Mathematical Foundations of Compositional Data Analysis

Carles Barceló-Vidal, Josep A. Martín-Fernández and Vera Pawlowsky-Glahn

Departament d'Informàtica i Matemàtica Aplicada, Universitat de Girona, Campus de Montilivi, E-17071 Girona, Spain. e-mail: carles.barcelo@udg.es

Summary. The definition of composition like a vector whose components are subject to the restriction of constant sum is revised from a mathematical point of view. Starting from the definition of an equivalence's relation on the positive orthant of the multidimensional real space $\mathbf{R}^{d+1}$, the compositions become equivalence classes of elements of $\mathbf{R}^{d+1}$, and the set of them —i.e., the corresponding quotient set— the compositional space $\mathcal{C}^d$. In this way, the simplex becomes one, among many, of the possible representations of this quotient space. The logarithmic transformation defines a one-to-one transformation between $\mathcal{C}^d$ and a suitable vector quotient space $\mathcal{L}^d$ defined on the multidimensional real space. This transformation allows to transfer the Euclidean structure easily defined on $\mathcal{L}^d$ to the compositional space. It is showed that, from a mathematical point of view, the methodology introduced by Aitchison is fully compatible with the nature of compositional data, and is independent of the representation used to manage the compositional data.

Keywords: Compositional data, closed data

1. Introduction

Traditionally, *compositional data* have been identified with *closed data*, and the simplex has been considered as the natural sample space of this kind of data.

We think that the emphasis on the constrained nature of compositional data has contributed to mask its real nature. In our opinion, this is the principal reason for most controversies (Watson and Philip, 1989; Aitchison, 1989, 1990a; Watson, 1990; Aitchison, 1991; Watson, 1991; Tangri and Wright, 1993; Baxter, 1993; Whitten, 1995; Bohling et al., 1996; Woronow, 1997a,b; Zier and Rehder, 1998; Barceló-Vidal et al., 1999; Tauber, 1999; Aitchison et al., 2001) generated by the methodology introduced by Aitchison (1986) in his monograph and pursued further in Aitchison (1990b, 1992, 1997, 1999, 2001), Aitchison and Thomas (1998), Aitchison and Bacon-Shone (1999), Aitchison and Greenacre (2001), Barceló-Vidal (1996), Barceló-Vidal and Pawlowsky-Glahn (1994), Barceló-Vidal et al. (1995, 1996), Martín-Fernández (2001), Martín-Fernández et al. (1997, 1998a,b,c, 1999, 2000), Mateu-Figueras et al. (1998), Pawlowsky-Glahn and Barceló-Vidal (1999), Pawlowsky-Glahn and Egozcue (2001a,b).

More crucial than the constraining property of compositional data is the *scale-invariant property* of this kind of data. Indeed, when we are considering only some parts of a full composition, our data are still compositional, although the constraint condition is not accomplished. This fact was acknowledged in Aitchison (1992, 1997) introducing his argument that any sensible statistical analysis on compositional data had to be based on logratios. From a mathematical point of view, we believe that it is necessary to give a wider definition of the concept of composition. This is done in Section 2 where the *compositional equivalence* *relation* in the positive orthant of the multidimensional real space $\mathbb{R}^D$ is introduced. In this manner, the space of all compositions —the *compositional space* $\mathcal{C}^d$, where $d = D - 1$— is a quotient space, and the d-dimensional *simplex* $\mathcal{S}^d$ is only one way, among many, to represent $\mathcal{C}^d$. In this manner, it is even more evident that any analysis performed on compositional data should and can be independent of their representation.

From this wider definition of composition and with the help of the logarithmic and the exponential transformations, we develop in Sections 3 to 6 the successive steps to define an Euclidean structure on the compositional space $\mathcal{C}^d$. First, an Euclidean structure is defined on a suitable real quotient vector space $\mathcal{L}^d$ of $\mathbb{R}^D$, and then this structure is translated to $\mathcal{C}^d$ by means of the exponential transformation. All of these results are more extensively developed in Barceló-Vidal (Barceló-Vidal).

2. The space of compositions

2.1. First definitions

Any $D \times 1$ real vector $\mathbf{w} = (w_1, \ldots, w_D)'$ with positive components or parts will be called a *D-observational vector*. The set of all these vectors is the D-dimensional positive real space $\mathbb{R}^D_+$, the positive orthant of $\mathbb{R}^D$.

DEFINITION 1. Two D-observational vectors $\mathbf{w}$ and $\mathbf{w}^*$ are *compositionally equivalent*, written $\mathbf{w} \sim \mathbf{w}^*$, if there exists a positive constant k such that $\mathbf{w} = k\mathbf{w}^*$. This equivalence relation on $\mathbb{R}^D_+$ partitions the space in equivalence classes, called *D- part compositions* or, briefly, *compositions*.
The composition generated by an observational vector $\mathbf{w}$ —i.e., the equivalence class of $\mathbf{w}$— is symbolized by $\underline{\mathbf{w}}$:

$$\underline{\mathbf{w}} = \{k\mathbf{w} : k \in \mathbb{R}^+\}.$$

The set of all D-part compositions —i.e., the quotient space $\mathbb{R}^D_+/\!\sim$— is called the *compositional space*, and is symbolized by $\mathcal{C}^d$, where $d = D - 1$
The quotient mapping from $\mathbb{R}^D_+$ to $\mathcal{C}^d$ which assigns the class $\underline{\mathbf{w}}$ to each D-observational vector $\mathbf{w}$ will be denoted by ccl (from $\underline{\text{c}}$ompositional $\underline{\text{cl}}$ass):

$$\operatorname{ccl}\mathbf{w} = \underline{\mathbf{w}} \qquad (\mathbf{w} \in \mathbb{R}^D_+).$$

□

In Section 4, we justify why the set of all D-part compositions is symbolized by $\mathcal{C}^d$ and not by $\mathcal{C}^D$. A D-part composition can be geometrically interpreted as a ray from the origin in the positive orthant of $\mathbb{R}^D$ (see Fig. 1a).

The equivalence relation given in Definition 1 can be reformulated from the ratios of the components of observational vectors.

PROPERTY 1. Two D-observational vectors $\mathbf{w} = (w_1, \ldots, w_D)'$ and $\mathbf{w}^* = (w_1^*, \ldots, w_D^*)'$ are compositionally equivalent if and only if

$$\frac{w_i}{w_j} = \frac{w_i^*}{w_j^*} \quad \text{for each } i, j = 1, \ldots, D.$$

□

2.2. Selection criteria

Any composition $\underline{\mathbf{w}}$ is determined by any observational vector belonging to the equivalence class $\underline{\mathbf{w}}$. Thus, different criteria can be used to select a specific observational vector to represent a composition, leading to interesting results.

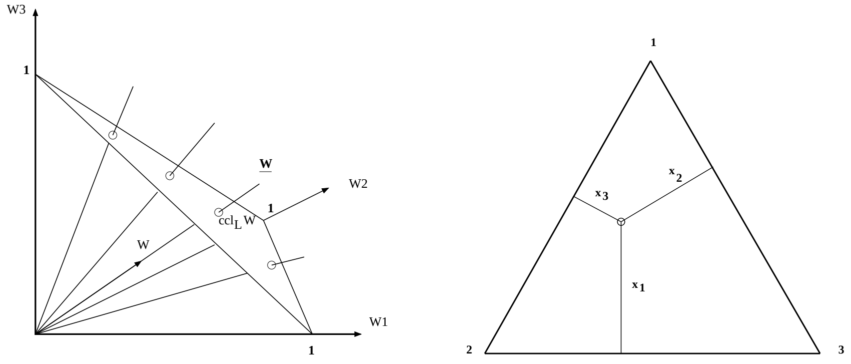
We symbolize by ccl_L the operator which transforms each D-observational vector $\mathbf{w}$ into the unit-sum vector $\mathbf{w}/\sum_{j=1}^{D} w_j$. This operator corresponds to the *constraining operator* $\mathcal{C}$ introduced in Aitchison (1986, p. 31). It is clear that $\mathbf{w} \sim \mathrm{ccl}_L\,\mathbf{w}$, and that the operator ccl_L is constant on the vectors belonging to the same compositional equivalence class, i.e., if $\mathbf{w} \sim \mathbf{w}^*$, then $\mathrm{ccl}_L\,\mathbf{w} = \mathrm{ccl}_L\,\mathbf{w}^*$.

DEFINITION 2. The operation which selects from each composition $\underline{\mathbf{w}}$ the compositionally equivalent unit-sum observational vector $\mathrm{ccl}_L\,\mathbf{w}$ is called *linear criterion*. □

Geometrically, $\mathrm{ccl}_L\,\mathbf{w}$ is the intersection of the ray going from the origin through $\mathbf{w}$ and the hyperplan of $\mathbb{R}^D$ defined by the equation $w_1 + \ldots + w_D = 1$ (see Fig. 1a). The set of all these points is the well-known d-dimensional regular *simplex*:

$$\mathcal{S}^d = \{(w_1, \ldots, w_D)' : w_1 > 0, \ldots, w_D > 0; w_1 + \ldots + w_D = 1\}.$$

The simplex $\mathcal{S}^2$ corresponds to *ternary diagram*, an equilateral triangle of unit altitude. For any point P in triangle 123 (see Fig. 1b) the perpendiculars w_1, w_2 and w_3 from P to the sides opposite 23, 13 and 12, respectively, satisfy $w_1 + w_2 + w_3 = 1$. Similarly, the simplex $\mathcal{S}^3$ corresponds to a regular tetrahedron 1234 of unit altitude.



(a) (b)

Fig. 1. (a) The 3-part compositions are interpreted as rays from the origin of $\mathbf{R}^3_+$. Linear selection criterion; (b) The simplex $\mathcal{S}^2$.

We symbolise by ccl_E the operator which transforms each D-observational vector $\mathbf{w}$ into the unit-length vector $\mathbf{w}/\|\mathbf{w}\|$. It is clear that $\mathbf{w} \sim \mathrm{ccl}_E\,\mathbf{w}$, and that the operator ccl_E is constant on the vectors belonging to the same compositional equivalence class.

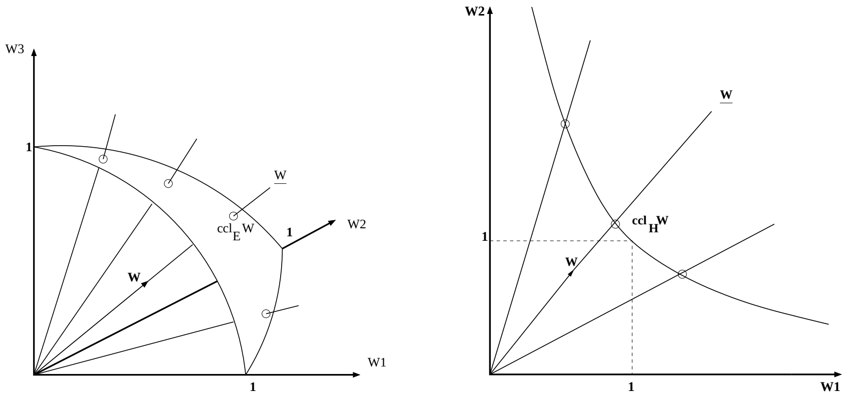
DEFINITION 3. The operation which selects from each composition $\underline{\mathbf{w}}$ the observational unit-length vector $\mathrm{ccl}_E\,\mathbf{w}$ is called *spherical criterion*. □

Geometrically, $\mathrm{ccl}_E\,\mathbf{w}$ is the intersection of the ray going from the origin through $\mathbf{w}$ and the unit sphere of $\mathbb{R}^D$ centered in the origin (see Fig. 2a).

We symbolise by ccl_H the operator which transforms each D-observational vector $\mathbf{w}$ into the unit-product vector $\mathbf{w}/g(\mathbf{w})$, where $g(\mathbf{w}) = (\prod_{j=1}^{D} w_j)^{1/D}$ is the geometric mean of the components of vector $\mathbf{w}$. It is clear that $\mathbf{w} \sim \mathrm{ccl}_H\,\mathbf{w}$, and that the operator ccl_H is constant on the vectors belonging to the same compositional equivalence class.

Definition 4. The operation which selects from each composition $\underline{\mathbf{w}}$ the unit-product observational vector $\mathrm{ccl}_H\,\mathbf{w}$ is called *hyperbolic criterion* . □

Geometrically, $\mathrm{ccl}_H\,\mathbf{w}$ is the intersection of the ray given from the origin through $\mathbf{w}$ and the hyperbolic surface Hip_D in $\mathbb{R}_+^D$ defined by the generic equation $\prod_{i=1}^{D} w_i = 1$ (see Fig. 2b).



(a) (b)

Fig. 2. (a) Spherical selection criterion (case D=3); (b) Hyperbolic selection criterion (case D=2).

Figure 3 shows the three criteria of selection for $D = 2$ in the same graph.

2.3. *Compositional nature of data*

When the components of the observational vectors of our data set represent ratios of a fixed total —i.e, when they represent *relative* magnitudes—, the data is clearly *compositional* because the components provide to us only relative information and not absolute information. In this case, only *ratios* between components are meaningful, and those ratios are independent from the arbitrary total.

Sometimes, the components of the observational vectors are themselves meaningful, i.e., they represent absolute magnitudes. But, in spite of that, we can decide to take only account the information given by the ratios between the components. If so, we are implicitly assuming that the vectors $\mathbf{w}$ and $k\mathbf{w}$, with $k > 0$, are providing to us the same information. Therefore, in this case, we are also handling our data as *compositional* data.

In any case, if our data set is a compositional data set, our analysis have to be independent of the observational vectors used to represent the compositions. This is equivalent

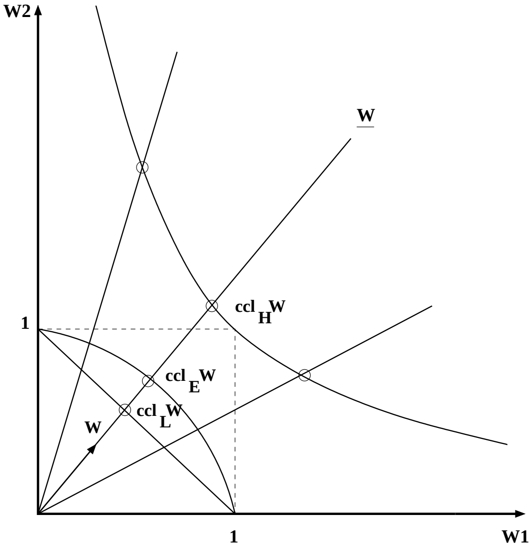


Fig. 3. The three selection criteria (case $D = 2$).

to say that any meaningful analysis can be expressed in terms of ratios of components of the observationals vectors. Or, in a more mathematical form, it is equivalent to work with the equivalence classes of the compositional space $\mathcal{C}^d$. Therefore, althoug the representation of the compositional space on the simplex might be considered the best and the simplest representation, we have to be able to perform our analysis independently of the representation.

2.4. Subcompositions

Sometimes, we focus our attention on the relative magnitudes of a subset of parts of a composition.

DEFINITION 5. Given a composition $\underline{\mathbf{w}} \in \mathcal{C}^d$, any composition obtained from the selection of two or more parts of the D-observational vector $\mathbf{w}$ is termed a *subcomposition* of $\underline{\mathbf{w}}$. □

Let be C the number of selected parts, with $2 \leq C < D$. We symbolize by S the ordered subset of indices of the selected parts of $\mathbf{w}$ to be included in the subcomposition. We write $\mathbf{w}_S$ to indicate the observational subvector formed from the corresponding parts of $\mathbf{w}$, and therefore $\underline{\mathbf{w}_S}$ represents the final subcomposition, which belongs to the compositional space $\mathcal{C}^c$, where $c = C - 1$.

DEFINITION 6. Given an ordered set S composed by C different indices from $\{1, \ldots, D\}$, the formation of a subcomposition can be considered as the transformation sub_S from $\mathcal{C}^d$ to $\mathcal{C}^c$ given by

$$\begin{array}{rccc} \mathrm{sub}_S : & \mathcal{C}^d & \to & \mathcal{C}^c \\ & \underline{\mathbf{w}} & \to & \underline{\mathbf{w}_S} \end{array} . \tag{1}$$

□

It is obvious that $\mathrm{sub}_S\underline{\mathbf{w}}$ is independent of the observational vector selected to represent the composition $\underline{\mathbf{w}}$.

Geometrically, the formation of a subcomposition $\underline{\mathbf{w}_S}$ from a D-part composition $\underline{\mathbf{w}}$ corresponds to the orthogonal projection of the ray associated to $\underline{\mathbf{w}}$ in $\mathbb{R}_+^D$ onto a subspace of dimension C. This subspace is generated by the coordinate axes associated to the parts selected to form the subcomposition (see Fig. 4).

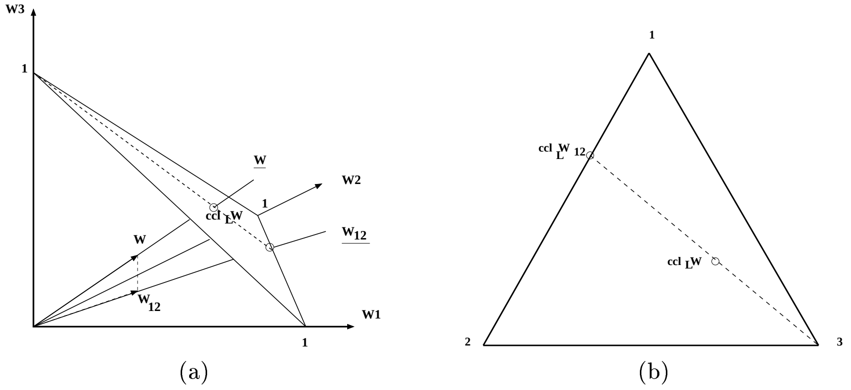


Fig. 4. Geometrical interpretation of the formation of the subcomposition $\underline{\mathbf{w}_{12}}$ from a composition $\underline{\mathbf{w}} \in \mathcal{C}^2$: (a) In $\mathbf{R}_+^3$; (b) In $\mathcal{S}^2$.

Obviously, the subcompositional mapping sub_S from $\mathcal{C}^d$ to $\mathcal{C}^c$ is not injective. In spite of that, it would be interesting to define an application from $\mathcal{C}^c$ into $\mathcal{C}^d$ which univocally sends each composition of $\mathcal{C}^c$ to a composition of $\mathcal{C}^d$. In order to not complicate unnecessarily the notation, we will assume that $S = \{D - C + 1, \ldots, D\}$.

Definition 7. We define the mapping inc_S from $\mathcal{C}^c$ to $\mathcal{C}^d$ as

$$\mathrm{inc}_S\underline{\mathbf{w}} = \mathrm{ccl}\left(1, \overset{D-C}{\ldots}, 1, \frac{w_1}{g(\mathbf{w})}, \ldots, \frac{w_C}{g(\mathbf{w})}\right)' \quad (\underline{\mathbf{w}} \in \mathcal{C}^c). \tag{2}$$

□

It is clear that $\mathrm{inc}_S\underline{\mathbf{w}}$ is independent of the observational vector selected to represent the composition $\underline{\mathbf{w}}$.

Property 2. The mapping inc_S from $\mathcal{C}^c$ to $\mathcal{C}^d$ is injective. Moreover, the composed mapping $\mathrm{sub}_S \circ \mathrm{inc}_S$ is the identity mapping on $\mathcal{C}^c$. □

3. The logarithmic transformation on the compositional space

3.1. A quotient space in $\boldsymbol{R}^D$

The logarithmic transformation from $\mathbb{R}_+^D$ to $\mathbb{R}^D$ suggests to define in $\mathbb{R}^D$ an equivalence relation in correspondence with the compositional equivalence relation defined in $\mathbb{R}_+^D$. Certainly, if $\mathbf{w} \sim \mathbf{w}^*$, then $\log \mathbf{w} - \log \mathbf{w}^*$ of $\mathbb{R}^D$ is a multiple of the vector of unities $\mathbf{1}_D = (1, \ldots, 1)' \in \mathbb{R}^D$.

DEFINITION 8. Two vectors $\mathbf{z}$ and $\mathbf{z}^*$ in $\mathbb{R}^D$ are *equivalent*, written $\mathbf{z} \equiv \mathbf{z}^*$, iff there exists a constant λ such that $\mathbf{z}^* = \mathbf{z} + \lambda \mathbf{1}_D$. If we consider the one-dimensional subspace $U = \{\lambda \mathbf{1}_D : \lambda \in \mathbb{R}\}$ of $\mathbb{R}^D$, the previous equivalence relation can be also defined as

$$\mathbf{z} \equiv \mathbf{z}^* \quad \Longleftrightarrow \quad \mathbf{z} - \mathbf{z}^* \in U.$$

□

Therefore, it is natural to symbolize by $\mathbf{z} + U$ the equivalence class or coset generated by the vector $\mathbf{z}$ in $\mathbb{R}^D$. The set of all these cosets —i.e., the quotient space $\mathbb{R}^D/U$— will be symbolized by $\mathcal{L}^d$, where $d = D - 1$.

The quotient mapping from $\mathbb{R}^D$ to $\mathcal{L}^d$ which assigns the class $\mathbf{z} + U$ to each vector $\mathbf{z} \in \mathbb{R}^D$ will be denoted by ucl:

$$\operatorname{ucl} \mathbf{z} = \mathbf{z} + U \qquad (\mathbf{z} \in \mathbb{R}^D).$$

From Figure 5 it is clear that the cosets $\mathbf{z} + U$ can be geometrically interpreted by straight lines parallel to $\mathbf{1}_D$. It seems "natural" to represent a equivalence class $\mathbf{z} + U$ by the point of intersection of the straight line associated to this coset and the hyperplan $V = \{\mathbf{z} \in \mathbb{R}^D : \mathbf{z}'\mathbf{1}_D = 0\}$ of $\mathbb{R}^D$ by the origin, orthogonal to $\mathbf{1}_D$. This point of intersection can also be interpreted as the common orthogonal projection onto V of all the vectors belonging to the equivalence class $\mathbf{z} + U$.

We symbolize by ucl_V the operator which transforms each vector $\mathbf{z}$ of $\mathbb{R}^D$ into its orthogonal projection onto the hyperplan V. It is clear that $\mathbf{z} \equiv \operatorname{ucl}_V \mathbf{z}$, and that the operator ucl_V is constant on the vectors belonging to the same coset. It is easy to prove that

$$\operatorname{ucl}_V \mathbf{z} = \mathbf{z} - \frac{\sum_{j=1}^{D} z_j}{D} \mathbf{1}_D = \mathbf{H}_D \mathbf{z},$$

where $\mathbf{H}_D$ is the well-known *centering matrix* of order $D \times D$ (see Mardia et al., 1979). We recall that this matrix is equal to $\mathbf{I}_D - D^{-1}\mathbf{J}_D$, where $\mathbf{I}_D$ is the identity matrix of order $D \times D$, and $\mathbf{J}_D = \mathbf{1}_D \mathbf{1}_D'$.

3.2. The logarithmic transformation between the quotient spaces

The logarithmic and exponential transformations from $\mathbb{R}_+^D$ to $\mathbb{R}^D$ are compatible with the equivalence relations $\sim$ and $\equiv$ defined in $\mathbb{R}_+^D$ and $\mathbb{R}^D$, respectively. That is to say,

$$\mathbf{w} \sim \mathbf{w}^* \text{ in } \mathbb{R}_+^D \implies \log \mathbf{w} \equiv \log \mathbf{w}^* \text{ in } \mathbb{R}^D,$$

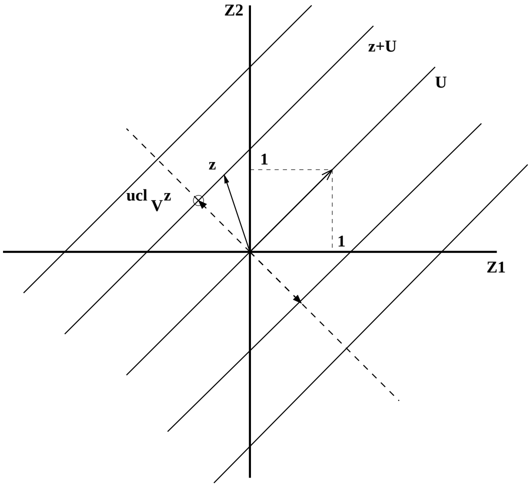


Fig. 5. Geometric interpretation of equivalence classes in $\mathcal{L}^1 = \mathbf{R}^2/U$

$$\mathbf{z} \equiv \mathbf{z}^* \text{ in } \mathbb{R}^D \quad \Longrightarrow \quad \exp \mathbf{z} \sim \exp \mathbf{z}^* \text{ in } \mathbb{R}^D_+.$$

Therefore, these transformations can be extended to the quotient spaces $\mathcal{C}^d$ and $\mathcal{L}^d$. We will symbolize by logc the transformation from $\mathcal{C}^d$ to $\mathcal{L}^d$:

$$\operatorname{logc} \underline{\mathbf{w}} = \log \mathbf{w} + U \qquad (\underline{\mathbf{w}} \in \mathcal{C}^d),$$

and by expc the inverse transformation from $\mathcal{L}^d$ to $\mathcal{C}^d$:

$$\operatorname{expc}(\mathbf{z} + U) = \operatorname{ccl}(\exp \mathbf{z}) \qquad (\mathbf{z} + U \in \mathcal{L}^d).$$

The point vector in V of the coset $\operatorname{logc} \underline{\mathbf{w}}$ will be given by

$$\operatorname{ucl}_V(\log \mathbf{w}) = \mathbf{H}_D \log \mathbf{w} = \log \frac{\mathbf{w}}{g(\mathbf{w})}.$$

Definition 9. The *centered logratio transformation* —denoted by clr— is the one-to-one function from the compositional space $\mathcal{C}^d$ to the subspace V of $\mathbb{R}^D$, defined by

$$\operatorname{clr} \underline{\mathbf{w}} = \log \frac{\mathbf{w}}{g(\mathbf{w})} \qquad (\underline{\mathbf{w}} \in \mathcal{C}^d).$$

The inverse transformation, from V to $\mathcal{C}^d$, is given by

$$\operatorname{clr}^{-1} \mathbf{z} = \operatorname{ccl}(\exp \mathbf{z}) \quad (\mathbf{z} \in V).$$

□

It is interesting to note that the logarithmic and the exponential transformations establish a one-to-one correspondence between the hyperbolic surface Hip_D in $\mathbb{R}^D_+$ and the hyperplan V in $\mathbb{R}^D$.

4. The compositional space as a real vector space

As U is a vector subspace of dimension one of the D-dimensional space $\mathbb{R}^D$, the quotient space $\mathcal{L}^d = \mathbb{R}^D/U$ can be structured as a real vector space of dimension $d = D-1$. To do so, we define the sum of two cosets $\mathbf{z}+U$ and $\mathbf{z}^*+U$ as

$$(\mathbf{z}+U)+(\mathbf{z}^*+U) = (\mathbf{z}+\mathbf{z}^*)+U,$$

and the product of a coset $\mathbf{z}+U$ by a constant $\lambda \in \mathbb{R}$ by

$$\lambda(\mathbf{z}+U) = \lambda\,\mathbf{z}+U.$$

The coset $\mathbf{1}_D+U$ is the neutral element and the opposite of $\mathbf{z}+U$ is the coset $(-\mathbf{z})+U$.

The one-to-one correspondence between $\mathcal{C}^d$ and $\mathcal{L}^d$ allows to define in $\mathcal{C}^d$ a real vector space isomorphic to $\mathcal{L}^d$.

Definition 10. In correspondence with the sum in $\mathcal{L}^d$, the inner operation $\otimes$ in $\mathcal{C}^d$ is defined as

$$\underline{\mathbf{w}} \otimes \underline{\mathbf{w}^*} = \text{expc}\ (\text{logc}\,\underline{\mathbf{w}} + \text{logc}\,\underline{\mathbf{w}^*}) = \text{ccl}\,(w_1 w_1^*, \ldots, w_D w_D^*)' \qquad (\underline{\mathbf{w}}, \underline{\mathbf{w}^*} \in \mathcal{C}^d).$$

Equally, in correspondence with the product by a constant in $\mathcal{L}^d$, the external operation $\odot$ in $\mathcal{C}^d$ is defined as

$$\lambda \odot \underline{\mathbf{w}} = \text{expc}\ (\lambda\,\text{logc}\,\underline{\mathbf{w}}) = \text{ccl}\,(w_1^\lambda, \ldots, w_D^\lambda)' \qquad (\underline{\mathbf{w}} \in \mathcal{C}^d)\ (\lambda \in \mathbb{R}).$$

□

Therefore, $(\mathcal{C}^d, \otimes, \odot)$ becomes a real vector space, isomorphic to the quotient space $\mathcal{L}^d$. In the conmutative group $(\mathcal{C}^d, \otimes)$, the composition $\underline{\mathbf{1}_D} = \text{ccl}\,(1, \ldots, 1)'$ is the neutral element, and the inverse composition $\underline{\mathbf{w}^{-1}}$ of $\underline{\mathbf{w}} = \text{ccl}\,(w_1, \ldots, w_D)'$ is the composition $\underline{\mathbf{w}^{-1}} = \text{ccl}\,(1/w_1, \ldots, 1/w_D)'$.

Provided that $(\mathcal{C}^d, \otimes, \odot)$ is a real vector space, it can be viewed as an affine space when the group $(\mathcal{C}^d, \otimes)$ operates on $\mathcal{C}^d$ as a group of transformations.

Definition 11. Given a composition $\mathbf{p} \in \mathcal{C}^d$, the *perturbation* associated to $\mathbf{p}$ is the transformation from $\mathcal{C}^d$ to $\mathcal{C}^d$ defined by

$$\mathbf{c} \to \mathbf{p} \otimes \mathbf{c} \qquad (\mathbf{c} \in \mathcal{C}^d).$$

Then, we say that $\mathbf{p} \otimes \mathbf{c}$ is the composition which results when the *perturbation* $\mathbf{p}$ is applied to the composition $\mathbf{c}$. □

Perturbations in the compositional space plays the same role as translations plays in the real space. Like it, the set of all perturbations in $\mathcal{C}^d$ is a commutative group isomorphic to $(\mathcal{C}^d, \otimes)$. Thus, the composition of two perturbations $\mathbf{p}_1$ and $\mathbf{p}_2$ is the perturbation associated to $\mathbf{p}_1 \otimes \mathbf{p}_2$. Furthermore, the perturbation associated to $\underline{\mathbf{1}_D}$ is the identity perturbation which does not produce any change when applies to a composition. Also, given any perturbation $\mathbf{p}$ there exists the inverse perturbation $\mathbf{p}^{-1}$ which undoes the changes produced by $\mathbf{p}$. Finally, given two compositions

$$\underline{\mathbf{w}} = \mathrm{ccl}\,(w_1, \ldots, w_D)' \text{ and } \underline{\mathbf{w}^*} = \mathrm{ccl}\,(w_1^*, \ldots, w_D^*)' \in \mathcal{C}^d,$$

there exists a unique perturbation $\mathbf{p}$ which transforms $\underline{\mathbf{w}}$ on $\underline{\mathbf{w}^*}$. This perturbation is no other than

$$\mathbf{p} = \underline{\mathbf{w}^*} \otimes \underline{\mathbf{w}}^{-1} = \mathrm{ccl}\,(\frac{w_1^*}{w_1}, \ldots, \frac{w_D^*}{w_D})'.$$

The assumption that the group of perturbations is the operating group on the compositional space is the keystone of the methodology introduced by Aitchison (1986). In fact, it means accepting that the "difference" between two compositions $\underline{\mathbf{w}} = \mathrm{ccl}\,(w_1, \ldots, w_D)'$ and $\underline{\mathbf{w}^*} = \mathrm{ccl}\,(w_1^*, \ldots, w_D^*)'$ is based on the ratios w_j^*/w_j between parts instead of on the differences $w_j^* - w_j$.

5. The compositional space as an Euclidean space

5.1. $\mathcal{L}^d$ as an Euclidean space

Given that the elements of $\mathcal{L}^d$ can be interpreted as straight lines parallel to vector $\mathbf{1}_D$, it seems "natural" to define the distance between two cosets $\mathbf{z} + U$ and $\mathbf{z}^* + U$ of $\mathcal{L}^d$ as the Euclidean distance between these two straight lines in $\mathrm{I\!R}^D$. This distance will be equal to the length of the difference vector $\mathrm{ucl}_V\,\mathbf{z}^* - \mathrm{ucl}_V\,\mathbf{z}$, where $\mathrm{ucl}_V\,\mathbf{z}$ and $\mathrm{ucl}_V\,\mathbf{z}^*$ are the intersection points of these straight lines with the orthogonal hyperplan V (see Fig. 6).

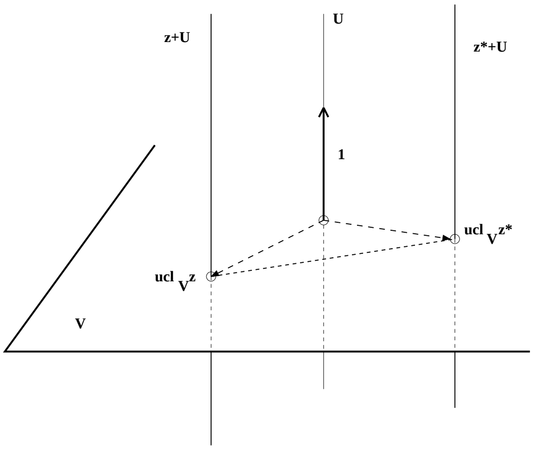


Fig. 6. Distance between two cosets $\mathbf{z} + U$ and $\mathbf{z}^* + U$ of $\mathcal{L}^2$.

Therefore, it is easy to translate to $\mathcal{L}^d$ the Euclidean structure on $V \subset \mathrm{I\!R}^D$.

DEFINITION 12. Given $\mathbf{z} + U \in \mathcal{L}^d$ and $\mathbf{z}^* + U \in \mathcal{L}^d$, we define the $\mathcal{L}$-inner product $<\mathbf{z} + U, \mathbf{z}^* + U>_{\mathcal{L}}$ as the usual inner product $<\mathrm{ucl}_V\,\mathbf{z}, \mathrm{ucl}_V\,\mathbf{z}^*>$ in $\mathrm{I\!R}^D$. □

It is easy to prove that

$$<\mathbf{z}+U, \mathbf{z}^*+U>_{\mathcal{L}}= \sum_{j=1}^{D} z_j z_j^* - \frac{1}{D}\left(\sum_{j=1}^{D} z_j\right)\left(\sum_{j=1}^{D} z_j^*\right) = \mathbf{z}'\mathbf{H}_D\mathbf{z}^*,$$

for each $\mathbf{z}+U, \mathbf{z}^*+U \in \mathcal{L}^d$.

It is possible to define a norm and a distance in $\mathcal{L}^d$ from the $\mathcal{L}$-inner product. The $\mathcal{L}$-norm of a coset $\mathbf{z}+U \in \mathcal{L}^d$ is given by

$$\|\mathbf{z}+U\|_{\mathcal{L}} = (<\mathbf{z}+U, \mathbf{z}+U>_{\mathcal{L}})^{1/2} = \left[\sum_{j=1}^{D} z_j^2 - \frac{1}{D}\left(\sum_{j=1}^{D} z_j\right)^2\right]^{1/2} = (\mathbf{z}'\mathbf{H}_D\mathbf{z})^{1/2},$$

and it holds that $\|\mathbf{z}+U\|_{\mathcal{L}} = \|\mathrm{ucl}_V\mathbf{z}\|$.

Similarly, the $\mathcal{L}$-distance between two cosets $\mathbf{z}+U$ and $\mathbf{z}^*+U$ in $\mathcal{L}^d$ is given by

$$d_{\mathcal{L}}(\mathbf{z}+U, \mathbf{z}^*+U) = \|(\mathbf{z}^*+U)-(\mathbf{z}+U)\|_{\mathcal{L}} = \left[\sum_{j=1}^{D}(z_j^*-z_j)^2 - \frac{1}{D}\left(\sum_{j=1}^{D}(z_j^*-z_j)\right)^2\right]^{1/2}.$$

This expression can be written in matrix form as

$$d_{\mathcal{L}}(\mathbf{z}+U, \mathbf{z}^*+U) = [(\mathbf{z}-\mathbf{z}^*)'\mathbf{H}_D(\mathbf{z}-\mathbf{z}^*)]^{1/2},$$

and it holds that $d_{\mathcal{L}}(\mathbf{z}+U, \mathbf{z}^*+U) = d(\mathrm{ucl}_V\mathbf{z}, \mathrm{ucl}_V\mathbf{z}^*)$.

In this manner, the quotient space $\mathcal{L}^d$ becomes an Euclidean space.

5.2. $\mathcal{C}^d$ as an Euclidean space

The one-to-one transformations logc and expc between $\mathcal{C}^d$ and $\mathcal{L}^d$ allow to transfer to $\mathcal{C}^d$ the real Euclidean structure defined on $\mathcal{L}^d$.

DEFINITION 13. We define the *compositional inner product* of two compositions $\underline{\mathbf{w}}$ and $\underline{\mathbf{w}^*}$ as

$$<\underline{\mathbf{w}}, \underline{\mathbf{w}^*}>_{\mathcal{C}}=<\log\mathbf{w}+U, \log\mathbf{w}^*+U>_{\mathcal{L}}.$$

□

It is easy to prove that

$$<\underline{\mathbf{w}}, \underline{\mathbf{w}^*}>_{\mathcal{C}}= \sum_{j=1}^{D}\log w_j \log w_j^* - \frac{1}{D}\left(\sum_{j=1}^{D}\log w_j\right)\left(\sum_{j=1}^{D}\log w_j^*\right) = (\log\mathbf{w})'\mathbf{H}_D\log\mathbf{w}^*,$$

and

$$<\underline{\mathbf{w}},\underline{\mathbf{w}^*}>_{\mathcal{C}}=\sum_{j=1}^{D}\log\frac{w_j}{g(\mathbf{w})}\log\frac{w_j^*}{g(\mathbf{w}^*)}=<\operatorname{clr}\underline{\mathbf{w}},\operatorname{clr}\underline{\mathbf{w}^*}>,$$

for each $\underline{\mathbf{w}},\underline{\mathbf{w}^*}\in\mathcal{C}^d$.

Thus, the $\mathcal{C}$-inner product of $\underline{\mathbf{w}}$ and $\underline{\mathbf{w}^*}$ in $\mathcal{C}^d$ is equal to the ordinary inner product of $\operatorname{clr}\underline{\mathbf{w}}$ and $\operatorname{clr}\underline{\mathbf{w}^*}$ in $\mathbb{R}^D$.

As usual, two compositions $\underline{\mathbf{w}}$, and $\underline{\mathbf{w}^*}$ are said to be *$\mathcal{C}$-orthogonal* if and only if $<\underline{\mathbf{w}},\underline{\mathbf{w}^*}>_{\mathcal{C}}=0$.

From this scalar product in $\mathcal{C}^d$ we can define a norm and a distance in the compositional space. The *compositional norm* of a composition $\underline{\mathbf{w}}\in\mathcal{C}^d$ will be given by

$$\|\underline{\mathbf{w}}\|_{\mathcal{C}}=(<\underline{\mathbf{w}},\underline{\mathbf{w}}>_{\mathcal{C}})^{1/2}=\left[\sum_{j=1}^{D}(\log w_j)^2-\frac{1}{D}\left(\sum_{j=1}^{D}\log w_j\right)^2\right]^{1/2}=\left[(\log\mathbf{w})'\mathbf{H}_D\log\mathbf{w}\right]^{1/2}.$$

The $\mathcal{C}$-norm of a composition of $\mathcal{C}^d$ is equal to the Euclidean norm in $\mathbb{R}^D$ of the clr-transformed vector:

$$\|\underline{\mathbf{w}}\|_{\mathcal{C}}=\|\operatorname{clr}\underline{\mathbf{w}}\|\qquad(\underline{\mathbf{w}}\in\mathcal{C}^d).$$

Another expression of the $\mathcal{C}$-norm of a composition is given by

$$\|\underline{\mathbf{w}}\|_{\mathcal{C}}^2=\frac{1}{D}\sum_{1\le i<j\le D}\left(\log\frac{w_i}{w_j}\right)^2\qquad(\underline{\mathbf{w}}\in\mathcal{C}^d).$$

A composition $\underline{\mathbf{w}}$ is said to be *$\mathcal{C}$-unitary* if and only if $\|\underline{\mathbf{w}}\|_{\mathcal{C}}=1$.

The *compositional distance* between two compositions $\underline{\mathbf{w}}$ and $\underline{\mathbf{w}^*}$ of $\mathcal{C}^d$ is given by the $\mathcal{C}$-norm of the composition $\underline{\mathbf{w}^*}\otimes\underline{\mathbf{w}}^{-1}=\operatorname{ccl}\,(w_1^*/w_1,\ldots,w_D^*/w_D)'$, i.e.,

$$d_{\mathcal{C}}(\underline{\mathbf{w}},\underline{\mathbf{w}^*})=\left[\sum_{j=1}^{D}\left(\log\frac{w_j^*}{w_j}\right)^2-\frac{1}{D}\left(\sum_{j=1}^{D}\log\frac{w_j^*}{w_j}\right)^2\right]^{1/2}.$$

This distance can be expressed in matrix form by

$$d_{\mathcal{C}}(\underline{\mathbf{w}},\underline{\mathbf{w}^*})=\left[(\log\mathbf{w}^*-\log\mathbf{w})'\mathbf{H}_D(\log\mathbf{w}^*-\log\mathbf{w})\right]^{1/2}.$$

Therefore, the $\mathcal{C}$-distance between two compositions in $\mathcal{C}^d$ is equal to the Euclidean distance in $\mathbb{R}^D$ between the corresponding clr-transformed vectors:

$$d_{\mathcal{C}}(\underline{\mathbf{w}},\underline{\mathbf{w}^*})=d(\operatorname{clr}\underline{\mathbf{w}},\operatorname{clr}\underline{\mathbf{w}^*})\qquad(\underline{\mathbf{w}},\underline{\mathbf{w}^*}\in\mathcal{C}^d).$$

Thus, the $\mathcal{C}$-distance just defined converts the compositional space $\mathcal{C}^d$ in an Euclidean space which is isometric to the Euclidean space $\mathcal{L}^d$.

Moreover, the centered logratio transformation clr is the natural isometry between $\mathcal{C}^d$ and the subspace V of $\mathbb{R}^D$.

This distance in $\mathcal{C}^d$ will accomplish the usual properties of all Euclidean distances. In particular, it is related to the operations in the compositional space by the following identities:

$$d_{\mathcal{C}}(\mathbf{a},\mathbf{b}) = d_{\mathcal{C}}(\mathbf{a}\otimes\mathbf{c},\mathbf{b}\otimes\mathbf{c}) \quad (\mathbf{a},\mathbf{b},\mathbf{c}\in\mathcal{C}^d),$$

and

$$d_{\mathcal{C}}(\lambda\odot\mathbf{a},\lambda\odot\mathbf{b}) = |\lambda| d_{\mathcal{C}}(\mathbf{a},\mathbf{b}) \quad (\mathbf{a},\mathbf{b}\in\mathcal{C}^d)\ (\lambda\in\mathbb{R}).$$

Another important property of this compositional distance holds in relation to subcompositions. It is based on the fact that the mapping sub_S introduced in Definition 6, which sends a composition $\underline{\mathbf{w}}\in\mathcal{C}^d$ to a subcomposition $\underline{\mathbf{w}_S}\in\mathcal{C}^c$, is a linear application between the real vector spaces $(\mathcal{C}^d,\otimes,\odot)$ and $(\mathcal{C}^c,\otimes,\odot)$.

PROPERTY 3. The compositional distance is *subcompositionally dominant*:

$$d_{\mathcal{C}}(\underline{\mathbf{w}},\underline{\mathbf{w}^*}) \geq d_{\mathcal{C}}(\underline{\mathbf{w}_S},\underline{\mathbf{w}_S^*}) \quad (\underline{\mathbf{w}},\underline{\mathbf{w}^*}\in\mathcal{C}^d).$$

Or, equivalently,

$$||\underline{\mathbf{w}}||_{\mathcal{C}} \ \geq\ ||\underline{\mathbf{w}_S}||_{\mathcal{C}} \quad (\underline{\mathbf{w}}\in\mathcal{C}^d).$$

□

This property means that the $\mathcal{C}$-distance between two subcompositions can never be greater than the $\mathcal{C}$-distance between the corresponding compositions. This is a reasonable property to be expected from any distance defined over the compositional space.

PROPERTY 4. The mapping inc_S introduced in Definition 7 is a linear application between the real vector spaces $(\mathcal{C}^c,\otimes,\odot)$ and $(\mathcal{C}^d,\otimes,\odot)$. This mapping preserves the $\mathcal{C}$-distance because

$$||\mathrm{ccl}\,(w_1,\ldots,w_C)'||_{\mathcal{C}} = \left\|\mathrm{ccl}\left(1,\overset{D-C}{\ldots},1,\frac{w_1}{g(\mathbf{w})},\ldots,\frac{w_C}{g(\mathbf{w})}\right)'\right\|_{\mathcal{C}} \quad (\underline{\mathbf{w}}\in\mathcal{C}^c).$$

□

6. Bases in the compositional space

6.1. Natural bases in $\mathcal{L}^d$

Let $\mathbf{e}_1=(1,0,\ldots,0)',\mathbf{e}_2=(0,1,0,\ldots,0)',\ldots,\mathbf{e}_D=(0,\ldots,0,1)'$ be the canonical basis of $\mathbb{R}^D$. Let $\mathcal{B}^{\mathcal{L}}$ be the ordered set $\{\mathbf{e}_1+U,\ldots,\mathbf{e}_D+U\}$ of cosets of $\mathcal{L}^d$. For each $j=1,\ldots,D$, we symbolize by $\mathcal{B}^{\mathcal{L}}_{-j}$ the ordered set $\mathcal{B}^{\mathcal{L}}-\{\mathbf{e}_j+U\}$. Clearly, the set $\mathcal{B}^{\mathcal{L}}_{-j}$ is a basis of the vector space $\mathcal{L}^d$, for each $j=1,\ldots,D$. In particular, if $\mathbf{z}=(z_1,\ldots,z_D)'\in\mathbb{R}^D$, the vector $\mathbf{y}$ whose components are those of $\mathbf{z}+U$ in the basis $\mathcal{B}^{\mathcal{L}}_{-D}$ is

$$\mathbf{y}=(z_1-z_D,\ldots,z_d-z_D)'=\mathbf{F}\mathbf{z},$$

where $\mathbf{F}$ is the $d \times D$ matrix $[\mathbf{I}_d : -\mathbf{1}_d]$. For this basis it hold that

$$\|\mathbf{e}_i + U\|^2_{\mathcal{L}} = \frac{D-1}{D} \quad (i = 1, \ldots, D),$$

and

$$< \mathbf{e}_i + U, \mathbf{e}_j + U >_{\mathcal{L}} = -\frac{1}{D} \quad (i, j = 1, \ldots, D;\ i \neq j).$$

Thus, none of the basis $\mathcal{B}^{\mathcal{L}}_{-j}$ of $\mathcal{L}^d$ is $\mathcal{L}$-orthonormal. The $d \times d$ matrix which expresses the $\mathcal{L}$-metric in the basis $\mathcal{B}^{\mathcal{L}}_{-j}$ of $\mathcal{L}^d$ is equal to $\mathbf{M} = \mathbf{I}_d - D^{-1}\mathbf{J}_d$, independently of the index $j = 1, \ldots, D$. This matrix can be expressed as a function of $\mathbf{F}$ because the following relationship holds:

$$\mathbf{M} = (\mathbf{F}\mathbf{F}')^{-1}. \tag{3}$$

6.2. Orthonormal bases in $\mathcal{L}^d$

The subspace $V = \{\mathbf{z} \in \mathrm{I\!R}^D : \mathbf{z}'\mathbf{1}_D = 0\}$ of $\mathrm{I\!R}^D$ has dimension $d = D - 1$. Let $\mathbf{v}_1 = (v_{11}, \ldots, v_{1D})', \ldots, \mathbf{v}_d = (v_{d1}, \ldots, v_{dD})'$ be an orthonormal basis of V, and let $\mathbf{V}$ be the $D \times d$ matrix $[\mathbf{v}_1 : \ldots : \mathbf{v}_d]$. It is straightforward to prove that this matrix verifies the following two identities:

$$\text{(i)}\ \ \mathbf{V}'\mathbf{V} = \mathbf{I}_d; \qquad \text{and} \qquad \text{(ii)}\ \ \mathbf{V}\mathbf{V}' = \mathbf{H}_D. \tag{4}$$

Inversely, if $\mathbf{V}$ is a $D \times d$ matrix verifying the two identities (4), their vector-columns constitute an orthonormal basis of the subspace V.

Then, the ordered set $\mathcal{V}^{\mathcal{L}} = \{\mathbf{v}_1 + U, \ldots, \mathbf{v}_d + U\}$ is an $\mathcal{L}$-orthonormal basis of $\mathcal{L}^d$. If $\mathbf{z} \in \mathrm{I\!R}^D$, the vector $\mathbf{u}$ of $\mathrm{I\!R}^d$ whose components are those of the class $\mathbf{z} + U$ in the basis $\mathcal{V}^{\mathcal{L}}$ is

$$\mathbf{u} = (\mathbf{F}\mathbf{V})^{-1}\mathbf{F}\mathbf{z}.$$

6.3. Natural bases in $\mathcal{C}^d$

From cosets $\mathbf{e}_1 + U, \ldots, \mathbf{e}_D + U$ of $\mathcal{L}^d$, we define the corresponding compositions in $\mathcal{C}^d$:

$$\tilde{\mathbf{e}}_1 = \operatorname{expc}(\mathbf{e}_1 + U) = \operatorname{ccl}(e, 1, \ldots, 1, 1)', \ldots, \tilde{\mathbf{e}}_D = \operatorname{expc}(\mathbf{e}_D + U) = \operatorname{ccl}(1, 1, \ldots, 1, e)'.$$

DEFINITION 14. If $\mathcal{B}$ symbolizes the ordered set $\{\tilde{\mathbf{e}}_1, \ldots, \tilde{\mathbf{e}}_D\}$, the set $\mathcal{B}_{-j} = \mathcal{B} - \{\tilde{\mathbf{e}}_j\}$ is a basis of the real vector space $(\mathcal{C}^d, \otimes, \odot)$, for any index $j = 1, \ldots, D$. These bases are called *natural basis* of $\mathcal{C}^d$. □

Then, if $\mathbf{w} = (w_1, \ldots, w_D)' \in \mathrm{I\!R}^D$, the vector $\mathbf{y}$ of $\mathrm{I\!R}^d$ whose components are those of the composition $\underline{\mathbf{w}\ \ }$ in the basis $\mathcal{B}_{-D}$ is equal to

$$\mathbf{y} = (\log \frac{w_1}{w_D}, \ldots, \log \frac{w_d}{w_D})'.$$

In general, if $\mathbf{w}_{-j}$ symbolizes the vector $\mathbf{w}$ without the component w_j, the components of $\underline{\mathbf{w}}$ in the basis $\mathcal{B}_{-j}$ are those of the vector $\log(\mathbf{w}_{-j}/w_j)$.

Definition 15. The *additive logratio transformation* of index j $(j = 1, \dots, D)$ — denoted by alr_j — is the one-to-one transformation from $\mathcal{C}^d$ to $\mathbb{R}^d$ which assigns to each composition $\underline{\mathbf{w}}$ its components in the basis $\mathcal{B}_{-j}$:

$$\underline{\mathbf{w}} \longrightarrow \mathrm{alr}_j\, \underline{\mathbf{w}} = \log \frac{\mathbf{w}_{-j}}{w_j}.$$

□

The inverse transformation of alr_j , from $\mathbb{R}^d$ to $\mathcal{C}^d$, is given by

$$\mathrm{alr}_j^{-1}\,\mathbf{y} = \mathrm{ccl}\,(\exp y_1, \dots, \exp y_{j-1}, 1, \exp y_j, \dots, \exp y_d)' \quad (\mathbf{y} \in \mathbb{R}^d).$$

In particular, when $j = D$, these transformations can be easily expressed in matrix form as

$$\mathrm{alr}_D\, \underline{\mathbf{w}} = \mathbf{F} \log \mathbf{w}, \qquad \text{and} \qquad \mathrm{alr}_D^{-1}\,\mathbf{y} = \mathrm{ccl}\left\{\exp\left[\mathbf{F}'(\mathbf{F}\mathbf{F}')^{-1}\mathbf{y}\right]\right\}.$$

6.4. Orthonormal bases in $\mathcal{C}^d$

As happens with the basis $\mathcal{B}_{-j}^{\mathcal{L}}$ of $\mathcal{L}^d$, none of the bases $\mathcal{B}_{-j}$ of $\mathcal{C}^d$ is $\mathcal{C}$-orthonormal. The matrix $\mathbf{M}$ introduced in (3) is the matrix which determines the $\mathcal{C}$-metric in these bases. This matrix $\mathbf{M}$ corresponds to matrix $\mathbf{H}^{-1}$ defined in Aitchison (1986, p. 343).

Definition 16. From any $D \times d$ matrix $\mathbf{V} = [\mathbf{v}_1 : \dots : \mathbf{v}_d]$ that verifies the two identities (4), we can define the compositions

$$\tilde{\mathbf{v}}_1 = \mathrm{expc}\,(\mathbf{v}_1 + U), \dots, \tilde{\mathbf{v}}_d = \mathrm{expc}\,(\mathbf{v}_d + U).$$

Then, the ordered set $\mathcal{V} = \{\tilde{\mathbf{v}}_1, \dots, \tilde{\mathbf{v}}_d\}$ is a $\mathcal{C}$-orthonormal basis of $\mathcal{C}^d$. □

Consequently, if $\mathbf{w}$ is an observational vector of $\mathbb{R}_+^D$, the vector $\mathbf{u}$ of $\mathbb{R}^d$ whose components are those of the composition $\underline{\mathbf{w}}$ in the basis $\mathcal{V}$ is equal to

$$\mathbf{u} = (\mathbf{F}\mathbf{V})^{-1}\mathbf{F} \log \mathbf{w}.$$

Definition 17. Given a $D \times d$ matrix $\mathbf{V} = [\mathbf{v}_1 : \dots : \mathbf{v}_d]$ that verifies conditions (4), the *isometric logratio transformation* —denoted by ilr_V — associated to this matrix $\mathbf{V}$, is the one-to-one transformation from $\mathcal{C}^d$ to $\mathbb{R}^d$ which assigns to each composition $\underline{\mathbf{w}}$ its components in the basis $\mathcal{V}$ just defined:

$$\underline{\mathbf{w}} \longrightarrow \mathrm{ilr}_\mathrm{V}\, \underline{\mathbf{w}} = (\mathbf{F}\mathbf{V})^{-1}\mathbf{F} \log \mathbf{w}.$$

□

The inverse transformation of ilr_V, from $\mathbb{R}^d$ to $\mathcal{C}^d$, is given by

$$\text{ilr}_{\text{V}}^{-1}\, \mathbf{x} = \text{ccl}\left(\exp\left\{\left[(\mathbf{F}\mathbf{V})^{-1}\mathbf{F}\right]'\mathbf{x}\right\}\right) \quad (\mathbf{x} \in \mathbb{R}^d).$$

Note that, by construction, the transformation ilr_{V} is an isometry between the metric spaces $\mathcal{C}^d$ and $\mathbb{R}^d$., thus justifying the term *isometric logratio transformation*. This term was first used by J.J. Egozcue (personal communication).

PROPERTY 5. If $\mathbf{V} = [\mathbf{v}_1 : \ldots : \mathbf{v}_d]$ and $\mathbf{V}^* = [\mathbf{v}_1^* : \ldots : \mathbf{v}_d^*]$ are two $D \times d$ matrices that verify conditions (4), then the isometric logratio transformations ilr_{V} and ilr_{V^*} associated to $\mathbf{V}$ and $\mathbf{V}^*$, respectively, are related by the following identity:

$$\text{ilr}_{\text{V}}\, \underline{\mathbf{w}} = \mathbf{V}'\mathbf{V}^*\text{ilr}_{V^*}\underline{\mathbf{w}} \quad (\underline{\mathbf{w}} \in \mathcal{C}^d).$$

□

6.5. Determination of a composition

As consequence of the above results, a composition $\underline{\mathbf{w}} \in \mathcal{C}^d$ can be expressed in several ways:

(i) Giving any D-observational vector belonging to $\underline{\mathbf{w}}$.

(ii) Giving the components $(y_1, \ldots, y_d) = \mathbf{y}'$ of $\underline{\mathbf{w}}$ in the basis $\mathcal{B}_{-D}$ of $\mathcal{C}^d$. If it is necessary, we can choose the components of any other logratio $\text{alr}_j\underline{\mathbf{w}}$ $(j \neq D)$.

(iii) Giving the components $(z_1, \ldots, z_D)' = \mathbf{z}$ of the centered logratio transformed vector $\text{clr}\, \underline{\mathbf{w}}$. Since $\mathbf{z}$ belongs to subspace V of $\mathbb{R}^D$, its components are related by the equality $z_1 + \ldots + z_D = 0$.

(iv) Or, finally, giving the components $(u_1, \ldots, u_d)' = \mathbf{u}$ of $\underline{\mathbf{w}}$ in an orthonormal basis $\mathcal{V}$ of $\mathcal{C}^d$. In this case, it is also necessary to know the matrix $\mathbf{V}$ which individualizes the basis $\mathcal{V}$.

Obviously, option (i) is the best in order to determine a composition $\underline{\mathbf{w}}$ because the components $w_1, \ldots, w_D$ of vector $\mathbf{w}$ are directly interpretable. Usually, we will choose the observational vector $\text{ccl}_L\, \mathbf{w}$ belonging to simplex $\mathcal{S}^d$ because, in this case, their components express parts of a total. In option (ii), the components y_j are also interpretable because they represent logratios —$y_j = \log(w_j/w_D)$ $(j = 1, \ldots, d)$—, and it is very easy to compute from them any other logratios:

$$\log\frac{w_i}{w_j} = y_i - y_j \ (i, j = 1, \ldots, d), \qquad \text{and} \qquad \log\frac{w_D}{w_j} = -y_j \ (j = 1, \ldots, d)$$

It is difficult to give a direct interpretation of the centred logratio components $z_j = \log[w_j/g(\mathbf{w})]$ $(j = 1, \ldots, D)$ because of the presence of the geometric mean $g(\mathbf{w})$ in the denominator of these logratios. The component z_j gives, in logarithmic scale, information about the value of part j with respect to overall value of the other parts. However it results very easy to calculate from the centred logratio components any logratio because $\log(w_i/w_j) = z_i - z_j$ $(i, j = 1 \ldots D)$.

Finally, the components of the vector $\mathbf{u}$ in option (iv) are not directly interpretable because they depend of an arbitrary matrix $\mathbf{V}$ satisfying conditions (4). However, this representation is very useful when we need to analyze the metric relations between a set of compositions because the relationship between components of a composition in the basis $\mathcal{V}$ are Euclidean, if we consider in $\mathcal{C}^d$ the metric structure associated to the $\mathcal{C}$-distance previously defined.

Property 6. Vectors $\mathbf{u}$, $\mathbf{y}$ and $\mathbf{z}$ associated to the same composition $\underline{\mathbf{w}}$ are related by the following relationships:

(a) $\mathbf{u} = (\mathbf{FV})^{-1}\mathbf{y}$, and $\mathbf{u} = (\mathbf{FV})^{-1}\mathbf{Fz}$.
(b) $\mathbf{y} = \mathbf{FVu}$, and $\mathbf{y} = \mathbf{Fz}$.
(c) $\mathbf{z} = \left[(\mathbf{FV})^{-1}\mathbf{F}\right]'\mathbf{u}$, and $\mathbf{z} = \mathbf{F}'(\mathbf{FF}')^{-1}\mathbf{y}$.

□

7. Conclusion

The methodology developed by Aitchison (1986) to perform the statistical analysis of compositional data is essentially based on the concept of perturbation and on the centered and additive logratio transformations. We have shown that these concepts and transformations are not arbitrary. In fact, from a mathematical point of view, they are induced by the nature of the compositional data provided that these kind of data is characterized by their scale-invariant property. Therefore, the methodology proposed by Aitchison (1986) cannot be mathematically refused, as it is fully compatible with the nature of compositional data and is independent of the representation used to manage the data. Moreover, the analysis of subcompositions is also coherent with the analysis of the full composition.

The detractors of this methodology who advocate the "standard" analysis of this kind of data are implicitly refusing the *ratio* as the natural form to compare two compositions or two parts of the same composition. They are also implicitly accepting the usual *difference* as the logical manner to perform these comparisons. This kind of analysis is completely depending of the representation used to manage the data, and many of the results can be misleading, as pointed out by Pearson (1897) more than one hundred years ago.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Chapman & Hall Ltd., London (UK), 416 p.

Aitchison, J. (1989). Measures of location of compositional data sets. *Mathematical Geology 21*(7), 787–790.

Aitchison, J. (1990a). Comment on "Measures of variability for geological data" by D. F. Watson and G. M. Philip. *Mathematical Geology 22*(2), 223–226.

Aitchison, J. (1990b). Relative variation diagrams for describing patterns of compositional variability. *Mathematical Geology 22*(4), 487–511.

Aitchison, J. (1991). Delusions of uniqueness and ineluctability. *Mathematical Geology 23*(2), 275–277.

Aitchison, J. (1992). On criteria for measures of compositional difference. *Mathematical Geology 24*(4), 365–379.

Aitchison, J. (1997). The one-hour course in compositional data analysis or compositional data analysis is simple. In: Pawlowsky-Glahn, Vera, Ed., *Proceedings of IAMG'97 — The Third Annual Conference of the International Association for Mathematical Geology.* International Center for Numerical Methods in Engineering (CIMNE), Barcelona (E), 3–35.

Aitchison, J. (1999). Logratios and natural laws in compositional data analysis. *Mathematical Geology 131*(5), 563–580.

Aitchison, J. (2001). Simplicial inference. In: M. Viana, and D. Richards, Eds., *Algebraic Methods in Statistics.* Contemporary Series of the American Mathematic Society, (in press).

Aitchison, J. and J. Bacon-Shone (1999). Convex linear combination of compositions. *Biometrika 86*(2), 351–364.

Aitchison, J., C. Barceló-Vidal, and V. Pawlowsky-Glahn (2001). Some comments on compositional data analysis in Archaeometry, in particular the fallacies in Tangri and Wright's dismissal of logratio analysis. *Archaeometry*, (submitted).

Aitchison, J. and M. Greenacre (2001). Biplots compositional data. *Applied Statistics*, (submitted).

Aitchison, J. and C. W. Thomas (1998). Differential perturbation processes: a tool for the study of compositional processes. In: A. Buccianti, G. Nardi, and R. Potenza, Eds., *Proceedings of IAMG'98, The Fourth Annual Conference of the International Association for Mathematical Geology,* De Frede, Naples (I), 499–504.

Barceló-Vidal, C. (2000). Fundamentación matemática del análisis de datos composicionales. *Technical Report IMA 00-02-RR.*

Barceló-Vidal, C. (1996). Mixturas de datos composicionales. *Ph. D., Universitat Politècnica de Catalunya,* Barcelona (E), 261 p.

Barceló-Vidal, C., J. A. Martín-Fernández, and V. Pawlowsky-Glahn (1999). Comment on "Singularity and nonnormality in the classification of compositional data". *Mathematical Geology 31*(5), 581–585.

Barceló-Vidal, C. and V. Pawlowsky-Glahn (1994). Finite mixtures of compositional data. *Science de la Terre, Ser. Inf. 32*, 29–48.

Barceló-Vidal, C., V. Pawlowsky-Glahn, and E. Grunsky (1995). Classification problems of samples of finite mixtures of compositions. *Mathematical Geology 27*(1), 129–148.

Barceló-Vidal, C., V. Pawlowsky-Glahn, and E. Grunsky (1996). Some aspects of transformations of compositional data and the identification of outliers. *Mathematical Geology 28*(4), 501–518.

Baxter, M. (1993). Comment on D. Tangri and R. V. S. Wright "Multivariate Analysis of Compositional Data ...". Archaeometry 35(1)(1993). *Archaeometry 35*(1), 112–115.

Bohling, G. C., J. C. Davis, R. A. Olea, and J. Harff (1996). Singularity and nonnormality in the classification of compositional data. *Mathematical Geology 30*(1), 5–20.

Mardia, K. V., J. T. Kent, and J. M. Bibby (1979). *Multivariate Analysis.* Academic Press, London (GB), 518 p.

Martín-Fernández, J. A. (2001). Medidas de diferencia y clasificación no paramétrica de datos composicionales. *Ph. D., Universitat Politècnica de Catalunya,* Barcelona (E), 233 p.

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (1997). Different classifications of the Darss Sill data set based on mixture models for compositional data. In: Pawlowsky-Glahn, Vera, Ed., *Proceedings of IAMG'97 — The Third Annual Conference of the International Association for Mathematical Geology.* International Center for Numerical Methods in Engineering (CIMNE), Barcelona (E), 151–156.

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

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (1998b). Medida de diferencia de Kullback-Leibler entre datos composicionales. In: *Actas del XXIV Congreso Nacional de la Sociedad de Estadística e Investigación Operativa (SEIO),* Almería (E), 291–292.

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

Martín-Fernández, J. A., M. Bren, C. Barceló-Vidal, and V. Pawlowsky-Glahn (1999). A measure of difference for compositional data based on measures of divergence. In: Lippard, S. J. and Næss, A. and Sinding-Larsen, R., Eds.*, Proceedings of IAMG'99, The Fifth Annual Conference of the International Association for Mathematical Geology,* Tapir, Trondheim (N), 211–216.

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (2000). Zero replacement in compositional data sets. In: Kiers, H.A.L:, Rasson, J.-P., Groenen, P.J.F., and Schader, M.*, Studies in Classification, Data Analysis, and Knowledge Organization (Proceedings of the 7th Conference of the International Federation of Classification Societies (IFCS'2000),* Namur (B), 155–160.

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

Pawlowsky-Glahn, V. and C. Barceló-Vidal (1999). Confidence regions in ternary diagrams. *Terra Nostra (Schriften der Alfred-Wegener-Stiftung) (99)*(1), 37–47.

Pawlowsky-Glahn, V. and J. J. Egozcue (2001a). About BLU estimators and compositional data. *Mathematical Geology*, (accepted for publication).

Pawlowsky-Glahn, V. and J. J. Egozcue (2001b). Geometric approach to statistical analysis on the simplex. *SERRA*, (in press).

Pearson, K. (1897). Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London LX*, 489–502.

Tangri, D. and R. Wright (1993). Multivariate analysis of compositional data: Applied comparison favour standard principal components analysis over Aitchison's loglinear contrast method. *Archaeometry 35*(1), 103–111.

Tauber, F. (1999). Spurious clusters in granulometric data caused by logratio transformation. *Mathematical Geology 31*(5), 491–504.

Watson, D. F. (1990). Reply to Comment on "Measures of variability for geological data" by D.F. Watson and G.M. Philip . *Mathematical Geology 22*(2), 227–231.

Watson, D. F. (1991). Reply to "Delusions of uniqueness and ineluctability" by J. Aitchison. *Mathematical Geology 23*(2), 279.

Watson, D. F. and G. M. Philip (1989). Measures of variability for geological data. *Mathematical Geology 21*(2), 233–254.

Whitten, E. H. T. (1995). Open and closed compositional data in petrology. *Mathematical Geology 27*(6), 789–806.

Woronow, A. (1997a). The elusive benefits of logrations. In: Pawlowsky-Glahn, Vera, Ed., *Proceedings of IAMG'97 — The Third Annual Conference of the International Association for Mathematical Geology*. International Center for Numerical Methods in Engineering (CIMNE), Barcelona (E), 97–101.

Woronow, A. (1997b). Regression and discrimination analysis using raw compositional data: Is it really a problem? In: Pawlowsky-Glahn, Vera, Ed., *Proceedings of IAMG'97 — The Third Annual Conference of the International Association for Mathematical Geology*. International Center for Numerical Methods in Engineering (CIMNE), Barcelona (E), 157–162.

Zier, U. and S. Rehder (1998). Grain-size analysis: A closed data set problem. In: A. Buccianti, G. Nardi, and R. Potenza, Eds., *Proceedings of IAMG'98, The Fourth Annual Conference of the International Association for Mathematical Geology,* De Frede, Naples, 555–558.