influence the axis of maximum variance, since the axes of maximum variance of any subset **D** of the parent population should be similar. All that this additional data added to **D** should do, if indeed the axis found represented the vector of maximum variance, is to *reinforce* this vector. In other words, neither the *direction* of **e** nor its length should vary greatly. Consequently **e**, in fact, represents an eigenvector of the covariance matrix **C**, since it can only be transformed into a multiple of itself by **C**, while its direction remains unchanged.

data values are almost identical for each variable, no scaling or standardization will be employed for the subsequent factor analysis. It must be pointed out, however, that this is bad practice if the variables have different units. Moreover, the covariance matrix will be used as the point of departure.

In order to extract from the covariance matrix **C** the eigenvalues λ and the corresponding eigenvectors **e** (and hence the factors), it is necessary to solve the following equation for λ

$$|\mathbf{C} - \lambda\mathbf{I}| = 0$$

where **I** is the identity matrix with diagonal elements equal to unity and the off-diagonal elements equal to zero and the vertical lines indicate that the determinant of the difference matrix between the lines should equal zero.

The above equation can be readily solved using standard matrix algebra and formulas for the solution of cubic equations.[6] In this case the roots extracted from **C** are $\lambda_1 = 30.17$, $\lambda_2 = 3.22$, and $\lambda_3 = 1.36$, i.e., these are the eigenvalues of the covariance matrix.

In order to obtain the corresponding eigenvectors **e** the following equation needs to be solved for the various λ's

$$(\mathbf{C} - \lambda\mathbf{I}) \cdot \mathbf{e} = 0$$

where **e** is the column vector of the three variables x, y, and z. Hence, the product of the difference matrix in brackets with the three-dimensional column vector **e** must equal zero.

On solution, the three eigenvalues yield eigenvectors

$$\mathbf{e}_1 = \begin{pmatrix} 2.03 \\ 1.00 \\ -1.36 \end{pmatrix} \quad \mathbf{e}_2 = \begin{pmatrix} 0.39 \\ 1.00 \\ 1.31 \end{pmatrix} \quad \mathbf{e}_3 = \begin{pmatrix} -0.84 \\ 1.00 \\ -0.51 \end{pmatrix}$$

for λ_1, λ_2, and λ_3, respectively.

The eigenvectors are ordinarily normalized, which in this case yields

$$\mathbf{e}_1 = \begin{pmatrix} 0.77 \\ 0.38 \\ -0.52 \end{pmatrix} \quad \mathbf{e}_2 = \begin{pmatrix} 0.23 \\ 0.59 \\ 0.77 \end{pmatrix} \quad \mathbf{e}_3 = \begin{pmatrix} -0.60 \\ 0.71 \\ -0.36 \end{pmatrix}$$

The matrix of eigenvectors **E** is hence

$$\mathbf{E} = \begin{pmatrix} 0.77 & 0.23 & -0.60 \\ 0.38 & 0.59 & 0.71 \\ -0.52 & 0.77 & -0.36 \end{pmatrix}$$

and that of the square roots of the eigenvalues is

$$\Lambda^{1/2} = \begin{pmatrix} 5.49 & 0 & 0 \\ 0 & 1.79 & 0 \\ 0 & 0 & 1.17 \end{pmatrix}$$

When these two matrices are combined as in eq 8, the factor matrix **F** emerges.

$$\mathbf{F} = \mathbf{E} \cdot \Lambda^{1/2} = \begin{pmatrix} 0.77 & 0.23 & -0.60 \\ 0.38 & 0.59 & 0.71 \\ -0.52 & 0.77 & -0.36 \end{pmatrix} \cdot \begin{pmatrix} 5.49 & 0 & 0 \\ 0 & 1.79 & 0 \\ 0 & 0 & 1.17 \end{pmatrix}$$

$$= \begin{pmatrix} 4.23 & 0.41 & -0.70 \\ 2.09 & 1.06 & 0.83 \\ -2.86 & 1.38 & -0.42 \end{pmatrix}$$

Consequently the three factors constituting the factor matrix are

$$\mathbf{f}_1 = 4.23x + 2.09y - 2.86z$$

$$\mathbf{f}_2 = 0.41x + 1.06y + 1.38z$$

$$\mathbf{f}_3 = -0.70x + 0.83y - 0.42z$$

A check on whether the correct eigenvalues have been found is afforded by a comparison of the sum of the λ's with the sum of the variances of the original variables. These should obviously be equal, since the total variance in the sample should be the same both *before* factor analysis and *after*. In this case the variances of the variables add up to 34.75 (= 18.45 + 6.15 + 10.15) as indeed do those of the factors (30.17 + 3.22 + 1.36) also! Furthermore, the proportion of the sample variance explained by each factor can be estimated from its eigenvalue. Thus, $\mathbf{f}_1$ has $\lambda = 30.17$, which represents ((30.17/34.75) × 100) percent of the variance. Hence $\mathbf{f}_1$, $\mathbf{f}_2$, and $\mathbf{f}_3$ describe respectively 86.8, 9.3, and 3.9% of the sample variance.

As pointed out earlier multiplication of the factor matrix **F** by its transpose $\mathbf{F}^T$ affords a means of checking whether the correct factors have been extracted from the covariance matrix **C**, since

$$\mathbf{C} = \mathbf{F} \cdot \mathbf{F}^T$$

if the eigenvectors making up **F** have been normalized.

In this case

$$\mathbf{F} \cdot \mathbf{F}^T = \begin{pmatrix} 18.55 & & \\ 8.69 & 6.18 & \\ -11.25 & -4.88 & 10.21 \end{pmatrix}$$

which is very close to the original covariance matrix obtained. Hence in this case the factors obtained are the correct ones within the limits of accuracy of these calculations, and we can therefore say that the first or principal factor adequately describes the variance in the sample. A reduction of dimensionality from three to one has consequently been achieved. As it is, these factors have no real significance, and rotation would thus be meaningless.

However, a graphical analysis of the results of the factor analysis might prove instructive. In order to accomplish this, it is necessary to transform the data matrix **D** into a factor score matrix **S** that represents the projection of each of the original observations onto the factor axes. This is done as in eq 9 and yields

$$\mathbf{S} = \begin{array}{c} \text{Case} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 12 \end{array} \begin{array}{c} \begin{array}{ccc} \mathbf{f}_1 & \mathbf{f}_2 & \mathbf{f}_3 \end{array} \\ \begin{pmatrix} -29.0 & -0.4 & 1.8 \\ -26.7 & -7.7 & 0.6 \\ -29.7 & 0.4 & -0.1 \\ -32.5 & 0.1 & -2.9 \\ -11.3 & 0.6 & 1.0 \\ -8.4 & -2.5 & -1.0 \\ 18.4 & 0.2 & 1.1 \\ 21.8 & 3.0 & 0.8 \\ 25.5 & -2.5 & -1.5 \\ 28.2 & 2.8 & -2.3 \\ 46.0 & -1.3 & 0.4 \\ 47.5 & -6.1 & -0.4 \end{pmatrix} \end{array}$$

The correlation matrix corresponding to **S** is

$$\mathbf{R} = \begin{pmatrix} 1.00 & & \\ 0.01 & 1.00 & \\ -0.06 & 0.05 & 1.00 \end{pmatrix}$$

Thus, within the context of the eigenanalysis performed and the truncation of numbers, the three factors are for all intents and purposes orthogonal to each other and consequently independent and uncorrelated, whereas the original variables x, y, and z were highly correlated. This emerges very clearly from an examination of Figures 2 and 3.

Figure 2 represents plots of the original variables against each other, and it reveals the approximately linear correlations between them. It is also easy to see, furthermore, that the data actually represent two clusters. Further information, however, cannot be gleaned from these plots.

Figure 3 represents plots of the factor scores against one another. From the scatter of the data points it becomes immediately obvious that $\mathbf{f}_1$, $\mathbf{f}_2$, and $\mathbf{f}_3$ are not correlated. Moreover, in Figure 3(a) and (c) the two clusters of points are very nicely separated, and it is this feature that makes factor analysis such a useful tool in more general and real cases. Additionally, both Figures 3(a) and (b) reveal two

[6] Any introductory text on matrix algebra will illustrate how to find eigenvectors and eigenvalues for simple 3 × 3 matrices.

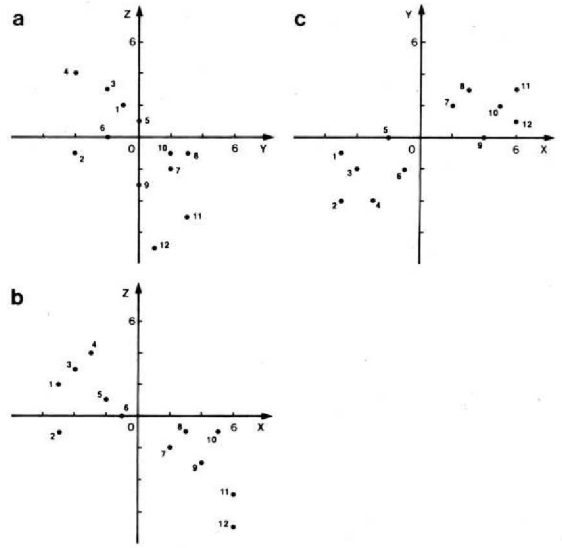

Figure 2. Scatterplots of data for worked example. Axes used are original variable axes.

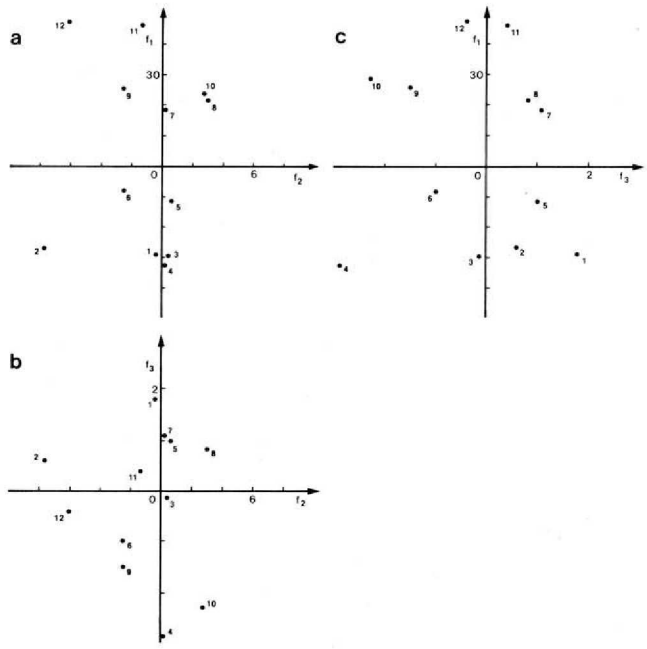

Figure 3. Scatterplots of factor scores for data of worked example.

possible outliers (observations 2 and 12). An examination of the data matrix reveals, in fact, that cases 2 and 12 (and possibly 11) are slight outliers.

In summary then, factor analysis is a method for mathematically, i.e., objectively, exploring the possibility of reducing the dimensionality of the problem being investigated, and it offers a useful graphical technique for the representation of both clusters and outliers in the sample. This point can also be intuitively appreciated as follows: consider two clusters of points. If the separation is sufficiently pronounced, then the line joining the two centroids will tend to become the major axis of variation, i.e., the principle component in the factor analysis. Projection onto that axis would help identify the two clusters.

Factor analysis of a more general n-dimensional data matrix would involve eigenanalysis of the $n \times n$ covariance or correlation matrix, and hence necessitate the solving of an equation of the nth power. The mathematical algorithms which have been developed for this purpose incorporate least-squares methods whereby the eigenvectors are consecutively calculated so as to minimize the residual error in each step. Thus each successive eigenvector accounts for a maximum of the variation in the data.

The procedure involves essentially the following steps. One, the eigenvector associated with the largest eigenvalue is orientated in the factor space so as to account in a least-squares sense for the greatest possible variance in the data. Two, the second eigenvector associated with the second largest eigenvalue is directed orthogonally away from the first and in the direction of maximum variance. These steps are then repeated for the $n - 2$ eigenvectors left, each step being subject to the conditions (1) that the eigenvector be orthogonal to each preceding one and (2) that it account for the maximum variance possible. In this way each eigenvector (or factor) that emerges from the iteration is orthogonal to all the previous ones and is oriented in the direction that maximizes the sum of squares of all projections onto that axis (factor) (*3*).

Since each successive eigenvector accounts for a smaller fraction of the total variance in the data, it often occurs that the first four factors, say, describe up to 90% of the variance, the last 10% being accounted for by the other $n - 4$ factors. In order, therefore, to know how many factors are necessary, some tests have been devised. For example, Kaiser's criterion, which has been incorporated into the BMDP factor analysis program, retains factors whose eigenvalues are greater than unity. An alternative criterion might be to retain all those factors that collectively account for, say, 90% of the variance and discard all others.

If the data were free of experimental error or superfluous information, then factor analysis would yield only c eigenvectors, one for each of the controlling "chemical" factors, where c is less than n. The computer algorithms cannot decide, however, which of the n eigenvectors have physical meaning. All that they can do is reject some of the more insignificant factors according to some pre-set condition. It would then be the task of the chemist to examine the results of the computer iteration judiciously and to interpret these according to his or her understanding of the chemical basis underlying the analysis.

Cluster Analysis

The Philosophical Basis

What all clustering algorithms essentially do is to cluster together similar or neighboring points into clusters in the n-dimensional space. Their differences lie mainly in the criteria used for establishing similarity and in the rationale according to which clusters are fused together. Generally, two types of algorithms are distinguished, these being hierarchical and nonhierarchical or relocation clustering. Both methods require the calculation of a similarity matrix, which contains a number indicating the "similarity" between each pair of observations of the original data set. This similarity, which is really a measure of the proximity of the pair of observations in the n-dimensional space, is usually expressed in terms of either the Euclidian or the Mahalanobis distance between the two points. Once this similarity matrix has been established the various clustering techniques can be applied to it.

Clustering Techniques

Hierarchical Clustering Methods. There are two opposing approaches to hierarchical clustering, these being agglomerative and divisive procedures. In agglomerative clustering each observation in a data set is initially considered as a

cluster on its own, and the hierarchical classification is built up by a series of linkages in which the most similar pairs of clusters are merged until all of the compounds are in a single cluster. Conversely, the divisive algorithm begins by placing all the observations into one cluster, which is then progressively subdivided into smaller ones until, finally, each observation is again in a cluster of its own. This approach may, consequently, be dubbed a "top down" technique, while the agglomerative algorithm represents a "bottom up" technique.

Nonhierarchical (Relocation) Clustering. Relocation (or partitioning) methods attempt to partition a data set into some number of disjoint clusters such that related or similar compounds fall into the same cluster, with compounds unrelated to that cluster being distributed among the other, well-separated clusters in the set (*4*, *7*). In general, the algorithm will generate a particular partition or clustering, determine the "goodness" of fit in some statistical sense and then relocate individual observations among the clusters until an optimum fit has been achieved. Since all the clusters are generated simultaneously, the resulting classification is nonhierarchical. Either the number of clusters to be generated can be specified in advance, or it may be optimized by the algorithm itself according to certain criteria.

Linkage Criteria

There are a large number of different criteria that have been developed to decide which individual elements and/or clusters should be merged together and in which way the similarity between a newly obtained cluster and other clusters or objects is defined. It is important to realize that the same algorithm may well give different results for a given data set depending on what linkage and similarity criteria are used. It is therefore important to apply different techniques or to complement the clustering method with graphical techniques (such as factor analysis) wherever possible.

Single linkage is the oldest and simplest procedure, and in it the distance between objects and/or clusters is simply considered to be equal to the shortest distance between two individual elements, one from each cluster.

Complete linkage is the opposite of single linkage, in that the distance between two clusters is now considered to be equal to the largest distance between two individual elements, one from each cluster.

Average linkage defines the intercluster distance as the average distance between all pairs formed by elements from each of the two clusters, respectively.

Centroid linkage focuses on the distance between the centroids of two clusters, or between the centroid of a cluster and an object outside of it.

Figure 4 shows graphically three of the linkage criteria outlined above.

The *K-means* clustering method has been devised for use with relocation algorithms exclusively, in contrast to the previous four linkage methods. Essentially this technique involves locating K centroids within the data space such that the sum of the distances from the data points to each nearest centroid is minimized. Obviously this will need to be done via an iterative procedure, since the first, usually randomly chosen distribution of the centroids is unlikely to correspond to that of the true centroids of the clusters in the data space. Depending on the algorithm used, however, the original K centroids can be specified or the number of clusters to be determined may be specified if there is some a priori notion of what the distribution is likely to be. The advantage of this method lies, of course, in lower computation times, although it would seem to be unfeasible for large data sets, unless these are highly ordered and the number of clusters is reasonably low. It must be emphasized, though, that different algorithms can yield different clusterings and that the results of a cluster analysis are therefore not necessarily unambiguous and, consequently, need to be supported by at least one other technique.

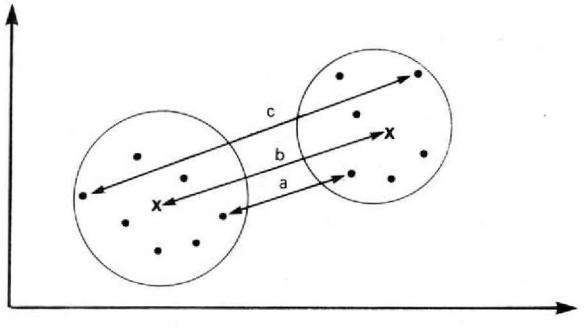

Figure 4. Diagram showing (a) single linkage, (b) centroid linkage, and (c) complete linkage.

A Worked Example

Two of the clustering techniques outlined in the previous section will be applied to a cluster analysis of the hypothetical data set used in the earlier worked example in order to demonstrate simply how the algorithms work. The two methods to be used are, firstly, agglomerative or "bottom up" clustering using the single linkage (nearest-neighbor) criterion and, secondly, divisive or "top down" clustering employing the complete linkage (furthest neighbor) criterion. The latter is of academic interest only at this stage, having never before been applied to a chemical analytical problem.

For both types of algorithms the point of departure is the similarity matrix. This has been established according to eq 6, and it contains simply the Euclidian distance between every pair of the 12 observations in the three-dimensional data space:

	Case											
Case	1	2	3	4	5	6	7	8	9	10	11	12
1	0											
2	4.2	0										
3	1.7	4.6	0									
4	4.1	5.4	2.4	0								
5	3.3	5.4	3.5	5.1	0							
6	4.6	4.6	4.2	4.9	2.4	0						
7	8.6	10.5	8.8	9.8	5.4	5.4	0					
8	9.4	10.6	9.5	10.5	6.2	6.5	1.7	0				
9	10.3	11.5	10.2	10.7	7.2	6.2	3.0	3.9	0			
10	10.9	11.7	10.6	11.2	7.5	6.4	3.2	2.2	3.0	0		
11	13.6	13.6	13.7	14.5	10.4	9.9	5.1	5.0	4.1	4.2	0	
12	14.4	13.5	14.5	15.1	11.4	10.3	6.5	7.0	4.6	6.4	2.8	0

From a cursory examination of the similarity matrix the following becomes obvious. First, the minimum distance between case 2 and any other member of the data set is 4.2, whereas most others have minimum distances considerably smaller than this, i.e., case 2 may be regarded as an outlier. Second, by a similar argument cases 11 and 12 (when taken as a pair) may be seen to be outliers, although they lie in close proximity to each other. Third, the data fall into two diffuse clusters, i.e., observations 1 to 6 and 7 to 12 in general being less than 5 units apart, with the elements of each cluster generally separated by more than 7 units. Indeed, a three-dimensional representation of the data distribution confirms these results as is shown in Figure 5, where it can be seen that the cases 1 to 6, with the exception of 2, fall into a quite different octant from those of cases 7 to 12.

Unfortunately, though, in practice the situation is seldom as unambiguous or as simple as this, and usually a skillful blend of cluster analysis, factor analysis, and graphical interpretation is required.

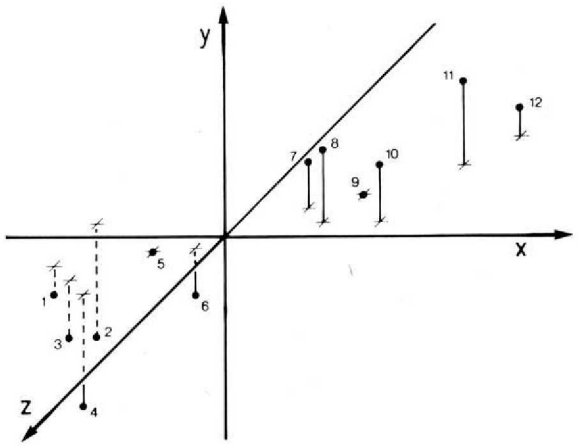

Figure 5. Diagram of the three-dimensional data space of the hypothetical data set used.

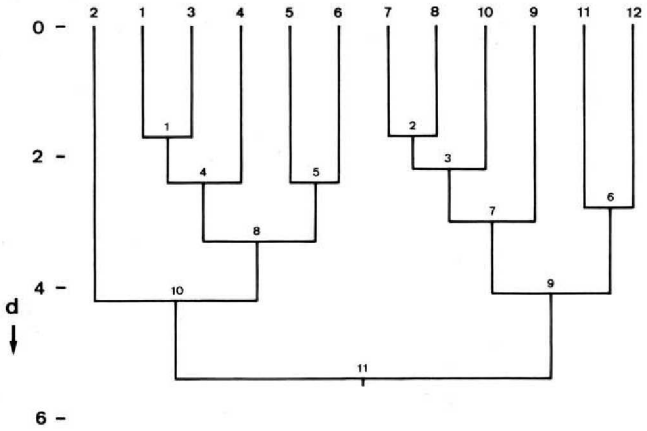

Figure 6. Dendrogram illustrating agglomerative clustering of the hypothetical data as a function of the intercluster distance (*d*). The level of clustering (*p*) at which a given cluster is formed is shown by the numbers within the dendrogram.

Single-Linkage Agglomerative Clustering. This technique begins by considering each observation initially to be in a cluster on its own, subsequently seeking to cluster together those compounds and/or clusters nearest to one another, until only one cluster remains. Applying this algorithm to the similarity matrix, it may be seen that observations 1 and 3, and 7 and 8, form the closest pairs (1.7 units) and will thus be first to be clustered together.[7] Thereafter case 10 is joined to the cluster (7, 8), since its distance from 8 (2.2 units) is now the shortest distance between any pair of elements of the data set. At the next stage, the fourth level of the clustering process, case 4 will be joined to the cluster (1, 3), since its distance to one of the elements (3) of that cluster is shorter than the distance between any other pair of observations and/or clusters at that stage.

This procedure is then repeated for a total of 11 stages, until all the elements have been joined together to form one cluster. The history of this clustering process is best represented in the form of a *dendrogram*, as shown in Figure 6.

Three important points emerge from the dendrogram. Firstly, observation 2 is only merged to the cluster (1, 3, 4, 5, 6) at the penultimate level of clustering ($p = 10$), thus easily identifying it as an outlier. Secondly, cases 11 and 12, joined together at the sixth level, are finally only joined to cluster (7, 8, 9, 10) at $p = 9$, thus also indicating that they may be treated as outliers of sorts. These observations echo the results obtained during the factor analysis of this data set, where these cases were also clearly identifiable as outliers from the plot of factor 2 against 3 and, more clearly, from that of factor 1 against 2 (Figures 3(b) and (a), respectively). Finally, the dendrogram graphically illustrates the notion that the data consist of essentially two clusters (1,2,3,4,5,6) and (7,8,9,10,11,12), since the elements of both are kept separate from each other until the final cluster is formed.

In the general case of a data set with m observations there would be $(m - 1)$ levels of clustering, and the algorithm would have difficulty in deciding at which stage to stop the process, i.e., at which stage are the clusters formed "meaningful" or "significant". In order to avoid confusion between this concept and that of "statistical significance" Massart (*8*) introduced the term "robust" cluster.

There have been some attempts at defining criteria for establishing the "correct" number of clusters. These are usually based on plots of some statistical or semistatistical measure, such as the average within-cluster distance, as a function of the number of clusters. Breaks in this curve are interpreted as indicating the emergence of robust clusters, or of the "correct" number of clusters. It must be pointed out,

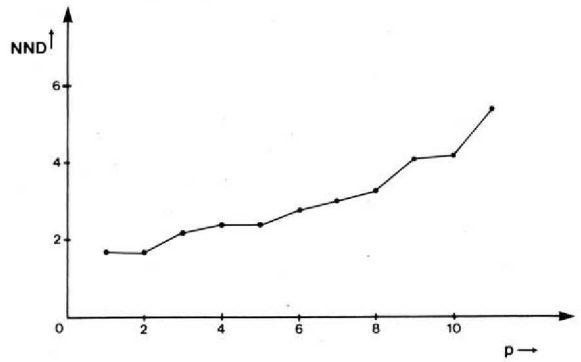

Figure 7. Plot of nearest-neighbor distance (NND) versus level of clustering (*p*).

however, that this approach has been the subject of some debate (*4*, *8*) and does not yet seem to have been resolved. Nevertheless, this method has been used successfully (*1*, *4*, *9*) and will therefore be demonstrated on this example. The measure chosen to indicate the break between "insignificant" and robust clusters is, for simplicity's sake, the nearest neighbor distance, since this would be expected to increase dramatically as the algorithm begins to cluster together robust, well separated clusters.

Figure 7 shows how the nearest-neighbor distance varies with the level of clustering p. Although the plot is not very striking, a discontinuity at the $p = 10$ level can be seen, implying that when the last two $(= 12 - 10)$ clusters are joined there is an increase in "heterogeneity" since the clusters joined are relatively far apart. It could similarly be argued that there is a slightly less obvious break in the graph at the level $p = 8$, suggesting that there are four $(= 12 - 8)$ clusters, these being the clusters (1, 3, 4, 5, 6), (7, 8, 9, 10), (11, 12) and (2). In this case, therefore, it would appear as if the "correct" number of clusters is either two or four.

Complete-linkage divisive clustering. In this approach the data points are initially all assumed to be in one (all embracing) cluster, and in subsequent steps the algorithm seeks to split those elements which are furthest apart from each other in any given cluster. The two resulting clusters

[7] Assuming that this algorithm is only capable of clustering two observations and/or clusters together at each level of clustering, this stage will necessitate two clusterings, i.e., the first level of clustering ($p = 1$) and the second ($p = 2$).

are then formed by the elements closest to those which were originally split apart.

From the similarity matrix it may be seen that cases 4 and 12, in fact, have the largest distance (15.1) between them. The algorithm will search the matrix for those elements closest to cases 4 and 12, respectively, and will then divide them up into two clusters according to their proximity. At the first stage this will therefore result in two clusters (1, 2, 3, 4, 5, 6) and (7, 8, 9, 10, 11, 12).

At the second stage the algorithm searches for the largest distance among the various pairs of elements in the two clusters and then splits that cluster which contains the most separated pair of elements. In this case it is observations 8 and 12, which are 7.0 units away from each other. The algorithm thus splits the cluster containing these two elements in such a way as to cluster around case 8 those elements closest to it, and similarly for case 12. Two clusters (7, 8, 9, 10) and (11, 12) emerge.

At the third stage the cluster (1, 2, 3, 4, 5, 6) is split, since two of its members (2 and 4) are now farthest apart (5.4 units). This division gives rise to the clusters (2, 6) and (1, 3, 4, 5).

This stepwise subdivision of the data set can again be summarized in the dendrogram shown in Figure 8.

The two dendrograms, the bottom-up and the top-down ones, offer interesting comparisons. For example, comparing the clusters at the eighth level in the former with those at the third level in the latter, i.e., where there are four clusters in both cases, one can see quite clearly that the two algorithms give significantly different answers. Thus the agglomerative technique yields clusters (7, 8, 9, 10), (11, 12), (1, 3, 4, 5, 6), and (2) while the divisive method results in the clusters (1, 3, 4, 5), (2, 6), (7, 8, 9, 10), and (11, 12). Moreover, whereas the former reveals case 2 as an outlier for nine successive clustering levels, the top-down approach isolates both observations 2 and 6 for only six successive clustering levels. The slightly different results obtained for the two techniques in this hypothetical example are intended as a "worst case" scenario[8] in order to indicate that the results of a cluster analysis need to be cautiously interpreted and that this interpretation will necessarily need to be guided by an understanding of the chemical basis underlying the analysis. In themselves the clustering algorithms do not yield an answer—this can only be arrived at through as judicious and objective an interpretation of their outcome as possible.

Conclusion

A complete analysis of multivariate chemical data cannot be satisfactorily achieved without resorting to multivariate statistical techniques. Factor analysis offers a completely objective mathematical technique for identifying important axes of variation of the data in the multidimensional data space. In this it can aid in identifying underlying factors which characterize the data distribution. It also affords a useful graphical tool, since scatterplots of the data onto planes described by the few most important axes will have the effect of separating data points from each other most effectively. Cluster analysis is a complementary technique that, although slightly less objective due to a large variety of possible linkage criteria, nonetheless is a considerable aid in identifying specific clusters of data points in the plots emerging from factor analysis.

Acknowledgment

I am indebted to John L. Fresen, Department of Statistics, University of the Western Cape, for helpful comments and discussions during the preparation of this manuscript.

[8] Indeed, in the example of the analysis of the molecular geometry of d^8 five-coordination (quoted earlier), hierarchical and nonhierarchical clustering techniques produced virtually identical results (*1*).

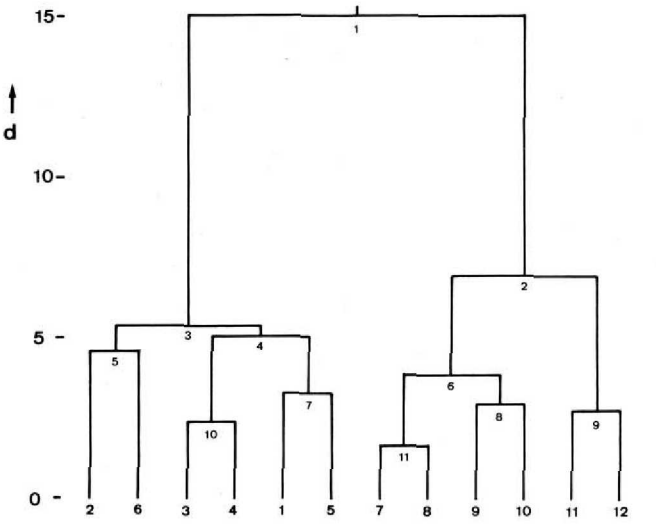

Figure 8. Dendrogram illustrating divisive clustering of the hypothetical data as a function of the intercluster distance (*d*). The level of clustering (*p*) at which a given cluster is split apart is shown by the numbers within the dendrogram.

Appendix

Suppose that n eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, and n corresponding eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ have been extracted from an $n \times n$ covariance matrix **C**. Then, for all i we have, by definition

$$\mathbf{C} \cdot \mathbf{e}_i = \lambda_i \cdot \mathbf{e}_i$$

Instead of using the individual $\mathbf{e}_i$ and λ_i, however, we can substitute the eigenvector matrix **E** composed of the column eigenvectors, and the eigenvalue matrix Λ, whose diagonal elements are the n eigenvalues and whose off-diagonal elements are zero. Thus

$$\mathbf{C} \cdot \mathbf{E} = \mathbf{E} \cdot \Lambda$$

Postmultiplying both sides of the equation by $\mathbf{E}^{-1}$, i.e., the inverse of **E**, we obtain

$$\begin{aligned} \mathbf{C} \cdot \mathbf{E} \cdot \mathbf{E}^{-1} &= \mathbf{E} \cdot \Lambda \cdot \mathbf{E}^{-1} \\ \mathbf{C} &= \mathbf{E} \cdot \Lambda \cdot \mathbf{E}^{-1} \\ &= \mathbf{E} \cdot \Lambda^{1/2} \cdot \Lambda^{1/2} \cdot \mathbf{E}^{-1} \end{aligned}$$

Now, if the eigenvector matrix is composed of normalized eigenvectors, then the inverse of **E** is just the transpose of **E**, i.e.,

$$\mathbf{E}^{-1} = \mathbf{E}^{\mathrm{T}}$$

Then, by eq 8 we have

$$\mathbf{C} = \mathbf{F} \cdot \mathbf{F}^{\mathrm{T}}$$

In other words, if the eigenvectors extracted from the covariance matrix **C** are normalized prior to computing the factor matrix **E**, then **C** should be recoverable from **F** and its transpose $\mathbf{F}^{\mathrm{T}}$ by multiplication of these two.

Literature Cited

1. (a) Auf der Heyde, T. P. E.; Bürgi, H. B. *Inorg. Chem.* **1989**, *28*, 3960–3969. (b) Auf der Heyde, T. P. E.; Bürgi, H. B. *Inorg. Chem.* **1989**, *28*, 3970–3981. (c) Auf der Heyde, T. P. E.; Bürgi, H. B. *Inorg. Chem.* **1989**, *28*, 3982.
2. Berry, R. S. *J. Chem. Phys.* **1960**, *32*, 933.
3. Malinowski, E. R.; Howery, D. G. *Factor Analysis in Chemistry*; Wiley: New York, 1980.
4. Massart, D. L.; Kaufman, L. *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*; Wiley: New York, 1983.
5. (a) BMDP: Dixon, W. J., Ed. "BMDP Statistical Software, 1983", Printing with Additions; University of California: Berkeley, 1983. (b) SPSS: Nie, N. H.; Hull, C. H.; Jenkins, J. G.; Steinbrenner, K.; Bent. D. H. *Statistical Package for the Social Sciences*, 2nd ed.; McGraw-Hill: New York, 1975. (c) CLUSTAN: Wishart, D. *CLUSTAN Users' Guide, the Clustan Project*; University College: London, 1975.
6. (a) Murray-Rust, P.; Motherwell, S. *Acta Cryst.* **1978**, *B34*, 2518. (b) Murray-Rust, P.; Bland, R. *Acta Cryst.* **1978**, *B34*, 2527. (c) Murray-Rust, P.; Motherwell, S. *Acta Cryst.* **1978**, *B34*, 2534. (d) Domenicano, A.; Murray-Rust, P.; Vaciago, A. *Acta Cryst.* **1983**, *B39*, 457. (e) Murray-Rust, P. *Acta Cryst.* **1982**, *B38*, 2765. (f) Murray-Rust, P.; Raftery, J. *J. Mol. Graphics* **1985**, *3*, 50. (g) Murray-Rust, P.; Raftery, J. *J. Mol. Graphics* **1985**, *3*, 60.
7. Willett, P. *J. Chem. Inf. Comput. Sc.* **1984**, *24*, 29.
8. Kaufman, L.; Massart, D. *Anal. Chem.* **1982**, *54*, 911.
9. Norskov-Lauritsen, L.; Bürgi, H. B. *J. Comput. Chem.* **1985**, *6*, 216.