Austrian Journal of Statistics
February 2018, Volume 47, 3–31.
http://www.ajs.or.at/
doi:10.17713/ajs.v47i1.689

Linear Association in Compositional Data Analysis

Juan José Egozcue Universitat Politècnica de Catalunya
Vera Pawlowsky-Glahn Universitat de Girona
Gregory B. Gloor University of Western Ontario

Abstract

With compositional data, ordinary covariation indices, designed for real random variables, fail to describe dependence. There is a need for compositional alternatives to covariance and correlation. Based on the Euclidean structure of the simplex, called Aitchison geometry, compositional association is identified to a linear restriction of the sample space when a log-contrast is constant. In order to simplify interpretation, a sparse and simple version of compositional association is defined in terms of balances which are constant across the sample. It is called b-association. This kind of association of compositional variables is extended to association between groups of compositional variables. In practice, exact b-association seldom occurs, and measures of degree of b-association are reviewed based on those previously proposed. Also, some techniques for testing b-association are studied. These techniques are applied to available oral microbiome data to illustrate both their advantages and difficulties. Both testing and measurements of b-association appear to be quite sensitive to heterogeneities in the studied populations and to outliers.

Keywords: Aitchison geometry, balances, CoDa-dendrogram, CoDa-biplot, hypothesis testing, log-ratio, log-contrast, simplex, oral microbiome.

1. Introduction

Measures of statistical dependence—association, concordance, covariation or correlation—have been important since the beginning of modern statistics. The introduction of the correlation coefficient by Galton (for historical details, see Stigler (1989)), nowadays known as the *Pearson correlation coefficient*, was a milestone in the development of multivariate statistics, though its value was discussed in varied frameworks. The controversy on the measure of association proposed by G. U. Yule (Pearson and Heron 1912) revealed both the importance of measures of association and the need of a clear comprehension on the assumptions underlaying the use of correlation-association indices. For instance, the Yule association coefficients, specially developed in the context of two way contingency tables (Yule 1903), were criticized from the point of view of the Pearson correlation coefficient, mainly designed for continuous quantitative variables. Throughout the 20^{th} century, and up to now, alternatives to Pearson correlation have been introduced, such as the Spearman and Kendall correlation coefficients (Spearman 1904; Kendall 1938), or the recently developed distance correlation (Székely, Rizzo, and Barikov 2007, 2009). Also, there were attempts of formalising these concepts and their respective indices (e.g. Schweizer and Wolff (1981); Scarsini (1984)). These indices are mainly

used to look for independence of random variables. However, the Pearson correlation coefficient is also commonly involved in modelling (linear) relationships between variables in the omnipresent linear regression. The extreme values (± 1) imply some restriction on the data, i.e. that the sample points are on a straight line or, in general, on a linear manifold of the sample space. This is also clear for distance correlation: a value of 1 implies that the sample is constant, thus lying on a single point of the sample space. Other cases are more involved, as the restrictions are not well defined, and they may depend on the underlying distribution of the data. For instance, both Spearman and Kendall coefficients of ± 1 imply a monotonic functional restriction between the sample points, but the precise relationship requires further exploration. The diversity of alternatives presented arise from the lack of specifications on the hypothesis and applicability of each method.

When multiple variables are at play, correlation coefficients describe bivariate relationships between pairs of variables. Correlation matrices can be difficult to interpret, and data simplifications are often needed to attain further insight. Principal component analysis and subsequent simplifications provide powerful interpretative tools. They consist in looking for linear combinations of the original variables so that the resulting correlations satisfy certain conditions on simplicity, sparsity or geometric orthogonality (Chipman and Gu 2005; Enki, Trendafilov, and Jolliffe 2013). These kind of approaches are of primary importance when dealing with a large number of variables (hundreds or thousands) and dimension reduction is usually required to gain insights into the phenomenon under study.

Compositional data demand careful handling of correlation-association. Pearson (1897) first observed that correlations between ratios with a common denominator were unreliable and coined the term *spurious correlation*. The problem remained at this level, ignored—or even denied (Fisher 1947)—until the sixties. At that time, problems of statistical analysis in Geology (Chayes 1960, 1962, 1971) and in Biology (Mosimann 1962; Connor and Mosimann 1969), motivated research on how to deal with spurious correlation in the analysis of compositional data. However, the discussion remained stuck at how to separate the *true* correlation part from the *spurious* one, something that we now realize makes no sense in view of the sample space of, and the information carried by, compositional data. No discussion on the concept of association was brought up. It was not until the eighties that J. Aitchison situated compositional data analysis on a premise (Aitchison 1982, 1986) by introducing the log-ratio approach, as summarized in Aitchison and Egozcue (2005).

With respect to association between compositional components or parts, Aitchison introduced the variation matrix, the matrix of sample variances of simple log-ratios between parts of a composition. These quantities

$$\mathrm{Var}\left(\ln \frac{x_i}{x_j}\right) \quad , \quad i,j = 1,2,\ldots,D \ ,$$

where x_i, x_j denote parts of a D-part composition, are in fact measures of lack of association between compositional variables. When the variance is null, x_i and x_j are strictly proportional; when it is large, proportionality is lost or it is too noisy to be considered. However, this simple measure lacks a meaningful scale; it lacks statistical techniques for testing; and is missing even a geometrical background that justifies the term *measures of (lack of) association*.

The publication of the Euclidean structure of the simplex as sample space of compositional data (Billheimer, Guttorp, and Fagan 2001; Pawlowsky-Glahn and Egozcue 2001) set the premises for further developments on association of parts of a composition. The goal of the present contribution is to give, within this framework, scale invariant measures of association for compositional parts or groups of parts. The main thesis is that proportionality between the samples of two compositional parts, or groups of parts, is an appropriate criterion of compositional association (Egozcue, Lovell, and Pawlowsky-Glahn 2013; Lovell, Pawlowsky-Glahn, Egozcue, Marguerat, and Bähler 2015).

Section 2 summarizes some main lines of the algebraic-geometric structure of the simplex as sample space of compositional data. Section 3 discusses details on geometric restrictions of the sample space when the association is exact. They inspire the measures of association proposed in Section 4 and testing techniques associated with them (Section 5). Section 6 presents some examples in the context of an *omics*-science using a well-known 16S rRNA tag sequencing data set.

2. Sample space for compositional data

The definition of compositional data has progressively evolved from vectors of positive components adding to a given constant (e.g. Aitchison (1982)), to more recent and general definitions based on equivalence classes (Barceló-Vidal, Martín-Fernández, and Pawlowsky-Glahn 2001). Here, a composition is defined as a D-component vector, all components being strictly positive, where the ratios between different components contain the relevant information. The components do not necessarily add up to a constant. The key concept is the relative character of the information of interest, as multiplication of the composition by any positive constant does not change the information contained in the ratios between components. This was stated as the principle of scale invariance by (Aitchison 1986). A direct consequence is that relevant features (distances, coordinates, sizes, relationships, ...) must be invariant under changes of scale of the compositions. The evolution of the concept of scale invariance lead to the definition of an equivalence relation between vectors of $\mathcal{R}^D_+$ (real vectors with D positive components): two vectors, $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_D)$, are compositionally equivalent, or c-equivalent, if there exists a positive constant α such that

$$x_i = \alpha\, y_i \quad , \quad i = 1, 2, \ldots, D \ .$$

The equivalence classes of this relation are called compositions. This definition was implicit in Aitchison (1986), but recognized explicitly only much later (see e.g. Aitchison (1997); Barceló-Vidal *et al.* (2001); Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn (2003); Barceló-Vidal and Martín-Fernández (2016)).

Equivalence classes are usually identified by choosing a representative of each class on which operations and reasoning can then be performed. A representative composition can be selected in many possible ways; one approach is to select the representative belonging to the simplex so that all the components add to a positive constant. In this representation compositions appear expressed in proportions. In some applications, compositions are represented in alternative ways that may not sum to a known constant: for instance, chemical concentrations are frequently expressed in mols per liter; or air pollutant composition is given in μg per m^3. However, the corresponding composition can always be represented on the simplex by a simple change of units. For a discussion of these issues, see Buccianti and Pawlowsky-Glahn (2005). That any composition can be represented on the simplex in a one-to-one way makes the simplex an appropriate representation of the sample space of compositional data. However, the simplex is not the only possible representation; for instance, a positive orthant of a hypersphere can also be used to represent a composition (Aitchison 1986; Wang, Liu, Mok, Fu, and Tse 2007). The advantage of the simplex representation is that it is mathematically easier to handle and has a simpler to interpret algebraic structure.

It is obvious that the sample space of a dataset needs a structure that is appropriate for its proper analysis. In the case of compositional data, the simplex, here denoted as $\mathcal{S}^D$, should be equipped with both internal operations and metrics. For instance, $\mathcal{S}^D$ needs to have an associated σ-field to handle random compositions and allow the assignment of probabilities to events that respects the constant sum (or closed representation) of the data. Moreover, to operate with compositions in some way, the required operation needs to be defined and closed in the sample space. The most common transformation of a composition is perturbation, i.e., componentwise multiplication. For instance, when some market shares, concentrations of

chemicals, or abundances of microbial species change, it is said that they increased, e.g. 3%, 1%, or −2%, meaning that previous shares, concentrations or abundances were multiplied by 1.03, 1.01, or 0.98. If the considered compositions are probabilities of a collection of incompatible events, perturbation is identified as Bayes' formula. Even a change of units of only some components of the composition can be viewed as a perturbation.

Perturbation of two compositions $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$, is defined as

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, x_2 y_2, \ldots, x_D y_D) \ ,$$

where $\mathcal{C}$ is the closure operator which selects the representative on the simplex; that is, $\mathcal{C}$ divides each component by the sum of all of them. Repetition of a perturbation, or multiplication by a real scalar $\alpha \in \mathcal{R}$, called powering, is

$$\alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha) \ .$$

These two operations in $\mathcal{S}^D$, introduced in Aitchison (1982), configure $\mathcal{S}^D$ as a $(D-1)$-dimensional vector space (Aitchison, Barceló-Vidal, Egozcue, and Pawlowsky-Glahn 2002). This means that it is natural to determine a basis of the space, to choose coordinates, to perform linear combinations or any of all those computations which are allowed and usual in a finite dimensional vector space and do not require a metric.

But compositional analysis also requires distances between compositions and projections of those distances. These requirements complete the structure of the simplex with metric concepts. The Aitchison distance in the simplex (Aitchison, Barceló-Vidal, Martín-Fernández, and Pawlowsky-Glahn 2000) is generated by an inner product given by

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ \ln \frac{y_i}{y_j} \ , \tag{1}$$

as $\mathrm{d}_a^2(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x} \ominus \mathbf{y}, \mathbf{x} \ominus \mathbf{y} \rangle_a$, and the corresponding norm $\|\mathbf{x}\|_a^2 = \langle \mathbf{x}, \mathbf{x} \rangle_a$. The subscripts a in the inner product, distance and norm, indicate that these expressions correspond to the Aitchison geometry of the simplex. The symbol $\ominus$ denotes perturbation-difference and can be defined as $\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus ((-1) \odot \mathbf{y})$. The simplex $\mathcal{S}^D$, endowed with perturbation, powering, and with the inner product defined above (Eq. 1) is a $(D-1)$-dimensional Euclidean space (Pawlowsky-Glahn and Egozcue 2001; Billheimer *et al.* 2001). For more details see Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado (2015).

Expressions of metric concepts in the Aitchison geometry can be simplified using the centered log-ratio (clr) transformation of a composition (Aitchison 1982). It is defined as

$$\mathrm{clr}(\mathbf{x}) = \left(\ln \frac{x_1}{\mathrm{g_m}(\mathbf{x})}, \ln \frac{x_2}{\mathrm{g_m}(\mathbf{x})}, \ldots, \ln \frac{x_D}{\mathrm{g_m}(\mathbf{x})} \right) \ ,$$

where $\mathrm{g_m}(\cdot)$ is the geometric mean of the components of its argument. The components of $\mathrm{clr}(\mathbf{x})$ add to zero by construction, that is $\sum_{i=1}^{D} \mathrm{clr}_i(\mathbf{x}) = 0$. The inverse transform is readily identified with $\mathbf{x} = \mathcal{C} \exp(\mathrm{clr}(\mathbf{x}))$, where the function exp operates componentwise on its arguments. The Aitchison inner product (Eq. 1) is then reduced to $\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle \mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{y}) \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the ordinary inner product in $\mathcal{R}^D$. From this property, the Aitchison distance and Aitchison norm can be expressed in terms of clr values as the ordinary Euclidean distance and norm of the respective clr values. Note that this expression only holds if the whole clr-vector is considered, as it changes when a reduced number of components are considered. It is said that the clr-vector is not subcompositionally coherent, as discussed below.

A crucial point in compositional data analysis is that subsets of compositional components are frequently analyzed, and the results of these analyses are to be taken as representative of what would be found with the complete set of variables. For instance, the chemical analysis of a drinking water is done using the dissolved matter after drying the sample; or a microbiome

sample is unlikely to contain proportions of all taxa present in the environment; or an RNA-seq experiment will exclude, by default, ribosomal RNA which forms the majority of the RNA in a cell; or a joint analysis of financial indices does not include all existing indices. As a consequence, a degree of coherence is required when comparing analyses of a composition and a subcomposition, the second being a strict subset of the first (Aitchison 1986; Egozcue 2009; Pawlowsky-Glahn *et al.* 2015). A reasonable required condition is that of subcompositional dominance: the distance between compositions must be greater than, or equal to, the distances between the respective subcompositions. It can be shown that taking a subcomposition in $\mathcal{S}^D$ is equivalent to an orthogonal projection in the Aitchison geometry of the simplex (Egozcue and Pawlowsky-Glahn 2005) and, therefore, within this geometry the required dominance of distances is automatically fulfilled.

In any Euclidean space, Cartesian coordinates and their corresponding orthonormal bases can be built, and the D-part simplex $\mathcal{S}^D$ endowed with the Aitchison geometry is not an exception. There are infinitely many orthonormal bases. Those made of balancing elements are nowadays of particular interest due to their easy interpretation in practical applications. A balancing element (Egozcue, Pawlowsky-Glahn, Mateu-Figueras, and Barceló-Vidal 2003; Egozcue and Pawlowsky-Glahn 2005) is a unitary composition $\mathbf{e}$, which $\mathrm{clr}(\mathbf{e})$ contains coefficients with only two non-null values. Up to permutation of components, the form of the clr of a balancing element is

$$\mathrm{clr}(\mathbf{e}) = (a_+, a_+, \ldots, a_+, a_-, a_-, \ldots, a_-, 0, 0, \ldots, 0) \ ,$$

with n_+ and n_- components with values a_+ and a_-, respectively. As the components of $\mathrm{clr}(\mathbf{e})$ add to zero, $n_+a_+ + n_-a_- = 0$, and we can assume, without loss of generality, that $a_+ > 0$ and $a_- < 0$. Moreover, as $\mathbf{e}$ is unitary, the sum of squares of the components is 1. This implies that

$$a_+ = \frac{1}{n_+}\sqrt{\frac{n_+n_-}{n_+ + n_-}} \quad , \quad a_- = -\frac{1}{n_-}\sqrt{\frac{n_+n_-}{n_+ + n_-}} \ .$$

The orthogonal projection of a composition $\mathbf{x}$ onto $\mathbf{e}$ is $b \odot \mathbf{e}$, where

$$b = \langle \mathbf{x}, \mathbf{e} \rangle_a = \sqrt{\frac{n_+n_-}{n_+ + n_-}} \ \ln \frac{\mathrm{g_m}(\mathbf{x}_+)}{\mathrm{g_m}(\mathbf{x}_-)} \ , \tag{2}$$

and $\mathrm{g_m}(\mathbf{x}_+)$, $\mathrm{g_m}(\mathbf{x}_-)$ are the geometric means of those components of $\mathbf{x}$ which correspond to positions of $\mathbf{e}$ with positive and negative components, respectively. In Equation (2), b is called balance between the groups of parts $\mathbf{x}_+$, $\mathbf{x}_-$.

An orthonormal basis of $\mathcal{S}^D$ made of balancing elements can be build using the sequential binary partition (SBP) procedure (Egozcue and Pawlowsky-Glahn 2005, 2006b, 2011). The assignment of balance-coordinates to a composition $\mathbf{x}$, corresponding to such a basis, is carried out using the isometric log-ratio transformation (ilr)

$$\mathrm{ilr}(\mathbf{x}) = V^\top \mathrm{clr}(\mathbf{x}) \quad , \quad \mathbf{x} = \mathcal{C}\exp(V \ \mathrm{ilr}(\mathbf{x})),$$

where the contrast $(D, D-1)$-matrix V has the clr values of $(D-1)$ orthonormal balancing elements as columns and $\mathrm{clr}(\mathbf{x})$, $\mathrm{ilr}(\mathbf{x})$ are considered as column matrices with D and $D-1$ components, respectively. For further mathematical details see Egozcue, Barceló-Vidal, Martín-Fernández, Jarauta-Bragulat, Díaz-Barrero, and Mateu-Figueras (2011).

3. Linear restrictions as exact association of variables

In multivariate real analysis, a group of g variables, with indices in G, is exactly correlated with another group of h variables, with indices in H, when the sample points are restricted

to lay on a hyperplane or straight-line (linear manifold) of the real space. The analytical equation for this restriction contains one or more (affine) linear relations like

$$\sum_{i\in G} \alpha_i x_i = \beta_0 + \sum_{j\in H} \beta_j x_j \ .$$

The particular case in which $g = 1$ and $h = 1$ is readily written as the simple linear model $x_1 = (\beta_0/\alpha_1) + (\beta_1/\alpha_1)x_2$. Exact correlation introduces linear constraints on the data points so that they are restricted to a linear manifold of the whole space. This idea can be translated to a compositional framework. These concepts on correlation can be formulated in two ways: a sample version in which it is assumed that a sample has been observed; and a random variable version in which all the sample space values with their respective probabilities are taken into account. In what follows, the sample version is used in the explanations, but all definitions and properties can be extended to the random variable version. A compositional n-sample is arranged in a (n, D)-matrix $\mathbf{X}$ which rows $\mathbf{x}_i$ are D-part compositions, possibly non-closed. The variables or parts are denoted X_j, $j = 1, 2, \ldots, D$, thus X_j denotes the columns of $\mathbf{X}$, while x_{ij} stands for the i, j-entry of $\mathbf{X}$.

The general expression of a one-dimensional linear restriction in the Aitchison geometry of the simplex (Pawlowsky-Glahn and Egozcue 2001) is that there is a log-contrast which is constant (k) across the sample data-points, that is

$$\sum_{i=1}^{D} \alpha_i \ln X_i = k \quad , \quad \sum_{i=1}^{D} \alpha_i = 0 \ , \tag{3}$$

where the condition on the α values assures that the log-contrast is invariant under scaling of the D-part composition $(X_1, X_2, \ldots, X_D)$. When a number $d < D - 1$ of independent one-dimensional restrictions are satisfied by the sample, the sample-points are confined within an affine subspace of dimension $D - 1 - d$. When Equation (3) holds, it is said that the group of parts with positive α are compositionally associated with the group of parts with negative α, which is called c-association for simplicity.

In compositional data analysis, this kind of general exact c-association, although important for dimension reduction, is not very useful for interpretation. Association, as a concept, tries to facilitate interpretation and this is commonly attained when statements can be easily formulated in terms of the original parts of the composition. Imagine a researcher deals with a composition with some hundreds of parts and he/she attains the result that there is a set of α values (some hundreds of them) satisfying Equation (3). The interpretation of such a log-contrast is cumbersome and the researcher possibly starts trying to find out which α values have large absolute values, identifying in this way which parts play an important role in the log-contrast. This is a typical situation in principal component analysis of high-dimensional multivariate samples or in factor analysis. At least two characteristics may be desirable for interpretation of such log-contrasts: sparsity and simplicity (Chipman and Gu 2005). Sparsity looks for a number of α values, as large as possible, that are equal to zero, so that a reduced number of parts participate in the association. When the number of α-values is restricted, the log-contrast is then simple in a given sense. The problem becomes more complicated when the analyst looks for low variability log-contrasts, since the simple selection of important contributions does not guarantee low variability of the simplified log-contrast.

The complexity of c-association, i.e. of the log-contrast in Equation (3), suggests a more restrictive concept of compositional association fulfilling requirements of sparsity and simplicity. Some log-contrasts are remarkably easier to interpret than the general version in Equation (3). They are the so called balances described in Section 2. When a balance is constant across a sample there is c-association in the sense of Equation (3), since a balance is a log-contrast. That the log contrast is a balance can be shown by defining balance association, or b-association for short.

For a more detailed definition of b-association, let us consider an n-sample D-part composition in $\mathbf{X}$. The parts of $\mathbf{X} \in \mathcal{S}^D$ can be classified in three disjoint groups of parts denoted G, H,

R; when in lower case, these letters denote the number of parts in each group. Without loss of generality, the first group G is placed in the first indices as $G = \{X_1, X_2, \ldots, X_g\}$ of the composition, and the second one, made of the following h parts, is $H = \{X_{g+1}, X_{g+2}, \ldots, X_{g+h}\}$. The remaining parts constitute a third group of parts $R = \{X_{g+h+1}, \ldots, X_D\}$ and contains $r = D - g - h$ parts. Consider the (non-normalized) balance

$$B(G/H) = \ln \frac{\mathrm{g_m}(G)}{\mathrm{g_m}(H)} \ , \tag{4}$$

where $\mathrm{g_m}(\cdot)$ denotes the geometric mean of its arguments. The balance $B(G/H)$ is a log-contrast for which the α-coefficients only take three different values: $1/g$ for group G, $1/h$ for H and 0 for group R, fulfilling simplicity requirements and sparsity when R is large compared to G and H. The groups of parts G and H are b-associated across a given sample, when the balance $B(G/H)$ is constant across that sample. This can be reformulated as G, H are b-associated when $\mathrm{Var}[B(G/H)] = 0$. Here, $\mathrm{Var}[\cdot]$ denotes the sample variance across the sample.

The particular case in which $G = \{X_1\}$ and $H = \{X_2\}$ reduces the balance to a simple log-ratio. This means that b-association in this case is expressed as any of the following equivalent conditions

$$B(\{X_1\}/\{X_2\}) = \ln \frac{X_1}{X_2} = k \quad , \quad \mathrm{Var}[B(\{X_1\}/\{X_2\})] = 0 \ . \tag{5}$$

The b-association of the groups G, H is easily interpreted in terms of proportionality of the geometric means of the parts in each group. When the balance in (4) is constant, then $\ln \mathrm{g_m}(G) = k_1 + \ln \mathrm{g_m}(H)$, for some real constant k_1; taking exponentials it yields $\mathrm{g_m}(G) = k_2 \ \mathrm{g_m}(H)$, where $k_2 = \exp(k_1)$ is a positive constant. This motivates the statement that b-association can also be called group-proportionality.

More than one b-association of groups of parts is possible, even involving overlapping groups. Each b-association implies a reduction of one unit in the total dimension $D-1$ of the simplex $\mathcal{S}^D$ whenever the restrictions are linearly independent in the Aitchison geometry of the simplex (Egozcue *et al.* 2011).

3.1. Synthetic example of b-association

In order to illustrate exact b-associations between groups, a synthetic case study of 5-part compositions has been conducted. One hundred compositions, $\mathbf{x}_i$, $i = 1, 2, \ldots, 100$, have been generated as follows:

$$\begin{aligned} \mathbf{x}_i &= (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}) \\ &= \left(0.25 \cdot i, \ 0.7 \cdot x_{i1}, \ (x_{i1})^{1.3}, \ (x_{i1})^{-1.3}, \ 2(x_{i1} \cdot x_{i4})^{1/2}\right) . \end{aligned}$$

These compositional samples are reduced to proportions taking closure and are plotted in Figure 1. The clr components of these proportions are shown in Figure 2, where the first component $\mathrm{clr}_1(\mathbf{x}_i)$ is taken on the x-axis and each of the other components on the y-axis. In Figure 2 all four dotted lines are straight-lines but only the line $(\mathrm{clr}_1(\mathbf{x}_i), \mathrm{clr}_2(\mathbf{x}_i))$ (blue) has slope equal to 1, as the parts x_{i1}, x_{i2} are proportional. Note that the power relations between x_{i1} and x_{i3} (green) (or x_{i4}, red) appear as straight lines with slopes different from 1. This tells us that X_1 is exactly b-associated with X_2 but not with X_3 and X_4. The variable X_5 also appears as a straight line (non-unitary slope, brown). Therefore, X_5 is not b-associated with X_1. However, X_5 was constructed so that $X_5 = 2 \ \mathrm{g_m}(X_1, X_4)$ and $B(\{X_5\}/\{X_1, X_4\}) = \ln 2$. Thus, there is an exact b-association between the groups $G = \{X_5\}$ and $H = \{X_1, X_4\}$, despite the fact that neither X_1 nor X_4 are separately associated with X_5. Consequently, an exact b-association between G and H does not imply b-association between the parts in these two groups. The correlation matrix of the clr-variables has all the entries equal to ± 1 in this case. For this example, the dimension of the sample space is $D - 1 = 4$, but it

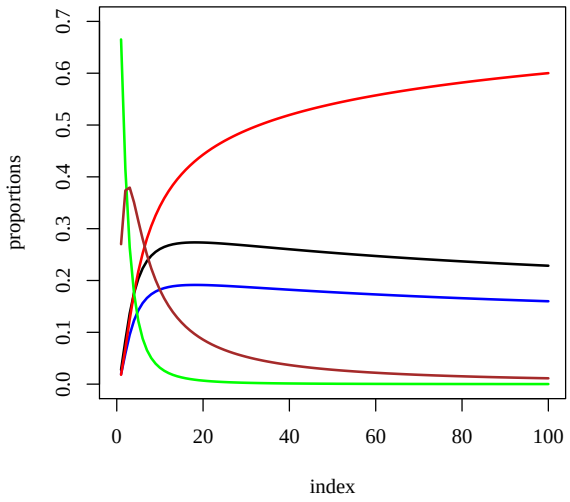

Figure 1: Proportions of one hundred compositions ordered by the subscript i as described in the text. Black, blue, green, red and brown lines correspond to $(x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})$ respectively, for $i = 1, 2, \ldots, 100$.

was constructed so that two exact b-associations take place. However, there is another exact b-association, namely $B(\{X_1, X_3, X_4\}/\{X_2\})$, which is also constant along the sample, since $x_{i3}x_{i4} = 1$. As each constant log-contrast reduces the dimension of the sample space in one unit, sample compositions are in a 1-dimensional subspace. This is easily shown centering the data matrix and performing the singular value decomposition (SVD): there is only one non-null singular value.

In the previous example, it is surprising that all correlations between clr-components are ± 1, that is, all the points in Figure 2 are aligned, yet only the case with slope equal to 1 is considered as b-association. To clarify this example it is convenient to think on the columns of $\mathbf{X}$, denoted X_j, as compositions that can be represented in $\mathcal{S}^n$. When two of these columns, say X_i, X_j, are proportional, the compositional equivalence indicates that they are equal as compositions, and the Aitchison distance in $\mathcal{S}^n$ is 0. One could also say that both parts are represented in equal proportions along the sample. For instance, this is the case of the first (black) and second (blue) parts of the example shown in Figure 1. This is related to the fact that $\mathrm{Var}(\ln(X_i/X_j)) = 0$, as this sample variance is proportional to $\mathrm{d}_a^2(X_i, X_j)$ (see appendix A). This can be seen in $\mathrm{clr}(\mathbf{X})$, which contains the clr values of the rows $\mathbf{x}_i$, $i = 1, 2, \ldots, n$ (see Figure 2, blue points). This clr-matrix can be centered by columns, which is equivalent to a shift of the origin of the simplex $\mathcal{S}^D$ to the center of the composition. The resulting matrix contains clr values both by rows for compositions in $\mathcal{S}^D$ and by columns for compositions in $\mathcal{S}^n$ and its expression is

$$\mathrm{clr}(\mathbf{X_c}) = \mathrm{clr}(\mathbf{X}) - \frac{1}{n}\mathrm{clr}(\mathbf{X})\mathbf{1}_D\mathbf{1}_n^\top ,$$

where $\mathbf{1}_k$ is a k-column vector with unitary entries. The interdistances $\mathrm{d}_a(X_i, X_j)$ can be computed as Euclidean distances between columns of $\mathrm{clr}(\mathbf{X_c})$, as they are not affected by the centering shift. The Euclidean inner product of the columns $\mathrm{clr}(\mathbf{X_c})$ is equal to the sample covariances of $\mathrm{clr}_i(\mathbf{X})$ and $\mathrm{clr}_j(\mathbf{X})$ which are the columns of $\mathrm{clr}(\mathbf{X_c})$ up to an additive constant. Therefore, if the square distance $\mathrm{d}_a^2(X_i, X_j) = 0$, then $\mathrm{Cov}(\mathrm{clr}_i(\mathbf{X}), \mathrm{clr}_j(\mathbf{X})) = 0$. But the reciprocal is not true. Intuitively, in $\mathcal{R}^{n-1}$ the fact that the angle, centered at the origin, between two points is zero does not imply that the distance between these two points is zero; it can be as large as desired. Proportionality of X_i, X_j is equivalent to $\mathrm{d}_a(X_i, X_j) = 0$. However, $\mathrm{Cov}(\mathrm{clr}_i(\mathbf{X}), \mathrm{clr}_j(\mathbf{X})) = 0$ alone is not sufficient for b-association.

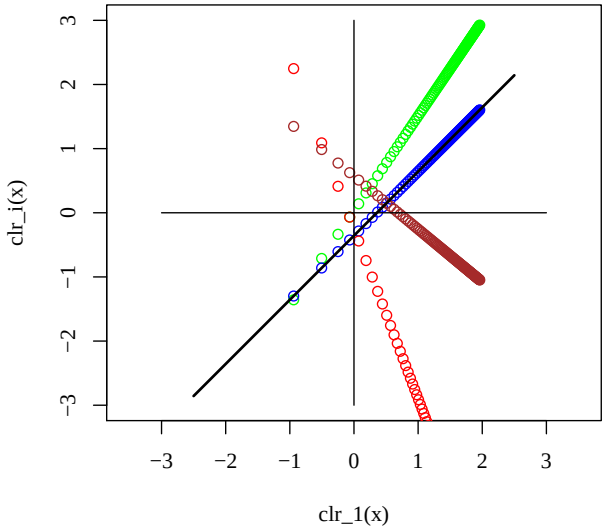

Figure 2: Components of clr of one hundred compositions of proportions shown in Figure 1. Components $\mathrm{clr}_2(\mathbf{x}_i)$ (blue), $\mathrm{clr}_3(\mathbf{x}_i)$ (green), $\mathrm{clr}_4(\mathbf{x}_i)$ (red) and $\mathrm{clr}_5(\mathbf{x}_i)$ (brown) are plotted against $\mathrm{clr}_1(\mathbf{x}_i)$. The black line has slope equal to 1 and passes through the mean value of $(\mathrm{clr}_1(\mathbf{x}_i), \mathrm{clr}_2(\mathbf{x}_i))$

3.2. Hardy-Weinberg law as a case of b-association

The Hardy-Weinberg (HW) law (Hardy 1908; Weinberg 1908) establishes that, under random mating conditions in an isolated population in which the alleles A and B present frequencies f_A, f_B, the genotype frequencies are $f_{AA} = f_A^2$, $f_{BB} = f_B^2$ (homozygote) and $f_{AB} = 2f_A f_B$ (heterozygote). This equilibrium is attained in one generation if there is parental genetic symmetry (Graffelman and Weir 2016). The allelic frequencies are arbitrary across a sample of populations in HW-equilibrium. However, the corresponding genotype frequencies satisfy

$$B(\{AA, BB\}/\{AB\}) = \ln \frac{(f_{AA} \cdot f_{BB})^{1/2}}{f_{AB}} = -\ln 2 \ ,$$

which implies an exact b-association involving the three parts. If balance-coordinates of the simplex $\mathcal{S}^3$ for genotype frequencies are defined as

$$b_1 = \sqrt{\frac{2}{3}} B(\{AA, BB\}/\{AB\}) \quad , \quad b_2 = \sqrt{\frac{1}{2}} B(\{AA\}/\{BB\}) \ ,$$

HW-equilibrium implies b_1 is constant across populations, but b_2 is not constant and can take any real value depending on the allelic frequencies in a population. This shows that if a balance involving several parts of a composition is constant, it does not imply that subsets of the parts are b-associated.

4. Approximate group-proportionality

Exact c-association seldom occurs in practice. A random composition with any exact compositional association between its parts has a degenerate probability distribution corresponding to the restriction imposed in Equation (3). Both for random compositions or sample compositions, interest is focussed in restrictions like those found in Equations (3) or (5) which only hold approximately. As a consequence, a measurement of the degree of association or proportionality is required for a consistent use in compositional analysis.

4.1. Aitchison's first approach

Aitchison (1986) introduced the first measure of (lack of) b-association when proposing the variation matrix as an exploratory tool. The entries of such a matrix are the sample variances $\mathrm{Var}[\ln(X_i/X_j)]$ for $i, j = 1, 2, \ldots, D$. The sum of all these entries, over $2D$, has been identified with the (sample) total variance or metric variance (Pawlowsky-Glahn and Egozcue 2001), that is

$$\mathrm{totVar}[\mathbf{X}] = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\mathrm{Var}\left[\ln\frac{X_i}{X_j}\right] ,$$

taking into account that $\mathrm{Var}[\ln(X_i/X_i)] = 0$. A large value for an entry of the variation matrix indicates a large contribution of the involved two parts to the total variance. Conversely, small values, close to zero, suggest that the two parts are nearly proportional. Therefore, $\mathrm{Var}[\ln(X_i/X_j)]$ is a measure of (lack of) b-association between X_i and X_j. However, the lack of a scale in $\mathrm{Var}[\ln(X_i/X_j)]$ makes its rigorous use difficult. There are different attempts to normalize the entries of a variation matrix.

A first approach reflects the proportion of total variance explained by a single log-ratio; these proportions are $\mathrm{Var}[\ln(X_i/X_j)]/(2D \cdot \mathrm{totVar}[\mathbf{X}])$. Although this normalization is sound, it strongly depends on the number of parts in the composition. A further advance consists of comparing the entries of the variation matrix with those obtained when the total variance is uniformly spread over all entries, a kind of maximal dissociation, which assigns to each non-null entry a value of $2D \cdot \mathrm{totVar}[\mathbf{X}]/(D(D-1))$. In this case, a possible normalization (Egozcue *et al.* 2013; Pawlowsky-Glahn *et al.* 2015) is

$$T_{ij} = \frac{D(D-1)\mathrm{Var}[\ln(X_i/X_j)]}{2D \cdot \mathrm{totVar}[\mathbf{X}]} = \frac{(D-1)\mathrm{Var}[\ln(X_i/X_j)]}{2 \cdot \mathrm{totVar}[\mathbf{X}]} . \tag{6}$$

Then, $T_{ij} < 1$ suggests association between the parts X_i and X_j. But experience points out that only values under 0.2 or even less can be considered candidates of effective b-association. Consequently, values larger then 0.2 can be rejected.

The mentioned lack of scale in the variances of simple log-ratios is moderated with this normalization by introducing a reference composition, $\mathbf{X}$, which includes all the parts involved in the association study. Changes of this reference composition by adding or removing parts will introduce changes in the scale.

4.2. Association between groups of parts

Recently, some additional measures of association between two parts have been introduced. For instance the ϕ-statistic (Lovell *et al.* 2015) and a modification of it (Erb and Notredame 2015). These measures of association are related to the approximate proportionality of single parts, but they can be generalised to association between groups of parts. In general, b-association is based on the proportionality of geometric means of groups of parts, that is $\mathrm{g_m}(G) \simeq k \cdot \mathrm{g_m}(H)$ or, taking logarithms,

$$\ln \mathrm{g_m}(G) \simeq k_1 + \beta_1 \ln \mathrm{g_m}(H) \quad , \quad k_1 = \ln k \quad , \quad \beta_1 = 1 . \tag{7}$$

For measuring b-association, interest is focussed on the slope β_1, which for b-association of G and H should be approximately 1. Equation (7) relates quantities such as $\ln \mathrm{g_m}(G)$ and $\ln \mathrm{g_m}(H)$, which are not scale invariant and, consequently, they are inappropriate for a compositional analysis, since the result would depend on the normalisation of the compositions. There are some options to transform Equation (7) into a scale invariant model preserving the value of β_1. A first choice is to subtract, from each term, $\ln \mathrm{g_m}(G \cup H \cup R)$, that is the logarithm of the geometric mean of all parts in the original composition; in the case that G and H reduce to a single part the equation would be $\mathrm{clr}_1(\mathbf{x}) \simeq k_2 + \beta_1 \mathrm{clr}_2(\mathbf{x})$. A second choice, which will be followed from now on, is to subtract $\ln \mathrm{g_m}(R)$ in Equation (7), thus yielding the linear model

$$B(G/R) = \beta_0 + \beta_1 B(H/R) + \epsilon , \tag{8}$$

where ϵ denotes residuals, and β_1 and the variability of the residuals are to be fitted to the available data. When $\beta_1 \simeq 1$, approximate proportionality holds for small residuals. The size of ϵ is relative to the balances $B(G/R)$ and $B(H/R)$. These balances depend on the original or reference composition which parts are included in the three groups G, H, R. Therefore, the scale of residuals are always relative to the reference composition.

4.3. Measuring b-association through a linear model

The next decision is how to fit the linear model (8) to available data. Both sides of the linear model play a symmetric role and are affected by variability. In such circumstances, ordinary least squares (OLS) regression is not an appropriate method. Following Warton, Wright, Falster, and Westoby (2006), and references therein, regression on the major axis (MA), also known as total least squares, and standardized major axis (SMA) (adopted in Lovell *et al.* (2015)) have been chosen for fitting the model (8). These approaches differ on the residual scores used to minimize their sum of squares. In OLS regression, the residual scores are equal to the ϵ values, that is the distance of data points to the fitted line in the direction of $B(G/R)$ (y-axis). In MA fitting, residual scores are chosen to be orthogonal to the fitted line. The SMA fitting procedure minimizes the sum of triangular areas comprising the data point and the fitted line limited by the projections of the data point on the fitted line following the response and explanatory axes. Figure 3 shows how residual scores are computed in OLS, MA and SMA regression. Note that, in these regression models, both axes have been taken as being orthogonal in the plane of Figure 3, a common practice in regression under the assumption that the sample space is the real space. However, when viewed in the Aitchison geometry of the simplex, these axes correspond to coordinates $B(G/R)$, $B(H/R)$ which are not orthogonal. Fortunately, this does not change the estimated model (i.e. estimation of β_0 and β_1), but it can slightly change the computation of residual scores and their sum of squares. Some tests associated with the sum of squares may be affected, but this details fall out of the scope of this contribution.

Both MA, SMA approaches coincide when the fitted line has slope equal to 1, as demanded for b-association. The estimation of slope in the SMA fit has a simpler expression than in the MA case. The SMA approach provides estimates of the slope slightly biased towards the unitary slope. In general, both approaches may be useful to evaluate b-association.

Our present goal is to discuss measures of (lack of) b-association as related to the model (8). The first one is the ϕ-statistic (Lovell *et al.* 2015) applied to balances appearing in the model given by Equation (8). That is

$$\phi(B(G/R), B(H/R)) = \frac{\mathrm{Var}(B(G/R) - B(H/R))}{\mathrm{Var}(B(G/R))} = 1 + \widehat{\beta_1}^2 - 2\widehat{\beta_1}|\widehat{r_{gh}}| \,, \tag{9}$$

where $\widehat{\beta_1}$ is the SMA-estimate of β_1 and $\widehat{r_{gh}}$ is the estimated correlation coefficient between $B(G/R)$ and $B(H/R)$.

It is obvious that $\phi(B(G/R), B(H/R))$ is equal to 0 when exact proportionality is attained with $\widehat{\beta_1} = 1$ and $\widehat{r_{gh}} = 1$; any departure from these values produces a positive increment of $\phi(B(G/R), B(H/R))$. Therefore, $\phi(B(G/R), B(H/R))$ can be taken as a measure of (lack of) b-association between G and H.

A second option measuring association (Erb and Notredame 2015), can be adapted to b-association between groups as

$$\rho(B(G/R), B(H/R)) = \frac{2\ \mathrm{Cov}(B(G/R), B(H/R))}{\mathrm{Var}(B(G/R)) + \mathrm{Var}(B(H/R))} = \frac{2\widehat{r_{gh}}}{\widehat{\beta_1} + \frac{1}{\widehat{\beta_1}}} \,, \tag{10}$$

where $\widehat{\beta_1}$ and $\widehat{r_{gh}}$ are the estimates appearing in Eq. (9). The statistic ρ is also a measure of b-association and satisfies $-1 \leq \rho(B(G/R), B(H/R)) \leq 1$, the latter value 1 corresponding

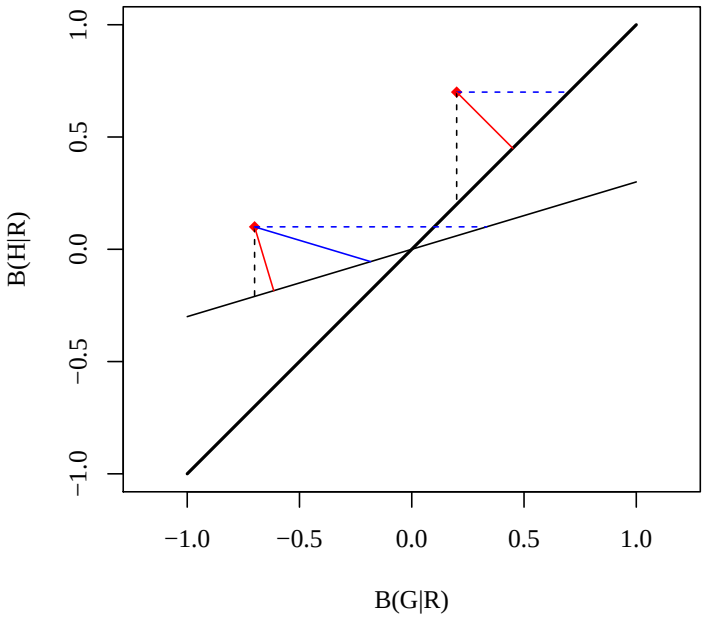

Figure 3: Visualization of residual scores for OLS, MA and SMA. Two artificial fitted lines (black full lines) with slopes 1 (thick line) and 0.3 (thin line). Two data points are shown (red marker): for each of them the OLS residual score (ϵ) is represented by a (vertical) black dashed segment to one of the fitted lines. Red lines are the residual scores used in MA, they are orthogonal to the respective fitted lines. The residual score for SMA for the fitted line with slope 0.3 is the blue segment. The SMA residual score coincides with the MA score (red) when the fitted line has slope equal to 1. The blue segment representing the SMA score is the half-diagonal of the rectangle when the triangle with vertical and horizontal sides is reflected by the fitted line; the square of this half-diagonal (blue) is proportional to the area of the triangle.

to exact b-association. Note again that the presence of the group R in both Eq. (9) and Eq. (10) and recall that these measures are defined with respect to a reference composition.

4.4. A generalisation of the variation matrix

A third possibility of measuring b-association is related to the generalisation of the entries of the variation matrix to the variances of log-ratios between the geometric means of components within each of the groups of parts considered. These variances correspond to the orthogonal projection of $\mathbf{x} \in \mathcal{S}^D$, $D = g + h + r$, such that all variance within the groups G, H, R are filtered out and the between-group variances are retained. The projected composition, up to a closure, is (Egozcue and Pawlowsky-Glahn 2005)

$$\mathbf{x}_b = (\underbrace{\mathrm{g_m}(G), \ldots, \mathrm{g_m}(G)}_{g \text{ components}}, \underbrace{\mathrm{g_m}(H), \ldots, \mathrm{g_m}(H)}_{h \text{ components}}, \underbrace{\mathrm{g_m}(R), \ldots, \mathrm{g_m}(R)}_{r \text{ components}}) \tag{11}$$

Consider the projected variation matrix

$$\mathbf{T}_b = \begin{pmatrix} 0 & gh\mathrm{Var}(B(G/H)) & gr\mathrm{Var}(B(G/R)) \\ hg\mathrm{Var}(B(H/G)) & 0 & hr\mathrm{Var}(B(H/R)) \\ rg\mathrm{Var}(B(R/G)) & rh\mathrm{Var}(B(R/H)) & 0 \end{pmatrix},$$

which entries add to $2D$ totVar($\mathbf{x}_b$). Inspired by the normalisation in Eq. (6), several normalisations can be proposed for the entries in $\mathbf{T}_b$. Here, the scaling

$$\frac{\mathbf{T}_b}{12\ D\ \mathrm{totVar}(\mathbf{x}_b)},$$

is proposed, and the entry

$$T_{GH} = \frac{gh\mathrm{Var}(B(G/H))}{12(g+h+r)\ \mathrm{totVar}(\mathbf{x}_b)}\ ,$$

is considered as a measure of (lack of) b-association between the groups G and H. Note that the constant 12 is twice the number of non-null entries of $\mathbf{T}_b$.

4.5. Comparison of measures of b-association

To compare the three measures of b-association, $\phi(B(G/R), B(H/R))$, $\rho(B(G/R),B(H/R))$ and T_{GH}, an experiment has been conducted. A set of 100 pairs $(B(G/R), B(H/R))$ has been simulated for different slopes and different orthogonal residuals as follows. First, one bivariate normal 100-sample has been simulated with mean at $(0, 0)$, with null covariance and standard deviations $\sigma_x = 5$ and $\sigma_y = 1$. These data points approximately fill an ellipse with principal axis on the x-axis. The y-axis values are then multiplied by values ranging from 0 to 5, and then rotated to attain different slopes between -3 and 3. This means that, when the standard deviation of the y-axis values of the original data is near to zero and the x-axis is rotated to slope 1, the b-association must be almost perfect. To compute T_{GH}, the number of parts grouped in G, H and R are assumed to be one, $g = h = r = 1$, and the standard deviation of $\ln \mathrm{g_m}(R) = 1$; this is needed to compute $\mathrm{totVar}(\mathbf{x}_b)$.

Figure 4 shows the results of the experiment. In the upper-left panel the values of T_{GH} are shown: exact association corresponds to slope equal 1 and standard deviation equal 0. The smaller values of T_{GH} correspond to better associations, which are illustrated by the white contour lines. The upper-right panel shows the values obtained for $\phi(B(G/R), B(H/R))$. Again the white contours are around the exact association point. Compared to T_{GH}, $\phi(B(G/R), B(H/R))$ has low sensitivity to the increase of standard deviation of orthogonal residuals, but both measures penalize symmetrically around the unitary slope. Contour lines for strong b-associations are very similar in the two cases. The bottom-left panel shows the results for $\rho(B(G/R), B(H/R))$. The main difference with the previous cases is that the values near to 1 correspond to the stronger b-associations, but changing high values into low values, the picture appears to be very similar to the one corresponding to T_{GH} (upper-left panel). For comparison, the correlation $\mathrm{Cor}(B(G/R), B(H/R))$ is presented in the bottom-right panel. Correlations near 1 include strong b-association, but it cannot be used as measure of association, as it is insensitive to changes of positive slopes. This example shows that the three measures of b-association are approximately equivalent except for their scale, and that simple correlation is inappropriate.

Introducing group R in the measures of b-association is important to fix a scale in the measures. However, there is a counterpart: the measures of b-association depend on the reference composition $G \cup H \cup R$ that is considered. Further comments on the subcompositional coherence of b-association are given in the next section on testing b-association.

Several reasons can explain a lack of proportionality even when T_{GH} and $\phi(B(G/R), B(H/R))$ are small, or when $\rho(B(G/R), B(H/R))$ is near to 1. First, these measures are not robust, and outlying samples can cause serious deviation in the results. Second, the presence of different populations in the sample may hide both b-association and the lack of it. The use of robust estimators of the variances in the variation matrix is recommended, but it does not remove the need to carefully inspect for different populations mixed in the sample. These points require further development.

5. Testing b-association

Testing b-association has limitations derived from the fact that the null hypothesis

$$H_0: \quad B(G/H) = k\ , \tag{12}$$

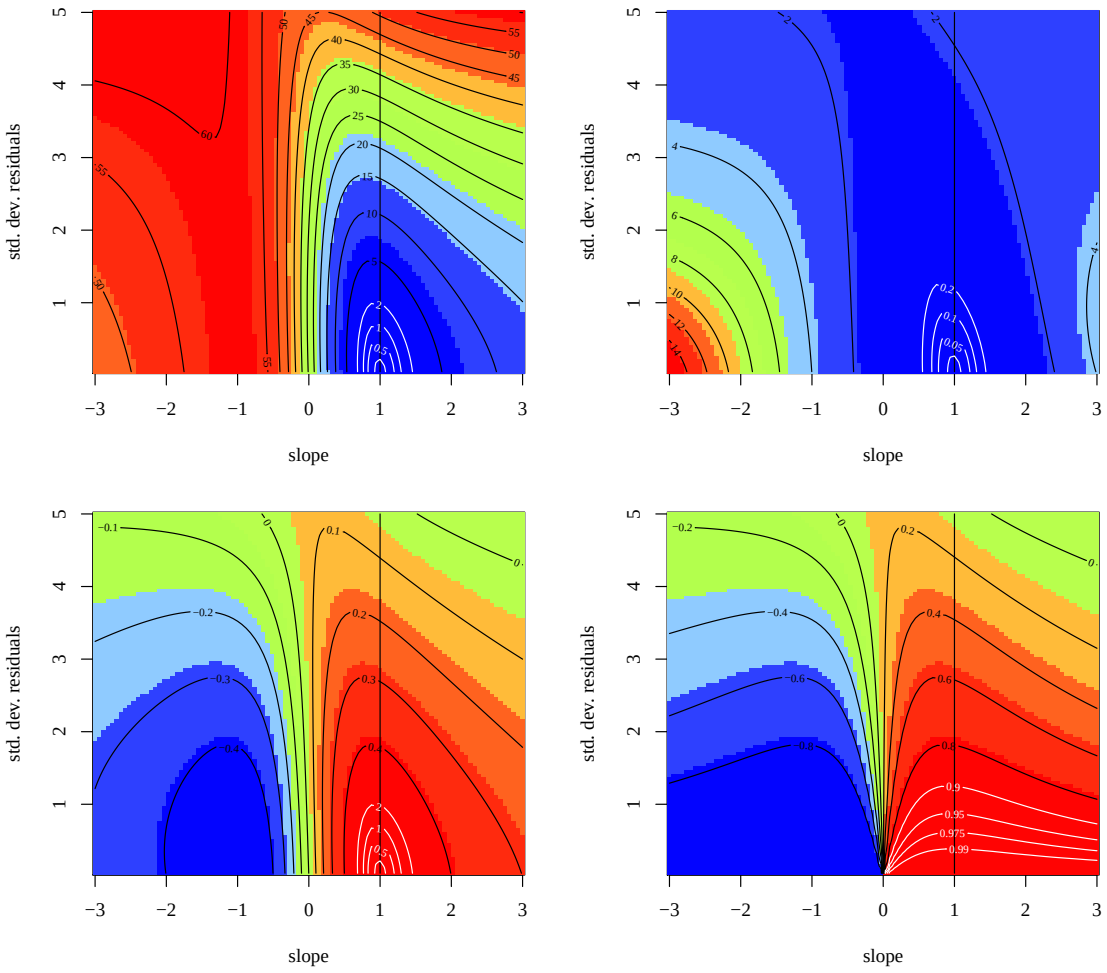

Figure 4: Magnitude of T_{GH} (upper-left), $\phi(B(G/R), B(H/R))$ (upper-right), $\rho(B(G/R), B(H/R))$ (bottom-left), and $\mathrm{Cor}(B(G/R), B(H/R))$ (bottom-right), for changing slope and standard deviation of orthogonal residuals. The scales of the four panels are not comparable as numbers on the isolines point out; only the shape of the isolines is relevant. Red and blue colors correspond to high and low values of the measure respectively. White isolines indicate approximate linear b-association. See text for elaboration.

does not admit variability for natural testing statistics. For instance, if $B(G/H)$ is used as a test statistic, its variance is null under H_0. In these circumstances, any deviation from H_0 leads to an immediate rejection of H_0 at any positive significance level. This situation is not unusual; it appears for example if for two real random variables, one tries to test that the Pearson correlation coefficient is exactly $+1$ or -1. There are some alternatives to test H_0, either modifying H_0, or making some assumption on the source of variability under H_0. Both alternatives are here adopted to propose two tests, one of them changing H_0 to $H_0' : \beta_1 = 1$, that is, testing for unitary slope in the model (8). The other alternative consists of considering a regression model of $B(G/H)$ predicted as a function of all or some balances orthogonal to $B(G/H)$ within a reference composition.

5.1. Test on unitary slope

Consider the linear model of Eq. (8). Exact b-association of the groups of parts G and H, implies H_0 as in Eq. (12). In its place, the hypothesis

$$H_0' : \quad \beta_1 = 1 \ ,$$

is tested. Note that, in case of approximate b-association, H_0' does not imply H_0, since H_0'

may hold even with large residuals. This means that H_0' is weaker than H_0.

Testing H_0' can be performed when model (8) is fitted using either major axis (MA) or standardized major axis approaches (SMA). The estimates of slope β_1 are (Warton *et al.* 2006)

$$\widehat{\beta_{MA}} = \frac{(s_{yy}^2 - s_{xx}^2) + \sqrt{(s_{yy}^2 - s_{xx}^2)^2 + s_{xy}^2}}{2s_{xy}}, \quad \widehat{\beta_{SMA}} = \text{sign}(s_{xy})\frac{s_{yy}}{s_{xx}}, \tag{13}$$

where subindices x, y denote the x-axis and the y-axis, which in model (8) are $B(G/R)$ and $B(H/R)$, respectively; the s denotes sample covariance between variables indicated by the subindices. Both estimates are very similar when the slope is near to one. Also, in both MA and SMA, the intercept term is estimated as $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$, where $\bar{x}$, $\bar{y}$ are the corresponding sample means.

The test statistic for H_0' is based on the correlation between fitted values and residuals when H_0' is assumed to hold, that is $\epsilon' = y - \widehat{\beta_0} - 1 \cdot x$ and the fitted values approach the response as

$$f'_{MA} = 1 \cdot (y - \widehat{\beta_0}) + x\,, \quad f'_{SMA} = (y - \widehat{\beta_0}) + 1 \cdot x\,.$$

For testing H_0', the proposed statistic is

$$F = (n-2)\frac{r^2_{\epsilon' f'}}{1 - r^2_{\epsilon' f'}}\,,$$

which, under H_0' and independent normality and homoscedasticity of ϵ, has an F-distribution with $(1,\, n-2)$ degrees of freedom.

This test has two additional inconveniences. The first, shared by most F-tests, is that the F-distribution is quite sensitive to departures of normality of residuals; however, even when the normality assumptions do not hold, the F-statistic is still a measure of departures from unit slopes although the obtained p-values using the F-distribution may be biased. The second inconvenience is that this test is actually testing for slope $\beta_1 = \pm 1$, and not simply for $\beta = +1$. When used for a small number of tests, it may be practical to examine by hand whether non rejections of H_0' are due to slopes near to -1. However, this strategy may be very inconvenient when performing some thousands of tests. A solution is to introduce an immediate rejection of H_0' whenever the slope estimate $\widehat{\beta_1}$ (see Eq. 13) is negative. The p-value in this situation is reported as -1.

5.2. Regression test

This testing option requires the establishment of a reference composition including the three groups of parts G, H and R. An appropriate option is to assume that the concatenation (G, H, R) is the reference and this is assumed in the discussion below. However, other subcompositions or orthogonal projections of the composition may be useful references. Consider the linear regression model

$$B(G/H) = \gamma_0 + \gamma_1 B(G, H/R) + \sum_{j=2}^{r} \gamma_j B_j + e\,, \tag{14}$$

where e represents the residuals, $B(G, H/R)$ is the balance of the concatenated G and H over R, and B_j are balances, possibly orthogonal, within the subcomposition R. The standard regression F-test, testing $\gamma_j = 0$ for $j = 1, 2, \ldots, r$, is equivalent to

$$H_0'':\ B(G/H) = \gamma_0 + e\,,$$

which is not H_0 in Eq. (12), but allows variability in H_0'' subject to the condition that it does not come from the reference composition. However, H_0'' is acceptable even with large residuals e, if they are not predictable from the reference composition. This means that the test will

be more powerful when a large number of terms are considered in the model (14). When R is a large composition containing tens or thousands of parts, this testing procedure may be too strict. This testing procedure was found to be demanding when $r = 5$ or greater (Egozcue *et al.* 2013). Here a reduced power test is proposed for its use in large compositions, both reducing power and computational effort. The approach consists of reducing the model (14) only to the first explanatory variable, the balance $B(G, H/R)$ and its coefficient γ_1, thus reducing the model to $B(G/H) = \gamma_0 + \gamma_1 B(G, H/R) + e$. Note that using this simplified approach, the projected composition in Equation (11) substitutes the original reference composition. The original model in Eq. (14) is only used when a possibly strong association between G and H is detected, and one wants to be more strict at reducing type II errors (false negatives). This simplified test gives p-values equal to the slope test whenever the estimated MA-slope is positive, as both test statistics reduce one to the other. The difference is that the simplified regression test do not distinguish between positive and negative slopes.

As in the case of testing slope (Section 5.1), the standard F-statistic is an acceptable test statistic even if the independent and homoscedastic normality of the residuals e fails. As in a routine exploration of a large composition the F-test will be used without further checking, intermediate p-values obtained (say 0.005 to 0.1) should be examined critically.

6. Example using an 16S rRNA gene profiling case

In the last decade, the interest in the study of microbial communities has grown spectacularly (e.g. Gilbert, Quinn, Debelius, Xu, Morton, Garg, Jansson, Dorrestein, and Knight (2016); Faust and Raes (2012)). In the analysis of a microbiota experiment, e.g. using 16S rRNA gene variable region sequencing techniques, there are several critical points that require further research in order to establish reliable inference procedures (Weiss, Van Treuren, Lozupone, Faust, Friedman, Deng, Xia, Xu, Ursell, Alm, Birmingham, Cram, Fuhrman, Raes, Fengzhu, Zhou, and Knight 2016). Among these points, association or correlation between taxa is an important issue. Most techniques compared in Weiss *et al.* (2016) do not take into account the compositional character of the data, while others are only partially compositional. Only few contributions to the field are consistently compositional, for instance, Fernandes, Reid, Macklaim, McMurrough, Edgell, and Gloor (2014); Gloor, Wu, Pawlowsky-Glahn, and Egozcue (2016); Tsilimigras and Fodor (2016); Silverman, Washburne, Mukherjee, and David (2017). Motivated by this situation, a data set (Human Microbiome Project Consortium 2012) of oral microbiome has been selected to illustrate the compositional techniques here proposed.

The final OTU table and metadata mapping file from the HMQCP v35 dataset were downloaded on Oct 22, 2015 from the base URL
http:\\hmpdacc.org\HMQCP\ site (Gevers, Pop, Schloss, and Huttenhower 2012). Only visit number one for each person's sample was used. The OTUs were aggregated to bacterial genus by name, and there are 229 bacterial genera sampled at 8 different sites in the mouth of $n = 1457$ human individuals, constituting the set of samples. For simplicity, only three of these sites are used: keratinzed gingiva (ak); buccal mucosa (bm); and supragingival plaque –outside plaque– (op). A common feature of 16S rRNA gene sequencing data sets, is that they are sparse, i.e., zero counts are extremely abundant in the data matrix. Log-ratio analysis of compositional data does not allow null proportions as an argument of a logarithm, thus requiring special treatment of such data (Martín-Fernández, Hron, Templ, Filzmoser, and Palarea-Albaladejo 2015a; Gloor, Macklaim, Pawlowsky-Glahn, and Egozcue 2017). The present contribution is not aimed at discussing procedures and methods for treating these zero count data, although Bayesian estimation methods show some promise (Fernandes *et al.* 2014). Therefore, the number of genera used in the examples has been reduced to 12 by removing all genera with a total count across all samples of less than 5000, or that have zero counts in more than 100 samples. Finally, samples were discarded if the 12 genera presented with more than 4 zeros. A non-defined (ND) taxon, initially included in the 12 selected ones, was also removed due to the mixed character of the ND category. The number of genera in

Table 1: Genera used in the present example. Sample size n and the number of zero counts are reported. The three last columns show the center (compositional mean) of the three selected sites attached keratinized gingiva (ak); buccal mucosa (bm); supragingival plaque –over plaque– (op).

ID	genera	n	n zeros	cen-ak	cen-bm	cen-op
1	*Actinomyces*	1436	32	0.0013	0.0140	0.2028
2	*Fusobacterium*	1436	22	0.0033	0.0087	0.0462
3	*Gemella*	1436	24	0.0314	0.0555	0.0057
4	*Granulicatella*	1436	26	0.0107	0.0082	0.0070
5	*Haemophilus*	1436	3	0.2481	0.1393	0.0878
6	*Leptotrichia*	1436	54	0.0006	0.0060	0.0617
7	*Neisseria*	1436	33	0.0044	0.0202	0.1072
8	*Porphyromonas*	1436	47	0.0086	0.0177	0.0280
9	*Prevotella*	1436	10	0.0229	0.0216	0.0423
10	*Streptococcus*	1436	0	0.6364	0.6807	0.3532
11	*Veillonella*	1436	0	0.0325	0.0280	0.0579

the final composition is then $D = 11$.

Table 1 shows the genera which are accounted for, the sample size, which is equal for each site, and the number of zero-counts for each genus. In order to estimate all proportions of taxa, the method GMB (Martín-Fernández *et al.* 2015a) has been used. It consists of estimating proportions as

$$x_{ij} = \frac{n_{ij} + \alpha_{ij}}{n_{i+} + \alpha_{i+}} \quad , \quad \alpha_{ij} = s_i t_{ij} \quad , \quad i = 1, 2, \ldots, n \ , \ j = 1, 2, \ldots, D \ ,$$

where n_{ij} are the counts, possibly zero, for the i-th sample in the j-th taxon, and a subscript $+$ indicates sum in the corresponding dimension. This estimation is a Bayesian point estimation of multinomial probabilities. The values of s_i (strength) and t_{ij} (share) should be assessed from the characteristics of the problem and the available prior knowledge. In this case, following the advice in the previous reference, these values are

$$t_{ij} = \frac{n_{+j} - n_{ij}}{n_{++} - n_{i+}} \quad , \quad s_i = \exp\left(-\frac{1}{D}\sum_{k=1}^{D} \log t_{ik}\right) \ .$$

Table 1 also reports the compositional means for the genera in the three sites ak, bm and op. For each site, they are computed as the compositional average (Pawlowsky-Glahn *et al.* 2015)

$$\widehat{\mathrm{Cen}}[\mathbf{X}] = \frac{1}{n} \odot \bigoplus_{i=1}^{n} \mathbf{x}_i = \mathcal{C}(\mathrm{g_m}(X_1), \mathrm{g_m}(X_2), \ldots, \mathrm{g_m}(X_D)) \ ,$$

which reduces to compute the geometric means of the columns of the data matrix. Observing the values of the proportions in the centres substantial differences between sites are detected. A compositional MANOVA (Martín-Fernández, i Estadella, and Mateu-Figueras 2015b) based on the Aitchison distance d_a could be conducted to find out substantial differences between centres (not presented here).

Principal component analysis of compositional data (Aitchison 1983) and the corresponding biplots (Aitchison and Greenacre 2002) are powerful tools for exploring compositional data. This is also the case in exploring microbiome data (Gloor *et al.* 2016). They are computed by taking clr of the data set, centring it, and then carrying out singular value decomposition. That is

$$\mathrm{clr}(\mathbf{X}) - \mathrm{clr}(\widehat{\mathrm{Cen}}[\mathbf{X}]) = U \ \Lambda \ V^\top \ ,$$

where Λ is a diagonal matrix containing the singular values λ_i, $i = 1, 2, \ldots, D$; the (n, D)-matrix U, called the score matrix, contains standardized coordinates of the compositions;

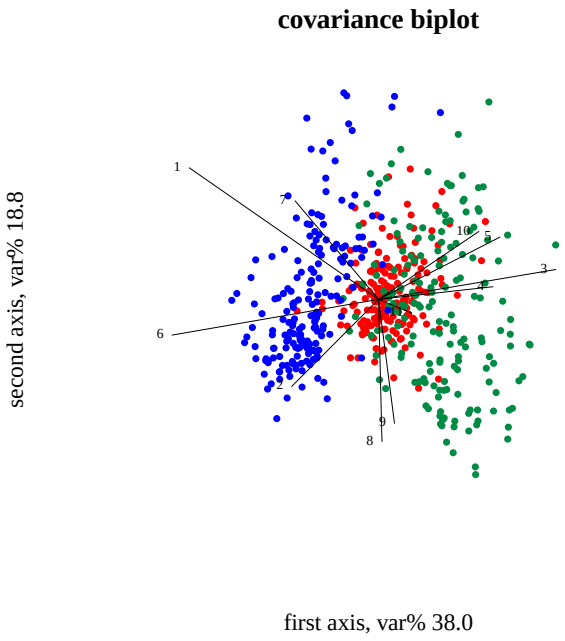

Figure 5: Covariance biplot of the data set projected on the first and second principal axes. Buccal sites: ak, green; bm, red; op, blue. Numbers represent clr-variables of taxa.

and the (D, D)-matrix V (loadings matrix), where the columns are the clr values of the new orthonormal basis of representation. The last singular value is always zero, as the components of the clr values add up to zero and $\text{clr}(\mathbf{X})$ has, at most, rank $D-1$. The matrix $U\Lambda$ (except its last column) contains centered ilr-coordinates with respect to the basis associated with V (except its last column). A biplot consists of a 2-dimensional plot, representing simultaneously either U and $V\Lambda$ (covariance biplot) or $U\Lambda$ and V (form biplot). In a two-dimensional plot of the first two principal components, the proportion of the total variance represented in the plot is $(\lambda_1^2 + \lambda_2^2)/\text{totVar}[\mathbf{X}]$ (for more details see e.g. Pawlowsky-Glahn *et al.* (2015)). Figures 5 and 6 show the covariance and form biplots projected on the first two principal components. The first observation is that the three sites are quite well separated by the first principal component (see this in the form biplot), which is roughly similar to the balance between the groups of parts $\{X_3, X_4, X_5, X_{10}\}$ and $\{X_1, X_6\}$, as revealed by the covariance biplot. The first two principal components explain 56.8% of the total variance and all observations made on the biplots are subject to this limited projection. In a covariance biplot like the one in Figure 5 the line linking two extremes of lines is, up to the projection, proportional to the standard deviation of the log-ratio between the corresponding variables. Therefore, rays where end-points are close suggest pairwise b-association between those variables. In Figure 5 the end-points of variables 8-*Porphyromonas* and 9-*Prevotella* are quite close, thus suggesting b-association. A similar case is visualised for the variables 5-*Haemophilus* and 10-*Streptococcus*. In both cases, the form biplot (Fig. 6) shows that both pairs of end-points are also close; this means that, in the projection, the corresponding unitary vectors are approximately equal, thus supporting the b-association between these couples of taxa. If these b-associations were confirmed, the result would be regarded as strong evidence for compositional association of these genera at the three different sites examined.

Biplots may be useful for a visual detection of pairwise b-associations but they need to be confirmed by examination of the variation matrix and other analyses. Another visual way of detecting possible b-associations, pairwise or involving larger groups, is the CoDa-dendrogram (Egozcue and Pawlowsky-Glahn 2006a) as shown in Figure 7. The tree structure indicates the SBP (sequential binary partition) used for the construction of a basis of balances. The length

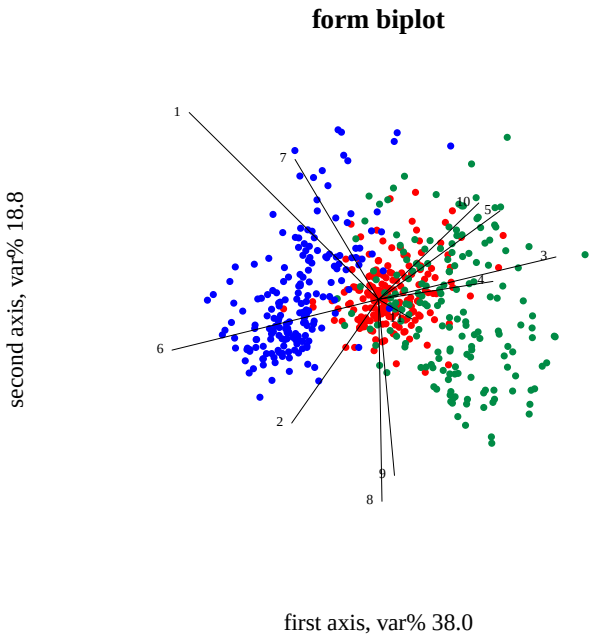

Figure 6: Form biplot of the data set projected on the first and second principal axes. Buccal sites: ak, green; bm, red; op, blue. Numbers represent clr-variables of taxa.

of the vertical bars illustrate the variance of the corresponding balance, and the length of the anchoring branch connecting the nodes is proportional to the mean variance of the balance. Vertical bars in black correspond to the aggregate of all samples at all sites (ak, bm, op), and the colored bars to individual sample groups. A relatively short vertical bar represents stronger evidence of b-association, because it represents a small variance of the balance. The advantage of the CoDa-dendrogram is that it is able to visualize associated groups of parts and not just pairwise b-associations. The counterpart is that only the balances obtained in the SBP are checked. Figure 7 shows the CoDa-dendrogram corresponding to an SBP which has been obtained by hierarchical clustering (Ward's method) of the genera (Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado 2011), using as the distance matrix, the square-roots of the variation matrix of the whole population (sites ak, bm, op), which is actually an Aitchison distance (Appendix A).

In Figure 7, only the balance between 5-*Haemophilus* and 10-*Streptococcus* seems to be approximately constant, that is, it exhibits a small variance of the log-ratio. The b-association seems particularly strong in the case of two individual sample groups (ak-green and bm-red). No b-association of larger groupings of genera is suggested in this dendrogram.

The suggested b-association between 5-*Haemophilus* and 10-*Streptococcus* was examined more closely. Figure 8 shows the fitted line of $B(5\text{-}Haemophilus/R)$ and $B(10\text{-}Streptococcus/R)$ for the three sites ak (left), bm (middle), op (right), where the p-values in the slope test are 0.51, 0.07, 0.0001. That is the b-association is not rejected in the ak samples, is near to rejection in the bm samples and is clearly rejected in op samples.

Similarly, potential b-association between 1-*Actinomyces* and 6-*Leptotrichia* is inspected in Figure 9. In this case, the obtained p-values are 0.82 and 0.11 for samples from ak and bm, while the slope was negative for the op samples thus producing an automatic rejection. In samples from the ak site, the slope test (p-value 0.86) did not reject b-association, but the scatterplot of the points indicates that b-association, if any, is quite noisy and the test is only driven by the slope, which in this case is approximately equal to 1. In these three cases, the obtained p-value is largely influenced by the presence of outliers, thus suggesting the future use of robust estimators for accurate testing.

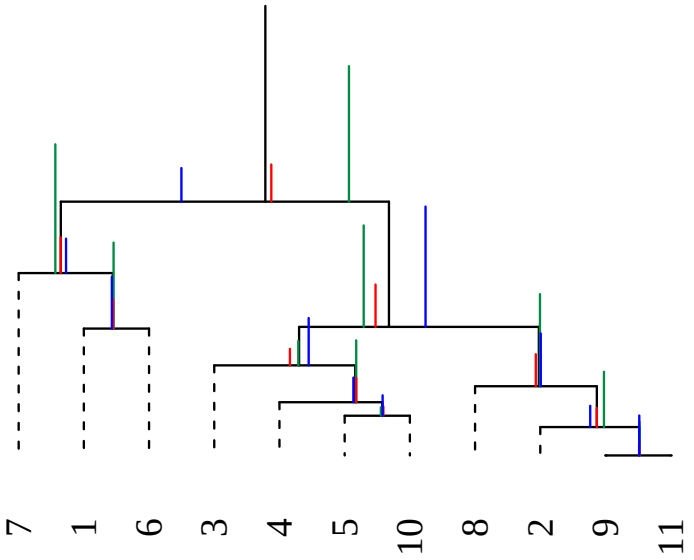

Figure 7: CoDa-dendrogram of approximately principal balances. Buccal sites: ak, green; bm, red; op, blue. Numbers represent clr-variables of taxa.

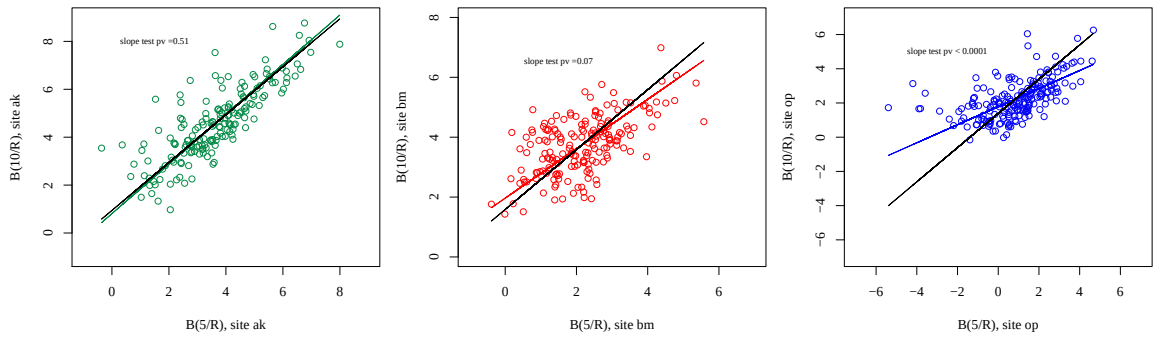

Figure 8: Scatterplot and fitted line of $B(5-Haemophilus/R)$ and $B(10-Streptococcus/R)$ corresponding to buccal sites ak (green), bm (red), op (blue). The null hypothesis corresponds to the unitary slope line (black).

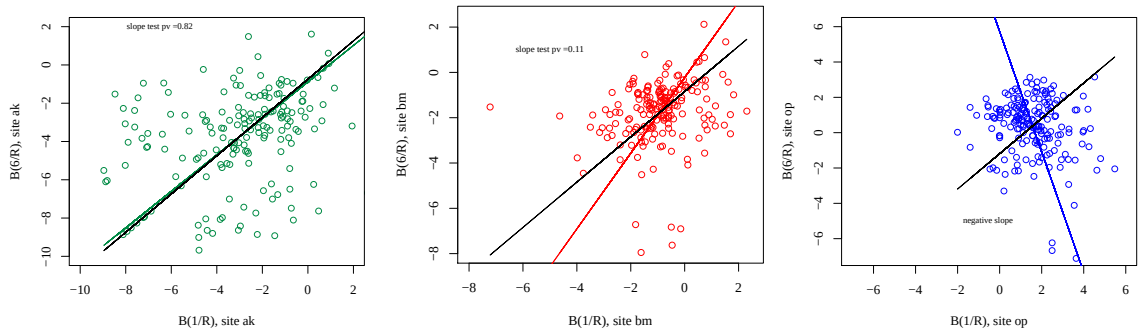

Figure 9: Scatterplot and fitted line of B(1-Actinomyces/R) and B(6-Leptotrichia/R) corresponding to buccal sites ak (green), bm (red), op (blue). The null hypothesis corresponds to the unitary slope line (black).

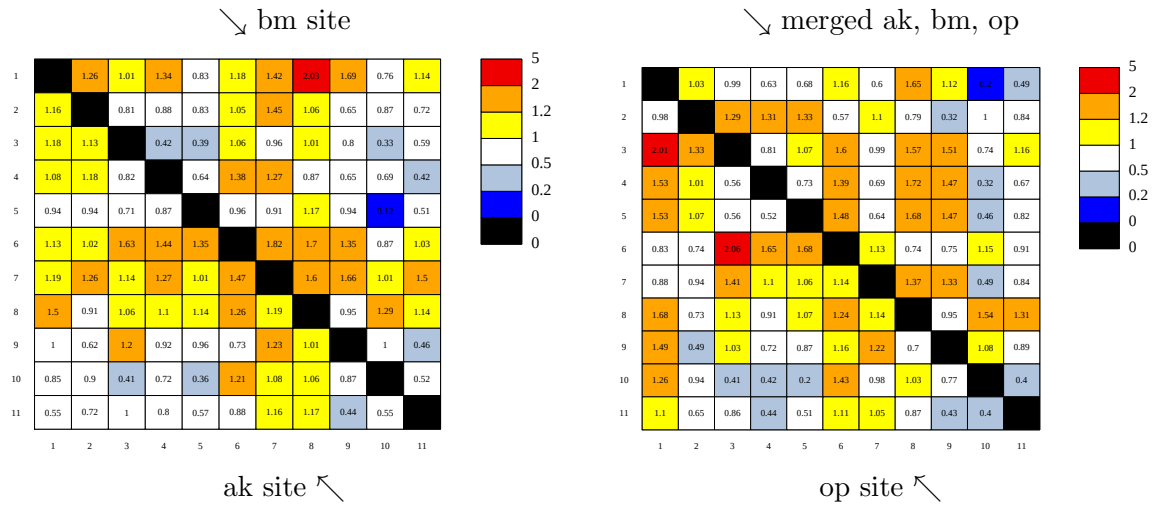

Figure 10: Normalized variation matrices for sites ak (left lower triangle), bm (left upper triangle), op (right lower triangle), for ak, bm, op merged in a single population (right upper triangle)

Figure 10 shows a summary heat map of pairwise b-association for the 11 genera in samples from the three selected sites (ak, bm, op). The top row shows the normalized variation matrices (Eq. 6) for the three sites and of the three sites together (four triangles). The middle and lower rows show the p-values of the slope test and the multiple regression test respectively for the genera at the same sites.

The first observation is that the suggestions of pairwise b-association by the three criteria (normalized variation, and the two p-value based methods) can differ substantially. However, if attention is paid to the strong b-association between 5-*Haemophilus* and 10-*Streptococcus* in bm site shown in Figure 8 (left panel), both the normalized variation matrix (0.12) and the slope test (p-value 0.51) suggest b-association. However, the regression test is clearly significant p-value $< 10^{-4}$, thus rejecting b-association of this pair by this test. In general, we observe that the multiple regression test is too strict, but instead suggests that b-association between larger groups of genera may be present if such groupings are examined. This situation is frequent, and is explored further below. In the lower panel row, there are only few non-significant cells and this number tends to decrease when the considered subcomposition is larger. As a counterpart, a large p-value, say greater than 0.05, in the multiple regression test often implies non-significant slope test and a small value of the normalized variation. This is the case in the joint population (ak, bm, op), for the b-association between 2-*Fusobacterium* and 9-*Prevotella*, which has a 0.05 p-value in the regression test, a 0.99 p-value in the slope test, and a moderate small value of normalized variation (0.49). This may be considered as a surprising case, as fewer b-associations are expected in the joint population of the three cases. However, the total variance is larger in the joint population and all tests and measures must be interpreted relative to the total variance. Together, these results indicate that further refinements in the tests and measures of b-association, and the use of robust statistics, may improve this approach.

As commented above, a rejection of b-association in the multiple regression test may suggest a b-association involving more than two variables. For this example with 11 parts, the number of non-significant cases (0.05 p-value or larger) in the slope test is large. For instance, in the ak site, there are 598 cases of G, H containing 2 parts each which give p-values larger than 0.05. Two of these b-associations, randomly selected, are studied in Figure 12. In the ak site (left panel), the balance of 10-*Streptococcus* and 11-*Veillonella* versus 4-*Granulicatella* and 5-*Haemophilus* is non-significant in the slope-test (p-value 0.409). Although it seems an acceptable b-association, it is quite noisy. The right panel of Figure 12 shows another case of b-association in the site bm (slope-test p-value 0.897). It involves 8-*Porphyromonas*

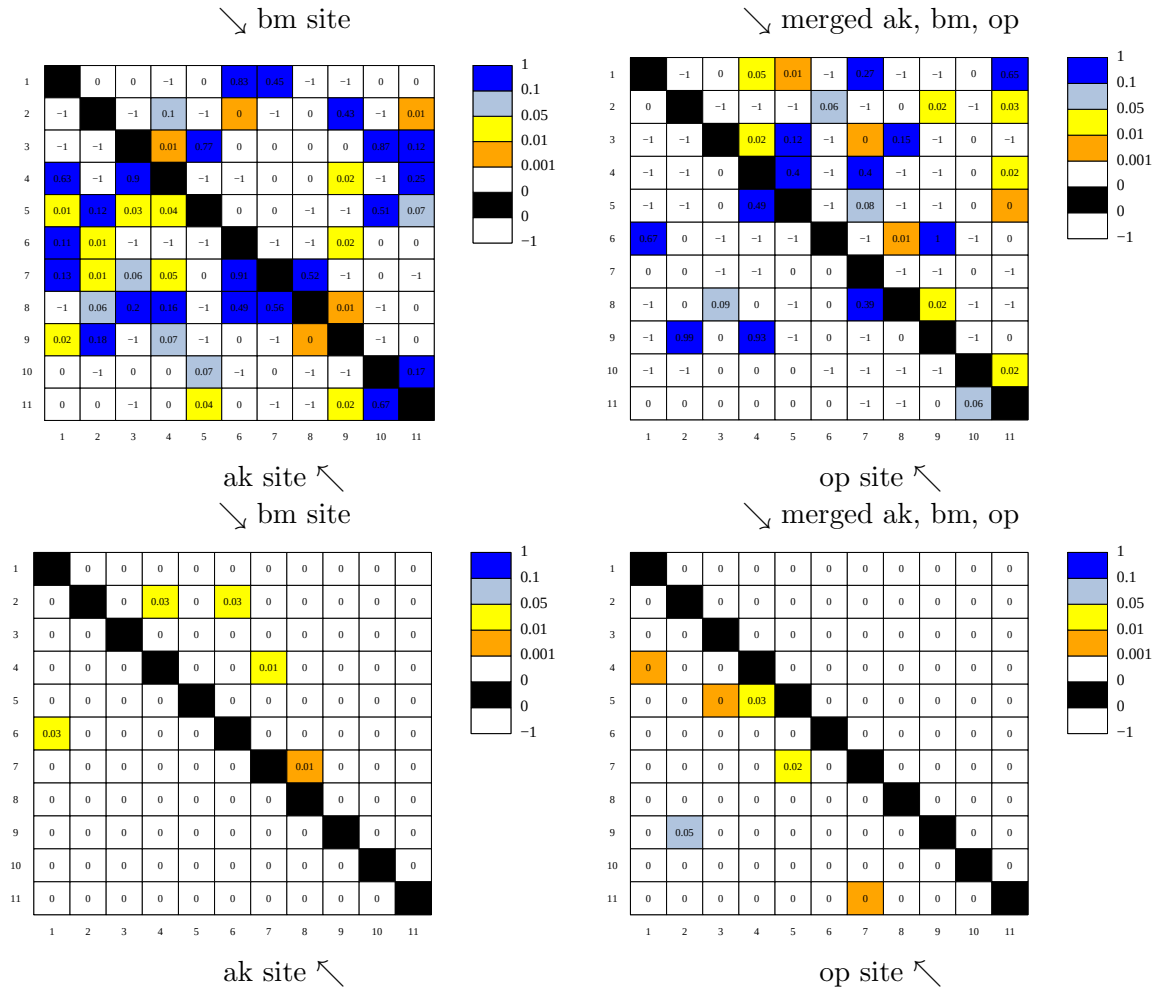

Figure 11: p-values for the slope test of association (panels in the upper row) and for the multiple regression test (panels in the lower row). They correspond to sites ak (left, lower triangles), bm (left, upper triangles), op (right, lower triangles), merged population ak, bm, op (right, upper triangles).

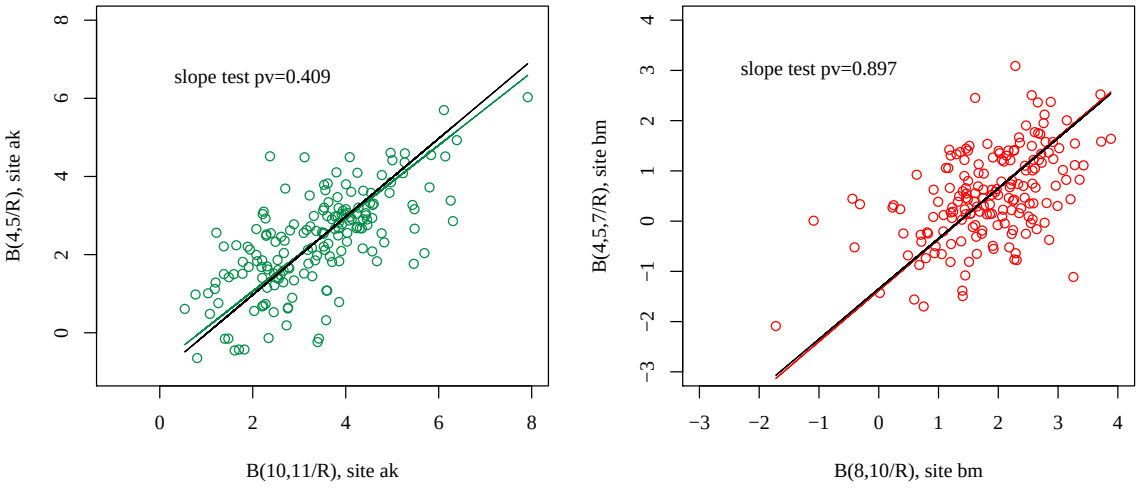

Figure 12: Testing unit slope for group association in two cases for which b-association is not rejected. Black line, unit slope; colored line, MA fitted line. Left panel: site ak, testing constant $B(10, 11/4, 5)$ in the 11-part simplex. Right panel: site bm, testing constant $B(8, 10/4, 5, 7)$

and 10-*Streptococcus* versus 4-*Granulicatella*, 5-*Haemophilus* and 7-*Neisseria*. Again the b-association is noisy. These two examples show that the slope and regression tests are useful to detect approximate unit slope between $B(G/R)$ and $B(H/R)$. However, the both the slope and regression tests are quite insensitive to large variability of residuals around the unit slope line. This is the reason why any of the proposed tests on b-association should be accompanied by some measurement of b-association as they, directly or indirectly, account for the scatter of data points around the unit slope line in plots like those shown in Figures 8, 9 and 12.

7. Conclusions and further research

Linear association of parts in a composition has been examined from a theoretical point of view, beyond the early proposals (Egozcue *et al.* 2013; Lovell *et al.* 2015). Proportionality of two parts in a composition was the idea for introducing b-association which leads to an approximately constant log-ratio. Maintaining this initial idea, the concept is extended to proportionality between geometric means of two groups of parts (b-association) which implies that the corresponding balance is constant. More generally, compositional linear association can be viewed as an approximately constant log-contrast between pairs or groups of parts.

The b-association, defined as the presence of a constant balance, is a simple generalisation of the proportionality between two parts, thus providing a way of thinking on groups of compositional components better than individual ones. That is, it is a way to go from pairwise relationships like that provided by standard pairwise correlations to more complex, but still linear in the simplex, relations involving more than two parts. The counterpart is that procedures to detect such group relations with a reasonable computing effort and their global interpretation is a pending research task.

Although the concept of linear compositional association is quite simple, its measurement and testing is more complex and requires further research. An important point is that the statistics here proposed, both for testing and measuring b-association, depend on the reference composition. Moreover, both measures of b-association and tests are severely affected by outliers and contamination in non-homogeneous populations. This strongly suggests the study of robust statistics both for measuring and testing.

When using b-association in fields like microbiome, or transcriptome, or other 'omic analysis, the presence of zeroes introduces a new source of contamination for detecting b-associations efficiently and accurately. This is especially true when such associations involve more than two parts. It is clear that the number of linear associations (b-associations) is less than predicted by other association criteria such as Pearson's correlation on proportions, or when using Spearman and Kendall correlations which include cases of non-linear associations. Nevertheless, the use of b-association provides a needed sanity-check, since the widely used correlation methods often provide many false positive associations.

Acknowledgements

This work was supported by grants MTM2015-65016-C2-1-R and MTM2015-65016-C2-2-R (MINECO/FEDER) of the Spanish Ministry of Economy and Competitiveness and European Regional Development Fund. The authors thank the constructive comments by the reviewers, which help improving the quality 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 (1983). "Principal Component Analysis of Compositional Data." *Biometrika*, **70**(1), 57–65.

Aitchison J (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press), London (UK). ISBN 0-412-28060-4. 416 p.

Aitchison J (1997). "The One-hour Course in Compositional Data Analysis or Compositional Data Analysis Is Simple." In V Pawlowsky-Glahn (ed.), *Proceedings of IAMG'97 - The III Annual Conference of the International Association for Mathematical Geology*, volume I, II and addendum, pp. 3–35. CIMNE, Barcelona, Spain ISBN 978-84-87867-76-7, Barcelona (E). ISBN: 84-87867-97-9.

Aitchison J, Barceló-Vidal C, Egozcue JJ, Pawlowsky-Glahn V (2002). "A Concise Guide for the Algebraic-geometric Structure of the Simplex, the Sample Space for Compositional Data Analysis." In U Bayer, H Burger, W Skala (eds.), *Proceedings of IAMG'02 – The VIII Annual Conference of the International Association for Mathematical Geology*, volume I and II, pp. 387–392. Selbstverlag der Alfred-Wegener-Stiftung, Berlin, 1106 p. ISSN 0946-8978.

Aitchison J, Barceló-Vidal C, Martín-Fernández JA, Pawlowsky-Glahn V (2000). "Logratio Analysis and Compositional Distance." *Mathematical Geology*, **32**(3), 271–275.

Aitchison J, Egozcue JJ (2005). "Compositional Data Analysis: where Are We and where Should We Be Heading?" *Mathematical Geology*, **37**(7), 829–850.

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

Barceló-Vidal C, Martín-Fernández JA (2016). "The Mathematics of Compositional Analysis." *Austrian Journal of Statistics*, **45**, 57–71. doi:10.17713/ajs.v45i4.142.

Barceló-Vidal C, Martín-Fernández JA, 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*, p. 20 p. Cancun (Mex).

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

Buccianti A, Pawlowsky-Glahn V (2005). "New Perspectives on Water Chemistry and Compositional Data Analysis." *Mathematical Geology*, **37**(7), 703–727.

Chayes F (1960). "On Correlation between Variables of Constant Sum." *Journal of Geophysical Research*, **65**(12), 4185–4193.

Chayes F (1962). "Numerical Correlation and Petrographic Variation." *The Journal of Geology*, **70**(4), 440–452.

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

Chipman HA, Gu H (2005). "Interpretable Dimension Reduction." *Journal of Applied Statistics*, **32**, 969–987.

Connor RJ, Mosimann JE (1969). "Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution." *Journal of the American Statistical Association*, **64**(325), 194–206.

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

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

Egozcue JJ, Lovell D, Pawlowsky-Glahn V (2013). "Testing Compositional Association." In PF K Hron, M Templ (eds.), *Proceedings of the 5th Workshop on Compositional Data Analysis –CoDaWork 2013*. ISBN: 978-3-200-03103-6, http://coda.data-analysis.at/.

Egozcue JJ, Pawlowsky-Glahn V (2005). "Groups of Parts and Their Balances in Compositional Data Analysis." *Mathematical Geology*, **37**(7), 795–828.

Egozcue JJ, Pawlowsky-Glahn V (2006a). "Exploring Compositional Data with the CoDa-Dendrogram." In E Pirard, A Dassargues, HB Havenith (eds.), *Proceedings of IAMG'06 – The XI Annual Conference of the International Association for Mathematical Geology*. University of Liège, Belgium, CD-ROM, Liège (B).

Egozcue JJ, Pawlowsky-Glahn V (2006b). "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. ISBN 978-1-86239-205-2.

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

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003). "Isometric Logratio Transformations for Compositional Data Analysis." *Mathematical Geology*, **35**(3), 279–300.

Enki HA, Trendafilov NT, Jolliffe IT (2013). "A Clustering Approach to Interpretable Principal Components." *Journal of Applied Statistics*, **40**(3), 583–599.

Erb I, Notredame C (2015). "How Should We Measure Proportionality on Relative Gene Expression Data?" *Theory in Biosciences*, **first online**, 1–16. doi:10.1007/s12064-015-0220-8.

Faust K, Raes J (2012). "Microbial Interactions: from Networks to Models." *Nature Reviews*, **August 2012**, 539–550.

Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB (2014). "Unifying the Analysis of High-throughput Sequencing Datasets: Characterizing RNA-seq, 16S rRNA Gene Sequencing and Selective Growth Experiments by Compositional Data Analysis." *Microbiome*, **2**, 15.1–15.13. doi:10.1186/2049-2618-2-15.

Fisher RA (1947). "The Analysis of Covariance Method for the Relation between a Part and the Whole." *Biometrics*, **3**(2), 65–68.

Gevers D, Pop M, Schloss PD, Huttenhower C (2012). "Bioinformatics for the Human Microbiome Project." *PLoS Comput Biol*, **8**(11), e1002779. doi:10.1371/journal.pcbi.1002779.

Gilbert JA, Quinn RA, Debelius J, Xu ZZ, Morton J, Garg N, Jansson JK, Dorrestein PC, Knight R (2016). "Microbiome-wide Association Studies Link Dynamic Microbial Consortia to Disease." *Nature*, **535**, 94–103. doi:10.1038/nature18850.

Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ (2017). "Microbiome Datasets Are Compositional: and This Is Not Optional." *Frontiers Microbiology*. doi:10.3389/fmicb.2017.02224.

Gloor GB, Wu JR, Pawlowsky-Glahn V, Egozcue JJ (2016). "It's All Relative: Analyzing Microbiome Data as Compositions." *Annals of Epidemiology*. doi:10.1016/j.annepidem.2016.03.003.

Graffelman J, Weir BS (2016). "Testing for Hardy-Weinberg Equilibrium at Biallelic Genetic Markers on the X Chromosome Heredity." *Heredity*, **116**(6), 558–568. doi:10.1038/hdy.2016.20.

Hardy GH (1908). "Mendelian Proportions in a Mixed Population." *Science*, **28**, 49–50.

Human Microbiome Project Consortium (2012). "Structure, Function and Diversity of the Healthy Human Microbiome." *Nature*, **486**(7402), 207–14. doi:10.1038/nature11234.

Kendall MG (1938). "A New Measure of Rank Correlation." *Biometrika*, **30**, 81–93.

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

Martín-Fernández JA, Barceló-Vidal C, Pawlowsky-Glahn V (2003). "Dealing With Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation." *Mathematical Geology*, **35**(3), 253–278.

Martín-Fernández JA, Hron K, Templ M, Filzmoser P, Palarea-Albaladejo J (2015a). "Bayesian-multiplicative Treatment of Count Zeros in Compositional Data Sets." *Statistical Modelling*, **15**(2), 134–158.

Martín-Fernández JA, i Estadella JD, Mateu-Figueras G (2015b). "On the Interpretation of Differences between Groups for Compositional Data." *Statistics & Operations Research Transactions, SORT*, **39**(2), 231–252.

Martín-Fernández JA, Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2017). "Advances on Principal Balances for Compositional Data."

Mosimann JE (1962). "On the Compound Multinomial Distribution, the Multivariate β-distribution and Correlations among Proportions." *Biometrika*, **49**(1-2), 65–82.

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

Pawlowsky-Glahn V, Egozcue JJ (2001). "Geometric Approach to Statistical Analysis on the Simplex." *Stochastic Environmental Research and Risk Assessment (SERRA)*, **15**(5), 384–398.

Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2011). "Principal Balances." In JJ Egozcue, R Tolosana-Delgado, MI Ortego (eds.), *Proceedings of the 4th International Workshop on Compositional Data Analysis (2011)*. CIMNE, Barcelona, Spain ISBN 978-84-87867-76-7.

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

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.

Pearson K, Heron D (1912). "On Theories of Association." *Journal of the Royal Statistical Society*, **LXXV**, 579–652.

Scarsini M (1984). "On Measures of Concordance." *Stochastica*, **VIII**(3), 201–218.

Schweizer B, Wolff EF (1981). "On Nonparametric Measures of Dependence for Random Variables." *The Annals of Statistics*, **9**(4), 879–885.

Silverman JD, Washburne AD, Mukherjee S, David LA (2017). "A Phylogenetic Transform Enhances Analysis of Compositional Microbiota Data." *eLife*, **6**(e21887). doi:10.7554/eLife.21887.

Spearman C (1904). "The Proof and Measurement of Association between Two Things." *American Journal of Psychology*, **15**, 72–101. doi:10.2307/1412159.

Stigler SM (1989). "Francis Galton's Account of the Invention of Correlation." *Statistical Science*, **4**(2), 73–79. URL http://www.jstor.org/stable/2245329.

Székely GJ, Rizzo ML, Barikov NK (2007). "Measuring and Testing Dependence by Correlation of Distances." *The Annals of Statistics*, **35**(6), 2769–2794.

Székely GJ, Rizzo ML, Barikov NK (2009). "Brownian Distance Covariance." *The Annals of Applied Statistics*, **3**(4), 1236–1265.

Tsilimigras MCB, Fodor AA (2016). "Compositional Data Analysis of the Microbiome: Fundamentals, Tools, and Challenges." *Annals Epidemiology*, **26**, 330–335. doi:10.1016/j.annepidem.2016.03.002.

Wang H, Liu Q, Mok HMK, Fu L, Tse WM (2007). "A Hyperspherical Transformation Forecasting Model for Compositional Data." *European Journal of Operational Research*, **179**(2), 459–468.

Warton DI, Wright IJ, Falster DS, Westoby M (2006). "Bivariate Line-fitting Methods for Allometry." *Biol. Rev.*, **81**, 259–291. doi:10.1017/S1464793106007007.

Weinberg W (1908). "Über den Nachweis der Vererbung beim Menschen." *Jahreshefte d. Vereins vaterl. Naturkunde in Württemb.*, **64**, 369–382.

Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, Xia LC, Xu ZZ, Ursell L, Alm EJ, Birmingham A, Cram JA, Fuhrman JA, Raes J, Fengzhu S, Zhou J, Knight R (2016). "Correlation Detection Strategies in Microbial Data Sets Vary Widely in Sensitivity and Precision." *International Society for Microbial Ecology (ISME) Journal*, **10**(7), 1669–1681.

Yule GU (1903). "Notes on the Theory of Association of Attributes in Statistics." *Biometrika*, **2**(2), 121–134.

A. Relation of compositional variation with Aitchison distance

A compositional data matrix $\mathbf{X}$ is considered. The n rows of $\mathbf{X}$ are compositions represented in $\mathcal{S}^D$. Using the notation introduced in Sections 3 and 4, the following theorem holds.

Theorem It holds

$$n \cdot \mathrm{Var}\left(\ln \frac{X_i}{X_j}\right) = \mathrm{d}_a^2(X_i, X_j)\,,$$

where Var is the n-sample variance and $\mathrm{d}_a^2(\cdot,\cdot)$ is the Aitchison distance in $\mathcal{S}^n$. The following proof is similar to that presented in Martín-Fernández, Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado (2017).

Proof: Denoting g_i for the geometric mean of column i, i.e. $g_i = (\prod_{k=1}^n x_{ki})^{1/n}$, each entry of the variation matrix can be written as

$$\mathrm{Var}\left(\ln \frac{X_i}{X_j}\right) = \frac{1}{n}\sum_{k=1}^{n}\left(\ln \frac{x_{ki}}{x_{kj}}\right)^2 - \left(\ln \frac{g_i}{g_j}\right)^2 .$$

On the other hand, if we take the transpose of the data matrix $\mathbf{X}$, one can consider each part of the composition as an observation of a composition with n parts. These n parts correspond to the n initial samples. The squared Aitchison distance between two of the parts, e.g. X_i and X_j, is then

$$\begin{aligned}
&\mathrm{d}_a^2(X_i, X_j) \\
&= \sum_{k=1}^{n} \left[\ln \frac{x_{ki}}{g_i} - \ln \frac{x_{kj}}{g_j} \right]^2 = \sum_{k=1}^{n} \left[\ln \frac{x_{ki}}{x_{kj}} - \ln \frac{g_i}{g_j} \right]^2 \\
&= \sum_{k=1}^{n} \left(\ln \frac{x_{ki}}{x_{kj}} \right)^2 + \sum_{k=1}^{n} \left(\ln \frac{g_i}{g_j} \right)^2 - 2 \sum_{k=1}^{n} \ln \frac{x_{ki}}{x_{kj}} \ln \frac{g_i}{g_j} \\
&= \sum_{k=1}^{n} \left(\ln \frac{x_{ki}}{x_{kj}} \right)^2 + n \left(\ln \frac{g_i}{g_j} \right)^2 - 2n \left(\ln \frac{g_i}{g_j} \right)^2 .
\end{aligned}$$

Thus, it holds that

$$n \cdot \mathrm{Var} \left(\ln \left(\frac{X_i}{X_j} \right) \right) = \mathrm{d}_a^2(X_i, X_j) \,.$$

As a consequence, we can define a *total squared distance* as follows

$$\begin{aligned}
\mathrm{totSDist}[\mathbf{X}] &= \mathrm{totVar}[\mathbf{X}] \\
&= \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \mathrm{Var} \left[\ln \frac{X_i}{X_j} \right] \qquad \text{(by definition)} \\
&= \frac{1}{2Dn} \sum_{i=1}^{D} \sum_{j=1}^{D} \mathrm{d}_a^2(X_i, X_j) \,.
\end{aligned}$$

Dividing each $\mathrm{d}_a^2(X_i, X_j)$ by $\mathrm{totSDist}[\mathbf{X}]$ we obtain a normalised matrix of inter-distances between compositional parts. In fact,

$$\frac{1}{2Dn} \sum_{i=1}^{D} \sum_{j=1}^{D} \frac{\mathrm{d}_a^2(X_i, X_j)}{\mathrm{totSDist}[\mathbf{X}]} = 1 \,.$$

Affiliation:

Juan José Egozcue
Dept. of Civil and Environmental Engineering
Universitat Politècnica de Catalunya
Barcelona, Spain
E-mail: juan.jose.egozcue@upc.edu

Vera Pawlowsky-Glahn
Dept. of Computer Science, Applied Mathematics, and Statistics
Universitat de Girona, Girona, Spain
E-mail: vera.pawlowsky@udg.edu

Gregory B. Gloor
Dept. of Biochemistry
University of Western Ontario
London, Ontario, Canada
E-mail: ggloor@uwo.ca

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 47 *Submitted:* 2017-03-27
February 2018 *Accepted:* 2018-01-02