Mathematical Geology, Vol. 32, No. 3, 2000

Logratio Analysis and Compositional Distance[1]

J. Aitchison,[2] C. Barceló-Vidal,[3] J. A. Martín-Fernández,[3] and V. Pawlowsky-Glahn[4]

The concept of distance between two compositions is important in the statistical analysis of compositional data, particularly in such activities as cluster analysis and multidimensional scaling. This paper exposes the fallacies in a recent criticism of logratio-based distance measures—in particular, the misstatements that logratio methods destroy distance structures and are denominator dependent. Emphasis is on ensuring that compositional data analysis involving distance concepts satisfies certain logically necessary invariance conditions. Logratio analysis and its associated distance measures satisfy these conditions.

KEY WORDS: Euclidean distance, permutation invariance, perturbation invariance, scale invariance, subcompositional dominance.

INTRODUCTION

In a paper presented at IAMG98 Zier and Rehder (1998) assert that what has come to be known as logratio analysis of compositional data is flawed when any concept of distance between compositions is involved. Their conclusions on page 558 contain such strong condemnations of logratio analysis as:

> The distance structure is entirely destroyed, even the ranks of the distances are not equal in S^2 and R^2. This implies that methods of distance statistics do not work properly (e.g., cluster analysis, MDS).
>
> There is strong dependence on the denominator, which can lead to strange results, what does a negative distance imply?

Since such censure may well deter researchers with compositional data problems from considering the correct use of appropriate logratio analysis and since the

[1] Received 2 November 1998; accepted 2 February 1999.

[2] Department of Statistics, University of Glasgow, Glasgow, G12 8QQ, UK. e-mail: John.Aitchison@btinternet.com

[3] Universitat de Girona, Escola Politècnica Superior, Dept. d'Informàtica i Matemàtica Aplicada, Avda. Lluis Santaló, s/n, 17971 Girona, Spain. e-mail: barcelo@ima.udg.es or jamf@ima.udg.es

[4] Universitat Politècnica de Catalunya, E. T. S. d'Eng. de Camins, Canals i Ports, Dept. de Matemàtica Aplicada III, Jordi Girona, 1–3, E-08034, Barcelona, Spain. e-mail: pawlowsky@etseccpb.upc.es

arguments in Zier and Rehder (1998) are based on false premises and contain both logical and mathematical fallacies, it seems important to make a clear restatement of the nature of distance measures in relation to compositions and the appropriateness of sensible logratio measures in particular.

THE IRRELEVANCE OF EUCLIDEAN DISTANCE

First, it is necessary to expose the flaws in the Zier and Rehder (1998) argument. This largely arises from their consideration of Euclidean distance as the only feasible distance or metric in either the simplex S^d or real space R^d. Thus their distance $\delta_S(x, X)$ between two D-part compositions x and X in the simplex S^d, where $d = D - 1$, is defined by

$$\delta_S(x, X) = \left[\sum_{i=1}^{D}(x_i - X_i)^2\right]^{1/2}$$

This is an unfortunate start since it offends several simple principles of compositional data analysis, such as scale invariance, perturbation invariance, and subcompositional dominance (Aitchison, 1992). For example, with this Euclidean definition the distance between the two 3-part compositions (0.65, 0.30, 0.05) and (0.20, 0.70, 0.10) is 0.604, less than the distance 0.653 between the associated two-part subcompositions (0.684, 0.316) and (0.222, 0.778), obtained from the first two components. This is an obvious absurdity, since the distance between full compositions, based on a larger amount of information, should be at least as great as the distance between any of their subcompositions (Aitchison, 1992).

There is a similar fixation on Euclidean distance when Zier and Rehder (1998) transform from the simplex S^d to real space R^d using the logratio transformations

$$y_i = \log(x_i/x_D) \quad \text{and} \quad Y_i = \log(X_i/X_D) \quad (i = 1, \ldots, d) \quad \text{(T)}$$

and adopt as distance between logratios y and Y:

$$\delta_R(y, Y) = \left[\sum_{i=1}^{d}(y_i - Y_i)^2\right]^{1/2} = \left[\sum_{i=1}^{d}\{\log(x_i/x_D) - \log(X_i/X_D)\}^2\right]^{1/2}$$

With this insistence on Euclidean distances in both S^d and R^d, it is not surprising that conclusions of “complete destruction” (lack of isometry) and denominator dependence (lack of permutation invariance) arise. These absurdities derive from a misunderstanding of the logratio method and relevant non-Euclidean measures of distance for compositional data analysis. Even within these inappropriate definitions their production of a negative distance results from a mathematical error. In the definitions of δ_S and δ_R the positive square root is intended and so how can a negative distance arise? In their example the final stage involves the following step:

$$\sqrt{\left(\log\frac{0.01-X}{0.01}\right)^2}=\log\frac{0.01-X}{0.01}$$

Since $(0.01 - X)/0.01 < 1$ the right hand side is clearly the negative square root; the correct step would have led to the positive distance $\log\{0.01/(0.01 - X)\}$. For interesting studies of the failure of some other suggested measures of compositional difference, see Martín (1996) and Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn (1998).

COMPOSITIONAL METRICS

All the above misconceptions are surprising since a fuller study of the literature would have provided sensible definitions of compositional distance. Even in the quoted reference (Aitchison, 1986) there is a clear definition on page 192, as well as in Aitchison (1992, 1994, 1997) and Martín (1996). Indeed, Aitchison (1992) provides a detailed discussion of criteria that any definition of distance must satisfy to provide a meaningful tool of compositional data analysis. There are two equivalent forms for a distance measure in the simplex space set out in these references:

$$\Delta_S(x,X)=\left[\sum_{i=1}^{D}\left\{\log\frac{x_i}{g(x)}-\log\frac{X_i}{g(X)}\right\}^2\right]^{1/2} \tag{1}$$

$$=\left[\frac{1}{D}\sum_{i<j}\left\{\log\frac{x_i}{x_j}-\log\frac{X_i}{X_j}\right\}^2\right]^{1/2} \tag{2}$$

where $g(x)$ denotes the geometric mean $(x_1\ldots x_D)^{1/D}$. In mathematical terms $\Delta_S(x, X)$ defines a metric on the simplex sample space and has all the necessary properties of scale invariance, permutation invariance, perturbation invariance, and subcompositional dominance as set out in Aitchison (1992) required for applications in compositional data analysis.

There is, of course, no need to move from the simplex to consider differences between compositions and applications such as cluster analysis or multidimensional scaling, but any analyst who so wishes may use the logratio transformation (T) to move from the simplex S^d to real space R^d and then use the distance measure based on the relevant quadratic form

$$\Delta_R(y,Y)=[(y-Y)^T H^{-1}(y-Y)]^{1/2} \tag{3}$$

where $H = [h_{ij}]$, with $h_{ij} = 2\ (i = j), h_{ij} = 1\ (i \neq j)$, as in Aitchison (1986, Section 4.7). Since Eq. (3) is equivalent to (1) and (2), the analyst can be assured that the transformation defines an isometry, with none of the problems attaching to the misplaced ideas of distance in Zier and Rehder (1998). The role of the H matrix here in "neutralizing" the choice of denominator in the logratio transformation is analogous to its use in principal component analysis of compositional data with the logratio covariance matrix (Aitchison, 1983, 1986, Section 8.3).

It should be emphasized that the metric or distance (1) or (2) is defined on the open simplex, so that compositions with zero components are excluded. It is also important not to imagine that overfamiliar ideas of geometry in R^d carry over into the simplex. In the ternary diagram, compositions equidistant from a fixed composition do not lie on a circle. Curves of constant logcontrast values often replace the Euclidean notion of straight line. When one of the components of a composition tends toward zero, then the distance of that composition from others will tend toward infinity.

There is nothing surprising about this feature; it is merely recognizing that a composition with one of the parts absent may be chemically, physically, or biologically completely different from compositions with all components positive. Doveton's (1998) perfect martini with its (gin, dry martini, sweet martini) composition is completely different from a cocktail with no gin but only dry and sweet martini present.

DISCUSSION

In developing a statistical methodology for compositional data analysis, Aitchison (1986) was aware that in presenting the ideas in terms of the logratio transformation (T) the obvious question of dependence of the inferences on the choice of divisor would arise. Great care was therefore taken to ensure that all the statistical procedures such as loglinear modeling in Aitchison (1986) and elsewhere were invariant under the group of permutations. Some of these essentially involve another measure of distance—namely the well-known Mahalanobis distance in its logratio forms. Techniques with the distances or metrics (1)–(3) are no exception in possessing this property of invariance under the group of permutations.

In their paper Zier and Rehder (1998) place their discussion against the background of grain size analysis. Whether such data sets should be considered as grouped data on some univariate distribution or as compositional data, possibly with zeros, or as a combination of the two, is another issue. All that this note seeks to address is to provide a rejoinder to their inappropriate ideas of compositional distance.

REFERENCES

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

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

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

Aitchison, J., 1994, Principles of compositional data analysis, *in* T. W. Anderson, I. Olkin, and K. T. Fang, 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* V. Pawlowsky-Glahn, ed., Proceedings of IAMG97, The Third Annual Conference of the International Association for Mathematical Geology: Universitat Politècnica de Catalunya, Barcelona, p. 3–35.

Doveton, J. H., 1998, Beyond the perfect martini: Teaching the mathematics of petrological logs, *in* A. Buccianti, G. Nardi, and R. Potenza, Proceedings of IAMG98, The Fourth Annual Conference of the International Association for Mathematical Geology: De Frede, Naples, p. 71–75.

Martín, M. C., 1996, Performance of eight dissimilarity coefficients to cluster a compositional data set, *in* Abstracts of IFCS-96, Fifth Conference of the International Federation of Classification Societies, Kobe, Japan, Abstracts, v. 1, p. 215–217.

Martín-Fernández, J. A., Barceló-Vidal, C., and Pawlowsky-Glahn, V., 1998, Measures of difference for compositional data and hierarchical clustering methods, *in* A. Buccianti, G. Nardi, and R. Potenza, eds., Proceedings of IAMG’98, The Fourth Annual Conference of the International Association for Mathematical Geology: De Frede, Naples, p. 526–531.

Zier, U., and Rehder, S., 1998, Grain-size analysis—A closed data problem, *in* A. Buccianti, G. Nardi, and R. Potenza, Proceedings of IAMG98, The Fourth Annual Conference of the International Association for Mathematical Geology: De Frede, Naples, p. 555–558.