Synopsis of Multivariate Methods in Geology

Richard A. Reyment

Department of Earth Sciences, Uppsala University, Sweden

Abstract: A synopsis of multivariate statistical methods as used in geological work is presented in the form of illustrations. Particular attention is given to the analysis of compositional data, the reification of multivariate results and questions such as stability and robustness of analyses.

Introduction

A very large part of the multivariate data accumulated in geological work derives from

analyses of rocks of various origins and complexity. These are usually chemical in nature, but sedimentology is also a common source of multidimensional observations. Some kinds of data encountered by geologists are:

1. Chemical analyses of rocks.
2. Petrographical determinations
3. Grain size classes of sediments
4. Frequencies of fossil species
5. Frequencies of mineral species
6. Content of ore minerals in a mining sample
7. Observations on physical properties of the crust

Six of these categories have a property in common that is not shared by the seventh. All are multidimensional, but the first six have a constant sum to their rows. This is the principal feature of most multivariate data in Geology, namely, that the data-matrix is *constrained*, or closed, so as to have rows that sum to the same constant. It may not be immediately obvious to you what this ingredient of "closure" implies for statistical analysis. Suffice it here to say that constrained data must be treated in a manner that is different from what pertains for "usual" data. The sixth category is more elusive. The content of ore in a mining sample is not evidently constrained, you might say. However, this is always expressed in relation to some measure of weight or volume, and it is this that imposes closure. The same situation arises in, for example, pollen analysis, where counts on frequencies of species are made on samples of constant weight or constant volume, or samples that are reduced, by a simple division, to a constant standard of reference.

For purposes of simplicity, the methods are ordered according to the number of samples to be analyzed and not according to theoretical statistical principles. The main structuring is as

follows:

One Sample:	Principal Component Analysis
	Principal Coordinate Analysis
	Q-*R*-mode combined analysis
Two Samples:	Linear Discriminant Function
	Quadratic Discriminant Function
	Generalized Statistical Distance
	Hotelling's T^2
Many Samples:	Canonical Variate Analysis
	Common Principal Components
	Multivariate analysis of variance
	"Discriminant Coordinates"
	Regional validity
Between-sets	Canonical Correlation Analysis

The Multivariate Sample

Some Basic Concepts

An array of data consisting of N specimens on which p characters have been measured is called a *data-matrix*, denoted here as **X**. Thus, each row of **X** is constituted by a vector of observations containing p components. You will seldom find this convention in statistical tests where the discussion is couched in terms of a p-dimensional vector of random observations. The data-matrix is, however, a very convenient concept in applied multivariate analysis, including geostatistical practice.

When the information contained in the matrix is analyzed so that the p variables are compared to each other, which is the most common approach, the analysis is said to be in the R-mode. If the alternative way is chosen, namely, to analyze the data-matrix so that one specimen is compared to another, the analysis is said to be in the Q-mode. In other words, R-mode applies to the treatment of the p columns of the data-matrix and Q-mode is specific to its N rows. The letter R comes from the standard representation of the sample correlation matrix. The letter Q has no other significance than that it precedes R in the alphabet. This usage comes from the field of Psychometrics. It is possible to unite both modes into a single graphical representation, referred to as a Q-R-mode figure.

The *mean vector* is the vector composed of the means of each of the p variables. It is then the sum of all the rows of that data-matrix, **X**, divided by N. It forms the centroid or barycentre of an empirical distribution.

The *covariance matrix* is a square $p \times p$ array formed from the variances and covariances of the p variables. The sample covariance matrix is usually written as **S**. It has the p variances ranged along its diagonal, the s_{ii} $(i=1, ..., p)$ and the covariances, s_{ij} in the off-diagonal positions. That is,

$$s_{ij} = s_{ji} \ (i = 1, ..., p; j = 1, ..., p)$$

which means that the matrix is a square symmetric matrix. The covariance matrix can be thought of as being a generalization of the variance of univariate statistics. By way of comparison, you will note that the data-matrix is neither symmetric nor square, other than by pure chance.

The matrix of correlations corresponding to **S** is written **R**. It has ones down the diagonal and correlations r_{ij} ($i \neq j$) in the off-diagonal positions. It is, then, also a square symmetric matrix. These quantities enter into almost all of the methods considered in the following pages. Before going any further, I should mention two conventions. The first concerns Greek letters for parameters. This was introduced by R. A. Fisher in order to make it easy to distinguish between theoretical, population quantities (Greek letters) and their sample estimates (Roman letters). Most statisticians tend to follow this but not even Fisher himself could be relied on to do so always. The second convention concerns the use of bold lettering. It has become more and more widespread, although by no means universal, to use bold type to denote vectors (lower case) and matrices (upper case) and to reserve italics for scalars.

Compositional Data

The most common kind of observations occurring in the geosciences are compositions. That is, the data are in the form of frequencies, proportions and percentages, all of which have the common property that the rows of the data-matrix sum to the same constant. This may not strike you as being much of an obstacle, but rest assured, there is no other area of data-analysis in which more incorrect applications of statistical methods have been perpetrated. Until Aitchison (1986) monographed the unit-sum constraint, the only avenues open to anyone wanting to try to get around the difficulty were to ignore it entirely, to wish it would go away, or to devise some totally inappropriate statistical analysis. In all cases, disastrous consequences are

habitually the outcome. Closed data require special methods for their correct analysis. Notwithstanding this sweeping declaration, you should be aware that the last word concerning the analysis of compositional data may not yet have been spoken.

Definition of Compositional Data

Any vector **x** with non-negative elements

$$x_1 + \ldots \ldots \ldots + x_D = 1 \qquad (1)$$

is subject to the unit-sum constraint. This condition is referred to as being a composition **x** composed of D parts summing to 1.

As geologists, we meet such data in geochemistry, petrochemistry, sedimentology, rock-analysis, palynology, palaeoecology, oceanography, environometrics, etc. In fact, it could almost be claimed that "closed data" are the most commonly occurring forms of measurements in general geology. Referring to biology, note that all serological data are of this kind, e.g. blood-group frequencies, counts of species occurrences, ecological observations, and many more.

The characteristic features of a compositional data-set are:

(a) each row of the data-matrix corresponds to a single specimen (i.e. rock sample);

(b) each column of the data-matrix represents a single chemical element, a mineral species, in short, a variable;

(c) each entry in the data-matrix is non-negative;

(d) each row of the data-matrix sums to 1 (proportions), respectively, 100 (percentages, frequencies). (N.B. you will sometimes find some other row-constant, owing to some manipulation or other);

A restricted part of real space, the *simplex*, constitutes the basic concept for the treatment of compositional data. The essential point is that although the vector **x** in (1) consists of D parts, the composition it represents is **completely specified** by the d components of a d-part subvector, defined as $d = D - 1$. Hence,

$$x_D = 1 - x_1 - \ldots \ldots - x_d. \qquad (2)$$

A D-part composition is therefore, to all intents and purposes, a d-dimensional vector. If you know the sizes of these d parts, x_D can be found by simple subtraction from the row-constant. The concept of the space of compositional data can then be simply defined as the d-dimensional simplex embedded in D-dimensional space.

The usual covariance matrix, when computed for a D-part composition, runs into interpretational difficulties. Some of these are:

1. Negative bias.

2. There is no relationship between the covariances of a subcomposition and those of the full composition. As the dimensionality of a subcomposition is decreased, so do the crude covariances between two specific parts fluctuate in sign. This is hardly a useful property by anybody's standards.

3. The way in which null-correlation is manifested is a further bugbear. A value of zero for the raw correlation coefficient of two parts of a composition is almost always an untrue representation of the real situation.

The Logratio Variance

The covariance structure of a *D*-part composition **x** is completely specified by the *½dD* logratio variances:

$$t_{ij} = var\{log(x_i/x_j)\} \quad (i=1, ..., d; j = i+1, ..., D) \qquad (3)$$

The inadvertency introduced by the logratio variance is that the logarithm of zero does not exist, so if there are such observations in the data, and they are common in geochemical work, special procedures must be devised to deal with this situation. The most direct way of dealing with the question is to see whether the part giving rise to zeros is necessary to the project; if not, one could consider excluding it from the analysis. I have seen many a table of analyses in which one component is always zero - it was analyzed for, but found to be lacking in all samples. A second alternative is to add some minute number to all the entries in the crude data-matrix. Such data are often referred to as *BDL* (below detection limit) observations Other, more advanced, procedures are discussed in Chapter 11 of the book by Aitchison (1986).

Centred Logratio Covariance Matrix

For a D-part composition $\mathbf{x}$, the $D \times D$ matrix

$$\Gamma = cov\left[\log\left(\frac{x_i}{g(x)}\right), \log\left(\frac{x_j}{g(x)}\right)\right] \quad i, j = 1, \ldots, D \qquad (4)$$

is termed the **centred logratio covariance matrix**. This matrix is the one used in the present connexion. Alternatives are discussed by Aitchison (1986) in case you should be interested. It is easy to interpret in that it has the advantage of being symmetric. The drawback is that it is singular (which means that its determinant is zero) which places a particular restriction on practical multivariate computational aspects. The most immediate obstacle in the path of many a multivariate analysis is how to obtain an inverse of the centred logratio covariance matrix (and correlation matrix). A generalized matrix inverse can be computed from the spectral relationship

$$\Gamma^{-} = \lambda_1^{-1} a_1 a_1^{T} + \ldots + \lambda_d^{-1} a_d a_d^{T} \qquad (5)$$

Equation (5) indicates that one computes the latent roots and vectors of $\mathbf{G}$, then performs the reconstitution indicated by formula (5) for the reciprocals of the d latent roots that are greater than zero:

BOX 1 *Example to illustrate Constrained Correlation*

DSDP data from the Sea of Japan; nine chemical elements (Usui, 1992)

Simplex Correlations

1.0000	.4308	-.3456	.6937	.3700	-.6411	.8709	-.5463	-.5072
.4308	1.0000	-.4696	.0638	.5206	-.6488	.2562	-.2360	-.4067
-.3456	-.4696	1.0000	-.133	-.1453	.2525	-.1642	-.1385	.1205
.6937	.0638	-.1337	1.0000	-.0277	-.2531	.4407	-.5742	-.1237
.3700	.5206	-.1453	-.0277	1.0000	-.7314	.3985	-.5431	-.2995
-.6411	-.6488	.2525	-.2531	-.7314	1.0000	-.5337	.6589	.1072
.8709	.2562	-.1642	.4407	.3985	-.5337	1.0000	-.3602	-.6613
-.5463	-.2360	-.1385	-.5742	-.5431	.6589	-.3602	1.0000	-.1244
-.5072	-.4067	.1205	-.1237	-.2995	.1072	-.6613	-.1244	1.0000

Raw Correlations

1.0000	.5729	-.3011	.7281	.3751	-.6532	.8778	-.5498	-.8654
.5729	1.0000	-.3892	.3439	.3667	-.7369	.3070	-.3724	-.6075
-.3011	-.3892	1.0000	-.0969	-.1728	.2023	-.1608	-.1991	.2509
.7281	.3439	-.0969	1.0000	-.0170	-.2966	.4680	-.5842	-.4717
.3751	.3667	-.1728	-.0170	1.0000	-.8023	.3937	-.6025	-.2119
-.6532	-.7369	.2023	-.2966	-.8023	1.0000	-.5799	.6390	.5217
.8778	.3070	-.1608	.4680	.3937	-.5799	1.0000	-.3962	-.8175
-.5498	-.3724	-.1991	-.5842	-.6025	.6390	-.3962	1.0000	.2090
-.8654	-.6075	.2509	-.4717	-.2119	.5217	-.8175	.2090	1.0000

==========

The differences between many of the entries in **Box 1** are obvious and any interpretations based entirely on the raw correlation coefficients can hardly inspire confidence. The present example is by no means an extreme case, granted that most of the differences in corresponding pairs of correlations are "supportable".

The stability of the logratio correlation

We now arrive at a subject which seems to be uncharted ground, namely, how stable are the logratio correlations? I have already told you that one of the advantages of the logratio covariance is that it is not affected by the number of parts in the composition, whereas the raw correlation coefficient when applied to constrained data is not invariant in that respect.

However, all is not gold that glisters, and I have found, empirically, that the logratio correlation coefficient tends to be sensitive to minor fluctuations in the parts. Returning to Usui's data on hydrothermal manganese minerals introduced in Box 1, I have altered slightly the proportions in just one of the compositions (specimens), and redone the computations.

We shall now compare the correlations of both kinds, before and after the alteration in the fifth observational vector. Consider the correlations for the simplex model, for the original data and the altered data in **Box 2**:

==========

Box 2: Stability of the log-ratio correlation coefficient
An empirical analysis.

Simplex Correlations for original data

1.0000	.4308	-.3456	.6937	.3700	-.6411	.8709	-.5463	-.5072
.4308	1.0000	-.4696	.0638	.5206	-.6488	.2562	-.2360	-.4067
-.3456	-.4696	1.0000	-.1337	-.1453	.2525	-.1642	-.1385	.1205
.6937	.0638	-.1337	1.0000	-.0277	-.2531	.4407	-.5742	-.1237
.3700	.5206	-.1453	-.0277	1.0000	-.7314	.3985	-.5431	-.2995
-.6411	-.6488	.2525	-.2531	-.7314	1.0000	-.5337	.6589	.1072
.8709	.2562	-.1642	.4407	.3985	-.5337	1.0000	-.3602	-.6613
-.5463	-.2360	-.1385	-.5742	-.5431	.6589	-.3602	1.0000	-.1244
-.5072	-.4067	.1205	-.1237	-.2995	.1072	-.6613	-.1244	1.0000

Simplex Correlations for altered data

1.0000	.5729	-.3011	.7281	.3751	-.6532	.8778	-.5498	-.8654
.5729	1.0000	-.3892	.3439	.3667	-.7369	.3070	-.3724	-.6075
-.3011	-.3892	1.0000	-.0969	-.1728	.2023	-.1608	-.1991	.2509
.7281	.3439	-.0969	1.0000	-.0170	-.2966	.4680	-.5842	-.4717
.3751	.3667	-.1728	-.0170	1.0000	-.8023	.3937	-.6025	-.2119
-.6532	-.7369	.2023	-.2966	-.8023	1.0000	-.5799	.6390	.5217
.8778	.3070	-.1608	.4680	.3937	-.5799	1.0000	-.3962	-.8175
-.5498	-.3724	-.1991	-.5842	-.6025	.6390	-.3962	1.0000	.2090

-.8654 -.6075 .2509 -.4717 -.2119 .5217 -.8175   .2090 1.0000

There are differences between the two covariance matrices, apart from those that should apply to the two variables the proportions of which were changed, namely, 2 and 9.

On the other hand, the situation for the crude covariances is different, as I have made you expect. Only the correlations for the two variables that were touched by the alteration have been influenced. All other entries remain unchanged. You will see that only the values listed in rows (and columns) 2 and 9 have been affected by the manipulation of observation 5 in the original data-matrix.

Crude correlation matrix for original data

1.0000	-.6425	-.4371	-.2175	-.1975	-.3677	.3313	-.8349	-.3044
-.6425	1.0000	-.1587	-.3216	.2067	-.0648	-.5004	.7270	-.0756
-.4371	-.1587	1.0000	.6140	-.0000	.4746	.1054	.0645	.4527
-.2175	-.3216	.6140	1.0000	-.1945	.7128	-.2395	-.1171	.4170
-.1975	.2067	-.0000	-.1945	1.0000	-.3802	-.1047	-.0848	.0211
-.3677	-.0648	.4746	.7128	-.3802	1.0000	-.1072	.3011	.0023
.3313	-.5004	.1054	-.2395	-.1047	-.1072	1.0000	-.2530	-.4085
-.8349	.7270	.0645	-.1171	-.0848	.3011	-.2530	1.0000	.0167
-.3044	-.0756	.4527	.4170	.0211	.0023	-.4085	.0167	1.0000

Crude correlation matrix for altered data

1.0000	-.0981	-.4371	-.2175	-.1975	-.3677	.3313	-.8349	-.6339
-.0981	1.0000	-.2780	-.3141	-.0577	-.2897	-.5497	.2860	-.1430
-.4371	-.2780	1.0000	.6140	-.0000	.4746	.1054	.0645	.7094
-.2175	-.3141	.6140	1.0000	-.1945	.7128	-.2395	-.1171	.6086
-.1975	-.0577	-.0000	-.1945	1.0000	-.3802	-.1047	-.0848	.1974
-.3677	-.2897	.4746	.7128	-.3802	1.0000	-.1072	.3011	.2403
.3313	-.5497	.1054	-.2395	-.1047	-.1072	1.0000	-.2530	-.2484
-.8349	.2860	.0645	-.1171	-.0848	.3011	-.2530	1.0000	.1631
-.6339	-.1430	.7094	.6086	.1974	.2403	-.2484	.1631	1.0000

==========

Methods for Analyzing a Single Sample

All methods considered in this section are concerned with analyzing a single multivariate sample. At the basic level of application, it is assumed that the data are multivariate normally distributed and interest is directed towards charting the geometrical and statistical properties of the sample. This is, however, not the only use that can be made of these methods and they can be profitably employed for seeking out heterogeneities in a data-set. This latter approach is often referred to as *exploratory data-analysis*.

Principal Component Analysis

This method is probably the most widely used one in applied multivariate statistics; it is also fundamental to the whole concept of multivariate statistics. In its most elementary form, the principal component analysis consists of the extraction of the latent roots and vectors of either the covariance matrix or the correlation matrix. The treatise by Jackson (1991) is recommended for all who wish to gain a deeper insight into the many ramifications of the method, notwithstanding that it roams off into all sorts of other fields and is regrettably deficient in its treatment of developments in applications in the Natural Sciences. Principal component analysis is an *R*-mode procedure. I have used the term "latent" above with respect to the roots and vectors of a matrix. Other terms used for latent roots and vectors are eigenvalues and eigenvectors, proper values and vectors, and characteristic roots and vectors, which latter amalgamation may just have priority over the designation chosen by me. *Eigenwert* is the German translation of characteristic root, which, in a period of anglophone linguistic decline, was back-translated as the truly horrible hybrid "eigenvalue". *Valeur propre* is the French rendition of *Eigenwert*, which also wandered to Britain as "proper value". The algebra of latent roots and vectors is rather complicated. If you wish to learn the basic principles, Reyment and Jöreskog (1993) have a section on the subject. Of the more technical accounts, Bellman (1960) is a mine of information.

The basic features of principal component analysis (but not the theoretical derivation thereof) depend on the extraction of the latent roots and vectors of a square symmetric matrix. Consider a real, square symmetric matrix, **R**. A latent vector **u** of **R** is given by

$$\mathbf{Ru} = \mathbf{u}? \qquad (6)$$

where ? is a scalar, called the latent root, to be estimated. Equation (6) can also be written as

$$(\mathbf{R} - ?\mathbf{I}) = \mathbf{0} \qquad (7)$$

where **I** is the *identity matrix*, a matrix with ones down its diagonal and zeros in all other positions, and **0** is the *null vector*, a vector with all its components equal to zero.

The first step in finding **u** and *?* is to solve the determinantal equation

$$|\mathbf{R} - ?\mathbf{I}| = 0 \qquad (8)$$

This expands to a polynomial with as many roots as there are dimensions. The lambdas enter into a diagonal matrix where they constitute the diagonal elements, all other positions being zero.
Thus,

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & \ldots & 0 \\ 0 & \lambda_2 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & & \\ 0 & 0 & 0 & \ldots. & \lambda_p \end{pmatrix}$$

with the lambdas ranged in descending order of magnitude. That is

$$?_1 > ?_2 > \ldots \ldots > ?_p.$$

By the same token, the latent vectors corresponding to the lambdas are found by solving

$$(\mathbf{R} - ?\mathbf{I})\mathbf{u} = \mathbf{0} \qquad (9)$$

for each lambda in turn. These vectors, when standardized to have unit length, can be grouped into a matrix **U**. This is an interesting, and useful, matrix because its columns are mutually orthogonal. In statistical terms, the latent vectors are linear combinations of the original *p* variables that are uncorrelated with each other.

Summing up what has been done, we have that in the most general form,

$$\mathbf{RU} = \mathbf{U?} \qquad (10)$$

which is equivalent to

$$\mathbf{R} = \mathbf{U?U}^{T} \qquad (11)$$

Let us ponder over this result for a moment. There are four interesting things to notice.

1. Equation (11) shows that a square symmetric matrix can be decomposed into two equal orthogonal matrices and a diagonal matrix. The superscript "T" attached to the second **U** indicates that it is transposed - it is lying on its side, as it were. (An equivalent, and more widely used, symbol is a dash placed after the symbol for the matrix to be transposed.)

2. The non-zero elements of **?** lie along a diagonal. The sum of these diagonal elements is exactly equal to the sum of the diagonal elements of **R**, which is a most valuable property. For one, it shows that the sum of the diagonal elements of **?** is equal to the sum of the variances of the original matrix. This sum is known as the *trace* of a matrix. The relationship implied is written:

$$\mathrm{tr}\,\mathbf{?} = \mathrm{tr}\,\mathbf{R}$$

3. Each diagonal element of **R** has a corresponding vector, a linear combination of the original variables. Hence, the *rotation* engendered in the principal component transformation makes *p* new uncorrelated variables, linear combinations of the *p* original, correlated ones. This is again a valuable property and one that can be used in many connexions in multivariate statistical analyses. Nothing has been altered, added or changed. All we have done is that we have transformed, by the rotation of axes, a set of correlated variables into an equivalent, new set of uncorrelated variables.

4. The product of the diagonal elements of **?** yield the numerical value of the matrix, its *determinant*. This is still another useful attribute in that this determinant is the same as that obtained from finding $|\mathbf{R}|$ directly. If you already have the latent roots of **R** it becomes a trivial matter to compute its determinant, a task that is onerous for larger matrices by the standard procedure.

5. There is a further point I shall mention now and return to later; it is a very important one. The *rank* of a square symmetric matrix is equal to the number of non-zero latent roots of that matrix.

The salient features of the calculations are displayed in **Box 3**.

==========

Box 3. A simple example of latent roots and vectors in statistics

	var 1	*var 2*	*var 3*	*var 4*
var 1	**.266433**	.085184	.182899	.055780
var 2	.085184	**.098469**	.082653	.041204
var 3	.182899	.082653	**.220816**	.073102
var 4	.055780	.041204	.073102	**.039106**

This matrix, which I shall call **R**, is square and it is symmetric. The diagonal elements, printed in bold type, are the (univariate) variances of the four variables in turn. The output you will get takes the following form:

simple principal component analysis for covariance matrix

latent roots = pca variances (the lambdas):

0.487875 0.072383 0.054776 0.009790

latent vectors = principal component loadings ($\mathbf{U}^{T}$):

component 1

0.686724 0.305346 0.623663 0.214984

```
   component   2
-0.669083    0.567479    0.343316    0.335305
   component   3
-0.265095   -0.729608    0.627171    0.063668
   component   4
 0.102281   -0.228919   -0.315968    0.915040
```

The latent roots are the lambdas found by solving (3). The latent vectors are the **u**'s obtained from equation (4), but with something extra done to them. The vectors, as computed, must be constrained so as to make them compatible from sample to sample since there is no unique solution to (4). If **u** is a solution, so is ***c*u** a solution, *c* being any scalar, and so on. The way out of this is to make the vectors have unit length and this has been done here. Hence, the four latent roots are the diagonal elements of **?** and the four rows of latent vectors are the *columns* of **U**. I have output the matrix in this form to show you what the transpose of **U** looks like.

If you add up the diagonal elements of **R**, and the four lambdas, you will get the sum 0.62482 in both cases. Another thing you can check is the sum of the squares of the components of the latent vectors. You should get a one, within rounding limits, in each case. A second instructive exercise is to multiply and add corresponding components in any two vectors and then take the cosine of this. The result should be 90°, within the limits of rounding, thus proving that the vectors are at right angles to each other, i.e. orthogonal, and hence, uncorrelated with each other.

==========

Scaling the latent vectors

I have called the latent vectors in the foregoing examples principal components. This is a common enough usage, but in many applications, the principal components are taken to be the latent vectors scaled by multiplication with the square root of the corresponding latent roots. This amounts to scaling each latent vector so that its squared length equals the matching latent root. This manipulation in no wise change the sense of direction of the vectors, since all vectorial elements retain the same proportionality. If **A** denotes the scaled latent vectors, the appropriate formula for producing these is:

$$\mathbf{A} = \mathbf{U}?^{½}.$$

It is appropriate at this point to consider some further properties of matrices insofar as they apply to the scaling of latent roots and vectors. In principle, there are three ways of scaling the vectors, of which we have already learned two. To summarize, these are:

(1) The vectors are scaled to unity. That is nothing is done to them other than to normalize them. These are the "raw" latent vectors obtained from the covariance or correlation matrix as in equation (9).

(2) The most commonly used scaling is the one we have just considered in this section, that is, the one that produces matrix **A**.

(3) There is a third method that is seldom found in applications in biology and geology, but which is widely used in quality control in industry (cf. Jackson, 1991). This is the scaling produced by *dividing* the elements of the latent vectors by the corresponding latent root

$$\mathbf{W} = \mathbf{U}\mathbf{L}^{-½}$$

These relationships bring to light some interesting identities of practical significance in statistical work.

The product

$$\mathbf{A}^{T}\mathbf{A} = \mathbf{L}$$

the diagonal matrix of latent roots.

The product

$$\mathbf{A}\mathbf{A}^{T} = \mathbf{S}$$

the covariance matrix with which we started.

The product

$$\mathbf{W}^{T}\mathbf{W} = \mathbf{L}^{-1}$$

the diagonal matrix of reciprocal latent roots.

The product

$$\mathbf{W}\mathbf{W}^{T} = \mathbf{S}^{-1}$$

the inverse of the matrix with which we started. This latter identity can be put to good use in some programming situations requiring a (not too accurate) matrix inversion.

z-scores and y-scores

The second method of scaling, the one that leads to z-scores, and which is the type commonly used in geology, leads to the scores defined by the relationship:

$$z^i = \boldsymbol{u}_{iT}[\boldsymbol{x} - \bar{\mathbf{x}}]$$

where $\mathbf{x}$ denotes any of the multivariate observations in the sample and $\bar{\mathbf{x}}$ is the mean vector of the sample. The principal components on z-scores have a practical advantage in geochemistry, for example, sine they are in the same units as the original variables. That is, grams per litre remain grams per litre. Likewise for the coefficients in matrix $\mathbf{A}$.

The *variances* of the scores are the squares of the corresponding latent roots, because

$$\mathbf{A}^T\mathbf{S}\mathbf{A} = \mathbf{L}^2.$$

The scores produced by the second type of scaling are computed as

$$y_i = \mathbf{w}_i^T[\mathbf{x} - \bar{\mathbf{x}}].$$

The attractive feature of these scores for quality control (which, of course, includes ore specimens) is that the variance is

$$\mathbf{W}^{T}\mathbf{S}\mathbf{W} = \mathbf{I}$$

the identity matrix. Hence, all variances are unity.

There is a simple relationship between the two kinds of scores, to wit:

$$y_i = \frac{z_i}{\sqrt{l_i}}$$

and

$$z_i = \sqrt{l_i}\, y_i$$

Principal Component Factor Analysis

For many years now, it has been customary to rotate the scores obtained from a standard principal component analysis, albeit with the benefit of some minor adjustments, of the correlation matrix by some appropriate technique. The idea comes from the realm of psychometry. This is usually known as "factor analysis" in geological work, although it is more correctly designated as principal component factor analysis, or principal component analysis with rotation of the axes to some kind of simple structure (Reyment and Jöreskog, 1993). There is a

considerable theoretical difference between the aims and methods of true factor analysis, as appropriate in psychometrical work, and almost all applications occurring under the name in the natural sciences. An example of true factor analysis in oceanology is given in Reyment and Jöreskog (1993) - the Ivorian oceanographical study.

In general terms the technique can be said to be concerned with sample quantities and there is no attempt at trying to estimate the population counterparts, such as pertains in true factor analysis in psychometry. Hence, the results obtained can only be interpreted at the level of the sample on which the calculations were performed. This is the **fixed model** as opposed to the **random model** (Reyment and Jöreskog, 1993, Chapter 4, section 4.2).

The procedure of rotating a principal component solution has in orthodox statistical spheres long been regarded as a manifestation of charlatanism. However, over the last decade, professional statisticians have seen the usefulness in doing this in some connexions and it is no longer considered poor form to rotate the axes of a principal component analysis in the search for enlightenment (Seber, 1984; Jackson, 1991; Preisendorfer, 1988).
==========

Box 4. A comprehensive principal component analysis

The data consist of a chemical analysis of 30 samples of impact glass from the Haitian bolide impact.

Number of dimensions = 9, being the following elements:

Si Al Mg Ca Na K S P H_2O

Correlation matrix

	1	2	3	4	5	6	7	8	9
1	1.000	.0286	.4699	-.8064	-.3469	-.8375	-.8298	.4252	.6836
2	.0286	1.0000	.3923	-.2956	-.0490	-.3207	-.4877	.6066	.5259
3	.4699	.3923	1.0000	-.7857	-.2858	-.7617	-.6701	.6462	.8156
4	-.8064	-.2956	-.7857	1.0000	.3388	.7919	.8250	-.6308	-.8970
5	-.3469	-.0490	-.2858	.3388	1.0000	.2911	.2619	-.3076	-.2762
6	-.8375	-.3207	-.7617	.7919	.2911	1.0000	.8994	-.6873	-.8023
7	-.8298	-.4877	-.6701	.8250	.2619	.8994	1.0000	-.7296	-.8966
8	.4252	.6066	.6462	-.6308	-.3076	-.6873	-.7296	1.0000	.6675
9	.6836	.5259	.8156	-.8970	-.2762	-.8023	-.8966	.6675	1.0000

Latent roots

5.82014 1.20923 .85460 .48070 .33069 .19719 .08106 .02402 .00237

latent vectors by columns

	1	2	3	4	5	6	7	8	9
1	.3285	-.4616	-.2044	-.3758	-.0757	.0001	.3068	.1393	.6138
2	.2041	.7228	.1584	-.2778	-.3826	-.1876	.3823	.0208	.0740
3	.3453	.0916	.0251	.7499	.1162	-.2875	.0857	.3881	.2350
4	-.3806	.1548	.0923	-.2184	.2237	-.5747	-.4345	.0483	.4510
5	-.1607	.3338	-.9083	.0540	.1387	.0984	.0452	.0026	.0616
6	-.3825	.1269	.1499	.0852	-.3140	.5771	-.1578	.4921	.3306
7	-.3927	-.0003	.1548	.3350	.0592	.0899	.4512	-.5902	.3823
8	.3290	.3139	.2101	-.1546	.7053	.4305	-.0825	-.0633	.1779
9	.3896	.0775	-.0916	.1585	-.4069	.1258	-.5712	-.4825	.2612

Latent vectors multiplied by the square root of the latent root

	1	2	3	4	5	6	7	8	9
1	.7924	-.5076	-.1889	-.2605	-.0435	.0001	.0874	.0216	.0299
2	.4924	.7948	.1464	-.1926	-.2200	-.0833	.1088	.0032	.0036
3	.8330	.1007	.0232	.5200	.0668	-.1277	.0244	.0601	.0114
4	-.9183	.1703	.0853	-.1515	.1287	-.2552	-.1237	.0075	.0219
5	-.3877	.3671	-.8397	.0374	.0798	.0437	.0129	.0004	.0030
6	-.9227	.1396	.1386	.0591	-.1805	.2563	-.0449	.0763	.0161
7	-.9474	-.0003	.1431	.2323	.0340	.0399	.1285	-.0915	.0186
8	.7936	.3452	.1942	-.1072	.4056	.1912	-.0235	-.0098	.0087
9	.9399	.0852	-.0847	.1099	-.2340	.0558	-.1626	-.0748	.0127

Test for equality of last *p*-1 roots of the correlation matrix

Chi-square = 357.9320

Degrees of freedom = 35

This value is highly significant, thus indicating that the latent roots are indeed all different.

Principal Component Factor Analysis by Varimax

varimax factor matrix

variable	communality	1	2	3	4	5	6
1	.9910	.9574	-.0942	.1845	.1448	.0797	.0649
2	.9881	.0737	.9577	-.0044	.1545	.2036	.0121
3	.9957	.3295	.1914	.1261	.8855	.2244	-.0094
4	.9842	-.6880	-.1097	-.1625	-.5365	-.1949	-.3829
5	.9998	-.1526	-.0102	-.9789	-.1027	-.0873	-.0079
6	.9919	-.7907	-.1394	-.1024	-.4510	-.2843	.2295
7	.9748	-.8293	-.3577	-.0789	-.2921	-.2582	-.0302
8	.9993	.3420	.3956	.1552	.2986	.7814	.0455
9	.9678	.6383	.4005	.1048	.5647	.1158	.2384
variance		37.38	15.99	12.07	20.27	10.17	2.93

Varimax factor score matrix

1	-1.1017	-.2256	1.1750	.4003	-1.2055
2	-.8601	-.7382	-.0588	.3871	-.7575
3	-.7540	-.4366	-.3136	.2209	-.9130
4	-.8258	-.5428	.1095	.2717	-.7371
5	-.5416	-.3064	-.1735	-.0336	-1.0822
6	-1.5024	-.7683	.5425	-2.2546	.0006
7	-1.1181	.2271	-.2287	-1.4631	-.3649
8	-1.3368	-.8879	.5203	-1.4147	-.2930
9	-.2434	-.1337	-4.7180	.0109	-.3307
10	-1.0168	-.0117	-.1945	.2907	-.9827
11	1.4190	2.0107	.3207	-2.2745	-1.1680
12	1.1527	-.0325	.2824	-.7228	-.5959
13	1.6414	-2.3496	-.4282	1.0744	-.4405
14	2.1218	-1.2958	-.1560	-.8125	-.3568
15	1.9338	-1.4104	.9151	-.6738	-.7237
16	-.8406	-.4176	.1096	.7512	1.4523
17	-.5460	-.2558	-.2089	.1021	1.4825
18	-.6179	-.3551	.0602	.7941	1.0726
19	-.5424	.0923	1.0368	.3849	.9859
20	-.2429	-.3611	-.2652	-.8325	2.3488
21	-.1455	2.7679	-.2526	-.1871	-.4140
22	-.5269	.4519	1.0818	1.7440	-.9332

23	.3418	1.1439	.0328	2.1013	-.8529
24	.4735	.5197	.1663	.6997	-.2862
25	.3598	1.1964	-.2618	.5553	-.1977
26	.5999	1.1232	.1886	-.3657	1.3388
27	.5940	.3071	-.0500	.0156	1.6095
28	.5081	-.1099	.8682	.6802	.9011
29	.6515	.5947	-.0599	-.0464	1.2182
30	.9656	.2042	-.0402	.5968	.2249

Zero latent roots

What do we do with latent roots that are zero (not due to the closure constraint), or almost zero? Should they be reified? The answer is yes because such roots explain linear relationships. The implication is, for z-scores for example, that

$$z_i = \mathbf{u}_i^T[\mathbf{x} - \bar{\mathbf{x}}] = 0 \text{ for any } \mathbf{x}.$$

In our present example, the coefficients associated with the almost zero latent root may be expressed as:

$$0.61x_1+0.07x_2+0.24x_3+0.45x_4+0.06x_5+0.33x_6+0.38x_7+0.18x_8+0.26x_9 = 0.$$

This is perhaps not such an exciting portrayal, given that all elements have the same sign.

Principal Components and Cross Validation

One of the aims of principal component analysis is to achieve a "parsimonious description" of a multivariate data-set. There is, therefore, a decision required as to how many principal components are to be retained in any given situation. There is no hard and fast rule for this,

although a "rule-of-thumb" exists. Compute the cumulative percentage variance contribution for successive values of the number of latent roots extracted; this we can call *k*. The appropriate level at which to stop extracting roots is usually taken at the point at which 95% of the trace of the covariance (correlation) matrix has been accumulated (Reyment and Jöreskog, 1993, p. 98). Another, less popular, rule of thumb is to retain all latent vectors that are at least as variable as the original variables, namely, equal to or greater than 1 for standardized variables (correlation matrix). There is also the "screes" method which indicates a cut-off for significant roots at a point where the graph of latent root against order falls off flatly (like the screes of débris on a mountain slope).

Applications of principal component analysis in chemometrics have shown that an *ad hoc* technique known as **cross-validation** can prove useful for obtaining answers to such practical questions as:

(a) How many principal components should be retained in an analysis?

(b) How many, if any, variables can be considered to be redundant?

(c) Do any of the specimens in the sample deviate from the others?

You will recognize the scope of cross-validation as being sample-oriented in that what you find pertains to the data-matrix on which you are operating, and there is no attempt at extrapolation to the population. Cross-validatory analysis is an exploratory technique that looks for interesting relationships in the sample. I find it to be most useful as a first step towards making a complete multivariate analysis. Analytical chemists seem to be completely convinced of the worthwhile nature of cross-validation in their work and it is therefore surprising, not to say alarming, that geochemists have not adopted the technique as far as I am aware.

Krzanowski (1987a,b) produced a synthesis of methods for obtaining an answer to the questions enumerated above, as well as other, more complicated ones and it is his combination of techniques that I employ here (example in **Box 5**).

==========

Box 5. A constrained cross-validational analysis of chemical data.

The data consist of chemical determinations made on the following constituents in alkaline Atlantic rocks, the oxides of: Si, Ti, Al, Fe^{+++}, Fe^{++}, Mn, Mg, Ca, Na, K, H_2O and P. The rows of the data matrix almost sum to 100%; that they do not do this exactly is due to minor elements having been deleted by the original analyst (cf. Borley *in* Sörensen, 1974). The input used here is, however, not the array of percentages but the log-ratio data-matrix (see below).

Number of Variables =12
Number of specimens = 23

Latent roots

7.1364 1.8310 1.1607 .8996 .4689 .1906 .1221 .0796 .0523 .0393 .0194 .0000

Note, that the last latent root is nought, as it should be for these data.

Principal components

Component	Si	Ti	Al	Fe^{+++}	Fe^{++}	Mn	Mg	Ca	Na	K	water	P
1	.34577	-.28402	.34318	-.00280	-.15014	.24722	-.36075	-.31475	.36565	.33888	.21029	-.27634
2	-.12799	.23875	-.11367	-.53873	-.62159	-.16651	.07893	-.28648	-.05720	.14191	.26281	.16909

3	.02239	-.04332	-.05970	.39814	-.07513	-.57385	.09979	.10889	-.04668	-.00334	.59970	-.34198
4	-.23045	-.50052	-.27044	.43555	-.19950	.15935	-.08402	-.28535	-.10545	-.02202	.12756	.50198
5	.02937	.34643	.08372	.45760	-.29910	-.27307	-.14908	-.02146	-.02971	.47150	-.49666	.07828
6	-.41022	.26741	-.13027	.23079	-.31458	.59718	.02228	.26362	.06469	.00757	.14021	-.37485
7	.13297	-.01698	.59804	.08260	-.40841	-.00056	-.15018	.35164	-.10234	-.47722	.04221	.25410
8	-.40849	-.48847	.09339	-.21171	-.11275	-.22686	.13076	.45146	.38813	.27509	-.17198	.00246
9	.53967	-.19878	-.28132	.08099	-.36731	.09016	.56049	.03537	.14073	-.14743	-.24915	-.14656
10	.35801	.03226	-.51714	-.11879	-.04472	.01492	-.51771	.50806	.03610	.10524	.12550	.18095
11	.10865	-.26780	.14590	-.11464	-.01725	.16527	.13200	.19189	-.77799	.42257	.03868	-.12192

Variable removed Residuals

	comp 1	comp 2	comp 3	comp 4	comp 5
1	.1549	1.1750	1.1825	1.2077	1.2078
2	1.2697	2.1061	2.1557	4.7810	3.3870
3	1.1645	1.2069	1.2766	1.2950	1.2834
4	.0002	**8.7016**	**13.2334**	**11.6639**	**12.3570**
5	.5673	13.7317	8.4548	5.3658	4.8385
6	1.0915	1.8202	**22.6235**	5.0779	4.5091
7	1.0126	.9982	1.0156	.9835	.9872
8	1.4178	2.1762	2.3459	1.6685	1.6662
9	.9460	.9298	.9365	.9169	.9169
10	1.1783	1.1892	1.1892	1.1924	1.7631
11	.8850	2.7908	**27.1701**	7.4366	7.8118
12	1.2323	1.8412	5.7259	4.0903	3.3889

Note: The values printed in bold type show where the highly significant changes occur on deletion of a variable.

Critical angles from deletion of specimens

Specimen *Maximum Angles*

Deleted	1	2	3	4	5	-5	-4	-3	-2	-1
1	.9559	2.1588	5.2389	5.1456	3.2456	3.3512	2.9749	2.0713	1.9854	.0668
2	1.5860	1.5958	8.8124	8.7488	4.7660	5.9461	6.6436	17.3737	11.2964	.0153
3	.9581	4.0379	3.8787	3.8873	9.4573	**16.1724**	6.0131	5.8305	4.5479	.0028
4	.7976	.7975	5.6314	2.6020	2.3632	4.0043	12.7931	14.7766	16.8919	.0028
5	.1363	.3044	2.4660	2.9045	5.4084	5.7249	5.9230	10.1374	.3065	.0011
6	4.3943	6.4606	14.2687	7.228	5.8251	7.7303	10.8821	10.3808	3.1632	.2203
7	.7324	.7817	.7827	.8478	4.7894	5.4253	17.4792	81.2251	9.6465	.0033
8	1.8094	2.4955	3.7011	9.8279	4.3709	5.0025	5.3468	1.7748	.8944	.0041

9	6.7428	9.2892	9.2896	10.1710	3.8076	**13.0346**	**13.0347**	**15.4255**	**13.1669**	2.8958
10	.8511	2.2301	2.2436	3.6214	4.4792	**76.1197**	**26.0967**	**26.0132**	12.9378	.0039
11	2.6076	11.8055	13.3077	4.4105	4.2425	6.8601	2.2562	1.2267	.5271	.0091
12	.1658	4.7613	5.3898	2.5071	2.4571	4.8253	5.0102	16.3280	6.0154	.0101
13	.0497	3.4406	2.7616	2.6747	3.6406	6.6312	11.7877	17.7691	12.2913	.0030
14	.9268	2.2425	2.7616	1.7005	1.7698	8.1458	9.1556	8.8263	8.8267	.0093
15	2.0611	4.5002	22.8387	10.0372	10.4071	**20.0983**	**19.2416**	**20.1018**	**21.7416**	.0013
16	4.4800	4.5918	28.3777	5.2141	5.1599	9.3412	9.7126	16.5213	88.0603	.0634
17	12.270	**17.7321**	**31.7687**	**80.4042**	**27.2568**	**29.4086**	**37.0641**	**44.6193**	**23.0932**	.6771
18	2.3529	5.2181	6.5107	4.7076	4.3923	**15.4335**	**14.0899**	**25.4795**	2.2717	.0038
19	.5083	2.4902	5.3361	2.2278	2.2173	4.5868	7.4909	7.7182	8.8497	.0030
20	.5551	**71.7444**	**31.6862**	**20.1174**	**15.7241**	**11.9063**	**12.3671**	**23.0049**	9.3480	.0265
21	.7759	**11.5738**	**33.2900**	**19.2942**	**14.7338**	**27.6782**	**17.6554**	**17.6554**	**17.7067**	.0087
22	1.8609	3.2423	3.5773	3.6792	**11.9640**	**13.9060**	**41.7238**	**34.5924**	**21.6445**	.0634
23	1.4797	2.0305	8.5720	4.5523	2.8010	4.8349	7.5434	5.9304	4.3779	.0065

Note: Deletion of specimens 3, 9, 15, 16, 17, 18, 20 and 21 perturb the results when deleted. These are candidates for closer inspection and interpretation.

Number of components to keep:

Number of components	*PRESS*	Test statistic
0	.9565	.0000
1	.5031	**6.3082**
2	.5071	-.0505
3	.4897	.2090
4	.3872	**1.4119**
5	.3372	.7063
6	.3149	.2957
7	.2990	.1900
8	.2942	.0480
9	.2914	.0219
10	.2820	.0533
11	.2768	.0159

==========

Interpretation of the cross-validational analysis

The main things to notice here are:

1. There are 11 latent roots, because of the constraint. You will see that the final value in the line for latent roots is zero. For this reason, there is no 12th principal component. Inspection of the array of latent vectors shows that the two iron species are not important in the first component, but they dominate, jointly, the second one.

2. The part that lists the residuals for "variables removed" shows what happens when each variable, in turn, is deleted from the analysis in the first four principal components. This gives you a good idea of whether some particular measure is really bringing essential information to the study. In the present example deleting the oxide of ferric iron (4) affects four of the principal components. Removal of the oxide of Mn (6) affects the third principal component strongly, as also does H_2O (11).

3. The section headed "minimum angles" lists the residuals when each specimen of the sample is deleted in turn. The table refers to five principal components, 1 through 5. The columns headed by integers with a negative sign usually indicate atypical observations that influence the pattern of correlations. The columns headed by positive integers show observations, the deletion of which from the sample has an influence on the variance pattern. The rule to be applied is that those observations that cause a marked increase in the magnitude of the residuals can be expected to be atypical, and, or, influential. Such observations are not always easy to detect in scatter plots of the raw data. This array is quite informative. Firstly, it tells us that most of the atypicality in the sample is due to covariances, to wit, specimens 9, 10, 17, 18, and 22. Specimens 20 and 21 influence both variances and covariances.

4. The section dealing with the number of principal components likely to contribute useful information indicates that four is a reasonable decision. This is also a rule-of-thumb technique, the rule being that values of the Prediction Sums of Squares (PRESS) is at least approximately one.

In conclusion, we have obtained much useful information about the sample of alkaline rocks. Firstly, that all variables contribute information, and that a few of them are more informative than others. Secondly, we have identified several observations that exert an undue influence on the analysis (and hence the stability of the principal component elements). The investigator would be well advised to pay special attention to these divergent observations and eventually repeat the analysis without them. Thirdly, the cross-validational synthesis gives a effective indication as to the number of components that can be assumed to be useful. In the present case, there are four.

Principal Component Analysis of Compositional Data

The calculations proceed in the same manner as outlined in the section on principal component analysis, but using the centred logratio covariance matrix or its correlational counterpart, introduced earlier on. In almost all applications known to me, the "crude" covariances or correlations are used, but this approach suffers from the crippling defect associated with interpreting crude covariance structures. The appropriate formulation of the principal component solution for compositional data is then to find the latent roots and vectors satisfying

$$(\mathbf{G} - ?_i\mathbf{I})\mathbf{a}_i = \mathbf{0} \qquad (12)$$

The logcontrast vector $\mathbf{a}^T_i$ log $\mathbf{x}$ is called the *i*-th logcontrast principal component.

Q-mode analysis

The next class of methods to be illustrated is that of Q-mode analysis. A Q-mode analysis is concerned with probing relationships between the objects of a sample. The reason for wanting to do this had not always been well understood, or appreciated by statisticians, who have, in the past, tended to regard it as a furtive procedure, best left unsung. This attitude has begun to change and there are few statisticians today who would negate the usefulness of "inverted factor analysis" for some aspects of data-analysis. The procedure I recommend here is that of *Principal Coordinate Analysis*, developed by Gower (1966). I usually rely on this method because Gower specifically designed it to accommodate the three classes of variables: quantitative, dichotomous, and qualitative. In this particular respect, it is superior to its competitors, at least in geological and biologiclal work, since they do not possess this quality. Reyment and Jöreskog (1993) provide an account of the area of Q-mode analysis (Chapter 5 in that book) to which you are referred for a more complete coverage than I aim at here, including what is rather inaccurately known as Q-mode factor analysis, and its appurtenances, but which enjoys considerable popularity in geochemical circles.

It may be necessary to stress that it is not a legitimate procedure to attempt to interpret the "loadings" of a Q-mode analysis, although I am well aware that this has become common practice in geochemistry, petrology and biology.

Gower's measure of distance between objects *i* and *j* is defined as the absolute difference between them for variable *p*, divided by the range of the variable and the sum for each variable subtracted from one. In essence, it is the same as the Euclidean metric and the city-block metric (Digby and Kempton, 1987, p. 20). In a situation where all variables are of the same kind and the association is of the correlation matrix kind, the Eckart-Young theorem applies and there may be no real advantage that makes one mode more effective than the other in

constructing an ordination. In cases where the variables are mixed and Pearsonian correlations are not applicable, the singular value decomposition cannot be applied.

For our present purposes, the essential features of the method can be summarized in the following words:

Given a set of N points in p-dimensional space, with coordinates given by the rows of the data-matrix for each of the points, the squared Euclidean distance d_{ij} between any pair of points, P_i and P_j, is expressed by-

$$d_{ij} = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \qquad (13)$$

The situation addressed by principal coordinate analysis is the converse of that described by equation (13) in that, if you have a set of N values that represent the squared distances within a set of N points in some Euclidean space, how do you go about finding the coordinates of these points. The idea is by no means new, having been around since 1935 (Schoenberg, 1935) under the rather vague designation "classical scaling" (Gordon, 1981). It was, however, Gower (1966) who made the method more generally known. The steps for doing the calculations are:

1. Make the centroid of the points lie at the origin of the coordinates.

2. Transform the matrix **D** of squared distances into a matrix **A** of inner products

$$a_{ij} = -\tfrac{1}{2}[d_{ij} - d_{i.} - d_{.j} + d_{..}]. \qquad (14)$$

3. Obtain the values of the coordinates x^{ij} from the distances d_{ij} by a standard principal component extraction.

$$\mathbf{A} = \mathbf{V?V'} \qquad (15)$$

4. The columns of the orthogonal matrix **V** contain the latent vectors ($\mathbf{v}_1$, $\mathbf{v}_2$, ..., $\mathbf{v}_N$) corresponding to the latent roots ($?_1$, $?_2$, ..., $?_N$).

5. The matrix of coordinates **X** is found by computing

$$\mathbf{X} = \mathbf{V?}^{\frac{1}{2}} \qquad (16)$$

6. Usually only the first 2 to 3 coordinates attract much interest (that is, k =2 or 3). The sum of the remaining latent roots, expressed as a percentage of the trace (the sum of all of the latent roots) supplies an indicator of how well the analysis has succeeded in preserving distance. If the residual is very large, then caution is advisable for interpreting the ordination obtained.

==========

Box 6. An example of a *Q-mode* analysis: principal coordinates

---

Example of Principal Coordinate Analysis

The Alkaline rock data

variables: the oxides of Si Ti Al Fe_2O_3 FeO_2 Mn Mg Ca Na K H_20 P_2O_3

Variables = 12
Rock samples = 23

Gower's Association Matrix

	1	2	3	4	5	6	7	8	9	10
1	1.0000									
2	.80528	1.00000								
3	.70787	.67111	1.00000							
4	.82483	.90029	.70665	1.00000						
5	.84202	.78440	.85567	.84154	1.00000					
6	.61028	.57889	.82102	.57749	.71201	1.00000				
7	.86419	.84622	.65039	.88521	.77734	.55345	1.00000			
8	.70662	.63498	.79774	.68967	.80336	.81585	.69020	1.00000		
9	.59313	.55710	.80316	.58571	.70401	.92868	.57288	.84317	1.00000	
10	.82730	.87799	.76039	.86308	.86926	.63709	.83257	.71885	.62212	1.00000
11	.65979	.64792	.88405	.66747	.78989	.86049	.60497	.76023	.82821	.70738
12	.83527	.71402	.80140	.73201	.86809	.71361	.76200	.76760	.70589	.78491
13	.83434	.76600	.77944	.80167	.87944	.73078	.79164	.83359	.72004	.83745
14	.89090	.83265	.63853	.83889	.77492	.56018	.90146	.65651	.54302	.82786
15	.84145	.72952	.56531	.74010	.70021	.49978	.82121	.62817	.48904	.71472
16	.64354	.67527	.43071	.68577	.55125	.26326	.69091	.39205	.27856	.62199
17	.51501	.50315	.79292	.53964	.67076	.81874	.47937	.72416	.82535	.56754
18	.73391	.81962	.58414	.84383	.71881	.43718	.79366	.54114	.43242	.75936
19	.87322	.75384	.69204	.78154	.79147	.62326	.84570	.74230	.61499	.80609
20	.64007	.60736	.75966	.61032	.70641	.74630	.63469	.73783	.73943	.62742
21	.71942	.61541	.79018	.69084	.78545	.63617	.66365	.69179	.65078	.71054
22	.74939	.82290	.61211	.77540	.70481	.50215	.81162	.59193	.49030	.81022
23	.77513	.76921	.57146	.77212	.68650	.44247	.82907	.57951	.46189	.75437

	11	12	13	14	15	16	17	18	19	20
	1.00000									
12	.75332	1.00000								
13	.74167	.83464	1.00000							
14	.59045	.81330	.81392	1.00000						
15	.51723	.71943	.73821	.81757	1.00000					
16	.35323	.51629	.51641	.69296	.65649	1.00000				
17	.82135	.60854	.60346	.44568	.37246	.25172	1.00000			
18	.51830	.59647	.69033	.76188	.69915	.81482	.40558	1.00000		
19	.62507	.78625	.85658	.85481	.82233	.61436	.49918	.72949	1.00000	
20	.80333	.73775	.69833	.60100	.54061	.33654	.72035	.46180	.61263	1.00000
21	.71170	.72214	.72274	.64799	.56087	.48333	.63205	.61404	.72489	.76081
22	.53694	.68432	.68929	.78106	.74174	.70098	.42501	.78293	.76963	.49861
23	.51782	.67724	.69398	.81355	.75984	.79483	.40073	.83526	.79357	.49896

	21	22	23
21	1.00000		
22	.57221	1.00000	
23	.65621	.86263	1.00000

Latent roots (first three) of transformed association matrix

2.5335870
.6946650
.4887996

Square root of the latent roots

1.59172 .83347 .69914

Specimen		Coordinates		
1	a	-.17281	-.22459	.06185
2	b	-.23604	.07957	-.10360
3	c	.34084	.08013	.10531
4	d	-.20739	.05608	-.00489
5	e	.07374	-.08381	.12184

6	f	.51856	.02685	-.22731
7	g	-.27216	-.09619	-.04094
8	h	.30386	-.12615	-.13475
9	i	.51537	.04860	-.22305
10	j	-.11726	-.02776	-.00869
11	k	.43332	.11414	.01824
12	l	.08962	-.23156	.09009
13	m	.04889	-.23072	-.01925
14	n	-.28212	-.18280	-.01990
15	o	-.30178	-.22874	-.10912
16	p	-.52459	.37200	.02847
17	q	.55493	.27980	-.07808
18	r	-.39774	.28576	.00337
19	s	-.16004	-.22907	-.00193
20	t	.37226	.00339	.22804
21	u	.15366	.06632	.46827
22	v	-.33495	.10827	-.16580
23	w	-.39818	.14049	.01182

The primary and perhaps sole interest in doing a principal coordinate analysis lies with the graphical representation of relationships between objects. A simple graph of the ordination obtained for the present data is given in **Fig. 1.**

values of residuals

roots exceeding	percentage residual
2	52.41

3 45.21

The conclusion prompted by this information concerning the residuals is that the analysis has been moderately successful in retaining distance relationships between the sample points since the percentage residuals are relatively small for this kind of analysis. There may be some uncertainty as to why a *Q*-mode principal coordinate analysis (or inverted *Q*-mode principal component factor analysis) could be useful, when a straight principal component analysis would probably do just about as well at disclosing eventual structure in the data. The way I look at the question is that if I have a mixture of categories, such as occur in many kinds of geological work, I prefer Gower's solution, providing that most of the trace of the Association Matrix used is concentrated to its first three latent roots. If the latent roots just drift off, becoming successively only a little smaller, then there is much random variation in the data-matrix and a principal coordinate ordination might not be giving you much of genuine value.

More on R- and Q-Mode Methods

A fundamental theorem of multivariate analysis is the *singular value decomposition*, first

enunciated in 1889 by the celebrated English mathematician J. J. Sylvester, expanded in scope by L. Autonne in 1913, and generalized to rectangular matrices by Eckart and Young in 1936. It is the key to the analysis of *Q*-*R*-mode relationships in the multivariate population. In statistics, the term often used (at least in psychometry) is the *basic structure of a rectangular matrix*.

Let **X** be a given data-matrix of order *N* by *p*, where *N*>*p*, and let *r* be the rank of **X**. In most cases, *r*=*p*, but we have already learned that the logratio data matrix is of rank *r*=*p*-*1*. The singular value decomposition of **X** states that

$$\mathbf{X} = \mathbf{VGU}^{\mathbf{T}} \qquad (13)$$

where $\mathbf{V}_{(N,r)}$ is a matrix with orthonormal columns, $\mathbf{U}_{(p,r)}$ is orthonormal, and $\mathbf{G}_{(r,r)}$ is a diagonal matrix with *r* positive diagonal elements $?_1 > ?_2 > ..... > ?_r$. These gammas are called the *singular values* of **X**.

Interpretation

1. The product $\mathbf{XX}^{\mathbf{T}}$ is sometimes referred to as the *major product moment*. It is of order *N* by *N* and has *r* positive latent roots $?_1^2, ..., ?_r^2$ and *(N-r)* zero latent roots. The corresponding latent vectors are $\mathbf{v}_1, ..., \mathbf{v}_r$.

2. The alternative multiplication, $\mathbf{X}^{\mathbf{T}}\mathbf{X}$, called the *minor product moment* in factor analysis, is of order *p* by *p*. It has *r* positive latent roots, $?_1^2, ..., ?_r^2$. These are the same as for the major product moment. There are *(p-r)* zero latent roots. The latent vectors corresponding to the positive latent roots are $\mathbf{u}_1, ..., \mathbf{u}_r$.

3. If $\mathbf{v}_m$ is a latent vector of $\mathbf{XX}^T$ and $\mathbf{u}_m$ is a latent vector of $\mathbf{X}^T\mathbf{X}$, both corresponding to the latent root $?_m^2$, then the following relationships hold:

$$\mathbf{v}_m = (1/?_m)\mathbf{X}\mathbf{u}_m \qquad (14)$$

and

$$\mathbf{u}_m = (1/?_m)\mathbf{X}^T\mathbf{v}_m. \qquad (15)$$

Thus, there is an easy path from the Q-mode state to the R-mode one.

4. More generally,

$$\mathbf{V} = \mathbf{XUG}^{-1}$$

and

$$\mathbf{U} = \mathbf{X}^T\mathbf{VG}^{-1}.$$

The singular value decomposition of a rectangular matrix is the best method for programming *Correspondence Analysis* and the *biplot* (see p. 000). The biplot is rather like the idea of correspondence analysis but differs in that it was designed by Gabriel (1971) for continuously distributed variables and there is no scaling of axes involved such as typifies Bénzécri's adaptation. It was originally exemplified by climatological data in Gabriel's paper, which clearly betokens its geoscientific importance.

As indicated by condition (4) above, you can work on very large data-matrices on which relatively few variables have been determined by the following algorithm:

1. Compute the minor product moment, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. This is p by p and $p < N$.

2. Compute the positive latent roots of the minor product moment and the corresponding latent vectors.

3. Compute then the vectors $\mathbf{V}$ from the first part of condition (4).

These vectors are of interest in *Q*-mode analysis.

==========

Box 7. *Example of Correspondence Analysis*

The Phenanthrene data from Telnaes *et al.* (1987).

The data consist of 11 phenanthrene determinations made on 36 samples of North Sea oils: monomethyl phenanthrenes (4 species), dimethyl phenanthrenes (6 species) and phenanthrene.

Similarity matrix

1	.1596	.1198	.1210	.1499	.1309	.0681	.0773	.1595	.1096	.0984	.0830
2	.1198	.0913	.0915	.1129	.0980	.0520	.0591	.1207	.0826	.0741	.0622
3	.1210	.0915	.0925	.1138	.0994	.0524	.0595	.1218	.0837	.0751	.0632
4	.1499	.1129	.1138	.1428	.1238	.0634	.0729	.1511	.1038	.0932	.0788
5	.1309	.0980	.0994	.1238	.1088	.0552	.0631	.1314	.0905	.0814	.0692
6	.0681	.0520	.0524	.0634	.0552	.0305	.0341	.0685	.0470	.0421	.0351
7	.0773	.0591	.0595	.0729	.0631	.0341	.0390	.0784	.0537	.0481	.0404

8	.1595	.1207	.1218	.1511	.1314	.0685	.0784	.1619	.1110	.0995	.0840
9	.1096	.0826	.0837	.1038	.0905	.0470	.0537	.1110	.0764	.0684	.0578
10	.0984	.0741	.0751	.0932	.0814	.0421	.0481	.0995	.0684	.0615	.0520
11	.0830	.0622	.0632	.0788	.0692	.0351	.0404	.0840	.0578	.0520	.0445

One latent root is lost because of the effects of the contingency table and the effect of scaling (see also principal coordinate analysis).

No.	Latent root	Percent	Cumulative percent
2	.42829E-02	47.877	47.877
3	.18375E-02	20.540	68.417
4	.13171E-02	14.723	83.140
5	.46818E-03	5.234	88.373
6	.41020E-03	4.586	92.959
7	.30684E-03	3.430	96.389
8	.14440E-03	1.614	98.003
9	.10267E-03	1.148	99.151

variable projections

1	-.000	-.071	.004	-.019	-.006	.020	-.006	.002
2	.080	-.022	-.036	.023	-.034	-.023	.006	.008
3	.047	-.013	.026	.002	.014	-.019	.000	-.002
4	-.055	-.002	-.066	.008	.021	.004	.009	-.004
5	-.085	-.027	.042	.022	.003	-.017	-.009	-.002
6	.201	-.015	.059	-.004	.025	.014	.037	-.019
7	.136	.055	-.007	.048	.033	.031	-.033	.008
8	.004	.053	-.010	-.020	-.022	.003	-.006	-.005

9	-.009	.043	.018	-.033	.026	-.015	.005	.023
10	-.027	.033	.023	-.004	.003	-.010	-.004	-.022
11	-.092	.050	.059	.038	-.025	.038	.024	.011

sample projections

1	-.092	-.017	.018	-.004	-.003	-.030	-.001	-.007
2	-.085	-.017	.029	-.008	-.008	-.024	-.006	-.001
3	.015	-.030	-.010	.019	.027	-.002	-.009	.007
4	.007	-.034	.023	-.007	-.021	-.017	-.007	-.012
5	-.016	-.003	-.025	-.026	-.030	-.015	.012	.011
6	-.026	-.028	-.005	-.000	-.006	-.016	.025	-.012
7	.004	-.015	-.081	.004	.013	-.001	.002	.018
8	-.060	-.056	.023	-.034	.020	.027	-.009	.015
9	-.019	-.031	-.019	-.027	.012	-.007	-.010	.000
10	-.014	-.016	-.008	-.029	.013	-.004	-.001	-.010
11	-.015	-.034	-.019	.010	.009	-.008	.008	-.014
12	-.011	-.017	.087	.041	-.046	.017	.013	-.005
13	.010	.019	-.007	-.029	-.044	.001	-.012	-.006
14	-.012	-.056	-.077	.037	.014	-.002	.015	-.006
15	.091	.053	-.062	-.015	.003	.037	.009	-.014
16	.026	.005	.001	-.005	.005	-.012	-.000	-.010
17	.012	-.003	-.006	.030	.013	.009	.010	.008
18	-.080	-.064	.030	-.033	-.002	.010	.004	.009
19	.162	-.073	.025	.020	-.051	.009	-.009	.006
20	.083	-.009	-.103	-.009	-.032	-.015	-.010	.002
21	-.058	-.038	-.013	-.006	.014	.006	-.001	-.015
22	.001	.043	-.021	.018	-.015	.009	-.014	-.003
23	-.016	.064	.017	-.035	-.003	.019	.007	-.002
24	-.009	.077	.019	.003	.014	-.008	.012	-.008
25	.173	.006	.049	.001	.035	-.043	-.000	.001
26	.100	.055	.028	-.006	.010	-.020	.017	.024

27	.116	-.078	.049	-.013	.034	.021	.002	-.007
28	-.005	.027	.009	.035	.017	.018	-.017	-.005
29	-.026	.090	.012	-.002	.000	.006	.011	-.014
30	-.006	.023	.009	.029	.011	.011	-.027	-.007
31	-.003	.029	.015	-.025	.004	.024	.001	.008
32	.009	.054	.010	-.009	-.010	.004	-.009	.004
33	.019	.054	.006	.010	.003	.003	-.003	.014
34	-.110	.043	-.000	.019	-.001	-.032	-.010	.009
35	-.087	-.020	-.006	.030	-.007	.020	.026	.014
36	-.078	-.002	.005	.015	.007	.004	-.020	.006

In **Fig. 2**, the crosses denote specimens, the numbers represent the variables.

Interpretation of the Correspondence Analysis

The mode of interpretation of correspondence analysis is entirely graphical in that the reason for doing the analysis is to

(a) Look for clustering in the data;
(b) Look for a relationship between the variables and the data-points.

It is this pairing of aims - variables coupled to observations - that justifies the designation - *Q*-*R*-method.

The Biplot

Although the biplot of Gabriel (1971) may seem to be just a simple variant of correspondence analysis due to the use made in both cases of the singular value decomposition, it is, in effect, not. Biplots provide plots of the n observations as well as the relative positions of the p variables, superimposed in the same figure. As originally conceived (hence the name) the technique was meant to be applied to just a single two-dimensional plot, under the assumption that the data matrix had rank 2 (or almost rank 2). However, the extension to further dimensions was quickly made by the innovator of the method. The usefulness of the original concept of the biplot is that the results are easy to interpret.

The basic concept in the formulation of the biplot is the fact that the decomposition of the (rectangular) data-matrix can be done in two ways. Firstly, there is the decomposition by means of the singular value decomposition

$$\mathbf{X} = \mathbf{VGU}^{\mathrm{T}}.$$

The second way is by the breakdown into a non-unique product of two matrices

$$\mathbf{X} = \mathbf{GH}^{\mathrm{T}} \qquad (16)$$

where $\mathbf{G}$ is $N \times r$ and $\mathbf{H}$ is $p \times r$. Here, N is the sample size (the number of rows in the data-matrix, p is the number of variables, and r is the rank of the data-matrix.

In computing the biplot, it is convenient to "standardize" $\mathbf{X}$ so that the mean on each variable is 0.

If the rank of $\mathbf{X}$ is exactly 2, the vectors constituting $\mathbf{G}$ and $\mathbf{H}$, $\mathbf{g}_i$ and $\mathbf{h}_j$, respectively, yield the biplot. If $r > 2$, the singular value decomposition can be invoked to construct a matrix $\mathbf{X}_2$ of rank 2 which is the best approximation to $\mathbf{X}$ in the sense that the sum of squares of the

elements of ($\mathbf{X}$ - $\mathbf{X}_2$) is a minimum. The indeterminancy in $\mathbf{G}$ and $\mathbf{H}$ can be removed by replacing each of the vectors forming $\mathbf{G}$ by a point located at the end of them. A further manipulation can be made which ensures that the distance between the i-th and j-th points in the geometrical representation is equal to the "distance" between the objects in the i-th and j-th rows of $\mathbf{X}$. This is done by defining $\mathbf{G}$ as

$$\mathbf{G}\mathbf{G}^{T} = \mathbf{X}_2\mathbf{M}\mathbf{X}_2^{T}$$

where the matrix $\mathbf{M}$ specifies the metric used on the rows of $\mathbf{X}_2$.

If $\mathbf{M} = \mathbf{I}_2$, the identity matrix, the Euclidean distance between the points at the end of each pair of $\mathbf{g}$-vectors corresponds to the Euclidean distance.

If $\mathbf{M} = \mathbf{S}^{-1}{}_2$, the inverse of the sample covariance matrix for the reduced data-matrix $\mathbf{X}_2$, then the Euclidean distance between the points of each pair of $\mathbf{g}$-vectors corresponds to the Mahalanobis generalized statistical distance between the corresponding objects in $\mathbf{X}_2$.

These two special properties of Gabriel's biplot make the technique attractive in theoretical developments, for example, in image-analytical studies of shape-variation, and other developments in multivariate statistics.

The variance of the k-th variable is given by the length of $\mathbf{h}_k$ and the correlation between any two variables is expressed by the cosine of the angle between them. More detailed discussion of the biplot is to be found in Gordon (1981), Jackson (1991) and Jolliffe (1986).

Comparing two samples: the Discriminant Function

One of the earliest multivariate statistical methods to appear was the Linear Discriminant Function, devised by R. A. Fisher to provide a quantitative expression of the multivariate difference residing in samples from two populations. The original proposal was made in reference to a taxonomic problem, the celebrated case of the species of *Iris*. The method was, however, quickly adopted for a wide range of other classes of data. Actually, Fisher got ahead of himself by helping an Australian doctoral candidate, Miss Barnard, analyze her Egyptian skulls in 1935 by his newly conceived method. This application, though biologically very unsound, is interesting for people concerned with linking time-differentiation to discrimination.

The linear discriminant function is computed from the covariance matrix formed by pooling the covariance matrices of each of the samples to yield matrix $\mathbf{S}_w$ and the difference **d** between the respective mean vectors. The equation for finding the *p* coefficients **f** of the linear discriminant function is simply:

$$\mathbf{f} = \mathbf{S}_w^{-1}\mathbf{d} \qquad (17)$$

==========

Box 8. Example of the discriminant function.

Example using data on *Leptocythere psammophila* - two stations (Reyment, 1996): the

observations consist of chemical determinations on the shell of a species of ostracods from northern Europe. The elements assessed are Ca Ba Cl S Sr Fe Mn Na Mg Al Si P

The discriminant function coefficients

	raw disc coeff.	standardized coefficients	Campbell's adjustment
1	33.028	15.690	7.443
2	566.635	1.559	127.692
3	41.807	4.463	9.421
4	162.117	6.391	36.533
5	-358.443	-4.264	-80.775
6	50.307	.885	11.337
7	155.380	1.508	35.015
8	23.807	5.065	5.365
9	7.797	1.571	1.757
10	256.883	1.985	57.889
11	-45.822	-2.197	-10.326
12	62.349	4.066	14.050

Most of the separation is being achieved by Ba, Sr and Al.

D**2 = 19.6917; D = 4.4375

F = 6.6270 for 12 and 15 degrees of freedom

Probability that the samples are from the same normal = .00054

```
Probability of misidentification =  .01325

     Analysis of discriminant scores

Sample 1
    Mean =   1622.5060 st.dev. =    3.8125
  14 correct   0 wrong for      .0 % wrong


Sample  2


    Mean =   1602.8150 St.dev. =   4.9698

  14 correct   0 wrong for   0 % wrong

Total N =  28  wrong =    0 % wrong =   0
```

Discussion

The discriminant analysis indicates that the two samples are highly significantly different and hence that the environmental conditions pertaining at the two localities were really distinct at the time at which the carapaces were secreted. The discriminant function coefficients suggest that variables 2(Ba), 5 (Sr) and 10 (Al) are responsible for most of the discriminating ability of the function which does not seem unreasonable. This example is worked in greater detail in a subsequent section.

==========

Box 9. *Example to exemplify quadratic discrimination*.

Terebratella retusa versus *T. septentrionalis*

The traits measured are length, height and width of the shell, two lengths on the foramen, and the weight of the shell, six characters in all (Endo *et al.*, 1995). Firstly a standard linear discriminant function is computed then a quadratic discriminant function, which is compared with the linear discriminant function.

D^2 = 5.3058, D = 2.3034

F = 21.3798 with 7 and 111 degrees of freedom.

The probability that the samples are not from same distribution > .00001

It is therefore quite clear that the two species are very different on the grounds of the characters determined. However, the probability of wrongly assigning individual specimens is quite high.

probability of misidentification = .12472 (based on ½D)

Dsq = 4.7076, D = 2.1697

misidentification probability = .1390

Results for Sample 1

mean = -15.3822 standard deviation = 2.1939

51 correct 7 wrong for 12.1 % wrong

Results for Sample 2

mean = -20.6880 standard deviation = 2.4028

54 correct 7 wrong for 11.5 % wrong

Quadratic discriminant analysis of the same data set as above

homogeneity test for matrices $\mathbf{S}_1$ and $\mathbf{S}_2$

chi-square = 79.2473 for 56 degrees of freedom

This value is not significant and it may be concluded that there is no significant heterogeneity in the covariance matrices.

sample 1 10.34 percent wrong of 58 specimens
sample 2   8.20 percent wrong of 61 specimens

==========

Discussion of the Quadratic Discrimination example

The quadratic discriminant classification in **Box 10** has produced a slight though important improvement in the efficiency of the discrimination. Granted that the covariance matrices are not obviously heterogeneous, the improvement could possibly be due to non-linearity in the data. The method of quadratic discrimination provides a means of proceeding when the covariance matrices are heterogeneous. It is only safe to use it when large samples are available, at least 50 specimens in each of both groups. If one variate has a very small variance and, or, several variates are highly correlated, spurious discrimination can be the result.

The Generalized Distance and Heterogeneous Dispersions

==========

Box 11. Example of discrimination and generalized distances for unequal covariance matrices. First Part of the Analysis.

T. retusa versus *T. septentrionalis*

Second suite of analyses for these data

$N_1 = 58, \quad N_2 = 61$

First sample

No significant values for the coefficients of skewness and kurtosis were obtained, which indicates that as far as the univariate distributions are concerned the data conform with the normal distribution.

Latent roots

.05114	.00663	.00376	.00152	.00057	.00031	.00025

Latent vectors

	1	2	3	4	5	6	7
1	.25724	-.17411	.13885	.40385	.33853	-.65068	.42794
2	.22318	-.34201	.15483	.72411	-.27510	.44225	-.11679
3	.33089	-.12842	-.14894	-.07739	.81523	.35938	-.22825
4	.04302	.85663	-.23876	.44314	.09012	.05223	-.00994
5	.28693	.20422	.61753	-.06436	-.03061	-.28546	-.63881
6	.28661	.24636	.54818	-.24000	-.01038	.39511	.58554
7	.77988	-.00312	-.44240	-.22016	-.36872	-.10747	.00950

Second sample

The second sample is less tidy than the first and almost all variables are not only significantly skewed but also have high kurtosis values.

Latent roots

.05730 .00535 .00149 .00064 .00042 .00034 .00020

Latent vectors

	1	2	3	4	5	6	7
1	.24886	.20172	-.02147	.53901	.15325	-.45933	-.60986
2	.27090	.09383	.01229	.65785	-.45424	.45685	.26430
3	.30046	.08469	.11265	.14140	.84641	.28226	.27171
4	-.06061	.60490	.76981	-.13022	-.11490	-.06779	.05536
5	.18676	.48727	-.46737	-.06661	-.05694	-.46699	.53241
6	.19830	.49730	-.36345	-.34925	-.04068	.50978	-.44467
7	.83458	-.30442	.20870	-.33665	-.18887	-.13445	-.01126

The two quantities now listed are for the non-central chi-squared distribution, for which you need a special table. It is, however, easy to transform to a standard chi-square value.

B^2 = 83.4497800

ß^2 = 1.7665050

For 28 degrees of freedom (chi-square is 41.33). The chi-square transformation of B^2 is 78.19, hence there is evidence of heterogeneity in covariances in respect of this test.

Orientation of Dispersion Ellipsoids

As noted, the foregoing result points to there being heterogeneity in covariances. We shall now see whether this can be ascribed to differing orientations of the hyperellipsoidal axes. Note, that in the following, a $?^2_6$ value of 12.59 indicates significance at the 5% level. There are only three axes of interesting length.

vector	chi-square
1	49.566
2	176.858
3	64.016

These values indicate that these axes are significantly rotated in relation to each other.

==========

Discussion of first part of the example for heterogeneous covariance matrices

The covariance matrices have been shown to differ in their geometrical properties, which means that the computations for producing a generalized distance and accompanying discriminant function should proceed via some appropriate method, such as the Anderson-Bahadur generalization of the Fisher-Behrens problem for heterogeneous variances in the analysis of variance (Anderson, 1984). It is a moot point as to whether this step can always be justified in practical situations since the improvement in the result is usually rather slight.

==========

Box 12. Example of discrimination and generalized distances for unequal covariance matrices. Second part of the analysis of the brachiopod data.

The method of estimation of the "heterogeneous" generalized statistical distance is iterative.

Five iterations were needed to arrive at the stable estimate of
D^2 = 7.05, which as F=27.33 on 7 and 111 degrees of freedom is highly significant.

Iterated estimate of the discriminant vector

Variable	Discriminant coefficient
1	57.03885
2	-10.16447
3	27.78253
4	20.33818
5	-54.21130
6	53.83771
7	-28.52486

Chernoff's separation T-criterion

In probing the properties of a generalized distance, it can be useful to determine how much of the distance is due to differences in the centroids and how much is to be put down to differences in the covariance matrices. A useful procedure is that of Chernoff (1973). Note, please, that this *T* is not the same as the generalized Student's *t* of Hotelling.

SEPARATION DUE TO MEANS = 1.7006110

SEPARATION DUE TO $\mathbf{S}^1$ P $\mathbf{S}_2$ = 1.9274760

CHERNOFF'S SEPARATION $T\_centroid^2$ = 3.6280870
$T\_covariance^2$ = 1.9047540

The Chernoff procedure suggests that about half of the dissimilarity between samples is due to the unequal covariances.

Discussion of the second part of the example

The intimation of heterogeneity in covariance matrices obtained earlier (the Box test is not a really good and reliable indicator) has been more completely exposed by a geometrically oriented analysis . There is no doubt in either case that the two samples differ highly significantly, but the reason for a large part of this is due to factors other than the difference in centroids.

The structural cause underlying the result seems to lie with the evidence for non-normality in the second sample exposed by the univariate analysis (skewness and kurtosis) which, in turn, could be ascribable to that sample actually being composed of more than one species (Endo *et al.*, 1995).

Analysis of Several Groups

Canonical Variate Analysis

If you want to analyze a several groups of observations, samples drawn from more than two statistical populations, then the method of canonical variates is an appropriate choice. In some respects, it can be thought of as being a generalization of the linear discriminant function, just reviewed, in others it can be thought of as a kind of generalized principal components. It has, of course, its deficiencies. These are mainly concerned with

1. The fact that the information supplied is in the form of *a priori* decided groups, which creates a tendency to reinforce the segregations already implied by the nature of the input. This is not a fatal fault and if there are really undecided specimens, they are efficiently disclosed.

2. It is a popular exercise to attempt to reify the canonical vectors. This can only done with extreme caution owing to the fact that the the components of the vectors tend to be unstable under repeated sampling. This condition is particularly noticeable when there are high between-groups correlations, such as occur in morphometrical work (Campbell, 1980).

In algebraic terms, the first canonical variate is that linear combination which maximizes the ratio of between groups sums of squares to the within-groups sums of squares for a one-way multivariate analysis of variance of the canonical variate scores. For k groups and p variables we have the canonical variate scores

$$y_{ij} = \mathbf{C}^{\mathrm{T}} x_{ij} \qquad (18)$$

where x_{ij} denotes the i-th of N observations for the j-th group. The first canonical vector is derived so as to maximize the ratio

$$f = \mathbf{c}^T\mathbf{B}\mathbf{c}/\mathbf{c}^T\mathbf{W}\mathbf{c}$$

where **B** is the between-groups matrix of sums of squares and cross products and **W** is the within-groups matrix of sums of squares and cross products. These matrices are formed as follows:

(1) Compute **T**, the matrix of sums and squares and cross products for all the groups pooled into a single data-matrix.

(2) Compute **W** by adding together the matris of sums of squares and cross products for each group.

(3) Find then **B** = **T** - **W**

The canonical vectors **c** and the canonical roots *f* satisfy equations (19) and (20)

$$(\mathbf{B} - f\mathbf{W})\mathbf{c} = \mathbf{0} \qquad (19)$$

and

$$^*\mathbf{B} - f\mathbf{W}^* = 0 \qquad (20)$$

The canonical vectors are usually scaled so that

$$\mathbf{c}^T\mathbf{W}\mathbf{c} = N_w \qquad (21)$$

where N_w is the within-groups degrees of freedom.

There are min(*k-1,p)* non-zero canonical roots to the solution of the determinantal equation (20). If there is closure in the data, there may be less than *p* non-zero roots when $p < k$. The

expression "min" says that whichever is smallest, *k+1* or *p*, decides the number of non+zero canonical roots.

The program for canonical variate analysis supplied on the diskette is an updated version of the one originally published in Blackith and Reyment (1971).

Canonical Variates for Compositional Data

An analysis for constrained data-sets follows on in the same manner as was done for principal component analysis. The log-ratio data-matrices are used as input in the place of the "usual" data-sets. Alternatively, you could use the within- and between-groups centred matrices as input. There is an alternative read-in available for doing this. Let us look at an example of canonical variate analysis (**Box 13)**. I have used compositional data on alkaline rocks sampled from mid-oceanic ridges. The specimens come from the North, Middle and South Atlantic ridges and a Pacific Ocean ridge (Sorensen, 1974). The oxides of the elements determined are of Si(1), Ti(2), Al(3), Fe^{+++}(4), Fe^{++}(5), Mn(6), Mg(7), Ca(8), Na(9), K(10), water(11), and P(12), in that order. The analysis was made on the raw data, that is, without accounting for the constraint.

==========

Box 13. An example of canonical variate analysis: **The raw alkaline rock data**

There are four groups and 12 variables, listed above.

Designation	Locality	Observations
A	The Canary Islands	15
B	The Pacific islands	15
C	The northern South Atlantic	15
D	The southern South Atlantic	14

** Sample Sizes and Mean Vectors **

1	15.	47.721	2.637	14.509	3.346	6.267	.194	7.993	8.419	4.448	2.076	1.829	.481
2	15.	47.841	2.918	15.042	3.093	8.234	.170	7.170	8.745	3.669	1.537	.978	.559
3	15.	44.012	2.297	14.938	3.845	6.224	.171	7.603	8.543	5.442	3.117	2.488	.586
4	14.	52.469	2.232	17.144	2.727	5.834	.176	3.734	6.447	4.615	3.103	1.119	.398

ANOVA

Variable	Among Msq	Within Msq	F	Prob
1	173.169	34.096	5.08	.0037
2	1.496	1.550	.96	.5826
3	19.864	14.538	1.37	.2619
4	3.214	1.609	2.00	.1235
5	17.293	7.315	2.36	.0798
6	.002	.003	.67	.5752
7	54.566	26.372	2.07	.1134
8	16.298	12.071	1.35	.2669

9	7.933	5.158	1.54	.2138
10	9.075	2.360	3.85	.0143
11	7.216	1.681	4.29	.0087
12	.104	.097	1.07	.3695

Canonical Root	1	3.0213240
Canonical Root	2	.5753111
Canonical Root	3	.2254926

Canonical Vectors of $\mathbf{W^{-1}B}$

	1	2	3
1	-1.245	-1.114	.165
2	-.373	-1.088	-.195
3	.036	-.712	-.077
4	.050	-.256	.213
5	-.805	1.498	-1.206
6	.200	-.675	.264
7	.842	-1.971	-.023
8	1.338	-.213	.809
9	1.454	-.825	-.692
10	.730	1.402	.720
11	.486	-.219	-.470
12	-.165	.234	-.338

Normalized Canonical Vectors

	1	2	3
1	-.446	-.317	.087
2	-.134	-.309	-.102
3	.013	-.202	-.041
4	.018	-.073	.112
5	-.288	.426	-.633
6	.071	-.192	.139
7	.302	-.560	-.012
8	.479	-.060	.425
9	.521	-.235	-.363
10	.262	.398	.378
11	.174	-.062	-.247
12	-.059	.067	-.177

rank of problem = 3

Canonical variate mean No. 1

1.79782 -3.04565 .45840

Canonical variate mean No. 2

-2.24667 .80927 -1.24249

Canonical variate mean No. 3

8.95587 2.23618 -.14291

Canonical variate mean No. 4

-9.11468 .00021 .99321

Dsquare above diagonal, D below diagonal

	1	2	3	4

1	.000	34.111	79.497	128.646
2	5.840	.000	128.742	52.822
3	8.916	11.346	.000	332.835
4	11.342	7.268	18.244	.000

F-ratio for D-square and probabilities (below diagonal)

	1	2	3	4
1	.000	17.056	39.749	62.105
2	.000	.000	64.371	25.501
3	.000	.000	.000	160.679
4	.000	.000	.000	.000

Significance of latent roots

Root No 1 = 2.0494070
chi-square = 102.4703000 on 36 degrees of freedom
Prob = .00

Root No 2 = .6577957
chi-square = 32.8897900 on 22 degrees of freedom
Prob = .06

Root No 3 = .2033429
chi-square = 10.1671400 on 10 degrees of freedom
Prob = .43

```
          test for equality of mean vectors

 Wilks Lambda =     0.12881

 chi-square =       79.927 on 36 degrees of freedom
      Prob =    .00

       Test of homogeneity of covariance matrices

  chi-square =     769.081100 on 234 degrees of freedom
      Prob =   .000
```

==========

Discussion of the canonical variate example

The ANOVA, the one-way analysis of variance for each variable on four samples, shows that only the means of three components differ significantly, to wit, Si (1), K (10) and H_2O (11).

The first canonical root dominates completely in magnitude over the other two (note that the rank of the problem is 3, one less than the number of groups and there are, therefore, only three valid canonical roots). If we confine attention to the first canonical vector, it will be seen that components (1), (5), (7), (8), (9), and (10) are influential. The rule of thumb I use here (cf. Reyment and Jöreskog, 1993) is that the vector component, when squared, should be 0.1, or greater. (N.B. the elements of a vector are referred to as its "components".)

The array of generalized statistical distances, D^2, and the accompanying array of variance ratios and probabilities, indicate that all samples are highly significantly different from each other. This is no doubt of considerable petrological interest.

In our present example, there are two roots that can be taken to be statistically significant. The MANOVA (multivariate one-way analysis of variance) indicates all mean vectors to differ highly significantly. Secondly, the test for homogeneity of covariance matrices shows them to be very different indeed. This latter result is a very common one and one that can have a negative effect on the outcome of an analysis. Notwithstanding the fact that canonical variate analysis has proven to be quite robust to deviations from normality and homogeneity of variances and covariances, there is a bound beyond which it would be unwise to pursue a study. By this I mean, an analysis of data that contain many strongly atypical values, for example, may not yield a scientifically useful result. For this reason, considerable attention has been given recently to attempting to find robust methods of estimation, that do not interfere too much with the integrity of the data. One of the leading exponents of the practical aspects of this kind of work is N. A. Campbell. Some references are Campbell (1980, 1984). I shall now discuss results for the application of canonical variates to a research problem in environmental work, some facets of which were presented in a foregoing section.

A Worked Example: A Study in Environmental Chemistry

Chemical composition of the shell of the ostracod Leptocythere psammophila

Bodergat, Carbonnel, Rio and Keyser (1993) studied the influence of environmental chemistry

on the composition of the shell secreted by *Leptocythere psammophila*. To this end, 41 live individuals were collected from the Baltic Sea (Kieler Förde), North Sea (Sahlenburg/Cuxhaven) and the English Channel (Roscoff). The Baltic locality is one complicated by the effects of industrial pollution. The sampling was carried out in Spring (April, 1987), Summer (September, 1988) and Winter (December, 1988). Unfortunately, the sample sizes are very small.

Thirteen elements were determined by electron microprobe analysis, to wit, Ca, Ba, Cl, S, Sr, Fe, Mn, Na, Mg, Al, Si, P, and O. Thirty analyses were made on each carapace at points situated according to a predetermined pattern. The results for each individual (being the average of 30 values for each element) were expressed as "element atomic percent." This is a vital piece of information which you need to keep in mind for what now follows. The main findings reported by Bodergat *et al.* (1993) were:

(a) There are no significant differences between chemical compositions of carapaces from the different stations.

(b) There are strongly manifested seasonal differences due, it was surmised, to the facilitated incorporation of Mg during the winter.

(c) The chemical composition of summer individuals was interpreted as being due to the influence of salinity fluctuations on the shell and to the supply of terrigenous sediment.

The data were analyzed by multivariate statistical methods, namely, discriminant functions and "normalized principal component analysis". The latter procedure does not seem to be principal component analysis, as is usually conceived, but rather, some kind of Q-R-mode application of latent roots and vectors. A quick appraisal of the results shows, moreover, that an inappropriate model for covariances was used with the consequence that the interpretations arrived at by Bodergat *et al.* (1993) are not unchallengeable.

The Data

The data are expressed as percentages. The reason why the data to not sum to the expected 100% is that they have been "doctored". There are 12 entries but 13 elements were actually determined. The missing entries are for oxygen and, possibly, carbon. Now that we have established the fact that the data are compositional, the next step is to establish rules for an appropriate multivariate model. This is where the analysis published by Bodergat *et al.* (1993) goes awry insofar as the vital consideration of closure does not seem to have been taken into account. Eliminating variables from a closed set does not help things, since compositional data are related in a manner which differs from that for "free" variables.

Principal Component Analyses

One of the problems bothering the original analysts was the difference in variability from variable to variable. The following computations were therefore made on the correlation matrix. We considered first the constrained model for the Spring data.

THE COVARIANCE MATRIX OF PROPORTIONS

OBSERVATIONS = 13 VARIABLES = 12

SIMPLEX CORRELATION MATRIX

This is the correlation matrix for the log-ratio method of Aitchison (1986).

	1	2	3	4	5	6	7	8	9	10
1	1.00000	.82711	-.56287	.36370	.82283	-.00127	-.18821	-.08733	.84578	-.56105
2	.82711	1.00000	-.41570	.24279	.72841	.16998	-.09945	.10689	.77737	-.67905
3	-.56287	-.41570	1.00000	.02460	-.55346	.16786	-.33877	.08770	-.56969	.15974
4	.36370	.24279	.02460	1.00000	-.05858	-.01237	-.05391	-.62764	.46888	-.15361
5	.82283	.72841	-.55346	-.05858	1.00000	.09116	-.27487	.12897	.65443	-.62325
6	-.00127	.16998	.16786	-.01237	.09116	1.00000	-.19714	-.16961	.11210	-.49993
7	-.18821	-.09945	-.33877	-.05391	-.27487	-.19714	1.00000	-.29800	-.13110	.10428
8	-.08733	.10689	.08770	-.62764	.12897	-.16961	-.29800	1.00000	-.16188	.01177
9	.84578	.77737	-.56969	.46888	.65443	.11210	-.13110	-.16188	1.00000	-.35535
10	-.56105	-.67905	.15974	-.15361	-.62325	-.49993	.10428	.01177	-.35535	1.00000
11	-.71703	-.79242	.21165	-.48079	-.41611	-.13811	.06426	.04787	-.84816	.32369
12	.78659	.63604	-.30322	.67925	.37678	.15109	-.06255	-.29344	.68766	-.55313

	11	12
1	-.71703	.78659
2	-.79242	.63604
3	.21165	-.30322
4	-.48079	.67925
5	-.41611	.37678
6	-.13811	.15109
7	.06426	-.06255
8	.04787	-.29344
9	-.84816	.68766
10	.32369	-.55313
11	1.00000	-.69994
12	-.69994	1.00000

USUAL (= raw) CORRELATION MATRIX

This is the correlation matrix obtained when the effects of the constraint are ignored.

	1	2	3	4	5	6	7	8	9	10
1	1.00000	.12433	-.68515	-.62422	.22566	-.50458	-.42289	-.66526	-.34475	-.83576
2	.12433	1.00000	-.21005	-.17241	.17465	.05163	.02032	-.01767	.21870	-.28906
3	-.68515	-.21005	1.00000	.50149	-.10343	.54126	.19069	.70312	-.33893	.47457
4	-.62422	-.17241	.50149	1.00000	-.67134	.32008	.49811	.07056	.20636	.47853
5	.22566	.17465	-.10343	-.67134	1.00000	-.01569	-.31001	.11642	-.17192	-.31454
6	-.50458	.05163	.54126	.32008	-.01569	1.00000	.05636	.23136	.10613	.11405
7	-.42289	.02032	.19069	.49811	-.31001	.05636	1.00000	.22314	-.00752	.43141
8	-.66526	-.01767	.70312	.07056	.11642	.23136	.22314	1.00000	-.21956	.63426
9	-.34475	.21870	-.33893	.20636	-.17192	.10613	-.00752	-.21956	1.00000	.31662
10	-.83576	-.28906	.47457	.47853	-.31454	.11405	.43141	.63426	.31662	1.00000
11	-.51704	-.39293	.86422	.31284	.07708	.22454	.20329	.71465	-.54386	.42347
12	-.64471	-.18001	.62240	.84361	-.66929	.54225	.53789	.29127	-.03007	.41941

	11	12
1	-.51704	-.64471
2	-.39293	-.18001
3	.86422	.62240
4	.31284	.84361
5	.07708	-.66929
6	.22454	.54225
7	.20329	.53789
8	.71465	.29127
9	-.54386	-.03007

10	.42347	.41941
11	1.00000	.42937
12	.42937	1.00000

Comparison of the above two correlation matrices is distressingly informative. Leaving possible effects of deviations from the Gaussian condition aside, it will be seen that almost all pairs of correlation coefficients differ strongly.

The log-ratio transformed data

latent roots

5.47058 1.93933 1.60300 1.15471 .71168 .57020 .29458 .16102 .07469 .01487
.00534

latent vectors

	1	2	3	4	5	6	7	8	9	10
1	.40614	-.06614	.10312	.09289	-.17667	.11051	.00725	.41703	-.12384	.29717
2	.38510	-.16283	-.00639	-.00982	.29110	.07343	-.27674	-.30185	-.69693	-.09298
3	-.22007	.08412	-.55893	.26318	.23596	.21639	-.44218	.29850	.15282	-.18952
4	.19997	.56158	-.13943	.23220	-.10786	.17762	.00823	-.52712	.18048	.43034
5	.32422	-.36225	.08414	-.09055	-.36666	.02361	-.36844	.13378	.34175	.22301
6	.08124	-.03148	-.56010	-.50552	.08327	-.50053	.22699	.01740	-.01625	.25464
7	-.05869	.25699	.47629	-.49807	.50412	.13808	-.13810	.09651	.23162	.08195
8	-.06288	-.59251	.01179	.30488	.41028	.13911	.39264	-.22342	.25933	.20855
9	.38964	.04876	.10544	.13777	-.00217	-.42373	-.00741	-.21144	.36176	-.60701

10	-.28033	.16304	.29867	.42574	-.01983	-.51617	.05246	.22857	-.23263	.19445
11	-.34947	-.09696	.03804	-.24632	-.49880	.27969	.18502	-.21151	-.13405	-.24220
12	.35591	.24285	-.08903	.04912	.06075	.30051	.57465	.38587	-.04600	-.22852

	11
1	.67922
2	-.13656
3	.06291
4	-.00936
5	-.49645
6	.02306
7	.03698
8	.05555
9	.21234
10	-.21177
11	.12220
12	-.39955

Principal component scores for specimens

1	-1.8938	.7073	.2026	.8567
2	-.5720	.7241	.2785	1.1127
3	-.6945	.9316	.5132	1.0574
4	-.6901	.7667	-.2914	.8375
5	-.3867	.7720	.2498	.8836
6	-.6847	.4514	.5010	.6062
7	-.8787	1.1659	.4540	.6244
8	-1.2145	.9560	.5674	1.1415

9	-.6712	.8023	.2706	.7700
10	-.8408	.8473	.4846	.6117
11	-.7291	.5552	.2933	1.0037
12	-.7997	.6228	.4985	.6155
13	-.6865	.6989	.3370	1.0061

ORDINARY PRINCIPAL COMPONENT VALUES

latent roots

5.06126 2.26917 1.46375 1.20591 .95820 .53203 .33083 .09519 .03625 .03161
.01243 .00337

latent vectors

	1	2	3	4	5	6	7	8	9	10
1	-.39628	.06296	-.33267	.08180	.09966	.06387	-.08853	.26208	.20361	.12836
2	-.12242	-.11203	.43486	.37890	.59658	.41337	.18158	.10690	.10663	.19004
3	.37473	.27772	-.01697	.20204	-.06970	.17398	.20369	.41330	-.54641	-.28435
4	.33525	-.32443	-.19909	.09956	-.01664	.04084	.57979	.06649	.37693	-.27705
5	-.18402	.43630	.37080	.04343	.01280	-.51250	.37489	-.12491	.27836	-.25028
6	.22132	.00120	.21892	.64638	-.27474	-.32188	-.35880	.23989	.15119	.18753
7	.23361	-.19163	-.08376	-.15344	.65563	-.58614	-.10299	.10688	-.24986	.03576
8	.29156	.34941	.29873	-.16847	.18148	.27839	-.34977	-.29918	.08916	-.28388
9	-.00868	-.47747	.49790	-.12894	-.28470	-.05024	.17280	-.21280	-.31847	.19981
10	.33532	-.07553	.24259	-.48760	-.07603	.05143	-.13388	.49891	.43804	.14405

11	.31539	.42391	-.14399	-.05476	.00433	.01328	.32037	-.18628	-.01438	.73945
12	.37516	-.19637	-.22615	.26460	.06122	.02094	-.16998	-.49160	.20565	-.05161

	11	12
1	.38854	.64977
2	-.14326	-.02467
3	-.11491	.31826
4	.39413	-.12691
5	-.15197	.23283
6	.21748	-.12077
7	.13497	.02792
8	.51092	.05856
9	.24246	.38543
10	-.25916	.17378
11	.11192	-.00723
12	-.41454	.45605

The disparities in the correlation coefficients carry over to the latent roots and vectors of the correlation matrix. You can check this by examining the two sets of output above. However, the principal component differences are not so extreme as are those for the bivariate correlations. Likewise for the summer and winter samples.

Bodergat *et al.* (1993) attempted to ordinate the 41 specimens by the principal component scores of the three samples pooled. This is inappropriate for the problem at hand given that in addition to the difficulty implied by closure, the data are not homogeneous with respect to variances and covariances; hence, any form of "common principal component analysis" is unsuitable (Flury, 1988). The appropriate procedure is then to proceed by means of a log-ratio

canonical variate analysis, using the results of Aitchison (1986), appropriately formulated.

1. In both cases, the results yielded by the correlations are not as decisive as those obtained from the covariances. The reason for this lies with the "levelling" effect produced by the log-ratio transformation with the result that taking correlations serves to lessen further the intrinsic unlikenesses between samples.

2. Look now at the generalized statistical distances. These all yield significant values of T^2, thus indicating that there are genuine differences between samples for both arrangements of the data; that is, for both seasons as well as sites. This is formally confirmed by the results for the multivariate analysis of variance (listed in the output as "test of equality of means").

3. The plots of the canonical variate scores (note, that there are only two canonical vectors, since there are only three groups) show excellent group-ordination. In the case of the seasons, there is no overlap between groups, whereas for the sites, there is slight overlap between the North Sea and the English Channel. Note, that this aspect of canonical variate analysis is, in effect, a multiple discriminant function analysis.

4. The pooled principal component analyses for the correlations and covariances and for both groupings of the data, do not yield an informative result. The points for the different categories merge over the whole field of the graph.

5. One of the points made by the original analysts was that the chemical composition of the shell is influenced by salinity (Na, Cl), terrigenous components (Si, Al, Fe) and the metabolic role of Mg. This may well be true, but the segregation into functional components that should be apparent in the appropriately formulated chemometric approach does not appear in the clearcut manner claimed by Bodergat *et al.*(1993).

Conclusions

We are now in a position to examine the validity of the conclusions arrived at by the original investigators. The element Ca was excluded from the original work owing to its "dominance" and, presumably, because it was felt that eventual environmental differences would be more specifically addressed by the minor elements.

1. *Each season sets its stamp upon the shell-chemistry of the species.* This is born out by the present analysis, and even more persuasively than in the original study.

2. *There are no significant differences between the samples taken from the three localities.* This is false. The canonical variate analysis shows quite clearly that the three samples are indeed different. This is supported by the MANOVA and by the generalized statistical distances, all of which transform to highly significant values of T^2.

I have underscored the importance of selecting an apt statistical model for analyzing compositional data, and this has been indeed one of the main themes pervading this tract. What happens in the case of canonical variate analysis if one just stumbles ahead by way of the inappropriate procedure.

You will find that as before, the ordination for sites gives excellent separation, in fact, slightly better than before. Moreover, the generalized statistical distances are highly significantly different. The separation for seasons is also excellent. This brings out an important practical point, namely, the robustness of the ordination aspect of the method of canonical variates. No matter whether you use the right method or the "wrong" one, the relationships between individual points will not be much altered by an analyses concerned with but a few groups. If you are also interested in the indications provided by the canonical

vectors, then it is important to select the appropriate procedure. In the present example, the original authors used a low-grade ordinating procedure (and an inappropriate statistical model) which accounts for their failure to identify fully the differentiation so well developed in their material. The case history briefly reviewed here is an example of the rapidly growing field of geological *environometrics.*

Common Principal Component Analysis

One of the newer methods of multivariate analysis to appear is known as *Common Principal Component Analysis* (Flury, 1988). Canonical variate analysis has a logical insufficiency in that group differences are exaggerated as a result of the groups having been established *a priori*. For this reason, various attempts have been made at producing a valid analysis of data occurring in groups by some procedure that does not have to rely on the difference relationship **T** = **W** + **B**, such as is done in canonical variate analysis. Flury (1988) developed a method which goes part of the way. Although common principal components can be computed both for covariance and correlation matrices, the computations for large-sample standard errors apply only in the former case.

Basis of Common Principal Component Analysis

Let there be two samples of an object on which multivariate observations have been made. The question asked is whether a unique common transformation for both groups can be estimated? Slight differences in covariance matrices could well be due to sampling differences and hence lack analytical significance. The model for two samples can be expressed simply as

$$\mathbf{S}_1 = \mathbf{B?_1B}^T$$

$$\mathbf{S}_2 = \mathbf{B?_2B}^T$$

where the $\mathbf{?}_i$ are diagonal.

By a theorem for the simultaneous decomposition if two positive definite symmetric matrices (Bellman, 1960) there exists a non-singular *p* x *p* matrix **?** such that:

$$\mathbf{B}^T\mathbf{S}_1\mathbf{B} = \mathbf{I}_p$$

and

$$\mathbf{B}^T\mathbf{S}_2\mathbf{B} = \mathbf{?}$$

where **?** is a diagonal matrix. Hence

$$\mathbf{S}_1^{-1}\mathbf{B}\mathbf{S}_2\mathbf{B} = \mathbf{B?} \qquad (22)$$

That is, the columns of **B** are the latent vectors of the matrix product $\mathbf{S}_1^{-1}\mathbf{S}_2$ and the diagonal of **?** contains the corresponding latent roots. The statistical implication of the simultaneous decomposition theorem is that it provides a convenient vehicule for obtaining uncorrelated variables in two populations. Thus if

$$\mathbf{U} = \mathbf{B}^T\mathbf{X}$$

then the covariance matrix of **U** is $\mathbf{I}_p$ in the first population and **?** in the second population. This is *almost* but not quite **exactly** a generalization of principal component analysis to two

groups. The catch is that **B** is not exactly orthonormal and hence the definition of the principal component transformation as a rotation of the coordinate system does not in general apply to the multiple group situation. Experience shows that strong departures from orthogonality of latent vectors are frequent in both geological and biological work, an unfortunate state of affairs that can invalidate what is otherwise an appealing way of dealing with the analysis of multivariate multi-group data.

The solution for two groups extends simply to several groups. Thus:

$$\mathbf{S}_i = \mathbf{B}?_i\mathbf{B}^T \qquad (i=1,...,k) \qquad (23)$$

B is an "almost orthonormal" matrix and **?** is diagonal.

The *sample covariance matrix* of **U** in group *i* is

$$\mathbf{F}_i = \mathbf{B}^T_i\mathbf{S}\mathbf{B}_i$$

and (24)

$$?_i = \text{diag } \mathbf{F}_i$$

holds.

A useful way of judging the suitability of the common principal component model is to examine the corresponding sample correlation matrices obtained from (24). These correlations are expected to be close to nought, that is, $\mathbf{I}_p$. Marked deviations indicate that the model is inappropriate.

Flury (1988) has also indicated how to compute large-sample standard errors for the latent roots and latent vectors of common principal component analysis.

The standard error for latent roots is given as

$$s(?_{ij}) = (2/n_i?_{ij}) \qquad (25)$$

where the n_i denote individual sample sizes. There will be a set of standard errors for the latent roots computed for each sample. On the other hand, there is only one set of standard errors for the principal components **B,** as indicated by equation (23). The term *n* is the sum of sample sizes.

$$s(\hat{\beta}) = (1/n\sum \hat{\phi}\beta^2)^{1/2} \qquad (26)$$

which is the sum over *p* variables of the product of the harmonic mean of the product of latent roots divided by their differences and the square of the common principal components. Obviously, the matrix resulting from producing the harmonic means cannot be inverted by standard procedures.

Common principal component analysis is not without its weaknesses, far from it. It is necessary to keep a close eye on what is happening and to make sure that the model is really appropriate. Moreover, I have not found many cases in geological work which fit well the model, particular compositional data. The departures from orthogonality of the common latent vectors can be very strong. The following example treats measurements on the carapace of the ostracod species *Veenia rotunda* Reyment from the Cenomanian-Turonian boundary in Morocco. The seven variables are (1) length of carapace; (2) height of carapace; (3) distance from the anterior margin to the adductor tubercle; (4) distance from the dorsal margin to the adductor tubercle; (5) distance from the eye-tubercle to the posteroventral margin; (6) posterior height of the carapace; (7) length of the caudal process. The samples were taken from four consecutive levels

near Ait Brahim, south-central Morocco.

Example of common principal component analysis

Groups = 4; dimensions = 7

The input covariance matrix No. 1

1	.00606	.00128	.00058	.00217	.00227	.00278	.00087
2	.00128	.00065	.00029	.00051	.00064	.00053	.00036
3	.00058	.00029	.00037	.00013	.00026	.00036	.00016
4	.00217	.00051	.00013	.00094	.00088	.00100	.00038
5	.00227	.00064	.00026	.00088	.00110	.00100	.00041
6	.00278	.00053	.00036	.00100	.00100	.00173	.00029
7	.00087	.00036	.00016	.00038	.00041	.00029	.00088

Corresponding usual principal components

Standard latent roots

.009573 .000896 .000463 .000408 .000212 .000118 .000063

Standard latent vectors

1	2	3	4	5	6	7
.788477	-.138563	-.259294	-.047444	.520838	-.087814	-.103131
.181744	.372493	.390224	-.463466	-.187499	-.540966	-.365160
.084727	.137184	.706465	-.002009	.390508	.125030	.553874
.290689	.035466	-.271949	-.081647	-.486230	-.312655	.706716
.308660	.164999	.029058	-.426099	-.332035	.762022	-.064607
.376641	-.360840	.430926	.552235	-.440914	.008272	-.206885
.128315	.815195	-.144572	.538288	.001681	.074249	-.053127

The input covariance matrix No. 2

1	.00597	.00138	.00090	.00240	.00179	.00326	.00118
2	.00138	.00068	.00039	.00053	.00056	.00070	.00048
3	.00090	.00039	.00034	.00033	.00036	.00053	.00024
4	.00240	.00053	.00033	.00114	.00073	.00132	.00050
5	.00179	.00056	.00036	.00073	.00076	.00094	.00045
6	.00326	.00070	.00053	.00132	.00094	.00204	.00053
7	.00118	.00048	.00024	.00050	.00045	.00053	.00054

Corresponding usual principal components

Standard latent roots

.010054	.000686	.000301	.000142	.000136	.000094	.000058

Standard latent vectors

1	2	3	4	5	6	7
.765718	-.168439	-.193021	-.035265	-.549075	-.195034	.085334
.189504	.624156	.240869	.162952	.002439	-.337945	-.612971
.124804	.297434	.541539	.061829	.095760	-.268885	.719305
.313765	-.106129	-.402951	-.005317	.757696	-.389348	.046855
.238656	.320910	.030645	-.826945	.145313	.366188	-.008555
.425924	-.352543	.518972	.267470	.304560	.466230	-.208094
.159284	.505032	-.422870	.461490	.036951	.521919	.232407

The input covariance matrix No. 3

1	.00580	.00121	.00055	.00248	.00191	.00274	.00161
2	.00121	.00050	.00020	.00057	.00049	.00058	.00044
3	.00055	.00020	.00030	.00018	.00023	.00036	.00021
4	.00248	.00057	.00018	.00130	.00088	.00116	.00072
5	.00191	.00049	.00023	.00088	.00087	.00090	.00060
6	.00274	.00058	.00036	.00116	.00090	.00149	.00079
7	.00161	.00044	.00021	.00072	.00060	.00079	.00067

Corresponding usual principal components

Standard latent roots

.009673 .000388 .000332 .000168 .000149 .000130 .000070

Standard latent vectors

1	2	3	4	5	6	7
.768064	-.336741	.174161	-.085277	.378167	-.293875	-.172355
.172853	.527195	-.277702	.155750	-.084559	-.677727	.352628
.081057	.579649	.442195	-.185489	-.211509	-.046834	-.616900
.339337	-.185725	-.526903	.149123	-.654191	.099208	-.335684
.265261	.308979	-.361424	-.705419	.133991	.382992	.203194
.371267	.047267	.511583	.131633	-.440214	.292677	.549030
.224225	.374810	-.164444	.629986	.406689	.458197	-.099881

The input covariance matrix No. 4

1	.00495	.00094	.00044	.00217	.00169	.00247	.00079
2	.00094	.00033	.00016	.00044	.00043	.00045	.00022
3	.00044	.00016	.00026	.00014	.00023	.00031	.00010
4	.00217	.00044	.00014	.00111	.00077	.00107	.00037
5	.00169	.00043	.00023	.00077	.00075	.00083	.00032
6	.00247	.00045	.00031	.00107	.00083	.00142	.00036
7	.00079	.00022	.00010	.00037	.00032	.00036	.00035

Corresponding usual principal components

Standard latent roots

.008146 .000356 .000304 .000159 .000110 .000051 .000044

Standard latent vectors

1	2	3	4	5	6	7
.775156	-.230666	-.068953	.069126	-.535442	-.187491	-.120596
.155564	.486340	-.090622	-.230062	.075126	-.588535	.571065
.077981	.462720	.627809	-.009030	.043934	-.239010	-.571425
.346886	-.100544	-.412390	-.244770	.693158	-.114497	-.382107
.274649	.461856	-.049451	-.461233	-.111856	.692220	.066299

.395128	-.219280	.554320	.233740	.453310	.217362	.425617
.130248	.473976	-.336314	.784174	.085197	.147730	-.035549

The first thing to notice here is that the covariance matrices are not greatly different from each other and it is therefore likely that they are identical and that any observed divergencies are due to sampling variation alone. This impression is further reinforced by the results for the usual principal components. The latent values do not differ greatly from each other and the components of the latent vectors present the same general pattern. The next step is to examine the correlations of the common principal components, displayed below as matrices. The off-diagonal elements should be close to zero if the common principal component model is a good fit.

Correlation matrices R(i) of CPCs

Sample No. 1

1	1.000000000	-.013816638	-.201536909	-.205134351	.176210914	.023208754	-.261130007
2	-.013816638	1.000000000	.101123132	-.022258553	-.080370420	-.124305527	.235476991
3	-.201536909	.101123132	1.000000000	-.323164943	-.103712231	-.009430030	.078494920
4	-.205134351	-.022258553	-.323164943	1.000000000	.060034638	.129515391	.048824623
5	.176210914	-.080370420	-.103712231	.060034638	1.000000000	-.142510262	.068536173
6	.023208754	-.124305527	-.009430030	.129515391	-.142510262	1.000000000	-.062726452
7	-.261130007	.235476991	.078494920	.048824623	.068536173	-.062726452	1.000000000

Sample No. 2

1	1.000000000	-.051349909	.006776333	.055871877	-.248794935	.242391939	.096925805
2	-.051349909	1.000000000	-.110417252	-.087038724	-.109924750	.076447636	-.065739502
3	.006776333	-.110417252	1.000000000	.189406448	-.027450398	-.225860363	-.061832414
4	.055871877	-.087038724	.189406448	1.000000000	.047216229	.060266155	.056195017
5	-.248794935	-.109924750	-.027450398	.047216229	1.000000000	-.145074080	-.211244983
6	.242391939	.076447636	-.225860363	.060266155	-.145074080	1.000000000	-.130357416
7	.096925805	-.065739502	-.061832414	.056195017	-.211244983	-.130357416	1.000000000

Sample No. 3

1	1.000000000	.166277624	.079553878	.006061947	.017303939	-.215468949	.207227021
2	.166277624	1.000000000	-.065226492	.233644910	.058376757	.101594673	-.020369769
3	.079553878	-.065226492	1.000000000	.117973819	.089660236	.052634477	-.031756343
4	.006061947	.233644910	.117973819	1.000000000	-.026360802	-.145310973	-.112001102
5	.017303939	.058376757	.089660236	-.026360802	1.000000000	.017170231	-.092486452
6	-.215468949	.101594673	.052634477	-.145310973	.017170231	1.000000000	.162743772

7	**.207227021**	**-.020369769**	**-.031756343**	**-.112001102**	**-.092486452**	**.162743772**	**1.000000000**

Sample No. 4

1	**1.000000000**	**-.176036378**	**.120231541**	**.135203515**	**.075117060**	**-.082756184**	**-.233748865**
2	**-.176036378**	**1.000000000**	**.067446188**	**.023290864**	**.209797826**	**.048678518**	**.018910973**
3	**.120231541**	**.067446188**	**1.000000000**	**.134785265**	**.064199603**	**.180765757**	**-.098005761**
4	**.135203515**	**.023290864**	**.134785265**	**1.000000000**	**-.045795271**	**-.087574016**	**-.123667879**
5	**.075117060**	**.209797826**	**.064199603**	**-.045795271**	**1.000000000**	**.231376954**	**-.067857797**
6	**-.082756184**	**.048678518**	**.180765757**	**-.087574016**	**.231376954**	**1.000000000**	**.047844402**
7	**-.233748865**	**.018910973**	**-.098005761**	**-.123667879**	**-.067857797**	**.047844402**	**1.000000000**

This is a satisfactory result (even impressive for geological data), taken as a whole, and there is only one moderately large correlation, namely that for r_{23} in Sample 1. It may therefore be inferred that the common principal component model is a good one and that the four sample covariance matrices really are identical and that any differences are just the result of sampling variation.

Common PCA latent vectors columnwise

	1	2	3	4	5	6	7
1	.772119	-.192938	.053569	-.545353	-.160741	-.174246	-.100639
2	.175500	.548101	-.443522	.170109	-.571109	.192243	-.282867
3	.096079	.245895	.663188	-.094357	-.063772	.690024	.037193
4	.322750	-.023137	.461585	.700686	-.169986	-.396283	-.073078
5	.267889	.350798	-.088205	.099042	.756605	.030879	-.462799
6	.396862	-.394080	-.372914	.399906	.142850	.492388	.354061
7	.174266	.568333	-.026358	-.063679	.148926	-.236870	.750728

The Common PCA latent roots, sample by sample

1	.00952	.00083	.00009	.00020	.00014	.00047	.00049
2	.01002	.00068	.00008	.00013	.00012	.00031	.00013
3	.00964	.00036	.00008	.00017	.00014	.00034	.00019
4	.00812	.00036	.00005	.00012	.00006	.00030	.00016

Considering now the common principal component latent roots, it will be seen that the first two values are little

different from those obtained by the usual method applied to each sample. The remaining roots diverge more, but seldom seriously. Nevertheless, it would be inadvisable to postulate the existence of more than two common principal components.

The log-likelihood test for the common principal component model yields the result:

chi-square = 106.5572 for 63 degrees of freedom.

This value is not significant, which indicates that the common principal component model is not inappropriate.

Standard errors of latent roots

Sample	Standard errors for latent roots						
1	.00161	.00014	.00001	.00003	.00002	.00008	.00008
2	.00169	.00011	.00001	.00002	.00002	.00005	.00002
3	.00163	.00006	.00001	.00003	.00002	.00006	.00003
4	.00158	.00007	.00001	.00002	.00001	.00006	.00003

Regional Validity of Data

Under this heading I place a procedure for checking the regional (geographical) validity of data. By this I mean the compatability of samples compared at two or more geographical locations. The interest in doing this comes from quantitative genetics in which the regional validity of the multidimensional phenotype is of paramount importance. The theme for geological observations is closely akin to what is done in a multivariate analysis of variance and pairwise tests of the properties of covariance matrices. The way in which the present analysis is constructed will now be described.

Testing for Regional Validity by Collinearity of Vectors

Blackith and Reyment (1971) demonstrated the use of the angles between latent vectors in morphometric work. This procedure is, however, useful only in an advisory capacity, since it cannot be connected to a statistical test. The latent roots and vectors computed from a covariance matrix can be used for assessing regional validity of a chemical data set by the following procedure.

1. Let $\mathbf{S}_1$ denote the covariance matrix of the data matrix of one of two samples and $\mathbf{S}_2$ that of the other. The first of these matrices is designated as being the *reference matrix*.

2. Compute the latent roots $\boldsymbol{\Lambda}$ and latent vectors $\mathbf{B}$ of the second matrix.

3. Using the simple relationships

$$\boldsymbol{\Lambda} = \mathbf{B}^T \mathbf{S}_1 \mathbf{B}$$

and

$$\boldsymbol{\Lambda}^{-1} = \mathbf{B}^T \mathbf{S}^{-1}_1 \mathbf{B} \qquad (27)$$

Anderson (1963) derived a large-sample test for collinearity of latent vectors. This was later adapted by Reyment (1969) for application to biometrical problems. The property of collinearity of latent vectors implies that the covariance matrices are equal, as indeed is indicated by (27) which says that the latent roots and vectors of the two matrices are in effect interchangeable.

4. Compute the approximate chi-square relationship

$$(N_2\text{-}1)[?_i\mathbf{b}^T_i\mathbf{S}_1^{-1}\mathbf{b}_i + ?^{-1}\mathbf{b}^T_i\mathbf{S}_1\mathbf{b}_i - 2) \qquad (28)$$

where N_2 is the size of the second weight matrix and $i = 1, ..., k$, where k denotes the number of partial warps represented in the weight matrix. The relationship (28) is approximately distributed as $?^2$ with $k\text{-}1$ degrees of freedom. The test is performed separately for as many of the latent vectors as are of interest.

5. The procedure relies rather strongly on the input matrices being "well behaved", by which I mean that the data-matrices should not deviate too markedly from multivariate Gaussian. It is therefore advisable to scan the data for atypical and influential observations by some such procedure as cross-validational principal component analysis (Krzanowski, 1987; Reyment and Jöreskog, 1993). This enables the analyst to isolate specimens that diverge significantly from their fellows. Such specimens become then candidates for deletion from the analysis. This action will usually lead to greater stability in estimation.

Exemplification of the Method

The data comprise two samples of the geochemical determinations on the shell of the ostracod species *Leptocythere psammophila*, used elsewhere in this article. I refer to page 000 for further details on the chemical elements determined. Our present aim is to see how well the two samples agree in their multivariate properties. Please note that in the example provided here, I have used the log-ratio data matrix transformation of Aitchison (1986).

Box 14. Regional Validity of Multivariate Analyses

The constrained *Leptocythere* data spring+summer compared with winter

The reference sample size is 26; the comparison sample size is 14 and there are 12 variables.

Distribution details for reference sample

Variable	Skewness	t_{skew}	Kurtosis	$t_{kurtosis}$
1	-.9700	**-2.1293**	.2972	.3353
2	-1.0677	**-2.3437**	1.1249	1.2689
3	.0460	.1010	-1.1309	-1.2756
4	-.0558	-.1225	-.7341	-.8281
5	-.6793	-1.4912	-.5581	-.6295
6	1.4424	**3.1662**	2.0763	**2.3421**
7	-.0809	-.1776	-.1954	-.2204
8	-.8488	-1.8632	-.0822	-.0927
9	-1.2720	**-2.7922**	.9471	1.0683
10	.6213	1.3637	-.2226	-.2510
11	.6616	1.4524	-1.0384	-1.1713
12	-1.1632	**-2.5533**	1.7417	**1.9647**

Distribution details for comparison sample

Variable	Skewness	t_{skew}	Kurtosis	$t_{kurtosis}$

1	-.6588	-1.1029	-2.2788	**-1.9746**
2	-.2823	-.4725	-.8401	-.7279
3	.2966	.4965	-1.5164	-1.3140
4	-.3333	-.5579	-.6806	-.5898
5	.1357	.2272	-.6491	-.5624
6	.8492	1.4216	1.8166	1.5741
7	.1247	.2087	-.8720	-.7556
8	-1.0394	-1.7399	.7555	.6546
9	.1113	.1862	-.8045	-.6971
10	.5876	.9837	-.5487	-.4754
11	.2205	.3691	-1.7779	-1.5406
12	.1856	.3108	-1.0156	-.8800

latent roots for reference matrix

3226.35000 626.16280 442.25390 324.93830 253.27770 163.02380 71.46946 45.61971 25.72011 23.36255 1.87733

latent vectors for reference matrix

	1	2	3	4	5	6	7	8	9	10
1	.25065	-.05863	-.07842	.06039	-.19022	-.02155	-.12115	-.29570	.16788	-.20468
2	.23352	-.09278	-.10240	-.02197	-.12226	.15998	.14135	.71963	.49190	.14909
3	-.20879	-.66188	.06126	.20284	.51330	-.21547	-.26787	.01608	.05673	.06988
4	.15740	.14566	.17933	-.04718	.04900	-.60645	.49932	-.09499	-.11018	.43951
5	.24735	-.03549	-.20979	.00236	-.26253	.17700	-.34509	-.44147	.14652	.47087
6	-.24763	.02606	-.17957	-.87097	.18561	.05104	.03099	-.08857	.06289	-.11830

7	.22167	.58299	.12104	.16847	.62642	.25197	-.16287	.01766	-.02698	-.03385
8	-.04531	-.27437	-.01403	.13713	-.00701	.59823	.54583	-.11470	-.38797	.01989
9	.20651	-.02453	.02096	-.11741	-.24584	-.13696	-.38321	.37295	-.69618	-.09446
10	-.40406	.11001	.76549	.00921	-.31277	.11065	-.12345	-.02302	.16611	-.02420
11	-.62403	.31297	-.51749	.33603	-.14520	-.13145	-.00754	.08289	-.02783	.02687
12	.21274	-.02997	-.04633	.14120	-.08842	-.23697	.19394	-.15071	.15743	-.70024

	11
1	-.79479
2	-.00506
3	.00907
4	-.05003
5	.39094
6	.00260
7	-.00284
8	-.01949
9	.02038
10	.03508
11	-.04255
12	.45718

latent roots for comparison matrix

701.40340 581.11000 369.92310 290.56610 148.97380 71.13950 56.84750 21.45061 13.63170 5.61000 2.48706

latent vectors for comparison matrix

	1	2	3	4	5	6	7	8	9	10
1	-.12393	-.19500	.15258	.03097	.16476	-.01613	.17124	.01892	-.18826	-.51814
2	-.23211	-.05968	.13349	.16448	.09991	.06193	.40413	.73779	.00003	.25485
3	.10349	-.50875	-.31675	.00655	-.71471	-.17143	.07237	-.00514	-.02254	.01933
4	-.15519	-.04954	-.11994	-.25561	.37102	-.74925	.03880	-.19124	.06088	.26296
5	-.15058	-.12120	.13486	.09042	.05626	.33609	.29129	-.44545	.66804	.07420
6	.86372	-.06042	-.11782	-.12346	.31091	.15241	.10316	.07276	-.00555	.04072
7	.09357	.61653	.37102	-.34702	-.43133	-.07103	.23829	-.08905	-.11096	.02103
8	.03853	.02758	.23509	-.06117	-.06867	-.10597	-.68911	.33715	.45651	-.20593
9	-.12640	-.20468	.23922	-.06513	.04182	.33919	-.38722	-.17730	-.40511	.55759
10	.10563	.29439	-.02383	.84508	-.01567	-.19130	-.09626	-.18274	-.14323	-.00220
11	-.24780	.39057	-.74693	-.14109	.04529	.29323	-.11532	.07567	.02017	-.01788
12	-.16885	-.12966	.05911	-.14393	.14062	.12269	-.03093	-.15110	-.32907	-.48560

	11
1	-.69079
2	.16798
3	.01963
4	.02164
5	.04404
6	.02087
7	-.03330
8	.01028
9	-.15488
10	.06433
11	-.13039
12	.66710

There are only 11 latent roots and vectors, owing to the constraint.

Test of collinearity of latent vectors

Angles between latent vectors

between vectors 1 = 71.7294 degr.

between vectors 2 = 30.1618 degr.

between vectors 3 = 70.2155 degr.

between vectors 4 = 89.9871 degr.

between vectors 5 = 49.9544 degr.

between vectors 6 = 69.3717 degr.

between vectors 7 = 71.2905 degr.

between vectors 8 = 48.6604 degr.

between vectors 9 = 84.8534 degr.

between vectors 10 = 55.0243 degr.

between vectors 11 = 29.0639 degr.

Collinearity of Latent vectors

significance of chi-square at 5 pct level for 11 degrees of freedom = 19.675

vector	chi-square
1	22.658
2	62.353
3	78.737
4	52.499
5	14.826
6	45.541
7	40.627

Findings

The univariate tests indicate that the larger sample shows several deviations from normality, both with respect to skewness as to kurtosis. There is only one such variable in the case of the

second sample. Significant values of *t* are indicated in bold type in the listing for the univariate tests.

All angles computed between pairs of latent vectors of the covariance matrices are high and it is therefore not surprising that the specific tests for collinearity indicate that all pairs of latent vectors differ in orientation. It may therefore be concluded that the geographical difference between samples is accompanied by a strongly accentuated difference in statistical properties of the contents of chemical elements.

Correlating between Sets

Relatively early in the history of multivariate analysis, interest was directed towards trying to quantify the association between sets of variables having a joint distribution. Most earlier research in multivariate analytical applications was concerned with the social sciences and this initial involvement has left an indelible mark on thinking and practice, even in work done in much different fields. Hotelling (1936) wanted to correlate between sets of different kinds of observations on the learning ability of schoolchildren. He called his procedure canonical correlation, pursuing in his presentation, a vague analogy with principal component interpretations. A suite of correlations is extracted from the data, analogously to the latent roots of principal components, but it is the interpretation of the results that has been the stumbling block ever since the method was introduced and even to this today, and despite numerous attempts at revamping the reification of the results, canonical correlation remains something of a maverick method in the minds of many practitioners. I proceed now to a quick and far from comprehensive presentation of the method in order to show that in its mathematical

structure, at least, canonical correlation is the same as canonical variate analysis of the previous section, or *vice versa*.

Just as in the method of canonical variates, the problem is that of the simultaneous reduction of two symmetric matrices to diagonal form. As we saw before, in canonical variate analysis, the two matrices we start with are **T**, the total sum of squares and deviations, and **W**, the within-groups sum of squares and deviations. These matrices are symmetric positive definite. Their difference, defined as **B** = **T** - **W**, is positive semidefinite. (This means that the determinants of **T** and **W** are positive, whereas that of **B** is nought.) We say how the required canonical roots and vectors of canonical variate analysis are obtained from the solution of the determinantal equation (29).

The structure of canonical correlation analysis is likewise based on the simultaneous reduction of two symmetric matrices to diagonal form, although the input is not the same in that the sample is assumed to be from one statistical population, not several, as in canonical variates. Consider now a correlation matrix **R** (the covariance matrix can be used instead if you wish) consisting of correlations computed between *p* variables. Imagine now that these variables are of two kinds, say, *q* chemical analyses in one set and *r* physical measures in the other. Such a matrix can be partitioned as follows:

$$R=\begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}$$

(29)

All the correlations pertinent to the chemical variables are sequestered into $\mathbf{R}_{11}$, all of those pertaining to the physical variables into $\mathbf{R}_{22}$. The matrix $\mathbf{R}_{12} = \mathbf{R}_{21}^{T}$ contains the correlations between the variables of the two sets, i.e., the associations between chemical and physical traits of the sample. In this representation, **T** and **W** are replaced respectively by $\mathbf{R}_{11}$ and $\mathbf{R}_{11} - \mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}$.

The roots of the equation are the squares of the required canonical correlations between two new variables. To each of these correlations correspond two sets of coefficients, one corresponding to the one set for *q* variables and the other corresponding to the other set for *r* variables, that is, a new coordinate system in the space of each set of variates. These are the linear combinations of variables in each set that have maximum correlation and which are the first coordinates in the new system of coordinates. Then a second linear combination in each set is sought such that the correlation between these is the maximum of correlations between such linear combinations as are uncorrelated with the first linear associations. The process is continued until the two new coordinate systems are completely specified. Difficulties begin to arise when an attempt is made to reify these coefficients and all too often, the results seem to lack scientific sense. There is also a difficulty with respect to the canonical correlations themselves. They do not represent successively smaller portions of the total correlation between sets, as one might expect from what is achieved by the partitioning of variance in principal component analysis. In fact, it is not uncommon that a very high canonical correlation can be associated with a very minor relationship between sets. Cooley and Lohnes (1971) tried to rectify this situation by a rather complicated and to a certain extent arbitrary *redundancy analysis*. My own experience indicates that fuzzy results with confusing interpretations can be expected when:

1. The data deviate markedly from multivariate normality.

2. There are very high correlations in $\mathbf{R}_{12}$, as forecast by N. A. Campbell (1980, 1984).

3. There is great disparity in the types of variables included in one or both of the sets.

4. Variances are greatly different.

There are things that can be done in order to iron out some of these difficulties, including a suitable transformation of the data before analysis (not forgetting the need to attend to compositional data in the right manner), and to scan the data for outliers, etc. In this latter respect, it can be useful to precede a canonical correlation analysis by scrutinizing the data-matrix, using Krzanowski's method for cross-validational principal components to ferret out atypical and influential observations.

The Vector Correlation Coefficient

People often ask if there is no way of expressing the correlation between sets by means of a single number, given the interpretational problems attaching to canonical correlation coefficients. Escoufier (1973) proposed a generalization of Pearson's "coefficient of determination", the square of the correlation coefficient, which he called the coefficient of *vector correlation*, denoted *RV.*

It is simply defined as

$$RV = \frac{Trace(S_{12}S_{21})}{\sqrt{Trace(S_{11}^2)\,Trace(S_{22}^2)}}$$

==========

Box 15. *Example of canonical correlation analysis: the Spanish palaeoenvironmental data*

variables in left set = 3 (length, height and breadth of the ostracod carapace); variables in right set = 3 (boron, chromium, vanadium). Data supplied by ELF, Pau, France.

number of observations = 10

Correlations for left hand set (R_{11})

1	1.0000		
2	-.9670	1.0000	
3	-.2142	-.0355	1.0000

correlations for right hand set (R_{22})

1	1.0000		
2	-.4270	1.0000	
3	-.2047	-.4830	1.0000

correlations between sets ($R_{12} = R_{21}$)

1	-.5147	.5546	-.0951
2	.3837	-.5490	.1308
3	.6058	-.1008	-.1798

latent root 1 .5870246

latent root 2 .2106163

latent root 3 .0650265

latent vectors

	number 1	number 2	number 3
1	.8155	.8750	.4931
2	-.3108	1.3233	.4940
3	-.1322	.4419	1.2539

Test for canonical correlation significance

Root 1 = .3048, chi-square = 7.7227

Degrees of freedom = 9.

Root 2 = .7381, chi-square = 1.9743

Degrees of freedom = 4.

Root 3 = .9350, chi-square = .4370

Degrees of freedom = 1.

canonical correlation 1 = .7662

coefficients for the right set

.8155 -.3108 -.1322

coefficients for the left set

5.2845 5.7917 2.0544

scores for canonical correlation 1

individual	left-hand scores	right-hand scores
1	-1.8319	-1.7508
2	-.8613	-.6072
3	.2173	-.1462
4	1.0349	.0814
5	1.2008	1.6321
6	-.3354	.7699
7	-.0726	-.6958
8	.4115	-.4240
9	-.8884	-.1250
10	1.1251	1.2660

Redundancy analysis for correlations between original variables and new canonical variables

correlations between original left-hand set and the new variables

-.7563 .6085 .7167

correlations between original right-hand set and the new variables

.9753 -.5951 -.1491

Note: It is often found that these two vectors give a more understandable reification of a canonical correlation analysis than the canonical variate vectors, from which they are computed in a manner reminiscent of the techniques of factor analysis.

Proportion of variance of left set
explained by canonical correlation 1 = .4853

proportion of variance of left-hand set explained by
canonical correlation 1 of the right set = .2849

proportion of variance of right set
explained by canonical correlation 1 = .4425

proportion of variance of right-hand set explained by
canonical correlation 1 of the left_set = .2598

Note: This is the asymmetry property of redundancy analysis.

Vector of correlations between original
left-set variables and canonical variates of right set
variables

-.5795 .4662 .5491

Vector of correlations between original
variables and canonical variates of left-set variables

.7472 -.4559 -.1142

==========

Discussion of the canonical correlation analysis

This simple little example presents several points of importance for understanding the way in which a canonical correlation analysis appears in practical work. The directly computed canonical vectors differ rather strongly from the corresponding redundancy vectors due no doubt to the fact that the data do not fit the multivariate normal distribution very well. Both do, however support the view that the first component of each vector is important in establishing the observed correlation. The example also brings out a sad fact of geostatistical life, namely, that geological data usually deviate from theoretical nicety for which reason it can be very difficult to get much sense out of the canonical correlation analysis.

The *redundancy analysis* is an attempt by Stewart and Love (1968) to clarify the interpretation of the results yielded by canonical correlation analysis. The redundancy is defined as the proportion of the variance extracted by the canonical factor R_{dx} multiplied by the proportion of shared variance between the factor and the corresponding canonical factor of the other set. It expresses the amount of overlap between the two sets that is contained in the first canonical relationship, and so on (Cooley and Lohnes, 1971, p. 170). The feeling that one is not doing something quite right with all the seemingly subjective jugglings with matrix multiplications, has created a good deal of hesitancy on the part of many practitioners in adopting the method. Gleason (1976) has, however, shown that the procedure is statistically correct. Providing the

data are not too different from the multivariate normal condition, the results yielded by the unembellished canonical correlation computations should prove satisfactory for most geological purposes. There is a geological example in Reyment (1991, p. 68) in which the redundancy vectors are reified to a scientifically appealing result and an analysis of the occurrence of lead and zinc in an earlier article (Reyment, 1972). The method has not caught on very well; Jackson (1991), for example, noted it as well as competitors, but did not exemplify the procedure.

A Problem in Petrology and Geochemistry

Introduction: It should by now be apparent to you that multivariate data abound in everday geochemistry and petrology. In fact, if a statistical survey were to be carried out it would probably be found that most multivariate situations in geology (excluding palaeontology) arise in the study of the compositions of rocks. It is, however, here that most errors in statistical applications occur, as has been well exposed by Aitchison (1986) in his fundamental monograph.

The example presented in this final chapter is rather typical of its genre. Demange *et al.* (1983) studied a recent N-S trending volcanic chain, the Jabal al Abyad, located along a fractural axis, classified as a rift-valley, the rocks of which have evolved from mildly alkaline basalts to phonolites under certain structural conditions, and to comendites under others. The analyses of the chemical data were made using simple graphs of ratios, the statistical soundness of which type of procedure is not unchallengeable. Chemical and mineralogical variations were explained as manifestations of fractional crystallization. There is an unexpected complication to understanding the importance of a major part of the publication in that there is no explanation of the symbols used in the graphs (the Fig. 4 of the authors lacks the key to which that figure and several others refer).

The data consist of determinations made on 11 samples of the major elements, expressed as oxides, SiO_2, TiO_2, Al_2O_3, Fe_2O_3, FeO, MnO, MgO, CaO, Na_2O, K_2O, P_2O_5 and an unspecified residue. The trace elements assessed were V, Cr, Co, Ni, Cu, Zn, Li, Rb, Sr, and Ba. There are 11 major elements and 10 trace elements in the data-matrix. It is quite obvious that the first set of determinations consists of parts which sum to a constant. What is the situation for the trace-elements? Although there does not seem to be an obvious constraint involved, there is, however, indeed. These data are expressed as parts per million and are no more or no less than the "parts" we met in the first chapter.

Although Demange *et al.* (1983) did not compute any statistics, other than a very few simple univariate ones, for their data, they did perform some graphical operations having a statistical import. We need only to consider the use of ratios. It is rather common praxis in petrology, it seems, to attempt to arrive at meaningful ordinations of data by plotting ratios that are interpreted as possessing special diagnostic significance. One such graph used in the study examined here is the ratio Ca/Sr plotted against Sr. Another is the graph of the ratio K/Ba plotted against Ba. The statistical, and logical, objection that comes to the fore here is that the same component enters into both axes of the graph; strontium is being compared with itself and barium likewise with itself. This is a questionable procedure at best. It is also difficult to desist from concluding that quantitatively based petrological classifications bear the stamp of arbitrariness (see Sörensen, 1974).

Suggested analytical approach

The principal value of the data-set lies with the possibility it offers of finding natural groupings in the samples, and hence, the rock-types subjected to chemical analysis. The first problem to be overcome is that there are more variables than samples, namely, 21 chemical parts

and only 11 samples. This indicates that a *Q*-mode strategy is going to be required, using a rather special arrangement of the data with the variables taking the role usually played by the specimens. It is, as it were, as though we had inverted the space occupied by the data. A good method with which to start is that of *principal coordinates.* We shall apply the method to the raw data-matrix and to its log-ratio equivalent as a start.

The second area of interest centres about the relationship between major elements and trace-elements. *Canonical correlation analysis* supplies a possible means of assessing this but, owing to the few of samples available, it will be necessary to cull the number of variables. In reducing the number of variables, I have relied on the elements betokened as being diagnostic in the article by Demange and his coworkers.

Principal coordinate analyses

The salient results obtained by applying principal coordinate analysis to various versions of the data-matrix for the samples of igneous rocks from the Saudi rift-valley (Demange *et al.,* 1983) are displayed in **Boxes 16** and **17**. The first box contains an abridged version of the results for the computations applied to the raw data-matrix. The second of these boxes presents a summary of the same computations implemented for the log-ratio data-matrix.

Box 16: Principal coordinate analysis of the raw data-matrix for the Saudi rift-valley

(a) The raw-data matrix in terms of the 11 rock-samples

Data from Demange (1983) SiO_2, TiO_2, Al_2O_3, Fe_2O_3, FeO, MnO, MgO, CaO, Na_2O, K_2O, P_2O_5

The Prim minimum spanning tree

Connected	to	by the	distance
hawaiite_1	basalt		1.592932
hawaiite_2	hawaiite_1		2.544794
mugearite	hawaiite_2		1.381281
benmoreite	mugearite		3.693137
quartz trachyte	benmoreite		2.383008
trachyphonolite	quartz trachyte		1.242190
alkaline phono-lite	trachyphonolite		1.346894
phonolite	alkaline phonolite		1.399943
comend. trachyte	alkaline phonolite		1.529463
comendite	comend. trachyte		0.794333

Latent roots of transformed association matrix

1.9094930
.5766517
.3747105

Specimen	Coordinates		
1	.61138	-.39597	.05916
2	.55055	-.27722	.01221
3	.51923	.23562	-.23695
4	.42266	.41319	-.10332
5	.01722	.17596	.50239
6	-.26899	.17259	-.03536
7	-.36128	-.01807	.01149
8	-.40077	.00370	-.07458
9	-.23541	.01897	.11587
10	-.42697	-.13929	-.12398
11	-.42760	-.18949	-.12692

The plot for the first pair of coordinates is shown in Fig. 3.

Values of residuals

roots exceeding	percentage residual
2	35.42
3	25.68

Discussion of results in ***Box 16***

The Prim minimum spanning tree, superimposed on the plot of the first two principal coordinates, gives a clear indication of progression in chemical properties as a function of time from basaltic rocks to phonolitic in reasonable accordance with the deductions of Demange *et al.* (1983). There is a high degree of integration in the data which explains the quite low residuals obtained. This condition is a favourable one in a *Q*-mode analysis, since it preserves distances between objects. Its statistical effects are further manifested in the concluding section of this section.

We now pass to the principal coordinate analysis of the log-ratio data-matrix of oxides, summarized in **Box 17**.

Box 17: Principal coordinate analysis of the constrained data-matrix for the Saudi rift data.

Variables = 11 Individuals = 11

MINIMUM SPANNING TREE

Connected	to by	distance

hawaiite_1	basalt	1.141441
hawaiite_2	hawaiite_1	1.841724
mugearite	hawaiite_2	.634479
benmoreite	mugearite	2.902132
quartzose trachyte	benmoreite	2.324739
trachyphonolite	quartzose trachyte	1.012213
comenditic trachyte	trachyphonolite	1.802150
phonolite	comenditic trachyte	1.580188
hyperalk. phonolite	comenditic trachyte	1.687741
comendite	hyperalk. phonolite	2.169437

LENGTH OF TREE = 17.096240

Latent roots of transformed association matrix

1.7819760
.4760598
.3112795

Specimen	Coordinates		
basalt	-.51810	.18779	-.21612
hawaiite_1	-.51791	.12274	-.21884
hawaiite_2	-.46363	.10088	.27345
mugearite	-.40106	.02029	.31308
benmoreite	-.15000	-.34961	-.18369
trachyphonolite	.16514	-.23012	.02877

phonolite	.35686	-.12786	.05374
hyperalk. phonolite	.46560	.20054	.02030
quartzose trachyte	.10969	-.25494	-.01640
comenditic trachyte	.40929	-.01353	.01906
comendite	.54413	.34383	-.07334

Fig. 4 presents the plot of the first two principal coordinates for this analysis.

Values of residuals

roots exceeding	percentage residual
2	35.71
3	26.85

Discussion of the constrained principal coordinate analysis

Although the general form of the results seems to be much the same as was obtained with the raw data-matrix, and the relatively small residuals are closely comparable, there are, nonetheless, some significant deviations that are worth noting. If you examine the plot of the points in the plane of the first two coordinates you will observe that the first two thirds of the path traced out by the Prim network is the same up to the location of the sample of trachyphonolite. Thereafter, the ordering of the samples differs, to end up, however, in both cases with comendite.

Whichever of the two results makes the best scientific sense requires expertise in "petromancy". The salient feature of the present analysis is that the statistically correct model leads to a somewhat different outcome from that yielded by the unsophisticated one. However. in both cases, the high correlations occurring between variables guarantee that the distances between objects are well preserved in the plane of the first two principal coordinates.

Relationships between sets

The original data set consists of observations on 11 "oxides" and 10 trace-elements. The discussion given by Demange *et al. (1983)* places emphasis on the interplay between the parts FeO, MnO, MgO, CaO, Na_2O and K_2O, for the oxides, and Rb, Sr and Ba for the trace-elements. I have, therefore, selected a subset of the data for analysis by the method of canonical correlation. I have done this in two ways for the purposes of comparison. One uses the model appropriate to compositional data, the other the raw correlations. Look now at the results listed in **Box 18** for the log-ratio covariance data-matrix.

Box 18. Canonical correlation analysis of a log-ratio data-matrix for the igneous rock samples from the Saudi rift-valley.

Canonical correlation analysis
Fe++(1), Mn(2), Mg(3), Ca(4), Na(5), K(6), against Rb(7) Sr(8) Ba(9) for the Jabal al Abyad data

variables in left set = 6, the six oxides

variables in right set = 3, the three trace-elements

number of observations = 11

Correlations for left-set (R_{11})

1	1.0000					
2	-.5498	1.0000				
3	.4844	-.0800	1.0000			
4	.2145	.1408	.8871	1.0000		
5	-.8972	.6187	-.5988	-.3548	1.0000	
6	-.8502	.4118	-.7831	-.5825	.9493	1.000

Some of these correlations are relatively high and there are no really poor performers.

Correlations for right-set (R_{22})

1	1.0000		
2	-.7662	1.0000	
3	-.7492	.6595	1.0000

These values are likewise high.

Correlations R_{12} between sets

1	-.8945	.5240	.7125
2	.5327	-.4667	-.7565
3	-.6261	.3674	.0066
4	-.4325	.2536	-.2185
5	.9859	-.7630	-.7937
6	.9669	-.7379	-.5830

The between-sets correlations point to there being an impressive level of integration between rock-types, major elements, and trace-elements. This is not unexpected bearing in mind the results of the principal coordinate analysis.

Weierstrass Diagonalization

Latent root	1	.9999996
Latent root	2	.9954983
Latent root	3	.7208668

Initial latent vectors

	Number 1	Number 2	Number 3
1	-1.2395	-1.1265	.6671
2	-.4451	.0033	1.5249
3	-1.3523	-.1820	-.7167

Wilks test of significance of canonical correlations

Root 1 = .0000; chi-square = 128.2139
Degrees of freedom = 18.

Root 2 = .0013; chi-square = 40.0762
Degrees of freedom = 10.

There are two significant canonical roots and hence two significant canonical correlations (the rank of the problem is 3).

Canonical Correlation 1 = 0.9999

Coefficients for right set

-1.2395 -.4451 -1.3523

Coefficients for left set

.8963 .3065 .8794 .4633 .6101 .9336

These values suggest a high level of association between all parts of the sets. We shall

however see what the redundancy analysis provides.

Stewart-Love Redundancy Analysis

Correlations between original LEFT-set and new variates

-.0880 .5704 .6037 .7186 .1909 -.0816

Correlations between original RIGHT-set and new variates

.1147 -.3872 -.7172

The redundancy analysis is possibly yielding a more informative picture of the relationships between sets with Mn, Mg and Ca entering into an association with Sr and Ba. Given that the second canonical correlation is also highly significant it behoves the analyst to examine it as well.

Canonical Correlation 2 = .9977

Coefficients for right set

-1.1265 .0033 -.1820

Coefficients for left set

.1362 .0468 -.2708 .2865 -.4979 -.4638

Correlations between original LEFT-set and new variates

.8818 -.4651 .7070 .5290 -.9709 -.9878

Correlations between original RIGHT-set and new variates

-.9928 .7465 .6643

The relationship represented here indicates a high level of integration in all variables. Note the marked difference between the redundancy vectors and the canonical correlation vectors.

The plot of the scores yields, as only to be expected, a straight line - the first canonical correlation is actually very slightly less than 1. This is indeed very high and incites a certain element of caution in the light of Campbell's (1980) results regarding instability in the components of latent vectors in multivariate analysis.

Discussion of canonical analysis of parts

The redundancy vectors for correlations between the original observations and the new variates underline the very high level of integration in the chemistry of the rock-samples. It is not an easy matter to say just how much reliance can be placed on results obtained for such remarkably high correlations. It can, therefore, be advisable to see what happens when the inappropriate data-matrix is analyzed in order to "gain a feeling" for the log-ratio based analysis. **Box 19** summarizes the calculations made on the raw data-matrix.

Box 19. Canonical correlation analysis of a raw data-matrix for the igneous rocks samples from the Saudi rift-valley

Canonical correlation analysis for data on Jabal al Abyad

Correlations for left-set (R_{11})

1	1.0000					
2	.6651	1.0000				
3	.7712	.5848	1.0000			
4	.8104	.5995	.9823	1.0000		
5	-.7655	-.5442	-.9253	-.9190	1.0000	
6	-.8360	-.5721	-.9455	-.9473	.9505	1.0000

If you compare this matrix with its companion in **Box 18**, you will see that there are very pronounced differences in some of the correlation coefficients, but not all of them.

Correlations for right-set (R_{22})

1	1.0000		
2	-.7315	1.0000	
3	-.4987	.4775	1.0000

These correlation coefficients are not greatly different from those of the companion set.

Correlations R_{12} between sets

1	-.7217	.6517	.1988
2	-.6157	.5183	.1565
3	-.5953	.4772	-.2276
4	-.6795	.5809	-.1410
5	.4990	-.4907	.1487
6	.5879	-.6148	.0916

The between-set correlation coefficients are mostly quite different from those of the companion set.

Weierstrass Diagonalization

Latent root 1 .9238565

Latent root 2 .6139144

Latent root 3 .3061815

Initial latent vectors

	Number 1	Number 2	Number 3
1	.8520	-.8656	.9046
2	-.3964	-1.0086	1.0284

3 .6271 -.6665 -.7374

The pattern displayed by the canonical vectors differs bluntly from that of the analysis by parts. There is only one significant canonical correlation.

Canonical Correlation 1 = .9612

Coefficients for right set

.8520 -.3964 .6271

Coefficients for left set

.1054 -.1660 .9647 -2.2709 -.7366 .2886

Correlations between original LEFT-set and new variates

-.7788 -.6574 -.8730 -.9338 .7417 .8344

Correlations between original RIGHT-set and new variates

.8292 -.7202 .0129

The redundancy configuration for the raw data set is quite different from that yielded by the appropriate model for constrained data.

Discussion of the Saudi Rift Data

The foregoing set of analyses indicates that the main conclusions arrived at by Demange *et al.* (1983) are supported by the multivariate analysis, albeit with certain reservations. Even a cursory perusal of the input material discloses that there are very pronounced trends in the observations and it is therefore not surprising that such obvious structure shows up in almost any kind of quantitative appraisal.

At the statistical level, the results achieved with the aid of the appropriate model in terms of compositional parts do not greatly differ from the inappropriate model for the Q-mode analysis. At the R-mode level, however, the differences are very strong indeed.

References

Aitchison, J. 1986. *The Statistical Analysis of Compositional Data.* Monographs on Statistics and Applied Probability. Chapman and Hall, London.

Anderson, T. W. 1963. Asymptotic theory for principal components analysis. *Ann. math. Statist.* 34, 122-148.

Anderson, T. W. 1984. *An Introduction to Multivariate Statistical Analysis.* Second Edition, Wiley and Sons, New York.

Bellman, R. 1960. *Introduction to Matrix Analysis.* McGraw-Hill, New York.

Bénzécri, J.-P. 1973. *L'Analyse des Données. 2. L'Analyses des Correspondances.* Dunod, Paris.

Blackith, R. E. and Reyment, R. A. 1971. *Multivariate Morphometrics*. Academic Press, London.

Bodergat,A. M, Carbonnel, G., Rio, M., Keyser, D. 1993. Chemical composition of *Leptocythere psammophila* (Crustacea: Ostracoda) as influenced by winter metabolism and summer supplies. *Marine Biology*, 117, 53-62.

Borley, G. D. 1974. Oceanic islands. Pp. 311-330 **In** *The Alkaline Rocks* (Editor: H. Sörensen), Wiley and Sons, New York.

Campbell, N. A. 1980. Shrunken estimators in discriminant and canonical variate analysis. *Applied Statistics*, 29: 5-14.

Campbell, N.A. 1984. Canonical analysis with unequal covariance matrices: generalizations of the usual solution. *Mathematical Geology*, 16: 109-124.

Cooley, W. W. and Lohnes, P. R. 1971. *Multivariate Data Analysis* Wiley and Sons, New York.

Davis, J. C. 1988. *Statistics and Data Analysis in Geology*. Second Edition, Wiley and Sons, New York

Davis, P. J. 1965. *The Mathematics of Matrices*. Blaisdell, New York.

Demange, J., Baubron, J.-C., Marcelot, G., Cotten, J., Maury, R. C. 1983. Cadre structural, pétrologie et géochimie de la série volcanique de Jabal al Abyad (Arabie Saoudite). *Bull.*

Centres Rech. Explor.-Prod. Elf-Aquitaine, 7, 233-248.

Digby, P. G. and Kempton, R. A. 1987. *Multivariate Analysis of Ecological Communities*. Chapman and Hall, London.

Eckart, C. and Young, G. 1936. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211-218.

Escoufier, Y. 1973. Le traitement des variables vectorielles *Biometrics*, 29, 751-760.

Flury, B. 1988. *Common Principal Components and Related Multivariate Models.* Wiley, New York.

Gabriel, K. R. 1971. The biplot display of matrices with application to principal components analysis. *Biometrika*, 58: 453-467.

Gleason, T. C. 1976. On redundancy in canonical analysis. *Psych. Bull.*, 83, 1004-1006.

Gordon,_A. D. 1981. *Classification*. Monographs on Applied Probability and Statistics. Chapman and Hall, London.

Gower, J. C. 1966. Some distance properties of latent root and vector methods used in multivariate methods. *Biometrika*, 55: 325-338

Hotelling, H. 1936. Relations between two sets of variates *Biometrika*, 28:321-377.

Jackson, J. E. 1991. *A User's Guide to Principal Components*. Wiley, New York.

Jolliffe, I. T. 1986. *Principal Component Analysis*. Springer Verlag, New York.

Krzanowski, W. J. 1987a. Cross-validation in principal component analyses. *Biometrics*, 43:575-584.

Krzanowski, W. J. 1987b. Selection of variables to preserve multivariate data structure, using principal components. *Applied Statistics*, 36:22-33.

Krzanowski, W. J. 1988. *Principles of Multivariate Analysis.* Oxford Science Publications, Oxford.

Preisendorfer, R. W. 1988. *Principal Component Analysis in Meteorology and Oceanography*. Elsevier, Amsterdam.

Reyment, R. A. 1969. A multivariate palaeontological growth problem. *Biometrics,* 22, 1-8.

Reyment, R. A. 1972. Models for studying the occurrence of lead and zinc in a deltaic environment. In *Mathematical Models of Sedimentary Processes*. Ed. T. W. Merriam, Plenum, New York.

Reyment, 1991. *Multidimensional Palaeobiology*, Pergamon Press.

Reyment, R.A. and Jöreskog, K.G. 1993. *Applied Factor Analysis in the Natural Sciences*. Cambridge University Press, New York.

Searle, S. R. 1966. *Matrix Algebra for the Biological Sciences.* Wiley, New York.

Schoenberg, I. J. 1935. Remarks to Maurice Fréchet's article "Sur la définition axiomatique d'une classe d'espaces distanciés vectoriellement applicable sur l'espace de Hilbert". *Ann. Math (Second Series)*, 36, 724-732.

Schönemann, P. H. and Carroll, R. M. 1970. Fitting one matrix to another under choice of a central dilation and a rigid motion. *Psychometrika*, 35, 245-256.

Seber, G. A. F. 1984. *Multivariate Observations*, Wiley, New York.

Stewart, D. K. and Love, W. A. 1968. A general canonical correlation index. *Psychological Bulletin*, 70: 160-163.

Sörensen, H. 1974. *The Alkaline Rocks*. Wiley and Sons, New York.

Telnaes, N., Björseth, A., Christy, A. A., and Kvalheim, O. M. 1987. Interpretation of multivariate data: Relationship between phenanthrenes in crude oils. *Chemometrics and Intelligent Laboratory Systems*, 2: 149-153.

Usui, A. 1992. Hydrothermal manganese minerals in Leg 126 cores. *Proc. Ocean Drilling Program, Scientific Results*, 126, 113-123.

Listing of Text-Boxes with Examples

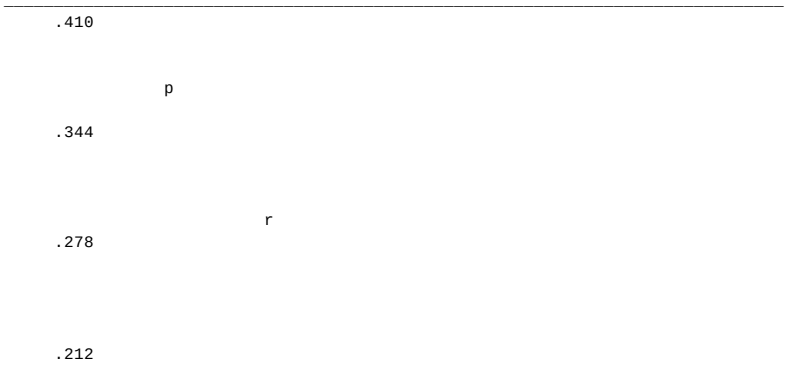


Fig. 1 **Plot of the first and second principal coordinates**

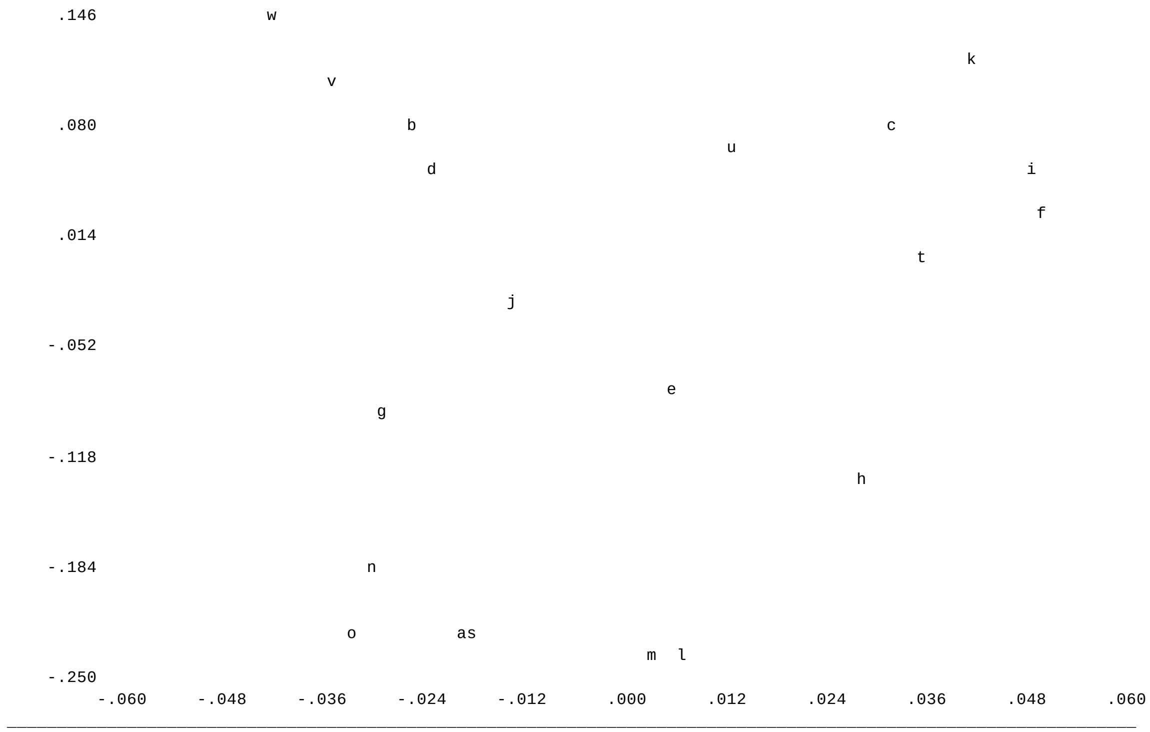
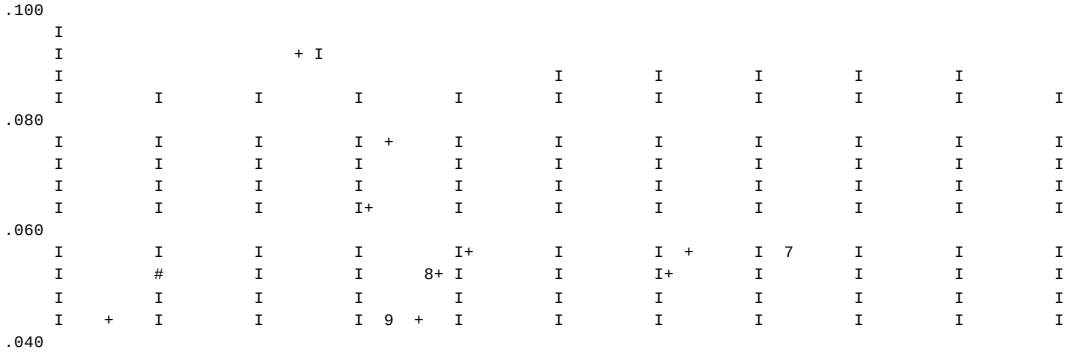
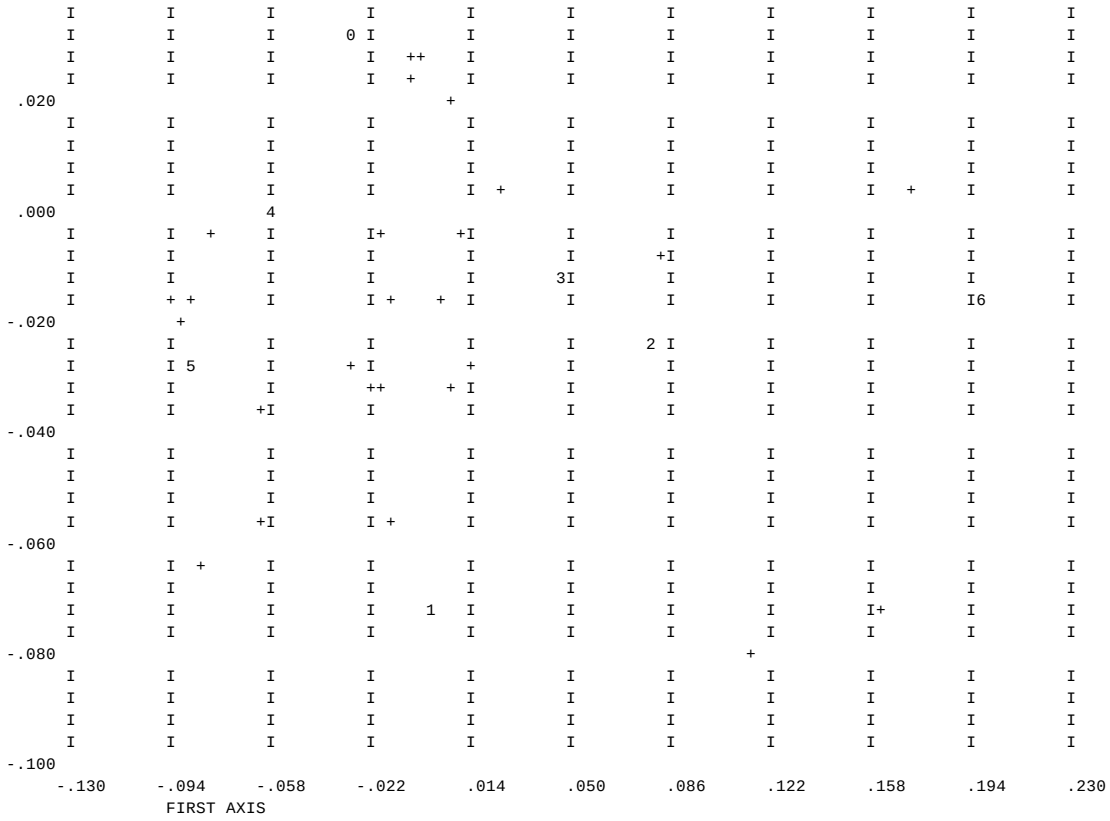

.146
.080
.014
-.052
-.118
-.184
-.250
-.060
-.048
-.036
-.024
-.012
.000
.012
.024
.036
.048
.060
w
k
v
b
c
u
d
i
f
t
j
e
g
h
n
o
as
m
l

Fig. 2 **Plot for first and second axes of a correspondence analysis**




.020
.000
-.020
-.040
-.060
-.080
-.100
-.130
-.094
-.058
-.022
.014
.050
.086
.122
.158
.194
.230
FIRST AXIS

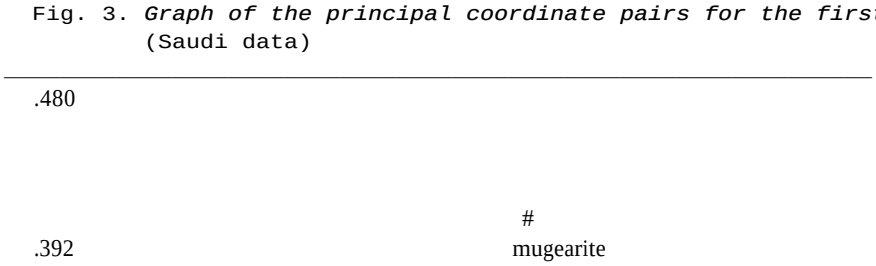


Fig. 3. *Graph of the principal coordinate pairs for the first two axes* (Saudi data)

.304

#
.216 hawaiite_2
#
benmoreite
trachyphonolite

.128

.040 hyperalkaline
phonolite #
quartzose trachyte
#
phonolite
-.048

#
-.136 comenditic trachyte

#
comendite
-.224

hawaiite_1

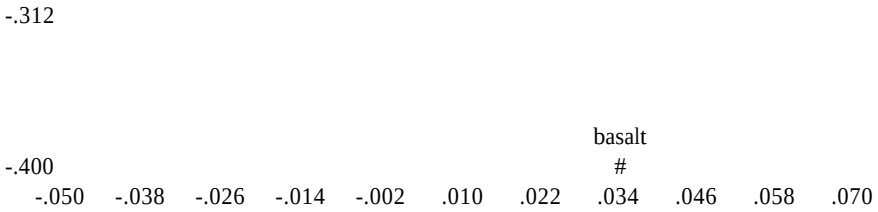
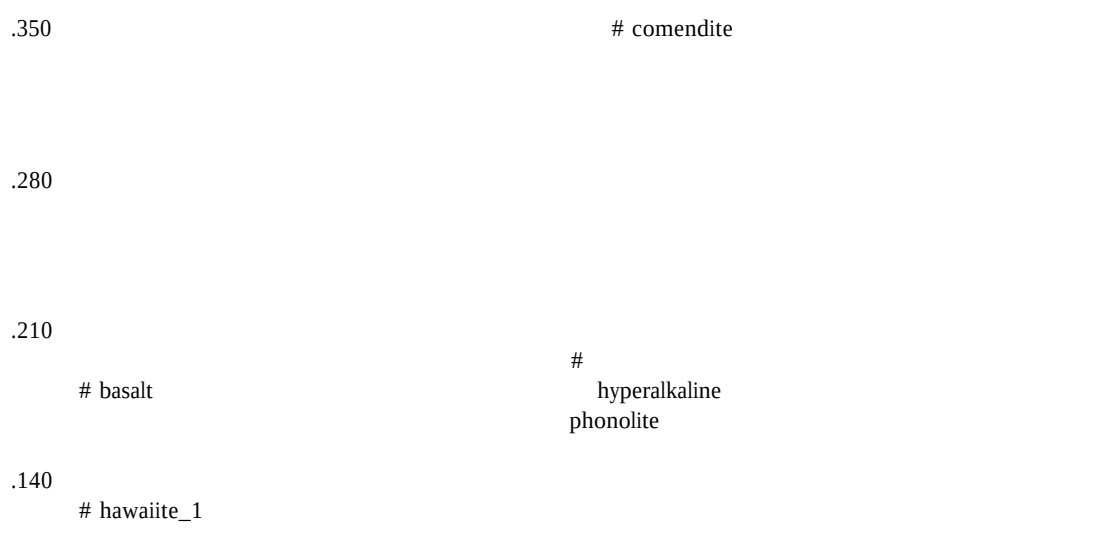
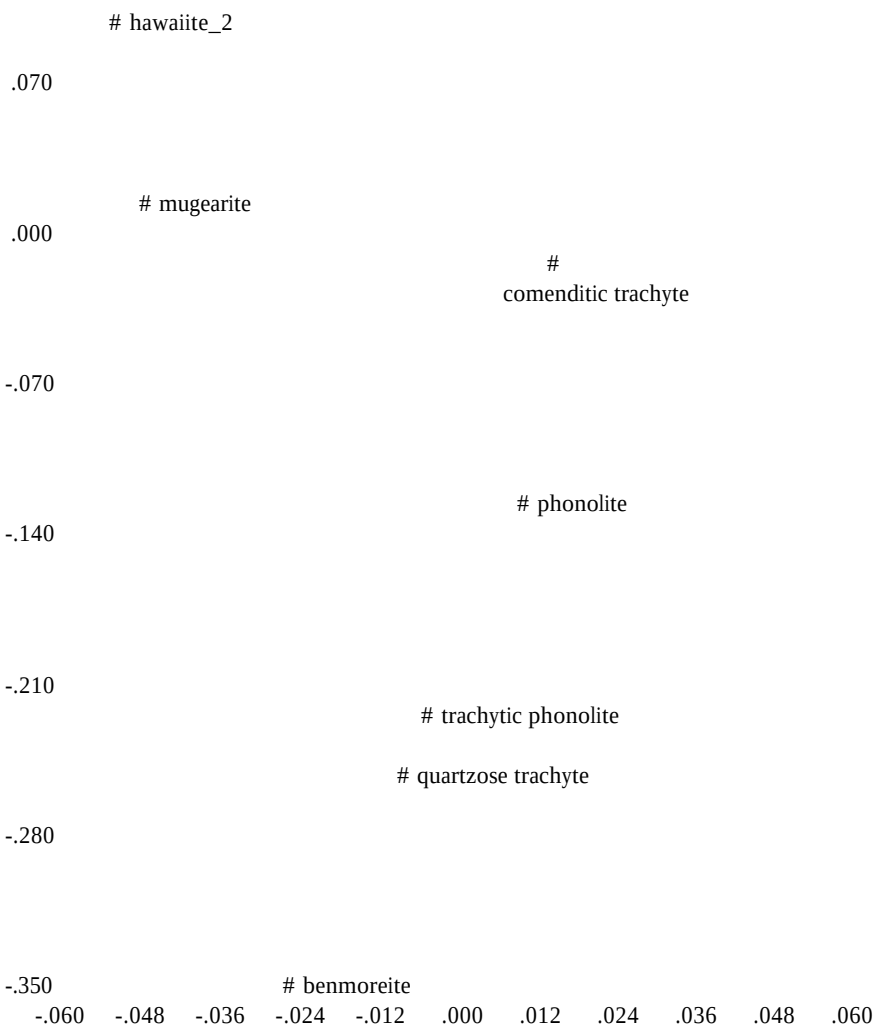


Fig. 4. *Plot of the first two principal coordinates* (Saudi data)




hawaiite_2
.070
mugearite
.000
#
comenditic trachyte
-.070
phonolite
-.140
-.210
trachytic phonolite
quartzose trachyte
-.280
-.350
benmoreite
-.060 -.048 -.036 -.024 -.012 .000 .012 .024 .036 .048 .060