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

The Statistical Analysis of Geochemical Compositions[1]

John Aitchison[2]

The analysis and interpretation of compositional data, such as major oxide compositions of rocks, has been traditionally plagued by the so-called constant-sum or closure problem. Particular difficulties have been the lack of a satisfactory, interpretable covariance structure and of rich, tractable, parametric classes of distributions on the simplex sample space. Consideration of logistic and logratio transformations between the simplex and Euclidan space has allowed the introduction of new concepts of covariance structure and of classes of logistic-normal distributions which have now opened up a substantial and meaningful array of statistical methodology for compositional data. From the motivation of a wide variety of practical geological problems we examine the range of possibilities with this new approach to the constant-sum problem.

KEY WORDS: Closure, closed number system, logistic, logratio, geochemistry.

1. INTRODUCTION

Any browser through the geological literature, past and present, must soon become aware of its pervasion by compositional data, such as major oxide percentages of rock specimens or sand-silt-clay compositions of sediments. The immediate inference, that the study of such compositions is fundamental to geology, would find ready support among geologists themselves. For example, Chayes (1962) affirms that "percentage data occur in every natural science but are possibly of more central importance to the petrographer than to most other naturalists." Reflecting that the study of patterns of variability in data of any kind is the bread-butter-and-jam of statisticians, our browser might reasonably expect to find a well-established and satisfactory methodology for the statistical analysis and interpretation of compositional data. He would be bitterly disappointed.

[1]Manuscript received 1 December 1983. Paper presented at the session on Current Statistical Developments in the Geosciences at the American Statistical Association Annual Meeting, 15–18 August 1983, Toronto, Canada.

[2]Department of Statistics, University of Hong Kong, Pokfulam Road, Hong Kong.

0200-5958/84/0800-0531$03.50/0

Instead of confident recommendations for statistical analysis he would discover confusion. From 1960 onward he would read many warnings, mainly by geologists to their colleagues, that there are apparently unyielding difficulties in the interpretation of compositional data caused by their constant-sum or the closure property which states that the sum of the proportions of a whole is unity. He would find analyses by these colleagues ignoring the warnings and even some by the warners themselves ignoring their own warnings. He would hear consequent cries of near-despair from some warners as, for example, from Reyment (1977) in his Presidential address to the International Association for Mathematical Geology: "... I feel I should express a certain amount of dismay at the relatively slow rate of penetration of quantitative thinking in the rank and file of geology. Petrologists are still largely unaware of the dangers of closure (that is the effect of the constant constraint) in their diagrams ...."

There indeed had been even earlier warnings from no less a statistical authority than Karl Pearson (1897), pointing out the dangers that may befall the analyst who attempts to interpret product-moment correlations between quotients whose numerators or denominators contain common parts, and thus implying that the analysis of proportions of some whole is likely to be fraught with difficulty. History has proved him correct, for over the succeeding years and indeed right up to the present day there has probably been no other form of data analysis where more confusion reigns and where improper or inadequate statistical methods are applied.

Unfortunately it must be admitted that the 'rank and file of geology' have had little more than warnings. In possession of a can of delicious compositional beans and wanting to reveal its contents they have naturally been unimpressed by the news that they are bound to make a mess with their neolithic flint of a tool. The sole purpose of this paper is to bring them the good news of the invention of a modest form of can opener. I hope that by its deliberately provocative style the paper may achieve five objectives:

(1) To persuade geologists and statisticians alike to abandon the use of "standard methods" quite inappropriate to "nonstandard" data sets such as compositions.

(2) To provide a clear exposition of new concepts and an associated methodology for meaningful statistical analysis of compositional data.

(3) To illustrate the simplicity and effectiveness of the new procedure in a series of applications.

(4) To convince geologists that they no longer need subscribe to defeatist attitudes toward the constant-sum problem but should direct their energies toward reformulating, or possibly even formulating for the first time, their geological hypotheses within the new framework.

(5) To persuade editors to reject all papers using "standard methods" or adopting the negative approach of extending the repertoire of variations on the theme of how standard methods fail.

2. COMPOSITIONAL DATA SETS

To illustrate the methodology of this paper the ideal would have been to present the reader with a series of published compositional data sets covering the whole range of problems considered. Unfortunately these would have been prohibitively large and so, after much hesitation, I have settled for four modest, presentable "unpublished" data sets of sufficient complexity to capture the

Table 1. Compositions of 25 Specimens of Hongite

Specimen no.	Percentages				
	A	B	C	D	E
H1	48.8	31.7	3.8	6.4	9.3
H2	48.2	23.8	9.0	9.2	9.8
H3	37.0	9.1	34.2	9.5	10.2
H4	50.9	23.8	7.2	10.1	8.0
H5	44.2	38.3	2.9	7.7	6.9
H6	52.3	26.2	4.2	12.5	4.8
H7	44.6	33.0	4.6	12.2	5.6
H8	34.6	5.2	42.9	9.6	7.7
H9	41.2	11.7	26.7	9.6	10.8
H10	42.6	46.6	0.7	5.6	4.5
H11	49.9	19.5	11.4	9.5	9.7
H12	45.2	37.3	2.7	5.5	9.3
H13	32.7	8.5	38.9	8.0	11.9
H14	41.4	12.9	23.4	15.8	6.5
H15	46.2	17.5	15.8	8.3	12.2
H16	32.3	7.3	40.9	12.9	6.6
H17	43.2	44.3	1.0	7.8	3.7
H18	49.5	32.3	3.1	8.7	6.3
H19	42.3	15.8	20.4	8.3	13.2
H20	44.6	11.5	23.8	11.6	8.5
H21	45.8	16.6	16.8	12.0	8.8
H22	49.9	25.0	6.8	10.9	7.4
H23	48.6	34.0	2.5	9.4	5.5
H24	45.5	16.6	17.6	9.6	10.7
H25	45.9	24.9	9.7	9.8	9.7

typical difficulties of analysis. Regretable though this device may be it does have the advantage that fresh data present a substantial challenge of discovery to the reader and may avoid any preconceptions he may have as to their pattern of variability. Moreover, for each statistical method illustrated we can indicate a reference providing an analysis of a published geological data set.

Tables 1-4 present four compositional data sets, each consisting of 25 five-part compositions of a particular type of rock. For convenience of reference, I have labeled the four different types *hongite*, *kongite*, *boxite*, *coxite* and the five geochemical parts, A, B, C, D, E. For each of the boxite and coxite compositions Tables 3 and 4 provide the depth at which it was sampled; in addition Table 4 provides a measure of porosity for each of the coxite compositions.

Table 2. Compositions of 25 Specimens of Kongite

Specimen no.	Percentages				
	A	B	C	D	E
K1	33.5	6.1	41.3	7.1	12.0
K2	47.6	14.9	16.1	14.8	6.6
K3	52.7	23.9	6.0	8.7	8.7
K4	44.5	24.2	10.7	11.9	8.7
K5	42.3	47.6	0.6	4.1	5.4
K6	51.8	33.2	1.9	7.0	6.1
K7	47.9	21.5	10.7	9.5	10.4
K8	51.2	23.6	6.2	13.3	5.7
K9	19.3	2.3	65.8	5.8	6.8
K10	46.1	23.4	10.4	11.5	8.6
K11	30.6	6.7	43.0	6.3	13.4
K12	49.7	28.1	5.1	8.0	9.1
K13	49.4	24.3	7.6	8.5	10.2
K14	38.4	9.5	30.6	14.8	6.7
K15	41.6	19.0	17.3	13.8	8.3
K16	42.3	43.3	1.6	5.9	6.9
K17	45.7	23.9	10.3	11.6	8.5
K18	45.5	20.3	13.6	10.9	9.7
K19	52.1	17.9	10.7	7.9	11.4
K20	46.2	14.3	18.5	12.2	8.8
K21	47.2	30.9	4.6	6.3	11.0
K22	45.4	33.3	4.0	11.9	5.4
K23	48.6	23.4	8.7	10.7	8.6
K24	31.2	4.5	47.0	10.2	7.1
K25	44.3	15.0	19.4	10.5	10.8

Table 3. Composition and Depths of 25 Specimens of Boxite

Specimen no.	Percentages					Depth
	A	B	C	D	E	
B1	43.5	25.1	14.7	10.0	6.7	1
B2	41.1	27.5	13.9	9.5	8.0	2
B3	41.5	20.1	20.6	11.1	6.7	3
B4	33.9	37.8	11.1	11.5	5.7	4
B5	46.5	16.0	15.6	14.3	7.6	5
B6	45.3	19.4	14.8	13.5	9.3	6
B7	33.2	25.2	15.2	17.1	9.3	7
B8	40.8	15.1	21.7	14.6	7.8	8
B9	33.0	30.8	15.1	12.9	8.2	9
B10	28.2	38.6	12.1	14.1	6.9	10
B11	33.9	31.5	15.4	12.0	7.2	11
B12	48.7	19.3	13.4	10.7	7.9	12
B13	37.8	37.1	10.4	8.6	6.1	13
B14	42.0	26.6	13.7	10.5	7.2	14
B15	44.2	26.5	12.9	9.6	6.8	15
B16	39.7	23.2	20.6	10.2	6.3	16
B17	39.3	28.1	13.0	13.6	6.0	17
B18	34.1	26.7	13.6	17.0	8.6	18
B19	36.2	35.3	11.2	11.9	5.4	19
B20	39.5	36.0	9.4	8.4	6.7	20
B21	39.5	22.5	18.7	11.4	7.9	21
B22	33.0	33.5	17.7	9.8	6.0	22
B23	42.3	16.6	16.9	17.0	7.2	23
B24	39.9	19.0	13.4	21.3	6.4	24
B25	37.8	30.9	11.9	12.9	6.5	25

Note that the compositional part of these data sets forms an array or a compositional data matrix $X = [x_{rc}]$, where x_{rc}, the entry in the rth row and cth column, denotes the cth component or proportion of the cth part of the rth replicate or rock specimen. In general X will be of order $n \times (d + 1)$ where n is the number of replicates and $d + 1$ is the number of components. The constant-sum constraint satisfied by compositions is then equivalent to the following condition on the data matrix

$$Xj_{d+1} = j_{d+1} \tag{1}$$

where j_{d+1} is a $(d + 1)$ vector of units.

Table 4. Compositions, Depths and Porosities of 25 Specimens of Coxite

Specimen no.	Percentages					Depth	Porosity
	A	B	C	D	E		
C1	44.2	31.9	5.4	10.5	8.0	1	43.5
C2	49.0	25.4	5.8	11.3	8.5	2	50.4
C3	50.2	24.8	5.7	11.1	8.2	3	52.3
C4	49.9	24.7	5.4	11.4	8.6	4	52.5
C5	48.5	27.8	5.9	10.2	7.6	5	45.2
C6	45.9	27.1	6.9	11.5	8.6	6	42.7
C7	44.1	31.9	6.0	10.2	7.8	7	44.0
C8	46.4	29.9	5.5	10.3	7.9	8	44.0
C9	45.7	27.0	6.2	12.0	9.1	9	46.0
C10	46.4	30.0	5.1	10.4	8.1	10	48.0
C11	41.7	30.2	7.7	11.6	8.8	11	36.7
C12	44.9	25.7	7.7	12.4	9.3	12	41.0
C13	48.6	27.7	5.8	10.2	7.7	13	45.7
C14	49.7	26.7	4.9	10.6	8.1	14	54.4
C15	49.6	24.4	6.4	11.2	8.4	15	46.8
C16	46.5	28.6	5.9	10.7	8.3	16	44.9
C17	47.3	24.2	7.9	11.8	8.8	17	43.1
C18	44.7	30.0	6.8	10.5	8.0	18	41.0
C19	48.0	25.6	7.0	11.1	8.3	19	45.4
C20	50.0	23.8	6.6	11.2	8.4	20	47.5
C21	51.4	24.2	5.7	10.7	8.0	21	52.5
C22	53.3	25.1	5.2	9.4	7.0	22	52.9
C23	47.9	25.4	6.7	11.4	8.6	23	44.4
C24	43.5	29.8	6.7	11.2	8.8	24	39.1
C25	44.5	29.2	6.5	11.2	8.6	25	42.6

3. TYPICAL PROBLEMS

The following is a typical, although by no means exhaustive, set of questions about compositional data sets in geology.

1. How can we satisfactorily describe the pattern of variability of the hongite compositions? For a new rock specimen with (A, B, C, D, E) composition (44.0, 19.7, 14.9, 9.1, 12.3) and claimed to be hongite, can we say whether it is fairly typical of hongite in composition or can we place some measure on its atypicality?

2. For a particular rock type can we suitably define a correlation structure that will allow us to pose and test meaningful hypotheses about that structure? What forms of independence are possible within the constant-sum constraint?

3. To what extent can we obtain insights into the pattern of variability of the compositions by partial analyses such as the commonly employed ternary diagrams? Are there any other ways of discovering the essential dimensionality of the pattern?

4. Are the patterns of variability of hongite and kongite essentially different and if so, can a convenient form of classification be devised on the basis of the composition? Can we investigate whether a subcomposition, such as a ternary diagram, would be as effective?

5. Are the compositions of boxite and coxite related in any way to depth; in other words, is there some trend in the compositions of boxite and coxite?

6. Does the porosity of a coxite specimen depend on its composition in any way?

These specific questions will be used to motivate the development of a simple methodology for the statistical analysis of compositional data and to illustrate its application.

4. THE SIMPLEX AS SAMPLE SPACE

The first task of a statistician when faced with modeling a new observational or experimental situation is surely to devise an appropriate sample space. For compositional data this is a *simplex*. For compositions $(x_1, \ldots, x_{d+1})$ of $d+1$ parts this is essentially a d-dimensional space, a subspace of R^d

$$S^d = \{(x_1, \ldots, x_d) : x_i \geqq 0 \ (i = 1, \ldots, d), x_i + \cdots + x_d \leqq 1\} \tag{2}$$

although it may often more conveniently be considered in a symmetric form as a d-dimensional subspace of R^{d+1}

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

The ternary diagrams and tetrahedral representations, so familiar to geologists, are the cases $d = 2$ and $d = 3$. The statistical problems of geochemical compositions are, therefore, those of investigating distributions over the simplex. Before we begin to consider these it is useful to define two simple algebraic operations on compositions.

Subcompositions

From a composition of 10 or 11 major oxides a geologist may select some, such as CaO, Na_2O, K_2O, and rescale to obtain a new composition, a *subcomposition* which he can then represent in the familiar CNK diagram. From any subvector, say $(x_1, \ldots, x_c)$, of a d-dimensional composition $(x_1, \ldots, x_{d+1})$ we

can form a subcomposition, denoted by $C(x_1, \ldots, x_c)$ with

$$C(x_1, \ldots, x_c) = (x_1, \ldots, x_c)/(x_1 + \cdots + x_c) \tag{4}$$

Formally the subcomposition operator (4) can be regarded as a projection from the full simplex S^d to a subsimplex S^{c-1}.

Partition

When we wish to divide the composition into a number of subvectors and examine the interrelationships, including the total amounts of the available unit taken up by the components of the various subvectors, we may find it convenient to consider the concept of a *partition* of a composition. We confine attention to a simple partition based on a single division of the composition into two subvectors $(x_1, \ldots, x_c)$ and $(x_{c+1}, \ldots, x_{d+1})$. A partition is then defined as (s_1, s_2, t) where s_1 and s_2 are the subcompositions based on the two subvectors and t is the total of one of the subvectors, say the first. Thus $s_1 = (s_{11}, \ldots, s_{1,c-1}), s_2 = (s_{21}, \ldots, s_{2,d-c})$ and

$$\begin{aligned} s_{1i} &= x_i/(x_1 + \cdots + x_c) && (i = 1, \ldots, c-1) \\ s_{2i} &= x_{c+i}/(x_{c+1} + \cdots + x_{d+1}) && (i = 1, \ldots, d-c) \\ t &= x_1 + \cdots + x_c \end{aligned} \tag{5}$$

An important aspect of this transformation is that it is one-to-one between S^d and $S^{c-1} \times S^{d-c} \times S^1$.

5. THE TRADITIONAL DIFFICULTIES

Analysts of compositional data have recognized three main aspects of the difficulty of interpretation.

High Dimensionality

Many geologists have perceived the major, possibly the only, difficulty with compositional data to be the inability of the human eye to see in more than three dimensions. For example, the view of Iddings (1903) that "since compositions involve so many components, projection into two-dimensional diagrams is needed for comparison of types and samples" has been frequently reiterated, and recently Barker (1978) has expressed this view in almost identical words. Such projections or subcompositional analyses are at best partial analyses, subject to substantial loss of information and to misinterpretation, and have been repeatedly condemned by geologist warners such as Chayes (1962) and Butler (1979). Thus Chayes (1962) asserts ". . . inferences based on intuitive geometrical examination or cookbook statistical testing of these diagrams will more often

be wrong than right." We find an excellent particular example of the futility of subcompositional analysis in Section 12.

Absence of an Interpretable Covariance Structure

The difficulties, foreseen by Pearson (1897), of interpreting product-moment correlations between components of a composition, were first brought to the attention of geologists by Chayes (1960, 1962), Krumbein (1962) and Sarmanov and Vistelius (1959) and have continued to be a matter of concern right up to the present (Chayes, 1983). There are even complete texts (Chayes, 1971; Le Maitre, 1982) describing analyses of compositions in terms of $\mathrm{cov}\,(x_i, x_j)$ $(i, j = 1, \ldots, d+1)$ and discussing the difficulties of interpretation. Since the difficulties are so well documented we confine attention here to two ways of expressing the awkward features.

(1) *Negative bias difficulty.* Since $\mathrm{cov}\,(x_1, x_1 + \cdots + x_{d+1}) = 0$ then

$$\mathrm{cov}\,(x_1, x_2) + \cdots + \mathrm{cov}\,(x_1, x_{d+1}) = -\mathrm{var}\,(x_1) < 0 \tag{6}$$

so that at least one of the covariances on the left must be negative. Hence correlations are not free to range over the usual interval $(-1, 1)$ and there are bound to be problems of interpretation.

(2) *Closure difficulty.* If the composition is formed from a *basis* ("open variables") of actual quantities $w_1, \ldots, w_{d+1}$ by closure or scaling, $x = C(w)$, independence of $w_1, \ldots, w_{d+1}$ does not correspond to any simple "null" structure of the $\mathrm{cov}\,(x_i, x_j)$. "Independence" of raw proportions, therefore, seems to be associated with ill-defined null, non-zero, correlations, an awkward concept. Chayes and Kruskal (1966) attempted to obtain tests of the hypothesis that the compositions could have arisen from such imagined bases with independent components. Apart from many technical and persisting interpretational difficulties (Miesch, 1969; Aitchison, 1981a) the method suffers from an insuperable conceptual difficulty that there are many bases corresponding to a single composition (Kork, 1977; Aitchison, 1981a, 1982).

Difficulty of Parametric Modeling

In complex situations such as those involving compositional data it is difficult to see how analysis of the patterns of variability can ever be wholly successful in the absence of a rich enough parametric class of distributions over the appropriate sample space. For example, the multinormal class $N^d(\mu, \Sigma)$ and the multivariate lognormal class $\Lambda^d(\mu, \Sigma)$ have proved themselves to be flexible instruments in the analysis of data in R^d and P^d, d-dimensional positive space. When the space is the simplex S^d any request of a statistical audience to declare what parametric classes of distributions its members know produces a standard

response: only the Dirichlet class and possibly some simple generalizations are offered. Unfortunately the Dirichlet distribution $D^d(\beta)$ with density function

$$\frac{\Gamma(\beta_1 + \cdots + \beta_{d+1})}{\Gamma(\beta_1) \cdots \Gamma(\beta_{d+1})} x_1^{\beta_1 - 1} \cdots x_{d+1}^{\beta_{d+1} - 1}$$

turns out to be totally inadequate for the description of compositional data for three main reasons.

(1) Its isoprobability contours are convex for $\beta_i > 1$ $(i = 1, \ldots, d+1)$ and so there is clearly no hope of a satisfactory fit to commonly occurring concave data patterns such as in Fig. 3.

(2) Its correlation structure is completely negative with $\operatorname{cov}(x_i, x_j) < 0$ for every $i \neq j$, and there are obviously data patterns for which some such correlations are definitely positive.

(3) A Dirichlet-distributed composition possesses all the imaginable independence properties of compositions (e.g., it can be visualized as the closure of a basis of independent, equally scaled, gamma variables). Generalizations have failed to widen the class substantially. Thus, recently James (1981) asserts that "there thus remains in the literature a lack of tractable rich distributions for random proportions which are not neutral." Neutrality is a particularly strong form of independence.

Having briefly retraced the difficulties, hopefully finally, let us now turn our attention more positively toward a fresh look at describing variability within the simplex.

6. COVARIANCE STRUCTURE

All the difficulties arising in the traditional analysis of geochemical compositions come from a lack of appreciation that to carry over ideas which are highly successful for one particular sample space, such as R^d, to another very different sample space, namely S^d, may be completely inappropriate. The spate of papers in the last 20 years setting out the difficulties and the absence of any significant progress surely speak for themselves and I pay my reader the compliment of having realized by now that the adoption of a crude covariance structure based on $\operatorname{cov}(x_i, x_j)$ causes more confusion than it removes. Any reader with a lingering nostalgia for an untampered $\operatorname{cov}(x_i, x_j)$ should ponder the following passage from Zukav (1979, p. 71)

> Things are not "correlated" in nature. In nature, things are as they are. Period. "Correlation" is a concept which *we* use to describe connections which *we* perceive. There is no word, "correlation," apart from people. There is no concept, "correlation," apart from people. This is because only people use words and concepts.

In other words we should not become hidebound in our approach to new problems to the extent of regarding any modeling concept as some embodiment of nature. As I have said elsewhere we would not expect that excellent tool of the wide open spaces (or R^d) of North America, namely the barbecue, necessaarily to be an appropriate concept for cooking in the confined space (or S^d) of a low-cost housing flatlet in Hong Kong. If our concepts fail to serve us in new situations we must invent new concepts.

A crucial sentence in Zukav's paragraph is surely that correlation is a concept we use to describe *connections we perceive*. In the plethora of papers enumerating the absence of connections within the crude covariance structure we find, for example in the comparison of variances of oxides in the complete composition with those in a subcomposition, namely, var (x_i) with var $\{x_i/(x_1 + \cdots + x_c\}$ $(i = 1, \ldots, c)$, very different and unrelatable rank orderings. And much is also made of the fact that from subcompositional information such as the crude variances and covariances the corresponding compositional quantities cannot be reconstructed. Yet some do perceive an elementary connection between subcompositions and the parent composition, namely that *ratios* of components are the same within the subcomposition and the composition, but fail to realize that the embodiment of this single wisp of a connection must surely form the foundation of a sensible study of compositions. This realization that the study of compositions is essentially concerned with the relative magnitudes of ingredients rather than in any sense their absolute values leads naturally to a concept of correlation structure based on product-moment covariances of ratios such as

$$\operatorname{cov}(x_i/x_j, x_k/x_l) \tag{7}$$

Experience with mathematical and statistical modeling has, however, led us to another perception, that in first attempts at modeling new situations there is little harm in opting for the tractable. Now, variances and covariances of ratios are awkward to manipulate and as any lecturer in statistics must grow weary of telling students, when stuck by complicated products and quotients take logs. Thus, as far as compositions are concerned, as far as living in the simplex is concerned, we should be able to think more clearly about relationships if we adopt a new concept of correlation (Aitchison, 1981a, 1982) based on

$$\operatorname{cov}(\log x_i/x_j, \log x_k/x_l) \tag{8}$$

This covariance structure can be more economically expressed in terms of a $d \times d$ *logratio covariance matrix*

$$\Sigma = \operatorname{cov}(\log x_1/x_{d+1}, \ldots, \log x_d/x_{d+1}) \tag{9}$$

Knowledge of $\Sigma = [\sigma_{ij}]$ allows us to construct any variance or covariance of logratios through the relationship

$$\operatorname{cov}(\log x_i/x_j, \log x_k/x_l) = \sigma_{ik} - \sigma_{il} - \sigma_{jk} + \sigma_{jl} \tag{10}$$

where any σ with a suffix $d+1$ is interpreted as zero. Also we can identify clearly and precisely what partial information about a compositional covariance structure Σ is provided by the knowledge of the covariance structure of a subcomposition. For example, the logratio covariance matrix of $C(x_{c+1}, \ldots, x_{d+1})$ is identical to the trailing $(d-c) \times (d-c)$ submatrix of Σ.

The introduction of logarithms, of course, requires the assumption that none of the components is zero, in other words, that the sample space is the strictly positive simplex

$$S_+^d = \{(x_1, \ldots, x_{d+1}) : x_i > 0 \ (i = 1, \ldots, d+1), x_1 + \cdots + x_{d+1} = 1\} \tag{11}$$

We continue with this assumption until Section 14 when we briefly discuss the problem of zeros.

Definition (9) apparently introduces an asymmetry in selecting one component for the special role of common divisor. The reader may, therefore, ask whether statistical procedures based on such a construct may lead to different conclusions depending on which component is chosen as divisor. Such is not the case for any of the procedures discussed in this paper: they are invariant under the group of permutations of the components. I do not digress to discuss the technicalities of this invariance but invite the reader to choose divisors different from the final component x_{d+1}, chosen here merely for convenience and to verify the accuracy of the claim.

As an alternative to the asymmetric logratio covariance matrix we can define a symmetric $(d+1) \times (d+1)$ *logcentered covariance matrix*

$$\Gamma = \operatorname{cov} \{\log x_1/g(x), \ldots, \log x_{d+1}/g(x)\} \tag{12}$$

where the common divisor $g(x)$ used to form the ratios is the geometric mean $(x_1, \ldots, x_{d+1})^{1/(d+1)}$ of the $d+1$ components of the composition. The aesthetic advantage of symmetry is bought at a price, the disadvantage of the singularity of Γ, although for many applications this is adequately overcome through the use of pseudo-inverses of singular matrices. For the special form of singularity of (12) a very convenient pseudo-inverse is usually

$$\Gamma^- = \left(\Gamma + \frac{\tau}{d+1} J_{d+1}\right)^{-1} - \frac{1}{\tau(d+1)} J_{d+1} \tag{13}$$

where $\tau = \operatorname{trace}(\Gamma)$ and J_{d+1} is the $(d+1) \times (d+1)$ matrix with every element 1. The choice between forms (9) and (12) is largely a matter of personal preference, of whether one hates asymmetry more or less than singularity.

To clarify the relationships of these various covariance matrices we examine the computation of the corresponding sample covariance matrices starting with data matrix $X = [x_{rc}]$. First we note that from any $n \times (d+1)$ data matrix X,

the sample covariance matrix S_X is formed by the operation

$$(n - 1)S_X = X^T G_n X \tag{14}$$

where

$$G_n = I_n - (1/n)J_n \tag{15}$$

and J_n is the $n \times n$ matrix with each element 1.

From the $n \times (d + 1)$ crude matrix X we can form three other data matrices, the $n \times (d + 1)$ *log data matrix*

$$W = [w_{rc}] = [\log x_{rc}] \tag{16}$$

the $n \times d$ *logratio data matrix*

$$T = [y_{rc}] = \left[\log \frac{x_{rc}}{x_{r,d+1}}\right] = [w_{rc} - w_{r,d+1}] = WB^T \tag{17}$$

where

$$B = [I_d, -j_d] \tag{18}$$

and the $n \times (d + 1)$ *logcentered data matrix*

$$Z = \left[w_{rc} - \frac{1}{d+1}(w_{r1} + \cdots + w_{r,d+1})\right] = WG_{d+1} \tag{19}$$

Then the relations between the sample log, logratio, and logcentered covariance matrices S_W, S_Y, and S_Z are

$$S_Y = BS_W B^T = BS_Z B^T \tag{20}$$

$$S_Z = G_{d+1} S_W G_{d+1} \tag{21}$$

Note that the centering process taking W to Z is a row-centering among the different components of individual rock specimens whereas the covariance matrix computation from W to S_W involves a column-centering of the same component within different rock specimens.

The natural first reaction of the geologist to the covariance structure (9) is probably one of resistance to what may seem an unnecessarily complicated descriptive tool. I think, however, that the geologist has to accept that for a constrained form of data this must be the simplest alternative to persistence with crude analysis and more decades of bewilderment. One should look to other restricted forms of vectors for some insights; for example, directional data with the circle and sphere as sample spaces. It is somewhat of a paradox that the simplex has been such a stumbling block to analysts whereas the sphere has not. No one doubts that statistical analysis of directional data should be very different from data in R^d. Why then should we expect methods in R^d to be successful in S^d? From our modest start with an appropriate covariance structure we find, as

we reexamine the questions posed in Section 3, that a set of concepts and methods will emerge, far easier to comprehend and to practice than their counterparts on the circle and sphere.

7. DIMENSION-REDUCING TECHNIQUES

In handling "ordinary" vectors in R^d we have become familiar with two forms of the dimension-reducing technique. The first is marginal analysis, where we simply select some of the components and look at their marginal distribution in fewer dimensions than d. For compositions the counterpart of marginal analysis is subcompositional analysis, as in the popular inspection of ternary diagrams. The second is principal component analysis with its attempt to capture most of the variability in a few linear combinations of the components. There have been many attempts to carry over this idea into the analysis of compositional data, mainly by a method advocated by Le Maitre (1968). With a new covariance structure for compositions we are in a position to study both techniques. Aitchison (1983, 1984a) has undertaken a detailed critical reappraisal of these two dimension-reducing techniques with applications to the major oxide compositions of rocks, thus answering question 3 posed in Section 3. Here we present the essence of the argument in relation to the hongite and boxite data sets.

Principal Component Analysis

The currently popular form of principal component analysis, advocated by Le Maitre (1968), uses the d positive eigenvalues and their d eigenvectors of the crude $(d+1) \times (d+1)$ covariance matrix S_X. Figure 1 shows the scattergrams of the first and second crude principal components for the hongite and boxite compositions. The persistent curvature in the hongite scattergram reflects a failure of the *linear* crude method to capture the essentially curved nature of the variability of hongite compositions in the simplex. Webb and Briggs (1966) provide an alternative method based on a ratio, not logratio, covariance matrix, but it also is linear in effect and is not invariant under different choices of divisor.

With the new covariance structures we can use either the asymmetric S_Y or the symmetric S_Z to obtain identical results. With S_Y we use the d eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_d$ and eigenvectors $b_1, \ldots, b_d$ which are solutions of

$$(S_Y - \lambda H_d)b = 0 \qquad b^T H_d b = 1 \tag{22}$$

where $H_d = I_d + J_d$ is the "isotropic" logratio covariance matrix (Aitchison, 1983). With S_Z we use the d positive eigenvalues λ_i and the corresponding eigenvectors a_i, satisfying

$$(S_Z - \lambda I_{d+1})a = 0 \qquad a^T a = 1 \tag{23}$$

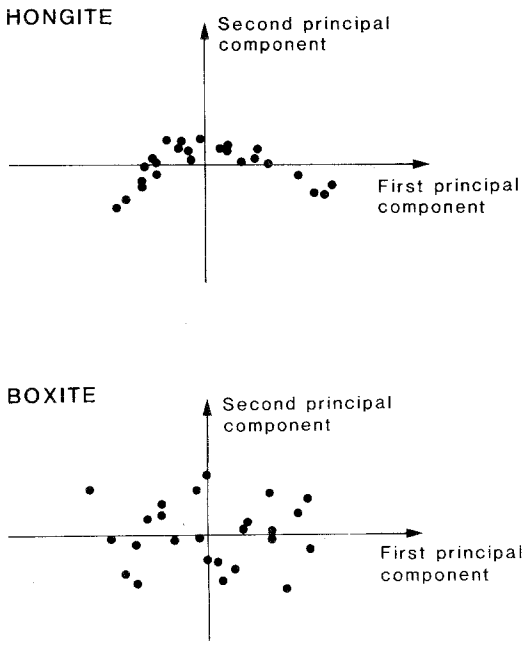

Fig. 1. Scattergram of the first and second crude principal components for hongite and boxite.

The eigenvectors a of (23) and b of (22) are simply related by $a = bB^T$ and lead to identical principal components which are loglinear contrasts of the composition components, that is, of the form

$$\sum_{i=1}^{d+1} a_i \log x_i \qquad \sum_{i=1}^{d+1} a_i = 0 \tag{24}$$

The total measure of variability is given by

$$\lambda_1 + \cdots + \lambda_d = \text{trace } S_Z \tag{25}$$

Here the symmetric form is probably the easier with standard algorithms for eigenanalysis.

Table 5 shows the 5×5 logcentered covariance matrix S_Z for hongite together with the four positive eigenvalues and their eigenvectors. The usual measure

$$(\lambda_1 + \cdots + \lambda_c)/(\lambda_1 + \cdots + \lambda_d) \tag{26}$$

of the proportion of "total variability" captured by the first c principal components applies. Thus, for hongite the first logratio or logcentered principal compo-

Table 5. Logcentered Covariance Matrix S_Z, Eigenvalues and Eigenvectors for Hongite

$$S_Z = \begin{bmatrix} 0.06350 & 0.17762 & -0.24063 & 0.01412 & -0.01461 \\ 0.17762 & 0.55134 & -0.72794 & 0.02537 & -0.02639 \\ -0.24063 & -0.72794 & 0.96718 & -0.03968 & 0.04105 \\ 0.01412 & 0.02537 & -0.39668 & 0.04693 & -0.04675 \\ -0.01461 & -0.02639 & 0.04105 & -0.04675 & 0.04670 \end{bmatrix}$$

Eigenvalues

λ_1	λ_2	λ_3	λ_4
1.579	0.0911	0.00566	0.000106

Eigenvectors

-0.194	-0.069	0.799	-0.345
-0.590	0.085	-0.549	-0.378
0.783	-0.013	-0.222	-0.372
-0.033	-0.704	-0.085	0.544
0.034	0.701	0.057	0.551

nent captures 94.2% of the total variability. The scattergrams of first and second logratio principal components, shown in Fig. 2 for hongite and boxite, demonstrate the capability of the new method to reduce successfully both the "curved" variability of the hongite compositions and the "linear" or "elliptical" variability of the boxite compositions to standard patterns for principal component variability. Figure 3 reinforces the success of the logratio method and the failure of the crude method by showing the ternary diagrams and the principal axes for the *ABC* subcompositions of the hongite compositions.

Subcompositional Analysis

A central question here is whether we can similarly obtain a measure of the proportion of total compositional variability retained by a subcomposition. Since a subcomposition such as $C(x_1, \ldots, x_{c+1})$ is technically a composition with dimension c, smaller than the dimension d of the original composition, its total measure of variability is given by the trace of its logcentered covariance matrix, say S. Then we can use

$$\text{trace } S / \text{trace } S_Z \tag{27}$$

to assess the proportion of the total compositional variability retained by the subcomposition. If the purpose of the subcompositional analysis is to retain as

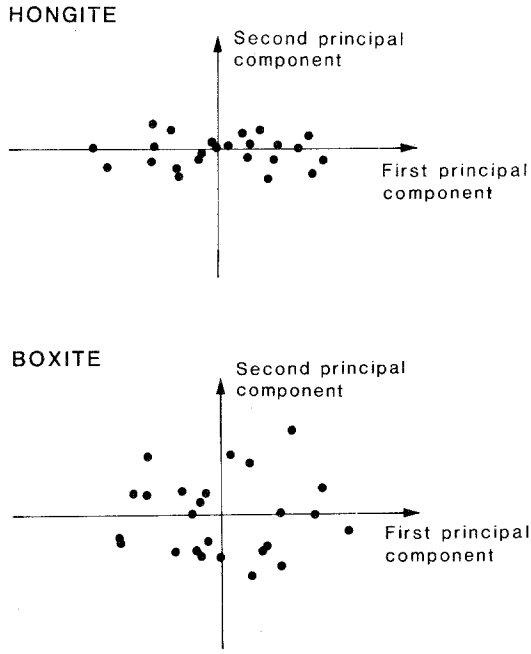

Fig. 2. Scattergram of first and second logcontrast principal components for hongite and boxite.

much variability as possible then we have to search for subcompositions which maximize (27); see Aitchison (1984a) for a simple algorithm for this search, and for applications to major oxide compositions of rocks.

Table 6 shows the rankings of all 10 two-dimensional subcompositions and the associated proportions (27) of total variability retained by each for hongite and boxite, and compares these with what is achievable by the first two principal components. Note the substantial differences between the proportions retained by the first and last ranked subcompositions for each type and the differences in the orders between hongite and boxite subcompositions. The first ranked ternary compositions for hongite and boxite capture 94.4 and 77.5% of total variability compared with 99.7 and 81.9% captured by the corresponding first two principal components. The good quality of the performance of these best subcompositions relative to the principal components should not delude us into imagining that this will be generally so, particularly with higher-dimensioned compositions. For example, Aitchison (1984a) gives an example of 11 major oxide compositions where a best ternary subcomposition retains only 60.5% of total variability compared with 90.6% for the first two principal components.

The importance of the technique is its provision of a quantitative measure of the effectiveness of a subcomposition in retaining variability displayed in the

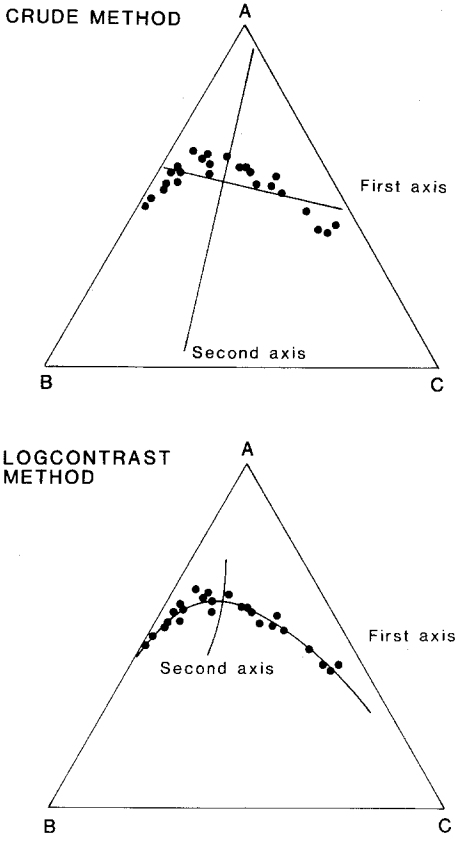

Fig. 3. Principal component axes for the subcomposition *ABC* of hongite by the crude and the logcontrast method.

Table 6. Subcompositions of Hongite and Boxite in Rank Order of Percentage Retention of Total Variability

Hongite		Boxite	
Subcomposition	Percentage	Subcomposition	Percentage
ABC	99.4	*BCD*	77.1
BCD	91.8	*ABD*	67.9
BCE	90.6	*ABC*	64.9
ACD	53.5	*BCE*	63.0
ACE	51.4	*BDE*	62.5
CDE	44.0	*ABE*	50.8
BDE	27.6	*ACD*	35.0
ABE	20.9	*CDE*	29.6
ABD	17.7	*ADE*	27.3
ADE	8.1	*ACE*	21.7

full composition. Sometimes the objective in subcompositional analysis, as, for example, in classification, is to discover subcompositions which display little variability within rock type but which are radically different between rock types. Although criterion (27) could be used as a guide for this purpose we find a more satisfactory procedure in Section 12.

8. PARAMETRIC CLASSES OF DISTRIBUTION ON S^d

The range of statistical analyses of compositional data is likely to be considerably extended if in addition to a sensible covariance structure we can find a rich parametric class of distributions on S^d which reflects the patterns of variability we observe in the simplex sample space. As indicated in Section 4 the familiar Dirichlet class and its generalizations to date are not sufficiently rich for the purposes of compositional data analysis, particularly in their inability to support a sufficient degree of compositional dependence.

Alternatives have been slow to emerge but there are now a number of practically useful classes based on a delightfully simple and old idea. When trying to find suitable means of describing patterns of variability over the positive real line P^1, for which at that time none existed, McAlister (1879) saw that it was sensible to transfer the highly successful normal pattern on the whole of the real line R^1 to P^1 through the obvious transformation

$$w = \exp(y) \qquad (w \in P^1, y \in R^1) \tag{28}$$

and so the lognormal class of distributions was invented. In the same way we can induce a class of distributions on the simplex from the class of multivariate normal distributions on R^d by use of any one-to-one transformation from R^d to S^d. We can do no better than to start with the simplest, which is already in use in other areas of statistical activity, namely the generalized logistic transformation, or as we prefer to term it, the additive logistic transformation between $x \in S^d$ and $y \in R^d$, specified by

$$x_i = \begin{cases} \exp(y_i)/[\exp(y_1) + \cdots + \exp(y_d) + 1] & (i = 1, \ldots, d) \\ 1/[\exp(y_1) + \cdots + \exp(y_d) + 1] & (i = d+1) \end{cases} \tag{29}$$

with inverse

$$y_i = \log(x_i/x_{d+1}) \qquad (i = 1, \ldots, d) \tag{30}$$

The distribution on S^d corresponding to $N^d(\mu, \Sigma)$ on R^d, will be denoted by $L^d(\mu, \Sigma)$. This class of distributions and its properties and uses have been studied by Aitchison and Shen (1980). One useful property in geochemical compositional analysis is that every subcomposition of a composition with a logistic-normal distribution has itself a logistic-normal distribution. Moreover

it is equally capable of contouring banana- and football-shaped patterns (Aitchison, 1982).

We meet another useful parametric class of distributions later but for the moment we can concentrate on the additive logistic-normal class. An important point to note is that the covariance parameter Σ of the distribution is precisely the logratio covariance matrix we arrived at in Section 6. Thus, if we can adopt a logistic-normal model for the description of our pattern of variability we will be in a position to test parametrically hypotheses concerning the covariance structure.

The great advantage of this approach to compositional data analysis is that it makes available the whole range of statistical procedures based on multivariate normality. All we have to do is transform the compositions x to the logratio compositions y and then work within R^d and on multivariate normal assumptions. In particular we can test the reasonableness of the logistic-normality assumptions through the battery of tests for multivariate normality. It would be preposterous to believe that all compositional data will turn out to be logistic normal and no doubt modifications to statistical procedures will have to be made in the light of further experience. But there do appear to be a sufficient number of geochemical data sets that are reasonably logistic normal in pattern and it is necessary to make a start with some parametric form. In addition to the investigation of covariance structure, particularly attractive procedures from the viewpoint of the statistical analysis of geochemical compositions are discriminant analysis for classification purposes and linear modeling of the mean for investigating the dependence of composition on other explanatory variables. Moreover, Bayesian methods, including statistical prediction analysis, are readily available through the substantial multivariate normal counterpart.

9. AN EXAMPLE OF MODEL FITTING

Let us consider the hongite data of Table 1, and investigate its pattern of variability.

Tests of Logistic Normality

We first transform the five-part compositions to their four-dimensional logratio counterparts and then submit these vectors to a battery of 22 tests of multivariate normality: the Kolmogorov-Smirnov and the Cramer-von Mises tests in their Stephens (1974) versions to

(i) each of the four univariate marginal distributions
(ii) each of the six bivariate angle distributions
(iii) the four-dimensional radius distribution

No significant departure from multivariate normality is found at the 5% level by any of these tests. A similar procedure applied to the kongite, boxite, and coxite data sets detected significant departures at the 5% level in some marginal tests only:

kongite: Cramer-von Mises test for y_2
coxite: Kolmogorov-Smirnov and Cramer-von Mises tests for y_1

In such a multiple hypothesis testing situation we must not be discouraged from continuing with a logistic-normal fit by one or two significances like this. Indeed the reader can now be let into the secret that the data were in fact simulated by an L^4 mechanism and yet we have a trace of apparently significant departure.

Estimation

The parameters μ and Σ are estimated in the usual way from the logratio data array Y as the sample mean vector y and the sample covariance matrix S_Y

$$y = \begin{bmatrix} 1.71 \\ 0.914 \\ 0.121 \\ 0.168 \end{bmatrix} \qquad S_Y = \begin{bmatrix} 0.134 & 0.255 & -0.212 & 0.117 \\ 0.255 & 0.624 & -0.668 & 0.139 \\ -0.212 & -0.668 & 0.895 & 0.012 \\ 0.117 & 0.139 & 0.012 & 0.180 \end{bmatrix} \tag{31}$$

Predictive Distribution

For purposes of embodying the experience we have gained from Table 1 of the compositional variability of hongite in a single distribution we recommend, for the various reasons presented in Aitchison and Dunsmore (1975), the predictive distributon based on the vaguest of priors. Within the simplex this would be what might be termed a logistic-Student distribution but we can work wholly within R^d in our subsequent discussion of typicality. For the y vector, therefore, the predictive distribution is simply, in the notation of Aitchison and Dunsmore (1975), the generalized Student distribution

$$St_4\{24, y, (1 + \tfrac{1}{25})S_Y\} \tag{32}$$

Atypicality Index

Is the composition (44.0, 19.7, 14.9, 9.1, 12.6) reasonably typical of the hongite experience or is it essentially an outlier? We can conveniently base our assessment on the following consideration. The predictive distribution assigns a probability density to each possible composition and the smaller the density

assigned to a composition the more it inclines to atypicality. First determine the density associated with the given composition. Then compute the probability, on the basis of the predictive distribution, that a hongite composition has a density greater than the density of the given composition and call this the atypicality index of the composition. The atypicality index, therefore, ranges between 0 and 1 and the closer it is to 1 the more atypical is the given composition. Fortunately the atypicality index associated with (32) is easily computed through the use of incomplete beta functions. See Aitchison and Dunsmore (1975, Sect. 11.4) for details. In the present case the atypicality index is 0.997 and so we must, therefore, express some doubt as to the specimen being an example of hongite.

Comparison of Hongite and Kongite

We may readily find the estimates of mean logratio vector and logratio covariance matrix for kongite as

$$y = \begin{bmatrix} 1.64 \\ 0.756 \\ 0.194 \\ 0.106 \end{bmatrix} \qquad S_y = \begin{bmatrix} 0.108 & 0.226 & -0.192 & 0.092 \\ 0.226 & 0.628 & -0.692 & 0.102 \\ -0.192 & -0.692 & 1.003 & 0.086 \\ 0.092 & 0.102 & 0.086 & 0.186 \end{bmatrix} \tag{33}$$

and we may then test hypotheses about similarities between hongite and kongite. For example, a standard multivariate normal test of equality of the two covariance matrices (Anderson, 1958, Sect. 10.6) shows no significant difference. If we follow this with a test of the hypothesis of the equality of the mean vectors we find with a pooled covariance matrix S, that the generalized T^2-statistic

$$\frac{n_1 n_2}{n_1 + n_2} (y^{(1)} - y^{(2)})^T S^{-1} (y^{(1)} - y^{(2)}) = 82.1 \tag{34}$$

where hongite and kongite sample sizes and logratio mean vectors are labeled by 1 and 2, respectively. The standard test (Anderson, 1958, Sect. 5.3.2) detects a significant difference between means at the 0.1% level of significance. This result suggests that it may be possible to devise a system of differential classification between hongite and kongite based on the geochemical compositions.

The reader may care to verify that the symmetric logcentered form of the Mahalanobis distance

$$(z^{(1)} - z^{(2)})^T S_Z^- (z^{(1)} - z^{(2)}) \tag{35}$$

with the pseudo-inverse S_Z^- computed as in (13), leads to the same result as the above asymmetric version. Reference to Section 3 will also show that we have

now effectively answered question 1, and made a start on the answer to question 4.

10. AN EXAMPLE OF LINEAR MODELING

The transformed, logratio composition vector can be subjected to linear modeling to examine its possible dependence on explanatory or concomitant variables through standard multivariate linear hypothesis testing. For the hongite four-dimensional logratio composition y found at depth u we may consider the model

$$y \text{ is } N^4(\alpha + \beta u, \Sigma) \tag{36}$$

and test the hypothesis of no dependence of composition on depth

$$\beta = 0 \tag{37}$$

by well-established and computer-packaged procedures; see, for example, Morrison (1976, Sect. 5.2). For boxite and coxite the test statistics (Morrison, 1976, p. 167, formula 41) have values 1.24 and 10.53, both to be compared against upper percentage points of $F(4,20)$. Thus, there is strong evidence, at the 0.1% level, that the coxite compositions depend on depth but no such evidence for boxite.

Note that there is no requirement for the modeling to be linear in u. For example, Aitchison (1982), in investigating the dependence of Arctic lake sediment compositions on depth, finds that forms $\alpha + \beta \log u$ or $\alpha + \beta u + \beta u^2$ are necessary for an adequate description of the dependence.

11. CONCEPTS OF COMPOSITIONAL INDEPENDENCE

Over-concentration on standard ideas and absence of an appropriate covariance structure have limited the scope of investigation by geologists of the possibilities of different forms of independence within the composition. The simplest approach to independence concepts for compositions is to consider various forms of independence in relation to a single division $(x_1, \ldots, x_c | x_{c+1}, \ldots, x_{d+1})$ of the composition and the corresponding transformation to the partition (s_1, s_2, t) given by (5). We can define three different forms of independence in descending order of strength of independence, which go a substantial way toward completing our answer to question 2 of Section 4.

(i) *Partition independence:* s_1, s_2, t independent

(ii) *Neutrality on the right:* (s_1, t) and s_2 independent
Neutrality on the left: s_1 and (s_2, t) independent
Neutrality: neutrality both on the right and on the left

(iii) *Subcompositional independence:* s_1 and s_2 independent

Partition independence and subcomposition independence were introduced by Aitchison (1982). Neutrality on the right, introduced by Connor and Mosimann (1969) in relation to problems of biological growth, can be expressed equivalently as independence of the left-hand subvector $(x_1, \ldots, x_c)$ and the right-hand subcomposition $C(x_{c+1}, \ldots, x_{d+1})$. If we imagine that the composition is determined by components $1, \ldots, c$, first assuming their values, then we have neutrality on the right if the relative magnitudes of the other components are quite uninfluenced by the actual values adopted by the first c components. Partition independence implies neutrality on the right (and equally its counterpart neutrality on the left) and neutrality of any form implies subcompositional independence. When we demand that these forms of independence hold for any partition and any permutation of the components of the composition then we have much more extreme forms of independence which we term *complete*.

Completeness is a strong form of independence and indeed provides characterizations of the Dirichlet class: any composition with complete partition independence or complete neutrality necessarily has a Dirichlet distribution. Since the L^d and D^d classes are separate this means that the logistic-normal class has a limitation in that such extreme forms of independence cannot be tested within its framework. Aitchison (1984b) has recently shown how to construct a more general class A^d on the simplex which includes as special cases the D^d and L^d classes and so provides a framework for testing these extreme (and, therefore, Dirichlet) forms. Since it seems unlikely that any geochemical compositions have such extreme forms of independence we do not pursue these tests here.

It is not possible within the scope of this paper to discuss in detail the variety of modeling and test procedures appropriate to the investigation of all these different forms of independence and, as illustration, we concentrate on one particular form, complete subcompositional independence. For a fuller discussion and applications of other forms of independence, see Aitchison (1982). Only one point of particular interest need be mentioned, namely that for the investigation of neutrality on the right a multiplicative form of logistic transformation from S^d to R^d is required (Aitchison, 1981b)

$$x_i = \begin{cases} \exp(y_i) \Big/ \prod_{j=1}^{i} [1 + \exp(y_j)] & (i = 1, \ldots, d) \\ 1 \Big/ \prod_{j=1}^{d} [1 + \exp(y_j)] & (i = d + 1) \end{cases} \tag{38}$$

with inverse

$$y_i = \log [x_i/(1 - x_1 - \cdots - x_i)] \qquad (i = 1, \ldots, d) \tag{39}$$

The form of independence which has been the goal of the work on relationships between open and closed sets, such as by Chayes and Kruskal (1966), is *complete subcompositional independence* which can be conveniently redefined, *in terms of the composition only*, as follows

> *Complete subcompositional independence:* A composition has complete subcompositional independence if the subcompositions formed from any set of nonoverlapping subvectors are mutually independent.

Aitchison (1982) has shown that complete subcompositional independence corresponds to a simple form of the logratio covariance matrix Σ, namely that its off-diagonal elements are all equal, or equivalently that

$$\Sigma = \text{diag}(\lambda_1, \ldots, \lambda_d) + \lambda_{d+1} J_d \tag{40}$$

Every member of the Dirichlet class has complete subcompositional independence and so yet again there is no way of testing this hypothesis within the Dirichlet class. On the other hand within the additive logistic normal class $L^d(\mu, \Sigma)$ complete subcompositional independence corresponds to the parametric hypothesis (40) which may be tested by standard test procedures such as generalized likelihood ratio tests. The method, reported in Aitchison (1982), uses the Wilks (1938) asymptotic test with critical region

$$2(l_M - l_H) > \text{upper percentage point of the } \chi^2\left[\tfrac{1}{2}(d+1)(d-2)\right] \text{ distribution} \tag{41}$$

where l_H and l_M are the maximized values of the loglikelihood function under the hypothesis H of complete subcompositional independence and under the model M of logistic normality. The only problem is the technical one of maximizing the likelihood with the special form (40) for the covariance structure but this need not concern us in this exposition.

The computed values of the test statistic (41) for hongite, kongite, boxite, and coxite are 262.7, 251.5, 3.64, 34.5, to be compared against upper percentage points of $\chi^2(5)$. There is, thus, overwhelming evidence against complete subcompositional independence of hongite, kongite, and coxite but it is certainly a tenable hypothesis for boxite. The differences in the forms of the hongite and boxite logratio covariance matrices, given in Table 7, are indeed fairly obvious to the naked eye.

Finally, we may note that there are forms of partial subcompositional independence which correspond to geological concepts such as *concretionary* and *metasomatic*, as introduced by Sarmanov and Vistelius (1959) within the context of open and closed variables and crude covariance structures. These correspond to hypotheses such as: s_1 and s_2 are independent and s_2, as a $(d - c)$-dimensional

Table 7. Logratio Covariance Matrices for Hongite and Boxite

Hongite $S_Y =$

$$\begin{bmatrix} 0.1339 & 0.2547 & -0.2116 & 0.1173 \\ 0.2547 & 0.6248 & -0.6681 & 0.1394 \\ -0.2116 & -0.6681 & 0.8945 & 0.0122 \\ 0.1173 & 0.1394 & 0.0122 & 0.1797 \end{bmatrix}$$

Boxite $S_Y =$

$$\begin{bmatrix} 0.0314 & 0.0122 & 0.0135 & 0.0020 \\ 0.0122 & 0.1308 & -0.0091 & -0.0012 \\ 0.0135 & -0.0091 & 0.0461 & 0.0098 \\ 0.0020 & -0.0012 & 0.0098 & 0.0510 \end{bmatrix}$$

composition, has complete subcompositional independence. Such independence hypotheses are readily tested within the framework outlined above.

12. CLASSIFICATION

A first step in answering question 4 of Section 3 is to investigate whether our past experience of the pattern of variability of the geochemical compositions of hongite and kongite as contained in Tables 1 and 2 will allow us to devise a process for the differential classification of new rock samples known to be of one of those types. The facts of Section 9 that we can fit additive logistic-normal models to each of these data sets and that we find a significant difference between the logratio vector means of the two models suggests that there is a reasonable chance of success. Since the hongite and kongite logratio covariance matrices are not significantly different we would be justified in applying standard discriminant analysis techniques to the logratio compositions. Many geologists are already familiar with this technique, albeit in relation to the questionable practice of applying it to crude proportions, and so it would serve little purpose to go over essentially familiar ground. Instead I will present an adaptation, suited to compositional data, of a well-established, although apparently, in geological circles, less familiar, alternative. In the analogous problem of differential diagnosis in medicine there are often substantial grounds for believing that this alternative method will provide more reliable indications of disease type and the interested reader may care to examine the arguments put forward simply and cogently in Dawid (1976) and to ask himself whether they are not equally valid in a context of geological classification. There is, however, another substantial reason for choosing this alternative approach. It provides a very simple technique

for investigating the extent to which use of only a subcomposition is an effective means of classification.

We briefly illustrate the techniques for two types only, namely hongite and kongite. The method readily extends to more than two types and the reader interested in greater detail and in applications to classifications of rock types from major oxide compositions may refer to Aitchison and Li (1985).

With the transformed logratio composition y we may write the logistic discriminant model in the following form

$$p(\text{type } 1 \mid x, \beta) = \frac{\exp(\beta_0 + \beta_1 y_1 + \cdots + \beta_d y_d)}{1 + \exp(\beta_0 + \beta_1 + \cdots + \beta_d y_d)}$$
$$= 1 - p(\text{type } 2 \mid x, \beta) \tag{42}$$

from which, by standard maximum likelihood estimation methods, we obtain maximum likelihood estimates of β as

$$\hat{\beta} = \begin{bmatrix} -258.5 \\ 100.8 \\ 113.1 \\ 110.0 \\ -162.1 \end{bmatrix} \tag{43}$$

We then regard classification as consisting of the assignment of type probabilities to new rock samples and for this purpose, for the many reasons advocated in Aitchison and Dunsmore (1975), we use the predictive diagnostic method which, for the logistic discriminant case, can be expressed in the approximate form (Lauder, 1978)

$$p(\text{hongite} \mid x, \text{data}) = \Phi[\hat{\beta}^T y^* / (2.89 + y^{*T} V y^*)^{1/2}] \tag{44}$$

where $\Phi(\cdot)$ is the $N(0, 1)$ distribution function, y^* is the extended logratio vector $[1\ y^T]^T$ and V is the estimated covariance matrix of the estimator of β, here

$$V = \begin{bmatrix} 7126 & -2785 & -3113 & -3023 & 4487 \\ -2785 & 1095 & 1203 & 1173 & -1747 \\ -3113 & 1203 & 1385 & 1336 & -1974 \\ -3023 & 1173 & 1336 & 1292 & -1912 \\ 4487 & -1747 & -1974 & -1912 & 2835 \end{bmatrix} \tag{45}$$

When reapplied to the 50 cases of the data set the probabilities assigned are distributed as shown in Table 8, and so it is seen that we have a reasonable

Table 8. Distributions of Predictive Probabilities for Classification as Hongite by the Crude and Logratio Methods

Probability interval	Crude method		Logratio method	
	Hongite	Kongite	Hongite	Kongite
0–0.05	–	–	–	–
0.05–0.10	–	–	–	4
0.10–0.15	–	–	–	6
0.15–0.20	–	–	1	4
0.20–0.25	–	–	–	1
0.25–0.30	–	–	–	2
0.30–0.35	–	–	–	2
0.35–0.40	–	–	1	1
0.40–0.45	–	–	–	1
0.45–0.50	13	17	1	2
0.50–0.55	12	8	1	1
0.55–0.60	–	–	–	1
0.60–0.65	–	–	–	–
0.65–0.70	–	–	2	–
0.70–0.75	–	–	1	–
0.75–0.80	–	–	4	–
0.80–0.85	–	–	3	–
0.85–0.90	–	–	6	–
0.90–0.95	–	–	5	–
0.95–1.00	–	–	–	–

method of classification, even making allowances for the well-known fact that assessing a system by resubstitution of the cases from which the system was constructed always gives an overly optimistic view. The complete failure of the corresponding crude analysis in which $y_1, \ldots, y_d$ in (42) are replaced by any d of the raw proportions is demonstrated in Table 8 by the very poor discrimination.

For the examination of the effectiveness of a subcomposition the model (42) can be expressed more conveniently in a symmetrical version

$$p(\text{hongite}\,|x,\beta) = \frac{\exp(\beta_0 + \beta_1 \log x_1 + \cdots + \beta_{d+1} \log x_{d+1})}{1 + \exp(\beta_0 + \beta_1 \log x_1 + \cdots + \beta_{d+1} \log x_{d+1})} \tag{46}$$

where $\beta_1 + \cdots + \beta_{d+1} = 0$. The hypothesis that the subcomposition $C(x_1, \ldots, x_c)$ is just as effective as the complete composition can then be regarded as the parametric hypothesis that $\beta_{c+1} = \cdots = \beta_{d+1} = 0$, and so can be tested by standard parametric hypothesis methods. All we have to do is to obtain the maximized loglikelihood l_M under the model M that the full composition is required and also the maximized loglikelihood l_H under the subcompositional hypothesis H, and this latter can be done by standard logistic discriminant analysis using the

Table 9. Maximized Loglikelihoods and Values of the Test Statistics for Investigating the Discriminatory Power of Subcompositions in the Classification of Hongite and Kongite

Composition or subcomposition	Maximized loglikelihood	$2(l_M - l_H)$
(A, B, C, D, E)	−11.42	–
Omitting A	−31.25	39.7
B	−34.00	45.2
C	−34.35	45.9
D	−34.06	45.3
E	−34.44	46.0

subcomposition in its logratio form as if it were the full composition. Then if $2(l_M - l_H)$ exceeds the upper percentage point of $\chi^2(d - c + 1)$ we have to conclude that the subcomposition is not a satisfactory substitute.

For the classification of hongite and kongite we can readily test the effectiveness of each of the five four-part subcompositions obtained by dropping out just one of the components. Table 9 shows the maximized loglikelihoods, and it is clear that we have highly significant test statistics when compared against $\chi^2(1)$ percentiles. It is, thus, all too clear that attempting to use any subcomposition for this classification purpose is fraught with disaster. This can easily be verified by assessing the classification probabilities obtained by using a subcomposition.

13. COMPOSITION AS A CONCOMITANT OR EXPLANATORY VECTOR

How may the porosity of coxite depend upon the composition? This is the final question 6 of Section 4 to be answered. If we regard porosity as the response to the mixture of geochemical components then we have a problem identical in form to what is traditionally classified under the heading of experiments with mixtures, as for example, in Becker (1968, 1978), Cox (1971), and Cornell (1981). Here we do not follow any of the traditional models but adopt a recent approach more consistent with the concepts of compositions advocated here, a method that allows easy testing of hypotheses of inactivity of some components and also certain forms of additivity. For a full discussion of these concepts and some applications see Aitchison and Bacon-Shone (1984). We here describe only briefly the approach that might be considered and illustrate it for coxite porosity.

Suppose that we adopt the usual least-squares model associated with mixtures but with the expected response $\eta(x)$ of quadratic logcontrast form in the

composition $(x_1, \ldots, x_{d+1})$

$$\eta(x) = \beta_0 + \sum_{i=1}^{d+1} \beta_i \log x_i + \sum_{i=1}^{d} \sum_{j=2}^{d+1} \beta_{ij} (\log x_i - \log x_j)^2 \tag{47}$$

with

$$\beta_1 + \cdots + \beta_{d+1} = 0 \tag{48}$$

and, thus, involving $\frac{1}{2}(d+1)(d+2)$ unconstrained parameters.

Then two hypotheses of interest with respect to the partition $(x_1, \ldots, x_c | x_{c+1}, \ldots, x_{d+1})$ may be the following

(i) *Inactivity of* $(x_1, \ldots, x_c)$: $\eta(x)$ does not depend on $(x_1, \ldots, x_c)$. This is expressible as the parametric hypothesis H_1

$$\beta_i = 0 \quad (i = 1, \ldots, c) \qquad \beta_{ij} = 0 \quad (i = 1, \ldots, c; j > i) \tag{49}$$

placing $\frac{1}{2}c(2d - c + 3)$ constraints on the parameters.

(ii) *Additivity with respect to the partition:* $\eta(x)$ can be expressed as a sum of two separate functions, one in $(x_1, \ldots, x_c)$ and the other in $C(x_{c+1}, \ldots, x_{d+1})$ or, equivalently for this symmetric expected response function, as one in $C(x_1, \ldots, x_c)$ and the other in $(x_{c+1}, \ldots, x_{d+1})$. This corresponds to the parametric hypothesis H_2

$$\beta_{ij} = 0 \quad (i = 1, \ldots, c; j = c+1, \ldots, d+1) \tag{50}$$

placing $c(d + 1 - c)$ constraints on the parameters.

Clearly inactivity H_1 implies additivity H_2, and so for the coxite porosity data and the partition $(A, B | C, D, E)$ we illustrate the testing procedure by testing the simpler hypothesis H_1 first and proceed to the more composite hypothesis H_2 only if we reject H_1. Here $c = 2$, $d = 4$, and the model M or (47) has parameter dimension 15, and hypotheses H_1 and H_2 have dimensions 6 and 9, respectively. Standard least-squares calculations give the following residual sums of squares

$$R_M = 13.2 \qquad R_{H_1} = 62.2 \qquad R_{H_2} = 19.4$$

Hence the usual F test of H_1 within M compares test statistic value 4.1 against upper percentage points of $F(6,10)$ and so rejects the inactivity hypothesis H_1 at the 5% level of significance. Subsequent testing of H_2 within M compares 0.78 against $F(9,10)$ values and so we have to conclude that partition additivity is a tenable hypothesis.

14. DISCUSSION

There remain aspects of current research and open questions in compositional data analysis on which it is possible to comment only briefly.

Zeros

When zero values in the components are persistent and cannot be ascribed to the recording of just a trace or to rounding off, conditional modeling along the lines of Aitchison (1982, Sect. 7.4) is necessary. Where the zeros are of the rounding-off type, consideration of the rounding-off process indicates a reasonable procedure. Any recorded composition in the interior of the simplex represents a whole polyhedron of compositions which round to the recorded composition at the center of the polyhedron. When there are zero components the recorded composition lies on the boundary of the simplex and also on the boundary of the set of compositions it represents. It is then reasonable to replace the recorded composition by a point interior to the set, say at the geometric center, in which process the zeros become positive. Robustness of the process can be investigated by varying the assumed maximum rounding error. See Aitchison (1982) for further possible approaches.

Measurement Error

Situations in which the observed composition X may differ from the true composition x may be readily investigated through the use of a *perturbation* error model

$$X = x \circ u = C(x_1 u_1, \ldots, x_{d+1} u_{d+1}) \tag{51}$$

a multiplicative form of error model which leads to tractable analysis (see Aitchison and Shen, 1984, for further details).

Nonparametric Methods

The parametric classes of transformed-normal distributions on S^d may fail to provide an adequate description and so the question of how to handle compositional data nonparametrically arises. For those nonparametric methods which depend on the idea of distance we suggest that the Euclidean squared distance $\Sigma(x_i - X_i)^2$ between two compositions may prove unsatisfactory because of its association with covariance structure involving raw proportions. As an alternative

$$\sum_{i=1}^{d+1} \{\log [x_i/g(x)] - \log [X_i/g(X)]\}^2 \qquad \text{or} \qquad \sum_{i=1}^{d+1} (x_i - X_i) \log (x_i/X_i) \tag{52}$$

may prove more useful. For problems, such as discrimination where density function estimates are required, it is possible to devise suitable kernel methods using Dirichlet or logistic-normal kernels (see Aitchison and Lauder, 1984, for details). An alternative procedure adapting projection pursuit methods to the special features of the simplex sample space is currently being investigated.

Another View of Covariance Structure

The logratio covariance structure $\Sigma = [\sigma_{ij}]$ is completed determined by the $\frac{1}{2}d(d+1)$ logratio variances

$$\tau_{ij} = \tfrac{1}{2}\,\mathrm{var}\,(\log x_i/x_j) \qquad (i \neq j = 1, \ldots, d+1) \tag{53}$$

through the relationships

$$\sigma_{ij} = \tau_{i,d+1} + \tau_{j,d+1} - \tau_{ij} \tag{54}$$

which can be expressed more compactly as

$$\Sigma = -BTB^T \tag{55}$$

where $T = [\tau_{ij}]$ is a $(d+1) \times (d+1)$ matrix of logratio *variances* with zero diagonal elements. There is some attraction toward discussing compositional covariance structure in terms of T since it treats the components symmetrically and provides a measure of the relative variation of every pair of components, with τ_{ij} being identical to the measure of total variability of the subcomposition $C(x_i, x_j)$ as discussed in Section 7. Another attractive feature is that complete subcompositional independence for this symmetrical form is equivalent to the expression of the τ_{ij} in additive form

$$\tau_{ij} = \lambda_i + \lambda_j \qquad (i \neq j) \tag{56}$$

The Future of Compositional Data Analysis

It must be clear to any reader that much remains to be done in developing and applying new and sound techniques for the analysis of compositional data. I hope that with the dismissal of crude forms of analysis, with a fair appraisal of the new methodology, and the development of even better concepts and tools we may make up for some of the lost 80 years of neglect of Pearson's 1897 warning.

REFERENCES

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

Aitchison, J., 1981b, Distributions on the simplex for the analysis of neutrality, *in*, C. Taillie, G. P. Patil, and B. Baldessari (Eds.), Statistical distributions in scientific work: D. Reidel, Dordrecht, Holland, p. 147–156.

Aitchison, J., 1982, The statistical analysis of compositional data (with discussion): 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.

Aitchison, J., 1984a, Reducing the dimensionality of compositional data sets: Jour. Math. Geol., v. 16, to appear.

Aitchison, J., 1984b, A general class of distributions on the simplex: Jour. Roy. Stat. Soc. Ser. B., v. 46, to appear.

Aitchison, J. and Bacon-Shone, J. H., 1984, A logcontrast approach to experiments with mixtures: Biometrika, v. 71, to appear.

Aitchison, J. and Dunsmore, I. R., 1975, Statistical prediction analysis: Cambridge University Press.

Aitchison, J. and Lauder, I. J., 1984, Kernel density estimation for compositional data: submitted to Appl. Stat.

Aitchison, J. and Li, C. K. T., 1985, A new approach to classification from compositional data: submitted to Jour. Math. Geol.

Aitchison, J. and Shen, S. M., 1980, Logistic-normal distributions: Some properties and uses: Biometrika, v. 67, p. 261–272.

Aitchison, J. and Shen, S. M., 1984, Measurement error in compositional data: Jour. Math. Geol., v. 16, p. 637–650.

Anderson, T. W., 1958, An Introduction to Multivariate Statistical Analysis: John Wiley & Sons, New York.

Barker, D. S., 1978, Magmatic trends on alkali–iron–magnesium diagrams: Amer. Min., v. 63, p. 531–534.

Becker, N. G., 1968, Models for the response of a mixture: Jour. Roy. Stat. Soc. Ser. B., v. 30, p. 349–358.

Becker, N. G., 1978, Models and designs for experiments with mixtures: Aust. Jour. Stat., v. 20, p. 195–208.

Butler, J. C., 1979, Trends in ternary petrologic variation diagrams–fact or fantasy?: Amer. Min., v. 64, p. 1115–1121.

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

Chayes, F., 1962, Numerical correlation and petrographic variation: Jour. Geol., v. 70, p. 440–452.

Chayes, F., 1971, Ratio correlation: University of Chicago Press, Illinois.

Chayes, F., 1983, Detecting nonrandom associations between proportions by tests of remaining-space variables: Jour. Math. Geol., v. 15, p. 197–206.

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

Connor, R. J. and Mosimann, J. E., 1969, Concepts of independence for proportions with a generalization of the Dirichlet distribution: Jour. Amer. Stat. Assoc., v. 64, p. 194–206.

Cornell, J. A., 1981, Experiments with mixtures: John Wiley & Sons, New York.

Cox, D. R., 1971, A note on polynomial response functions for mixtures: Biometrika, v. 58, p. 155–159.

Dawid, A. P., 1976, Properties of diagnostic data distributions: Biometrics, v. 32, p. 647–658.

Iddings, J. P., 1903, Chemical composition of igneous rocks: U.S. Geol. Survey Prof. Pap. 18.

James, I. R., 1981, Distributions associated with neutrality properties for random proportions, *in*, C. Taillie, G. P. Patil, and B. Baldessari (Eds.), Statistical distributions in scientific work: D. Reidel, Dordrecht, Holland, p. 125–136.

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

Krumbein, W. C., 1962, Open and closed number systems in stratigraphic mapping: Bull. Amer. Assoc. Pet. Geol., v. 46, p. 2229–2245.

Lauder, I. J., 1978, Computational problems in predictive diagnosis: Compstat 1978, p. 186–192.

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

Le Maitre, R. W., 1982, Numerical petrography: Elsevier, Amsterdam.

McAlister, D., 1879, The law of the geometric mean: Proc. Roy. Soc., v. 29, p. 367.

Miesch, A. T., 1969, The constant sum problem in geochemistry, *in* D. F. Merriam (Ed.), Computer Applications in the earth sciences: Plenum Press, New York, p. 161–167.

Morrison, D. F., 1976, Multivariate statistical methods: New York, McGraw-Hill.

Pearson, K., 1897, Mathematical contributions to the theory of evolution. On a form of spurious correlations which may arise when indices are used in the measurement of organs: Proc. Roy. Soc., v. 60, p. 489–498.

Reyment, R., 1977, Presidential address to the International Association for Mathematical Geology: Jour. Math. Geol., v. 9, p. 451–454.

Sarmanov, O. V. and Vistelius, A. B., 1959, On the correlation of percentage values: Dokl. Akad. Nauk. SSSR, v. 126, p. 22–25.

Stephens, M. A., 1974, EDF statistics for goodness of fit and some comparisons: Jour. Amer. Stat. Assoc., v. 69, p. 730–737.

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.

Wilks, S. S., 1938, The large-sample distribution of the likelihood ratio for testing composite hypotheses: Ann. Math. Stat., v. 9, p. 60–62.

Zukav, G., 1979, The dancing Wu-Li masters: Bantam, New York.