

The Mathematics of Compositional Analysis

Carles Barceló-Vidal
University of Girona, Spain

Josep-Antoni Martín-Fernández
University of Girona, Spain

Abstract

The term *compositional data analysis* is historically associated to the approach based on the logratio transformations introduced in the eighties. Two main principles of this methodology are scale invariance and subcompositional coherence. New developments and concepts emerged in the last decade revealed the need to clarify the concepts of compositions, compositional sample space and subcomposition. In this work the mathematics of compositional analysis based on equivalence relation is presented. A logarithmic isomorphism between quotient spaces induces a metric space structure for compositions. The *logratio compositional analysis* is the statistical analysis of compositions based on this structure, consisting of analysing logratio coordinates.

Keywords: composition, compositional analysis, equivalence class, logratio, quotient space, simplex.

1. Introduction

The term *compositional data* (CoDa) was first introduced by Aitchison (1982) and later developed in Aitchison (1986). In these publications CoDa is identified with vectors of strictly positive components whose sum is always equal to one; that is, vectors of the unit simplex

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

The term *compositional data analysis* (CoDA) has been implicitly associated with the methodology proposed by Aitchison (1986), which is based on applying the logratio transformations to the CoDa and describing, analysing and modelling them statistically from the logratios of their components. The main aim of this methodology is to free the CoDa from the constraints of the constant sum in order to be able to use the standard distributions in the real space to model the CoDa, e.g., the multivariate normal distribution. This strategy has two fundamental concepts in the so-called *principles of CoDA* (Aitchison 1986), namely, ‘scale invariance’ and ‘subcompositional coherence’. From Aitchison (1986), “scale invariance merely reinforces the intuitive idea that a composition provides information only about relative values not about absolute values and therefore ratios of componentes are the relevant entities to study”; and “subcompositional coherence demands that two scientist, one using full composition and the other using subcompositions of these full compositions, should make the same inference about relations within the common parts”. Later it was seen that the methodology initiated by Aitchison is more than a simple transformation of the CoDa, because it is in

fact a way to provide the simplex with a structure of Euclidean space. The interested reader can refer to [Egozcue, Barceló-Vidal, Martín-Fernández, Jarauta-Bragulat, Díaz-Barrero, and Mateu-Figueras \(2011\)](#) for further information.

The identification of the term CoDA with the methodology based on the logratio transformations developed by Aitchison has meant that other possible methods for analysing CoDa have made little impact. [Watson and Philip \(1989\)](#), [Wang, Liu, Mok, Fu, and Tse \(2007\)](#) or [Scealy and Welsh \(2011\)](#), for example, prefer to apply the techniques characteristic of directional data, given that they take the positive orthant of the unit hypersphere centred at the origin as the sample space of the CoDa. At that time this alternative method for analysing CoDa was cause for intense epistolary exchanges between D. F. Watson (and G. M. Philip) and J. Aitchison (see [Aitchison 1990](#); [Watson 1990](#); [Aitchison 1991](#); [Watson 1991](#)). Recently, [Scealy and Welsh \(2014\)](#), have returned to the controversial questioning of the principles of CoDa, which they consider to be made to specifically exclude any methodology other than the one developed by Aitchison. As [Scealy and Welsh \(2014\)](#) recognise, the crux of the controversy lies in the definitions of composition and sample space in CoDA, both of which were introduced by [Aitchison \(1986\)](#) and based on constant sum vectors. The lack of clarity in the presentation of the properties scale invariance and subcompositional coherence is also a matter for discussion.

The main aim of this paper is to provide a precise and unequivocal definition of the concepts of composition, CoDa sample space and subcomposition, on which compositional analysis (CoAn) is based. Contrary to [Scealy and Welsh \(2014\)](#), we turn to mathematics to introduce these concepts with maximum precision. Thus, in Section 2 we define the quotient space of the compositions and we provide a precise definition of the concept of subcomposition. We also define what we understand by CoAn, distinguishing it from the traditional concept CoDA. In Section 3 we show how the logarithmic and exponential functions allow us to structure the sample space as a Euclidean space and to operate with the logratio coordinates of the data as if we were doing so in the real space. In the last section we compile the advantages and limitations of CoAn based on logratio coordinates and of the analysis based on transformations that take the positive orthant of the unit hypersphere as the sample space. Finally, we present the main conclusions.

2. The sample space in a compositional analysis

2.1. A composition is an equivalence class

We assume that our data and observations materialise in vectors $\mathbf{w} = (w_1, \dots, w_D)'$ with strictly positive components, that is vectors from real space \mathbb{R}_+^D , the positive orthant of \mathbb{R}^D . Note that we are eluding to the case of zero values in the data. We consider the zero as a special value that deserves a particular analysis according to its nature ([Palarea-Albaladejo and Martín-Fernández 2015](#)); that is, the reason why a zero value is present in a CoDa set is informative and determines the approach to be applied. The interested reader is referred to [Martín-Fernández, Palarea-Albaladejo, and Olea \(2011\)](#) for further information. In the discussion we outline some of the approaches and discuss some kinds of zero.

Sometimes the observational vectors \mathbf{w} are constant sum vectors. Typical examples are the data from time-use surveys where the sum equals to 24 in hours, 1440 in minutes or 100 in percentages. This case of CoDa is known as ‘closed data’. In other situations, the components of the observational vectors are themselves meaningful, that is, they represent absolute magnitudes. However, in spite of that, we can decide to take only the relative information into account for our analysis. For example, in the analysis of household expenditure on D commodity groups, we can decide to analyze the distribution of the expenditure regardless of the total. In both scenarios we are implicitly assuming that the vectors \mathbf{w} and $k\mathbf{w}$, with $k \in \mathbb{R}^+$, are providing us with the same compositional information, that is, the information given

by the ratios between the components. For example, the vectors $(0.3, 0.5, 0.2)$, $(30, 50, 20)$, $(7.2, 12, 4.8)$, and $(3/2, 5/2, 1)$ provide the same compositional information. In both cases we are assuming that our data are CoDa and our analysis will be a CoAn. Moreover, from a strictly mathematical point of view this implies that in a CoAn the sample space is not \mathbb{R}_+^D .

DEFINITION 2.1. Two D -observational vectors \mathbf{w} and \mathbf{w}^* are *compositionally equivalent*, written $\mathbf{w} \sim \mathbf{w}^*$, if there is a positive constant k such that $\mathbf{w} = k\mathbf{w}^*$. This equivalence relation on \mathbb{R}_+^D splits the space into equivalence classes, called *D -compositions* or, simply, *compositions*. The composition generated by an observational vector \mathbf{w} , i.e. the equivalence class of \mathbf{w} , is symbolized by $\underline{\mathbf{w}}$:

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

Following Aitchison (1986), it is clear that a D -part composition can be geometrically interpreted as a ray from the origin in the positive orthant of \mathbb{R}^D (Figure 1). Therefore, from a strictly mathematical point of view, the term CoAn is the equivalent of assuming that the sample space is the set of all D -compositions.

DEFINITION 2.2. The set of all D -compositions, that is, the quotient space \mathbb{R}_+^D/\sim is called the *D -compositional space* or, in brief, *compositional space*, and is symbolized by \mathcal{C}^D . We symbolize by ccl (from compositional class) the mapping from \mathbb{R}_+^D to \mathcal{C}^D which assigns each D -observational vector \mathbf{w} to the composition $\underline{\mathbf{w}}$, i.e.,

$$\begin{aligned} \text{ccl} : \mathbb{R}_+^D &\longrightarrow \mathcal{C}^D \\ \mathbf{w} &\longmapsto \underline{\mathbf{w}} = \{k\mathbf{w} : k \in \mathbb{R}^+\} . \end{aligned} \quad (1)$$

PROPERTY 2.1. Two D -observational vectors $\mathbf{w} = (w_1, \dots, w_D)'$ and $\mathbf{w}^* = (w_1^*, \dots, w_D^*)'$ are compositionally equivalent when the information provided by their ratios is the same, that is

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

Any D -composition $\underline{\mathbf{w}}$ is completely determined by its ratios w_i/w_j of their components. Therefore, in a CoAn the relevant information provided by the observational vector \mathbf{w} is found not in its components w_i , but rather in its ratios w_i/w_j . This is what we mean when we say that a composition only contains ‘relative information’ about its components. Note that all the observational vectors in the same ray (Figure 1) have the same ratios, providing the same relative information. That is, any point in the ray can be selected as representative of the equivalence class, and any statistical analysis has to provide the same information regardless of the representative selected. Importantly, if one applies a statistical method that does not take into account this essential attribute of the compositions, the application of different criteria to select the representatives will give different results and, likely, one will extract different conclusions.

To conclude, when we decide to do a CoAn we are assuming that the sample space of our data is the compositional space \mathcal{C}^D , which means in fact an acceptance of the ‘scale invariance’ principle.

2.2. Representatives of compositions

Any composition $\underline{\mathbf{w}}$ is determined by any observational vector \mathbf{w} that belongs to the equivalence class. Thus, many different criteria can be used to select a representative of a composition. Each criterion gives rise to a different reference frame where projecting the compositions of \mathcal{C}^D . Here we present the most commonly-used criteria that facilitate the interpretation and have relevant mathematical properties.

DEFINITION 2.3. The *linear criterion* selects the unit-sum vector $\mathbf{w}/\sum_{j=1}^D w_j$ to represent the composition $\underline{\mathbf{w}}$. We symbolize by r_1 the mapping from \mathcal{C}^D to the subset \mathcal{S}^D of \mathbb{R}_+^D , that is

$$\begin{aligned} r_1 : \mathcal{C}^D &\longrightarrow \mathcal{S}^D \subset \mathbb{R}_+^D \\ \underline{\mathbf{w}} &\longmapsto \mathbf{w}/\sum_{j=1}^D w_j, \end{aligned} \quad (2)$$

where \mathcal{S}^D is the well-known *unit simplex*, historically considered as the sample space of CoDa.

The mapping r_1 corresponds to the *constraining operator* or *closure operator* \mathcal{C} introduced by Aitchison (1986). Geometrically, $r_1(\underline{\mathbf{w}})$ is the intersection of the ray going from the origin through \mathbf{w} and the hyperplane of \mathbb{R}^D defined by the equation $w_1 + \dots + w_D = 1$ (Figure 1). This criterion can be generalized to representatives with a sum equal to 100 or any other positive value.

DEFINITION 2.4. We symbolize by r_s the mapping from \mathcal{C}^D to the subset Sph_+^D of \mathbb{R}_+^D which assigns to composition $\underline{\mathbf{w}}$ the intersection of the ray going from the origin through \mathbf{w} and the unit hypersphere of \mathbb{R}^D centred in the origin, i.e.,

$$\begin{aligned} r_s : \mathcal{C}^D &\longrightarrow \text{Sph}_+^D \subset \mathbb{R}_+^D \\ \underline{\mathbf{w}} &\longmapsto \mathbf{w}/\|\mathbf{w}\|, \end{aligned} \quad (3)$$

where Sph_+^D is the strictly positive orthant of the unit hypersphere of \mathbb{R}^D centred in the origin. We call this selection criterion the *spherical criterion* (Figure 1) because the representatives are unit-norm vectors using the classical Euclidean norm. This selection criterion was proposed by Watson and Philip (1989).

DEFINITION 2.5. The *hyperbolic criterion*, r_h , assigns to composition $\underline{\mathbf{w}}$ the intersection of the ray going from the origin through \mathbf{w} and the hyperbolic surface Hip_+^D in \mathbb{R}_+^D implicitly defined by the equation $\prod_{i=1}^D w_i = 1$:

$$\begin{aligned} r_h : \mathcal{C}^D &\longrightarrow \text{Hip}_+^D \subset \mathbb{R}_+^D \\ \underline{\mathbf{w}} &\longmapsto \mathbf{w}/g(\mathbf{w}), \end{aligned} \quad (4)$$

where $g(\mathbf{w}) = (\prod_{j=1}^D w_j)^{1/D}$ is the geometric mean of the components of vector \mathbf{w} (Fig. 1).

Note that the function composition $\log \circ r_h$ is equivalent to the *centred logratio* transformation (clr) introduced by Aitchison (1986): $\text{clr}(\mathbf{w}) = \log(\mathbf{w}/g(\mathbf{w}))$.

The mappings r_1 , r_s and r_h can also be viewed as scale-invariant functions from \mathbb{R}_+^D to \mathcal{C}^D . Recall that a function $f(\cdot)$ from \mathbb{R}_+^D is said to be ‘scale invariant’ if for any positive constant k and for any observational vector \mathbf{w} , the function verifies $f(k\mathbf{w}) = f(\mathbf{w})$.

These criteria to select a representative of a composition can be extended to any surface defined in \mathbb{R}_+^D using a bijective function. Indeed, making each composition correlate with the intersection of the corresponding ray with the surface is sufficient.

2.3. Subcompositions

In a CoAn, attention is usually focused on a determinate subset of the components of our observations of \mathbb{R}_+^D . For example, in time-use surveys we might only be interested in those activities that are different from the sleeping hours. If the analysis to be carried out on the components selected from our observations must also be compositional, then the sample space also needs to interpret it as a quotient space. This brings us to the need to introduce the concept of *subcomposition*.

DEFINITION 2.6. Given a composition $\underline{\mathbf{w}} \in \mathcal{C}^D$, any composition obtained from the selection of two or more components of the D -observational vector \mathbf{w} is termed a *subcomposition* of $\underline{\mathbf{w}}$. More precisely, let s be the number of selected components, with $2 \leq s < D$, and $i_1 < \dots < i_s$ the sub indexes of these components (we implicitly assume that the sub indexes

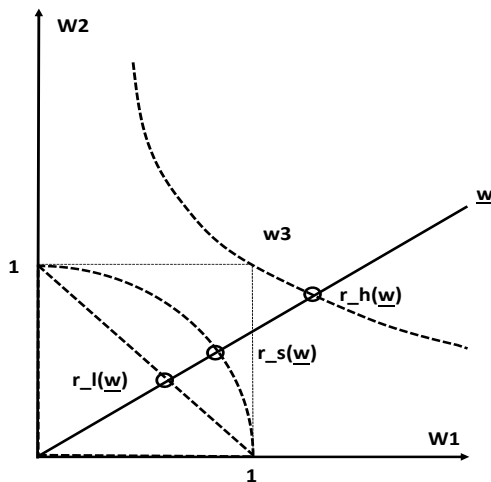


Figure 1: Linear, spherical and hyperbolic selection criteria (case $D = 2$).

of the D -observational vectors are $1, \dots, D$). Let \mathbf{S} be the $s \times D$ matrix with the ones in the positions $(1, i_1), \dots, (s, i_s)$ of the matrix and zeros in the remaining positions. Making a subcomposition can be viewed as the transformation sub_S from \mathcal{C}^D to \mathcal{C}^s given by

$$\begin{aligned} \text{sub}_S : \mathcal{C}^D &\rightarrow \mathcal{C}^s \\ \underline{\mathbf{w}} &\rightarrow \underline{\mathbf{S}\mathbf{w}} . \end{aligned} \tag{5}$$

The symbol \mathbf{w}_S indicates the observational subvector $\mathbf{S}\mathbf{w} = (w_{i_1}, \dots, w_{i_s})'$, and $\underline{\mathbf{w}}_S$ represents the final subcomposition which belongs to the compositional space \mathcal{C}^s . The transformation sub_S is compatible with the equivalence relation \sim , that is, equivalent observational vectors are transformed into equivalent subvectors. Importantly, the selected components $(w_{i_1}, \dots, w_{i_s})'$ provide the same relative information regardless they belong to $\underline{\mathbf{w}}$ or they form the subcomposition $\underline{\mathbf{w}}_S$. This ‘subcompositional coherence’ is an inherent attribute of the compositions rather a required principle. The formation of a subcomposition $\underline{\mathbf{w}}_S$ from a D -composition $\underline{\mathbf{w}}$ can be geometrically interpreted as the orthogonal projection of the ray associated to $\underline{\mathbf{w}}$ onto the subspace of \mathbb{R}_+^D generated by the positive coordinate axes associated to the components in the subcomposition. Figure 2 shows the subcompositions for the case $D = 3$ and the relationship with the corresponding representatives.

3. The Euclidean compositional space

Any statistical analysis with data from the sample space \mathcal{C}^D needs this space to have an algebraic and metric structures. Remember that such basic concepts as the mean and the variance of a set of data are based on the algebraic and metric structure of the sample space of the data. The strategy that we develop is to define an isomorphism between \mathcal{C}^D and another Euclidean space using the logarithmic function. Despite this isomorphism may not be the unique feasible isomorphism, the rest of the possibilities are still unknown to us.

3.1. A quotient Euclidean space in the real space

The well known classical Euclidean space \mathbb{R}^D is based on the addition and subtraction operations. Because we need to connect the relative information provided by the ratios of components with an existing Euclidean space, the logarithmic function becomes a useful option.

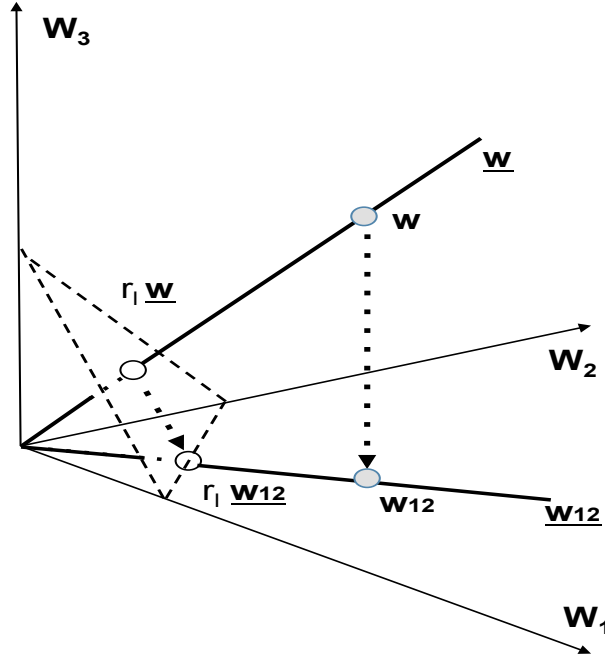


Figure 2: Geometrical interpretation in \mathbb{R}_+^3 of a subcomposition \underline{w}_{12} of a composition $\underline{w} \in \mathcal{C}^3$. Filled circles are the observational vectors. Empty circles their corresponding linear representatives.

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

DEFINITION 3.1. Two vectors \mathbf{z} and \mathbf{z}^* in \mathbb{R}^D are *equivalent*, written $\mathbf{z} \equiv \mathbf{z}^*$, if a constant λ exists such that $\mathbf{z}^* - \mathbf{z} = \lambda \mathbf{1}_D$. The equivalence class $\{\mathbf{z} + \lambda \mathbf{1}_D : \lambda \in \mathbb{R}\}$ generated by the vector \mathbf{z} in \mathbb{R}^D is denoted by $\underline{\underline{\mathbf{z}}}$. The set of all these classes is the quotient space \mathbb{R}^D / \equiv , denoted by \mathcal{L}^D . We denote by *ocl* (from ones class) the mapping from \mathbb{R}^D to \mathcal{L}^D which assigns each vector $\mathbf{z} \in \mathbb{R}^D$ to the class $\underline{\underline{\mathbf{z}}}$

$$\begin{aligned} \text{ocl} : \mathbb{R}^D &\rightarrow \mathcal{L}^D \\ \mathbf{z} &\rightarrow \underline{\underline{\mathbf{z}}} . \end{aligned} \quad (6)$$

Figure 3 shows that the classes $\underline{\underline{\mathbf{z}}}$ can be geometrically interpreted by straight lines parallel to $\mathbf{1}_D$. A simple criterion for selecting a representative of an equivalence class $\underline{\underline{\mathbf{z}}}$ is to assign the intersection point of the straight line associated to this class and the orthogonal hyperplane by the origin

$$V_D = \{\mathbf{z} \in \mathbb{R}^D : \mathbf{z}' \mathbf{1}_D = 0\} . \quad (7)$$

DEFINITION 3.2. We denote by r_{V_D} the one-to-one mapping which assigns each class $\underline{\underline{\mathbf{z}}}$ to this representative

$$\begin{aligned} r_{V_D} : \mathcal{L}^D &\rightarrow V_D \\ \underline{\underline{\mathbf{z}}} &\rightarrow \mathbf{z} - \frac{\sum_{j=1}^D z_j}{D} \mathbf{1}_D = \mathbf{H}_D \mathbf{z} , \end{aligned} \quad (8)$$

where \mathbf{H}_D is the $D \times D$ centering matrix, that is $\mathbf{H}_D = \mathbf{I}_D - D^{-1} \mathbf{J}_D$ (\mathbf{I}_D is the identity matrix of order $D \times D$, and $\mathbf{J}_D = \mathbf{1}_D \mathbf{1}_D'$).

DEFINITION 3.3. The sum of two classes $\underline{\underline{\mathbf{z}}}$ and $\underline{\underline{\mathbf{z}^*}}$ in \mathcal{L}^D is defined as $\underline{\underline{\mathbf{z}}} + \underline{\underline{\mathbf{z}^*}} = \underline{\underline{\mathbf{z} + \mathbf{z}^*}}$, and the product of an equivalence class $\underline{\underline{\mathbf{z}}}$ by a constant $\alpha \in \mathbb{R}$ is defined by $\alpha \underline{\underline{\mathbf{z}}} = \underline{\underline{\alpha \mathbf{z}}}$.

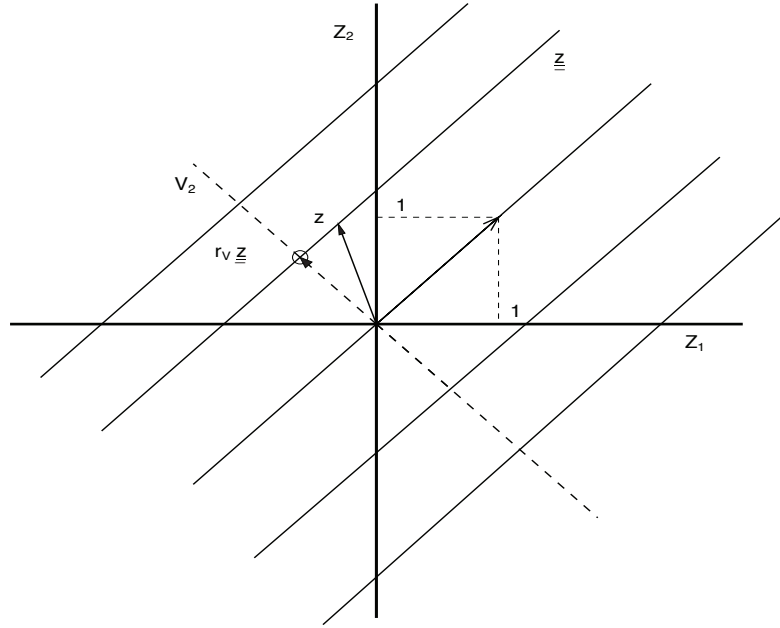


Figure 3: Selection of the representative $r_{V_D} \underline{z}$ (empty circle) for an equivalence class \underline{z} in $\mathcal{L}^2 = \mathbb{R}^2 / \equiv$. Dashed line is the orthogonal hyperplane to vector $\mathbf{1}_D$.

With these definitions, the quotient space \mathcal{L}^D becomes a real vector space. The class of $\underline{\mathbf{0}_D}$ is the neutral element and the opposite of \underline{z} is the class $\underline{-z}$. Moreover, the mapping r_{V_D} is an isomorphism between the vector space $(\mathcal{L}^D, +, \cdot)$ and the subspace V_D of \mathbb{R}^D (Equation 7). Since the dimension of V_D is $D - 1$, the dimension of the vector space \mathcal{L}^D will be also equal to $D - 1$.

The vector space structure defined in \mathcal{L}^D is coherent with a subcompositional analysis because one can define subvectors in the space V_D and reproduce them in \mathcal{L}^D using the inverse mapping $r_{V_D}^{-1}$. More precisely, the mapping sub_S (Equation 5) corresponds to orthogonal projection of the hyperplane V_D (Equation 7) onto the subspace of \mathbb{R}^D defined implicitly by

$$\{z \in \mathbb{R}^D : z' \mathbf{1}_D = 0; z_{j_1} = 0; \dots, z_{j_{(D-s)}} = 0\},$$

where $j_1, \dots, j_{(D-s)}$ are the sub indexes of the no-selected components in the subvector.

Given that the elements of \mathcal{L}^D can be interpreted as straight lines parallel to vector $\mathbf{1}_D$, one can define the distance between the two classes \underline{z} and $\underline{z^*}$ of \mathcal{L}^D as the Euclidean distance between these two straight lines in \mathbb{R}^D . This distance is equal to the length of the difference vector $r_{V_D}(\underline{z}) - r_{V_D}(\underline{z^*})$ (Figure 4).

Following this strategy, it is possible to reproduce the Euclidean structure defined on $V_D \subset \mathbb{R}^D$ on \mathcal{L}^D .

DEFINITION 3.4. For each $\underline{z}, \underline{z^*} \in \mathcal{L}^D$, we define the \mathcal{L} -inner product $\langle \underline{z}, \underline{z^*} \rangle_{\mathcal{L}}$ as the usual inner product $\langle r_{V_D} \underline{z}, r_{V_D} \underline{z^*} \rangle$ in \mathbb{R}^D .

It follows that $\langle \underline{z}, \underline{z^*} \rangle_{\mathcal{L}} = z' \mathbf{H}_D z^*$. Then it is possible to define a norm and a distance in \mathcal{L}^D from the \mathcal{L} -inner product.

DEFINITION 3.5. The \mathcal{L} -norm of an equivalence class $\underline{z} \in \mathcal{L}^D$ is given by

$$\|\underline{z}\|_{\mathcal{L}} = (\langle \underline{z}, \underline{z} \rangle_{\mathcal{L}})^{1/2} = (z' \mathbf{H}_D z)^{1/2}, \tag{9}$$

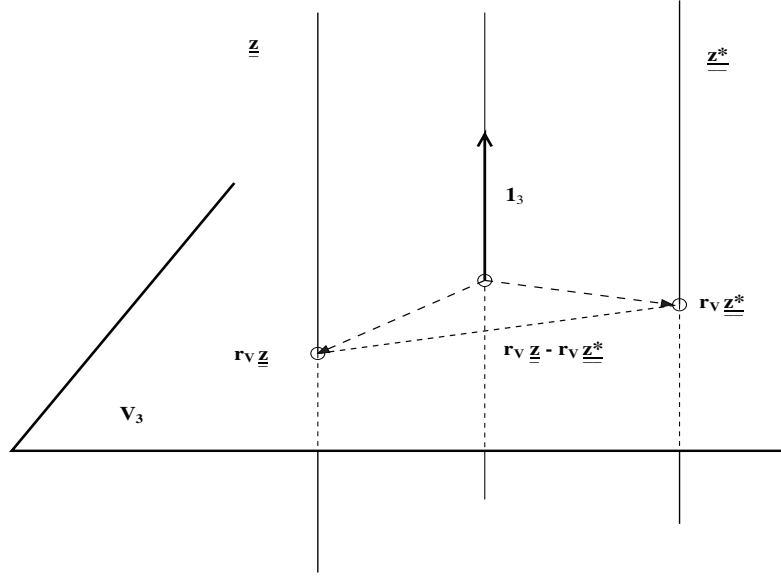


Figure 4: Two equivalence classes \underline{z} and \underline{z}^* of \mathcal{L}^D , its corresponding representatives $r_{V_D} \underline{z}$ and $r_{V_D} \underline{z}^*$, and the distance between them (case $D=3$).

and the \mathcal{L} -distance between two classes \underline{z} and \underline{z}^* in \mathcal{L}^D is given by

$$d_{\mathcal{L}}(\underline{z}, \underline{z}^*) = \|\underline{z} - \underline{z}^*\|_{\mathcal{L}} = [(\mathbf{z} - \mathbf{z}^*)' \mathbf{H}_D (\mathbf{z} - \mathbf{z}^*)]^{1/2}. \quad (10)$$

Since $d_{\mathcal{L}}(\underline{z}, \underline{z}^*) = d(r_{V_D} \underline{z}, r_{V_D} \underline{z}^*)$, the following property holds.

PROPERTY 3.1. From the definitions (9) and (10), the quotient space \mathcal{L}^D becomes an Euclidean space isometric to the subspace V_D of \mathbb{R}^D .

3.2. The logarithmic isomorphism 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, i.e.,

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

and

$$\mathbf{z} \equiv \mathbf{z}^* \text{ in } \mathbb{R}^D \iff \exp \mathbf{z} \sim \exp \mathbf{z}^* \text{ in } \mathbb{R}_+^D.$$

Therefore, these transformations can be extended to the quotient spaces \mathcal{C}^D and \mathcal{L}^D .

DEFINITION 3.6. We will symbolize by logc the transformation from \mathcal{C}^D to \mathcal{L}^D , i.e.,

$$\begin{aligned} \text{logc} : \mathcal{C}^D &\longrightarrow \mathcal{L}^D \\ \underline{\mathbf{w}} &\longmapsto \underline{\log \mathbf{w}}, \end{aligned} \quad (11)$$

and by expc the inverse transformation from \mathcal{L}^D to \mathcal{C}^D , i.e.,

$$\begin{aligned} \text{expc} : \mathcal{L}^D &\longrightarrow \mathcal{C}^D \\ \underline{\mathbf{z}} &\longmapsto \underline{\exp \mathbf{z}}. \end{aligned} \quad (12)$$

The representative in V_D of the equivalence class $\underline{\log \mathbf{w}}$ is

$$r_{V_D}(\underline{\log \mathbf{w}}) = \mathbf{H}_D \log \mathbf{w} = \log \frac{\mathbf{w}}{g(\mathbf{w})} ,$$

where $g(\mathbf{w})$ is the geometric mean of the vector \mathbf{w} .

Importantly, the function composition $r_{V_D} \circ \log c$ is equivalent to the transformation clr (Aitchison 1986). This one-to-one correspondence between \mathcal{C}^D and \mathcal{L}^D allows a real vector space isomorphic to \mathcal{L}^D to be defined in \mathcal{C}^D .

DEFINITION 3.7. In correspondence with the sum in \mathcal{L}^D , the inner operation \otimes in \mathcal{C}^D is defined as

$$\underline{\mathbf{w}} \otimes \underline{\mathbf{w}^*} = \text{expc}(\underline{\log \mathbf{w}} + \underline{\log \mathbf{w}^*}) = \text{expc}(\underline{\log \mathbf{w} + \log \mathbf{w}^*}) = \underline{(w_1 w_1^*, \dots, w_D w_D^*)}' ,$$

for any $\underline{\mathbf{w}}, \underline{\mathbf{w}^*} \in \mathcal{C}^D$.

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

$$\alpha \odot \underline{\mathbf{w}} = \text{expc}(\alpha \underline{\log \mathbf{w}}) = \text{expc}(\underline{\alpha \log \mathbf{w}}) = \underline{(w_1^\alpha, \dots, w_D^\alpha)'} \quad (\underline{\mathbf{w}} \in \mathcal{C}^D) \ (\alpha \in \mathbb{R}) .$$

The operations \otimes and \odot are respectively the perturbation and power operations introduced by Aitchison (1986).

Therefore, $(\mathcal{C}^D, \otimes, \odot)$ becomes a real vector space of dimension $D - 1$, isomorphic to the quotient space \mathcal{L}^D and to the subspace V_D of \mathbb{R}^D . In the commutative group (\mathcal{C}^D, \otimes) , the composition $\underline{\mathbf{1}}_D = (1, \dots, 1)'$ is the neutral element, and the inverse composition $\underline{\mathbf{w}}^{-1}$ of $\underline{\mathbf{w}}$ is the composition $\underline{\mathbf{w}}^{-1} = (1/w_1, \dots, 1/w_D)'$.

Moreover, the structure of real vector space of $(\mathcal{C}^D, \otimes, \odot)$ is compatible with the concept of subcomposition.

PROPERTY 3.2. The mapping sub_S defined in Equation 5 is a linear function between the vector spaces $(\mathcal{C}^D, \otimes, \odot)$ and $(\mathcal{C}^S, \otimes, \odot)$. Therefore, it holds that

$$\text{sub}_S(\underline{\mathbf{w}} \otimes \underline{\mathbf{w}^*}) = \underline{\mathbf{w}}_S \otimes \underline{\mathbf{w}^*}_S \quad \text{and} \quad \text{sub}_S(\alpha \odot \underline{\mathbf{w}}) = \alpha \odot \underline{\mathbf{w}}_S , \tag{13}$$

for any $\underline{\mathbf{w}}, \underline{\mathbf{w}^*} \in \mathcal{C}^D$ and $\alpha \in \mathbb{R}$.

3.3. The compositional space as an affine Euclidean space

Because $(\mathcal{C}^D, \otimes, \odot)$ 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 3.8. Given a composition $\underline{\mathbf{p}} \in \mathcal{C}^D$, the *perturbation* associated to $\underline{\mathbf{p}}$ is the transformation from \mathcal{C}^D to \mathcal{C}^D defined by

$$\underline{\mathbf{w}} \rightarrow \underline{\mathbf{p}} \otimes \underline{\mathbf{w}} \quad (\underline{\mathbf{w}} \in \mathcal{C}^D) .$$

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

Perturbations in the compositional space play the same role as translations play in the real space. Like them, 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 $\underline{\mathbf{p}}_1$ and $\underline{\mathbf{p}}_2$ is the perturbation associated to $\underline{\mathbf{p}}_1 \otimes \underline{\mathbf{p}}_2$. Furthermore, the perturbation associated to $\underline{\mathbf{1}}_D$ is the identity perturbation which does not produce any change when applied to a composition. Also, for any given perturbation

$\underline{\mathbf{p}}$ there is the inverse perturbation $\underline{\mathbf{p}}^{-1}$ which undoes the changes produced by $\underline{\mathbf{p}}$. Finally, given two compositions $\underline{\mathbf{w}}'$ and $\underline{\mathbf{w}}^* \in \mathcal{C}^D$, a unique perturbation $\underline{\mathbf{p}}$ exists which transforms $\underline{\mathbf{w}}$ on $\underline{\mathbf{w}}^*$. This perturbation is

$$\underline{\mathbf{p}} = \underline{\mathbf{w}}^* \otimes \underline{\mathbf{w}}^{-1} = \left(\frac{w_1^*}{w_1}, \dots, \frac{w_D^*}{w_D} \right)'$$

the *perturbation difference* between $\underline{\mathbf{w}}$ and $\underline{\mathbf{w}}^*$. Thus, the measurement of the ‘difference’ between two compositions is defined from the ratios between the components of compositions. The one-to-one transformations logc (Equation 11) and expc (Equation 12) between \mathcal{C}^D and \mathcal{L}^D allow the real Euclidean structure defined on \mathcal{L}^D to be transferred to \mathcal{C}^D .

DEFINITION 3.9. The *compositional inner product* of two compositions $\underline{\mathbf{w}}$ and $\underline{\mathbf{w}}^*$ will be equal to

$$\langle \underline{\mathbf{w}}, \underline{\mathbf{w}}^* \rangle_{\mathcal{C}} = \langle \underline{\text{log } \underline{\mathbf{w}}}, \underline{\text{log } \underline{\mathbf{w}}^*} \rangle_{\mathcal{L}} = (\text{log } \underline{\mathbf{w}})' \mathbf{H}_D \text{log } \underline{\mathbf{w}}^*.$$

Importantly, $\langle \underline{\mathbf{w}}, \underline{\mathbf{w}}^* \rangle_{\mathcal{C}} = \langle \text{clr } \underline{\mathbf{w}}, \text{clr } \underline{\mathbf{w}}^* \rangle$, i.e., the standard inner product of the clr-transformed vectors.

From this inner product in \mathcal{C}^D we can define a norm and a distance in the compositional space.

DEFINITION 3.10. The *compositional norm* of a composition $\underline{\mathbf{w}} \in \mathcal{C}^D$ will be given by

$$\|\underline{\mathbf{w}}\|_{\mathcal{C}} = (\langle \underline{\mathbf{w}}, \underline{\mathbf{w}} \rangle_{\mathcal{C}})^{1/2} = [(\text{log } \underline{\mathbf{w}})' \mathbf{H}_D \text{log } \underline{\mathbf{w}}]^{1/2},$$

and the *compositional distance* between two compositions $\underline{\mathbf{w}}$ and $\underline{\mathbf{w}}^*$ of \mathcal{C}^D is given by

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

The distance $d_{\mathcal{C}}(\underline{\mathbf{w}}, \underline{\mathbf{w}}^*)$ defined on \mathcal{C}^D is equivalent to the Aitchison distance (Aitchison, Barceló-Vidal, Martín-Fernández, and Pawłowsky-Glahn 2000) that can be expressed as the typical Euclidean distance between the corresponding clr-transformed vectors.

PROPERTY 3.3. In relation to subcompositions, the distance $d_{\mathcal{C}}$ satisfies what is known as *subcompositional dominance*, according to which

$$d_{\mathcal{C}}(\underline{\mathbf{w}}_S, \underline{\mathbf{w}}_S^*) \leq d_{\mathcal{C}}(\underline{\mathbf{w}}, \underline{\mathbf{w}}^*),$$

for any $\underline{\mathbf{w}}, \underline{\mathbf{w}}^* \in \mathcal{C}^D$ and for any subcomposition S .

PROOF. It is sufficient to demonstrate that the compositional norm of a composition $\underline{\mathbf{w}}$ is greater or equal to the compositional norm of a subcomposition $\underline{\mathbf{w}}_S$ obtained by removing one of its components. If, without lack of generality, we assume that $\underline{\mathbf{w}}_S = \underline{(w_1, \dots, w_{D-1})}'$, then it holds that

$$\|\underline{\mathbf{w}}\|_{\mathcal{C}}^2 = \|\underline{\mathbf{w}}_S\|_{\mathcal{C}}^2 + \frac{1}{D(D-1)} (\text{log } w_1 + \dots + \text{log } w_{D-1} - (D-1) \text{log } w_D)^2,$$

and $\|\underline{\mathbf{w}}\|_{\mathcal{C}} \geq \|\underline{\mathbf{w}}_S\|_{\mathcal{C}}$.

The subcompositional dominance property of the Euclidean space \mathcal{C}^D correlates with the traditional property at the real space \mathbb{R}^D , according to which the distance between the orthogonal projections of two points on any subspace is never greater than the original distance between the points. In practical terms, this property also admits the following interpretation: given two observational vectors $\underline{\mathbf{w}}_S$ and $\underline{\mathbf{w}}_S^*$, if one adds supplementary components to both vectors to form, respectively, the vectors $\underline{\mathbf{w}}$ and $\underline{\mathbf{w}}^*$, then the difference between the new vectors must be at least equal to the difference between the initial vectors.

Palarea-Albaladejo, Martín-Fernández, and Soto (2012) present examples to illustrate that other usual distances, like the typical Euclidean or the angular distances, do not verify this property. As a consequence, when one applies these distances or one calculates related statistics (e.g., correlation coefficient), some misleading results can be obtained.

Since \mathcal{C}^D is a real vector space of dimension $D - 1$, any composition $\underline{\mathbf{w}}$ could be identified with its $D - 1$ coordinates relative to a basis of \mathcal{C}^D . In practice, we can obtain a basis of \mathcal{C}^D from a basis of the subspace V_D of \mathbb{R}^D . Indeed, if $\mathbf{v}_1, \dots, \mathbf{v}_{D-1}$ is a basis of V_D then $\text{expc}(\mathbf{r}_{V_D}^{-1} \mathbf{v}_1), \dots, \text{expc}(\mathbf{r}_{V_D}^{-1} \mathbf{v}_{D-1})$ is a basis of \mathcal{C}^D , and the coordinates of a composition $\underline{\mathbf{w}}$ relative to this basis coincide with the coordinates of $\mathbf{r}_{V_D}(\underline{\log \mathbf{w}})$ relative to $\mathbf{v}_1, \dots, \mathbf{v}_{D-1}$.

DEFINITION 3.11. Let $\mathbf{v}_1, \dots, \mathbf{v}_{D-1}$ be a basis of V_D , and let \mathbf{V} be the $D \times (D - 1)$ matrix $[\mathbf{v}_1 : \dots : \mathbf{v}_{D-1}]$. Then the coordinates of the composition $\underline{\mathbf{w}}$ relative to the basis $\text{expc}(\mathbf{r}_{V_D}^{-1} \mathbf{v}_1), \dots, \text{expc}(\mathbf{r}_{V_D}^{-1} \mathbf{v}_{D-1})$ are the components of the vector $(\mathbf{FV})^{-1} \mathbf{F} \log \mathbf{w}$, where $\mathbf{F} = [\mathbf{I}_{D-1} : -\mathbf{1}_{D-1}]$.

Note the expression of the coordinates of $\underline{\mathbf{w}}$ will depend on the matrix \mathbf{V} we selected. These coordinates are usually known as *logratio coordinates* because they are always expressed in terms of logarithms of ratios of components. For example, for

$$\mathbf{V} = \begin{bmatrix} 1 - 1/D & -1/D & -1/D & \dots & -1/D \\ -1/D & 1 - 1/D & -1/D & \dots & -1/D \\ -1/D & -1/D & 1 - 1/D & \dots & -1/D \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1/D & -1/D & -1/D & \dots & 1 - 1/D \\ -1/D & -1/D & -1/D & \dots & -1/D \end{bmatrix}, \quad (14)$$

the coordinates of $\underline{\mathbf{w}}$ relative to \mathbf{V} are equal to $(\log(w_1/w_D), \dots, \log(w_{D-1}/w_D))'$. In this case, the logratio coordinates coincide with the *additive logratio* transformation (alr) introduced by Aitchison (1986).

When one is making a statistical analysis it is recommendable to select orthonormal basis in \mathcal{C}^D because the metrics properties are preserved under a change of basis. This fact guarantees the invariance of the results under a change of basis. To select an orthonormal basis it suffices that the matrix \mathbf{V} verifies the two identities $\mathbf{V}'\mathbf{V} = \mathbf{I}_{D-1}$ and $\mathbf{V}\mathbf{V}' = \mathbf{H}_D$. In this case, the mapping that assigns composition $\underline{\mathbf{w}}$ to its logratio coordinates is the *isometric logratio* transformation $\text{ilr}_{\mathbf{V}}$ relative to matrix \mathbf{V} (Egozcue, Pawlowsky-Glahn, Mateu-Figueras, and Barceló-Vidal 2003), that is

$$\begin{aligned} \text{ilr}_{\mathbf{V}} : \mathcal{C}^D &\longrightarrow \mathbb{R}^{d-1} \\ \underline{\mathbf{w}} &\longmapsto \text{ilr}_{\mathbf{V}} \underline{\mathbf{w}} = (\mathbf{FV})^{-1} \mathbf{F} \log \mathbf{w} . \end{aligned}$$

In practice, it is very useful to select a basis that facilitates the interpretation of the logratio coordinates. Egozcue and Pawlowsky-Glahn (2005) describe a stepwise procedure to make an orthonormal basis of \mathcal{C}^D from sequential binary partitions of components of the observational vectors of \mathbb{R}_+^D .

4. Final remarks and conclusions

Because all Euclidean spaces of the same dimension are isometric, the sample space of CoDa \mathcal{C}^D is isometric to \mathbb{R}^{D-1} . This fact allows all the statistical procedures that we naturally apply on the real space \mathbb{R}^{D-1} to be applied to CoDa. The isomorphism presented in this article is based on the logarithmic function. From a theoretical point of view, other approaches could be possible but are unknown to us. With our approach, the compositional quotient space \mathcal{C}^D has an algebraic and a metric structure induced by the isomorphism. Consequently, it suffices to work with the logratio coordinates of the compositions with respect to an orthonormal

basis on \mathcal{C}^D (Mateu-Figueras, Pawlowsky-Glahn, and Egozcue 2011). That is, our CoAn is in essence a *logratio* CoAn, that is an analysis of CoDa based on the logarithm of the information provided by the ratios.

The fact that our analysis focuses on ratios means that it can be applied directly to the original data of \mathbb{R}_+^D , to the simplex \mathcal{S}^D , to the strictly positive orthant of the unit hypersphere Sph_+^D , to the hyperbolic surface Hip_+^D or to any other representative. Moreover, when working with logratio coordinates all of the statistical procedures that are defined in \mathbb{R}^{D-1} , both descriptive and inferential, are transferred to the space \mathcal{C}^D . The application of CoAn leads to the assumption that the group of perturbations is the operating group on the compositional space, in the same manner as we assume that the translations is the operating group in the real space. This is the keystone of the methodology introduced by Aitchison (1986). In fact, it means accepting that the ‘difference’ between two compositions $\underline{\mathbf{w}} = (w_1, \dots, w_D)'$ and $\underline{\mathbf{w}}^* = (w_1^*, \dots, w_D^*)'$ is based on the ratios w_j^*/w_j between parts instead of on the arithmetic differences $w_j^* - w_j$, according to the ‘relative scale’ property. Therefore, for example, the difference between the compositions $(0.980, 0.010, 0.010)'$ and $(0.970, 0.002, 0.028)'$ is more than three times greater than the distance between $(0.300, 0.200, 0.500)'$ and $(0.200, 0.300, 0.500)'$. The relative scale property of CoAn justifies the choice of the logarithm transformation to measure the difference between two compositions.

The CoAn applies only in the open orthant \mathbb{R}_+^D . That is, the components of the observational vectors must be strictly positive. This limitation is certainly a difficulty because often the observations contain zeros. However, when the zeros are rounded zeros or are count zeros they can be preprocessed using techniques inspired by techniques for missing data (Palarea-Albaladejo and Martín-Fernández 2015) that make a replacement by a small value. When the zero is an essential zero, that is, the zero value is a *true* value, it makes no sense to replace the zero by a small value. In this case, the analysis should take into account the presence and absence of zeros, that is, the pattern of zeros. Both descriptive and inferential analysis should be performed among the groups defined by the pattern of zeros.

Some researchers, for example Watson and Philip (1989), consider that the appropriate group to operate on compositions is the rotations on the sphere and not the perturbations on which the logratio CoAn is based. Watson and Philip (1989) represent a composition $\underline{\mathbf{w}}$ from the components of the unitary vector $\underline{\mathbf{w}}/\|\underline{\mathbf{w}}\|$, that is, from the cosine of the different angles that $\underline{\mathbf{w}}$ forms with the axes of coordinates. Then, the angle formed by two observation vectors $\underline{\mathbf{w}}$ and $\underline{\mathbf{w}}^*$ is taken as the appropriate measure from which to define the distance between the two compositions. Others, for example Wang *et al.* (2007) and Scealy and Welsh (2011), also apply the methodology of Watson and Philip (1989) after applying the scale-invariant transformation $\underline{\mathbf{w}} \rightarrow (\underline{\mathbf{w}}/\sum_{j=1}^D w_j)^{1/2}$ to the observations. Thus, they work with the components of the unitary vector $(\underline{\mathbf{w}}/\sum_{j=1}^D w_j)^{1/2}$ rather than the coordinates of the vector $\underline{\mathbf{w}}/\|\underline{\mathbf{w}}\|$. From these approaches, which are based on the representation of the compositions on the positive orthant of the unit hypersphere centred at the origin, the authors apply the statistical analysis (characteristic) of *directional statistics*, based on the von Mises-Fisher distribution. As stated in Aitchison (1982)’s final discussion, the problems of this approach derives from the fact that the von Mises-Fisher distribution is defined on the whole unit hypersphere and not only on the positive orthant. This leads to problems when the components of $\underline{\mathbf{w}}$ are too close to 0. Aitchison (1982) also points out the difficulties that CoAn based on the spherical representation of the compositions encounters when dealing with problems related to independence and regression. Neither is it possible from this representation to easily relate the statistics that describe a set of compositions $\underline{\mathbf{w}}_1, \dots, \underline{\mathbf{w}}_n$ of \mathcal{C}^D to the statistics of the subcompositions $\underline{\mathbf{w}}_{S,1}, \dots, \underline{\mathbf{w}}_{S,n}$.

To conclude, the most relevant results shown in this article are:

- A *composition* is an equivalence class and its sample space is the quotient space \mathcal{C}^D . Geometrically, the compositions are semi-straight lines by the origin of the positive orthant of the space \mathbb{R}_+^D . We refer any analysis of these equivalence classes as *compositional*

analysis (CoAn). Regardless the use of the logarithm function or a transformation, when an analyst decides to do a CoAn he or she is assuming that the sample space of the data is the compositional space \mathcal{C}^D , which means in fact an acceptance of the ‘scale invariance’ principle of CoDA.

- The logarithmic and exponential transformations provide the space \mathcal{C}^D with an Euclidean space structure. We denominate *logratio CoAn* the compositional analysis developed from this structure of \mathcal{C}^D . It agrees with the methodology introduced by Aitchison (1982), based on a logratio relative scale of measurement of the difference between two compositions.
- The logratio CoAn allows us to carry out the standard statistical analyses on the logratio coordinates.
- The logratio CoAn allows us to apply the subcompositional analysis in a natural and intuitive way, giving results which are coherent with those obtained from the whole compositions.
- The logratio CoAn has the drawback of being unable to operate directly with compositions with zero values. Applying preprocessing techniques to replace rounded and count zeros is then recommended. A statistical analysis in the presence of essential zeros must take into account the groups defined by the pattern of zeros.
- When the techniques for analysing directional data are restricted to compositions, they must be considered to be a CoAn. Even though these analyses do not have the problem of the zeros it is still impossible to guarantee that coherent results will always be obtained in inferential studies (e.g., confidence regions), that is, strictly contained within the positive orthant, because the sample space of these analyses is the whole sphere. Moreover, this approach does not guarantee that a subcompositional analysis will produce results that concur with the results of the analysis of the whole composition.

5. Acknowledgements

This work has been partially financed by the Ministerio de Economía y Competitividad (Ref: MTM2015-65016-C2-1-R) and the Agència de Gestió d’Ajuts Universitaris i de Recerca (AGAUR), Generalitat de Catalunya (Ref: 2014 SGR 551).

References

- Aitchison J (1982). “The Statistical Analysis of Compositional Data (with Discussion).” *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **44**, 139–177.
- Aitchison J (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.
- Aitchison J (1990). “Comment on *Measures of Variability for Geological Data*, by D.F. Watson and G.M. Philip.” *Mathematical Geology*, **22**, 223–226.
- Aitchison J (1991). “Delusions of Uniqueness and Ineluctability.” *Mathematical Geology*, **23**, 275–277.
- Aitchison J, Barceló-Vidal C, Martín-Fernández JA, Pawłowsky-Glahn V (2000). “Logratio Analysis and Compositional Distance.” *Mathematical Geology*, **32**, 271–275.

- 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, Pawlowsky-Glahn V (2005). “Groups of Parts and their Balances in Compositional Data Analysis.” *Mathematical Geology*, **37**, 795–828.
- Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003). “Isometric Logratio Transformations for Compositional Data Analysis.” *Mathematical Geology*, **35**, 279–300.
- Martín-Fernández JA, Palarea-Albaladejo J, Olea RA (2011). *Dealing with Zeros*. John Wiley & Sons, Ltd.
- Mateu-Figueras G, Pawlowsky-Glahn V, Egozcue JJ (2011). “The Principle of Working on Coordinates.” In Pawlowsky-Glahn and Buccianti (2011), pp. 31–42. 378 p.
- Palarea-Albaladejo J, Martín-Fernández JA (2015). “zCompositions - R Package for Multivariate Imputation of Nondetects and Zeros in Compositional Data sets.” *Chemometrics and Intelligent Laboratory Systems*, **143**, 85–96.
- Palarea-Albaladejo J, Martín-Fernández JA, Soto JA (2012). “Dealing with Distances and Transformations for Fuzzy C-Means Clustering of Compositional Data.” *Journal of Classification*, **29**, 144–169.
- Pawlowsky-Glahn V, Buccianti A (eds.) (2011). *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons. 378 p.
- Scealy JL, Welsh AH (2011). “Regression for Compositional Data by Using Distributions Defined on the Hypersphere.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 351–375.
- Scealy JL, Welsh AH (2014). “Colours and cocktails: Compositional Data Analysis 2013 Lancaster Lecture.” *Australian & New Zealand Journal of Statistics*, **56**, 145–169.
- Wang H, Liu Q, Mok HM, Fu L, Tse WM (2007). “A Hyperspherical Transformation Forecasting Model for Compositional Data.” *European Journal of Operational Research*, **179**, 459–468.
- Watson DF (1990). “Reply to Comment on Measures of Variability for Geological Data, by D.F. Watson and G.M. Philip.” *Mathematical Geology*, **22**, 227–231.
- Watson DF (1991). “Reply to Delusions of Uniqueness and Ineluctability, by J. Aitchison.” *Mathematical Geology*, **23**, 279–279.
- Watson DF, Philip GM (1989). “Measures of Variability for Geological Data.” *Mathematical Geology*, **21**, 233–254.

Affiliation:

Carles Barceló-Vidal
Department of Computer Science, Applied Mathematics and Statistics
University of Girona
E-17071 Girona, Spain
E-mail: carles.barcelo@udg.edu

Josep-Antoni Martín-Fernández
Department of Computer Science, Applied Mathematics and Statistics
University of Girona
E-17071 Girona, Spain
E-mail: josepantoni.martin@udg.edu
URL: <http://ima.udg.edu/~jamf/>