Lecture Notes on Compositional Data Analysis

Vera Pawlowsky-Glahn
University of Girona, Spain

Juan Jos Egozcue
Technical University of Catalonia, Spain

Raimon Tolosana-Delgado
Technical University of Catalonia, Spain

March 2011

Prof. Dr. Vera Pawlowsky-Glahn
Catedrática de Universidad (full professor)
University of Girona
Dept. of Computer Science and Applied Mathematics
Campus Montilivi — P-4, E-17071 Girona, Spain
vera.pawlowsky@udg.edu

Prof. Dr. Juan José Egozcue
Catedrático de Universidad (full professor)
Technical University of Catalonia
Dept. of Applied Mathematics III
Campus Nord c/Jordi Girona 1-3, C-2, E-08034 Barcelona, Spain
juan.jose.egozcue@upc.edu

Dr. Raimon Tolosana-Delgado
Investigador Juan de la Cierva
Technical University of Catalonia
Maritime Engineering Laboratory (LIM-UPC)
Campus Nord c/Jordi Girona 1-3, D-1, E-08034 Barcelona, Spain
raimon.tolosana@upc.edu

Preface

These notes have been prepared as support to a short course on compositional data analysis. The first version dates back to the year 2000. Their aim is to transmit the basic concepts and skills for simple applications, thus setting the premises for more advanced projects. The notes have been updated over the years. But one should be aware that frequent updates will be still required in the near future, as the theory presented here is a field of active research.

The notes are based both on the monograph by John Aitchison, *Statistical analysis of compositional data* (1986), and on recent developments that complement the theory developed there, mainly those by Aitchison (1997); Barceló-Vidal et al. (2001); Billheimer et al. (2001); Pawlowsky-Glahn and Egozcue (2001, 2002); Aitchison et al. (2002); Egozcue et al. (2003); Pawlowsky-Glahn (2003) and Egozcue and Pawlowsky-Glahn (2005). To avoid constant references to mentioned documents, only complementary references will be given within the text.

Readers should take into account that for a thorough understanding of compositional data analysis, a good knowledge in standard univariate statistics, basic linear algebra and calculus, complemented with an introduction to applied multivariate statistical analysis, is a must. The specific subjects of interest in multivariate statistics in real space can be learned in parallel from standard textbooks, like for instance Krzanowski (1988) and Krzanowski and Marriott (1994) (in English), Fahrmeir and Hamerle (1984) (in German), or Peña (2002) (in Spanish). Thus, the intended audience goes from advanced students in applied sciences to practitioners.

Concerning notation, it is important to note that, to conform to the standard praxis of registering samples as a matrix where each row is a sample and each column is a variate, vectors will be considered as row vectors to make the transfer from theoretical concepts to practical computations easier.

Most chapters end with a list of exercises. They are formulated in such a way that they have to be solved using an appropriate software. CoDaPack is a user friendly freeware to facilitate this task and it can be downloaded from the web. Details about this package can be found in Thió-Henestrosa and Martín-Fernández (2005) or Thió-Henestrosa et al. (2005). Those interested in working with R (or S-plus) may use the full-fledged package "compositions" by van den Boogaart and Tolosana-Delgado (2005).

Girona, *Vera Pawlowsky-Glahn*
Barcelona, *Juan José Egozcue*
March 2011 *Raimon Tolosana-Delgado*

Acknowledgements. We acknowledge the many comments made by readers, pointing at small and at important errors in the text. They all have contributed to improve the Lecture Notes presented here. We appreciate also the support received from our Universities, research groups, from the *Spanish Ministry of Education and Science* under projects: 'Ingenio Mathematica (i-MATH)' Ref. No. CSD2006-00032 and 'CODA-RSS' Ref. MTM2009-13272; and from the *Agència de Gestió d'Ajuts Universitaris i de Recerca* of the *Generalitat de Catalunya* under the project with Ref: 2009SGR424.

Contents

Chapter 1

Introduction

The awareness of problems related to the statistical analysis of compositional data analysis dates back to a paper by Karl Pearson (1897) which title began significantly with the words *"On a form of spurious correlation ... "*. Since then, as stated in Aitchison and Egozcue (2005), the way to deal with this type of data has gone through roughly four phases, which they describe as follows:

> The pre-1960 phase rode on the crest of the developmental wave of standard multivariate statistical analysis, an appropriate form of analysis for the investigation of problems with real sample spaces. Despite the obvious fact that a compositional vector—with components the proportions of some whole—is subject to a constant-sum constraint, and so is entirely different from the unconstrained vector of standard unconstrained multivariate statistical analysis, scientists and statisticians alike seemed almost to delight in applying all the intricacies of standard multivariate analysis, in particular correlation analysis, to compositional vectors. We know that Karl Pearson, in his definitive 1897 paper on spurious correlations, had pointed out the pitfalls of interpretation of such activity, but it was not until around 1960 that specific condemnation of such an approach emerged.
>
> In the second phase, the primary critic of the application of standard multivariate analysis to compositional data was the geologist Felix Chayes (1960), whose main criticism was in the interpretation of product-moment correlation between components of a geochemical composition, with negative bias the distorting factor from the viewpoint of any sensible interpretation. For this problem of negative bias, often referred to as the closure problem, Sarmanov and Vistelius (1959) supplemented the Chayes criticism in geological applications and Mosimann (1962) drew the attention of biologists to it. However, even conscious researchers, instead of working towards an appropriate methodology, adopted what can only be described as

a pathological approach: distortion of standard multivariate techniques when applied to compositional data was the main goal of study.

The third phase was the realisation by Aitchison in the 1980's that compositions provide information about relative, not absolute, values of components, that therefore every statement about a composition can be stated in terms of ratios of components (Aitchison, 1981, 1982, 1983, 1984). The facts that logratios are easier to handle mathematically than ratios and that a logratio transformation provides a one-to-one mapping on to a real space led to the advocacy of a methodology based on a variety of logratio transformations. These transformations allowed the use of standard unconstrained multivariate statistics applied to transformed data, with inferences translatable back into compositional statements.

The fourth phase arises from the realisation that the internal simplicial operation of perturbation, the external operation of powering, and the simplicial metric, define a metric vector space (indeed a Hilbert space) (Billheimer et al., 1997, 2001; Pawlowsky-Glahn and Egozcue, 2001). So, many compositional problems can be investigated within this space with its specific algebraic-geometric structure. There has thus arisen a staying-in-the-simplex approach to the solution of many compositional problems (Mateu-Figueras, 2003; Pawlowsky-Glahn, 2003). This staying-in-the-simplex point of view proposes to represent compositions by their coordinates, as they live in an Euclidean space, and to interpret them and their relationships from their representation in the simplex. Accordingly, the sample space of random compositions is identified to be the simplex with a simplicial metric and measure, different from the usual Euclidean metric and Lebesgue measure in real space.

The third phase, which mainly deals with (log-ratio) transformation of raw data, deserves special attention because these techniques have been very popular and successful over more than a century; from the Galton-McAlister introduction of such an idea in 1879 in their logarithmic transformation for positive data, through variance-stabilising transformations for sound analysis of variance, to the general Box-Cox transformation (Box and Cox, 1964) and the implied transformations in generalised linear modeling. The logratio transformation principle was based on the fact that there is a one-to-one correspondence between compositional vectors and associated logratio vectors, so that any statement about compositions can be reformulated in terms of logratios, and vice versa. The advantage of the transformation is that it removes the problem of a constrained sample space, the unit simplex, to one of an unconstrained space, multivariate real space, opening up all available standard multivariate techniques. The original transformations were principally the additive logratio transformation (Aitchison, 1986, p.113) and the

> centred logratio transformation (Aitchison, 1986, p.79). The logratio transformation methodology seemed to be accepted by the statistical community; see for example the discussion of Aitchison (1982). The logratio methodology, however, drew fierce opposition from other disciplines, in particular from sections of the geological community. The reader who is interested in following the arguments that have arisen should examine the Letters to the Editor of Mathematical Geology over the period 1988 through 2002.

The notes presented here correspond to the fourth phase. They pretend to summarise the state-of-the-art in the staying-in-the-simplex approach. Therefore, the first part will be devoted to the algebraic-geometric structure of the simplex, which we call *Aitchison geometry*.

Chapter 2

Compositional data and their sample space

2.1 Basic concepts

DEFINITION 2.1.1 *A row vector,* $\mathbf{x} = [x_1, x_2, \ldots, x_D]$*, is defined as a D-part composition when all its components are* strictly positive real numbers *and they carry only* relative information*.*

Indeed, that compositional information is relative is implicitly stated in the units, as they are usually parts of a whole, like weight or volume percent, ppm, ppb, or molar proportions. The most common examples have *a constant sum* κ and are known in the geological literature as *closed data* (Chayes, 1971). Frequently, $\kappa = 1$, which means that measurements have been made in, or transformed to, parts per unit, or $\kappa = 100$, for measurements in percent. Other units are possible, like ppm or ppb, which are typical examples for compositional data where only a part of the composition has been recorded; or, as recent studies have shown, even concentration units (mg/L, meq/L, molarities and molalities), where no constant sum can be feasibly defined (Buccianti and Pawlowsky-Glahn, 2005; Otero et al., 2005).

DEFINITION 2.1.2 *The sample space of compositional data is the simplex, defined as*

$$\mathcal{S}^D = \left\{ \mathbf{x} = [x_1, x_2, \ldots, x_D] \,\middle|\, x_i > 0, i = 1, 2, \ldots, D; \sum_{i=1}^{D} x_i = \kappa \right\} . \qquad (2.1)$$

However, this definition does not include compositions in e.g. meq/L. Therefore, a more general definition, together with its interpretation, is given in Section 2.2.

The components of a vector in $\mathcal{S}^D$ are called **parts** to remark their compositional character.

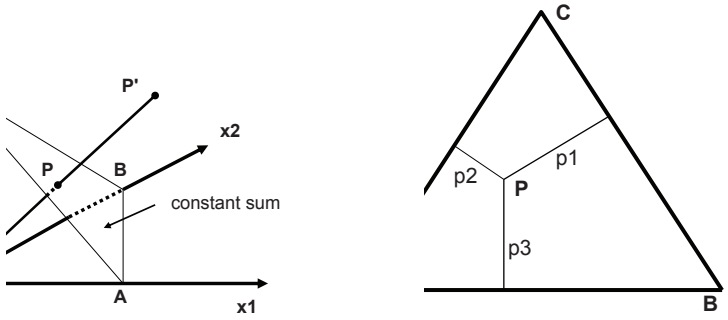

Figure 2.1: Left: Simplex imbedded in $\mathbb{R}^3$. Right: Ternary diagram.

DEFINITION 2.1.3 *For any vector of D real positive components*

$$\mathbf{z} = [z_1, z_2, \ldots, z_D] \in \mathbb{R}_+^D$$

($z_i > 0$ for all $i = 1, 2, \ldots, D$), the closure of $\mathbf{z}$ is defined as

$$\mathcal{C}(\mathbf{z}) = \left[\frac{\kappa \cdot z_1}{\sum_{i=1}^D z_i}, \frac{\kappa \cdot z_2}{\sum_{i=1}^D z_i}, \cdots, \frac{\kappa \cdot z_D}{\sum_{i=1}^D z_i} \right].$$

The result is the same vector re-scaled so that the sum of its components is κ. This operation is required for a formal definition of subcomposition and for inverse transformations. Note that κ depends on the units of measurement: usual values are 1 (proportions), 100 (%), 10^6 (ppm) and 10^9 (ppb).

DEFINITION 2.1.4 *Given a composition $\mathbf{x}$, a subcomposition $\mathbf{x}_s$ with s parts is obtained applying the closure operation to a subvector $[x_{i_1}, x_{i_2}, \ldots, x_{i_s}]$ of $\mathbf{x}$. Subindexes $i_1, \ldots, i_s$ tell which parts are selected in the subcomposition, not necessarily the first s ones.*

Very often, compositions contain many parts; e.g., the major oxide bulk composition of igneous rocks has around 10 elements, and they are but a few of the total possible. Nevertheless, one seldom represents the full composition. In fact, most of the applied literature on compositional data analysis (mainly in geology) restricts their figures to 3-part (sub)compositions. For 3 parts, the simplex can be represented as an equilateral triangle (Figure 2.1 left), with vertices at $A = [\kappa, 0, 0]$, $B = [0, \kappa, 0]$ and $C = [0, 0, \kappa]$. But this is commonly visualised in the form of a *ternary diagram*—which is an equivalent representation—. A ternary diagram is an equilateral triangle such that a generic sample $\mathbf{p} = [p_1, p_2, p_3]$ will plot at a distance p_1 from the opposite side of vertex A, at a distance p_2 from the opposite side of vertex B, and at a distance p_3 from the opposite side of vertex C (Figure 2.1 right). The triplet $[p_1, p_2, p_3]$ is commonly called the *barycentric coordinates* of $\mathbf{p}$, easily interpretable but useless in plotting (plotting them would yield the three-dimensional left-hand

plot of Figure 2.1). What is needed to get the right-hand plot of Figure 2.1) is the expression of the coordinates of the vertices and of the samples in a 2-dimensional Cartesian coordinate system $[u, v]$, and this is given in Appendix A.

Finally, if only some parts of the composition are available, a fill up or residual value can be defined, or simply the observed subcomposition can be closed. Note that, since one seldom analyses every possible part, in practice only subcompositions are analysed. In any case, both methods (fill-up or closure) should lead to identical, or at least compatible, results.

2.2 Principles of compositional analysis

Three conditions should be fulfilled by any statistical method to be applied to compositions: scale invariance, permutation invariance, and subcompositional coherence (Aitchison, 1986).

2.2.1 Scale invariance

The most important characteristic of compositional data is that *they carry only relative information*. Let us explain this concept with an example. In a paper with the suggestive title of "unexpected trend in the compositional maturity of second-cycle sands", Solano-Acosta and Dutta (2005) the lithologic composition of a sandstone and of its derived recent sands is analysed looking at the percentage of grains made up of only quartz, of only feldspar, or of rock fragments. For medium sized grains coming from the parent sandstone, they report an average composition $[Q, F, R] = [53, 41, 6]\,\%$, whereas for the daughter sands the mean values are $[37, 53, 10]\,\%$. One expects that feldspar and rock fragments decrease as the sediment matures, thus they should be less important in a second generation sand. "Unexpectedly" (or apparently), this does not happen in their example. To pass from the parent sandstone to the daughter sand, several different changes are possible, yielding exactly the same final composition. Assume those values were weight percent (in g/100 g of bulk sediment). Then, one of the following might have happened:

- Q suffered no change passing from sandstone to sand, but per 100 g parent sandstone 35 g F and 8 g R were added to the sand (for instance, due to comminution of coarser grains of F and R from the sandstone),
- F was unchanged, but per 100 g parent sandstone 25 g Q were depleted and at the same time 2 g R were added (for instance, because Q was better cemented in the sandstone, thus it tends to form coarser grains),
- any combination of the former two extremes.

The first two cases yield, per 100 g parent sandstone, final masses of $[53, 76, 14]$ g, respectively $[28, 41, 8]$ g. In a purely compositional data set, we do not know whether mass was added or subtracted from the sandstone to the sand. Thus,

which of these cases really occurred cannot be decided. Without further (non-compositional) information, there is no way to distinguish between $[53, 76, 14]$ g and $[28, 41, 8]$ g, as we only have the value of the sand composition *after closure*. Closure is a projection of any point in the positive orthant of D-dimensional real space onto the simplex. All points on a ray starting at the origin (*e.g.*, $[53, 76, 14]$ and $[28, 41, 8]$) are projected onto the same point of $\mathcal{S}^D$ (*e.g.*, $[37, 53, 10]\,\%$). The ray is an *equivalence class* and the point on $\mathcal{S}^D$ a *representant* of the class: Figure 2.2 shows this relationship. Moreover, to change the units of the data

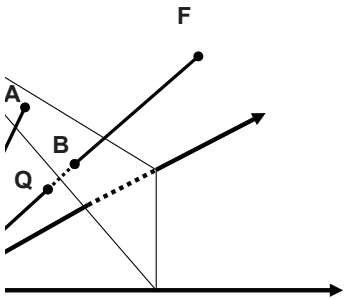

Figure 2.2: Representation of the compositional equivalence relationship. A represents the original sandstone composition, B the final sand composition, F the amount of each part if feldspar was added to the system (first hypothesis), and Q the amount of each part if quartz was depleted from the system (second hypothesis). Note that the points B, Q and F are compositionally equivalent.

(for instance, from % to ppm), simply multiply all the points by the constant of change of units, moving them along their rays to the intersections with another triangle, parallel to the plotted one.

DEFINITION 2.2.1 *Two vectors of D positive real components $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D_+$ ($x_i, y_i \geq 0$ for all $i = 1, 2, \ldots, D$), are* compositionally equivalent *if there exists a positive scalar $\lambda \in \mathbb{R}^+$ such that $\mathbf{x} = \lambda \cdot \mathbf{y}$ and, equivalently, $\mathcal{C}(\mathbf{x}) = \mathcal{C}(\mathbf{y})$.*

It is highly reasonable to expect analyses to yield the same results, independently of the value of λ. This is known as *scale invariance* (Aitchison, 1986):

DEFINITION 2.2.2 *A function $f(\cdot)$ is scale-invariant if for any positive real value $\lambda \in \mathbb{R}^+$ and for any composition $\mathbf{x} \in \mathcal{S}^D$, the function satisfies $f(\lambda\mathbf{x}) = f(\mathbf{x})$,* i.e. *it yields the same result for all vectors compositionally equivalent.*

Mathematically speaking, this is achieved if $f(\cdot)$ is a 0-degree homogeneous function of the parts in $\mathbf{x}$. Practical choices of such functions are log-ratios of the parts in $\mathbf{x}$ (Aitchison, 1997; Barceló-Vidal et al., 2001). For instance, assume that $\mathbf{x} = [x_1, x_2, \ldots, x_D]$ is a composition given in percentages. The ratio $f(\mathbf{x}) = x_1/x_2 = (\lambda \cdot x_1)/(\lambda \cdot x_2)$ is scale invariant and yields the same results if the composition is given in different units, e.g. in parts per unit or in

parts per million, because units cancel in the ratio. However, ratios depend on the ordering of parts because $x_1/x_2 \neq x_2/x_1$. A convenient transformation of ratios is the corresponding log-ratio, $f(\mathbf{x}) = \ln(x_1/x_2)$. Now, the inversion of the ratio only produces a change of sign, thus giving a symmetry to $f(\cdot)$ with respect to the ordering of parts.

More complicated log-ratios are useful. For instance, define

$$f(\mathbf{x}) = \ln \frac{x_1^{\alpha_1} \cdot x_2^{\alpha_2} \cdots x_s^{\alpha_s}}{x_{s+1}^{-\alpha_{s+1}} \cdot x_{s+2}^{-\alpha_{s+2}} \cdots x_D^{-\alpha_D}} = \sum_{i=1}^{D} \alpha_i \ln x_i \ ,$$

where powers α_i are real constants (positive or negative). In the ratio expression, for $i = s+1, s+2, \ldots, D$, α_i are assumed negative, thus appearing in the denominator with a positive value $-\alpha_i$. For this log-ratio to be scale invariant, the sum of all powers should be null. Scale invariant log-ratios are called log-contrasts (Aitchison, 1986).

DEFINITION 2.2.3 *Consider a composition* $\mathbf{x} = [x_1, x_2, \ldots, x_D]$. *A log-contrast is a function*

$$f(\mathbf{x}) = \sum_{i=1}^{D} \alpha_i \ln x_i \ , \quad \textit{with} \quad \sum_{i=1}^{D} \alpha_i = 0 \, .$$

In applications, some log-contrasts may be easily interpreted. A typical example is chemical equilibrium. Consider a chemical D-part composition, denoted $\mathbf{x}$ expressed in ppm of mass. A chemical reaction involving four species may be

$$\alpha_1 x_1 + \alpha_2 x_2 \rightleftharpoons \alpha_3 x_3 + \alpha_4 x_4 \ ,$$

where other parts are not involved. The α_i's, called stoichiometric coefficients, are normally known. If the reaction is mass preserving, then $\alpha_1 + \alpha_2 = \alpha_3 + \alpha_4$. Whenever this chemical reaction is in equilibrium, the log-contrast

$$\ln \frac{x_1^{\alpha_1} \cdot x_2^{\alpha_2}}{x_3^{\alpha_3} \cdot x_4^{\alpha_4}}$$

should be constant and, therefore, it is readily interpreted.

2.2.2 Permutation invariance

A function is *permutation-invariant* if it yields equivalent results when the ordering of the parts in the composition is changed. Two examples might illustrate what "*equivalent*" means here. The *distance* between the initial sandstone and the final sand compositions should be the same working with $[Q, F, R]$ or working with $[F, R, Q]$ (or any other *permutation* of the parts). On the other side, if interest lies in the *change* occurred from sandstone to sand, results should be equal after reordering. A classical way to get rid of the singularity of the classical covariance matrix of compositional data is to erase one component: this

procedure is not permutation-invariant, as results will largely depend on which component is erased.

However, ordered compositional data are frequent. A typical case corresponds to the discretisation of a continuous variable. Some interval categories are defined on the span of the variable, and then the number of occurrences in each category are recorded as frequencies. These frequencies can be considered as a composition although categories are still ordered. The information concerning the ordering will be lost in a standard compositional analysis.

2.2.3 Subcompositional coherence

The final condition is *subcompositional coherence*: subcompositions should behave like orthogonal projections in conventional real analysis. The size of a projected segment is less than or equal to the size of the segment itself. This general principle, though shortly stated, has several practical implications, explained in the next chapters. The most illustrative, however, are the following.

- The distance measured between two full compositions must be greater than (or at least equal to) the distance between them when considering any subcomposition. This particular behaviour of the distance is called *subcompositional dominance*. Exercise 2.3.4 proves that the Euclidean distance between compositional vectors does not fulfill this condition, and it is thus ill-suited to measure distance between compositions.

- If a non-informative part is erased, results should not change; for instance if hydrogeochemical data are available, and interest lies in classifying the kind of rocks washed by the water, in general the relations between some major oxides and ions will be used (SO_4^{2+}, HCO_3^-, Cl^-, to mention a few), and the same results should be obtained taking meq/L (including implicitly water content), or weight percent of the ions of interest.

Subcompositional coherence can be summarized as: (a) distances between two compositions should decrease when subcompositions of the original ones are considered; (b) scale invariance is preserved within arbitrary subcompositions (Egozcue, 2009). This means that the ratios between any parts in the subcomposition should be equal to the ratios in the original composition.

2.3 Exercises

EXERCISE 2.3.1 *If data are measured in ppm, what is the value of the constant κ in definition (2.1.2)?*

EXERCISE 2.3.2 *Plot a ternary diagram using different values for the constant sum κ.*

EXERCISE 2.3.3 *Verify that data in table 2.1 satisfy the conditions for being compositional. Plot them in a ternary diagram.*

Table 2.1: Simulated data set (3 parts, 20 samples).

	1	2	3	4	5	6	7	8	9	10
x_1	79.07	31.74	18.61	49.51	29.22	21.99	11.74	24.47	5.14	15.54
x_2	12.83	56.69	72.05	15.11	52.36	59.91	65.04	52.53	38.39	57.34
x_3	8.10	11.57	9.34	35.38	18.42	18.10	23.22	23.00	56.47	27.11

	11	12	13	14	15	16	17	18	19	20
x_1	57.17	52.25	77.40	10.54	46.14	16.29	32.27	40.73	49.29	61.49
x_2	3.81	23.73	9.13	20.34	15.97	69.18	36.20	47.41	42.74	7.63
x_3	39.02	24.02	13.47	69.12	37.89	14.53	31.53	11.86	7.97	30.88

EXERCISE 2.3.4 *Compute the Euclidean distance between the first two vectors of table 2.1. Imagine originally a fourth variable x_4 was measured, constant for all samples and equal to 5%. Take the first two vectors, close them to sum up to 95%, add the fourth variable to them (so that they sum up to 100%) and compute the Euclidean distance between the closed vectors. If the Euclidean distance is subcompositionally dominant, the distance measured in 4 parts must be greater or equal to the distance measured in the 3 part subcomposition.*

Chapter 3

The Aitchison geometry

3.1 General comments

In real space we are used to add vectors, to multiply them by a constant or scalar value, to look for properties like orthogonality, or to compute the distance between two points. All this, and much more, is possible because the real space is a linear vector space with an Euclidean metric structure. We are familiar with its geometric structure, the Euclidean geometry, and we represent our observations within this geometry. But this geometry is not a proper geometry for compositional data.

To illustrate this assertion, consider the compositions

$$[5, 65, 30],\ [10, 60, 30],\ [50, 20, 30],\ \text{and}\ [55, 15, 30].$$

Intuitively we would say that the difference between $[5, 65, 30]$ and $[10, 60, 30]$ is not the same as the difference between $[50, 20, 30]$ and $[55, 15, 30]$. The Euclidean distance between them is certainly the same, as there is a difference of 5 units both between the first and the second components, but in the first case the proportion in the first component is doubled, while in the second case the relative increase is about 10%, and this relative difference seems more adequate to describe compositional variability.

This is not the only reason for discarding Euclidean geometry as a proper tool for analysing compositional data. Problems might appear in many situations, like those where results end up outside the sample space, e.g. when translating compositional vectors, or computing joint confidence regions for random compositions under assumptions of normality, or using hexagonal confidence regions. This last case is paradigmatic, as such hexagons are often naively cut when they lay partly outside the ternary diagram, and this without regard to any probability adjustment. This kind of problems are not just theoretical: they are practical and interpretative.

What is needed is a sensible geometry to work with compositional data. In the simplex, things appear not as simple as they (apparently) are in real space,

but it is possible to find a way of working in it that is completely analogous. In fact, it is possible to define two operations which give the simplex a vector space structure. The first one is perturbation, which is analogous to addition in real space, the second one is powering, which is analogous to multiplication by a scalar in real space. Both require in their definition the closure operation; recall that closure is nothing else but the projection of a vector with positive components onto the simplex. Moreover, it is possible to obtain a Euclidean vector space structure on the simplex, just adding an inner product, a norm and a distance to the previous definitions. With the inner product compositions can be projected onto particular directions, one can check for orthogonality and determine angles between compositional vectors; with the norm the *length* of a composition can be computed; the possibilities of a distance should be clear. With all together one can operate in the simplex in the same way as one operates in real space.

3.2 Vector space structure

The basic operations required for a vector space structure of the simplex follow. They use the closure operation given in Definition 2.1.3.

DEFINITION 3.2.1 *Perturbation of a composition* $\mathbf{x} \in \mathcal{S}^D$ *by a composition* $\mathbf{y} \in \mathcal{S}^D$*,*

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}\left[x_1 y_1, x_2 y_2, \ldots, x_D y_D\right].$$

DEFINITION 3.2.2 *Power transformation or powering of a composition* $\mathbf{x} \in \mathcal{S}^D$ *by a constant* $\alpha \in \mathbb{R}$*,*

$$\alpha \odot \mathbf{x} = \mathcal{C}\left[x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha\right].$$

For an illustration of the effect of perturbation and powering on a set of compositions, see Figure 3.1.

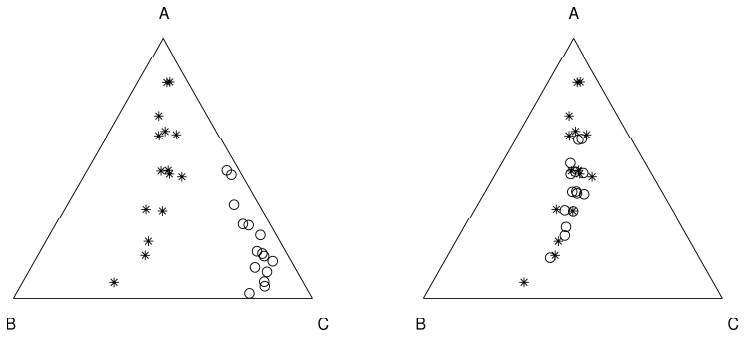

Figure 3.1: Left: Perturbation of initial compositions ($\circ$) by $\mathbf{p} = [0.1, 0.1, 0.8]$ resulting in compositions ($\star$). Right: Powering of compositions ($\star$) by $\alpha = 0.2$ resulting in compositions ($\circ$).

The simplex , $(\mathcal{S}^D, \oplus, \odot)$, with perturbation and powering, is a vector space. This means the following properties hold, making them analogous to translation and scalar multiplication:

PROPERTY 3.2.1 *$(\mathcal{S}^D, \oplus)$ has a commutative group structure; i.e., for $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z} \in \mathcal{S}^D$ it holds*

1. *commutative property:* $\mathbf{x} \oplus \mathbf{y} = \mathbf{y} \oplus \mathbf{x}$;
2. *associative property:* $(\mathbf{x} \oplus \mathbf{y}) \oplus \mathbf{z} = \mathbf{x} \oplus (\mathbf{y} \oplus \mathbf{z})$;
3. *neutral element:*

$$\mathbf{n} = \mathcal{C}[1, 1, \ldots, 1] = \left[\frac{1}{D}, \frac{1}{D}, \ldots, \frac{1}{D}\right];$$

   $\mathbf{n}$ *is the barycentre of the simplex and is unique;*

4. *inverse of* $\mathbf{x}$*:* $\mathbf{x}^{-1} = \mathcal{C}\left[x_1^{-1}, x_2^{-1}, \ldots, x_D^{-1}\right]$*; thus,* $\mathbf{x} \oplus \mathbf{x}^{-1} = \mathbf{n}$*. By analogy with standard operations in real space, we will write* $\mathbf{x} \oplus \mathbf{y}^{-1} = \mathbf{x} \ominus \mathbf{y}$.

PROPERTY 3.2.2 *Powering satisfies the properties of an external product. For $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$, $\alpha, \beta \in \mathbb{R}$it holds*

1. *associative property:* $\alpha \odot (\beta \odot \mathbf{x}) = \alpha \cdot \beta) \odot \mathbf{x}$;
2. *distributive property 1:* $\alpha \odot (\mathbf{x} \oplus \mathbf{y}) = (\alpha \odot \mathbf{x}) \oplus (\alpha \odot \mathbf{y})$;
3. *distributive property 2:* $(\alpha + \beta) \odot \mathbf{x} = (\alpha \odot \mathbf{x}) \oplus (\beta \odot \mathbf{x})$;
4. *neutral element:* $1 \odot \mathbf{x} = \mathbf{x}$*; the neutral element is unique.*

Note that the closure operation cancels out any constant and, thus, the closure constant itself is not important from a mathematical point of view. This fact allows us to omit the closure in intermediate steps of any computation without problem. It has also important implications for practical reasons, as shall be seen during simplicial principal component analysis. We can express this property for $\mathbf{z} \in \mathbb{R}_+^D$ and $\mathbf{x} \in \mathcal{S}^D$ as

$$\mathbf{x} \oplus (\alpha \odot \mathbf{z}) = \mathbf{x} \oplus (\alpha \odot \mathcal{C}(\mathbf{z})). \tag{3.1}$$

Nevertheless, one should be always aware that the closure constant is very important for the correct interpretation of the units of the problem at hand. Therefore, controlling for the right units should be the last step in any analysis.

3.3 Inner product, norm and distance

To obtain a Euclidean vector space structure, we take the following inner product, with associated norm and distance:

DEFINITION 3.3.1 *Inner product of* $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$,

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

DEFINITION 3.3.2 *Norm of* $\mathbf{x} \in \mathcal{S}^D$,

$$\|\mathbf{x}\|_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} \right)^2}.$$

DEFINITION 3.3.3 *Distance between* $\mathbf{x}$ *and* $\mathbf{y} \in \mathcal{S}^D$,

$$d_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{x}\|_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2}.$$

In practice, alternative but equivalent expressions of the inner product, norm and distance may be useful. Three possible alternatives for the inner product follow:

$$\begin{aligned} \langle \mathbf{x}, \mathbf{y} \rangle_a &= \frac{1}{D} \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j} \\ &= \sum_{i=1}^{D} \ln x_i \ln y_i - \frac{1}{D} \left(\sum_{j=1}^{D} \ln x_j \right) \left(\sum_{k=1}^{D} \ln y_k \right) \\ &= \sum_{i=1}^{D} \ln \frac{x_i}{\mathrm{g}(\mathbf{x})} \cdot \ln \frac{y_i}{\mathrm{g}(\mathbf{y})} \ . \end{aligned}$$

where $\mathrm{g}(\cdot)$ denotes the geometric mean of the arguments. The last expression in 3.2 corresponds to an ordinary inner product of two real vectors. These vectors are called centered log-ratio (clr) of $\mathbf{x}$, $\mathbf{y}$, as defined in Chapter 4. Note that notation $\sum_{i<j}$ means exactly $\sum_{i=1}^{D-1} \sum_{j=i+1}^{D}$. Moreover, in the previous expressions, simple logratios, $\ln(x_i/x_j)$, are null whenever $i = j$; in these circumstances, $\sum_{i=1}^{D-1} \sum_{j=i+1}^{D} = (1/2) \sum_{i=1}^{D} \sum_{j=1}^{D}$.

To refer to the properties of $(\mathcal{S}^D, \oplus, \odot)$ as an Euclidean linear vector space, we shall talk globally about the *Aitchison geometry on the simplex*, and in particular about the Aitchison distance, norm and inner product. Note that in mathematical textbooks, such a linear vector space is called either real Euclidean space or finite dimensional real Hilbert space.

The algebraic-geometric structure of $\mathcal{S}^D$ satisfies standard properties, like compatibility of the distance with perturbation and powering, i.e.

$$\mathrm{d}_a(\mathbf{p} \oplus \mathbf{x}, \mathbf{p} \oplus \mathbf{y}) = \mathrm{d}_a(\mathbf{x}, \mathbf{y}), \quad \mathrm{d}_a(\alpha \odot \mathbf{x}, \alpha \odot \mathbf{y}) = |\alpha| \mathrm{d}_a(\mathbf{x}, \mathbf{y}) \ ,$$

for any $\mathbf{x}, \mathbf{y}, \mathbf{p} \in \mathcal{S}^D$ and $\alpha \in \mathbb{R}$. Other typical properties of metric spaces are valid for $\mathcal{S}^D$. Some of them follow:

1. Cauchy-Schwartz inequality:

$$|\langle \mathbf{x}, \mathbf{y} \rangle_a| \leq \|\mathbf{x}\|_a \cdot \|\mathbf{y}\|_a \ ;$$

2. Pythagoras: If $\mathbf{x}$, $\mathbf{y}$ are orthogonal, i.e. $\langle \mathbf{x}, \mathbf{y} \rangle_a = 0$, then

$$\|\mathbf{x} \ominus \mathbf{y}\|_a^2 = \|\mathbf{x}\|_a^2 + \|\mathbf{y}\|_a^2 \ ;$$

3. Triangular inequality:

$$\mathrm{d}_a(\mathbf{x}, \mathbf{y}) \leq \mathrm{d}_a(\mathbf{x}, \mathbf{z}) + \mathrm{d}_a(\mathbf{y}, \mathbf{z}) \ .$$

For a discussion of these and other properties, see Billheimer et al. (2001) or Pawlowsky-Glahn and Egozcue (2001). For a comparison with other measures of difference obtained as restrictions of distances in $\mathbb{R}^D$ to $\mathcal{S}^D$, see Martín-Fernández et al. (1998, 1999); Aitchison et al. (2000) or Martín-Fernández (2001). The Aitchison distance is subcompositionally coherent, as all this set of operations induce the same linear vector space structure in the subspace corresponding to the subcomposition. Finally, the distance is subcompositionally dominant (Exercise 3.5.7).

3.4 Geometric figures

Within this framework, we can define lines in $\mathcal{S}^D$, which we call *compositional lines*, as $\mathbf{y} = \mathbf{x}_0 \oplus (\alpha \odot \mathbf{x})$, with $\mathbf{x}_0$ the starting point and $\mathbf{x}$ the leading vector. Note that $\mathbf{y}$, $\mathbf{x}_0$ and $\mathbf{x}$ are elements of $\mathcal{S}^D$, while the coefficient α varies in $\mathbb{R}$. To illustrate what we understand by *compositional lines*, Figure 3.2 shows two families of parallel lines in a ternary diagram, forming a square, orthogonal grid of side equal to one Aitchison distance unit. Recall that parallel lines have the same leading vector, but different starting points, like for instance $\mathbf{y_1} = \mathbf{x}_1 \oplus (\alpha \odot \mathbf{x})$ and $\mathbf{y_2} = \mathbf{x}_2 \oplus (\alpha \odot \mathbf{x})$, while orthogonal lines are those for which the inner product of the leading vectors is zero, i.e., for $\mathbf{y_1} = \mathbf{x}_0 \oplus (\alpha_1 \odot \mathbf{x}_1)$ and $\mathbf{y_2} = \mathbf{x}_0 \oplus (\alpha_2 \odot \mathbf{x}_2)$, with $\mathbf{x}_0$ their intersection point and $\mathbf{x}_1$, $\mathbf{x}_2$ the corresponding leading vectors, it holds $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_a = 0$. Thus, *orthogonal* means here that the inner product given in Definition 3.3.1 of the leading vectors of two lines, one of each family, is zero, and one Aitchison distance unit is measured by the distance given in Definition 3.3.3.

Once we have a well defined geometry, it is straightforward to define any geometric figure, like for instance circles, ellipses, or rhomboids, as illustrated in Figure 3.3.

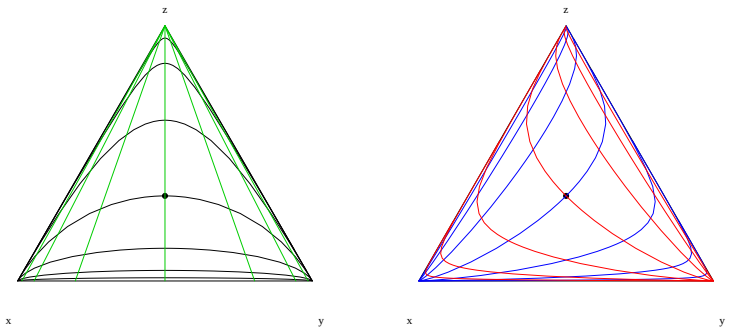

Figure 3.2: Orthogonal grids of compositional lines in $\mathcal{S}^3$, equally spaced, 1 unit in Aitchison distance (Def. 3.3.3). The grid in the right is rotated 45^o with respect to the grid in the left.

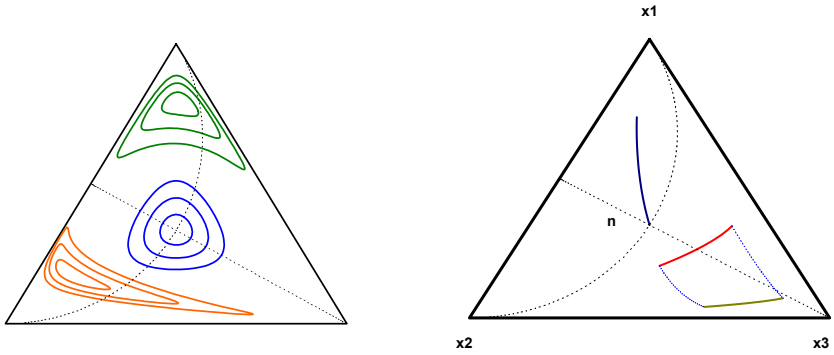

Figure 3.3: Circles and ellipses (left) and perturbation of a segment (right) in $\mathcal{S}^3$.

3.5 Exercises

EXERCISE 3.5.1 *Consider the two vectors* $[0.7, 0.4, 0.8]$ *and* $[0.2, 0.8, 0.1]$. *Perturb one vector by the other with and without previous closure. Is there any difference?*

EXERCISE 3.5.2 *Perturb each sample of the data set given in Table 2.1 with* $\mathbf{x}_1 = \mathcal{C}[0.7, 0.4, 0.8]$ *and plot the initial and the resulting perturbed data set. What do you observe?*

EXERCISE 3.5.3 *Apply powering with* α *ranging from* -3 *to* $+3$ *in steps of* 0.5 *to* $\mathbf{x}_1 = \mathcal{C}[0.7, 0.4, 0.8]$ *and plot the resulting set of compositions. Join them by a line. What do you observe?*

EXERCISE 3.5.4 *Perturb the compositions obtained in Ex. 3.5.3 by* $\mathbf{x}_2 = \mathcal{C}[0.2, 0.8, 0.1]$. *What is the result?*

EXERCISE 3.5.5 *Compute the Aitchison inner product of* $\mathbf{x}_1 = \mathcal{C}[0.7, 0.4, 0.8]$ *and* $\mathbf{x}_2 = \mathcal{C}[0.2, 0.8, 0.1]$. *Are they orthogonal?*

EXERCISE 3.5.6 *Compute the Aitchison norm of* $\mathbf{x}_1 = \mathcal{C}[0.7, 0.4, 0.8]$ *and call it* a. *Compute* $\alpha \odot \mathbf{x}_1$ *with* $\alpha = 1/a$. *Compute the Aitchison norm of the resulting composition. How do you interpret the result?*

EXERCISE 3.5.7 *Re-do Exercise 2.3.4, but using the Aitchison distance given in Definition 3.3.3. Is it subcompositionally dominant?*

EXERCISE 3.5.8 *In a 2-part composition* $\mathbf{x} = [x_1, x_2]$, *simplify the formula for the Aitchison distance, taking* $x_2 = 1 - x_1$ *(using* $\kappa = 1$*). Use it to plot 7 equally-spaced points on the segment* $(0, 1) = \mathcal{S}^2$, *from* $x_1 = 0.014$ *to* $x_1 = 0.986$.

EXERCISE 3.5.9 *In a mineral assemblage, several radioactive isotopes have been measured, obtaining* $\left[^{238}\mathrm{U}, ^{232}\mathrm{Th}, ^{40}\mathrm{K}\right] = [150, 30, 110]$*ppm. Which will be the composition after* $\Delta t = 10^9$ *years? And after another* Δt *years? Which was the composition* Δt *years ago? And* Δt *years before that? Close these 5 compositions and represent them in a ternary diagram. What do you see? Could you write the evolution as an equation? (Half-life disintegration periods:* $\left[^{238}\mathrm{U}, ^{232}\mathrm{Th}, ^{40}\mathrm{K}\right] = [4.468; 14.05; 1.277] \cdot 10^9$ *years)*

Chapter 4

Coordinate representation

4.1 Introduction

J. Aitchison (1986) used the fact that for compositional data size is irrelevant—as interest lies in relative proportions of the components measured—to introduce transformations based on ratios, the essential ones being the additive log-ratio transformation (alr) and the centred log-ratio transformation (clr). Then, he applied classical statistical analysis to the transformed observations, using the alr transformation for modeling, and the clr transformation for those techniques based on a metric. The underlying reason was, that the alr transformation does not preserve distances, whereas the clr transformation preserves distances but leads to a singular covariance matrix. In mathematical terms, we say that the alr transformation is an isomorphism, but not an isometry, while the clr transformation is an isometry, and thus also an isomorphism, but between $\mathcal{S}^D$ and a subspace of $\mathbb{R}^D$, leading to degenerate distributions. Thus, Aitchison's approach opened up a rigorous strategy, but care had to be applied when using either of both transformations.

Using the Euclidean vector space structure, it is possible to give an algebraic-geometric foundation to his approach, and it is possible to go even a step further. Within this framework, a transformation of coefficients is equivalent to express observations in a different coordinate system. We are used to work in an orthogonal system, known as a Cartesian coordinate system; we know how to change coordinates within this system and how to rotate axis. But neither the clr nor the alr transformations can be directly associated with an orthogonal coordinate system in the simplex, a fact that lead Egozcue et al. (2003) to define a new transformation, called ilr (for *isometric logratio*) transformation, which is an isometry between $\mathcal{S}^D$ and $\mathbb{R}^{D-1}$, thus avoiding the drawbacks of both the alr and the clr. The ilr stands actually for the association of coordinates with compositions in an orthonormal system in general, and this is the framework we are going to present here, together with a particular kind of coordinates, named balances, because of their usefulness for modeling and interpretation.

4.2 Compositional observations in real space

Compositions in $\mathcal{S}^D$ are usually expressed in terms of the canonical basis $\{\vec{\mathbf{e}}_1, \vec{\mathbf{e}}_2, \ldots, \vec{\mathbf{e}}_D\}$ of $\mathbb{R}^D$. In fact, any vector $\mathbf{x} \in \mathbb{R}^D$ can be written as

$$\mathbf{x} = x_1 \, [1, 0, \ldots, 0] + x_2 \, [0, 1, \ldots, 0] + \cdots + x_D \, [0, 0, \ldots, 1] = \sum_{i=1}^{D} x_i \cdot \vec{\mathbf{e}}_i \,, \quad (4.1)$$

and this is the way we are used to interpret it. The problem is, that the set of vectors $\{\vec{\mathbf{e}}_1, \vec{\mathbf{e}}_2, \ldots, \vec{\mathbf{e}}_D\}$ is neither a generating system nor a basis with respect to the vector space structure of $\mathcal{S}^D$ defined in Chapter 3. In fact, not every combination of coefficients gives an element of $\mathcal{S}^D$ (negative and zero values are not allowed), and the $\vec{\mathbf{e}}_i$ do not belong to the simplex as defined in Equation (2.1). Nevertheless, in many cases it is interesting to express results in terms of compositions (4.1), so that interpretations are feasible in usual units, and therefore one of our purposes is to find a way to state statistically rigorous results in this coordinate system.

4.3 Generating systems

A first step for defining an appropriate orthonormal basis consists in finding a generating system which can be used to build the basis. A natural way to obtain such a generating system is to take $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_D\}$, with

$$\mathbf{w}_i = \mathcal{C}\left(\exp(\vec{\mathbf{e}}_i)\right) = \mathcal{C}\left[1, 1, \ldots, e, \ldots, 1\right] \,, \quad i = 1, 2, \ldots, D \,, \quad (4.2)$$

where in each $\mathbf{w}_i$ the number e is placed in the i-th column, and the operation $\exp(\cdot)$ is assumed to operate component-wise on a vector. In fact, taking into account Equation (3.1) and the usual rules of precedence for operations in a vector space, i.e., first the external operation, $\odot$, and afterwards the internal operation, $\oplus$, any vector $\mathbf{x} \in \mathcal{S}^D$ can be written

$$\begin{aligned} \mathbf{x} &= \bigoplus_{i=1}^{D} \ln x_i \odot \mathbf{w}_i = \\ &= \ln x_1 \odot [e, 1, \ldots, 1] \oplus \ln x_2 \odot [1, e, \ldots, 1] \oplus \cdots \oplus \ln x_D \odot [1, 1, \ldots, e] \,. \end{aligned}$$

It is known that the coefficients with respect to a generating system are not unique; thus, the following equivalent expression can be used as well,

$$\begin{aligned} \mathbf{x} &= \bigoplus_{i=1}^{D} \ln \frac{x_i}{\mathrm{g}(\mathbf{x})} \odot \mathbf{w}_i = \\ &= \ln \frac{x_1}{\mathrm{g}(\mathbf{x})} \odot [e, 1, \ldots, 1] \oplus \cdots \oplus \ln \frac{x_D}{g(\mathbf{x})} \odot [1, 1, \ldots, e] \,, \end{aligned}$$

where

$$\mathrm{g}(\mathbf{x}) = \left(\prod_{i=1}^{D} x_i \right)^{1/D} = \exp\left(\frac{1}{D} \sum_{i=1}^{D} \ln x_i \right) \,,$$

is the component-wise geometric mean of the composition. One recognises in the coefficients of this second expression the centred logratio transformation defined by Aitchison (1986). Note that one could indeed replace the denominator *by any constant*. This non-uniqueness is consistent with the concept of compositions as equivalence classes (Barceló-Vidal et al., 2001).

We will denote by clr the transformation that gives the expression of a composition in centred logratio coefficients

$$\mathrm{clr}(\mathbf{x}) = \left[\ln \frac{x_1}{\mathrm{g}(\mathbf{x})}, \ln \frac{x_2}{\mathrm{g}(\mathbf{x})}, \ldots, \ln \frac{x_D}{\mathrm{g}(\mathbf{x})}\right] = \boldsymbol{\xi}. \tag{4.3}$$

The inverse transformation, which gives us the coefficients in the canonical basis of real space, is then

$$\mathrm{clr}^{-1}(\boldsymbol{\xi}) = \mathcal{C}\left[\exp(\xi_1), \exp(\xi_2), \ldots, \exp(\xi_D)\right] = \mathbf{x}. \tag{4.4}$$

The centred logratio transformation is symmetrical in the components, but the price is a new constraint on the transformed sample: the sum of the components has to be zero. This means that the transformed sample will lie on a plane, which goes through the origin of $\mathbb{R}^D$ and is orthogonal to the vector of unities $[1, 1, \ldots, 1]$. But, more importantly, it means also that for random compositions the covariance matrix of $\boldsymbol{\xi}$ is singular, i.e. the determinant is zero. Certainly, generalised inverses can be used in this context when necessary, but not all statistical packages are designed for it and problems might arise during computation. Furthermore, clr coefficients are not subcompositionally coherent, because the geometric mean of the parts of a subcomposition $\mathrm{g}(\mathbf{x}_s)$ is not necessarily equal to that of the full composition, and thus the clr coefficients are in general not the same. A formal definition of the clr coefficients follows.

DEFINITION 4.3.1 *For a composition* $\mathbf{x} \in \mathcal{S}^D$*, the* clr *coefficients are the components of* $\boldsymbol{\xi} = [\xi_1, \xi_2, \ldots, \xi_D] = \mathrm{clr}(\mathbf{x})$*, the unique vector satisfying*

$$\mathbf{x} = \mathrm{clr}^{-1}(\boldsymbol{\xi}) = \mathcal{C}\left(\exp(\boldsymbol{\xi})\right) \ , \ \sum_{i=1}^{D} \xi_i = 0 \ .$$

The i*-th* clr *coefficient is*

$$\xi_i = \ln \frac{x_i}{\mathrm{g}(\mathbf{x})} \ ,$$

being $\mathrm{g}(\mathbf{x})$ *the geometric mean of the components of* $\mathbf{x}$*.*

Although the clr coefficients are not coordinates with respect to a basis of the simplex, they have very important properties. Among them the translation of operations and metrics from the simplex into the real space deserves special attention. Denote ordinary distance, norm and inner product in $\mathbb{R}^{D-1}$ by $\mathrm{d}(\cdot,\cdot)$, $\|\cdot\|$, and $\langle\cdot,\cdot\rangle$ respectively. The following property holds.

PROPERTY 4.3.1 *Consider* $\mathbf{x}_k \in \mathcal{S}^D$ *and real constants* α, β; *then*

$$\begin{aligned} \operatorname{clr}(\alpha \odot \mathbf{x}_1 \oplus \beta \odot \mathbf{x}_2) &= \alpha \cdot \operatorname{clr}(\mathbf{x}_1) + \beta \cdot \operatorname{clr}(\mathbf{x}_2) \ ; \\ \langle \mathbf{x}_1, \mathbf{x}_2 \rangle_a &= \langle \operatorname{clr}(\mathbf{x}_1), \operatorname{clr}(\mathbf{x}_2) \rangle \ ; \\ \|\mathbf{x}_1\|_a = \|\operatorname{clr}(\mathbf{x}_1)\| \quad , \quad \mathrm{d}_a(\mathbf{x}_1, \mathbf{x}_2) &= \mathrm{d}(\operatorname{clr}(\mathbf{x}_1), \operatorname{clr}(\mathbf{x}_2)) \ . \end{aligned} \tag{4.5}$$

4.4 Orthonormal coordinates

Omitting one vector of the generating system given in Equation (4.2) a basis is obtained. For example, omitting $\mathbf{w}_D$ results in $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{D-1}\}$. This basis is not orthonormal, as can be shown computing the inner product of any two of its vectors. But a new basis, orthonormal with respect to the inner product, can be readily obtained using the well-known Gram-Schmidt procedure (Egozcue et al., 2003). The basis thus obtained will be just one out of the infinitely many orthonormal basis which can be defined in any Euclidean space. Therefore, it is convenient to study their general characteristics.

Let $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$ be a generic orthonormal basis of the simplex $\mathcal{S}^D$ and consider the $(D-1, D)$-matrix $\mathbf{\Psi}$ whose rows are $\operatorname{clr}(\mathbf{e}_i)$. An orthonormal basis satisfies that $\langle \mathbf{e}_i, \mathbf{e}_j \rangle_a = \delta_{ij}$ (δ_{ij} is the Kronecker-delta, which is null for $i \neq j$, and one whenever $i = j$). This can be expressed using (4.5),

$$\langle \mathbf{e}_i, \mathbf{e}_j \rangle_a = \langle \operatorname{clr}(\mathbf{e}_i), \operatorname{clr}(\mathbf{e}_j) \rangle = \delta_{ij} \ .$$

It implies that the $(D-1, D)$-matrix $\mathbf{\Psi}$ satisfies $\mathbf{\Psi}\mathbf{\Psi}' = I_{D-1}$, being I_{D-1} the identity matrix of dimension $D-1$. When the product of these matrices is reversed, then $\mathbf{\Psi}'\mathbf{\Psi} = I_D - (1/D)\mathbf{1}_D'\mathbf{1}_D$, with I_D the identity matrix of dimension D, and $\mathbf{1}_D$ a D-row-vector of ones; note this is a matrix of rank $D-1$. The compositions of the basis are recovered from $\mathbf{\Psi}$ using clr^{-1} in each row of the matrix. Recall that these rows of $\mathbf{\Psi}$ also add up to 0 because they are clr coefficients (see Definition 4.3.1).

Once an orthonormal basis has been chosen, a composition $\mathbf{x} \in \mathcal{S}^D$ is expressed as

$$\mathbf{x} = \bigoplus_{i=1}^{D-1} x_i^* \odot \mathbf{e}_i \ , \ x_i^* = \langle \mathbf{x}, \mathbf{e}_i \rangle_a \ , \tag{4.6}$$

where $\mathbf{x}^* = \left[x_1^*, x_2^*, \ldots, x_{D-1}^*\right]$ is the vector of coordinates of $\mathbf{x}$ with respect to the selected basis. The function $\operatorname{ilr} : \mathcal{S}^D \to \mathbb{R}^{D-1}$, assigning the coordinates $\mathbf{x}^*$ to $\mathbf{x}$ has been called ilr (isometric log-ratio) transformation, as it is an isometric isomorphism of vector spaces. For simplicity, sometimes this function is also denoted by h, i.e. $\operatorname{ilr} \equiv h$ and also the asterisk (*) is used to denote coordinates if convenient. The following properties hold.

PROPERTY 4.4.1 *Consider* $\mathbf{x}_k \in \mathcal{S}^D$ *and real constants* α, β; *then*

$$h(\alpha \odot \mathbf{x}_1 \oplus \beta \odot \mathbf{x}_2) = \alpha \cdot h(\mathbf{x}_1) + \beta \cdot h(\mathbf{x}_2) = \alpha \cdot \mathbf{x}_1^* + \beta \cdot \mathbf{x}_2^* \ ;$$

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_a = \langle h(\mathbf{x}_1), h(\mathbf{x}_2) \rangle = \langle \mathbf{x}_1^*, \mathbf{x}_2^* \rangle \ ;$$

$$\|\mathbf{x}_1\|_a = \|h(\mathbf{x}_1)\| = \|\mathbf{x}_1^*\| \quad , \quad \mathrm{d}_a(\mathbf{x}_1, \mathbf{x}_2) = \mathrm{d}(h(\mathbf{x}_1), h(\mathbf{x}_2)) = \mathrm{d}(\mathbf{x}_1^*, \mathbf{x}_2^*) \ .$$

The main difference between Property 4.3.1 for clr and Property 4.4.1 for ilr is that the former refers to vectors of coefficients in $\mathbb{R}^D$, whereas the latter deals with vectors of coordinates in $\mathbb{R}^{D-1}$, thus matching the actual dimension of $\mathcal{S}^D$.

Taking into account Properties 4.3.1 and 4.4.1, and using the clr image matrix of the basis, $\mathbf{\Psi}$, the coordinates of a composition $\mathbf{x}$ can be expressed in a compact way. As written in (4.6), a coordinate is an Aitchison inner product, and it can be expressed as an ordinary inner product of the clr coefficients. Grouping all coordinates in a vector

$$\mathbf{x}^* = \mathrm{ilr}(\mathbf{x}) = h(\mathbf{x}) = \mathrm{clr}(\mathbf{x}) \cdot \mathbf{\Psi}' \ , \tag{4.7}$$

a simple matrix product is obtained.

Inversion of ilr, i.e. recovering the composition from its coordinates, corresponds to Equation (4.6). In fact, taking clr coefficients in both sides of (4.6) and taking into account Property 4.3.1,

$$\mathrm{clr}(\mathbf{x}) = \mathbf{x}^* \mathbf{\Psi} \ , \ \mathbf{x} = \mathcal{C}\left(\exp(\mathbf{x}^* \mathbf{\Psi})\right) \ . \tag{4.8}$$

A suitable algorithm to recover $\mathbf{x}$ from its coordinates $\mathbf{x}^*$ consists of the following steps: (i) construct the clr-matrix of the basis, $\mathbf{\Psi}$; (ii) carry out the matrix product $\mathbf{x}^* \mathbf{\Psi}$; and (iii) apply clr^{-1} to obtain $\mathbf{x}$.

There are several ways to define orthonormal bases in the simplex. The main criterion for the selection of an orthonormal basis is that it enhances the interpretability of the representation in coordinates. For instance, when performing principal component analysis an orthogonal basis is selected so that the first coordinate (principal component) represents the direction of maximum variability, etc. Particular cases deserving our attention are those bases linked to a sequential binary partition of the compositional vector (Egozcue and Pawlowsky-Glahn, 2005). The main interest of such bases is that they are easily interpreted in terms of grouped parts of the composition. The Cartesian coordinates of a composition in such a basis are called *balances* and the compositions of the basis *balancing elements*. A *sequential binary partition* is a hierarchy of the parts of a composition. In the first order of the hierarchy, all parts are split into two groups. In the following steps, each group is in turn split into two groups, and the process continues until all groups have a single part, as illustrated in Table 4.1. For each order of the partition, it is possible to define the *balance* between the two sub-groups formed at that level: if $i_1, i_2, \ldots, i_r$ are the r parts of the first sub-group (coded by $+1$), and $j_1, j_2, \ldots, j_s$ the s parts of the second (coded by -1), the balance is defined as the normalised logratio of the geometric mean of each group of parts:

$$b = \sqrt{\frac{rs}{r+s}} \ln \frac{(x_{i_1} x_{i_2} \cdots x_{i_r})^{1/r}}{(x_{j_1} x_{j_2} \cdots x_{j_s})^{1/s}} = \ln \frac{(x_{i_1} x_{i_2} \cdots x_{i_r})^{a_+}}{(x_{j_1} x_{j_2} \cdots x_{j_s})^{a_-}} \ , \tag{4.9}$$

Table 4.1: Example of sign matrix, used to encode a sequential binary partition and build an orthonormal basis. The lower part of the table shows the matrix $\mathbf{\Psi}$ of the basis.

order	x_1	x_2	x_3	x_4	x_5	x_6	r	s
1	+1	+1	−1	−1	+1	+1	4	2
2	+1	−1	0	0	−1	−1	1	3
3	0	+1	0	0	−1	−1	1	2
4	0	0	0	0	+1	−1	1	1
5	0	0	+1	−1	0	0	1	1

order	x_1	x_2	x_3	x_4	x_5	x_6
1	$\frac{1}{4}\sqrt{\frac{4\cdot2}{4+2}}$	$\frac{1}{4}\sqrt{\frac{4\cdot2}{4+2}}$	$\frac{-1}{2}\sqrt{\frac{4\cdot2}{4+2}}$	$\frac{-1}{2}\sqrt{\frac{4\cdot2}{4+2}}$	$\frac{1}{4}\sqrt{\frac{4\cdot2}{4+2}}$	$\frac{1}{4}\sqrt{\frac{4\cdot2}{4+2}}$
2	$+\frac{\sqrt{3}}{2}$	$-\frac{1}{\sqrt{12}}$	0	0	$-\frac{1}{\sqrt{12}}$	$-\frac{1}{\sqrt{12}}$
3	0	$+\frac{\sqrt{2}}{\sqrt{3}}$	0	0	$-\frac{1}{\sqrt{6}}$	$-\frac{1}{\sqrt{6}}$
4	0	0	0	0	$+\frac{1}{\sqrt{2}}$	$-\frac{1}{\sqrt{2}}$
5	0	0	$+\frac{1}{\sqrt{2}}$	$-\frac{1}{\sqrt{2}}$	0	0

where

$$a_+ = +\frac{1}{r}\sqrt{\frac{rs}{r+s}}\,, \quad a_- = -\frac{1}{s}\sqrt{\frac{rs}{r+s}} \quad \text{or} \quad a_0 = 0, \tag{4.10}$$

a_+ for parts in the numerator, a_- for parts in the denominator, and a_0 for parts not involved in that splitting. The balance is then

$$b_i = \sum_{j=1}^{D} a_{ij} \ln x_j \ ,$$

where a_{ij} equals a_+ if the code, at the i-th order partition, is $+1$ for the j-th part; the value is a_- if the code is -1; and $a_0 = 0$ if the code is null, using the values of r and s at the i-th order partition. Note that the matrix with entries a_{ij} is just the matrix $\mathbf{\Psi}$, as shown in the lower part of Table 4.1.

EXAMPLE 4.4.1 *In Egozcue et al. (2003) an orthonormal basis of the simplex was obtained using a Gram-Schmidt technique. It corresponds to the sequential binary partition shown in Table 4.2. The main feature is that the entries of the* $\mathbf{\Psi}$ *matrix can be easily expressed as*

$$\mathbf{\Psi}_{ij} = a_{ji} = +\sqrt{\frac{1}{(D-i)(D-i+1)}} \ , \ j \le D-i \ ,$$

$$\mathbf{\Psi}_{ij} = a_{ji} = -\sqrt{\frac{D-i}{D-i+1}} \ , \ j = D-i+1 \ ;$$

and $\mathbf{\Psi}_{ij} = 0$ *otherwise. This matrix is closely related to Helmert matrices.*

Table 4.2: Example of sign matrix for $D = 5$, used to encode a sequential binary partition in a standard way. The lower part of the table shows the matrix $\mathbf{\Psi}$ of the basis.

level	x_1	x_2	x_3	x_4	x_5	r	s
1	$+1$	$+1$	$+1$	$+1$	-1	4	1
2	$+1$	$+1$	$+1$	-1	0	3	1
3	$+1$	$+1$	-1	0	0	2	1
4	$+1$	-1	0	0	0	1	1
1	$+\frac{1}{\sqrt{20}}$	$+\frac{1}{\sqrt{20}}$	$+\frac{1}{\sqrt{20}}$	$+\frac{1}{\sqrt{20}}$	$-\frac{2}{\sqrt{5}}$		
2	$+\frac{1}{\sqrt{12}}$	$+\frac{1}{\sqrt{12}}$	$+\frac{1}{\sqrt{12}}$	$-\frac{\sqrt{3}}{\sqrt{4}}$	0		
3	$+\frac{1}{\sqrt{6}}$	$+\frac{1}{\sqrt{6}}$	$-\frac{\sqrt{2}}{\sqrt{3}}$	0	0		
4	$+\frac{1}{\sqrt{2}}$	$-\frac{1}{\sqrt{2}}$	0	0	0		

The interpretation of balances relays on some of its properties. The first one is the expression itself, specially when using geometric means in the numerator and denominator as in

$$b = \sqrt{\frac{rs}{r+s}} \ln \frac{(x_1 \cdots x_r)^{1/r}}{(x_{r+1} \cdots x_D)^{1/s}} \ .$$

The geometric means are central values of the parts in each group of parts; its ratio measures the relative weight of each group; the logarithm provides the appropriate scale; and the square root coefficient is a normalising constant which allows to compare numerically different balances. A positive balance means that, in (geometric) mean, the group of parts in the numerator has more weight in the composition than the group in the denominator (and conversely for negative balances).

A second interpretative element is related to the intuitive idea of balance. Imagine that in an election, the parties have been divided into two groups, the left and the right wing ones (there are more than one party in each wing). If, from a journal, you get only the percentages within each group, you are unable to know which wing, and obviously which party, has won the elections. You probably are going to ask for the balance between the two wings as the information you need to complete the actual state of the elections. The balance, as defined here, permits you to complete the information. The balance is the remaining relative information about the elections once the information within the two wings has been removed. To be more precise, assume that the parties are six and the composition of the votes is $\mathbf{x} \in \mathcal{S}^6$; assume the left wing contested with 4 parties represented by the group of parts $\{x_1, x_2, x_5, x_6\}$ and only two parties correspond to the right wing $\{x_3, x_4\}$. Consider the sequential binary partition in Table 4.1. The first partition just separates the two wings and thus the balance informs us about the equilibrium between the two wings. If one

leaves out this balance, the remaining balances inform us only about the left wing (balances 3,4) and only about the right wing (balance 5). Therefore, to retain only balance 5 is equivalent to know the relative information within the subcomposition called right wing. Similarly, balances 2, 3 and 4 only inform about what happened within the left wing. The conclusion is that the balance 1, the forgotten information in the journal, does not inform us about relations within the two wings: it only conveys information about the *balance* between the two groups representing the wings.

Many questions can be stated which can be handled easily using the balances. For instance, suppose we are interested in the relationships between the parties within the left wing and, consequently, we want to remove the information within the right wing. A traditional approach to this is to remove parts x_3 and x_4 and then close the remaining subcomposition. However, this is equivalent to project the composition of 6 parts orthogonally onto the subspace associated with the left wing, what is easily done by setting $b_5 = 0$. If we do so, the obtained projected composition is

$$\mathbf{x}_{\text{proj}} = \mathcal{C}[x_1, x_2, \text{g}(x_3, x_4), \text{g}(x_3, x_4), x_5, x_6] \ , \ \text{g}(x_3, x_4) = (x_3 x_4)^{1/2} \ ,$$

i.e. each part in the right wing has been substituted by the geometric mean within the right wing. This composition still has the information on the left-right balance, b_1. If we are also interested in removing it ($b_1 = 0$), the remaining information will be only that within the left-wing subcomposition which is represented by the orthogonal projection

$$\mathbf{x}_{\text{left}} = \mathcal{C}[x_1, x_2, \text{g}(x_1, x_2, x_5, x_6), \text{g}(x_1, x_2, x_5, x_6), x_5, x_6] \ ,$$

with $\text{g}(x_1, x_2, x_5, x_6) = (x_1, x_2, x_5, x_6)^{1/4}$. The conclusion is that the balances can be very useful to project compositions onto special subspaces just by retaining some balances and making other ones null.

4.5 Working in coordinates

Coordinates with respect to an orthonormal basis in a linear vector space underly standard rules of operation in real space. As a consequence, perturbation in $\mathcal{S}^D$ is equivalent to translation in real space, and power transformation in $\mathcal{S}^D$ is equivalent to multiplication. Thus, if we consider the vector of coordinates $h(\mathbf{x}) = \mathbf{x}^* \in \mathbb{R}^{D-1}$ of a compositional vector $\mathbf{x} \in \mathcal{S}^D$ with respect to an arbitrary orthonormal basis, it holds (Property 4.4.1)

$$h(\mathbf{x} \oplus \mathbf{y}) = h(\mathbf{x}) + h(\mathbf{y}) = \mathbf{x}^* + \mathbf{y}^* \, , \quad h(\alpha \odot \mathbf{x}) = \alpha \cdot h(\mathbf{x}) = \alpha \cdot \mathbf{x}^* \, , \qquad (4.11)$$

and we can think about perturbation as having the same properties in the simplex as translation has in real space, and of the power transformation as having the same properties as multiplication.

Furthermore,

$$\text{d}_a(\mathbf{x}, \mathbf{y}) = \text{d}(h(\mathbf{x}), h(\mathbf{y})) = \text{d}(\mathbf{x}^*, \mathbf{y}^*),$$

where d stands for the usual Euclidean distance in real space. This means that, when performing analysis of compositional data, results that could be obtained using compositions and the Aitchison geometry are exactly the same as those obtained using the coordinates of the compositions and using the ordinary Euclidean geometry. This latter possibility reduces the computations to the ordinary operations in real spaces thus facilitating the applied procedures. The duality of the representation of compositions, in the simplex and by coordinates, introduces a rich framework where both representations can be interpreted to extract conclusions from the analysis (see Figures 4.1, 4.2, 4.3, and 4.4, for illustration). The price is that the basis selected for representation should be carefully selected for an enhanced interpretation.

Working on coordinates can be also done in a blind way, just selecting a default basis and coordinates and, when the results in coordinates are obtained, translating the results back into the simplex for interpretation. This blind strategy, although acceptable, hides to the analyst features of the analysis that may be relevant. For instance, when detecting a linear dependence of compositional data on an external covariate, data can be expressed in coordinates and then the dependence estimated using standard linear regression. Back in the simplex, data can be plotted with the estimated regression line in a ternary diagram. The procedure is completely acceptable but the visual picture of the residuals and a possible non-linear trend in them can be hidden or distorted in the ternary diagram. A plot of the fitted line and the data in coordinates may reveal new interpretable features.

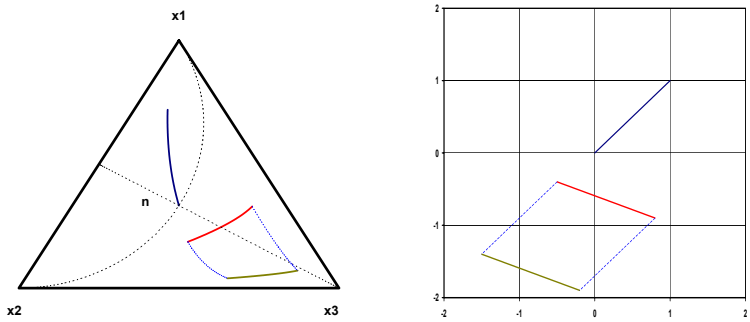

Figure 4.1: Perturbation of a segment in $\mathcal{S}^3$ (left) and in coordinates (right).

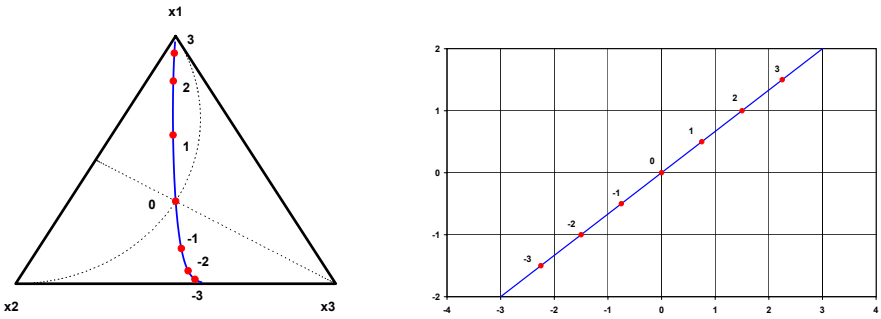

Figure 4.2: Powering of a vector in $\mathcal{S}^3$ (left) and in coordinates (right).

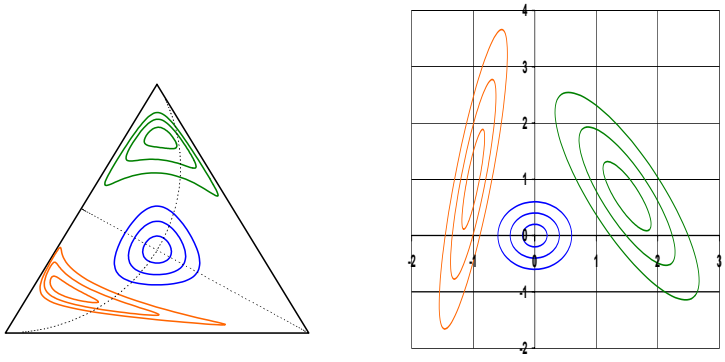

Figure 4.3: Circles and ellipses in $\mathcal{S}^3$ (left) and in coordinates (right).

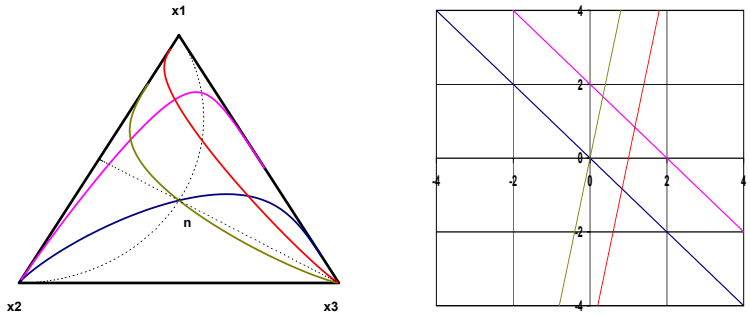

Figure 4.4: Couples of parallel lines in $\mathcal{S}^3$ (left) and in coordinates (right).

One point is essential in the proposed approach: no zero values are allowed, as neither division by zero is admissible, nor taking the logarithm of zero. We are not going to discuss this subject here. Methods on how to approach the problem have been discussed by Aitchison (1986); Aitchison and Kay (2003); Bacon-Shone (2003); Fry et al. (1996); Martín-Fernández (2001) and Martín-Fernández et al. (2000; 2003).

4.6 Additive log-ratio coordinates

In section 4.3 we considered the generating system of the simplex (4.2). One of the elements, e.g. the last one, can be suppressed to obtain a basis: $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{D-1}\}$. Then, any composition $\mathbf{x} \in \mathcal{S}^D$ can be written

$$\begin{aligned} \mathbf{x} &= \bigoplus_{i=1}^{D-1} \ln \frac{x_i}{x_D} \odot \mathbf{w}_i = \\ &= \ln \frac{x_1}{x_D} \odot [e, 1, \ldots, 1, 1] \oplus \cdots \oplus \ln \frac{x_{D-1}}{x_D} \odot [1, 1, \ldots, e, 1] \ . \end{aligned}$$

The coordinates correspond to the well known additive log-ratio transformation (alr) introduced by Aitchison (1986). We will denote by alr the transformation that gives the expression of a composition in additive log-ratio coordinates

$$\mathrm{alr}(\mathbf{x}) = \left[\ln \frac{x_1}{x_D}, \ln \frac{x_2}{x_D}, ..., \ln \frac{x_{D-1}}{x_D} \right] = \mathbf{y}.$$

Note that the alr transformation is not symmetrical in the components. But the essential problem with alr coordinates is the non-isometric character of this transformation. In fact, they are coordinates in an oblique basis, something that affects distances if the usual Euclidean distance is computed from the alr coordinates. This approach is frequent in many applied sciences and should be avoided (see for example Albarède (1995), p. 42).

4.7 Simplicial matrix notation

Many operations in real spaces are expressed in matrix notation. Since the simplex is an Euclidean space, matrix notations may be also useful. However, in this framework a vector of real constants cannot be considered in the simplex although in the real space they are readily identified. This produces two kind of matrix products which are introduced in this section. The first is simply the expression of a perturbation-linear combination of compositions which appears as a power-multiplication of a real vector by a compositional matrix whose rows are in the simplex. The second one is the expression of a linear transformation in the simplex: a composition is transformed by a matrix, involving perturbation and powering, to obtain a new composition. The real matrix implied in this case is not a general one but when expressed in coordinates it is completely general.

Perturbation-linear combination of compositions

For a row vector of ℓ scalars $\mathbf{a} = [a_1, a_2, \ldots, a_\ell]$ and an array of row vectors $\mathbf{V} = (\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v}_\ell)'$, i.e. an (ℓ, D)-matrix,

$$\mathbf{a} \odot \mathbf{V} = [a_1, a_2, \ldots, a_\ell] \odot \begin{pmatrix} \mathbf{v_1} \\ \mathbf{v_2} \\ \vdots \\ \mathbf{v}_\ell \end{pmatrix}$$

$$= [a_1, a_2, \ldots, a_\ell] \odot \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1D} \\ v_{21} & v_{22} & \cdots & v_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ v_{\ell 1} & v_{\ell 2} & \cdots & v_{\ell D} \end{pmatrix} = \bigoplus_{i=1}^{\ell} a_i \odot \mathbf{v}_i .$$

The components of this matrix product are

$$\mathbf{a} \odot \mathbf{V} = \mathcal{C} \left[\prod_{j=1}^{\ell} v_{j1}^{a_j} \ , \ \prod_{j=1}^{\ell} v_{j2}^{a_j} \ , \ldots , \ \prod_{j=1}^{\ell} v_{jD}^{a_j} \right] \ .$$

In coordinates this simplicial matrix product takes the form of a linear combination of the coordinate vectors. In fact, if h is the function assigning the coordinates,

$$h(\mathbf{a} \odot \mathbf{V}) = h\left(\bigoplus_{i=1}^{\ell} a_i \odot \mathbf{v_i} \right) = \sum_{i=1}^{\ell} a_i \ h(\mathbf{v}_i) \ .$$

EXAMPLE 4.7.1 *A composition in $\mathcal{S}^D$ can be expressed as a perturbation-linear combination of the elements of the basis $\mathbf{e}_i$, $i = 1, 2, \ldots, D-1$ as in Equation (4.6). Consider the $(D-1, D)$-matrix $\mathbf{E} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1})'$ and the vector of coordinates $\mathbf{x}^* = \mathrm{ilr}(\mathbf{x})$. Equation (4.6) can be re-written as*

$$\mathbf{x} = \mathbf{x}^* \odot \mathbf{E} \ .$$

Perturbation-linear transformation of $\mathcal{S}^D$: endomorphisms

Consider a row vector of coordinates $\mathbf{x}^* \in \mathbb{R}^{D-1}$ and a general $(D-1, D-1)$-matrix A^*. In the real space setting, $\mathbf{y}^* = \mathbf{x}^* A^*$ expresses an endomorphism, obviously linear in the real sense. Given the isometric isomorphism of the real space of coordinates onto the simplex, the A^* endomorphism has an expression in the simplex. Taking $\text{ilr}^{-1} = h^{-1}$ in the expression of the real endomorphism and using Equation (4.8)

$$\mathbf{y} = \mathcal{C}(\exp[\mathbf{x}^* A^* \boldsymbol{\Psi}]) = \mathcal{C}(\exp[\text{clr}(\mathbf{x}) \boldsymbol{\Psi}' A^* \boldsymbol{\Psi}]) \tag{4.12}$$

where $\boldsymbol{\Psi}$ is the clr matrix of the selected basis and the right-most member has been obtained applying Equation (4.7) to $\mathbf{x}^*$. The (D, D)-matrix $A = \boldsymbol{\Psi}' A^* \boldsymbol{\Psi}$ has entries

$$a_{ij} = \sum_{k=1}^{D-1} \sum_{m=1}^{D-1} \Psi_{ki} \Psi_{mj} a^*_{km} \ , \ i, j = 1, 2, \ldots, D \ .$$

Substituting $\text{clr}(\mathbf{x})$ by its expression as a function of the logarithms of parts, the composition $\mathbf{y}$ is

$$\mathbf{y} = \mathcal{C} \left[\prod_{j=1}^{D} x_j^{a_{j1}} \ , \ \prod_{j=1}^{D} x_j^{a_{j2}} \ , \ldots, \ \prod_{j=1}^{D} x_j^{a_{jD}} \right] \ ,$$

which, taking into account that products and powers match the definitions of $\oplus$ and $\odot$, deserves the definition

$$\mathbf{y} = \mathbf{x} \circ A = \mathbf{x} \circ (\boldsymbol{\Psi}' A^* \boldsymbol{\Psi}) \ , \tag{4.13}$$

where $\circ$ is the perturbation-matrix product representing an endomorphism in the simplex. This matrix product in the simplex should not be confused with that defined between a vector of scalars and a matrix of compositions and denoted by $\odot$.

An important conclusion is that endomorphisms in the simplex are represented by matrices with a peculiar structure given by $A = \boldsymbol{\Psi}' A^* \boldsymbol{\Psi}$, which have some remarkable properties:

(a) it is a (D, D) real matrix;

(b) each row and each column of A adds to 0;

(c) $\text{rank}(A) = \text{rank}(A^*)$; particularly, when A^* is full-rank, $\text{rank}(A) = D - 1$;

(d) the identity endomorphism corresponds to $A^* = I_{D-1}$, the identity in $\mathbb{R}^{D-1}$, and to $A = \boldsymbol{\Psi}' \boldsymbol{\Psi} = I_D - (1/D) \mathbf{1}'_D \mathbf{1}_D$, where I_D is the identity (D, D)-matrix, and $\mathbf{1}_D$ is a row vector of ones.

The matrix A^* can be recovered from A as $A^* = \mathbf{\Psi} A \mathbf{\Psi}'$. However, A is not the only matrix corresponding to A^* in this transformation. Consider the following (D, D)-matrix

$$A = A_0 + \sum_{i=1}^{D} c_i (\vec{e}_i)' \mathbf{1}_D + \sum_{j=1}^{D} d_j \mathbf{1}'_D \vec{e}_j \ ,$$

where, A_0 satisfies the above conditions, $\vec{e}_i = [0, 0, \ldots, 1, \ldots, 0, 0]$ is the i-th row-vector in the canonical basis of $\mathbb{R}^D$, and c_i, d_j are arbitrary constants. Each additive term in this expression adds a constant row or column, being the remaining entries null. A simple development proves that $A^* = \mathbf{\Psi} A \mathbf{\Psi}' = \mathbf{\Psi} A_0 \mathbf{\Psi}'$. This means that $\mathbf{x} \circ A = \mathbf{x} \circ A_0$, i.e. A, A_0 define the same linear transformation in the simplex. To obtain A_0 from A, first compute $A^* = \mathbf{\Psi} A \mathbf{\Psi}'$ and then compute

$$A_0 = \mathbf{\Psi}' A^* \mathbf{\Psi} = \mathbf{\Psi}' \mathbf{\Psi} A \mathbf{\Psi}' \mathbf{\Psi} = (I_D - (1/D)\mathbf{1}'_D \mathbf{1}_D) A (I_D - (1/D)\mathbf{1}'_D \mathbf{1}_D) \ ,$$

where the second member is the required computation and the third member explains that the computation is equivalent to add constant rows and columns to A.

EXAMPLE 4.7.2 *Consider the matrix*

$$A = \begin{pmatrix} 0 & a_2 \\ a_1 & 0 \end{pmatrix}$$

representing a linear transformation in $\mathcal{S}^2$. *The matrix* $\mathbf{\Psi}$ *is*

$$\mathbf{\Psi} = \left[\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}} \right] \ .$$

In coordinates, this corresponds to a $(1,1)$*-matrix* $A^* = (-(a_1 + a_2)/2)$. *The equivalent matrix* $A_0 = \mathbf{\Psi}' A^* \mathbf{\Psi}$ *is*

$$A_0 = \begin{pmatrix} -\frac{a_1+a_2}{4} & \frac{a_1+a_2}{4} \\ \frac{a_1+a_2}{4} & -\frac{a_1+a_2}{4} \end{pmatrix} \ ,$$

whose columns and rows add to 0.

4.8 Exercises

EXERCISE 4.8.1 *Consider the data set given in Table 2.1. Compute the* clr *coefficients (Eq. 4.3) to compositions with no zeros. Verify that the sum of the transformed components equals zero.*

EXERCISE 4.8.2 *Using the sign matrix of Table 4.1 and Equation (4.10), compute the coefficients for each part at each level. Arrange them in a* 6×5 *matrix. Which are the vectors of this basis?*

EXERCISE 4.8.3 *Consider the 6-part composition*

$$[x_1, x_2, x_3, x_4, x_5, x_6] = [3.74, 9.35, 16.82, 18.69, 23.36, 28.04]\,\%\,.$$

Using the binary partition of Table 4.1 and Eq. (4.9), compute its 5 balances. Compare with what you obtained in the preceding exercise.

EXERCISE 4.8.4 *Consider the log-ratios* $c_1 = \ln x_1/x_3$ *and* $c_2 = \ln x_2/x_3$ *in a simplex* $\mathcal{S}^3$*. They are coordinates when using the* alr *transformation. Find two unitary vectors* $\mathbf{e}_1$ *and* $\mathbf{e}_2$ *such that* $\langle \mathbf{x}, \mathbf{e}_i \rangle_a = c_i$*,* $i = 1, 2$*. Compute the inner product* $\langle \mathbf{e}_1, \mathbf{e}_2 \rangle_a$ *and determine the angle between them. Does the result change if the considered simplex is* $\mathcal{S}^7$*?*

EXERCISE 4.8.5 *When computing the* clr *of a composition* $\mathbf{x} \in \mathcal{S}^D$*, a* clr *coefficient is* $\xi_i = \ln(x_i/g(\mathbf{x}))$*. This can be consider as a balance between two groups of parts, which are they and which is the corresponding balancing element?*

EXERCISE 4.8.6 *Six parties have contested elections. In five districts they have obtained the votes in Table 4.3. Parties are divided into left (L) and right (R) wings. Is there some relationship between the L-R balance and the relative votes of R1-R2? Select an adequate sequential binary partition to analyse this question and obtain the corresponding balance coordinates. Find the correlation matrix of the balances and give an interpretation to the maximum correlated two balances. Compute the distances between the five districts; which are the two districts with the maximum and minimum inter-distance. Are you able to distinguish some cluster of districts?*

Table 4.3: Votes obtained by six parties in five districts.

	L1	L2	R1	R2	L3	L4
d1	10	223	534	23	154	161
d2	43	154	338	43	120	123
d3	3	78	29	702	265	110
d4	5	107	58	598	123	92
d5	17	91	112	487	80	90

EXERCISE 4.8.7 *Consider the data set given in Table 2.1. Check the data for zeros. Apply the* alr *transformation to compositions with no zeros. Plot the transformed data in* $\mathbb{R}^2$*.*

EXERCISE 4.8.8 *Consider the data set given in table 2.1 and take the components in a different order. Apply the* alr *transformation to compositions with no zeros. Plot the transformed data in* $\mathbb{R}^2$*. Compare the result with those obtained in Exercise 4.8.7.*

EXERCISE 4.8.9 *Consider the data set given in table 2.1. Apply the* ilr *transformation to compositions with no zeros. Plot the transformed data in* $\mathbb{R}^2$. *Compare the result with the scatterplots obtained in exercises 4.8.7 and 4.8.8 using the* alr *transformation.*

EXERCISE 4.8.10 *Compute the* alr *and* ilr *coordinates, as well as the* clr *coefficients of the 6-part composition*

$$[x_1, x_2, x_3, x_4, x_5, x_6] = [3.74, 9.35, 16.82, 18.69, 23.36, 28.04]\,\% \,.$$

EXERCISE 4.8.11 *Consider the 6-part composition of the preceding exercise. Using the binary partition of Table 4.1 and Equation (4.9), compute its 5 balances. Compare with the results of the preceding exercise.*

Chapter 5

Exploratory data analysis

5.1 General remarks

In this chapter we are going to address the first steps that should be performed whenever the study of a compositional data set $\mathbf{X}$ is initiated. Essentially, these steps are five. They consist in (1) computing descriptive statistics, i.e. the centre and variation matrix of a data set, as well as its total variability; (2) centring the data set for a better visualisation of subcompositions in ternary diagrams; (3) looking at the biplot of the data set to discover patterns; (4) defining an appropriate representation in orthonormal coordinates and computing the corresponding coordinates; and (5) compute the summary statistics of the coordinates and represent the results in a balance-dendrogram. In general, the last two steps will be based on a particular sequential binary partition, defined either a priori or as a result of the insights provided by the preceding three steps. The last step consist of a graphical representation of the sequential binary partition, including a graphical and numerical summary of descriptive statistics of the associated coordinates.

Before starting, let us make some general considerations. The first thing in standard statistical analysis is to check the data set for errors, and we assume this part has been already done using standard procedures (e.g. using the minimum and maximum of each component to check whether the values are within an acceptable range). Another, quite different thing is to check the data set for outliers, a point that is outside the scope of this short-course. See Barceló et al. (1994, 1996) for details. Recall that outliers can be considered as such only with respect to a given distribution. Furthermore, we assume there are no zeros in our samples. Zeros require specific techniques (Aitchison and Kay, 2003; Bacon-Shone, 2003; Fry et al., 1996; Martín-Fernández, 2001; Martín-Fernández et al., 2000; Martín-Fernández et al., 2003) and will be addressed in future editions of this short course.

5.2 Centre, total variance and variation matrix

Standard descriptive statistics are not very informative in the case of compositional data. In particular, the arithmetic mean and the variance or standard deviation of individual components do not fit with the Aitchison geometry as measures of central tendency and dispersion. The skeptic reader might convince himself/herself by doing exercise 5.8.1 immediately. These statistics were defined as such in the framework of Euclidean geometry in real space, which is not a sensible geometry for compositional data. Therefore, it is necessary to introduce alternatives, which we find in the concepts of *centre* (Aitchison, 1997), *variation matrix*, and *total variance* (Aitchison, 1986).

DEFINITION 5.2.1 *A measure of central tendency for compositional data is the closed geometric mean. For a data set of size n it is called* centre *and is defined as*

$$\mathbf{g} = \mathcal{C}\left[g_1, g_2, \ldots, g_D\right] ,$$

with $g_i = (\prod_{j=1}^{n} x_{ij})^{1/n}$, $i = 1, 2, \ldots, D$.

Note that in the definition of centre of a data set the geometric mean is considered column-wise (i.e. by parts), while in the clr transformation, given in equation (4.3), the geometric mean is considered row-wise (i.e. by samples).

DEFINITION 5.2.2 *Dispersion in a compositional data set can be described either by the* variation matrix*, originally defined by Aitchison (1986) as*

$$\mathbf{T} = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1D} \\ t_{21} & t_{22} & \cdots & t_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ t_{D1} & t_{D2} & \cdots & t_{DD} \end{pmatrix} , \quad t_{ij} = \operatorname{var}\left(\ln \frac{x_i}{x_j}\right) ,$$

or by the normalised variation matrix

$$\mathbf{T}^* = \begin{pmatrix} t^*_{11} & t^*_{12} & \cdots & t^*_{1D} \\ t^*_{21} & t^*_{22} & \cdots & t^*_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ t^*_{D1} & t^*_{D2} & \cdots & t^*_{DD} \end{pmatrix} , \quad t^*_{ij} = \operatorname{var}\left(\frac{1}{\sqrt{2}} \ln \frac{x_i}{x_j}\right) .$$

As can be seen, t_{ij} stands for the usual experimental variance of the log-ratio of parts i and j, while t^*_{ij} stands for the usual experimental variance of the normalised log-ratio of parts i and j, so that the log ratio is a balance.

Note that

$$t^*_{ij} = \operatorname{var}\left(\frac{1}{\sqrt{2}} \ln \frac{x_i}{x_j}\right) = \frac{1}{2} t_{ij} ,$$

and thus $\mathbf{T}^* = \frac{1}{2}\mathbf{T}$. Normalised variations have squared Aitchison distance units (see Figure 3.3).

DEFINITION 5.2.3 *A measure of global dispersion is the* total variance *given by*

$$\text{totvar}[\mathbf{X}] = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\text{var}\left(\ln\frac{x_i}{x_j}\right) = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}t_{ij} = \frac{1}{D}\sum_{i=1}^{D}\sum_{j=1}^{D}t^*_{ij}\,.$$

By definition, $\mathbf{T}$ and $\mathbf{T}^*$ are symmetric and their diagonal will contain only zeros. Furthermore, neither the total variance nor any single entry in both variation matrices depend on the constant κ associated with the sample space $\mathcal{S}^D$, as constants cancel out when taking ratios. Consequently, rescaling has no effect. These statistics have further connections. From their definition, it is clear that the total variation summarises the variation matrix in a single quantity, both in the normalised and non-normalised version, and it is possible (and natural) to define it because all parts in a composition share a common scale (it is by no means so straightforward to define a total variation for a pressure-temperature random vector, for instance). Conversely, the variation matrix, again in both versions, explains how the total variation is split among the parts (or better, among all log-ratios).

5.3 Centring and scaling

A usual way in geology to visualise data in a ternary diagram is to rescale the observations in such a way that their range is approximately the same. This is nothing else but applying a perturbation to the data set, a perturbation which is usually chosen by trial and error. To overcome this somehow arbitrary approach, note that, as mentioned in Proposition 3.2.1, for a composition $\mathbf{x}$ and its inverse $\mathbf{x}^{-1}$ it holds that $\mathbf{x}\oplus\mathbf{x}^{-1} = \mathbf{n}$, the neutral element. This means that perturbation allows to move any composition to the barycentre of the simplex, in the same way that translation moves real data in real space to the origin. This property, together with the definition of centre, allows to design a strategy to better visualise the structure of the sample. In fact, computing the centre $\mathbf{g}$ of the data set, as in Definition 5.2.1, and perturbing each sample composition by the inverse $\mathbf{g}^{-1}$, the centre of a data set is shifted to the barycentre of the simplex, and the sample will gravitate around the barycentre.

This property was first introduced by Martín-Fernández et al. (1999) and used by Buccianti et al. (1999). An extensive discussion can be found in von Eynatten et al. (2002), where it is shown that a perturbation transforms straight lines into straight lines. This allows the inclusion of gridlines and compositional fields in the graphical representation without the risk of a nonlinear distortion. See Figure 5.1 for an example of a data set before and after perturbation with the inverse of the closed geometric mean and the effect on the gridlines.

In the same way in real space a centred variable can be scaled to unit variance dividing it by the standard deviation, a (centred) compositional data set $\mathbf{X}$ can be scaled by powering it with $\text{totvar}[\mathbf{X}]^{-1/2}$. In this way, a data set with unit total variance is obtained, but with the same relative contribution of each log-ratio in the variation array. This is a significant difference with conventional

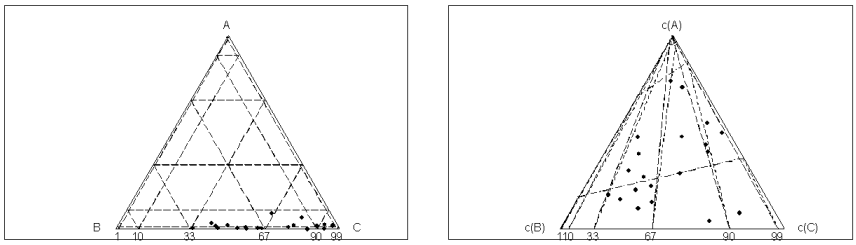

Figure 5.1: Simulated data set before (left) and after (right) centring.

standardisation: with real vectors, the relative contributions are an artifact of the units of each variable, and most usually should be ignored; in contrast, in compositional vectors, all parts share the same "units", and their relative contribution to the total variation is a rich information.

5.4 The biplot: a graphical display

Gabriel (1971) introduced the biplot to represent simultaneously the rows and columns of any matrix by means of a rank-2 approximation. Aitchison (1997) adapted it for compositional data and proved it to be a useful exploratory and expository tool. Here we briefly describe first the philosophy and mathematics of this technique, and then its interpretation in depth.

5.4.1 Construction of a biplot

Consider the data matrix $\mathbf{X}$ with n rows and D columns. Thus, D measurements have been obtained from each one of n samples. Centre the data set as described in Section 5.3, and find the coefficients $\mathbf{Z}$ in clr coordinates (Eq. 4.3). Note that $\mathbf{Z}$ is of the same order as $\mathbf{X}$, i.e. it has n rows and D columns and recall that clr coordinates preserve distances. Thus, standard results can be applied to $\mathbf{Z}$, and in particular the fact that the best rank-2 approximation $\mathbf{Y}$ to $\mathbf{Z}$, in the least squares sense, is provided by the singular value decomposition of $\mathbf{Z}$ (Krzanowski, 1988, p. 126-128).

The singular value decomposition of a matrix $\mathbf{Z}$ is obtained from the matrix of eigenvectors $\mathbf{U}$ of $\mathbf{Z}\mathbf{Z}'$, the matrix of eigenvectors $\mathbf{V}$ of $\mathbf{Z}'\mathbf{Z}$ and the square roots of the s, $s \leq \min\{(D-1), n\}$ positive eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_s$ of either $\mathbf{Z}\mathbf{Z}'$ or $\mathbf{Z}'\mathbf{Z}$, which are equal up to additional null eigenvalues. As a result, taking $k_i = \lambda_i^{1/2}$, we can write

$$\mathbf{Z} = \mathbf{U} \begin{pmatrix} k_1 & 0 & \cdots & 0 \\ 0 & k_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & k_s \end{pmatrix} \mathbf{V}', \tag{5.1}$$

where s is the rank of $\mathbf{Z}$ and the singular values $k_1, k_2, \ldots, k_s$ are in descending order of magnitude. The matrix $\mathbf{U}$ has dimensions (n, s) and $\mathbf{V}$ is a (D, s)-matrix. Both matrices $\mathbf{U}$ and $\mathbf{V}$ are orthonormal, i.e. $\mathbf{U}\mathbf{U}' = I_s$, $\mathbf{V}\mathbf{V}' = I_s$. When $\mathbf{Z}$ is made of centered clr's of compositional data, its rows add to zero and consequently its rank is $s \leq D - 1$ being the common case $s = D - 1$. The interpretation of SVD (5.1) is straightforward. Each row of matrix $\mathbf{V}'$ is the clr of an element of an orthonormal basis of the simplex. This kind of matrices have been denoted Ψ in chapter 4, section 4.4. The matrix product $\mathbf{U}\ \mathrm{diag}(k_1, k_2, \cdots, k_s)$ is an (n, s)-matrix whose n rows contain the coordinates of each compositional data point with respect to the orthonormal basis described by $\mathbf{V}'$. Therefore, $\mathbf{U}\ \mathrm{diag}(k_1, k_2, \cdots, k_s)$ contains ilr-coordinates of the (centered)-compositional data set. Note that these ilr-coordinates are not balances but general orthonormal coordinates. Singular values $\lambda_1 = k_1^2$, $\lambda_2 = k_2^2$, $\ldots$, $\lambda_s = k_s^2$, are proportional to the sample variance of the coordinates.

In order to reduce the dimension of the compositional data set, we can suppress some orthogonal coordinates, typically those with associated low variance. This can be thought as a deletion of small square-singular values. Assume that we retain singular values k_1, k_2, $\ldots$, k_t, $(t \leq s)$. Then the proportion of retained variance is

$$\frac{k_1^2 + k_2^2 + \cdots + k_t^2}{k_1^2 + k_2^2 + \cdots + k_s^2} = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_t}{\lambda_1 + \lambda_2 + \cdots + \lambda_s} \ .$$

The biplot is normally drawn in two dimensions, or at most three dimensions, and then we normally take $t = 2$, provided that the proportion of explained variance is high. This rank-2 approximation is then obtained by simply substituting all singular values with index larger then 2 by zero. As a result we get a rank-2 approximation of $\mathbf{Z}$

$$\mathbf{Y} = \begin{pmatrix} u_{11} & u_{21} \\ u_{12} & u_{22} \\ \vdots & \vdots \\ u_{1n} & u_{2n} \end{pmatrix} \begin{pmatrix} k_1 & 0 \\ 0 & k_2 \end{pmatrix} \begin{pmatrix} v_{11} & v_{21} & \cdots & v_{D1} \\ v_{12} & v_{22} & \cdots & v_{D2} \end{pmatrix} . \tag{5.2}$$

The proportion of variability retained by this approximation is $(\lambda_1+\lambda_2)/(\sum_{i=1}^{s} \lambda_i)$.

To obtain a biplot, it is first necessary to write $\mathbf{Y}$ as the product of two matrices $\mathbf{G}\mathbf{H}'$, where $\mathbf{G}$ is an $(n, 2)$ matrix and $\mathbf{H}$ is an $(D, 2)$ matrix. There are different possibilities to obtain such a factorisation, one of which is

$$\mathbf{Y} = \begin{pmatrix} \sqrt{n}\ u_{11} & \sqrt{n}\ u_{21} \\ \sqrt{n}\ u_{12} & \sqrt{n}\ u_{22} \\ \vdots & \vdots \\ \sqrt{n}\ u_{1n} & \sqrt{n}\ u_{2n} \end{pmatrix} \begin{pmatrix} \frac{k_1 v_{11}}{\sqrt{n}} & \frac{k_1 v_{21}}{\sqrt{n}} & \cdots & \frac{k_1 v_{D1}}{\sqrt{n}} \\ \frac{k_2 v_{12}}{\sqrt{n}} & \frac{k_2 v_{22}}{\sqrt{n}} & \cdots & \frac{k_2 v_{D2}}{\sqrt{n}} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{g_1} \\ \mathbf{g_2} \\ \vdots \\ \mathbf{g_n} \end{pmatrix} \begin{pmatrix} \mathbf{h_1} & \mathbf{h_2} & \cdots & \mathbf{h_D} \end{pmatrix} .$$

The biplot consists simply in representing the vectors $\mathbf{g}_i$, $i = 1, 2, ..., n$ (row vectors of two components), and $\mathbf{h}_j$, $j = 1, 2, ..., D$ (column vectors of two components), on a plane. The vectors $\mathbf{g}_1$, $\mathbf{g}_2$, ..., $\mathbf{g}_n$ are termed the row markers of $\mathbf{Y}$ and correspond to the projections of the n samples on the plane defined by the first two eigenvectors of $\mathbf{ZZ}'$. The vectors $\mathbf{h}_1$, $\mathbf{h}_2$, ..., $\mathbf{h}_D$ are the column markers, which correspond to the projections of the D clr-parts on the plane defined by the first two eigenvectors of $\mathbf{Z}'\mathbf{Z}$. Both planes can be superposed for a visualisation of the relationship between samples and parts.

5.4.2 Interpretation of a compositional biplot

The biplot graphically displays the rank-2 approximation $\mathbf{Y}$ to $\mathbf{Z}$ given by the singular value decomposition. A biplot of compositional data consists of

1. an *origin* O which represents the centre of the compositional data set,
2. a *vertex* at position $\mathbf{h}_j$ for each of the D parts, and
3. a *case marker* at position $\mathbf{g}_i$ for each of the n samples or cases.

We term the join of O to a vertex $\mathbf{h}_j$ the *ray* $\overline{O\mathbf{h}_j}$ and the join of two vertices $\mathbf{h}_j$ and $\mathbf{h}_k$ the *link* $\overline{\mathbf{h}_j\mathbf{h}_k}$. These features constitute the basic characteristics of a biplot with the following main properties for the interpretation of compositional variability.

1. Links and rays provide information on the relative variability in a compositional data set, as

$$|\overline{\mathbf{h}_j\mathbf{h}_k}|^2 \approx \mathrm{var}\left(\ln\frac{x_j}{x_k}\right) \quad \text{and} \quad |\overline{O\mathbf{h}_j}|^2 \approx \mathrm{var}\left(\ln\frac{x_j}{g(x)}\right).$$

   Nevertheless, one has to be careful in interpreting rays, which cannot be identified neither with $\mathrm{var}(x_j)$ nor with $\mathrm{var}(\ln x_j)$, as they depend on the full composition through $\mathrm{g}(x)$ and vary when a subcomposition is considered.

2. Links provide information on the correlation of subcompositions: if links $\overline{\mathbf{h}_j\mathbf{h}_k}$ and $\overline{\mathbf{h}_i\mathbf{h}_\ell}$ intersect at M then

$$\cos(\mathbf{h}_j M \mathbf{h}_i) \approx \mathrm{corr}\left(\ln\frac{x_j}{x_k}, \ln\frac{x_i}{x_\ell}\right).$$

   Furthermore, if the two links are at right angles, then $\cos(\mathbf{h}_j M \mathbf{h}_i) \approx 0$, and zero correlation of the two log-ratios can be expected. This is useful in investigation of subcompositions for possible independence.

3. Subcompositional analysis: The centre O is the centroid (centre of gravity) of the D vertices $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_D$; ratios are preserved under formation of subcompositions; it follows that the biplot for any subcomposition is

simply formed by selecting the vertices corresponding to the parts of the subcomposition and taking the centre of the subcompositional biplot as the centroid of these vertices.

4. Coincident vertices: If vertices $\mathbf{h}_j$ and $\mathbf{h}_k$ coincide, or nearly so, this means that $\text{var}(\ln(x_j/x_k))$ is zero, or nearly so, and the ratio x_j/x_k is constant, or nearly so. Then, the two involved parts, x_j and x_k can be assumed to be redundant. If the proportion of variance captured by the biplot is not very high, two coincident vertices suggest that $\ln(x_j/x_k)$ is orthogonal to the plane of the biplot, and this might be an indication of the possible independence of that log-ratio and the two first principal directions of the singular value decomposition.

5. Collinear vertices: If a subset of vertices is collinear, it might indicate that the associated subcomposition has a biplot that is one-dimensional, which might mean that the subcomposition has one-dimensional variability, i.e. compositions plot along a compositional line.

From the above aspects of interpretation, it should be clear that links are fundamental elements of a compositional biplot. The lengths of links are (approximately) proportional to variance of simple log-ratios between single elements, as they appear in the variation matrix. The complete constellation of links informs about the compositional covariance structure of simple log-ratios and provides hints about subcompositional variability and independence. Interpretation of the biplot is concerned with its internal geometry and is unaffected by any rotation or mirror-imaging of the diagram. For an illustration, see Section 5.6.

For some applications of biplots to compositional data in a variety of geological contexts see Aitchison (1990), and for a deeper insight into biplots of compositional data, with applications in other disciplines and extensions to conditional biplots, see Aitchison and Greenacre (2002).

5.5 Exploratory analysis of coordinates

Either as a result of the preceding descriptive analysis, or due to a priori knowledge of the problem at hand, we may consider a given sequential binary partition as particularly interesting. In this case, its associated orthonormal coordinates, being a vector of real variables, can be treated with the existing battery of conventional descriptive analysis. If $\mathbf{X}^* = h(\mathbf{X})$ represents the coordinates of the data set—rows contain the coordinates of an individual observation—then its experimental moments satisfy

$$\begin{aligned} \bar{\mathbf{y}}^* &= h(\mathbf{g}) = \mathbf{\Psi} \cdot \text{clr}(\mathbf{g}) = \mathbf{\Psi} \cdot \ln(\mathbf{g}) \\ \mathbf{S}_y &= -\mathbf{\Psi} \cdot \mathbf{T}^* \cdot \mathbf{\Psi}' \end{aligned}$$

with $\mathbf{\Psi}$ the matrix whose rows contain the clr coefficients of the orthonormal basis chosen (see Section 4.4 for its construction); $\mathbf{g}$ the centre of the dataset as

defined in Definition 5.2.1, and $\mathbf{T}^*$ the normalised variation matrix as introduced in Definition 5.2.2.

There is a graphical representation, with the specific aim of representing a system of coordinates based on a sequential binary partition: the CoDa- or balance-dendrogram (Egozcue and Pawlowsky-Glahn, 2006; Pawlowsky-Glahn and Egozcue, 2006). A balance-dendrogram is the joint representation of the following elements:

1. a sequential binary partition, in the form of a tree structure;
2. the sample mean and variance of each balance;
3. a box-plot, summarising the order statistics of each balance.

Each coordinate is represented in a horizontal axis, which limits correspond to a certain range (the same for every coordinate). The vertical bar going up from each one of these coordinate axes represents the variance of that specific coordinate, and the contact point is the coordinate mean. Figure 5.2 shows these elements in an illustrative example.

Given that the range of each coordinate is symmetric (in Figure 5.2 it goes from -3 to $+3$), the box plots closer to one part (or group) indicate that part (or group) is more abundant. Thus, in Figure 5.2, SiO_2 is slightly more abundant than Al_2O_3, there is more FeO than Fe_2O_3, and much more structural oxides (SiO_2 and Al_2O_3) than the rest. Another feature easily read from a balance-dendrogram is symmetry: it can be assessed both by comparison between the several quantile boxes, and looking at the difference between the median (marked as "Q2" in Figure 5.2 right) and the mean.

5.6 Illustration

We are going to use, both for illustration and for the exercises, the data set $\mathbf{X}$ given in table 5.1. They correspond to 17 samples of chemical analysis of rocks from Kilauea Iki lava lake, Hawaii, published by Richter and Moore (1966) and cited by Rollinson (1995).

Originally, 14 parts had been registered, but H_2O^+ and H_2O^- have been omitted because of the large amount of zeros. CO_2 has been kept in the table, to call attention upon parts with some zeros, but has been omitted from the study precisely because of the zeros. This is the strategy to follow if the part is not essential in the characterisation of the phenomenon under study. If the part is essential and the proportion of zeros is high, then we are dealing with two populations, one characterised by zeros in that component and the other by non-zero values. If the part is essential and the proportion of zeros is small, then we can look for input techniques, as explained in the beginning of this chapter.

The centre of this data set is

$$\mathbf{g} = (48.57, 2.35, 11.23, 1.84, 9.91, 0.18, 13.74, 9.65, 1.82, 0.48, 0.22)\,,$$

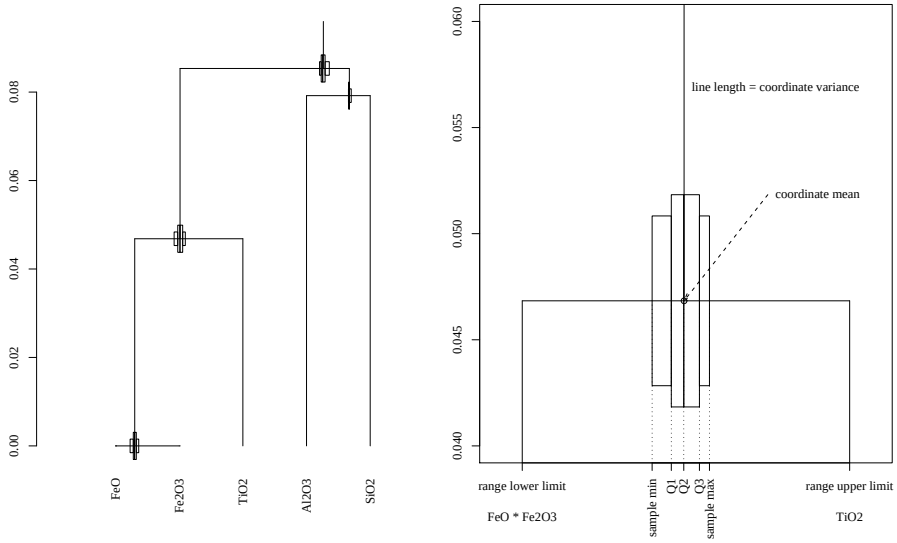

Figure 5.2: Illustration of elements included in a balance-dendrogram. The left subfigure represents a full dendrogram, and the right figure is a zoomed part, corresponding to the balance of (FeO,Fe_2O_3) against TiO_2.

the total variance is totvar[$\mathbf{X}$] = 0.3275 and the normalised variation matrix $\mathbf{T}^*$ is given in Table 5.2.

The biplot (Fig. 5.3), shows an essentially two dimensional pattern of variability, two sets of parts that cluster together, A = [TiO_2, Al_2O_3, CaO, Na_2O, P_2O_5] and B = [SiO_2, FeO, MnO], and a set of one dimensional relationships between parts.

The two dimensional pattern of variability is supported by the fact that the first two axes of the biplot reproduce about 90% of the total variance, as captured in the scree plot in Fig. 5.3, left. The orthogonality of the link between Fe_2O_3 and FeO (i.e., the oxidation state) with the link between MgO and any of the parts in set A might help in finding an explanation for this behaviour and in decomposing the global pattern into two independent processes.

Concerning the two sets of parts we can observe short links between them and, at the same time, that the variances of the corresponding log-ratios (see the normalised variation matrix $\mathbf{T}^*$, Table 5.2) are very close to zero. Consequently, we can say that they are essentially redundant, and that some of them could be either grouped to a single part or simply omitted. In both cases the dimensionality of the problem would be reduced.

Another aspect to be considered is the diverse patterns of one-dimensional variability that can be observed. Examples that can be visualised in a ternary diagram are Fe_2O_3, K_2O and any of the parts in set A, or MgO with any of

Table 5.1: Chemical analysis of rocks from Kilauea Iki lava lake, Hawaii

SiO_2	TiO_2	Al_2O_3	Fe_2O_3	FeO	MnO	MgO	CaO	Na_2O	K_2O	P_2O_5	CO_2
48.29	2.33	11.48	1.59	10.03	0.18	13.58	9.85	1.90	0.44	0.23	0.01
48.83	2.47	12.38	2.15	9.41	0.17	11.08	10.64	2.02	0.47	0.24	0.00
45.61	1.70	8.33	2.12	10.02	0.17	23.06	6.98	1.33	0.32	0.16	0.00
45.50	1.54	8.17	1.60	10.44	0.17	23.87	6.79	1.28	0.31	0.15	0.00
49.27	3.30	12.10	1.77	9.89	0.17	10.46	9.65	2.25	0.65	0.30	0.00
46.53	1.99	9.49	2.16	9.79	0.18	19.28	8.18	1.54	0.38	0.18	0.11
48.12	2.34	11.43	2.26	9.46	0.18	13.65	9.87	1.89	0.46	0.22	0.04
47.93	2.32	11.18	2.46	9.36	0.18	14.33	9.64	1.86	0.45	0.21	0.02
46.96	2.01	9.90	2.13	9.72	0.18	18.31	8.58	1.58	0.37	0.19	0.00
49.16	2.73	12.54	1.83	10.02	0.18	10.05	10.55	2.09	0.56	0.26	0.00
48.41	2.47	11.80	2.81	8.91	0.18	12.52	10.18	1.93	0.48	0.23	0.00
47.90	2.24	11.17	2.41	9.36	0.18	14.64	9.58	1.82	0.41	0.21	0.01
48.45	2.35	11.64	1.04	10.37	0.18	13.23	10.13	1.89	0.45	0.23	0.00
48.98	2.48	12.05	1.39	10.17	0.18	11.18	10.83	1.73	0.80	0.24	0.01
48.74	2.44	11.60	1.38	10.18	0.18	12.35	10.45	1.67	0.79	0.23	0.01
49.61	3.03	12.91	1.60	9.68	0.17	8.84	10.96	2.24	0.55	0.27	0.01
49.20	2.50	12.32	1.26	10.13	0.18	10.51	11.05	2.02	0.48	0.23	0.01

the parts in set A and any of the parts in set B. Let us select one of those subcompositions, e.g. Fe_2O_3, K_2O and Na_2O. After closure, the samples plot in a ternary diagram as shown in Figure 5.4 and we recognise the expected trend and two outliers corresponding to samples 14 and 15, which require further explanation. Regarding the trend itself, notice that it is in fact a line of isoproportion Na_2O/K_2O: thus we can conclude that the ratio of these two parts is independent of the amount of Fe_2O_3.

As a last step, we compute the conventional descriptive statistics of the orthonormal coordinates in a specific reference system (either a priori chosen or derived from the previous steps). In this case, due to our knowledge of the typical geochemistry and mineralogy of basaltic rocks, we choose a priori the set of balances of Table 5.3, where the resulting balance will be interpreted as

1. an oxidation state proxy (Fe^{3+} against Fe^{2+});
2. silica saturation proxy (when Si is lacking, Al takes its place);
3. distribution within heavy minerals (rutile or apatite?);
4. importance of heavy minerals relative to silicates;
5. distribution within plagioclase (albite or anortite?);
6. distribution within feldspar (K-feldspar or plagioclase?);
7. distribution within mafic non-ferric minerals;
8. distribution within mafic minerals (ferric vs. non-ferric);

Table 5.2: Normalised variation matrix of data given in table 5.1. For simplicity, only the upper triangle is represented, omitting the first column and last row.

$\mathrm{var}(\frac{1}{\sqrt{2}}\ln\frac{x_i}{x_j})$	TiO_2	Al_2O_3	Fe_2O_3	FeO	MnO	MgO	CaO	Na_2O	K_2O	P_2O_5
SiO_2	0.012	0.006	0.036	0.001	0.001	0.046	0.007	0.009	0.029	0.011
TiO_2		0.003	0.058	0.019	0.016	0.103	0.005	0.002	0.015	0.000
Al_2O_3			0.050	0.011	0.008	0.084	0.000	0.002	0.017	0.002
Fe_2O_3				0.044	0.035	0.053	0.054	0.050	0.093	0.059
FeO					0.001	0.038	0.012	0.015	0.034	0.017
MnO						0.040	0.009	0.012	0.033	0.015
MgO							0.086	0.092	0.130	0.100
CaO								0.003	0.016	0.004
Na_2O									0.024	0.002
K_2O										0.014

Table 5.3: A possible sequential binary partition for the data set of table 5.1.

balance	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	FeO	MnO	MgO	CaO	Na_2O	K_2O	P_2O_5
v1	0	0	0	+1	-1	0	0	0	0	0	0
v2	+1	0	-1	0	0	0	0	0	0	0	0
v3	0	+1	0	0	0	0	0	0	0	0	-1
v4	+1	-1	+1	0	0	0	0	0	0	0	-1
v5	0	0	0	0	0	0	0	+1	-1	0	0
v6	0	0	0	0	0	0	0	+1	+1	-1	0
v7	0	0	0	0	0	+1	-1	0	0	0	0
v8	0	0	0	+1	+1	-1	-1	0	0	0	0
v9	0	0	0	+1	+1	+1	+1	-1	-1	-1	0
v10	+1	+1	+1	-1	-1	-1	-1	-1	-1	-1	+1

9. importance of mafic minerals against feldspar;

10. importance of cation oxides (those filling the crystalline structure of minerals) against frame oxides (those forming that structure, mainly Al and Si).

One should be aware that such an interpretation is totally problem-driven: if we were working with sedimentary rocks, it would have no sense to split MgO and CaO (as they would mainly occur in limestones and associated lithologies), or to group Na_2O with CaO (as they would probably come from different rock types, e.g. siliciclastic against carbonate).

Using the sequential binary partition given in Table 5.3, Figure 5.5 represents the balance-dendrogram of the sample, within the range $(-3, +3)$. This range translates for two part compositions to proportions of $(1.4, 98.6)\%$; i.e. if we look at the balance MgO-MnO the variance bar is placed at the lower extreme of the balance axis, which implies that in this subcomposition MgO represents in average more than 98%, and MnO less than 2%. Looking at the lengths of the

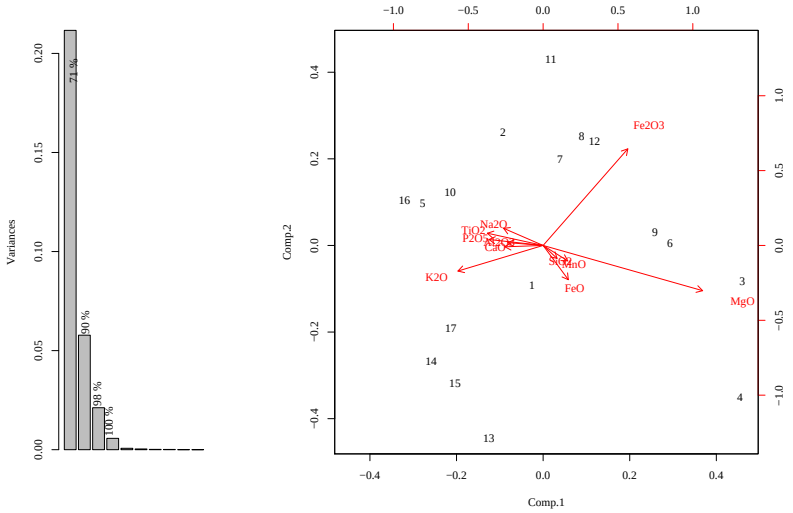

Figure 5.3: Biplot of data of Table 5.1 (right), and scree plot of the variances of all principal components (left), with indication of cumulative explained variance.

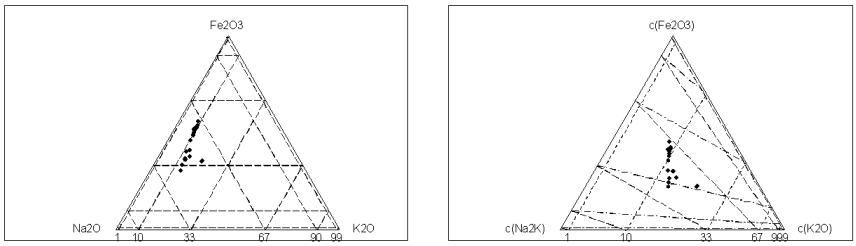

Figure 5.4: Plot of subcomposition (Fe_2O_3,K_2O,Na_2O). Left: before centring. Right: after centring.

several variance bars, one easily finds that the balances P_2O_5-TiO_2 and SiO_2-Al_2O_3 are almost constant, as their bars are very short and their box-plots extremely narrow. Again, the balance between the subcompositions (P_2O_5, TiO_2) vs. (SiO_2, Al_2O_3) does not display any box-plot, meaning that it is above +3 (thus, the second group of parts represents more than 98% with respect to the first group). The distribution between K_2O, Na_2O and CaO tells us that Na_2O and CaO keep a quite constant ratio (thus, we should interpret that there are no strong variations in the plagioclase composition), and the ratio of these two against K_2O is also fairly constant, with the exception of some values below the first quartile (probably, a single value with an particularly high K_2O content). The other balances are well equilibrated (in particular, see how centred is the proxy balance between feldspar and mafic minerals), all with moderate dispersions.

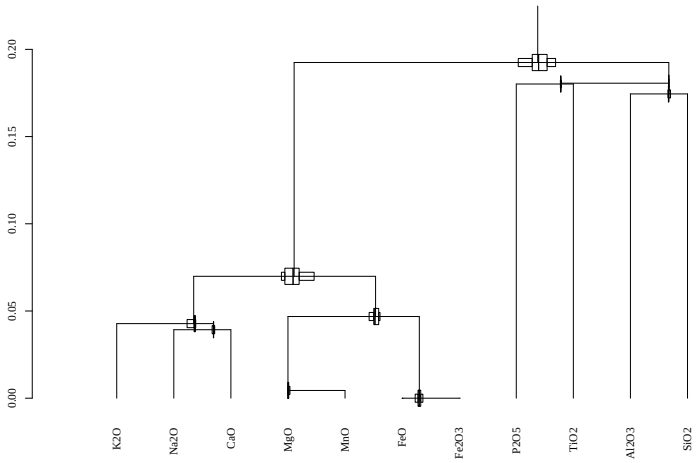

Figure 5.5: Balance-dendrogram of data from Table 5.1 using the balances of Table 5.3.

Table 5.4: Covariance (lower triangle) and correlation (upper triangle) matrices of balances

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10
v1	0.047	0.120	0.341	0.111	-0.283	0.358	-0.212	0.557	0.423	-0.387
v2	0.002	0.006	-0.125	0.788	0.077	0.234	-0.979	-0.695	0.920	-0.899
v3	0.002	-0.000	0.000	-0.345	-0.380	0.018	0.181	0.423	-0.091	0.141
v4	0.003	0.007	-0.001	0.012	0.461	0.365	-0.832	-0.663	0.821	-0.882
v5	-0.004	0.000	-0.000	0.003	0.003	-0.450	-0.087	-0.385	-0.029	-0.275
v6	0.013	0.003	0.000	0.007	-0.004	0.027	-0.328	-0.029	0.505	-0.243
v7	-0.009	-0.016	0.001	-0.019	-0.001	-0.011	0.042	0.668	-0.961	0.936
v8	0.018	-0.008	0.001	-0.011	-0.003	-0.001	0.021	0.023	-0.483	0.516
v9	0.032	0.025	-0.001	0.031	-0.001	0.029	-0.069	-0.026	0.123	-0.936
v10	-0.015	-0.013	0.001	-0.017	-0.003	-0.007	0.035	0.014	-0.059	0.032

Once the marginal empirical distribution of the balances has been analysed, the biplot can be used to explore their relations (Figure 5.6), and the conventional covariance or correlation matrices (Table 5.4). From these, we can see, for instance:

- The constant behaviour of v3 (balance TiO_2-P_2O_5), with a variance below 10^{-4}, and in a lesser degree, of v5 (anortite-albite relation, or balance CaO-Na_2O).

- The orthogonality of the pairs of rays v1-v2, v1-v4, v1-v7, and v6-v8, suggests the lack of correlation of their respective balances, confirmed by Table 5.4, where correlations of less than ± 0.3 are reported. In particular, the pair v6-v8 has a correlation of -0.029. These facts would imply that silica saturation (v2), the presence of heavy minerals (v4) and the MnO-

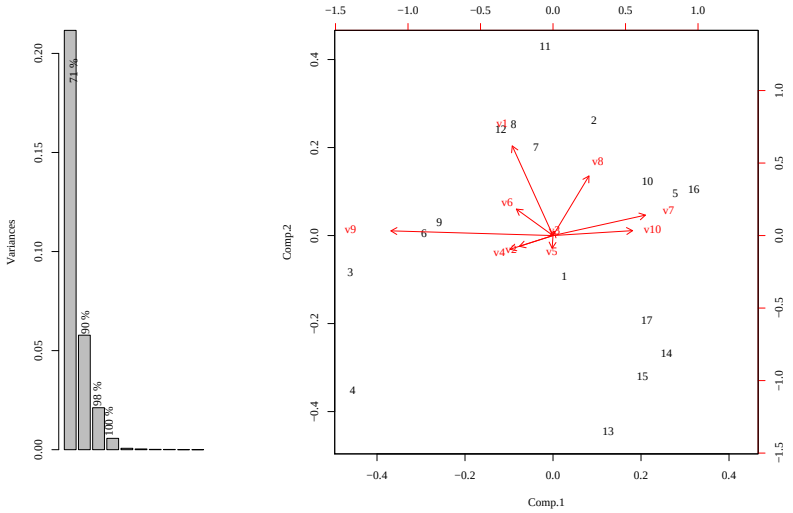

Figure 5.6: Biplot of data of table 5.1 expressed in the balance coordinate system of Table 5.3 (right), and scree plot of the variances of all principal components (left), with indication of cumulative explained variance. Compare with Figure 5.3, in particular: the scree plot, the configuration of data points, and the links between the variables related to balances v1, v2, v3, v5 and v7.

MgO balance (v7) are uncorrelated with the oxidation state (v1); and that the type of feldspars (v6) is unrelated to the type of mafic minerals (v8).

- The balances v9 and v10 are opposite, and their correlation is -0.936, implying that the ratio mafic oxides/feldspar oxides is high when the ratio Silica-Alumina/cation oxides is low, i.e. mafics are poorer in Silica and Alumina.

A final comment regarding balance descriptive statistics: since the balances are chosen due to their interpretability, we are no more just "describing" patterns here. Balance statistics represent a step further towards modeling: all our conclusions in these last three points heavily depend on the preliminary interpretation (="model") of the computed balances.

5.7 Linear trend using principal components

Singular value decomposition (SVD) applied to centered clr-data has been presented in section 5.4 as a technique for dimension reduction of compositional data. In a statistical framework, this technique is known as principal component analysis (PC). As a result the compositional biplot has been shown to be a powerful exploratory tool. Additionally, PC-SVD can be used as the appropriate modeling tool whenever the presence of a trend in the compositional data

set is suspected, but no external variable has been measured on which it might depend (Otero et al., 2003; Tolosana-Delgado et al., 2005). To illustrate this fact let us consider the most simple case, in which one PC (i.e. one square-singular value) explains a large proportion of the total variance, e.g. more then 98%, like the one in Figure 5.7, where the subcomposition [Fe_2O_3,K_2O,Na_2O] from Table 5.1 has been used without samples 14 and 15. Consider the SVD

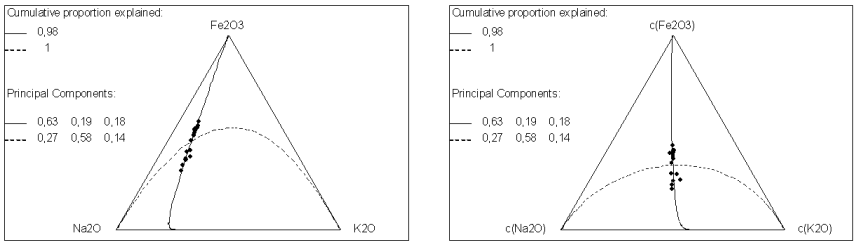

Figure 5.7: Principal components in $\mathcal{S}^3$. Left: before centring. Right: after centring

of the centered clr-data as in eq. (5.1). The first PC are the clr-coordinates on the unitary clr-vector $\mathbf{v}_1$ (first column of $\mathbf{V}$). The composition $\mathbf{a_1} = \mathrm{clr}^{-1}(\mathbf{v}_1)$ determines the direction of the first PC in the simplex. The compositional line going through the barycentre of the simplex, $\alpha \odot \mathbf{a}_1$, describes the trend shown by the centred sample, and $\mathbf{g} \oplus \alpha \odot \mathbf{a}_1$, with $\mathbf{g}$ the centre of the sample, describes the trend shown in the non-centred data set. The evolution of the proportion per unit volume of each part, as described by the first principal component, is shown in Figure 5.8 left. The cumulative proportion is drawn in Figure 5.8 right.

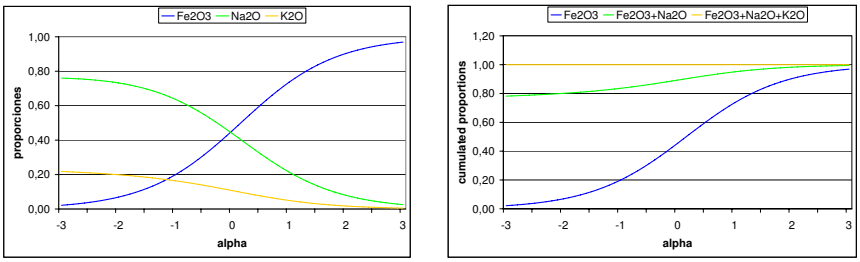

Figure 5.8: Evolution of proportions as described by the first principal component. Left: proportions. Right: cumulated proportions.

To interpret a trend we can use Equation (3.1), which allows us to re-scale the direction of the first PC, assuming whatever is convenient according to the process under study, e.g. that one part is stable. A representation of the result is also possible, as can be seen in Figure 5.9. The part assumed to be stable, K_2O, has a constant, unit perturbation coefficient. We see that under this assumption, within the range of variation of the observations, Na_2O has only a very small increase, while Fe_2O_3 shows a considerable increase compared to

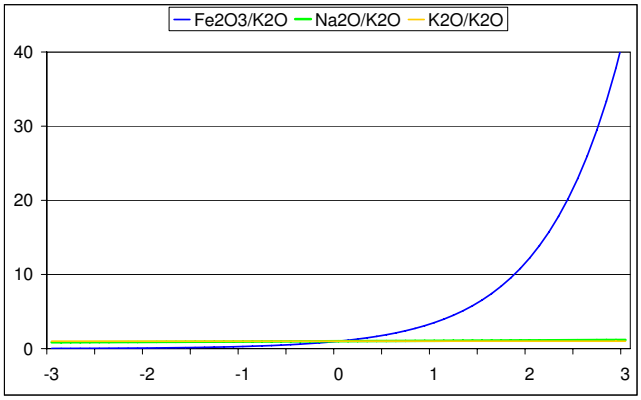

Figure 5.9: Interpretation of a principal component in $\mathcal{S}^2$ under the assumption of stability of K_2O.

the other two parts. In other words, one possible explanation for the observed pattern of variability is that Fe_2O_3 varies significantly, while the other two parts remain stable.

The graph gives additional information: the relative behaviour will be preserved under any assumption. If the assumption is that K_2O increases (decreases), then Na_2O will show the same behaviour as K_2O, while Fe_2O_3 will always change from *below* to *above*.

Note that, although we can represent a perturbation process described by a PC only in a ternary diagram, we can extend the representation in Figure 5.9 to as many parts as we might be interested in.

5.8 Exercises

EXERCISE 5.8.1 *This exercise pretends to illustrate the problems of classical statistics if applied to compositional data. Using the data given in Table 5.1, compute the classical correlation coefficients between the following pairs of parts: (MnO vs. CaO), (FeO vs. Na_2O), (MgO vs. FeO) and (MgO vs. Fe_2O_3). Now ignore the structural oxides Al_2O_3 and SiO_2 from the data set, reclose the remaining variables, recompute the same correlation coefficients as above. Compare the results. Compare the correlation matrix between the feldspar-constituent parts (CaO,Na_2O,K_2O), as obtained from the original data set, and after closing this 3-part subcomposition.*

EXERCISE 5.8.2 *For the data given in Table 2.1 compute and plot the centre with the samples in a ternary diagram. Compute the total variance and the variation matrix.*

EXERCISE 5.8.3 *Perturb the data given in table 2.1 with the inverse of the centre. Compute the centre of the perturbed data set and plot it with the samples in*

a ternary diagram. Compute the total variance and the variation matrix. Compare your results numerically and graphically with those obtained in exercise 5.8.2.

EXERCISE 5.8.4 *Make a biplot of the data given in Table 2.1 and give an interpretation.*

EXERCISE 5.8.5 *Figure 5.3 shows the biplot of the data given in Table 5.1. How would you interpret the different patterns that can be observed?*

EXERCISE 5.8.6 *Select 3-part subcompositions that behave in a particular way in Figure 5.3 and plot them in a ternary diagram. Do they reproduce properties mentioned in the previous description?*

EXERCISE 5.8.7 *Do a scatter plot of the log-ratios*

$$\frac{1}{\sqrt{2}} \ln \frac{\mathrm{K_2O}}{\mathrm{MgO}} \quad \text{against} \quad \frac{1}{\sqrt{2}} \ln \frac{\mathrm{Fe_2O_3}}{\mathrm{FeO}} ,$$

identifying each point. Compare with the biplot. Compute the total variance of the subcomposition (K_2O,MgO,Fe_2O_3,FeO) and compare it with the total variance of the full data set.

EXERCISE 5.8.8 *How would you recast the data in table 5.1 from mass proportion of oxides (as they are) to molar proportions? You may need the following molar weights. Any idea of how to do that with a perturbation?*

SiO_2	TiO_2	Al_2O_3	Fe_2O_3	FeO	MnO
60.085	79.899	101.961	159.692	71.846	70.937

MgO	CaO	Na_2O	K_2O	P_2O_5
40.304	56.079	61.979	94.195	141.945

EXERCISE 5.8.9 *Re-do all the descriptive analysis (and the related exercises) with the Kilauea data set expressed in molar proportions. Compare the results.*

EXERCISE 5.8.10 *Compute the vector of arithmetic means of the* ilr *transformed data from table 2.1. Apply the* ilr^{-1} *backtransformation and compare it with the centre.*

EXERCISE 5.8.11 *Take the parts of the compositions in table 2.1 in a different order. Compute the vector of arithmetic means of the* ilr *transformed sample. Apply the* ilr^{-1} *backtransformation. Compare the result with the previous one.*

EXERCISE 5.8.12 *Centre the data set of table 2.1. Compute the vector of arithmetic means of the* ilr *transformed data. What do you obtain?*

EXERCISE 5.8.13 *Compute the covariance matrix of the* ilr *transformed data set of table 2.1 before and after perturbation with the inverse of the centre. Compare both matrices.*

Chapter 6

Distributions on the simplex

The usual way to pursue any statistical analysis after an exhaustive exploratory analysis consists in assuming and testing distributional assumptions for the random phenomena. This can be easily done for compositional data, as the linear vector space structure of the simplex allows us to express observations with respect to an orthonormal basis, a property that guarantees the proper application of standard statistical methods. The only thing that has to be done is to perform any standard analysis on orthonormal coefficients and to interpret the results in terms of coefficients of the orthonormal basis. Once obtained, the inverse can be used to get the same results in terms of the canonical basis of $\mathbb{R}^D$ (i.e. as compositions summing up to a constant value). The justification of this approach lies in the fact that standard mathematical statistics relies on real analysis, and real analysis is performed on the coefficients with respect to an orthonormal basis in a linear vector space, as discussed by Pawlowsky-Glahn (2003).

There are other ways to justify this approach coming from the side of measure theory and the definition of density function as the Radon-Nikodým derivative of a probability measure (Eaton, 1983), but they would divert us too far from practical applications.

Given that most multivariate techniques rely on the assumption of multivariate normality, we will concentrate on the expression of this distribution in the context of random compositions and address briefly other possibilities.

6.1 The normal distribution on $\mathcal{S}^D$

DEFINITION 6.1.1 *Given a random vector* $\mathbf{x}$ *which sample space is* $\mathcal{S}^D$*, we say that* $\mathbf{x}$ *follows a* normal distribution *on* $\mathcal{S}^D$ *if, and only if, the vector of orthonormal coordinates,* $\mathbf{x}^* = h(\mathbf{x})$*, follows a multivariate normal distribution on* $\mathbb{R}^{D-1}$*.*

To characterise a multivariate normal distribution we need to know its parameters, i.e. the vector of expected values μ and the covariance matrix Σ. In practice, they are seldom, if ever, known, and have to be estimated from the sample. Here the maximum likelihood estimates will be used, which are the vector of arithmetic means $\bar{\mathbf{x}}^*$ for the vector of expected values, and the sample covariance matrix $S_{\mathbf{x}^*}$ with the sample size n as divisor. Remember that, in the case of compositional data, the estimates are computed using the orthonormal coordinates $\mathbf{x}^*$ of the data and not the original measurements.

As we have considered coordinates $\mathbf{x}^*$, we will obtain results in terms of coefficients of $\mathbf{x}^*$ coordinates. To obtain them in terms of the canonical basis of $\mathbb{R}^D$ we have to backtransform whatever we compute by using the inverse transformation $h^{-1}(\mathbf{x}^*)$. In particular, we can backtransform the arithmetic mean $\bar{\mathbf{x}}^*$, which is an adequate measure of central tendency for data which follow reasonably well a multivariate normal distribution. But $h^{-1}(\bar{\mathbf{x}}^*) = \mathbf{g}$, the centre of a compositional data set introduced in Definition 5.2.1, which is an unbiased, minimum variance estimator of the expected value of a random composition (Pawlowsky-Glahn and Egozcue, 2002). Also, as stated in Aitchison (2002), $\mathbf{g}$ is an estimate of $\mathcal{C}\left[\exp(\mathrm{E}[\ln(\mathbf{x})])\right]$, which is the theoretical definition of the closed geometric mean, thus justifying its use.

6.2 Other distributions

Many other distributions on the simplex have been defined (using on $\mathcal{S}^D$ the classical Lebesgue measure on $\mathbb{R}^D$), like e.g. the additive logistic skew normal, the Dirichlet and its extensions, the multivariate normal based on Box-Cox transformations, among others. Some of them have been recently analysed with respect to the linear vector space structure of the simplex (Mateu-Figueras, 2003). This structure has important implications, as the expression of the corresponding density differs from standard formulae when expressed in terms of the metric of the simplex and its associated Lebesgue measure (Pawlowsky-Glahn, 2003). As a result, appealing invariance properties appear: for instance, a normal density on the real line does not change its shape by translation, and thus a normal density in the simplex is also invariant under perturbation; this property is not obtained if one works with the classical Lebesgue measure on $\mathbb{R}^D$. These densities and the associated properties shall be addressed in future extensions of this short course.

6.3 Tests of normality on $\mathcal{S}^D$

Testing distributional assumptions of normality on $\mathcal{S}^D$ is equivalent to test multivariate normality of h transformed compositions. Thus, interest lies in the following test of hypothesis:

$\mathcal{H}_0$**:** the sample comes from a normal distribution on $\mathcal{S}^D$,

$\mathcal{H}_1$**:** the sample comes not from a normal distribution on $\mathcal{S}^D$,

which is equivalent to

$\mathcal{H}_0$**:** the sample of h coordinates comes from a multivariate normal distribution,

$\mathcal{H}_1$**:** the sample of h coordinates comes not from a multivariate normal distribution.

Out of the large number of published tests, for $\mathbf{x}^* \in \mathbb{R}^{D-1}$, Aitchison selected the Anderson-Darling, Cramer-von Mises, and Watson forms for testing hypothesis on samples coming from a uniform distribution. We repeat them here for the sake of completeness, but only in a synthetic form. For clarity we follow the approach used by Pawlowsky-Glahn and Buccianti (2002) and present each case separately; in Aitchison (1986) an integrated approach can be found, in which the orthonormal basis selected for the analysis comes from the singular value decomposition of the data set.

The idea behind the approach is to compute statistics which under the initial hypothesis should follow a uniform distribution in each of the following three cases:

1. all $(D-1)$ marginal, univariate distributions,
2. all $\frac{1}{2}(D-1)(D-2)$ bivariate angle distributions,
3. the $(D-1)$-dimensional radius distribution,

and then use mentioned tests.

Another approach is implemented in the R "compositions" library (**?**), where all pair-wise log-ratios are checked for normality, in the fashion of the variation matrix. This gives $\frac{1}{2}(D-1)(D-2)$ tests of univariate normality: for the hypothesis $\mathcal{H}_0$ to hold, all marginal distributions must be also normal. This condition is thus necessary, but not sufficient (although it is a good indication). Here we will not explain the details of this approach: they are equivalent to marginal univariate distribution tests.

6.3.1 Marginal univariate distributions

We are interested in the distribution of each one of the components of $h(\mathbf{x}) = \mathbf{x}^* \in \mathbb{R}^{D-1}$, called the marginal distributions. For the i-th of those variables, the observations are given by $\langle \mathbf{x}, \mathbf{e}_i \rangle_a$, which explicit expression can be found in Equation 4.7. To perform mentioned tests, proceed as follows:

1. Compute the maximum likelihood estimates of the expected value and the variance:

$$\hat{\mu}_i = \frac{1}{n}\sum_{r=1}^{n} x^*_{ri}, \qquad \hat{\sigma}_i^2 = \frac{1}{n}\sum_{r=1}^{n}(x^*_{ri} - \bar{\mu}_i)^2.$$

Table 6.1: Critical values for marginal test statistics.

Significance level	(%)	10	5	2.5	1
Anderson-Darling	Q_a	0.656	0.787	0.918	1.092
Cramer-von Mises	Q_c	0.104	0.126	0.148	0.178
Watson	Q_w	0.096	0.116	0.136	0.163

2. Obtain from the corresponding tables or using a computer built-in function the values
$$\Phi\left(\frac{x^*_{ri} - \hat{\mu}_i}{\hat{\sigma}_i}\right) = z_r, \quad r = 1, 2, ..., n,$$
where $\Phi(.)$ is the $\mathcal{N}(0; 1)$ cumulative distribution function.

3. Rearrange the values z_r in ascending order of magnitude to obtain the ordered values $z_{(r)}$.

4. Compute the Anderson-Darling statistic for marginal univariate distributions:
$$Q_a = \left(\frac{25}{n^2} - \frac{4}{n} - 1\right)\left(\frac{1}{n}\sum_{r=1}^{n}(2r-1)\left[\ln z_{(r)} + \ln(1 - z_{(n+1-r)})\right] + n\right).$$

5. Compute the Cramer-von Mises statistic for marginal univariate distributions
$$Q_c = \left(\sum_{r=1}^{n}\left(z_{(r)} - \frac{2r-1}{2n}\right)^2 + \frac{1}{12n}\right)\left(\frac{2n+1}{2n}\right).$$

6. Compute the Watson statistic for marginal univariate distributions
$$Q_w = Q_c - \left(\frac{2n+1}{2}\right)\left(\frac{1}{n}\sum_{r=1}^{n} z_{(r)} - \frac{1}{2}\right)^2.$$

7. Compare the results with the critical values in table 6.1. The null hypothesis will be rejected whenever the test statistic lies in the critical region for a given significance level, i.e. it has a value that is larger than the value given in the table.

The underlying idea is that if the observations are indeed normally distributed, then the $z_{(r)}$ should be approximately the order statistics of a uniform distribution over the interval $(0, 1)$. The tests make such comparisons, making due allowance for the fact that the mean and the variance are estimated. Note that to follow the **?** approach, one should apply this scheme to all pair-wise log-ratios, $y = \log(x_i/x_j)$, with $i < j$, instead of to the x^* coordinates of the observations.

A visual representation of each test can be given in the form of a plot in the unit square of the $z_{(r)}$ against the associated order statistic $(2r-1)/(2n)$, $r = 1, 2, ..., n$, of the uniform distribution (a PP plot). Conformity with normality on $\mathcal{S}^D$ corresponds to a pattern of points along the diagonal of the square.

6.3.2 Bivariate angle distribution

The next step consists in analysing the bivariate behaviour of the ilr coordinates. For each pair of indices (i, j), with $i < j$, we can form a set of bivariate observations (x^*_{ri}, x^*_{rj}), $r = 1, 2, ..., n$. The test approach here is based on the following idea: if (u_i, u_j) is distributed as $\mathcal{N}^2(\mathbf{0}; \mathbf{I}^2)$, called a circular normal distribution, then the radian angle between the vector from $(0, 0)$ to (u_i, u_j) and the u_1-axis is distributed uniformly over the interval $(0, 2\pi)$. Since any bivariate normal distribution can be reduced to a circular normal distribution by a suitable transformation, we can apply such a transformation to the bivariate observations and ask if the hypothesis of the resulting angles following a uniform distribution can be accepted. Proceed as follows:

1. For each pair of indices (i, j), with $i < j$, compute the maximum likelihood estimates

$$\begin{pmatrix} \hat{\mu}_i \\ \hat{\mu}_j \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{r=1}^{n} x^*_{ri} \\ \frac{1}{n}\sum_{r=1}^{n} x^*_{rj} \end{pmatrix},$$

$$\begin{pmatrix} \hat{\sigma}_i^2 & \hat{\sigma}_{ij} \\ \hat{\sigma}_{ij} & \hat{\sigma}_j^2 \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{r=1}^{n}(x^*_{ri} - \bar{x}^*_i)^2 & \frac{1}{n}\sum_{r=1}^{n}(x^*_{ri} - \bar{x}^*_i)(x^*_{rj} - \bar{x}^*_j) \\ \frac{1}{n}\sum_{r=1}^{n}(x^*_{ri} - \bar{x}^*_i)(x^*_{rj} - \bar{x}^*_j) & \frac{1}{n}\sum_{r=1}^{n}(x^*_{rj} - \bar{x}^*_j)^2 \end{pmatrix}.$$

2. Compute, for $r = 1, 2, ..., n$,

$$u_r = \frac{1}{\sqrt{\hat{\sigma}_i^2\hat{\sigma}_j^2 - \hat{\sigma}_{ij}^2}}\left[(x^*_{ri} - \hat{\mu}_i)\hat{\sigma}_j - (x^*_{rj} - \hat{\mu}_j)\frac{\hat{\sigma}_{ij}}{\hat{\sigma}_j}\right],$$
$$v_r = (x^*_{rj} - \hat{\mu}_j)/\hat{\sigma}_j.$$

3. Compute the radian angles $\hat{\theta}_r$ required to rotate the u_r-axis anticlockwise about the origin to reach the points (u_r, v_r). If $\arctan(t)$ denotes the angle between $-\frac{1}{2}\pi$ and $\frac{1}{2}\pi$ whose tangent is t, then

$$\hat{\theta}_r = \arctan\left(\frac{v_r}{u_r} + \frac{(1 - \mathrm{sgn}(u_r))\,\pi}{2} + \frac{(1 + \mathrm{sgn}(u_r))\,(1 - \mathrm{sgn}(v_r))\,\pi}{4}\right).$$

4. Rearrange the values of $\hat{\theta}_r/(2\pi)$ in ascending order of magnitude to obtain the ordered values $z_{(r)}$.

Table 6.2: Critical values for the bivariate angle test statistics.

Significance level	(%)	10	5	2.5	1
Anderson-Darling	Q_a	1.933	2.492	3.070	3.857
Cramer-von Mises	Q_c	0.347	0.461	0.581	0.743
Watson	Q_w	0.152	0.187	0.221	0.267

5. Compute the Anderson-Darling statistic for bivariate angle distributions:

$$Q_a = -\frac{1}{n}\sum_{r=1}^{n}(2r-1)\left[\ln z_{(r)} + \ln(1 - z_{(n+1-r)})\right] - n.$$

6. Compute the Cramer-von Mises statistic for bivariate angle distributions

$$Q_c = \left(\sum_{r=1}^{n}\left(z_{(r)} - \frac{2r-1}{2n}\right)^2 - \frac{3.8}{12n} + \frac{0.6}{n^2}\right)\left(\frac{n+1}{n}\right).$$

7. Compute the Watson statistic for bivariate angle distributions

$$Q_w = \left(\sum_{r=1}^{n}\left(z_{(r)} - \frac{2r-1}{2n}\right)^2 - \frac{0.2}{12n} + \frac{0.1}{n^2} - n\left(\frac{1}{n}\sum_{r=1}^{n} z_{(r)} - \frac{1}{2}\right)^2\right)\left(\frac{n+0.8}{n}\right).$$

8. Compare the results with the critical values in Table 6.2. The null hypothesis will be rejected whenever the test statistic lies in the critical region for a given significance level, i.e. it has a value that is larger than the value given in the table.

The same representation as mentioned in the previous section can be used for visual appraisal of conformity with the hypothesis tested.

6.3.3 Radius test

To perform an overall test of multivariate normality, the radius test is going to be used. The basis for it is that, under the assumption of multivariate normality of the orthonormal coordinates, $\mathbf{x}_r^*$, the radii−or squared deviations from the mean−are approximately distributed as $\chi^2(D-1)$; using the cumulative function of this distribution we can obtain again values that should follow a uniform distribution. The steps involved are:

1. Compute the maximum likelihood estimates for the vector of expected values and for the covariance matrix, as described in the previous tests.

2. Compute the radii $u_r = (\mathbf{x}_r^* - \hat{\mu})'\hat{\Sigma}^{-1}(\mathbf{x}_r^* - \hat{\mu})$, $r = 1, 2, ..., n$.

Table 6.3: Critical values for the radius test statistics.

Significance level	(%)	10	5	2.5	1
Anderson-Darling	Q_a	1.933	2.492	3.070	3.857
Cramer-von Mises	Q_c	0.347	0.461	0.581	0.743
Watson	Q_w	0.152	0.187	0.221	0.267

3. Compute $z_r = F(u_r)$, $r = 1, 2, ..., n$, where F is the distribution function of the $\chi^2(D-1)$ distribution.

4. Rearrange the values of z_r in ascending order of magnitude to obtain the ordered values $z_{(r)}$.

5. Compute the Anderson-Darling statistic for radius distributions:

$$Q_a = -\frac{1}{n}\sum_{r=1}^{n}(2r-1)\left[\ln z_{(r)} + \ln(1 - z_{(n+1-r)})\right] - n.$$

6. Compute the Cramer-von Mises statistic for radius distributions

$$Q_c = \left(\sum_{r=1}^{n}\left(z_{(r)} - \frac{2r-1}{2n}\right)^2 - \frac{3.8}{12n} + \frac{0.6}{n^2}\right)\left(\frac{n+1}{n}\right).$$

7. Compute the Watson statistic for radius distributions

$$Q_w = \left(\sum_{r=1}^{n}\left(z_{(r)} - \frac{2r-1}{2n}\right)^2 - \frac{0.2}{12n} + \frac{0.1}{n^2} - n\left(\frac{1}{n}\sum_{r=1}^{n} z_{(r)} - \frac{1}{2}\right)^2\right)\left(\frac{n+0.8}{n}\right).$$

8. Compare the results with the critical values in table 6.3. The null hypothesis will be rejected whenever the test statistic lies in the critical region for a given significance level, i.e. it has a value that is larger than the value given in the table.

Use the same representation described before to assess visually normality on $\mathcal{S}^D$.

6.4 Exercises

EXERCISE 6.4.1 *Test the hypothesis of normality of the marginals of the* ilr *transformed sample of table 2.1.*

EXERCISE 6.4.2 *Test the bivariate normality of each variable pair* (x_i^*, x_j^*), $i < j$, *of the* ilr *transformed sample of table 2.1.*

EXERCISE 6.4.3 *Test the variables of the* ilr *transformed sample of table 2.1 for joint normality.*

Chapter 7

Statistical inference

7.1 Testing hypothesis about two groups

When a sample has been divided into two or more groups, interest may lie in finding out whether there is a real difference between those groups and, if it is the case, whether it is due to differences in the centre, in the covariance structure, or in both. Consider for simplicity two samples of size n_1 and n_2, which are realisation of two random compositions $\mathbf{x}_1$ and $\mathbf{x}_2$, each with an normal distribution on the simplex. Consider the following hypothesis:

1. there is no difference between both groups;
2. the covariance structure is the same, but centres are different;
3. the centres are the same, but the covariance structure is different;
4. the groups differ in their centres and in their covariance structure.

Note that if we accept the first hypothesis, it makes no sense to test the second or the third; the same happens for the second with respect to the third, although these two are exchangeable. This can be considered as a lattice structure in which we go from the bottom or lowest level to the top or highest level until we accept one hypothesis. At that point it makes no sense to test further hypothesis and it is advisable to stop.

To perform tests on these hypothesis, we are going to use coordinates $\mathbf{x}^*$ and to assume they follow each a multivariate normal distribution. For the parameters of the two multivariate normal distributions, the four hypothesis are expressed, in the same order as above, as follows:

1. the vectors of expected values and the covariance matrices are the same: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and $\Sigma_1 = \Sigma_2$;
2. the covariance matrices are the same, but not the vectors of expected values: $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ and $\Sigma_1 = \Sigma_2$;

3. the vectors of expected values are the same, but not the covariance matrices:
   $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and $\Sigma_1 \neq \Sigma_2$;
4. neither the vectors of expected values, nor the covariance matrices are the same:
   $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ and $\Sigma_1 \neq \Sigma_2$.

The last hypothesis is called the model, and the other hypothesis will be confronted with it, to see which one is more plausible. In other words, for each test, the model will be the alternative hypothesis $\mathcal{H}_1$.

For each single case we can use either unbiased or maximum likelihood estimates of the parameters. Under assumptions of multivariate normality, they are identical for the expected values and have a different divisor of the covariance matrix (the sample size n in the maximum likelihood approach, and $n-1$ in the unbiased case). Here developments will be presented in terms of maximum likelihood estimates, as those have been used in the previous chapter. Note that estimators change under each of the possible hypothesis, so each case will be presented separately. The following developments are based on Aitchison (1986, p. 153-158) and Krzanowski (1988, p. 323-329), although for a complete theoretical proof Mardia et al. (1979, section 5.5.3) is recommended. The primary computations from the coordinates, $h(\mathbf{x}_1) = \mathbf{x}_1^*$, of the n_1 samples in one group, and $h(\mathbf{x}_2) = \mathbf{x}_2^*$, of the n_2 samples in the other group, are

1. the separate sample estimates
   (a) of the vectors of expected values:
   $$\hat{\boldsymbol{\mu}}_1 = \frac{1}{n_1}\sum_{r=1}^{n_1}\mathbf{x}_{1r}^*, \quad \hat{\boldsymbol{\mu}}_2 = \frac{1}{n_2}\sum_{s=1}^{n_2}\mathbf{x}_{2s}^*,$$
   (b) of the covariance matrices:
   $$\hat{\Sigma}_1 = \frac{1}{n_1}\sum_{r=1}^{n_1}(\mathbf{x}_{1r}^* - \hat{\boldsymbol{\mu}}_1)(\mathbf{x}_{1r}^* - \hat{\boldsymbol{\mu}}_1)',$$
   $$\hat{\Sigma}_2 = \frac{1}{n_2}\sum_{s=1}^{n_2}(\mathbf{x}_{2s}^* - \hat{\boldsymbol{\mu}}_2)(\mathbf{x}_{2s}^* - \hat{\boldsymbol{\mu}}_2)',$$
2. the pooled covariance matrix estimate:
   $$\hat{\Sigma}_p = \frac{n_1\hat{\Sigma}_1 + n_2\hat{\Sigma}_2}{n_1 + n_2},$$
3. the combined sample estimates
   $$\begin{aligned}\hat{\boldsymbol{\mu}}_c &= \frac{n_1\hat{\boldsymbol{\mu}}_1 + n_2\hat{\boldsymbol{\mu}}_2}{n_1 + n_2},\\ \hat{\Sigma}_c &= \hat{\Sigma}_p + \frac{n_1 n_2(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)'}{(n_1 + n_2)^2}.\end{aligned}$$

To test the different hypothesis, we will use the generalised likelihood ratio test, which is based on the following principles: consider the maximised likelihood function for data $\mathbf{x}^*$ under the null hypothesis, $L_0(\mathbf{x}^*)$ and under the model with no restrictions (case 4), $L_m(\mathbf{x}^*)$. The test statistic is then $R(\mathbf{x}^*) = L_m(\mathbf{x}^*)/L_0(\mathbf{x}^*)$, and the larger the value is, the more critical or resistant to accept the null hypothesis we shall be. In some cases the exact distribution of this cases is known. In those cases where it is not known, we shall use Wilks asymptotic approximation: under the null hypothesis, which places c constraints on the parameters, the test statistic $Q(\mathbf{x}^*) = 2\ln(R(\mathbf{x}^*))$ is distributed approximately as $\chi^2(c)$. For the cases to be studied, the approximate generalised ratio test statistic then takes the form:

$$Q_{0m}(\mathbf{x}^*) = n_1 \ln\left(\frac{|\hat{\Sigma}_{10}|}{|\hat{\Sigma}_{1m}|}\right) + n_2 \ln\left(\frac{|\hat{\Sigma}_{20}|}{|\hat{\Sigma}_{2m}|}\right).$$

1. Equality of centres and covariance structure: The null hypothesis is that $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and $\Sigma_1 = \Sigma_2$, thus we need the estimates of the common parameters $\boldsymbol{\mu} = \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and $\Sigma = \Sigma_1 = \Sigma_2$, which are, respectively, $\hat{\boldsymbol{\mu}}_c$ for $\boldsymbol{\mu}$ and $\hat{\Sigma}_c$ for Σ under the null hypothesis, and $\hat{\boldsymbol{\mu}}_i$ for $\boldsymbol{\mu}_i$ and $\hat{\Sigma}_i$ for Σ_i, $i = 1, 2$, under the model, resulting in a test statistic

$$Q_{1vs4}(\mathbf{x}^*) = n_1 \ln\left(\frac{|\hat{\Sigma}_c|}{|\hat{\Sigma}_1|}\right) + n_2 \ln\left(\frac{|\hat{\Sigma}_c|}{|\hat{\Sigma}_2|}\right) \sim \chi^2\left(\frac{1}{2}D(D-1)\right),$$

   to be compared against the upper percentage points of the $\chi^2\left(\frac{1}{2}D(D-1)\right)$ distribution.

2. Equality of covariance structure with different centres: The null hypothesis is that $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ and $\Sigma_1 = \Sigma_2$, thus we need the estimates of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and of the common covariance matrix $\Sigma = \Sigma_1 = \Sigma_2$, which are $\hat{\Sigma}_p$ for Σ under the null hypothesis and $\hat{\Sigma}_i$ for Σ_i, $i = 1, 2$, under the model, resulting in a test statistic

$$Q_{2vs4}(\mathbf{x}^*) = n_1 \ln\left(\frac{|\hat{\Sigma}_p|}{|\hat{\Sigma}_1|}\right) + n_2 \ln\left(\frac{|\hat{\Sigma}_p|}{|\hat{\Sigma}_2|}\right) \sim \chi^2\left(\frac{1}{2}(D-1)(D-2)\right).$$

3. Equality of centres with different covariance structure: The null hypothesis is that $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and $\Sigma_1 \neq \Sigma_2$, thus we need the estimates of the common centre $\boldsymbol{\mu} = \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and of the covariance matrices Σ_1 and Σ_2. In this case no explicit form for the maximum likelihood estimates exists. Hence the need for a simple iterative method which requires the following steps:

   (a) Set the initial value $\hat{\Sigma}_{ih} = \hat{\Sigma}_i$, $i = 1, 2$;

   (b) compute the common mean, weighted by the variance of each group:

$$\hat{\boldsymbol{\mu}}_h = (n_1\hat{\Sigma}_{1h}^{-1} + n_2\hat{\Sigma}_{2h}^{-1})^{-1}(n_1\hat{\Sigma}_{1h}^{-1}\hat{\boldsymbol{\mu}}_1 + n_2\hat{\Sigma}_{2h}^{-1}\hat{\boldsymbol{\mu}}_2);$$

(c) compute the variances of each group with respect to the common mean:

$$\hat{\Sigma}_{ih} = \hat{\Sigma}_i + (\hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}}_h)(\hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}}_h)' \, , i = 1, 2 \, ;$$

(d) Repeat steps 2 and 3 until convergence.

Thus we have $\hat{\Sigma}_{i0}$ for Σ_i, $i = 1, 2$, under the null hypothesis and $\hat{\Sigma}_i$ for Σ_i, $i = 1, 2$, under the model, resulting in a test statistic

$$Q_{3vs4}(\mathbf{x}^*) = n_1 \ln \left(\frac{|\hat{\Sigma}_{1h}|}{|\hat{\Sigma}_1|} \right) + n_2 \ln \left(\frac{|\hat{\Sigma}_{2h}|}{|\hat{\Sigma}_2|} \right) \sim \chi^2(D-1).$$

7.2 Probability and confidence regions for compositional data

Like confidence intervals, confidence regions are a measure of variability, although in this case it is a measure of joint variability for the variables involved. They can be of interest in themselves, to analyse the precision of the estimation obtained, but more frequently they are used to visualise differences between groups. Recall that for compositional data with three components, confidence regions can be plotted in the corresponding ternary diagram, thus giving evidence of the relative behaviour of the various centres, or of the populations themselves. The following method to compute confidence regions assumes either multivariate normality, or the size of the sample to be large enough for the multivariate central limit theorem to hold.

Consider a composition $\mathbf{x} \in \mathcal{S}^D$ and assume it follows a normal distribution on $\mathcal{S}^D$ as defined in section 6.1.1. Then, the $(D-1)$-variate vector $\mathbf{x}^* = h(\mathbf{x})$ follows a multivariate normal distribution.

Three different cases might be of interest:

1. we know the true mean vector and the true variance matrix of the random vector $\mathbf{x}^*$, and want to plot a probability region;
2. we do not know the mean vector and variance matrix of the random vector, and want to plot a confidence region for its mean using a sample of size n,
3. we do not know the mean vector and variance matrix of the random vector, and want to plot a probability region (incorporating our uncertainty).

In the first case, if a random vector $\mathbf{x}^*$ follows a multivariate normal distribution with known parameters $\boldsymbol{\mu}$ and Σ, then

$$(\mathbf{x}^* - \boldsymbol{\mu})\Sigma^{-1}(\mathbf{x}^* - \boldsymbol{\mu})' \sim \chi^2(D-1),$$

is a chi square distribution of $D-1$ degrees of freedom. Thus, for given α, we may obtain (through software or tables) a value κ such that

$$1 - \alpha = \mathrm{P}\left[(\mathbf{x}^* - \boldsymbol{\mu})\Sigma^{-1}(\mathbf{x}^* - \boldsymbol{\mu})' \leq \kappa\right]. \tag{7.1}$$

This defines a $(1-\alpha)100\%$ probability region centred at $\boldsymbol{\mu}$ in $\mathbb{R}^D$, and consequently $\mathbf{x} = h^{-1}(\mathbf{x}^*)$ defines a $(1-\alpha)100\%$ probability region centred at the mean in the simplex.

Regarding the second case, it is well known that for a sample of size n (and $\mathbf{x}^*$ normally-distributed or n big enough), the maximum likelihood estimates $\bar{\mathbf{x}}^*$ and $\hat{\Sigma}$ satisfy that

$$\frac{n-D+1}{D-1}(\bar{\mathbf{x}}^* - \boldsymbol{\mu})\hat{\Sigma}^{-1}(\bar{\mathbf{x}}^* - \boldsymbol{\mu})' \sim \mathcal{F}(D-1, n-D+1),$$

follows a Fisher $\mathcal{F}$ distribution on $(D-1, n-D+1)$ degrees of freedom (Krzanowski, 1988, see p. 227-228 for further details). Again, for given α, we may obtain a value c such that

$$\begin{aligned} 1-\alpha &= \mathrm{P}\left[\frac{n-D+1}{D-1}(\bar{\mathbf{x}}^* - \boldsymbol{\mu})\hat{\Sigma}^{-1}(\bar{\mathbf{x}}^* - \boldsymbol{\mu})' \leq c\right] \\ &= \mathrm{P}\left[(\bar{\mathbf{x}}^* - \boldsymbol{\mu})\hat{\Sigma}^{-1}(\bar{\mathbf{x}}^* - \boldsymbol{\mu})' \leq \kappa\right], \end{aligned} \tag{7.2}$$

with $\kappa = \frac{D-1}{n-D+1}c$. But $(\bar{\mathbf{x}}^* - \boldsymbol{\mu})\hat{\Sigma}^{-1}(\bar{\mathbf{x}}^* - \boldsymbol{\mu})' = \kappa$ (constant) defines a $(1-\alpha)100\%$ confidence region centred at $\bar{\mathbf{x}}^*$ in $\mathbb{R}^D$, and consequently $\xi = h^{-1}(\boldsymbol{\mu})$ defines a $(1-\alpha)100\%$ confidence region around the centre in the simplex.

Finally, in the third case, one should actually use the multivariate Student-Siegel predictive distribution: a new value of $\mathbf{x}^*$ will have as density

$$f(\mathbf{x}^*|data) \propto \left[\mathbf{1} + (n-1)\left(1-\frac{1}{n}\right)(\mathbf{x}^* - \bar{\mathbf{x}}^*)\cdot\Sigma\right]^{-1} \cdot [(\mathbf{x}^* - \bar{\mathbf{x}}^*)']^{n/2}.$$

This distribution is unfortunately not commonly tabulated, and it is only available in some specific packages. On the other side, if n is large with respect to D, the differences between the first and third options are negligible.

Note that for $D=3$, $D-1=2$ and we have an ellipse in real space, in any of the first two cases: the only difference between them is how the constant κ is computed. The parameterisation equations in polar coordinates, which are necessary to plot these ellipses, are given in Appendix B.

7.3 Exercises

EXERCISE 7.3.1 *Divide the sample of Table 5.1 into two groups (at your will) and perform the different tests on the centres and covariance structures.*

EXERCISE 7.3.2 *Compute and plot a confidence region for the* ilr *transformed mean of the data from table 2.1 in* $\mathbb{R}^2$.

EXERCISE 7.3.3 *Transform the confidence region of exercise 7.3.2 back into the ternary diagram using* ilr^{-1}.

EXERCISE 7.3.4 *Compute and plot a 90% probability region for the* ilr *transformed data of table 2.1 in* $\mathbb{R}^2$*, together with the sample. Use the chi square distribution.*

EXERCISE 7.3.5 *For each of the four hypothesis in section 7.1, compute the number of parameters to be estimated if the composition has D parts. The fourth hypothesis needs more parameters than the other three. How many, with respect to each of the three simpler hypothesis? Compare with the degrees of freedom of the* χ^2 *distributions of page 65.*

Chapter 8

Compositional processes

Compositions can evolve depending on an external parameter like space, time, temperature, pressure, global economic conditions and many other ones. The external parameter may be continuous or discrete. In general, the evolution is expressed as a function $\mathbf{x}(t)$, where t represents the external variable and the image is a composition in $\mathcal{S}^D$. In order to model compositional processes, the study of simple models appearing in practice is very important. However, apparently complicated behaviours represented in ternary diagrams may be close to linear processes in the simplex. The main challenge is frequently to identify compositional processes from available data. This is done using a variety of techniques that depend on the data, the selected model of the process and the prior knowledge about them. Next subsections present three simple examples of such processes. The most important is the linear process in the simplex, that follows a straight-line in the simplex. Other frequent process are the complementary processes and mixtures. In order to identify the models, two standard techniques are presented: regression and principal component analysis in the simplex. The first one is adequate when compositional data are completed with some external covariates. Contrarily, principal component analysis tries to identify directions of maximum variability of data, i.e. a linear process in the simplex with some unobserved covariate.

8.1 Linear processes: exponential growth or decay of mass

Consider D different species of bacteria which reproduce in a rich medium and assume there are no interaction between the species. It is well-known that the mass of each species grows proportionally to the previous mass and this causes an exponential growth of the mass of each species. If t is time and each component of the vector $\mathbf{x}(t)$ represents the mass of a species at the time t, the model is

$$\mathbf{x}(t) = \mathbf{x}(0) \cdot \exp(\boldsymbol{\lambda} t) \ , \tag{8.1}$$

where $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_D]$ contains the rates of growth corresponding to the species. In this case, λ_i will be positive, but one can imagine $\lambda_i = 0$, the i-th species does not vary; or $\lambda_i < 0$, the i-th species decreases with time. Model (8.1) represents a process in which both the total mass of bacteria and the composition of the mass by species are specified. Normally, interest is centred in the compositional aspect of (8.1) which is readily obtained applying a closure to the equation (8.1). From now on, we assume that $\mathbf{x}(t)$ is in $\mathcal{S}^D$.

A simple inspection of (8.1) permits to write it using the operations of the simplex,

$$\mathbf{x}(t) = \mathbf{x}(0) \oplus t \odot \mathbf{p} \ , \ \mathbf{p} = \exp(\boldsymbol{\lambda}) \ , \tag{8.2}$$

where a straight-line is identified: $\mathbf{x}(0)$ is a point on the line taken as the origin; $\mathbf{p}$ is a constant vector representing the direction of the line; and t is a parameter with values on the real line (positive or negative).

The linear character of this process is enhanced when it is represented using coordinates. Select a basis in $\mathcal{S}^D$, for instance, using balances determined by a sequential binary partition, and denote the coordinates $\mathbf{u}(t) = \text{ilr}(\mathbf{x})(t)$, $\mathbf{q} = \text{ilr}(\mathbf{p})$. The model for the coordinates is then

$$\mathbf{u}(t) = \mathbf{u}(0) + t \cdot \mathbf{q} \ , \tag{8.3}$$

a typical expression of a straight-line in $\mathbb{R}^{D-1}$. The processes that follow a straight-line in the simplex are more general than those represented by Equations (8.2) and (8.3), because changing the parameter t by any function $\phi(t)$ in the expression, still produces images on the same straight-line.

EXAMPLE 8.1.1 (GROWTH OF BACTERIA POPULATION) *Set $D = 3$ and consider species 1, 2, 3, whose relative masses were 82.7%, 16.5% and 0.8% at the initial observation ($t = 0$). The rates of growth are known to be $\lambda_1 = 1$, $\lambda_2 = 2$ and $\lambda_3 = 3$. Select the sequential binary partition and balances specified in Table 8.1.*

Table 8.1: Sequential binary partition and balance-coordinates used in the example *growth of bacteria population*

order	x_1	x_2	x_3	balance-coord.
1	+1	+1	−1	$u_1 = \frac{1}{\sqrt{6}} \ln \frac{x_1 x_2}{x_3^2}$
2	+1	−1	0	$u_2 = \frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2}$

The process of growth is shown in Figure 8.1, both in a ternary diagram (left) and in the plane of the selected coordinates (right). Using coordinates it is easy to identify that the process corresponds to a straight-line in the simplex. Figure 8.2 shows the evolution of the process in time in two usual plots: the one on the left shows the evolution of each part-component in per unit; on the right,

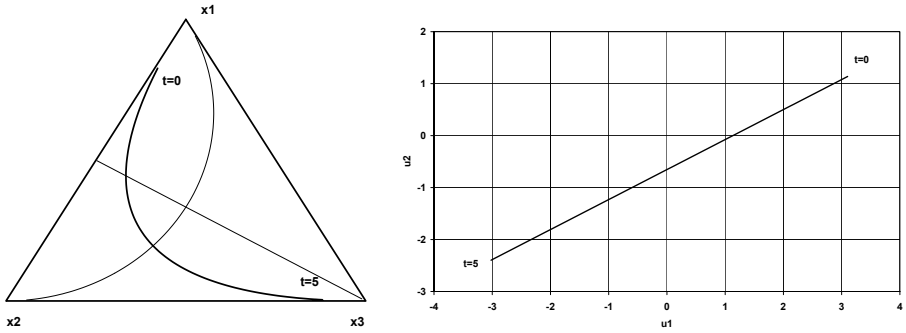

Figure 8.1: Growth of 3 species of bacteria in 5 units of time. Left: ternary diagram; axis used are shown (thin lines). Right: process in coordinates.

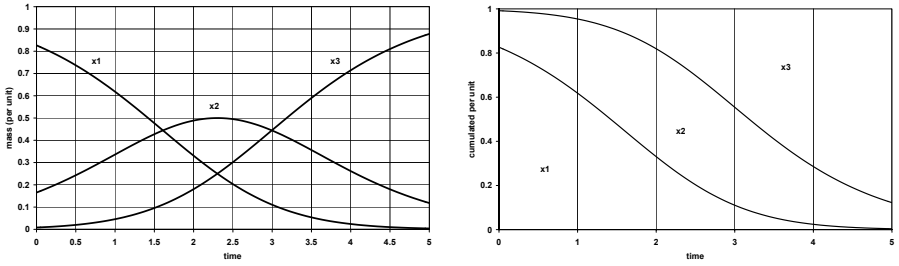

Figure 8.2: Growth of 3 species of bacteria in 5 units of time. Evolution of per unit of mass for each species. Left: per unit of mass. Right: cumulated per unit of mass; x_1, lower band; x_2, intermediate band; x_3 upper band. Note the inversion of abundances of species 1 and 3.

the same evolution is presented as parts adding up to one in a cumulative form. Normally, the graph on the left is more understandable from the point of view of evolution.

EXAMPLE 8.1.2 (WASHING PROCESS) *A liquid reservoir of constant volume V receives an input of the liquid of Q (volume per unit time) and, after a very active mixing inside the reservoir, an equivalent output Q is released. At time $t = 0$, volumes (or masses) $x_1(0)$, $x_2(0)$, $x_3(0)$ of three contaminants are stirred in the reservoir. The contaminant species are assumed non-reactive. Attention is paid to the relative content of the three species at the output in time. The output concentration is proportional to the mass in the reservoir (Albarède, 1995, p.346),*

$$x_i(t) = x_i(0) \cdot \exp\left(-\frac{t}{V/Q}\right) \ , \ i = 1, 2, 3 \ .$$

After closure, this process corresponds to an exponential decay of mass in $\mathcal{S}^3$. The peculiarity is that, in this case, $\lambda_i = -Q/V$ for the three species. A repre-

sentation in orthogonal balances, as functions of time, is

$$u_1(t) = \frac{1}{\sqrt{6}} \ln \frac{x_1(t)x_2(t)}{x_3^2(t)} = \frac{1}{\sqrt{6}} \ln \frac{x_1(0)x_2(0)}{x_3^2(0)} \ ,$$

$$u_2(t) = \frac{1}{\sqrt{2}} \ln \frac{x_1(t)}{x_2(t)} = \frac{1}{\sqrt{2}} \ln \frac{x_1(0)}{x_2(0)} \ .$$

Therefore, from the compositional point of view, the relative concentration of the contaminants in the subcomposition associated with the three contaminants is constant. This is not in contradiction to the fact that the mass of contaminants decays exponentially in time.

EXERCISE 8.1.1 *Select two arbitrary 3-part compositions,* $\mathbf{x}(0)$, $\mathbf{x}(t_1)$, *and consider the linear process from* $\mathbf{x}(0)$ *to* $\mathbf{x}(t_1)$. *Determine the direction of the process normalised to one and the time,* t_1, *necessary to arrive to* $\mathbf{x}(t_1)$. *Plot the process in a) a ternary diagram, b) in balance-coordinates, c) evolution in time of the parts normalised to a constant.*

EXERCISE 8.1.2 *Chose* $\mathbf{x}(0)$ *and* $\mathbf{p}$ *in* $\mathcal{S}^3$. *Consider the process* $\mathbf{x}(t) = \mathbf{x}(0) \oplus t \odot \mathbf{p}$ *with* $0 \leq t \leq 1$. *Assume that the values of the process at* $t = j/49$, $j = 1, 2, \ldots, 50$ *are perturbed by observation errors,* $\mathbf{y}(t)$ *distributed as a normal on the simplex* $\mathcal{N}_s(\mu, \Sigma)$, *with* $\mu = \mathcal{C}[1, 1, 1]$ *and* $\Sigma = \sigma^2 I_3$ *(*I_3 *unit* (3×3)*-matrix). Observation errors are assumed independent of* t *and* $\mathbf{x}(t)$. *Plot* $\mathbf{x}(t)$ *and* $\mathbf{z}(t) = \mathbf{x}(t) \oplus \mathbf{y}(t)$ *in a ternary diagram and in a balance-coordinate plot. Try with different values of* σ^2.

8.2 Complementary processes

Apparently simple compositional processes appear to be non-linear in the simplex. This is the case of systems in which the mass from some components are transferred into other ones, possibly preserving the total mass. For a general instance, consider the radioactive isotopes $\{x_1, x_2, \ldots, x_n\}$ that disintegrate into non-radioactive materials $\{x_{n+1}, x_{n+2}, \ldots, x_D\}$. The process in time t is described by

$$x_i(t) = x_i(0) \cdot \exp(-\lambda_i t) \ , \ x_j(t) = x_j(0) + \sum_{i=1}^{n} a_{ij}(x_i(0) - x_i(t)) \ , \ \sum_{i=1}^{n} a_{ij} = 1 \ ,$$

with $1 \leq i \leq n$ and $n + 1 \leq j \leq D$. From the compositional point of view, the subcomposition corresponding to the first group behaves as a linear process. The second group of parts $\{x_{n+1}, x_{n+2}, \ldots, x_D\}$ is called complementary because it sums up to preserve the total mass in the system and does not evolve linearly despite of its simple form.

EXAMPLE 8.2.1 (ONE RADIOACTIVE ISOTOPE) *Consider the radioactive isotope x_1 which is transformed into the non-radioactive isotope x_3, while the element x_2 remains unaltered. This situation, with $\lambda_1 < 0$, corresponds to*

$$x_1(t) = x_1(0) \cdot \exp(\lambda_1 t) \ , \ x_2(t) = x_2(0) \ , \ x_3(t) = x_3(0) + x_1(0) - x_1(t) \ ,$$

that is mass preserving. The group of parts behaving linearly is $\{x_1, x_2\}$, and a complementary group is $\{x_3\}$. Table 8.2 shows parameters of the model and Figures 8.3 and 8.4 show different aspects of the compositional process from $t = 0$ to $t = 10$.

Table 8.2: Parameters for Example 8.2.1: *one radioactive isotope.* Disintegration rate is $\ln 2$ times the inverse of the half-lifetime. Time units are arbitrary. The lower part of the table represents the sequential binary partition used to define the balance-coordinates.

parameter	x_1	x_2	x_3
disintegration rate	0.5	0.0	0.0
initial mass	1.0	0.4	0.5
balance 1	+1	+1	−1
balance 2	+1	−1	0

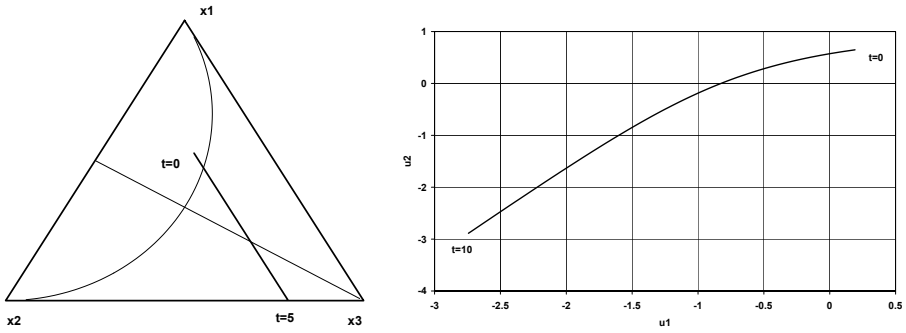

Figure 8.3: Disintegration of one isotope x_1 into x_3 in 10 units of time. Left: ternary diagram; axis used are shown (thin lines). Right: process in coordinates. The mass in the system is constant and the mass of x_2 is constant.

A first inspection of the Figures reveals that the process appears as a segment in the ternary diagram (Fig. 8.3, right). This fact is essentially due to the constant mass of x_2 in a conservative system, thus appearing as a constant per-unit. In figure 8.3, left, the evolution of the coordinates shows that the process is not linear; however, except for initial times, the process may be approximated by a linear one. The linear or non-linear character of the process is hardly detected in Figures 8.4 showing the evolution in time of the composition.

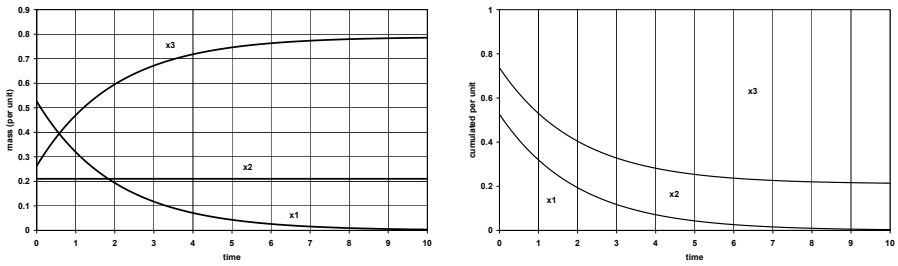

Figure 8.4: Disintegration of one isotope x_1 into x_3 in 10 units of time. Evolution of per unit of mass for each species. Left: per unit of mass. Right: cumulated per unit of mass; x_1, lower band; x_2, intermediate band; x_3 upper band. Note the inversion of abundances of species 1 and 3.

EXAMPLE 8.2.2 (THREE RADIOACTIVE ISOTOPES) *Consider three radioactive isotopes that we identify with a linear group of parts, $\{x_1, x_2, x_3\}$. The disintegrated mass of x_1 is distributed on the non-radioactive parts $\{x_4, x_5, x_6\}$ (complementary group). The whole disintegrated mass from x_2 and x_3 is assigned to x_4 and x_5 respectively. The values of the parameters considered are shown in Table*

Table 8.3: Parameters for Example 8.2.2: *three radioactive isotopes.* Disintegration rate is $\ln 2$ times the inverse of the half-lifetime. Time units are arbitrary. The middle part of the table corresponds to the coefficients a_{ij} indicating the part of the mass from x_i component transformed into the x_j. Note they add to one and the system is mass conservative. The lower part of the table shows the sequential binary partition to define the balance coordinates.

parameter	x_1	x_2	x_3	x_4	x_5	x_6
disintegration rate	0.2	0.04	0.4	0.0	0.0	0.0
initial mass	30.0	50.0	13.0	1.0	1.2	0.7
mass from x_1	0.0	0.0	0.0	0.7	0.2	0.1
mass from x_2	0.0	0.0	0.0	0.0	1.0	0.0
mass from x_3	0.0	0.0	0.0	0.0	0.0	1.0
balance 1	+1	+1	+1	−1	−1	−1
balance 2	+1	+1	−1	0	0	0
balance 3	+1	−1	0	0	0	0
balance 4	0	0	0	+1	+1	−1
balance 5	0	0	0	+1	−1	0

8.3. Figure 8.5 (left) shows the evolution of the subcomposition of the complementary group in 20 time units; no special conclusion is got from it. Contrarily, Figure 8.5 (right), showing the evolution of the coordinates of the subcomposition, reveals a loop in the evolution with a double point (the process passes two times through this compositional point); although less clearly, the same fact can be observed in the representation of the ternary diagram in Figure 8.6. This is a quite surprising and involved behaviour despite of the very simple character

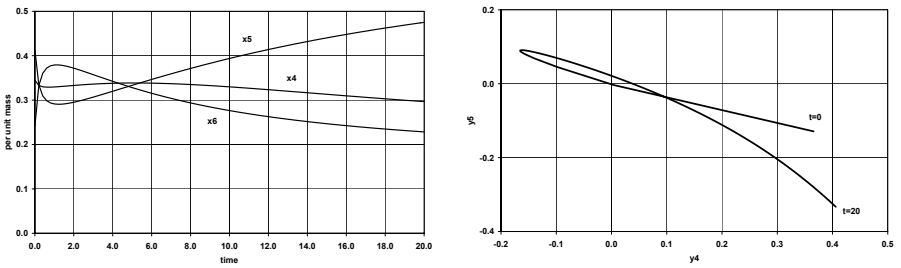

Figure 8.5: Disintegration of three isotopes x_1, x_2, x_3. Disintegration products are masses added to x_4, x_5, x_6 in 20 units of time. Left: evolution of per unit of mass of x_4, x_5, x_6. Right: x_4, x_5, x_6 process in coordinates; a loop and a double point are revealed.

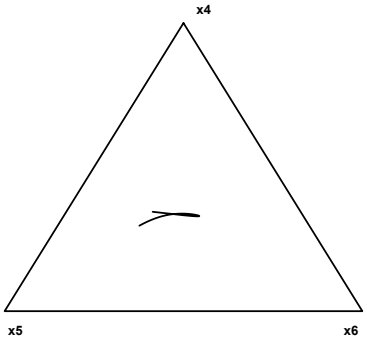

Figure 8.6: Disintegration of three isotopes x_1, x_2, x_3. Products are masses added to x_4, x_5, x_6, in 20 units of time, represented in the ternary diagram. Loop and double point are visible.

of the complementary process. Changing the parameters of the process one can obtain more simple behaviours, for instance without double points or exhibiting less curvature. However, these processes only present one possible double point or a single bend point; the branches far from these points are suitable for a linear approximation.

EXAMPLE 8.2.3 (WASHING PROCESS (CONTINUED)) *Consider the washing process. Let us assume that the liquid is water with density equal to one and define the mass of water* $x_0(t) = V \cdot 1 - \sum x_i(t)$*, that may be considered as a complementary process. The mass concentration at the output is the closure of the four components, being the closure constant proportional to* V*. The compositional process is not a straight-line in the simplex, because the new balance now needed to represent the process is*

$$y_0(t) = \frac{1}{\sqrt{12}} \ln \frac{x_1(t)x_2(t)x_3(t)}{x_0^3(t)} \ ,$$

that is neither a constant nor a linear function of t*.*

EXERCISE 8.2.1 *In the washing process example, set* $x_1(0) = 1.$, $x_2(0) = 2.$, $x_3(0) = 3.$, $V = 100.$, $Q = 5.$. *Find the sequential binary partition used in the example. Plot the evolution in time of the coordinates and mass concentrations including the water* $x_0(t)$. *Plot, in a ternary diagram, the evolution of the subcomposition* x_0, x_1, x_2.

8.3 Mixture process

Another kind of non-linear process in the simplex is that of the mixture processes. Consider two large containers partially filled with D species of materials or liquids with mass (or volume) concentrations given by $\mathbf{x}$ and $\mathbf{y}$ in $\mathcal{S}^D$. The total masses in the containers are m_1 and m_2 respectively. Initially, the concentration in the first container is $\mathbf{z}_0 = \mathbf{x}$. The content of the second container is steadily poured and stirred in the first one. The mass transferred from the second to the first container is ϕm_2 at time t, i.e. $\phi = \phi(t)$. The evolution of mass in the first container, is

$$(m_1 + \phi(t)m_2) \cdot \mathbf{z}(t) = m_1 \cdot \mathbf{x} + \phi(t)m_2 \cdot \mathbf{y} \ ,$$

where $\mathbf{z}(t)$ is the process of the concentration in the first container. Note that $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ are considered closed to 1. The final composition in the first container is

$$\mathbf{z}_1 = \frac{1}{m_1 + m_2}(m_1\mathbf{x} + m_2\mathbf{y}) \tag{8.4}$$

The mixture process can be alternatively expressed as mixture of the initial and final compositions (often called end-points):

$$\mathbf{z}(t) = \alpha(t)\mathbf{z}_0 + (1 - \alpha(t))\mathbf{z}_1 \ ,$$

for some function of time, $\alpha(t)$, where, to fit the physical statement of the process, $0 \leq \alpha \leq 1$. But there is no problem in assuming that α may take values on the whole real-line.

EXAMPLE 8.3.1 (OBTAINING A MIXTURE) *A mixture of three liquids is in a large container A. The numbers of volume units in A for each component are* $[30, 50, 13]$, *i.e. the composition in ppu (parts per unit) is* $\mathbf{z}_0 = \mathbf{z}(0) = [0.3226, 0.5376, 0.1398]$. *Another mixture of the three liquids,* $\mathbf{y}$, *is in container B. The content of B is poured and stirred in A. The final concentration in A is* $\mathbf{z}_1 = [0.0411, 0.2740, 0.6849]$. *One can ask for the composition* $\mathbf{y}$ *and for the required volume in container B. Using the notation introduced above, the initial volume in A is* $m_1 = 93$, *the volume and concentration in B are unknown. Equation (8.4) is now a system of three equations with three unknowns:* m_2, y_1, y_2 *(the closure condition implies* $y_3 = 1 - y_1 - y_2$*):*

$$m_1 \begin{pmatrix} z_1 - x_1 \\ z_2 - x_2 \\ z_3 - x_3 \end{pmatrix} = m_2 \begin{pmatrix} y_1 - z_1 \\ y_2 - z_2 \\ 1 - y_2 - y_3 - z_3 \end{pmatrix} \ , \tag{8.5}$$

which, being a simple system, is not linear in the unknowns. Note that (8.5) involves masses or volumes and, therefore, it is not a purely compositional equation. This situation always occurs in mixture processes. Figure 8.7 shows the process of mixing (M) both in a ternary diagram (left) and in the balance-coordinates $u_1 = 6^{-1/2}\ln(z_1z_2/z_3)$, $u_2 = 2^{-1/2}\ln(z_1/z_2)$ *(right). Fig. 8.7 also*

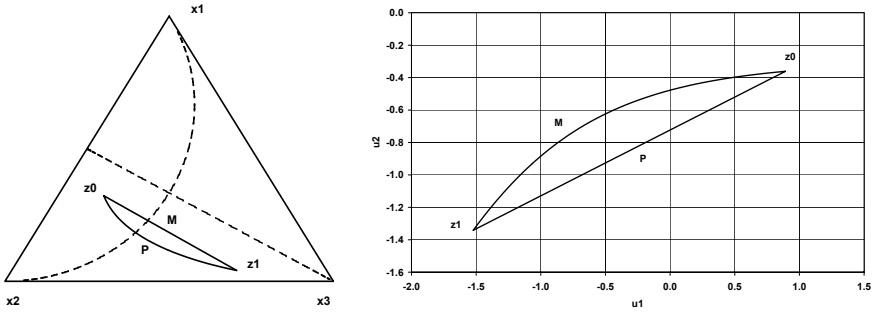

Figure 8.7: Two processes going from $\mathbf{z}_0$ to $\mathbf{z}_1$. (M) mixture process; (P) linear perturbation process. Representation in the ternary diagram, left; using balance-coordinates $u_1 = 6^{-1/2}\ln(z_1z_2/z_3)$, $u_2 = 2^{-1/2}\ln(z_1/z_2)$, right.

shows a perturbation-linear process, i.e. a straight-line in the simplex, going from $\mathbf{z}_0$ *to* $\mathbf{z}_1$ *(P).*

EXERCISE 8.3.1 *In the example obtaining a mixture find the necessary volume* m_2 *and the composition in container B,* $\mathbf{y}$*. Find the direction of the perturbation-linear process to go from* $\mathbf{z}_0$ *to* $\mathbf{z}_1$*.*

EXERCISE 8.3.2 *A container has a constant volume* $V = 100$ *volume units and initially contains a liquid whose composition is* $\mathbf{x}(0) = \mathcal{C}[1,1,1]$*. A constant flow of* $Q = 1$ *volume unit per second with volume composition* $\mathbf{x} = \mathcal{C}[80,2,18]$ *gets into the box. After a complete mixing there is an output whose flow equals* Q *with the volume composition* $\mathbf{x}(t)$ *at the time* t*. Model the evolution of the volumes of the three components in the container using ordinary linear differential equations and solve them (Hint: these equations are easily found in textbooks, e.g. Albarède (1995, p. 345–350)). Are you able to plot the curve for the output composition* $\mathbf{x}(t)$ *in the simplex without using the solution of the differential equations? Is it a mixture?*

Chapter 9

Linear compositional models

Linear models are intended to relate two sets of random variables using linear relationships. They are very general and appear routinely in many statistical applications. A first set of variables, called response variables, are to be predicted from a second set of variables, called predictors or covariates. Linear combinations of the predictors, transformed by some non-linear function, frequently called *link function*, are used to get a predictor function approaching responses. Errors or residuals are measured as Euclidean differences between responses and the predictor function. There is a extensive literature on general linear models, for instance Anderson (1984). The two sets of variables may have very different characteristics (categorical, real, discrete) and the link functions choices are also multiple. We are here interested in cases where responses or predictors have compositional character. When the response is compositional we must be aware that residuals should be measured with compositional distances, i.e. within the framework of Aitchison geometry of the simplex. Section 9.1 treats the case where the response is compositional and the covariates are real or discrete thus corresponding to a multiple regression. Response is handled using its coordinates. The ilr transformation plays the role of a link function. Section 9.2 assumes a single real response and a compositional predictor. Again ilr plays an important role. Section 9.3 discusses the case in which the response is compositional and the predictors reduce to a categorical variable indicating a treatment or a subpopulation. The goal of such an analysis of variance (ANOVA) model is to decide whether the center of compositions across the treatments are equal or not. Section 9.4 deals with discriminant analysis. In this case, response is a category to which a compositional observation (predictor) is assigned. The model is then a rule to assign an observed composition to categories or treatments and it provides a probability of belonging for each category.

9.1 Linear regression with compositional variables

Linear regression is intended to identify and estimate a linear model from response data that depend linearly on one or more covariates. The assumption is that responses are affected by errors or random deviations of the mean model. The most usual methods to fit the regression coefficients are the well-known least-squares techniques.

The problem of regression when the response is compositional is stated as follows. A compositional sample in $\mathcal{S}^D$ is available and it is denoted by $\mathbf{x}_1$, $\mathbf{x}_2, \ldots,$ $\mathbf{x}_n$. The i-th data-point $\mathbf{x}_i$ is associated with one or more external variables or covariates grouped in the vector $\mathbf{t}_i = [t_{i0}, t_{i1}, \ldots, t_{ir}]$, where $t_{i0} = 1$. The goal is to estimate the coefficients of a curve or surface in $\mathcal{S}^D$ whose equation is

$$\hat{\mathbf{x}}(\mathbf{t}) = \boldsymbol{\beta}_0 \oplus (t_1 \odot \boldsymbol{\beta}_1) \oplus \cdots \oplus (t_r \odot \boldsymbol{\beta}_r) = \bigoplus_{j=0}^{r} (t_j \odot \boldsymbol{\beta}_j) \ , \tag{9.1}$$

where $\mathbf{t} = [t_0, t_1, \ldots, t_r]$ are real covariates and are identified as the parameters of the curve or surface; the first parameter is defined as the constant $t_0 = 1$, as assumed for observations. The compositional coefficients of the model, $\boldsymbol{\beta}_j \in \mathcal{S}^D$, are to be estimated from the data. The model (9.1) is very general and takes different forms depending on how the covariates t_j are defined. For instance, defining $t_j = t^j$, being t a covariate, the model is a polynomial, particularly, if $r = 1$, it is a straight-line in the simplex (8.2).

The most popular fitting method of the model (9.1) is the least-squares deviation criterion. As the response $\mathbf{x}(\mathbf{t})$ is compositional, it is natural to measure deviations also in the simplex using the concepts of the Aitchison geometry. The deviation of the model (9.1) from the data is defined as $\hat{\mathbf{x}}(\mathbf{t}_i) \ominus \mathbf{x}_i$ and its size by the Aitchison norm $\|\hat{\mathbf{x}}(\mathbf{t}_i) \ominus \mathbf{x}_i\|_a^2 = \mathrm{d}_a^2(\hat{\mathbf{x}}(\mathbf{t}_i), \mathbf{x}_i)$. The target function (sum of squared errors, SSE) is

$$\mathrm{SSE} = \sum_{i=1}^{n} \|\hat{\mathbf{x}}(\mathbf{t}_i) \ominus \mathbf{x}_i\|_a^2 \ ,$$

to be minimised as a function of the compositional coefficients $\boldsymbol{\beta}_j$ which are implicit in $\hat{\mathbf{x}}(\mathbf{t}_i)$. The number of coefficients to be estimated in this linear model is $(r + 1) \cdot (D - 1)$.

This least-squares problem is reduced to $D - 1$ ordinary least-squares problems when the compositions are expressed in coordinates with respect to an orthonormal basis of the simplex. Assume that an orthonormal basis has been chosen in $\mathcal{S}^D$ and that the coordinates of $\hat{\mathbf{x}}(\mathbf{t})$, $\mathbf{x}_i$ and $\boldsymbol{\beta}_j$ are $\mathbf{x}_i^* = [x_{i1}^*, x_{i2}^*, \ldots, x_{i,D-1}^*]$, $\hat{\mathbf{x}}^*(\mathbf{t}) = [\hat{x}_1^*(\mathbf{t}), \hat{x}_2^*(\mathbf{t}), \ldots, \hat{x}_{D-1}^*(\mathbf{t})]$ and $\boldsymbol{\beta}_j^* = [\beta_{j1}^*, \beta_{j2}^*, \ldots, \beta_{j,D-1}^*]$, being these vectors in $\mathbb{R}^{D-1}$. Since perturbation and powering in the simplex are translated into the ordinary sum and product by scalars in the coordinate real space, the

model (9.1) is expressed in coordinates as

$$\hat{\mathbf{x}}^*(\mathbf{t}) = \boldsymbol{\beta}_0^* + \boldsymbol{\beta}_1^* \; t_1 + \cdots + \boldsymbol{\beta}_r^* \; t_r = \sum_{j=0}^{r} \boldsymbol{\beta}_j^* \; t_j \; .$$

For each coordinate, this expression becomes

$$\hat{x}_k^*(\mathbf{t}) = \beta_{0k}^* + \beta_{1k}^* \; t_1 + \cdots + \beta_{rk}^* \; t_r \; , \; k = 1, 2, \ldots, D-1 \; . \tag{9.2}$$

Also Aitchison norm and distance become the ordinary norm and distance in real space. Then, using coordinates, the target function is expressed as

$$\text{SSE} = \sum_{i=1}^{n} \|\hat{\mathbf{x}}^*(\mathbf{t}_i) - \mathbf{x}_i^*\|^2 = \sum_{k=1}^{D-1} \left\{ \sum_{i=1}^{n} |\hat{x}_k^*(\mathbf{t}_i) - x_{ik}^*|^2 \right\} \; , \tag{9.3}$$

where $\|\cdot\|$ is the norm of a real vector. The last right-hand member of (9.3) has been obtained permuting the order of the sums on the components of the vectors and on the data. All sums in (9.3) are non-negative and, therefore, the minimisation of SSE implies the minimisation of each term of the sum in k,

$$\text{SSE}_k = \sum_{i=1}^{n} |\hat{x}_k^*(\mathbf{t}_i) - x_{ik}^*|^2 \; , \; k = 1, 2, \ldots, D-1 \; .$$

This is, the fitting of the compositional model (9.1) reduces to the $D-1$ ordinary least-squares problems in (9.2).

Example: Vulnerability of a system.
A system is subjected to external actions. The response of the system to such actions is frequently a major concern in engineering. For instance, the system may be a dike under the action of ocean-wave storms; the response may be the level of service of the dike after one event. In a simplified scenario, three responses of the system may be considered: θ_1, service; θ_2, damage; θ_3 collapse. The dike can be designed for a design action, e.g. wave-height, d, ranging $3 \leq d \leq 20$ (metres wave-height). Actions, parameterised by some wave-height of the storm, h, also ranging $3 \leq d \leq 20$ (metres wave-height). Vulnerability of the system is described by the conditional probabilities

$$p_k(d,h) = \text{P}[\theta_k | d, h] \; , \; k = 1, 2, 3 = D \; , \; \sum_{k=1}^{D} p_k(d,h) = 1 \; ,$$

where, for any d, h, $\mathbf{p}(d,h) = [p_1(d,h), p_2(d,h), p_3(d,h)] \in \mathcal{S}^3$. In practice, $\mathbf{p}(d,h)$ only is approximately known for a limited number of values $\mathbf{p}(d_i, h_i)$, $i = 1, \ldots, n$. The whole model of vulnerability can be expressed as a regression model

$$\hat{\mathbf{p}}(d,h) = \boldsymbol{\beta}_0 \oplus (d \odot \boldsymbol{\beta}_1) \oplus (h \odot \boldsymbol{\beta}_2) \; , \tag{9.4}$$

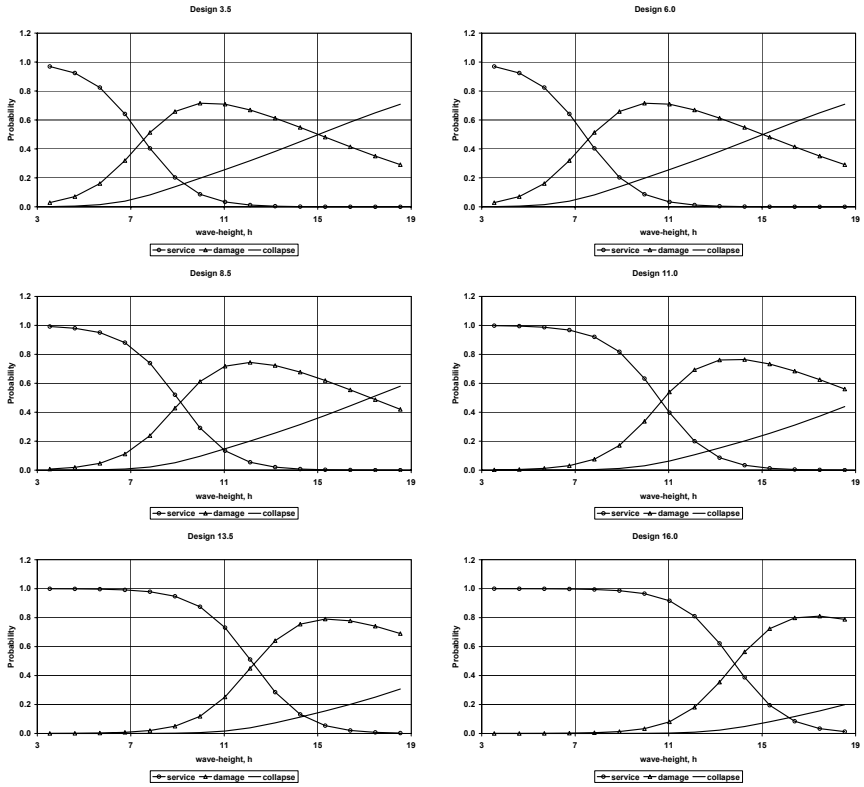

Figure 9.1: Vulnerability models obtained by regression in the simplex from the data in the Table 9.1. Horizontal axis: incident wave-height in m. Vertical axis: probability of the output response. Shown designs are 3.5, 6.0, 8.5, 11.0, 13.5, 16.0 (m design wave-height).

so that it can be estimated by regression in the simplex.

Consider the data in Table 9.1 containing $n = 9$ probabilities. Figure 9.1 shows the vulnerability probabilities obtained by regression for six design values. An inspection of these Figures reveals that a quite realistic model has been obtained from a really poor sample: service probabilities decrease as the level of action increases and conversely for collapse. This changes smoothly for increasing design level. Despite the challenging shapes of these curves describing the vulnerability, they come from a linear model as can be seen in Figure 9.2 (left). In Figure 9.2 (right) these straight-lines in the simplex are shown in a ternary diagram. In these cases, the regression model has shown its smoothing capabilities.

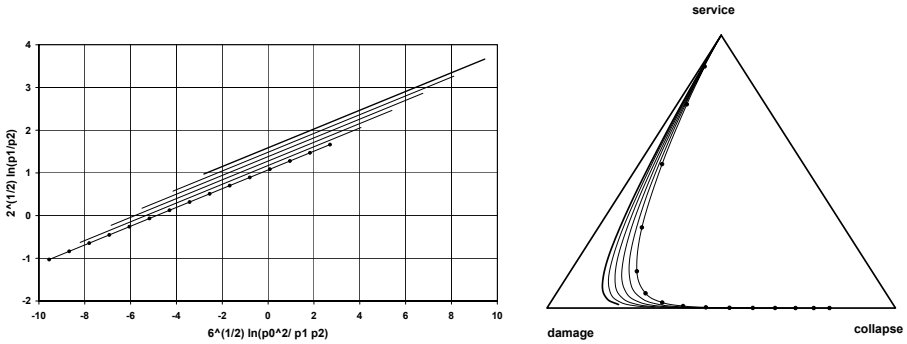

Figure 9.2: Vulnerability models in Figure 9.1 in coordinates (left) and in the ternary diagram (right). Design 3.5 (circles); 16.0 (thick line).

9.2 Regression with compositional covariates

The model with compositional covariates appears when the goal is to predict one external variable as a function of a composition. Assume a compositional data set $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ is available, and that the i-th data-point $\mathbf{x}_i$ is associated with an observation y_i of an external response variable (with support the *whole real line*, i.e. possibly transformed with logs or any other suitable transformation). The goal is to estimate a surface on $\mathcal{S}^D \times \mathbb{R}$ with equation

$$\hat{y}(\mathbf{x}) = \beta_0 + \langle \boldsymbol{\beta}, \mathbf{x} \rangle_a \tag{9.5}$$

where $\boldsymbol{\beta} \in \mathcal{S}^D$ is the (simplicial) gradient of y with respect to $\mathbf{x}$, and β_0 is a real intercept. In this case, since the response is a real value, the classical least squares fitting criterion may be applied, which yields the target function

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \beta_0 - \langle \boldsymbol{\beta}, \mathbf{x}_i \rangle_a)^2 \ .$$

As the Aitchison inner product can be computed easily from clr coefficients or ilr coordinates of $\mathbf{x}_i$, this sum of squares becomes

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \beta_0 - \langle \text{clr}(\boldsymbol{\beta}), \text{clr}(\mathbf{x}_i) \rangle)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \langle \text{ilr}(\boldsymbol{\beta}), \text{ilr}(\mathbf{x}_i) \rangle)^2.$$

This suggests that the actual fitting can be done using ilr coordinates without further ado. One simply fits a linear regression to the response y as a linear function of $\text{ilr}(\mathbf{x})$. The estimated version of $\text{ilr}(\boldsymbol{\beta})$ contains slope coefficients of the response y with respect the coordinates $\text{ilr}(\mathbf{x})$. The simplicial gradient $\boldsymbol{\beta}$ is then easily computed using ilr^{-1}. The clr transformation should be avoided in this case, as its correct handling in the scope of regression requires the generalized inversion of singular matrices (something that most statistical packages are not able to do).

Tests on the coefficients of the model (9.5) may be used as usual, for instance to obtain a simplified model depending on less variables. However, one should be aware that they are related to the particular basis used, and that these simplified models will be different depending on the basis. One should thus carefully select the basis: for instance, a basis of balances may be adequate to check the dependence of y on a particular subcomposition of $\mathbf{x}$.

EXERCISE 9.2.1 (SAND-SILT-CLAY FROM A LAKE) *Consider the data in Table 9.2. They are sand-silt-clay compositions from an Arctic lake taken at different depths (adapted from Coakley and Rust (1968) and cited in Aitchison (1986)). The goal is to check whether there is some trend in the composition related to the depth. Particularly, using the standard hypothesis testing in regression, check the constant and the straight-line models*

$$\hat{\mathbf{x}}(\mathbf{t}) = \boldsymbol{\beta}_0 \ , \ \hat{\mathbf{x}}(\mathbf{t}) = \boldsymbol{\beta}_0 \oplus (t \odot \boldsymbol{\beta}_1) \ ,$$

being t =depth. Plot both models, the fitted model and the residuals, in coordinates and in the ternary diagram.

EXERCISE 9.2.2 (SAND-SILT-CLAY FROM A LAKE: SECOND SIGHT) *One can equivalently check whether the sediment composition brings any information about the depth at which that sample was taken. Using the data from the previous excercise, fit a linear model to explain depth as a function of the composition. Analyse the residuals as usual, as they may be considered real values.*

To display the model, you can follow these steps. Split the range of observed depths in several segments of the same length (four to six will be enough in CoDaPack), and give each sample a number corresponding to its depth cathegory. Plot the compositional data in a ternary diagram, using colors for each depth interval. Draw a line on the simplex, from the center of the data set along the gradient of the fitted model

9.3 Analysis of variance with compositional response

ANalysis Of the VAriance (ANOVA) is the name given to a linear model where a continuous response is explained as a function of a (set of) discrete variable(s). Compositional ANOVA follows the same steps that were used to predict a composition from a continuous covariable. Notation in multi-way ANOVA (with more than one discrete covariable) can become quite difficult, therefore only one-way compositional ANOVA will be addressed here. Textbooks on multivariate ANOVA may be then useful to extend this material.

As in the preceding section, assume a compositional sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ in $\mathcal{S}^D$ is available. These observations are classified into K categories across an external categorical variable z. Category z may represent different treatments

or subpopulations. In other words, for each composition one has also available a category z_i. ANOVA deals with the centres (compositional means) of the composition for each category, $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$. Following the classical notation, a compositional ANOVA model, for a given z, is

$$\hat{\mathbf{x}} = \boldsymbol{\beta}_1 \oplus (I(z=2)) \odot \boldsymbol{\beta}_2) \oplus \cdots \oplus (I(z=K) \odot \boldsymbol{\beta}_K) \ , \mathbf{x} \ominus \hat{\mathbf{x}} = \boldsymbol{\epsilon},$$

where the indicator $I(z=k)$ equals 1 when the condition is true and 0 otherwise. This means that only one of the $K-1$ terms may be taken into account for each possible z value (as the other will be powered by zero). Note that the first category does not explicitly appear in the equation: when $z=1$, then the predictor is only $\boldsymbol{\beta}_1 = \boldsymbol{\mu}_1$ the centre of the first group. The remaining coefficients are then interpreted as increments of the composition from the reference first level to the category actually observed, $\hat{\mathbf{x}}|k = \boldsymbol{\mu}_k = \boldsymbol{\beta}_1 \oplus \boldsymbol{\beta}_k$, or equivalently $\boldsymbol{\beta}_k = \boldsymbol{\mu}_k \ominus \boldsymbol{\mu}_1$. The variable $\boldsymbol{\epsilon}$ is the compositional residual of the model.

ANOVA notation may be quite cumbersome, but its fitting is straightforward. For each observation z_i one just constructs a vector $\mathbf{t}_i$ with the $K-1$ indicators, i.e. with many zeroes and a single 1 in the position of the observed category. With these vectors, one applies the same steps as with linear regression with compositional response (section 9.1). An important consequence of this procedure is that we implicitly assume equal total variance within each category (compare with chapter 7).

The practice of compositional ANOVA follows the principle of working on coordinates. One first selects a basis and computes the corresponding coordinates of the observations. Then classical ANOVA is applied to explain each ilr coordinate as a mean reference level plus $K-1$ mean differences between categories. Tests may be applied to conclude that some of these differences are not significative, i.e. that the mean of a particular coordinate at two categories may be taken as equal. Finally, all coefficients associated with category k can be back-transformed to obtain its associated compositional coefficient $\boldsymbol{\beta}_k$.

EXERCISE 9.3.1 (DISTINGUISHING MEAN RIVERS HYDROGEOCHEMISTRY) *In the course accessory files you can find a file named "Hydrochem.csv", containing an extensive data set of the geochemical composition of water samples from several rivers and tributaries of the Llobregat river, the most important river in the Barcelona province (NE Spain). This data was studied in detail by Otero et al. (2005); Tolosana-Delgado et al. (2005), and placed, with the authors' consent, in the public domain within the R package "compositions" (van den Boogaart and Tolosana-Delgado, 2008). Table 9.3 provides a random sample, in case the whole data set is not accessible.*

Fit an ANOVA model to this 3-part composition. Draw the data set using colours to distinguish between the four rivers. Plot the means of the four groups, as estimated by the ANOVA fit.

Advanced excercise: *Extract the variance-covariance matrix of each mean. Draw confidence regions for them, as explained in section B.*

9.4 Linear discrimination with compositional predictor

A composition can also be used to predict a categorical variable. Following the preceding section notation, the goal is now to estimate $\mathbf{p}(\mathbf{x})$, the probability that z takes each of its possible K values given an observed composition $\mathbf{x}$. There are many techniques to obtain this result, like the Fisher rule, linear or quadratic discriminant analysis, or multinomial logistic regression. But essentially, we would always apply the principle of working on coordinates (take ilr's, apply your favourite method to the coordinates, and back-transform the coefficients if they may be interpreted as compositions). This section illustrates this procedure with linear discriminant analysis (LDA), a technique available in most basic statistical packages.

First, LDA assumes some prior probabilities p_k^0 of data-points corresponding to each one of the K categories. These are typically taken as $p_k = 1/K$ (equally probable) or $p_k = n_k/n$ (where n_k is the number of samples in kth-category). Then, it assumes that the ilr-transformed composition has a normal distribution, with ilr mean $\boldsymbol{\mu}_k^* = \text{ilr}(\boldsymbol{\mu}_k)$ and common ilr-coordinates covariance $\boldsymbol{\Sigma}$ (i.e., all categories have the same covariance and possibly different mean). Applying Bayes' Theorem, it is verified that the posterior probability vector of belonging to each class for a particular composition $\mathbf{x}$, can be derived from the discriminant functions

$$d_{jk}(\mathbf{x}) = \ln \frac{p_j}{p_k} = A_{ij} + (\boldsymbol{\mu}_j^* - \boldsymbol{\mu}_k^*)' \cdot \boldsymbol{\Sigma}^{-1} \cdot \mathbf{x}^*.$$

with

$$A_{ij} = \ln \frac{p_j^0}{p_k^0} - \frac{1}{2}(\boldsymbol{\mu}_j^* - \boldsymbol{\mu}_k^*)' \cdot \boldsymbol{\Sigma}^{-1} \cdot (\boldsymbol{\mu}_j^* - \boldsymbol{\mu}_k^*)$$

Again, as happened with ANOVA, one category is typically placed as a sort of reference level. For instance, take $j = K$ fix. Then LDA just computes the log-odds of the the other $K-1$ categories with respect to the last one. The desired probabilities can be obtained with the inverse alr transformation, as explained in section 4.6.

Obtained probabilities can be then used to decide which category is more probable for each possible composition $\mathbf{x}$: typically we classify each point in $\mathcal{S}^D$ into the most likely group, the one with largest probability. In this sense, the discriminant functions can be used to draw the boundaries between regions j and k, by identifying the set of points where $d_{jk}(\mathbf{x}) = 0$. Some linear algebra shows that this boundary is the affine hyperplane of $\mathcal{S}^D$ orthogonal to the vector

$\mathbf{v}_{dj}$ and passing through the point $\mathbf{x}^0_{jk}$ obtained as

$$\begin{aligned}\mathbf{v}_{jk} &= [\boldsymbol{\Sigma}^{-1} \cdot (\boldsymbol{\mu}^*_j - \boldsymbol{\mu}^*_k)] \\ \mathbf{x}^0_{jk} &= (\boldsymbol{\mu}^*_j + \boldsymbol{\mu}^*_k)/2 - \frac{A_{ij}}{2(\boldsymbol{\mu}^*_j - \boldsymbol{\mu}^*_k)' \cdot \mathbf{v}_{jk}} \cdot (\boldsymbol{\mu}^*_j - \boldsymbol{\mu}^*_k).\end{aligned}$$

Note that these equations are only useful to draw boundaries between neighbouring categories, i.e. between the two most probable categories of a given point in $\mathcal{S}^D$. For more than 2 categories, care should be taken to draw them by segments.

EXERCISE 9.4.1 (DRAWING BORDERS BETWEEN THREE GROUPS) *Between three groups we can draw three borders (A with B, B with C, A with C). Show that these three boundaries intersect in one single point (a* triple junction*). Find the equation of that point. Now assume that the discriminating composition has three components: note that in this case, the boundaries could be drawn as segments from the triple junction along the directions of some vectors* $\mathbf{v}^{\perp}_{jk}$ *orthogonal to* $\mathbf{v}_{jk}$*.*

EXERCISE 9.4.2 (THE HYDROCHEMICAL DATA SET REVISITED) *Using the data set from exercise 9.3.1, obtain the discriminant functions between rivers A, U and L (remove C for this exercise). This may be easily done by computing the ilr coordinates in CoDaPack, and exporting them to your favourite statistical software. Draw the data in the ilr plane and in a ternary diagram, using colours to distinguish between rivers. Add the group centers, and the boundaries between groups. If you use R, linear discriminant analysis is available with function "lda" in the package "MASS".*

Table 9.1: Assumed vulnerability for a dike with only three outputs or responses. Probability values of the response θ_k conditional to values of design d and level of the storm h.

d_i	h_i	service	damage	collapse
3.0	3.0	0.50	0.49	0.01
3.0	10.0	0.02	0.10	0.88
5.0	4.0	0.95	0.049	0.001
6.0	9.0	0.08	0.85	0.07
7.0	5.0	0.97	0.027	0.003
8.0	3.0	0.997	0.0028	0.0002
9.0	9.0	0.35	0.55	0.01
10.0	3.0	0.999	0.0009	0.0001
10.0	10.0	0.30	0.65	0.05

Table 9.2: Sand, silt, clay composition of sediment samples at different water depths in an Arctic lake.

sample no.	sand	silt	clay	depth (m)	sample no.	sand	silt	clay	depth (m)
1	77.5	19.5	3.0	10.4	21	9.5	53.5	37.0	47.1
2	71.9	24.9	3.2	11.7	22	17.1	48.0	34.9	48.4
3	50.7	36.1	13.2	12.8	23	10.5	55.4	34.1	49.4
4	52.2	40.9	6.9	13.0	24	4.8	54.7	40.5	49.5
5	70.0	26.5	3.5	15.7	25	2.6	45.2	52.2	59.2
6	66.5	32.2	1.3	16.3	26	11.4	52.7	35.9	60.1
7	43.1	55.3	1.6	18.0	27	6.7	46.9	46.4	61.7
8	53.4	36.8	9.8	18.7	28	6.9	49.7	43.4	62.4
9	15.5	54.4	30.1	20.7	29	4.0	44.9	51.1	69.3
10	31.7	41.5	26.8	22.1	30	7.4	51.6	41.0	73.6
11	65.7	27.8	6.5	22.4	31	4.8	49.5	45.7	74.4
12	70.4	29.0	0.6	24.4	32	4.5	48.5	47.0	78.5
13	17.4	53.6	29.0	25.8	33	6.6	52.1	41.3	82.9
14	10.6	69.8	19.6	32.5	34	6.7	47.3	46.0	87.7
15	38.2	43.1	18.7	33.6	35	7.4	45.6	47.0	88.1
16	10.8	52.7	36.5	36.8	36	6.0	48.9	45.1	90.4
17	18.4	50.7	30.9	37.8	37	6.3	53.8	39.9	90.6
18	4.6	47.4	48.0	36.9	38	2.5	48.0	49.5	97.7
19	15.6	50.4	34.0	42.2	39	2.0	47.8	50.2	103.7
20	31.9	45.1	23.0	47.0					

Table 9.3: Main anion composition of some water samples from 4 different rivers in Barcelona province (NE Spain).

river	Cl	SO4	HCO3	river	Cl	SO4	HCO3
A	197.43	857.99	348.39	U	16.54	71.88	182.20
A	312.37	487.83	377.13	U	27.29	93.35	197.97
A	15.49	239.93	146.00	U	26.00	96.81	176.96
A	118.09	445.63	341.50	U	29.15	76.87	188.60
A	352.84	341.68	557.50	U	37.14	94.72	179.60
A	309.78	371.71	538.50	U	22.86	84.46	244.80
A	432.24	357.35	393.70	U	33.29	116.76	180.10
L	142.80	120.34	210.30	U	9.57	42.96	197.31
L	305.74	199.97	222.45	U	7.79	25.75	171.29
L	309.67	164.40	206.32	U	6.07	36.85	174.20
L	325.76	151.63	201.90	U	108.14	96.16	180.45
L	256.18	145.33	189.20	U	24.79	109.86	209.70
L	242.42	196.08	187.10	C	15.22	83.35	177.40
L	373.26	166.62	249.70	C	265.84	116.69	188.70
L	382.45	222.31	219.96	C	385.13	118.58	191.70
L	228.30	181.83	368.40	C	634.93	164.80	232.56
L	14.02	55.52	245.90	C	519.88	397.32	220.10
L	445.39	455.62	286.67	C	844.45	154.68	175.10
L	300.05	469.89	287.40	C	10.22	83.98	180.44
L	1133.39	581.08	613.60	C	194.83	228.07	293.60
L	652.03	517.47	410.78				

Appendix A

Plotting a ternary diagram

Denote the three vertices of the ternary diagram counter-clockwise from the upper vertex as A, B and C (see Figure A.1). The scale of the plot is arbitrary

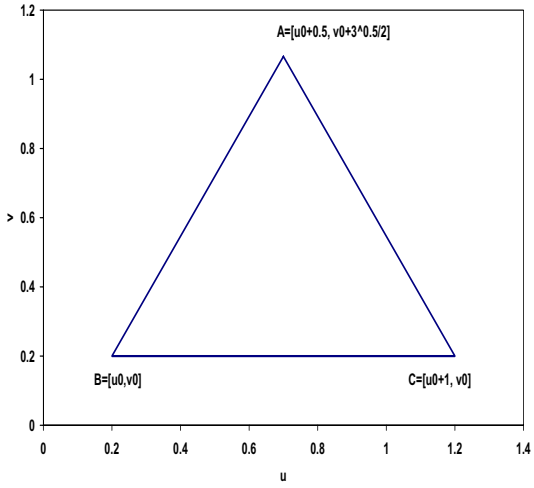

Figure A.1: Plot of the frame of a ternary diagram. The shift plotting coordinates are $[u_0, v_0] = [0.2, 0.2]$, and the length of the side is 1.

and a unitary equilateral triangle can be chosen. Assume that $[u_0, v_0]$ are the plotting coordinates of the B vertex. The C vertex is then $C = [u_0 + 1, v_0]$; and the vertex A has abscissa $u_0 + 0.5$ and the square-height is obtained using Pythagorean Theorem: $1^2 - 0.5^2 = 3/4$. Then, the vertex $A = [u_0 + 0.5, v_0 + \sqrt{3}/2]$. These are the vertices of the triangle shown in Figure A.1, where the origin has been shifted to $[u_0, v_0]$ in order to centre the plot. The figure is obtained plotting the segments AB, BC, CA.

To plot a sample point $\mathbf{x} = [x_1, x_2, x_3]$, closed to a constant κ, the corresponding plotting coordinates $[u, v]$ are needed. They are obtained as a convex

linear combination of the plotting coordinates of the vertices

$$[u, v] = \frac{1}{\kappa}(x_1 A + x_2 B + x_3 C) \ ,$$

with

$$A = [u_0 + 0.5, v_0 + \sqrt{3}/2] \ , \ B = [u_0, v_0] \ , \ C = [u_0 + 1, v_0] \ .$$

Note that the coefficients of the convex linear combination must be closed to 1 as obtained dividing by κ. Deformed ternary diagrams can be obtained just changing the plotting coordinates of the vertices and maintaining the convex linear combination.

Appendix B

Parametrisation of an elliptic region

To plot an ellipse in $\mathbb{R}^2$, and to plot its backtransform in the ternary diagram, we need to give to the plotting program a sequence of points that it can join by a smooth curve. This requires the points to be in a certain order, so that they can be joint consecutively. The way to do this is to use polar coordinates, as they allow to give a consecutive sequence of angles which will follow the border of the ellipse in one direction. The degree of approximation of the ellipse will depend on the number of points used for discretisation.

The algorithm is based on the following reasoning. Imagine an ellipse located in $\mathbb{R}^2$ with principal axes not parallel to the axes of the Cartesian coordinate system. What we have to do to express it in polar coordinates is (a) translate the ellipse to the origin; (b) rotate it in such a way that the principal axis of the ellipse coincide with the axis of the coordinate system; (c) stretch the axis corresponding to the shorter principal axis in such a way that the ellipse becomes a circle in the new coordinate system; (d) transform the coordinates into polar coordinates using the simple expressions $x^* = r\cos\rho$, $y^* = r\sin\rho$; (e) undo all the previous steps in inverse order to obtain the expression of the original equation in terms of the polar coordinates. Although this might sound tedious and complicated, in fact we have results from matrix theory which tell us that this procedure can be reduced to a problem of eigenvalues and eigenvectors.

In fact, any symmetric matrix can be decomposed into the matrix product $Q\Lambda Q'$, where Λ is the diagonal matrix of eigenvalues and Q is the matrix of orthonormal eigenvectors associated with them. For Q we have that $Q' = Q^{-1}$ and therefore $(Q')^{-1} = Q$. This can be applied to either the first or the second options of the last section.

In general, we are interested in ellipses whose matrix is related to the sample covariance matrix $\hat{\Sigma}$, particularly its inverse. We have $\hat{\Sigma}^{-1} = Q\Lambda^{-1}Q'$ and substituting into the equation of the ellipse (7.1), (7.2):

$$(\bar{\mathbf{x}}^* - \boldsymbol{\mu})Q\Lambda^{-1}Q'(\bar{\mathbf{x}}^* - \boldsymbol{\mu})' = (Q'(\bar{\mathbf{x}}^* - \boldsymbol{\mu})')'\Lambda^{-1}(Q'(\bar{\mathbf{x}}^* - \boldsymbol{\mu})') = \kappa \ ,$$

where $\bar{\mathbf{x}}^*$ is the estimated centre or mean and $\boldsymbol{\mu}$ describes the ellipse. The vector $Q'(\bar{\mathbf{x}}^* - \boldsymbol{\mu})'$ corresponds to a rotation in real space in such a way, that the new coordinate axis are precisely the eigenvectors. Given that Λ is a diagonal matrix, the next step consists in writing $\Lambda^{-1} = \Lambda^{-1/2}\Lambda^{-1/2}$, and we get:

$$\begin{aligned}&(Q'(\bar{\mathbf{x}}^* - \boldsymbol{\mu})')'\Lambda^{-1/2}\Lambda^{-1/2}(Q'(\bar{\mathbf{x}}^* - \boldsymbol{\mu})')\\ &= (\Lambda^{-1/2}Q'(\bar{\mathbf{x}}^* - \boldsymbol{\mu})')'(\Lambda^{-1/2}Q'(\bar{\mathbf{x}}^* - \boldsymbol{\mu})') = \kappa.\end{aligned}$$

This transformation is equivalent to a re-scaling of the basis vectors in such a way, that the ellipse becomes a circle of radius $\sqrt{\kappa}$, which is easy to express in polar coordinates:

$$\Lambda^{-1/2}Q'(\bar{\mathbf{x}}^* - \boldsymbol{\mu})' = \begin{pmatrix}\sqrt{\kappa}\cos\theta\\ \sqrt{\kappa}\sin\theta\end{pmatrix}, \quad \text{or} \quad (\bar{\mathbf{x}}^* - \mu)' = Q\Lambda^{1/2}\begin{pmatrix}\sqrt{\kappa}\cos\theta\\ \sqrt{\kappa}\sin\theta\end{pmatrix}.$$

The parametrisation that we are looking for is thus given by:

$$\boldsymbol{\mu}' = (\bar{\mathbf{x}}^*)' - Q\Lambda^{1/2}\begin{pmatrix}\sqrt{\kappa}\cos\theta\\ \sqrt{\kappa}\sin\theta\end{pmatrix}.$$

Note that $Q\Lambda^{1/2}$ is the upper triangular matrix of the Cholesky decomposition of $\hat{\Sigma}$:

$$\hat{\Sigma} = Q\Lambda^{1/2}\Lambda^{1/2}Q' = (Q\Lambda^{1/2})(\Lambda^{1/2}Q') = UL;$$

thus, from $\hat{\Sigma} = UL$ and $L = U'$ we get the condition:

$$\begin{pmatrix}u_{11} & u_{12}\\ 0 & u_{22}\end{pmatrix}\begin{pmatrix}u_{11} & 0\\ u_{12} & u_{22}\end{pmatrix} = \begin{pmatrix}\hat{\Sigma}_{11} & \hat{\Sigma}_{12}\\ \hat{\Sigma}_{12} & \hat{\Sigma}_{22}\end{pmatrix},$$

which implies

$$\begin{aligned}u_{22} &= \sqrt{\hat{\Sigma}_{22}},\\ u_{12} &= \frac{\hat{\Sigma}_{12}}{\sqrt{\hat{\Sigma}_{22}}},\\ u_{11} &= \sqrt{\frac{\hat{\Sigma}_{11}\hat{\Sigma}_{22} - \hat{\Sigma}_{12}^2}{\hat{\Sigma}_{22}}} = \sqrt{\frac{|\hat{\Sigma}|}{\hat{\Sigma}_{22}}},\end{aligned}$$

and for each component of the vector $\boldsymbol{\mu}$ we obtain:

$$\begin{aligned}\mu_1 &= \bar{x}_1^* - \sqrt{\frac{|\hat{\Sigma}|}{\hat{\Sigma}_{22}}}\sqrt{\kappa}\cos\theta - \frac{\hat{\Sigma}_{12}}{\sqrt{\hat{\Sigma}_{22}}}\sqrt{\kappa}\sin\theta\\ \mu_2 &= \bar{x}_2^* - \sqrt{\hat{\Sigma}_{22}}\sqrt{\kappa}\sin\theta.\end{aligned}$$

The points describing the ellipse in the simplex are $\text{ilr}^{-1}(\boldsymbol{\mu})$ (see Section 4.4).

The procedures described apply to the three cases studied in section 7.2, just using the appropriate covariance matrix $\hat{\Sigma}$. Finally, recall that κ will be obtained from a chi-square distribution.

Bibliography

Aitchison, J. (1981). A new approach to null correlations of proportions. *Mathematical Geology 13*(2), 175–189.

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. (1984). The statistical analysis of geochemical compositions. *Mathematical Geology 16*(6), 531–564.

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

Aitchison, J. (1990). Relative variation diagrams for describing patterns of compositional variability. *Mathematical Geology 22*(4), 487–511.

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 third annual conference of the International Association for Mathematical Geology*, Volume I, II and addendum, pp. 3–35. International Center for Numerical Methods in Engineering (CIMNE), Barcelona (E), 1100 p.

Aitchison, J. (2002). Simplicial inference. In M. A. G. Viana and D. S. P. Richards (Eds.), *Algebraic Methods in Statistics and Probability*, Volume 287 of *Contemporary Mathematics Series*, pp. 1–22. American Mathematical Society, Providence, Rhode Island (USA), 340 p.

Aitchison, J., C. Barceló-Vidal, J. J. Egozcue, and V. Pawlowsky-Glahn (2002). A concise guide for the algebraic-geometric structure of the simplex, the sample space for compositional data analysis. In U. Bayer, H. Burger, and W. Skala (Eds.), *Proceedings of IAMG'02 — The eigth annual conference of*

the International Association for Mathematical Geology, Volume I and II, pp. 387–392. Selbstverlag der Alfred-Wegener-Stiftung, Berlin, 1106 p.

Aitchison, J., C. Barceló-Vidal, J. A. Martín-Fernández, and V. Pawlowsky-Glahn (2000). Logratio analysis and compositional distance. *Mathematical Geology 32*(3), 271–275.

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

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

Aitchison, J. and J. W. Kay (2003). Possible solution of some essential zero problems in compositional data analysis. See Thió-Henestrosa and Martín-Fernández (2003).

Albarède, F. (1995). *Introduction to geochemical modeling*. Cambridge University Press (UK). 543 p.

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis*. Second ed., John Wiley and Sons, New York, USA.

Bacon-Shone, J. (2003). Modelling structural zeros in compositional data. See Thió-Henestrosa and Martín-Fernández (2003).

Barceló, C., V. Pawlowsky, and E. Grunsky (1994). Outliers in compositional data: a first approach. In C. J. Chung (Ed.), *Papers and extended abstracts of IAMG'94 — The First Annual Conference of the International Association for Mathematical Geology*, Mont Tremblant, Québec, Canadá, pp. 21–26. IAMG.

Barceló, C., V. Pawlowsky, and E. Grunsky (1996). Some aspects of transformations of compositional data and the identification of outliers. *Mathematical Geology 28*(4), 501–518.

Barceló-Vidal, C., J. A. Martín-Fernández, and V. Pawlowsky-Glahn (2001). Mathematical foundations of compositional data analysis. In G. Ross (Ed.), *Proceedings of IAMG'01 — The sixth annual conference of the International Association for Mathematical Geology*, pp. 20 p. CD-ROM.

Billheimer, D., P. Guttorp, and W. Fagan (1997). Statistical analysis and interpretation of discrete compositional data. Technical report, NRCSE technical report 11, University of Washington, Seattle (USA), 48 p.

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

Box, G. E. P. and D. R. Cox (1964). The analysis of transformations. *Journal of the Royal Statistical Society, Series B (Statistical Methodology) 26*(2), 211–252.

Buccianti, A. and V. Pawlowsky-Glahn (2005). New perspectives on water chemistry and compositional data analysis. *Mathematical Geology 37*(7), 703–727.

Buccianti, A., V. Pawlowsky-Glahn, C. Barceló-Vidal, and E. Jarauta-Bragulat (1999). Visualization and modeling of natural trends in ternary diagrams: a geochemical case study. See Lippard et al. (1999), pp. 139–144.

Chayes, F. (1960). On correlation between variables of constant sum. *Journal of Geophysical Research 65*(12), 4185–4193.

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

Coakley, J. P. and B. R. Rust (1968). Sedimentation in an Arctic lake. *Journal of Sedimentary Petrology 38*, 1290–1300.

Eaton, M. L. (1983). *Multivariate Statistics. A Vector Space Approach*. John Wiley & Sons.

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

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

Egozcue, J. J. and V. Pawlowsky-Glahn (2006). Exploring compositional data with the coda-dendrogram. In E. Pirard (Ed.), *Proceedings of IAMG'06 — The XIth annual conference of the International Association for Mathematical Geology*.

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

Fahrmeir, L. and A. Hamerle (Eds.) (1984). *Multivariate Statistische Verfahren*. Walter de Gruyter, Berlin (D), 796 p.

Fry, J. M., T. R. L. Fry, and K. R. McLaren (1996). Compositional data analysis and zeros in micro data. Centre of Policy Studies (COPS), General Paper No. G-120, Monash University.

Gabriel, K. R. (1971). The biplot — graphic display of matrices with application to principal component analysis. *Biometrika 58*(3), 453–467.

Galton, F. (1879). The geometric mean, in vital and social statistics. *Proceedings of the Royal Society of London 29*, 365–366.

Krzanowski, W. J. (1988). *Principles of Multivariate Analysis: A user's perspective*, Volume 3 of *Oxford Statistical Science Series*. Clarendon Press, Oxford (UK). 563 p.

Krzanowski, W. J. and F. H. C. Marriott (1994). *Multivariate Analysis, Part 2 - Classification, covariance structures and repeated measurements*, Volume 2 of *Kendall's Library of Statistics*. Edward Arnold, London (UK). 280 p.

Lippard, S. J., A. Næss, and R. Sinding-Larsen (Eds.) (1999). *Proceedings of IAMG'99 — The fifth annual conference of the International Association for Mathematical Geology*, Volume I and II. Tapir, Trondheim (N), 784 p.

Mardia, K. V., J. T. Kent, and J. M. Bibby (1979). *Multivariate Analysis*. Academic Press, London (GB). 518 p.

Martín-Fernández, J., C. Barceló-Vidal, and V. Pawlowsky-Glahn (1998). A critical approach to non-parametric classification of compositional data. In A. Rizzi, M. Vichi, and H.-H. Bock (Eds.), *Advances in Data Science and Classification (Proceedings of the 6th Conference of the International Federation of Classification Societies (IFCS'98), Università "La Sapienza", Rome, 21–24 July*, pp. 49–56. Springer-Verlag, Berlin (D), 677 p.

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

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (2000). Zero replacement in compositional data sets. In H. Kiers, J. Rasson, P. Groenen, and M. Shader (Eds.), *Studies in Classification, Data Analysis, and Knowledge Organization (Proceedings of the 7th Conference of the International Federation of Classification Societies (IFCS'2000), University of Namur, Namur, 11-14 July*, pp. 155–160. Springer-Verlag, Berlin (D), 428 p.

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology 35*(3), 253–278.

Martín-Fernández, J. A., M. Bren, C. Barceló-Vidal, and V. Pawlowsky-Glahn (1999). A measure of difference for compositional data based on measures of divergence. See Lippard et al. (1999), pp. 211–216.

Mateu-Figueras, G. (2003). *Models de distribució sobre el símplex*. Ph. D. thesis, Universitat Politècnica de Catalunya, Barcelona, Spain.

McAlister, D. (1879). The law of the geometric mean. *Proceedings of the Royal Society of London 29*, 367–376.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution and correlations among proportions. *Biometrika 49*(1-2), 65–82.

Otero, N., R. Tolosana-Delgado, and A. Soler (2003). A factor analysis of hidrochemical composition of Llobregat river basin. See Thió-Henestrosa and Martín-Fernández (2003).

Otero, N., R. Tolosana-Delgado, A. Soler, V. Pawlowsky-Glahn, and A. Canals (2005). Relative vs. absolute statistical analysis of compositions: a comparative study of surface waters of a mediterranean river. *Water Research Vol 39*(7), 1404–1414.

Pawlowsky-Glahn, V. (2003). Statistical modelling on coordinates. See Thió-Henestrosa and Martín-Fernández (2003).

Pawlowsky-Glahn, V. and A. Buccianti (2002). Visualization and modeling of subpopulations of compositional data: statistical methods illustrated by means of geochemical data from fumarolic fluids. *International Journal of Earth Sciences (Geologische Rundschau) 91*(2), 357–368.

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

Pawlowsky-Glahn, V. and J. J. Egozcue (2002). BLU estimators and compositional data. *Mathematical Geology 34*(3), 259–274.

Pawlowsky-Glahn, V. and J. J. Egozcue (2006). Anlisis de datos composicionales con el coda-dendrograma. In J. Sicilia-Rodríguez, C. González-Martín, M. A. González-Sierra, and D. Alcaide (Eds.), *Actas del XXIX Congreso de la Sociedad de Estadística e Investigación Operativa (SEIO'06)*, pp. 39–40. Sociedad de Estadística e Investigación Operativa, Tenerife (ES), CD-ROM.

Peña, D. (2002). *Análisis de datos multivariantes*. McGraw Hill. 539 p.

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.

Richter, D. H. and J. G. Moore (1966). Petrology of the Kilauea Iki lava lake, Hawaii. U.S. Geol. Surv. Prof. Paper 537-B, B1-B26, cited in Rollinson (1995).

Rollinson, H. R. (1995). *Using geochemical data: Evaluation, presentation, interpretation*. Longman Geochemistry Series, Longman Group Ltd., Essex (UK). 352 p.

Sarmanov, O. V. and A. B. Vistelius (1959). On the correlation of percentage values. *Doklady of the Academy of Sciences of the USSR – Earth Sciences Section 126*, 22–25.

Solano-Acosta, W. and P. K. Dutta (2005). Unexpected trend in the compositional maturity of second-cycle sand. *Sedimentary Geology 178*(3-4), 275–283.

Thió-Henestrosa, S. and J. A. Martín-Fernández (Eds.) (2003). *Compositional Data Analysis Workshop – CoDaWork'03, Proceedings*. Universitat de Girona, ISBN 84-8458-111-X, http://ima.udg.es/Activitats/CoDaWork03/.

Thió-Henestrosa, S. and J. A. Martín-Fernández (2005). Dealing with compositional data: the freeware codapack. *Mathematical Geology 37*(7), 773–793.

Thió-Henestrosa, S., R. Tolosana-Delgado, and O. Gómez (2005). New features of codapack—a compositional data package. Volume 2, pp. 1171–1178.

Tolosana-Delgado, R., N. Otero, V. Pawlowsky-Glahn, and A. Soler (2005). Latent compositional factors in the Llobregat river basin (Spain) hydrogeoeochemistry. *Mathematical Geology 37*(7), 681–702.

van den Boogaart, G. and R. Tolosana-Delgado (2005). A compositional data analysis package for R providing multiple approaches. In G. Mateu-Figueras and C. Barceló-Vidal (Eds.), *Compositional Data Analysis Workshop – CoDaWork'05, Proceedings*. Universitat de Girona, ISBN 84-8458-222-1, http://ima.udg.es/Activitats/CoDaWork05/.

van den Boogaart, K. G. and R. Tolosana-Delgado (2008). "compositions": a unified R package to analyze compositional data. *Computers & Geosciences 34*(4), 320–338.

von Eynatten, H., V. Pawlowsky-Glahn, and J. J. Egozcue (2002). Understanding perturbation on the simplex: a simple method to better visualise and interpret compositional data in ternary diagrams. *Mathematical Geology 34*(3), 249–257.