Biometrika (1983), **70**, 1, *pp.* 57–65
Printed in Great Britain

Principal component analysis of compositional data

BY J. AITCHISON
Department of Statistics, University of Hong Kong

SUMMARY

Compositional data, consisting of vectors of proportions, have proved difficult to handle statistically because of the awkward constraint that the components of each vector must sum to unity. Moreover such data sets frequently display marked curvature so that linear techniques such as standard principal component analysis are likely to prove inadequate. From a critical reexamination of previous approaches we evolve, through adaptation of recently introduced transformation techniques for compositional data analysis, a log linear contrast form of principal component analysis and illustrate its advantages in applications.

Some key words: Compositional data; Isotropic covariance structure; Logistic normal distribution; Log linear contrasts; Principal components; Proportions; Spherical normal distribution.

1. THE DIFFICULTIES

1·1. *Introduction*

The often high dimensionality of compositional data, consisting of vectors of proportions of some unit, has encouraged the use of principal component analysis in attempts to find adequate low-dimensional descriptions of compositional variability (Webb & Briggs, 1966; Le Maitre, 1968; Butler, 1976). These attempts have so far been thwarted by two main difficulties with compositional data,

(i) the marked curvature often displayed by such data sets,

(ii) the constant-sum constraint which each compositional vector must satisfy.

The difficulties and their eventual resolution are best discussed against the background of two specific data sets consisting of three-part compositions and therefore capable of visual representation in standard triangular coordinates. Figure 1(a) shows the relative proportions of urinary excretions of three steroid metabolites, here simply labelled 1, 2, 3, of 37 healthy adults. Figure 1(b) shows the so-called AFM compositions of 23 aphyric Skye lavas and is reproduced from Aitchison (1982, Fig. 1).

1·2. *The curvature difficulty*

The difference between these two data sets is striking. The first is decently elliptical, a linear set in the sense of Gnanadesikan (1977, Chapter 2). The second has a decidedly nonlinear pattern but with little variation about a conceptual curved line. In purely geometrical terms standard principal component analysis, being a linear reduction technique, may be adequate for the first set but will fail to capture successfully the essentially one-dimensional curved variability of the second set. This is confirmed in Fig. 1 by the relation of the data points to the axes found by the linear principal component method of Le Maitre (1968), to be discussed in §2. Such problems of curvature are not confined to compositional data sets. For example, Gnanadesikan (1977, §2·4) uses a

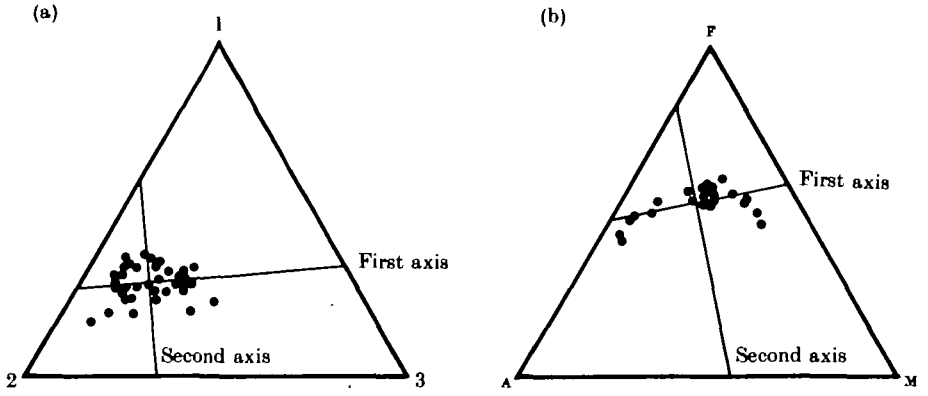

Fig. 1. Compositional data sets for (a) steroid metabolites, (b) aphyric Skye lavas, showing axes obtained by the linear principal component analysis of Le Maitre (1968).

three-dimensional paraboloid data set, albeit simulated, to demonstrate the effectiveness of quadratic principal components. Moreover, it is easy to visualize multivariate log normal data sets in the positive orthant with hyperboloid-shaped variability, for which the standard practice of working with logarithmically transformed data would lead to successful principal component analysis.

To deal with curvature in compositional data we may similarly search for appropriate nonlinear functions of the proportions from which to build up a suitable principal component system. For the compositional data sets of most practical problems, of dimensions higher than our simple illustrative examples, there is no easy way of detecting whether or not there is curvature, but our own detailed study of a number of such sets suggests that it would be unwise to ignore the possibility of curvature. Ideally then we require a form of principal component analysis which will cope with both linear and nonlinear data patterns. Such a method obviously has to be nonlinear and sufficiently flexible to approximate linearity within its range.

1·3. *The constraint difficulty*

Principal component analysis of data sets in d-dimensional Euclidean sample space R^d has been variously conceived, analytically as the determination of a sequence of variance-maximizing orthonormal linear combinations (Hotelling, 1933), geometrically as the identifying of lines and hyperplanes of closest fit (Pearson, 1901; Gower, 1967). However conceived, these procedures turn out to be equivalent to, and are now usually effected by, the successive determination of the eigenvectors of the sample covariance matrix in decreasing order of magnitude of the eigenvalues. This equivalence arises from the conformity of the definitions of distance, orthogonality and covariance in R^d.

For compositional data the appropriate sample space is the positive simplex

$$S^d = \{x^{(d)} = (x_1, \ldots, x_d): x_i > 0 \ (i = 1, \ldots, d), \ x_1 + \ldots + x_d < 1\}$$

in its proper d-dimensional, though asymmetric, setting as a subset of R^d, or as

$$S^d = \{x^{(d+1)} = (x_1, \ldots, x_{d+1}): x_i > 0 \ (i = 1, \ldots, d+1), \ x_1 + \ldots + x_{d+1} = 1\}$$

in symmetric form as a d-dimensional subspace of R^{d+1}. Although a principal component analysis which regards a data set in S^d as embedded in R^d or R^{d+1} and finds lines and

hyperplanes of closest fit using Euclidean distance and orthogonality (Gower, 1967, §5·1) may seem geometrically attractive it suffers not only from the curvature problem already discussed but also from a substantial interpretative difficulty. The role of orthogonality in such a procedure is commonly associated with the supposedly nice property that there is zero correlation between the principal components, here linear combinations of the raw proportions $x_1, \ldots, x_{d+1}$. Any attempt to argue that zero correlation is a useful statistical property when dealing with raw proportions and compositional data is, however, questionable.

It is now well known, perhaps more so in geological than in either biometrical or statistical circles, that covariances of raw proportions do not have the simple interpretations which can be placed on their counterparts used in describing variability in R^d. For examples of discussion of interpretative difficulty, see Aitchison (1981), Chayes (1960, 1962, 1971), Chayes & Kruskal (1966), Darroch (1969), Darroch & Ratcliff (1970, 1978), Miesch (1969), Mosimann (1962), Sarmanov & Vistelius (1959) and Vistelius & Sarmanov (1961). In particular the constraint places restrictions on the correlation structure of raw proportions, giving a bias towards negative values, so that the correlations are not free to range unrestrictedly over the usual parameter space associated with correlations of variables in R^d. Indeed definitions of maximum possible independence of raw proportions have come to be associated with nonzero correlations, and much effort has been directed towards determination of appropriate values of these so-called null, as opposed to zero, correlations (Chayes & Kruskal, 1966; Darroch, 1969; Darroch & Ratcliff, 1970; Mosimann, 1962). Thus there seems little point in insisting on Euclidean orthogonality and zero correlation of linear combinations of raw proportions if we wish to discuss even weak distributional properties of principal components.

Aitchison (1981, 1982) has proposed a resolution of the constraint difficulty in the analysis of compositional data. In practical terms this consists of first transforming each compositional vector $x^{(d+1)} = (x_1, \ldots, x_{d+1})$ in S^d to a vector $y^{(d)} = (y_1, \ldots, y_d)$ in R^d by a log ratio transformation

$$y^{(d)} = \log (x_{-j}/x_j),$$

where x_{-j} denotes the vector $x^{(d+1)}$ with x_j omitted, and then treating the transformed vectors by standard multivariate methods available in R^d. The success of this device in a number of applications, and in particular the successful capture of a curved data set by a predictive region (Aitchison, 1982, Fig. 1), encourages the view that it may also prove effective in principal component analysis of data sets with and without curvature.

2. The proposed method

2·1. *Three previous approaches*

Previous approaches to principal component analysis have used the eigenvalues and eigenvectors of one of three different covariance matrices:

$$\text{(i) } \operatorname{cov}(x_{-j}), \quad \text{(ii) } \operatorname{cov}(x^{(d+1)}), \quad \text{(iii) } \operatorname{cov}(x_{-j}/x_j).$$

All three approaches lead to straight line principal axes and so fail to cope with the curvature difficulty as already illustrated in Fig. 1(b).

Method (i) involves a common, though naive, approach to the analysis of compositional data through the omission of one of the proportions, presumably in the mistaken belief that the remaining proportions are relatively unrestricted. The effect on

the interpretation of correlations persists through the inequality $\Sigma_{i \neq j} x_i < 1$. Moreover, the principal components obtained differ with the choice of the omitted proportion x_j and so the method fails an obvious invariance criterion.

Covariance matrix (ii) is singular and of order $d+1$, and its zero eigenvalue has corresponding eigenvector a vector of constants. Le Maitre (1968) advocates the use of principal components based on the d nonzero eigenvalues and their corresponding eigenvectors. The principal components are then linear contrasts in the raw proportions and thus also subject to the difficulty of any correlation interpretation caused by the constant-sum constraint. It is interesting that one of Le Maitre's applications is to a markedly curved data set in a 17-dimensional simplex. Instead of investigating the possibility of regarding nonlinearity in the simplex as a form of natural variability he gives the departure from the linear an interpretation of a geological trend, in contrast to more recent views that visual trends in compositional data may be a matter of fantasy rather than fact (Butler, 1979).

The use of (iii) by Webb & Briggs (1966), with its implicit exchange of the restrictions of the simplex for the simpler restrictions of the positive orthant through the ratio transformation, is certainly a step in the direction of our own solution. Unfortunately it still suffers from its inherent linearity and the dependence of the resulting principal components on the choice of common divisor x_j.

2·2. *The log linear contrast approach*

The starting point for principal component analysis of compositional data suggested by an earlier resolution (Aitchison, 1981, 1982) of the constraint difficulty is, as indicated in § 1, the covariance matrix

$$\Omega_j = \text{cov}\,\{\log\,(x_{-j}/x_j)\}. \tag{2·1}$$

The nonlinearity of the logarithmic function opens up the possibility of coping with curvature in data sets and its approximate linearity over parts of its range also makes feasible the modelling of linear data sets. Even so, a major difficulty remains since different choices of the divisor x_j lead to different principal components. We shall find the underlying reason for this lack of invariance is our failure so far to identify the isotropic covariance structure in S^d, the counterpart of the spherically symmetric covariance matrix I_d associated with R^d.

We can gain insight into this invariance difficulty by working towards a more symmetric approach. First we note that in the use of Ω_j principal components take the form

$$\sum_{i \neq j} a_i \log\,(x_i/x_j) = \sum_{i=1}^{d+1} a_i \log x_i,$$

where $a_1 + \ldots + a_{d+1} = 0$. In other words, principal components can be expressed symmetrically as log linear contrasts of the $d+1$ proportions. A way of gaining symmetry, while retaining the log ratios so essential to the removal of the constraint difficulty, is to notice that, with $a_1 + \ldots + a_{d+1} = 0$,

$$\sum_{i=1}^{d+1} a_i \log x_i = \sum_{i=1}^{d+1} a_i \log\,\{x_i/g(x)\},$$

where $g(x) = (x_1 \dots x_{d+1})^{1/(d+1)}$, the geometric mean of the $d+1$ proportions. Thus study of the eigenvalues and eigenvectors of the covariance matrix of order $d+1$,

$$\Omega = \text{cov}\,[\log\{x^{(d+1)}/g(x)\}], \tag{2·2}$$

may be an appropriate principal component analysis procedure; see also the discussion of Aitchison (1982).

The matrix Ω is positive-semidefinite, its one zero eigenvalue having an associated eigenvector u_{d+1} consisting of $d+1$ units. The other d eigenvalues $\lambda_1, \dots, \lambda_d$, labelled in descending order of magnitude, are positive and the corresponding eigenvectors $a_1, \dots, a_d$, being orthogonal to u_{d+1}, yield log linear contrasts of the proportions as required. Thus we adopt a principal component system based on these d solutions of

$$(\Omega - \lambda I_{d+1})\, a = 0, \tag{2·3}$$

with $a_1, \dots, a_d$ scaled to form an orthonormal set.

We can now easily discover an equivalent, invariant asymmetric form based on the covariance matrix Ω_j. First we note a simple relationship between Ω and Ω_j, namely

$$\Omega_j = B_j \Omega B_j^{\mathrm{T}}, \tag{2·4}$$

where B_j is the unit matrix of order d widened to order $d \times (d+1)$ by the insertion of a column vector with each element -1 as the jth column. More specifically,

$$B_j = \begin{bmatrix} I_{j-1} & -u_{j-1} & 0 \\ 0 & -u_{d-j+1} & I_{d-j+1} \end{bmatrix}, \quad B_j B_j^{\mathrm{T}} = H_d, \tag{2·5}$$

the $d \times d$ constant matrix with all diagonal elements equal to 2 and all off-diagonal elements equal to 1, whatever the value of j. Premultiplying (2·3) by B_j and noting that $a = B_j^{\mathrm{T}} a_{-j}$, we obtain

$$(B_j \Omega - \lambda B_j) B_j^{\mathrm{T}} a_{-j} = 0,$$

that is $(\Omega_j - \lambda H_d)\, a_{-j} = 0$. Thus the asymmetric eigenvalue problem

$$(\Omega_j - \mu H_d)\, b = 0 \tag{2·6}$$

and the symmetric problem (2·3) produce identical eigenvalues and identical log linear principal components, through the relations $a = B_j^{\mathrm{T}} b$ and $b = a_{-j}$. Note that the eigenvectors $b_1, \dots, b_d$ of the asymmetric version are not orthonormal but are chosen to satisfy

$$b_i^{\mathrm{T}} H_d b_k = \begin{cases} 1 & (i = k), \\ 0 & (i \neq k). \end{cases}$$

In principal component analysis in R^d scalar multiples of the identity matrix I_d play a central role as the only covariance structures which remain invariant under the group of orthogonal transformations. Another way of viewing this is that for spherically symmetric distributions there can be no dimension-reducing explanation of variability through the use of principal component analysis. We can easily discover the corresponding isotropic covariance structure for compositional variability. The only form of Ω which is symmetric in all the components and which satisfies the condition of zero eigenvalue with u_{d+1} eigenvector is a scalar multiple of

$$G_{d+1} = I_{d+1} - \{1/(d+1)\}\, U_{d+1}, \tag{2·7}$$

where U_{d+1} is a $(d+1)\times(d+1)$ matrix with each element equal to 1. Then the corresponding form for Ω_j is given by (2·4) as a scalar multiple of $B_j B_j^T = H_d$ since $B_j U_{d+1} = 0$. The isotropic nature of G_{d+1} and H_d is confirmed by application of (2·3) and (2·6) to these covariance structures. Each yields d equal unit eigenvalues with freedom to choose log linear contrasts based on any d $(d+1)$-dimensional orthonormal eigenvectors orthogonal to u_{d+1}. Thus the equivalent methods of principal components analysis have the sensible feature that nothing is gained in terms of explanation of variability by their application to isotropic covariance structures.

As in standard principal component analysis the sum of the eigenvalues can be used as a measure of the total variability and the ratio $(\lambda_1 + \ldots + \lambda_c)/(\lambda_1 + \ldots + \lambda_d)$ as an indication of the proportion of this total variability embodied in the first c principal components. We note that the measure of total variability for compositional data can be expressed in the form

$$\sum_{i=1}^{d+1} \mathrm{var}\,[\log\{x_i/g(x)\}].$$

3. APPLICATIONS

We adopt here the symmetric version (2·3) and note that the only computational tool required is an algorithm for the determination of eigenvalues and eigenvectors of a symmetric matrix. For a given data set we replace Ω by its sample estimate. Let x_{ri} $(r = 1, \ldots, n;\ i = 1, \ldots, d+1)$ denote the proportion of the ith part in the rth replicate and let

$$z_{ri} = \log x_{ri}, \quad z_{r.} = \sum_{i=1}^{d+1} z_{ri}/(d+1), \quad y_{ri} = z_{ri} - z_{r.}.$$

The required estimate is then the sample covariance matrix S_y of the n vectors

$$y_r = (y_{r1}, \ldots, y_{r,d+1}) \quad (r = 1, \ldots, n).$$

Note that this is different from the sample covariance matrix S_z of the n vectors $z_r = (z_{r1}, \ldots, z_{r,d+1})$, the relationship being $S_y = G_{d+1} S_z G_{d+1}$, where G_{d+1} is defined in (2·7).

Table 1 gives the estimates S_y for the two data sets of Fig. 1 together with the nonzero eigenvalues and their corresponding eigenvectors. It is then a trivial mathematical exercise to convert the principal axes in the y-space to their triangular coordinate forms, as shown in Fig. 2. It is clear that the log linear contrast method faithfully depicts the curved one-dimensional pattern of data set (b), and gives as convincing a near-linear representation of the data set (a) as the specifically linear method shown in Fig. 1a.

Table 1. *Log linear contrast analysis for data of Fig.* 1

	Data set					
	(a) Steroid metabolites			(b) Aphyric Skye lavas		
Covariance matrix S_y	0·03790	0·00919	−0·04709	0·00593	0·01668	−0·02261
		0·06139	−0·07058		0·28370	−0·30038
			0·11767			0·32299
Eigenvalues	0·179	0·0375		0·606	0·00695	
Eigenvectors	−0·302	0·759		−0·046	0·815	
	−0·506	−0·641		−0·683	−0·448	
	0·808	−0·118		0·729	−0·367	

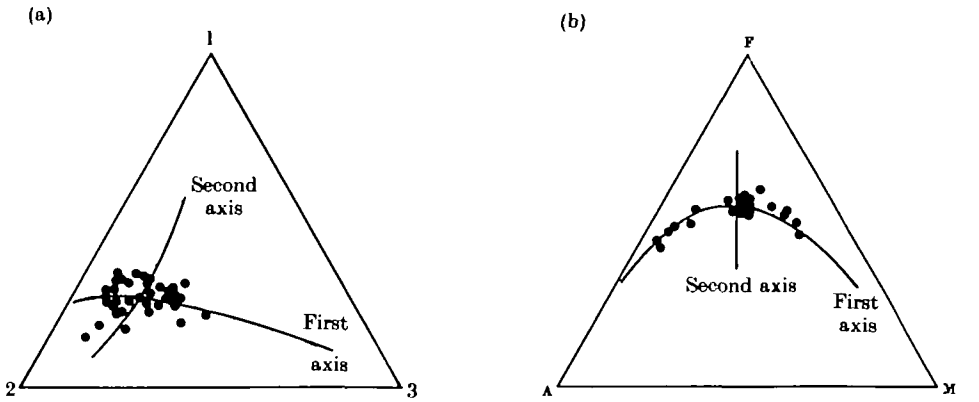

Fig. 2. Principal axes obtained by the log linear contrast method for the data sets of Fig. 1.

We now turn to two higher-dimensional applications. The first is to an extended form of the data set of Fig. 1a, fourteen-part compositions of steroid metabolites in urinary excretions of 37 healthy adults. The second data set consists of the breakdown of 26 days in the life of an academic statistician into the five activities of teaching, research including consultation, administration, leisure and sleep. The record was kept over a year with the 26 days selected randomly from workdays in alternate weeks to avoid carryover effects such as lack of sleep from one day to another and so to ensure approximate independence of the vectors. Space does not permit publication of these data sets but data listings of these and the two illustrative data sets are available on request from the author.

The simplest way of comparing the effectiveness of the log linear and linear contrast approaches of (2·3) and §2·1 is in the plots in Fig. 3 of first against second principal

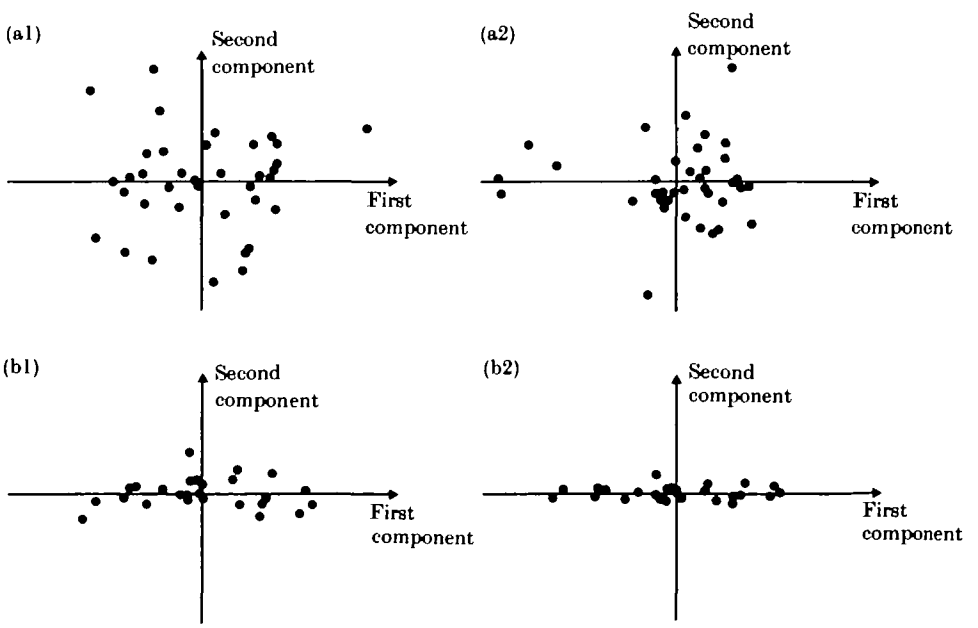

Fig. 3. Scattergrams of first against second principal components for (a) steroid metabolite compositions, (b) statistician's activities, obtained by (1) linear contrast method, (2) log linear contrast method.

components. We have plotted the components on the same scale so that the relation of horizontal to vertical scatter indicates the relative extent of the variability captured by the first and second components. For the apparently linear first data set the log linear contrast approach provides as adequate a description of the major variability as the linear contrast approach. Since the plot by the linear contrast method is simply a particular linear projection of the higher-dimensional data it is clear that the second data set has a curvature which the first two linear-contrast components fail to capture. On the other hand, the linear or elliptical nature of the log linear principal component plots indicates the success of the log linear contrast method in capturing the essential curvature in this compositional variability.

4. Discussion

Although success in four applications hardly constitutes an argument for the general advocacy of a procedure, we can report similar success in the study of ten geological data sets of 10- to 17-part major-oxide compositions, some decidedly curved, others apparently linear. Some of these results are reported elsewhere (Aitchison, 1983). It is certainly our impression that curved data sets are sufficiently frequent to demand that any general method should be able to cope with such sets. While it is true that the linear contrast method through its failures will often demonstrate that data sets are curved there seems no particular merit in this ability. Curved compositional data sets in a simplex sample space seem no less natural than linear or ellipsoidal data sets.

We have based our advocacy of the log linear contrast method on its ability to overcome the two difficulties of curvature and constant-sum constraint. There are, however, three further aspects of log linear contrast principal components worth noting.

(i) There is no known class of distributions for $x^{(d+1)}$ in S^d which gives tractable distributions for linear contrasts $\Sigma a_i x_i$. There is, however, a class over S^d of additive logistic normal distributions (Aitchison & Shen, 1980; Aitchison, 1982), that is with $\log(x_{-j}/x_j)$ having d-dimensional normal distributions, which seems to have a reasonable validity in modelling variability of compositions. For such models any set of log linear contrasts has a known distributional form, namely multivariate normal. Although this feature in itself may be unnecessary in any purely geometric view of data analysis it does have advantages if further investigation or modelling of the principal components is undertaken.

(ii) Even without this distributional background the log ratio covariance structure of compositions is readily interpretable (Aitchison, 1981, 1982) and such interpretations carry over to log linear contrasts. To say this is, of course, simply a reiteration of the fact that the log ratio approach fully removes the effect of the constraint in its formulation of a useful definition of covariance. This definition implies particular definitions of distance and orthogonality within S^d if we wish to link the log linear contrast approach with more geometric approaches to principal component analysis. For example, the appropriate distance $d(x, X)$ between two compositions x and X is given by

$$d^2(x, X) = \sum_{i=1}^{d+1} [\log\{x_i/g(x)\} - \log\{X_i/g(X)\}]^2.$$

We have here preferred the approach through covariance structure since this has proved the familiar stumbling block in compositional data analysis and it seems natural to attack other aspects of compositional analysis from a consolidated bridgehead.

(iii) A desirable feature of any form of compositional data analysis is an ability to study subcompositions (Aitchison, 1982), that is subvectors rescaled to give unit sum. One important requirement is an ability to quantify the extent to which a subcomposition retains a picture of the variability of the whole composition. One way of doing this is to compare the variability of the subcomposition, say of dimension c, with the extent of the total compositional variability captured by the first c principal components. The log ratio and log linear contrast approach is admirably suited to this task since for a subcomposition based on the subvector $x^{(c)} = (x_1, \ldots, x_c)$ of $x^{(d+1)}$ it does not involve the awkward ratio $x_i/(x_1 + \ldots + x_c)$ $(i = 1, \ldots, c)$, but the log ratios $\log(x_1/x_c), \ldots, \log(x_{c-1}/x_c)$ which are as easy to analyse as the parent composition. The development and application of this form of assessment of subcompositional analysis in geochemistry is reported elsewhere (Aitchison, 1983).

References

Aitchison, J. (1981). A new approach to null correlations of proportions. *J. Math. Geol.* **13**, 175–89.

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *J. R. Statist. Soc.* B **44**, 139–77.

Aitchison, J. (1983). Reducing the dimensionality of compositional data sets. *J. Math. Geol.* **15**. To appear.

Aitchison, J. & Shen, S. M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika* **67**, 261–72.

Butler, J. C. (1976). Principal component analysis using the hypothetical closed array. *J. Math. Geol.* **8**, 25–36.

Butler, J. C. (1979). Trends in ternary petrologic variation diagrams—fact or fantasy? *Am. Mineral.* **64**, 1115–21.

Chayes, F. (1960). On correlation between variables of constant sum. *J. Geophys. Res.* **65**, 4185–93.

Chayes, F. (1962). Numerical correlation and petrographic variation. *J. Geol.* **70**, 440–52.

Chayes, F. (1971). *Ratio Correlation*. University of Chicago Press.

Chayes, F. & Kruskal, W. (1966). An approximate statistical test for correlations between proportions. *J. Geol.* **74**, 692–702.

Darroch, J. N. (1969). Null correlations for proportions. *J. Math. Geol.* **1**, 467–83.

Darroch, J. N. & Ratcliff, D. (1970). Null correlations for proportions II. *J. Math. Geol.* **2**, 307–12.

Darroch, J. N. & Ratcliff, D. (1978). No association of proportions. *J. Math. Geol.* **10**, 361–8.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. New York: Wiley.

Gower, J. C. (1967). Multivariate analysis and multidimensional geometry. *Statistician* **17**, 13–28.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psych.* **24**, 417–41.

Le Maitre, R. W. (1968). Chemical variation within and between volcanic rock series—a statistical approach. *J. Petrol.* **9**, 220–52.

Miesch, A. T. (1969). The constant sum problem in geochemistry. In *Computer Applications in the Earth Sciences*, Ed. D. F. Merriam, pp. 161–77. New York: Plenum.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution and correlation among proportions. *Biometrika* **49**, 65–82.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space *Phil. Mag.* **2** (6th Series), 559–72.

Sarmanov, O. V. & Vistelius, A. B. (1959). On the correlation of percentage values. *Dokl. Akad. Nauk. SSSR* **126**, 22–5.

Vistelius, A. B. & Sarmanov, O. V. (1961). On the correlation between percentage values: Major component correlation in ferromagnesium micas. *J. Geol.* **69**, 145–53.

Webb, W. M. & Briggs, L. I. (1966). The use of principal component analysis to screen mineralogical data. *J. Geol.* **74**, 716–20.

[*Received September* 1981. *Revised June* 1982]