Austrian Journal of Statistics
September 2016, Volume 45, 25–44.
http://www.ajs.or.at/
doi:10.17713/ajs.v45i4.126

Changing the Reference Measure in the Simplex and Its Weighting Effects

Juan José Egozcue
Universitat Politècnica de Catalunya, Spain

Vera Pawlowsky-Glahn
Universitat de Girona, Spain

Abstract

Under the assumption that the Aitchison geometry holds in the simplex, standard analysis of compositional data assumes a uniform distribution as reference measure of the space. Changing the reference measure induces a weighting of parts. The changes that appear in the algebraic-geometric structure of the simplex are analysed, as a step towards understanding the implications for elementary statistics of random compositions. Some of the standard tools in exploratory analysis of compositional data, such as center, variation matrix and biplots are studied in some detail, although further research is still needed. The main result is that through a progressive down-weighting of some parts, the geometry of the space approaches that of the corresponding subcomposition. In this way, the coherence between standard and down-weighted analyses is preserved.

Keywords: simplex, sigma-additive measures, subcomposition, weighting, Bayes space, biplot, center, variability.

1. Introduction

When analysing a composition, some parts may heavily influence the results. A typical example are inaccuracies in the measurements in some not fully relevant parts. They can dominate the analysis, producing a large contribution to variability or to distances. Also, relevance of some parts in a given problem can call for weighting techniques to adapt the simplex geometry accordingly. There are a number of weighting techniques that can be useful in this sense (e.g. Filzmoser and Hron 2015). Among them, the change of reference measure of the simplex has several implications that need to be fully understood for a consistent analysis. This contribution is aimed at showing changes that appear in the algebraic-geometric structure of the simplex, as well as some effects in elementary statistics and exploratory tools.

One of the most fruitful concepts in compositional analysis is that of subcomposition (Aitchison 1986). In Aitchison (1992), some reasonable principles for a coherent analysis of subcompositions were established. Beyond the idea that compositional analyses should be scale invariant, those principles included the assumption that distances between compositions should be greater than or equal to those observed in a subcomposition. This principle, called subcompositional dominance (Aitchison 1992; Aitchison, Barceló-Vidal, Martín-Fernández, and Pawlowsky-Glahn 2000; Egozcue 2009), highlights a change of the geometry of subcomposi-

tions (for instance, a change in inter-distances between two data-points in the subcomposition) with respect to the original geometry of the full composition. Taking a subcomposition can be considered as an extreme case of down-weighting, since the influence of some parts of the composition is removed from the analysis. However, there are cases in which the complete removal of the influence of some parts of the original composition is not desirable. This motivates the idea of weighting compositions as a continuous transition from the full composition, endowed with the corresponding Aitchison geometry (Pawlowsky-Glahn and Egozcue 2001), to a subcomposition, endowed with the induced Aitchison geometry, which differs in dimension and metrics (distances, inner product, norm).

Apparently, there are many ways of weighting compositions so that a transition from a full composition to a subcomposition is performed. However, fulfilling all coherence requirements is quite challenging. One option deserving attention is the one proposed for Bayes spaces (Boogaart, Egozcue, and Pawlowsky-Glahn 2010; Egozcue, Pawlowsky-Glahn, Tolosana-Delgado, Ortego, and Boogaart 2013b) and, more specifically, for Bayes Hilbert spaces (Egozcue, Díaz-Barrero, and Pawlowsky-Glahn 2006; Boogaart, Egozcue, and Pawlowsky-Glahn 2014). Bayes Hilbert spaces are spaces of measures and densities, and their algebraic-geometric structure is an extension of the Aitchison geometry of the simplex. In fact, in (Boogaart *et al.* 2014), it is shown that the simplex, endowed with the Aitchison geometry, is a particular case of a Bayes Hilbert space. In the development of Bayes Hilbert spaces, a reference probability measure is introduced as a parameter regulating the geometry of the measures and densities in the space. This kind of approach provides a way of coherently introducing weighting strategies, both in the simplex and in the analysis of compositional data. The present aim is to start studying the change of reference measure in the simplex, being conscious that there is a long way from the general theory of Bayes Hilbert spaces to applications in compositional data analysis. Special attention is paid to the transition from the geometry of the simplex $\mathcal{S}^D$ for compositions to the geometry of $\mathcal{S}^d$, $d < D$, where subcompositions are defined. The main difficulties are interpretative, as usual in compositional data analysis.

The structure of the paper is as follows: Section 2 translates the milestones of Bayes Hilbert spaces into the case of compositions, with special emphasis on the role of the reference measure. Section 3 introduces the centered log-ratio transformation (clr) with respect to an arbitrary reference measure in the simplex, following the definition in Boogaart *et al.* (2014) for general Bayes Hilbert spaces. Section 4 gets into details of metric concepts under a change of the reference measure, such as orthogonality, bases, and balances. A proposition on dominance of distances is there stated (see proof in Appendix A). Section 5 gives an introduction to distributions of random compositions, their variability and centre under a weighted geometry of the simplex. Section 6 shows how variation matrix and biplots work under weighting using an example of electoral results.

2. Change of reference measure for compositions

Consider D categories $c_1, c_2, \ldots, c_D$; they represent a partition of a measurable space Ω. A D-part composition $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ in the D-part simplex $\mathcal{S}^D$ assigns a proportion x_i to the category c_i. Assuming that the composition $\mathbf{x}$ is closed to 1, the proportion assigned to the whole space Ω is just 1. For any subset of categories, the proportion assigned is the sum of the corresponding proportions. For instance, the proportion assigned to the subset $\{c_1\}$ is x_1, and the proportion assigned to the subset $\{c_1, c_2\}$ is $x_1 + x_2$. From this point of view, the composition $\mathbf{x}$ defines a finite additive measure on Ω, which is denoted $\mu_{\mathbf{x}}\{\cdot\}$. The argument of this measure is any subset of Ω. Examples are $\mu_{\mathbf{x}}\{\Omega\} = 1$, $\mu_{\mathbf{x}}\{\emptyset\} = 0$, $\mu_{\mathbf{x}}\{c_1\} = x_1$, $\mu_{\mathbf{x}}\{c_1, c_2\} = x_1 + x_2$.

Measures can be represented by densities. The idea is that sums (integrals) on a subset of Ω give the measure of this subset. In the case of the simplex $\mathcal{S}^D$, the density is identified with

the composition $\mathbf{x}$, as for any subset $A \subseteq \Omega$ it satisfies

$$\mu_{\mathbf{x}}\{A\} = \sum_{c_i \in A} x_i P_0\{c_i\} \quad , \quad P_0\{c_i\} = 1 \ , \ i = 1, 2, \ldots, D \ ,$$

where the uniform measure $P_0\{\cdot\}$ on Ω has been made explicit as reference measure. Note that $P_0\{\Omega\} = D$ and addends of sums (integrals) along the composition are equally weighted with $1 = P_0\{c_i\}$. The reference measure specified as $\mathbf{p}_0 = (P_0\{c_1\}, P_0\{c_2\}, \ldots, P_0\{c_D\})$ is a non-closed uniform measure. Therefore, it is compositionally equivalent to the neutral element of the simplex $\mathbf{n} = (1/D, 1/D, \ldots, 1/D)$. The conclusion is that a composition $\mathbf{x} \in \mathcal{S}^D$ defines a measure $\mu_{\mathbf{x}}$ on Ω specifying the measure of each elementary subset $\{c_i\}$ and, at the same time, $\mathbf{x}$ is the density of $\mu_{\mathbf{x}}$ with respect to the uniform reference measure P_0, which density is $\mathbf{p}_0$. In mathematical terms, the density (composition) $\mathbf{x}$ is the Radon-Nikodym derivative of $\mu_{\mathbf{x}}$ with respect to the reference measure P_0 which can be written as

$$\mathbf{x} = \frac{d\mu_{\mathbf{x}}}{dP_0} \quad , \quad \mu_{\mathbf{x}}\{A\} = \int_A \frac{d\mu_{\mathbf{x}}}{dP_0} \, dP_0 = \sum_{c_i \in A} x_i \ P_0\{c_i\} \ ,$$

for any $A \subseteq \Omega$. When P_0 is the unitary and uniform reference measure, there is no need to distinguish between $\mathbf{x}$ as a composition, as a measure or as a density. These facts change when weights are introduced through the reference measure.

To analyse the effects of a change of reference measure as a means to introduce weights, consider an arbitrary array of positive weights, $\mathbf{p} = (p_1, p_2, \ldots, p_D)$. The corresponding measure P is then characterised by $P\{c_i\} = p_i$, for $i = 1, 2, \ldots, D$, and by the measure of the whole space, $P\{\Omega\} = \sum_{i=1}^{D} p_i$. Note that $\mathbf{p}$ is the density of P with respect to the uniform measure P_0. A question is now to look for the density of the measure $\mu_{\mathbf{x}}$ with respect to the new reference measure P. This density is $\mathbf{y} = \mathbf{x}/\mathbf{p} = (x_1/p_1, x_2/p_2, \ldots, x_D/p_D)$. In fact, for $A \subseteq \Omega$,

$$\mu_{\mathbf{x}}\{A\} = \sum_{c_i \in A} x_i = \sum_{c_i \in A} y_i \ p_i = \sum_{c_i \in A} \frac{x_i}{p_i} \ p_i \, . \tag{1}$$

The measure $\mu_{\mathbf{x}}$ is thus retrieved from two different densities, $\mathbf{x}$ when considering the uniform reference P_0, and $\mathbf{y}$ for a reference P. Note that $\mathbf{y}$ is a vector which components do not add to one, i.e. it is not closed. However, it is compositionally equivalent to $\mathcal{C}\mathbf{y} = \mathbf{x} \ominus \mathbf{p}$, as its components are proportional (Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado 2015).

If the reference measure P is represented by the vector of weights $\mathbf{p}$, the composition $\mathcal{C}\mathbf{y}$ is just a perturbation of $\mathbf{x}$, a shift in the simplex, recalling that the perturbation-difference $\ominus$ includes the closure, $\mathcal{C}$, and, consequently, $\mathcal{C}\mathbf{y} = \mathbf{x} \ominus \mathbf{p} = \mathbf{x} \ominus \mathcal{C}\mathbf{p}$. From now on, the non-closed version of $\mathbf{y}$ is denoted $\mathbf{y}^{(\mathbf{p})}$ when the reference measure needs to be specified. Following Boogaart *et al.* (2010) and Boogaart *et al.* (2014), a weighted perturbation and powering can be defined for densities like $\mathbf{y}^{(\mathbf{p})}$ such that they operate linearly in the weighted simplex. However, their use is not recommended in this context as standard perturbation ($\oplus$) and powering ($\odot$) are easily interpreted and computed in the applications. This avoids linear operations with the shifted densities $\mathcal{C}\mathbf{y}^{(\mathbf{p})} = \mathbf{x} \ominus \mathbf{p}$. In practice, weighted compositions will be used only in the computation of distances and inner products, as explained below.

3. Centred log-ratio with respect to a reference measure

In Boogaart *et al.* (2014), the clr-transformation of a density f with respect to a given reference measure P, is defined as

$$\mathrm{clr}_P(f)(x) = \log f(x) - \frac{1}{P\{\Omega\}} \int_\Omega \log f(\xi) \ dP\{\xi\} \ , \quad x \in \Omega \ , \tag{2}$$

where Ω is the measurable set where the density f is defined. In the present case, Ω is the set of the D parts or categories of $\mathcal{S}^D$, namely c_i, $i = 1, 2, \ldots, D$. Therefore, the values of x

in such an expression correspond to the c_i's. Since f is a density of a measure with respect to the reference measure P, it can be identified with the density $\mathbf{y} = \mathbf{x}/\mathbf{p}$, as introduced in Section 2. With these identifications, the $\mathrm{clr}_\mathbf{p}$-transformation of the simplex with respect to the measure P, represented by $\mathbf{p} = (p_1, p_2, \ldots, p_D)$, is

$$\mathrm{clr}_\mathbf{p}(\mathbf{y}) = \left(\log \frac{y_1}{\mathrm{g}_\mathbf{p}(\mathbf{y})} \ , \ \log \frac{y_2}{\mathrm{g}_\mathbf{p}(\mathbf{y})} \ , \ldots , \log \frac{y_D}{\mathrm{g}_\mathbf{p}(\mathbf{y})} \right) , \quad \mathrm{g}_\mathbf{p}(\mathbf{y}) = \exp\left(\frac{1}{s_\mathbf{p}} \sum_{i=1}^{D} p_i \log y_i \right) \ , \quad (3)$$

where $s_\mathbf{p} = \sum_{i=1}^{D} p_i$, and $\mathrm{g}_\mathbf{p}(\cdot)$ denotes a weighted geometric mean of the parts y_i. It is remarkable that $\mathbf{p}$, the reference measure of the categories c_i, is not closed to D, and that $P\{\Omega\} = s_\mathbf{p}$, while for P_0 the uniform reference measure $s_{\mathbf{p}_0} = D$. Note also that $\mathbf{y}$ can be closed or not, as Equation 3 is scale invariant.

An important characteristic of $\mathrm{clr}_\mathbf{p}(\mathbf{y})$ is that the weighted sum of its D components is zero, that is

$$\sum_{i=1}^{D} p_i \log \frac{y_i}{\mathrm{g}_\mathbf{p}(\mathbf{y})} = 0 \ , \quad (4)$$

generalising the ordinary clr in $\mathcal{S}^D$, for which the sum of its components (weights equal to 1) is zero. This has a geometric interpretation in the space $\mathbb{R}^D$, where a point has coordinates $\log(\mathbf{y}) = (\log y_1, \log y_2, \ldots, \log y_D)$. As illustrated in Figure 1, which shows a scheme for $D = 2$, to obtain the ordinary clr of a generic point $\log(\mathbf{y})$, the point is orthogonally projected onto a hyperplane through the origin whose orthogonal vector is $(1, 1, \ldots, 1)$ (Aitchison 1986; Pawlowsky-Glahn *et al.* 2015). When using a non-uniform $\mathbf{p} = (p_1, p_2, \ldots, p_D)$ the procedure to get $\mathrm{clr}_\mathbf{p}(\mathbf{y})$ is to orthogonally project the point $\log(\mathbf{y})$ onto a hyperplane whose orthogonal vector is $\mathbf{p}$, as shown by the inner product in $\mathbb{R}^D$ implicit in Equation 4. Summarising, $\mathrm{clr}_\mathbf{p}$ is a projection of $\log(\mathbf{y})$ on a hyperplane whose normal vector is $\mathbf{p}$.

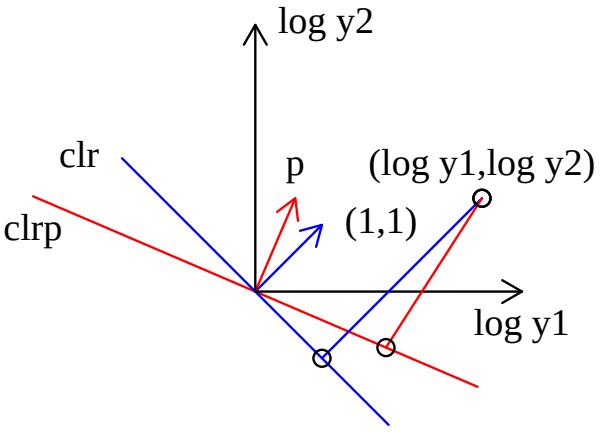

Figure 1: Generic 2-part composition (y_1, y_2), log-transformed into $(\log y_1, \log y_2)$. Two reference measures with densities $(1, 1)$ (uniform, blue arrow) and $\mathbf{p}$ (red arrow) are considered. The point $(\log y_1, \log y_2)$ is projected, parallel to the reference arrow, on the clr-plane (blue) and on the $\mathrm{clr}_\mathbf{p}$-plane (red), thus obtaining the respective transformations.

A particular case of interest is that of

$$p_i = 1 \ , \ i = 1, 2, \ldots, D-1 \ , \ p_D = \epsilon \ , \quad (5)$$

for which $P\{\Omega\} = (D-1) + \epsilon$. When $\epsilon \to 0$, the D-th part is down-weighted from 1 to $\epsilon \ll 1$. For small enough ϵ, the weighted geometric mean $\mathrm{g}_\mathbf{p}$ in Equation 3 approaches the ordinary geometric mean of the first $D-1$ parts of $\mathbf{y}$. A consequence is that the first $D-1$ components of $\mathrm{clr}_\mathbf{p}(\mathbf{y})$ approach the ordinary clr of the subcomposition formed by $(y_1, y_2, \ldots, y_{D-1})$. This

suggests that this kind of reference measures may approach the induced Aitchison geometry on the subcomposition.

4. Metrics under change of reference

The clr transformation can be used to define the inner product in $\mathcal{S}^D$, as was done in Bayes Hilbert spaces (Boogaart *et al.* 2014, Def. 2). There, the proposed definition was

$$\langle \mathbf{y}_1, \mathbf{y}_2 \rangle_{B^2} = \frac{1}{P\{\Omega\}} \langle \mathrm{clr}_P(\mathbf{y}_2), \mathrm{clr}_P(\mathbf{y}_1) \rangle \ ,$$

where $\langle \cdot, \cdot \rangle$ is the ordinary inner product in $\mathbb{R}^D$. This definition leads to an inner product in $\mathcal{S}^D$ which, for a uniform reference measure P_0, with weights $\mathbf{p}_0 = (1, 1, \ldots, 1)$, is

$$\langle \mathbf{y}_1, \mathbf{y}_2 \rangle_a = \frac{1}{D} \langle \mathrm{clr}(\mathbf{y}_2), \mathrm{clr}(\mathbf{y}_1) \rangle \ , \tag{6}$$

which is not the standard in compositional data analysis due to the factor $1/D$. This inner product is not suitable for compositional data analysis, as it does not fulfill the principle of subcompositional dominance of distances. For instance, consider the 3-part compositions $\mathbf{u} = (0.1, 0.7, 0.2)$ and $\mathbf{v} = (1/3, 1/3, 1/3)$. Their distance, in the geometry induced by the inner product (6) in $\mathcal{S}^3$, is $\mathrm{d}_3(\mathbf{u}, \mathbf{v}) = 0.805$. Taking the subcomposition formed by the first and second part and computing the distance in $\mathcal{S}^2$ according to (6), the result is $\mathrm{d}_2(\mathbf{u}, \mathbf{v}) = 0.973$. Since $\mathrm{d}_3(\mathbf{u}, \mathbf{v}) < \mathrm{d}_2(\mathbf{u}, \mathbf{v})$, the principle of subcompositional dominance is violated.

The discussion about the role of the constant $1/D$ in the inner product is related to the fact that in Boogaart *et al.* (2014) the reference measure was assumed to satisfy $P_0\{\Omega\} = 1$. If $0 < P_0\{\Omega\} < +\infty$, the value $P_0\{\Omega\}$ is irrelevant when one does not try to compare results of an analysis using different reference measures, as was the case in that contribution. On the contrary, in Egozcue *et al.* (2006) the reference is implicitly assumed to be proportional to the length of the interval supporting the densities of the Hilbert space, that is $P_0\{\Omega\}$ is adapted for each support Ω. Here this second strategy has been adopted so that analytical results using different references become comparable, fulfilling the subcompositional coherence requirements. This strategy of normalizing the reference measures has a consequence which might be uncomfortable for some readers, namely that $\mathbf{p}_0$, or in general $\mathbf{p}$, are not only non-closed compositions, but convey also information about the size of Ω, $P\{\Omega\} = \sum_{i=1}^{D} P\{c_i\}$. In the following development, $\mathbf{p}$ or $\mathbf{p}_0$ appear to be closed when represented as elements of the simplex, but retain their absolute values when the components are used as weights in sums (integrals) along compositions or clr images.

To match the present definition to the standard practice in compositional data analysis (Aitchison 1986; Aitchison and Egozcue 2005; Egozcue, Barceló-Vidal, Martín-Fernández, Jarauta-Bragulat, Díaz-Barrero, and Mateu-Figueras 2011; Pawlowsky-Glahn *et al.* 2015) and to the subcompositional dominance of distances, the factor $1/D$ in (6) is suppressed. Remember that multiplication by a real scalar in an inner product does not change its character.

In the case of using a reference measure represented by the weights $\mathbf{p}$, the appropriate definition of the weighted Aitchison inner product is

$$\langle \mathbf{y}_1, \mathbf{y}_2 \rangle_{\mathbf{p}} = \sum_{i=1}^{D} p_i \log \frac{y_{1i}}{\mathrm{g}_{\mathbf{p}}(\mathbf{y}_1)} \log \frac{y_{2i}}{\mathrm{g}_{\mathbf{p}}(\mathbf{y}_2)} \ , \tag{7}$$

where $\mathbf{y}_k = \mathbf{y}_k^{(\mathbf{p})}$, $k = 1, 2$ are in $\mathcal{S}^D$. The expression in the right hand side of Equation (7) is an inner product of the $\mathrm{clr}_{\mathbf{p}}$ as real vectors with respect to the measure P.

The weighted Aitchison norm, derived from the inner product, is $\|\mathbf{y}\|_{\mathbf{p}}^2 = \langle \mathbf{y}, \mathbf{y} \rangle_{\mathbf{p}}$, and an explicit expression of the distance is

$$\mathrm{d}_{\mathbf{p}}^2(\mathbf{y}_1, \mathbf{y}_2) = \langle \mathrm{clr}_{\mathbf{p}}(\mathbf{y}_1) - \mathrm{clr}_{\mathbf{p}}(\mathbf{y}_2), \mathrm{clr}_{\mathbf{p}}(\mathbf{y}_1) - \mathrm{clr}_{\mathbf{p}}(\mathbf{y}_2) \rangle_{\mathbf{p}} = \sum_{i=1}^{D} p_i \left(\log \frac{y_{1i}}{\mathrm{g}_{\mathbf{p}}(\mathbf{y}_1)} - \log \frac{y_{2i}}{\mathrm{g}_{\mathbf{p}}(\mathbf{y}_2)} \right)^2 .$$

This expression of weighted distance can be written in matrix notation

$$\mathrm{d}_{\mathbf{p}}^2(\mathbf{y}_1, \mathbf{y}_2) = (\mathrm{clr}_{\mathbf{p}}(\mathbf{y}_1) - \mathrm{clr}_{\mathbf{p}}(\mathbf{y}_2))\ \mathrm{diag}(\mathbf{p})\ (\mathrm{clr}_{\mathbf{p}}(\mathbf{y}_1) - \mathrm{clr}_{\mathbf{p}}(\mathbf{y}_2))^\top ,$$

where the $\mathrm{clr}_{\mathbf{p}}$ are row vectors and $\mathrm{diag}(\mathbf{p})$ is a diagonal (D, D)-matrix containing the weights $\mathbf{p}$. These definitions coincide with those of the ordinary Aitchison geometry of $\mathcal{S}^D$ whenever $\mathbf{p} = \mathbf{p}_0 = (1, 1, \ldots, 1)$. When $\mathbf{p} \neq \mathbf{p}_0$, the inner product differs from the ordinary Aitchison inner product and, consequently, also norm and distance are different.

To get a further intuition of what is changing with $\mathbf{p}$, it is instructive to build orthonormal basis of the simplex according to the change of reference. It allows to show how these bases appear under a change of $\mathbf{p}$ in particular cases.

A straightforward technique for obtaining orthonormal basis of the simplex and their respective coordinates (Egozcue, Pawlowsky-Glahn, Mateu-Figueras, and Barceló-Vidal 2003) is that of sequential binary partitions (SBP) (Egozcue and Pawlowsky-Glahn 2005, 2006). Like in the standard case (reference measure P_0), when using a reference measure with the weights $\mathbf{p}$, the procedure is based on a partition coded as in Table 1, but the formulae to obtain the contrast matrix are modified. Table 1 shows a generic sign code for an SBP, adding weights $\mathbf{p}$ as column labels (second row) for further comment on the generalised technique.

Table 1: A generic table of an SBP for a five-part composition. Weights from the reference measure are placed in the second row, under the part label. Rows are labelled as balances b_i for further reference.

parts	y_1	y_2	y_3	y_4	y_5
weights	p_1	p_2	p_3	p_4	p_5
b_1	+1	−1	−1	−1	+1
b_2	+1	0	0	0	−1
b_3	0	+1	−1	−1	0
b_4	0	0	+1	−1	0

Denote the entries of the matrix code as θ_{ij}, $i = 1, 2, \ldots, D-1$, $j = 1, 2, \ldots, D$. For the case in Table 1, $D = 5$ and, for instance, $\theta_{32} = +1$. When using the standard reference measure $\mathbf{p}_0 = (1, 1, \ldots, 1)$, the clr coefficients of an element of the basis, that is of a balancing element, are given by

$$\psi_{ij} = \begin{cases} +\frac{1}{n_i^+} \sqrt{\frac{n_i^+ n_i^-}{n_i^+ + n_i^-}} & \text{if } \theta_{ij} = +1 \\ -\frac{1}{n_i^-} \sqrt{\frac{n_i^+ n_i^-}{n_i^+ + n_i^-}} & \text{if } \theta_{ij} = -1 \\ 0 & \text{if } \theta_{ij} = 0 \ , \end{cases} \tag{8}$$

where n_i^+ denotes the number of $+1$, respectively n_i^- of -1, in the i-th row of the code table. When using the reference measure which weights p_j are not unity, these formulas for the $\mathrm{clr}_{\mathbf{p}}$ of balancing elements are the same except that n_i^+, n_i^- are

$$n_i^+ = \sum_{\theta_{ij}=+1} p_j \quad , \quad n_i^- = \sum_{\theta_{ij}=-1} p_j \ .$$

The contrast matrix Ψ, with entries ψ_{ij}, $i = 1, 2, \ldots, D$, $j = 1, 2, \ldots, D-1$, fulfills the conditions

$$\Psi\ \mathrm{diag}(\mathbf{p})\ \Psi^\top = I_{D-1} \quad , \quad \mathrm{diag}(\mathbf{p})\ \Psi^\top \Psi = I_D - \frac{1}{D} \mathbf{p}^\top \mathbf{1} \ , \tag{9}$$

where I_m is the (m,m)-identity matrix; $\mathbf{p}$ and $\mathbf{1} = (1,1,\ldots,1)$ are taken as row D-vectors, and $\mathrm{diag}(\mathbf{p})$ is a (D,D) diagonal matrix with entries equal to the components of $\mathbf{p}$. The first condition is equivalent to saying that balancing elements are unitary compositions mutually orthogonal. In fact, their $\mathrm{clr}_{\mathbf{p}}$ are unitary and orthogonal in the weighted Euclidean geometry.

Coordinates of a density $\mathbf{y} \in \mathcal{S}^D$ with respect to an orthonormal basis are found carrying out the inner product of a balancing element in the basis with the density $\mathbf{y} = \mathbf{x}/\mathbf{p}$. In general, these coordinates are termed weighted isometric log-ratio coordinates and denoted by $\mathrm{ilr}_{\mathbf{p}}$. In the particular case in which they are obtained using an SBP, they are called weighted balances. For simplicity, these weighted balances are denoted b_i, $i = 1,2,\ldots,D-1$, with no reference to the weights associated with the change of measure (as shown in Table 1). The $\mathrm{ilr}_{\mathbf{p}}$ coordinates can be obtained using the matrix expression

$$\mathrm{ilr}_{\mathbf{p}}(\mathbf{y}) = \mathbf{b} = \mathrm{clr}_{\mathbf{p}}(\mathbf{y})\ \mathrm{diag}(\mathbf{p})\ \Psi^{\top}\ , \tag{10}$$

where compositions and their $\mathrm{clr}_{\mathbf{p}}$ and $\mathrm{ilr}_{\mathbf{p}}$ transforms are considered row-vectors. Note that each component of $\mathbf{b} = (b_1, b_2, \ldots, b_{D-1})$ is a weighted inner product of $\mathrm{clr}_{\mathbf{p}}(\mathbf{y})$ with the corresponding $\mathrm{clr}_{\mathbf{p}}$ of a balancing element. The inverse $\mathrm{ilr}_{\mathbf{p}}$ transformation is readily obtained using the properties (9) of Ψ

$$\mathcal{C}\mathbf{y} = \mathcal{C}\exp(\mathrm{ilr}_{\mathbf{p}}(\mathbf{y})\Psi) \quad , \quad \mathrm{clr}_{\mathbf{p}}(\mathbf{y}) = \mathrm{ilr}_{\mathbf{p}}(\mathbf{y})\Psi\ ,$$

being the first of these relations formally identical to the standard inverse ilr with reference measure P_0. The relationship of $\mathrm{ilr}_{\mathbf{p}}(\mathbf{y})$ and $\mathrm{ilr}(\mathbf{x})$ is developed in Appendix B.

Although Equation 10 is useful from a computational point of view, an explicit expression of balances gives a deeper insight into the meaning of weighted balances. Consider a sign code of a step in an SBP, for which n_i^+, n_i^- are given. The corresponding weighted balance is

$$b_i = \sqrt{\frac{n_i^+ n_i^-}{n_i^+ + n_i^-}}\ \log\left(\frac{\prod_{(\theta_{ij}=+1)}\ y_j^{p_j/n_i^+}}{\prod_{(\theta_{ij}=-1)}\ y_j^{p_j/n_i^-}}\right)\ , \tag{11}$$

where the products span over the parts corresponding to the sign code θ_{ij}. When the weights $p_j = 1$, the balance reduces to the standard balances, as n_i^+, n_i^- are then the number of $+1$ and -1 in the i-th row of the sign code, respectively. The main feature, when the reference is not $\mathbf{p}_0$, is that the ratios within the logarithm are ratios of a kind of weighted geometric means. Note that, in general, n_i^+, n_i^- are not integers and each part is powered to the weight corresponding to that part. When some p_j is small, relative to other weights, it plays a minor role in these weighted geometric means. Furthermore, the weighted balances are scale invariant log-contrasts, that is, if the composition $\mathbf{y}$ is multiplied by a positive constant, the weighted balance remains unaltered.

Expressing inner products, norms, and distances as functions of weighted coordinates $\mathrm{ilr}_{\mathbf{p}}$ can be useful, because they are exactly those of the standard Euclidean geometry. For the inner product and square-distance they are

$$\langle \mathbf{y}_1, \mathbf{y}_2 \rangle_{\mathbf{p}} = \langle \mathrm{ilr}_{\mathbf{p}}(\mathbf{y}_1), \mathrm{ilr}_{\mathbf{p}}(\mathbf{y}_2) \rangle \quad , \quad \mathrm{d}^2_{\mathbf{p}}(\mathbf{y}_1, \mathbf{y}_2) = \mathrm{d}^2(\mathrm{ilr}_{\mathbf{p}}(\mathbf{y}_1), \mathrm{ilr}_{\mathbf{p}}(\mathbf{y}_2))\ , \tag{12}$$

where $\langle \cdot, \cdot \rangle$, $\mathrm{d}(\cdot,\cdot)$, are the ordinary Euclidean inner product and distance.

Whenever there is a change in the geometry of compositions, the subcompositional dominance of distances is a critical point. In the standard approach, the distance between any two compositions $\mathbf{x}_1$, $\mathbf{x}_2 \in \mathcal{S}^D$ is $\mathrm{d}_a(\mathbf{x}_1, \mathbf{x}_2)$. After taking a given subcomposition in $\mathcal{S}^d$, $d < D$, the distance between the respective subcompositions, $\mathbf{x}_1^{(d)}, \mathbf{x}_2^{(d)}$, satisfies $\mathrm{d}_a(\mathbf{x}_1^{(d)}, \mathbf{x}_2^{(d)}) \leq \mathrm{d}_a(\mathbf{x}_1, \mathbf{x}_2)$. In this case, both spaces have integer reference measures with $P\{\Omega_D\} = D$ and $P\{\Omega_d\} = d$ and, for $D = 3$, $d = 2$ the corresponding weights are $(1,1,1)$ and $(1,1,0)$, respectively. When changing the reference measure by down weighting some of the weights, a dominance of distances is expected, as it occurs when taking subcompositions. The dominance of distances

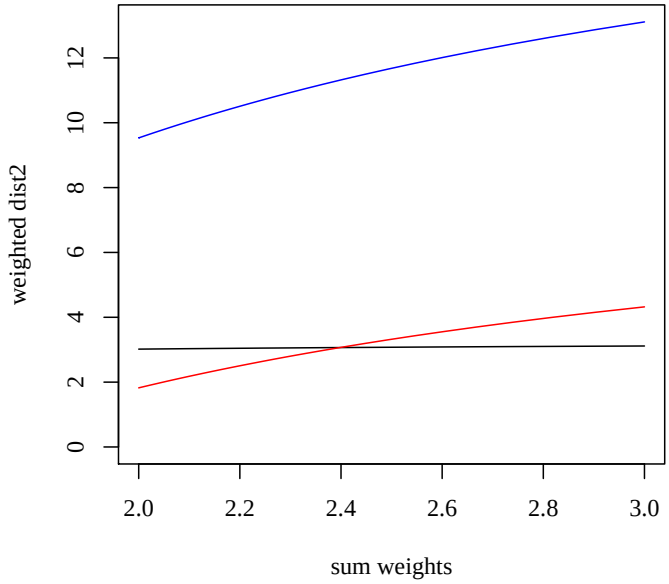

Figure 2: Evolution of weighted square-distances between three measures represented by the compositions $\mathbf{x}_1 = (0.1, 0.7, 0.2), \mathbf{x}_2 = (0.5, 0.3, 0.2), \mathbf{x}_3 = (0.9, 0.08, 0.02) \in \mathcal{S}^3$ with respect to the reference measure P_0 with weights $\mathbf{p}_0 = (1, 1, 1)$. With $\mathbf{y}_i = \mathbf{x}_i/\mathbf{p}$, square-distance curves are $\mathrm{d}_{\mathbf{p}}(\mathbf{y}_1, \mathbf{y}_2)$ (black), $\mathrm{d}_{\mathbf{p}}(\mathbf{y}_1, \mathbf{y}_3)$ (blue), $\mathrm{d}_{\mathbf{p}}(\mathbf{y}_2, \mathbf{y}_3)$ (red). Reference measure is $\mathbf{p} = (1, 1, \epsilon)$ and x-axis is scaled as $P\{\Omega\} = 1 + 1 + \epsilon$. The three square-distances monotonically increase from $P\{\Omega\} = 2$ to $P\{\Omega\} = 3$. The end points of the curves at $P\{\Omega\} = 2$ and $P\{\Omega\} = 3$ are equal to standard Aitchison square-distances in $\mathcal{S}^2$ and $\mathcal{S}^3$ respectively.

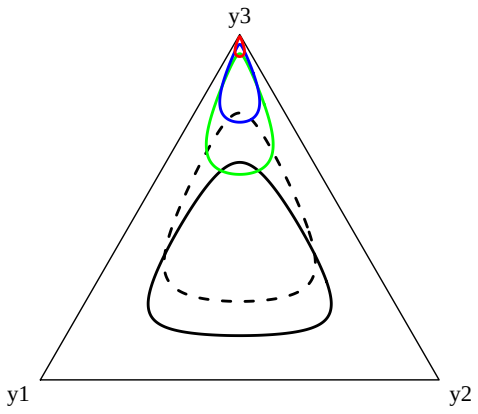

Figure 3: The unit circle (black, full line) in the uniform reference. After change of origin to $(1, 1, \epsilon)$, $\epsilon = 0.5$ (black, dashed), 0.1 (green), 0.05 (blue), and 0.01 (red), the circle is shifted towards the vertex y_3.

can be stated as follows.

PROPOSITION (dominance of distances) *Let* $\mathbf{x}_1$, $\mathbf{x}_2$ *be two compositions in* $\mathcal{S}^D$, *endowed with the reference measure* P_0, *with weights* $\mathbf{p}_0 = (1, 1, \ldots, 1)$. *Consider two reference measures,* P_1 *and* P_2, *represented by their respective weights* $\mathbf{p}_1 = (p_{11}, p_{12}, \ldots, p_{1D})$ *and* $\mathbf{p}_2 = (p_{21}, p_{22}, \ldots, p_{2D})$, *such that all their components are* $0 < p_{ki} \leq 1$, *for* $k = 1, 2$, $i = 1, 2, \ldots, D$ *and* $P_k\{\Omega\} = \sum_{i=1}^{D} p_{ki}$. *Define* $\mathbf{y}_j^{(\mathbf{p}_k)} = \mathbf{x}_j / \mathbf{p}_k$ *for* $k = 1, 2$ *and* $j = 1, 2$. *Then,*

$$p_{1i} \leq p_{2i},\ i = 1, 2, \ldots, D \quad \Rightarrow \quad \mathrm{d}_{\mathbf{p}_1}(\mathbf{y}_1^{(\mathbf{p}_1)}, \mathbf{y}_2^{(\mathbf{p}_1)}) \leq \mathrm{d}_{\mathbf{p}_2}(\mathbf{y}_1^{(\mathbf{p}_2)}, \mathbf{y}_2^{(\mathbf{p}_2)}) \ .$$

It is worth to remark that the notation of distances like $\mathrm{d}_{\mathbf{p}_1}(\mathbf{y}_1^{(\mathbf{p}_1)}, \mathbf{y}_2^{(\mathbf{p}_1)})$ could be changed to $\mathrm{d}_{\mathbf{p}_1}(\mathbf{x}_1, \mathbf{x}_2)$, as distances assigned to shifted $\mathbf{y}$'s are equal to those of the original compositions $\mathbf{x}$'s. This is due to the fact that $\mathbf{x}$ and $\mathbf{y}$ are densities of the same measure, namely $\mu_{\mathbf{x}}$, with respect to different reference measures.

Figure 2 shows the evolution of square-distances between three compositions $\mathbf{x}_1 = (0.1, 0.7, 0.2)$, $\mathbf{x}_2 = (0.5, 0.3, 0.2)$, $\mathbf{x}_3 = (0.9, 0.08, 0.02)$ with respect to the uniform reference in $\mathcal{S}^3$ when the reference measure changes progressively. The reference measure is $(1, 1, \epsilon)$, with ϵ going from 0 to 1. The plot is scaled according the $P\{\Omega\} = 1+1+\epsilon$. The square-distances increase monotonically, from distances corresponding to the subcomposition (y_1, y_2) to square-distances with the standard reference $\mathbf{p}_0 = (1, 1, 1)$. This result is expected after the previous proposition.

An experiment has been conducted to show how the changes of reference modify distances and shapes. Five different reference measures $\mathbf{p} = (1, 1, \epsilon)$ have been considered with ϵ equal to 1, 0.5, 0.1, 0.05, 0.01, so that they approach progressively the geometry of the subcomposition of the two first parts. The unit circle centered at the neutral element was shifted by the five reference measures. Figure 3, shows this unit circle (black) and the sequence of perturbations as a consequence of the change of origin. Note that the transformed circle is shifted to the vertex which weight is reduced, as expected after dividing each part by the corresponding weight.

After the change of origin, each point on the circles was ilr-transformed using the corresponding weights according to the SBP sign code

y_1	y_2	y_3
+1	−1	−1
0	+1	−1

,

which has been selected to avoid a balance representing the subcomposition (y_1, y_2). Figure 4 (left panel) shows the coordinates of the circles, to show the changes of the distances between points on the same circle. Note that the centers of the ellipses do not coincide, as they correspond to the closure of the reference measure $(1, 1, \epsilon)$. The main feature is the progressive stretch of the original circle. For very small ϵ the ellipse tends to degenerate into a segment following the direction of the subcomposition (y_1, y_2). Similarly, Figure 4 (right panel) shows the deformation of a grid originally at $-1, 0, 1$ in both axes (black). The new references are $\epsilon = 0.1$ (blue), and 0.01 (red). The grid is progressively tilted and distances between nodes decrease as ϵ decreases. Although straight-lines are preserved, their angles change, thus showing the change of geometry when changing the reference.

5. Elementary statistics

The change of reference measure and its associated weighting have consequences in the definitions of elementary concepts of compositional statistics. Variability and center are the two main concepts examined below. Both concepts are redefined following previous developments in the statistical analysis of compositional data, just looking for the influence of the weighting. These new definitions are intended to match the standard concepts whenever the weights are unity over the categories defining the composition.

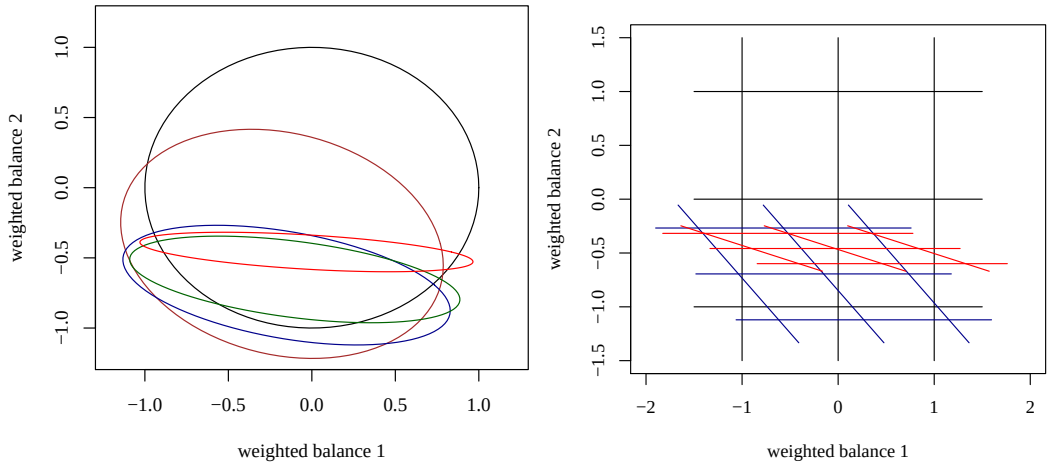

Figure 4: Left panel: the five circles in Figure 3 after weighted ilr-transformation. Reference measures are $(1, 1, \epsilon)$, $\epsilon = 1$ (black), $\epsilon = 0.5$ (brown), 0.1 (blue), 0.05 (green), and 0.01 (red). Right panel: a regular grid at points -1, 0, -1 in both axes after change of origin and weighted ilr transformation. Weights are $(1, 1, \epsilon)$, $\epsilon = 1$ (black), 0.1 (blue), and 0.01 (red).

Let $\mathbf{X}$ be a random composition (density) in $\mathcal{S}^D$ (Pawlowsky-Glahn *et al.* 2015, ch. 6) which, for some selected ilr coordinates denoted $\mathbf{X}^*$ in $\mathbb{R}^{D-1}$, is absolutely continuous with joint probability density (pdf) $f^*_{\mathbf{X}}$. Therefore, $f^*_{\mathbf{X}}(\mathbf{x})$ is a function defined on $\mathbb{R}^{D-1}$, the space of the ilr coordinates, with the standard definitions from probability theory. Assume also that a new reference measure is chosen and it is represented by a set of positive weights $\mathbf{p}$. Accordingly, the random composition $\mathbf{Y} = \mathbf{X} \ominus \mathbf{p}$ corresponds to the change of reference and its distribution only differs from that of $\mathbf{X}$ in a shift of the center. The $\text{ilr}_{\mathbf{p}}$ coordinates of $\mathbf{Y}$, denoted $\mathbf{Y}^*$, are also random, but their distribution on $\mathbb{R}^{D-1}$ is a transformation of the previous pdf $f^*_{\mathbf{X}}$, here denoted as f^*, where the subscript is dropped when it corresponds to the composition $\mathbf{Y}$. In Appendix B it is shown that the transformation from $\mathbf{X}^* = \text{ilr}(\mathbf{X})$ to $\mathbf{Y}^* = \text{ilr}_{\mathbf{p}}(\mathbf{Y})$ is a linear (affine) transformation. For instance, this means that, if $\mathbf{X}$ has a normal distribution on the simplex (Mateu-Figueras, Pawlowsky-Glahn, and Egozcue 2013; Pawlowsky-Glahn *et al.* 2015) and, thus, $\mathbf{X}^*$ is multivariate normal on $\mathbb{R}^{D-1}$, the distribution of $\mathbf{Y}^*$ is also a multivariate normal on $\mathbb{R}^{D-1}$. As a conclusion, the normality of $\text{ilr}_{\mathbf{p}}$ coordinates is maintained when the weights $\mathbf{p}$ of the reference measure change.

Following the general formalism developed by Fréchet (1948) for metric spaces, the first milestone to be defined is the (total) variability of $\mathbf{Y}$ with respect to an arbitrary point $\boldsymbol{\eta} \in \mathcal{S}^D$. It is defined as

$$\text{totVar}_{\mathbf{p}}[\mathbf{Y}; \boldsymbol{\eta}] = \text{E}[\text{d}^2_{\mathbf{p}}(\mathbf{Y}, \boldsymbol{\eta})] \ ,$$

provided that the expectation exists. The distance $\text{d}^2_{\mathbf{p}}(\mathbf{Y}, \boldsymbol{\eta})$ is a function of the coordinates $\mathbf{Y}^* = \text{ilr}_{\mathbf{p}}(\mathbf{Y})$ and the expectation $\text{E}[\cdot]$ is taken with respect to their pdf f^*. Since $\text{d}^2_{\mathbf{p}}(\mathbf{Y}, \boldsymbol{\eta}) = \sum (Y^*_i - \eta^*_i)^2$ (Equation 12), the minimum of $\text{Var}_{\mathbf{p}}[\mathbf{Y}; \boldsymbol{\eta}]$ is attained for $\boldsymbol{\eta}^* = \text{E}[\mathbf{Y}^*]$, a standard result in real multivariate statistics. Based on this result, the weighted center and total variance are

$$\text{Cen}_{\mathbf{p}}[\mathbf{Y}] = \text{ilr}^{-1}_{\mathbf{p}}(\text{E}[\mathbf{Y}^*]) = \text{clr}^{-1}_{\mathbf{p}}(\text{E}[\mathbf{Y}^*]) \quad , \quad \text{totVar}_{\mathbf{p}}[\mathbf{Y}] = \text{E}[\text{d}^2_{\mathbf{p}}(\mathbf{Y}, \text{Cen}_{\mathbf{p}}[\mathbf{Y}])] \ . \tag{13}$$

Note that this kind of approach has been used in Pawlowsky-Glahn and Egozcue (2001) and in Boogaart and Tolosana-Delgado (2013), but total variance is there called metric variance.

Despite the previous expression of $\text{Cen}_{\mathbf{p}}[\mathbf{Y}]$ in Equation 13, the weighted center of a random composition only depends on the weights in $\mathbf{p}$ through the shift applied, that is

$$\text{Cen}_{\mathbf{p}}[\mathbf{Y}] = \text{Cen}[\mathbf{X}] \ominus \mathbf{p} \ , \quad \text{or, equivalently,} \quad \text{Cen}[\mathbf{X}] = \text{Cen}_{\mathbf{p}}[\mathbf{Y}] \oplus \mathbf{p} \ ,$$

where Cen and $\oplus$ are the ordinary center and perturbation of random compositions, respectively, and $\mathbf{Y} = \mathbf{X} \ominus \mathbf{p}$, thus enhancing the linearity of expectations.

Decompositions of total variance underlays many standard statistical methods, thus remarking its upmost importance. Equation 13 leads to decompositions of the total variance when the reference measure $\mathbf{p}$ is not $\mathbf{p}_0$. Similarly to those described in Egozcue *et al.* (2011), we obtain

$$\begin{aligned} \text{totVar}_{\mathbf{p}}[\mathbf{Y}] &= \sum_{i=1}^{D-1} \text{Var}[\ \text{ilr}_{\mathbf{p},i}(\mathbf{Y})\ ] \\ &= \sum_{i=1}^{D} p_i \text{Var}[\ \text{clr}_{\mathbf{p},i}(\mathbf{Y})\ ] \\ &= \frac{1}{2s_{\mathbf{p}}} \sum_{i=1}^{D}\sum_{j=1}^{D} p_i p_j \text{Var}\left[\ln \frac{Y_i}{Y_j}\right] \ , \end{aligned} \tag{14}$$

where $s_{\mathbf{p}} = \sum_{i=1}^{D} p_i$, $\text{ilr}_{\mathbf{p},i}(\mathbf{Y}) = y_i^*$ and $\text{clr}_{\mathbf{p},i}(\mathbf{Y})$ is the i-th component of $\text{clr}_{\mathbf{p}}(\mathbf{Y})$. Note that the decomposition of $\text{totVar}_{\mathbf{p}}[\mathbf{Y}]$ into $\text{ilr}_{\mathbf{p}}$ variance components points out that $\text{totVar}_{\mathbf{p}}[\mathbf{Y}]$ is the trace of the covariance matrix of $\text{ilr}_{\mathbf{p}}(\mathbf{Y})$, and that $\text{totVar}_{\mathbf{p}}[\mathbf{Y}]$ is not the sum of $\text{clr}_{\mathbf{p}}$ variances, but a weighted sum of them.

The decompositions of the total variance are closely related to the relationships between the covariance matrices of the $\text{ilr}_{\mathbf{p}}$ coordinates and the $\text{clr}_{\mathbf{p}}$ coefficients. These relationships can be summarized as

$$\Sigma_{\mathbf{p}} = \Psi\ \text{diag}(\mathbf{p})\ \Sigma_{\mathbf{p}}^{c}\ \text{diag}(\mathbf{p})\ \Psi^{\top} \quad , \quad \Sigma_{\mathbf{p}}^{c} = \Psi\ \Sigma_{\mathbf{p}} \Psi^{\top} \ ,$$

where Ψ is the $(D-1, D)$-contrast matrix of the $\text{ilr}_{\mathbf{p}}$, $\Sigma_{\mathbf{p}}$ is the covariance matrix of $\mathbf{Y}^*$ and $\Sigma_{\mathbf{p}}^{c}$ is the covariance matrix of $\text{clr}_{\mathbf{p}}(\mathbf{Y})$.

Also, the variation matrix (Aitchison 1986) plays an important role in the statistics of compositional data. Its entries are variances of simple log-ratios, $\ln(X_i/X_j)$. At least, it has two important uses: (a) it constitutes a simple and interpretable representation of the variability (second order moments) of the random composition, identifying the binary sources of variability relative to the total variance; and (b) each entry of the variation matrix is a measure of the compositional dissociation, as opposite of association, between the two parts involved. Point (a) is reflected in the fact that the covariance matrices of ilr-coordinates and clr coefficients can be retrieved from the variation matrix (Pawlowsky-Glahn *et al.* 2015, Appendix A). Concerning point (b), large entries, relative to other entries, point out most dissociated pairs of parts. The measurement of compositional association of two parts, understood as proportionality between them, is motivated by the fact that $\text{Var}[\ln(X_i/X_j)] = 0$ implies that X_i and X_j are strictly proportional (Egozcue, Lovell, and Pawlowsky-Glahn 2013a; Lovell, Pawlowsky-Glahn, Egozcue, Marguerat, and Bähler 2015).

Inspired by the third decomposition of weighted total variance in Equation 14, a weighted variation matrix can be defined as a (D, D)-matrix $T_{\mathbf{p}}$ with entries

$$t_{\mathbf{p},i,j} = p_i p_j \text{Var}\left[\ln \frac{Y_i}{Y_j}\right] \quad , \quad i, j = 1, 2, \ldots, D \ .$$

The relationship of $T_{\mathbf{p}}$ with the covariance matrix of $\text{ilr}_{\mathbf{p}}$ coordinates is

$$\Sigma_{\mathbf{p}} = -\frac{1}{2} \Psi T_{\mathbf{p}} \Psi^{\top} \ .$$

The decomposition of weighted total variance and the relationships between covariance matrices reduce to the standard ones whenever the reference measure is $P = P_0$, that is, whenever $\mathbf{p} = (1, 1, \ldots, 1)$.

Table 2: Weights, $\mathbf{p}$ (second row) and $p_i^{(sub)}$ (third row), for each part used in the analysis of the Cat10 data set. Weights $p_i^{(sub)}$ are only used in an example of biplot. Center of the composition, expressed in percent, for the original composition (forth row), and for the shifted composition $\mathbf{Y} = \mathbf{X} \ominus \mathbf{p}$ (fifth row).

party	abs	nota	null	C's	CiU	ERC	ICV	PSC	PP	other
p_i	0.1	0.3	0.3	1	1	1	1	1	1	0.5
$p_i^{(sub)}$	0.001	0.001	0.001	1	1	1	1	1	1	0.001
Cen[$\mathbf{X}$] (%)	38.9	1.9	0.5	0.9	27.6	5.5	3.0	9.6	5.2	6.8
Cen$_{\mathbf{p}}$[$\mathbf{Y}$] (%)	84.1	1.4	0.4	0.2	6.0	1.2	0.6	2.1	1.1	2.9

6. Exploratory tools

In compositional data analysis, the main specific exploratory tools are the variation matrix (Aitchison 1986), principal component analysis of the clr transformed compositional sample (Aitchison 1983) and its corresponding biplots (Aitchison and Greenacre 2002), and the compositional dendrogram (Pawlowsky-Glahn and Egozcue 2011). These three tools take slightly different forms when taking a reference measure different from P_0. In order to show how to use and interpret the weighted versions in an exploratory analysis, the data from the Catalan parliament (Spain) elections in November 2010 (Cat10) have been selected. This data set was previously analysed in Egozcue and Pawlowsky-Glahn (2011) (see also Pawlowsky-Glahn *et al.* 2015).

The data set Cat10 contains the number of votes obtained by several parties, including abstention (abs), null (null) and none of the above or blank votes (nota) in $n = 41$ electoral districts. The major parties contesting the elections were *Convergència i Unió* (CiU), *Partit dels Socialistes de Catalunya* (PSC), *Ciutadans-Partido de la Ciudadanía* (C's), *Esquerra Republicana de Catalunya* (ERC), *Iniciativa per Catalunya Verds-Esquerra Unida i Alternativa* (ICV) and *Partit Popular* (PP). Other minor parties are amalgamated in *other*. The present analysis focusses on the whole composition of votes, that is, the $D = 10$ parts of the composition: abs, nota, null, CiU, C's, ERC, ICV, PP, PSC, other.

A first step in exploratory analysis is to choose suitable weights for the 10 parts involved. The situation in most political elections is that votes to parties show a homogeneous preference to a given party, meanwhile "abs", "nota", "null" and "other" mix non-homogeneous support to democratic elections or other situations, thus suggesting to weight them differently. Well defined parties were weighted by 1. The abstention is the more heterogeneous group of electors and the choice for its weight was 0.1. The electors that choose blank vote (nota) and null vote (null) can be considered less heterogeneous than abstention, as they express something similar to "I want to vote, but none of the contesting parties convinced me"; these two categories have been weighted by 0.3. Votes to parties included in "other" are well defined, but directed to different parties with different programmes; there is a well defined intention in the vote, but the amalgamation of different parties makes the group heterogeneous; the category "other" is weighted by 0.5. The vector of weights $\mathbf{p}$ chosen is shown in the second row of Table 2. These weights have been chosen to show the effects of weighting, and not to carry out a sound analysis of the data set. Methods to establish suitable weights should be object of further research. The third row of Table 2 shows an alternative set of weights $p_i^{(sub)}$ that will be used only for illustrating how these weights make the analysis to be close to that of a subcomposition of the well defined parties. The forth row of Table 2 shows the center of the composition, expressed in percent. The fifth row is the center Cen$_{\mathbf{p}}$[$\mathbf{Y}$] (also in percent), which is not useful for interpretation, but for comparison with Cen[$\mathbf{X}$]. Note how the percent of "abs", with weight 0.1, increased when dividing by the weight. The same fact may occur for all parts with weights less than one, but closure hides this fact. Note that the center is

Table 3: Weighted variation matrix for Cat10 data. Last column: weighted $\text{clr}_{\mathbf{p}}$ variances, $p_i \text{Var}(\text{clr}_{\mathbf{p},i}[\mathbf{Y}])$, adding to weighted total variance. Upper triangle: elements of the weighted variation matrix (values greater than or equal to 0.30 are highlighted in boldface). Lower triangle: product of weights $p_i p_j$. Two last rows: weighted total variance and total variance (uniform reference).

	Abs	Nota	Null	C's	CiU	ERC	ICV	PSC	PP	other	$\text{clr}_{\mathbf{p}}$ var.
Abs		0.002	0.005	0.029	0.007	0.020	0.008	0.006	0.011	0.007	0.001
Nota	0.03		0.008	0.164	0.009	0.027	0.040	0.031	0.073	0.017	0.014
Null	0.03	0.09		0.238	0.022	0.015	0.078	0.064	0.111	0.020	0.039
C's	0.10	0.09	0.30		**0.563**	**0.870**	0.270	**0.320**	0.160	**0.303**	**0.308**
CiU	0.10	0.30	0.30	1.00		0.077	0.157	0.142	0.268	0.034	0.054
ERC	0.10	0.30	0.30	1.00	1.00		0.249	0.262	**0.459**	0.052	0.153
ICV	0.10	0.30	0.30	1.00	1.00	1.00		0.101	0.189	0.085	0.051
PSC	0.10	0.30	0.30	1.00	1.00	1.00	1.00		0.122	0.133	0.051
PP	0.10	0.30	0.30	1.00	1.00	1.00	1.00	1.00		0.193	0.113
other	0.05	0.15	0.15	0.50	0.50	0.50	0.50	0.50	0.50		0.054
$\text{totVar}_{\mathbf{p}}$											0.836
totVar											1.020

a composition of a "mean electoral district", and that variability around this center may be large. This can be checked, for instance, on C's, which minimum percentage is 0.3% and its maximum is 2.8% across the sample of electoral districts, what in turns may represent a number of electors from 3046 up to 1,572,425 for the surroundings of Barcelona. Therefore, reporting mean values or centers needs to be complemented with the analysis of variability.

The weighted variation matrix is shown in the upper triangle of Table 3. In the lower triangle of Table 3, the cross products of weights $p_i p_j$ are specified. When the entries of the weighted variation matrix are divided by the corresponding $p_i p_j$ they result in the corresponding entry of the traditional variation matrix with reference P_0. Terms in the weighted variation matrix larger than or equal to 0.30 are highlighted in boldface. They constitute the larger sources of variability in the data set. Most of them correspond to C's, whose votes are irregularly distributed over electoral districts. This fact is confirmed by the weighted $\text{clr}_{\mathbf{p}}$ variances, as the largest value corresponds to C's as well. Small values in the weighted variation matrix suggest association between parts, i.e. approximate proportionality, although this needs further analysis to be confirmed (Egozcue *et al.* 2013a; Lovell *et al.* 2015). The strongest associations appear between abs, nota, null, with traditionally nationalist parties in Catalonia, i.e. CiU, ERC, and even with PSC. Compared to the variation matrix published in Egozcue and Pawlowsky-Glahn (2011), the possible associations appear stronger in Table 3. This is due to the fact that the 2011 analysis was performed without any weighting in the reference. Differences in the variances of simple log-ratios of not down-weighted parts are the consequence of dividing entries in Table 3 by $n-1=40$, while in 2011 the divisor was $n=41$. The weighted total variance is 0.836, smaller than that obtained with unit weights (1.020), using in both cases the same divisor ($n-1=40$).

In compositional data analysis, principal component analysis (PCA) is commonly performed using the singular value decomposition (SVD) of the clr-transformed data set (Aitchison 1983). The scores, multiplied by the singular values, are proportional to ilr-coordinates, such that their variances are proportional to the square singular values. The loadings matrix contains the clr representation of the principal directions. The last singular value is zero, as the clr data sum to zero for each data point. Similar features are expected for a PCA performed on a weighted composition using its weighted $\text{clr}_{\mathbf{p}}$ transformed values. However, when the $\text{clr}_{\mathbf{p}}$-transformed data set is SVD-decomposed, the square singular values are no longer proportional to $\text{ilr}_{\mathbf{p}}$ variances and they do not provide a decomposition of the weighted total variance. The way proposed here consists of dividing the $\text{clr}_{\mathbf{p}}$ data previous to SVD, so

that resulting square singular values add to the total variance.

Let X be a compositional data set in $\mathcal{S}^D$; therefore, X is a (n, D)-matrix and n is the size of the sample. After selecting some positive weights, $\mathbf{p}$, each row of the data matrix is accordingly shifted and is written as $Y = X \ominus \mathbf{p}$. Applying the $\mathrm{clr}_{\mathbf{p}}$ transformation to each row yields $\mathrm{clr}_{\mathbf{p}}(Y)$. This $\mathrm{clr}_{\mathbf{p}}$-transformed data set is centered and weighted with the square-root of the weights in $\mathbf{p}$, that is

$$A = [\mathrm{clr}_{\mathbf{p}}(Y) - (\overline{\mathrm{clr}_{\mathbf{p}}(Y)})]\ \mathrm{diag}(\sqrt{\mathbf{p}})\ ,$$

where $\overline{\mathrm{clr}_{\mathbf{p}}(Y)}$ denotes the average by columns of $\mathrm{clr}_{\mathbf{p}}(Y)$. The SVD of A,

$$A = U\Lambda V^{\top}\ ,$$

has the standard properties of an SVD. Among these properties, some of them are reinterpreted in the compositional framework. The singular values contained in the diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{D-1}, 0)$ are positive and in decreasing order of magnitude; the last one is zero due to the property of the $\mathrm{clr}_{\mathbf{p}}(Y)$ that the weighted sum of its components adds to zero (Equation 4). The non-standardized scores $U\Lambda$ are $\mathrm{ilr}_{\mathbf{p}}$-coordinates whose sample variances are $\lambda_i^2/(n-1)$. The sample total variance is $\mathrm{totVar}_{\mathbf{p}}(Y) = \sum_{i=1}^{D-1} \lambda_i^2/(n-1)$. The $D-1$ first columns of $V^{\top}$ contain the contrast matrix corresponding to the $\mathrm{ilr}_{\mathbf{p}}$. The loadings are given by the columns of $\mathrm{diag}(1/\sqrt{\mathbf{p}})V\Lambda$, where $\mathrm{diag}(1/\sqrt{\mathbf{p}})$ appears to compensate the previous weighting in A.

A covariance biplot (Aitchison and Greenacre 2002) is a simultaneous projection of U (scores) and $V\Lambda$ (loadings) onto two principal directions, usually the first two. The percent of weighted total variance explained in such a projection is given by

$$100\ \frac{\lambda_1^2 + \lambda_2^2}{\sum_{i=1}^{D-1} \lambda_i^2}\ .$$

This kind of biplots have been obtained for the data set Cat10. Figure 5 shows four different cases: top-left panel shows the biplot when the reference is $\mathbf{p}_0 = (1, 1, \ldots, 1)$; top-right panel adopts the weights p_i shown in Table 2; bottom-left panel shows the biplot when using $p_i^{(sub)}$ also shown in Table 2 (third row). Finally, the bottom-right panel shows the biplot obtained using the subcomposition of individual parties, excluding "abs", "nota", "null", and "other", and using the reference $\mathbf{p}_0$ for the subcomposition.

The first impression is that the two biplots in the upper part of Figure 5 appear to be quite similar, as the main features are preserved. In fact, the clr-variables corresponding to well defined parties are projected very similarly. For instance, the first principal axis is dominated by the clr-variables corresponding to C's on one side, and CiU and ERC on the opposite side, which can be identified with a balance of non-nationalist *versus* nationalist Catalan parties; this fact was previously observed in the weighted variation matrix. The second principal axis is mainly influenced by the links between PP-ICV and PSC-ICV, leading to identify the second principal axis with a balance of right *versus* left wing parties. In fact, the three parties involved are perceived by electors as right wing (PP), very moderate social-democratic (PSC) and left wing (ICV). However, when looking at the clr-variables corresponding to down-weighted parts (abs, null, nota, other), the shortening of the corresponding parts proportional to $\sqrt{p_i}$ is apparent. For example, the role of clr-other in the projection has been reduced in an appreciable way.

In the bottom-left panel of Figure 5, the weights $p_i^{(sub)}$ (Table 2) have been used in order to approach a subcompositional analysis of the parties C's, CiU, ERC, ICV, PP, PSC. As the rest of the parts are severely down-weighted, they appear as very short rays from the origin (labels are overlapping). Compared with the subcompositional analysis (bottom-right panel, Figure 5), it is clear that, exception made of these short rays, the rest is almost identical in the two bottom biplots. See, for instance, that the total variance of the two cases are, respectively, 0.7020 (weights $p_i^{(sub)}$) and 0.7016 (unit weights in the subcomposition) and

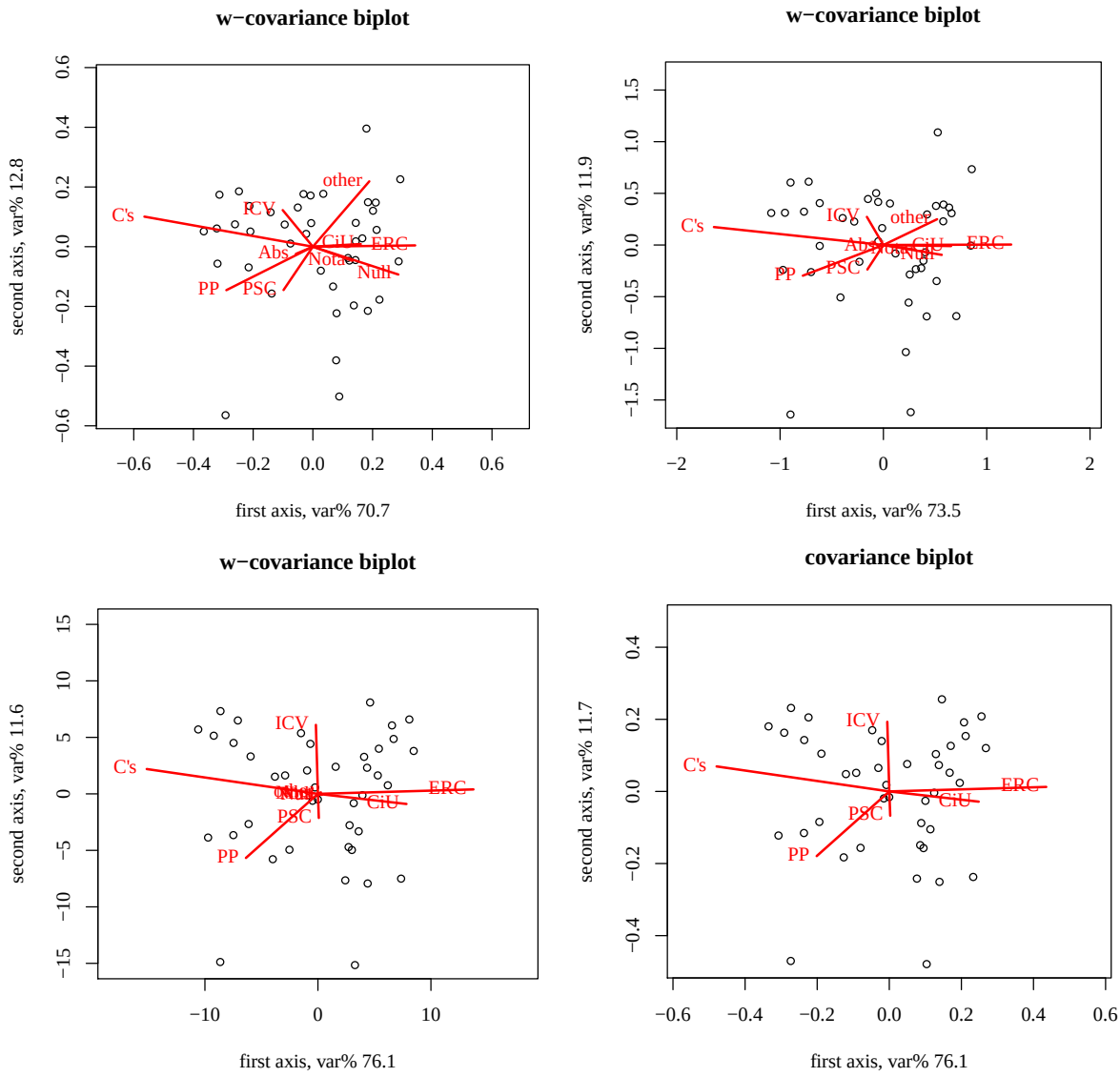

Figure 5: Covariance biplots of Cat10 dataset. Top-left panel: uniform reference $\mathbf{p}_0 = (1, 1, \ldots, 1)$, total variance 1.020. Top-right panel: weights given in Table 2, weighted total variance 0.836. Bottom-left panel: extreme weighting, given in Table 2, weighted total variance 0.7020. Bottom-right panel: subcomposition of parties, total variance 0.7016.

the corresponding proportions of explained total variance in the two dimensional projections are very close. This illustrates the fact, that down-weighting some parts is a path towards subcompositional analysis.

The fact that the projection changes only slightly from top to bottom of Figure 5 indicates that most of the variance introduced by "abs" is small (see Table 3) and that of "nota" and "null" is not well represented in the first and second principal axes. A feature that is clear in the weighted biplot (top-right panel) is that the link "null-other" is almost parallel to the second axis and to the link PSC-ICV: the variance of this two log-ratios are mainly included in the second principal component. The "nota" and "null" votes are quite associated one to each other across electoral districts as the rays appear almost parallel (see also Table 3). When they are down weighted (top-right panel) the main effect is that the corresponding rays are equally shortened as the weights were equal for these two parts.

The so called balance-dendrogram is not discussed here in detail, as the changes to be incorporated when using weights are quite obvious. Firstly, a balance-dendrogram presents

a hierarchical structure describing an SBP, which in the weighted case is identical to the standard case. The decomposition of the total variance changes quantitatively with weighting, as indicated in Equation 14 (second member). Finally, the position of mean balances is substituted by the new mean weighted balances. However, the qualitative structure of the dendrogram remains the same.

The present study of different exploratory tools for compositional data analysis is only preliminary. Details on interpretation and methods to assess weights require further study.

7. Conclusions and further research

A weighting strategy for the analysis of compositions is proposed. It is based on the theory of Bayes Hilbert spaces. However, some modifications have been introduced to fulfill the principle of dominance of distances when down-weighting some parts of the composition. When the weights considered are unitary in each part, that is, when there is no down or up-weighting, the approach is reduced to the standard compositional data analysis. If some parts are down-weighted approaching zero, the weighted geometry of the simplex tends to the ordinary Aitchison geometry of the corresponding subcomposition.

In order to use the proposed weighting approach, it is advisable to deal with compositional data as usual for linear operations, using the standard perturbation and powering. When distances or inner products are involved in the analysis, they are computed in two steps: first, shifting the compositional data by $\ominus\mathbf{p}$, that is, dividing each part by the corresponding weight; and second, computing $\mathrm{clr}_{\mathbf{p}}$ (Equations 3 or 16) or $\mathrm{ilr}_{\mathbf{p}}$ (Equation 10) to find the required distances or inner products in a straightforward way.

Statistical consequences of weighting compositions need to be studied in the future. Standard tools of exploratory analysis, as variation matrix, biplots or balance-dendrogram, clustering and others, will be influenced by weighting. The reason is that distances between compositions and computation of variances-covariances are influenced as well. Thus, the proposed weighting approach is only a first step towards developing effective weighting techniques applicable to compositional data analysis.

Acknowledgements

This research has been supported by the *Spanish Ministry of Education and Science* under projects 'METRICS' (Ref. MTM2012-33236) and 'CODA-RETOS' (Ref. MTM2015-65016-C2-1-R); and by the *Agència de Gestió d'Ajuts Universitaris i de Recerca* of the *Generalitat de Catalunya* under the project Ref. 2009SGR424. We thank the deep and detailed revision by K. Hron and an anonymous reviewer, which lead to improvements of the original contribution.

References

Aitchison J (1983). "Principal component analysis of compositional data." *Biometrika*, **70**(1), 57–65.

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), London (UK). ISBN 0-412-28060-4. 416 p.

Aitchison J (1992). "On Criteria for Measures of Compositional Difference." *Mathematical Geology*, **24**(4), 365–379.

Aitchison J, Barceló-Vidal C, Martín-Fernández JA, Pawlowsky-Glahn V (2000). "Logratio

analysis and compositional distance." *Mathematical Geology*, **32**(3), 271–275. ISSN 0882-8121.

Aitchison J, Egozcue JJ (2005). "Compositional data analysis: where are we and where should we be heading?" *Mathematical Geology*, **37**(7), 829–850.

Aitchison J, Greenacre M (2002). "Biplots for compositional data." *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **51**(4), 375–392.

Boogaart KGvd, Egozcue JJ, Pawlowsky-Glahn V (2010). "Bayes linear spaces." *SORT - Statistics and Operations Research Transactions*, **34**(2), 201–222. ISSN 1696-2281.

Boogaart KGvd, Egozcue JJ, Pawlowsky-Glahn V (2014). "Bayes Hilbert Spaces." *Australian and New Zealand Journal of Statistics*, **56**(2), 171–194. doi:10.1111/anzs.12074.

Boogaart KGvd, Tolosana-Delgado R (2013). *Analysing compositional data with R*. Springer, Heidelberg. 280 pp.

Egozcue JJ (2009). "Reply to "On the Harker variation diagrams;..." by J. A. Cortés." *Mathematical Geosciences*, **41**(7), 829–834.

Egozcue JJ, Barceló-Vidal C, Martín-Fernández JA, Jarauta-Bragulat E, Díaz-Barrero JL, Mateu-Figueras G (2011). "Elements of simplicial linear algebra and geometry." In Pawlowsky-Glahn and Buccianti (2011), pp. 141–157. 378 p.

Egozcue JJ, Díaz-Barrero JL, Pawlowsky-Glahn V (2006). "Hilbert space of probability density functions based on Aitchison geometry." volume 22, pp. 1175–1182. DOI: 10.1007/s10114-005-0678-2.

Egozcue JJ, Lovell D, Pawlowsky-Glahn V (2013a). *Testing compositional association*. In: Proceedings of the 5th Workshop on compositional data analysis, CoDaWork 2013, ISBN: 978-3-200-03103-6. Pp 28–36.

Egozcue JJ, Pawlowsky-Glahn V (2005). "Groups of parts and their balances in compositional data analysis." *Mathematical Geology*, **37**(7), 795–828.

Egozcue JJ, Pawlowsky-Glahn V (2006). "Simplicial geometry for compositional data." In *Compositional Data Analysis in the Geosciences: From Theory to Practice*, pp. 145–159.

Egozcue JJ, Pawlowsky-Glahn V (2011). "Basic concepts and procedures." In Pawlowsky-Glahn and Buccianti (2011), pp. 12–28. 378 p.

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003). "Isometric logratio transformations for compositional data analysis." *Mathematical Geology*, **35**(3), 279–300. ISSN 0882-8121.

Egozcue JJ, Pawlowsky-Glahn V, Tolosana-Delgado R, Ortego MI, Boogaart KGvd (2013b). "Bayes spaces: use of improper distributions and exponential families." *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales, Serie A, Matemáticas (RACSAM)*, **107**, 475–486. DOI 10.1007/s13398-012-0082-6.

Filzmoser P, Hron K (2015). "Robust coordinates for compositional data using weighted balances." *In Nordhausen, K. and Taskinen, S., (eds.), Modern Nonparametric, Robust and Multivariate Methods. Springer, Heidelberg.*

Fréchet M (1948). "Les éléments Aléatoires de Nature Quelconque dans une Espace Distancié." *Annales de l'Institut Henri Poincaré*, **10**(4), 215–308.

Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bähler J (2015). "Proportionality: A Valid Alternative to Correlation for Relative Data." *PLoS Comput Biol*, **11**(3), e1004075. doi:10.1371/journal.pcbi.1004075. URL http://dx.doi.org/10.1371%2Fjournal.pcbi.1004075.

Mateu-Figueras G, Pawlowsky-Glahn V, Egozcue JJ (2013). "The normal distribution in some constrained sample spaces." *SORT - Statistics and Operations Research Transactions*, **37**(1), 29–56. ISSN 1696-2281.

Pawlowsky-Glahn V, Buccianti A (eds.) (2011). *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons. ISBN 978–0–470–71135–4. 378 p.

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

Pawlowsky-Glahn V, Egozcue JJ (2011). "Exploring Compositional Data with the Coda-Dendrogram." *Austrian Journal of Statistics*, **40**(1 & 2), 103–113. ISSN 1026-597X.

Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015). *Modeling and analysis of compositional data*. Statistics in practice. John Wiley & Sons, Chichester UK. ISBN 9781118443064. 272 pp.

A. Dominance of distances under change of reference

In Section 4 the following proposition was stated:

PROPOSITION (dominance of distances). *Let* $\mathbf{x}_1$, $\mathbf{x}_2$ *be two compositions in* $\mathcal{S}^D$, *endowed with the reference measure* P_0, *which weights are* $\mathbf{p}_0 = (1, 1, \ldots, 1)$. *Consider two reference measures,* P_1 *and* P_2, *represented by their respective weights* $\mathbf{p}_1 = (p_{11}, p_{12}, \ldots, p_{1D})$ *and* $\mathbf{p}_2 = (p_{21}, p_{22}, \ldots, p_{2D})$, *such that all their components are* $0 < p_{ki} \leq 1$, *for* $k = 1, 2$, $i = 1, 2, \ldots, D$ *and* $P_k(\Omega) = \sum_{i=1}^{D} p_{ki}$. *Define* $\mathbf{y}_j^{(\mathbf{p}_k)} = \mathbf{x}_j / \mathbf{p}_k$ *for* $k = 1, 2$ *and* $j = 1, 2$. *Then,*

$$p_{1i} \leq p_{2i},\ i = 1, 2, \ldots, D \quad \Rightarrow \quad \mathrm{d}_{\mathbf{p}_1}(\mathbf{y}_1^{(\mathbf{p}_1)}, \mathbf{y}_2^{(\mathbf{p}_1)}) \leq \mathrm{d}_{\mathbf{p}_2}(\mathbf{y}_1^{(\mathbf{p}_2)}, \mathbf{y}_2^{(\mathbf{p}_2)}) \ .$$

Proof: The change of reference from $\mathbf{p}_2$ to $\mathbf{p}_1$ with $p_{1i} \leq p_{2i}$, $i = 1, 2, \ldots, D$, can be conceived as a sequence of intermediate changes of reference for which only one weight p_{2i} is changed to p_{1i} at each step. These steps can be ordered, for instance, with the index $i = 1, 2, \ldots, D$. The sequence of weights can be the following.

step	initial reference		final reference
1-st	$\mathbf{p}_2 = (p_{21}, p_{22}, \ldots, p_{2D})$	to	$\mathbf{q}_1 = (p_{11}, p_{22}, \ldots, p_{2D})$
2-nd	$\mathbf{q}_1 = (p_{11}, p_{22}, \ldots, p_{2D})$	to	$\mathbf{q}_2 = (p_{11}, p_{12}, \ldots, p_{2D})$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
i-th	$\mathbf{q}_{i-1} = (p_{11}, p_{12}, \ldots, p_{1,i-1}, p_{2i}, \ldots, p_{2D})$	to	$\mathbf{q}_i = (p_{11}, p_{12}, \ldots, p_{1i}, p_{2,i+1}, \ldots, p_{2D})$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
D-th	$\mathbf{q}_{D-1} = (p_{11}, p_{12}, \ldots, p_{1,D-1}, p_{2D})$	to	$\mathbf{p}_1 = (p_{11}, p_{12}, \ldots, p_{1D})$

As one weight decreases at each step, the statement is proven if the distance $\mathrm{d}_{\mathbf{q}_i}(\mathbf{y}_1^{(\mathbf{q}_i)}, \mathbf{y}_2^{(\mathbf{q}_i)})$ is less than or equal to $\mathrm{d}_{\mathbf{q}_{i-1}}(\mathbf{y}_1^{(\mathbf{q}_{i-1})}, \mathbf{y}_2^{(\mathbf{q}_{i-1})})$, for $i = 1, 2, \ldots, D$, where $\mathbf{q}_D = \mathbf{p}_1$. The i-th step consists of changing the weight p_{2i} into p_{1i}, while all other weights remain equal. Consider that $\mathrm{ilr}_{\mathbf{q}_i}$ corresponds to a partition (SBP) that separates the i-th part of the composition

from the other $D-1$ parts. For both sets of weights $\mathbf{q}_i$ and $\mathbf{q}_{i-1}$ all weighted balances are equal except the first one, denoted $b_i^{(\mathbf{q}_k)}$, $k = i-1, i$. Equation 12 implies that

$$\mathrm{d}^2_{\mathbf{q}_{i-1}}(\mathbf{y}_1^{(\mathbf{q}_{i-1})}, \mathbf{y}_2^{(\mathbf{q}_{i-1})}) - \mathrm{d}^2_{\mathbf{q}_i}(\mathbf{y}_1^{(\mathbf{q}_i)}, \mathbf{y}_2^{(\mathbf{q}_i)})$$

$$= \left[b_i^{(\mathbf{q}_{i-1})}(\mathbf{y}_1^{(\mathbf{q}_{i-1})}) - b_i^{(\mathbf{q}_{i-1})}(\mathbf{y}_2^{(\mathbf{q}_{i-1})})\right]^2 - \left[b_i^{(\mathbf{q}_i)}(\mathbf{y}_1^{(\mathbf{q}_i)}) - b_i^{(\mathbf{q}_i)}(\mathbf{y}_2^{(\mathbf{q}_i)})\right]^2 . \tag{15}$$

Using the expression of balances (11), it holds

$$b_i^{(\mathbf{q}_k)}(\mathbf{y}_\ell^{(\mathbf{q}_k)}) = \sqrt{\frac{q_{ki} n_i^-}{q_{ki} + n_i^-}} \log \frac{y_{\ell i}}{\prod_{j \neq i} y_{\ell j}^{q_{ki}/n_i^-}} , \quad k = i-1, i, \quad \ell = 1, 2,$$

where $q_{ki} = p_{2i}$ if $k = i-1$, and $q_{ki} = p_{1i}$ if $k = i$. Moreover, the values of the parts of the compositions are $y_{\ell j} = x_j / p_{2j}$ if $j < i$ and $y_{\ell j} = x_j / p_{1j}$ if $j > i$. As a result, the differences of balances in Equation 15 simplify to

$$b_i^{(\mathbf{q}_k)}(\mathbf{y}_1^{(\mathbf{q}_k)}) - b_i^{(\mathbf{q}_k)}(\mathbf{y}_2^{(\mathbf{q}_k)}) = \sqrt{\frac{q_{ki} n_i^-}{q_{ki} + n_i^-}} \log \left(\frac{x_{1i} x_{2i}}{\prod_{j \neq i} (x_j / q_{kj})^{q_{kj}/n_i^-}} \right) ,$$

where the closure constants associated with the change $\mathbf{x}_\ell = \mathbf{y}_\ell \oplus \mathbf{q}_k$ cancel within the balance, as it is scale invariant. Remarkably, the logarithmic term does not depend on k, as the weights q_{kj}, $j \neq i$, are equal for $\mathbf{q}_k$, $k = i-1, i$. Substituting these differences of balances in Equation 15 it yields

$$\mathrm{d}^2_{\mathbf{q}_{i-1}}(\mathbf{y}_1^{(\mathbf{q}_{i-1})}, \mathbf{y}_2^{(\mathbf{q}_{i-1})}) - \mathrm{d}^2_{\mathbf{q}_i}(\mathbf{y}_1^{(\mathbf{q}_i)}, \mathbf{y}_2^{(\mathbf{q}_i)}) = \frac{p_{2i} n_i^-}{p_{2i} + n_i^-} - \frac{p_{1i} n_i^-}{p_{1i} + n_i^-} = \frac{n_i^- (p_{2i} - p_{1i})}{(p_{2i} + n_i^-)(p_{1i} + n_i^-)} \geq 0 ,$$

since it was assumed that $p_{2i} \geq p_{1i}$. This proves the statement. □

B. Relationship between ordinary and weighted clr and ilr

In this appendix the relationship between ordinary and weighted clr and ilr is studied. The main goal is to prove that this relationship is linear up to additive terms. The expressions obtained are not central in the developed theory but they help to understand how the probability distributions of random compositions change under change of reference.

Let $\mathbf{x}$ be a composition in $\mathcal{S}^D$, taken as a density of a measure μ with respect to the uniform reference measure P_0, given by $\mathbf{p}_0 = (1, 1, \ldots, 1)$. An alternative reference measure represented by the weights $\mathbf{p} = (p_1, p_2, \ldots, p_D)$ is considered, and the corresponding density of μ is then $\mathbf{y} = \mathbf{x}/\mathbf{p}$. The weighted centered log-ratio of $\mathbf{y}$ (Equation 3) is

$$\mathrm{clr}_\mathbf{p}(\mathbf{y}) = \log \mathbf{y} - \log(\mathrm{g}_\mathbf{p}(\mathbf{y}))\mathbf{1} ,$$

where $\mathrm{g}_\mathbf{p}(\cdot)$ denotes the weighted geometric mean of the arguments (Equation 3), $\mathbf{1}$ is a row vector of D ones and $\log \mathbf{x}$, $\log \mathbf{y}$ and $\log \mathbf{p}$ are taken as row vectors. Then, $\log(\mathrm{g}_\mathbf{p}(\mathbf{y}))\mathbf{1}$ is a row vector with all components equal to $\mathrm{g}_\mathbf{p}(\mathbf{y})$. Moreover, $\log(\mathbf{y}) = \log(\mathbf{x}) - \log(\mathbf{p})$ and $\log(\mathrm{g}_\mathbf{p}(\mathbf{y})) = (1/\mathbf{s}_\mathbf{p}) \sum p_i (\log x_i - \log p_i)$, with $s_\mathbf{p} = \sum p_i$. Using matrix notation this leads to

$$\log(\mathrm{g}_\mathbf{p}(\mathbf{y}))\mathbf{1} = \frac{1}{s_\mathbf{p}} (\log \mathbf{x} - \log \mathbf{p}) \, \mathbf{p}^\top \mathbf{1} .$$

Substitution into the definition of $\mathrm{clr}_\mathbf{p}(\mathbf{y})$ yields

$$\mathrm{clr}_\mathbf{p}(\mathbf{y}) = (\log \mathbf{x} - \log \mathbf{p}) \left[I_D - \frac{1}{s_\mathbf{p}} \mathbf{p}^\top \mathbf{1} \right] , \tag{16}$$

where I_D is the (D, D)-identity matrix. Equation 16 shows that $\mathrm{clr}_{\mathbf{p}}(\mathbf{y})$ is a linear transformation of $\log \mathbf{x}$ up to additive terms depending on $\mathbf{p}$. The ordinary $\mathrm{clr}(\mathbf{x})$ can be written

$$\mathrm{clr}(\mathbf{x}) = \log \mathbf{x} \left[I_D - \frac{1}{D} \mathbf{1}^\top \mathbf{1} \right] ,$$

which can be substituted into Equation 16. The resulting expression is

$$\mathrm{clr}_{\mathbf{p}}(\mathbf{y}) = \mathrm{clr}(\mathbf{x}) - \log(\mathbf{p}) - \log(\mathrm{g}_{\mathbf{p}}(\mathbf{x}))\mathbf{1} + \log(\mathrm{g}_{\mathbf{p}}(\mathbf{p}))\mathbf{1} + \log(\mathrm{g}(\mathbf{x}))\mathbf{1} . \tag{17}$$

In order to relate ordinary and weighted ilr, assume that ilr-coordinates of $\mathbf{x}$ are

$$\mathrm{ilr}(\mathbf{x}) = \mathrm{clr}(\mathbf{x})\ \Psi_0^\top \quad , \text{ with } \quad \Psi_0 \Psi_0^\top = I_{D-1} \ , \ \Psi_0^\top \Psi_0 = I_D - \frac{1}{D} \mathbf{1}^\top \mathbf{1} \ ,$$

that is, Ψ_0 is an ordinary contrast matrix (Egozcue *et al.* 2011). The weighted $\mathrm{ilr}_{\mathbf{p}}$-coordinates are computed as in Equation 10 using the weighted contrast matrix Ψ,

$$\mathrm{ilr}_{\mathbf{p}}(\mathbf{y}) = \mathrm{clr}_{\mathbf{p}}(\mathbf{y})\ \mathrm{diag}(\mathbf{p})\ \Psi^\top .$$

Substituting Equation 17 and taking into account that $\mathbf{1}\ \mathrm{diag}(\mathbf{p})\ \Psi^\top = \mathbf{0}$, it yields

$$\mathrm{ilr}_{\mathbf{p}}(\mathbf{y}) = (\mathrm{clr}(\mathbf{x}) - \log \mathbf{p})\ \mathrm{diag}(\mathbf{p})\ \Psi^\top .$$

Inserting $I_D = \Psi_0^\top \Psi_0 + (1/D)\mathbf{1}^\top \mathbf{1}$ after $\mathrm{clr}(\mathbf{x})$, the desired relationship is

$$\mathrm{ilr}_{\mathbf{p}}(\mathbf{y}) = \mathrm{ilr}(\mathbf{x})\ \Psi_0\ \mathrm{diag}(\mathbf{p})\ \Psi^\top - \log \mathbf{p}\ \mathrm{diag}(\mathbf{p})\ \Psi^\top , \tag{18}$$

which shows that $\mathrm{ilr}_{\mathbf{p}}(\mathbf{y})$ is a linear transformation of $\mathrm{ilr}(\mathbf{x})$, up to additive terms depending only on $\mathbf{p}$ and the selected weighted contrast matrix Ψ.

Affiliation:

Juan José Egozcue
Universitat Politecnica de Catalunya
Jordi Girona 1-3, C2-UPC
E-08034 Barcelona, Spain
E-mail: juan.jose.egozcue@upc.edu

Vera Pawlowsky-Glahn
Universitat de Girona
Campus Montilivi, P4
E-17071 Girona, Spain
E-mail: vera.pawlowsky@udg.edu

http://www.ajs.or.at/
http://www.osg.or.at/
Submitted: 2015-09-23
Accepted: 2016-04-20