CoDa-dendrogram: A new exploratory tool

J.J. Egozcue[1], **and V. Pawlowsky-Glahn**[2]
[1]Dept. Matemàtica Aplicada III, Universitat Politècnica de Catalunya, Barcelona, Spain;
juan.jose.egozcue@upc.edu
[2]Dept. Informàtica i Matemàtica Aplicada, Universitat de Girona, Spain;
vera.pawlowsky@udg.es

Abstract

The use of orthonormal coordinates in the simplex and, particularly, balance coordinates, has suggested the use of a dendrogram for the exploratory analysis of compositional data. The dendrogram is based on a sequential binary partition of a compositional vector into groups of parts. At each step of a partition, one group of parts is divided into two new groups, and a balancing axis in the simplex between both groups is defined. The set of balancing axes constitutes an orthonormal basis, and the projections of the sample on them are orthogonal coordinates. They can be represented in a dendrogram-like graph showing: (a) the way of grouping parts of the compositional vector; (b) the explanatory role of each subcomposition generated in the partition process; (c) the decomposition of the total variance into balance components associated with each binary partition; (d) a box-plot of each balance. This representation is useful to help the interpretation of balance coordinates; to identify which are the most explanatory coordinates; and to describe the whole sample in a single diagram independently of the number of parts of the sample.

Key words: balance, simplex, Aitchison geometry, composition, orthonormal basis.

1 Introduction

Graphic tools for exploratory analysis of multivariate real data are usually based on a reduction of dimensions. A simple way of reducing dimension from n components to 1 is to take marginals and to describe the sample using univariate tools such as histograms, box-plots, Pareto-plots or simple descriptive statistics (e.g. average, quantiles, standard deviation). The set of the n marginal univariate exploratory analyses may give a preliminary idea of what is in the sample. A pair-wise analysis of the correlation normally complements the marginal analysis. More elaborate tools, such as principal component analysis and the corresponding projection in bivariate plots (bi-plots, scatter plots in two-dimensions) also reduce to one, two, or at most three, dimensional representations. Comparison of these two families of tools for exploratory analysis reveals a difference between them: in the marginal analysis, the meaning and interpretation of each marginal distribution is direct: the analysis is related to the corresponding marginal variable. The situation differs when performing principal component analysis. The meaning of each component is explained either graphically or using the algebraic expression, taking into account the optimality criterion used.

Compositional Data (CoDa) can be represented using different transformations. Traditional transformations are de additive-log-ratio (alr) and the centered-log-ratio (clr) transformations (Aitchison, 1982, 1986). They transform a D-part composition into a $(D-1)$-real-vector and a D-real-vector, respectively, and they can be considered as a kind of coordinates of the composition (Egozcue and Pawlowsky-Glahn, 2005, Appendix). A transformed CoDa sample can then be explored using marginal techniques. However, the variability of the alr coordinates do not give an easy-to-interpret decomposition of the total variability, and the clr coefficients are constrained to add to zero and, consequently, marginal exploratory analyses are also linked by this constraint. Nevertheless, the clr coefficients lead to a decomposition of the variability of the sample, although the directions associated with each part of the variability are not orthogonal (there are D of these directions while the space of the data is only $(D-1)$-dimensional).

Another representation of CoDa is obtained using orthogonal coordinates, like those given by the isometric-log-ratio transformation (ilr) (Egozcue et al., 2003). It assigns a $(D-1)$-real-vector to each compositional vector. These orthogonal coordinates are log-ratios and can be interpreted similarly to alr and clr coefficients, with analogous properties. In particular, the variability of the sample is easily decomposed into variabilities associated with each coordinate. Thus, ilr coordinates can also be used in an exploratory marginal analysis. Nevertheless, although a single coordinate can be easily interpreted, the whole set of them may be difficult to interpret jointly.

Problems related to amalgamation of parts motivated a development of the concept of balances, a specific kind of ilr coordinates (Egozcue et al., 2003). Balances are associated with groups of parts, and they allow a simultaneous interpretation of all the balance-coordinates preserving the decomposition of variability (Egozcue and Pawlowsky-Glahn, 2005). The main result is that each term in the decomposition of the total variability can be assigned either to intra-group (subcompositional) variability or to a balance between two groups of parts. To visualize this, together with other univariate characteristics, a specific tool, the CoDa-dendrogram, has been developed. This CoDa-dendrogram can be built up for any orthonormal basis associated with groups of parts. In particular, similarly to principal component analysis, we can find out an optimal partition that makes the balances as uncorrelated as possible. The CoDa-dendrogram is thus a marginal analysis of balances between groups of parts, but needs a specification of the orthonormal basis used. The dendrogram itself explains the meaning of this basis and the corresponding balances.

2 Groups of parts and balances

The concepts of orthonormal basis in the simplex and balancing element were introduced in Egozcue et al. (2003). The association with a sequential binary partition of the compositional vector can be found in Egozcue and Pawlowsky-Glahn (2005), where balances are discussed and interpreted. Here we summarize those ideas that will be used in the CoDa-dendrogram.

Consider a sample of a random composition of D parts $\mathbf{X} \in \mathcal{S}^D$. Assume the parts are assembled into three non-overlapping groups, represented by the sets of their subscripts. Let $Q = \{1, 2, \ldots, r_1 + r_2\}$ and $\bar{Q} = \{r_1 + r_2 + 1, \ldots, D\}$ be a partition of the subscripts into two groups; and now divide the Q-group into two non-overlapping groups defined by $R_1 = \{1, 2, \ldots, r_1\}$ and $R_2 = \{r_1 + 1, \ldots, r_1 + r_2\}$, so that $R_1 \cup R_2 = Q$, $R_1 \cap R_2 = \emptyset$. We are assuming the grouped parts are consecutive for simplicity, but this is not necessary.

Assume that, at this instance, we are only interested in the information related to the subcomposition defined by Q. The information coming from ratios involving some part with index in $\bar{Q}$ can be removed by a simple orthogonal projection (Egozcue and Pawlowsky-Glahn, 2005), which is equivalent to classical subcompositional analysis (Aitchison, 1986). If $\mathbf{X}$ is partitioned into the three subvectors associated with R_1, R_2 and $\bar{Q}$, $\mathbf{X} = [\mathbf{X}_{R_1}, \mathbf{X}_{R_2}, \mathbf{X}_{\bar{Q}}]$, this projection is

$$\mathcal{C}[\mathbf{X}_{R_1}, \mathbf{X}_{R_2}, \underbrace{A, \ldots, A}_{D-r_1-r_2 \text{ terms}}] \ , \ A = \mathrm{g}(\mathbf{X}_{R_1}, \mathbf{X}_{R_2}) \ ,$$

where $\mathrm{g}(\cdot)$ denotes the geometric mean of the arguments, and $\mathcal{C}$ denotes the closure to a constant κ. Note that $A = \mathrm{g}(\mathbf{X}_{R_1}, \mathbf{X}_{R_2})$ plays de same role as the zero appearing in an orthogonal projection in real space, as the parts in $\bar{Q}$ do not appear.

The information within a group of parts can also be removed by an orthogonal projection. For instance, the vector

$$\mathcal{C}[\underbrace{B_1,\ldots,B_1}_{r_1\text{ terms}},\mathbf{X}_{R_2},\underbrace{A,\ldots,A}_{D-r_1-r_2\text{ terms}}]\ ,\ B_1=\mathrm{g}(\mathbf{X}_{R_1})\ ,$$

does not convey any information about the variability of the subcomposition associated with R_1. In this case, $B_1 = \mathrm{g}(\mathbf{X}_{R_1})$ does not play the role of the zero, but the role of a mean value of the parts. Similarly, we can remove the information about the subcomposition associated with R_2, obtaining

$$\mathbf{X}_{R_1,R_2}=\mathcal{C}[\underbrace{B_1,\ldots,B_1}_{r_1\text{ terms}},\underbrace{B_2,\ldots,B_2}_{r_2\text{ terms}},\underbrace{A,\ldots,A}_{D-r_1-r_2\text{ terms}}]\ ,\ B_2=\mathrm{g}(\mathbf{X}_{R_2})\ .$$

This compositional vector, called *balance* between groups R_1 and R_2, still retains some information. It is essentially one-dimensional, and corresponds to the ratio $B_1/B_2 = \mathrm{g}(\mathbf{X}_{R_1})/\mathrm{g}(\mathbf{X}_{R_2})$. The balance between groups R_1 and R_2 is thus the remaining information from the subcomposition associated with the parts in $Q = R_1 \cup R_2$, after removing the information corresponding to the subcompositions associated with the R_1 and R_2 groups of parts. The balance is itself a projection onto the *balancing* element in $\mathcal{S}^D$

$$\mathbf{e}_{R_1,R_2}=\mathcal{C}\left[\exp\left(\underbrace{\sqrt{\frac{r_2}{r_1(r_1+r_2)}}}_{r_1\text{ equal terms}},\underbrace{-\sqrt{\frac{r_1}{r_2(r_1+r_2)}}}_{r_2\text{ equal terms}},\underbrace{0,\ldots,0}_{D-r_1-r_2\text{ terms}}\right)\right]\ , \tag{1}$$

i.e. $\mathbf{X}_{R_1,R_2} = \langle\mathbf{X},\mathbf{e}_{R_1,R_2}\rangle_a \odot \mathbf{e}_{R_1,R_2}$, where $\langle\cdot,\cdot\rangle_a$ is the Aitchison inner product in the simplex and $\odot$ stands for powering. The quantity $C_{R_1,R_2} = \langle\mathbf{X},\mathbf{e}_{R_1,R_2}\rangle_a$ is the signed magnitude of the projection of $\mathbf{X}_{R_1,R_2}$ onto the balancing element, and is also called balance. When the balancing element $\mathbf{e}_{R_1,R_2}$ is an element of an orthonormal basis we say that C_{R_1,R_2} is a coordinate or a balance-coordinate of $\mathbf{X}$. Balances are easily computed. For the partition of Q into R_1 and R_2 the balance is

$$C_{R_1,R_2}=\langle\mathbf{X},\mathbf{e}_{R_1,R_2}\rangle_a=\sqrt{\frac{r_1r_2}{r_1+r_2}}\ \ln\frac{(\prod_{i\in R_1}X_i)^{1/r_1}}{(\prod_{j\in R_2}X_j)^{1/r_2}}\ , \tag{2}$$

where we easily identify the log-ratio of geometric means of the parts in each group.

As mentioned, a sequential binary partition (SBP) of a compositional vector has an associated orthonormal basis. Its main property is that each coordinate in such a basis is a balance between two groups of parts. The idea of SBP is that the original compositional vector is divided into two non-overlapping groups of parts. In a similar way, each of these two groups is divided again, and so on until all groups contain only a single part. The number of partitions is then $(D-1)$, being D the number of parts of the original composition. Each of the binary partitions has the form of the above partition: a group containing the parts in Q is divided into R_1 and R_2 and the balancing element is then (1). The set of balancing elements generated by a SBP constitutes an orthonormal basis of $\mathcal{S}^D$ and the respective balances are the coordinates of the composition (Egozcue and Pawlowsky-Glahn, 2005).

In order to describe a SBP we normally use a code as the one shown in Table 1. It indicates, e.g., that the first order partition separates the groups $\{1,4,5\}$ from $\{2,3\}$, and that the second order partition subdivides $\{1,4,5\}$ into $\{1,4\}$ and $\{5\}$. At the 4^{th} order partition all groups contain only a single part and the SBP stops.

Given a SBP, a random compositional vector $\mathbf{X}$ in $\mathcal{S}^D$ can be represented by its balance-coordinates in the associated basis. These balance-coordinates are real random variables and we denote them by C_j, $j = 1,2,\ldots,D-1$, arranged in the random array $\mathbf{C}$. Now, the random composition is

$$\mathbf{X}=\bigoplus_{j=1}^{D-1}(C_j\odot\mathbf{e}_j)\ , \tag{3}$$

Table 1: Code for a SBP. The first row represents the labels of parts of the compositional vector. At each order of partition, +1 means that the part is assigned to the first group, −1 to the second group, and 0 that this part is not in the group which is divided at this order.

	Parts of the composition				
order	x_1	x_2	x_3	x_4	x_5
1	+1	−1	−1	+1	+1
2	+1	0	0	+1	−1
3	+1	0	0	−1	0
4	0	+1	−1	0	0

where $\mathbf{e}_j$ are the balancing elements that constitute the orthonormal basis. Since the representation of $\mathbf{X}$ by its coordinates in an orthonormal basis is an isometric transformation called ilr (Egozcue et al., 2003), we denote $\text{ilr}(\mathbf{X}) = \mathbf{C}$, and its inverse $\text{ilr}^{-1}(\mathbf{C}) = \mathbf{X}$, that is equivalent to Equation (3). Relevant properties of ilr-transformations are the expression of the center and the metric or total variance of $\mathbf{X}$ (Pawlowsky-Glahn and Egozcue, 2001; Egozcue and Pawlowsky-Glahn, 2005),

$$\boldsymbol{\gamma} = \text{Cen}[\mathbf{X}] = \text{ilr}^{-1}(\text{E}[\text{ilr}(\mathbf{X})]) = \mathcal{C}[\exp(\text{E}[\ln(\mathbf{X})])] \ ,$$

$$\text{Mvar}[\mathbf{X}] = \sum_{j=1}^{D-1} \text{Var}[C_j] \ . \qquad (4)$$

Equation (4) is an additive decomposition of the total variance into variances associated with each balance-coordinate.

Each balance term, $C_j \odot \mathbf{e}_j$, in (3) is in the one-dimensional space generated by the balancing element $\mathbf{e}_j$ and, consequently, can be represented in a simplex of two parts, $\mathcal{S}^2$, just taking the two-part composition $C_j \odot \mathcal{C}(\exp(1/\sqrt{2}, -1/\sqrt{2}))$. The center of such a composition in $\mathcal{S}^2$ is simply $\mathcal{C}(\exp(\sqrt{2}\text{E}[C_j]), 1)$ and the total variance is $\text{Var}[C_j]$, the variance associated with the j-th balance in (4). The parametric description of the balances may include the correlations or covariances between balance-coordinates, $\text{Corr}[C_j, C_k]$, for $j, k = 1, 2, \ldots, D-1$.

Whenever a n-sample of $\mathbf{X}$ is available, all mentioned moments of the coordinate-balances, i.e. center, total variance and correlations, can be estimated from the sample and we obtain their sample versions. Given a SBP or, equivalently, an orthonormal basis, the compositional sample $\mathbf{x}_k = [x_{1k}, x_{2k}, \ldots, x_{Dk}]$, $k = 1, \ldots, n$, can be ilr-transformed into the respective sample coordinates $\mathbf{c}_k = [c_{1k}, c_{2k}, \ldots, c_{D-1,k}]$, $k = 1, \ldots, n$, using (2). These sample-coordinate-balances allow the estimation of the coordinates of the center, variances and covariances using standard techniques, for instance, for the j-balance variance one usual estimator is

$$\widehat{\text{Var}[C_j]} = \frac{1}{n} \sum_{k=1}^{n} c_{jk}^2 - \bar{c}_j^2 \ ,$$

where $\bar{c}_j = (1/n) \sum_i c_{ij}$ is the standard estimator of the mean.

The CoDa-dendrogram is intended to represent in a comprehensive plot most of the information contained in the SBP. This is the goal of the next section.

3 Elements of a CoDa-dendrogram

The goal of the CoDa-dendrogram is to represent simultaneously the following features: (a) the SBP defining the balances, (b) the sample-center, (c) the decomposition of the sample total variance, and (d) some characteristics of the empirical balance distribution. Correlations between balances could also be represented, but such a representation is very difficult to visualize, and are therefore not included.

Figure 1 shows a scheme of a CoDa-dendrogram. At the left-hand side the names of the parts have been reordered to easily visualize the grouping of parts defined by the SBP given in Table 1. The vertical segments describe the groups formed at each order of the SBP. In this sense, it is similar to a typical cluster dendrogram. Therefore, the length of the vertical lines do not contain any quantitative information; they are as long as required to connect the groups of parts. These vertical segments can be associated with the order of the SBP in which the groups have been separated, as shown below.

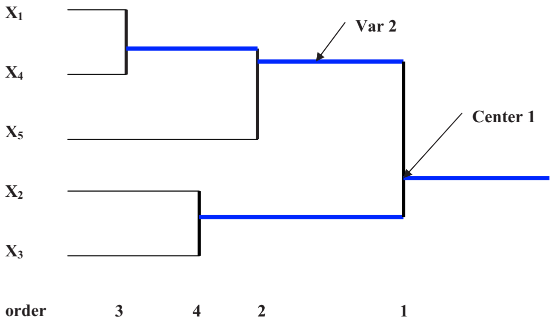

Figure 1: Scheme of a CoDa-dendrogram. Sequential binary partition defined in Table 1.

Vertical segments can be used to visualize mean balances and other distributional characteristics. There are two different scaling options for the vertical segments: (a) each vertical segment is identified with a 2-part simplex $\mathcal{S}^2$, i.e. the segment (0,1); (b) vertical segments are scaled as a part of the real axis of the balance. In option (a) the lower end of each vertical line is identified with 0 and the upper end with 1, or with any other suitable closure constant like 100. For each j-th order of the SBP, the center in $\mathcal{S}^2$ of the j-th balance,

$$\frac{\exp(\sqrt{2}\ \bar{c}_j)}{1+\exp(\sqrt{2}\ \bar{c}_j)}\ ,$$

can be plotted in the j-th vertical segment. Each horizontal segment starts from this center point, goes rightwards, and finishes at the end point of a vertical segment of a lower order of the SBP. Accordingly, we visualize the center of the compositional sample as the points from which horizontal segments, going to the right, start (see Fig. 1). We have to be aware that this representation corresponds to the Aitchison metrics in $\mathcal{S}^2$. For instance, a center-point placed at 0.5 means that the two groups balance to 0 or, equivalently, that the overall geometric means within the two groups of parts are equal. A center-point near the upper limit of the segment means, that the geometric mean of the parts within the group hanging from the upper limit is much greater than the geometric mean of the parts within the group hanging from the lower limit of the vertical segment.

Table 2: Sequential binary partition used in benchmark problem 2 by J. Aitchison, described by ± code.

	Parts of the composition					
order	X_1	X_2	X_3	Y_1	Y_2	Y_3
1	+1	+1	+1	−1	−1	−1
2	+1	−1	−1	0	0	0
3	0	+1	−1	0	0	0
4	0	0	0	+1	−1	−1
5	0	0	0	0	+1	−1

Option (b) treats the vertical lines as an interval, $(-u, u)$, of the real axis of a balance-coordinate. The starting point of the horizontal rightward segment is placed at the mean balance if it is in $(-u, u)$ or just in $\pm u$ if the mean balance is outside. The interval $(-u, u)$ is also used to plot quantiles or box-plots of the sample-balances. The two options represent the same information in different scales: option (a) in a binary compositional diagram with its Aitchison metric scale, and option (b) in the real balance scale.

Horizontal segments are used to represent the decomposition of the total variance given in Equation (4). The length of each horizontal segment to the right is the sample variance of the balance associated with the separation of the corresponding groups. The total variance is then the sum of all horizontal segments to the right. A short horizontal segment means that the balance has a small variability in the sample, thus explaining only a little bit of the total variance. Conversely, a long horizontal segment implies a balance explaining a good deal of the total variance.

Finally, vertical segments can also be useful to visualize the sample distribution of balances. For instance, we can plot distinguished quantiles, e.g. the median, the quartiles and the 0.05 and 0.95 quantiles (used here) or other kinds of box-plot. Differences between the median and the center (the starting point of the horizontal segments) gives an idea of asymmetry of the distribution of the balance.

The whole CoDa-dendrogram may be viewed as a *mobile*. Imagine you pick up the right end of the CoDa-dendrogram and you maintain it hanging downwards. Assume that sample-balances are unit masses on the vertical lines (they are horizontal when hanging), placed on their corresponding values in the scaled segment. The links between vertical and horizontal bars (mean balances) are placed so that the mobile is in equilibrium. In this image, the variance bars (now vertical) have no mass. On the other hand, the lengths on the segments where the masses are placed should be measured properly: in the balance scale (b) they are just as they appear; in option (a) the distances should be measured as Aitchison distances. Now, moments of masses, arm times mass, add to zero for equilibrium.

4 Example

J. Aitchison proposed some benchmark problems to CoDaWork'05. Problem 2, *A differential diagnostic problem using blood compositions*, is used to illustrate the CoDa-dendrogram. The problem assumes that the blood of two samples, of 25 individuals each, is analyzed and that 6 different blood extracts have been quantified for each patient. The extracts are classified into two groups of parts, namely X_1, X_2, X_3 and Y_1, Y_2, Y_3. The main goal is to determine which parts distinguish between the two populations, A-patients and B-patients, from which the samples have been taken. In order to represent the data using a CoDa-dendrogram, we define a first order partition following the groups suggested in the problem, the X's and the Y's. Afterwards, we proceed with arbitrary partitions. Table 2 shows the code of the selected SBP. The structure of this SBP is also shown in the CoDa-dendrogram of Figure 2. This CoDa-dendrogram shows the

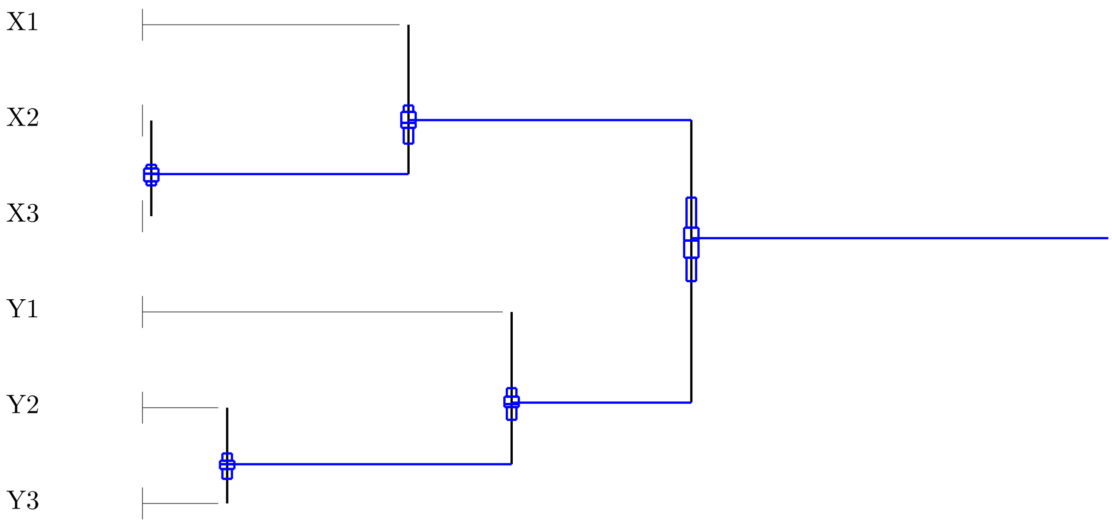

Figure 2: CoDa-dendrogram of A-patients sample. Vertical bars scaled in (-3,3).

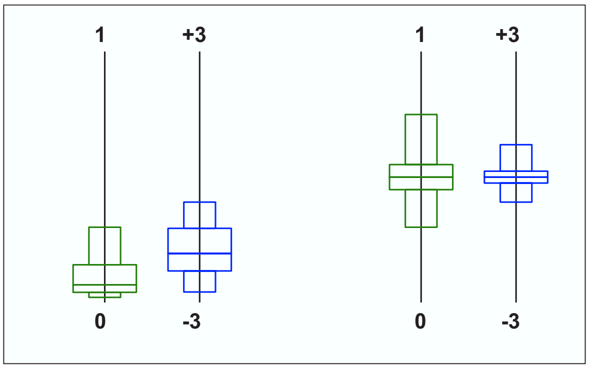

Figure 3: Comparison of 0.05, 0.25, 0.50, 0.75 and 0.95 quantiles in $\mathcal{S}^2 \equiv (0,1)$ scale (green) and balance scale $(-3,+3)$ (blue). Small proportions (left); about 0.5 proportion (right).

characteristics of the A-patients sample. Blue horizontal bars are proportional to the variance of each balance and their length add to the total variance of the sample. The largest variance corresponds to the first balance comparing the X's and the Y's. Other balances also exhibit large variabilities. The vertical segments are assumed to represent the interval $(-3,3)$ of the real balance axis. Accordingly, the intersect point with the blue segment to the right is the sample mean balance. On the vertical bars, blue box-plots are defined by the 0.05, 0.25, 0.50, 0.75 and 0.95 quantiles of the sample balances. For instance, the mean balance between the X's and the Y's points out larger relative presence of X-extracts (short arm) than of Y-extracts (long arm). However, the variability of this balance indicates that this relationship is reversed for some A-patients, as the box-plot covers an interval of negative values (lower half-bar). We also see that the sample distribution of this first balance is quite symmetric, as so is the box-plot and there is no large difference between the mean and the median. The balance between Y_2 and Y_3, which is proportional to their log-ratio, shows the larger presence of Y_3 compared to Y_2.

The CoDa-dendrogram in Figure 2 could be presented with vertical bars scaled as a $\mathcal{S}^2$ simplex, but the visual differences would not be very large. Figure 3 compares quantile plots in both scales.

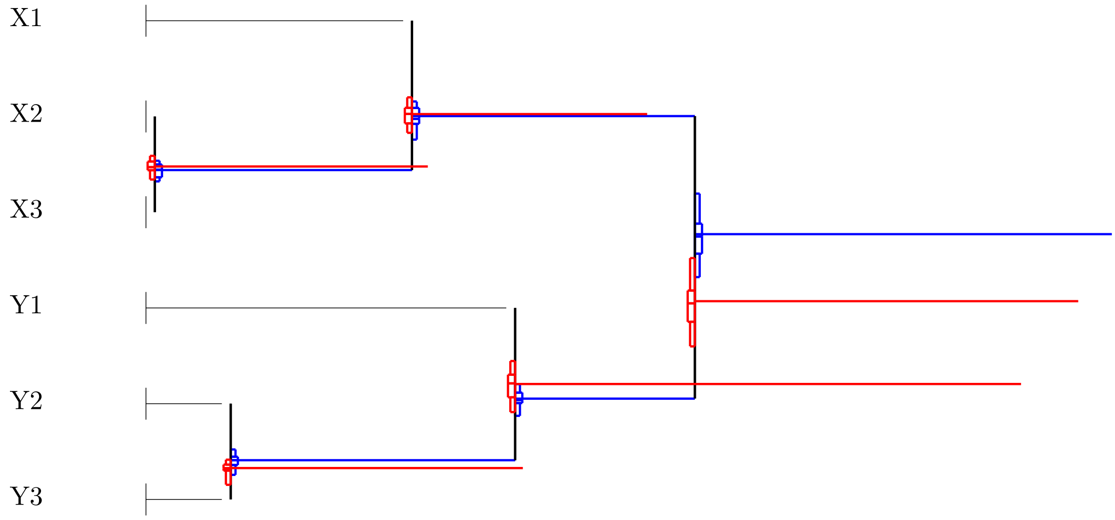

Figure 4: CoDa-dendrogram of two samples: A-patients, in blue, and B-patients in red. Vertical bars scaled in (-3,3).

At the left hand side, a box-plot with median 0.07 (per-unit scale, green) is compared with its equivalent in balance scale (median -1.83, blue). Low quantiles in per-unit scale are closer than the corresponding ones in the balance scale, and the latter appear as a more symmetric distribution. At the right-hand side a similar comparison is presented, but now the median is 0.5 in per-unit scale, and 0.0 in the balance scale. Per-unit scale in $\mathcal{S}^2$ may be useful because any value in per-units remains in $(0, 1)$, but at the price of fictitious asymmetry and compression of quantiles. The balance scale is easier to interpret, but it cannot cover the whole real line and some values may fall outside the scaled segment.

As the stated problem aims at the comparison between A-patients and B-patients, a superposition of the two corresponding CoDa-dendrograms may be useful. Figure 4 shows the original CoDa-dendrogram for A-patients (blue) and, superimposed, the characteristics of the B-patients in red. The second CoDa-dendrogram does not exactly match the tree of the previous one because balance means, variances and sample distributions may be different. The differences are now easily visualized. For instance, the balance between $\{Y_1\}$ and $\{Y_2, Y_3\}$ has more than double variance in B-patients than in A-patients, whereas other balance variances are quite similar in both samples. Therefore, a first observation is that the increase of total variance in B-patients is mainly due to the mentioned balance.

In order to classify a new patient into one of the two classes A and B, we are interested in the mean behavior of the two samples. Figure 4 reveals that there are two sample mean balances apparently different: the first balance between X's and Y's and also the balance between $\{Y_1\}$ and $\{Y_2, Y_3\}$. The significance of different mean balances can be checked using standard ANOVA techniques. The mentioned balances are clearly significant if tested for equal means, as expected from the respective box-plots; other balances are less significant. The classification of a new patient can be based on the values of these two balances, as the other balances are almost irrelevant.

The problem also asks for discrimination of patients based on only four extracts: the X's, as they are, and an amalgam of the Y's, $AY = Y_1 + Y_2 + Y_3$. Obviously, the discrimination based on the balance of $\{Y_1\}$ against $\{Y_2, Y_3\}$ is lost in this approach, but the balance of the X's against AY may retain a good deal of discrimination power. Figure 5 shows two-sample CoDa-dendrograms for this four-part composition of blood extracts (A-patients in blue, B-patients in red). The first balance, X's against AY, still shows significant differences in mean, and the subcomposition of the X maintains its low discriminatory power.

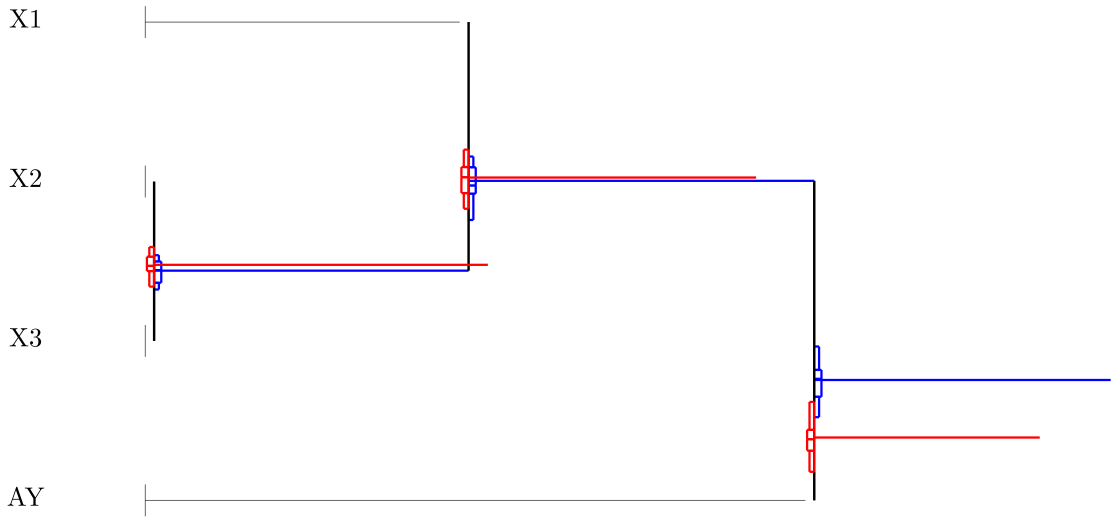

Figure 5: CoDa-dendrogram of two samples: A-patients, in blue, and B-patients in red, after amalgamation of Y-parts. Vertical bars scaled in (-3,3).

5 Conclusion

Compositional data analysis is mainly based on the study and interpretation of log-ratios. Balances are a particular case of log-ratios with a particular interpretation due to their relationship with groups of parts. The CoDa-dendrogram uses the representation of compositions by balances. It is a powerful descriptive tool, as it allows the simultaneous visualization of an orthonormal basis of balancing elements, the induced decomposition of the total variance, the sample mean values, and some quantiles of the balance distribution. The CoDa-dendrogram is able to summarize sample information even when compositional vectors have a large number of parts.

Balances can be selected by the user in order to improve interpretability of the results when some kind of affinity between parts is *a priori* stated or desired. The groups of parts are the result of a sequential binary partition of the whole compositional vector. This process of partition into groups may be difficult to describe for high-dimensional problems. The CoDa-dendrogram visualizes this sequential binary partition as a binary clustering of parts, leading to a more intuitive representation.

Two-sample CoDa-dendrograms can be used for a preliminary comparison of sample balance means and variances. For mean comparisons between two samples, it can identify which balances have significant different means and which may be irrelevant.

Acknowledgements

This research has received financial support from the *Dirección General de Investigación* of the Spanish Ministry for Science and Technology through the project BFM2003-05640/MATE.

REFERENCES

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology) 44*(2), 139–177.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

Egozcue, J. J. and V. Pawlowsky-Glahn (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology 37*(7), 799–832.

Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology 35*(3), 279–300.

Pawlowsky-Glahn, V. and J. J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment (SERRA) 15*(5), 384–398.