SORT 47 (2) July-December 2023, 229-244 DOI: 10.57645/20.8080.02.7

Subcompositional coherence and a novel proportionality index of parts

Juan José Egozcue[1] and Vera Pawlowsky-Glahn[2]

Abstract

Research in compositional data analysis was motivated by spurious (Pearson) correlation. Spurious results are due to semantic incoherence, but the question of ways to relate parts in a statistically consistent way remains open. To solve this problem we first define a coherent system of functions with respect to a subcomposition and analyze the space of parts. This leads to understanding why measures like covariance and correlation depend on the subcomposition considered, while measures like the distance between parts are independent of the same. It allows the definition of a novel index of proportionality between parts.

MSC: *62-07.*

Keywords: *Compositional data analysis, Aitchison geometry, simplex, compositional parts, proportionality, dominance, correlation.*

1. Introduction

Research in compositional data analysis (CoDA) was motivated by the so-called spurious (Pearson) correlation (Pearson, 1897; Chayes, 1971). It appears as correlations changing when considering the same variables, or parts, as parts of different compositions represented in closed form. For instance, in the example below, the Pearson correlation coefficient between milk and eggs changes from -1 when they are considered as a two part composition, to more than 0.6 when represented as proportions of a subcomposition including also sugar, fat, juices and non-alcoholic drinks (see also

[1] Dep. de Ingeniería Civil y Ambiental, Universidad Politécnica de Cataluña, Barcelona, Spain; juan.jose.egozcue@upc.edu

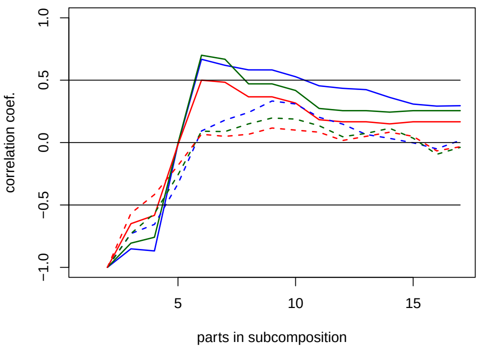
[2] Dep. Informática, Matemática Aplicada y Estadística, Universidad de Girona, Spain; vera.pawlowsky@udg.edu

Received: November 2022

Accepted: May 2023

Fig. 1). This led to the principle of *subcompositional coherence* as a requirement for a consistent methodology (Aitchison, 1986, 1992). Nowadays, we know that most measures between two compositional parts are in this sense also spurious, e.g. Spearman and Kendall correlations, or copulas (Ortego and Egozcue, 2013; Egozcue and Pawlowsky-Glahn, 2019; Pawlowsky-Glahn and Egozcue, 2022). We also know that those spurious results are actually due to a misnomer of the parts in different compositions, as they really involve all the parts considered through the operation of closure (Pawlowsky-Glahn and Egozcue, 2022). There is in fact a semantic incoherence inherent to the common practice of assigning identical labels to different functions.

From the initial developments of CoDA (Aitchison, 1986) up to now there were several reformulations of the principle of subcompositional coherence, (for instance, Aitchison, 1992; Aitchison and Egozcue, 2005; Egozcue, 2009; Boogaart and Tolosana-Delgado, 2013; Pawlowsky-Glahn, Egozcue and Tolosana-Delgado, 2015). However, no formal definition of coherence has been clearly stated. There are other quite different, but sound, interpretations of this principle, for instance, Bear and Billheimer (2017).



1 milk	2 eggs	3 sugar
4 fat	5 juices	6 non-alcoh
7 alcoholic	8 water	9 herbs
10 grains	11 vegetables	12 roots
13 nuts	14 fruit	15 meat
16 fish	17 snacks	

Figure 1. *Correlation coefficients (Pearson, blue; Spearman, green; Kendall, red) between* milk *and* eggs*. Using raw data (full lines), and* clr *transformed data (dashed lines). Coefficients are computed on closed subcompositions including the number of parts annotated in the x-axis and ordered as enumerated in the legend.*

Nevertheless, the question of ways to relate two parts in a statistically consistent way remains open. Certainly, many analysts try to evaluate the co-variation of two parts in a compositional sample using traditional tools of statistics conceived for real random variables. For instance, the correlation between raw parts, or between clr-components (see A.3 for definition). These two examples are spurious in the sense described above since the values of such correlations can change dramatically with the subcomposition considered, even from extremely positive to extremely negative values.

In order to illustrate the drawbacks of some correlation functions on compositional data samples, we selected the *EFSA Nutrition consumption* data for *adults* as reported in the R-package *robCompositions* (Templ, Hron and Filzmoser, 2010) (See also Appendix D). Figure 1 shows the well-known effects on correlation coefficients between milk and eggs when changing the subcomposition. The correlation coefficients change from -1, when only these two parts are in the subcomposition, up to more than 0.6 for a subcomposition including 6 parts. These changes in all correlation coefficients illustrate their spurious character.

After the next Section 2 on the background of CoDA, the present goal is to formally define subcompositionally coherent sets of functions constituting a CoDA and to discuss the desirable properties of the many functions frequently appearing in CoDA (Section 3). This goal leads to the search for measures of co-variation between two parts that are admissible in a coherent CoDA, thus able to substitute the traditional, but spurious, correlation coefficients (Section 4). This last point is based on the idea that a compositional sample can be viewed as a sample of observations by rows and also as a set of compositional parts shared out on observations. The compositional space of parts (columns) is called $\mathcal{P}$-space (Pawlowsky-Glahn and Egozcue, 2022), whereas the traditional view of observations (rows) constitute the space of observations ($\mathcal{O}$-space, Section 4 and Appendix B),

2. Background

The root idea underlying CoDA is that the compositional information conveyed by a K-part composition $\mathbf{x} = (x_1, x_2, \ldots, x_K)$ does not change when it is multiplied by a real positive constant. This was formulated in the seminal works by Aitchison (1982, 1986) as the principle of scale invariance. It leads to the following concept of compositional equivalence (Aitchison, 1992; Barceló-Vidal, Martín-Fernández and Pawlowsky-Glahn, 2001; Barceló-Vidal and Martín-Fernández, 2016; Pawlowsky-Glahn et al., 2015).

Definition 2.1. [Compositional equivalence, proportionality] Two K-tuples $\mathbf{x} = (x_1, x_2, \ldots, x_K)$, $\mathbf{y} = (y_1, y_2, \ldots, y_K)$, with strictly positive components, are *compositionally equivalent* if there exists a constant $c > 0$ such that $x_i = cy_i$ for all $i = 1, 2, \ldots, K$. The equivalence classes are compositions.

This definition implies that a composition can be conveniently represented by K-tuples whose components add up to a given constant $\kappa > 0$, e.g. $\kappa = 1$ or $\kappa = 100$, as the concept of compositional equivalence exactly matches the principle of scale invariance formulated in the early 1980s (Aitchison, 1982, 1986). In a compositional exploratory analysis, proportionality between K-tuples, or an approximation of it, has been considered a linear association in the simplex (Lovell et al., 2015; Erb and Notredame, 2015; Egozcue, Pawlowsky-Glahn and Gloor, 2018). After Definition 2.1, the proportionality of two compositions is really equivalence and equality of compositions.

If compositions are considered to be equivalence classes, the usual way to work with them is to select representatives. One way of doing this is the closure operation (Aitchison, 1982).

Definition 2.2. [Closure to κ] The *closure* of $\mathbf{x} = (x_1, x_2, \ldots, x_K) \in \mathbb{R}_+^K$ to a constant sum $\kappa > 0$ is

$$\mathcal{C}_\kappa(\mathbf{x}) = \left[\frac{\kappa \cdot x_1}{\sum_{k=1}^K x_k}, \frac{\kappa \cdot x_2}{\sum_{k=1}^K x_k}, \cdots, \frac{\kappa \cdot x_K}{\sum_{k=1}^K x_k} \right] \in \mathbb{S}^K .$$

In what follows, $\kappa = 1$ for simplicity and without loss of generality.

The fact that any composition can be represented in the simplex by its closed representative suggests defining the K-part simplex, $\mathbb{S}^K$, as the sample space of the K-part compositions (Aitchison, 1982, 1986). The *perturbation* and the *powering* (Eq. A.1) were defined between closed compositions in the above references, although powering appeared there as a marginal concept. Also, distance, norm, and inner product between compositions (Eq. A.2) are defined so that $\mathbb{S}^K$ is structured as a $(K-1)$-dimensional Euclidean vector space (Billheimer, Guttorp and Fagan, 2001; Pawlowsky-Glahn and Egozcue, 2001). This structure was called *Aitchison geometry of the simplex* in the latter reference. A consequence is that any composition can be represented by Cartesian orthogonal coordinates obtained by using an isometric log-ratio transformation (ilr, also known as orthonormal log-ratio transformation (olr) after Martín-Fernández (2019)). Details of the Aitchison geometry are presented in Appendix A.

3. Definition of subcompositional coherence

The main concept of interest in the present framework is that of subcomposition.

Definition 3.1. [Subcomposition] Let $\mathbf{x}$ be a composition in $\mathbb{S}^K$. A subset of k parts, $1 < k < K$, is a *subcomposition* denoted $\mathrm{sub}(\mathbf{x})$. Its representative is chosen to be, by convention, its closure $\mathcal{C}\mathrm{sub}(\mathbf{x}) \in \mathbb{S}^k$.

The selection of parts included in a subcomposition is arbitrary. In order to simplify the notation, the k common parts in $\mathbf{x}$ and $\mathrm{sub}(\mathbf{x})$ are taken to be the first k parts ordered in the same way as in $\mathbf{x}$.

Let $\mathbf{x}$ be a K-part composition and $\mathbf{y} = \mathrm{sub}(\mathbf{x})$ a k-part subcomposition of $\mathbf{x}$, $1 < k < K$. The number of parts D is used to denote one of K or k.

Definition 3.2. [Scale invariant function in $\mathbb{S}^D$] A function $f : \mathbb{S}^D \to \mathbb{R}$, here called generically function, is *scale invariant* if, for any positive real constant $\alpha > 0$ and for any $\mathbf{x} \in \mathbb{S}^D$, it satisfies

$$f(\alpha \cdot \mathbf{x}) = f(\mathbf{x}) ,$$

that is, f is a zero-degree homogeneous function.

Any analysis of compositions is required to be based on scale invariant functions (Aitchison, 1992). When studying subcompositional coherence, this is the first requirement for any functions used. Compositional equivalence (Definition 2.1) can be considered as the main reason to require the functions involved in compositional analyses to be scale invariant.

Definition 3.3. [Invariant function under subcomposition] Let $f_D : \mathbb{S}^D \to \mathbb{R}$, $D = K, k$, be two scale invariant functions from compositions onto real values. The functions f_D are *invariant under the subcomposition* $\mathbf{y} = \mathrm{sub}(\mathbf{x})$ (IfS) if, for any composition $\mathbf{x} \in \mathbb{S}^K$, $f_K(\mathbf{x}) = f_k(\mathbf{y})$.

A trivial consequence is that, if the arguments of f_K are only those of $\mathbf{y}$, then f_K and f_k are IfS. A more subtle point refers to the labels assigned to $\mathbf{x}$, $\mathcal{C}\mathbf{x} \in \mathbb{S}^K$ and $\mathbf{y}$, $\mathcal{C}\mathbf{y} \in \mathbb{S}^k$. The common parts are normally labeled equally, although their numerical value and the space in which they are defined are different. However, the value of the i-th part in $\mathcal{C}\mathbf{y}$, depends on the whole parent composition $\mathcal{C}\mathbf{x}$ since

$$y_i = \frac{x_i}{\sum_{j=1}^{k} x_j} = \frac{x_i}{1 - \sum_{j=d+1}^{K-k} x_j} \ .$$

To provide the reader with further insight into this topic, below we examine some examples of functions that are and are not invariant to subcompositions.

- Any simple log ratio between parts in the subcomposition $\mathbf{y} = \mathrm{sub}(\mathbf{x})$ is an IfS under $\mathbf{y}$.

- Any log contrast including parts of $\mathbf{x}$ not included in $\mathbf{y} = \mathrm{sub}(\mathbf{x})$, is not an IfS under $\mathbf{y}$. For instance, $\log(x_1/\mathrm{g_m}(\mathbf{x}))$, $\mathrm{g_m}(\mathbf{x}) = (\prod_{j=1}^{K} x_j)^{1/K}$, is not an IfS under $\mathbf{y}$, as it includes parts which are not in the subcomposition and can take arbitrary values.

- Balances, logratios of geometric means of groups of parts, defined on parts within $\mathbf{y} = \mathrm{sub}(\mathbf{x})$ are IfS under $\mathbf{y}$. For instance, $\log(\mathrm{g_m}(x_1, x_2)/\mathrm{g_m}(x_3, \ldots, x_k))$ is an IfS under $\mathbf{y}$ if $k \geq 3$.

There are families of functions that change with the considered subcomposition in a monotonic way with respect to the number of parts. When these functions are used in an analysis, although changing with the subcomposition, they preserve a type of consistency. In practice, these functions use a second composition/subcomposition as a parameter. The Aitchison distance between $\mathbf{x}_1$ and a reference composition $\mathbf{x}_0$ will be non-increasing when taking any subcomposition. This subcomposition affects both $\mathbf{x}_1$ and the reference $\mathbf{x}_0$, that is, the distances to be compared are

$$\mathrm{d}_a(\mathbf{x}_1, \mathbf{x}_0) \quad \text{and} \quad \mathrm{d}_a(\mathrm{sub}(\mathbf{x}_1), \mathrm{sub}(\mathbf{x}_0)) \ .$$

In the following, compositional references, as arguments of functions, are transformed accordingly when taking a subcomposition.

Definition 3.4. [Dominant function under subcomposition] Let $f_K : \mathbb{S}^K \to \mathbb{R}$, $f_k : \mathbb{S}^k \to \mathbb{R}$, $1 < k < K$, be two scale-invariant functions from a simplex onto real values. The function f_K is *dominant* with respect to f_k (DfS) *under the subcomposition* $\mathbf{y} = \mathrm{sub}(\mathbf{x})$

if, for any composition $\mathbf{x} \in \mathbb{S}^K$, it holds either $f_K(\mathbf{x}) \geq f_k(\mathbf{y})$ (non-increasing dominance) or $f_K(\mathbf{x}) \leq f_k(\mathbf{y})$ (non-decreasing dominance). If the definition of f_K includes a compositional parameter $\mathbf{x}_0 \in \mathbb{S}^K$, the corresponding parameter in f_k, after taking subcomposition, is assumed to be $\mathbf{y}_0 = \mathrm{sub}(\mathbf{x}_0)$.

This kind of definition has been applied to Aitchison distances from the early works of Aitchison (1992) and then reported by many authors. This property was called *subcompositional dominance* of the distance since the Aitchison distance of the whole composition to a reference is always larger than or equal to that of the subcomposition. Here *dominance* refers both to a non-increasing and a non-decreasing monotonic behavior of f_K.

Some examples of functions that are and are not DfS follow:

- All IfS under a subcomposition are also DfS with respect to that subcomposition.

- Consider $\mathbf{x}_0$ and its corresponding subcomposition $\mathbf{y}_0 = \mathrm{sub}(\mathbf{x}_0)$, and take them as references in $\mathbb{S}^K$ and $\mathbb{S}^k$ respectively. Then, Aitchison distances $\mathrm{d}_a(\mathbf{x}, \mathbf{x}_0)$ dominates $\mathrm{d}_a(\mathbf{y}, \mathbf{y}_0)$, as functions of $\mathbf{x}$ and $\mathbf{y}$. However, they are not IfS under the subcomposition.

- Taking the neutral elements as references, $\mathbf{x}_0 = \mathbf{n}_K$, $\mathbf{y}_0 = \mathbf{n}_k$ in the previous example, it yields that the Aitchison norm is a decreasing DfS under the subcomposition, that is $\|\mathbf{x}\|_a \geq \|\mathbf{y}\|_a$. The inverse of the norm is an increasing DfS since

$$\|\mathbf{x}\|_a \geq \|\mathbf{y}\|_a \Rightarrow \|\mathbf{x}\|_a^{-1} \leq \|\mathbf{y}\|_a^{-1} .$$

- Consider $\mathbf{x}_0$ and its corresponding subcomposition $\mathbf{y}_0 = \mathrm{sub}(\mathbf{x}_0)$ as references in $\mathbb{S}^K$ and $\mathbb{S}^k$ respectively. Then $\langle \mathbf{x}, \mathbf{x}_0 \rangle_a$ and $\langle \mathbf{y}, \mathbf{y}_0 \rangle_a$ are functions of $\mathbf{x}$ and $\mathbf{y}$. However, the Aitchison inner product $\langle \cdot, \cdot \rangle_a$ is not DfS (see Proposition B.5).

- Consider a component of $\mathrm{clr}(\mathbf{x})$, say $f_K(\mathbf{x}) = \log(x_1/\mathrm{g_m}(\mathbf{x}))$. The function f_K is not DfS, since taking the subcomposition $\mathbf{y}$, the function $|f_k(\mathbf{y})| = |\log(x_1/\mathrm{g_m}(\mathbf{y}))|$ can take values smaller than, equal to or larger than $|f_K(\mathbf{x})|$, depending on the removed parts from $\mathbf{x}$, thus changing $\mathrm{g_m}(\mathbf{y})$ arbitrarily. Therefore, the components of a clr transformation depend in a non-dominant way on the particular subcomposition in which they are computed.

From the previous examples, some important points can be summarized in the following proposition.

Proposition 3.1. *Let* $\mathbf{x} \in \mathbb{S}^K$ *and consider a subcomposition* $\mathbf{y} = \mathrm{sub}(\mathbf{x}) \in \mathbb{S}^k$*,* $1 < k < K$*. Let* $\mathbf{x}_0 \in \mathbb{S}^K$ *and* $\mathbf{y}_0 = \mathrm{sub}(\mathbf{x}_0)$ *be reference compositions in* $\mathbb{S}^K$ *and* $\mathbb{S}^k$ *respectively. Then,*
(a) $\mathrm{d}_a(\mathbf{x}, \mathbf{x}_0)$*, as a function of* $\mathbf{x}$*, is dominant under the subcomposition* $\mathbf{y} = \mathrm{sub}(\mathbf{x})$*;*
(b) $\langle \mathbf{x}, \mathbf{x}_0 \rangle_a$*, as a function of* $\mathbf{x}$*, is not dominant under subcomposition* $\mathbf{y} = \mathrm{sub}(\mathbf{x})$.

Any CoDA consists of a set of functions $f : \mathbb{S}^K \to \mathbb{R}$. In many instances, this set of functions is applied to sample compositions and some of their subcompositions. Consistency of the analyses on the original composition and their subcompositions imposes certain requirements on the functions which define the idea of subcompositional coherence.

Definition 3.5. [Subcompositional coherence] Let $f_{K,\ell} : \mathbb{S}^K \to \mathbb{R}$ be a collection of functions, labeled by ℓ, used in a compositional analysis (CoDA). The set of functions $f_{K,\ell}$ is *subcompositionally coherent* with respect to a given subcomposition $\mathrm{sub}(\mathbf{x})$, if each function in the collection $f_{K,\ell}$ satisfies the following properties:
(a) it is scale invariant;
(b) it is subcompositionally dominant with respect to the subcomposition $\mathrm{sub}(\mathbf{x})$.
Whenever all the functions in the collection $f_{K,\ell}$ are invariant under the subcomposition, the collection is said *strictly subcompositionally coherent*.

Note that, in an informal framework, the term subcompositionally coherent was applied to invariant functions under subcompositions. Also, the term dominant was mainly used for distances as a separate concept from invariance. After Definition 3.5, IfS is a particular case of DfS. The two concepts, invariance (IfS) and dominance (DfS), are here applied to functions, whereas coherence is reserved to sets of functions constituting an analysis. For instance, the balance $\log(X_1/\sqrt{X_2X_3}) \cdot \sqrt{2/3}$ is an IfS under the subcomposition (X_1, X_2, X_3); the Aitchison distance $\mathrm{d}_a(\mathbf{x}_1, \mathbf{x}_2)$ between two samples of the original composition, $\mathbf{x}_1, \mathbf{x}_2$, is (decreasingly) dominant (DfS) under the subcomposition (X_1, X_2, X_3). Following Definition 3.5, the set of the mentioned balance and the distance between the samples is a subcompositionally coherent CoDA.

Some examples of the coherence of a CoDA follow.

- Any scale invariant function involving exclusively parts of the subcomposition is an IfS under the subcomposition.

- Expressing a composition $\mathbf{x} \in \mathbb{S}^K$ in ilr-coordinates provides $(K-1)$-coordinates which are real functions. These functions can be IfS or not depending on the coordinate system selected and the subcomposition considered. Assume that in the subcomposition $\mathbf{y} = \mathrm{sub}(\mathbf{x}) \in \mathbb{S}^k$ the k first parts remain in $\mathbf{y}$. A sequential binary partition (SBP) (Egozcue and Pawlowsky-Glahn, 2005) can always separate the first k parts and the $(K-k)$ ones not in the subcomposition as the first step of the SBP. The balance coordinates from an SBP of $(x_1, x_2, \dots, x_k)$ are then IfS. The information contained in the remaining coordinates, including the one defined in the first step of the SBP, is lost when taking subcomposition. Therefore, these latter balance coordinates are not computable from the subcomposition and they are not IfS under the subcomposition. If CoDA is based on the first $k-1$ coordinates it is coherent; if some of the $K-k+1$ coordinates are included in the CoDA, then it is not coherent.

- There is no subset of the components of $\mathrm{clr}(\mathbf{x}) \in \mathbb{R}^K$ which is coherent under any subcomposition $\mathbf{y} = \mathrm{sub}(\mathbf{x})$. This is due to the presence of all components of $\mathbf{x}$ in the geometric mean appearing in the denominator of all components of $\mathrm{clr}(\mathbf{x})$. The same applies to pivot coordinates, as they are proportional to the clr coefficients.

- Most measures of entropy or information are computed on probability distributions normalized to 1. When the distributions are considered as compositions scale invariance should allow arbitrary normalization. Then, these measures are not scale invariant and, consequently any CoDA including these measures is not coherent. However, the scalar measure of evidence information $\mathcal{I}_e(\mathbf{x}) = \|\mathbf{x}\|_a$ is a DfS under any subcomposition (Egozcue and Pawlowsky-Glahn, 2018). Also, the symmetrized compositional Kullback-Leibler divergence (Martín-Fernández, 2001) is shown to be a DfS.

4. Searching for a coherent co-variability measure of parts

4.1. Requirements and proposal

Consider a compositional sample, denoted by $\mathbf{x}_i$, $i = 1,2,\dots,N$, called observations. They form a compositional data (N,D)-matrix $\mathbf{X}$, whose entries are denoted x_{ij}, $i = 1,2,\dots,N$, $j = 1,2,\dots,D$. The columns of $\mathbf{X}$, here denoted X_j, $j = 1,2,\dots,D$ are called parts and are also N-part compositions (in general not closed). Using matrix notation

$$\mathbf{X} = (X_1,X_2,\dots,X_D) = (\mathbf{x}_1^\top,\mathbf{x}_2^\top,\dots,\mathbf{x}_N^\top)^\top ,$$

where $(^\top)$ denotes transposition. Assume that the entries of $\mathbf{X}$ are positive. Such tables can be analyzed in three different ways: (1) row-wise; (2) column-wise; (3) as a single realization of a compositional (N,D)-table (see e.g. Egozcue et al., 2015; Pawlowsky-Glahn, Egozcue and Planes-Pedra, 2019). Cases (1) and (2) are studied here and correspond to the well known R-mode and Q-mode analysis (e.g. Zhou, Chang and Davis, 1983; Grunsky, 2001).

Herein, the goal is to evaluate the relationship between two parts, for instance, without loss of generality, X_1 and X_2. Parts in $\mathbf{X}$ are N-part compositions which can be represented in $\mathbb{S}^N$, called the space of parts $\mathcal{P}^N$ (Pawlowsky-Glahn and Egozcue, 2022). In contrast, the observations $\mathbf{x}_i$ are D-part compositions in the *space of observations* which can be represented in $\mathbb{S}^D$.

The relation, co-variation, or linear association between two parts X_1 and X_2 corresponds to the comparison of two elements in $\mathcal{P}^N$, which are N-part compositions. The functions to be used to this end should be of the kind $f_N : \mathbb{S}^N \to \mathbb{R}$. If X_1 is the argument of the function, X_2 can be included as a reference in the function. Then the notation can be $f_N(X_1,X_2)$.

The requirements on f_N are:

A When taking a subcomposition in the space of observations, including a closure, $f_N(X_1, X_2)$ remains unaltered. That is, the measure of co-variation does not depend on the subcomposition of observations considered.

B When taking a sub-sample of observations, $f_N(X_1, X_2)$ should be dominant with respect to any subsample of observations which is a subcomposition in $\mathcal{P}^N$.

These two requirements are equivalent to that the analysis of co-variation of X_1 and X_2 must be IfS in $\mathcal{P}^N$ [A], and DfS in $\mathcal{O}^D$ [B]. The relations between $\mathcal{P}^N$ and $\mathcal{O}^D$ (App. B, Pawlowsky-Glahn and Egozcue (2022)) are crucial for the following discussion.

The main result is that taking a subcomposition in $\mathcal{O}^D$ results in a perturbation in $\mathcal{P}^N$. An elementary but key result is that closure in $\mathcal{O}^D$ is a perturbation in $\mathcal{P}^N$ (Proposition B.1) and vice versa, closure in $\mathcal{P}^N$ is a perturbation in $\mathcal{O}^D$. When taking a subcomposition of observations, the closure of observations should not alter the result of the analysis. This implies that the measure of co-variation $f_N(X_1, X_2)$ must be invariant under perturbation in $\mathcal{P}^N$. Denoting $\oplus_p$ the perturbation in $\mathcal{P}^N$, it means that

$$f_N(X_1, X_2) = f_N(X_1 \oplus_p P, X_2 \oplus_p P) \quad , \quad P \in \mathbb{S}^N \ ,$$

for any N-part perturbation P. An obvious solution to this requirement is that $f_N(X_1, X_2) = \varphi_N(X_1 \ominus_p X_2)$, where φ_N is a function from $\mathbb{S}^N$ into $\mathbb{R}$ and $\ominus_p$ is the perturbation-subtraction in $\mathcal{P}^N$ as suggested by Proposition B.6. The more intuitive guess for φ_N is the Aitchison norm in $\mathcal{P}^N$ or any power of it as

$$f_N(X_1, X_2) = \varphi_N(X_1 \ominus_p X_2) = \|X_1 \ominus_p X_2\|_a^w \quad , \quad w \in \mathbb{R} \ .$$

The case $w = 0$ is trivial and useless; the cases $w = 1, 2$ correspond to the Aitchison distance and its square, $f_N(X_1, X_2) = \mathrm{d}_p(X_1, X_2)$, $f_N(X_1, X_2) = \mathrm{d}_p^2(X_1, X_2)$ respectively. Taking into account that $\mathrm{d}_p(X_1, X_2) = \sqrt{N\tau_{12}^o}$, where $\tau_{12}^o = \mathrm{Var}_o(\log(X_1/X_2))$ is the $(1,2)$-entry of the variation matrix of observations (Proposition B.4), a promising scaling of the co-variation measure is

$$f_N^*(X_1, X_2) = \frac{1}{1 + \mathrm{d}_p(X_1, X_2)/\sqrt{N}} = \frac{1}{1 + \sqrt{\tau_{12}^o}} \ , \tag{1}$$

which can be named *proportionality index of parts* (PIP). It ranges from ~ 0 for large Aitchison distances to ~ 1 for small distances. Note that when $\mathrm{d}_p(X_1, X_2) = 0$ or, equivalently, $f_N^*(X_1, X_2) = 1$, the parts X_1, X_2 are proportional (Egozcue et al., 2018; Erb and Notredame, 2015; Lovell et al., 2015; Egozcue, Lovell and Pawlowsky-Glahn, 2013), and they are equivalent as compositions (Def. 2.1; see Appendices A and B). Note that the number of observations, N, is trivially invariant under p-perturbation and, given a value of τ_{12}^o, $f_N^*(X_1, X_2)$ does not change with N.

Condition B in the requirements on f_N is easily checked for f_N^* due to the relationships between distances in $\mathcal{P}^N$ and $\mathcal{O}^D$ (Propositions B.3, B.4). In fact, taking a sub-sample of observations is equivalent to taking a subcomposition in $\mathcal{P}^N$. Also in $\mathcal{P}^N$, the Aitchison distances are DfS, thus satisfying condition B.

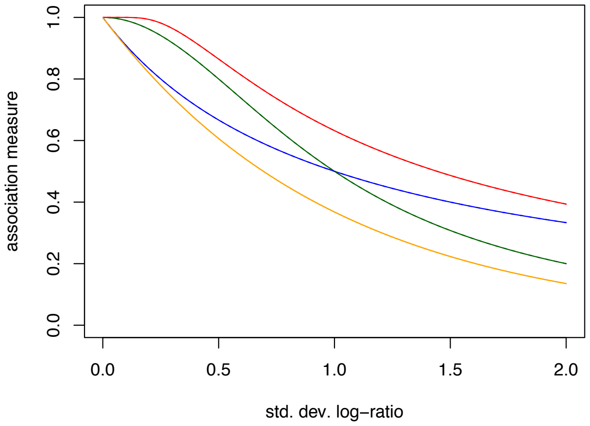


Figure 2. *Measures of association as functions of the standard deviation of the logratio. Blue: f_N^* (PIP); Green: $\tilde{f}_N$; Red: f_N^{a1}; Orange: f_N^{a2} . These functions do not depend on N.*

There are alternative functions of co-variation satisfying the same requirements. The following three are used for comparison with the PIP; they are

$$\tilde{f}_N(X_1,X_2) = \frac{1}{1+\mathrm{d}_p^2(X_1,X_2)/N} = \frac{1}{1+\tau_{12}^o} ,$$

$$f_N^{a1}(X_1,X_2) = 1-\exp\left(-\frac{1}{\sqrt{\tau_{12}^o}}\right) ,$$

$$f_N^{a2}(X_1,X_2) = \exp\left(-\sqrt{\tau_{12}^o}\right) .$$

The first one is similar to f_N^* (PIP) but uses the square p-distance in place of the p-distance; the second and the third are inspired in that proposed by Aitchison (1997) and corrected according to his explanation but in discordance from the equation shown. The curves in Figure 2 are the measures of association as functions of the standard deviation of the corresponding log-ratio $\sqrt{\tau_{12}^o}$. The main characteristic is that they attain the unit value for null standard deviation and decay to zero for large values. It is worth paying attention to the behavior for small values of the standard deviation. When the standard deviation goes to zero, f_N^{a1} attains values near to one (high linear association) before $\tilde{f}_N$. The latter goes to one clearly more quickly than f_N^* and f_N^{a2}. This is the reason to propose the PIP (Eq. 1) or f_N^{a2} as appropriate measures of proportionality as they better distinguish between small standard deviations from the exact proportionality of parts. The PIP is preferred to f_N^{a2} because the latter gives very small values for values of $\sqrt{\tau_{12}^o}$ which appear quite frequently in practice. However, the reasons to adopt f_N^* (PIP) in front f_N^{a2} are quite subjective and require further study.

The measure of association PIP has been computed for a subcomposition of nutrition data (App. D). For the whole set of countries, the PIP attains the maximum value of 0.80 for grains (*grn*) and sugar (*sug*). This value only depends on the number of countries (observations, $N = 16$) through the variability of the estimation of the standard deviation of the log-ratio between grains and sugar. Although $f_N^* = 0.80$, the proportionality between the grains and sugar is doubtful. Figure 3 (left) shows the relation between the balance $B(grn/R)$ (x-axis) and $B(sug/R)$ (y-axis), where R denotes the parts of the composition after removing grn and sug. The exact proportionality of the parts corresponds to the unit slope (red line) (Egozcue et al., 2018). The blue line has been fitted in a regression of $B(sug/R)$ on $B(grn/R)$. The p-value (0.11) on the hypothesis of the unit slope is indicated in the Figure. Note that this test depends on the subcomposition selected, thus lacking subcompositional coherence. The left panel of Figure 3, shows the same analysis for fish (fsh) and snacks (snk) when the sample is restricted to 8 countries in North Europe. In this case, $f_N^* = 0.88$ and the fitted line (blue) seem to better approach the unit slope, but the p-value (0.026) suggests a rejection of the unit-slope hypothesis.

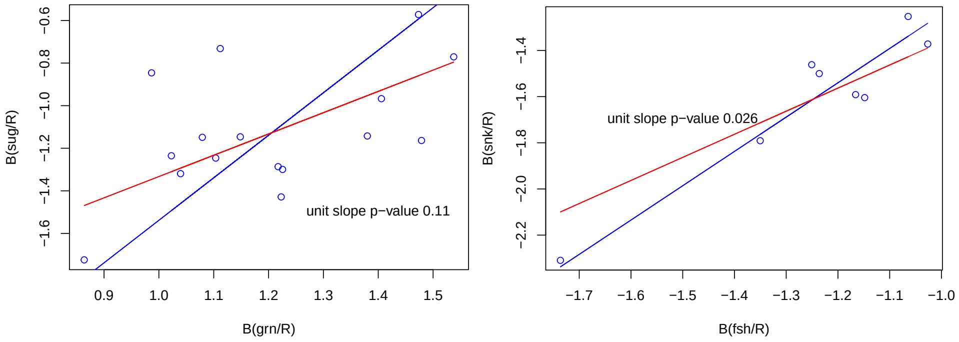


Figure 3. *Left: Balance of grain (grn) over the rest (R) of the composition against the balance of sugar (sug) over R using the whole EFSA nutrition-consumption (adults) data. Right: Balances of fsh (fish) and snk (snacks) over the rest (R) for the subsample North European countries. Red line: unit slope. Blue line: fitted to data. Both figures show a poor linear association and a doubtful unit slope.*

4.2. Non invariant under subcomposition measures of co-variation

The fact that functions used in a CoDA are not IfS or DfS (non-coherent CoDA) does not invalidate its use. However, this fact should be remarked indicating which is the subcomposition for which the result was obtained. There are many cases of standard compositional tools using non IfS or non DfS functions. Most results in a compositional biplot (Aitchison and Greenacre, 2002) are examples since the projection into two or three dimensions depends on the subcomposition. However, the explained variance is (increasingly) DfS. Also, the test of linear association used below depends on the subcomposition and is not a DfS.

There are a number of approaches to measuring co-variation or proportionality between parts. In general, they are not DfS, and hence not IfS, as the resulting values depend on the subcomposition considered. This is the case of measures proposed in

Egozcue et al. (2018), Erb and Notredame (2015), Lovell et al. (2015), Kynčlová, Hron and Filzmoser (2017) and Erb (2020). Here, the correlation in the space of parts provides another measure of co-variation but once again its DfS fails as it is based on the inner product between parts which is known to be non DfS (Prop. B.5)

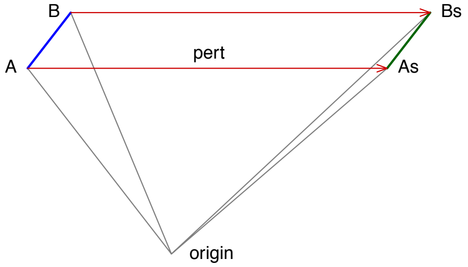


Figure 4. *In a two dimensional setting, a segment AB is shifted (pert) to AsBs maintaining its length and orientation. However, the angles AOB and AsOBs change, thus revealing that the inner product $\langle A, B\rangle$ is not invariant under shifting (pert).*

Consider an o-centred composition $\mathbf{W} = \mathbf{X} \ominus_o \mathrm{Cen}_o(\mathbf{X})$, where the observation centre is $\mathrm{Cen}_o(\mathbf{X}) = (1/N) \odot_o \bigoplus_o \mathbf{x}_i$ (Eq. 3). For two parts of $\mathbf{W}$, W_1, W_2, the p-covariance is

$$\mathrm{Cov}_p(W_1, W_2) = \frac{1}{N} \langle W_1, W_2 \rangle_p = \frac{1}{N} \langle \mathrm{clr}_p(W_1), \mathrm{clr}_p(W_2) \rangle_e \ ,$$

where $\langle \cdot, \cdot \rangle_e$ is the ordinary inner product in $\mathbb{R}^N$. The p-correlation is (Pawlowsky-Glahn and Egozcue, 2022)

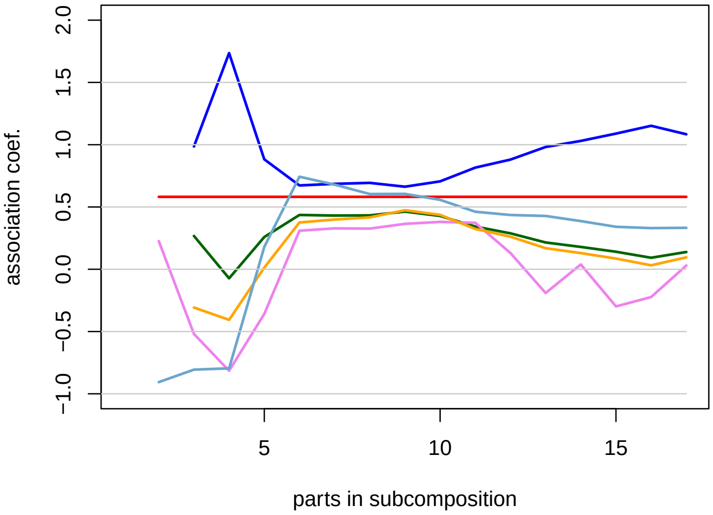
$$\mathrm{Corr}_p(W_1, W_2) = \frac{\mathrm{Cov}_p(W_1, W_2)}{\sqrt{\mathrm{Cov}_p((W_1, W_1)) \cdot \mathrm{Cov}_p((W_2, W_2))}} \ . \tag{2}$$

Although the computation of $\mathrm{Corr}_p(W_1, W_2)$ only involves the parts W_1 and W_2, it is not an IfS, as the p-inner product is neither invariant nor dominant under perturbation (Prop. B.5) and, consequently, produces non-coherent analyses. Figure 4 intuitively explains, in two dimensions, why this happens: a perturbation of two compositions shifts them in a parallel way. However, the origin of the space is unaltered and the angles subtended by the vectors before and after the shift change in a non-monotonic way.

Figure 5 shows the behavior of different measures of linear association between milk and eggs when changing the subcomposition as in Figure 1. All these measures of co-variability depend on the subcomposition considered except f_N^* (PIP) which appears constant (red line) along the set of subcompositions.

5. Conclusion

A formal definition of subcompositional coherence in a CoDA is given (Section 3). This is based on the properties of the functions which configure the analysis. In a coherent



color	range	perfect assoc.	reference
blue	$0 \leq \phi < +\infty$	$\phi = 0$	Lovell et al. (2015)
green	$-1 \leq \rho \leq 1$	$\rho = 1$	Erb and Notredame (2015)
orange	$-1 \leq \rho_{symb} \leq 1$	$\rho_{symb} = 1$	Kynčlová et al. (2017)
violet	$-1 \leq \rho_{\|R} \leq 1$	$\rho_{\|R} = 1$	Erb (2020)
turquoise	$-1 \leq \mathrm{Corr}_p \leq 1$	$\mathrm{Corr}_p = 1$	here Eq. (2)
red	$0 \leq f^*_N \leq 1$	$f^*_N = 1$	here Eq. (1)

Figure 5. *Association metrics between* milk *and* eggs*. Colors: ϕ, blue; ρ, green; symmetric balance correlation ρ_{symb}, orange; partial correlation $\rho_{|R}$, violet; p-correlation, turquoise; f^*_N, red. Different metrics are computed on closed subcompositions including the number of parts annotated in the x-axis and ordered as enumerated in the legend of Figure 1. The table below the Figure shows some characteristics of these association metrics.*

analysis, all used functions are either invariant or dominant functions under the given subcomposition. When all functions are invariant under subcomposition, the coherence is termed strict. An important point is that it is necessary to specify under which subcompositions an analysis is coherent or not. Under these definitions, there is no global coherence for all possible subcompositions.

The lack of subcompositional coherence affected the correlation analysis between parts from the beginning of CoDA, resulting in the rejection of correlation between parts as being spurious. Consequently, the study of coherent alternatives to the correlation of parts is a sensible topic. Based on the properties linking the space of parts and the space of observations, a coherent alternative called PIP, defined in the interval $(0,1)$ and based on the Aitchison distance in the space of parts, is proposed.

The lack of coherence in many CoDA, has promoted biased interpretations. They try to identify certain functions to parts. If the functions are not invariant under a subcomposition, this identification is simply a misnomer, and the conclusion of the analysis can be wrong. The most typical biased identification is that of a part to the clr coefficient which has that part in the numerator of the log-ratio. Certainly, the clr coefficient changes with the subcomposition considered.

Acknowledgements.

Grants PID2021-125380OB-I00 and PID2021-123833OB-I00 funded by MCIN/AEI/ 10.13039/501100011033 and by ERDF A way of making Europe. The authors also thank the two anonymous reviewers for their detailed and constructive review of the manuscript.

References

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

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

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

Aitchison, J. (1997). *The one-hour course in compositional data analysis or compositional data analysis is simple.* in V. Pawlowsky-Glahn (ed.), Proceedings of the III Annual Conference of the International Association for Mathematical Geology (vol. I), CIMNE, Barcelona, Spain.

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

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

Barceló-Vidal, C. and Martín-Fernández, J. A. (2016). The mathematics of compositional analysis. *Austrian Journal of Statistics 45*, 57–71.

Barceló-Vidal, C., Martín-Fernández, J. A. and Pawlowsky-Glahn, V. (2001). Mathematical foundations of compositional data analysis. In G. Ross (Ed.), *Proceedings of IAMG'01 – The VII Annual Conference of the International Association for Mathematical Geology*, Cancun (Mex), pp. 20 p.

Bear, J. and Billheimer, D. (2017). Zeros and subcompositionally coherent estimators. In K. Hron and R. Tolosana-Delgado (Eds.), *Proceedings of the 6th International Workshop on Compositional Data Analysis (CoDaWork 2017)*, pp. 1–10. Association for Compositional Data: CoDA, http://www.coda-association.org/en/.

Billheimer, D., Guttorp, P. and Fagan, W. (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association 96*(456), 1205–1214.

Boogaart, K. G. and Tolosana-Delgado, R. (2013). *Analysing Compositional Data with R*. Springer-Verlag, Berlin. pp. 258.

Chayes, F. (1971). *Ratio Correlation*. University of Chicago Press, Chicago, IL (USA). 99 p.

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

Egozcue, J. J., Lovell, D. and Pawlowsky-Glahn, V. (2013). Testing compositional association. In Hron, K., P. Filzmoser, and M. Templ (eds.) (2013). *Proceedings of the 5th Workshop on Compositional Data Analysis-CoDaWork 2013*, pp. 28-36. https://upcommons.upc.edu/bitstream/handle/2117/22147/CoDaWork2013Proceedings.pdf. Last accessed 9 October 2023.

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

Egozcue, J. J. and Pawlowsky-Glahn, V. (2006). Simplicial geometry for compositional data. In *Compositional Data Analysis in the Geosciences: From Theory to Practice*, Volume 264 of *Special Publications*, pp. 145–159. Geological Society, London.

Egozcue, J. J. and Pawlowsky-Glahn, V. (2011). Basic concepts and procedures. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), *Compositional Data Analysis: Theory and Applications*, pp. 12–28.

Egozcue, J. J. and Pawlowsky-Glahn, V. (2018). Evidence functions: A compositional approach to information (invited paper). *SORT-Statistics and Operations Research Transactions 42*(2), 1–24.

Egozcue, J. J. and Pawlowsky-Glahn, V. (2019). Compositional data: the sample space and its structure (with discussion). *TEST 28*(3), 599–638.

Egozcue, J. J., Pawlowsky-Glahn, V. and Gloor, G. B. (2018). Linear association in compositional data analysis. *Austrian Journal of Statistics 47*(1), 3–31.

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

Egozcue, J. J., Pawlowsky-Glahn, V., Templ, M. and Hron, K. (2015). Independence in contingency tables using simplicial geometry. *Communications in Statistics –Theory and Methods 44*(18), 3978–3996.

Erb, I. and Notredame, C. (2015). How should we measure proportionality on relative gene expression data? *Theory in Biosciences 135*(1-2), 21–36.

Erb, J. (2020). Partial correlations in compositional data analysis. *Applied Computing and Geosciences 6*, 9p.

Grunsky, E. (2001). A program for computing rq-mode principal components analysis for s-plus and r. *Computers & Geosciences 27*, 229–235.

Kynčlová, P., Hron, K. and Filzmoser, P. (2017). Correlation between compositional parts based on symmetric balances. *Math Geosci 49*, 777–796. doi 10.1007/s11004-016-9669-3.

Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J., Marguerat, S. and Bähler, J. (2015). Proportionality: A valid alternative to correlation for relative data. *PLoS Comput Biol 11*(3), e1004075.

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

Martín-Fernández, J. A. (2019). Comments on: Compositional data: the sample space and its structure, by Egozcue and Pawlowsky-Glahn. *TEST 28*(3), 653–657.

Martín-Fernández, J. A., Pawlowsky-Glahn, V., Egozcue, J. J. and Tolosona-Delgado, R. (2018). Advances in principal balances for compositional data. *Math Geosci 50*(3), 273–298.

Ortego, M. I. and Egozcue, J. J. (2013). Spurious copulas. In Hron, K., P. Filzmoser, and M. Templ (eds.) (2013). *Proceedings of the 5th Workshop on Compositional Data Analysis -CoDaWork 2013*, pp. 123–130. https://upcommons.upc.edu/bitstream/handle/2117/22147/CoDaWork2013Proceedings.pdf. Last accessed 9 October 2023.

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

Pawlowsky-Glahn, V. and Egozcue, J. J. (2022). Notes on the space of parts and subcompositional coherence. In C. Thomas-Agnan and V. Pawlowsky-Glahn (Eds.), *Proceedings of the 9th International Workshop on Compositional Data Analysis – CoDaWork2022–*, pp. 39–44. Association for Compositional Data.

Pawlowsky-Glahn, V., Egozcue, J. J. and Planes-Pedra, M. (2019). Survey data on perceptions of contraceptive measures as compositional tables. *Revista Latinoamericana de Psicología 50*(3), 179–186.

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

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

Templ, M., Hron, K. and Filzmoser, P. (2010). *robCompositions: Robust Estimation for Compositional Data. Manual and package, version 1.4.1.*

Zhou, D., Chang, T. and Davis, J. C. (1983). Dual extraction of r-mode and q-mode factor solutions. *Mathematical Geosciences 15*, 581–606.