A Critical Approach to Non-Parametric Classification of Compositional Data

J. A. Martín-Fernández[1], C. Barceló-Vidal[1], V. Pawlowsky-Glahn[2]
[1] Dept. d'Informàtica i Matemàtica Aplicada, Escola Politècnica Superior, Universitat de Girona, Lluis Santaló, s/n, E-17071 Girona, Spain
[2] Dept. de Matemàtica Aplicada III, ETS de Eng. de Camins, Canals i Ports, Universitat Politècnica de Catalunya, Jordi Girona Salgado, 1 i 3, E-08034 Barcelona, Spain

Abstract: The application of hierarchic methods of classification needs to establish in advance some or all of the following measures: difference, central tendency and dispersion, in accordance with the nature of the data. In this work, we present the requirements for these measures when the data set to classify is a compositional data set. Specific measures of difference, central tendency and dispersion are defined to be used with the most usual non-parametric methods of classification.

Key words: Compositional Data, Cluster analysis, Classification.

1. Introduction

Any vector $\mathbf{x} = (x_1,...,x_D)$ with non-negative elements $x_1,...,x_D$ representing proportions of some whole is subject to the unit-sum-constraint $x_1+...+x_D = 1$. Compositional data, consisting of such vectors of proportions (compositions), play an important role in many disciplines. 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 modelling 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.

In the next section we present the ideas proposed by Aitchison (1992) about the conditions that have to be satisfied by any distance between two compositions, and by any measure of central tendency and dispersion of a compositional data set. Next, we propose a modification of the most standard hierarchic agglomerative classification methods to make them suitable for the classification of a compositional data set. Finally, we present two examples where the proposed methodology is applied to simulated compositional data sets.

2. Statistical analysis of compositional data

If a vector $\mathbf{w} = (w_1,\ldots,w_D) \in \Re^D$ with non-negative components is compositional, we are implicitly recognising that the total size $w_1+\ldots+w_D$ of the composition is irrelevant. Therefore, a suitable sample space for compositional data is the unit simplex

$$\mathbf{S}^{D-1} = \left\{(x_1,\ldots,x_D) : x_j > 0\ (j = 1,\ldots,D),\ x_1+\ldots+x_D = 1\right\},$$

and any meaningful function f of a composition must be invariant under the group of scale transformations; i.e. $f(\lambda\mathbf{w}) = f(\mathbf{w})$, for every $\lambda>0$. Note that only functions expressed in terms of ratios of the components of the composition satisfy this condition (Aitchison, 1992).

As an analogy to the role played by the group of translations when the sample space is the real space $\Re^D$, Aitchison (1986, Section 2.8) introduces the group of perturbations as a means to characterize the `difference' between two compositions. If we denote the perturbation operation by `$\circ$', then the perturbation $\mathbf{p} = (p_1,\ldots,p_D) \in \mathbf{S}^{D-1}$ applied to a composition $\mathbf{x}$ produces the new composition

$$\mathbf{p} \circ \mathbf{x} = (p_1x_1,\ldots,p_Dx_D) \Big/ \sum_j p_j\, x_j\ .$$

If $\mathbf{x}, \mathbf{x}^* \in \mathbf{S}^{D-1}$ are two compositions it is easy to prove that the perturbation

$$\mathbf{x}^* \circ \mathbf{x}^{-1} = (x^*_1/x_1,\ldots,\ x^*_D/x_D) \Big/ \sum_j x^*_j/x_j\ ,$$

moves $\mathbf{x}$ to $\mathbf{x}^*$.

2.a Distance between two compositions

The requirements which any scalar measure of distance between two compositions should verify, according to the definitions given by Aitchison (1992), are scale invariance, permutation invariance, subcompositional dominance and perturbation invariance. These requirements are sensible, as they acknowledge the compositional nature of the data. A feasible distance between two compositions $\mathbf{x}, \mathbf{x}^* \in \mathbf{S}^{D-1}$ is given by

$$\Delta(\mathbf{x},\mathbf{x}^*) = \left[D^{-1} \sum_{j<k} \left\{ \log\left(\frac{x_j}{x_k}\right) - \log\left(\frac{x^*_j}{x^*_k}\right) \right\}^2 \right]^{\frac{1}{2}}, \qquad (2)$$

which is equivalent to the distance proposed by Aitchison (1992). Using the definition of *centred logratio transformation* (clr) from $\mathbf{S}^{D-1}$ to $\Re^D$ given by

$$\mathrm{clr}(\mathbf{x}) = \left(\log\left(\frac{x_1}{g(\mathbf{x})} \right), \ldots, \log\left(\frac{x_D}{g(\mathbf{x})} \right) \right), \tag{3}$$

where $g(\mathbf{x})$ is the geometric mean of the composition $\mathbf{x}$, it is easy to establish that

$$\Delta(\mathbf{x}, \mathbf{x}^*) = d_{eu}(\mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{x}^*)), \tag{4}$$

where d_{eu} represents the euclidean distance. Moreover, since $\mathrm{clr}(\mathbf{p} \circ \mathbf{x}) = \mathrm{clr}(\mathbf{p}) + \mathrm{clr}(\mathbf{x})$, for any $\mathbf{p}, \mathbf{x} \in \mathbf{S}^{D-1}$, it is clear that the distance (4) is perturbation invariant.

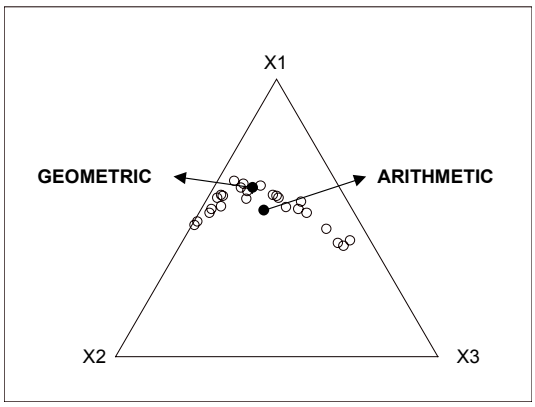
2.b Measure of central tendency of a compositional data set

If $\mathbf{X} = \left\{ \mathbf{x}_i = (x_{i1}, \ldots, x_{iD}) \in \mathbf{S}^{D\text{-}\mathbf{1}} : i = 1, \ldots, N \right\}$ represents a set of compositions, the arithmetic mean $\overline{\mathbf{X}}$ of the data set is usually not representative of the 'centre' of the set, and neither is compatible with the group of perturbations. Aitchison (1997) proposes the geometric mean $\mathrm{cen}(\mathbf{X})$ as more representative of the central tendency of a compositional data set. It is defined as

$$\mathrm{cen}(\mathbf{X}) = \frac{(g_1, \ldots, g_D)}{g_1 + \ldots + g_D}, \tag{5}$$

where $g_j = \left(\prod_{i=1}^{N} x_{ij} \right)^{\frac{1}{N}}$ is the geometric mean of the jth component of the data set. Figure 1 shows the ternary diagram of a simulated data set (adapted from Aitchison, 1986, data set 1) where it can be observed that the geometric mean lies inside the bulk of data, while the arithmetic mean lies slightly apart.

Figure 1: *Data set in the simplex (geometric mean = (0.6058, 0.2719, 0.1223) and arithmetic mean = (0.54, 0.2756, 0.1844))*



It is easy to prove that $\text{cen}(\mathbf{p} \circ \mathbf{X}) = \mathbf{p} \circ \text{cen}(\mathbf{X})$ for any perturbation $\mathbf{p} \in \mathbf{S}^{D-1}$, and that $\text{clr}(\text{cen}(\mathbf{X})) = \overline{\text{clr}(\mathbf{X})}$. Therefore, it will be true that

$$\Delta(\mathbf{x}, \text{cen}(\mathbf{X})) = d_{eu}(\text{clr}(\mathbf{x}), \overline{\text{clr}(\mathbf{X})}). \tag{6}$$

2.c Measure of dispersion of a compositional data set

It is sensible to assume that any measure of dispersion of a compositional data set should be invariant under the group of perturbations. The measure of dispersion defined by Aitchison (1992,1997) satisfies this condition; it is based on the trace of the covariance matrix of the centred logratio transformed compositions. In accordance with his definition, a measure of total variability of a compositional data set $\mathbf{X}$ can be defined as

$$\text{totvar}(\mathbf{X}) = \sum_{j=1}^{D} \sum_{i=1}^{N} \left\{ \log\left(\frac{x_{ij}}{g(\mathbf{x}_i)} \right) - m_j \right\}^2, \tag{7}$$

where $m_j = \frac{1}{N} \sum_i \log\left(\frac{x_{ij}}{g(\mathbf{x}_i)} \right)$ $(j = 1, \ldots, D)$. It is easy to prove that

$$\text{totvar}(\mathbf{X}) = \sum_{i=1}^{N} d_{eu}^2\left(\text{clr}(\mathbf{x}_i), \overline{\text{clr}(\mathbf{X})}\right) = \sum_{i=1}^{N} \Delta^2\left(\mathbf{x}_i, \text{cen}(\mathbf{X})\right), \tag{8}$$

which proves that the proposed measure of total variability (7) is compatible with the distance defined in (2). It can also be proved that $\text{totvar}(\mathbf{p} \circ \mathbf{X}) = \text{totvar}(\mathbf{X})$, for any perturbation $\mathbf{p} \in \mathbf{S}^{D-1}$, which proves that the measure of total variability (7) is invariant under perturbations.

3. Hierarchic cluster analysis of compositional data

Before applying any hierarchic method of classification to a data set, it is necessary to establish in advance some or all of the following measures: of difference, central tendency and dispersion, to be used in accordance with the nature of the data. Therefore, if we are using a hierarchic method to classify a compositional data set, we have to take into account that all these measures must be invariant under the group of scale transformations. It is clear that the definitions given in (2), (5) and (8) are scale-invariant, while the euclidean distance is not. Therefore, from this point of view, it is wrong to use the euclidean distance between two compositions to calculate the matrix of distances associated with hierarchic methods like single linkage, complete linkage and average linkage. We propose to use the distance defined in (2). By property (4), this distance is equivalent to the euclidean distance between the compositions transformed by

the centred logratio transformation *clr* defined in (3).

Likewise, any method of classification which reduces the distance from a composition to a set of compositions to the distance between the composition and the `centre' of the group, has to take into account the considerations made in the previous section *2.b*. We propose to use (5) as a definition of the `centre' of a set of compositions, in addition to the distance defined in (2). Then, by property (6), it is easy to conclude that the centroid classification method can be used if it is applied to the data transformed by the centred logratio transformation.

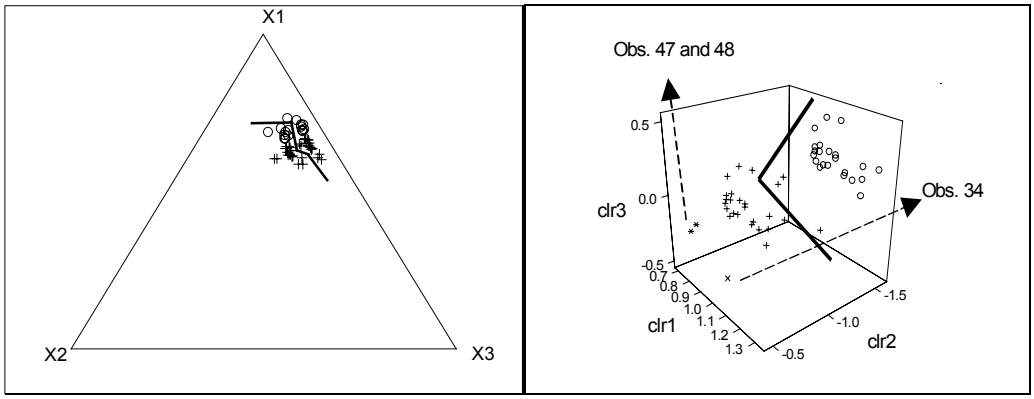
On the other hand, the well-known method of Ward is a hierarchic method which uses the measure of dispersion to classify the data. In essence, this method is based on the concept of variability on a cluster C. This variability is defined by the sum $\sum_{\mathbf{x}\in C} d_{eu}^2\left(\mathbf{x}, \overline{C}\right)$ (see Everitt, 1993, Section 5.2), where $\overline{C}$ denotes the centre of the class. When the data set is compositional, we suggest replacing the squared Euclidean distance $d_{eu}^2\left(\mathbf{x}, \overline{C}\right)$, which appears in the previous sum, by $\Delta_{eu}^2\left(\mathbf{x}, \mathrm{cen}(C)\right)$ defined in (2). Then, by definition (7) and property (8), the variability of a cluster C of compositions will be equal to $\mathrm{totvar}(C)$. Thus, modifying the method of Ward to make it suitable for the classification of a compositional data set $\mathbf{X}$ is equivalent to applying the standard procedure to $\mathrm{clr}(\mathbf{X})$.

4. Two examples

Consider the 50 points plotted on a ternary diagram in Figure 2a, corresponding to a simulated compositional data set **X1** characterised by three components. Samples belong to two groups obtained one from the other by the application of a perturbation. Figure 3a shows the ternary diagram of a second simulated data set **X2** (adapted from Aitchison, 1986, data set 1) with 50 elements, which has been generated and labelled in a similar manner. Figures 2b-3b show the plots in $\Re^3$ of the *clr*- transformed data set clr(**X1**) and clr(**X2**), respectively. In each case, original groups are separated by a line and the groups resulting of a cluster method are distinguished by a different symbol.

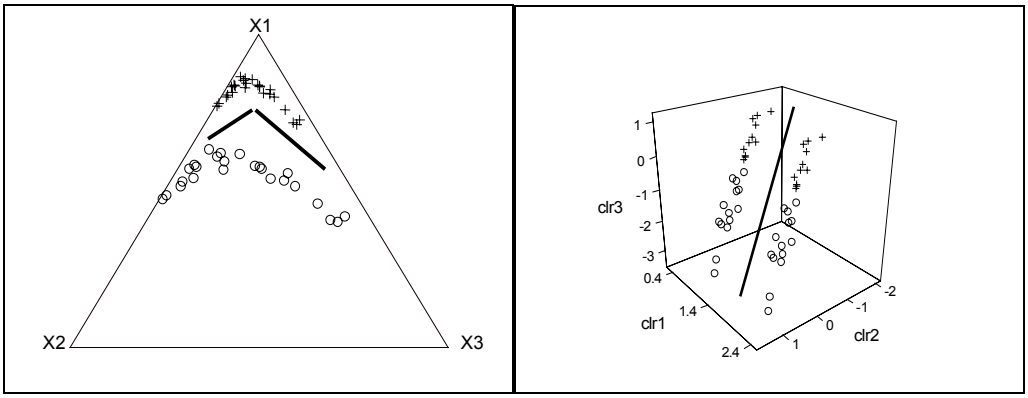
As can be observed, original groups show no overlapping neither in $\mathbf{S}^2$ nor in $\Re^3$ but, while Figures 3a and 3b show also a clear visual separation of the two groups, this is not the case for the data represented in Figure 2a and 2b. For the sake of comparison, different standard classification methods have been applied to the four sets using the Euclidean distance. Misclassification rates are listed in Table 1.

Figure 2: *Example 1 (a) Plot in the simplex (groups from Ward's method) (b) clr-transformed set (groups obtained using single linkage)); classification results distinguished by symbols '+' and 'o'.*



(a) *(b)*

Figure 3: *Example 2 (a) Plot in the simplex (groups obtained using single linkage) (b) clr-transformed set (groups from Ward's method); classification results distinguished by symbols '+' and 'o'.*



(a) *(b)*

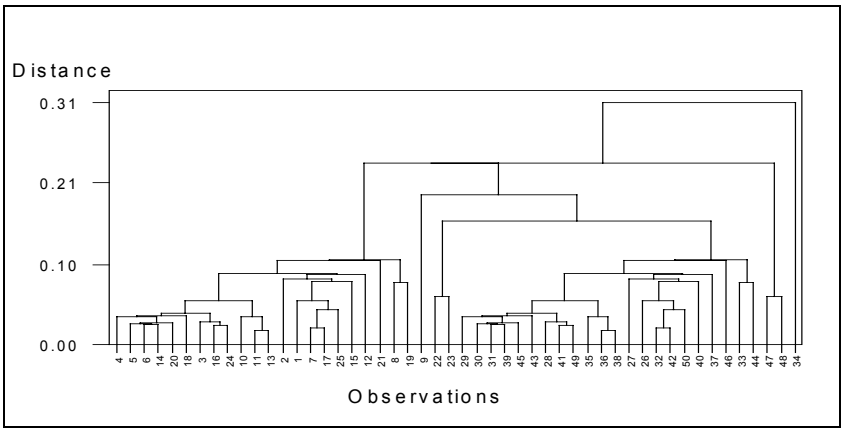
Table 1 *Misclassification rates on the two examples*

	Example 1		Example 2	
Method	**X1**	*clr*(**X1**)	**X2**	*clr*(**X2**)
Single Linkage	48%	12%	0%	8%
Ward	50%	6%	8%	50%
Complete Linkage	22%	6%	42%	50%
Centroid	48%	6%	42%	50%
Average Linkage	50%	6%	28%	50%

As could be expected, a poor classification power is obtained for **X1**, because the two groups are *close* from an euclidean point of view (only the complete linkage method gives an acceptable classification). However the classification power seems to be reasonable for *clr*(**X1**), because only the three ele-

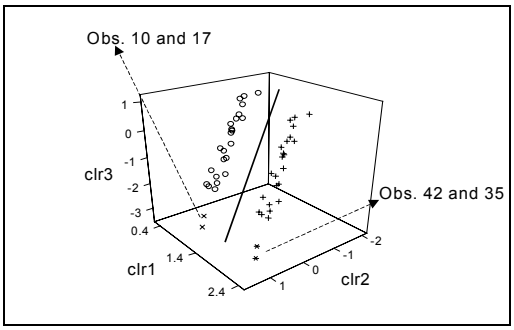
ments close to the border are misclassified. For *clr* **(X1)** the poorest result is obtained when the single linkage is used. Figure 4 shows the associated dendrogram. The samples of the first group are labelled from 1 to 25, while the others are labelled from 26 to 50. This dendrogram is similar to the dendrograms associated to the others methods whit only one difference: compositions labelled by 34, 47 and 48 are considered by the single linkage method as separated groups (see Figure 2(b)).

Figure 4: *Dendrogram of the single linkage method applied to clr* **(X1)**.

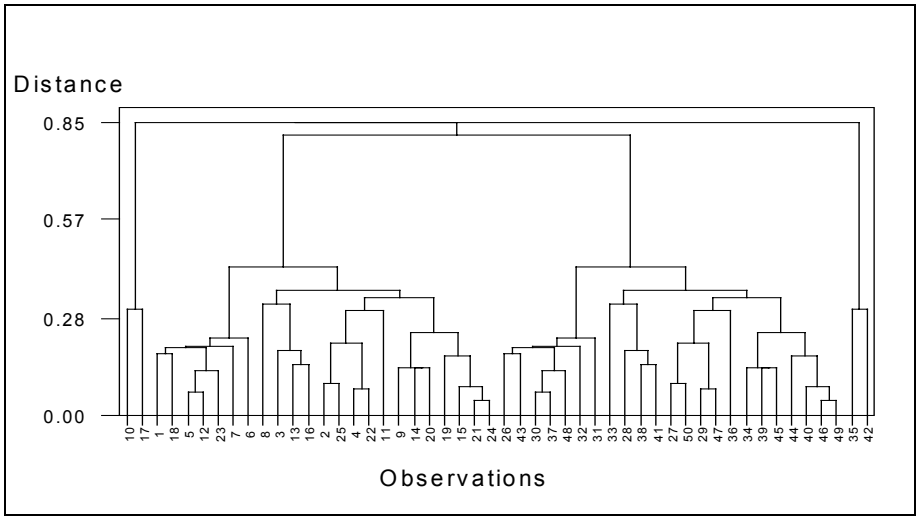


Results are more striking for **X2** and *clr*(**X2)**, given the clear separation between groups both in the simplex and in real space: only single linkage and Ward's methods show a high classification power for **X2,** while complete linkage, centroid and average linkage methods have a poor classification power. They work still worse when applied to *clr*(**X2)** because the two groups are parallel and elongated. Figure 5(b) shows the associated dendrogram when the single linkage method is applied to the data set *clr*(**X2).** Numbers from 1 to 25 correspond to the first group and 26 to 50 to the second. It can be observed that the almost all samples are well classified: only observations labelled 10 and 17, and observations 35 and 42 are misclassified as separated groups (see Figure 5(a)).

Figure 5: *Single linkage method applied to clr* **(X2)**: *(a) Plot (b) Dendrogram*



(a)



(b)

5. Conclusions

- There are theoretical objections to the application of the standard hierarchic classification methods to compositional data sets because they don't take into account the nature of this kind of data.
- To classify a compositional data set, we suggest adapting the usual hierarchic methods using the definitions of distance, centre and variability defined in (2), (5), and (7), which are compatible with the compositional nature of the data. This is equivalent to applying standard methods to the centred logratio transformed data set.
- Further research is needed to understand more thoroughly the performance of modified standard parametric classification methods when they are applied to compositional data.

References

Aitchison, J., (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall, New York (USA), 416 p.

Aitchison, J., (1992). On Criteria for Measures of Compositional Difference, *Mathematical Geology*, vol. 24, No. 4, pp. 365-379

Aitchison, J., (1997). 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, pp. 3-35.

Everitt, BS., (1993). *Cluster Analysis*. Edward Arnold, Cambridge (UK), 170 p.