A MEASURE OF DIFFERENCE FOR COMPOSITIONAL DATA BASED ON MEASURES OF DIVERGENCE

J. A. Martín-Fernández[(1)], M. Bren[(2)], C. Barceló-Vidal[(1)] and V. Pawlowsky-Glahn[(3)]

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

[(3)] 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 many statistical methods it is necessary to establish the measure of difference to be used. This measure has to be defined in accordance with the nature of the data. In this study we propose a measure of difference when the data set is compositional. We analyze its properties and we present examples to illustrate its performance.

1. INTRODUCTION

It is well known that the usual dissimilarities and distances are inadequate to measure the difference between two compositonal data (see [6] for further details). In [2] Aitchison proposes that a suitable measure of difference defined on the simplex $\mathcal{S}^D$ should verify two essential requirements: perturbation invariance and subcompositional dominance.

One of the most widely used measures of divergence between two multinomial probability distributions is the *Kullback-Leibler information number*. The purpose of this paper is to propose a measure of difference between two compositional data based on the Kullback-Leibler divergence.

In the next section we define the new measure, we analyze its propierties, and we show that the compositional requirements are verified. Then we expose an interpretation of the measure and present an example to illustrate its performance.

2. MEASURE OF DIFFERENCE BETWEEN TWO COMPOSITIONS

In [2] Aitchison proposes that any scalar measure of difference between two compositions $\mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D$ can be expressed in terms of the ratios of the components. More accurately, a suitable measure of difference should be a function of the compositions $\mathbf{x} \circ \mathbf{x}^{*-1}$ and $\mathbf{x}^* \circ \mathbf{x}^{-1}$, where "$\circ$" simbolizes the perturbation operation introduced in [1]. We call these compositions as the *perturbation differences* between $\mathbf{x}$ and $\mathbf{x}^*$. Note that to the special case $\mathbf{x} = \mathbf{x}^*$ we obtain the perturbation difference $\mathbf{x} \circ \mathbf{x}^{*-1} = \mathbf{e}$, where $\mathbf{e} = (1/D,\ 1/D,\ \ldots,\ 1/D)$ is the center of the simplex $\mathcal{S}^D$.

It is well knwon that a suitable measure is the distance d_A (squared) called Aitchison distance

$$d_A^2(\mathbf{x}, \mathbf{x}^*) \ = \ D \cdot d_E^2(\mathrm{clr}(\mathbf{x}),\ \mathrm{clr}(\mathbf{x}^*)), \tag{1}$$

where d_E represents the Euclidean distance and clr the clr-transformation (see [1] for more details). Because the distance d_A is perturbation invariant we can express (1) as function

of the perturbation differences

$$\begin{aligned} d_A^2(\mathbf{x},\mathbf{x}^*) &= \frac{1}{2}\left(d_A^2\left(\mathbf{e},\mathbf{x}^*\circ\mathbf{x}^{-1}\right) + d_A^2\left(\mathbf{e},\mathbf{x}\circ\mathbf{x}^{*-1}\right)\right) \\ &= \frac{D}{2}\left(d_E^2\left(\mathrm{clr}(\mathbf{e}),\ \mathrm{clr}(\mathbf{x}^*\circ\mathbf{x}^{-1})\right) + d_E^2\left(\mathrm{clr}(\mathbf{e}),\ \mathrm{clr}(\mathbf{x}\circ\mathbf{x}^{*-1})\right)\right). \end{aligned} \quad (2)$$

On the other hand, we can express the Kullback-Leibler information number (see [5]) between two compositonal data $\mathbf{x},\mathbf{x}^*\in\mathcal{S}^D$ as the expression

$$\mathcal{I}(\mathbf{x},\mathbf{x}^*) = \sum_{k=1}^{D} x_k\ \log\left(\frac{x_k}{x_k^*}\right). \quad (3)$$

In equation (3) the factor $\log(x_k/x_k^*)$ is interpreted as the *information gain* in predicting the event $\mathbf{E}_k$ whose probability is x_k by the estimation x_k^*. Then, $\mathcal{I}(\mathbf{x},\mathbf{x}^*)$ is the average information gain given by D events $\mathbf{E}_k$ $(k=1,2,\ldots,D)$. In [3], [8] and [9] related measures for compositional data can be found.

By analogy to the equation (2) we can define on $\mathcal{S}^D$ a measure of difference based on the K-L index (3) as

$$d_K(\mathbf{x},\mathbf{x}^*) = \frac{D}{2}\left(\mathcal{I}(\mathbf{e},\mathbf{x}^*\circ\mathbf{x}^{-1}) + \mathcal{I}(\mathbf{e},\mathbf{x}\circ\mathbf{x}^{*-1})\right). \quad (4)$$

Consequently d_K is a measure proportional to the average information gain of D events $\mathbf{E}_k$ $(k=1,2,\ldots,D)$ whose probability is $1/D$ by the estimation of the components of the perturbation differences. Actually it is possible to expand the definition (4) to other measures of divergence.

Then, we can prove that the measure of difference d_K verifies the following properties

P1. $d_K(\mathbf{x},\mathbf{x}^*) = \frac{D}{2}\log\left(A(\mathbf{x}/\mathbf{x}^*)\cdot A(\mathbf{x}^*/\mathbf{x})\right)$, where $A(\mathbf{x}/\mathbf{x}^*)$ simbolizes the arithmetic mean of the vector of ratios $\mathbf{x}/\mathbf{x}^* = \left(\frac{x_1}{x_1^*},\frac{x_2}{x_2^*},\ldots,\frac{x_D}{x_D^*}\right)$.

P2. Definite: $d_K(\mathbf{x},\mathbf{x}^*)\geq 0, \quad \forall\mathbf{x},\mathbf{x}^*\in\mathcal{S}^D$, and, $d_K(\mathbf{x},\mathbf{x}^*)=0 \iff \mathbf{x}=\mathbf{x}^*$.

P3. Symmetry: $d_K(\mathbf{x},\mathbf{x}^*) = d_K(\mathbf{x}^*,\mathbf{x}), \quad \forall\mathbf{x},\mathbf{x}^*\in\mathcal{S}^D$.

P4. Perturbation invariance: $d_K(\mathbf{p}\circ\mathbf{x},\mathbf{p}\circ\mathbf{x}^*) = d_K(\mathbf{x},\mathbf{x}^*), \quad \forall\mathbf{x},\mathbf{x}^*,\mathbf{p}\in\mathcal{S}^D$.

P5. Subcompositional dominance: $d_K(\mathbf{x_s},\mathbf{x^*_s})\leq d_K(\mathbf{x},\mathbf{x}^*), \quad \forall\mathbf{x},\mathbf{x}^*\in\mathcal{S}^D$, for any subcomposition with $\mathbf{s}$ components.

Therefore the measure d_K defined in (4) is a definite dissimilarity (see [4]) that verifies all the compositional requirements.

3. COMPOSITIONAL PERFORMANCE

A composition can be understood as the vector of probabilities associated to a multinomial distribution, and thus as a probability statement about a finite number of possible states or hypotheses. Consequently, the counterpart of a subcomposition is a conditional multinomial distribution or a conditional probability statement on a subset of the states.

Therefore, compositional dominance could be a desirable and a wellcome property. We note that $\mathcal{I}$ does not have it, but d_K has. Looking for perturbation invariance, we note that the perturbing vector $\mathbf{p}$ can be regarded as the likelihood vector in a Bayesian updating of a probability statement. If two scientists start with probability statements $\mathbf{x}$ and $\mathbf{x}^*$ and both update correctly and Bayesianly with likelihood $\mathbf{p}$, then surely we would not expect any change in the measure of difference between them. Thus, perturbation invariance seems to be a sensible requirement. We see that $\mathcal{I}$ has not perturbation invariance, but d_K has.

On the other hand, in Figure 1 we can see that the d_K dissimilarity has a reasonable behaviour. Note that if the center of the neighbourhoods is near to a border or near to a corner they are crushed, pressed together. This happens because the compositions near to the border or near to the corner have some components close to zero and any small change in its magnitude produces a great change in the ratio of the components.

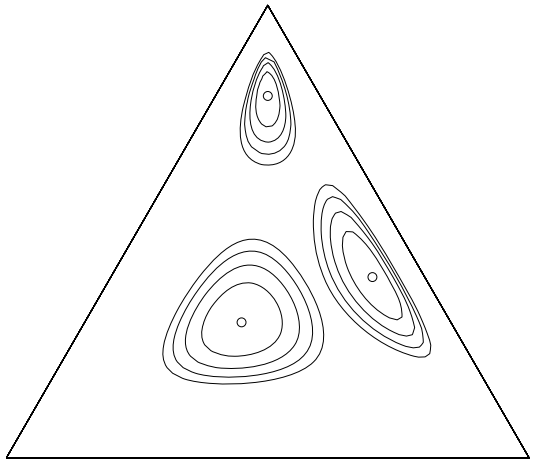

Figure 1: Neighbourhoods in the simplex $\mathcal{S}^3$ with distance d_K.

Now we consider the Metabol data set $\mathbf{X}$ presented in [1] formed by the urinary excretions (mg/24 hours) of 37 normal adults and 30 normal children of

1. x_1: total cortisol metabolites;
2. x_2: total corticosterone metabolites;
3. x_3: total pregnanetriol and Λ-5-pregnentriol.

In Figure 2a we can observe that the data set $\mathbf{X}$ is near to the x_1-corner. This fact happens when one of the components of the data set is near to 1. We consider the center of the data set $\mathbf{X}$ as the compositional geometric mean $cen(\mathbf{X})$. This center is defined as

$$cen(\mathbf{X}) = \frac{(g_1, g_2, \ldots, g_D)}{g_1 + g_2 + \ldots + g_D} \quad , \tag{5}$$

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 $\mathbf{X}$. In our case we obtain $cen(\mathbf{X}) = (0.8230, 0.0886, 0.0884)$. Therefore it is very difficult to establish if there are differences between adults and children in the

realtive patterns of their excretions. If we perturb the data set $\mathbf{X}$ by the perturbation $cen(\mathbf{X})^{-1}$ the result data set are centered, i.e. the center of the set $cen(\mathbf{X})^{-1} \circ \mathbf{X}$ is $\mathbf{e}$, the center of the simplex. Now we can observe in Figure 2b that urinary excretions of adults have a different pattern than children's metabolite.

Then if we choose a measure of difference which is perturbation invariant and we apply a hierarchic method of cluster we obtain the same results for the original data set $\mathbf{X}$ and for the centered data set $cen(\mathbf{X})^{-1} \circ \mathbf{X}$. Figure 3 shows the dendrogram of Ward's method applied to the set $\mathbf{X}$ using the d_K dissimilarity. The classification power is aproximatly equal to 88%. For the Aitchison distance similar results are obtained (see [7] for further examples). If we use the Euclidean distance to cluster the set $\mathbf{X}$ we obtain that the classification power decreases to 64%.

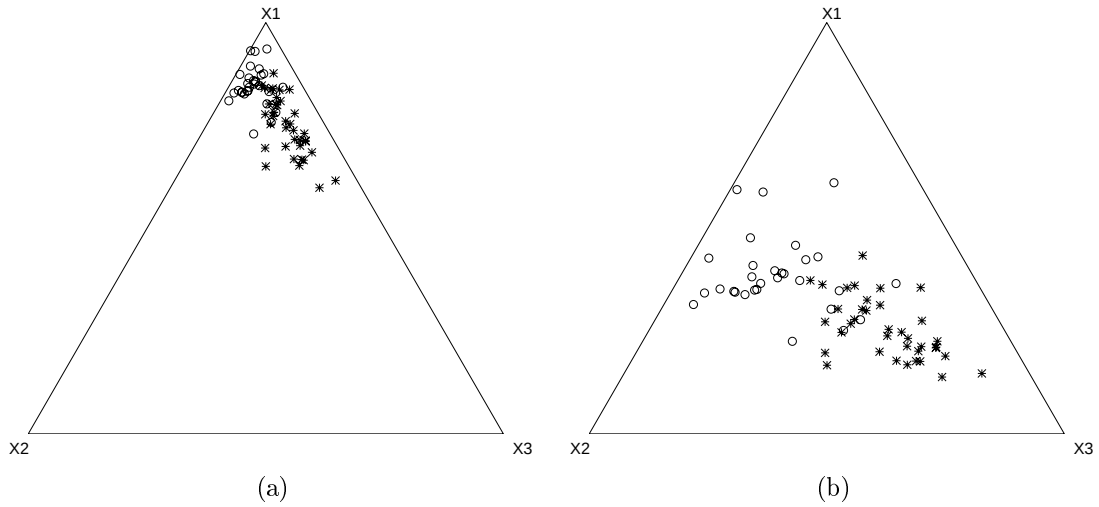

Figure 2: Metabol data set X in the ternary diagram (simbols: '$*$'-adult and '$\circ$'-children) where (a): *Original data;* (b): *Centered data.*

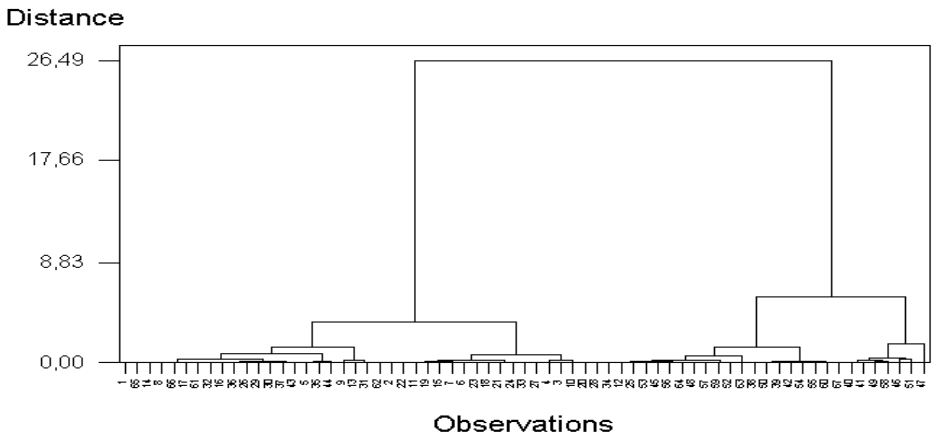

Figure 3: Dendrogram of Ward's method applied to Metabol data set X.

4. CONCLUSIONS

- This new dissimilarity is a possible choice for a measures of difference to compositional data.
- It is necessary to study more deeply the performance of other related measures 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. BREN, M., BATAGELJ, V - The Metric Index, University of Ljubljana (IMMF), Preprint Series, Vol. 35, 561, (1997).
5. 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.
6. MARTÍN-FERNÁNDEZ, J. A., BARCELÓ-VIDAL, C. and PAWLOWSKY-GLAHN, V. - 'Measures of Difference for Compositional Data and HierarchicalClustering Methods', in Proceedings of IAMG'98. The 1998 Annual Conference of the International Association for Mathematical Geology, Ed. Buccianti, A., Nardi,G. and Potenza, R., Napoli (I), Part 2, 1998, pp. 526-531.
7. MARTÍN-FERNÁNDEZ, J. A., BARCELÓ-VIDAL, C. and PAWLOWSKY-GLAHN, V. - 'A Critical Approach to Non-parametric Classification of Compositional Data', in Advances in Data Science and Classification. Proceedings of IFCS'98. The Sixth Conference of the International Federation of Classification Societies, Ed. Rizzi, A, Vichi, M. and Bock, H.H., Roma (I), 1998, pp. 49-56.
8. MEDAK, F., CRESSIE, N.- Confidence regions in ternary diagrams based on the power-divergence statistics, *Math. Geology*, vol. 23, No. 8, pp. 1045-1057, (1991).
9. RAYENS, W.S., SRINIVASAN, C.- Estimation in compositional dat analysis, *Journal of Chemometrics*, vol. 5, pp. 361-374, (1991).