Multivariate Analysis and Its Applications
IMS Lecture Notes – Monograph Series (1994) Volume 24

PRINCIPLES OF COMPOSITIONAL DATA ANALYSIS

BY JOHN AITCHISON

University of Virginia

Compositional data consisting of vectors of positive components subject to a unit-sum constraint arise in many disciplines, for example in geology as major-oxide compositions of rocks, in economics as budget share patterns of household expenditures, in medicine as compositions of renal calculi, in psychology as activity patterns of subjects. 'Standard' multivariate techniques, designed for unconstrained data, are wholly inappropriate and uninterpretable for such data and yet are still being commonly misapplied. Recognition that the study of compositions must satisfy simple principles has led recently to the advocacy of new forms of analysis of compositional data. The nature of the absurdities arising from applying traditional multivariate techniques to compositions is briefly highlighted and a description of the essential aspects and the advantages of the new methodology is provided.

1. Introduction: the Nature of Compositions. An alternative title for this paper could have been *Compositional Data Analysis is Easy*, though the history of the subject would hardly support this view. Almost a century ago Pearson (1897) warned us to beware of naive interpretations of correlations of his product-moment correlation $\mathrm{corr}(u_1, u_2)$, when u_1, u_2 are of the form $(u_1, u_2) = (x_1, x_2)/(x_1 + x_2 + x_3)$, that is when u_1, u_2 are essentially components of a composition. Statisticians and non-statisticians alike have largely disregarded the warning. A recent statistician-created disregard is in the software package Execustat Student Edition (1991) where the introductory tutorial unfortunately uses compositional data consisting of proportions of sand, silt and clay in sediments and refers to correlation coefficients of such proportions. For non-statistician examples the reader has only to browse through geological research journals abounding in arguments which depend on the interpretation of such uninterpretable correlation coefficients. For a detailed account of the sad history of compositional data see Aitchison (1986, Chapter 3).

Compositional data consisting of vectors of positive components subject

---

AMS 1980 Subject Classifications: Primary 62H, Secondary 62P.
Key words and phrases: Logratio analysis, logistic-normal distributions, perturbation, scale invariance, simplex sample space and subcomposition.

to a unit-sum constraint commonly arise in many disciplines, for example in geology as major-oxide compositions of rocks, in economics as budget share patterns of household expenditures, in medicine as compositions of renal calculi, in psychology as activity patterns of subjects. Thus a typical D-part composition is a vector $u = (u_1, \cdots, u_D)$ of D positive components or proportions $u_i\,(i = 1, \cdots, D)$ satisfying the unit-sum constraint $u_1 + \cdots + u_D = 1$. This unit-sum constraint reduces the effective dimension of D-part compositions to $d = D - 1$, and an appropriate sample space for the study of compositions is then the d-dimensional unit simplex

$$S^d = \{(u_1, \cdots, u_D) : \; u_i > 0 \; (i = 1, \cdots, D), \; u_1 + \cdots + u_D = 1\}.$$

More generally a composition may be regarded as a D-dimensional vector $x = (x_1, \cdots, x_D)$ in positive space R^D_+, where each component x_i $(i = 1, \cdots, D)$ is measured in the same units. It is then widely recognized that the particular unit chosen, for example ounces or grams, should make no difference in any statistical analysis of such specimens, nor should the size of the experimental object. Thus two compositions x and X are regarded as equivalent, $x \sim X$, if there is some $a > 0$ such that $X = ax$. We are here involved with the group of scale transformations $a : x \to ax$ from R^D_+ to R^D_+ and with equivalent compositions falling into a compositional class $c = \{ax : a > 0\}$. Any such compositional class, geometrically represented by a ray from the origin in R^D_+, can be specified by its unit-sum composition $u = x/(x_1 + \cdots + x_D)$, the intersection of the ray with the simplex S^d, where x is any composition in c. Since it should not matter which composition in a compositional class is chosen to represent the physical entity it follows that any meaningful function f of a composition must satisfy the requirement of scale invariance

$$f(ax) = f(x) \text{ for every } x \in c \text{ and for every } a > 0.$$

A maximal invariant is the set of ratios x_i/x_D $(i = 1, \cdots, d)$ and so it follows that any meaningful function of a composition must be expressible in terms of a set of such ratios or some equivalent set. This rather obvious perception about compositions seems to be a real stumbling block in some disciplines, as demonstrated by some recent published correspondence (Aitchison 1990a, 1991, 1992a; Watson 1990, 1991). In what follows we shall confine our attention to the unit-sum representations within the simplex and note that since the ratios of every pair of components in any equivalent compositions are the same $u_i/u_j = x_i/x_j$, and so any meaningful function of a unit-sum composition u is expressible in terms of ratios such as u_i/u_D $(i = 1, \cdots, d)$.

2. Subcompositional Invariance. The fact that a natural sample space for compositional data is the unit simplex S^d, not the whole of real space R^d, should have been sufficient to warn against the use in compositional data

analysis of "standard" multivariate methodology designed for the statistical analysis of unconstrained vectors. Although the difficulty of interpreting correlations of crude components has been well documented (Chayes 1948, 1960; Mosimann 1963; Sarmanov and Vistelius 1959) and usually expressed in terms of negative bias, much effort has been diverted into describing in great detail the pathological effects of standard methods applications rather than finding a methodology suitable to simplex sample spaces.

One of the principles of compositional data analysis must be a form of subcompositional coherence between scientists. Suppose that scientist A can measure all the parts of a D-part composition $(u_1, \cdots, u_D)$ but scientist B has facilities only for measuring the first C parts. Thus scientist B has available a *subcomposition* $(s_1, \cdots, s_C)$ related to the full D-part composition by $(s_1, \cdots, s_C) = (u_1, \cdots, u_C)/(u_1 + \cdots + u_C)$. A requirement of any sensible methodology is surely that any statements about the parts $1, \cdots, C$ made by A and B must be consistent. A simple illustrative example shows the folly of the use of product-moment correlation of the crude components, namely $\operatorname{corr}(u_1, u_2)$ for A and $\operatorname{corr}(s_1, s_2)$ for B, as a means of communication. For example, for the 4-part compositions (0.1, 0.2, 0.1, 0.6), (0.2, 0.1, 0.1, 0.6), (0.3, 0.3, 0.2, 0.2) and the 3-part subcompositions formed from parts 1, 2, 3, namely (0.25, 0.50, 0.25), (0.50, 0.25, 0.25), (0.375, 0.375, 0.25), we have $\operatorname{corr}(u_1, u_2) = 0.5$ and $\operatorname{corr}(s_1, s_2) = -1$.

Our knowledge that any meaningful function of a composition must be expressible in terms of ratios of components and the obvious fact that ratios are unaltered in the process of forming subcompositions $(s_i/s_j = u_i/u_j)$ lead us inevitably to consideration of some form of covariance structure for compositions based upon ratios of components. It should be noted here that subcompositions play a central role in compositional data analysis, replacing the concept of marginals in unconstrained multivariate data analysis.

3. A Covariance Structure for Compositions. A natural step towards defining some dependence structure for compositions would now seem to be the introduction of such characteristics as $\operatorname{var}(u_i/u_j)$ and $\operatorname{cov}(u_i/u_j, u_k/u_l)$. This, however, has the drawback that there is no simple relationship between, for example, $\operatorname{var}(u_i/u_j)$ and $\operatorname{var}(u_j/u_i)$, so that a large number of these characteristics would be required for a full description at this second moment level. This difficulty disappears if we consider logratios such as $\log(u_i/u_j)$ instead of ratios, for which we have simple relationships such as $\operatorname{var}\{\log(u_i/u_j)\} = \operatorname{var}\{\log(u_j/u_i)\}$. Indeed it is easy to see that in a specification of such a logratio covariance structure only the $d(d+1)/2$ logratio variances $\tau_{ij} = \operatorname{var}\{\log(u_i/u_j)\}$ $(i < j)$ are required, since the general logratio covariance is determined through the relationship $\operatorname{cov}\{\log(u_i/u_j), \log(u_k/u_l)\} = \frac{1}{2}(\tau_{il} + \tau_{jk} - \tau_{ik} - \tau_{jl})$.

If we wish for a compositional data set something equivalent to the mean

vector and the covariance matrix for a data set of unconstrained vectors then we can do no better than set out in a $D \times D$ variation array the obvious sample estimates of $E\{\log(u_i/u_j)\}$ below the diagonal and the sample estimates of $\text{var}\{\log(u_i/u_j)\}$ above the leading diagonal of the array.

4. Parametric Classes of Distributions on the Simplex. The well-known Dirichlet class of distributions on the simplex with typical density function proportional to $u^{\alpha_1 - 1} \cdots u^{\alpha_D - 1}$ is incapable of describing the vast majority of compositional variability. The main reason for this is that the Dirichlet distribution has the maximum degree of independence available to compositions. For example, every subcomposition is independent of any other non-overlapping subcomposition. To describe real variability, and to allow the investigation of hypotheses of independence, some parametric class richer in dependence structure is required. An answer to this is to be found in the old idea (McAlister, 1879) of inducing a distribution (the lognormal distribution) on an awkward space (the positive real line) from one (the normal distribution) on a more familiar space (the real line) by way of transformations (the exponential and logarithmic) between the two spaces. Our situation with awkward space S^d and familiar space R^d and its multivariate normal class is hardly more difficult than this early use of the transformation technique. Probably the simplest transformation from $y \in R^d$ to $u \in S^d$ is the 'additive' logistic transformation $u = alg(y)$, defined by

$$u_i = e^{y_i}/(e^{y_1} + \cdots + e^{y_d} + 1) \qquad i = 1, \cdots, d,$$
$$u_D = 1/(e^{y_1} + \cdots + e^{y_d} + 1)$$

with inverse transformation from S^d to R^d the logratio transformation $y = alr(u)$, defined by

$$y_i = \log(u_i/u_D), \qquad i = 1, \cdots, d.$$

There are, of course, many other possible such transformations which may be of relevance to certain aspects of compositional data analysis (Aitchison (1986a, Chapter 6)) but we shall confine attention here to the above.

5. A Methodology for Compositional Data Analysis. The considerations of Sections 3 and 4 suggest a simple methodology for compositional data analysis.

Transform each composition $(u_1, \cdots, u_D)$ to it logratio vector $y = (\log(u_1/u_D, \cdots, u_d/u_D)\}$, after reformulating your problem about compositions in terms of the corresponding logratio vectors, then apply the appropriate, standard multivariate procedures to the logratio vectors.

Since any meaningful function of a composition must always be expressible in terms of ratios, and therefore logratios, of components, the required reformulation can always be achieved. The fact that the final component is used as

divisor raises the question of whether the choice of another divisor might lead to different conclusions. It can readily be established (Aitchison 1986a, Chapter 6) that standard multivariate statistical procedures are invariant under the group of permutations of the parts $1, \cdots, D$ of the composition, in particular with respect to a common divisor x_j different from x_D.

Note that in the use of the logratio vector y above the covariance structure is being defined in terms of the covariance matrix of y with typical elements $\sigma_{ij} = \text{cov}\{\log(u_i/u_D, u_j/u_D\}$ $(i, j = 1, \cdots, d)$. The relationship of the σ_{ij} to the basic logratio variances τ_{ij} is simply obtained as $\tau_{ij} = \sigma_{ii} + \sigma_{jj} - 2\sigma_{ij}$. It is indeed possible to treat all the components symmetrically through the use of the centered logratio transformation $z = [\log\{u_i/g(u)\}, \cdots, \log\{u_D/g(u)\}]$ with $D \times D$ covariance matrix with typical element $\gamma_{ij} = \text{cov}[\log\{u_i/g(u)\}, \log\{u_j/g(u)\}]$, where $g(u) = (u_1 \cdots u_D)^{1/D}$ is the geometric mean of the components of u, with a trade-off of semipositive definiteness of the covariance matrix for the symmetry achieved. The alternative specifications $\tau_{ij}, \sigma_{ij}, \gamma_{ij}$ allow a convenient flexibility in the statistical analysis of compositional data. For example, some of the interesting definitions of forms of independence associated with compositional variability are more readily expressed in terms on one specification rather than another. Complete subcompositional independence, defined as the independence of every set of non-overlapping subcompositions, is characterized by τ_{ij} taking the additive form $\alpha_i + \alpha_j$, whereas the characterization in terms of γ_{ij} is $\delta_{ij}\alpha_i - D^{-1}(\alpha_i + \alpha_j - \alpha.)$, where δ_{ij} is the Kronecker delta and $\alpha.$ is the arithmetic mean of α_i $(i = 1, \cdots, D)$.

For details of problems special to compositions and a wide variety of applications see Aitchison (1986a). For an effective graphical display of composition of data along the lines of the familiar biplot for unconstrained data see Aitchison (1990b).

6. Perturbation : the Fundamental Operation in the Simplex. The basic operations of translation in R^D and rotation on the sphere and the role of the corresponding groups are well known in the statistical analysis of unconstrained multivariate vectors and directional data. The answers to the questions of whether there is an analogous basic operation in the simplex and whether it can be used to characterize the difference between two compositions are less familiar, but are fundamental for an understanding of compositional data analysis. The answers are to be found in the basic operation of perturbation (Aitchison, 1986a), which may be motivated in relation to the wider notions of compositional classes as described in Section 1.

For any two equivalent compositions x and X, in the same compositional class, there is a scale relationship $(X_1, \cdots, X_D) = (ax_1, \cdots, ax_D)$ for some $a > 0$, where each component of x is scaled by the same factor a to obtain the corresponding component of X. For any two compositions x and X in different compositional classes c and C a similar, but differential,

scaling relationship $(X_1,\cdots,X_D) = (p_1x_1,\cdots,p_Dx_D)$ can always be found, simply by taking $p_i = X_i/x_i (i = 1,\cdots,D)$. Denoting the operation between the positive perturbing vector $p = (p_1,\cdots,p_D)$ and the composition x by $\circ$ we have $p \circ x = (p_1x_1,\cdots,p_Dx_D)$ and $X = p \circ x$. Such a perturbation operator is then easily adapted to the simplex simply by defining $p \circ u = (p_1u_1,\cdots,p_Du_D)/(p_1u_1 + \cdots + p_Du_D)$. Note that the roles of p and u are interchangeable in this definition and we can conveniently restrict p to lie in the simplex S^d. Perturbations thus defined form a group, with p^{-1}, the inverse of p, defined as $(p_1^{-1},\cdots,p_D^{-1})/(p_1^{-1} + \cdots + p_D^{-1})$ and the identity perturbation as $(1/D,\cdots,1/D)$. Moreover, for any two compositions u, U there is a unique perturbation $p \in S^d$ such that $U = p \circ u$ and $u = p^{-1} \circ U$, where $p = U \circ u^{-1}$. Thus the perturbation $U \circ u^{-1}$, or equivalently $X \circ x^{-1}$ characterizes the change from c to C; the change from X to x is simply the inverse perturbation $u \circ U^{-1}$. The question of whether the group of perturbations is unique in this role of describing compositional change has been addressed by Aitchison (1992b) who shows that under some simple requirements the answer is in the affirmative.

When we recall the importance of the logratio vectors in compositional data analysis, perturbation is seen to be the natural operation within the simplex, since the logratio vectors y, Y corresponding to u, U are related in R^d by the translation operation $Y = r + y$, where r is the logratio vector of $p = U \circ u^{-1}$. The fact that translation is used in the statistical modeling of error or imprecision in R^d directs attention to its use for such purposes for compositional modeling. For example, when the purpose of experimentation is to combine measurements of some 'true' composition v the natural measurement model relating an observed composition u to v is $u = v \circ p$, where p is the imprecision perturbation. Similarly if we wish to relate the variability of a composition u to a covariate vector v in a manner analogous to generalized linear modeling we can express the relationship as $u = alg(\beta^T v) \circ p$ where the random perturbation p plays the role of the error and the 'link function' is the additive logistic function alg.

7. Some Further Principles of Compositional Data Analysis. We can now turn our attention to the role of the group of perturbations in determining principles for compositional data analysis. Aitchison (1992b) has discussed in detail the rationale for defining a sensible scalar measure of difference between two compositions u and U. We confine attention here to the part of the argument involving perturbations. From the above discussion the full difference between u and U is $U \circ u^{-1}$. Now if u and U are both subjected to the same perturbation p to produce new compositions $u^* = p \circ u$ and $U^* = p \circ U$ the difference between u^* and U^* is also $U \circ u^{-1}$ since $U^* = U^* \circ u^{*-1} \circ u^* = (p \circ U) \circ (p^{-1} \circ u^{-1}) \circ u^* = (U \circ u^{-1}) \circ u^*$. Thus we would require any function purporting to be a scalar measure of difference $f(u, U)$

between u and U to be perturbation invariant in the sense that

$$f(p \circ u, p \circ U) = f(u, U) \qquad \text{for every} \quad p,\ u,\ U \in S^d.$$

A maximal perturbation invariant function is $U \circ u^{-1}$ so that any perturbation invariant function must be a function of the ratios $(u_1/U_1, \cdots, u_D/U_D)$. This together with the requirement of scale invariance limits the functions available as scalar measure of difference to functions of the ratios of ratios of the form $(u_i/u_j)/(U_i/U_j)$. With some other simple requirements Aitchison (1992b) derives

$$f(u, U) = \left[\sum_{i<j} \left\{ \log (u_i/u_j) - \log (U_i/U_j) \right\}^2 \right]^{1/2}$$

as the simplest and most tractable measure of difference or distance between two compositions. If for a compositional data set a total measure of variability is taken as the sum of squares of the distances between also possible pairs of compositions it is easy to see that this is in conformity with total measures of variability based on covariance structure, for example the sum of sample estimates of all possible logratio variances τ_{ij}.

Argument concerning the suitability of arithmetic, geometric or other means as a measure of central location can be traced far back into statistical history. The tradition in compositional data analysis seems to be the arithmetic mean, but some simple considerations suggest that it is an unreliable tradition. Consider what properties are desirable in such a measure, say $cen(u)$. If we imagine the distribution of compositions perturbed by a constant perturbation p then surely a minimum requirement is that the centers of the original and the perturbed distributions should be related by the perturbation; in other words, $cen(p \circ u) = p \circ cen(u)$. This is clearly not satisfied by the arithmetic mean. On the other hand, the natural consequence of modeling the imprecision process in the form $u = v \circ p$ and the use of logratio analysis leads to the identification of v with $alg\{alr(E(y)\}$. If we use $G(u)$ =exp $\{E(\log u)\}$ to denote the geometric means vector then the central measure is simply the geometric mean vector scaled to form a unit-sum composition. In practical terms, for a compositional data set, we simply compute the geometric means g_i of each component and then take $(g_1, \cdots, g_D)/(g_1 + \cdots + g_D)$ as a center of the data set. We may finally remark that many compositional data sets are curved or concave in the naive geometry of the simplex and it is possible for the arithmetic mean to fall outside the data set and indeed be quite an atypical composition; see Aitchison (1989) for such a situation.

8. Discussion. Is the initial claim of this paper that compositional data analysis is easy justified ? The computational aspect of converting compositions to logratio vectors and applying standard multivariate techniques could

hardly be simpler, and a software package (Aitchison, 1986b) is available for the special forms of problem that arise in compositional data analysis. Nevertheless there appears to be some reluctance to change from the bad habits and meaningless consequences of ignoring the special nature of compositional data. In the unconstrained world the concept of product-moment correlation is so ingrained into statistical argument as a useful and straightforward tool for the description of dependence that it is difficult to conceive of other ways of describing dependence. For example, within logratio analysis the simplest construct of two components is the logratio variance $\tau_{ij} = \text{var}\{\log(u_i/u_j)\}$ and this can range over all non-negative values. The value zero, in which case u_i and u_j are in constant proportion, replaces the concept of 'perfect positive correlation' whereas large values, corresponding to the components departing substantially from constant proportionality, replaces the concept of 'negative correlation'.

Acknowledgement. This research was supported by National Science Foundation grant DMS-9011822.

REFERENCES

AITCHISON, J. (1986a). *The Statistical Analysis of Compositional Data.* Chapman and Hall, London.

AITCHISON, J. (1986b). *CODA: A Microcomputer Package for 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. Geol.* **21**, 787–790.

AITCHISON, J. (1990a). Comment on "Measures of Variability for Geological Data" by D.F. Watson and G.M. Philip, *Math. Geol.* **22**, 223–226.

AITCHISON, J. (1990b). Relative variation diagrams for describing patterns of compositional variability, *Math. Geol.* **22**, 487–512.

AITCHISON, J. (1991). Delusions of uniqueness and ineluctability, *Math. Geol.* **23**, 275–277.

AITCHISON, J. (1992a). A plea for precision in Mathematical Geology, *Math. Geol.* **24**, 1085–1086.

AITCHISON, J. (1992b). On criteria for measures of difference between compositions, *Math. Geol.* **24**, 365–380.

CHAYES, F. (1948). A petrographic criterion for the possible replacement origin of rocks, *Amer. J. Science* **246**, 413–429.

CHAYES, F. (1960). On correlation between variables of constant sum, *J.*

Geophys. Res. **65**, 4185–4193.

EXECUSTAT (1991). *Execustat Student Edition.* Strategy Plus inc.

MCALISTER, D. (1879). The law of the geometric mean, *Proc. Roy. Soc.* **29**, 367–376.

MOSIMANN, J. E. (1963). On the compound distribution, the multivariate β-distribution and correlations among proportions, *Biometrika* **49**, 65–82.

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 measurement of organs, *Proc. Roy. Soc. Lond.* **60**, 489–498.

SARMANOV, O. V. and VISTELIUS, A. B. (1959). On the correlation of percentage values, *Dokl. Akad. Nauk. SSSR* **126**, 22–25.

WATSON, D. F. (1990). Reply to Comment on "Measures of Variability for Geological Data" by D.F. Watson and G.M. Philip, *Math. Geol.* **22**, 227–231.

WATSON, D. F. (1991). Reply to "Delusions of Uniqueness and Ineluctability" by J. Aitchison, *Math. Geol.* **23**, 279.

DIVISION OF STATISTICS
HALSEY HALL
UNIVERSITY OF VIRGINIA
CHARLOTTESVILLE
VA 22901, USA