COMPOSITIONAL DATA ANALYSIS: WHERE ARE WE AND WHERE SHOULD WE BE HEADING?

John Aitchison

Department of Statistics, University of Glasgow, Glasgow G12 8QQ, Scotland
Address for correspondence: Rosemount, Carrick Castle, Argyll PA24 8AF, Scotland
Email: john.aitchison@btinternet.com

Abstract

We take stock of the present position of compositional data analysis, of what has been achieved in the last 20 years, and then make suggestions as to what may be sensible avenues of future research. We take an uncompromisingly applied mathematical view, that the challenge of solving practical problems should motivate our theoretical research; and that any new theory should be thoroughly investigated to see if it may provide answers to previously abandoned practical considerations. Indeed a main theme of this lecture will be to demonstrate this applied mathematical approach by a number of challenging examples.

1. A personal note

In the United Kingdom recently resigning government ministers have been allowed to make personal statements on their reasons for resignation, pulling no punches as to the nature of their disagreement with their leaders. Resignation and retirement, though different in nature, have many similarities. Though my main concern in yet another attempt to retire is more concerned with the opposition, those pockets of resistance and confusion that I referred to in my IAMG97 lecture, I feel this is nevertheless an occasion for taking stock of what has been achieved and, perhaps more importantly, how I feel the subject should develop to meet the challenge of the many, indeed very many, problems that remain unanswered or not fully answered.

2. Some history: the four phases of compositional history

The statistical analysis of compositional data has gone through roughly four phases. The pre-1960 phase rode on the crest of the developmental wave of standard multivariate statistical analysis, an appropriate form of analysis for the investigation of problems with sample space R^D. Despite the obvious fact that a compositional vector, with components the proportions of some whole, is subject to a constant-sum constraint and so is entirely different from the unconstrained vector of standard unconstrained multivariate statistical analysis, scientists and statisticians alike seemed almost to delight in applying all the intricacies of standard multivariate analysis, in particular correlation analysis, to compositional vectors. We know that Karl Pearson, in his definitive 1897 paper on spurious correlations, had pointed out the pitfalls of

interpretation of such activity, but it was not until around 1960 that specific condemnation of such an approach emerged.

In this second phase, the primary critic of the application of standard multivariate analysis to compositional data was the geologist Felix Chayes, whose main criticism was in the interpretation of product-moment correlation between components of a geochemical composition, with negative bias the distorting factor from the viewpoint of any sensible interpretation. For this problem of negative bias, often referred to as the closure problem, Vistelius and Sarmanov supplemented the Chayes criticism in geological applications and Mosimann drew the attention of biologists. These warnings were largely ignored and the same silly and meaningless analysis persisted in the application of a methodology inappropriate to the special format of compositional data. Unfortunately the warners, instead of working towards an appropriate methodology, adopted what can only be described as a pathological approach. What was the nature of the distortion when standard multivariate techniques were applied to compositional data. Doctorates were obtainable on such topics as the effect of closure on the structure of principal component or on the measures of similarity between samples.

The third phase was the realisation by Aitchison in the 1980's that compositions provide information about relative, not absolute, values of components, that therefore every statement about a composition can be stated in terms of ratios of components. The facts that logratios are easier to handle mathematically than ratios and that a logratio transformation provides a one-to-one mapping on to a real space led to the advocacy of a methodology based on a variety of logratio transformations. These transformations allowed the use of standard unconstrained multivariate statistics, with inferences translatable back into compositional statements.

The fourth phase arises from the realisation that the internal simplicial operation of perturbation, the external operation of powering and the simplicial metric define a metric vector space (indeed a Hilbert space), so many compositional problems can be investigated within this space. There has thus arisen a staying-in-the-simplex approach to the solution of many compositional problems.

3. A comment on statistical modelling

In the mid 1940's as a mathematical student in the University of Edinburgh I attended two courses touching on statistics. First from Sir Edmund Whittaker a chronological development of mathematics including an appealing section on probability, with a Bayesian argument the only suggested form of inference. A simple example of a marksman firing at a target with his skill, probability of a hit, having a beta prior assigned and the outcome of a binomial trial allowing an updating of an assessment of his skill, elementary BUGS in modern parlance. Second, Professor A C Aitken, in a course on Statistical Mathematics, laid out in great elegance all sorts of mathematical tools associated with current statistical thought but left me blind as far as its application to statistical inference was concerned. It was only later in Cambridge when I decided to study for the Diploma in Mathematical Statistics and read Kolmogorov's treatise on axiomatic probability that I recognised that a clearly defined and appropriate reference set (what we now refer to as a sample space) is the essential first step in statistical model building.

4. Sample space and probabilistic structure.

In probabilistic and statistical model building the role of the sample space is, I believe, not widely understood and we have seen this misunderstanding in recent controversy in the geological compositional community. The sample space is nothing more nor less than a convenient reference space in which to record unambiguously the possible outcomes of the experiment of interest. As long as there is a one-to-one correspondence between possible outcomes of the experiment and the elements or points of the sample space the basic modelling condition has been met. This can often give an amount of freedom in the choice of sample space and which of these sample spaces the modeller chooses may depend entirely on personal choice or considerations of mathematical simplicity or tractability. At the second stage of modelling a probability structure or indeed a class of probability structures is placed on the sample space. It is important to realise that these are separate steps in the model building, and as we shall see have special relevance to the problem of essential or structural zeros in compositional modelling.

5. Transformation methodology

The original, largely intuitive, approach to compositional data analysis in my 1986 monograph was by way of a logratio transformation methodology. Transformation techniques have been very popular and successful over more than a century, from the Galton-McAllister introduction of such an idea in 1879 in their logarithmic transformation for positive data, through variance-stabilising transformations for sound analysis of variance, to the general Box-Cox transformation and the implied transformations in generalised linear modelling. The logratio transformation principle was based on the fact that there is a one-to-one correspondence between compositional vectors and associated logratio vectors, so that any statement about compositions can be reformulated in terms of logratios, and vice versa. The advantage of the transformation is that it removes the problem of a constrained sample space, the unit simplex, to one of an unconstrained space, multivariate real space, opening up all available standard multivariate techniques. The original transformations were principally the additive logratio transformation

$$alr(x) = [\log(x_1 / x_{D)}) \;\; \log(x_2 / x_D), \ldots \log(x_{D-1} / x_D)]$$

and the centred logratio transformation

$$clr(x) = [\log(x_1 / g(x)) \;\; \log(x_2 / g(x)) \ldots \log(x_D / g(x))],$$

where $g(x)$ denotes the geometric mean of the components of x.

Either or both, with a little care, can be used to analyse a wide variety of compositional problems. An important aspect of such transformations from the viewpoint of interpretation is that the logarithmic function is monotonic increasing. If a logratio increases, the ratio increases.

6. TRS (Transformation resistance syndrome) and other pockets of resistance and confusion

There is a serious condition of scientists and indeed statisticians, some of them eminent, that I refer to as TRS, transformation resistance syndrome. It's been around for a long time, seems to be highly infectious and so far no really effective cure is available. That's not quite true since the effective cure involves sensible thought though not all sufferers seem willing to accept that therapy.

The logratio transformation methodology seemed to be accepted by the statistical community; see for example the discussion of Aitchison (1982). The logratio methodology, however, drew fierce opposition from other disciplines, in particular from sections of the geological community. The reader who is interested in following the arguments that have arisen should examine the letters to the Editor of *Mathematical Geology* over the period 1988 through 2002; in particular, see Watson and Philip (1989), Aitchison (1989, 1990a), Watson (1990), Aitchison (1991a), Watson (1991), Aitchison (1991b, 1992b), Woronow (1997a, 1997b), Aitchison (1999), Zier and Rehder (1998), Aitchison et al (2000), Rehder and Zier (2001), Aitchison et al (2001) and Aitchison, Barceló-Vidal and Pawlowsky-Glahn (2002). The transformation methodology has withstood these attacks, and in many ways the adverse responses have helped to clarify the important principles underlying compositional data analysis and to consolidate knowledge of the underlying algebraic-geometric structure of the simplex sample space.

7. Principles of compositional data analysis

Two main principles of compositional data analysis are scale invariance and subcompositional coherence. Scale invariance merely reinforces the intuitive idea that a composition provides information only about relative values not about absolute values, and therefore ratios of components are the relevant entities to study. This concept is easily formalised into a statement that all meaningful functions of a composition can be expressed in terms of a set of component ratios (Aitchison 1997, 2001). Subcompositions of compositions are the analogue of marginals or subvectors in unconstrained multivariate analysis (Aitchison 1986, p.33). Subcompositional coherence demands that two scientists, one using full compositions and the other using subcompositions of these full compositions, should make the same inference about relations within the common parts. Working with ratios, or equivalently logratios, involves not only scale invariance but automatically subcompositional coherence since ratios within a subcomposition are equal to the corresponding ratios within the full composition. For details of these arguments associated with subcompositional coherence see Aitchison (1992a, 1994, 1997, 2001).

8. The algebraic-geometric structure of the simplex sample space

Time has revealed the great importance of the basic operation of perturbation within the simplex S^D (Aitchison, 1986, p.27) in the analysis of compositional data. We recall that given two *D*-part compositions x and y the perturbation $x \oplus y$ is defined by

$$x \oplus y = [x_1 y_1, \ldots, x_D y_D] / (x_1 y_1 + \ldots + x_D y_D) = C[x_1 y_1, \ldots, x_D y_D],$$

where C is the closure or constraining operator, standardising the contents of a positive vector to unit sum by division by the sum of the components. The inverse operation Θ is easily defined by

$$x \Theta y = C[x_1 / y_1 \ldots, x_D / y_D].$$

The underlying reason for this is that perturbation plays in the simplex a role precisely analogous to displacement or translation in real space; it is a mechanism for recording change. For example, if a D-part composition x changes through whatever process to a D-part composition X the change can be ascribed to a perturbation p satisfying $X = p \oplus x$ with solution provided in terms of the inverse perturbation operator Θ as

$$p = X \Theta x = C[X_1 / x_1, \ldots, X_D / x_d].$$

Perturbation thus plays an important role not only in simple change as just described but also in describing imprecision, in characterising error in compositional regression and in the computation of residual compositions in compositional regression and in other compositional fitting techniques.

It is important to realise that the perturbation operation on the simplex defines an abelian group on the simplex, with identity $e = (1/D)[1, \ldots, 1]$ and inverse $p^{-1} = C[1/p_1 \ldots, 1/p_D]$.

There is a second operation in the simplex, that of powering, the analogue of scalar multiplication in real space, which is playing an increasingly important role in compositional data analysis. Given a D-part composition $x \in S^D$ and a real number $a \in R^1$ the power transformed composition is

$$a \otimes x = C[x_1^a, \ldots, x_D^a].$$

Note that we have used the operator symbols $\oplus$ and $\otimes$ to emphasize the analogy with the familiar operations of translation and scalar multiplication of vectors in the vector space R^D. It is trivial to establish that the operations $\oplus$ and $\otimes$ define a vector or linear space structure on S^d.

The structure can be extended to produce a metric vector space by the introduction of the simplicial metric $\Delta_S : S^D \times S^D \to R_{\geq 0}$ and defined by Aitchison (1983; p.193) as

$$\Delta_S(x, y) = \left[\sum_{i=1}^{D} \left\{ \log \frac{x_i}{g(x)} - \log \frac{y_i}{g(y)} \right\}^2 \right]^{1/2} \quad (x, y \in S^d),$$

where $g(.)$ denotes the geometric mean of the components of the enclosed vector. The fact that this metric has also desirable properties, such as permutation and perturbation invariance, a powering effect analogous to a scalar multiplication effect in R^D and

subcompositional dominance, relevant and indeed logically necessary for meaningful statistical analysis of compositional data, has been spelt out in detail, for example in Aitchison (1992b). The norm $\|x\|$ consistent with the metric Δ_S is defined by

$$\|x\|^2 = \Delta_S^2(x,e) = \sum_{i=1}^{D}\left(\log\frac{x_i}{g(x)}\right)^2,$$

where $e = [1/D,\ldots,1/D]$ is the identity of the perturbation group; and the associated inner product $\langle x,y\rangle$ is defined by

$$\langle x,y\rangle = \sum_{i=1}^{D}\log\frac{x_i}{g(x)}\log\frac{y_i}{g(y)}.$$

As for any vector space, generating vectors, bases, linear dependence, orthonormal bases and subspaces play a fundamental role and this is equally true for the simplex metric vector space. In such concepts the counterpart of 'linear combination' is a power-perturbation combination such as

$$x = (u_1 \otimes b_1) \oplus \ldots \oplus (u_C \otimes b_C),$$

and such combinations play a central role. In such a specification the b's are compositions regarded as generators, and the combination generates some subspace of the unit simplex as the real number u-coefficients vary. When this subspace is the whole of the unit simplex then the b's form a basis. Generally a basis should be chosen such that the generators are 'linearly independent' in the sense that $b_1,\ldots,b_C$ are linearly independent if and only if

$$(u_1 \otimes b_1) \oplus \ldots \oplus (u_C \otimes b_C) = e \Rightarrow u_1 = \ldots = u_C = 0,$$

where $e = [1/D,\ldots,1/D]$ is the identity composition. For S^D which is essentially a $(D-1)$-dimensional space, a linearly independent basis has $D-1$ generators. Important among such bases are those which form an orthonormal basis, say with generators $b_1,\ldots,b_{D-1}$, which have unit norm $\|b_i\| = 1 \quad (i = 1,\ldots,D-1)$, and are orthogonal in the sense that $\langle b_i, b_j\rangle = 0 \quad (i \neq j)$.

The coefficients of a D-part composition x relative to an orthonormal basis $b_1,\ldots,b_{D-1}$ are $\langle x,b\rangle,\ldots\langle x,b_{D-1}\rangle$ and are logratios, termed isometric logratios since the corresponding *ilr* transformation preserves the simplicial metric as the Euclidean metric in R^{D-1}. Within the *ilr* framework we can get different transformations corresponding to different orthonormal bases.

As on any vector space a set of C orthonormal generators can easily be extended to form an orthonormal basis of S^D. Later we shall see that orthonormal bases play a central role in a data-analytic sense in terms of the simplicial singular value decomposition of a

compositional data set.

Clearly in compositional processes rates of change of compositions are important and here we define the basic ideas. Suppose that a composition $x(t)$ depends on some continuous variable t such as time or depth. Then the rate of change of the composition with respect to t can be defined as the limit

$$Dx(t) = \lim_{dt \to 0} \frac{1}{dt} \otimes \{x(t+dt) \ominus x(t)\} = C[\exp\{\frac{d}{dt} \log x(t)\}]$$

where d/dt denotes 'ordinary' differentiation with respect to t. Thus, for example, if $x(t) = x \oplus h(t) \otimes b$ then $Dx(t) = h'(t) \otimes b$.? There are obvious extensions through partial differentiation to compositional functions of more than one variable. We note also that the inverse operation of integration of a compositional function $x(t)$ over an interval (T_0, T) can be expressed as

$$\oint x(t)dt = C[\exp\{\int_{T_0}^{T} \log x(t)dt\}].$$

For further details of this algebraic-geometric structure of the simplex see Aitchison (2001), Aitchison et al (2002), Barceló-Vidal, Martin-Fernández, and Pawlowsky-Glahn, (2001), Pawlowsky-Glahn and Egozcue (2001).

9. Limitations in the interpretability of compositional data

There is a tendency in some compositional data analysts to expect too much in their inferences from compositional data. For these the following situation may show the nature of the limitations of compositional data.

Outside my home I have a planter consisting of water, soil and seed. One evening before bedtime I analysed a sample and determined its (water, soil, seed) composition as x = [3/6 2/6 1/6]. I slept soundly and in the morning again analysed a sample, finding X = [6/9 2/9 1/9]. I measured the change as the perturbation

$$X \ominus x = C[(6/9)/(3/6) \quad (2/9)/(2/6) \quad (1/9)/(1/6)] = [1/2 \quad 1/4 \quad 1/4].$$

Now I can picture two simple scenarios which could describe this change. Suppose that the planter last evening actually contained [18 12 6] kilos of (water, soil, seed), corresponding to the evening composition [3/6 2/6 1/6], and it rained during the night increasing the water content only, so that the morning content was [36 12 6] kilos, corresponding to the morning composition [6/9 2/9 1/9]. Although this *rain only* explanation may be true, is it the only explanation? Obviously not, because the change could equally be explained by a *wind only* scenario, in which the overnight wind had swept away soil and seed resulting in content of [18 6 3] kilos and the same morning composition [6/9 2/9 1/9]. Even more complicated scenarios will produce a similar change. For example a combination of *rain and wind* might have resulted in a combination of increased water and decreased soil and seed, say to a content of [27 9 4.5] kilos, again with morning composition [6/9 2/9 1/9].

The point here is that compositions provide information only about the relative magnitudes of the compositional components and so interpretations involving absolute values as in the above example cannot be justified. Only if there is evidence external to the compositional information would such inferences be justified. For example, if I had been wakened by my bedroom windows rattling during the night and I found my rain gauge empty in the morning I would be justified in painting the wind only scenario. But I slept soundly during the night.

A consequence of this example is that we must learn to phrase our inferences from compositional data in terms which are meaningful and we have seen that the meaningful operations are perturbation and power.

10. Characteristics of compositional variability

For statistical modelling we have to consider distributions on the simplex and their characteristics. The well-established 'measure of central tendency' $x \in S^D$, which minimizes $E\{\Delta_S(x,x)\}$, is the 'centre'

$$x = cen(x) = C[\exp\{E(\log x)\}].$$

Conforming with this mean value there is a variety of equivalent forms of dispersion and covariance characteristics, the logratio covariance matrix Σ (Aitchison 1986, p.77), the centred logratio covariance matrix Γ (Aitchison, 1986, p.79) and the variation matrix T (Aitchison, 1986, p.76). Importantly these dispersion characteristics are consistent with the simplicial metric defined above.

11. Consequential results for compositional data sets

In what follows we shall be concerned with compositional data sets, typically an $N \times D$ matrix X with nth row composition x_n. First we note that the estimate $\hat{x}$ of x is given by

$$\hat{x} = C[g_1, \ldots, g_D],$$

where the g's are the geometric means of the individual components. There is for such a compositional data matrix a central result, analogous to the singular value decomposition for data sets associated with the sample space R^D, on which much of multivariate statistical methodology is based. Any compositional data matrix X can be decomposed in a power-perturbation form as follows

$$x_n = \hat{x} \oplus (u_{n1} p_1 \otimes b_1) \oplus \ldots \oplus (u_{nR} p_R \otimes b_R),$$

where $\hat{x}$ is the estimate of the centre of the data set, and p_i $(i = 1, \ldots, R)$ are positive 'singular values' in descending order of magnitude, the b_i $(i = 1, \ldots, R)$ are orthogonal

compositions, R is a readily defined rank of the compositional data set and the u's are power components specific to each composition. In practice R is commonly $D-1$, the full dimension of the simplex. In a way similar to that for data sets in R^D we may consider an approximation of order $r < R$ to the compositional data set given by

$$x_n^{(r)} = \hat{x} \oplus (u_{n1} p_1 \otimes b_1) \oplus \ldots \oplus (u_{nr} p_r \otimes b_r).$$

Such an approximation retains a proportion

$$(p_1^2 + \ldots + p_r^2) / (p_1^2 + \ldots + p_R^2)$$

of the total variability of the $N \times D$ compositional data matrix as measured by the trace of the estimated centered logratio covariance matrix or equivalently in terms of the total mutual squared distances as

$$\{N(N-1)\}^{-1} \sum_{m<n}^{D} \Delta_S^2 (x_m, x_n).$$

This increased understanding of the algebraic-geometric structure of the underlying simplex sample space has opened up the possibility of a staying-in-the-simplex approach to compositional data analysis, an alternative to the logratio analysis approach. It is important to realise that the approaches are equivalent in the sense that each, properly used and interpreted, will lead to identical inferences. Which is preferred by a particular analyst will, I believe, depend largely on personal choice, with the more mature mathematically probably favouring the stay-in-the-simplex approach. With the simplex as a metric vector space ideas such as minimum variance unbiasedness and least squares estimation, are available in compositional data analysis, as demonstrated by Pawlowsky-Glahn and Egozcue (2002).

12. Probability measures on the simplex

A welcome addition to the various classes of parametric distributions on the simplex – the additive logistic normal (Aitchison and Shen, 1980; Aitchison, 1986, p.113), the multiplicative logistic normal (Aitchison, 1986, p.130), partitioned classes (Aitchison, 1986, p.132) and the Dirichlet-embracing generalisation *Aitchison, 1985, 1986) – is the multivariate logistic skew normal based on the multivariate skew normal class on R^D introduced by Azzalini and Dalle Valle (1996) and further developed by Azzalini and Capitanio (1999). This allows for skewness in the logratio transformed data and promises to allow more extensive study of methods which depend on distributional form. For some uses of this distribution in compositional data analysis see Aitchison and Bacon-Shone (1999), Mateu-Figueras, Barceló-Vidal and Pawlowsky-Glahn (1998). An underlying difficulty with this parametric class may turn out to be the complex relationship among the parameters, for example in the sense that the correlations involve the skewness parameters.

The characteristic and moment generating functions for distributions in R^D are familiar useful tools of distributional analysis. Study of the properties of simplicial distributions has been greatly eased by the introduction of the appropriate transform or generating

function. The transform which seems to be most suited to this purpose is a multivariate adaptation of the Mellin transform. Let

$$U^D = \{(u_1, \ldots u_D): u_1 + \ldots + u_D = 0\}.$$

Suppose that a composition $x \in S^D$ has density function $f(x)$. Then define its Mellin generating function $M_x: U^d \to R^1$ by the relationship

$$M_x(u) = \int_{S^d} x_1^{u_1} \ldots x_D^{u_D} f(x) dx .$$

Note that the restriction of the vector u to the hyperplane U^D rather than R^D is dictated by the need to meet the requirement of scale invariance, here ensured by the fact that integrand is expressible in terms of ratios of the components of x.

The Mellin generating function has perturbation, power and limit properties similar to additive and scale properties of characteristic and moment generating functions for distributions in R^D. For further details of its properties and uses see Aitchison (2001).

Testing for distributional form and outlier detection has also been developed by Barceló, Pawlowsky and Grunsky (1996) and an attempt at a definitive form which overcomes the problem of choice of divisor and allows insight into the extent of concurrence has recently been obtained in Aitchison, Mateu-Figueras and Ng (2003).

It is also worth remembering that kernel density estimation is also available for compositional data; see, for example, Aitchison and Lauder (1985).

14. Compositional processes

Most scientists are interested in the nature of the process which has led to the data they observe. For example, geological language contains many terms to describe a whole variety of envisaged geochemical processes, such as denudation, diagenesis, erosion, gravity transport, metasomatism, metamorphism, mixing, orogenesis, polymetamorphism, sedimentation, transportation, weathering. Unfortunately the scientist is seldom in the position of observing a closed system where fundamental principles such as conservation of mass and energy apply. Commonly the only data available take the form of compositional data providing information only on relative magnitudes of the constituents of the specimens. Thus there is a need to extend compositional data analysis to provide satisfactory models to describe such processes. We direct attention here to two such processes: differential perturbation processes and convex linear mixing processes.

Differential perturbation processes.

Many of the terms used to describe the compositional processes appear to envisage some kind of differential change in the components of the composition. Since differential change in compositions is simply characterised by the simplex operation of perturbation this seems the sensible tool for the mathematical statistical study of such processes. The fundamentals for such a study were set out in Aitchison and Thomas

(1998). Briefly the argument went as follows.

Consider a process which results in an observable *D*-part composition $x(t) = [x_1(t), \ldots, x_D(t)]$ which varies with some ordered variable such as time *t*. Since processes are commonly assumed to take place continuously over time we can attempt to describe such a process in a time-differential way by relating the composition $x(t + dt)$ at time $t + dt$ to the composition $x(t)$ at previous time t in terms of a small perturbation. Since such an infinitesimal perturbation will be a slight departure from the identity perturbation [1/*D*, . . . , 1/*D*] the process can be set out as

$$x(t + dt) = x(t) \oplus (1/D)\{1 + d_1(t)dt, \ldots, 1 + d_D(t)dt\}.$$

Sometimes it is convenient to assume that such a perturbation is in the *D*-part simplex but since the perturbation operation is invariant with respect to scale there is strictly no need for such a requirement. The original development then moved to a set of differential equations in logratios for which the solution is

$$x(t) - x(t_0) \oplus [\exp\left\{\int_{t_0}^{t} d_i(u)du\right\} \ (i = 1, \ldots, D)],$$

where $x(t_0)$ is the known or assumed composition at time t_0.With differentiation now defined on the simplex we note that an alternative expression of the process is in terms of the simple differential equation $Dx(t) = C[\exp(d_i(t)\colon\ i = 1, \ldots, D]$ with the known value at t_0 being the 'boundary condition'.

An interesting and important special case is where $d_i(t) = g_i h(t)$, when the relationship takes the form of a simple compositional regression in a power-perturbation form as

$$x(t) = x(t_0) \oplus H(t) \otimes b,$$

where

$$H(t) = \int_{t_0}^{t} h(t)dt \ \text{ and } \ b = C[(\exp(g_i)\colon\ i = 1, \ldots, D]$$

With actual compositional data the regression either in logratio terms or in staying-in-the-simplex mode is easily accomplished. The important feature here is the possibility of alternative approaches to interpretation. For further details and an application see Aitchison and Thomas (1998) and for further developments see Aitchison and Barceló-Vidal (2002)..

A great disappointment here is that while scientists are very ready to argue that a main avenue of study is in process of change, (we've already seen the jargon of geology above) such as in biologal developmental processes, environmental processes, there seems to be little attempt to structure these in sensible probabilistic compositional terms.

15. Rest and be thankful

To reach my home from the University of Glasgow I have to drive on a climbing road which was once so treacherous that the summit was accorded the name 'Rest and be thankful'. I think in compositional data analysis we have reached such a position. However the road ahead to my home is a twisting , undulating, single track with passing and overtaking slots, and it is well equipped with ditches for the unwary. Driving needs concentration with particular attention to the variable road conditions and weather. Each problem, snow and ice, driving rain, reflection glare from a low sun, needs it own solution. I think the sensible way ahead with compositional data analysis is probably to be found in facing up to the applied problems which face us. This workshop with its emphasis on application is an opportunity to face the challenge of new problems in a great variety of disciplines.

Some comments at the resting place

Most of us here would probably accept most of the previous rather theoretical part of this lecture with its theme that some logical consequences of two simple principles of compositional data analysis, namely scale invariance and subcompositional coherence. To an extent some of our problems seem to arise from an embarrassment of riches, in that we have available three different transformation possibilities – *alr*, *clr* ad *ilr* - each with advantages and disadvantages, so that we have to make choices to suit the applied problem. The transformation *alr* is simple in that the logratios involve only two components and so are relatively easy to interpret; but care has to be taken to note that the simplicial metric is based on the norm $alr(x)\,H\,alr(x)^T$ and so careless consideration of the Euclidean metric obtained after the transformation to R^{D-1} is incorrect. The transformation *clr* has been criticised because, while isometric and treating the components symmetrically, it transforms onto a hyperplane of R^D and that the associated centred logratio covariance matrix is singular; these criticisms are more fussy than real since modern matrix algebra provides generalised inverses and determinants. For example, I have now designed my compositional regression program so that I regress $clr(x)$ on the concomitants and then it is easy to pick out any two-component logratio by a simple subtraction. The *ilr* transformations are especially useful for theoretical work where the simplicial metric is particularly important such as in establishing 'least squares' properties. I have found, however, that they are not particularly suited to providing simple modelling of applied problems.

We now have also a staying-in-the-simplex possibility. While this is elegant and certainly satisfying to the mathematically adequate there must be some doubts about how we can serve the various disciplines in the use of the mathematical ideas in interpreting inferences in consultative work.

Now I would like to focus on some real applied problems and see how this may alter our modelling, even our sample space. My argument from now on is driven by an applied mathematical approach.

16. Caveat 1: Is my problem compositional? A study of the crustacean *Tripartus Aitchisonii*

A biological friend, BF, has discovered, and is fish-farming a new thin-shelled crustacean, which he has named Tripartus Aitchisonii. It is similar to those Plymouth shrimps that Karl Pearson and Weldon discussed in Pearson (1997). It has a head, body and tail, with only the body edible. BF has just brought me the results of an experiment he has carried out to investigate the effects of a new hormone addition to diet, which he anticipates may increase the proportion of body at the expense of head and tail. His experiment consisted of separating a randomly selected sample of 100 Triparti Aitchisonii at random into two sets, each of 50. The first set was sacrificed to determine their compositions as proportions by weight. The second set was fed the hormone-enhanced diet over a period of eight weeks and then sacrificed to determine their compositions as proportions by weight. BF is numerate and had plotted these compositions in the triangular diagram of Figure 1, where blue denotes before and red after the hormone treatment. He was excited and ready to bulk purchase the hormone. 'Steady', I said. Statisticians experienced in having to deal with data from experiments in which they have had no part in designing will appreciate my caution.

'Have you any data other than the compositions', I asked. 'Oh, yes', he replied, 'I have the weights of each specimen. Here's the complete data set.' You can see my train of thought. We know that the shape (head, trunk, leg) of children change as they grow taller. May the obvious change in composition be solely due to natural growth in weight of Tripartus. It's obviously a badly designed experiment with confounding between treatment and natural developmental effects. But let's investigate the data.

A histogram of the before and after weights is shown in Figure 2. Note the substantial differences between before and after weights and the skewness of the distributions. I decided to construct a lattice of hypotheses to investigate the situation. (Fig. 3). The maximum model M considered was

$$M: \quad x_b = x_b \oplus (t \otimes b_b) \oplus \ error, \quad x_a = x_a \oplus (t \otimes b_a \oplus \ error),$$

where *b*, *a* denote before and after, *t* denotes logarithm of weight, and b_b, b_a are the compositional form of regression coefficients. The simplest hypothesis H_0 (at the bottom of the lattice) is one of 'no difference, no size effect), namely

$$H_0: \quad x_b = x \oplus \ error, \quad x_a = x \oplus \ error,$$

with at higher levels the hypotheses H_1 and H_2 of 'no size effect' and 'equal size effect', respectively

$$H_1: \quad x_b = x_b \oplus \ error, \quad x_a = x_a \oplus \ error,$$
$$H_2: \quad x_b = x \oplus (t \otimes b) \oplus \ error. \quad x_a = x \oplus (t \otimes b) \oplus \ error.$$

Use of generalased likelihood ratio tests shows that we must successively reject H_0, H_1 but that H_2 cannot be rejected, the test statistic value of 11.8 to be compared against the 5 per cent critical value 14.07 of chi-squared at 7 degrees of freedom.

Thus we conclude that the apparent compositional change so obvious in Figure 1 can be wholly explained by the increase in size. Of course, it may be that the hormone additive to diet is responsible for the increase in size but because of the confounding of possible effects within this experiment there is no way of establishing the truth. We can only recommend to our biologist that he conducts a properly designed experiment as in Aitchison and Ng (2003, in a later session in CODAWORK03). And it would obviously be better to have a larger number of specimens. After all I'm enjoying meals of Tripartus Aitchisonii.

Note: The data set and Figures 1- 3 will be available at the workshop after this lecture.

17. Caveat 2: Is the Hilbert space simplex the appropriate sample space? A study of how the lesser goilbird spends its time

Given the elegance of the algebraic-geometric (Hilbert space) structure of the simplex it is easy to fall into the pure-mathematical trap that all compositional problems must depend on this structure, that all statistical problems should be addressed in terms of coordinates associated with orthonormal, isometric bases, that orthogonality is closely associated with statistical independence. Let me say here that I think that many of these ideas are important in establishing useful results. For example, such a structure is obviously central to establishing the counterparts of the well known Markov least squares theory associated with R^D. But while we recognise the simplex as our compositional sample space we must ensure that the ways we place probability measures or distributions on that sample space are appropriate to the applied compositional problem we face. I take an example similar to the statistician's day problem in Aitchison (1986, Sections 1.9, 10.3) for illustrative purposes. Time budgets have become a regular source of information in analysing behaviour patterns in many disciplines. Our example concerns the behaviour pattern of the lesser goilbird, a garden bird whose territory is confined to a particular garden. Its four activities (feeding, fighting | perching, sleeping) divide themselves into two natural divisions: active, including feeding and fighting, and passive, including perching and sleeping. Obvious behavioural questions are whether active and passive patterns are independent and whether these patterns are independent of the division of the day between active and passive.

The time budgets of 60 goilbirds observed in 60 gardens over random days is given in Table 1.

In terms of the generic composition $[x_1\ x_2\ x_3\ x_4]$ we are here dealing with a partition $[x_1\ x_2\ |x_3\ x_4]$ of order 1' The relevant question in terms of logratios is whether

$$y_1 = \log(x_1 / x_2),\ \ y_2 = \log(x_3 / x_4),\ \ y_3 = \log\{(x_1 + x_2) / (x_3 + x_4)$$

are distributed independently.

Now it has been put to me that the way to tackle such problems is to consider an isometric logratio transformation, acknowledging that an appropriate representation of the composition is in terms of coordinates with respect to an orthonormal basis, resulting in

$$x = \log(x_1 / x_2) \otimes e_1 \oplus \log(x_3 / x_4) \otimes e_2 \oplus \log(x_1 x_2 / x_3 x_4) \otimes e_3,$$

even suggesting that establishing that

$$z_1 = \log(x_1 / x_2),\ z_2 = \log(x_3 / x_4),\ z_3 = \log(x_1 x_2 / x_3 x_4)_3$$

are independent would imply independence of y_1, y_2, y_3. This is simply not true, as our the data set will demonstrate.

The correlation matrices of y_1, y_2, y_3 and z_1, z_2, z_3 are as follows

1.0000	-0.0022	-0.0861
-0.0022	1.0000	-0.2457
-0.0861	-0.2457	1.0000

and

1.0000	-0.0022	-0.6227
-0.0022	1.0000	-0.6404
-0.6227	-0.6404	1.0000

demonstrating clearly that there is independence associated with the real question whereas the pseudo-question suggests dependence between the subcompotions and the partition.

Another line of the orthonormalists is that the appropriate modelling must indeed be in terms of the orthonormal coefficients z_1, z_2, z_3 and then it is simply a case of expressing the relevant variables y_1, y_2, y_3 in terms of these coordinates. The first two relations are obviously straight forward but

$$y_3 = \frac{\exp\{\frac{1}{2}(z_1 + z_2 + z_3) + \exp\{\frac{1}{2}(z_1 - 3z_2 + z_3)}{\exp(z_2) + 1}.$$

This will, of course, lead to a correct analysis but my point is why go to all this complexity, not addressing the problem of interest in its simplest terms. Statisticians have over the past century addressed problems of statistical independence correctly without being aware of any algebraic-geometric structure of their sample spaces. My complaint is not that such structure is unimportant but that we must not let pure-mathematical ideas drive us into making the statistical modelling more complicated that is necessary. Simplicity in modelling is important, particularly when we have to explain the inferences to less numerate colleagues.

18. Caveat 3. Is the Hilbert space simplex the appropriate sample space? A study of multiplicative subjective probability assessments.

A less well known niche of compositional problems is where *subjects* are presented with a series of *cases* of unknown *category,* given information about each case and

asked to assign probabilities to each of the possible categories. See, for example, Taylor, Aitchison and McGirr (1970), Aitchison and Kay (1976). for some typical situations. In one diagnostic situation involving three possible categories (a diagnosis of malignancy in adrenal carcinoma, or one of two benign adrenal conditions, namely adenoma or hyperplasia) each subject was asked first to divide the available unit of probability between malignancy and benignancy, say as x_1 and $1-x_1$; then to divide the remaining $1-x_1$ between adenoma and hyperplasia as x_2 and x_3. The natural way of investigating the resulting composition $[x_1\ x_2\ x_3]$ is in terms of the ratios $x_1/(1-x_1)$, x_2/x_3, or their logratios, and leading as above to a much simpler and direct analysis than what would be attained by insisting on working within the Hilbert space coordinate systems

I suspect that there are many problems of this 'remaining space' nature waiting to be tackled along compositional lines, for example in developmental biology and in the earth sciences. See, for example, the discussion of Niggli remaining space in Chayes (1983)..

19. Convex linear mixing

A popular way in some disciplines, such as sedimentology and environmental pollution studies, of studying compositional data is in terms of convex linear modelling processes. Such an approach is based on some such assumption as conservation of mass. There is, of course, no way that compositional data can be used to support such a mass conservation hypothesis since compositions carry no information about mass. Compositions can, however, be analyzed *within models which assume conservation of mass.* All these models assume that there are source compositions, say $x_1,\ldots,x_C$, from which a generic observed composition x arises as a convex linear combination

$$x = p_1x_1 + \ldots + p_Cx_C\ ,$$

where $p = [p_1,\ldots,p_C] \in S^C$ is the vector of mixing proportions. The form of modelling obviously depends on the extent of the information about the number of sources and the source compositions. At the 'ignorance end' neither the number of sources nor their compositions are known – the so-called *endmember problem* as presented, for example, in Renner (1993) and Weltje (1997). At the opposite extreme the problem may be to test a hypothesis that the sources are specified compositions $x_1,\ldots,x_C$. Many intermediate situations can be visualised: an example is the pollution problem analysed by Aitchison and Bacon-Shone (1999), where there are not only samples from the target set but also sampled compositions from the sources.

Note that the basic operation here is an additive one, so that all the nice distributional properties associated with perturbation and power are not available. For example, given that $x_1,\ldots,x_C$ are independently distributed as $L^{D(}(a_1,\Omega_1),\ldots,L^D(a_C,\Omega_C)$ and that p is a constant or has some given logistic normal distribution, no explicit form for the distribution of the convex linear mixture x can be found. It is only by the determination of good approximations to the distribution that Aitchison and Bacon-Shone (1999) can

resolve their pollution problem. I suspect that the full solution to other problems in this area will depend on our ability to construct such approximations.

The additive nature of such modelling does not mean that basic principles of compositional data analysis should be neglected. In solutions of the endmember problem there has been a tendency to avoid the simplicial metric and to revert to Euclidean distance and classical least squares in estimating mixture vectors. This is certainly not necessary and the more appropriate simplicial metric may be used. For example an approach to the so-called endmember problem where a set of say C endmember compositions $x_1, \ldots, x_C$ is sought such that each composition x_n ($n = 1, \ldots, N$) of the data set can be expressed as a convex linear combination x_n of $b_1, \ldots, b_C$, uses as criterion of success the magnitude of

$$\sum_{n=1}^{N} \Delta_S^2(x_n, x_n)$$

while monitoring the magnitude of

$$\sum_{b<c} \Delta_S^2(x_b, x_c) .$$

See Aitchison and Barceló-Vidal (2002) for further details and an example of a method of comparing the adequacy of differential perturbation and convex linear mixing processes. In the computation for such analysis a basic algorithm is obviously required for the maximisation or minimisation of a function on the simplex and we now have efficient search algorithms based on perturbation techniques.

20. Perturbation and subcompositional stability analysis

In standard multivariate statistical analysis common hypotheses of interest concern changes in mean vectors and subvectors and there is a substantial methodology to deal with such applied problems. In compositional data analysis it is now well established that compositional change is most readily described in terms of the simplicial operation of perturbation and that subcompositions replace the marginal concept of subvectors. Since it is obvious that hypotheses concerning perturbations and subcompositions are manageable within the framework of either logratio or staying-in-the-simplex analysis it is surprising that little application has been undertaken in this area. We have seen a simple application of perturbation hypothesis testing in Section 12. Since a later paper in this workshop (Aitchison and Ng, 2003a) will present the challenge of such hypothesis testing in the analysis of two experiments in food production I will not go into further details here.

21. Joint compositional distributions

Some years ago, having been asked by several geologists, whether logratio analysis would apply to bicompositions such as major oxide by trace element compositions I submitted a paper to Math Geology showing how the analysis could be done on the

basis of partition modelling, with a number of examples to illustrate the methodology. The first referee, a geologist, could not see any real geological problem in practice (despite the fact of the requests from geologists) and damned the paper with faint praise. The second appeared to be a statistician of sorts, with no idea about the nature of compositional problems and made objections on the basis of what would arise from application of standard multivariate analysis. I complained to the editor about the quality of the refereeing and the paper was put to a third referee, an arbiter who 'sat so firmly on the fence' .that I withdrew the paper. I think many of us here may have had similar experiences and I'll have something to say about the quality of refereeing later. The form of analysis used is still, in my view valid, and of course would apply equally to other situations, where for example the bicomposition consists of (hair colour, eye colour) proportions within different sections of a population as in the study of Tocher (1908), in the study of (blood, urine) compositions in clinical medicine, in psephology in trying to relate the US Presidential vote composition by state to the (ethnic, rural) bicomposition.

One point which is worth making is that in geochemistry major oxides and trace element compositions are essentially a single composition though measured usually in different units, percentages by weight and parts per million. The conversion to common units can be simply made in terms of a perturbation of the data. Since perturbation affects only the centre of a compositional distribution and not the dispersion the methodology for analysing such compositional variability is essentially invariant under perturbation..

.

There are still issues involved, and many areas of application to major oxides and trace elements in geology, to blood and urine compositions in clinical medicine, to various genotype compositions.

22. Multiway compositional problems

A school pupil has a hair colour and an eye colour. In the Tocher (1908) study hair colour and eye colour of every Scottish schoolboy and girl was recorded and for each of 33 regions the hair and colour compositions for the regional population were recorded separately. While this is obviously of interest so much more information might have been obtained if the two-way composition had been recorded. What proportion of pupils have blue eyes and red hair? I suspect that Tocher may have such detail in mind because he records such a two-way composition for one of the regions. Perhaps his research money ran out?

Aware of the possibilities of investigating hypotheses associated with such multiway tables I had a PhD student (C. K. Li) in Hong Kong investigate the nature of such problems. It was to an extent theory looking for an application since the only daa set we could find in the literature was a two-way classification of a small sea-bed study and there were no obvious hypotheses of interest. There is no doubt that other data sets and associated problems exist. Granulometric data appear to be not only classified by diameter of particle but also by nature of particle. And in the US Presidentail Election data it would obviously be interesting to have the state compositions presented in a two-way manner in terms of ethnic x rural/urban.

So we have here an example of adequate theory awaiting a real application.

23. Graphical aids

Harker and related diagrams. It is now over four decades since Felix Chayes warned geologists of the dangers of attempting to interpret Harker and similar diagrams where one component of a composition is plotted against another. Yet a recent search of the web under 'Harker diagram' produced some 60 sites, many of them instructing students in the use of such 'graphical aids'. The only legitimate use of such diagrams is in terms of the ratios, that is in terms of the rays from the origin to the data points. In my view Harker diagrams are best condemned as misleading and best left out of any attempts to interpret compositional variability.

Ternary diagrams. Like Harker diagrams these should be treated with caution. For example, in the past there has been substantial discussion on the nature of data sets with apparent curvature within a ternary diagram (Butler 1979) Are these trends or not? With our knowledge of the algebraic structure of the simplex we now know that constant logcontrast 'curves' are indeed the 'straight lines' of the simplex and so any interpretation of curvature within the ternary diagram should be treated with substantial caution. See Aitchison and Thomas (1998) for an example where such curvature can indeed be interpreted as a trend or compositional process.

Ratio and logratio scattergrams. If scattergrams are to be used in interpreting compositional data then because of the necessity to meet the demands of the principle of scale invariance they should involve ratios or logratios. A good example of how such diagrams can be used for exposition is to be found in the discrimatory example in Thomas and Aitchison (1998).

Compositional biplots. The development of biplot techniques for compositional data is a substantial advance in the study of compositional data sets.

The biplot (Gabriel, 1971, 1981) is a well established graphical aid in other branches of statistical analysis. Its adaptation for compositional data is simple and can prove a useful exploratory and expository tool. For a compositional data set the biplot is based on a singular value decomposition of the doubly centered logratio matrix. For details of biplot construction see Aitchison (1990b, 1997, 2001) and Aitchison and Greenacre (2002). Such biplots, consisting of vertices, rays, links and case markers, allow an overall view of compositional covariance structure, subcompositional analysis, the relationship of individual compositions to parts, and provide useful interpretations of near-coincident vertices, collinear vertices and orthogonal links.

There are obviously extensions of biplot methodology to bicompositions and to conditional biplots.

24, Compositional regression

Compositional regression, where the composition is the regressand and we seek an explanation of its variability in terms of factors and/or concomitant variable, has been

extensively discussed and illustrated in Aitchison (1986, 7.6–7.9) and need not be further discussed here. Such linear modelling within the alr-transformation methodology is simple and can rely on standard multivariate techniques. The expression of compositional regression by the staying-in-the-simplex approach is by way of power-perturbation combinations. A composition x depends on concomitants $t_1, t_2, \ldots$ as

$$x = a \oplus (t_1 \otimes b_1) \oplus (t_2 \otimes b_2) \oplus \ldots \oplus p,$$

where the composition a is the analogue of 'intercept' in ordinary regression, the compositions $b_1, b_2, \ldots$ are the analogues of the 'regression coefficients' and p is the perturbation error. Clearly interpretation here is dependent on a sound mathematical appreciation of the algebraic-geometric structure of the simplex.

25. Ordination

A popular pursuit in some disciplines is that of ordination whereby some statistical means is sought to place the multivariate specimens in some linear ordering which may have some significance within the discipline. A standard method os attempting this is to order on the basis of the magnitude of the first principal component. In compositional terms this takes the form of a principal logcontrast analysis and a good example is to be found in von Eynatten, Barcelo-Vidal and Pawlowsky-Glahn (2003). An alternative staying in the simplex is to perform a compositional singular value decomposition of the data set as in Section 11 above and use an ordering of the u_{n1} $(n = 1, \ldots, N)$.

26. Subcompositions and logcontrasts

It is worth pointing out that a subcomposition can be simply identified with a special set of logcontrasts. For example there is a one-to-one relationship between the (1, 2, 3, 4) subcomposition and the values of the logcontrasts:

$$\log x_1 - \log x_2. \quad \log x_1 + \log x_2 - 2\log x_3, \ \log x_1 + \log x_2 + \log_3 - 3\log x_4.$$

The reader will see here the elements of a Helmert transformation. One feature to note is that in such a representation the parts of the composition are in a specific order.

27. Natural laws

The discovery of any 'natural law' from compositional observations has been the subject of debate recently. The tools for such discovery are again either principal logcontrast analysis or, equivalently, singular value decomposition. For details of such discoveries through principal logcontrast analysis see Aitchison (1999) and through biplot analysis see Aitchison and Greenacre (2002).

28. Compositions in an explanatory or regressor role

Aitchison (1986, Chapter 12) gave a number of practical situations where compositions play an explanatory or regressor role, where we may wish to see how a composition is changed by different treatments, where in experiments with mixtures we may attempt to determine the mixture which will provide the optimum response, and in classification or diagnostic problems where we may wish to use a composition as a convenient or efficient means of determining type or to find out if any subcomposition accounts for the substantive difference between the types.

Binary logistic discrimination. I take the classification-diagnostic problem to illustrate how simple the technique here may be developed. For two types ($t = 0$, $t = 1$) a useful model is the binary logistic model, using a logcontrast

$$lc(\alpha, x) = \alpha_0 + \alpha_1 \log x_1 + \ldots + \alpha_D \log x_D \quad (\alpha_1 + \ldots + \alpha_D = 0)$$

as the regressor. More specifically,

$$pr(t = 0 \,|\, x) = pr(t = 1 \,|\, x) = \frac{\exp\{lc(\alpha, x)\}}{1 + \exp\{lc(\alpha, x)\}} \ .$$

Maximum likelihood estimation of the parameter α is straightforward. The beauty of this model is that the adequacy of a subcomposition say (1, . . . , C) can readily be tested since this hypothesis can be expressed as $\alpha_{C+1} = \ldots = \alpha_D = 0$. Thus the whole lattice of subcompositional hypotheses can be investigate and any adequate subcomposition identified. Examples of this procedure can be found for hongite-kongite discrimination and Permian and post-Permian: discrimination (Aitchison 1986, Sections 12.6, 12.7). Such reduction to subcomposition; can be important because it may eliminate expensive determinations.

Probably the most dramatic example of such discrimination is in the Thomas and Aitchison (1998) analysis of Scottish metamorphosed limestones, where out of a 17-part geochemical composition a 3-part subcomposition is found to be an adequate discriminator. A further discussion of this will be given later in this workshop (Thomas and Aitchison, 2003).

With such a powerful tool available it is disappointing that no other applications seem to have been undertaken.

While I have confined attention above to two types the modelling is easily extended to more than two types.

Sequential discrimination. Even with more than two types the above binary logistic regression approach may be possible, even sensible. For example, in cliinical medicine when a possible case of Cushing's syndrome presents itself, the possibilities are five: the patient's condition is (1) normal, (2) ectopic carcinoma, (3) adrenal carcinoma, (4) adrenal adenoma, (5) adrenal hyperplasia. The compositional problem here is that

diagnosis is based on 14-part urine steroid metabolite compositions. Following a suggested model of the clinician's thought processes we might consider a four binary step sequence towards a diagnosis: discriminate between

(i) 1 and (2, 3, 4, 5),
(ii) If not 1, then (2, 3) against ((4,5),
(iii) If (2, 3) then (2) against (3),
(iv) If (4, 5) then (4) against (5).

A possible advantage of this sequential process is that it may be found that different subcompositions are important at different stages and this may have some importance in the treatment of the disease.

It would be interesting to see if such sequential processes have any bearing in other disciplines, for example in geology in the classification of rock types.

29. Experiments with mixtures

Within this category of compositional problems is a large set where the aim is to investigate some response, commonly univariate and quantitative, but possibly even compositional, to different mixtures (and so compositions) of ingredients. For example, how does the micro-hardness of glass depend on the composition of the rare element additive. Here the simplex is the design space. Within this niche the question arises of which mixtures should be used, essentially the question of the efficient or optimum design of the experiment.

30. Problems of zero components, in particular essential or structural zeros

The replacement method (Aitchison, 1986, p. 266) of rounded or trace zeros is not subcompositionally coherent and should now be replaced by the method arrived at independently by Fry, Fry and McLaren (2000) and Martin-Fernández, Barceló-Vidal and Pawlowsky-Glahn (2000), which preserve the ratios of non-zero components. Such replacement procedures still appear to be the most viable methods available provided sensitivity analysis over a sensible range of replacement values is used as a check.

One of the tantalising remaining problems in compositional data analysis lies in how to deal with data sets in which there are components which are essential zeros. By an essential zero we mean a component which is truly zero, not something recorded as zero simply because the experimental design or the measuring instrument has not been sufficiently sensitive to detect a trace of the part. Such essential zeros occur in many compositional situations, such as household budget patterns, time budgets, pollen zonation studies. Devices such as non-zero replacement and amalgamation are almost invariably ad hoc and unsuccessful. An alternative approach through ranking of components is given by Bacon-Shone (1992).

For some essential or structural zeros careful consideration of the questions being asked can sometimes remove the problem; see for example the predator-prey example in Aitchison (1986, Section 11.7)

Research is under way to attempt to construct two-stage models for the treatment of essential or structural zeros. In such modelling it seems sensible to build up a model in two stages, the first determining where the zeros will occur and the second how the unit available is distributed among the non-zero parts. Two reports on this promising line of research will be presented later in this workshop by Aitchison and Kay (2003) and Bacon-Shone (2003).

31. Implications of compositional data analysis for other sample spaces

The experience of researchers in compositional data analysis has some lessons for workers with other forms of data. The importance of the identification of the principles such as scale invariance and subcompositional coherence, the clear definition of an appropriate sample space and recognition of the basic operations of change such as perturbation and power, have led us to meaningful systems of statistical inference. The same has been true of the analysis of directional data based on the special algebraic-geometric structure of the sphere. It is now being recognised that many, even most, standard multivariate data problems are concerned with positive (or non-negative) vectors and that perhaps we should pay particular attention to the peculiar properties of the appropriate sample space. Included within this category would be ratio data. See Aitchison and Ng (2003b) for a discussion of this.

32. Implications of compositional data analysis for simplex parametric spaces

Multinomial and contingency table data depend for their analysis on the assignment of probabilistic parameters within a model, or by way of a hypothesis, to the categories or boxes of the contingency table. Such assignments are mathematically similar to compositions since they are divisions of the unit of probability to the categories or boxes. The contribution of compositional analysis here is through forms of Bayesian analysis, simple or hierarchical, where logistic normal distributions are assigned in various ways to the parameter vector. For an excellent example of such an approach, see Billheimer, Guttorp and Fagan (1997):.

33. A personal view of the future of compositional data analysis

I think the reader will have reached the conclusion that I think that the interesting future of compositional data analysis will lie in statisticians searching for real applied problems in as many disciplines as possible. A recent search of the web under ‘compositional data’ located over 3000 sites varying over a wide variety of disciplines, so there is plenty of challenges in this direction. Equally important is that applied workers in these disciplines should search out statisticians and present them with the challenge of answering their compositional questions. Tchebycheff, in his Theory of Maps has the fundamental idea:

Real progress is made when theory and the needs of application go hand in hand.

A substantial problem for those of us who have tried to promote understanding of the special features of compositional data analysis has been, to put it crudely, the closed minds of referees of journals and, to an extent, editors. I have a collection of referees' reports ranging from the head in the sand, who think that the simplex is nothing more than a subset of real space and 'There isn't a special problem', through those who insist that the new methodology should be doing little more than corroborating views already obtained and firmly held from previous incorrect analysis, to some who have probably spent a lifetime looking at raw correlations and pitifully know that their life's work is being attacked. I am not sure how we counter all the ignorance and prejudice. It is in my view a general trend in the quality of refereeing.

Finally one thought out of fifty years of statistical consultative work. Take time for patient discussions between statistician and person with a compositional problem. My experience is that consultees often have great difficulty in formulating precisely the purpose of their experiment or observational study. It's worth the effort. I end with my favourite quotation from Sir Harold Jeffreys, a mathematician-scientist, who preferred the simple to the complicated, and the first quotation in my 1986 monograph.

It is sometimes considered a paradox that the answer depends not only on the observations but on the question: it should be a platitude,

References

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

Aitchison, J., 1982, The statistical analysis of compositional data (with discussion): J. R. Statist. Soc. B, v.44, p. 139-177.

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

Aitchison, J., 1985, A general class of distributions on the simplex: J. R. Statist. Soc. B, v. 47, p. 136-146.

Aitchison, J., 1986, The Statistical Analysis of Compositional Data: Chapman and Hall, London. Reprinted in 2003 with additional material by The Blackburn Press.

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. Geology, 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., 1991a,. Letter to the Editor. Delusions of uniqueness and ineluctability: Math Geology, v. 23, p. 275-277.

Aitchison, J., 1991b, A plea for precision in Mathematical Geology: Math Geology, v. 23, p. 1081-1084.

Aitchison, J., 1992a, The triangle in statistics *in* Mardia, K.V., ed., The Art of Statistical Science. A Tribute to G.S.Watson : Wiley, New York, p. 89-104.

Aitchison, J., 1992b, On criteria for measures of compositional differences: Math. Geol.ogy, 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., 1997, The one-hour course in compositional data analysis or compositional data analysis is easy, *in* Pawlowsky Glahn, V., ed., Proceedings of the Third Annual Conference of the International Association for Mathematical Geology: CIMNE, Barcelona, p. 3-35.

Aitchison, J., 1999, Logratios and natural laws in compositional data analysis: Math. Geology, v. 31, p. 563-589.

Aitchison, J., 2001, Simplicial inference, *in* Viana, M.A.G. and Richards, D.St.P., eds., Algebraic Methods in Statistics and Probability: Contemporary Mathematics Series 287, American Mathematical Society, Providence, Rhode Island, p. 1-22.

Aitchison, J. and Bacon-Shone. J., 1999, Convex linear combinations of compositions: Biometrika, v. 86, p. 351-364.

Aitchison, J. and Barceló-Vidal, C., 2002,. Compositional processes: a statistical search for understanding, *in* Proceedings of the Eighth Annual Conference of the International Association for Mathematical Geology, to appear.

Aitchison, J., Barceló-Vidal, C., Egozcue, J.J. and Pawlowsky-Glahn, V., 2002, A concise guide to the algebraic-geometric structure of the simplex, the sample space for compositional data analysis, *in* Proceedings of the Eighth Annual Conference of the International Association for Mathematical Geology, to appear,

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J.A. and Pawlowsky-Glahn, V., 2000, Logratio analysis and compositional distance: Math. Geology, v. 32**,** p. 271-275.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J.A. and Pawlowsky-Glahn, V., 2001, Reply to Letter to the Editor by S. Rehder and U. Zier on 'Logratio analysis and compositional distance' by J. Aitchison, C. Barceló-Vidal, J. A. Martín-Fernández and V. Pawlowsky-Glahn: Math. Geol.ogy, v. 33, p. 849-860.

Aitchison, J. Barceló-Vidal, C., and Pawlowsky-Glahn, V., 2002, Some comments on compositional data analysis in archaeometry, in particular the fallacies in Tangri and Wright's dismissal of logratio analysis: Archaeometry, v. 44, p. 295-304...

Aitchison, J. and Greenacre, M., 2002, Biplots for compositional data: Appl. Statist., v. 51.. p. 375-382.

Aitchison, J. and Kay, J.W., 1975,. Principles, practice and performance in decision-making in clinical medicine, *in* Bowen, K. C. and White, D.G., eds., Proceedings of the 1973 NATO on The Role and Effectiveness of Decision Theories in Practice: English Universities Press, London..

Aitchison, J. and Kay, J.W., 2003, Possible solutions of some essential zero problems in compositional data analysis: paper in CODAWORK03.

Aitchison, J. and Lauder. I.J., 1985, Kernel density estimation for compositional data: Appl. Statist., v. 34, p. 129-137.

Aitchison, J., Mateu-Figueras, G. and Ng, K.W., 2003, Characterisation of distributional forms for compositional data and associated distributional tests: Math Geology, to appear..

Aitchison, J. and Ng, KW., 2003a, Aitchison, J. and Ng, KW., 2003a, Compositional hypotheses of subcompositional stability and specific perturbation change and their testing: paper in CODAWORK03.

Aitchison, J. and Ng, KW., 2003b, The statistical analysis of positive data: a review: in preparation.

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

Aitchison, J. and Thomas, C.W., 1998, Differential perturbation processes: a tool for the study of compositional processes, *in* Buccianti, A., Nardi, G. and Potenza, R., eds., Proceedings of IAMG98, The Fourth Annual Conference of the International Association for Mathematical Geology: De Frede, Naples, p. 499-504.

Azzalini, A. and Dalla Valle, A., 1996, The multivariate skew-normal distribution: Biometrika, v. 83, p. 715-726.

Azzallini, A. and Capitanio, A., 1999, Statistical application of the multivariate skew-normal distribution: J. R. Statist. Soc. B, v. 61, p. 579-602..

Bacon-Shone, J., 1992, Ranking methods for compositional data: Appl. Statist., v. 41, p. 533-537.

Bacon-Shone, J., 2003, Modelling structural zeros in compositional data analysis: paper in CODAWORK03.

Barceló, C., Pawlowsky, V. and Grunsky, E., 1996, Some aspects of transformations of compositional data and the identification of outliers: Math. Geology, v. 28, p. 501-518.

Barceló-Vidal, C., Martin-Fernández, J.A. and Pawlowsky-Glahn, V., 2001, Mathematical foundations of compositional data analysis, *in* Ross, G, ed., Proceedings of the Seventh Annual Conference of the International Association for Mathematical Geology: Volume CD, electronic publication.

Billheimer, D,. Guttorp, P. and Fagan, W.F., 1997, Statistical analysis and interpretation of discrete compositional data: NRCSE technical report **11**, University of Washington

Butler, J.C., 1979, Trends in ternary petrologic variation diagrams: J. Amer. Mineral., v. 64, p. 1115-1121.

Chang, T.C., 1988,.Spherical regression: Ann. Statist., v. 14, p. 907-924.

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

Fry, J.M., Fry, T.R.L and McLaren, K.R., 2000, Compositional data analysis and zeros in micro data: Appl. Economics, v. 2, p. 953-959.

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.

McAlister, D., 1879, The law of the geometric mean: Proc. Roy. Soc., v. 29, p. 367-375.

Martin- Fernández, J.A. Barceló-Vidal, C. and Pawlowsky-Glahn, V., 2000, Zero replacement in compositional data sets, *in* Kiers, H., Rasson, J,, Groenen, P. and M. Shader, M., eds., Studies in Classifiacation, Data Analysis and Knowledge Organisation. Proceedings of 7th Conferencee of the International Federation of Classification Societies: Springer-Verlag, Berlin, p. 155-160.

Mateu-Figueras, G., Barcelo-Vidal, C. and Pawlowsky-Glahn. V., 1998, Modeling compositional data with multivariate skew-normal distributions, *in* Buccianti, A. Nardi, G. and Potenza, R., eds., Proceedings of IAMG98, The Fourth Annual Conference of the International Association for Mathematical Geology: De Frede, Naples, p. 532-537.

Pawlowsky-Glahn, V. and Egozcue, J.J., 2001,. Geometric approach to statistical analysis on the simplex: SERRA, v. 15**,** p. 384-398.

Pawlowsky-Glahn, V. and Egozcue, J.J., 2002, About BLU estimators and compositional data: Math Geology., v.34. p. 259-274..

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.

Rehder, U. and Zier, S., 2001, Comment on "Logratio analysis and compositional

distance by Aitchison et al. (2000)": Math. Geology, v. 33, p. 845-848.

Renner, R.M., 1993, The resolution of a compositiional data set into mixtures of fixed source components: Appl. Statist., v. 42, p. 615-631.

Taylor, T.R., Aitchison, J. and McGirr, E.M., 1971, Doctors as decision-makers: a computer-assisted study of diagnosis as a cognitive skill: Brit. Med. J., v. 3, p. 35-40.

Thomas, C.W. and Aitchison, J., 1998, The use of logratios in subcompositional analysis and geochemical discrimination of metamorphosed limestones from the northeast and central Scottish Highlands, *in* Buccianti, A., Nardi, G. and Potenza, R., eds., Proceedings of IAMG98, The Fourth Annual Conference of the International Association for Mathematical Geology: De Frede, Naples, p. 549-554..

Thomas, C.W. and Aitchison, J., 2003, Exploration of geological variability and possible processes through the use of compositional data analysis: the example of Scottish metamorphosed limestones: Paper in CODAWORK03.

Tocher, J.F., 1908, Pigmentation survey of school children in Scotland: Biometrika, v, 6, p. 129-235.

von Eynatten, H., Barcelo-Vidal, C. and Pawlowsky-Glahn, V., Modelling compositional change: the example of chemical weathering of granitoid rocks: Math. Geology.

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.

Weltje, G.J., 1997, End-member modeling of compositional data: numerical-statistical algorithms for solving the explicit mixing problem: Math. Geology, v. 29, p. 503-549.

Woronow, A., 1997a, The elusive benefits of logratios, *in* Pawlowsky-Glahn, V., ed., Proceedings of IAMG97, The Third Annual Conference of the International Association for Mathematical Geology: CIMNE, Barcelona, p. 97-101.

Woronow, A., 1997b, Regression and discrimination analysis using raw compositional data - is it really a problem?, *in*: Pawlowsky-Glahn, V., ed., Proceedings of IAMG97, The Third Annual Conference of the International Association for Mathematical Geology: CIMNE, Barcelona, p. 157-162.

Zier, U. and Rehder, S., 1998, Grain-size analysis - a closed data problem, *in* Buccianti, A., Nardi, G. and Potenza, R., eds., Proceedings of the Fourth Annual Conference of the International Association for Mathematical Geology: Naples: De Frede, p. 555-558..

Table 1. Time budgets of 50 goilbirds

feeding	fighting	perching	sleeping
0.5476	0.0107	0.0113	0.4303
0.5385	0.0253	0.0090	0.4271
0.4712	0.0175	0.0211	0.4902
0.4830	0.0091	0.0553	0.4526
0.4340	0.0031	0.1003	0.4627
0.5220	0.0169	0.0321	0.4290
0.5939	0.0027	0.0115	0.3919
0.5781	0.0229	0.0222	0.3767
0.4733	0.0047	0.0122	0.5098
0.4863	0.0309	0.0096	0.4732
0.5277	0.0220	0.0058	0.4445
0.4440	0.0128	0.0044	0.5389
0.5106	0.0076	0.0215	0.4603
0.5264	0.0016	0.0406	0.4313
0.5323	0.0088	0.0262	0.4327
0.4396	0.0119	0.0258	0.5227
0.5981	0.0067	0.0191	0.3761
0.5453	0.0312	0.0121	0.4115
0.3141	0.0063	0.1560	0.5236
0.4096	0.0049	0.0227	0.5628
0.4630	0.0112	0.0068	0.5190
0.3388	0.0073	0.0235	0.6304
0.6120	0.0095	0.0107	0.3679
0.5121	0.0063	0.0205	0.4611
0.5489	0.0020	0.0149	0.4341
0.4105	0.0011	0.0129	0.5755
0.5107	0.0048	0.0046	0.4798
0.5914	0.0396	0.0116	0.3574
0.5500	0.0071	0.0050	0.4378
0.5452	0.0171	0.0190	0.4186
0.5218	0.0257	0.0477	0.4048
0.4907	0.0046	0.1617	0.3429
0.4085	0.0047	0.0442	0.5425
0.6490	0.0143	0.0231	0.3136
0.3846	0.0101	0.0721	0.5333
0.5142	0.0218	0.0323	0.4317
0.4805	0.0504	0.0682	0.4009
0.6062	0.0520	0.0137	0.3281
0.4494	0.0251	0.0280	0.4975
0.5978	0.0162	0.0100	0.3759
0.4533	0.0070	0.0128	0.5269
0.5091	0.0075	0.0133	0.4701
0.5280	0.0314	0.0428	0.3978
0.4216	0.0040	0.0290	0.5454
0.5417	0.0066	0.0039	0.4478
0.6328	0.0029	0.0801	0.2842
0.4924	0.0146	0.0418	0.4512
0.6818	0.0126	0.0035	0.3021
0.4337	0.0131	0.0186	0.5346
0.7006	0.0065	0.0167	0.2762
0.4954	0.0032	0.0118	0.4895
0.5156	0.0059	0.0206	0.4579
0.4277	0.0006	0.0367	0.5350
0.3431	0.0073	0.0761	0.5734
0.4692	0.0057	0.0068	0.5183

```
0.4886    0.0578    0.0083    0.4453
0.5483    0.0169    0.0114    0.4234
0.3339    0.0367    0.0348    0.5946
0.3455    0.0070    0.0980    0.5495
0.4376    0.0279    0.1273    0.4072
```