The One-hour Course in Compositional Data Analysis or Compositional data analysis is simple.

John Aitchison

Department of Statistics, The University of Glasgow, Glasgow G12 8QQ, Scotland

Compositional data analysis has long been regarded as difficult because of the so-called constant sum constraint. The contrary is the case. Attention to a few simple logical necessities of compositional data analysis, such as scale invariance, subcompositional coherence, perturbation as the mechanism of change of a composition, leads inevitably to the recognition of the unit simplex as a sensible sample space with its associated and relevant statistical methodology. Such a recognition dictates radical changes in compositional thinking. For example, use of the arithmetic mean of a set of compositions as the centre of the compositional data set can be absurd; a more meaningful centre is based on geometric means. The methodology is most simply described in terms of a transformation technique involving logratios of the components of the compositions. Since any meaningful statement about a composition can be expressed in terms of logratios the method is applicable throughout the complete range of compositional problems. A number of compositional data sets is used to illustrate a variety of practical situations and appropriate methods such as biplots, predictive distributions, atypicality indices, joint variability models, conditional models and convex linear combinations of compositions are demonstrated. Finally extensions are considered and a plea is made to geologists for more precision in specifying geological hypotheses.

KEY WORDS: biplot, conditional model, endmember problem, joint variability, logratio analysis, perturbation, scale invariance, simplex, subcomposition.

A LITTLE BIT OF HISTORY

This is 1997 and we must look back to 1897 for our starting point. One hundred years ago Karl Pearson, undoubtedly one of the great-grandfathers of modern statistics, published in Pearson (1897) one of the clearest warnings ever issued to statisticians and other scientists beset with uncertainty and variability: Beware of attempts to interpret correlations between ratios whose numerators and denominators contain common parts. And of such is the world of compositional data, where for example some rock specimen, of total weight w, is broken down into mutually exclusive and exhaustive parts with component weights $w_1, ..., w_D$ and then transformed into a composition $(x_1, ..., x_D) = (w_1, ..., w_D)/(w_1 + \cdots + w_D)$. Our reason for forming such a composition is that in many problems composition is the relevant entity. For example the comparison of rock specimens of different weights can only be achieved by some form of standardisation and composition (per unit weight) is a simple and obvious concept for achieving this. Equivalently we could say that any meaningful statement about the rock specimens should not depend on the largely accidental weights of the specimens.

It is a great pity that Pearson's warning went unheeded for so long and that so many are still unaware of it and of possible remedies. I hope I can assume that you, as mathematical geologists, accept the deductive power of mathematics and also, dare I hope, the inductive strength of statistics. Such an acceptance and some very simple mathematics lead inevitably to certain theoretical requirements of any meaningful discussion of compositions. The role of statistics is then to develop an appropriate methodology for the practical analysis of compositional data (Aitchison, 1994).

SCALE INVARIANCE: THE FUNDAMENTAL PRINCIPLE OF COMPOSITIONAL DATA ANALYSIS

When we say that a problem is compositional we are recognising that the sizes of our specimens are irrelevant. When you as geologists talk about the composition of an object such as the major oxide composition of a rock you imply that you are interested in a dimensionless problem. You are not concerned whether the rock weighs one gm or one lb. This trivial admission has far-reaching consequences. Let us apply some clear thinking to acceptance of this fundamental scale invariance principle.

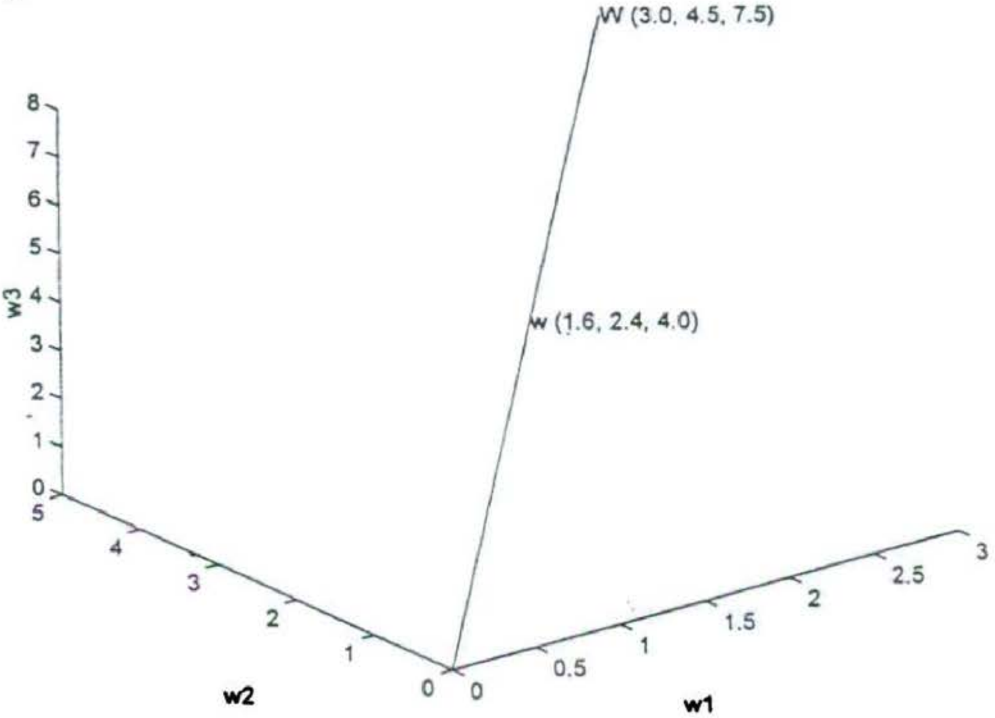


Figure 1. Representation of equivalent compositions as points on rays of the positive orthant.

A simple example can illustrate the argument. In Fig. 1, which shows three-dimensional positive space R^3_+, the two points $w(1.6, 2.4, 4.0)$ and $W(3.0, 4.5, 7.5)$ represent the weights of the three parts (a, b, c) of two specimens of total weight 8 gm and 15 gm, respectively. If we are interested in compositional problems we recognise that these are of the same composition with the difference in weight being taken account of by the scale relationship $W = (15/8)w$. More generally two compositions w and W are compositionally equivalent, written $W \sim w$, when there exists a positive proportionality constant p such that $W = pw$. The fundamental requirement of compositional data analysis can then be stated as follows: any meaningful construct or function f of a composition must be such that $f(W) = f(w)$ when $W \sim w$, or equivalently

$$f(pw) = f(w) \text{ for every } p > 0. \tag{1}$$

This is a common problem in group theory: the requirement (1) is that the function f must be *invariant under the group of scale transformations*. A general result of group theory is that any group invariant function can be expressed as a function of a maximal invariant. Now a function g is a maximal invariant when $g(W) = g(w)$

implies $W \sim w$. Here it is trivial to show that the $(D-1)$-vector function

$$f(w) = \left(\frac{w_1}{w_D}, ..., \frac{w_{D-1}}{w_D} \right)$$

is a maximal invariant. The important consequence of this is the following.

Any meaningful (scale-invariant) function of a composition can be expressed in terms of ratios of the components of the composition.

Note the one-to-one correspondence between the components of w and a set of independent and exhaustive ratios such as

$$\begin{aligned} r_i &= \frac{w_i}{w_1 + \cdots + w_D} \quad (i = 1, ..., D-1), \\ r_D &= \frac{1}{w_1 + \cdots + w_D}, \end{aligned} \tag{2}$$

with the components of w determined by these ratios as

$$\begin{aligned} w_i &= \frac{r_i}{r_1 + \cdots + r_{D-1} + 1} \quad (i = 1, ..., D-1), \\ w_D &= \frac{1}{r_1 + \cdots + r_{D-1} + 1}. \end{aligned} \tag{3}$$

Note also that there are many equivalent sets of ratios which may be used for the purpose of creating meaningful functions of compositions. For example, a more symmetric set of ratios such as $w/g(w)$, where $g(w) = (w_1 \cdots w_D)^{1/D}$ is the geometric mean of the components of w, would equally meet the scale-invariant requirement.

All that this blinding by mathematics is saying is surely the obvious. Compositions are concerned with relative values and so ratios of components. When I first became interested in compositional data I thought that this was self-evident, but apparently not; see, for example, the sequence of letters (Aitchison, 1990a, b; 1991; Watson, 1990; 1991) in *Mathematical Geology* arising from Watson and Philip (1989) and ending with Aitchison (1992).

CHOICE OF SAMPLE SPACE

In my own teaching over the last 45 years I have issued a warning to all my students, similar to that of Pearson. Ignore the clear definition of your sample space at your peril. When faced with a new situation the first thing you must resolve before you do anything else is an appropriate sample space. On occasions when I have found some dispute between students over some statistical issue the question of which of them had appropriate sample spaces has almost always determined which students are correct in their conclusions. If, for example, it is a question of association between the directions of departure and return of migrating Barcelona swallows then an appropriate sample space is a doughnut.

Now I am sure that, with the exception of that man who mistook his wife for his hat, we recognise that a rectangular box, a tetrahedron, a sphere and a doughnut look rather different. It should come as no surprise to us therefore that problems with these different sample spaces might require completely different statistical methodologies. It has always seemed surprising to me that the direction data analysts had little difficulty in seeing that the sphere and the torus require their own special methodology, whereas for so long statisticians, geologists and all other scientists seemed to think that what was good enough for a box was good enough for a tetrahedron. And there certainly can be little doubt that most geologists are aware of the true nature of the sample space. Which of you has not constructed or at least studied a ternary diagram?

In the first step of statistical modelling, namely specifying a sample space, the choice is with the modeller. It is how the sample space is used or exploited to answer relevant problems that is important. We might, as in our study of scale invariance above, take the set of rays through the origin and in the positive orthant as our sample space. The awkwardness here is that the notion of placing a probability measure on a set of rays is less familiar than on a set of points. Moreover we know that as far as the study of compositions is concerned any point on a ray can be used to represent the corresponding composition. The selection of each representative point x where the rays meet the unit hyperplane $w_1 + \cdots + w_D = 1$ with $x = w/(w_1 + \cdots + w_D)$ is surely the simplest form of standardisation possible. We shall thus adopt the unit simplex

$$\mathcal{S}^D = \{(x_1, ..., x_D) : x_i > 0 \quad (i = 1, ..., D), \quad x_1 + \cdots + x_D = 1\}. \tag{4}$$

To avoid any confusion on terminology for our generic composition x we refer to the labels $1, ..., D$ as the *parts* and the proportions $x_1, ..., x_D$ as the *components* of the composition x. With this representation we shall continue to ensure scale invariance by formulating all our statements concerning compositions in terms of ratios of components. Our next logical requirement will reinforce the good sense of this action.

SUBCOMPOSITIONAL COHERENCE

The concept of a subcomposition such as the (Na_2O, K_2O, Al_2O_3) subcomposition of a major-oxide composition of a rock is familiar to geologists. Formally the subcomposition based on parts $(1, 2, ..., C)$ of a D-part composition $(x_1, ..., x_D)$ is the $(1, 2, ..., C)$-subcomposition $(s_1, ..., s_C)$ defined by

$$(s_1, ..., s_C) = (x_1, ..., x_C)/(x_1 + \cdots + x_C), \tag{5}$$

the *closure* operation, so familiar to geologists and filling the role that marginals (subvectors or projections) play in the study of unconstrained vector data. Less familiar is another logical necessity of compositional analysis, namely subcompositional coherence.

Let us consider two scientists A and B who are interested in soil samples, which have been divided into aliquots. For each aliquot A records a 4-part composition (animal, vegetable, mineral, water); B first dries each aliquot without recording the water content and arrives at 3-part composition (animal, vegetable, mineral). Now let us suppose for simplicity the ideal situation where the aliquots in each pair are identical and where the two scientists are absolutely accurate in their determinations. Then clearly B's 3-part composition (s_1, s_2, s_3) for an aliquot will be a subcomposition of A's 4-part composition (x_1, x_2, x_3, x_4) for the corresponding aliquot related as above with $C = 3$, $D = 4$. It is surely obvious that any compositional statements that A and B make about the common parts, animal, vegetable and mineral, must agree. This is the nature of subcompositional coherence.

A simple example illustrates the lack of subcompositional coherence between scientists who use product moment correlation of raw components. Consider the simple data set:

Full compositions (x_1, x_2, x_3, x_4)	**Subcompositions (s_1, s_2, s_3)**
$(0.1, 0.2, 0.1, 0.6)$	$(0.25, 0.50, 0.25)$
$(0.2, 0.1, 0.1, 0.6)$	$(0.50, 0.25, 0.25)$
$(0.3, 0.3, 0.2, 0.2)$	$(0.375, 0.375, 0.25)$

Scientist A would report the correlation between animal and vegetable as $\mathrm{corr}(x_1, x_2) = 0.5$ whereas B would report $\mathrm{corr}(s_1, s_2) = -1$. There is thus incoherence of the product-moment correlation between raw components as a measure of dependence.

Note, however, that the ratio of two components remains unchanged when we move from full composition to subcomposition: $s_i/s_j = x_i/x_j$ so that, as long as we work with scale invariant functions, or equivalently express all our statements about compositions in terms of ratios, we shall be subcompositionally coherent.

PERTURBATION: THE FUNDAMENTAL COMPOSITIONAL OPERATION

The role of group operations in statistics

For every sample space there are basic operations which when recognised dominate clear thinking about data analysis. For example in the use of D-dimensional real space R^D as a sample space for unconstrained vectors, two such vectors y and Y can always be fully related by asking what transformation is required to change y into Y. The answer is in the operation of a *translation* t where $Y = y + t$, or equivalently by the inverse translation $y = Y - t$. Moreover this relationship between y_1 and Y_1 is the same as that between y_2 and Y_2 if and only if Y_1 and Y_2 are equal translations t of y_1 and y_2. Any definition of a difference or a distance measure must thus be such that the measure is the same for (y_1, Y_1) as for $(y_1 + t, Y_1 + t)$ for every translation t. Technically this is a requirement of invariance under the group of translations. It is the reason, though seldom expressed because of its obviousness in this simple space, for the use of the mean vector $\mu = \mathrm{E}(y)$ and the covariance matrix $\Sigma = \mathrm{V}(y) = \mathrm{E}\{(y - \mu)(y - \mu)^T\}$ as meaningful measures of central tendency and dispersion. Recall also for further reference two basic properties: for a fixed translation t,

$$\mathrm{E}(y + t) = \mathrm{E}(y) + t, \quad \mathrm{V}(y + t) = \mathrm{V}(y). \tag{6}$$

Similar considerations of groups of fundamental operations are essential for other sample spaces. For example, in the analysis of directional data, as in the study of the movement of tectonic plates, it was recognition that the group of rotations on the sphere plays a central role and the use of a satisfactory representation of that group that led (Chang, 19??) to the production of the essential statistical tool for spherical regression.

Rationale for the recognition of the perturbation operation

As we have seen above a fundamental requirement of numerate investigation is to be able to characterise change in the selected sample space. In the consideration of the differences between compositions the obvious first questions are whether there is an operation on a composition x, analogous to translation of a vector in real space, which transforms it into X, and whether this can be used to characterise the relationship or 'difference' between two compositions. The answers are to be found in the *perturbation* operator defined by Aitchison (1986, Section 2.8) and already implicitly used in a geological application by Woronow (1990). The argument is only slightly more complicated than that for real space.

The perturbation operator can be motivated by the following observation within the positive orthant representation of compositions. For any two equivalent compositions w and W on the same ray there is a scale relationship $W = pw$ for some $p > 0$, where each component of w is scaled by the *same* factor p to obtain the corresponding component of W. For any two non-equivalent compositions w and W on different rays a similar, but differential, scaling relationship $W_1 = p_1 w_1$, ..., $W_D = p_D w_D$ reflects the change from w to W. Such a unique differential scaling can always be found by taking $p_i = W_i/w_i$ $(i = 1, ..., D)$. Translating this into terms of the compositional representations x and X within the unit simplex sample space $\mathcal{S}^D$ requires only an appropriate scaling: If we denote the perturbation operation by '$\circ$' then the perturbation $p = (p_1, ..., p_D)$ applied to the composition $x = (x_1, ..., x_D)$ produces the composition X given by

$$X = p \circ x = (p_1x_1, ..., p_Dx_D)/(p_1x_1 + \cdots + p_Dx_D). \tag{7}$$

Note that the perturbing vector p can, without loss of generality, be chosen to be of compositional form, with $p_1 + \cdots + p_D = 1$.

In mathematical terms the set of perturbations in $\mathcal{S}^D$ form a group with the identity perturbation $e = (1/D, ..., 1/D)$ and the inverse of a perturbation p being the closure $\mathcal{C}(p_1^{-1}, ..., p_D^{-1})$. The relation between any two compositions x and X can always be expressed as a perturbation operation $X = (X \circ x^{-1}) \circ x$, where $X \circ x^{-1}$ is a perturbation in the group of perturbations in the unit simplex $\mathcal{S}^D$. The change from X to x is simply the inverse perturbation defined by $(X \circ x^{-1})^{-1} = x^{-1} \circ X$. Thus any measure of difference between compositions x and X must be expressible in terms of one or other of these perturbations. A consequence of this is that if we wish to define any *scalar measure of distance* between two compositions x and X, say $\Delta(x, X)$, then we must ensure that it is a function of the ratios $x_1/X_1, ..., x_D/X_D$. This together with attention to the need for scale invariance, subcompositional coherence and some other simple requirements has led Aitchison (1992) to advocate the use of

$$\Delta(x, X) = \sum_{i<j} \left\{ \log\left(\frac{x_i}{x_j}\right) - \log\left(\frac{X_i}{X_j}\right) \right\}^2, \tag{8}$$

reinforcing an intuitive equivalent choice in Aitchison (1986, Section 8.3).

Some familiar perturbations

Perturbations are not some esoteric mathematical entity. They are already in use in other branches of statistics. If you have ever engaged in Bayesian inference you have perturbed the prior probability assessment x on a finite number D of hypotheses by the likelihood p to obtain the posterior assessment X through the use of Bayes's formula (7). Again, in genetic selection, the population composition x of genotypes of one generation is perturbed by differential survival probabilities represented by a perturbation p to obtain the composition X at the next generation, again by the perturbation probabilistic mechanism (7). May it not be that certain geological processes, such as metamorphic change, sedimentation, crushing in relation to particle size distributions, may be best modelled by such perturbation mechanisms, where an initial specimen of composition x_0 is subjected to a sequence of perturbations $p_1, ..., p_n$ in reaching its current state x_n:

$$x_1 = p_1 \circ x_0, \quad x_2 = p_2 \circ x_1, \quad ..., \quad x_n = p_n \circ x_{n-1},$$

so that

$$x_n = (p_1 \circ p_2 \circ \cdots \circ p_n) \circ x_0. \tag{9}$$

We now look at the logical consequences of such a process assumption.

The central limit theorem for compositions

It is well known that sequences of additive and multiplicative changes lead to normal and lognormal variability through the magnificent central limit theorems and (9) is crying out for such an interpretation. We can very simply relate (9) to an additive central limit theorem by rewriting it in terms of logratios:

$$\log\left(\frac{x_{ni}}{x_{nD}}\right) = \left\{\log\left(\frac{p_{1i}}{p_{1D}}\right) + \cdots + \log\left(\frac{p_{ni}}{p_{nD}}\right)\right\} + \log\left(\frac{x_{0i}}{x_{0D}}\right)$$

$$(i = 1, ..., D-1).$$

If the perturbations are random then sums within the brackets will, under certain regularity conditions which need not divert us here, tend for large n towards a multivariate normal pattern of variability. It is a simple application of distribution theory to deduce the form of the probability density function $f(x)$ on the unit simplex as

$$f(x) = \det(2\pi\Sigma)^{-1/2}(x_1 \cdots x_D)^{-1} \times \exp\left[-\frac{1}{2}\left\{\log\left(\frac{x_{-D}}{x_D}\right) - \mu\right\}\Sigma^{-1}\left\{\log\left(\frac{x_{-D}}{x_D}\right) - \mu\right\}^T\right], \qquad (10)$$

$(x \in \mathcal{S}^D)$, where $x_{-D} = (x_1, ..., x_{D-1})$, μ is a $D-1$ row vector and Σ is a positive definite square matrix of order $D-1$. This is the parametric class of additive logistic normal distributions $\mathcal{L}^{D-1}(\mu, \Sigma)$ described by Aitchison and Shen (1980).

CHARACTERISTICS OF PATTERNS OF COMPOSITIONAL VARIABILITY

Measure of central tendency

In describing variability of vectors there are two related aspects. How can we describe characteristics which in meaningful ways define (1) a centre around which the variability takes place and (2) measures of dispersion around this centre. Within (2) we include measures of the dependence between the various components of the composition. It is worth recalling the arguments which determine sensible centres and dispersions in R^D. In such a sample space, in which ideas of Euclidean distance dominate, it is claimed to be sensible to consider as centre μ which minimises the average squared distance $\mathrm{E}\{||y-\mu||^2\}$ and this turns out to be simply $\mathrm{E}(y)$. For compositions and the simplex we have no long-established distance measure such as Euclidean distance (though see Aitchison (1992)) but we can discover a sensible definition of centre by a simple optimising argument.

Suppose that we have a probability distribution of compositions in the simplex $\mathcal{S}^D$. Denote by x a generic composition. What should we use as centre, or measure of central tendency, say $\xi = \mathrm{cen}(x)$, of this distribution. A well-established and commonly used information-theoretic measure of the divergence of x from ξ is the Kullback-Leibler (1951) directed divergence $\mathrm{E}\{\sum_i \xi_i \log(\xi_i/x_i)\}$. It seems reasonable therefore to investigate the consequences of choosing ξ to minimise this measure, subject to the condition that ξ is a composition in $\mathcal{S}^D$ so that $\xi_1 + \cdots + \xi_D = 1$. This simple mathematical exercise yields the result that $\xi_i \propto \exp(\mathrm{E}\{\log x_i\})$. Choosing the factor of proportionality to ensure that ξ is a composition leads to the definition of centre as

$$\xi = \mathrm{cen}(x) = \mathcal{C}\left[\exp\left(\mathrm{E}\{\log x\}\right)\right]. \qquad (11)$$

At first sight this seems a very unfamiliar object until we realise that for any positive random variable z the formal definition of the *geometric* mean is $\exp(\mathrm{E}\{\log z\})$. Note here that although it seems that we have abandoned in the use of $\log(x)$

our scale-invariant directive to use only ratios the complete expression for cen(x) involves a closure operation $\mathcal{C}$ which ensures ratios. Indeed an alternative and equivalent definition of centre could be used involving ratios at the first stage of the computation, namely

$$\text{cen}(x) = \mathcal{C}\left[\exp\left(\mathrm{E}\left\{\log\left(\frac{x}{g(x)}\right)\right\}\right)\right], \tag{12}$$

where $g(x)$ denotes the geometric mean of the components of x. Any geologist who has used the lognormal distribution $\Lambda(\mu, \Sigma)$ to describe the pattern of variability of a positive quantity will have used the geometric mean $\exp(\mathrm{E}\{\log z\}) = \exp(\mu)$ in their analysis, and it is worth noting that advocacy of the use of geometric means (McAlister, 1879) even precedes Pearson's 1897 warning. We shall refer to ξ as the *geometric centre* and note that for any fixed perturbation p, cen$(p \circ x) = p \circ$cen(x), in analogy with $\mathrm{E}(t + y) = \mathrm{E}(y) + t$ in (6) for unconstrained variability in R^D.

It is worth digressing here to demonstrate the practical implications of this simple result. For a compositional data set

$$\begin{array}{cccc} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{ND} \end{array} \tag{13}$$

standard practice seems to be to take the arithmetic center $\bar{x} = (x_{\cdot 1}, ..., x_{\cdot D})$, where $x_{\cdot i} = \sum_r x_{ri}/N$. What we are advocating is the use of

$$\mathcal{C}(g_1, ..., g_D) = (g_1, ..., g_D)/(g_1 + \cdots + g_D) \tag{14}$$

as centre of the compositional data set, where $g_i = (\prod_i x_{ri})^{1/N}$ is the geometric mean of the ith component over all N cases. And there can be a substantial difference as is illustrated by the three different but not untypical 3-part compositional data sets of Fig. 2, where G and A denote the geometric and arithmetic centres. Note particularly Fig. 2c where the arithmetic centre is clearly an atypical composition.

Measures of dispersion and dependence

There are a number of criteria which dictate the choice of any measure $\mathrm{V}(x)$ of dispersion and dependence which forms the basis of characteristics of compositional variability in terms of second order moments.

(a) Is the measure interpretable in relation to the specific hypotheses and problems of interest in fields of application?

(b) Is the measure conformable with the definition of centre associated with the sample space and basic algebraic operation?

(c) Is the measure invariant under the group of basic operations, in our case the group of perturbations? Is $\mathrm{V}(p \circ x) = \mathrm{V}(x)$ for every constant perturbation p? (Recall the result in (6) that for $y \in R^D$ the covariance matrix V is invariant under translation: $\mathrm{V}(t + y) = \mathrm{V}(y)$.)

(d) Is the measure tractable mathematically?

To ensure a positive answer to (a) we must clearly work in terms of ratios of the components of compositions to ensure scale invariance. At first thought this might suggest the use of variances and covariances of the form var(x_i/x_j) and cov$(x_i/x_j, x_k/x_l)$. Unfortunately these are mathematically intractable because, for example, there is no exact or even simple approximate relationship between

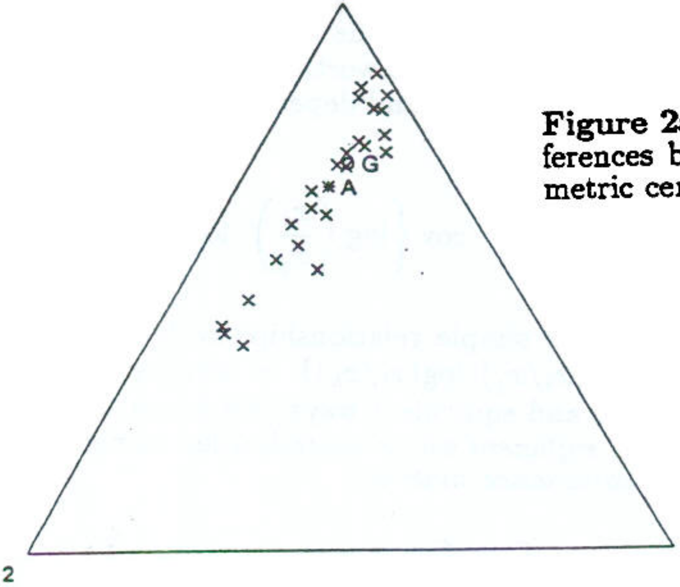
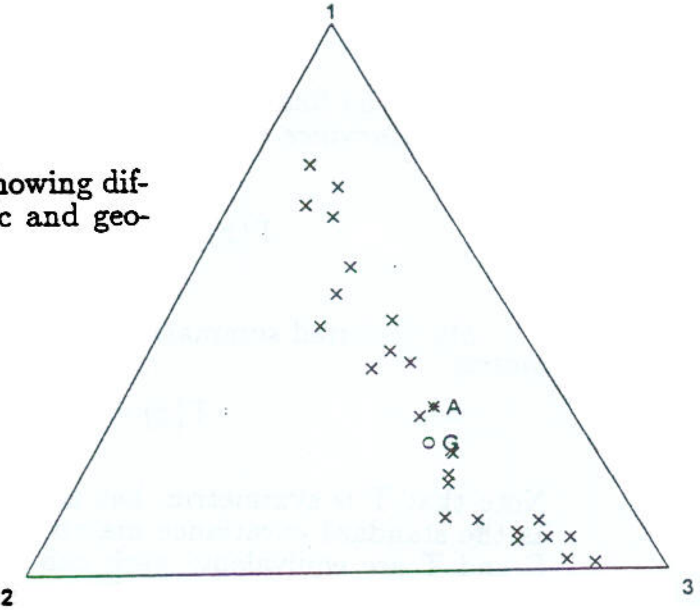


Figure 2a. Ternary diagram showing differences between the arithmetic and geometric centres.

Figure 2b. Ternary diagram showing differences between the arithmetic and geometric centres.



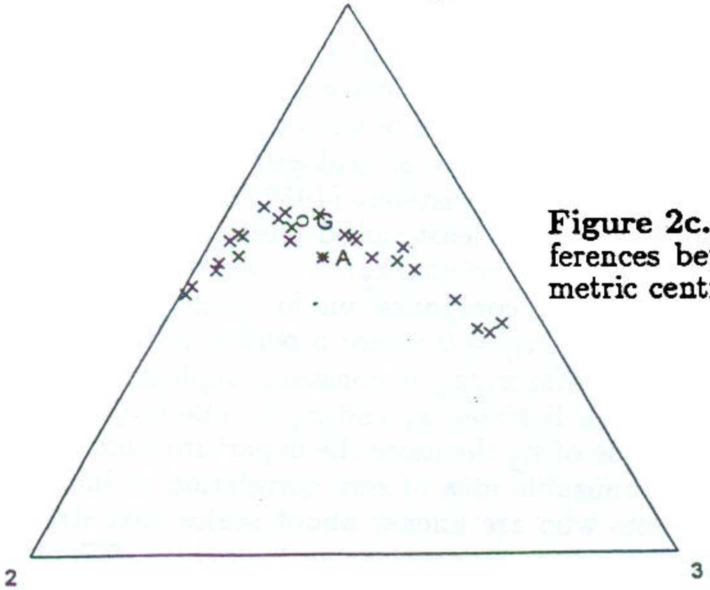


Figure 2c. Ternary diagram showing differences between the arithmetic and geometric centres.

$\mathrm{var}(x_i/x_j)$ and $\mathrm{var}(x_j/x_i)$. Fortunately we already have a clue as to how to overcome this difficulty in the appearance of logarithms of ratios of components both in the central limit theorem at (10) and in the definition of the centre of compositional variability at (11) or (12). It seems worth the risk therefore of apparently complicating the definition of dispersion and dependence by considering such dispersion characteristics as

$$\mathrm{var}\left\{\log\left(\frac{x_i}{x_j}\right)\right\}, \qquad \mathrm{cov}\left\{\log\left(\frac{x_i}{x_j}\right), \log\left(\frac{x_k}{x_l}\right)\right\}. \tag{15}$$

Obvious advantages of this are simple relationships such as $\mathrm{var}\{\log(x_i/x_j)\} = \mathrm{var}\{\log(x_j/x_i)\}$ and $\mathrm{cov}\{\log(x_i/x_j), \log(x_l/x_k)\} = \mathrm{cov}\{\log(x_j/x_i), \log(x_k/x_l)\}$. There are a number of useful and equivalent ways (Aitchison, 1986, Chapter 4) in which to summarise such a *sufficient* set of second-order moment characteristics. For example, the *logratio covariance matrix*

$$\Sigma(x) = [\sigma_{ij}] = \left[\mathrm{cov}\left\{\log\left(\frac{x_i}{x_D}\right), \log\left(\frac{x_j}{x_D}\right)\right\}\right] \tag{16}$$

using only the final component x_D as the common ratio divisor, or the *centred logratio covariance matrix*

$$\Gamma(x) = \left[\mathrm{cov}\left\{\log\left(\frac{x_i}{g(x)}\right), \log\left(\frac{x_j}{g(x)}\right)\right\}\right] \tag{17}$$

My preferred summarising characteristic is what I have termed the *variation matrix*

$$\mathrm{T}(x) = [\tau_{ij}] = \left[\mathrm{var}\left\{\log\left(\frac{x_i}{x_j}\right)\right\}\right] \tag{18}$$

Note that T is symmetric, has zero diagonal elements, and cannot be expressed as the standard covariance matrix of some vector. It is a fact, however, that Σ, Γ and T are equivalent: each can be derived from any other by simple matrix operations (Aitchison, 1986, Chapter 4). A first reaction to this variation matrix characterisation is surprise because it is defined in terms of variances only. The simplest statistical analogue is in the use of a completely randomised block design in, say, an industrial experiment. From such a situation information about $\mathrm{var}(y_i - y_j)$ for all i, j is a sufficient description of the variability for purposes of inference.

Hopefully by now early warners of geologists such as Chayes (1960, 1962), Krumbein (1962), Sarmanov and Vistelius (1959) have reinforced Karl Pearson's century-old warning and have at least raised uneasiness about interpretations of product-moment correlations $\mathrm{cov}(x_i, x_j)$. Relative variances such as $\mathrm{var}\{\log(x_i/x_j)\}$ provide some compensation for such deprivation of correlation interpretations. For example, $\tau_{ij} = 0$ means a perfect relationship between x_i and x_j in the sense that the ratio x_i/x_j is constant, replacing the unusable idea of perfect positive correlation between x_i and x_j by one of perfect proportionality. Again, the larger the value of τ_{ij} the more the departure from proportionality with $\tau_{ij} = \infty$ replacing the unusable idea of zero correlation or independence between x_i and x_j. For scientists who are uneasy about scales that stretch to infinity we can easily provide a finite scale by considering $1 - \exp(-\sqrt{\tau_{ij}})$ as a measure of relationship between components x_i and x_j. The scale is now from 0 (corresponding to lack of proportional relationship) and 1 (corresponding to perfect proportional relationship). Note that if we are really interested in hypotheses of independence these are most appropriately expressed in terms of independence of subcompositions. For example independence of the $(1,2,3)$- and $(4,5)$-subcompositions would be reflected in the following statements:

$$\operatorname{cov}\left\{\log\left(\frac{x_1}{x_3}\right), \log\left(\frac{x_4}{x_5}\right)\right\}=0, \quad \operatorname{cov}\left\{\log\left(\frac{x_2}{x_3}\right), \log\left(\frac{x_4}{x_5}\right)\right\}=0. \tag{19}$$

Finally we can provide an analogue of the rough-and-ready normal 95 percent range of mean plus and minus two standard deviations. This is expressed in terms of ratios x_i/x_j and a signed version of a *coefficient of variation*:

$$\mathrm{cv}=\frac{\sqrt{\operatorname{var}\{\log(x_i/x_j)\}}}{\mathrm{E}\{\log(x_i/x_j)\}}, \tag{20}$$

giving

$$\left(\frac{g_i}{g_j}\right)^{1-2\mathrm{cv}} \leq \frac{x_i}{x_j} \leq \left(\frac{g_i}{g_j}\right)^{1+2\mathrm{cv}}, \tag{21}$$

where g_i, g_j are the geometric means of the ith and jth components.

So far we have emphasised criteria (a), (b) and (d). Fortunately criterion (c) is automatically satisfied since, for example, $\mathrm{T}(p \circ x)=\mathrm{T}(x)$ for any constant perturbation p. Fig. 3 shows how the 3-part compositional data sets of Fig. 2 are effected by constant perturbations. We should also note here that the dimensionality of the covariance parameter T is $\frac{1}{2}D(D-1)$ and so is as parsimonious as corresponding definitions in other essentially $(D-1)$-dimensional spaces.

In the study of unconstrained variability in R^D it is often convenient to have available a measure of total variability, for example in principal component analysis and in biplots. For such a sample space the trace of the covariance matrix is the appropriate measure. Here we might consider trace(Γ), the trace of the symmetric centred logratio covariance matrix Γ. Equally we might argue on common sense grounds that the sum of all the possible relative variances in T, namely $\sum_{i<j} \operatorname{var}\{\log(x_i/x_j)\}$, would be equally good. These two measures indeed differ only by a constant factor and so we can define totvar(x), a measure of total variability, as

$$\operatorname{totvar}(x)=\operatorname{trace}(\Gamma)=\frac{1}{D}\sum_{i<j}\operatorname{var}\left\{\log\left(\frac{x_i}{x_j}\right)\right\}. \tag{22}$$

We may also note here that the scalar measure of distance (8) is compatible with the above definitions of covariance analogous to the compatibility of Euclidean distance with the covariance matrix of an unconstrained vector. As an illustration of this consider how we might construct a measure of the total variability for the compositional data set (13). The definition at (22) suggests that we may obtain such a total measure, totvar1 say, by replacing each $\operatorname{var}\{\log(x_i/x_j)\}$ in (22) by its standard estimate. An alternative intuitive measure of total variation is surely the sum of all the possible distances between the N compositions, namely totvar2 $= \sum_{r<s} \Delta(x_r, x_s)$, where here x_r, x_s denote the rth and sth compositions in (13). The easily established proportional relationship totvar1 $= \{D/\{N(N-1)\}\}$totvar2 confirms the compatibility of the defined covariance structures and scalar measures of distance for compositional variability.

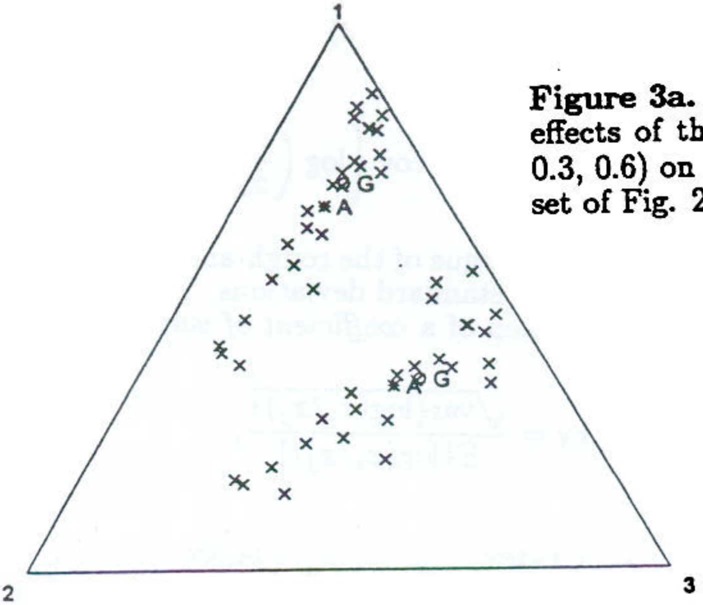


Figure 3a. Ternary diagram showing the effects of the constant perturbation (0.1, 0.3, 0.6) on the 3-part compositional data set of Fig. 2a.

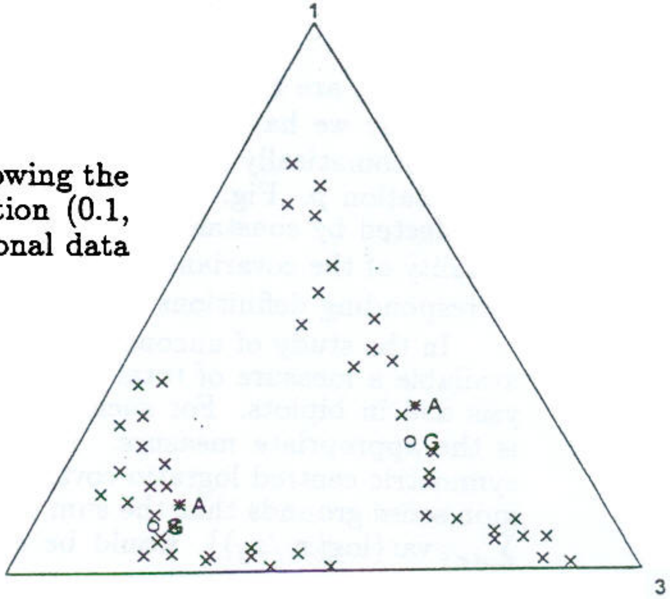


Figure 3b. Ternary diagram showing the effects of the constant perturbation (0.1, 0.8, 0.1) on the 3-part compositional data set of Fig. 2b.

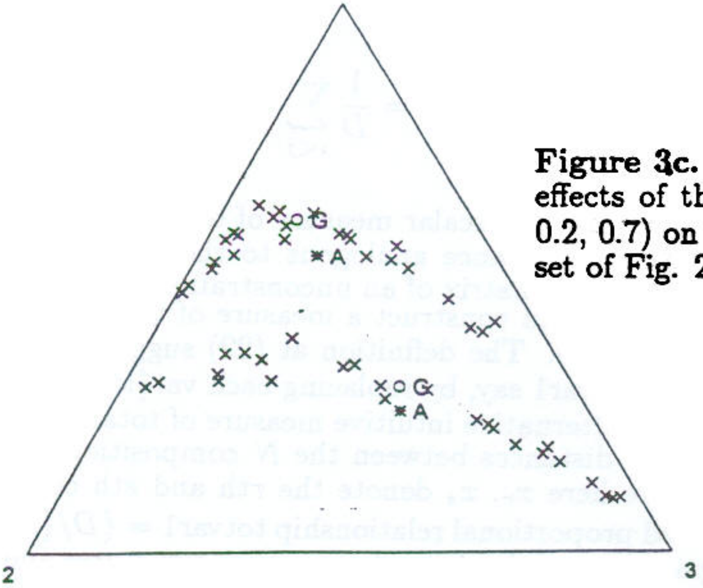


Figure 3c. Ternary diagram showing the effects of the constant perturbation (0.1, 0.2, 0.7) on the 3-part compositional data set of Fig. 2c.

FROM THEORY TO PRACTICE

The logical necessities of scale invariance, subcompositional coherence and the recognition of perturbation as the fundamental operation in the simplex have led us to the adoption of certain logratio forms of summarising characteristics for patterns of compositional variability. Not surprisingly these are compatible with the logistic-normal class of distributions on the simplex which emerge from a study of the genesis of compositional data through sequences of random perturbations. Let us see then what practical tools we now have at our disposal for the statistical analysis of compositional data. Earlier we said that any problem concerning compositions can be expressed in terms of ratios and now we extend that claim to logratios. Any convenient sufficient set of logratios can be used. For example the set of final divisor logratios

$$y_i = \log\left(\frac{x_i}{x_D}\right) \quad (i = 1, ..., D-1) \tag{23}$$

can be used with inverse transformation

$$\begin{aligned} x_i &= \frac{\exp(y_i)}{\exp(y_1) + \cdots + \exp(y_{D-1}) + 1} \quad (i = 1, ..., D-1), \\ x_D &= \frac{1}{\exp(y_1) + \cdots + \exp(y_{D-1}) + 1}. \end{aligned} \tag{24}$$

Note that transformations (23) and (24), which are asymmetric in the components of the composition, form a mapping between the unit simplex $\mathcal{S}^D$ and (D-1)-dimensional real space R^{D-1}. If we insist on a symmetric set of logratios then we may take

$$z_i = \log\left(\frac{x_i}{g(x)}\right) \quad (i = 1, ..., D) \tag{25}$$

with inverse

$$x_i = \frac{\exp(z_i)}{\exp(z_1) + \cdots + \exp(z_D)} \quad (i = 1, ..., D). \tag{26}$$

This is a transformation between the unit simplex $\mathcal{S}^D$ and the unit hyperplane $y_1 + \cdots + y_D = 0$ in D-dimensional real space R^D. The new constraint on the transformed composition is not a transfer of the so-called constant-sum constraint but a penalty for the insistence on a symmetric treatment of the components of the composition. It is linked to the use of the singular centred logratio covariance matrix Γ at (17). In practice this singularity causes no interpretational or computational problem in these days when pseudo-inverses of matrices abound in statistical analysis and software.

Compositional data analysis is then easy. The simplest recipe for success consists of four steps.

(1) Reformulate your compositional problem in terms of logratios of the components.

(2) Transform your compositional data set into compatible logratios.

(3) Since you are now in real space and free of the constant-sum constraint, simply apply the appropriate multivariate methodology associated with unconstrained vectors.

(4) Reinterpret the inference from the statistical analysis of the logratios into terms of the compositions.

Aitchison (1986) has already set out the wide variety of compositional problems which can be studied through the above logratio transformation techniques. And clearly any problem of compositional data analysis can be studied through this methodology. We shall examine some of these important problems and also some new ones after we have studied various forms of resistance to, and confusion of, this form of statistical methodology for compositional data analysis.

POCKETS OF RESISTANCE AND CONFUSION

There are a number of well-defined categories of response to the problems of compositional data analysis. I hope readers do not recognise their position in any of the categories.

The wishful thinkers

No problem exists (Gower, 1987) or, at worst, it is some esoteric mathematical statistical curiosity which has not worried our predecessors and so should not worry us. Let us continue to calculate and interpret correlations of raw components. After all if we omit one of the parts the constant-sum constraint no longer applies. Someday, somehow, what we are doing will be shown by someone to have been correct all the time.

The describers

As long as we are just describing a compositional data set we can use any characteristics. In describing compositional data we can use arithmetic means, covariance matrices of raw components and indeed any linear methods such as principal components of the raw components. After all we are simply describing the data set in summary form, not analysing it (Le Maitre, 1982).

The openers

The fact that most compositions are recorded by first arriving experimentally at an 'open vector' of quantities of the D parts constituting some whole and then forming a 'closed vector', the composition, seems to have led to a particular form of wishful thinking. All will be resolved if we can reopen the closed vector in some ideal way and then perform some statistical analysis on the open vectors to reveal the inner secrets of the compositions. The notion that there is some magic powder which can be sprinkled on closed data to make them open and unconstrained dies hard. Most recently Whitten (1995) takes as closed vectors major-oxide compositions of rocks expressed as percentages by weight, scales by whole rock specific gravities to obtain 'open vectors' recorded in g/100cc. His argument depends on attempts to establish that whole rock specific gravity is independent of the composition of the rock (to someone with virtually no knowledge of geology a seemingly naive concept) by a series of regression studies in which whole rock specific gravities are regressed against at most two of the constituent major oxides. Percentages of explanation of over 50 per cent are cavalierly regarded as indications of independence. And why we may ask was not a regression on the complete set of major oxides considered. These would certainly have led to even higher percentages of explanation. Apart from this statistical criticism the consequent open vectors are peculiarly placed geometrically, being only minor displacements from a different constraining hyperplane. If only such openers would realise that in any opened composition the ratios of components are the same as in the closed composition so that any *scale invariant* procedure applied to the opened composition will be identical to that procedure applied to the closed composition. Opening compositions is indeed superfluous folly.

The null correlationists

Pearson was the originator of this school. The idea developed from a study of the composition (shape) of Plymouth shrimps; see Aitchison (1986, Chapter 3) for an account of his ingenious early bootstrap experiment. Others, in particular

Chayes and Kruskal (1966) and Darroch and Ratcliff (1970, 1978) have attempted this approach. The basic idea here is related to the openers' ideas. Because of the 'negative bias' in correlations of raw components of compositions, zero correlation obviously does not have its usual meaning in relation to independence. There must be some non-zero value of such a correlation, called the null correlation, which corresponds to 'independence'. Usually the null correlation is surmised by some opening out procedure, as for example the oft-quoted Chayes-Kruskall method. The concept of null correlation is spurious and indeed unnecessary. All meaningful concepts of compositional dependence and independence can be studied within the simplex and in relation to the logratio covariance structures already specified.

The pathologists

A study of the compositional literature suggests that much of compositional data analysis in the period 1965-85 was directed at trying to find some inspiration from calculation of crude correlations and other linear methods. Those who were aware that things go wrong with crude correlations attempted to describe the nature of the disease instead of trying to find a cure. Thus we have many papers with titles such as 'An effect of closure on the structure of principal component' (Chayes and Trochimczyk, 1978) and 'The effect of closure on the measure of similarity between samples' (Butler, 1979).

The non-transformists

Despite his warning about the spuriousness of correlations of crude proportions, Pearson would have been unhappy about the solution through logratio transformations. He had bitter arguments (Pearson, 1905, 1906) with some of the rediscoverers (for example, Kapteyn, 1903) of the lognormal distribution. This lay in his distrust of transformations: what can possibly be the meaning of the logarithm of weight? I had hoped that we were now sufficiently convinced, particularly in geology, that the lognormal distribution has a central role to play in many geological applications. But the mention of a logratio of components still brings forth that same resistance. What is the meaning of such a logratio is a question posed by Fisher in the discussion of Aitchison (1982) and even more recently by Whitten (1995). We hope that the analogy with the lognormal distribution and the comments earlier that every piece of compositional statistical analysis can be carried out within the simplex may mean that this resistance will soon collapse.

The sphericists

There have been various attempts to escape from the unit simplex to what are thought to be simpler or more familiar sample spaces. One popular idea (Atkinson and Stephens in the discussion of Aitchison (1982), and Stephens(1982)) is to move from the unit simplex $\mathcal{S}^D$ to the positive orthant of the unit hypersphere by the transformation $z_i = \sqrt{u_i}$ $(i = 1, ..., D)$ and then to use established theory of distributions on the hypersphere. There are two insuperable difficulties about such a transformation. First, the transformation is only onto part of the hypersphere and so established distributional theory, associated as it is with the whole hypersphere, does not apply. There is clearly no way round this since the simplex and hypersphere are topologically different: there is no way of transforming a triangle to the surface of a two-dimensional sphere. As serious a difficulty is the impossibility of representing the fundamental operation of perturbation on the simplex as something tractable on the hypersphere. This is not surprising since the fundamental algebraic operation on the hypersphere is rotation and this bears no relationship to the structure of perturbation. The additional step of Stanley (1990) in transforming z to spherical polar co-ordinates further complicates such issues. Although the angles involved are scale invariant functions of the composition their relationship to the composition is bewilderingly complicated. Moreover there would be no subcompositional coherence since in terms of our previous discussion scientist B would be transforming onto a hypersphere of lower dimension with impossibly complicated relationships between the angles used by scientist A and B.

The Dirichlet extenders

Many statisticians are attempting to extend the Dirichlet class of distributions on the simplex in the hope that greater generality will bring greater realism than the simple Dirichlet class. Unfortunately I think they are likely to fail, since even the simple Dirichlet class with all its elegant mathematical properties does not have any exact perturbation properties. A further point on this will be made later in the Discussion.

Conclusion

The only sensible conclusion, it seems to me, is to reiterate my advice to my students. Recognise your sample space for what it is. Pay attention to its properties and follow through any logical necessities arising from these properties. The solution here to the apparent awkwardness of the sample space is not so difficult. The difficulty is facing up to reality and not imagining that there is some esoteric panacea.

BIPLOTS OF COMPOSITIONAL DATA

The biplot (Gabriel, 1971, 1981) is a well established graphical aid in other branches of statistical analysis. Its adaptation for compositional data (Aitchison, 1990b, 1998) can also prove a useful exploratory and expository tool. Its great strength is that it provides an approximate picture of the complete compositional variability, not only of the dependence structure but also the relationship of individual cases to the compositional parts. The biplot is based on a fundamental result of matrix theory, the singular value decomposition, but we can avoid the mathematical and computational technicalities because software can do the work for us and we can concentrate on the simple and important aspects of interpretation.

For our practical purposes here of exploring a compositional data set such as (13) consisting of N cases of D-part compositions a biplot such as Fig. 4 consists of an *origin* O which represents the centre of the compositional data set, a *vertex* for each of the parts, labelled $1, ..., D$ and a *marker* for each of the N cases, labelled $c_1, ..., c_N$. We term the join of O to a vertex i a *ray* Oi and the join of two vertices i and j the *link* ij. These features constitute a biplot with the following the main properties for the interpretation of the compositional variability.

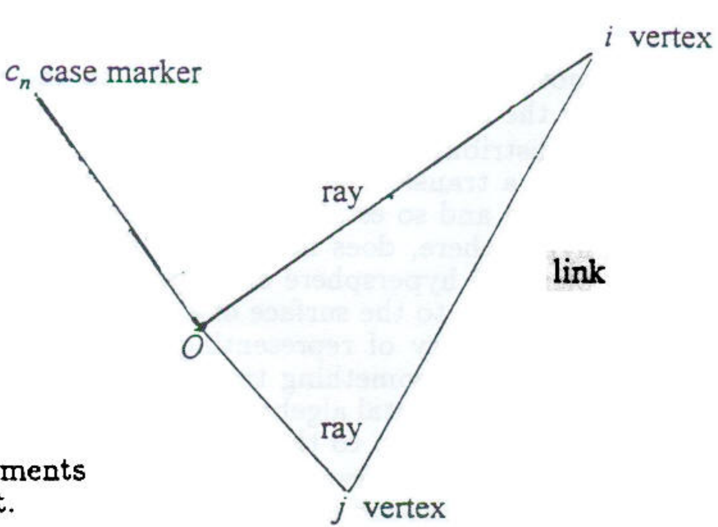


Figure 4. The basic elements of a compositional biplot.

Links, rays and covariance structure

The links and rays provide information on the covariance structure of the compositional data set.

$$|ij|^2 = \text{var}\left\{\log\left(\frac{x_i}{x_j}\right)\right\}, \tag{27}$$

$$|Oi|^2 = \text{var}\left\{\log\left(\frac{x_i}{g(x)}\right)\right\}, \tag{28}$$

$$\cos(iOj) = \text{corr}\left\{\log\left(\frac{x_i}{g(x)}\right), \log\left(\frac{x_j}{g(x)}\right)\right\}. \tag{29}$$

It is tempting to imagine that (29) can be used to replace discredited $\text{corr}(x_i, x_j)$ as a measure of the dependence between two components. Unfortunately this measure does not have subcompositional coherence.

I have found a more useful result is the following. If links ij and kl intersect in M the

$$\cos(iMk) = \text{corr}\left\{\log\left(\frac{x_i}{x_j}\right), \log\left(\frac{x_k}{x_l}\right)\right\}. \tag{29}$$

A particular case of this is when the two links are at right angles so that $\cos(iMk) = 0$, implying that there is zero correlation (independence) of the two logratios. This is useful in investigation of subcompositions for possible independence.

Subcompositional analysis

The centre O is the *centroid* (centre of gravity) of the D vertices $1, ..., D$. Since ratios are preserved under formation of subcompositions it follows that the biplot for any subcomposition s is simply formed by selecting the vertices corresponding to the parts of the subcomposition and taking the centre O_s of the subcompositional biplot as the centroid of these vertices.

Coincident vertices

If vertices i and j coincide or nearly so this means that $\text{var}\{\log(x_i/x_j)\}$ is zero or nearly so, so that the ratio x_i/x_j is constant or nearly so.

Collinear vertices

If a subset of vertices, say $1, ..., C$ is collinear then we know from our comment on subcompositional analysis that the associated subcomposition has a biplot that is one-dimensional, and then a technical argument leads us to the conclusion that the subcomposition has one-dimensional variability. Technically this one-dimensionality is described by the constancy of $C-2$ logcontrasts of the components $x_1, ..., x_C$. Such a logcontrast is simply a linear combination of the logarithms of the components, $a_1 \log x_1 + \cdots + a_C \log x_C$ with the constraint $a_1 + \cdots + a_C = 0$ ensuring that this linear form can be expressed as a function of component ratios. Inspection of these constant logcontrasts may then give further insights into the nature of the compositional variability.

Case markers and recovery of data

Such markers have the easily established property that $Oc_n.ji$ represents the departure of $\log(x_i/x_j)$ for case c_n from the average of this logratio over all the cases. Let P and P_n in Fig. 5 denote the projections of the centre O and the compositional marker c_c on the possibly extended link ji. Then $Oc_n.ji = \pm|PP_n||ji|$, where the positive sign is taken if the directions of PP_n and ji are the same, otherwise the negative sign is taken. A simple interpretation can be obtained as follows. Consider the extended line ji as divided into positive and negative parts by the point P, the positive part being in the direction of ji from P. If P_n falls on the positive (negative) side of this line then the logratio of $\log(x_{ni}/x_{nj})$ of the nth composition exceeds (falls short of) the average value of this logratio over all cases and the further P_n is from P the greater is this exceedance (shortfall); if P_n coincides with P then the compositional logratio coincides with the average. In Fig. 5 the nth composition clearly has a logratio $\log(x_{ni}/x_{nj})$ which falls short of the overall average of this logratio.

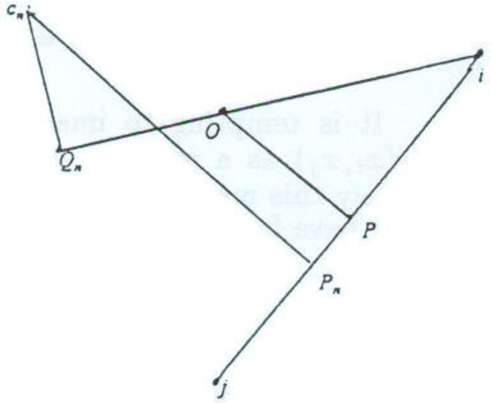


Figure 5.
Interpretation of case markers in a compositional biplot.

A similar form of interpretation can be obtained from the fact that $Oc_n.Oi$ represents the departure of the centered logratio $\log(x_{ni}/g(x_n))$ of the nth composition from the average of this centered logratio over all replicates. In Fig. 5 let Q_n be the projection of the composition marker c_n on the possibly extended ray. Then $Oc_n.Oi = \pm|OQ_n|.|Oi|$, the positive or negative sign depending on whether Q_n and the vertex i lie on the same side or opposite sides from O. We then have the following simple interpretation. If Q_n lies on the same (opposite) side of the divided line as the vertex i then the centered logratio $\log(x_{ni}/g(x_n))$ of the nth composition c_n exceeds (falls short of) the average of this logratio over all cases, and so we can infer that the ith component of the nth composition is higher (lower) than average relative to the other components. Obviously also the further Q_n is from O the greater is the divergence from the average.

Table 1 reports a compositional data set which will be new to everyone and so no preconceived ideas will dictate our analysis. It consists of 20 6-part mineral compositions of goilite rocks from a site on the edge of Loch Goil near Carrick Castle. I am told that this is an interesting site so let us see what we can discover about it.

Inspection of the variation array of Table 2 provides little insight into the nature of variability of the goilite compositions of Table 1. In contrast, the relative variation biplot of Fig. 6, retaining 98.2 per cent of the total compositional variability, allows easy identification of a number of characteristics. For simplicity in our interpretation we shall use only the initial letters to identify the mineral parts. First, we see that the de link is by far the longest indicating the greatest relative variation in the ratios of components is between d and e. Secondly, the near coincidence of the vertices a and c implies that the a and c are in almost constant proportion with the approximate relationship of $a/c = 0.55$ easily obtained from Table 1 or from the estimate -0.605 for $\mathrm{E}\{\log(a/c)\}$ in the variation array of Table 2. Note that in the ternary diagram of the abc subcomposition in Fig. 7 the representative compositional points lie roughly on a ray through the vertex b. Applying the approximate 95 percent range formula at (20) and noting that

$$[g_a\ g_b\ g_c\ g_d\ g_e\ g_f] = [0.157\ 0.207\ 0.288\ 0.102\ 0.055\ 0.162]$$

and coefficients of variation for $\log(e/f)$ and $\log(a/e)$ are -0.716 and -0.214 we

Table 1. Six-part mineral compositions of 22 specimens of goilite.

	a	b	c	d	e	f
1	0.125	0.353	0.266	0.163	0.031	0.181
2	0.143	0.224	0.313	0.111	0.051	0.159
3	0.147	0.231	0.303	0.058	0.129	0.133
4	0.164	0.209	0.282	0.120	0.047	0.178
5	0.197	0.151	0.299	0.132	0.033	0.188
6	0.157	0.256	0.246	0.072	0.116	0.153
7	0.153	0.232	0.282	0.101	0.062	0.170
8	0.115	0.249	0.259	0.176	0.025	0.176
9	0.178	0.167	0.347	0.048	0.143	0.118
10	0.164	0.183	0.281	0.158	0.027	0.186
11	0.175	0.211	0.283	0.070	0.104	0.157
12	0.168	0.192	0.305	0.120	0.044	0.171
13	0.155	0.251	0.257	0.091	0.085	0.161
14	0.126	0.273	0.269	0.045	0.156	0.131
15	0.199	0.170	0.318	0.080	0.076	0.158
16	0.163	0.196	0.335	0.107	0.054	0.144
17	0.136	0.185	0.304	0.162	0.020	0.193
18	0.184	0.152	0.350	0.110	0.039	0.165
19	0.169	0.207	0.300	0.111	0.057	0.156
20	0.146	0.240	0.250	0.141	0.038	0.184
21	0.200	0.172	0.313	0.059	0.120	0.136
22	0.135	0.225	0.217	0.217	0.019	0.187

a: arkaigite	b: broomite	c: carronite
d: dhuite	e: eckite	f: fyneite

Table 2. Variation array for goilite compositional data set.

		Column *j*					
		a	b	c	d	e	f
	a	0	0.307	0.129	0.502	0.617	0.225
	b	-0.275	0	0.270	0.465	0.646	0.221
Row *i*	c	-0.605	-0.330	0	0.486	0.628	0.213
	d	0.432	0.706	1.037	0	1.071	0.314
	e	1.047	1.322	1.652	0.615	0	0.769
	f	-0.027	0.247	0.578	-0.459	-1.074	0

Estimates below the diagonal are of $E(\log(x_j/x_i)$ and above the diagonal of $\sqrt{}\mathrm{var}\{(\log(x_i/x_j)\}$

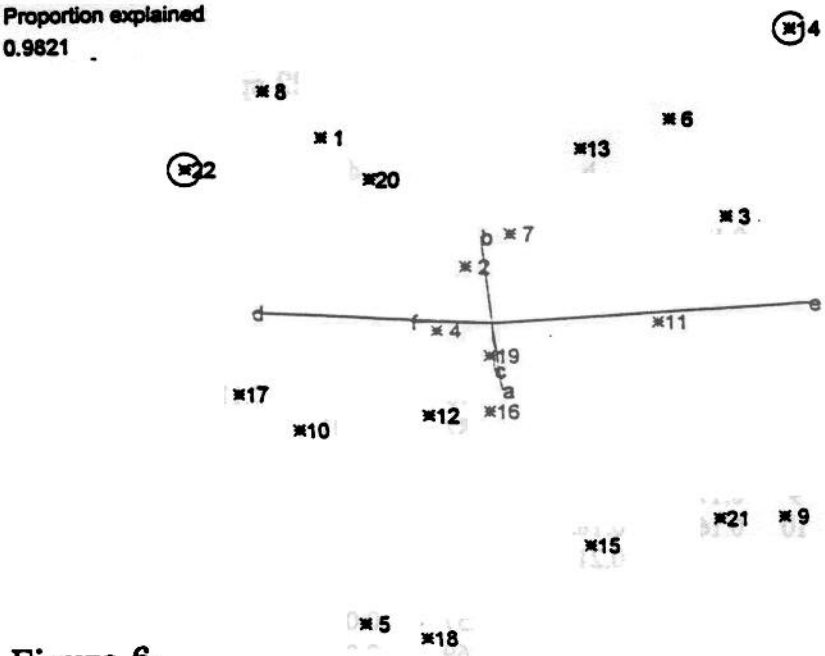
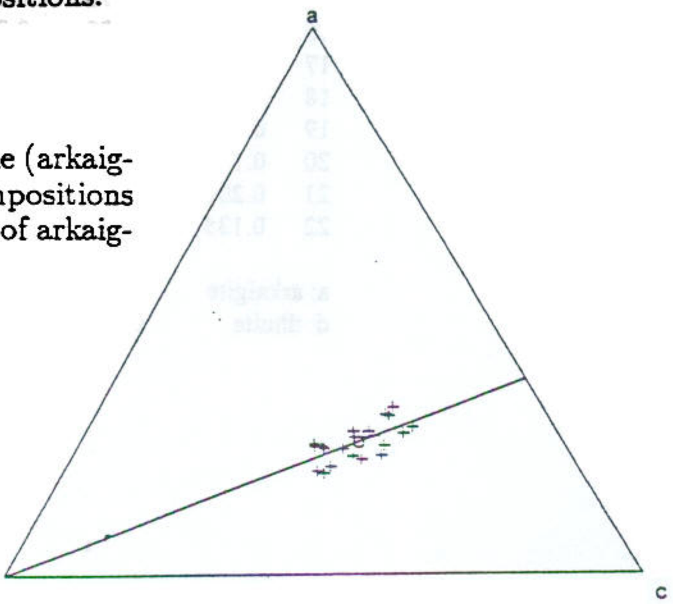


Figure 6.
Biplot for goilite mineral compositions.

Figure 7. Ternary diagram of the (arkaigite, broomite, arronite)- subcompositions showing the near proportionality of arkaigite to carronite.



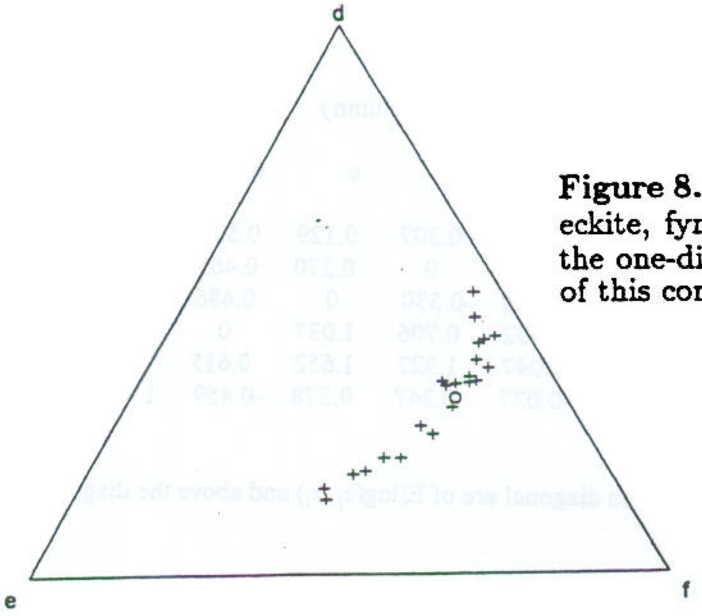


Figure 8. Ternary diagram of the (dhuite, eckite, fyneite)- subcompositions showing the one-dimensional pattern of variability of this composition.

obtain the ranges

$$0.073 < e/f < 1.59, \quad 0.42 < a/c < 0.71.$$

Thirdly and most strikingly we see the near-orthogonality of the *ab* (or *cb*) link and the links *de*, *df* and *ef*. We can immediately infer that the ratios d/e, d/f and e/f are independent of the ratio of a/b or c/b. *Another* way of expressing this feature is to state that the subcompositions (c, d, e) and (a, b) are independent. A formal test of this hypothesis of subcompositional independence (Aitchison, 1986, Section 10.3) results in a significance probability 0.27 confirming our conclusion. Fourthly, the collinearity of the three mineral links *de*, *df* and *ef* and the consequent one-dimensionality of the pattern of variability of this (d, e, f)-subcomposition, confirmed by the corresponding subcompositional ternary diagram of Fig. 8, implies some relationship between the proportions of the minerals d, e, and f. Direct investigation by logcontrast principal component analysis leads to the following eigenvalues and corresponding logcontrast principal components:

$$\begin{aligned} \lambda_1 &= 12.79, & 0.587\log d - 0.785\log e + 0.194\log f \\ \lambda_2 &= 0.0625, & -0.567\log d - 0.225\log e + 0.792\log f \end{aligned} \qquad (31)$$

2,50 log d + 1 log e - 3.5 log f

The near-constant logcontrast arises from the near-zero second eigenvalue. Moreover the fact that the coefficients are roughly in the ratios of $-2 : -1 : 3$ suggests that we can make a substantial simplification to our interpretation if we consider the constant logcontrast

2,5 log d + log e - 3,5 log f

$$-3\log d - \log e + 4\log f = \text{ constant } = 2.46,$$

5 log d + 2 log e - 7 log f

where the constant value is estimated from the sample average of the logcontrast. This can be simply converted into the approximate relationship;

$$\frac{e}{f} = 0.85\left(\frac{f}{d}\right)^3. \qquad \left(\frac{e}{f}\right)^2 = 0. \quad \left(\frac{f}{d}\right)^5 \qquad (32)$$

Whether this suggested ‘cubic hypothesis’ is worth further investigation as a geological finding is a matter for geologists not an ingeolate statistician.

As a final comment here we note that any subcomposition can be viewed as a set of logcontrasts (Aitchison, 1984) and so are included in any logcontrast principal component analysis for study of the dimensionality of the pattern of compositional variability.

SOME OTHER USEFUL TOOLS

As we have seen, the biplot can be a very useful data-exploratory tool. It should, however, be used with caution and supported by appropriate statistical analysis, as for example in the appropriate test of subcompositional independence referred to in the previous section. Let me here provide two further related analytical tools which I have found useful in all my practical work on compositional data analysis.

The predictive distribution as the fitted distribution

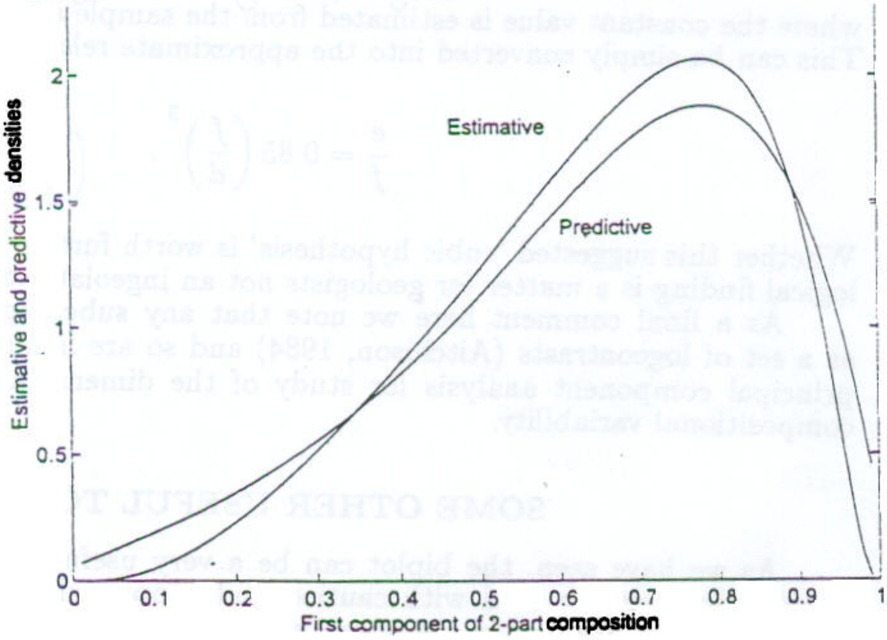
In much of statistical work we fit models to describe patterns of variability of our observed data and there has been much discussion in statistical circles as to what the appropriate distribution should be. It is clearly beyond the scope of this paper to argue any case here but let me direct your attention to the use of what have become known as the *predictive distributions*. Instead of simply inserting the maximum-likelihood estimates in the logistic-normal $\mathcal{L}^{D-1}(\mu, \Sigma)$ density function (the estimative method), as it were putting all our eggs in one basket, we average all the possible logistic-normal density functions taking account of the relative plausibilities of the various (μ, Σ) parametric combinations. The resulting predictive distribution is what can be termed a logistic-Student distribution with density function

$$f(x|data) \propto (x_1 ... x_D)^{-1}$$

$$\times \left[1 + \left\{\log\left(\frac{x_{-D}}{x_D}\right) - \mu\right\}\{(N-1)(1+N^{-1})\Sigma\}^{-1}\left\{\log\left(\frac{x_{-D}}{x_D}\right) - \mu\right\}^T\right]^{N/2} \tag{33}$$

for compositional data set (13). For large data sets there is little difference between estimative and predictive fitted distributions, but for moderate compositional data sets the difference can be substantial. The fact that geological sets often have N small (a few rock specimens) and D large (ten or more major oxides) should recommend the use of the predictive distribution in applications to compositional geology. Fig. 9 shows the difference in the estimative and predictive density functions for the 2-part compositional data set consisting of the first eight (d, e)-subcompositions of the goilite data set, arguably large in comparison with commonly analysed geological data sets. Note the sensibly conservative predictive approach with its more disperse density function.

Figure 9. Estimate and predictive density functions based on eight (dhuite, eckite)-subcompositions of the goilite data set.



Atypicality indices

The fitted density function assigns different plausibilities to different compositions. Fig. 10 shows a 3-part compositional data set in a ternary diagram with some contour lines of the fitted predictive distribution. A composition such as C near the centre is clearly more probable than one such as B in the less dense area. B is more atypical than C of the past experience of goilite. We can express this in terms of an atypicality index, which is, roughly speaking, the probability that a future composition will be more typical (be associated with a higher probability density) than the considered composition. Technically the atypicality index $A(x^*)$ of a composition x^* is given by

$$A(x) = \int_R f(x|data)dx \text{ where } R = \{x : f(x|data) > f(x^*|data)\}, \qquad (34)$$

and this is easily evaluated in terms of standard incomplete beta functions; for details see Aitchison (1986, Section 7.10). Atypicality indices lie between 0 and 1, with near-zero corresponding to a composition near the centre of the distribution and near 1 corresponding to an extremely atypical composition lying in a region of very low density. Atypicality indices are therefore useful in detecting possible outliers or anomalous compositions. For inspection of a given data set it is advisable to use the now standard jack-knife or leaving-one-out technique to avoid resubstitution bias in assessing the atypicality index of any composition in the data set. Again atypicality indices for such a procedures are readily computable.

In the goilite example above two compositions 14 and 22, circled in Fig. 5 have atypicality indices 0.999 and 0.95? greater than 0.95. From their positions in Fig. 5 and the interpretation of case markers as described above it is clear that, for composition 14 this is probably due to a combination of its maximally high ratios of f to d, e to d, and b to a; and for composition 22 its minimally low ratios of f to d, e to d, and e to f; facts easily confirmed from Table 1.

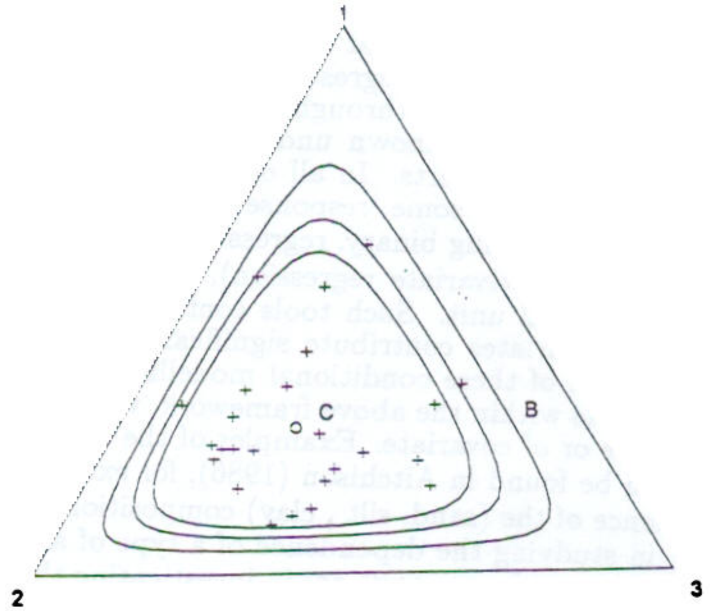


Figure 10. Ternary diagram showing a 3-part compositional data set and contour lines associated with the fitted predictive density function.

MORE COMPLEX PROBLEMS

Within the structure set up for discussion of problems within the simplex we could claim that it is possible to tackle any compositional problem. Such a claim might require substantial statistical research but that is a role that statisticians are very willing to undertake. Let me indicate with a number of practical examples just how some of the problems I have encountered in geology may be approached.

Joint variability

Geochemical compositions are commonly reported in terms of percentages by weight of major oxides and parts per million of trace elements. Seldom, if ever, is the joint variability of these two aspects fully studied by statistical techniques which take full cognisance of the constrained nature of the data. No methodology seems to have emerged for the systematic study of the pattern of joint variability of the major oxides and the trace elements of compositions within a rock type and the determination of the geochemical nature of difference between rock types. Indeed, since introducing and advocating the logratio approach to the statistical analysis of compositional data I have been repeatedly asked if and how the logratio approach can be applied to such major-oxide, trace-element compositions. The present deficiency in the methodology probably arises from a combination of two inhibiting factors: the well-known failure of standard statistical techniques, devised for unconstrained data, in the analysis of 'closed data' and the awkwardness of the different units in which the major oxides and trace elements are reported. This is surprising since there is no difficulty in accommodating compositions involving components measured in different units. If there is a conceptual perturbation which would bring all components to the same units of measurement then, because of the invariance of logratio covariance structures under the group of perturbations, we can simply treat the composition as if the units were in common units. Only the mean or central position is altered by perturbation; covariance structure including the biplot remains unchanged. Again a biplot may aid in exploring the nature of the dependencies inter and intra the major and trace aspects of the compositional variability.

Conditional variability

It could be argued that the greatest practical tool that statisticians have ever produced is that of regression analysis in its multitude of forms, from Galton's basic ideas of 1889 through analysis of variance and covariance to the more recent generalisations known under the title of general linear models and their multivariate counterparts. In all of these we are attempting to explain the pattern of variability of some 'response' of an experimental unit, such as a category (categorical, including binary, regression), quantity or vector of quantities (simple univariate or multivariate regression), depends on other factors or covariates of the experimental unit. Such tools contain the facility for deciding which of the factors or covariates contribute significantly to the pattern of variability of the response. All of these conditional modelling tools are available to compositional data analysts within the above framework, whether the composition plays the role of response or of covariate. Examples of the application of such conditional modelling can be found in Aitchison (1986), for example in studying the nature of the dependence of the (sand, silt , clay) composition of sediments in an Arctic lake on depth, in studying the dependence of a type of a rock, Permian or post-Permian, on its major-oxide composition; in investigating the dependence of the (flesh, skin, stone) composition of this year's crop of yatquats on the composition of last year's yatquats and on the nature of the treatment (hormone or placebo) of the trees.

The last example is of course non-geological. We can, however, illustrate the exploratory and interpretative power of this technique through an adaptation of the biplot to a conditional biplot which in its representation concentrates on the dependence of response to covariate through a geological example. The data set consists of 21 tektites (Chao, 1963; Miesch et al, 1966), for which the two compositions are 8-part major-oxide compositions and 8-part mineral compositions as described in Table 3. These are subcompositions of the original data set, this reduction being adopted only for the sake of simpler exposition. While experimentally these two types of compositions are determined by completely different processes they are obviously chemically related since the minerals are themselves more complicated major oxide compounds. The challenge of the conditional biplot of Fig. 11 is whether it can at least identify these relationships from the compositional data alone, without any additional information about the chemical formulae of the minerals, and hopefully provide other meaningful interpretations of the data.

Table 3. Major-oxide and mineral compositions of 21 tektites.

Major oxide compositions

Case	SiO_2	K_2O	Na_2O	CaO	MgO	Fe_2O_3	TiO	P_2O_5
1	70.83	1.86	1.20	0.52	0.46	0.030	0.65	0.005
2	80.10	1.99	1.37	0.49	0.42	0.110	0.66	0.020
3	80.17	2.24	1.53	0.56	0.37	0.180	0.60	0.030
4	78.40	1.90	1.36	0.55	0.59	0.050	0.69	0.010
5	78.37	2.43	1.84	0.78	0.70	0.050	0.59	0.020
6	77.21	2.42	1.80	0.96	0.50	0.060	0.62	0.060
7	78.19	2.23	1.71	0.65	0.73	0.230	0.74	0.040
8	76.11	2.38	1.59	0.81	0.59	0.220	0.74	0.040
9	76.68	1.81	1.27	0.59	0.56	0.005	0.83	0.010
10	76.09	2.04	1.60	0.67	0.54	0.230	0.80	0.040
11	76.25	2.22	1.63	0.74	0.74	0.270	0.74	0.050
12	76.23	2.03	1.50	0.51	0.58	0.330	0.77	0.050
13	75.59	2.42	1.72	0.79	0.66	0.200	0.73	0.050
14	75.58	2.40	1.84	0.79	0.95	0.210	0.71	0.050
15	75.38	2.21	1.77	0.79	0.95	0.320	0.78	0.060
16	75.51	2.25	1.61	0.74	0.67	0.350	0.75	0.050
17	75.13	1.84	1.42	0.54	0.61	0.170	0.90	0.050
18	74.94	1.84	1.50	0.66	0.43	0.130	0.86	0.040
19	73.36	1.93	1.44	0.61	0.75	0.310	0.87	0.030
20	72.70	1.63	1.43	0.41	0.70	0.320	0.99	0.070
21	71.89	1.60	1.28	0.045	0.78	0.270	1.05	0.040

Mineral compositions

Case	qu	or	al	an	en	ma	il	ap
1	62.02	10.99	10.15	2.58	1.15	0.040	1.23	0.010
2	61.13	11.76	11.59	2.30	1.05	0.160	1.25	0.050
3	59.17	13.25	12.94	2.58	0.92	0.260	1.14	0.070
4	58.93	11.23	11.50	2.66	1.47	0.070	1.31	0.020
5	53.79	14.36	15.56	3.74	1.74	0.070	1.12	0.050
6	52.54	14.30	15.22	4.37	1.24	0.090	1.18	0.140
7	55.20	13.17	14.46	2.96	1.82	0.330	1.41	0.090
8	52.78	14.06	13.45	3.76	1.47	0.320	1.41	0.090
9	57.90	10.69	10.74	2.86	1.39	0.010	1.58	0.020
10	54.19	12.05	13.53	3.06	1.34	0.330	1.52	0.090
11	53.22	13.12	13.79	3.34	1.84	0.390	1.41	0.120
12	55.38	11.99	12.69	2.20	1.44	0.480	1.46	0.120
13	51.24	14.30	14.55	3.59	1.64	0.290	1.39	0.120
14	50.15	14.18	15.56	3.59	2.37	0.300	1.35	0.120
15	50.97	13.06	14.97	3.53	2.37	0.460	1.48	0.140
16	52.39	13.29	13.62	3.34	1.67	0.510	1.42	0.120
17	54.92	10.87	12.01	2.35	1.52	0.250	1.71	0.120
18	54.01	10.87	12.69	3.01	1.07	0.190	1.63	0.090
19	51.99	11.40	12.18	2.83	1.87	0.450	1.65	0.070
20	52.95	9.63	12.09	1.58	1.74	0.460	1.88	0.170
21	52.79	9.45	10.83	1.97	1.94	0.390	1.99	0.090

qu: quartz	or: orthoclase	al: albite	an: anorthite
en: enstatite	ma: magnetite	il: ilmenite	ap: apatite

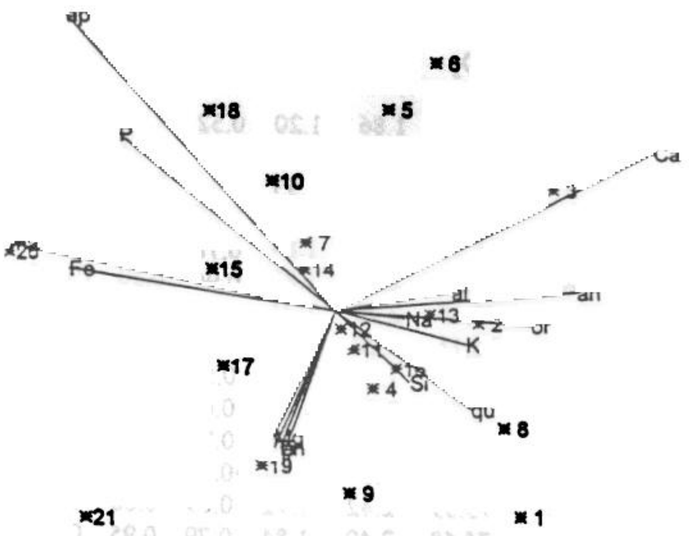


Figure 11. Conditional biplot showing the dependence of the mineral compositions on the major-oxide compositions for the tektite compositional data set.

Table 4. Oxides and associated minerals in tektite study.

Oxide	Mineral	Abbreviation	Formula
SiO_2	Quartz	qu	SiO_2
K_2O	Orthoclase	or	$KAlSi_3O_8$
Na_2O	Albite	al	$NaAlSi_3O_8$
CaO	Anorthite	an	$CaAl_2Si_2O_8$
MgO	Enstatite	en	$MgSiO_3$
Fe_2O_3	Magnetite	ma	Fe_3O_4
TiO	Ilmenite	il	$FeTiO_3$
P_2O_5	Apatite	ap	$Ca_5(F,Cl)(PO_4)_3$

First we report that the covariance aspects of the biplots constructed separately for the major-oxide and mineral compositions are almost identical with the relevant parts of the conditional relative variation biplot. A second interesting feature of the diagram is that it is indeed successful in identifying which oxides are associated with which minerals. From Table 4 we see that, apart from SiO_2, each of the other seven major oxides is associated with only one of the minerals, for example MgO is contained only in enstatite. In the biplot diagram the cosine of the angle between the rays of a major oxide and a mineral is, relatively speaking, a measure of association between oxide and mineral. A striking feature of the diagram is the way in which it demonstrates the close association of each of the seven major oxides, for example Fe_2O3, with its corresponding mineral, magnetite and P_2O_5 with it corresponding mineral apatite. Moreover, even SiO_2, which is a constituent of all eight minerals, is nevertheless primarily identified with quartz which is simply its oxide self. Note the fundamental lesson of this analysis. Here am I, a statistician with no geological knowledge, analysing a compositional data set and discovering, without any geological input simple connections between minerals and major oxides.

Miesch et al. (1966) put forward a theory of the formation of tektites which they identify as the independence of the set of three ratios (Fe_2O, MgO, P_2O_2) / SiO_2 from the set of three ratios (CaO, Na_2O, K_2O)/SiO_2. This in turn implies that the subcompositions (Fe_2O, MgO, P_2O_5), (CaO, Na_2O, K_2O) must be independent and we would then be disappointed in failing to identify the necessary approximate right angles within Fig. 10. For example, the Fe_2O-P_2O_5 and CaO-Na_2O links are approximately parallel instead of orthogonal, throwing considerable doubt on the theory. Rejection of the theory can be confirmed by a full statistical analysis testing the hypothesis of independence of the two subcompositions by the procedure described in Aitchison (1986, Section 10.3) for which the significance probability is 0.04.

A similar example appears in Aitchison (1990b), where the artefact CIPW norms (mineral compositions constructed by a complicated formula from major oxide compositions) are related to the major oxide compositions. Success here is not surprising but raises the question of whether a fitted statistical conditioning model relating minerals to major oxides as described above might be a better means of constructing mineral norms from major oxides.

Irregular data

A brief comment on irregular data problems in compositional data analysis is perhaps worth making here since it appears from recent papers in *Mathematical Geology* that an important general technique, namely the EM algorithm, has gone unnoticed by mathematical geologists in relation to irregular data, in particular in advocating so-called replacement strategies (Sanford, Pierson and Crovelli, 1993; Chang, 1993). The data set of Table 5 contains ten 5-part compositions suffering from irregularities of different types. The notes to Table 5 specify the nature of these irregularities. We can identify two basic questions associated with such an irregular compositional data set.

1. Can we test whether the irregular compositions (whether trace, missing, amalgamated, or indeed combinations of these) conform with the pattern of variability of the full compositions?
2. Given that we have conformity, can we estimate the characteristics of the pattern of variability and obtain the important fitted predictive distribution?

The answer to both these questions is *yes* and the technique is the easily applicable EM algorithm. This is simply a two-step iterative process, whereby trace, missing and amalgamated components are replaced at an E-step by estimated values based on current iterates of the characteristics; then new iterates of the characteristics are obtained at the M-step by maximisation of the likelihood based on the current completed compositional data set. These are straightforward computational procedures which take full account of the multivariate nature of the

Table 5. An illustrative irregular compositional data set.

	a	b	c	d	e
1	0.370	0.091	0.342	0.095	0.102
2	0.442	0.383	0.029	0.077	0.069
3	0.446	0.330	0.046	0.122	0.056
4	0.412	0.117	0.267	0.096	0.108
5	0.414	0.129	0.234	0.158	0.065
6	0.486	0.340	0.025	0.094	0.055
7	0.455	0.166	0.176	0.096	0.107
8	0.429	0.469	trace	0.057	0.045
9	0.453	0.465	trace	0.082	trace
10	missing	0.549	0.088	0.262	0.101
11	0.767	missing	trace	0.135	0.098
12	0.446	(b+c = 0.3530)		0.116	0.085

Notes

trace: This indicates in the preliminary determination process that led to reporting of the composition the quantity of the part fell below the minimum detection value. What is reported is the composition of the non-trace subcomposition.

missing: This indicates that the preliminary determination failes for whatever reason record the quantity of this part. What is reported is the composition of the non-missing subcomposition.

amalgamation: This indicates that the preliminary determination process could record only the combined quantity of the amalgamated parts.

Table 6. Typical river and fishing location pollutant compositions

	pollutant			
	a	*b*	*c*	*d*
River 1	0.6541	0.1553	0.1129	0.0777
	0.5420	0.3497	0.0349	0.0734
River 2	0.2450	0.2924	0.2450	0.2176
	0.2503	0.0420	0.5571	0.1506
River 3	0.3334	0.1704	0.2026	0.2936
	0.4332	0.1409	0.1352	0.2907
Location A	0.4014	0.1864	0.2619	0.1503
	0.3820	0.1169	0.3480	0.1531
Location B	0.4033	0.2300	0.2168	0.1498
	0.4706	0.2207	0.1594	0.1493
Location C	0.3140	0.1060	0.3896	0.1904
	0.2460	0.2278	0.3488	0.1774

data, unlike Sanford, Pierson and Crovelli, (1993) and Chang (1993), who adopt an unnecessarily univariate approach. The technical details need not concern us here since the computational procedures are available within a new software package NEWCODA (Aitchison, 1997b).

An endmember problem

By an endmember problem I mean one in which some *target* D-composition X is visualised as arising from some convex linear combination $\pi = (\pi_1, ..., \pi_C)$ of C *source* or endmember compositions $x_1, ..., x_C$, as

$$X = \pi_1 x_1 + \cdots + \pi_C x_C. \tag{35}$$

One such problem is where information is available only on the target in the form of a compositional data set and the objective is to attempt to find fixed endmember source compositions from which the target compositions could have arisen through varying mixtures of these source compositions. For a successful approach to this difficult problem, see Renner (1991, 1992, 1993, 1995, 1996). An interesting endmember problem of a different nature arises in some forms of pollution detection (Aitchison and Bacon-Shone, 1998). The following example illustrates how such a problem can be resolved.

Sources of pollution in a Scottish loch

A Scottish loch is supplied by three rivers, here labelled 1, 2, 3. At the mouth of each 10 water samples have been taken at random times and analysed into 4-part compositions of pollutants a, b, c, d. Also available are 20 samples, again taken at random times, at each of three fishing locations A,B,C. Space does not allow the publication of the full data set of 90 4-part compositions but Table 6, which records the first and last compositions in each of the rivers and fishing locations, gives a picture of the variability and the statistical nature of the problem. The problem here is to determine whether the compositions at a fishing location may be regarded as mixtures of compositions from the three sources, and what can be inferred about the nature of such a mixture. Although there is information about the source compositions this is not precise since each source is defined only by an observed cluster of compositions. As a first approach to this problem Aitchison and Bacon-Shone (1998) propose that the variability in each of the source compositions be summarised by using assessments of their distributions obtained from the available samples. Thus a basic distributional problem that faces us is to find the distribution of X in (35) for a given π and for given independent (logistic normal) distributions of $x_1, ..., x_C$. The existence of an excellent approximation to this distribution allows the computation of the likelihood function, from which all statistical analyses can proceed. It is then possible within this framework to test whether the mixing vector p is fixed; if so, to estimate it and, if not, to describe the nature of its variability, all with a view to making inferences about the relative responsibilities of the rivers as sources of pollution.

DISCUSSION

Addition (and subtraction) and multiplication (and division) play an intriguing role in many areas of statistics. Compositional data analysis is no exception. Scale invariance, subcompositional coherence and perturbation invariance all lead us to ratios of components and division. The central limit theorem and simplifying power of the logarithmic function which, by its basic property $\log(x_i/x_j) = \log x_i - \log x_j$, converts awkward division into simpler subtraction, lead us to consideration of logratios. Such a commitment makes consideration of the addition of components more awkward; for example, if we assume that a

D-part composition is logistic-normal there is no simple form for the distribution of $\log((x_i + x_j)/x_k)$. This is similar to the awkward problem of determining the distribution of the sum of two lognormal variables. But these are mathematical and computational difficulties and should not detract us from consideration of geological problems of an additive or linear form in the components. For example, for a major-oxide compositional data set there is no conceptual difficulty in testing a linear hypothesis such as A + 2B = 3C + 2D within the methodology described above. An exact test of such a hypothesis is not available but we can provide, as in most practical statistical work, reliable approximations. An excellent example of the practical use of such approximations is in the pollution problem just discussed.

Although for most analysts the simplest way of handling compositional data will be to immediately transform to logratios diehard non-transformists can be readily accommodated, by way of basic operations in the simplex. For example, analogous to models arising in general linear modelling, an appropriate model for the study of conditional variability of a composition x on a covariate t is the following:

$$x = \mathrm{agl}(tB) \circ p, \tag{35}$$

where agl is the multivariate additive generalised logistic function

$$\mathrm{agl}(u) = (\exp(u_1), ..., \exp(u_C), 1) \,/\, (\exp(u_1) + ... + \exp(u_C) + 1)\,, \tag{36}$$

as defined in Aitchison (1986, Section 6.15), B is a matrix of regression coefficients and p is a random perturbation playing a role similar to additive random error in simple regression: see also Aitchison and Shen (1984).

I have confined discussion of classes of distributions on the simplex to the additive logistic normal class simply because I have found it the most useful in my own compositional data analysis. Such an assumption about the pattern of variability of a compositional data set, or about the perturbation in any study of conditional variability, should be tested, and there are many such tests available (Aitchison, 1986, Section 7.3). Where tests of the validity of the distributional form fail the multivariate extension of the Box-Cox transformation may be explored (Aitchison, 1986, Section 13.2; Barceló, Pawlowsky and Grunsky, 1996) and will often yield a closer description of the variability. Such a better description, however, is achieved at the expense of interpretation, particularly in relation to aspects of independence hypotheses and subcompositional studies. There are other logistic-normal classes within the simplex suited to certain types of application (Aitchison, 1986, Sections 6.13-14). For example the multiplicative logistic-normal class is equivalent to the multivariate normality of the logratios

$$\log\left(\frac{x_1}{1-x_1}\right), \log\left(\frac{x_2}{1-x_1-x_2}\right), ..., \log\left(\frac{x_D}{1-x_1-\cdots-x_D}\right) \tag{37}$$

is obviously highly relevant to the study of Niggli or remaining space hypotheses about the nature of the formation of compositions, as in Chayes (1983), though as yet unexplored. There are other even more general parametric classes available, for example the A-distribution and its associates (Aitchison, 1985). The great advantage of this class is that it includes as a special case the Dirichlet class, the class having the ultimate in compositional independence properties, though requiring more computational effort in application. With recent advances in computational techniques, in particular the powerful MCMC procedures, this is increasingly more viable as a basis for compositional data analysis.

Much has been left untouched here. A question which will surely arise in discussion is the problem of zero components: you cannot take the logarithm of zero so how can logratio analysis cope. If the zeros can be regarded as minute trace components then a way of proceeding is by a series of finite but small replacements accompanied by a sensitivity analysis on the effects of varying replacement values.

If the zeros truly imply that some parts are absent from the composition then the scientific problem is more fundamental. Do compositions with absent parts differ in the nature of their variability, or in their effect on some response for example in a melting experiment, from that of full compositions. Such questions can be addressed within the logratio framework, as for example in an admittedly non-geological problem in Aitchison (1986, Section 11.6). I would welcome consultation on any such problems from any reader with such a zero problem. Other major areas of compositional interest are in compositional time processes; and in spatial or regionalized compositions, where the work of Pawlowsky and her colleagues has been fundamental; see for example, Pawlowsky (1986) and Pawlowsky, Olea and Davis (1995). There are limitations to a simple statistician's comprehension of the complexity of geological problems.

Finally geologists should be aware that they are by no means the only discipline faced with compositional data analysis. In nutritional studies physiologists meet whole-body compositions into water, protein, fat and other parts; in economics household budget pattern of proportions of income spent on food, housing, transport, and other commodity groups is a composition; in psychology and sociology there is increasing interest in time budgets, where for example the proportions of the day spent by an academic statistician in teaching, consultation, research, administration and so on form the composition of interest. What many or indeed all of these disciplines including geology seem reluctant to do is to formulate their compositional problems in firm numerate form. I have in particular scoured the geological literature and found hardly any precise statements of the hypotheses of interest, at least not anything a statistician would recognise as a testable hypothesis. Instead there is a tendency to be happy with vague, non-specific, qualitative talk with descriptive data-exploratory statements. What are the hard hypotheses of geology? For example, tell us precisely in terms of compositional components what a particular cogenetic hypothesis is and we may be able to pin down the appropriate testing procedure. Spell out for us in numerate terms your petrogenetic mixing hypothesis and we can attempt to approximate the distributions required to perform the necessary tests and estimation. Describe to us more clearly what you mean by a trend in a suite of compositions and we may be able to provide you with a means of detecting whether there is trend or non-trend variability. Make clear to us the true purpose of melting experiments with mixtures and we may be in a position to advise you on the design of efficient experiments and how to determine the dependence of the final product on the ingredients and conditions of the experiments. The problems here I believe are a question of patient communication and I hope that this conference may provide an excellent forum for the encouragement of such collaboration.

The proof of any pudding is in the recipe and in the skill of the cooks to implement that recipe. So it is here. I think that I have a promising recipe. I hope you have the energy to cook and give your verdict on the meal. Happy logratioing!

REFERENCES

Aitchison, J., 1982, The statistical analysis of compositional data (with discussion): J. R. Statist. Soc., v. B44, p.139-177.

Aitchison, J., 1984, Reducing the dimensionality of compositional data sets: Math. Geology, v. 16, p. 635.

Aitchison, J., 1986, The Statistical Analysis of Compositional Data: Chapman and Hall, London.

Aitchison, J., 1989, Letter to the Editor. Measures of location of compositional data sets: Math. Geology, v. 21, p.787-790.

Aitchison, J., 1990a, Comment on "Measures of Variability for Geological Data" by D. F. Watson and G. M. Philip: Math. Geol, v. 22, p. 223-226.

Aitchison, J., 1990b, Relative variation diagrams for describing patterns of variability of compositional data: Math. Geology, v. 22, p. 487-512.

Aitchison, J., 1991, Delusions of uniqueness and ineluctability: Math. Geol., v. 23, p. 275-277.

Aitchison, J., 1992, On criteria for measures of compositional differences: Math. Geology, v. 24, p. 365-380.

Aitchison, J., 1994, Principles of compositional data analysis, in Anderson, T. W., Olkin, I., and Fang, K. T., eds, Multivariate Analysis and its Applications: California Institute of Mathematical Statistics, Hayward, p. 73-81.

Aitchison, J., 1997a, Biplots for compositional data: Available from author.

Aitchison, J. 1997b, NEWCODA: a software package for compositional data analysis: available from Social Science Research Centre, University of Hong Kong, Pokfulam Road, Hong Kong.

Aitchison, J., and Bacon-Shone, J. H., 1997, Convex linear combinations of compositions: submitted to Biometrika.

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: Math. Geology, v.16, p. 637-650.

Barcelo, C., Pawlowsky, V. and Grunsky, E., 1996, Detecting outliers in compositional data sets: Math Geology,

Butler, J. C., 1979, The effect of closure on the measure of similarity between samples: Math. Geology, v. 11, p. 73-84.

Chang, C-J. F., 1993, Estimates of covariance matrix from geochemical data with observations below detection levels: Math. Geology, v. 25, p. 852-865..

Chang, T. C., 19??, Spherical regression: Annals of Statistics, v. ??, p. ???-???.

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

Chayes, F., 1962, Numerical correlation and petrographic variation: J. Geology, v. 70, p. 440-552.

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

Chayes, F. and Kruskal, W., 1966, An approximate statistical test for correlation between proportions: J. Geology, v. 74, p. 692-702.

Chayes, F. and Trochimczyk, J., 1978, The effect of closure on the structure of principlal components: Math. Geology, v. 10, p. 323-333.

Darroch, J. N. and Ratcliff, D. , 1970, Null correlations for proportions: Math. Geology, v. 2, p. 307-312,

Darroch, J. N. and Ratcliff, D. 1978, No association of proportions: Math. Geology, v.10, p. 361-368.

Gabriel, K. R. , 1971, The biplot-graphic display of matrices with application to principal component analysis: Biometrika, v. 58, p. 453-467.

Gabriel, K. R. , 1981, Biplot display of multivariate matrices for inspection of data and diagnosis: in Barnett, V., ed., Interpreting Multivariate Data: Wiley, New York:, p. 147-173.

Gower, J. C. ,1987, Introduction to ordination techniques, in Legendre, P. and Legendre, L., eds., Developments in Numerical Ecology: Springer-Verlag, Berlin, p. 35-??.

Kapteyn, J. C., 1903, Skew Frequency Curves in Biology and Statistics: Astronomical Laboratory, Groningen, Noordhoff.

Krumbein, C., 1962, Open and closed number systems: stratigraphic mapping: Bull. Amer. Assoc. Petrol. Geologists, v. 46, p. 322-337.

Kullback, S.and Leibler, R. A., 1951, On information and sufficiency: Ann. Math. Statist., v.22, p. 525-540.

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.

Pawlowsky, V., 1986, Rumliche Strukturanalyse und Schtzung ortsabhngiger Kompositionen mit Anwendungsbeispeilen aus der Geologie: unpublished dissertation, FB Geowissenschaften, Freie Universitt Berlin, 120.

Pawlowsky, V., Olea, R. A., and Davis, J. C., 1995, Estimation of regionalized compositions: a comparison of three methods: Math. Geology, v. 27, p. 105-148.
Pearson, K., 1897, Mathematical contributions to the theory of evolution: on a form of spurious correlation which may arise when indices are used in the measurements of organs: Proc. R. Soc., v. 60, p. 489-498.
Pearson, K., 1905, Das Fehlergetz und seine erallgemeinerungen durch Fechner und Pearson. A rejoinder: Biometrika, v. 4, p. 169=???..
Pearson, K., 1906, Skew frequency curves. A rejoinder to Professor Kapteyn: Biometrika, v. 5, p. 168.
Renner, R. M., 1991, An examination of the use of the logratio transformation for the testing of endmember hypotheses: Math. Geology, v. 23, p. 549-563.
Renner, R. M., 1992, Endmember graphics: Math. Geology, v. 24, p.287-303.
Renner, R. M., 1993, The resolution of a compositiional data set into mixtures of fixed source components: Appl. Statist., v. 42, p. 615-631.
Renner, R. M., 1995, The construction of extreme compositions: Math. Geology, v. 27, p. 485-497.
Renner, R. M., 1996, An algorithm for computing extreme compositions: Computers and Geosciences, v. 21, p. 15-23.
Sanford, R. F., Pierson, C. T. and Crovelli, R. A., An objective replacement method for censored geochemical data. Math.Geology, v. 25, p. 59-89.
Sarmanov, O. V.and Vistelius, A. B., 1959, On the correlation of percentage values: Dokl. Akad. Nauk. SSSR, v. 126, p. 22-25.:
Stanley, C. R., 1990, Descriptive statistics for N-dimensional closed arrays: a spherical coordinate approach: Math. Geology, v. 22, p. 933-956.
Stephens, M.A., 1982, Use of the von Mises distribution to analyse continuous proportions: Biometrika, v. 69, p. 197-203.
Watson, D. F., 1990, Reply to Comment on "Measures of Variability for Geological Data" by D. F. Watson and G. M. Philip: Math.Geology, v. 22, p. 227-231.
Watson, D. F., 1991, Reply to "Delusions of Uniqueness and Ineluctability" by J. Aitchison: Math. Geology., v. 23, p. 279.
Watson, D. F. and Philip, G. M., 1989, Measures of Variability for Geological Data: Math. Geology., v. 21, p. 233-254.
Whitten, E. H. T., 1995, Open and closed compositional data in petrology: Math. Geology, v. 27, p. 789-806.
Woronow, A., 1990, Quantifying and testing differences among means of compositional data suites: Math.Geology,