Some Practical Aspects on Multidimensional Scaling of Compositional Data

J. A. Martín-Fernández[1] and M. Bren[2]

To visualize the data with Multidimensional Scaling methods we approximate a given dissimilarity matrix –matrix of differences among observations– to obtain a configuration of points in low (two) dimensional real (usually) Euclidean space. The Multidimensional Scaling methods input is a dissimilarity matrix and to construct such a matrix a suitable measure of difference between observations is needed.

In our work we discuss applications of different dissimilarity measures, relations between them and their (un)suitability in case of compositional data. We present results of Multidimensional Scaling methods applied to real compositional data sets to visualize all these relations. Visualizations also confirm our theoretical results and show which dissimilarity measures are coherent with the compositional nature of the data.

KEY WORDS: Euclidean distance, Aitchison distance, Kullback-Leibler information index, Bhattacharyya distance, centering operation, stress index.

[1]Universitat de Girona, Depart. Informàtica i Matemàtica Aplicada, Campus Montilivi P-I, Girona, SPAIN, e-mail: josepantoni.martin@udg.es

[2]University of Maribor, Faculty of Organizational Sciences, Kidričeva cesta 55[a], Kranj, SLOVENIA, and University of Ljubljana, Institute of mathematics, Phisics, and Mechanics, Jadranska 19, Ljubljana, SLOVENIA, e-mail: matevz.bren@fov.uni-mb.si

1 INTRODUCTION

1.1 The sample space for compositional data

An observation $\mathbf{x} \in \mathbb{R}_+^D$ is compositional (Aitchison, 1986) if its components are proportions of some whole. Thus, their natural sample space is the *simplex* denoted

$$\mathcal{S}^D = \{\mathbf{x} = (x_1, x_2, \ldots, x_D):\ x_j > 0;\ j = 1, 2, \ldots, D;\ x_1 + x_2 + \cdots + x_D = 1\}. \quad (1)$$

Actually, any vector of positive components $\mathbf{w} \in \mathbb{R}_+^D$ can be projected onto the simplex by the *closure operation* $\mathcal{C}(\mathbf{w}) = (w_1/\sum w_j, w_2/\sum w_j, \ldots, w_D/\sum w_j)$. It is easy to see that the *perturbation operation*

$$\mathbf{p} \circ \mathbf{x} = \mathcal{C}(p_1 x_1, p_2 x_2, \ldots, p_D x_D) \quad (2)$$

defined on $\mathcal{S}^D \times \mathcal{S}^D$, and the *power transformation* $\alpha \cdot \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha)$ defined on $\mathbb{R} \times \mathcal{S}^D$, induce a vector space structure in to the unit simplex. The neutral element of this vector space is $\mathbf{e}_D := \left(\frac{1}{D}, \frac{1}{D}, \cdots, \frac{1}{D}\right)$ and the inverse element of a composition $\mathbf{x} \in \mathcal{S}^D$ is $\mathbf{x}^{-1} = \mathcal{C}\left(\frac{1}{x_1}, \frac{1}{x_2}, \cdots, \frac{1}{x_D}\right) \in \mathcal{S}^D$.

It is important to recall that the vector space structure $\left(\mathcal{S}^D, \circ, \cdot\right)$, its algebraic and geometrical concepts: vector, norm, scalar product, and distance plays a central role in most of the statistical methods.

1.2 Multidimensional scaling (MDS) techniques

Graphical representations of multivariate data are widely used in research and applications of many disciplines. Methods used are based on techniques of representing a set of observations by a set of points in a low-dimensional real (usually) Euclidean vector space, so that observations that are similar to one another are represented by points that are close together. Multidimensional scaling (MDS) techniques belong to these family of graphical representations (Cox and Cox, 1994).

We begin with a set of n observations under consideration and between each pair of observations (i, j) there is a measurement δ_{ij} of the dissimilarity between them. MDS techniques search for a low dimensional (usually) Euclidean space and a set of *points* in the space that represent the observations, each point represents one observation, in such manner that the distances $\{d_{ij}\}$ between points in the space approximate as well as possible the original dissimilarities $\{\delta_{ij}\}$. It is the different notions of approximation that give rise to the different techniques of MDS: Metric MDS and Nonmetric MDS (for more details see Cox and Cox, 1994). For our purpose, we focus on the metric MDS techniques called *Classical Scaling* or *Principal Coordinates Analysis*. These techniques assume that the dissimilarities $\{\delta_{ij}\}$ between observations are distances within a set of n points in some D-dimensional Euclidean space. Then, it is possible to find (or to reconstruct) a configuration of points in a low-dimensional Euclidean space where the equality $d_{ij} = \delta_{ij}$ holds.

1.3 Measures of difference

A measure of difference between observations plays a central role in all methods of MDS: the input of MDS is a dissimilarity matrix and to construct such a matrix we need a suitable measure of difference. Therefore, when the data are compositional a measure

of difference coherent with the nature of compositions has to be applied. In the literature different measures of dissimilarity between compositions have been proposed: Euclidean, Aitchison's distance, compositional Kullback-Leibler divergence... Recent studies (Aitchison et al., 2000; Martín-Fernández et al.; 1998a, 1998b, 1999, 2001) have shown that some of these measures are not suitable for compositional data (Euclidean, Bhattacharyya...) and some are coherent with compositional nature (Aitchison, compositional KL divergence...). Nevertheless, results obtained applying different measures sometimes differ much and sometimes are closely related.

In our work we'll consider four measures: Aitchison's distance, compositional KL divergence, Bhattacharyya distance and Euclidean distance.

The Aitchison's distance (squared), deeply analyzed in mentioned recent works, is given by

$$d^2_{\mathrm{A}}(\mathbf{x},\mathbf{y}) = \sum \left[\frac{\log(x_k)}{g(\mathbf{x})} - \frac{\log(y_k)}{g(\mathbf{y})}\right]^2, \tag{3}$$

where $g(\mathbf{x})$ denotes the *geometric mean* of compositional vector $\mathbf{x}$.

The compositional KL divergence (squared), introduced in Martín-Fernández et al. (1999), is given by

$$d^2_{\mathrm{M}}(\mathbf{x},\mathbf{y}) = \frac{D}{2}\log\left[\overline{\frac{\mathbf{x}}{\mathbf{y}}}\,\overline{\frac{\mathbf{y}}{\mathbf{x}}}\right], \tag{4}$$

where $\overline{\frac{\mathbf{x}}{\mathbf{y}}}$ denotes the *arithmetical Mean* of the vector of ratios $\frac{\mathbf{x}}{\mathbf{y}} = (x_1/y_1,\ldots,x_D/y_D)$. In the cited works it is shown that both measures are compatible with the compositional nature of the data. The main difference between them is that d_{A} is a distance but d_{M} is a dissimilarity – it is not metric.

Following Rao (1982) the most suitable measure of divergence between multinomial probability distributions is the Bhattacharyya's distance given by

$$d_{\mathrm{B}}(\mathbf{x},\mathbf{y}) = \arccos\left[\sum \sqrt{x_k}\sqrt{y_k}\right], \tag{5}$$

where implicitly we consider a composition $\mathbf{x}$ as a vector of probabilities of a multinomial distribution.

The Euclidean distance is most commonly used measure of difference in Classical Scaling. Squared is given by

$$d^2_{\mathrm{E}}(\mathbf{x},\mathbf{y}) = \sum (x_k - y_k)^2. \tag{6}$$

The Bhattacharyya and Euclidean distances are not compatible (Martín-Fernández et al., 1998a) with the real vector space structure defined on a simplex and therefore are not coherent with the nature of compositional data. But still in some cases for some compositional data we obtain reasonable results of MDS even when we apply these not suitable measures of difference. In our communication we'll found out answers to questions: When such cases arise and what are the reasons for these exceptions?

Our strategy consists of combining the *centering operation* and the multidimensional scaling. The *centering operation,* for the first time introduced in Martín-Fernández et al. (1999), is a useful tool to treat the compositional data sets located near to the border or to the corner of the simplex. First we consider the center of the data set $\mathbf{X}$ as the compositional geometric mean $\mathrm{cen}(\mathbf{X})$ defined by $\mathrm{cen}(\mathbf{X}) = \mathcal{C}(g_1, g_2, \ldots, g_D)$, where $g_j = \left(\prod_{i=1}^{N} x_{ij}\right)^{1/N}$ is the geometric mean of the j-th components of all compositions $\mathbf{x}_1$,

$\mathbf{x}_2$,..., $\mathbf{x}_N$ in $\mathbf{X}$. If we perturb the data set $\mathbf{X}$ by the vector $\mathrm{cen}(\mathbf{X})^{-1}$ the resulting data set is centered, i.e. the center of the perturbed set $\mathrm{cen}(\mathbf{X})^{-1} \circ \mathbf{X}$ is $\mathbf{e}_D$, the center of the simplex. The data set is now located around the *barycenter* $\mathbf{e}_D$ of the simplex. For the four measures of difference described above we'll calculate and compare the dissimilarity matrices obtained for centered and non centered data sets. With this comparison we'll find out answers to above posed questions.

Also we'll carry out the evaluation of the performances of the four measures of difference described above using the *stress* (*standarized residual sum of squares*) index defined by

$$stress = \frac{\sum_{i<j} \left(d_1(\mathbf{x}_i, \mathbf{x}_j) - d_2(\mathbf{x}_i, \mathbf{x}_j)\right)^2}{\sum_{i<j} d_1^2(\mathbf{x}_i, \mathbf{x}_j)}. \tag{7}$$

Here d_1 and d_2 denotes the distances between observations calculated before and after the MDS or/and before and after performing the centering operation. This *stress* measure, applied in the same manner in Martín-Fernández et al. (2001), is one of the basic elements of multidimensional scaling theory (Cox and Cox, 1994).

2 RELATIONS BETWEEN CONSIDERED MEASURES OF DIFFERENCE

2.1 Aitchison's distance versus compositional KL divergence

In Martín-Fernández (2001) it is shown that Aitchison's distance d_{A} (3) and compositional KL divergence d_{M} (4) are closely related. For any two compositions $\mathbf{x}$, $\mathbf{y} \in \mathcal{S}^D$, it holds that

$$d_{\mathrm{A}}(\mathbf{x}, \mathbf{y}) \;\approx\; \sqrt{2}\, d_{\mathrm{M}}(\mathbf{x}, \mathbf{y}). \tag{8}$$

The consequence of this property is that for any compositional data set both measures produce similar matrices of distances. Thus, when we apply MDS with d_{A} or d_{M} the results obtained will be very similar.

2.2 Aitchison's distance versus Euclidean distance

Figure 1 presents neighborhoods of different points in the simplex calculated with the Aitchison's distance (3). We see that the shape of the neighborhoods are nearly spherical when the center of the neighborhoods is located near to the barycenter $\mathbf{e}_D$ of the simplex. On the other hand, when the center of the neighborhoods is near to an edge or to a corner of the simplex the shape of neighborhoods extremely differs from a sphere. This fact suggests that only when the data are near to barycenter $\mathbf{e}_D$ we can expect some close relation between the Aitchison's and the Euclidean distance (6).

Indeed. From an Euclidean point of view, when two observations $\mathbf{x}$ and $\mathbf{y}$ are "near" to the center we can express them as $\mathbf{x} = \mathbf{e}_D + \delta_{\mathbf{x}}$ and $\mathbf{y} = \mathbf{e}_D + \delta_{\mathbf{y}}$, where $\delta_{\mathbf{x}}$ and $\delta_{\mathbf{y}}$ are vectors with values approximately equal to 0 and $\sum(\delta_{\mathbf{x}})_k = \sum(\delta_{\mathbf{y}})_k = 0$. Thus, it is obvious that $d_{\mathrm{E}}(\mathbf{x}, \mathbf{y}) = d_{\mathrm{E}}(\delta_{\mathbf{x}}, \delta_{\mathbf{y}})$.

From compositional point of view, when two observations $\mathbf{x}$ and $\mathbf{y}$ are "near" to the center we can express them as $\mathbf{x} = \mathbf{e}_D \circ \varepsilon_{\mathbf{x}}$ and $\mathbf{y} = \mathbf{e}_D \circ \varepsilon_{\mathbf{y}}$ where $\varepsilon_{\mathbf{x}}$ and $\varepsilon_{\mathbf{y}}$ are vectors with all components approximately equal to $1/D$ and $\circ$ denotes the perturbation operation (2). Because the Aitchison distance is perturbation invariant we have $d_{\mathrm{A}}(\mathbf{x}, \mathbf{y}) = d_{\mathrm{A}}(\varepsilon_{\mathbf{x}}, \varepsilon_{\mathbf{y}})$.

Then, if we take $\delta_{\mathbf{x}} = \varepsilon_{\mathbf{x}} - \mathbf{e}_D$ and $\delta_{\mathbf{y}} = \varepsilon_{\mathbf{y}} - \mathbf{e}_D$ with Taylor expansion of the logarithm we get the approximate relation:

$$d_{\mathrm{A}}(\mathbf{x}, \mathbf{y}) \approx D\, d_{\mathrm{E}}(\mathbf{x}, \mathbf{y}). \tag{9}$$

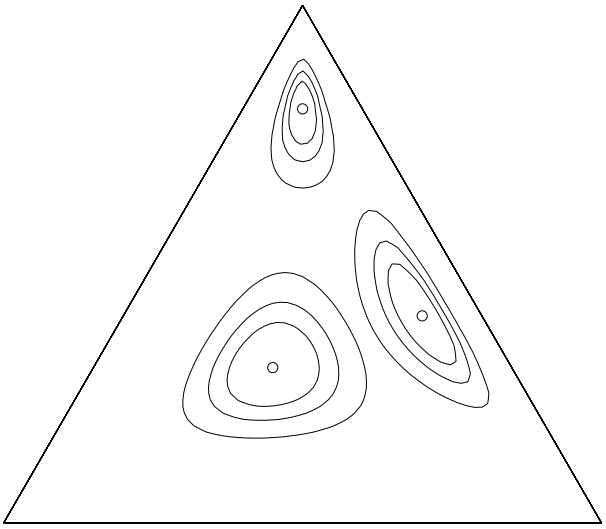


Figure 1: *Neighborhoods in the ternary diagram calculated with Aitchison's distance.*

Thus, when we apply Classical Scaling to a compositional data set located near to the barycenter of the simplex results using Euclidean or Aitchison's distance will be similar.

2.3 Bhattacharyya distance versus Euclidean distance

Also in the case of the Bhattacharyya distance the Figure 2 now shows that the shape of the neighborhoods is in the same relation to the Euclidean neighborhoods as in the Aitchison's distance case. Therefore we'll analyze the relation between the Bhattacharyya distance (5) and the Euclidean distance (6) only when the observations are "near" to the barycenter $\mathbf{e}_D$.

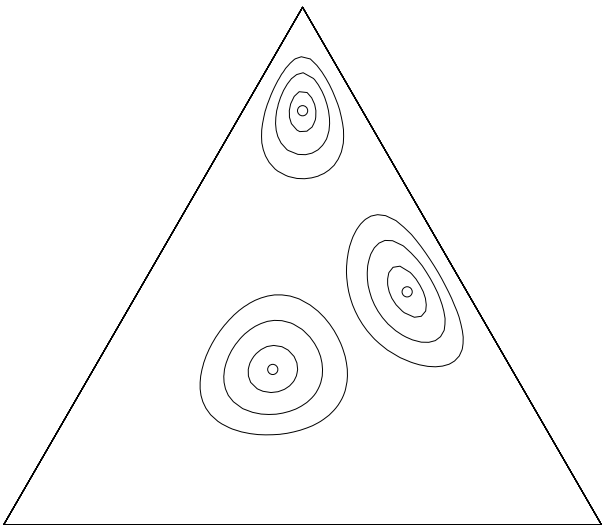


Figure 2: *Neighborhoods in the ternary diagram calculated with Bhattacharyya distance.*

Let's rewrite the Bhattacharyya distance in terms of the angle between vectors. From the definition of the Bhattacharyya distance $d_{\mathrm{B}}(\mathbf{x},\mathbf{y}) = \arccos\left[\sum \sqrt{x_k}\sqrt{y_k}\right]$ we get

$$\cos(d_{\mathrm{B}}(\mathbf{x},\mathbf{y})) = \sum \sqrt{x_k}\sqrt{y_k} = \langle \sqrt{\mathbf{x}}, \sqrt{\mathbf{y}} \rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product of vectors and $\sqrt{\mathbf{x}} = (\sqrt{x_1}, \ldots, \sqrt{x_D})$ is the vector of square roots of the components. By the definition of the scalar product we have

$$\langle \sqrt{\mathbf{x}}, \sqrt{\mathbf{y}} \rangle = \| \sqrt{\mathbf{x}} \| \, \| \sqrt{\mathbf{y}} \| \, \cos \widehat{\sqrt{\mathbf{x}} \sqrt{\mathbf{y}}},$$

where $\| \cdot \|$ denotes the Euclidean norm and $\widehat{\cdot \ \cdot}$ the angle between vectors. Note that if $\mathbf{x}$ is a composition then the Euclidean norm of the vector $\sqrt{\mathbf{x}}$ is equal to 1 and therefore $\cos(d_{\mathrm{B}}(\mathbf{x}, \mathbf{y})) = \cos \widehat{\sqrt{\mathbf{x}} \sqrt{\mathbf{y}}}$. Since $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D_+$ we obtain that $d_{\mathrm{B}}(\mathbf{x}, \mathbf{y}) = \widehat{\sqrt{\mathbf{x}} \sqrt{\mathbf{y}}}$. We see that Bhattacharyya distance between compositions $\mathbf{x}$ and $\mathbf{y}$ can be interpreted as the angle between their projections $\sqrt{\mathbf{x}}$ and $\sqrt{\mathbf{y}}$ on the unit sphere. Therefore we can establish a relation between d_{B} and d_{E}: When $\mathbf{x}$ and $\mathbf{y}$ are two observations "near" to the barycenter $\mathbf{e}_D$, the angle between $\sqrt{\mathbf{x}}$ and $\sqrt{\mathbf{y}}$ is close to the Euclidean distance between them. Applying the Taylor expansion we can show that these distances verify the approximate relation:

$$d_{\mathrm{E}}(\mathbf{x}, \mathbf{y}) \approx \sqrt{\frac{8}{D}(1 - \cos(d_{\mathrm{B}}(\mathbf{x}, \mathbf{y})))}. \tag{10}$$

Thus, when we apply Classical Scaling to a compositional data set located near to the barycenter using Euclidean or Bhattacharyya distance similar results are expected.

3 CASE STUDIES

3.1 Lyons West data set

The data set Lyons West, analyzed recently in Martín-Fernández et al. (2001), are gathered from 76 wells in the Lyons West oil field, Rice County, Kansas (for more details see Martín-Fernández et al., 2001). In this compositional data set we consider three components: oil, water, and rock denoted $\mathbf{O}$, $\mathbf{W}$, and $\mathbf{R}$ respectively. Figure 3 shows the data set in the ternary diagram (units are represented by '$*$'). The third components $\mathbf{R}$ of the data takes large values and that's why the data set is located near to the $\mathbf{R}$ corner. The shape of the data set suggests that some linear relation exists between the components of the data. This kind of relation called *logcontrast* was analyzed deeply in Aitchison's book (Aitchison, 1986). In our case we can consider that there exists a vector $\mathbf{a} = (a_1, a_2, a_3)$ with $\sum a_i = 0$ such that the *logcontrast equality*

$$a_1 \log(\mathbf{O}) + a_2 \log(\mathbf{W}) + a_3 \log(\mathbf{R}) = \lambda \tag{11}$$

holds. It is easy to see that if an observation $\mathbf{x}$ verifies the logcontrast equation (11) then any perturbed composition $\mathbf{p} \circ \mathbf{x}$ verifies the logcontrast equation

$$a_1 \log(\mathbf{O}) + a_2 \log(\mathbf{W}) + a_3 \log(\mathbf{R}) = \lambda + \sum a_i p_i.$$

Thus, any "linear" data set preserves its linearity after an arbitrary perturbation. In particular, the linearity is preserved by the centering operation. In Figure 3 we can see that the centered data set (units are now represented by '$\circ$') preserves the linear pattern.

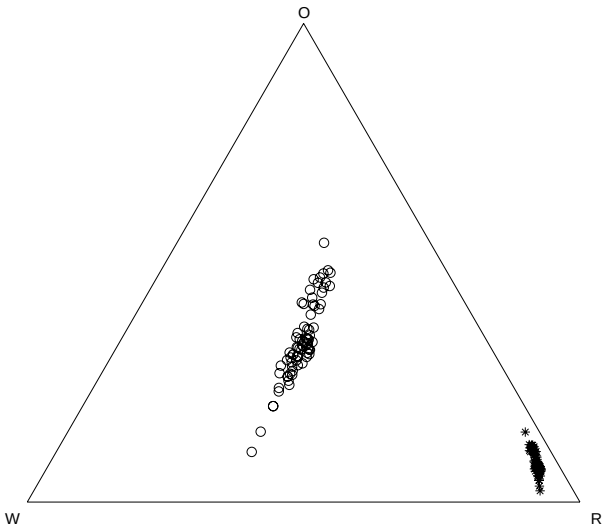


Figure 3: *Lyons West data set in the ternary diagram ('∗': initial data set; '∘': centered data set).*

First, to illustrate the suitability of the selected measures we have calculated the values of the distances between any pair of observations of the Lyons West data set, before and after centering operation is applied (see Figures 4 and 5).

In Figure 4A and 4B for the Aitchison's distance and the compositional KL divergence, respectively, we see that these two measures are *perturbation invariant*: for any pair of observations the values of distances before and after centering are the same.

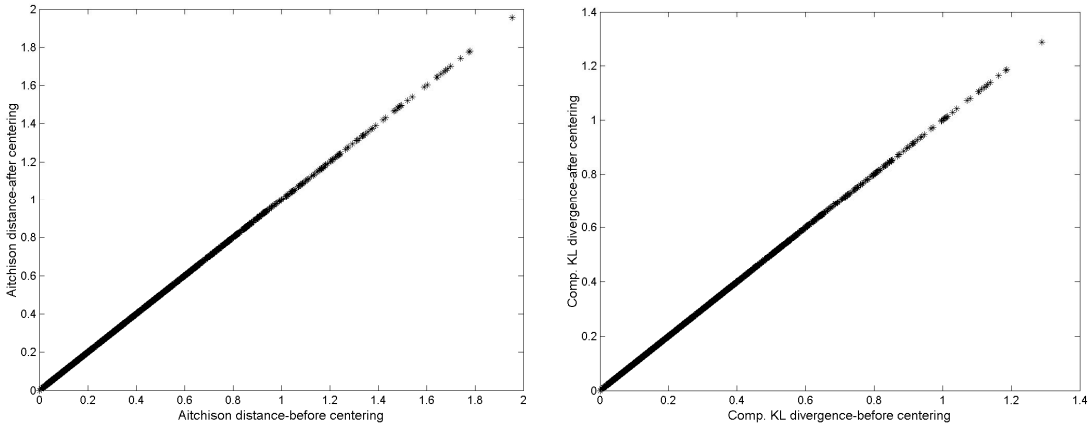


Figure 4: Distances before and after centeringLyons of West data set. A: d_{A}; B: d_{M}.

But in the plots 5A and 5B of the next figure we can see, respectively, that this is not the case for the Euclidean and Bhattacharyya distance: the distances calculated for the same pair of observations before and after centering operation differs. Therefore the Aitchison's distance and the compositional KL divergence are, but the Euclidean and Bhattacharyya distance are not compatible with compositional nature of the data.

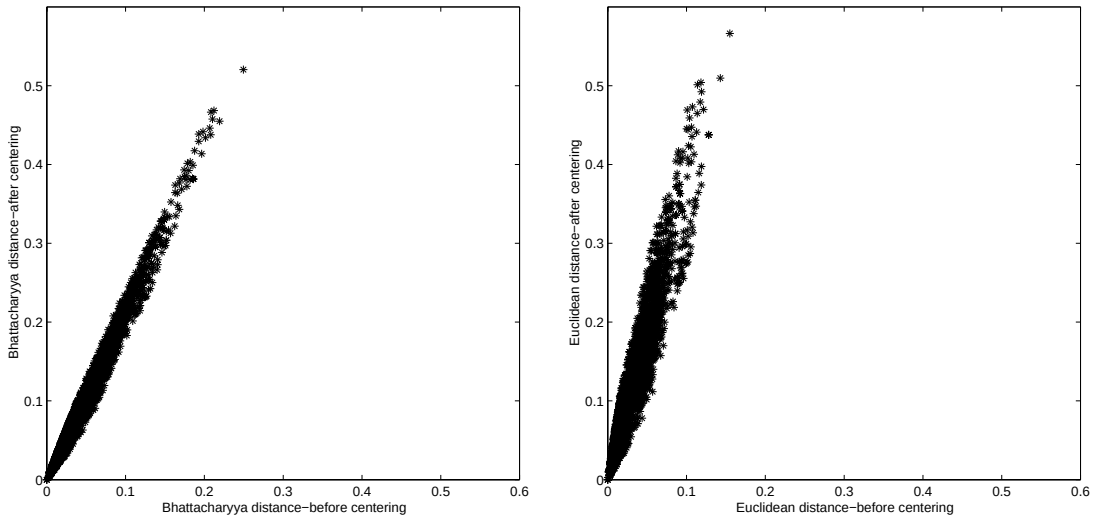


Figure 5: Distances before and after centering of Lyons West data set. A: d_{B}; B: d_{E}.

Second, to illustrate the above derived relations (8, 9, 10) between the discussed distances we compare, for the four measures, the values of the distances between all pairs of observations of the centered Lyons West data set. In a plot we represent the relations between the values calculated with one measure versus another – Figure 6 shows these plots. We can observe in Figure 6A how closely related is the Aitchison's distance d_{A} to compositional KL divergence d_{M} following the relation (8). In the Figure 6B we have plotted the compositional KL divergence (squared) versus the Aitchison's distance. Obviously, because (8) holds the shape is very similar to the line $y = \sqrt{2x}$. This figure shows that squaring the dissimilarity d_{M} we obtain a measure d^2_{M} also closely related to the metric d_{A}. Thus, we expect reasonable results of Classical Scaling applied to the dissimilarity matrix calculated with d^2_{M}.

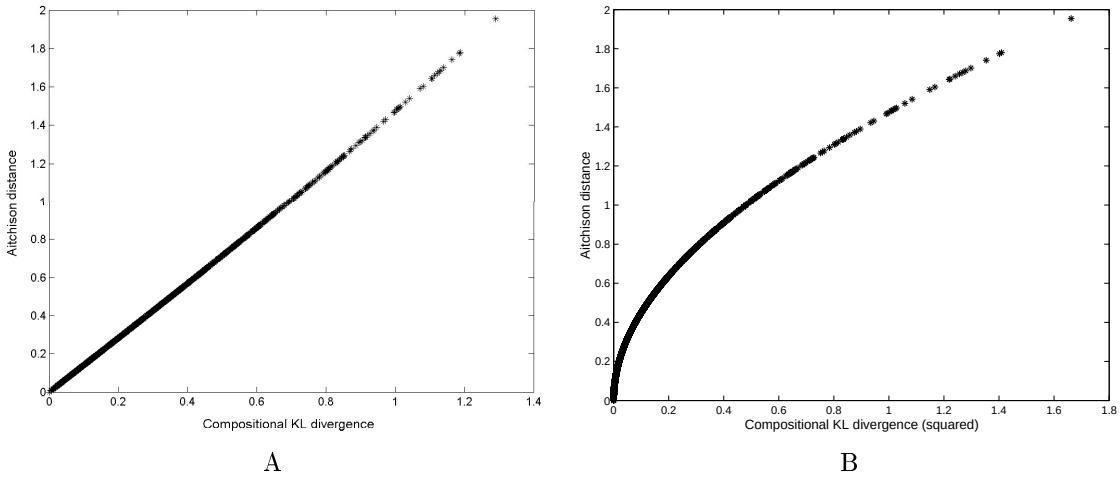
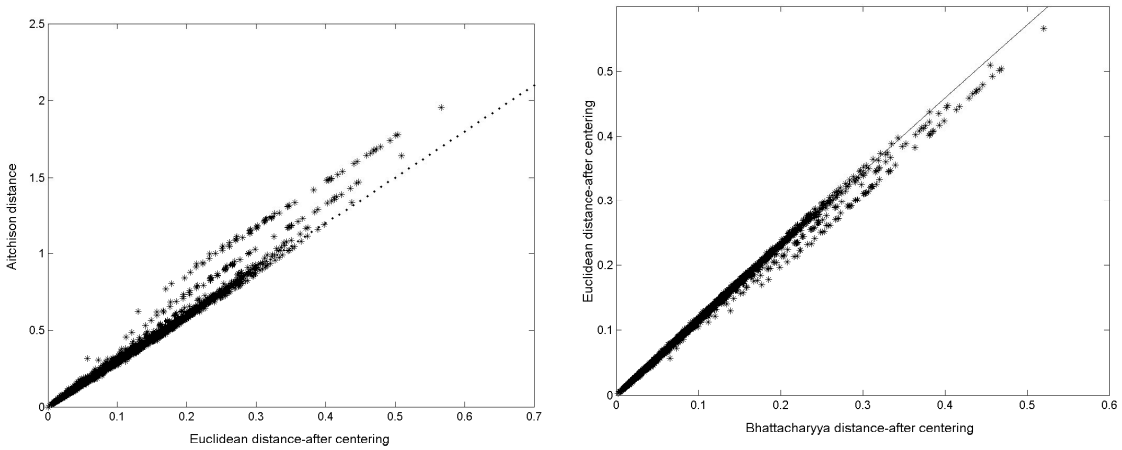


Figure 6: Distances between observations of Lyons West data set. A: d_{M} versus d_{A}; B: Squared d_{M} versus d_{A}.

In Figure 7A we can observe that the relation (9) is verified (when previously, we have centered the data set). In this figure we have drawn the straight line $y = 3x$ (dotted line)

to emphasize this relation. Figure 7B illustrates the relation (10) between the distances d_{B} and d_{E}. Here we have represented with a line the function $y = \sqrt{\frac{8}{3}(1-\cos(x))}$ – note that this line appears very similar to straight line $y = x$ for the values near to zero.



A B

Figure 7: Distances between observations of Lyons West data set. A: d_{E} (centered data set) versus d_{A}; B: d_{B} (centered data set) versus d_{E} (centered data set).

With the four measures d_{A}, d_{M}, d_{B}, and d_{E} we have calculated, respectively, the values of the distances between any pair of observations of the Lyons West data set and in this manner we have obtained four dissimilarity matrices. To these dissimilarity matrices we have applied the Classical Scaling method resulting four configurations of points in three dimensional Euclidean space. In Figure 8 we represent the first two coordinates of these configurations. To carry out all this analysis we've used the functions included in the package S-plus®.

Figure 8A corresponds to the Aitchison's distance. If we just visually compare obtained new configuration with the centered data set of the Figure 3 we see that in this case Classical Scaling gives us reasonable results. In Figure 8B configuration obtained using the compositional KL divergence d_{M} is shown. Here we observe that the results are not reasonable – linearity is lost, what could be a consequence of the fact that d_{M} is not metric.

Figures 8C and 8D show results applying, respectively, the Bhattacharyya distance d_{B} and Euclidean distance d_{E}. With suitable left-right transformation of these figures we obtain the same shape as in the case of Aitchison's distance.

We conclude that in this two last cases the obtained results are reasonable. But still it is important to compare also the scales of the axis at these three figures.

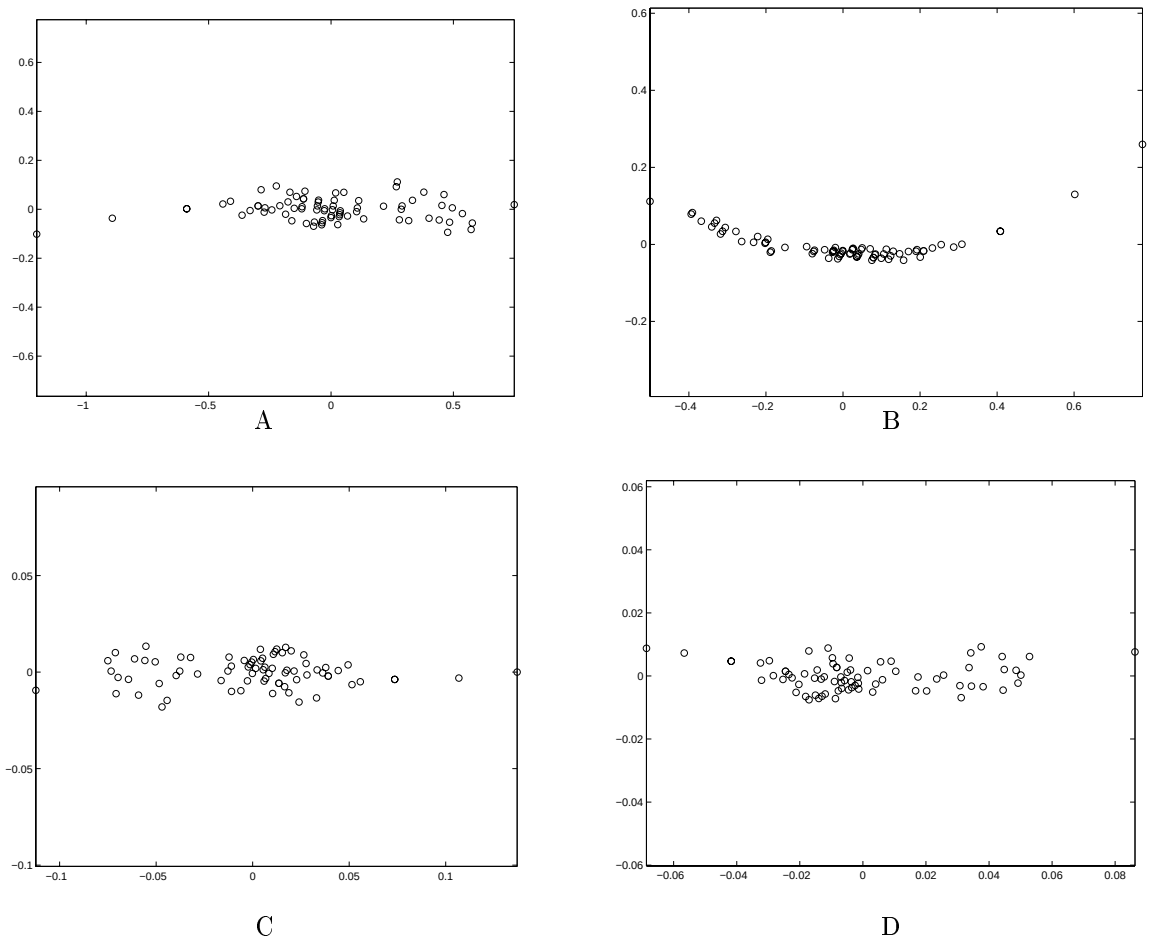


Figure 8: Results of Classical Scaling applied to Lyons West data set – first two coordinates of the resulting configuration. A: d_{A} distance; B: d_{M} dissimilarity; C: d_{B} distance; D: d_{E} distance.

Following Gordon's advise (Gordon, 1999) for nonmetric dissimilarities applied in Classical Scaling we apply Classical Scaling also to the squared compositional KL divergence d^2_{M}. In addition, we apply Classical Scaling to the centered Lyons West data set using the dissimilarity matrices calculated with distances d_{B} and d_{E}. Obtained results are shown in the Figure 9. For straightforward comparisons we have included in the figure also the plot obtained with Aitchison's distance – Figure 9A. In this four figures we can observe confirmations of the approximate relations (8, 9, 10) between the measures, we can see that now also the d^2_{M} gives reasonable results and differs from the configuration obtained applying Aitchison's distance only on the scales of axis. We note also that results obtained applying d_{B} and d_{E} are now even more similar.

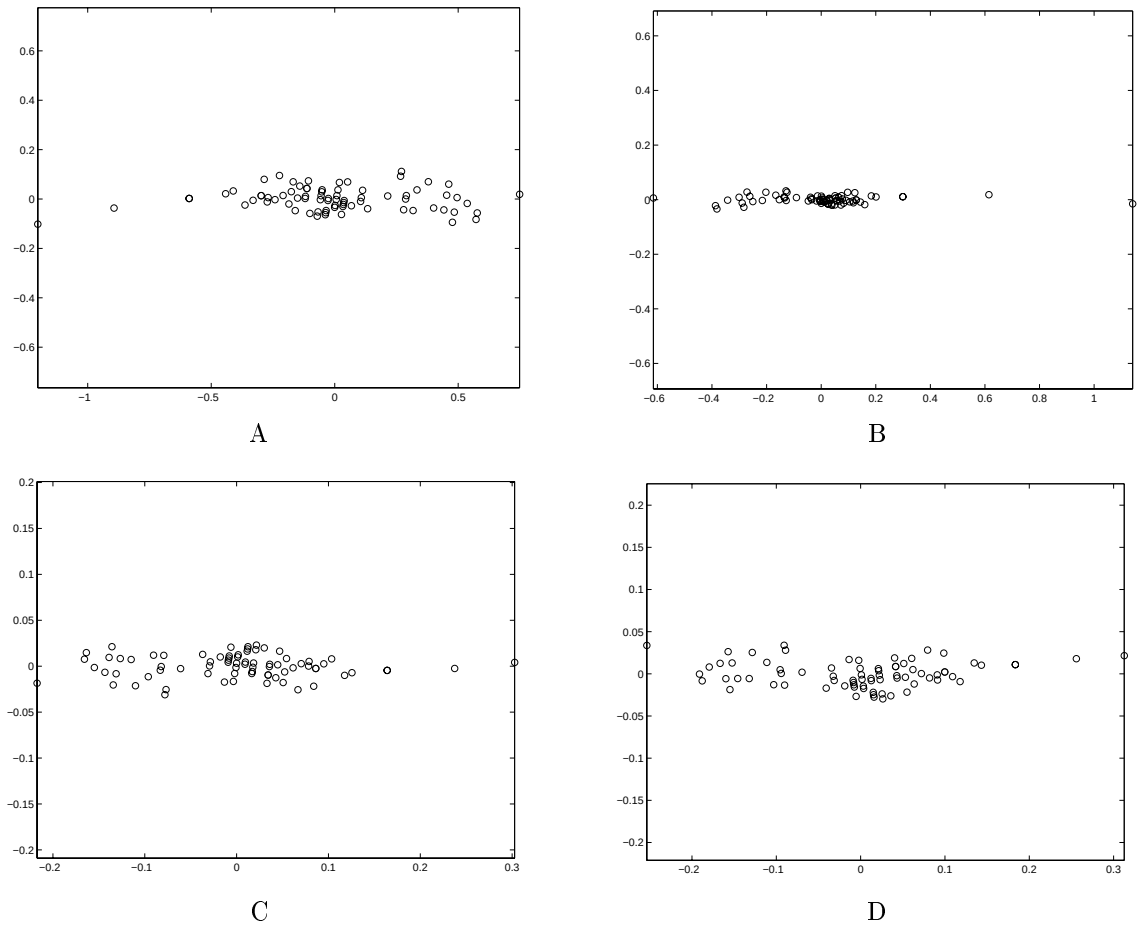


Figure 9: Results of Classical Scaling applied to Lyons West data set – first two coordinates of the resulting configuration. A: d_{A} distance; B: d_{M} dissimilarity (squared); C: d_{B} distance (centered data set); D: d_{E} distance (centered data set).

The main feature of results on Lyons West data set is that we obtain reasonable results applying non suitable measures d_{B} and d_{E}. We'll see that this is an exception and that the reason for this exceptional behavior is the linear pattern of the data. Linearity is preserved also by the centering operation and thus results obtained on centered data are reasonable for the same reason – linearity!

These results and all above mentioned aspects are confirmed also by the stress coefficient (7). This coefficient is a dissimilarity measures between two matrices of distances and it takes values between 0 and 1. Therefore the stress values near to 0 means that considered matrices of distances are very similar. Table 1 shows the value of the stress between matrices of distances calculated with the compositional KL divergence and Euclidean distance versus the matrice calculated with Aitchison distance, and Bhattacharyya distance versus Euclidean distance. We observe that accordance between Aitchison's distance d_{A} and Euclidean distance d_{E} is better after the centering operation. The same goes for the Bhattacharyya distance d_{B} versus Euclidean distance d_{E}. Table 2 presents, for each observed measure respectively, the value of the stress between the matrice of distances of the original data set and the matrice of distances in Euclidean configuration resulting the Classical Scaling method. These stress values measure an "error" of the MDS technique. We see that the error of Classical Scaling is extremely

low for all four measures considered. Note that with the non-metric Compositional KL divergence d_{M} the *worst* result is obtained.

Table 1: Performance of relation between considered dissimilarity measures as measured by the stress for Lyons West data set.

Data set	d_{A} *vs* d_{M}	d_{A} *vs* d_{E} before centering	d_{A} *vs* d_{E} after centering	d_{B} *vs* d_{E} before centering	d_{B} *vs* d_{E} after centering
Lyons West	0.3043	0.9197	0.6861	0.3804	0.1394

Table 2: Performance of Classical Scaling applied to Lyons West data set measured by the stress for considered dissimilarity measures.

Data set	d_{A}	d_{M}	d_{B} before centering	d_{B} after centering	d_{E} before centering	d_{E} after centering
Lyons West	1.7×10^{-6}	0.0372	2.4×10^{-5}	7.2×10^{-5}	2.0×10^{-6}	1.1×10^{-6}

To confirm that the linearity of the Lyons West data set is the main cause of above results, we'll analyze other data set that has no linear pattern.

3.2 Halimba data set

The data set called Halimba described in Mateu-Figueras et al. (1998) corresponds to the subcomposition of the first three components $(Al_2O_3, SiO_2, Fe_2O_3)$ of a composition $(Al_2O_3, SiO_2, Fe_2O_3, TiO_2, H_2O, Res_6)$ of 332 samples from 34 core-bore holes in the Halimba bauxite deposit in Hungary. Figure 10 shows this data set in the ternary diagram (units are represented by '$*$') and the data set resulting the centering operation (units are now denoted by '$\circ$').

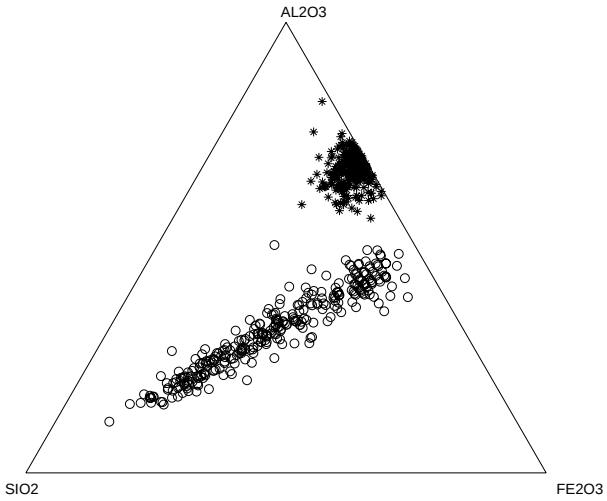


Figure 10: *Halimba data set in the ternary diagram ('$*$': initial data set; '$\circ$': centered data set).*

Note that the second component "SiO_2" of the data takes low values, thus the data set is located near to the edge. Because of our *Euclidean seeing* we can recognize the true pattern of the data set only when it is centrally located and not when its near to the border (see Figure 1). Thus, we can detect the true pattern only after applying the centering operation. In the centered data set we see that the largest variability appears in the second component "SiO_2" and that the ratio between others two components is nearly constant.

To compare performances of discussed measures: d_{A}, d^2_{M}, d_{B}, and d_{E} we have calculated, respectively, the values of the distances between any pair of observations of the Halimba data set and to the dissimilarity matrices obtained we have applied the Classical Scaling method. Figure 11 shows four resulting configurations.

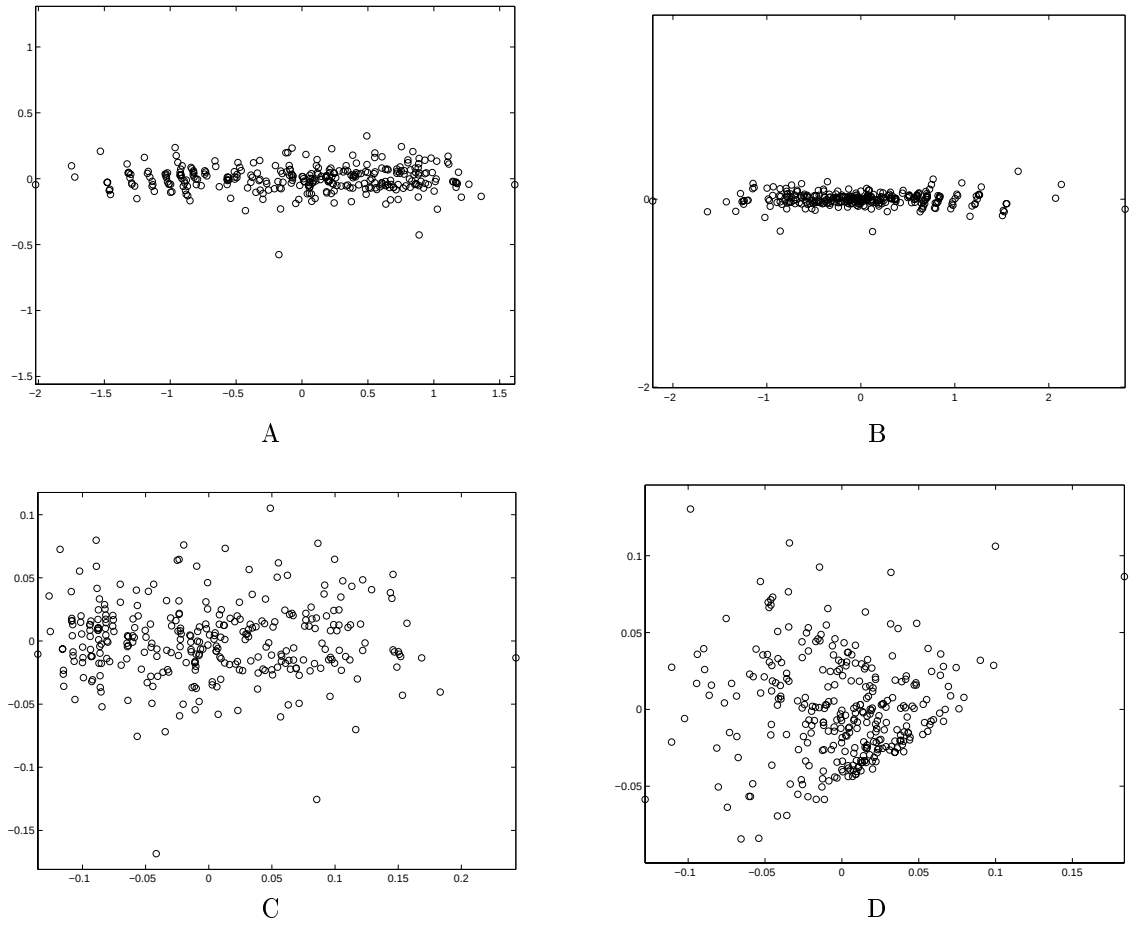
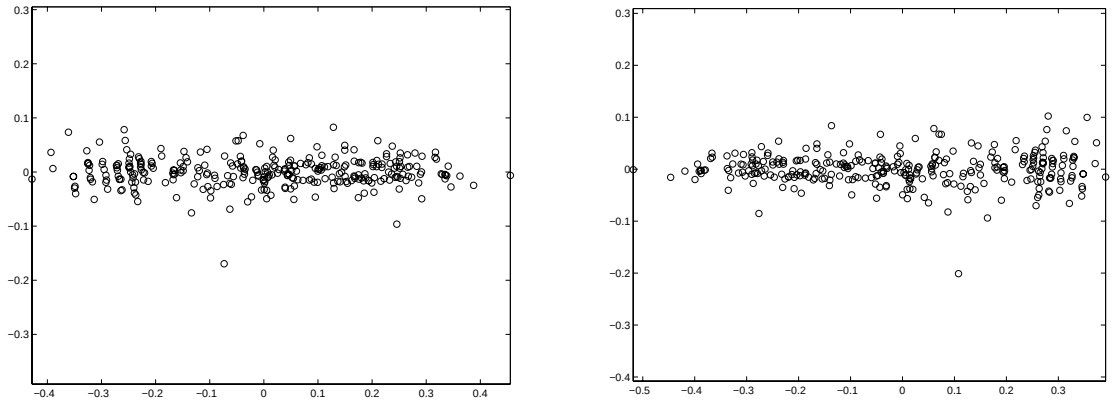


Figure 11: Results of Classical Scaling applied to Halimba data set – first two coordinates of the resulting configuration. A: d_{A}; B: d^2_{M}; C: d_{B}; D: d_{E}.

It is very clear that the Bhattacharyya distance d_{B} and the Euclidean distance d_{E} do not give reasonable results. This two measures do not take into account the ratios between components of the data and therefore are not suitable for measuring distances between compositions (Martín-Fernández et al., 1998a).

To illustrate further this phenomena we have applied Classical Scaling also to the centered Halimba data set – distance matrices calculated with the distances d_{B} and d_{E}. Figure 12 shows obtained configurations. We see here that the results are reasonable

now, extremely similar to the results obtained with the Aitchison's distance – see Figure 11A, or to the original centered configuration – see Figure 10.



A B

Figure 12: Results of Classical Scaling applied to Halimba data set – first two coordinates of the resulting configuration. A: d_{B} distance (centered data set); B: d_{E} distance (centered data set).

Table 3 and Table 4 presents the values of the stress coefficient in the case of Halimba data set. We observe similar behavior to the Lyons West data set. In Table 3 we observe that the accordance between different matrices of distances is better after the centering operation. The stress values shown in Table 4 confirm that all measures considered perform good results applied in the Classical Scaling procedures.

Table 3: Performance of relation between considered dissimilarity measures as measured by the stress for Halimba data set.

Data set	d_{A} *vs* d_{M}	d_{A} *vs* d_{E} before centering	d_{A} *vs* d_{E} after centering	d_{B} *vs* d_{E} before centering	d_{B} *vs* d_{E} after centering
Halimba	0.3297	0.9400	0.7080	0.4387	0.1247

Table 4: Performance of Classical Scaling applied to Halimba data set measured by the stress for considered dissimilarity measures.

Data set	d_{A}	d_{M}	d_{B} before centering	d_{B} after centering	d_{E} before centering	d_{E} after centering
Halimba	8.0×10^{-7}	0.0487	3.1×10^{-4}	4.9×10^{-4}	2.4×10^{-6}	5.0×10^{-6}

4 CONCLUSIONS

Throughout examples of multidimensional scaling applied to real compositional data we have shown

- which measures of difference are compatible (Aitchison's distance, compositional KL dissimilarity) with compositional nature of the data, and
- when and why sometimes also inappropriate, with compositional nature of the data incompatible measures can give reasonable results.

With these examples we have confirmed in practice the theoretical results on (un)suitability of some concrete dissimilarity measures in cases when compositional data are considered. We have shown that cases when inappropriate measures give reasonable results although applied to compositional data are exceptional. Therefore the general conclusion is that only the use of suitable measures (Aitchison's distance and/or compositional KL dissimilarity) will give us adequate results of multidimensional scaling techniques in any case and for any data set of compositions.

We believe that further deep studies are needed to develop all features of measures of difference on compositional data in relation to applications in multidimensional scaling.

5 ACKNOWLEDGMENTS

This work has been partially supported by the

- Dirección General de Enseñanza Superior e Investigación Científica (DGE-SIC) of the Spanish Ministry for Education and Culture through the project BFM-2000-0540;
- Sloven Ministry of Education, Science, and Sport through the research project J1-2391.

In particular, we wish to acknowledge the help of Dr. Ricardo Olea from the Kansas Geological Survey (USA) and of Dr. G. Bárdossy from the Hungarian Academy Sciences for providing, respectively, the Lyons West data set and the Halimba data set.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London (GB): Chapman and Hall. 416 p.

Aitchison, J., C. Barceló-Vidal, J. A. Martín-Fernández, and V. Pawlowsky-Glahn (2000). Logratio analysis and compositional distance. *Mathematical Geology 32*(3), 271–275.

Cox, T. F. and M. A. Cox (1994). *Multidimensional Scaling*. Monographs on statistics and applied probability. London (GB): Chapman & Hall Ltd. 213 p.

Gordon, A. D. (1999). *Classification*. London (GB): Chapman & Hall. (second edition). 256 p.

Martín-Fernández, J. A. (2001, March). *Measures of Difference and Non-parametric Cluster Analysis to Compositional Data*. PhD thesis, Universitat Politècnica de Catalunya, Barcelona (Spain). [Avaible in http://www.tdcat.cesca.es/TDCat-0516101-135345/; and in http://ima.udg.es/∼jamf/].

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (1998b). A critical approach to non-parametric classification of compositional data. In A. Rizzi, M. Vichi, and H. H. Bock (Eds.), *Advances in Data Science and Classification. Proceedings of the 6th Conference of the International Federation of Classification Societies (IFCS-98)*, Berlin, Heidelberg, New York, pp. 49–56. Università La Sapienza, Roma: Springer-Verlag.

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (1998a). Measures of difference for compositional data and hierarchical clustering methods. In A. Buccianti, G. Nardi, and R. Potenza (Eds.), *Proceedings of IAMG'98*, Volume 2, Napoli (Italia), pp. 526–531. The Fourth Annual Conference of the International Association for Mathematical Geology: De Frede Editore.

Martín-Fernández, J. A., M. Bren, C. Barceló-Vidal, and V. Pawlowsky-Glahn (1999). A measure of difference for compositional data based on measures of divergence. In S. Lippard, A. Næss, and R. Sinding-Larsen (Eds.), *Proceedings of IAMG'99*, Volume 1, Trondheim (Norway), pp. 211–215. The Fifth Annual Conference of the International Association for Mathematical Geology: Tapir, Trondheim (N).

Martín-Fernández, J. A., R. Olea-Meneses, and V. Pawlowsky-Glahn (2001). Criteria to compare estimation methods of regionalized compositions. *Mathematical Geology*. (In press).

Mateu-Figueras, G., C. Barceló-Vidal, and V. Pawlowsky-Glahn (1998). Modeling compositional data with multivariate skew-normal distributions. In A. Buccianti, G. Nardi, and R. Potenza (Eds.), *Proceedings of IAMG'98*, Volume 1, Napoli (Italia), pp. 532–537. The Fourth Annual Conference of the International Association for Mathematical Geology: De Frede Editore.

Rao, C. R. (1982). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology 21*, 24–43.