Please cite this paper as: R. Pérez-Fernández, M. Gagolewski, B. De Baets,
On the aggregation of compositional data, Information Fusion 73, 103–110, 2021, doi:10.1016/j.inffus.2021.02.021

On the aggregation of compositional data

Raúl Pérez-Fernández[a,b,*], Marek Gagolewski[c,d], Bernard De Baets[b]

[a] *Department of Statistics and O.R. and Mathematics Didactics, University of Oviedo, Calle Federico García Lorca 18, 33007 Oviedo, Spain*
[b] *KERMIT, Department of Data Analysis and Mathematical Modelling, Ghent University, Coupure links 653, 9000 Gent, Belgium*
[c] *School of Information Technology, Deakin University, Geelong, VIC 3220, Australia*
[d] *Systems Research Institute, Polish Academy of Sciences, ul. Newelska 6, 01-447 Warsaw, Poland*

Abstract

Compositional data naturally appear in many fields of application. For instance, in chemistry, the relative contributions of different chemical substances to a product are typically described in terms of a compositional data vector. Although the aggregation of compositional data frequently arises in practice, the functions formalizing this process do not fit the standard order-based aggregation framework. This is due to the fact that there is no intuitive order that carries the semantics of the set of compositional data vectors (referred to as the standard simplex). In this paper, we consider the more general betweenness-based aggregation framework that yields a natural definition of an aggregation function for compositional data. The weighted centroid is proved to fit within this definition and discussed to be linked to a very tangible interpretation. Other functions for the aggregation of compositional data are presented and their fit within the proposed definition is discussed.

Keywords: Aggregation; Compositional data; Beset; Centroid.

1. Introduction

This paper is devoted to the specific problem of aggregating compositional data. A tangible illustration of the problem is given by the act of mixing liquids with known compositions, say coffee and milk, resulting in milk coffee.

*Corresponding author; email: perezfernandez@uniovi.es

If both liquids are mixed in a one-to-one proportion, the composition of milk coffee will be given by the componentwise arithmetic mean (often referred to as the centroid in multivariate statistics) of the compositions of both coffee and milk. In case of a different mixing ratio, a componentwise weighted arithmetic mean is to be considered, the vector of weights being dependent on the considered mixing ratio.

The statistical analysis of compositional data owes its formalization as a scientific discipline to seminal works by Aitchison [1, 2] in the 1980s. Nevertheless, it is admittedly true that a basic statistical analysis of compositional data had already been performed by practitioners as can be derived from Aitchison's introduction in [1] where three examples of preceding articles in geology dealing with compositional data are given [11, 22, 34]. Additional to the field of geology, the study of compositional data naturally arises in almost all fields of application, e.g., the fields of geochemistry [31] and microbiology [18].

Probably due to its natural interpretation, the aggregation of compositional data is dominated by countless papers just considering the componentwise arithmetic mean (see, e.g., [14]). However, one could sporadically find some other componentwise functions, some of them being enumerated by Rock [31]. Unfortunately, the structure of compositional data is way more complex than just a mere restriction to vectors of unit sum. This was already pointed out by Pearson [28] back in 1897 and further explored by Chayes [10] in 1960 when describing some difficulties with the measurement of correlation for compositional data (actually, for vectors of constant sum). For this very reason, most componentwise functions are to be disregarded as encouraged by Aitchison [3]: "A composition provides information only about the relative, not the absolute, values of its components. No component therefore can be considered in isolation. [...] Any statistical analysis must recognize the multivariate nature of the composition and treat it as whole, not as a set of univariate measurements." Interestingly, the componentwise geometric mean has been proposed as a natural aggregation function by Aitchison. It must be admitted though that, since the componentwise geometric mean does not respect the unit-sum constraint, the resulting vector needs to be rescaled by its sum – thus treating the composition as a whole and not as a set of univariate measurements.

As can be seen, there are many issues to address here and the topic has not yet received attention within the aggregation theory community. Thus, in this paper we shall further explore the aggregation of compositional data

and its interaction with the venerated property of monotonicity. Hopefully, this exploration will also be of interest to practitioners, who will benefit from a careful mathematical study of most existing functions for the aggregation of compositional data. The remainder of this paper is structured as follows. Section 2 is devoted to the description of the structure of compositional data (typically referred to as the simplex). In Section 3, we connect the field of aggregation theory with the analysis of compositional data. In particular, we describe what an aggregation function for compositional data should be and introduce a natural example of such an aggregation function: the weighted centroid. In Section 4, we discuss a typical transformation for compositional data vectors and study a specific function related to such transformation. Finally, we discuss how to obtain aggregation functions from convex-hull internal and componentwisely monotone functions in Section 5. We end with some concluding remarks in Section 6.

2. The structure of compositional data: The simplex

2.1. Definition

In this paper, we are dealing with compositional data of the type described in the introductory example to Aitchinson's seminal paper [1] (taken from [34]).

Example 1. The chemical compositions of 32 basalt specimens from the Isle of Skye (Scotland) are given in the form of the proportion of ten major oxides. An example of such type of composition is the following:

SiO_2	Al_2O_3	Fe_2O_3	MgO	CaO	Na_2O	K_2O	TiO_2	P_2O_5	MnO
0.4631	0.1418	0.1232	0.1274	0.0962	0.0251	0.0034	0.0153	0.0016	0.0018

In this paper, we will treat this composition as a vector

$$\mathbf{x} = (0.4631, 0.1418, 0.1232, 0.1274, 0.0962, 0.0251, 0.0034, 0.0153, 0.0016, 0.0018)\,.$$

For any $j \in \{1, \ldots, k\}$, the j-th component of the composition $\mathbf{x}$ will be denoted by $\mathbf{x}(j)$. When a list of n compositions is studied, we will make use of superindices, as in $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$.

It is assumed that the sum of all components should be equal to one (in the above example it is not the case either due to the use of numerical approximation or to the absence of some residually-significant oxides). The set of all possible compositions is called the simplex and will be the object of interest throughout this paper.

The standard simplex is the most prominent type of simplex in which the vertices are the standard unit vectors. Since throughout this paper we are only interested in the standard simplex, we will just refer to *the* simplex meaning the standard simplex. Formally, for fixed $k \in \mathbb{N}$, the simplex is defined as[1]

$$\mathcal{S}_k = \left\{ \mathbf{x} \in [0,1]^k \;\middle|\; \sum_{j=1}^{k} \mathbf{x}(j) = 1 \right\}.$$

The simplex admits a representation in three dimensions whenever $k \leq 4$. In particular, $\mathcal{S}_1$ is a point, $\mathcal{S}_2$ is a line segment, $\mathcal{S}_3$ is a triangle and $\mathcal{S}_4$ is a tetrahedron. In order to illustrate some notions later on in this paper, we will often consider the two-dimensional representation of $\mathcal{S}_3$ on the right-hand side of Figure 1 instead of its more cumbersome three-dimensional representation on the left-hand side of the figure.

As an illustrative example, we refer to Figure 1 for visualizing the coordinates of the compositional data vector $(0.55, 0.32, 0.13)$. In the two-dimensional representation, the (barycentric) coordinates of any point are obtained by projecting the given point to each of the medians of the triangle[2]. Note that barycentric coordinates are also called areal coordinates because of an equivalent representation in which the coordinates represent the areas of the three triangles obtained when drawing a line from the point to each of the three vertices.

2.2. The simplex as a poset

An order relation $\leq$ on a set X is a binary relation that is reflexive ($x \leq x$, for any $x \in X$), antisymmetric ($x \leq y$ and $y \leq x$ imply $x = y$, for any $x, y \in X$) and transitive ($x \leq y$ and $y \leq z$ imply $x \leq z$, for any

[1] It is important to note that we index the summation starting from 1 and ending at k (and not from 0 to $k-1$) and that we do not omit the last component (and thus we do not require the sum of all components to be smaller than one). There do exist some variations in notation in the literature that possibly explain the differences in the understanding of the terms "n-simplex" and "n-dimensional simplex", where n is sometimes understood as k and othertimes as $k-1$. Although there is little doubt that the dimension of the simplex $\mathcal{S}_k$ is $k-1$ if understood as a subset of $\mathbb{R}^k$, it is also true that k easily refers to the number of components of any vector belonging to the simplex. This latter number matters most when dealing with compositional data, thus the reason why this notation has been considered here.

[2] A median of a triangle is a line segment joining a vertex to the midpoint of the opposite side.

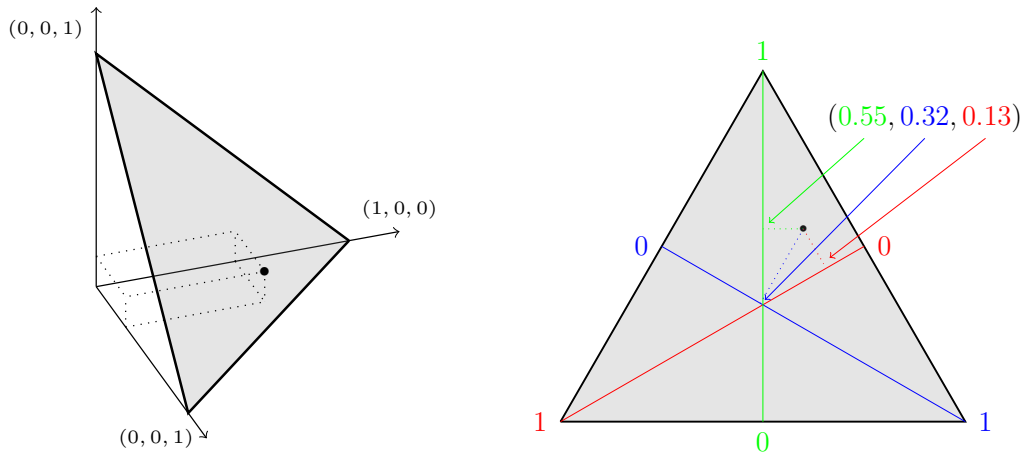

Figure 1: Graphical representation of $\mathcal{S}_3$ in three dimensions (left) and in two dimensions (right). The point $(0.55, 0.32, 0.13)$ is marked within both representations.

$x, y, z \in X$). A set X equipped with an order relation $\leq$ is called a partially ordered set (poset, for short) and is denoted by $(X, \leq)$. A poset $(X, \leq)$ is called bounded if it has both a lower bound (an element $0 \in X$ such that $0 \leq x$, for any $x \in X$) and an upper bound (an element $1 \in X$ such that $x \leq 1$, for any $x \in X$), thus leading to the notation $(X, \leq, 0, 1)$.

By construction, the simplex is incompatible with the product order relation $\leq_k$ on $\mathbb{R}^k$, defined by $\mathbf{x} \leq_k \mathbf{y}$ if $\mathbf{x}(j) \leq \mathbf{y}(j)$ for all $j \in \{1, \ldots, k\}$. In particular, one cannot find two compositional data vectors that are comparable with regard to this order relation. This is due to the fact that if one compositional data vector is smaller than another one in at least one of its components, then it must be greater in at least another component. Formally, there do not exist $\mathbf{x}, \mathbf{y} \in \mathcal{S}_k$ such that $\mathbf{x} \leq_k \mathbf{y}$ for $\mathbf{x} \neq \mathbf{y}$. Thus, $(X, \leq_k)$ is not an interesting poset; it actually is an anti-chain.

Aside from the product order relation, one could think of another natural order relation that fits well with the simplex, the majorization order relation [6], which has been largely considered in the field of economics for measuring the notion of inequality [8, 12, 17, 25]. Formally, a vector $\mathbf{y} \in \mathcal{S}_k$ majorizes another vector $\mathbf{x} \in \mathcal{S}_k$ (denoted $\mathbf{x} \leq_M \mathbf{y}$) if, for any $j \in \{1, \ldots, k\}$, it holds that

$$\sum_{i=1}^{j} \mathbf{x}(\sigma_{\mathbf{x}}(i)) \leq \sum_{i=1}^{j} \mathbf{y}(\sigma_{\mathbf{y}}(i)) \,,$$

where $\sigma_{\mathbf{x}}$ and $\sigma_{\mathbf{y}}$ are the permutations that sort in descending order the components of $\mathbf{x}$ and $\mathbf{y}$, respectively. Note that there exists a natural generalization of the majorization order relation for probability distributions called the Lorenz order relation that can be easily derived from the Lorenz curves [25]. If the set on which the probability distribution is defined is finite, then the majorization order relation and the Lorenz order relation coincide.

Unfortunately, as these latter order relations are conceived for dealing with the notion of inequality in economics, all components are presumed to possibly be rearranged. This means that, for instance, both $(0.25, 0.25, 0.5)$ and $(0.5, 0.25, 0.25)$ carry the same meaning. Thus, $\leq_M$ actually defines a preorder relation[3] on $\mathcal{S}_k$ rather than an order relation. We end by noting that $(\frac{1}{k}, \ldots, \frac{1}{k})$ is the unique lower bound of $(\mathcal{S}_k, \leq_M)$ (i.e., the vector associated with the least possible degree of inequality), whereas all vectors formed by a single one and many zeros are upper bounds of $(\mathcal{S}_k, \leq_M)$ (i.e., the vectors associated with the greatest possible degree of inequality).

Example 2. Consider the chemical composition in Example 1, denoted by $\mathbf{x}^{(1)}$, and the chemical composition below, denoted by $\mathbf{x}^{(2)}$.

SiO_2	Al_2O_3	Fe_2O_3	MgO	CaO	Na_2O	K_2O	TiO_2	P_2O_5	MnO
0.8	0.1	0.09	0.005	0.004	0.0002	0.0002	0.0002	0.0002	0.0002

The proportion of SiO_2 is smaller in $\mathbf{x}^{(1)}$ (0.4631) than in $\mathbf{x}^{(2)}$ (0.8), whereas the proportion of Al_2O_3 is greater in $\mathbf{x}^{(1)}$ (0.1418) than in $\mathbf{x}^{(2)}$ (0.1). Thus, it holds that $\mathbf{x}^{(1)} \not\leq_{10} \mathbf{x}^{(2)}$ and $\mathbf{x}^{(2)} \not\leq_{10} \mathbf{x}^{(1)}$. However, it does hold that $\mathbf{x}^{(1)} \leq_M \mathbf{x}^{(2)}$ since $\mathbf{x}^{(2)}$ is more unequal than $\mathbf{x}^{(1)}$. Note that the composition $\mathbf{x}^{(2)'}$ below in which the porportions of the first two compounds are switched would have led to exactly the same results.

SiO_2	Al_2O_3	Fe_2O_3	MgO	CaO	Na_2O	K_2O	TiO_2	P_2O_5	MnO
0.1	0.8	0.09	0.005	0.004	0.0002	0.0002	0.0002	0.0002	0.0002

2.3. The simplex as a beset

A betweenness relation B on a set X is a ternary relation that satisfies the following properties: symmetry in the end-points ($(x, y, z) \in B$ holds if

[3] A preorder relation is a binary relation that is reflexive and transitive.

and only if $(z, y, x) \in B$, for any $x, y, z \in B$), closure (both $(x, y, z) \in B$ and $(x, z, y) \in B$ hold if and only if $y = z$, for any $x, y, z \in X$) and end-point transitivity ($(o, x, y) \in B$ and $(o, y, z) \in B$ imply that $(o, x, z) \in B$, for any $o, x, y, z \in X$). A set X equipped with a betweenness relation B is called a beset [29] and denoted by (X, B). A non-empty subset S of X is called a set of bounds of a beset (X, B) if none of its elements is in-between two elements that do not belong to the set of bounds (for any $y \in S$ and any $x, z \in X \backslash S$, it holds that $(x, y, z) \notin B$). The triplet (X, B, S) is thus referred to as a bounded beset.

A betweenness relation can be understood as a family of order relations [37]. This connection between order relations and betweenness relations is key for defining an interesting betweenness relation $B_{\mathcal{S}_k}$ on $\mathcal{S}_k$, defined as follows:

$$B_{\mathcal{S}_k} = \left\{ (\mathbf{x}, \mathbf{y}, \mathbf{z}) \in (\mathcal{S}_k)^3 \;\middle|\; \begin{array}{l} (\forall j \in \{1, \ldots, k\}) \\ \left(\min\left(\mathbf{x}(j), \mathbf{z}(j)\right) \leq \mathbf{y}(j) \leq \max\left(\mathbf{x}(j), \mathbf{z}(j)\right)\right) \end{array} \right\}.$$

This betweenness relation is illustrated in Figure 2. Note that, in the right-hand side of the figure, we represent the triplets of the betweenness relation for which one of the two end-points is fixed to be a standard unit vector. Interestingly, whenever an end-point is fixed, the betweenness relation induces a natural order relation representing how close the other two points are to the fixed end-point [37]. In the setting of compositional data, this carries the very appealing meaning of getting closer to being a pure substance (i.e., a vertex of the simplex). This interpretation will be used further on to define a natural property for the aggregation of compositional data (monotonicity).

Example 3. Consider the chemical compositions in Example 2 and the chemical composition below, denoted by $\mathbf{x}^{(3)}$.

SiO_2	Al_2O_3	Fe_2O_3	MgO	CaO	Na_2O	K_2O	TiO_2	P_2O_5	MnO
1	0	0	0	0	0	0	0	0	0

Note that $\mathbf{x}^{(1)}(1) \leq \mathbf{x}^{(2)}(1) \leq \mathbf{x}^{(3)}(1)$ and $\mathbf{x}^{(3)}(j) \leq \mathbf{x}^{(2)}(j) \leq \mathbf{x}^{(1)}(j)$ for any other $j \in \{2, \ldots, k\}$. Thus, it holds that $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}) \in B_{\mathcal{S}_{10}}$, which implies that $\mathbf{x}^{(2)}$ is closer to being pure SiO_2 than $\mathbf{x}^{(3)}$ is. Note that, in case $\mathbf{x}^{(2)'}$ instead of $\mathbf{x}^{(2)}$ would have been considered, it would hold that $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)'}, \mathbf{x}^{(3)}) \notin B_{\mathcal{S}_{10}}$.

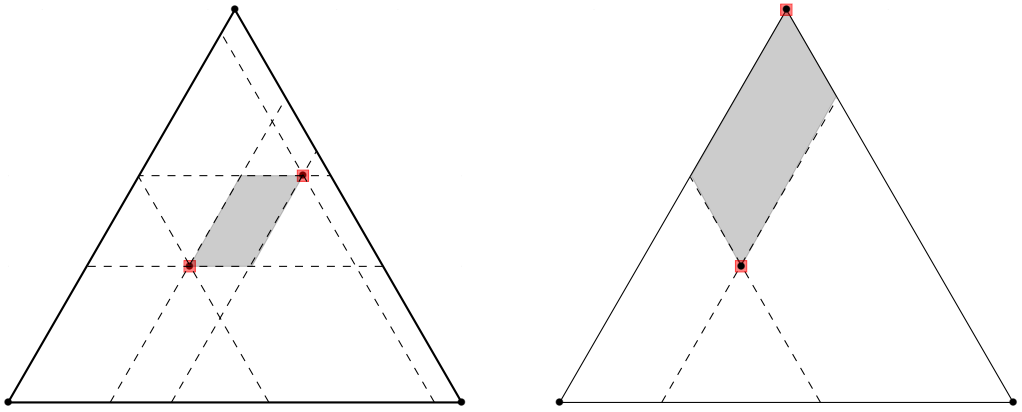

Figure 2: Illustration of the compositional data vectors (highlighted in grey) that are strictly in-between the two compositional data vectors that are highlighted in red according to the betweenness relation $B_{\mathcal{S}_3}$.

It is easy to verify that the standard basis of $\mathbb{R}^k$, denoted by $\mathbb{E} = \{\mathbf{e}^{(j)}\}_{j=1}^{k}$ and formed by all standard unit vectors formed by a one and many zeros, i.e., of the form $(0, \ldots, 0, 1, 0, \ldots, 0)$, is a set of bounds of the beset $(\mathcal{S}_k, B_{\mathcal{S}_k})$. Therefore, $(\mathcal{S}_k, B_{\mathcal{S}_k}, \mathbb{E})$ is a bounded beset.

3. Aggregation functions for compositional data

3.1. Classical order-based aggregation theory for compositional data

The study of aggregation processes is the core topic of aggregation theory and is formalized by the notion of an aggregation function [9, 19]. Until recently, it has been typically assumed that aggregation is a process that operates on a bounded poset, typically a compact real interval.

Definition 4. Consider a bounded poset $(X, \leq, 0, 1)$ and $n \in \mathbb{N}$. A function $A : X^n \to X$ is called an (n-ary) aggregation function on $(X, \leq, 0, 1)$ if

(i) it satisfies the boundary conditions, i.e., $A(0, \ldots, 0) = 0$ and $A(1, \ldots, 1) = 1$;

(ii) it is increasing, i.e., for any $(x^{(1)}, \ldots, x^{(n)}), (y^{(1)}, \ldots, y^{(n)}) \in X^n$ such that $x^{(i)} \leq y^{(i)}$ for any $i \in \{1, \ldots, n\}$, it holds that $A(x^{(1)}, \ldots, x^{(n)}) \leq A(y^{(1)}, \ldots, y^{(n)})$.

The above definition has been proven to fit many different scenarios from the most basic ones such as the aggregation of real numbers to more involved ones such as the aggregation of labels of an ordinal linguistic scale [7] and the aggregation of elements of a bounded lattice [23].

However, this definition clearly does not fit the aggregation of compositional data when considering the poset $(\mathcal{S}_k, \leq_M)$. More specifically, let $k = 3$, $(1, 0, 0)$ and $(0, 0, 1)$ obviously carry a very different meaning (although they are equivalent in terms of $\leq_M$). Moreover, think of the most natural function $A : (\mathcal{S}_k)^3 \to \mathcal{S}_k$ for aggregating compositional data vectors by computing the centroid, i.e., the arithmetic mean in each of their components:

$$A(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}) = \frac{1}{3}\mathbf{x}^{(1)} + \frac{1}{3}\mathbf{x}^{(2)} + \frac{1}{3}\mathbf{x}^{(3)}\,.$$

This function cannot be considered an aggregation function on $(\mathcal{S}_k, \leq_M)$ since it is not increasing:

$$\begin{aligned}
(1,0,0) &\leq_M (1,0,0)\,,\\
(1,0,0) &\leq_M (0,1,0)\,,\\
(1,0,0) &\leq_M (0,0,1)\,,\\
A\Big((1,0,0),(1,0,0),(1,0,0)\Big) = (1,0,0) &\not\leq_M (\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3}) = A\Big((1,0,0),(0,1,0),(0,0,1)\Big)\,.
\end{aligned}$$

3.2. Betweenness-based aggregation theory

In a recent work by some of the present authors [29], a natural extension of the notion of an aggregation function to besets was proposed.

Definition 5. Consider a bounded beset (X, B, S) and $n \in \mathbb{N}$. A function $A : X^n \to X$ is called an (n-ary) aggregation function on (X, B, S) if

(i) it satisfies the boundary conditions, i.e., $A(o, \ldots, o) = o$, for any $o \in S$;

(ii) it is monotone, i.e., for any $o \in S$ and any $(x^{(1)}, \ldots, x^{(n)}), (y^{(1)}, \ldots, y^{(n)}) \in X^n$, the fact that $\Big(o, x^{(i)}, y^{(i)}\Big) \in B$ for any $i \in \{1, \ldots, n\}$ implies that $\Big(A(o, \ldots, o), A(x^{(1)}, \ldots, x^{(n)}), A(y^{(1)}, \ldots, y^{(n)})\Big) \in B$.[4]

[4]Note that, due to the boundary conditions, monotonicity can be rewritten as, for any $o \in S$ and any $(x^{(1)}, \ldots, x^{(n)}), (y^{(1)}, \ldots, y^{(n)}) \in X^n$, the fact that $\big(o, x^{(i)}, y^{(i)}\big) \in B$ for any $i \in \{1, \ldots, n\}$ implies that $\big(o, A(x^{(1)}, \ldots, x^{(n)}), A(y^{(1)}, \ldots, y^{(n)})\big) \in B$.

If one considers the bounded beset $(\mathcal{S}_k, B_{\mathcal{S}_k}, \mathbb{E})$, the requirements above can be intuitively described as: (i) the aggregation of n times the same pure substance should result in the same pure substance; (ii) the closer all the compositions to be aggregated are to being the same pure substance, the closer the result of the aggregation should be to being this pure substance.

Probably the most natural interpretation of aggregation for compositional data arises from mixing liquids with known compositions. We can then understand the composition of the resulting liquid as the aggregation of n compositional data vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)} \in \mathcal{S}_k$. This aggregation results in a new compositional data vector and can be formalized using the function $\mathbf{C}_{\mathbf{w}} : (\mathcal{S}_k)^n \to \mathcal{S}_k$ defined by

$$\mathbf{C}_{\mathbf{w}}(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})(j) = \sum_{i=1}^{n} w_i \mathbf{x}^{(i)}(j)\,, \tag{1}$$

for any $j \in \{1, \ldots, k\}$, where $\mathbf{w} = (w_1, \ldots, w_n)$ is a suitable weighing vector (in the case of the liquids, reflecting the mixing ratio associated with each of the different liquids in the mixture). The above function is well known in the field of multivariate statistics and is referred to as the weighted centroid. The special case in which $\mathbf{w} = (\frac{1}{n}, \ldots, \frac{1}{n})$ is called the centroid.

Example 6. As an illustrative example, consider the compositional data vector $\mathbf{x}^{(1)} = (0.55, 0.32, 0.13)$ used in Figure 1 and the pure substance $\mathbf{x}^{(2)} = (1, 0, 0)$ representing the composition of two liquids in terms of three compounds. In case both liquids are mixed in a one-to-one ratio, the resulting mixture will have the following composition:

$$\mathbf{C}_{(\frac{1}{2}, \frac{1}{2})}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = (0.775, 0.16, 0.065)\,.$$

In case different quantities of each of the liquids are mixed, a weighted centroid rather than the centroid is to be used. For instance, consider mixing both liquids such that the quantity of $\mathbf{x}^{(1)}$ is the quadruple of the quantity of $\mathbf{x}^{(2)}$. The associated weighing vector will thus be $(\frac{4}{5}, \frac{1}{5})$, as follows:

$$\mathbf{C}_{(\frac{4}{5}, \frac{1}{5})}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = (0.64, 0.256, 0.104)\,.$$

The weighted centroid is easily proved to be an aggregation function in the sense of Definition 5.

Proposition 7. *The function $\mathbf{C}_{\mathbf{w}} : (\mathcal{S}_k)^n \to \mathcal{S}_k$ defined by Eq. (1) is an aggregation function on $(\mathcal{S}_k, B_{\mathcal{S}_k}, \mathbb{E})$.*

Proof. We first prove that $\mathbf{C}_{\mathbf{w}}$ satisfies the boundary conditions. Consider any $\mathbf{e}^{(\ell)} \in \mathbb{E}$. Since $\mathbf{e}^{(\ell)}(\ell) = 1$, it follows that

$$\mathbf{C}_{\mathbf{w}}(\mathbf{e}^{(\ell)}, \dots, \mathbf{e}^{(\ell)})(\ell) = \sum_{i=1}^{n} w_i \, 1 = 1 \, .$$

Similarly, for any $j \neq \ell$, since $\mathbf{e}^{(\ell)}(j) = 0$, it holds that

$$\mathbf{C}_{\mathbf{w}}(\mathbf{e}^{(\ell)}, \dots, \mathbf{e}^{(\ell)})(j) = \sum_{i=1}^{n} w_i \, 0 = 0 \, .$$

We conclude that

$$\mathbf{C}_{\mathbf{w}}(\mathbf{e}^{(\ell)}, \dots, \mathbf{e}^{(\ell)}) = \mathbf{e}^{(\ell)} \, ,$$

and, thus, $\mathbf{C}_{\mathbf{w}}$ satisfies the boundary conditions.

We now prove that $\mathbf{C}_{\mathbf{w}}$ is monotone. Consider any $\mathbf{e}^{(\ell)} \in \mathbb{E}$ and any $(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}), (\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)}) \in (\mathcal{S}_k)^n$ such that $\left(\mathbf{e}^{(\ell)}, \mathbf{x}^{(i)}, \mathbf{y}^{(i)}\right) \in B_{\mathcal{S}_k}$, for any $i \in \{1, \dots, n\}$. Since $\mathbf{e}^{(\ell)}(\ell) = 1$, it follows that $\mathbf{y}^{(i)}(\ell) \leq \mathbf{x}^{(i)}(\ell) \leq \mathbf{e}^{(\ell)}(\ell) = 1$. Similarly, since $\mathbf{e}^{(\ell)}(j) = 0$ for any $j \neq \ell$, it holds that $0 = \mathbf{e}^{(\ell)}(j) \leq \mathbf{x}^{(i)}(j) \leq \mathbf{y}^{(i)}(j)$ for any $j \neq \ell$. Thus, since the weighted centroid is defined componentwisely and due to the increasingness of the weighted arithmetic mean as a univariate function, it holds that

$$\mathbf{C}_{\mathbf{w}}(\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)})(\ell) \leq \mathbf{C}_{\mathbf{w}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})(\ell) \leq \mathbf{C}_{\mathbf{w}}(\mathbf{e}^{(\ell)}, \dots, \mathbf{e}^{(\ell)})(\ell) \, ,$$

and that, for any $j \neq \ell$,

$$\mathbf{C}_{\mathbf{w}}(\mathbf{e}^{(\ell)}, \dots, \mathbf{e}^{(\ell)})(j) \leq \mathbf{C}_{\mathbf{w}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})(j) \leq \mathbf{C}_{\mathbf{w}}(\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)})(j) \, .$$

We conclude that

$$\left(\mathbf{C}_{\mathbf{w}}(\mathbf{e}^{(\ell)}, \dots, \mathbf{e}^{(\ell)}), \mathbf{C}_{\mathbf{w}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}), \mathbf{C}_{\mathbf{w}}(\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)})\right) \in B_{\mathcal{S}_k} \, ,$$

and, thus, $\mathbf{C}_{\mathbf{w}}$ is monotone. □

4. Interaction with a popular transformation for compositional data

When dealing with compositional data, it is quite common to apply some transformations to the elements of the simplex in order to (1) deal with an easier structure, and (2) properly study correlations between the components.

There exist several such transformations, but here, we will only pay attention to the most prominent and simplest one: the additive logistic transformation [4]. As mentioned by Aitchison in [2] (page 113), this transformation was already "in use in other areas of statistical activity" at the time of his seminal book and can be traced back to the problem of transferring patterns of variability from the real line to the positive real line [26].

Let $\mathcal{S}_k^+$ be the positive simplex, defined as

$$\mathcal{S}_k^+ = \left\{ \mathbf{x} \in]0,1[^k \;\middle|\; \sum_{j=1}^{k} \mathbf{x}(j) = 1 \right\}.$$

Given a compositional data vector $\mathbf{x} \in \mathcal{S}_k^+$, one may obtain an element $\mathbf{y} \in \mathbb{R}^{k-1}$ (assuming $k > 1$) by applying the transformation $\phi : \mathcal{S}_k^+ \to \mathbb{R}^{k-1}$ defined as follows:

$$\phi(\mathbf{x})(j) = \mathbf{y}(j) = \ln\left(\frac{\mathbf{x}(j)}{\mathbf{x}(k)}\right), \quad \text{for any } j \in \{1,\ldots,k-1\}.$$

Similarly, given an element $\mathbf{y} \in \mathbb{R}^{k-1}$, one may obtain a compositional data vector $\mathbf{x} \in \mathcal{S}_k^+$ by applying the inverse transformation $\phi^{-1} : \mathbb{R}^{k-1} \to \mathcal{S}_k^+$ defined as follows:

$$\phi^{-1}(\mathbf{y})(j) = \mathbf{x}(j) = \frac{e^{\mathbf{y}(j)}}{e^{\mathbf{y}(1)} + \ldots + e^{\mathbf{y}(k-1)} + 1}, \quad \text{for any } j \in \{1,\ldots,k-1\},$$
$$\phi^{-1}(\mathbf{y})(k) = \mathbf{x}(k) = \frac{1}{e^{\mathbf{y}(1)} + \ldots + e^{\mathbf{y}(k-1)} + 1}.$$

Note that the transformation ϕ is not well-defined on the boundaries of the simplex, thus the reason why the positive simplex is considered here.

The aim of this section is to explore the following statement of Aitchison [4]: "Transform each composition to it logratio vector, after reformulating your problem about compositions in terms of the corresponding logratio vectors, then apply the appropriate, standard multivariate procedures to the logratio vectors."

In particular, we consider the function that transforms the compositional data vectors in logratio vectors, next applies the centroid, and, finally, returns to compositional data vectors. Such a function $A_\phi : (\mathcal{S}_k^+)^n \to \mathcal{S}_k^+$ is defined as follows:

$$A_\phi\left(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}\right) = \phi^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\phi\left(\mathbf{x}^{(i)}\right)\right). \tag{2}$$

The resemblance of the above expression with a classical (univariate) quasi-arithmetic mean is obvious [20]. Actually, as explained by Aitchison [3], linear algebra shows that this function can be equivalently expressed as the result of applying the geometric mean componentwisely and then dividing by the total sum of the vector of geometric means assuring the function to stay within the simplex.

Interestingly, as can be seen in Figure 3 (as in Fig. 1 in [3]), the function defined in Eq. (2) might yield a better representative of a dataset of compositional data vectors than the centroid in case of curved datasets (which apparently are not untypical datasets in geochemical studies).

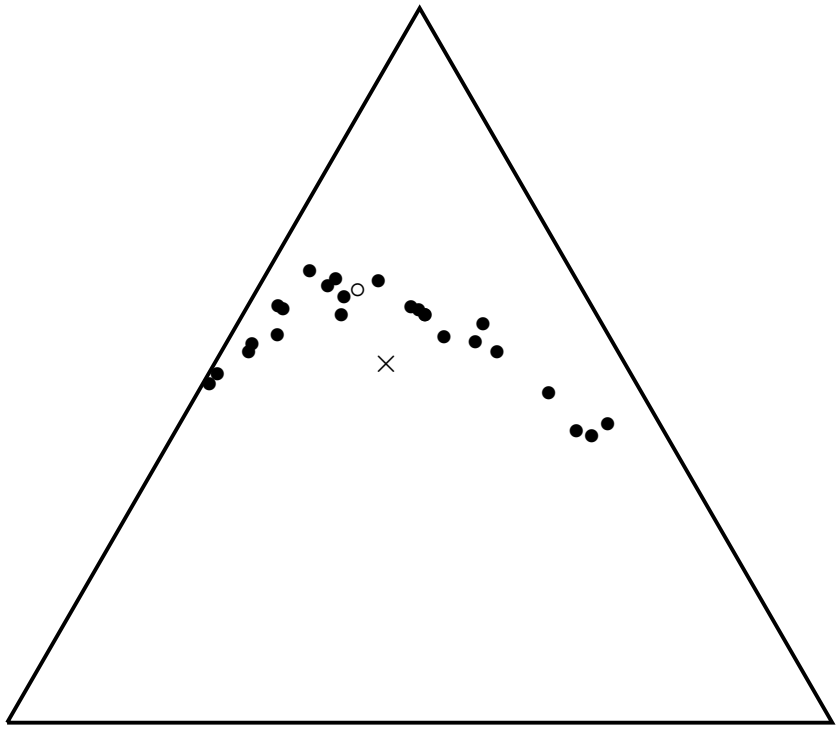

Figure 3: Figure constructed from the same data as Fig. 1 in [3]. Illustration of a "not untypical data set of 3-part compositions". The point marked with the symbol $\times$ represents the centroid, whereas the point marked with the symbol $\circ$ represents the result of applying the function in Eq. (2).

Unfortunately, it turns out that A_ϕ is not monotone (see Footnote 4), as can be understood from the following example.

Example 8. Consider $\mathbf{x}^{(1)} = (0.4, 0.15, 0.45)$ and $\mathbf{x}^{(2)} = (0.05, 0.5, 0.45)$. It holds that $A_\phi\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\right) = (0.1634, 0.3165, 0.5201)$. Consider now $\mathbf{x}^{(1)'} = (0.41, 0.145, 0.445)$. We can easily verify that $\left(\mathbf{x}^{(1)}, \mathbf{x}^{(1)'}, (1,0,0)\right) \in B_{\mathcal{S}_3}$ and it obviously holds that $\left(\mathbf{x}^{(2)}, \mathbf{x}^{(2)}, (1,0,0)\right) \in B_{\mathcal{S}_3}$. However, since $A_\phi\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)'}\right) = (0.1665, 0.3131, 0.5204)$, we have that[5]

$$\left(A_\phi\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\right), A_\phi\left(\mathbf{x}^{(1)'}, \mathbf{x}^{(2)}\right), (1,0,0)\right) \notin B_{\mathcal{S}_3}.$$

We end the section by noting that, since A_ϕ is not monotone on the positive simplex, there is no point in trying to extend the function A_ϕ to the simplex in order to find an aggregation function on $(\mathcal{S}_k, B_{\mathcal{S}_k}, \mathbb{E})$.

5. Aggregation functions for multidimensional data

In this section, we discuss three classical properties of aggregation functions for multidimensional data. Firstly, convex-hull internality assures that the aggregate remains within the simplex. Secondly, componentwise monotonicity – in combination with convex-hull internality – assures that we have an aggregation function in the sense of Definition 5. Thirdly, we point out that the addition of orthogonal equivariance restricts the family of aggregation functions for multidimensional data to the weighted centroid.

5.1. On convex-hull internality

Researchers in the field of multivariate statistics have extensively studied how to generalize classical aggregation functions to higher dimensions. For instance, as has already been mentioned, a weighted arithmetic mean can be componentwisely extended to higher dimensions, thus defining a weighted centroid [16]. Similarly, one could also think of extending the median to higher dimensions componentwisely. Interestingly, this componentwise median is not the only extension of the median to higher dimensions that has called the attention of the scientific community [32]. Just to name a few, one could find other examples such as the spatial median [36], Tukey's halfspace median [35], the convex hull peeling median [13], Oja's simplex median [27], the simplicial depth median [24] and the orthomedian [21].

[5]Note that A_ϕ is not explicitly defined for vertices of the simplex. Due to the idempotence of A_ϕ on the positive simplex, it seems natural to extend this property to the simplex and define $A_\phi((1,0,0),(1,0,0)) = (1,0,0)$.

An interesting property for functions aiming at combining several points of $\mathbb{R}^k$ into a single one is that of convex-hull internality [15]. This property requires the result of aggregating the points $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)} \in \mathbb{R}^k$ to belong to their convex hull, i.e., the set

$$\mathrm{CH}\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right) = \left\{ \mathbf{x} = \sum_{i=1}^{n} \lambda_i \mathbf{x}^{(i)} \in \mathbb{R}^k \mid (\lambda_1, \ldots, \lambda_n) \in \mathcal{S}_n \right\}.$$

Definition 9. A function $A : (\mathbb{R}^k)^n \to \mathbb{R}^k$ is called convex-hull internal if, for any $\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right) \in (\mathbb{R}^k)^n$, it holds that $A\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right) \in \mathrm{CH}\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right)$.

Remark 10. Any convex-hull internal function $A : (\mathbb{R}^k)^n \to \mathbb{R}^k$ is idempotent, i.e., $A(\mathbf{x}, \ldots, \mathbf{x}) = \mathbf{x}$ for any $\mathbf{x} \in \mathbb{R}^k$.

Typical examples of convex-hull internal functions are the aforementioned weighted centroid and the spatial median. Some other generalizations of the univariate median are also known to be convex-hull internal (e.g., the convex hull peeling median [13]), whereas some other ones are known to fail this intuitive property (e.g., the componentwise median in case $k \geq 3$).

The property of convex-hull internality is of relevance to the aggregation of compositional data because, since the simplex is a convex set, any convex combination of vectors of unit sum will still be a vector of unit sum. More precisely, all convex-hull internal functions for the aggregation of multidimensional data can be used for the aggregation of compositional data although – unlike with the weighted centroid – an intuitive physical meaning could be lacking.

Proposition 11. *Consider a convex-hull internal function $A : (\mathbb{R}^k)^n \to \mathbb{R}^k$. For any $\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right) \in (\mathcal{S}_k)^n$, it holds that $A\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right) \in \mathcal{S}_k$.*

Proof. Consider $\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right) \in (\mathcal{S}_k)^n$. Since $A : (\mathbb{R}^k)^n \to \mathbb{R}^k$ is convex hull internal, it follows that $A\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right)$ belongs to the convex hull of $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$. Since all $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ belong to $\mathcal{S}_k$, the convex hull of $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ needs to be a subset of the convex hull of $\mathcal{S}_k$. Finally, since $\mathcal{S}_k$ is already a convex set, it coincides with its convex hull and, thus, we conclude that $A\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right) \in \mathcal{S}_k$. □

5.2. On componentwise monotonicity

Another interesting property is that of componentwise monotonicity [15, 30]. This property assures that, whenever the values in one of the components are increased, the aggregated value for this same component increases.

Definition 12. A function $A : (\mathbb{R}^k)^n \to \mathbb{R}^k$ is called componentwisely monotone if, for any $j \in \{1, \ldots, k\}$ and any $\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right), \left(\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)}\right) \in (\mathbb{R}^k)^n$ satisfying that $x^{(i)}(j) \leq y^{(i)}(j)$ for any $i \in \{1, \ldots, n\}$, it holds that $A\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right)(j) \leq A\left(\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)}\right)(j)$.

Interestingly, componentwise monotonicity of a function $A : (\mathbb{R}^k)^n \to \mathbb{R}^k$ is equivalent to A being a componentwise extension of some k monotone functions $A_1, \ldots, A_k : \mathbb{R}^n \to \mathbb{R}^n$ (see Proposition 15 in [15]).

It is easy to prove that any convex-hull internal and componentwisely monotone function is an aggregation function on $(\mathcal{S}_k, B_{\mathcal{S}_k}, \mathbb{E})$.

Proposition 13. *If a function $A : (\mathbb{R}^k)^n \to \mathbb{R}^k$ is convex-hull internal and componentwisely monotone, then it is an aggregation function on $(\mathcal{S}_k, B_{\mathcal{S}_k}, \mathbb{E})$.*

Proof. From Proposition 11, we conclude that A is well defined for aggregating compositional data. The boundary conditions follow directly from the convex-hull internality of A (since convex-hull internality implies idempotence). We now prove that A is monotone. Consider any $\mathbf{e}^{(\ell)} \in \mathbb{E}$ and any $(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}), (\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)}) \in (\mathcal{S}_k)^n$ such that $\left(\mathbf{e}^{(\ell)}, \mathbf{x}^{(i)}, \mathbf{y}^{(i)}\right) \in B_{\mathcal{S}_k}$, for any $i \in \{1, \ldots, n\}$. Since $\mathbf{e}^{(\ell)}(\ell) = 1$, it follows that $\mathbf{y}^{(i)}(\ell) \leq \mathbf{x}^{(i)}(\ell) \leq \mathbf{e}^{(\ell)}(\ell) = 1$. Similarly, since $\mathbf{e}^{(\ell)}(j) = 0$ for any $j \neq \ell$, it holds that $0 = \mathbf{e}^{(\ell)}(j) \leq \mathbf{x}^{(i)}(j) \leq \mathbf{y}^{(i)}(j)$ for any $j \neq \ell$. Due to the componentwise monotonicity of A, it holds that

$$A(\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)})(\ell) \leq A(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})(\ell) \leq A(\mathbf{e}^{(\ell)}, \ldots, \mathbf{e}^{(\ell)})(\ell)\,,$$

and that, for any $j \neq \ell$,

$$A(\mathbf{e}^{(\ell)}, \ldots, \mathbf{e}^{(\ell)})(j) \leq A(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})(j) \leq A(\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)})(j)\,.$$

We conclude that

$$\left(A(\mathbf{e}^{(\ell)}, \ldots, \mathbf{e}^{(\ell)}), A(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}), A(\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)})\right) \in B_{\mathcal{S}_k}\,,$$

and, thus, A is monotone. □

Note that Proposition 7 could well have come as a corollary of the proposition above. Weighted centroid aside, the componentwise median is a prominent example of convex-hull internal and componentwisely monotone function in case $k \leq 2$. Unfortunately, the componentwise median is easily proven not to be convex-hull internal if $k \geq 3$.

5.3. On orthogonal equivariance

The field of multivariate statistics is overpopulated with functions $A : (\mathbb{R}^k)^n \to \mathbb{R}^k$ that are convex-hull internal and orthogonally equivariant[6]. Examples of such functions are the centroid, most generalizations of the median [32] (e.g., the spatial median [36], Tukey's halfspace median [35], the convex hull peeling median [13], Oja's simplex median [27], the simplicial depth median [24] and the orthomedian [21]) and the Euclidean center [33]. We end the section by noting that only a weighted centroid among such functions is an aggregation function for multidimensional data. This result is not surprising since orthogonal transformations are not natural when dealing with compositional data, where the coordinate system is associated with the different components.

Proposition 14. *A convex-hull internal and orthogonally equivariant function $A : (\mathbb{R}^k)^n \to \mathbb{R}^k$ is an aggregation function on $(\mathcal{S}_k, B_{\mathcal{S}_k}, \mathbb{E})$ if and only if it is a weighted centroid.*

Proof. It was proven in Proposition 7 that a weighted centroid is an aggregation function on $(\mathcal{S}_k, B_{\mathcal{S}_k}, \mathbb{E})$. We now prove the converse implication. First, one should note that monotonicity for compositional data in the sense of Definition 5 is equivalent to $\leq_k$-nondecreasingness as presented in [15], but for a different orthant. From the orthogonal equivariance of A, it follows that A is $\leq_k$-nondecreasing. From Proposition 29 in [15] (and since orthogonal equivariance implies equivariance to reflections), it follows that A is componentwisely monotone (referred to as componentwisely nondecreasing in [15]). Thus, it holds that A is convex-hull internal, orthogonally equivariant and componentwisely monotone. From [16], we conclude that A necessarily is a weighted centroid. □

[6] A function $A : (\mathbb{R}^k)^n \to \mathbb{R}^k$ is called orthogonally equivariant if $A\left(\mathbf{O}\mathbf{x}^{(1)}, \ldots, \mathbf{O}\mathbf{x}^{(n)}\right) = \mathbf{O}A\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right)$ for every orthogonal matrix $\mathbf{O}$ (i.e., for every matrix such that $\mathbf{O}^{-1} = \mathbf{O}^T$).

6. Conclusions

The field of aggregation theory builds upon the notion of a poset, however, the set of compositional data vectors has been shown not to fit within this order-based framework. Following the natural generalization of aggregation functions to besets proposed in [29], we have discussed an intuitive definition for an aggregation function for compositional data.

The weighted centroid, which is inherently linked to the process of mixing, has been proved to be an example of such aggregation function. Unfortunately, a prominent function based on the geometric mean suggested by Aitchison for describing the location of a set of compositional data vectors has been shown not to fit the proposed definition of an aggregation function.

Finally, we have discussed how functions from the field of multivariate statistics could be used for aggregating compositional data. We have presented a sufficient condition (Proposition 13) for a function to be an aggregation function for compositional data. Unfortunately, in case $k \geq 3$, we have not identified any function satisfying this sufficient condition other than the weighted centroid. For this very reason, we conjecture that the weighted centroid could possibly be the unique convex-hull internal and componentwisely monotone function in case $k \geq 3$. This would not come as a surprise if one bears in mind that the weighted centroid is the unique such function that additionally is orthogonally equivariant [16].

We end by noting that the results presented here are given for the specific case of compositional data and for the particular betweenness relation that we understand is the most meaningful when dealing with this kind of data. However, when we are dealing with other types of data that can also be represented as a simplex (see, e.g., [5]), other betweenness relations might be of interest. For instance, for the case of probability distributions defined on a finite chain, it might be more meaningful to consider a betweenness relation induced by the stochastic dominance order.

Acknowledgements. Raúl Pérez-Fernández acknowledges the support of the Research Foundation of Flanders (FWO17/PDO/160) and the Spanish MINECO (TIN2017-87600-P). Marek Gagolewski acknowledges the support of the Australian Research Council Discovery Project ARC DP210100227.

References

[1] Aitchison, J., 1982. The statistical analysis of compositional data (with discussion). Journal of the Royal Statistical Society B44, 139–177.

[2] Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman and Hall, London.

[3] Aitchison, J., 1989. Measures of location of compositional data sets. Mathematical Geology 21, 787–790.

[4] Aitchison, J., 1994. Principles of compositional data analysis, in: Lecture Notes - Monograph Series. volume 24 of *Multivariate Analysis and Its Applications*, pp. 73–81.

[5] Angelova, M., Beliakov, G., Shelyag, S., Zhu, Y., 2020. Density estimates on the unit simplex and calculation of the mode of a sample. International Journal of Intelligent Systems 35, 850–868.

[6] Arnold, B.C., 1987. Majorization and the Lorenz Order: A Brief Introduction. Springer-Verlag, Berlin.

[7] Balinski, M., Laraki, R., 2007. A theory of measuring, electing and ranking. PNAS 104, 8720–8725.

[8] Beliakov, G., Gagolewski, M., James, S., 2016. Penalty-based and other representations of economic inequality. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 24, 1–23.

[9] Beliakov, G., Pradera, A., Calvo, T., 2007. Aggregation Functions: A Guide for Practitioners. volume 221 of *Studies in Fuzziness and Soft Computing*. Springer, Berlin, Heidelberg.

[10] Chayes, F., 1960. On correlation between variables of constant sum. Journal of Geophysical Research 65, 4185–4193.

[11] Coakley, J.P., Rust, B.R., 1968. Sedimentation in an Arctic lake. Journal of Sedimentary Petrology 38, 1290–1300.

[12] Dalton, H., 1920. The measurement of the inequality of incomes. The Economic Journal 30, 348–361.

[13] Eddy, W.F., 1982. Convex hull peeling, in: Proceedings of the COMPSTAT Symposium, Toulouse. pp. 42–47.

[14] Edjabou, M.E., Martín-Fernández, J.A., Scheutz, C., Astrup, T.F., 2017. Statistical analysis of solid waste composition data: Arithmetic mean, standard deviation and correlation coefficients. Waste Management 69, 13–23.

[15] Gagolewski, M., 2017. Penalty-based aggregation of multidimensional data. Fuzzy Sets and Systems 325, 4–20.

[16] Gagolewski, M., Pérez-Fernández, R., De Baets, B., 2020. An inherent difficulty in the aggregation of multidimensional data. IEEE Transactions on Fuzzy Systems 28, 602–606.

[17] Gini, C., 1921. Measurement of inequality of incomes. The Economic Journal 31, 124–126.

[18] Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., Egozcue, J.J., 2017. Microbiome datasets are compositional: And this is not optional. Frontiers in Microbiology 8, 1–6.

[19] Grabisch, M., Marichal, J.L., Mesiar, R., Pap, E., 2009. Aggregation Functions. Cambridge University Press, Cambridge.

[20] Grabisch, M., Marichal, J.L., Mesiar, R., Pap, E., 2011. Aggregation functions: Means. Information Sciences 181, 1–22.

[21] Grübel, R., 1996. Orthogonalization of multivariate location estimators: The orthomedian. Annals of Statistics 24, 1457–1473.

[22] Kaiser, R.F., 1962. Composition and origin of glacial till, Mexico and Kasoag quadrangles, New York. Journal of Sedimentary Research 32, 502–513.

[23] Karaçal, F., Mesiar, R., 2017. Aggregation functions on bounded lattices. International Journal of General Systems 46, 37–51.

[24] Liu, R.Y., 1990. On a notion of data depth based on random simplices. The Annals of Statistics 18, 405–414.

[25] Lorenz, M.O., 1905. Methods for measuring the concentration of wealth. Publications of the American Statistical Association 9, 209–219.

[26] McAlister, D., 1879. The law of the geometric mean. Proceedings of the Royal Society of London 29, 367–376.

[27] Oja, H., 1983. Descriptive statistics for multivariate distributions. Statistics and Probability Letters 1, 327–332.

[28] 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. Proceeding of the Royal Society of London 60, 489–498.

[29] Pérez-Fernández, R., De Baets, B., 2020. Aggregation theory revisited. IEEE Transactions on Fuzzy Systems, in press. DOI: 10.1109/TFUZZ.2020.2965904 .

[30] Pérez-Fernández, R., De Baets, B., Gagolewski, M., 2019. A taxonomy of monotonicity properties for the aggregation of multidimensional data. Information Fusion 52, 322–334.

[31] Rock, N.M.S., 1988. Summary statistics in geochemistry: A study of the performance of robust estimates. Mathematical Geology 20, 243–275.

[32] Small, C.G., 1990. A survey of multidimensional medians. International Statistical Review 58, 263–277.

[33] Sylvester, J.J., 1857. A question in the geometry of situation. The Quarterly Journal of Mathematics 1, 79.

[34] Thompson, R.N., Esson, J., Duncan, A.C., 1972. Major element chemical variation in the eocene lavas of the Isle of Skye, Scotland. Journal of Petrology 13, 219–253.

[35] Tukey, J.W., 1975. Mathematics and the picturing of data, in: Proceedings of the International Congress of Mathematicians, Vancouver. pp. 523–531.

[36] Weber, A., 1909. Ueber den Standort der Industrien. Mohr Siebeck Verlag, Tübingen.

[37] Zhang, H.P., Pérez-Fernández, R., De Baets, B., 2019. Topologies induced by the representation of a betweenness relation as a family of order relations. Topology and its Applications 258, 100–114.

Please cite this paper as: R. Pérez-Fernández, M. Gagolewski, B. De Baets,
On the aggregation of compositional data, Information Fusion 73, 103–110, 2021, doi:10.1016/j.inffus.2021.02.021