Chapter 4
Modelling Compositional Data. The Sample Space Approach

Juan José Egozcue and Vera Pawlowsky-Glahn

Abstract Compositions describe parts of a whole and carry relative information. Compositional data appear in all fields of science, and their analysis requires paying attention to the appropriate sample space. The log-ratio approach proposes the simplex, endowed with the Aitchison geometry, as an appropriate representation of the sample space. The main characteristics of the Aitchison geometry are presented, which open the door to statistical analysis addressed to extract the relative, not absolute, information. As a consequence, compositions can be represented in Cartesian coordinates by using an isometric log-ratio transformation. Standard statistical techniques can be used with these coordinates.

Keywords Compositional data analysis · Aitchison geometry
Simplex · Variation matrix · Biplot · Balance dendrogram · ilr · clr

AMS Subjet classifications 62-07 · 62-02

4.1 Introduction

The difficulties when dealing with compositional data have been known for more than a century. Indirectly, Pearson (1897) described some of these problems and coined the term spurious correlation. They are easily illustrated using the early characterizations of compositional data, which relay on the constant sum constraint (CSC). For instance, Chayes (1960, 1962) and Connor and Mosimann (1969) based

J. J. Egozcue (✉)
Department of Civil and Environmental Engineering, Universidad Politécnica de Cataluña,
Barcelona, Spain
e-mail: juan.jose.egozcue@upc.edu

V. Pawlowsky-Glahn
Department of Computer Science, Applied Mathematics and Statistics, University of Girona,
Girona, Spain
e-mail: vera.pawlowsky@udg.edu

B. S. Daya Sagar et al. (eds.), *Handbook of Mathematical Geosciences*,
https://doi.org/10.1007/978-3-319-78999-6_4

their analysis on the fact that a vector of proportions $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ satisfies the CSC,

$$\sum_{i=1}^{D} x_i = \kappa > 0\,, \quad x_i > 0\,, \quad i = 1, 2, \ldots, D\,. \tag{4.1}$$

It defines the κ-simplex of D components or parts. Here the simplex is denoted $\mathbb{S}^D$, with no reference to the positive constant κ. Data fulfilling the CSC were called *constrained or closed data*. In the eighties, promoted by J. Aitchison, this kind of data were recognized as *compositional data* (Aitchison and Shen 1980; Aitchison 1982, 1986). In the last reference, additional conditions were added to the original CSC characterisation, leading to the formulation of some principles for compositional data analysis. They were the starting point on which the log-ratio approach to compositional data is based. These principles have been reformulated several times in order to depurate and to clarify them for users (Aitchison and Egozcue 2005; Egozcue 2009; Egozcue and Pawlowsky-Glahn 2011a; Pawlowsky-Glahn et al. 2015). Nonetheless, they have been contested from different points of view (e.g. Scealy and Welsh 2014), arguing that they match the conditions for the application of log-ratio methods. But not all data satisfying the CSC (4.1), for instance admitting that some parts can be zero, are automatically adequate for a log-ratio analysis. In the last decade, in which the log-ratio approach has shown to be useful in a large number of applications, it also became clear that it can be rigorously applied to problems in which the CSC is not fulfilled, or where the components do not represent proportions. The key point for this change of the paradigm represented by the CSC, is the conception of compositions as equivalence classes of vectors which positive components are proportional (Barceló-Vidal et al. 2001; Martín-Fernández et al. 2003; Pawlowsky-Glahn et al. 2015; Barceló-Vidal and Martín-Fernández 2016), and the related idea that the simplex is just a representation of the sample space of compositions. This fact is a direct consequence of the scale invariance of compositions (Aitchison 1986) but, up to now, its implications have not been completely recognised.

This contribution aims at a reformulation of the principles of compositional data analysis in their log-ratio version, presenting them as a practical and natural need in many situations of data analysis. Section 4.2 discusses scale invariance and compositional equivalence and Sect. 4.3 presents the simplex as an appropriate sample space for compositional data. Perturbation, the group operation between compositions, is shown to be a natural operation in Sect. 4.4. The Aitchison distance and the requirements on it are discussed in Sect. 4.5. The consequence of the previous sections is the Euclidean space structure of the simplex, which has been termed Aitchison geometry (Pawlowsky-Glahn and Egozcue 2001). The Aichison geometry has been shown to be useful for the modelling and analysis of compositions, centring the interest in the relative information contained in the data. Some of these elements are commented in Sect. 4.6.

4.2 Scale Invariance, Key Principle of Compositions

When somebody records the composition of a product, material, shares of a market, species in an ecosystem or a kitchen recipe, he or she implicitly recognizes that the total amount is irrelevant for the description of the product, material, shares, species or recipe. This does not mean that the size or the amount is not informative, it only tells us that, whichever is the size, the elements of the total are distributed according to the specified composition. Essential are the ratios between the components of the described system. One can say that for any system that can be decomposed into parts its description has, at least, two types of information: one that is referred to as *size*, and another one that concerns the relations between the parts irrespective of the size. This latter one is called compositional information and, when the system is a geometric object, it is called *shape*. Beyond size (total amount) and composition (shape), there may be other properties of the system which can be quantified (color, sound, complexity, strength, …) and again these additional properties may be decomposed into size and composition. Here, attention is paid to systems which are formed by parts, while their size or total amount is either analysed in another way or is irrelevant. For a discussion of a possible approach to a problem where interest lies in the relative information and in the total, see Pawlowsky-Glahn et al. (2015), Olea et al. (2016), Ferrer-Rossell et al. (2016).

Think about the map of a region; even changing the scale of the map, the same region is identified. If the distance between two mountain peaks was 12 cm, and a lake between the two was 4 cm broad, halving the scale new lengths of 6 and 2 cm will be obtained. The distance between the two peaks and the width of the lake can be identified as equal in the two maps, as the ratio is in both cases $12/4 = 6/2 = 3$. Only when the maps are to be transformed into an actual region, the size becomes relevant and it is revealed taking into account the scale of the maps. Note that in the case of the peaks and the lake, the considered parts, the distance between peaks and the width of the lake, are not disjoint, as the first includes the second. In fact, the previous comments did not imply that the parts of the system had to be non-overlapping or disjoint.

The irrelevance of the total led J. Aitchison (1986) to introduce the principle of scale invariance for compositions. A composition is assumed to be represented by an array of positive numbers which quantitatively represent the parts of the system. Let $\mathbf{x} = (x_1, x_2, \dots, x_D)$, $x_i > 0$ for $i = 1, 2, \dots, D$, be such a composition. Consider any positive constant $c > 0$. The scale invariance principle can be stated as: $\mathbf{x}$ and $c\mathbf{x}$ contain the same compositional information. From this point of view, *compositional equivalence* can be defined (Aitchison 1997; Barceló-Vidal et al. 2001; Barceló-Vidal and Martín-Fernández 2016; Pawlowsky-Glahn et al. 2015).

Definition 4.2.1 (*Compositional equivalence*) Let $\mathbf{x} = (x_1, x_2, \dots, x_D)$ and $\mathbf{y} = (y_1, y_2, \dots, y_D)$ be two arrays of D positive components. They are compositionally equivalent if there exists a positive constant c such that, for $i = 1, 2, \dots, D$, $y_i = cx_i$.

Two equivalent arrays **x**, **y** represent the same composition. Both the equivalence class generated and its representative are called compositions.

Figure 4.1 shows some artificial, arbitrary data of Ca and Mg in mg/l from a fictitious water analysis (circles). Each pair (Ca,Mg) can be considered as a two part composition. A line from the origin through each data point consists of compositionally equivalent points, thus visualising a composition, strictly speaking an equivalence class. Any point on these rays can be chosen as a representative of the composition. Particularly, they can be selected so that the sums of the two components add to 100, which correspond to the triangles on the 2-part simplex (full line). This means that compositions are equivalence classes of compositionally equivalent arrays. Equivalence classes are handled by selecting a representative of each class and operating with these representatives. The selection of representative of a class is arbitrary, but imposes a condition on any further analysis. This condition is the principle of scale invariance formulated in Aitchison (1986).

Principle 4.2.1 (Scale invariant analysis) *Any analysis or operation with compositions must be expressed by scale invariant functions of the components. Scale invariant functions are identified with real,* 0*-degree homogeneous functions, that is, satisfying the condition* $f(\mathbf{x}) = f(c\mathbf{x})$ *for any positive constant c and for any composition* **x**.

Consequently, for any composition given by the array **x** it is possible to choose another compositionally equivalent array, denoted $\mathcal{C}\mathbf{x}$, such that it is in the simplex, that is, it fulfills the CSC (4.1). To this end, the constant in CSC (4.1) $\kappa = 1$ is chosen, thus yielding

$$\mathcal{C}\mathbf{x} = \left(\frac{x_1}{\sum_{i=1}^{D} x_i}, \frac{x_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{x_D}{\sum_{i=1}^{D} x_i} \right).$$

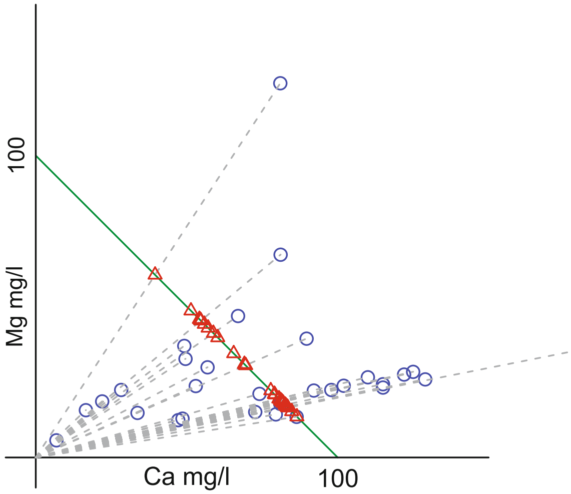


Fig. 4.1 Some two-component data points with positive components (circles), are compositionally equivalent to all points on the dashed lines from the origin through the data points. Triangles are the representatives of each equivalence class on the 2-part simplex in which components add to 100

The symbol $\mathcal{C}$ is called closure operator. It assigns a representative in the simplex (closed form of $\mathbf{x}$, satisfying the CSC) to the equivalence class where $\mathbf{x}$ is included. Due to the scale invariance analysis principle, any analysis on the elements in the simplex (closed) must lead to identical results as that performed using the non-closed representatives.

The scale invariance principle is familiar to any scientist. For instance, an array of probabilities as $(0.1, 0.3, 0.2)$, originally expressed as values between 0 and 1, can be expressed in percentages as $(10, 30, 20)$ without any confusion; a set of concentrations given in percentages of mass can be translated into ppm (parts per million of mass) just multiplying by $10,000$ and the geologist does not get confused provided that he/she is informed about which units are in use.

Despite the intuitive character of the scale invariance principle, in practice it is frequently violated. For instance, when performing a cluster analysis of geochemical samples given in ppm using the Euclidean distance between the samples. In fact, assume that we have two samples $\mathbf{x}$ and $\mathbf{y}$, and the square distance between them is taken as the square-Euclidean distance $\mathrm{d}^2(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{D}(x_i - y_i)^2$. Imagine that $\mathbf{y}$ is now expressed in ppb (parts per billion). This is a valid operation as $\mathbf{y}$ in ppm and in ppb are compositionally equivalent, but $\mathrm{d}^2(\mathbf{x}, \mathbf{y})$ changes dramatically as the square-differences $(x_i - y_i)^2$ become $(x_i - 1000 \cdot y_i)^2$ which constitutes a violation of the scale invariance principle.

Similarly, given a set of geochemical samples in ppm, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, the Pearson correlation coefficient between two components also violates the principle of scale invariance. This coefficient between $x_{\cdot 1}$ and $x_{\cdot 2}$ is

$$r_{12} = \frac{\sum_{j=1}^{n}(x_{j1} - \bar{x}_1)(x_{j2} - \bar{x}_2)}{\sqrt{\sum_{j=1}^{n}(x_{j1} - \bar{x}_1)^2 \sum_{j=1}^{n}(x_{j2} - \bar{x}_2)^2}}, \tag{4.2}$$

where $\bar{x}_k$ is the average of the k-th component along the sample. Now suppose that the first sample $\mathbf{x}_1$ is expressed in ppb. This should not change the analysis as preconized by the scale invariance principle. However, everything changes: the average values $\bar{x}_k = (1/n)\sum_{j=1}^{n} x_{jk}$ are now dominated by the first term $1000 \cdot x_{1k}$ which replaced the initial term x_{1k}. The global effect is evident after a simple inspection of Eq. (4.2). When the change of closure affects all the samples, the effect is the *spurious correlation* studied by Chayes (1960), although without any successful solution. Nowadays, after J. Aitchison's work, spurious correlation just corresponds to a violation of the scale invariance principle. Or, in other words, if a data set is assumed scale invariant, covariance or Pearson correlation are meaningless and spurious, and should not be used.

4.3 The Simplex as Sample Space of Compositions

In any data analysis, the first modeling step is to establish an appropriate sample space. In general, this step conditions all subsequent steps, and may affect dramatically the conclusions. Dealing with compositional data is not an exception. However, the choice and structure of the sample space is usually not explicit, and its consequences remain hidden in practice. Even the analyst is frequently not aware of the choice he or she has made when taking a decision on which methodology to apply.

The sample space of an observation (variable, vector, function or, in general, object) is a set where all the possible outcomes can be represented. However, the sample space may contain elements which do not correspond to any possible observation. When the considered object is a random one, the sample space must contain subsets, called events, which can be assigned a probability. Technically, if $\mathcal{S}$ is the sample space, a σ-field in $\mathcal{S}$ (e.g. Ash 1972; Feller 1968) needs to be defined. This is the minimum structure of a sample space for a random object. There are many qualitatively different random objects in practice. Multivariate real random d-vectors may be thought of as taking values in real space $\mathbb{R}^d$; a discrete time, real valued stochastic process, can be represented in the space ℓ^∞ of all real, bilaterally bounded sequences; if the observation is a random set on a plane, like paint stains on the floor, the sample space can be the set of compact sets in the plane; there are many more examples. It should be noted that the sample space is a choice of the analyst and it must be selected according to the stated questions from the beginning of the analysis. Commonly, beyond probability statements, the data analysis requires performing operations (sums, differences, averages, scaling), metric computations (distances or divergences, projections, approximations), or computing functionals (averages of components, extraction of extremes). All these procedures must be defined on the sample space. Consequently, the structure of the sample space is richer than that provided by the σ-field of events.

When dealing with D-part compositional data, the simplex $\mathbb{S}^D$ as the sample space is a valid choice, given that any composition can be assigned a representative in it. However, there are many alternatives. Figure 4.1 suggests that any curve intersecting once, and only once, all rays from the origin in the positive orthant might be taken as sample space. For instance, for two dimensional data points like those shown in Fig. 4.1, a possible choice is a quarter of a circumference, or two segments completing a square with the axes, as shown in Fig. 4.2. In the case of compositional data, the analyst is mainly interested in proportions and ratios, thus suggesting the choice of the simplex as an appropriate and intuitive representation. However, a key point for the choice of an adequate sample space is the decision on which is a translation or shift relevant for the analysis.

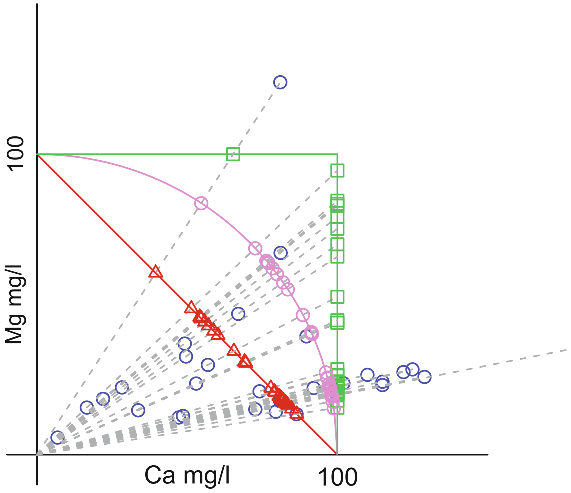


Fig. 4.2 Some two-component data points with positive components (blue circles), are compositionally equivalent to all points on the dashed lines from the origin through the data points. Red triangles are the representatives of each equivalence class on the 2-part simplex in which components add to 100. Violet circles are representatives of these data points on a quarter of circumference. Green squares are representatives on the 100-square

4.4 Perturbation, a Natural Shift Operation on Compositions

Perturbation, as operation in the simplex, was introduced by Aitchison (1986) on an intuitive basis. It can be stated as follows.

Definition 4.4.1 (*perturbation*) Let $\mathbf{x}$, $\mathbf{y}$ be two elements in the D-part simplex $\mathbb{S}^D$, $\mathbf{x} = (x_1, x_2, \ldots, x_D)$, $\mathbf{y} = (y_1, y_2, \ldots, y_D)$. The perturbation between them is

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, x_2 y_2, \ldots, x_D y_D)\,. \tag{4.3}$$

Some properties of perturbation are quite immediate. They can be summarized as that perturbation is a commutative group operation in $\mathbb{S}^D$ (Aitchison 1997). The neutral element is the composition with equal components $\mathbf{n} = \mathcal{C}(1, 1, \ldots, 1)$. The opposite to $\mathbf{x}$ is

$$\ominus \mathbf{x} = \mathcal{C}((1/x_1), (1/x_2), \ldots, (1/x_D)),$$

where each component is inverted.

Repeated perturbation, like $\mathbf{x} \oplus \mathbf{x} \oplus \mathbf{x}$, suggests the definition of a multiplication by a real scalar, so that $\mathbf{x} \oplus \mathbf{x} \oplus \mathbf{x} = 3 \odot \mathbf{x}$. Following this idea, multiplication by real scalars, called powering, is defined as follows.

Definition 4.4.2 (*powering*) Let $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ be an element in the D-part simplex $\mathbb{S}^D$ and let α be a real scalar. The powering of $\mathbf{x}$ by α is

$$\alpha \odot \mathbf{x} = \mathcal{C}(x_1^{\alpha}, x_2^{\alpha}, \ldots, x_D^{\alpha})\,. \tag{4.4}$$

These definitions present perturbation and powering as operations on elements of the simplex. However, as the simplex can be taken as the sample space of compositions and its elements are representatives of compositions, perturbation and powering are also operations on compositions. The simplex, endowed with perturbation and powering is a $(D-1)$-dimensional vector space. Perturbation plays the role of the sum in real space, and powering is multiplication by a real scalar. Perturbing a composition $\mathbf{x}$ by another composition $\mathbf{y}$ is thus a shift of $\mathbf{x}$ in the direction of $\mathbf{y}$.

Despite the mathematical aspect of Definition 4.4.1, perturbation is a common place in real life and scientific activity. To begin with, imagine a water filtering device which is fed with an inflow with disolved matter characterised by the concentrations (mg/l) of the major ions specified in Table 4.1, first row. Suppose that the filtering device has been designed to filter out sulphur, SO_4, iron, Fe, and phosphorus, P; SO_4 is ideally reduced by 75%, Fe by 10%, and P by 5%, meanwhile other ions remain unaltered. In order to compute the outflow concentrations, the filter factor or transfer function (4th row) is computed as $1-(10/100)=0.9$ in the case of Fe. Then, the filter factor multiplies the inflow concentrations to obtain the outflow concentrations in mg/l. Notably, when the inflow concentrations are represented in closed form, as percentages (second row), then, once multiplied by the filter factor, the same outflow concentrations in percent are obtained. In fact, the outflow concentrations in mg/l, when closed to 100, are those in the last row of the table. The closed form of the filter factor, labelled filter perturbation, can be used to obtain the same outflow concentrations. That is the filter factor is a composition. Although elementary, this example shows that inflow and outflow concentrations and the filter factor can be represented by different, but compositionally equivalent, arrays; and that the traditional form of expressing change of concentrations by percentages is nothing else than a way of expressing a perturbation. Also, one may be confronted with the estimation of the filter factor (perturbation) from the inflow and outflow concentrations. From the example, it is clear that a ratio of outflow over inflow concentrations gives a factor compositionally equivalent to the filter perturbation. This suggests the

Table 4.1 Inflow concentrations of some ions disolved in water are filtered reducing Fe, SO_4 and P by a given percentage. Outflow concentrations are obtained by multiplication of inflow concentration by the filter factor (closed or not). Inflow, outflow concentrations and filter factor are presented also in closed form as they are treated as compositions

	Ca	Fe	K	Mg	Na	P	SO_4
Inflow (mg/l)	0.760	0.225	5.30	1.54	2.00	0.079	2.40
Inflow (closed to 100)	6.177	1.829	43.08	12.52	16.25	0.642	19.51
Filter effect (%)	0	−10	0	0	0	−5	−75
Filter factor	1	0.9	1	1	1	0.95	0.25
Filter perturbation	0.164	0.148	0.164	0.164	0.164	0.156	0.041
Outflow (mg/l)	0.760	0.203	5.30	1.54	2.00	0.075	0.60
Outflow (closed to 100)	7.254	1.933	50.58	14.70	19.09	0.716	5.73

definition of the difference-perturbation, the opposite operation to perturbation, as

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

which is the natural difference for perturbation as a group operation.

In the context of probability theory, arrays of probabilities can be considered as compositions. Consider a family of non overlapping events A_i, $i = 1, 2, \ldots, D$, which are assigned probabilities $p_i = \mathrm{P}[A_i]$. Observing the result R of an experiment, the conditional probabilities $q_i = \mathrm{P}[R|A_i]$ allow to update the probabilities p_i —according to the information obtained from the observation R— using Bayes' formula

$$\mathrm{P}[A_i|R] = \frac{\mathrm{P}[A_i] \cdot \mathrm{P}[R|A_i]}{\sum_{j=1}^{D} \mathrm{P}[A_j] \cdot \mathrm{P}[R|A_j]} = \mathcal{C}\left(\mathbf{p} \oplus \mathbf{q}\right) ,$$

where $\mathbf{p} = (p_1, p_2, \ldots, p_D)$ and $\mathbf{q} = (q_1, q_2, \ldots, q_D)$. Bayes' formula states that the final probabilities, conditioned to the result R, are the perturbation of the initial or prior probabilities $\mathbf{p}$ and the probabilities of the result given the events A_i, denoted q_i, also known as the likelihood of R. In this way perturbation becomes a very natural way of operating vectors of probabilities and likelihood, as it is the paradigm of incorporating information from observations. This interpretation of perturbation was proposed in Aitchison (1986, 1997) and developed in other contexts (Egozcue and Pawlowsky-Glahn 2011b; Egozcue et al. 2013).

Perturbation also appears as a natural operation on compositions when changing units. For instance, consider a grain size distribution for different sieve diameters. It may be expressed as proportions of volume corresponding to each sieve or as proportions of mass assigned to the same sieves. Both distributions can be considered as compositions. Transforming volume to mass consists of multiplication by the density of the material in each sieve, possibly different from one sieve to the other. This componentwise multiplication is a perturbation (Parent et al. 2012). Also, changing the concentrations of chemical elements from mg/kg to molar concentration consists of dividing each component by its molar mass, thus performing a perturbation. In all these examples, the secondary role of the closure and the CSC is remarkable: closure might only be necessary to facilitate interpretation.

Exponential decay of mass is frequent in nature. The typical example is the decay of mass of radioactive isotopes in time. These type of processes describe straight lines in the simplex (Egozcue et al. 2003; Pawlowsky-Glahn et al. 2015; Tolosana-Delgado 2012). This supports that perturbation is a natural operation in the simplex and between compositions. To sketch the argument, consider the masses of $D = 3$ fictitious radioactive isotopes $\mathbf{x}(t) = (x_1(t), x_2(t), x_3(t))$, which decay rates in time are $\lambda_1 = 3$, $\lambda_2 = 0.5$, $\lambda_3 = 0.1$, respectively. Initially, at $t = 0$, there are masses $\mathbf{x}(0) = (0.9, 0.04, 0.01)$ which disintegrate into other non considered isotopes. The total mass decreases in time, and the mass of each isotope changes as

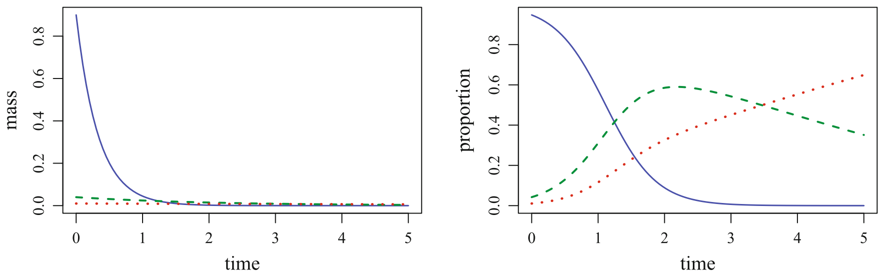
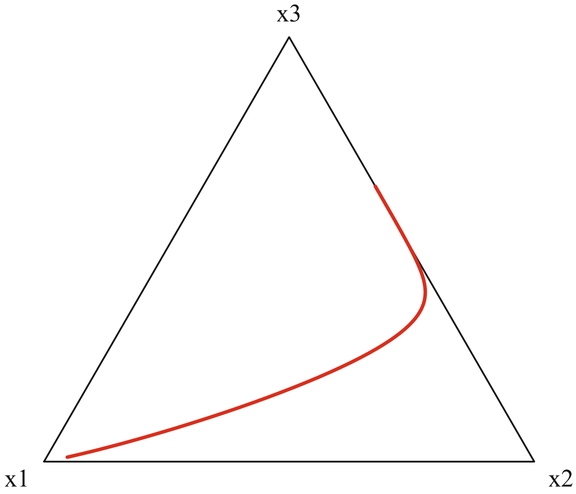


Fig. 4.3 Evolution of masses (left panel) and proportions (right panel) of three isotopes which disintegrate at rates $3, 0.5, 0.1$ in time, respectively. Initial masses are $0.9, 0.04, 0.01$

Fig. 4.4 Evolution of proportions in time of three isotopes which disintegrate at rates $3, 0.5, 0.1$, respectively, represented in a ternary diagram. The initial masses are $0.9, 0.04, 0.01$, and they change as a function of time



$$x_i(t) = x_i(0) \cdot \exp[-\lambda_i\, t] , \quad i = 1, 2, 3 . \tag{4.5}$$

This evolution of mass is shown in Fig. 4.3, left panel, where the decreasing mass is clearly observed. Figure 4.3, right panel, shows the evolution of proportions of the isotopes after the closure, which corresponds to

$$\mathcal{C}\mathbf{x}(t) = \mathcal{C}\,(\mathbf{x}(0) \oplus (-t \odot \exp[\lambda])) , \tag{4.6}$$

where $\exp[\lambda] = (\exp(\lambda_1), \exp(\lambda_2), \exp(\lambda_3))$. Figure 4.4 shows the evolution of the isotopes in a ternary diagram. The main fact on this exponential decay of isotopes is that it is naturally expressed using perturbation and powering, as in Eq. (4.6). In the simplex, this compositional evolution is a linear one. If proportions are thought as real variables, as they are shown in Fig. 4.3 (right panel), or in Fig. 4.4, then they are taken as non-linear thus ignoring their simplicity as compositional evolution.

The fact that perturbation is easily interpreted on vectors of proportions supports the idea that the simplex is a suitable sample space for compositions. Think, for instance, how perturbation could be interpreted when taking representatives of compositions as projections on the positive orthant of a hypersphere, or on the surface of a unit hypercube. It is not intuitive at all. Obviously, if the operation that is considered relevant for the stated problem is a rotation, the representation on the hypersphere may be a sensible choice of sample space.

4.5 Conditions on Metrics for Compositions

In many applications a distance between data points is a central issue. Cluster analysis is a typical example of this. Other metric concepts are crucial, like the size of a vector, the norm, or the possibility of performing orthogonal projections. Note that all these metric concepts are used in the omnipresent regression analysis. Compositional data analysis has the same need of introducing metrics, distances, norms and orthogonality. From the early developments by J. Aitchison (1983), a distance between compositions was introduced and developed (Aitchison 1992; Aitchison et al. 2000). Nowadays, that distance between compositions is called Aitchison distance, and the corresponding Euclidean geometry is named Aitchison geometry (Pawlowsky-Glahn and Egozcue 2001).

The need of a distance between compositions can be motivated from the most basic statistics. For instance, concepts as elementary as mean and variance are based on a choice of a distance in the sample space. Following Fréchet (1948) (see also Pawlowsky-Glahn et al. 2015, Chap. 6), mean and variance of a sample can be introduced in a metric space (sample space endowed with a distance). Consider a compositional sample $\mathbf{x}_i$, $i = 1, 2, \ldots, n$, represented in the D-part simplex $\mathbb{S}^D$. The data matrix $\mathbf{X}$ has the compositions $\mathbf{x}_i$ as rows. Suppose that a distance in $\mathbb{S}^D$ is $\mathrm{d}_a(\cdot, \cdot)$ (this notation corresponds to the Aitchison distance, although here it is used in a generic sense). A first step is to define variability of the sample with respect to a given composition $\mathbf{z}$ as

$$\mathrm{Var}[\mathbf{X}, \mathbf{z}] = \frac{1}{n} \sum_{i=1}^{n} \mathrm{d}_a^2(\mathbf{x}_i, \mathbf{z}) , \quad \mathbf{z} \in \mathbb{S}^D . \tag{4.7}$$

The sample mean, called center for compositions, and the total variance are then defined as

$$\mathrm{Cen}[\mathbf{X}] = \operatorname*{argmin}_{\mathbf{z} \in \mathbb{S}^D} \{\mathrm{Var}[\mathbf{X}, \mathbf{z}]\} , \tag{4.8}$$

$$\mathrm{totVar}[\mathbf{X}] = \min_{\mathbf{z} \in \mathbb{S}^D} \{\mathrm{Var}[\mathbf{X}, \mathbf{z}]\} = \mathrm{Var}[\mathbf{X}, \mathrm{Cen}[\mathbf{X}]] . \tag{4.9}$$

Equations (4.7), (4.8) and (4.9) show that elementary statistics like mean and variance depend critically on the distance used in the sample space.

The Aitchison distance can be defined in different ways (see Pawlowsky-Glahn et al. 2015). One of them is

$$\mathrm{d}_a^2(\mathbf{x},\mathbf{y}) = \frac{1}{2D}\sum_{j=1}^{D}\sum_{k=1}^{D}\left(\ln\frac{x_j}{x_k} - \ln\frac{y_j}{y_k}\right)^2 , \tag{4.10}$$

where it is worth to realize that $\ln(x_k/x_k) = 0$. The distance has been subscripted as d_a to emphasize that it is the Aitchison distance. The first observation on the Aitchison distance is that it is scale invariant, as required by Principle 4.2.1. In fact, any multiplicative constant in $\mathbf{x}$ or $\mathbf{y}$ cancels out in the log-ratios in Eq. (4.10). After accepting the Aitchison distance as a proper one for compositions, a simple but tedious computation drives us to the expression of the sample center

$$\mathrm{Cen}[\mathbf{X}] = \frac{1}{n}\odot\bigoplus_{i=1}^{n}\mathbf{x}_i ,$$

where $\bigoplus$ stands for repeated perturbation, similar to a summation for real addition. At a first glance, just dropping the circles in the signs $\oplus$ and $\odot$, this expression is an average where the traditional sum has been changed to perturbation. Thus, the computation of Cen[$\mathbf{X}$] consists of computing the geometric mean of the columns of $\mathbf{X}$ and closing the resulting vector if a representation on the simplex is desired.

An interesting question is which are desirable and intuitive properties of a metric (distance, norm, inner product) for compositions. Our geometric intuition comes from our experience in the Euclidean space $\mathbb{R}^3$ and we try to translate these observations to a geometry of the simplex. In this way, if we have a rigid object on the table and we move this to another position, for instance on the floor, we expect that distances between points of the object are equal to those observed previous to the movement. Also, we observe that projecting a segment on the floor ($\mathbb{R}^2$), perhaps the edge of a roof, produces a segment with length shorter than the original one. If the points delimiting the segment are expressed in Cartesian coordinates, x and y, on the floor, and z vertical or orthogonal to the floor, the projection of the points consists in suppressing the z-coordinate. That is, our experience tells us that suppressing coordinates makes the resulting projected distances shorter than or equal to the original ones. Being a little bit more subtle, we realize that suppressing the z-coordinate is a special projection (orthogonal projection), but there are other kinds of projections. For instance, the shadow projected by the edge of the roof on the floor may be larger than the length of the edge depending on the position of the sun. This is because the shadow is not an orthogonal projection unless the floor is tilted orthogonal to the sun rays. These daily experiences with Euclidean geometry may inspire the following properties of the geometry in the simplex that we take as requirements.

A. **Equidistance on shift**: The distance between two compositions $\mathbf{x}_1$ and $\mathbf{x}_2$ in $\mathbb{S}^D$ is equal to their distance after a shift $\mathbf{z}$, that is

$$\mathrm{d}_a(\mathbf{x}_1 \oplus \mathbf{z}, \mathbf{x}_2 \oplus \mathbf{z}) = \mathrm{d}_a(\mathbf{x}_1, \mathbf{x}_2) \; ; \tag{4.11}$$

B. **Dominance on subcompositions**: From the composition $\mathbf{x}_k = (x_{k1}, x_{k2}, \dots, x_{kD})$, a subcomposition $\mathbf{x}_k^{sub}$ is extracted by suppressing some components, for instance, $\mathbf{x}_k^{sub} = (x_{k1}, x_{k2}, \dots, x_{kd})$, with $D > d > 1$. Then, for $k = 1, 2$, compositional distance should satisfy $\mathrm{d}_a(\mathbf{x}_1, \mathbf{x}_2) \geq \mathrm{d}_a(\mathbf{x}_1^{sub}, \mathbf{x}_2^{sub})$;

C. **Subcomposition as orthogonal projection**: The geometry on the simplex is an Euclidean geometry, that is, there is an inner product from which the norm and distance derive. Particularly, geometry on subcompositions in $\mathbb{S}^d$, $D > d > 1$, is equivalent to that of the orthogonal projection of $\mathbb{S}^D$ onto $\mathbb{S}^d$.

Point A is essential for defining sensible elementary statistics as shown in Eqs. (4.8) and (4.9). To show the importance of this property a subset of water analyses in Bangladesh has been selected. It comes from a survey conducted in the 1990s as a joint effort by the British Geological Survey and the Department of Public Health Engineering of Bangladesh (British Geological Survey 2001a, b). The subset, called hereafter Northern Bangladesh data, includes 13 disolved ions in Northern Bangladesh (latitude greater than 26 °N) and has been selected with the only purpose to serve as illustration. This data set was also used in several studies (see Pawlowsky-Glahn et al. 2015 and references therein). Concentrations of As, Fe and P (mg/l) are shown in a ternary diagram (Fig. 4.5). In the left panel they appear close to the border Fe-P due to the small concentrations of As relative to Fe and P. Right panel of Fig. 4.5 shows the same data set after centering it, that is $\mathbf{X} \ominus \mathrm{Cen}[\mathbf{X}]$. Now details are made visible; for instance, the rounding of As to 1 μg/l is now visible in form of straight bands extending from the Fe vertex. Although the aspect of the data points is more disperse in the left panel than the right one, the total variance is equal in the two representations, as perturbation does not change the total variance; that is, $\mathrm{totVar}[\mathbf{X}] = \mathrm{totVar}[(\mathbf{X} \ominus \mathrm{Cen}[\mathbf{X}])]$. This points out the inconvenience of using the visual distance (Euclidean distance) in the ternary diagram.

Requirement B is a consequence of point C, and is to be discussed at the end of this section. Requirement C is a bit technical but is again inspired by the real multivariate geometry. Suppose that a sample of d real variables has been observed and the corresponding data set is arranged in an (n, d) matrix. One may be interested in a multiple scatter-plot of each couple of variables, similar to that shown in Fig. 4.6. The fact that the axes of such plots are perpendicular does not surprise anybody. The assumption is that adding a real variable to a previous set is naturally represented by adding a new coordinate on an axis orthogonal to the previous ones.

Requirement C is implicitly claiming for an orthogonality relation, usually given by an inner product between compositions, namely $\langle \mathbf{x}, \mathbf{y} \rangle_a$, where $\mathbf{x}$ and $\mathbf{y}$ are compositions represented in the same simplex, say $\mathbb{S}^D$. From this inner product two

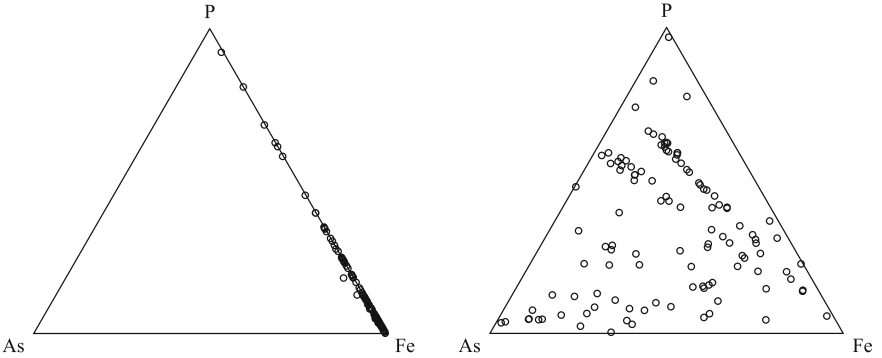
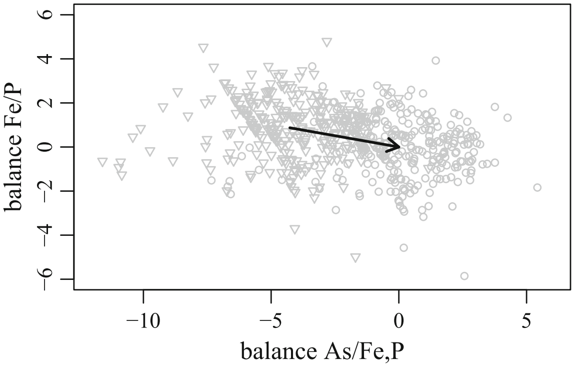


Fig. 4.5 Disolved As, Fe, P data set. Left panel, data expressed in mg/l. Right panel, same data after centering

Fig. 4.6 Disolved As, Fe, P data set represented in orthonormal coordinates. Triangles: original data; Circles: centered data. The arrow indicates the centering perturbation and it is anchored in the sample mean of coordinates



compositions are orthogonal if they satisfy $\langle \mathbf{x}, \mathbf{y} \rangle_a = 0$. All metric elements can be derived from the inner product. The square-norm (square size) is $\|\mathbf{x}\|_a^2 = \langle \mathbf{x}, \mathbf{x} \rangle_a$; and square-distance is $\mathrm{d}_a^2(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_a^2$. A general property of Euclidean spaces (Queysanne 1973) is that there exists an orthonormal basis constituted by $D-1$ compositions $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}$. Orthonormal coordinates are then computed as

$$\phi_k(x_1, x_2, \ldots, x_D) = \langle \mathbf{x}, \mathbf{e}_k \rangle_a, \quad k = 1, 2, \ldots, D-1,$$

and, consequently,

$$\|\mathbf{x}\|_a^2 = \sum_{k=1}^{D-1} \phi_k^2(x_1, x_2, \ldots, x_D).$$

The question is which form can the coordinates ϕ_k take, so that they satisfy requirements A, B, C, and so that they are compatible with perturbation and powering. These latter conditions lead to the following additional requirement.

D. The coordinates in $\mathbb{S}^D$, ϕ_k, $k = 1, 2, \ldots, D-1$ satisfy

$$\phi_k(\mathbf{x} \oplus (\alpha \odot \mathbf{y})) = \phi_k(\mathbf{x}) + \alpha \cdot \phi_k(\mathbf{y}) \,, \tag{4.12}$$

for any compositions $\mathbf{x}, \mathbf{y}$, and any real constant α.

From requirements A and D, the ϕ_k can be deduced. Consider first a two part subcomposition of $\mathbf{x}$, denoted $\mathbf{x}^{(2)}$. These subcompositions constitute a Euclidean space of dimension 1, and two part compositions can be represented by a single coordinate $\phi_1 = \phi_1(x_1^{(2)}, x_2^{(2)})$. This function must be scale invariant and such that it can take all real values. A simple log-ratio, $\phi_1 = a_1 \ln(x_1^{(2)}/x_2^{(2)})$, where a_1 is a real constant to be determined, is a possible choice. The ratio argument within the logarithm guarantees scale invariance, and the logarithm allows ϕ_1 to range over all real numbers. The superscripts denoting the number of parts of the subcomposition are superfluous due to the scale invariance property and, from now on, it is assumed that $x_i^{(k)} = x_i$, being the latter the value of the i-th component in the large composition $\mathbf{x}$.

Consider now a 3-part subcomposition $\mathbf{x}^{(3)} = (x_1, x_2, x_3)$ in a 2-dimensional subspace which includes subcompositions $\mathbf{x}^{(2)}$, that is $(x_1^{(3)}, x_2^{(3)}) = (x_1, x_2)$. The additional dimension corresponds to a new coordinate ϕ_2 in an orthogonal direction to that ϕ_1 as proposed by requirement C. Again this coordinate needs to be scale invariant and taking any real value. A simple choice can be $\phi_2 = a_2 \ln(x_3/\mathrm{g_m}(\mathbf{x}^{(2)}))$ where $\mathrm{g_m}$ denotes geometric mean of the arguments. Iterating the reasoning for increasing number of parts of the subcomposition the k-th coordinate takes the form

$$\phi_k = a_k \ln \frac{x_{k+1}}{\mathrm{g_m}(\mathbf{x}^{(k)})} \,, \quad k = 1, 2, \ldots, D-1 \,.$$

These expressions for the coordinates fulfill conditions A–D.

The inner product in a Euclidean space can be expressed using Cartesian coordinates as

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \sum_{k=1}^{D-1} \phi_k \psi_k \,, \tag{4.13}$$

where ϕ_k and ψ_k are the coordinates of the D-part compositions $\mathbf{x}$, $\mathbf{y}$ respectively. A tedious exercise consists of substituting the expression of the coordinates in Eq. (4.13) and carrying out the sum for values of a_k such that all components of $\mathbf{x}$, $\mathbf{y}$ appear in a symmetric way. Up to a multiplicative constant, the result is

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \sum_{j=1}^{D} \ln \frac{x_j}{\mathrm{g_m}(\mathbf{x})} \, \ln \frac{y_j}{\mathrm{g_m}(\mathbf{y})} \,, \quad a_j = \sqrt{\frac{j}{j+1}} \,,$$

where the a_js appear as normalizing constants homogenizing the scale of the different axes. The inner product $\langle \mathbf{x}, \mathbf{y} \rangle_a$ is the ordinary inner product of the $\mathbb{R}^D$ vectors $\mathrm{clr}(\mathbf{x})$ and $\mathrm{clr}(\mathbf{y})$, which are

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

and analogously for $\mathrm{clr}(\mathbf{y})$.

The square Aitchison distance expressed in coordinates is the ordinary Euclidean distance in $\mathbb{R}^{D-1}$, which can be compared to the expression using the clr coefficients in $\mathbb{R}^D$:

$$\mathrm{d}_a^2(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{D-1} (\phi_k - \psi_k)^2 = \sum_{j=1}^{D} (\mathrm{clr}_j(\mathbf{x}) - \mathrm{clr}_j(\mathbf{y}))^2 . \tag{4.14}$$

Requirement B on dominance of distance of a subcomposition is now evident. From the expression of the distance in coordinates (Eq. 4.14, central term), computing distances within a subcomposition consists of removing some positive terms from the sum.

Apparently, there are many possible choices for the form of coordinates ϕ_k, but most of them are discarded by requirements A and D on compatibility with perturbation (Eqs. 4.11, 4.14). For instance, $\phi_k = \ln(x_{k+1}/(x_1 + x_2 + \cdots + x_k))$, implicitly proposed in Aitchison (1986), Sect. 10.3, does not lead to a distance and coordinate expressions satisfying A and D. The critical point is that amalgamation or sum of compositional parts is not a linear operation for compositions.

Figure 4.6 shows the sample of disolved As, Fe, P previously represented in Fig. 4.5 in ilr-coordinates. These coordinates are the balances

$$\phi_1 = \sqrt{\frac{2}{3}} \ln \frac{\mathrm{As}}{(\mathrm{Fe} \cdot \mathrm{P})^{(1/2)}} , \quad \phi_2 = \sqrt{\frac{1}{2}} \ln \frac{\mathrm{Fe}}{\mathrm{P}} .$$

The visual distances between the data points are now the Aitchison distances. The triangles correspond to the original data set. Its center, expressed in coordinates, is the point where the arrow is anchored. A shift (perturbation) is applied in order to center the data set (circles), so that the new center is the origin of coordinates (end of the arrow). Importantly, the distances between data points after shifting (requirement A) are equal to the previous ones. The fact that the axes are drawn orthogonally, exactly corresponds to the fact that these coordinates are orthogonal in the Aitchison geometry for compositional data.

The historical way of defining the centered log-ratio transformation of $\mathbf{x}$ and the whole structure was the reverse of the one here presented. The definitions of perturbation, powering and clr can be found in Aitchison (1986), although the Aitchison distance was already introduced in Aitchison (1983) and discussed in Aitchison et al. (2000). The inner product as such, and the corresponding Euclidean space structure (Aitchison geometry), was introduced independently in Pawlowsky-Glahn and

Egozcue (2001), and in Billheimer et al. (2001), although there is a previous definition of orthogonal log-contrasts in Aitchison (1986). Orthogonal coordinates were introduced in Egozcue et al. (2003), and in Egozcue and Pawlowsky-Glahn (2005).

4.6 Consequences of the Aitchison Geometry in the Sample Space of Compositional Data

The consequences of the Euclidean character of the Aitchison geometry for compositional data are multiple and relevant. Once the principles and requirements on the sample space are assumed, they appear as a guidance in most, if not all, statistical models. The main idea is that compositions are advantageously represented as vectors in coordinates, better than as proportions. Standard operations, sum and multiplication, on appropriate coordinates are equivalent to perturbation and powering on compositions in the simplex. The fact that Aitchison distances, norms and orthogonal projections are transformed into the ordinary Euclidean distances, norms and orthogonal projections opens the door to use on ilr coordinates all mathematical and statistical methods designed for real variables. The recommendation of working on coordinates has been formulated as *the principle of working on coordinates* (Mateu-Figueras et al. 2011). The specific exploratory tools for compositional data are examples of the usefulness of ilr coordinates.

Principal component analysis for compositional data (CoDa-PCA) and its graphical representation, the CoDa-biplot, were studied before ilr-coordinates were available (Aitchison 1983; Aitchison and Greenacre 2002), but they are a wonderful example of their usefulness. A D-part compositional data set, $\mathbf{X}$ in a (n, D)-matrix, is clr-transformed and centered; then, the singular value decomposition is carried out. This can be summarized as

$$\mathrm{clr}(\mathbf{X}_c) = \mathrm{clr}(\mathbf{X} \ominus \mathbf{1}_n \mathrm{Cen}[\mathbf{X}]) = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top , \tag{4.15}$$

where clr is applied to each composition (row) of the centered matrix, and $\mathbf{1}_n$ is a column vector of n ones. The diagonal matrix $\boldsymbol{\Lambda}$ contains $D-1$ singular values ordered from the largest one to the smallest. The D-th singular value is always null, since the rows of $\mathrm{clr}(\mathbf{X}_c)$ add to zero, and can be removed. The $(D, D-1)$-matrix $\mathbf{V}$ (loadings matrix), once the last column corresponding to the null singular value is removed, is orthogonal and satisfies $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_{D-1}$, $\mathbf{V}\mathbf{V}^\top = \mathbf{I}_D - (1/D)\mathbf{1}_D\mathbf{1}_D{}^\top$. Therefore, it is a contrast matrix like that used to compute ilr-coordinates of a composition $\mathbf{x}$ (column vector) (Egozcue et al. 2011)

$$\mathbf{z} = \mathrm{ilr}(\mathbf{x}) = \mathbf{V}^\top \mathrm{clr}(\mathbf{x}) , \quad \mathbf{x} = c \cdot \exp[\mathbf{V}\mathbf{z}] .$$

This means that the rows of the $(n, D-1)$-matrix $\mathbf{U}\boldsymbol{\Lambda}$ are ilr-coordinates of the centered compositional data set. A form biplot represents simultaneously the rows of

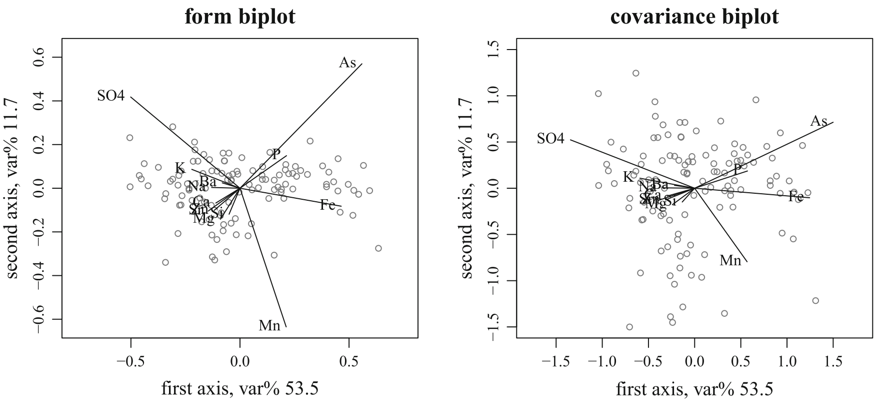


Fig. 4.7 Biplots of Northern Bangladesh data set, representing 13 disolved ions. Left: form biplot showing that the projection is mainly dominated by the clr coefficients of As, Mn, and SO_4; up to the projection (65.2% of total variance), Aitchison distances between data points are approximately those visualized. Right: covariance biplot adequate for interpretation. Up to the projection, length of links between vertices of rays are proportional to the standard deviation of the corresponding logratio. The length of the rays are approximately proportional to the standard deviation of the corresponding clr-coefficients. Variability is largely dominated by the log ratios of SO_4 over As, Fe and Mn

$\mathbf{U}\boldsymbol{\Lambda}$ (coordinates of the compositions) and the columns of $\mathbf{V}$ (clrunitary vectors of the ilr-basis) in an optimal bi-dimensional projection for visualization.

Figure 4.7 shows the form biplot of the Northern Bangladesh data set. Form biplots (Fig. 4.7, left) and scatter-plots of coordinates (Fig. 4.6) can replace plots on ternary diagrams, as distances between compositions are not distorted in an uncontroled manner. They are only affected by the orthogonal projections.

The ilr coordinates are real variables and their exploratory analysis relies on standard exploratory analysis tools (mean, standard deviation, quantiles, correlations). However, interpretable coordinates are desirable. They can be designed by the analyst to get insight in some aspects of the data he/she may be interested in. Other times a data driven technique may be used to design suitable coordinates (Pawlowsky-Glahn et al. 2011; Martín-Fernández et al. 2017). In these cases, the CoDa-dendrogram (Pawlowsky-Glahn and Egozcue 2011) can be useful to summarize properties of the coordinate sample jointly with an interpretable description of the coordinates used. The definition of the coordinates is based on a sequential binary partition (SBP) of the parts of the composition (Egozcue and Pawlowsky-Glahn 2005, 2006). Each coordinate is associated with a partition of a group of parts into two new groups. For instance, Table 4.2 shows this kind of partitions for the Northern Bangladesh data set. The second row of Table 4.2, indicates the separation of As (+1) from the group constituted by Fe, Mn and P (−1). This separation is associated with the second ilr coordinate

Table 4.2 Sign code for a SBP of the 13 disolved ions, obtained by clustering variables of the Northern Bangladesh data set

As	Ba	Ca	Fe	K	Mg	Mn	Na	P	Si	SO4	Sr	Zn
+1	−1	−1	+1	−1	−1	+1	−1	+1	−1	−1	−1	−1
+1	0	0	−1	0	0	−1	0	−1	0	0	0	0
0	0	0	−1	0	0	+1	0	−1	0	0	0	0
0	0	0	+1	0	0	0	0	−1	0	0	0	0
0	−1	−1	0	−1	−1	0	−1	0	−1	+1	−1	−1
0	0	0	0	0	0	0	+1	0	−1	0	+1	−1
0	0	0	0	0	0	0	+1	0	0	0	−1	0
0	0	0	0	0	0	0	0	0	+1	0	0	−1
0	−1	−1	0	−1	−1	0	+1	0	+1	0	+1	+1
0	−1	−1	0	+1	−1	0	0	0	0	0	0	0
0	+1	−1	0	0	−1	0	0	0	0	0	0	0
0	0	+1	0	0	−1	0	0	0	0	0	0	0

$$z_2 = \sqrt{\frac{3}{4}} \ln \frac{\mathrm{As}}{(\mathrm{Fe} \cdot \mathrm{Mn} \cdot \mathrm{P})^{1/3}} \, .$$

These kinds of coordinates are called balances between two groups of parts (Egozcue and Pawlowsky-Glahn 2005) as they are logratios of the geometric mean of the elements in each group; the coefficient in front of the logarithm is a normalization coefficient which takes into account the number of elements in each group of parts. Figure 4.8 shows the CoDa-dendrogram for the Northern Bangladesh data set. The tree-dendrogram itself follows the partition in Table 4.2. The length of the lines perpendicular to the labels, say vertical lines, are proportional to the variance of the balance separating the groups of elements at left and right hand sides. These vertical lines are anchored to horizontal segments joining the two groups of parts. All these segments are scaled in such a way that the zero value is placed in the center of the segment, and the length represents the same length in all cases. The fulcrum of the vertical line is placed at the average value of the balance; it can be compared to the median indicated in the box-plot under the horizontal line. In this way, the CoDa-dendrogram combines the interpretation of the balance-coordinates given by the SBP and their mean, variance and quantiles (box-plots).

In Fig. 4.8, the variances within the subcomposition (Zn, Si, Sr, Na, SO_4) are small compared to other variances, thus pointing out a possible compositional association between these elements; it suggests that these elements change proportionally along the considered sample. At the same time, most of the total variance is driven by As, Fe, Mn and P, as indicated by longer vertical lines.

The explanatory power of the CoDa-biplot and the CoDa-dendrogram relies on the fact that they are based on Cartesian coordinates for plotting data-points and that the represented variables are orthonormal in a geometric sense. The key in interpret-

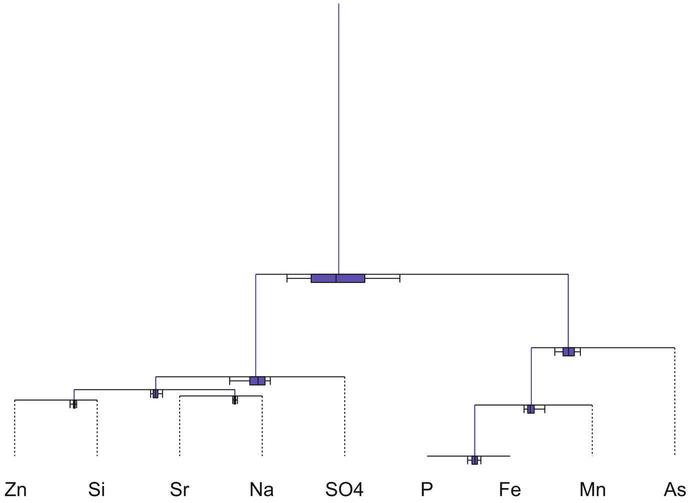


Fig. 4.8 CoDa-dendrogram following the sign code in Table 4.2 obtained by clustering variables of the Northern Bangladesh data set. Vertical bars describe the decomposition of the total variance given in Eq. (4.16). Anchoring points of vertical bars indicate the mean value of the corresponding coordinate

ing the results is the decomposition of the total variance of the data set into variances of the ilr-coordinates (Egozcue and Pawlowsky-Glahn 2011a)

$$\text{totVar}[\mathbf{X}] = \sum_{k=1}^{D-1} \text{Var}[\phi_k] \,. \tag{4.16}$$

4.7 Conclusions

The first step in any data modelling is to establish a sample space able to give answers to the questions stated by the analyst. If these questions involve probabilistic statements, the sample space needs a sigma field of events for which probabilities can be defined. However, most analysts search for statements implying operations, distances, projections between data points or variables. All these concepts need to be defined in the sample space for useful computations and interpretations. These definitions are not intrinsic, but are adapted to the questions stated by the analyst in a subjective way. Therefore, the choice of a sample space has always a subjective character, which is only validated by the ability in giving useful answers to sound questions.

Compositional data require defining a sample space with a rich structure. The log-ratio approach to the analysis of compositional data is based on a set of principles and conditions. The approach here presented is a modification of the standard principles introduced by J. Aitchison in the eighties and reformulated afterwards. Scale invariance and compositional equivalence are maintained exactly as they were introduced, but additional conditions are to be discussed in relation to perturbation, which is assumed to be the main operation between compositions. The Euclidean structure of compositional data represented in the simplex, called Aitchison geometry, is here motivated using the idea that reduction to a subcomposition should be an orthogonal projection.

The Aitchison geometry is thought as a powerful mathematical tool which consistently completes the previous Aitchisonian ideas on the log-ratio approach. The main points are the conception of compositions as equivalence classes (Barceló-Vidal and Martín-Fernández 2016) thus overcoming the early definitions based on the constant sum constraint; and the introduction of coordinates in the Aitchison geometry (Pawlowsky-Glahn and Egozcue 2001; Egozcue et al. 2003; Egozcue and Pawlowsky-Glahn 2005) thus overcoming the idea that taking log-ratios is just a transformation which circumvents the constant sum constraint.

Acknowledgements The research on compositional data analysis has been continuously supported by the *International Association for Mathematical Geosciences* (formerly *for Mathematical Geology*), *IAMG*. The authors appreciate this support, in many cases unconditional, and recognize that the development of the compositional data analysis would have been a lot harder without it. This research has been funded by the Spanish Ministry of Economy and Competitiveness under the project CODA-RETOS/TRANS-CODA (Ref: MTM2015-65016-C2-1/2-R).

References

Aitchison J (1982) The statistical analysis of compositional data (with discussion). J R Stat Soc Ser B (Stat Methodol) 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, p 416. Chapman & Hall Ltd, London (Reprinted in 2003 with additional material by The Blackburn Press)

Aitchison J (1992) On criteria for measures of compositional difference. Math Geol 24(4):365–379

Aitchison J (1997) The one-hour course in compositional data analysis or compositional data analysis is simple. In Pawlowsky-Glahn V (ed) Proceedings of IAMG'97, pp 3–35 Barcelona (E). CIMNE, Barcelona, Spain ISBN 978-84-87867-76-7

Aitchison J, Barceló-Vidal C, Martín-Fernández JA, Pawlowsky-Glahn V (2000) Logratio analysis and compositional distance. Math Geol 32(3):271–275

Aitchison J, Egozcue JJ (2005) Compositional data analysis: where are we and where should we be heading? Math Geol 37(7):829–850

Aitchison J, Greenacre M (2002) Biplots for compositional data. J R Stat Soc, Ser C 51(4):375–392

Aitchison J, Shen S (1980) Logistic-normal distributions. Some properties and uses. Biometrika 67(2):261–272

Ash RB (1972) Real analysis and probability. Academic Press Inc, New York, NY (USA), p 476

Barceló-Vidal C, Martín-Fernández JA (2016) The mathematics of compositional analysis. Aust J Stat 45:57–71

Barceló-Vidal C, Martín-Fernández JA, Pawlowsky-Glahn V (2001) Mathematical foundations of compositional data analysis. In: Ross G (ed) Proceedings of IAMG'01 – The VII annual conference of the international association for mathematical geology, p 20. Cancun (Mex)

Billheimer D, Guttorp P, Fagan W (2001) Statistical interpretation of species composition. J Am Stat Assoc 96(456):1205–1214

British Geological Survey (2001a) Arsenic contamination of groundwater in Bangladesh. Technical Report (WC/00/019), Dep. of Public Health Engineering (Bangladesh). p 630

British Geological Survey (2001b). Arsenic contamination of groundwater in bangladesh: data. Technical Report (WC/00/019), Dep. of Public Health Engineering (Bangladesh)

Chayes F (1960) On correlation between variables of constant sum. J Geophys Res 65(12):4185–4193

Chayes F (1962) Numerical correlation and petrographic variation. J Geol 70(4):440–452

Connor RJ, Mosimann JE (1969) Concepts of independence for proportions with a generalization of the Dirichlet distribution. J Am Stat Assoc 64(325):194–206

Egozcue JJ (2009) Reply to "On the Harker variation diagrams;..." by J. A. Cortés. Math Geosci 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. See Pawlowsky-Glahn Buccianti, pp 141–157, p 378

Egozcue JJ, Pawlowsky-Glahn V (2005) Groups of parts and their balances in compositional data analysis. Math Geol 37(7):795–828

Egozcue JJ, Pawlowsky-Glahn V (2006) Simplicial geometry for compositional data. Compositional data analysis in the geosciences: from theory to practice, vol 264, pp 145–159, Special Publication, Geological Society, London

Egozcue JJ, Pawlowsky-Glahn V (2011a) Basic concepts and procedures. See Pawlowsky-Glahn and Buccianti, pp 12–28

Egozcue JJ, Pawlowsky-Glahn V (2011b) Evidence information in bayesian updating. In: Egozcue JJ, Tolosana–Delgado R, Ortego MI (eds.)Proceedings of CoDaWork-2011, Sant Feliu de Guixols, Girona, Spain

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric logratio transformations for compositional data analysis. Math Geol 35(3):279–300

Egozcue JJ, Pawlowsky-Glahn V, Tolosana-Delgado R, Ortego MI, van den Boogaart KG (2013) Bayes spaces: use of improper distributions and exponential families. Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales, Serie A. Matemáticas (RACSAM) 107:475–486. https://doi.org/10.1007/s13398-012-0082-6

Feller W (1968) An introduction to probability theory and its applications, p 501 1950 (1st edn.), 1968 (3rd edn.), Vol I. Wiley, New York, NY (USA)

Ferrer-Rossell B, Coenders G, Mateu-Figueras G, Pawlowsky-Glahn V (2016) Understanding low-cost airline users' expenditure patterns and volume. Tour Econ 22(2):269–291

Fréchet M (1948) Les éléments Aléatoires de Nature Quelconque dans une Espace Distancié. Annales de l'Institut Henri Poincaré 10(4):215–308

Martín-Fernández JA, Barceló-Vidal C, Pawlowsky-Glahn V (2003) Dealing with zeros and missing values in compositional data sets using nonparametric imputation. Math Geol 35(3):253–278

Martín-Fernández J-A, Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2017) Principal balances for compositional data. Math Geosci under review

Mateu-Figueras G, Pawlowsky-Glahn V, Egozcue JJ (2011) The principle of working on coordinates. See Pawlowsky-Glahn and Buccianti, pp 31–42

Olea RA, Luppens JA, Egozcue JJ, Pawlowsky-Glahn V (2016) Calorific value and compositional ultimate analysis with a case study of a Texas lignite. J Coal Geol 162:27–33

Parent LE, de Almeida CX, Hernandes A, Egozcue JJ, Gülser C, Bolinder MA, Kätterer T, Andrén O, Parent SE, Anctil F, Centurion JF, Natale W (2012) Compositional analysis for an unbiased measure of soil aggregation. Geoderma 179–180:123–131

Pawlowsky-Glahn V, Buccianti A (eds) (2011) Compositional data analysis: theory and applications, p 378. Wiley

Pawlowsky-Glahn V, Egozcue JJ (2001) Geometric approach to statistical analysis on the simplex. Stoch Environ Res Risk Assess (SERRA) 15(5):384–398

Pawlowsky-Glahn V, Egozcue JJ (2011) Exploring compositional data with the coda-dendrogram. Aust J Stat 40(1 & 2):103–113

Pawlowsky-Glahn V, Egozcue JJ, Olea RA, Pardo-Igúzquiza E (2015) Cokriging of compositional balances including a dimension reduction and retrieval of original units. J South Afr Inst Min Metal, SAIMM 115:59–72

Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2011) Principal balances to analyse the geochemistry of sediments. In Marschallinger R, Zobel F (eds) Proceedings of IAMG 2011–The XVth annual conference of the international association for mathematical geology, p 10. University of Salzburg, Austria

Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015) Modeling and analysis of compositional data. Statistics in practice. Wiley, Chichester UK, p 272

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. In: Proceedings of the Royal Society of London LX, pp 489–502

Queysanne M (1973) Álgebra Básica. Editorial Vicens Vives, Barcelona (E), p 669

Scealy JL, Welsh AH (2014) Colors and cocktails: compositional data analysis. Aust New Zealand J Stat 56(2):145–169

Tolosana-Delgado R (2012) Uses and misuses of compositional data in sedimentology. Sediment Geol 280:60–79