MEASURES OF DIFFERENCE FOR COMPOSITIONAL DATA AND HIERARCHICAL CLUSTERING METHODS

J. A. Martín-Fernández[1], C. Barceló-Vidal[1] and V. Pawlowsky-Glahn[2]

[1] Universitat de Girona
Escola Politècnica Superior
Dept. d'Informàtica i Matem. Aplicada
Avda. Lluis Santaló, s/n, 17071 Girona
SPAIN

[2] Universitat Politècnica de Catalunya
E.T.S. de Eng. de Camins, Canals i Ports
Dept. de Matemàtica Aplicada III
E-08034 Barcelona
SPAIN

ABSTRACT

For the application of a hierarchic classification method it is necessary to establish the measure of difference to be used, as well the appropriate measures of central tendency and dispersion in accordance with the nature of the data. In this study we review the requirements for the measure of difference when the data set is compositional and we present specific measures of central tendency and dispersion to be used with hierarchical clustering methods.

1. INTRODUCTION

Any vector $\mathbf{x} = (x_1, x_2, ..., x_D)$ representing proportions of some whole is subject to the unit-sum-constraint $x_1 + x_2 + ... + x_D = 1$. Therefore, a suitable sample space for compositional data, consisting of such vectors of proportions (compositions), is the unit simplex $\mathbf{S}^D$ (see [1] for further details).

Frequently, some form of statistical analysis is essential for the adequate analysis and interpretation of the data. Nevertheless, all too often the unit-sum-constraint is either ignored or improperly incorporated into the statistical modeling giving rise to an erroneous or irrelevant analysis. The purpose of this paper is to revise the specific statistical requirements of standard hierarchic agglomerative classification methods when they are performed on compositional data. We place emphasis on the measure of difference between two compositions and the measures of central tendency and dispersion of a data set.

In the next section we analyze some possible distances and dissimilarities between two compositions. For these measures, the requirements proposed by Aitchison [2] to define a measure of difference between two compositions are considered, and an example is presented to illustrate their performance. Then, we propose a modification of the most standard hierarchic agglomerative classification methods to make them suitable for the classification of a compositional data set.

2. MEASURES OF DIFFERENCE BETWEEN TWO COMPOSITIONS

Some of the most usual dissimilarities and distances to measure the difference between two compositions are listed in Table 1. The performance of some of these measures and others dissimilarities is analysed in [8]. In the formula of Aitchison's distance the divisor $g(\mathbf{x}) = \left(\prod_{k=1}^{D} x_k\right)^{1/D}$ represents the geometric mean of the composition $\mathbf{x}$. This

distance is equivalent to the Euclidean distance between the transformed compositions by the centered logratio function *clr*. The angular distance gives the angle between two compositions, i.e. between their projection in the unit hypersphere. This measure was proposed by Watson and Philip in [11]. The Bhattacharyya (arccos) distance between two compositions $\mathbf{x}_i$ and $\mathbf{x}_j$ can be interpreted as the angle between the unit vectors $\sqrt{\mathbf{x}_i}$ and $\sqrt{\mathbf{x}_j}$. This distance is directly related to the Matusita distance, which is also known as the *Hellinger distance*. The Bhattacharyya (arccos) and Matusita distances, and the Bhattacharyya (log) and J-divergence dissimilarities can be considered particular cases of a more general class of dissimilarities commonly called *Jeffreys divergences*. More specific information can be found in [6].

The Mahalanobis distances - crude and clr - between any two compositions are referred to a compositional data set $\mathbf{X}$. The matrix $\mathbf{K}^+$, which appears in Mahalanobis (crude) distance, symbolizes the Moore-Penrose pseudo-inverse of the covariance matrix $\mathbf{K}$ of the compositional data set $\mathbf{X}$. Equally, the Mahalanobis (clr) distance uses the Moore-Penrose pseudo-inverse $\mathbf{\Gamma}^+$ of the covariance matrix $\mathbf{\Gamma}$ of the transformed data set $clr(\mathbf{X})$.

In [2] Aitchison proposes that any scalar measure of difference between two compositions should verify four specific requirements: scale invariance, permutation invariance, perturbation invariance and subcompositional dominance. The scale invariance is not an essential requirement if it is implicitly assumed that any scalar measure is always applied to compositional observations previously normalized to one. The permutation invariance is a logic requirement which is satisfied by all the measures of Table 1. The perturbation invariance requirement plays the same role as the translation invariance requirement plays in the Euclidean space. Similarly, the subcompositional dominance requirement is in correspondence with the subspace dominance of the Euclidean distance.

Table 2 summarizes which of these requirements are verified by the distances and dissimilarities of Table 1. The proof, that Aitchison's and the Mahalanobis (clr) distances accomplish the four requirements can be found in [2], [4] and [5].

A simple example will serve to illustrate this assertions; at the same time, it can be considered as a counterexample for those cases where the properties are not fulfilled.

Let $\mathbf{X}$ the compositional data set formed by the four observations in $\mathbf{S}^3$:

$$\mathbf{x}_1 = (0.1, 0.2, 0.7), \quad \mathbf{x}_2 = (0.2, 0.1, 0.7), \quad \mathbf{x}_3 = (0.3, 0.4, 0.3) \quad \text{and} \quad \mathbf{x}_4 = (0.4, 0.3, 0.3).$$

We symbolize by $\mathbf{x}_i^*$ the perturbed composition $\mathbf{p} \circ \mathbf{x}_i$, where $\mathbf{p} = (0.8, 0.1, 0.1)$. Similarly, $\mathbf{s}_i$ symbolizes the subcomposition of the observation $\mathbf{x}_i$ formed by the first two components. Figure 1 shows the location of these elements on the ternary diagram and Table 3 summarizes the values of the distances and dissimilarities of Table 1 between some of these compositions in $\mathbf{S}^3$.

The results in Table 3 confirm that only the distances of Aitchison and Mahalanobis (clr) verify all the requirements.

This example is also intended to convince sceptic people that it is not reasonable to apply Euclidean thinking to measure the difference between two compositions. Certainly, the translation $\mathbf{t} = (0.2, 0.2, -0.4)$ transforms the observation $\mathbf{x}_1$ into $\mathbf{x}_3$ and the observation $\mathbf{x}_2$ into $\mathbf{x}_4$, i.e., $\mathbf{x}_1 + \mathbf{t} = \mathbf{x}_3$ and $\mathbf{x}_2 + \mathbf{t} = \mathbf{x}_4$. This fact implies that the Minkowski's, City Block and Euclidean distance between $\mathbf{x}_1$ and $\mathbf{x}_2$ is the same as between $\mathbf{x}_3$ and $\mathbf{x}_4$, because these measures of difference are translation invariant. However, from a compositional point of view, the difference between $\mathbf{x}_1$ and $\mathbf{x}_2$ must be greater than the difference

Table 1: Some measures of difference between two compositions

Distance/Dissimilarity	$d(\mathbf{x}_i, \mathbf{x}_j)$
Aitchison	$\left[\sum_{k=1}^{D} (log(\frac{x_{ik}}{g(\mathbf{x}_i)}) - log(\frac{x_{jk}}{g(\mathbf{x}_j)}))^2\right]^{\frac{1}{2}}$
Angular	$arccos\left(\sum_{k=1}^{D} \sqrt{\frac{x_{ik}^2}{\sum x_{ik}^2}}\sqrt{\frac{x_{jk}^2}{\sum x_{jk}^2}}\right)$
Bhattacharyya (arccos)	$arccos\left(\sum_{k=1}^{D} \sqrt{x_{ik}}\sqrt{x_{jk}}\right)$
Bhattacharyya (log)	$-log\left(\sum_{k=1}^{D} \sqrt{x_{ik}}\sqrt{x_{jk}}\right)$
City Block	$\sum_{k=1}^{D} \lvert x_{ik} - x_{jk}\rvert$
J-divergence	$\left[\sum_{k=1}^{D} (log(x_{ik}) - log(x_{jk}))(x_{ik} - x_{jk})\right]^{\frac{1}{2}}$
Euclidean	$\left[\sum_{k=1}^{D} (x_{ik} - x_{jk})^2\right]^{\frac{1}{2}}$
Mahalanobis (crude)	$[(\mathbf{x}_i - \mathbf{x}_j)'\mathbf{K}^+(\mathbf{x}_i - \mathbf{x}_j)]^{\frac{1}{2}}$
Mahalanobis (clr)	$\left[(clr(\mathbf{x}_i) - clr(\mathbf{x}_j))'\mathbf{\Gamma}^+(clr(\mathbf{x}_i) - clr(\mathbf{x}_j))\right]^{\frac{1}{2}}$
Matusita	$\left[\sum_{k=1}^{D} (\sqrt{x_{ik}} - \sqrt{x_{jk}})^2\right]^{\frac{1}{2}}$
Minkowski	$\left[\sum_{k=1}^{D} (x_{ik} - x_{jk})^p\right]^{\frac{1}{p}}$

Table 2: Aitchison's requirements verified by the measures of Table 1

Distance/Dissimilarity	Scale invariance	Permutation invariance	Perturbation invariance	Subcompositional dominance
Aitchison	Yes	Yes	Yes	Yes
Angular	Yes	Yes	No	No
Bhattacharyya (arccos)	No	Yes	No	No
Bhattacharyya (log)	No	Yes	No	No
City Block	No	Yes	No	No
J-divergence	No	Yes	No	No
Euclidean	No	Yes	No	No
Mahalanobis (crude)	Yes	Yes	No	No
Mahalanobis (clr)	Yes	Yes	Yes	Yes
Matusita	No	Yes	No	No
Minkowski	No	Yes	No	No

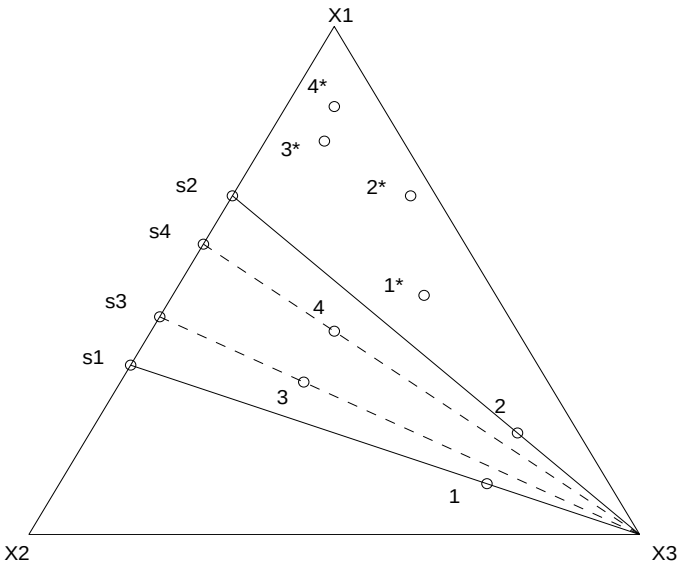

Figure 1: The four observations 1 to 4, their subcompositions s_1 to s_4 and perturbations 1^* to 4^* on the ternary diagram

Table 3: Distances and dissimilarities of Table 1 between some compositions of Figure 1

Distance/Dissimilarity	$d(\mathbf{x}_1, \mathbf{x}_2)$	$d(\mathbf{x}_1^*, \mathbf{x}_2^*)$	$d(\mathbf{s}_1, \mathbf{s}_2)$	$d(\mathbf{x}_3, \mathbf{x}_4)$	$d(\mathbf{x}_3^*, \mathbf{x}_4^*)$	$d(\mathbf{s}_3, \mathbf{s}_4)$
Aitchison	0.98	0.98	0.98	0.41	0.41	0.41
Angular	0.19	0.33	0.64	0.24	0.08	0.28
Bhattacharyya (arccos)	0.19	0.22	0.34	0.12	0.09	0.14
Bhattacharyya (log)	0.02	0.02	0.06	0.01	0	0.01
City Block	0.2	0.4	0.67	0.2	0.14	0.29
J-divergence	0.37	0.43	0.68	0.24	0.18	0.29
Euclidean	0.14	0.24	0.47	0.14	0.09	0.2
Mahalanobis (crude)	3	4.46	5.07	3	1.63	0.93
Mahalanobis (clr)	5.12	5.12	5.12	0.88	0.88	0.88
Matusita	0.19	0.22	0.34	0.12	0.09	0.14
Minkowski (p=3)	0.13	0.21	0.42	0.13	0.08	0.18

between $\mathbf{x}_3$ and $\mathbf{x}_4$. Observe that $\mathbf{x}_1$ and $\mathbf{x}_2$ only differ in ± 0.1 in the two first components, and the same occurs to $\mathbf{x}_3$ and $\mathbf{x}_4$. But in the first case, the difference ± 0.1 is produced over a residual of $0.3\ (= 1 - 0.7)$, whereas in the second case the same difference ± 0.1 is over a residual of $0.7\ (= 1 - 0.3)$. This argument is equivalent to compare the corresponding subcompositions

$$\mathbf{s}_1 = (\frac{1}{3}, \frac{2}{3}), \quad \mathbf{s}_2 = (\frac{2}{3}, \frac{1}{3}), \quad \mathbf{s}_3 = (\frac{3}{7}, \frac{4}{7}) \quad \text{and} \quad \mathbf{s}_4 = (\frac{4}{7}, \frac{3}{7}),$$

which are also plotted in Figure 1. From this graphic it is clear that the difference between subcompositions $\mathbf{s}_1$ and $\mathbf{s}_2$ is greater than the difference between subcompositions $\mathbf{s}_3$ and $\mathbf{s}_4$. It is also important to point out from this example that the Angular distance doesn't have a compositional coherent behavior, because the Angular distance between $\mathbf{x}_3$ and $\mathbf{x}_4$ results greater than the distance between $\mathbf{x}_1$ and $\mathbf{x}_2$.

3. HIERARCHIC CLUSTER ANALYSIS OF COMPOSITIONAL DATA

Before applying any hierarchic method of classification to a data set $\mathbf{X}$, it is necessary to establish in advance which are the measures of difference, central tendency and dispersion, to be used in accordance with the nature of data to be classified (see examples in [9],[10]). Thus, if we are using a hierarchic method to classify a compositional data set, we have to use an appropriate measure of difference, like Aitchison's or the Mahalanobis (clr) distances.

Consequently, to calculate the matrix of differences associated to hierarchic methods like single linkage, complete linkage and average linkage, when they are applied to a compositional data set, only Aitchison's distance of Table 1 will be suitable. It is not appropriate to use any one of the other measures of differences recorded in Table 1.

Likewise, any method of classification which reduces the measure of difference from a composition to a cluster $\mathbf{C}$ of compositions to the difference between the composition and the 'center' of the group, would have to take into account that the arithmetic mean $\overline{\mathbf{C}}$ of the data set is usually not representative of the 'center' of the set, and neither is compatible with the group of perturbations. Aitchison [3] proposes the geometric mean $cen(\mathbf{C})$ as a more representative point of the central tendency of a compositional data set $\mathbf{C}$ in $\mathbf{S}^D$. It is defined as

$$cen(\mathbf{C}) = \frac{(g_1, g_2, ..., g_D)}{g_1 + g_2 + ... + g_D} \quad , \tag{1}$$

where $g_j = \left(\prod_{i=1}^{N} x_{ij}\right)^{1/N}$ is the geometric mean of the jth component of the compositions $\mathbf{x}_1$, $\mathbf{x}_2$,..., $\mathbf{x}_N$ in $\mathbf{C}$. Thus, we recommend to use (1) as a definition of the 'center' of a set of compositions, in addition to Aitchison's distance.

On the other hand, the well-known method of Ward is a hierarchic method which uses the measure of dispersion to classify the observations of a data set. In essence, this method is based on the concept of variability on a cluster $\mathbf{C}$. This variability is defined (see [7], Section 5.2) as follows

$$\sum_{\mathbf{x}\in\mathbf{C}} d_{eu}^2\left(\mathbf{x}, \overline{\mathbf{C}}\right), \tag{2}$$

where $\overline{\mathbf{C}}$ denotes the center of the group. When the data set is compositional, we suggest replacing this measure by

$$\sum_{\mathbf{x}\in\mathbf{C}} d_{at}^2\left(\mathbf{x}, cen(\mathbf{C})\right), \tag{3}$$

where d_{at} symbolizes Aitchison's distance. This measure is equivalent to the measure of total variability of a compositional data set proposed by Aitchison in [2] and [3].

The above adaptations are introduced to make the standard hierarchic clustering methods compatible with the compositional nature of a data set $\mathbf{X}$. All these adaptations can be omitted if these methods are directly applied to the transformed data set $clr(\mathbf{X})$. This equivalence is discussed in detail in [10].

4. CONCLUSIONS

- When the standard hierarchic classification methods are applied to compositional data sets, they should be adapted to take into account the nature of the data.
- The distance of Aitchison between two compositions and the center and variability of a compositional data set defined in (1) and (3) are compatible with the compositional nature of the data. We suggest performing the usual hierarchic methods of classification using these measures. This is equivalent to the application of standard clustering methods to the centered logratio transformed data set.
- It is necessary to study more deeply the performance of other usual non-parametric and parametric classification methods when they are applied to compositional data.

REFERENCES

1. AITCHISON, J. - *"The Statistical Analysis of Compositional Data"*. Chapman and Hall, New York (USA), 416 p., 1986.
2. AITCHISON, J.- On Criteria for Measures of Compositional Difference, *Math. Geology*, vol. 24, No. 4, pp. 365-379, (1992).
3. AITCHISON, J.- 'The one-hour course in compositional data analysis or compositional data analysis is simple', in: Proceedings of IAMG'97 . The 1997 Annual Conference of the International Association for Mathematical Geology, Ed. Pawlowsky-Glahn, V., CIMNE, Barcelona (E), Part I, 1997, pp. 3-35.
4. BARCELÓ-VIDAL, C. - *"Mixtures of Compositional Data"*. Universitat Politècnica de Catalunya, Barcelona (E), Ph. D. thesis, 1996.
5. BARCELÓ-VIDAL, C., MARTÍN-FERNÁNDEZ, J. A. and PAWLOWSKY-GLAHN, V.- Comment on "Singularity and Nonnormality in the Classification of Compositional Data" by G.C. Bohling et al. (submitted to *Math. Geology*), (1998).
6. BURBEA, J. - 'J-Divergences and Related Concepts', in: Encyclopedia of Statistical Sciences. John Wiley and Sons, New York (USA), Vol. 4, 1983, pp. 290-296.
7. EVERITT, B. S.- *"Cluster Analysis"*. Edward Arnold, Cambridge (UK), 170 p., 1993.
8. MARTÍN, M. C. - 'Performance of Eight Dissimilarity Coefficients to Cluster a Compositional Data Set' ,in Abstracts of IFCS-96. Fifth Conference of International Federation of Classification Societies, Kobe, Japan, Abstracts, Vol. 1, 1996, pp. 215-217.
9. MARTÍN-FERNÁNDEZ, J. A., BARCELÓ-VIDAL, C. and PAWLOWSKY-GLAHN, V. - 'Different Classifications of the Darss Sill Data Set Based on Mixture Models for Compositional Data', in Proceedings of IAMG'97. The 1997 Annual Conference of the International Association for Mathematical Geology, Ed. Pawlowsky-Glahn, V., CIMNE, Barcelona (E), Part I, 1997, pp. 151-156.
10. MARTÍN-FERNÁNDEZ, J. A., BARCELÓ-VIDAL, C. and PAWLOWSKY-GLAHN, V. - 'A Critical Approach to Non-parametric Classification of Compositional Data', in: Proceedings of IFCS'98. The Sixth Conference of the International Federation of Classification Societies.(1998, in press).
11. WATSON,D.F. and PHILIP, G.M. - Measures of Variability for Geological Data, *Math. Geology*, vol. 21, 2, pp. 233-254, (1989).