Mathematical Geology, Vol. 16, No. 6, 1984

Reducing the Dimensionality of Compositional Data Sets[1]

J. Aitchison[2]

The high-dimensionality of many compositional data sets has caused geologists to look for insights into the observed patterns of variability through two dimension-reducing procedures: (i) the selection of a few subcompositions for particular study, and (ii) principal component analysis. After a brief critical review of the unsatisfactory state of current statistical methodology for these two procedures, this paper takes as a starting point for the resolution of persisting difficulties a recent approach to principal component analysis through a new definition of the covariance structure of a composition. This approach is first applied for expository purposes to a small illustrative compositional data set and then to a number of larger published geochemical data sets. The new approach then leads naturally to a method of measuring the extent to which a subcomposition retains the pattern of variability of the whole composition and so provides a criterion for the selection of suitable subcompositions. Such a selection process is illustrated by application to geochemical data sets.

KEY WORDS: principal-component analysis, subcomposition, geochemistry, closed arrays, compositional data.

CURRENT PROCEDURES AND THEIR DIFFICULTIES

It is well recognized that the analysis of compositional data, consisting of vectors of proportions adding to unity, for example the percentages of various oxides in rock specimens, should play a central role in the understanding of geological processes. The fact that many of these geochemical compositions are high-dimensional, for example, consisting of 10 or more oxides, and the unfortunate inability of the human eye to see in more than three dimensions have naturally led geologists to consider devices for obtaining lower-dimensional insights into the nature of their data sets. Awareness of the difficulties of interpretation associated with the constant-sum constraint or the so-called effect of "closure" seems to reinforce this wish to reduce the effective dimensionality of the data set.

[1]Manuscript received 10 February 1982; revised 2 February 1983.

[2]Department of Statistics, University of Hong Kong, Hong Kong (Permanent address) and Princeton University.

0020-5958/84/0800-0617$03.50/0

Table 1. Five-Part Compositions of Fifteen Specimens of Hongkongite

	Percentages of five components				
Specimen no.	1	2	3	4	5
1	43.4	40.8	1.9	9.4	4.5
2	47.4	18.3	13.8	7.7	12.8
3	34.1	8.9	36.0	6.5	14.5
4	45.7	23.4	11.1	10.7	9.1
5	52.4	30.0	2.6	10.3	4.7
6	49.8	22.0	9.2	11.6	7.4
7	45.4	9.8	25.3	8.9	10.6
8	43.2	42.3	1.8	5.8	6.9
9	46.7	11.3	21.9	12.6	7.5
10	51.1	20.7	9.5	9.7	9.0
11	54.6	21.2	6.7	10.2	7.3
12	31.4	8.6	40.2	10.7	9.1
13	37.2	11.8	30.2	12.2	8.6
14	52.5	28.6	3.1	11.0	4.8
15	49.3	30.1	4.4	6.8	9.4

We consider here two popular dimension-reducing procedures, subcompositional analysis and principal component analysis. So that the reader may easily follow the new techniques of these analyses and appreciate the simplicity of the computations involved we develop the statistical argument against the background of a specific, manageable data set. Moreover, in order that the geologist reader may come to the problems with no preconceptions we use a data set unfamiliar to him, the analyses of fifteen specimens of a new "rock" (let us name it for convenience, hongkongite) into five constituent components labelled simply 1, 2, 3, 4, 5. Table 1 records the detailed compositions.

Subcomposition

A standard dimension-reducing artefact is to consider a subcomposition of the complete composition. The ternary and tetrahedral diagrams, so familiar in the geological literature, are examples of such devices. If $(x_1, \ldots, x_{d+1})$ is a complete $(d+1)$-part composition with the constraint

$$x_1 + \cdots + x_{d+1} = 1 \tag{1}$$

then any subvector with its elements rescaled so that their sum is 1 is a subcomposition. For example, the subvector $(x_1, \ldots, x_{c+1})$ gives a subcomposition

$$x_i/(x_1 + \cdots + x_{c+1}) \qquad (i = 1, \ldots, c+1) \tag{2}$$

There appears to have been no attempt to study criteria for the selection of subcompositions such as (CaO, Na_2O, K_2O) in the familiar *CNK* ternary diagram. Indeed, the objective of the whole subcompositional analysis is often left vague and seems to range from attempts to retain within the subcompositional data as much as possible of the variability in the complete compositional data to attempts to discover subcompositions which display little variability within a rock type but which are radically different between rock types, the latter objective being appropriate for classification purposes. Our interest here is the former purpose, specifically to try to identify subcompositions that retain, in some sensibly definable way, as much of the total variability in the entire composition as possible. For example, we may ask the extent to which the subcompositions based on components 1, 2, 3, and 2, 4, 5 of Table 1 and displayed in the ternary diagrams of Fig. 1 retain the variability of the complete five-part compositions of Table 1.

Any composition can be represented by a point in a sample space which is mathematically the d-dimensional simplex

$$S^d = \{(x_1, \ldots, x_d): x_i \geqq 0 \qquad (i = 1, \ldots, d), \qquad x_1 + \cdots + x_d < 1\} \tag{3}$$

The dimension of this space is d, not $d + 1$, because knowledge of the values of just d of the components determines the value of the remaining component through the constant-sum constraint (1). It is, however, sometimes convenient to express this simplex in a symmetric form

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

as a subspace of $(d + 1)$-dimensional real space. The process of formation of a subcomposition from a composition is then simply a process of projecting the

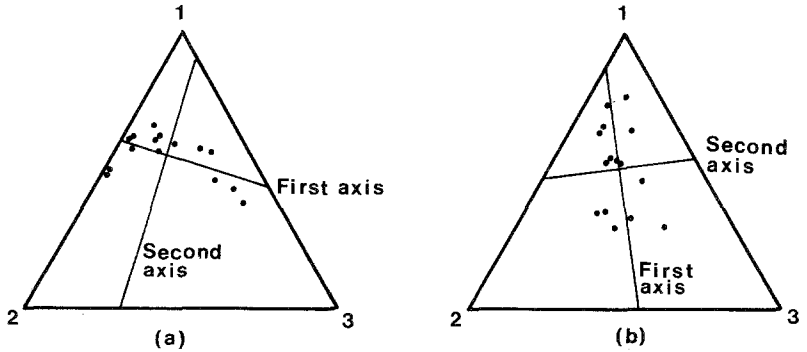

Fig. 1. Ternary diagrams for the subcompositions (1, 2, 3) and (2, 4, 5) of the hongkongite data sets, with the principal axes obtained by linear principal-component analysis.

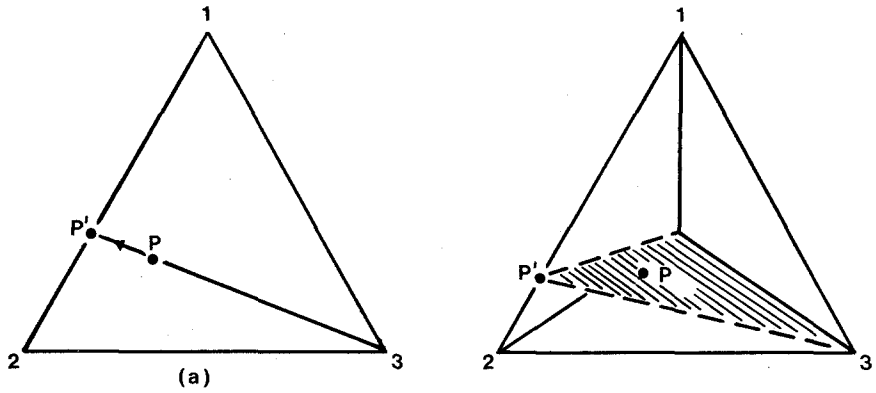

Fig. 2. Formation of a subcomposition as a projection of a composition on to a subsimplex from the complementary subsimplex.

point representing the composition onto a suitable subsimplex of S^d from the complementary subsimplex. Figure 1 shows two simple examples. In Fig. 2(a) P' is the projection of the three-part, or two-dimensional, composition P from the (trivial) subsimplex 3 to the subsimplex 12, and so represents the subcomposition $\{x_1/(x_1 + x_2), x_2/(x_1 + x_2)\}$. In Fig. 2(b) P' is the point of intersection of the plane through the subsimplex 34 and the compositional point P with the subsimplex 12, and so represents the subcomposition $\{x_1/(x_1 + x_2), x_2/(x_1 + x_2)\}$ within the simplex 12.

The fact that subcompositions are interpretable as linear projections in the above way means that any convex set of compositions must produce a convex set of subcompositions. A concave set, however, may project into either a concave or a convex set depending on the orientation of the set relative to the two subsimplices involved in the projecting process. A subcompositional scattergram of concave shape implies a concave complete composition scattergram, whereas a convex-shaped subcompositional pattern may have arisen from a concave compositional pattern. Obviously a substantial amount of information may be being lost in the process of focusing on subcompositions.

Principal-Components Analysis

The device of examining subcompositions has a long history dating back at least to Becke (1897) and the introduction of ternary diagrams. Although a version of principal-component analysis dates back to Pearson (1901), the full development of the methodology is of much more recent origin (Hotelling, 1933), and it is only in the last two decades that it has been explored as a possible dimension-reducing device in the study of patterns of variability of high-dimensional com-

positional data. For geochemical applications see, for example, Butler (1976), Chayes and Trochimczyk (1978), Hawkins and Rasmussen (1973), Le Maitre (1962, 1968, 1976), Miesch (1980), Roth, et al. (1972), Till and Colley (1973), Trochimczyk and Chayes (1977, 1978), Vistelius et al. (1970), Webb and Briggs (1966).

We regard principal-component analysis simply as a dimension-reducing tool, asking the question: Are there a few functions of the many original variables which in some sense capture the essential variability in the data? Any subject-based interpretation which may be placed on the computed functions would then be regarded only as a speculative hypothesis to be subjected to proper testing with a further, independent data set.

Like so many other statistical procedures for compositional data, principal-component analysis has been snarled by that supposedly insuperable difficulty of analyzing and interpreting such data, variously known as the constant-sum problem (Chayes, 1960) and the problem of closure (Chayes and Kruskal, 1966). The fact that each compositional vector $(x_1, \ldots, x_{d+1})$ is subject to the constraint (1) has wrought the customary and now expected havoc with the unmodified carry-over of a standard statistical procedure, designed for, and doing noble service in the analysis of, data sets in a real or Euclidean sample space, into problems of data sets confined to a completely different and awkward sample space, namely a simplex. Since, in particular, any analysis based on covariances between *raw* proportions is of doubtful value (Aitchison, 1981; Kork, 1977) the popular current method of principal-component analysis (Le Maitre, 1968) which uses the d nonzero eigenvalues and corresponding eigenvectors of the singular covariance matrix computed from raw proportions must be open to the same criticism. Moreover, such a procedure is mathematically linear in nature and so cannot hope to capture patterns of curved variability which are commonly present in many compositional data sets. For example, this procedure produces the principal axes shown for the data sets of Fig. 1 and the plots of first against second and first against third principal components in Fig. 3 for the complete compositions of Table 1. It is clear from Fig. 1a and Fig. 3 that the method is failing to describe the inherent curvature of the data set.

Aitchison (1983) has reviewed the unsatisfactory nature of the versions of principal-component analysis currently used on geochemical data sets, and has proposed a radically different approach which has the capability of capturing the structure of the variability of the data set, whereas the other methods do not.

The purpose of this paper is to determine the strength of this new method in geological work with compositional data. We find that not only does it provide a promising new approach to principal-component analysis but also contains a key which opens a door to the quantitative assessment of the effectiveness of subcompositional analysis.

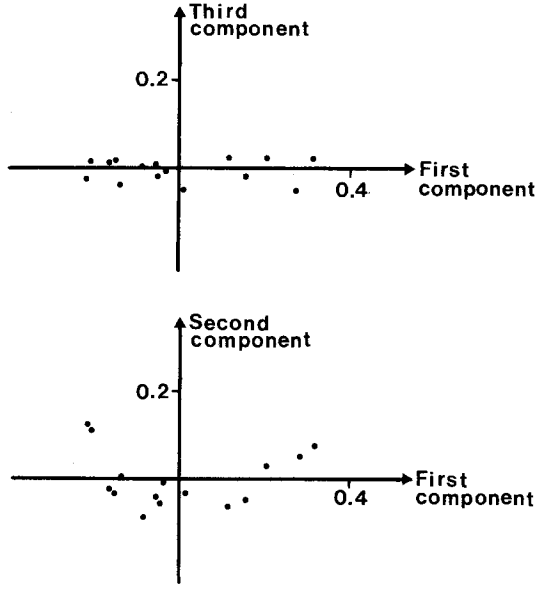

Fig. 3. Scattergrams of first against second, and first against third, linear principal components for the hongkongite data set.

LOGLINEAR-CONTRAST PRINCIPAL COMPONENTS

A solution to the well-known constant-sum and covariance-interpretation problem of compositional data is to study the dependence structure of compositions through the covariance matrix of logarithms of the ratios of components (Aitchison, 1981, 1982). To see exactly how this may be implemented in principal-components analysis it is important to appreciate how the covariance matrix used relates to the original data matrix.

Table 1 is a typical data array or matrix X, in general having n rows and $d + 1$ columns and with entry x_{rc} in the rth row and cth column. The rth row of the table then constitutes the rth *replicate* composition with the row sum equal to 1, the constant-sum constraint. The entries in the cth column all refer to a particular component c and there is no specific constraint on this column sum. The sample covariance matrix S_X formed from this array X of raw proportions can then be expressed conveniently in a form for comparison with later constructs as

$$(n - 1)\, S_X = X^T G_n X \tag{4}$$

where

$$G_n = I_n - (1/n)\, J_n \tag{5}$$

with I_n the usual identity matrix and J_n the $n \times n$ matrix with every element 1. The d nonzero eigenvalues of S_X and their corresponding eigenvectors form the principal component approach of Le Maitre (1968).

From the original array X we first form an intermediate array Z of logarithms of the raw proportions, so that

$$z_{rc} = \log x_{rc}$$

and then form our required $n \times (d + 1)$ array Y by subtracting from each entry in Z the corresponding *row* average, and so

$$y_{rc} = z_{rc} - \sum_{b=1}^{d+1} z_{rb}/(d + 1) \tag{6}$$

$$= \log \{x_{rc}/(x_{r1}, \ldots, x_{r,d+1})^{1/(d+1)}\} \tag{7}$$

We assume that all entries in X are positive so that the arrays Z and Y can be formed, and we delay until the final section a discussion of how to deal with zero components. The array Y consists of the required logratio entries and in a symmetric form, the common divisor for the composition in a row being the geometric mean of the different components of the composition. The covariance matrices S_Y and S_Z associated with the arrays Y and Z can be computed, as in (4), by

$$\begin{aligned}(n - 1)\, S_Y &= Y^T G_n Y \\ (n - 1)\, S_Z &= Z^T G_n Z\end{aligned} \tag{8}$$

Note that S_Y and S_Z are not identical since in the formation of the corrected cross-product matrix associated with Z the corrections are the column averages, not the row averages used in the formation of the y_{rc} in (6). Since in fact

$$Y = ZG_{d+1} \tag{9}$$

the difference between S_Y and S_Z can be seen in their relationship

$$S_Y = G_{d+1} S_Z G_{d+1} \tag{10}$$

For the array X of Table 1, we give in Table 2 the corresponding logratio array Y and the covariance matrix S_Y. It is important to realize that the zero row sums of the array Y are not transformed versions of the constant row-sum constraints of X but are simply an overspecification consequence of wishing to work symmetrically in terms of logratios. A completely equivalent asymmetric logratio approach without any such restrictions is possible (Aitchison, 1983), but for principal-component analysis the symmetric approach seems preferable. As in Aitchison (1981, 1982) the logratio covariance approach effectively disposes of the constant-sum problem.

The row sums of S_Y are also all zero, so that one of the $d + 1$ eigenvalues is

Table 2. Data Array Y Corresponding to Data Array X of Table 1 and the Covariance Matrix S_Y

Specimen no. r	y_{rc}				
	$c = 1$	2	3	4	5
1	1.397	1.336	−1.731	−0.132	−0.869
2	1.062	0.111	−0.172	−0.755	−0.247
3	0.760	−0.583	0.815	−0.897	−0.095
4	1.030	0.361	−0.385	−0.422	−0.584
5	1.520	0.962	−1.484	−0.107	−0.892
6	1.174	0.357	−0.515	−0.283	−0.733
7	1.040	−0.493	0.456	−0.589	−0.414
8	1.408	1.387	−1.770	−0.600	−0.426
9	1.063	−0.356	0.306	−0.247	−0.766
10	1.197	0.293	−0.486	−0.465	−0.540
11	1.347	0.401	−0.751	−0.331	−0.665
12	0.673	−0.622	0.920	−0.404	−0.566
13	0.787	−0.361	0.579	−0.328	−0.677
14	1.478	0.871	−1.351	−0.085	−0.914
15	1.310	0.816	−1.107	−0.671	−0.348

$$S_Y = 10^{-2} \times \begin{bmatrix} 7.089 & 16.071 & -23.001 & 2.907 & -3.066 \\ 16.071 & 46.173 & -61.826 & 5.841 & -6.259 \\ -23.001 & -61.826 & 84.267 & -8.773 & 9.333 \\ 2.907 & 5.841 & -8.773 & 5.890 & -5.865 \\ -3.066 & -6.259 & 9.333 & -5.865 & 5.857 \end{bmatrix}$$

zero with corresponding eigenvector the vector j_{d+1} of $d + 1$ units. The method of principal-component analysis advocated by Aitchison (1983) then uses the d positive eigenvalues $\lambda_1, \ldots, \lambda_d$ of S_Y, arranged in descending order of magnitude, and the corresponding unit-length eigenvectors $a_1, \ldots, a_d$, easily obtainable by any standard eigenvalue package. Each eigenvector is then orthogonal to j_{d+1}, so that the sum of its elements is zero. The consequence of this is that any principal component $a^T y$ is expressible as a loglinear contrast of the raw proportions, that is, a linear combination of the logarithms of the proportions with coefficients summing to zero.

Table 3 shows the four positive eigenvalues of S_Y and their associated eigenvectors for the full compositions of hongkongite data of Table 1, and Fig. 4 shows the plots of first against second and first against third principal components. In contrast to Fig. 3 there is no evidence of curvature in the pattern in Fig. 4; the scatter is now decently elliptical. The ability of the loglinear-contrast approach to principal-component analysis to describe curved data patterns is, of

Table 3. Positive Eigenvalues and Associated Eigenvectors of Covariance Matrix S_Y for the Hongkongite Data and the Percentage of Total Variability Captured by the First c Principal Components

c	1	2	3	4
Eigenvalue λ_c	1.38	0.0987	0.0136	0.000098
Eigenvector a_c	0.212	0.073	0.789	0.356
	0.574	-0.151	-0.559	0.367
	-0.781	0.067	-0.218	0.372
	0.086	0.701	-0.097	-0.540
	-0.091	-0.690	0.085	-0.554
Percentage of total variability	92.5	99.1	99.99	100

course, due to the nonlinear logarithmic transformation involved. This is well illustrated on the compositions of Fig. 1a; the principal axes when transformed into the triangle by the logistic transformation $x_c = \exp(y_c)/\{\exp(y_1) + \exp(y_2) + \exp(y_3)\}$ $(c = 1, 2, 3)$ are, as seen in Fig. 5a, very effective in capturing the curvature of the data. The logarithmic transformation is, however, approximately linear over part of its range, and so it is still capable of capturing the more linear data pattern of Fig. 1b, as shown in Fig. 5b.

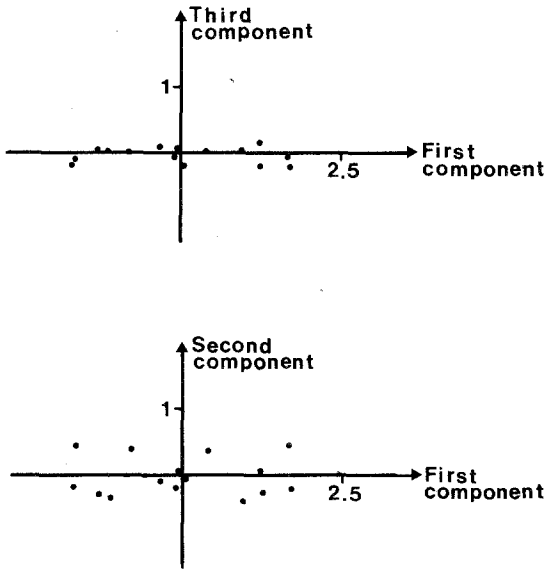

Fig. 4. Scattergrams of first against second, and first against third, loglinear-contrast principal components for the hongkongite data set.

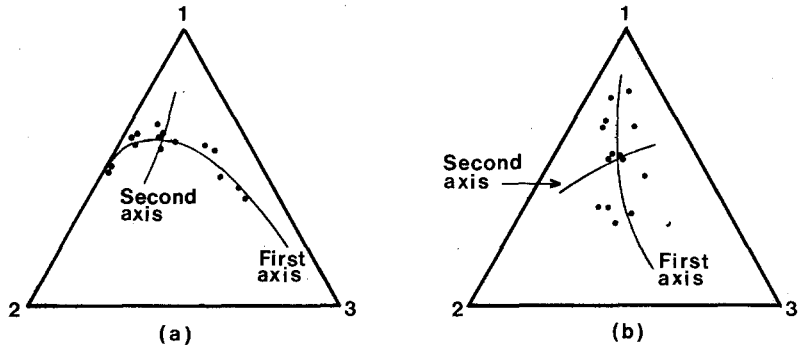

Fig. 5. Principal axes by the loglinear-contrast method for subcompositions (1, 2, 3) and (2, 4, 5) of the hongkongite data set.

As a measure of the total variability of a compositional data set we may still take the sum $\lambda_1 + \cdots + \lambda_d$ of the eigenvalues of S_Y, or equivalently, trace S_Y. This measure is a sample estimate of the sum

$$\sum_{c=1}^{d+1} \text{var}\left\{\log \frac{x_c}{g(x)}\right\} \tag{11}$$

of the variances of logratios, where the common divisor $g(x)$ is the geometric mean of the components. The proportion of the total variability captured by use of only the first c principal components is then

$$(\lambda_1 + \cdots + \lambda_c)/(\lambda_1 + \cdots + \lambda_d) \tag{12}$$

Table 3 also shows the way in which this proportion increases with c for the hongkongite data set. Thus the first principal component alone captures 92.5%, and together with the second principal component 99.1%, of the total variability. There is, of course, no point of comparing the corresponding performance, 85.4 and 97.6%, of linear principal components with those of our loglinear approach since, as we hope is now amply evident, the measure of total variability used in the linear approach is not appropriate to the description of curved or concave data sets.

APPLICATIONS OF PRINCIPAL COMPONENT ANALYSIS

It could readily be argued that the illustrative hongkongite data set has been contrived to produce the dramatic advantages of the loglinear-contrast method just described. We now, therefore, turn to four published geochemical compositional data sets to allow a comparison between the uses of S_X and S_Y as the covariance structures underlying the analyses, in other words between the Le Maitre (1968) linear and the Aitchison (1983) loglinear-contrast forms of principal-component analysis. Details of the data sets used are shown in Table 4.

Table 4. The Four Data Sets Analyzed

Data set no.	Reference	Description	Number of	
			Specimens	Major oxides
1	Nisbet, Bickle, and Martin (1977)	Lavas of the Belingwe Greenstone Belt, Rhodesia: Table 3	60	11
2	Carr (1981)	Igneous rocks of the southern Sydney Basin of New South Wales	102	10
3	Steiner (1958)	Effusive rocks of the Taupo volcanic association: Table 1, nos. 1–45, omitting no. 10, which has information missing	44	11
4	Thompson, Esson, and Duncan (1972)	Eocene lavas of the Isle of Skye, Scotland: Table 2 basalts	32	10

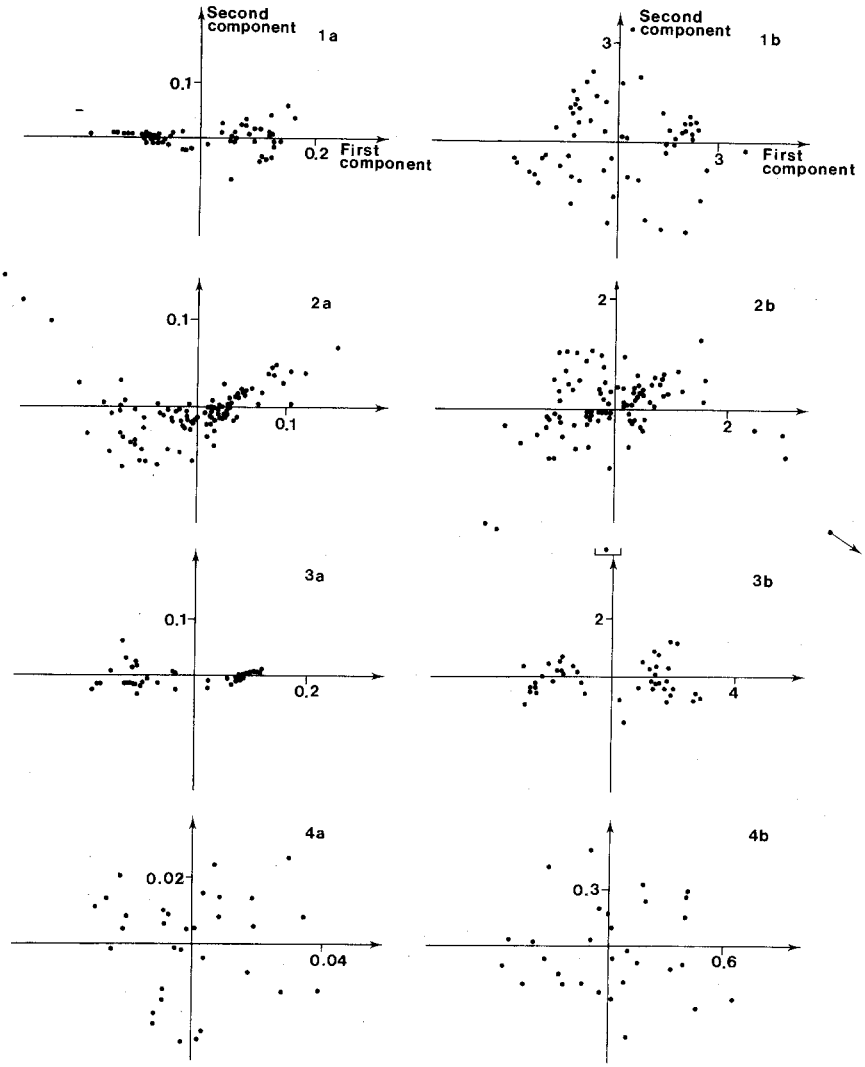

Fig. 6. Scattergrams of first against second principal components for the four data sets of Table 4 by (a) the linear method, and (b) the loglinear-contrast method.

The simplest way of providing a comparison is by presenting side by side the scattergrams of first against second principal components for the covariance matrices S_X and S_Y.

Only a short comment need be made here. Whereas for data sets 1, 2, 3, there is obvious curvature persisting in the principal component diagram by the linear-contrast method, the fact that the loglinear-contrast method captures this

Table 5. Subcompositions of Dimensions 2 and 3 Retaining the Highest Percentages of the Total Variability and the Corresponding Percentages for the First Two and Three Loglinear-Contrast Principal Components

Data set	Subcompositions of dimension 2	Percentage retained	Subcompositions of dimension 3	Percentage retained
1	MgO, K_2O, Na_2O	75.4	MgO, K_2O, Na_2O, P_2O_5	83.3
	FeO, K_2O, Na_2O	63.2	MgO, K_2O, Na_2O, TiO_2	81.4
	SiO_2, K_2O, Na_2O	62.9	MgO, K_2O, Na_2O, CaO	80.8
	Principal components	88.2	Principal components	95.2
2	MgO, K_2O, TiO_2	53.1	MgO, K_2O, TiO_2, MnO	66.4
	MgO, K_2O, P_2O_5	48.9	MgO, K_2O, TiO_2, P_2O_5	62.7
	MgO, K_2O, MnO	47.8	MgO, K_2O, TiO_2, Na_2O	62.2
	Principal components	74.4	Principal components	84.9
3	MgO, K_2O, FeO	60.5	MgO, K_2O, FeO, Na_2O	72.2
	MgO, K_2O, Na_2O	59.3	MgO, K_2O, FeO, SiO_2	70.8
	MgO, K_2O, Fe_2O_3	58.5	MgO, K_2O, FeO, Fe_2O_3	69.0
	Principal components	90.6	Principal components	95.3
4	MgO, K_2O, TiO_2	63.1	MgO, K_2O, TiO_2, P_2O_5	75.4
	MgO, K_2O, P_2O_5	58.5	MgO, K_2O, TiO_2, MnO	74.9
	MgO, K_2O, MnO	52.6	MgO, K_2O, TiO_2, CaO	71.1
	Principal components	82.4	Principal components	95.0

curved pattern of variability is demonstrated by the more elliptical scatter in the principal component diagram. For data set 4 for which the principal-component scattergram of the linear method is elliptical in character it can be seen that the loglinear-contrast method is just as satisfactory in this respect.

The percentages of total variability captured by the first two and first three loglinear-contrast principal components are reported later in Table 5.

ASSESSING THE EFFECTIVENESS OF SUBCOMPOSITIONAL ANALYSIS

We are now in a position to reexamine the nature of subcompositional analysis, through a simple relationship that it bears to principal-component analysis. We first recall that loglinear principal-component analysis of compositions simply searches out particular forms from among all loglinear contrasts of the raw proportions. Without loss of generality we may consider the subcomposition (2) formed from the leading subvector $(x_1, \ldots, x_{c+1})$ of the complete composition $(x_1, \ldots, x_{d+1})$. Such a subcomposition is technically a composition with dimension c, smaller than the dimension d of the original composition, and so has a logratio form $Y_1, \ldots, Y_{c+1}$ similar to (7) with

$$Y_i = \log \{x_i/(x_1, \ldots, x_{c+1})^{1/(c+1)}\} \qquad (i = 1, \ldots, c+1) \qquad (13)$$

These logratios $Y_1, \ldots, Y_{c+1}$ are obviously loglinear contrasts of the raw proportions and so belong to the class of functions from which our loglinear-contrast principal components are selected. It therefore follows that the proportion of total compositional variability retained by a subcomposition of dimension c will be at most equal to, and generally less than, that retained by the first c loglinear-contrast principal components.

The measure of variability retained by the subcomposition is simply the total variability of $(Y_1, \ldots, Y_{c+1})$ regarded as a composition and so is, by (11), the trace of the covariance matrix of $(Y_1, \ldots, Y_{c+1})$. Since

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_{c+1} \end{bmatrix} = [G_{c+1}\ 0] \begin{bmatrix} y_1 \\ \vdots \\ y_{d+1} \end{bmatrix}, \tag{14}$$

where $y_i = \log \{x_i/(x_1, \ldots, x_{d+1})^{1/(d+1)}\}$ $(i = 1, \ldots, d+1)$, this measure of variability is estimated by the trace of

$$[G_{c+1}\ 0]\ S_Y \begin{bmatrix} G_{c+1} \\ 0 \end{bmatrix} = G_{c+1}\ S_Y^{(c+1)}\ G_{c+1} \tag{15}$$

where $S_Y^{(c+1)}$ is the appropriate submatrix of S_Y, for this subcomposition the leading $(c+1) \times (c+1)$ submatrix of S_Y.

Thus, for any subcomposition of dimension c we can arrive at two measures of its effectiveness in retaining the inherent variability of the complete composition. The first is the proportion of the total variability retained, namely

$$\text{trace } [G_{c+1}\ S_Y^{(c+1)}\ G_{c+1}]/(\lambda_1 + \cdots + \lambda_d) \tag{16}$$

The second is the proportion of the maximum variability attainable by c loglinear contrasts, that is by the first c principal components, and so takes the form

$$\text{trace } [G_{c+1}\ S_Y^{(c+1)}\ G_{c+1}]/(\lambda_1 + \cdots + \lambda_c) \tag{17}$$

We can now see that if we wish to select from all the possible subcompositions of dimension c the one which is best by either of the above criteria we have to find a subcomposition with maximizes

$$\text{trace } [G_{c+1}\ S_Y^{(c+1)}\ G_{c+1}] = \text{trace } S_Y^{(c+1)} - (c+1)^{-1}\ j_{c+1}^T\ S_Y^{(c+1)}\ j_{c+1} \tag{18}$$

This last form is very simple to compute since trace $S_Y^{(c+1)}$ is the sum of the diagonal elements of the submatrix and $j_{c+1}^T S_Y^{(c+1)} j_{c+1}$ is the sum of all its elements. For a given matrix S_Y it is, therefore, easy to organize an algorithm to compute the value of (18) for each possible subcomposition of a given dimension.

For example, for the subcomposition (2, 4, 5) of the hongkongite data of

Table 1, the appropriate $S_Y^{(c+1)}$ matrix is the 3 × 3 matrix

$$10^{-2} \times \begin{bmatrix} 46.173 & 5.841 & -6.259 \\ 5.841 & 5.890 & -5.865 \\ -6.259 & -5.865 & 5.857 \end{bmatrix}$$

formed from rows and columns 2, 4, and 5 of S_Y of Table 2, with trace 0.57960 and sum of elements 0.45394. Thus, for this subcomposition the measure (18) of retained variability is $0.57960 - \frac{1}{3} \times 0.45394 = 0.428$. From Table 3 the total measure of variability for the whole composition is 1.493, the sum of all four eigenvalues, and so by (16) this subcomposition retains only 28.7% of the total variability. Similarly, by (17) with $c = 2$, we can say that the subcomposition attains only 28.9% of what is attainable by the first two principal components. In this particular example the subcompositions retaining the greatest amount of variability are (1, 2, 3), (2, 3, 4), (2, 3, 5) with percentage measure (16) at 92.1, 89.8, 87.1, compared with 99.1% for the first two loglinear-contrast principal components.

Since in (14) the vector y could be replaced by the vector $z = (\log x_1, \ldots, \log x_{d+1})$, it follows that in the above method of determining the measure (18) of variability of a subcomposition, S_Z could be used instead of S_Y. There seems nothing to be gained, however, by such a substitution since S_Y is required in the determination of the eigenvalues.

APPLICATION TO GEOCHEMICAL DATA SETS

We have applied the criterion for the selection of subcompositions to find the top subcompositions of dimensions 2 and 3 for the four geochemical data sets of Table 4. For these two dimensions and for data sets 1 and 3 this involved the calculation of (18) for $\binom{11}{3} = 165$ and $\binom{11}{4} = 310$ submatrices of S_Y; the corresponding numbers for data sets 2 and 4 are 120 and 210 submatrices. A BASIC program on a Wang 2200 PCS-4 desk minicomputer completed these tasks in between five and fifteen minutes.

The top three subcompositions, together with the proportion (16) of total variability retained by each and, for comparison purposes, the proportions of total variability retained by the first two and first three loglinear-contrast principal components, are shown for each of the four data sets in Table 5. It is interesting that on the whole the best subcompositions perform rather poorly as retainers of variability compared with the corresponding set of principal components. Moreover the popular (CaO, Na_2O, K_2O) subcomposition displayed in *CNK* ternary diagrams, as for example in Le Maitre (1962), does not appear in Table 5, and indeed retains only 62, 23, 28, 29% of total variability for the data sets 1, 2, 3, 4, respectively. The top subcompositions (MgO, K_2O, FeO) and (MgO, K_2O,

FeO, Na_2O) for data set 3 are recognizable as forms of (alkali, F, M) diagrams but retain only 61 and 72% of total variability, compared with the 91 and 95% retained by the corresponding sets of principal components.

Some further insights into subcompositional analysis may be gained by confining attention to the selection of the top subcomposition of dimension 2. Note that for data set 1 the selection of (MgO, K_2O, Na_2O) is further supported by the appearance of these three oxides in the top three (indeed the top seven) subcompositions of dimension 3. A similar comment applies to the other three data sets.

Although the dominant role of MgO and K_2O in all of these top subcompositions may appear very reasonable, the presence of TiO_2 as the third oxide for data sets 2 and 4 probably comes as a surprise to many geologists. They might argue that TiO_2 contributes only minutely to the total variance of any set of rock analyses. For example, for data set 2 the 10 oxides in descending order of the variances of their raw proportions are

$$SiO_2, MgO, CaO, Al_2O_3, FeO, K_2O, Na_2O, TiO_2, MnO, P_2O_5$$

with the variance of the 9th-placed TiO_2 just over 2% of that of SiO_2. The surprise is, therefore, probably ascribable to a conditioned mode of thinking in terms of variance and covariances of raw proportions and we now know, from the wide literature on the constant-sum problem, that very little of interpretable value comes out of such a mode.

It is important to realize that there are no physical entities corresponding to "variance" and "covariance." They are simply concepts, inventions of our own minds, in our attempts to find satisfactory explanations of natural phenomena: what definitions we try out are entirely at our disposal. If we are to grasp the advantage that the logratio and logcontrast approach to compositional data analysis provides in its removal of constant-sum difficulties, then we must begin to think in terms of a new concept of *relative* variance. For component x_c of a composition $(x_1, \ldots, x_{d+1})$ this is defined as

$$\text{var}\,\{\log [x_c/(x_1, \ldots, x_{d+1})^{1/(d+1)}]\} \tag{19}$$

where variation of x_c is measured relative to the variation of the geometric mean of all $d + 1$ components. The reader who finds difficulty in accepting an apparently strange definition for a measure of variability should reflect that for over a century the use of lognormal models in describing the variability of a positive measurement x has successfully accepted var (log x) as a measure of variability. The simplex is more complicated than the set of positive numbers, and it is not surprising that a successful measure of variability should turn out to be a little more sophisticated.

The changed viewpoint can make a radical difference to the interpretation of data. For example, for data set 2 the 10 oxides in descending order of magni-

tude of relative variance (19) are

$$MgO, TiO_2, K_2O, MnO, P_2O_5, CaO, Na_2O, FeO, Al_2O_3, SiO_2$$

with TiO_2 now in second position and with a relative variance about eight times that of SiO_2, which is now last in order. Note also that after MgO, TiO_2, and K_2O, the oxides MnO and P_2O, small in absolute variation, are now contenders for consideration in terms of relative variation.

Data set 1 is a mixture of rock types with

(1.1) 23 peridotitic komatites

(1.2) 23 basaltic komatites and magnesium basalts

(1.3) 14 in three other categories

Similarly data set 2 consists of

(2.1) 65 Permian igneous rocks

(2.2) 37 post-Permian igneous rocks

It could be argued, therefore, that the analyses should be carried out on the separate subsets. We have in fact done this, except for subset (1.3) which is too small, and can report the results briefly. The principal-component analyses display broadly the same features as for the full data sets. In the analysis of this section to determine the best variance-retaining subcomposition of dimension 2 the result is the same as in the complete data sets for subsets (1.2) and (2.2). For subset (1.1) the best subcomposition is (K_2O, Na_2O, P_2O_5), and so P_2O_5 replaces MgO, and for subset (2.1) the best subcomposition is (MgO, K_2O, CaO) with CaO replacing TiO_2.

DISCUSSION

Although the loglinear-contrast approach to principal-component and subcompositional analyses has undoubtedly attractive features and appears successful in the applications of the previous sections, it is far from being a complete answer to dimension-reducing problems in compositional data analysis. One drawback is that we cannot take logarithms of zero proportions. We avoided this issue in all our applications by choosing data sets without zero entries. The treatment of zeros in such analyses must depend largely on their nature. If there are only a few of the trace or rounding-off variety then their replacement by half, or some fraction of, the lowest recordable value will allow an analysis, and variation of the fraction will allow an assessment of the sensitivity of the conclusions to this replacement value. If there is a substantial number of zeros, say in one particular component, then some form of conditional modeling along the lines suggested in Aitchison (1982, Sect. 7.4) would seem essential to describe adequately this

probability concentration on zero in the pattern of variability, and this requirement would apply equally to the linear approach. This seems to be the best advice available at present. We hope to report later on an extensive study in progress on the effects of different treatments of zeros in compositional data analysis.

The loglinear-contrast approach is not a panacea for the treatment of all curved data sets. While it may succeed in straightening out curved sets where the linear approach fails, as in data sets 1, 2, 3 of Table 4, it can prove just as inadequate. For example, for the 11-part major-oxide compositions of 28 Gough Island volcanic rocks (Le Maitre, 1962, Table 10) analyzed into linear principal components by Butler (1976) the scattergram of first against second loglinear-contrast principal components displays as much curvature as Fig. 1 of Butler (1976). It thus remains an open problem to discover other forms of functions for principal components which will adequately describe such variability.

ACKNOWLEDGMENTS

The content and presentation of this paper has been greatly improved by the usual penetrating and constructive comments of Felix Chayes. Thanks are also due to Paul Carr for making available data set 2.

REFERENCES

Aitchison, J., 1981, A new approach to null correlations of proportions: Jour. Math. Geol., v. 13, p. 175–189.

Aitchison, J., 1982, The statistical analysis of compositional data: Jour. Roy. Stat. Soc. Ser. B, v. 44, p. 139–177.

Aitchison, J., 1983, Principal component analysis of compositional data: Biometrika, v. 70, p. 57–65.

Becke, F., 1897, Gesteine der Columbretes, Tschermak's Mineral: Petrogr. Mitt., v. 16, p. 308–336.

Butler, J. C., 1976, Principal component analysis using the hypothetical closed array: Jour. Math. Geol., v. 8, p. 25–36.

Chayes, F., 1960, On correlation between variables of constant sum: Jour. Geophys. Res., v. 65, p. 4185–4193.

Chayes, F. and Kruskal, W., 1966, An approximate statistical test for correlation between two proportions: Jour. Geol., v. 74, p. 692–702.

Chayes, F. and Trochimczyk, J., 1978, An effect of closure on the structure of principal components: Jour. Math. Geol., v. 10, p. 323–333.

Hawkins, D. M. and Rasmussen, S. E., 1973, Use of discriminant analysis for classification on strata in sedimentary successions: Jour. Math. Geol., v. 5, p. 163–177.

Hotelling, H., 1933, Analysis of a complex of statistical variables into principal components: Jour. Educ. Psych., v. 24, p. 417–441.

Kork, J. O., 1977, Examination of the Chayes-Kruskal procedure for testing correlations between proportions: Jour. Math. Geol., v. 9, p. 543–562.

Le Maitre, R. W., 1962, Petrology of volcanic rocks, Gough Island South Atlantic: Geol. Soc. Amer. Bull., v. 73, p. 1309–1340.

Le Maitre, R. W., 1968, Chemical variation within and between volcanic rock series–a statistical approach: Jour. Petrol., v. 9, p. 220–252.

Le Maitre, R. W., 1976, A new approach to the classification of igneous rocks using the basalt–andesite–dacite–rhyolite suite as an example: Contrib. Min. Petrol., v. 56, p. 191–203.

Miesch, A. T., 1980, Scaling variables and interpretation of eigenvalues in principal component analysis of geologic data: Jour. Math. Geol., v. 12, p. 523–538.

Pearson, K., 1901, On lines and planes of closest fit to systems of points in space: Phil. Mag. Ser. 6, v. 2, p. 559–572.

Roth, H. D., Pierce, J. W., and Huang, T. C., 1972, Multivariate discriminant analysis of bioclastic turbidites: Jour. Math. Geol., v. 4, p. 249–261.

Till, R. and Colley, H., 1973, Thoughts on use of principal component analysis in petrogenetic problems: Jour. Math. Geol., v. 4, p. 341–350.

Trochimczyk, J. and Chayes, F., 1977, Sampling variation of principal components: Jour. Math. Geol., v. 9, p. 497–506.

Trochimczyk, J. and Chayes, F., 1978, Some properties of principal component scores: Jour. Math. Geol., v. 10, p. 43–52.

Vistelius, A. B., Ivanov, D. N., Kuroda, Y., and Fuller, C. R., 1970, Variation of model composition of granitic rocks in some regions around the Pacific: Jour. Math. Geol., v. 2, p. 63–80.

Webb, W. M. and Briggs, L. I., 1966, The use of principal component analysis to screen mineralogical data: Jour. Geol., v. 74, p. 716–720.