

TEXTBOOK

Compositional Analysis of Data
with *CoDaPack*
Online Course

CoDa-Research Group

University of Girona
Spain

2021 (updated)

Course presentation

This material draws together the collective experience of the research group on compositional data ([CoDa-Research Group](#)) from the many semesters of our course in Compositional Analysis of Data (CoDa-course) at the University of Girona ([UdG](#)). The CoDa-course provides an introduction to the theoretical and practical aspects of the compositional analysis of data, as well as an informal discussion forum on more advanced modeling topics. This material comprises lectures and exercises. In the lectures, the theoretical aspects are not presented in detail. For a detailed study of these aspects, the reader is referred to the list of references. Each chapter finishes with a specific bibliography list that completes the general bibliography index provided at the end of this presentation. All of these references were used to partially elaborate most of the contents of this material. On the other hand, the exercises included in this material have been developed in detail, with close attention being paid to all of the aspects related to the practical application of the compositional methodology.

Rather than “Analysis of Compositional Data”, this book purposely has the title “Compositional Analysis of Data”. As we will see in some of the examples, what makes the data compositional is sometimes the researcher’s objectives and analysis interests. Once the researcher decides that her/his analysis is compositional then the data are considered compositional data (CoDa).

The first part of these materials is devoted to presenting the theoretical basis of the compositional analysis of data. Next, we introduce the features of statistical techniques used to explore and model CoDa, such as PCA or the CoDa-dendrogram. After the presentation of the techniques employed to deal with irregular data (e.g. zeros), some common multivariate techniques are described when they are used to analyse CoDa sets. The presentation of theoretical aspects is interspersed with problem-solving exercises aimed at facilitating the understanding of the most important concepts. Students enrolled in the course must submit continuous evaluation activities. They are encouraged to contact the tutor who has been assigned to solve their doubts.

The course includes two itineraries: *basic* and *advanced*. The concepts, examples and exercises of the *advanced* part will be highlighted in blue. All students must do the *basic* part. The *advanced* part is voluntary. The student, depending on her/his interests can choose to make the *advanced* parts that she/he wishes.

Statisticians and applied scientists in any field, in particular, engineers, geologists, economists, bioenvironmental scientists and environmental engineers, working in academic or industrial institutions, are strongly encouraged to take this course. It is recommended that students enrolling will have taken some first year courses on statistics, algebra and calculus. Basic knowledge of multivariate statistics may also be helpful, however, basic knowledge about the statistical package **R** and the software **CoDaPack** is not required.

We hope that these materials will help any scientist to master the basic techniques to undertake statistical analysis of CoDa.

General references

- [Ait86] J. Aitchison, *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], 1986, 416 p.
- [FHT18] P. Filzmoser, K. Hron and T. Templ, *Applied Compositional Data Analysis. With Worked Examples in R*, Springer International Publishing, 2018, 280 p.
- [Fil+21] P. Filzmoser, K. Hron, J.A. Martín-Fernández and J. Palarea-Albaladejo, *Advances in Compositional Data Analysis: Festschrift in Honour of Vera Pawlowsky-Glahn*, Springer International Publishing, 2021, 404 p.
- [PET15] V. Pawlowsky-Glahn, J.J. Egozcue and R. Tolosona-Delgado, *Modeling and Analysis of Compositional Data*, Wiley, 2015, 272 p.
- [BT13] G.K. Van den Boogaart and R. Tolosona-Delgado, *Analyzing compositional data with R*, Springer-Verlag Berlin Heidelberg, 2013, (122) 258 p.

Contents

Chapter 1: The geometric structure of the sample space

- 1.1 The sample space of compositional data
- 1.2 Principles of CoDa analysis
- 1.3 Perturbation and power operations in the simplex
- 1.4 Original logratio analysis: alr and clr transformations
- 1.5 The algebraic-geometric structure of the simplex
- 1.6 Compositional-linear dependence, basis and coordinates
- 1.7 Scale invariant logratios or logcontrasts
- 1.8 Representation of compositions by orthonormal coordinates
- 1.9 olr-basis associated to a sequential binary partition

Chapter 2: Exploratory analysis and distributions on the simplex

- 2.1 Centre of a compositional data set
- 2.2 Covariance structure of a compositional data set
- 2.3 CoDa-dendrogram
- 2.4 Reduced-dimensionality representation of a compositional data set: clr-biplot
- 2.5 Principal balances
- 2.6 Distributions on the simplex

Chapter 3: Data pre-processing: irregular data

- 3.1 Missing data
- 3.2 Essential zeros
- 3.3 Count zeros
- 3.4 Censored data: rounded zeros
- 3.5 Dealing with missing values and zeros
- 3.6 Potential outliers

Chapter 4: Linear regression models

- 4.1 LRM for a compositional response and scalar predictor
- 4.2 LRM for a scalar response and compositional predictor
- 4.3 LRM extensions

Chapter 5: On the analysis of grouped data

- 5.1 Cluster analysis
- 5.2 Discriminant analysis
- 5.3 MANOVA

Appendix: Some typical compositional problems

- A.1 Chemical compositions of Roman-British pottery
- A.2 Arctic lake sediments at different depths
- A.3 Household budget patterns
- A.4 Milk composition study
- A.5 A statistician's time budget
- A.6 The MN blood system
- A.7 Mammal's milk
- A.8 Calc-alkaline and tholeiitic volcanic rocks
- A.9 Concentration of minor elements in carbon ashes
- A.10 Paleocological compositions
- A.11 Pollen composition in fossils
- A.12 Food consumption in European countries
- A.13 Household expenditures
- A.14 Serum proteins
- A.15 Physical activity and body mass index
- A.16 Hotel posts in social media
- A.17 The waste composition in Catalonia
- A.18 Employment distribution in EUROSTAT countries

The geometric structure of the sample space

Contents

- 1.1 The sample space of compositional data
 - 1.1.1 The simplex \mathcal{S}^D as sample space
 - 1.1.2 The special difficulties of CoDa analysis
- 1.2 Principles of CoDa analysis
 - 1.2.1 The principle of scale invariance
 - 1.2.2 Subcompositional coherence
- 1.3 Perturbation and power operations in the simplex
 - 1.3.1 The role of group operations in statistics
 - 1.3.2 Perturbation (\oplus): a fundamental group operation in the simplex
 - 1.3.3 Power (\odot) as a subsidiary operation in the simplex
- 1.4 Original logratio analysis: alr and clr transformations
 - 1.4.1 The additive logratio transformation (alr)
 - 1.4.2 The centred logratio transformation (clr)
 - 1.4.3 Philosophy of the logratio analysis
- 1.5 The algebraic-geometric structure of the simplex
 - 1.5.1 The simplex $(\mathcal{S}^D, \oplus, \odot)$ real vector space
 - 1.5.2 The simplex \mathcal{S}^D Euclidean space: distance, inner product and norm
- 1.6 Compositional-linear dependence, basis and coordinates
- 1.7 Scale invariant logratios or logcontrasts
- 1.8 Representation of compositions by orthonormal coordinates
 - 1.8.1 Orthonormal logratio coordinates (olr)
 - 1.8.2 The isometric logratio transformation (ilr)
- 1.9 olr-basis associated to a sequential binary partition
 - 1.9.1 Sequential binary partition (SBP)
 - 1.9.2 Balances

Objectives

- ✓ To show the nature of compositional data along with the inconsistency and difficulties involved in applying standard statistical analysis to this type of data.
- ✓ To present the simplex as the *natural* sample space of compositional data.
- ✓ To introduce the principles on which the statistical analysis of compositional data should be based according to their nature.
- ✓ To define the two basic operations on the simplex —perturbation and powering— on which the statistical analysis of compositional data is based.
- ✓ To learn how to structure the simplex \mathcal{S}^D in a Euclidean space of dimension $D - 1$.
- ✓ To introduce the concept of *logcontrast* on the simplex \mathcal{S}^D with special emphasis on the additive and the centred logratio transformations.
- ✓ To show the procedure for calculating the coordinates of a composition with respect to an orthonormal basis of \mathcal{S}^D introducing isometric logratio transformations.
- ✓ To show a procedure for selecting a suitable orthonormal basis that allows the coordinates of a composition to be easily interpreted.

1.1. The sample space of compositional data

1.1.1. The simplex \mathcal{S}^D as sample space[†]. The concept of compositional data (CoDa) is the starting point for the development of all the geometric and algebraic results that are necessary for building up reliable probabilistic and statistical models for such data.

Following on from earlier developments in CoDa [Ait86], a compositional vector of D parts, $\mathbf{x} = [x_1, \dots, x_D]$, is defined as a vector in which the only relevant information is contained in the ratios between its components. All components of the vector are assumed positive. Throughout the text, components are called *parts* and a compositional vector is called a *composition*. The notation of the vector with square brackets means that this vector is considered to be a row vector.

The assertion that all the relevant information is contained in the ratios implies that, if a is a real positive number, the vectors $[x_1, \dots, x_D]$ and $[ax_1, \dots, ax_D]$ essentially convey the same information and are thus indistinguishable. Therefore, a composition is a class of equivalent compositional vectors [BMP03, BM16]. That is, proportional vectors with positive parts are *compositionally equivalent*.

A way to simplify the use of compositions is to represent them in closed form, that is, as positive vectors, the parts of which add up to a positive constant, κ . Common values of the closure constant κ are 1 for parts per unit, 100 for percentages, or 10^6 for parts per million. A consequence of this is that a composition of D

[†]This section is an adaptation of [EP06, p. 145-147]. Further information in [Ait86, Sections 2.1-2.6, p. 24-38] and [PET15, Section 2.1, p. 8-12].

parts, $[x_1, \dots, x_D]$, can be identified with a closed vector

$$(1.1) \quad \mathbf{x} = \mathcal{C}[x_1, \dots, x_D] = \left[\frac{x_1 \cdot \kappa}{\sum_{i=1}^D x_i}, \dots, \frac{x_D \cdot \kappa}{\sum_{i=1}^D x_i} \right],$$

where \mathcal{C} is called the *closure operation* to the constant κ [Ait86]. For this reason, historically, CoDa are defined as closed data.

Example 1. The daily time devoted by an academic statistician (see *statisticiantimebudget* CoDa set description in the [Appendix](#)) to teaching (T); consultation (C); administration (A); research (R); other wakeful activities (O); and sleep can be expressed equivalently in hours or minutes. More formally, the vector $\mathbf{w} = [3.50, 2.25, 4.25, 2.50, 6.25, 5.25]$ (hours) is compositionally equivalent to vector $\mathbf{y} = [210, 135, 255, 150, 375, 315]$ (minutes). We can close to $\kappa = 100$ the vector \mathbf{w} :

$$\begin{aligned} \mathbf{x} &= \mathcal{C}[3.50, 2.25, 4.25, 2.50, 6.25, 5.25] \\ &= \left[\frac{3.50 \cdot 100}{24}, \frac{2.25 \cdot 100}{24}, \frac{4.25 \cdot 100}{24}, \frac{2.50 \cdot 100}{24}, \frac{6.25 \cdot 100}{24}, \frac{5.25 \cdot 100}{24} \right] \\ &= [14.6, 9.4, 17.7, 10.4, 26.0, 21.9]. \end{aligned}$$

The components of the new vector \mathbf{x} give the percentages of daily time spent in the six activities. Formally, the vectors \mathbf{w} , \mathbf{y} and \mathbf{x} are closed vectors to different constant ($\kappa = 24, 1440$ or 100 respectively). They are compositionally equivalent, that is, they are different representations of the same composition.

Example 2. The expenditures on four commodity groups (housing, foodstuff, services and others) of forty single persons living alone in a rented accommodation (see [Appendix](#)) is not a closed data set as the total amount of expenditures is not the same for each individual. Nevertheless, if our interest is on the pattern or composition of expenditures, the only relevant information is contained in the ratios between expenditures in each group. Then, the vector $\mathbf{w} = [64.13, 76.26, 37.55, 19.74]$ (US\$) is compositionally equivalent to the closed vector $\mathbf{y} = [32, 39, 19, 10]$ (%).

Some mathematical properties:

- Let i be any index from $\{1, \dots, D\}$. Then it holds that two compositions $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ are compositionally equivalent if and only if it holds that $x_i/x_j = y_i/y_j$ for all indexes $i, j = 1, \dots, D, i \neq j$.
- For any $\mathbf{x} \in \mathbb{R}_+^D$ and $a \in \mathbb{R}^+$ it holds
 - $\mathcal{C}(\mathcal{C}\mathbf{x}) = \mathcal{C}\mathbf{x}$;
 - $\mathcal{C}(a\mathbf{x}) = \mathcal{C}\mathbf{x}$.

More recently, the definition of a composition as an equivalence class is introduced [BM16, Bar00]. When we say that two vectors \mathbf{w} and \mathbf{W} are compositionally equivalent we are assuming that they are proportional, that is, there is a positive constant a such that $\mathbf{W} = a\mathbf{w}$. This equivalence relation on \mathbb{R}_+^D splits the space into equivalence classes that we called compositions. The equivalence class of \mathbf{w} , or the composition \mathbf{w} , is the set $\{a\mathbf{w} : a \in \mathbb{R}^+\}$. This is a semi-straight line from the origin of the positive orthant of \mathbb{R}^D (see Fig. 1.1 for $D = 2$ and Fig. 1.2 for $D = 3$). The concept of compositionally equivalent and the equivalence class will be implicitly used in Section 1.2 to introduce a fundamental principle of CoDa, the scale invariance principle.

From a mathematical point of view, the set of all compositions is a quotient vector space. When we close our data to a constant k , we select a representative of our composition, say \mathbf{W} . Geometrically this corresponds to the intersection of the corresponding semi-straight line and the hyperplane of \mathbb{R}^D defined by equation $W_1 + \dots + W_D = k$ (Fig. 1.1 and 1.2).

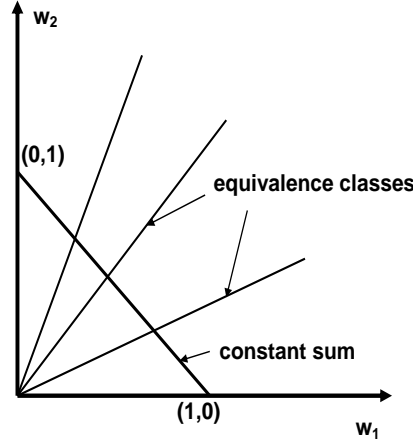


Figure 1.1. All points situated on the same ray of \mathbb{R}_+^2 are compositionally equivalent, giving rise to a composition or equivalent class. A representative is obtained with the intersection of the rays with the straight line $x_1 + x_2 = 1$.

The set of real positive vectors closed to a constant κ is called the *simplex* of D parts and is denoted \mathcal{S}^D from now on:

$$(1.2) \quad \mathcal{S}^D = \left\{ \mathbf{x} = [x_1, \dots, x_D] \in \mathbb{R}_+^D : \sum_{i=1}^D x_i = \kappa \right\}.$$

This notation is not completely standard and sometimes the superscript D for the number of parts is changed to $D-1$, indicating dimension or degrees of freedom instead of number of parts. Unless the value of κ is explicitly mentioned, we implicitly assume $\kappa = 1$.

The representation of compositions by their parts gives rise to useful graphical diagrams. Compositions of two parts can be plotted as a point in the interval $(0, \kappa)$.

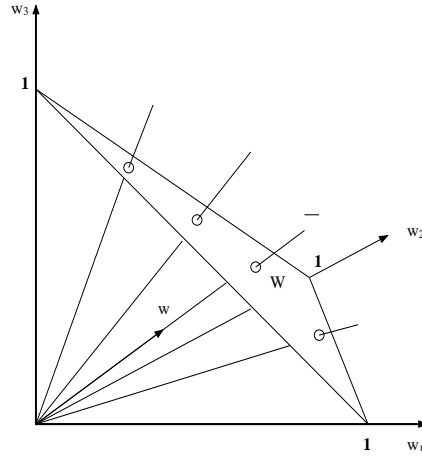


Figure 1.2. All vectors situated on the same ray of \mathbb{R}_+^3 are compositionally equivalent, giving rise to equivalence class. A representative is obtained with the intersection of the rays with the hyperplane forming the ternary diagram.

Compositions of three or four parts can be represented in ternary or tetrahedral diagrams, respectively. Figure 1.3 shows a ternary diagram in which a point represents

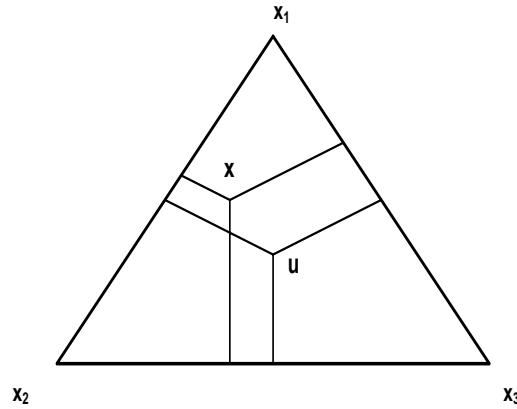


Figure 1.3. Ternary diagram. Representation of two 3-part compositions, $\mathbf{u} = C[1, 1, 1] = [1/3, 1/3, 1/3]$ and $\mathbf{x} = [0.50, 0.35, 0.15]$, and the segments proportional to their parts.

a composition. The parts of such a composition are proportional to the lengths of the perpendicular segments from the point to each side of the ternary diagram. The sum of such perpendiculars from any point within the equilateral triangle equals the length of the triangle's altitude (Viviani's theorem). The vertices of the ternary

diagram correspond to the three parts. It is customary to designate them by labels associated with the parts (in Figure 1.3, x_1 , x_2 , and x_3). The tetrahedral diagrams can be similarly interpreted replacing 'sides' (of triangle) by 'faces' (of tetrahedra).

Frequently, attention is centred on a group of parts of a composition. For example, a geologist may only be interested in the parts (Na_2O , K_2O , Al_2O_3) of a ten-part major oxide composition of a rock. Ratios of parts within the selected group are considered relevant; ratios involving some part, not in the group, are ignored. This corresponds to the definition of a subcomposition including only the parts in this group. Formally, if $S = \{i_1, \dots, i_C\}$ is a subset of the parts $1, \dots, D$ of a D -part composition \mathbf{x} , and \mathbf{x}_S is the subvector formed from the corresponding components of \mathbf{x} , then

$$\text{sub}(\mathbf{x}; S) = \mathcal{C}(\mathbf{x}_S) = [x_{i_1} \dots, x_{i_C}] / (x_{i_1} + \dots + x_{i_C})$$

is termed the *subcomposition* of the parts S . If C is the number of elements in S , then $\text{sub}(\mathbf{x}; S)$ is an element of simplex \mathcal{S}^C .

Example 3. In the example of the statistician's time budget (see Appendix), if we are only interested in the working activities T, C, A, and R, the *full* (or *complete*) composition $\mathcal{C}[3.50, 2.25, 4.25, 2.50, 6.25, 5.25]$ must be substituted by the following subcomposition ($\kappa = 100$):

$$\begin{aligned} & \mathcal{C}[3.50, 2.25, 4.25, 2.50] \\ &= \left[\frac{3.50 \times 100}{12.50}, \frac{2.25 \times 100}{12.50}, \frac{4.25 \times 100}{12.50}, \frac{2.50 \times 100}{12.50} \right] \\ &= [28.00, 18.00, 24.00, 20.00], \end{aligned}$$

where $12.50 = 3.50 + 2.25 + 4.25 + 2.50$. The components of the subcomposition vector give the percentages of daily working time devoted by the statistician to T, C, A and R, respectively.

Formation of a subcomposition may be considered as a transformation from the simplex \mathcal{S}^D to a simplex \mathcal{S}^C of lower dimension. Graphically it can be interpreted as a projection from the simplex \mathcal{S}^D to the subsimplex \mathcal{S}^C . Note that the simplicial projection involves the closure operator \mathcal{C} and, therefore, it is slightly more complex than the one for unconstrained vectors in \mathbb{R}^D , where a marginal vector is simply a subvector of the full D -dimensional vector. The concept of subcomposition is compatible with the equivalence class and quotient space. In fact, equivalent vectors are transformed into equivalent subvectors and, the selected parts provide the same relative information regardless they belong to the original or complete class of the subcomposition class [BM16]. This property is formulated in Section 1.2 as the subcompositional coherence principle but note that it is an inherent attribute to compositions viewed as an equivalence class.

1.1.2. The special difficulties of CoDa analysis[§]. We must go back to 1897 for our starting point. Over a century ago Karl Pearson published one of the

[§]This section is an adaptation of [Ait03, section 1.2, p. 13-15]. Further information in [Ait86, sections 3.1-3.3, p. 48-58] and [PET15, Chapter 1, p. 1-7].

clearest warnings [Pea97] ever issued to statisticians and other scientists beset with uncertainty and variability:

Beware of attempts to interpret correlations between ratios whose numerators and denominators contain common parts.

And such is the world of CoDa, where for example some rock specimen, of total weight w , is broken down into mutually exclusive and exhaustive parts with component weights w_1, \dots, w_D and then transformed into a composition

$$[x_1, \dots, x_D] = [w_1, \dots, w_D] / (w_1 + \dots + w_D).$$

Our reason for forming such a composition is that in many problems the composition is the relevant entity. For example the comparison of rock specimens of different weights can only be achieved by some form of standardization and composition (per unit weight) is a simple and obvious concept for achieving this. Equivalently we could say that any meaningful statement about the rock specimens should not depend on the largely accidental weights of the specimens.

It appears that Pearson's warning went unheeded, with raw components of CoDa being subjected to product moment correlation analysis with unsound interpretation based on methods of *standard* multivariate analysis designed for unconstrained multivariate data. In the 1960's there emerged a number of scientists who warned against such methodology and interpretation: mainly Chayes, Krumbein, Sarmanov and Vistelius in geology, and mainly Mosimann in biology; see, for example, [Cha56, Cha60, Cha62, Cha71], [Kru62], [SV59], [Mos62, Mos63]. The main problem was perceived as the impossibility of interpreting the product moment correlation coefficients between the raw components and was commonly referred to as the *negative bias problem*. For a D -part composition $[x_1, \dots, x_D]$ with the component sum $x_1 + \dots + x_D = 1$, since

$$\text{cov}(x_1, x_1 + \dots + x_D) = 0,$$

we have

$$\text{cov}(x_1, x_2) + \dots + \text{cov}(x_1, x_D) = -\text{var}(x_1).$$

The right hand side here is negative except for the trivial case where the first component is constant. Thus at least one of the covariances on the left must be negative or, equivalently, there must be at least one negative element in the first row of the raw covariance matrix. The same negative bias must occur in the same way in each of the other rows so that at least D elements of the raw covariance matrix are negative due to the constant sum constraint.

Example 4. Let's consider a simple 3-part CoDa set with 4 individuals

$$\mathbf{X} = \begin{bmatrix} 0.2 & 0.3 & 0.5 \\ 0.4 & 0.4 & 0.2 \\ 0.3 & 0.4 & 0.3 \\ 0.5 & 0.4 & 0.1 \end{bmatrix},$$

the corresponding covariance matrix is

$$\text{cov}(\mathbf{X}) = \begin{bmatrix} 0.0167 & 0.0050 & -0.0217 \\ 0.0050 & 0.0025 & -0.0075 \\ -0.0217 & -0.0075 & 0.0292 \end{bmatrix}.$$

We can see that the sum of the elements of each row is equal 0. Thus, at least one covariance of each row must be negative due to the constant sum constraint.

Hence correlations are not free to range over the usual interval $(-1, 1)$ subject only to the non-negative definiteness of the covariance or correlation matrix, and there are bound to be problems of interpretation.

The problem was described under different headings: the constant-sum problem, the closure problem, the negative bias problem, the null correlation difficulty. Strangely no attempt was made to try and establish principles of CoDa analysis. The approach was essentially pathological with attempts to see what went wrong when standard multivariate analysis was applied to CoDa in the hope that some corrective treatments could be applied; see, for example, [But79], [Cha71, Cha72], [ChK66], [ChT78], [DaJ74], and [DaR70, DaR78].

An appropriate methodology, taking into account some logically necessary principles of CoDa analysis and the special nature of compositional sample spaces, began to emerge in the 1980's with the work of John Aitchison. For example, contributions from [AiS80], [Ait82, Ait83, Ait85], culminating in the methodological monograph *The Statistical Analysis of Compositional Data* [Ait86].

Activities for Section 1.1

 [Click here to get the activities of this section](#)

1.2. Principles of CoDa analysis^{||}

CoDa analysis is based on the following general principle:

A composition quantitatively describes the parts of some whole and they provide only relative information between their components

1.2.1. The principle of scale invariance. When we say that a problem is compositional we are recognising that the *sizes* of our D -part compositions are irrelevant, that is, our analysis must be *scale invariant*. In other words, two compositions \mathbf{w} and \mathbf{W} are compositionally equivalent, written $\mathbf{W} \sim \mathbf{w}$, when a positive proportionality constant k exists, so that $\mathbf{W} = k\mathbf{w}$. This trivial admission has far-reaching consequences.

^{||}This section is an adaptation of [Ait03, Sections: 1.4, p. 17-19; 1.7, p. 21-22]. See also [EP11, Section 2.3, p. 15-17] and [PET15, Section 2.2, p. 12-16].

Example 5. A simple example can illustrate the argument. Consider two specimen vectors

$$\mathbf{w} = [1.6, 2.4, 4.0] \quad \text{and} \quad \mathbf{W} = [3.0, 4.5, 7.5]$$

representing the weights of the three parts A , B and C of two specimens of total weight of 8 g and 15 g, respectively. If we are interested in compositional problems we recognise that these are of the same composition, the difference in weight being accounted for by the scale relationship $\mathbf{W} = (15/8)\mathbf{w}$.

The fundamental requirement of CoDa analysis can then be stated as follows:

Any meaningful function f of a composition must be such that $f(\mathbf{W}) = f(\mathbf{w})$ when $\mathbf{W} \sim \mathbf{w}$, or equivalently

$$f(k\mathbf{w}) = f(\mathbf{w}) \quad \text{for every } k > 0.$$

In other words, the function f must be *invariant under the group of scale transformations*.

Since any group invariant function can be expressed as a function of any maximal invariant h and since

$$h(\mathbf{w}) = [w_1/w_D, \dots, w_{D-1}/w_D]$$

is such a maximal invariant we have the following important consequence:

Any meaningful (scale-invariant) function of a composition can be expressed in terms of ratios of the components of the composition.

Example 6. We present some examples of real scale-invariant functions defined on \mathbb{R}_+^D .

- The simple ratio of two components:

$$f : \mathbf{w} \rightarrow f(\mathbf{w}) = \frac{w_i}{w_j},$$

for any $i, j = 1, \dots, D$.

- The ratio between one component and the geometric mean of all components:

$$f : \mathbf{w} \rightarrow f(\mathbf{w}) = \frac{w_i}{g(\mathbf{w})},$$

for any $i = 1, \dots, D$, where $g(\mathbf{w}) = (w_1 \cdot w_2 \cdot \dots \cdot w_D)^{1/D}$.

- The ratio between the geometric mean of two subsets of components:

$$f : \mathbf{w} \rightarrow f(\mathbf{w}) = \frac{(w_{i_1} \cdot w_{i_2} \cdot \dots \cdot w_{i_{C_1}})^{1/C_1}}{(w_{j_1} \cdot w_{j_2} \cdot \dots \cdot w_{j_{C_2}})^{1/C_2}},$$

where $\{i_1, i_2, \dots, i_{C_1}\}$ and $\{j_1, j_2, \dots, j_{C_2}\}$ are two subsets of set $\{1, \dots, D\}$ of indices. By way of example,

$$f : \mathbf{w} \rightarrow f(\mathbf{w}) = \frac{(w_1 \cdot w_2)^{1/2}}{(w_2 \cdot w_3 \cdot w_5)^{1/3}}.$$

- Any linear combination of logarithms of components:

$$f : \mathbf{w} \rightarrow f(\mathbf{w}) = a_1 \ln w_1 + \dots + a_D \ln w_D ,$$

where $a_1 + \dots + a_D = 0$.

Such linear combinations are denominated *logcontrasts* [Ait86] and we can express it in terms of ratios of the components. In fact, using the logarithm properties, we have that $f(\mathbf{w}) = \ln w_1^{a_1} + \dots + \ln w_D^{a_D} = \ln(w_1^{a_1} \cdot \dots \cdot w_D^{a_D})$. We can write this product as a ratio because, as $a_1 + \dots + a_D = 0$, thus, at least one a_i is negative. That is, a logcontrast is a *logratio*

$$a_1 \ln w_1 + \dots + a_D \ln w_D = \ln \frac{\prod_{a_k > 0} w_k^{a_k}}{\prod_{a_j < 0} w_j^{|a_j|}}$$

1.2.2. Subcompositional coherence. Less familiar than scale invariance, but linked to it, is another logical necessity of CoDa analysis, namely *subcompositional coherence*. This means that studies performed on two parts of a subcomposition should not stand in contradiction with those performed when the two parts are considered in the full composition. Ignoring this principle of subcompositional coherence has been a source of great confusion in CoDa analysis. The literature, even currently, is full of attempts to explain the dependence of parts of compositions in terms of product moment correlation of raw components.

Example 7. Consider two scientists A and B interested in soil samples, which have been divided into aliquots. For each aliquot A records a 4-part composition (animal, vegetable, mineral, water); B first dries each aliquot without recording the water content and arrives at a 3-part composition (animal, vegetable, mineral). Let us further assume for simplicity the ideal situation where the aliquots in each pair are identical and where the two scientists are accurate in their determinations. Then clearly B 's 3-part composition $[s_1, s_2, s_3]$ for an aliquot will be a subcomposition of A 's 4-part composition $[x_1, x_2, x_3, x_4]$.

Scientist A	Scientist B
$[x_1, x_2, x_3, x_4]$	$[s_1, s_2, s_3]$
$[0.100, 0.200, 0.100, 0.600]$	$[0.250, 0.500, 0.250]$
$[0.200, 0.100, 0.100, 0.600]$	$[0.500, 0.250, 0.250]$
$[0.300, 0.300, 0.200, 0.200]$	$[0.375, 0.375, 0.250]$

It is then obvious that any compositional statements that A and B make about the common parts —animal, vegetable and mineral— must agree. This is the nature of subcompositional coherence. As regards the Pearson correlation coefficient, scientist A would report the correlation between animal and vegetable as $\text{corr}(x_1, x_2) = 0.5$ whereas B would report $\text{corr}(s_1, s_2) = -1$. Thus, there is incoherence in the product-moment correlation between raw components as a measure of dependence.

Note, however, that the ratio of two components remains unchanged when we move from full composition to subcomposition: $x_i/x_j = s_i/s_j$. So, as long as we

work with scale invariant functions, or equivalently express all our statements about compositions in terms of ratios, we shall be subcompositionally coherent.

Activities for Section 1.2

 [Click here to get the activities of this section](#)

1.3. Perturbation and power operations in the simplex^{††}

1.3.1. The role of group operations in statistics. For every sample space there are basic group operations which, when recognized, dominate clear thinking about data analysis. In \mathbb{R}^D the two operations, translation (or displacement) and scalar multiplication, are so familiar that their fundamental role is often overlooked. Yet the change from \mathbf{y} to $\mathbf{Y} = \mathbf{y} + \mathbf{t}$ by the translation \mathbf{t} , or to $\mathbf{Y} = a\mathbf{y}$ by the scalar multiple a are at the heart of statistical methodology for \mathbb{R}^D sample spaces. For example, since the translation relationship between \mathbf{y}_1 and \mathbf{Y}_1 is the same as that between \mathbf{y}_2 and \mathbf{Y}_2 if and only if \mathbf{Y}_1 and \mathbf{Y}_2 are equal translations \mathbf{t} of \mathbf{y}_1 and \mathbf{y}_2 , any definition of a difference or a distance measure must be such that the measure is the same for $(\mathbf{y}_1, \mathbf{Y}_1)$ as for $(\mathbf{y}_1 + \mathbf{t}, \mathbf{Y}_1 + \mathbf{t})$, for every translation \mathbf{t} . Technically this is a requirement of invariance under the group of translations. This is the reason, though seldom expressed because of its obviousness in this simple space, for the use of the mean vector $\boldsymbol{\mu} = E(\mathbf{y})$ and the covariance matrix $\boldsymbol{\Sigma} = \text{cov}(\mathbf{y}) = E\{(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^t\}$ as meaningful measures of 'central tendency' and 'dispersion'. Recall also, for further reference, two basic properties:

$$E(\mathbf{y} + \mathbf{t}) = E(\mathbf{y}) + \mathbf{t}, \quad \text{var}(\mathbf{y} + \mathbf{t}) = \text{var}(\mathbf{y}),$$

for any fixed translation $\mathbf{t} \in \mathbb{R}^D$. The second operation, that of scalar multiplication, also plays a substantial role in, for example, linear forms of statistical analysis such as principal component analysis, where linear combinations $a_1y_1 + \dots + a_Dy_D$ with certain properties are sought. Recall, again for further reference, that for a fixed scalar multiple a ,

$$E(a\mathbf{y}) = aE(\mathbf{y}), \quad \text{var}(a\mathbf{y}) = a^2\text{var}(\mathbf{y}).$$

Similar considerations of groups of fundamental operations are essential for other sample spaces. For example, in the analysis of directional data, as in the study of the movement of tectonic plates, it was recognition that the group of rotations on the sphere plays a central role and the use of a satisfactory representation of that group that led [Cha86] to the production of the essential statistical tool for spherical regression.

^{††}This section is an adaptation of [Ait03, Sections 1.8-1.10, p. 22-28]. See also [EP11, Section 2.4, p. 17-18]; [Ait86, Section 2.8, p. 42-43].

1.3.2. Perturbation (\oplus): a fundamental group operation in the simplex.

By analogy with the group operation arguments for \mathbb{R}^D the obvious questions for a simplex sample space are whether there is an operation on a composition \mathbf{x} , analogous to translation in \mathbb{R}^D , which transforms it into \mathbf{X} , and whether this can be used to characterize ‘difference’ between compositions or change from one composition to another. The answer is to be found in the *perturbation*.

For any $\mathbf{x} = [x_1, \dots, x_D]$ and $\mathbf{p} = [p_1, \dots, p_D] \in \mathcal{S}^D$, we define the perturbation operation and denote it by \oplus as

$$\mathbf{X} = \mathbf{p} \oplus \mathbf{x} = \mathcal{C}[p_1 x_1, \dots, p_D x_D].$$

Note that perturbation \mathbf{p} applied to the composition \mathbf{x} produces the composition \mathbf{X} . Remember that \mathcal{C} is the so-called *closure* operation which divides each component of a vector by the sum of the components, thus scaling the vector to the constant sum 1. Note that because of the nature of the scaling in this relationship it is not strictly necessary for the perturbation \mathbf{p} to be a vector in \mathcal{S}^D .

The perturbation operator can be motivated by the following observation within the positive orthant representation of compositions. For any two equivalent compositions \mathbf{w} and \mathbf{W} on the same ray there is a scale relationship $\mathbf{W} = p\mathbf{w}$ for some $p > 0$, where each component of \mathbf{w} is scaled by the same factor p to obtain the corresponding component of \mathbf{W} . For any two non-equivalent compositions \mathbf{w} and \mathbf{W} on different rays a similar, but differential, scaling relationship $W_1 = p_1 w_1, \dots, W_D = p_D w_D$ reflects the change from \mathbf{w} to \mathbf{W} . Such a unique differential scaling can always be found by taking $p_i = W_i/w_i$ ($i = 1, \dots, D$). That is, it holds $\mathbf{W} = \mathbf{p} \cdot \mathbf{w}$, where the product is component-wise. We can translate this into terms of the perturbation and the compositional representations \mathbf{x} and \mathbf{X} within the unit simplex sample space \mathcal{S}^D .

In Fig 1.4 (left) we can see the effect of the perturbation on the ternary diagram. We have a 3-part CoDa set $\mathbf{x} = [x_A, x_B, x_C]$ and we apply a perturbation to each composition taking $\mathbf{p} = [0.1, 0.1, 0.8]$. Note that the resulting data set is moved towards the vertex C , as the change in component C is eight times the change in A or B .

Example 8. Imagine a population composed of three different species of animals, A , B and C . Suppose that fifty years ago the percentages of species A , B and C were 20%, 60%, and 20%, respectively. At present, these percentages are 40%, 30%, and 30%, respectively. The change from the initial composition $\mathbf{x} = \mathcal{C}[20, 60, 20]$ to the current composition $\mathbf{y} = \mathcal{C}[40, 30, 30]$ can be interpreted in terms of a *perturbation difference*. Thus, if $\mathbf{p} = \mathcal{C}[40/20, 30/60, 30/20] = [0.500, 0.125, 0.375]$, it holds that $\mathbf{y} = \mathbf{p} \oplus \mathbf{x}$. The perturbation \mathbf{p} just gives information about the relative changes on the three parts A , B and C , from \mathbf{x} to \mathbf{y} . Thus, for example, since $p_1 = 4p_2$, which means that the change (in percentage) in species A in these fifty years is four times the change undergone by species B . Similarly, since $p_2 = p_3/3$, the change in B is one-third of the change undergone in C .

Perturbations can be found in very different contexts:

- In relation to probability statements the perturbation operation is a standard process. Thus, *Bayesians* perturb the prior probability assessment x on a finite number D of hypotheses by the likelihood p to obtain the posterior assessment X through the use of Bayes's formula.
- In genetic selection, the population composition \mathbf{x} of genotypes of one generation is perturbed by differential survival probabilities represented by a perturbation \mathbf{p} to obtain the composition \mathbf{X} at the next generation, again by the perturbation probabilistic mechanism.
- In certain geological processes, such as metamorphic change, sedimentation, crushing in relation to particle size distributions, change may be best modelled by such perturbation mechanisms, where an initial specimen of composition \mathbf{x}_0 is subjected to a sequence of perturbations $\mathbf{p}_1, \dots, \mathbf{p}_n$ in reaching its current state \mathbf{x}_n :

$$\mathbf{x}_1 = \mathbf{p}_1 \oplus \mathbf{x}_0, \mathbf{x}_2 = \mathbf{p}_2 \oplus \mathbf{x}_1, \dots, \mathbf{x}_n = \mathbf{p}_n \oplus \mathbf{x}_{n-1},$$

so that

$$\mathbf{x}_n = (\mathbf{p}_1 \oplus \mathbf{p}_2 \oplus \dots \oplus \mathbf{p}_n) \oplus \mathbf{x}_0.$$

It is clear that in this mechanism we have the makings of some form of central limit theorem but we will delay consideration of this until we have completed the more mathematical aspects of the simplex sample space.

- A further role which perturbation plays in simplicial inference is in characterizing imprecision or error. Thus, in the process of replicate analyses of aliquots of some specimen in an attempt to determine its composition ζ , we may obtain different compositions $\mathbf{x}_1, \dots, \mathbf{x}_n$ because of the imprecision of the analytic process. We can model this situation by setting

$$\mathbf{x}_i = \zeta \oplus \mathbf{p}_i \quad (i = 1, \dots, n),$$

where the \mathbf{p}_i ($i = 1, \dots, n$) are independent error perturbations characterising the imprecision.

1.3.3. Power (\odot) as a subsidiary operation in the simplex. The simplicial operation analogous to scalar multiplication in real space is the *power* operation. First we define the power operation and then consider its relevance in CoDa analysis. For any real number $a \in \mathbb{R}$ and any composition $\mathbf{x} \in \mathcal{S}^D$ we define

$$\mathbf{X} = a \odot \mathbf{x} = \mathcal{C}[x_1^a, \dots, x_D^a]$$

as the a -power transformation of \mathbf{x} .

In Fig 1.4 (right) we can see the effect of the power transformation on the ternary diagram. We have a 3-part CoDa set $\mathbf{x} = [x_A, x_B, x_C]$ and we apply a power transformation with $a = 0.2$. Note that the resulting data set is more concentrated because we choose $a < 1$.

Example 9. Sometimes, the nature of the sampling process is directly related to the power operation. Thus, in grain size studies of sediments, sediment samples may be successively sieved through meshes of different diameters and the weights of these successive separations converted into compositions based on proportions by

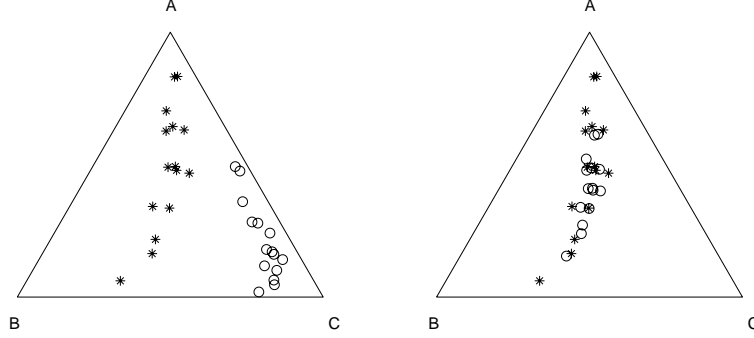


Figure 1.4. (left): perturbation of initial compositions (*) by $\mathbf{p} = [0.1, 0.1, 0.8]$ resulting in compositions (o); (right): power transformation of compositions (*) by $a = 0.2$ resulting in compositions (o)

weight. Thus though separation is based on the linear measurement *diameter* the composition is based essentially on a *weight*, or equivalently a *volume* measurement, with a power transformation being the natural connecting concept.

Imagine that we have mixed 500 balls of diameters 6, 4 and 2 cm. We will use two meshes M_5 and M_3 of diameters 5 and 3 cm, respectively, to separate the balls according to their diameter. We will use mesh M_5 to separate the 6 cm diameter balls from the others balls. Then, the mesh M_3 will allow us to separate the 4 cm diameter balls from those of 2 cm diameter. Suppose that we obtain 50, 350 and 100 balls of diameters 6, 4 and 2 cm, respectively. Thus, $\mathbf{x} = \mathcal{C}[50, 350, 100] = [0.1, 0.7, 0.2]$ reflects the composition of the *number* of balls according to their size. Assuming that all balls are made of the same material, the total weight of 6 cm diameter balls will be equal to $k \cdot 50 \cdot 6^3$, where the factor k depends on the density of the material. Similarly, the total weight of 4 and 2 cm balls will be equal to $k \cdot 350 \cdot 4^3$ and $k \cdot 100 \cdot 2^3$, respectively. Therefore, the composition $\mathbf{w} = \mathcal{C}[k \cdot 50 \cdot 6^3, k \cdot 350 \cdot 4^3, k \cdot 100 \cdot 2^3] = [0.317, 0.659, 0.024]$ reflects the composition by *weight* of the three sizes of balls. Observe that we can use the compositional operations \odot and \oplus to express \mathbf{w} in terms of \mathbf{x} and $\mathcal{C}[6, 4, 2]$, that is $\mathbf{w} = (3 \odot \mathcal{C}[6, 4, 2]) \oplus \mathbf{x}$. Similarly, it is possible to express the composition \mathbf{x} in terms of \mathbf{w} , since $\mathbf{x} = ((-3) \odot \mathcal{C}[6, 4, 2]) \oplus \mathbf{w}$.

It must be clear that together the operations perturbation \oplus and power \odot play roles in the geometry of \mathcal{S}^D analogous to translation and scalar multiplication in \mathbb{R}^D and indeed can be used to define a vector space in \mathcal{S}^D . We shall take up the full algebraic-geometric structure of the simplex sample space later in the next section.

Some mathematical properties:


- For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^D$, and $a \in \mathbb{R}$, it holds
 - $(\mathcal{C}\mathbf{x}) \oplus (\mathcal{C}\mathbf{y}) = \mathcal{C}[x_1y_1, \dots, x_Dy_D]$;
 - $a \odot (\mathcal{C}\mathbf{x}) = \mathcal{C}(a \odot \mathbf{x})$.

- Let n be a positive integer and \mathbf{x} be a composition. It holds that

$$\underbrace{\mathbf{x} \oplus \mathbf{x} \oplus \dots \oplus \mathbf{x}}_n = n \odot \mathbf{x} .$$

- Let $S = \{i_1, \dots, i_C\}$ be a subset of the parts $1, \dots, D$ of the simplex \mathcal{S}^D . For any $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ and $a \in \mathbb{R}$, it holds
 - $\text{sub}(\mathbf{x} \oplus \mathbf{y}; S) = \text{sub}(\mathbf{x}; S) \oplus \text{sub}(\mathbf{y}; S)$;
 - $\text{sub}(a \odot \mathbf{x}; S) = a \odot \text{sub}(\mathbf{x}; S)$.

Activities for Section 1.3

 [Click here to get the activities of this section](#)

1.4. Original logratio analysis: alr and clr transformations ^{§§}

What has come to be known as *log-ratio analysis for compositional problems* arose in the 1980's out of the realization of the importance of the principle of scale invariance (see Section 1.2.1) and that its practical implementation required working with ratios of components. This, together with an awareness that logarithms of ratios are mathematically more tractable than ratios, led to the advocacy of a transformation technique involving logratios of the components. In this section we will introduce the two main transformations on the simplex on which this analysis is based.

1.4.1. The additive logratio transformation (alr). Let $\mathbf{x} = [x_1, \dots, x_D] \in \mathcal{S}^D$ be a typical D -part composition. Then the so-called *additive logratio* transformation $\text{alr} : \mathcal{S}^D \rightarrow \mathbb{R}^{D-1}$ is defined by

$$(1.3) \quad \mathbf{y} = \text{alr } \mathbf{x} = [\text{alr}_1 \mathbf{x}, \text{alr}_2 \mathbf{x}, \dots, \text{alr}_{D-1} \mathbf{x}] = [\ln(x_1/x_D), \ln(x_2/x_D), \dots, \ln(x_{D-1}/x_D)],$$

where the ratios involve the division of each of the first $D - 1$ components by the final component.

The *alr* transformation is one-to-one. The inverse transformation $\text{alr}^{-1} : \mathbb{R}^{D-1} \rightarrow \mathcal{S}^D$ is

$$\mathbf{x} = \text{alr}^{-1} \mathbf{y} = \mathcal{C}[\exp y_1, \dots, \exp y_{D-1}, 1],$$

where \mathcal{C} denotes the closure operation.

Note that the *alr* transformation takes the D -part composition into the whole of the \mathbb{R}^{D-1} space and so we have the prospect of using standard unconstrained multivariate analysis on the transformed data, and because of the one-to-one nature of this transformation, of transferring any inferences back to the simplex and to the components of the composition.

One apparent drawback to this technique is the choice of the final component as the divisor, with the frequently asked question: would we obtain the same inference if we chose another component as divisor, or more generally if we permuted the

^{§§}This section is an adaptation of [Ait03, Sections 2.1-2.2, p. 29-32].

parts? The answer to this question is *yes*. We will not go into any details here that prove this assertion, but the interested reader may find these in [Ait86, Chapter 5]. Nevertheless, the principal drawback of the *alr* transformation is that distances are not preserved. We will see this in detail in Section 1.5.2, once a compositional distance is defined.

1.4.2. The centred logratio transformation (clr). The *alr* transformation is asymmetric in the parts and it is sometimes convenient to treat the parts symmetrically. This can be achieved by the so-called *centred logratio* transformation $\text{clr} : \mathcal{S}^D \rightarrow \mathbb{R}^D$:

$$(1.4) \quad \mathbf{z} = \text{clr } \mathbf{x} = [\text{clr}_1 \mathbf{x}, \text{clr}_2 \mathbf{x}, \dots, \text{clr}_D \mathbf{x}] = \left[\ln \frac{x_1}{g(\mathbf{x})}, \ln \frac{x_2}{g(\mathbf{x})}, \dots, \ln \frac{x_D}{g(\mathbf{x})} \right],$$

where $g(\mathbf{x}) = (x_1 \cdot x_2 \cdot \dots \cdot x_D)^{1/D}$ is the geometric mean of the components of the composition \mathbf{x} . The *clr* transformation maps \mathcal{S}^D into the $(D-1)$ -dimensional subspace U of \mathbb{R}^D defined as

$$(1.5) \quad U = \{\mathbf{z} \in \mathbb{R}^D : z_1 + \dots + z_D = 0\}.$$

The *clr* transformation is one-to-one from \mathcal{S}^D to U . The inverse transformation $\text{clr}^{-1} : U \rightarrow \mathcal{S}^D$ takes the form

$$(1.6) \quad \mathbf{x} = \text{clr}^{-1} \mathbf{z} = \mathcal{C}[\exp z_1, \dots, \exp z_D].$$

Note that the *clr*-transformed vector has D components but belongs to a $D-1$ dimensional subspace. For this reason, the resulting vector is also constrained because the sum of its parts equals to 0. Consequently, the covariance matrix of a *clr*-transformed data set is singular and the Pearson correlation coefficient $\text{corr}(\text{clr}_k \mathbf{x}, \text{clr}_j \mathbf{x})$, $k, j = 1, \dots, D$ is not informative. This is the principal drawback of the *clr* transformation.

But, despite the drawbacks, these two transformations of \mathcal{S}^D to a real space open up the possibility of using standard multivariate methods. Thus, the mean vector $\boldsymbol{\mu} = \text{E}(\text{alr } \mathbf{x})$ and covariance matrix $\boldsymbol{\Sigma} = \text{cov}(\text{alr } \mathbf{x})$ of the logratio vector $\text{alr } \mathbf{x}$ will play an important role in our CoDa analysis, as will the centred logratio analogues $\boldsymbol{\lambda} = \text{E}(\text{clr } \mathbf{x})$ and $\boldsymbol{\Gamma} = \text{cov}(\text{clr } \mathbf{x})$.

Some mathematical properties:

- If $D = 2$, the *alr* and *clr* transformations from \mathcal{S}^2 (closed to $\kappa = 1$) are well related with the *logit* function. Thus,

$$\text{alr}[x, 1-x] = \ln \frac{x}{1-x} = \text{logit } x,$$

and

$$\text{clr}[x, 1-x] = \left[\frac{1}{2} \ln \frac{x}{1-x}, -\frac{1}{2} \ln \frac{x}{1-x} \right] = \frac{1}{2} [\text{logit } x, -\text{logit } x].$$

- Given any composition $\mathbf{x} \in \mathcal{S}^D$, the real vectors $\text{alr } \mathbf{x} \in \mathbb{R}^{D-1}$ and $\text{clr } \mathbf{x} \in \mathbb{R}^D$ are linearly related:

$$(1.7) \quad \text{clr } \mathbf{x} = (\text{alr } \mathbf{x}) \cdot \mathbf{K}^t \quad \text{alr } \mathbf{x} = (\text{clr } \mathbf{x}) \cdot \mathbf{F}^t,$$

where \mathbf{K} is the $D \times (D - 1)$ matrix defined as

$$(1.8) \quad \mathbf{K} = \begin{bmatrix} 1 - 1/D & -1/D & -1/D & \dots & -1/D \\ -1/D & 1 - 1/D & -1/D & \dots & -1/D \\ -1/D & -1/D & 1 - 1/D & \dots & -1/D \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1/D & -1/D & -1/D & \dots & 1 - 1/D \\ -1/D & -1/D & -1/D & \dots & -1/D \end{bmatrix},$$

and \mathbf{F} is the $(D - 1) \times D$ matrix $[\mathbf{I}_{D-1} : -\mathbf{1}_{D-1}]$, that is the identity matrix \mathbf{I}_{D-1} with an extra last column of -1's. More explicitly,

$$(1.9) \quad \mathbf{F} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & -1 \\ 0 & 1 & 0 & \dots & 0 & -1 \\ 0 & 0 & 1 & \dots & 0 & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & -1 \\ 0 & 0 & 0 & \dots & 1 & -1 \end{bmatrix}$$

Therefore, given any of the two vectors —alr \mathbf{x} or clr \mathbf{x} — we can easily calculate the other vector.

- $\mathbf{z} = \text{clr } \mathbf{x}$ can be interpreted as the row-centred vector of $\ln \mathbf{x}$. Indeed, it holds

$$\mathbf{z} = \left[\ln \frac{x_1}{g(\mathbf{x})}, \dots, \ln \frac{x_D}{g(\mathbf{x})} \right] = \ln \mathbf{x} - (\ln g(\mathbf{x}), \dots, \ln g(\mathbf{x})) = \ln \mathbf{x} - (\overline{\ln \mathbf{x}}, \dots, \overline{\ln \mathbf{x}}),$$

because $\ln g(\mathbf{x}) = \frac{1}{D} \sum_{j=1}^D \ln x_j$.

- $\text{clr}_k \mathbf{x}$, for $k = 1, \dots, D$, can be interpreted as an average of the pairwise logratios of the part x_k against the other parts. Indeed, it holds

$$\text{clr}_k \mathbf{x} = \ln \frac{x_k}{g(\mathbf{x})} = \ln x_k - \frac{1}{D} \sum_{j=1}^D \ln x_j = \frac{1}{D} \left(\ln \frac{x_k}{x_1} + \dots + \ln \frac{x_k}{x_D} \right),$$

informing about the relative importance, in average, of the part x_k as regards the other parts.

1.4.3. Philosophy of the logratio analysis. The philosophy of log-ratio analysis can be stated simply:

- (1) Formulate the compositional problem in terms of the components of the composition.
- (2) Translate this formulation into terms of the log-ratio vector of the composition.
- (3) Transform the CoDa into log-ratio vectors.
- (4) Analyse the log-ratio data by an appropriate standard multivariate statistical method.
- (5) Translate back into terms of the compositions the inference obtained at step 4.

Later we will see many examples of this compositional methodology.


Log-ratio analysis has been successfully applied in a wide variety of disciplines. Since, however, there seem to be an appreciable number of statisticians and scientists who seem, for whatever reason, uncomfortable with transformation techniques, it is worth considering what the alternatives are. In the discussion of [Ait82], Fisher made the following comment:

.../... Clearly the speaker has been very successful in fitting simple models to normal transformed data, the counterpart to the simplicity of these models is the complexity of corresponding relationships among the untransformed components. This is hardly an original observation. Yet there are certain aromas rising from the murky potage of CoDa problems which are redolent of some aspects of problems with directional data, and herein lies the point. When attacking these latter problems, one is ultimately better off working within the confines of the original geometry (of the circle, sphere cylinder, ...) and with techniques particular thereto (vector methods, etc), in terms of perceiving simple underlying ideas and modelling them in a natural way. Mapping from, say, the sphere into the plane, and then back, rarely produces these elements, and usually introduces unfortunate distortion. I still hold out some hope that simple models of dependence can be found, peculiar to the simplex .../... Meanwhile, I shall analyse data with the normal-transform methods.

The lack of success in transforming the sphere into the plane is that the spaces are topologically different whereas the simplex and real space are topologically equivalent. Nevertheless, it is a challenge to confine the statistical argument to the geometry of the simplex, and this approach has been emerging over the last decade, based on the operations of perturbation and power and on the simplicial metric.

Alternatively, it is now certainly possible to analyse CoDa entirely within simplicial geometry that is using the particular operations for compositions. This is, what we call the *stay-in-the-simplex approach*. Clearly the success of such an approach must depend largely on the mathematical sophistication of the user. In this online course we shall adopt a bilateral approach, attempting to interpret inferences from our CoDa problems both from the log-ratio analysis approach and the *staying-in-the-simplex approach*.

Activities for Section 1.4

 [Click here to get the activities of this section](#)

1.5. The algebraic-geometric structure of the simplex ***

We begin this section by introducing in greater detail the two operations —perturbation and powering— previously addressed in Section 1.3. Then, we define a distance on the simplex, giving it the structure of Euclidean space.

1.5.1. The simplex $(\mathcal{S}^D, \oplus, \odot)$ real vector space. It is easy to prove that the perturbation and the powering operators satisfy the properties required to give a vector-space-structure to the simplex \mathcal{S}^D .

The perturbation operator \oplus defined in Section 1.3.2 can be viewed as an internal operation on the simplex \mathcal{S}^D . Thus, if $\mathbf{x} = [x_1, \dots, x_D]$, $\mathbf{y} = [y_1, \dots, y_D]$ are in \mathcal{S}^D , their *perturbation* is

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}[x_1 y_1, \dots, x_D y_D] .$$

Given $\mathbf{x} \in \mathcal{S}^D$, the perturbation operation allows us to define the opposite or inverse element of \mathbf{x} . It is the composition $\mathcal{C}[1/x_1, \dots, 1/x_D]$, denoted as $\ominus \mathbf{x}$ such that $\mathbf{x} \oplus (\ominus \mathbf{x}) = (\ominus \mathbf{x}) \oplus \mathbf{x} = \mathcal{C}[1, \dots, 1]$. The composition $\mathcal{C}[1, \dots, 1]$ is called the neutral element of the simplex and denoted as \mathbf{u} .

Note that the closure operation is implicit in the perturbation. Perturbation can be carried out using compositions closed to different constants and no previous closure to the same constant is required. For instance, let $\mathbf{x} = [x_1, \dots, x_D]$ and $\mathbf{y} = [y_1, \dots, y_D]$ be compositions closed to κ_x and κ_y , and we need to compute their perturbation $\mathbf{z} = \mathbf{x} \oplus \mathbf{y}$ closed to κ . The result can be obtained as

$$\mathbf{z} = [z_1, \dots, z_D] = \frac{\kappa}{\sum_{i=1}^D x_i y_i} [x_1 y_1, \dots, x_D y_D] = \mathcal{C}[x_1 y_1, \dots, x_D y_D] ,$$

where the closure constants, κ_x and κ_y , do not play any role.

The perturbation of a composition with the opposite of a composition \mathbf{y} will be denoted by the symbol \ominus , that is

$$\mathbf{x} \oplus (\ominus \mathbf{y}) = \mathbf{x} \ominus \mathbf{y} .$$

This operation is equivalent to subtraction in real vector spaces and will be called *perturbation-difference*.

The powering defined in Section 1.3.3 can be viewed as an external operation in \mathcal{S}^D by a real constant. In \mathcal{S}^D it plays the same role as multiplication by real constants in real space. Let a be a scalar and \mathbf{x} be a composition in \mathcal{S}^D . *Powering* \mathbf{x} by a is defined as

$$a \odot \mathbf{x} = \mathcal{C}[x_1^a, \dots, x_D^a] .$$

Note that powering by -1 can be used to define the opposite element: $(-1) \odot \mathbf{x} = \ominus \mathbf{x}$.

Without going into mathematical details here, note that the perturbation, powering operations allows as to define the concept of a compositional line with starting compositional point \mathbf{x}_0 and leading vector \mathbf{x} as $\mathbf{y} = \mathbf{x}_0 \oplus (\alpha \odot \mathbf{x})$ with $\alpha \in \mathbb{R}$. In Fig 1.5 we can see some compositional lines represented on the ternary diagram.

***This section is an adaptation of [EP06, p. 147-148, p. 150-151] and [Ait03, Section 2.3.3, p. 23-36]. Further information in [PET15, Chapter 3, p. 23-31].

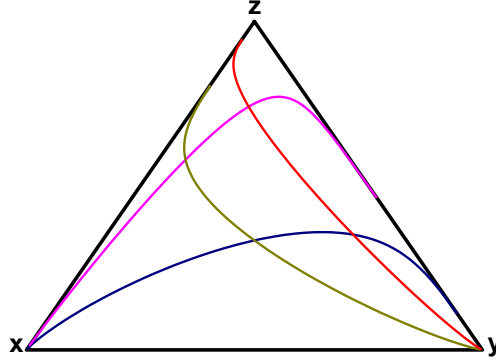


Figure 1.5. Compositional lines.

As regards the operators “ \oplus ” and “ \odot ”, it is easy to prove that the additive and the centred logratio transformations defined in Eq. 1.3 and 1.4 have analogous properties fulfilled by the logarithm: $\ln(x \cdot y) = \ln x + \ln y$ and $\ln(x^a) = a \ln x$. That is, *alr* and *clr* transformations fulfill the following equalities,

$$\text{alr}(\mathbf{x} \oplus \mathbf{y}) = \text{alr } \mathbf{x} + \text{alr } \mathbf{y} \quad \text{and} \quad \text{alr}(a \odot \mathbf{x}) = a \cdot \text{alr } \mathbf{x} ,$$

$$\text{clr}(\mathbf{x} \oplus \mathbf{y}) = \text{clr } \mathbf{x} + \text{clr } \mathbf{y} \quad \text{and} \quad \text{clr}(a \odot \mathbf{x}) = a \cdot \text{clr } \mathbf{x} ,$$

This means that using the *alr* or *clr*-vectors we can apply the standard operations, translation and scalar multiplication, we use in real space. The results, obviously, are real vectors. To obtain the corresponding composition we have only to apply the inverse transformation, alr^{-1} or clr^{-1} .

Some mathematical properties:

- The operator perturbation (\oplus) fulfills the standard properties of a commutative group operation, which are:
 - a) Internal operation: $\mathbf{x} \oplus \mathbf{y}$ is in \mathcal{S}^D .
 - b) Commutative: $\mathbf{x} \oplus \mathbf{y} = \mathbf{y} \oplus \mathbf{x}$.
 - c) Neutral element: $\mathbf{u} = \mathcal{C}[1, \dots, 1]$, satisfying $\mathbf{x} \oplus \mathbf{u} = \mathbf{u} \oplus \mathbf{x} = \mathbf{x}$.
 - d) Opposite element: given $\mathbf{x} \in \mathcal{S}^D$ there is an opposite element, namely $\ominus \mathbf{x} = \mathcal{C}[1/x_1, \dots, 1/x_D]$, such that $\mathbf{x} \oplus (\ominus \mathbf{x}) = (\ominus \mathbf{x}) \oplus \mathbf{x} = \mathbf{u}$.
- The main properties of powering (\odot) are:
 - a) Unitary element: the real number 1 satisfies $1 \odot \mathbf{x} = \mathbf{x}$.
 - b) Distributive with respect to perturbation: $a \odot (\mathbf{x} \oplus \mathbf{y}) = (a \odot \mathbf{x}) \oplus (a \odot \mathbf{y})$.
- The additive logratio transformation is a *linear map* from the vector space \mathcal{S}^D to the real vector space \mathbb{R}^{D-1} . Similarly, the centred logratio transformation is a linear map from the vector space \mathcal{S}^D to the real $(D - 1)$ -dimensional subspace U of \mathbb{R}^D defined in Eq. (1.5).

1.5.2. The simplex \mathcal{S}^D Euclidean space: distance, inner product and norm.

[Ait86] introduced a simplicial distance suitable for the analysis of CoDa. If $\mathbf{x} = [x_1, x_2, \dots, x_D]$ and $\mathbf{y} = [y_1, y_2, \dots, y_D]$ are compositions in \mathcal{S}^D , the *compositional distance* (or Aitchison distance) between them is defined as

$$(1.10) \quad \begin{aligned} d_a(\mathbf{x}, \mathbf{y}) &= \left(\frac{1}{D} \sum_{i=1}^{D-1} \sum_{j=i+1}^D \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2 \right)^{1/2} \\ &= \left(\sum_{i=1}^D \left(\ln \frac{x_i}{g(\mathbf{x})} - \ln \frac{y_i}{g(\mathbf{y})} \right)^2 \right)^{1/2}, \end{aligned}$$

This compositional distance is called Aitchison distance and in the literature it is denoted as d_a (*hereinafter we will use the subindex “a” for the basic elements of the Aitchison geometry: distance, inner product and norm*).

The last member of Eq. (1.10) demonstrates that the compositional distance between two compositions is equal to the Euclidean distance between the *clr*-transformed compositions, that is

$$(1.11) \quad d_a(\mathbf{x}, \mathbf{y}) = d(\text{clr } \mathbf{x}, \text{clr } \mathbf{y}) = \left(\sum_{i=1}^D (\text{clr }_i \mathbf{x} - \text{clr }_i \mathbf{y})^2 \right)^{1/2}.$$

Consequently, the *clr* transformation is an *isometry*, it preserves the distances. This equality is not true for the *alr*-transformed compositions, that is

$$d_a(\mathbf{x}, \mathbf{y}) \neq d(\text{alr } \mathbf{x}, \text{alr } \mathbf{y}).$$

This is the principal drawback of the *alr* transformation, it is not an isometry. For this reason, its use is not recommended.

Remember that the philosophy of the log-ratio analysis is to translate our compositional problem in terms of the log-ratio vectors and apply a standard statistical method. Note that if we combine the additive-logratio vectors with a standard method that uses distances, our results will not be the same as we stay in the simplex and use the compositional distance.

The compositional distance has desirable, relevant, and logically necessary properties, such as scale and perturbation invariance, and subcompositional dominance:

a) Perturbation invariance: if $\mathbf{p} \in \mathcal{S}^D$,

$$(1.12) \quad d_a(\mathbf{x} \oplus \mathbf{p}, \mathbf{y} \oplus \mathbf{p}) = d_a(\mathbf{x}, \mathbf{y}).$$

b) Subcompositional dominance: let S be a set of C subscripts, $1 < C < D$, then

$$d_a(\text{sub}(\mathbf{x}; S), \text{sub}(\mathbf{y}; S)) \leq d_a(\mathbf{x}, \mathbf{y}).$$

The Aitchison distance (Eq. (1.10)) can be derived from an inner product defined in the simplex. The *compositional inner product* (or Aitchison inner product) of two compositions \mathbf{x}, \mathbf{y} in \mathcal{S}^D can be defined as

$$(1.13) \quad \langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{D} \sum_{i=1}^{D-1} \sum_{j=i+1}^D \ln \frac{x_i}{x_j} \cdot \ln \frac{y_i}{y_j} = \sum_{i=1}^D \ln \frac{x_i}{g(\mathbf{x})} \cdot \ln \frac{y_i}{g(\mathbf{y})},$$

where $g(\cdot)$ is again the geometric mean of the components of the vector in the argument.

As occurs with the Aitchison distance, the Aitchison inner product of two compositions is equal to the standard inner product of the *clr*-transformed compositions, that is

$$(1.14) \quad \langle \mathbf{x}, \mathbf{y} \rangle_a = \langle \text{clr } \mathbf{x}, \text{clr } \mathbf{y} \rangle .$$

The Aitchison inner product allows us to define the *compositional norm* (or Aitchison norm) of a composition and to re-define the compositional distance between compositions,

$$(1.15) \quad \|\mathbf{x}\|_a = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_a} \quad , \quad d_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_a .$$

It is also clear that the Aitchison norm of a composition is equal to the standard norm of the *clr*-transformed composition, that is

$$(1.16) \quad \|\mathbf{x}\|_a = \|\text{clr } \mathbf{x}\| = \left(\sum_{i=1}^D \left(\ln \frac{x_i}{g(\mathbf{x})} \right)^2 \right)^{1/2} .$$

Finally, we say that a composition is *unitary* if its Aitchison norm is equal to one. It follows that \mathbf{x} will be unitary if and only if $\text{clr } \mathbf{x}$ is unitary in \mathbb{R}^D , that is if

$$\sum_{i=1}^D \left(\ln \frac{x_i}{g(\mathbf{x})} \right)^2 = 1 .$$

As in the real space, we define the *compositional angle* (Aitchison angle) ϑ_a between two compositions \mathbf{x} and \mathbf{y} (with $\mathbf{x} \neq \mathbf{u}$ and $\mathbf{y} \neq \mathbf{u}$) as

$$\vartheta_a = \cos^{-1} \frac{\langle \mathbf{x}, \mathbf{y} \rangle_a}{\|\mathbf{x}\|_a \cdot \|\mathbf{y}\|_a} \quad (\text{with } 0 \leq \vartheta_a \leq \pi) .$$

Therefore, it holds that

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \|\mathbf{x}\|_a \cdot \|\mathbf{y}\|_a \cdot \cos \vartheta_a .$$

Thereby, we say that two compositions ($\neq \mathbf{u}$) are *orthogonal* if $\vartheta_a = \pi/2$, that is if their Aitchison inner product is zero. It follows that $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ will be orthogonal if and only if $\text{clr } \mathbf{x}$ and $\text{clr } \mathbf{y}$ are orthogonal in \mathbb{R}^D , that is if

$$(1.17) \quad \sum_{i=1}^D \ln \frac{x_i}{g(\mathbf{x})} \cdot \ln \frac{y_i}{g(\mathbf{y})} = 0 .$$

In general, given two compositions \mathbf{x} and \mathbf{y} (with $\mathbf{x} \neq \mathbf{u}$ and $\mathbf{y} \neq \mathbf{u}$), the (signed) *orthogonal projection* of \mathbf{x} onto \mathbf{y} is equal to $\|\mathbf{y}\|_a \cdot \cos \vartheta_a$, and $\|\mathbf{x}\|_a \cdot \cos \vartheta_a$ is (the signed) orthogonal projection of \mathbf{y} onto \mathbf{x} .

Some mathematical properties:

- The compositional distance satisfies the usual metric axioms:
 - a) Positivity: $d_a(\mathbf{x}, \mathbf{y}) \geq 0$; $d_a(\mathbf{x}, \mathbf{y}) = 0 \leftrightarrow \mathbf{x} = \mathbf{y}$.
 - b) Symmetry: $d_a(\mathbf{x}, \mathbf{y}) = d_a(\mathbf{y}, \mathbf{x})$.
 - c) Powering relationship: $d_a(a \odot \mathbf{x}, a \odot \mathbf{y}) = |a| \cdot d_a(\mathbf{x}, \mathbf{y})$.
 - d) Triangular inequality: $d_a(\mathbf{x}, \mathbf{y}) + d_a(\mathbf{y}, \mathbf{z}) \geq d_a(\mathbf{x}, \mathbf{z})$.

- The Aitchison inner product satisfies the following standard properties:
 - a) Positivity: $\langle \mathbf{x}, \mathbf{x} \rangle_a > 0$ if $\mathbf{x} \neq \mathbf{u}$.
 - b) Commutativity: $\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle \mathbf{y}, \mathbf{x} \rangle_a$.
 - c) Distributive with respect to perturbation: $\langle \mathbf{x} \oplus \mathbf{z}, \mathbf{y} \rangle_a = \langle \mathbf{x}, \mathbf{y} \rangle_a + \langle \mathbf{z}, \mathbf{y} \rangle_a$.
 - d) Powering relationship: $\langle a \odot \mathbf{x}, \mathbf{y} \rangle_a = a \cdot \langle \mathbf{x}, \mathbf{y} \rangle_a$.

The compositional distance together with the inner product from which it is derived allows us to define the concept of orthogonal and parallel lines (see Fig 1.6). Finally we could define other mathematical objects as compositional circumferences or ellipses (see Fig 1.7).

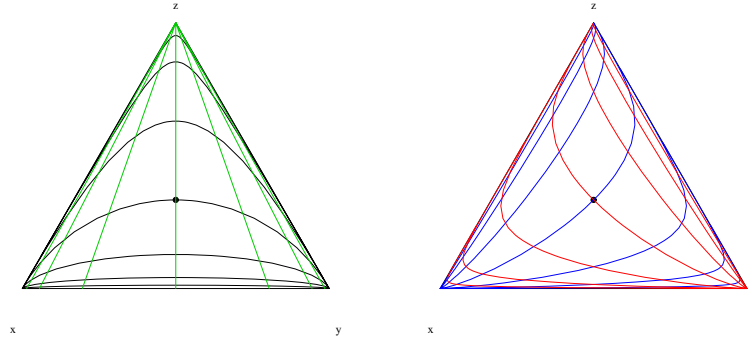


Figure 1.6. Compositional lines; (left) green lines are orthogonal to black lines; (right) Blue lines are orthogonal to red lines. In both figures, the lines of the same color are parallel.

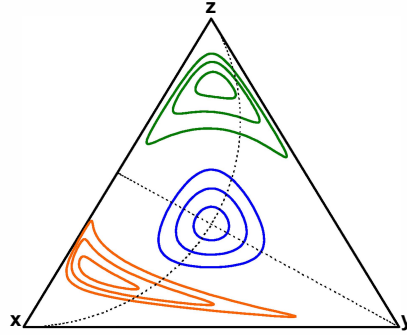


Figure 1.7. Compositional circumferences (in blue) and compositional ellipses (in green and orange).

A Euclidean space is a vector space in which an inner product is defined satisfying the above mentioned properties (a), (b), (c) and (d). Therefore, the simplex \mathcal{S}^D has the structure of a Euclidean space.

From a mathematical perspective this means that \mathcal{S}^D is completely equivalent to \mathbb{R}^{D-1} . In order to use this fact in practice, compositions have to be represented by coordinates that are actually real vectors, the values of which are not constrained to be positive or less than one.

Activities for Section 1.5

🔗 [Click here to get the activities of this section](#)

1.6. Compositional-linear dependence, basis and coordinates †††

The vector space structure of \mathcal{S}^D allows the concepts of linear dependence and independence to be used.

A set of m compositions in \mathcal{S}^D , $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m$, and m scalars, a_1, a_2, \dots, a_m , is combined (linearly) in a *compositional linear combination* as

$$(1.18) \quad (a_1 \odot \mathbf{x}_1) \oplus (a_2 \odot \mathbf{x}_2) \oplus \dots \oplus (a_m \odot \mathbf{x}_m) = \bigoplus_{i=1}^m (a_i \odot \mathbf{x}_i),$$

which is a perturbation-powering version of the traditional linear combination in real vector spaces. The symbol \bigoplus represents repeated perturbation on the subscripts.

A set of m compositions $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m$ from \mathcal{S}^D is called *linear dependent* if and only if there exists a set of m scalars a_1, a_2, \dots, a_m , not all zero, such that the linear combination (1.18) is equal to the neutral element \mathbf{u} . If such scalars do not exist, then the compositions are said to be *linear independent*.

Since the additive log-ratio transformation alr is an isomorphism between the vector spaces \mathcal{S}^D and \mathbb{R}^{D-1} , the dimension of \mathcal{S}^D is $D-1$. Therefore, the maximum number of independent compositions in \mathcal{S}^D is $D-1$. If the D -part compositions $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{D-1}$ are linear independent, they constitute a *basis* of \mathcal{S}^D . This implies that each composition $\mathbf{x} \in \mathcal{S}^D$ can be expressed as a combination

$$(1.19) \quad \mathbf{x} = (x_1^* \odot \mathbf{e}_1) \oplus (x_2^* \odot \mathbf{e}_2) \oplus \dots \oplus (x_{D-1}^* \odot \mathbf{e}_{D-1}) = \bigoplus_{i=1}^{D-1} (x_i^* \odot \mathbf{e}_i),$$

for some coefficients x_i^* that are then termed *coordinates* of composition \mathbf{x} with respect to the basis $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{D-1}$ of \mathcal{S}^D .

Given a basis, the coordinates x_1^*, \dots, x_{D-1}^* of \mathbf{x} are uniquely determined. This allows the representation of a composition by its coordinates with respect to a given basis. Hereafter the asterisk superscript shall be used to denote coordinates with respect to a given basis. In turn the vector $\mathbf{x}^* = [x_1^*, \dots, x_{D-1}^*]$ in \mathbb{R}^{D-1} is the vector of coordinates of a composition \mathbf{x} in \mathcal{S}^D .

Perturbation and powering can be easily expressed in terms of coordinates. In fact, let \mathbf{e}_i ($i = 1, 2, \dots, D-1$) be a basis of \mathcal{S}^D and x_i^*, y_i^* ($i = 1, 2, \dots, D-1$) the coordinates of \mathbf{x} and \mathbf{y} , respectively, that is

$$\mathbf{x} = \bigoplus_{i=1}^{D-1} (x_i^* \odot \mathbf{e}_i), \quad \mathbf{y} = \bigoplus_{i=1}^{D-1} (y_i^* \odot \mathbf{e}_i).$$

††† This section is an adaptation of [EP06, p. 148-149] and [Ait03, Section 2.3.4, p. 36-37].

Perturbation and powering are now

$$\mathbf{x} \oplus \mathbf{y} = \bigoplus_{i=1}^{D-1} [(x_i^* + y_i^*) \odot \mathbf{e}_i], \quad a \odot \mathbf{x} = \bigoplus_{i=1}^{D-1} [(a \cdot x_i^*) \odot \mathbf{e}_i],$$

and the coordinates of $\mathbf{x} \oplus \mathbf{y}$ are the ordinary sum $\mathbf{x}^* + \mathbf{y}^*$ of the vectors of coordinates, and the coordinates of $a \odot \mathbf{x}$ is reduced to the ordinary product $a \cdot \mathbf{x}^*$.

Consequently, compositional operations are reduced to ordinary vector operations when representing compositions by their coordinates.

Example 10.


Let $\mathbf{u}_1 = [1, 0, 0, \dots, 0, 0]$, $\mathbf{u}_2 = [0, 1, 0, \dots, 0, 0]$, \dots , $\mathbf{u}_{D-1} = [0, 0, 0, \dots, 0, 1]$ be the canonical basis of \mathbb{R}^{D-1} . The compositions

$$(1.20) \quad \begin{aligned} \mathbf{e}_1 &= \text{alr}^{-1} \mathbf{u}_1 = \mathcal{C}[e, 1, 1, \dots, 1, 1, 1], \\ \mathbf{e}_2 &= \text{alr}^{-1} \mathbf{u}_2 = \mathcal{C}[1, e, 1, \dots, 1, 1, 1], \\ &\dots \\ \mathbf{e}_{D-1} &= \text{alr}^{-1} \mathbf{u}_{D-1} = \mathcal{C}[1, 1, 1, \dots, 1, e, 1], \end{aligned}$$

form a basis of \mathcal{S}^D . Thus the vector \mathbf{x}^* of coordinates of a composition \mathbf{x} with respect to this basis is no other than $\text{alr } \mathbf{x} = [\ln(x_1/x_D), \dots, \ln(x_{D-1}/x_D)]$.


Because we are dealing with the Aitchison geometry, hereafter we will not write the prefix *Aitchison* in expressions, such as Aitchison distance, Aitchison norm or similar, as long as there is no possibility of confusion.

Activities for Section 1.6

 [Click here to get the activities of this section](#)

1.7. Scale invariant logratios or logcontrasts

The properties of a *standard* linear combination of the components of a random vector are commonly used in typical multivariate analysis, with sample space \mathbb{R}^D . For example, among others, techniques involving eigen-analysis and the linear functions in discriminant analysis are based on linear combinations. When the sample space is the simplex, any meaningful function of a random composition should be compatible with the scale invariance requirement. This requirement leads to the use of ratios of parts in our *linear* combinations so that scale constants are cancelled. Furthermore, ratios can be considered in a relative scale and taking their logarithms is then a natural choice. Remember that the analysis of CoDa is essentially based on the statistical analysis of logratios of parts.

 This section is an adaptation of [Ait86, Section 4.10, p. 83-86] and [Ait03, Section 2.5, p. 46-47].

The simplest logratios are those comparing two parts of a composition $\mathbf{x} \in \mathcal{S}^D$ as, that is, $\ln(x_i/x_j)$. More complex linear combinations of logratios can be useful in the analysis, but they must be scale invariant.

A *logcontrast* of parts of a composition \mathbf{x} is a log-linear combination

$$(1.21) \quad \sum_{i=1}^D a_i \ln x_i = \ln \left(\prod_{i=1}^D x_i^{a_i} \right), \quad \text{with } \sum_{i=1}^D a_i = 0,$$

where the condition $\sum_i a_i = 0$ on the real coefficients a_i guarantees scale invariance. This condition ensures that a logcontrast can always be expressed equivalently as a linear combination of simple logratios with a common part divisor, for example as linear combination of *alr*-scores, where $a_1 + \dots + a_{D-1} = -a_D$,

$$\sum_{i=1}^D a_i \ln x_i = a_1 \ln(x_1/x_D) + \dots + a_{D-1} \ln(x_{D-1}/x_D) = a_1 \text{alr}_{1\mathbf{x}} + \dots + a_{D-1} \text{alr}_{D-1\mathbf{x}}$$

or as linear combination of *clr*-scores, with the geometric $g(\mathbf{x})$ as divisor

$$\sum_{i=1}^D a_i \ln x_i = a_1 \ln(x_1/g(\mathbf{x})) + \dots + a_D \ln(x_D/g(\mathbf{x})) = a_1 \text{clr}_{1\mathbf{x}} + \dots + a_{D-1} \text{clr}_{D\mathbf{x}}.$$

In addition, although any log-linear combination $\sum_{i=1}^D a_i \ln x_i$ can be expressed as a logratio

$$\sum_{i=1}^D a_i \ln x_i = \ln \frac{\prod_{a_k > 0} x_k^{a_k}}{\prod_{a_k < 0} x_k^{|a_k|}},$$

only a logcontrast (i.e. $\sum_{i=1}^D a_i = 0$) becomes an scale-invariant logratio, where $\sum_{a_k > 0} a_k = \sum_{a_k < 0} |a_k|$.

Thus, the logcontrasts of parts of a composition in \mathcal{S}^D are the compositional version of standard linear combinations of the components of a vector in \mathbb{R}^D . In particular, the compositional inner product of two compositions \mathbf{x} and \mathbf{y} defined in (1.13) can be expressed equivalently as a logcontrast:

$$(1.22) \quad \langle \mathbf{x}, \mathbf{y} \rangle_a = a_1 \ln x_1 + \dots + a_D \ln x_D,$$

where $a_i = \ln(y_i/g(\mathbf{y}))$ ($i = 1, \dots, D$) and so $a_1 + \dots + a_D = 0$.

Logcontrast emerge naturally in CoDa analysis. Just as linear combinations can be used to define subspaces of the vector space \mathbb{R}^D by way of null spaces or range spaces, so logcontrasts can be used to identify subspaces of the vector space \mathcal{S}^D . Thus, for example, the subspace orthogonal to a given composition $\mathbf{v} \in \mathcal{S}^D$

$$\mathbf{v}^\perp = \{\mathbf{x} \in \mathcal{S}^D : \langle \mathbf{x}, \mathbf{v} \rangle_a = 0\},$$

can be expressed as a logcontrast because it holds that

$$\mathbf{v}^\perp = \{\mathbf{x} \in \mathcal{S}^D : a_1 \ln x_1 + \dots + a_D \ln x_D = 0\},$$

where $a_i = \ln(v_i/g(\mathbf{v}))$ ($i = 1, \dots, D$).

Activities for Section 1.7

🔗 [Click here to get the activities of this section](#)

1.8. Representation of compositions by orthonormal coordinates¹⁸

1.8.1. Orthonormal logratio coordinates (olr). Because \mathcal{S}^D is a vector space of dimension $D - 1$ then $D - 1$ independent vectors in \mathcal{S}^D constitute a basis. If these vectors are unitary and mutually orthogonal they form an orthonormal basis, that is if the compositions $\mathbf{e}_1, \dots, \mathbf{e}_{D-1}$, in \mathcal{S}^D , satisfy

$$(1.23) \quad \|\mathbf{e}_i\|_{\mathbf{a}}^2 = \langle \mathbf{e}_i, \mathbf{e}_i \rangle_{\mathbf{a}} = 1, \quad \langle \mathbf{e}_i, \mathbf{e}_j \rangle_{\mathbf{a}} = 0, \quad i, j = 1, \dots, D - 1, \quad i \neq j,$$

they constitute an orthonormal basis of \mathcal{S}^D . Importantly, Eq. (1.23) is equivalent to

$$(1.24) \quad \|\text{clr } \mathbf{e}_i\|^2 = 1, \quad \langle \text{clr } \mathbf{e}_i, \text{clr } \mathbf{e}_j \rangle = 0, \quad i, j = 1, \dots, D - 1, \quad i \neq j,$$

that is, $\text{clr } \mathbf{e}_1, \dots, \text{clr } \mathbf{e}_{D-1}$ constitute an orthonormal basis of the $(D-1)$ -dimensional subspace U of \mathbb{R}^D ($U = \{\mathbf{z} \in \mathbb{R}^D : z_1 + \dots + z_D = 0\}$) defined in Eq. (1.5). In other words, $\mathbf{e}_1, \dots, \mathbf{e}_{D-1}$ constitute an *orthonormal log-ratio basis* (olr) of \mathcal{S}^D .

Example 11. The *alr*-basis $\mathbf{e}_1, \dots, \mathbf{e}_{D-1}$ defined in Eq. (1.20) is **not** an orthonormal log-ratio basis because

$$\|\mathbf{e}_i\|_{\mathbf{a}}^2 = 1 - \frac{1}{D} \neq 1 \quad (i = 1, \dots, D - 1),$$

and,

$$\langle \mathbf{e}_i, \mathbf{e}_j \rangle_{\mathbf{a}} = -\frac{1}{D} \neq 0 \quad (i, j = 1, \dots, D - 1) \quad (i \neq j).$$

Consequently, the vector $\text{alr } \mathbf{x}$ of *alr*-scores is not a vector of orthonormal log-ratio coordinates.

In real vector spaces, an example of an orthonormal basis is easily found. For instance, in \mathbb{R}^3 , the vectors $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$ form an orthonormal basis which, because of its simplicity, is termed *canonical*. This is not so simple in \mathcal{S}^D . Because

$$(1.25) \quad \begin{aligned} \mathbf{u}_1 &= \left[\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0 \right], \\ \mathbf{u}_2 &= \left[\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{-2}{\sqrt{6}} \right], \end{aligned}$$

is an orthonormal basis in subspace U of \mathbb{R}^3 then an orthonormal log-ratio basis in \mathcal{S}^3 is

$$(1.26) \quad \begin{aligned} \mathbf{e}_1 = \text{clr}^{-1} \mathbf{u}_1 &= \mathcal{C} \left(\exp \left[\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0 \right] \right), \\ \mathbf{e}_2 = \text{clr}^{-1} \mathbf{u}_2 &= \mathcal{C} \left(\exp \left[\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{-2}{\sqrt{6}} \right] \right). \end{aligned}$$

¹⁸This section is an adaptation of [EP06, p. 152–153] and [EB11, 149–151]. Further information in [PET15, Sections 4.1–4.4, p. 32–38].

Although expression (1.26) is not very complicated, it does not exhibit the simplicity of the canonical basis in a real space. Moreover, other possible bases have similar expressions. Hence, there is no reason to refer the compositions in (1.26) as a canonical basis of \mathcal{S}^3 .

A practical approach is to have some simple rules to identify unitary and orthogonal compositions as demanded in equations (1.23). The compositions in the basis (1.26) are expressed using their clr coefficients as defined in (1.6). General characteristics of an orthonormal basis can be seen in (1.25): the squared clr coefficients add to 1; and the ordinary inner product of the clr terms, as real vectors, is null. These properties are general for any orthonormal basis in the simplex, as they are equivalent to conditions in (1.23).

Once an orthonormal log-ratio basis $\mathbf{e}_1, \dots, \mathbf{e}_{D-1}$ of \mathcal{S}^D is given, a composition \mathbf{x} in \mathcal{S}^D can be expressed as a compositional linear combination of the vectors forming the basis. Thus the expression of the *orthonormal log-ratio coordinates* ($\text{olr } \mathbf{x}$) becomes simplified, that is

$$(1.27) \quad \mathbf{x} = \bigoplus_{i=1}^{D-1} (x_i^* \odot \mathbf{e}_i), \quad x_i^* = \langle \mathbf{x}, \mathbf{e}_i \rangle_a.$$

Note that the coordinate x_i^* is the signed-size of the orthogonal projection of the composition \mathbf{x} onto the direction defined by the unitary composition \mathbf{e}_i . The vector $\mathbf{x}^* = \text{olr } \mathbf{x} = (x_1^*, \dots, x_{D-1}^*)$ is the vector of orthonormal log-ratio (*olr*) coordinates or simply the coordinates of \mathbf{x} [Mar19].

All operations and metric relationships between compositions \mathbf{x} and \mathbf{y} in \mathcal{S}^D are translated into ordinary vector operations between the corresponding coordinate vectors \mathbf{x}^* and \mathbf{y}^* in \mathbb{R}^{D-1} . Moreover, if the basis of \mathcal{S}^D is orthonormal, the Aitchison distance becomes equal to

$$d_a(\mathbf{x}, \mathbf{y}) = d(\text{olr } \mathbf{x}, \text{olr } \mathbf{y}) = d(\mathbf{x}^*, \mathbf{y}^*) = \sqrt{\sum_{i=1}^{D-1} (x_i^* - y_i^*)^2},$$

where $d(\cdot, \cdot)$ is the Euclidean distance in real vector spaces.

Analogous results hold for the norm and the inner product, that is all compositional operations are reduced to ordinary vector operations when compositions are represented by their *olr*-coordinates. This is comfortable for working with CoDa, as all known techniques designed for real data hold for their coordinates. Note that this is not true when the coordinates of compositions refer to a non orthonormal basis of \mathcal{S}^D .

1.8.2. The isometric logratio transformation (ilr). Let $\mathbf{e}_1, \dots, \mathbf{e}_{D-1}$ be an *olr*-basis in \mathcal{S}^D . In [EPM03], the function assigning coordinates with respect to $\mathbf{e}_1, \dots, \mathbf{e}_{D-1}$ to a compositions $\mathbf{x} \in \mathcal{S}^D$ is called the *isometric logratio transformation* $\text{ilr} : \mathcal{S}^D \rightarrow \mathbb{R}^{D-1}$,

$$\text{ilr } \mathbf{x} = [\langle \mathbf{x}, \mathbf{e}_1 \rangle_a, \dots, \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a].$$

The word *isometric* in *ilr* refers to the distance preservation. In [Mar19] was introduced the name *olr* to avoid confusion because the *clr* transformation is also an isometric log-ratio transformation.

Some mathematical properties

- Let $\mathcal{B} = \{\mathbf{e}_1, \dots, \mathbf{e}_{D-1}\}$ be a *olr*-basis of the vector space \mathcal{S}^D . Let Φ be the $(D-1) \times D$ -matrix whose i -th row is the vector $\text{clr } \mathbf{e}_i$, for $i = 1, \dots, D-1$.

The matrix Φ satisfies the following properties:

- $\Phi \cdot \Phi^t = \mathbf{I}_{D-1}$, where \mathbf{I}_{D-1} is the identity matrix of dimension $D-1$.
- $\Phi^t \cdot \Phi = \mathbf{I}_D - \frac{1}{D} \mathbf{1}_D^t \cdot \mathbf{1}_D$, with $\mathbf{1}_D$ a D -row-vector of ones.

Moreover, the matrix Φ plays a decisive role relating a composition to its *olr*-coordinates.

- The coordinates of a composition $\mathbf{x} \in \mathcal{S}^D$ with respect to the basis \mathcal{B} , that is the *olr* \mathbf{x} vector, can be expressed from the *clr* \mathbf{x} vector as

$$(1.28) \quad \text{olr } \mathbf{x} = \mathbf{x}^* = (\text{clr } \mathbf{x}) \cdot \Phi^t = (\ln \mathbf{x}) \cdot \Phi^t .$$

- Let $\mathbf{x}^* = \text{olr } \mathbf{x}$ be the coordinates of a composition \mathbf{x} with respect to the basis \mathcal{B} . The composition \mathbf{x} can be recovered from its coordinates \mathbf{x}^* since it holds that

$$(1.29) \quad \text{clr } \mathbf{x} = \mathbf{x}^* \cdot \Phi \quad \text{and therefore} \quad \mathbf{x} = \text{olr}^{-1} \mathbf{x}^* = \mathcal{C}(\exp(\mathbf{x}^* \cdot \Phi)) .$$

Therefore the *olr* transformation is one-to-one from \mathcal{S}^D to \mathbb{R}^{D-1} . The inverse transformation $\text{olr}^{-1} : \mathbb{R}^{D-1} \rightarrow \mathcal{S}^D$ takes the form

$$\text{olr}^{-1} \mathbf{v} = \mathcal{C}(\exp(\mathbf{v} \cdot \Phi)) .$$


- From the identities (1.7), (1.28) and (1.29), for any $\mathbf{x} \in \mathcal{S}^D$ it holds that

$$(1.30) \quad \text{olr } \mathbf{x} = (\text{alr } \mathbf{x}) \cdot \mathbf{K}^t \cdot \Phi^t \quad \text{alr } \mathbf{x} = (\text{olr } \mathbf{x}) \cdot \Phi \cdot \mathbf{F}^t ,$$

\mathbf{K} and \mathbf{F} being the matrices defined in (1.8) and (1.9), respectively.

Therefore, if we know any of the two vectors of coordinates —*alr* \mathbf{x} or *olr* \mathbf{x} — we can easily calculate the other vector.

Activities for Section 1.8

 [Click here to get the activities of this section](#)

1.9. *olr*-basis associated to a sequential binary partition²⁰

There are several ways to define *olr*-bases in the simplex. The main criterion of selection of compositions to be included in the basis is that it enhances the

²⁰This section is an adaptation of [EP06, p. 152–153]. Further information in [PET15, Section 4.5, p. 32–42]

interpretability of the representation in coordinates. For instance, when performing principal components analysis, an *olr*-basis is created so that the first *olr*-coordinate (first principal component: PC_1) represents the direction of maximum variability.

1.9.1. Sequential binary partition. Particular cases of orthonormal bases are those linked to a *sequential binary partition* (SBP) of the parts of the compositional vector [EP05]. A SBP is a hierarchy selection of the parts of a composition. Each step of the partition, of a total of $D - 1$ steps, gives rise to a vector of the *olr*-basis. In a first step, SBP consists of dividing the composition into two groups of parts which are indicated by +1 and -1, respectively, as shown in the first row of *sign matrix* \mathbf{S} . In further steps, each previously obtained group of parts is again subdivided into two groups until all groups are made of a single part. Thus we obtain a $(D - 1) \times D$ dimensional sign matrix $\mathbf{S} = [s_{ij}]$ with +1, -1 and 0 entries (0's correspond to parts not included in the partition).

The matrix \mathbf{S} serves to build the $(D - 1) \times D$ dimensional matrix Φ of the *clr*-transformed vectors of the *olr*-basis $\mathbf{e}_1, \dots, \mathbf{e}_{D-1}$ associated with the SBP. The ϕ_{ij} entry of Φ is defined as follows:

$$(1.31) \quad \begin{aligned} \phi_{ij} &= 0, & \text{if } s_{ij} &= 0; \\ \phi_{ij} &= +\frac{1}{p_i} \sqrt{\frac{p_i \cdot n_i}{p_i + n_i}}, & \text{if } s_{ij} &> 0; \\ \phi_{ij} &= -\frac{1}{n_i} \sqrt{\frac{p_i \cdot n_i}{p_i + n_i}}, & \text{if } s_{ij} &< 0, \end{aligned}$$

where p_i and n_i are the number of parts in the i -th row of \mathbf{S} coded by +1 (*positive*) and -1 (*negative*), which respectively will be in the numerator and denominator of the corresponding logratio.

Example 12. An SBP of parts in S^6 to build an *olr*-basis.

Sign matrix \mathbf{S} of the SBP								
i	x_1	x_2	x_3	x_4	x_5	x_6	p	n
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

Matrix Φ of the <i>clr</i> -transformed vectors of the <i>olr</i> -basis defined by \mathbf{S}						
i	x_1	x_2	x_3	x_4	x_5	x_6
1	$+\frac{1}{4}\sqrt{\frac{4 \cdot 2}{4+2}}$	$+\frac{1}{4}\sqrt{\frac{4 \cdot 2}{4+2}}$	$-\frac{1}{2}\sqrt{\frac{4 \cdot 2}{4+2}}$	$-\frac{1}{2}\sqrt{\frac{4 \cdot 2}{4+2}}$	$+\frac{1}{4}\sqrt{\frac{4 \cdot 2}{4+2}}$	$+\frac{1}{4}\sqrt{\frac{4 \cdot 2}{4+2}}$
2	$+\frac{1}{1}\sqrt{\frac{1 \cdot 3}{1+3}}$	$-\frac{1}{3}\sqrt{\frac{1 \cdot 3}{1+3}}$	0	0	$-\frac{1}{3}\sqrt{\frac{1 \cdot 3}{1+3}}$	$-\frac{1}{3}\sqrt{\frac{1 \cdot 3}{1+3}}$
3	0	$+\frac{1}{1}\sqrt{\frac{1 \cdot 2}{1+2}}$	0	0	$-\frac{1}{2}\sqrt{\frac{1 \cdot 2}{1+2}}$	$-\frac{1}{2}\sqrt{\frac{1 \cdot 2}{1+2}}$
4	0	0	0	0	$+\frac{1}{1}\sqrt{\frac{1 \cdot 1}{1+1}}$	$-\frac{1}{1}\sqrt{\frac{1 \cdot 1}{1+1}}$
5	0	0	$+\frac{1}{1}\sqrt{\frac{1 \cdot 1}{1+1}}$	$-\frac{1}{1}\sqrt{\frac{1 \cdot 1}{1+1}}$	0	0

Thus, $\text{clr } \mathbf{e}_i = \phi_i = [\phi_{i1}, \dots, \phi_{iD}]$ ($i = 1, \dots, D-1$). It is easy to prove that the $D-1$ vectors $\text{clr } \mathbf{e}_1, \dots, \text{clr } \mathbf{e}_{D-1}$ (the rows of Φ) are unitary and two to two orthogonal in \mathbb{R}^D . Moreover, these vectors belong to the $(D-1)$ -dimensional subspace U of \mathbb{R}^D defined in 1.5. Therefore, they constitute an orthonormal basis of U . Hence, the compositions $\mathbf{e}_1, \dots, \mathbf{e}_{D-1}$ are an *olr*-basis of \mathcal{S}^D . Moreover, the i -th coordinate x_i^* of a composition \mathbf{x} with respect to this basis is equal to the logcontrast

$$(1.32) \quad x_i^* = \sum_{j=1}^D \phi_{ij} \ln x_j.$$

According to equation (1.31), this logcontrast can be rewritten equivalently as

$$(1.33) \quad x_i^* = \sqrt{\frac{p_i \cdot n_i}{p_i + n_i}} \ln \frac{(x_{k_1} \cdots x_{k_{p_i}})^{1/p_i}}{(x_{l_1} \cdots x_{l_{n_i}})^{1/n_i}} = \ln \frac{(x_{k_1} \cdots x_{k_{p_i}})^{a_i^+}}{(x_{l_1} \cdots x_{l_{n_i}})^{a_i^-}},$$

where k_1, \dots, k_{p_i} are the labels of parts in the numerator (coded by $+1$ in the i -th row of \mathbf{S}); l_1, \dots, l_{n_i} are the labels of parts in the denominator (coded by -1 in the same row); and

$$a_i^+ = \frac{1}{p_i} \sqrt{\frac{p_i \cdot n_i}{p_i + n_i}} \quad \text{and} \quad a_i^- = \frac{1}{n_i} \sqrt{\frac{p_i \cdot n_i}{p_i + n_i}}.$$

1.9.2. Balances. The i -th *olr*-coordinate x_i^* of a composition \mathbf{x} with respect to a basis linked to an SBP is

$$(1.34) \quad x_i^* = \sqrt{\frac{p_i \cdot n_i}{p_i + n_i}} \ln \frac{(x_{k_1} \cdots x_{k_{p_i}})^{1/p_i}}{(x_{l_1} \cdots x_{l_{n_i}})^{1/n_i}},$$

where k_1, \dots, k_{p_i} are the labels of parts in the numerator (coded by $+1$ in the i -th row of \mathbf{S}) and l_1, \dots, l_{n_i} are the labels of parts in the denominator (coded by -1 in the same row).

According to Eq. (1.34) the i -th coordinate x_i^* of \mathbf{x} can be interpreted as a *normalised balance* (the square root is the *normalising factor*) between the logratio of the geometric mean of parts coded by $+1$ in the i -th row of \mathbf{S} and the geometric mean of parts coded by -1 in the same row. In general, we refer to these coordinates as *balances*.

Focus your attention again on the Eq. (1.34): each one of the two geometric means represents the central value of the parts in its coded group; the ratio between the two geometric means measures the relative weight of each group in the composition; the logarithm provides the appropriate scale; and the square root coefficient is a normalising constant which allows different balances to be compared numerically. A positive balance means that, in (geometric) mean, the group of parts in the numerator (coded by $+1$) has more weight in the composition than the group in the denominator (coded by -1). And conversely for negative balances.

Example 12 (cont.). According to the sign matrix \mathbf{S} of the SBP of this example, the first coordinate x_1^* gives the balance between the subcompositions $\text{sub}(\mathbf{x}; 1, 2, 5, 6)$ and $\text{sub}(\mathbf{x}; 3, 4)$ of \mathbf{x} . Coordinates x_2^*, x_3^* and x_4^* are exclusively associated to $\text{sub}(\mathbf{x}; 1, 2, 5, 6)$, whereas coordinate x_5^* is associated to $\text{sub}(\mathbf{x}; 3, 4)$. Thus, x_2^* gives the balance between the first part and the group of parts $\{2, 5, 6\}$ in $\text{sub}(\mathbf{x}; 1, 2, 5, 6)$, whereas x_5^* gives the balance between the third and fourth parts in $\text{sub}(\mathbf{x}; 3, 4)$.

Sometimes we need to calculate the *olr*-coordinates of a set of compositions relative to any orthonormal basis of \mathcal{S}^D , without regard for a particular basis. If so, we can choose the *Default Partition* option which CoDaPack offers in the menus involving *olr*-coordinates. Depending on the dimension D , the default *olr*-basis chosen by CoDaPack is:

$$D = 2$$

i	x_1	x_2	<i>olr</i> -coordinates
1	+1	-1	$x_1^* = \sqrt{\frac{1}{2}} \ln \frac{x_1}{x_2}$

$$D = 3$$

i	x_1	x_2	x_3	<i>olr</i> -coordinates
1	+1	+1	-1	$x_1^* = \sqrt{\frac{2}{3}} \ln \frac{(x_1 \cdot x_2)^{1/2}}{x_3}$
2	+1	-1	0	$x_2^* = \sqrt{\frac{1}{2}} \ln \frac{x_1}{x_2}$

$$D = 4$$

i	x_1	x_2	x_3	x_4	<i>olr</i> -coordinates
1	+1	+1	-1	-1	$x_1^* = \ln \frac{(x_1 \cdot x_2)^{1/2}}{(x_3 \cdot x_4)^{1/2}}$
2	+1	-1	0	0	$x_2^* = \sqrt{\frac{1}{2}} \ln \frac{x_1}{x_2}$
3	0	0	+1	-1	$x_3^* = \sqrt{\frac{1}{2}} \ln \frac{x_3}{x_4}$

$$D = 5$$

i	x_1	x_2	x_3	x_4	x_5	<i>olr</i> -coordinates
1	+1	+1	+1	-1	-1	$x_1^* = \sqrt{\frac{6}{5}} \ln \frac{(x_1 \cdot x_2 \cdot x_3)^{1/3}}{(x_4 \cdot x_5)^{1/2}}$
2	+1	+1	-1	0	0	$x_2^* = \sqrt{\frac{2}{3}} \ln \frac{(x_1 \cdot x_2)^{1/2}}{x_3}$
3	+1	-1	0	0	0	$x_3^* = \sqrt{\frac{1}{2}} \ln \frac{x_1}{x_2}$
4	0	0	0	+1	-1	$x_4^* = \sqrt{\frac{1}{2}} \ln \frac{x_4}{x_5}$

$$D = 6$$

i	x_1	x_2	x_3	x_4	x_5	x_6	<i>olr</i> -coordinates
1	+1	+1	+1	-1	-1	-1	$x_1^* = \sqrt{\frac{3}{2}} \ln \frac{(x_1 \cdot x_2 \cdot x_3)^{1/3}}{(x_4 \cdot x_5 \cdot x_6)^{1/3}}$
2	+1	+1	-1	0	0	0	$x_2^* = \sqrt{\frac{2}{3}} \ln \frac{(x_1 \cdot x_2)^{1/2}}{x_3}$
3	+1	-1	0	0	0	0	$x_3^* = \sqrt{\frac{1}{2}} \ln \frac{x_1}{x_2}$
4	0	0	0	+1	+1	-1	$x_4^* = \sqrt{\frac{2}{3}} \ln \frac{(x_4 \cdot x_5)^{1/2}}{x_6}$
5	0	0	0	+1	-1	0	$x_5^* = \sqrt{\frac{1}{2}} \ln \frac{x_4}{x_5}$

$$D = 7$$

i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	<i>olr</i> -coordinates
1	+1	+1	+1	+1	-1	-1	-1	$x_1^* = \sqrt{\frac{12}{7}} \ln \frac{(x_1 \cdot x_2 \cdot x_3 \cdot x_4)^{1/4}}{(x_5 \cdot x_6 \cdot x_7)^{1/3}}$
2	+1	+1	-1	-1	0	0	0	$x_2^* = \ln \frac{(x_1 \cdot x_2)^{1/2}}{(x_3 \cdot x_4)^{1/2}}$
3	+1	-1	0	0	0	0	0	$x_3^* = \sqrt{\frac{1}{2}} \ln \frac{x_1}{x_2}$
4	0	0	+1	-1	0	0	0	$x_4^* = \sqrt{\frac{1}{2}} \ln \frac{x_3}{x_4}$
5	0	0	0	0	+1	+1	-1	$x_5^* = \sqrt{\frac{2}{3}} \ln \frac{(x_5 \cdot x_6)^{1/2}}{x_7}$
6	0	0	0	0	+1	-1	0	$x_6^* = \sqrt{\frac{1}{2}} \ln \frac{x_5}{x_6}$

In contrast to other spaces, in a CoDa space it is not possible to define a canonical basis. Because of this, different types of compositional basis are defined to meet certain properties. In this section we define an special case of basis, the *Pivot coordinates* [FH11], which are based on SBP, and therefore, their coordinates are balances. Pivot coordinates are constructed using an SBP such that the first binary partition separates one component (called pivot part) from the rest of parts. Then, using the non-pivot components new binary partitions are constructed by splitting a new pivot part from the rest of non-pivot components. Without loss of generality, one can assume that the first pivot part is x_1 , followed by x_2 and continuing up to x_{D-1} . If this is not the case, then one can accordingly order the parts as $x_{[1]}, \dots, x_{[D]}$ in advance to create pivot coordinates.

The i -th pivot coordinate is defined as

$$(1.35) \quad x_i^* = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{(x_{i+1} \cdots x_D)^{D-i}}, \quad i = 1, \dots, D-1.$$

Using this basis, we keep all relative information coming from the first pivot against the rest of variables in the first coordinate. In addition, the first pivot is proportional to the corresponding *clr*-score, that is, it holds


$$x_1^* = \sqrt{\frac{D-1}{D}} \ln \frac{x_1}{(x_2 \cdots x_D)^{D-1}} = \sqrt{\frac{D}{D-1}} \ln \frac{x_1}{(x_1 \cdots x_D)^D} = \text{clr}_1(\mathbf{x}).$$

For this reason, given two different ordinations of the parts, the Pearson correlation coefficient of the two first pivot coordinates is spurious because it is equivalent to the spurious correlation coefficient of the two counterparts *clr*-variables (see Section 1.4.2).

Example 13. Pivot coordinates in \mathcal{S}^6 for creating an *olr*-basis.

$D = 6$							Pivot coordinates
i	x_1	x_2	x_3	x_4	x_5	x_6	
1	+1	-1	-1	-1	-1	-1	$x_1^* = \sqrt{\frac{5}{6}} \ln \frac{x_1}{(x_2 \cdots x_6)^{1/5}}$
2	0	+1	-1	-1	-1	-1	$x_2^* = \sqrt{\frac{4}{5}} \ln \frac{x_2}{(x_3 \cdots x_6)^{1/4}}$
3	0	0	+1	-1	-1	-1	$x_3^* = \sqrt{\frac{3}{4}} \ln \frac{x_3}{(x_4 \cdots x_6)^{1/3}}$
4	0	0	0	+1	-1	-1	$x_4^* = \sqrt{\frac{2}{3}} \ln \frac{x_4}{(x_5 \cdot x_6)^{1/2}}$
5	0	0	0	0	+1	-1	$x_5^* = \sqrt{\frac{1}{2}} \ln \frac{x_5}{x_6}$

Activities for Section 1.9

 [Click here to get the activities of this section](#)

The chapter's key concepts

- ✓ CoDa provide only relative information between their components.
- ✓ Any meaningful function of a composition can be expressed in terms of ratios of their components, that is the function must be scale-invariant.
- ✓ Any analysis involving CoDa must be subcompositionally coherent.

- ✓ The simplex is the sample space of CoDa.
- ✓ Perturbation and powering are the basic operations in the simplex.
- ✓ The analysis of CoDa is based on the logratios of their components.
- ✓ The simplex \mathcal{S}^D can be viewed as a real vector space of dimension $D - 1$, perturbation being the internal operation and powering the external one.
- ✓ A logcontrast is any real scale invariant log-ratio function of the components of a composition.
- ✓ The *alr* and the *clr* transformations are linear maps from the vector space \mathcal{S}^D to \mathbb{R}^{D-1} and \mathbb{R}^D , respectively. These transformations and their inverses let us operate easily with compositions.
- ✓ The simplex \mathcal{S}^D is also a Euclidean space because the *clr* transformation allows the Euclidean metric of the real space to be exported to \mathcal{S}^D .
- ✓ Given a composition $\mathbf{x} \in \mathcal{S}^D$, the components in \mathbb{R}^{D-1} of the real vector *alr* \mathbf{x} can be interpreted as the coordinates of \mathbf{x} in a (not orthonormal) basis of \mathcal{S}^D .
- ✓ Given an *olr*-basis \mathcal{B} of \mathcal{S}^D , the isometric logratio transformation *ilr* (associated to \mathcal{B}) send each composition \mathbf{x} to vector \mathbf{x}^* in \mathbb{R}^{D-1} whose components are the coordinates of \mathbf{x} in relation to the basis \mathcal{B} .
- ✓ The SBP is a procedure to easily obtain the *olr*-basis of \mathcal{S}^D . In this case, the coordinates of a composition (relative to one of these bases) can be easily interpreted as *balances* between the geometric means of two subsets of parts.

Specific references in Chapter 1

- [Ait82] J. Aitchison, *The statistical analysis of compositional data (with discussion)*, J. R. Statist. Soc. B **44** (1982), no. 2, 139–177.
- [Ait83] J. Aitchison, *Principal component analysis of compositional data*, Biometrika **70** (1983), no. 1, 57–65.
- [Ait85] J. Aitchison, *A general class of distributions on the simplex*, J. R. Statist. Soc. B **47** (1985), no. 1, 136–146.
- [Ait03] J. Aitchison, *A concise guide to compositional data analysis*, Available from <http://www.compositionaldata.com>, 2003.
- [AiS80] J. Aitchison and S.M. Shen, *Logisticnormal distributions: Some properties and uses*, Biometrika **67** (1980), no. 2, 261–271.
- [Bar00] C. Barceló-Vidal, *Fonamentació Matemàtica de l'Anàlisi de Dades Composicionals*, Departament d'Informàtica i Matemàtica Aplicada. Universitat de Girona, 2000.
- [BM16] C. Barceló-Vidal and J. A. Martín-Fernández *The mathematics of compositional analysis*, Austrian Journal of Statistics. **45** (2016), 57–71.
- [BMP03] C. Barceló-Vidal, J. A. Martín-Fernández, and V. Pawlowsky-Glahn, *Mathematical foundations of compositional data analysis*, Proceedings of IAMG'01 —

- The sixth annual conference of the International Association for Mathematical Geology (G. Ross, ed.), vol. CD-ROM, 2003, p. 20.
- [But79] J. C. Butler, *The effect of closure on the measure of similarity between samples*, J. Math. Geol. **11** (1979), 73–84.
 - [Cha56] F. Chayes, *Petrographic modal analysis*, Wiley, New York (USA), 1956.
 - [Cha60] F. Chayes, *On correlation between variables of constant sum*, J. Geophys. Res. **65** (1960), 4185–4193.
 - [Cha62] F. Chayes, *Numerical correlation and petrographic variation*, Journal of Geology **70** (1962), 440–452.
 - [Cha71] F. Chayes, *Ratio correlation: A manual for students of petrology and geochemistry*, University of Chicago Press, Chicago, London, 1971.
 - [Cha72] F. Chayes, *Effect of the proportion transformation on central tendency*, J. Math. Geol. **4** (1972), 269–70.
 - [ChK66] F. Chayes and W. Kruskal, *An approximate statistical test for correlation between proportions*, J. Geology, **74** (1966), 692–702.
 - [ChT78] F. Chayes and J. Trochimczyk, *The effect of closure on the structure of principal components*, J. Math. Geol. **10** (1978), 323–333.
 - [Cha86] T.C. Chang, *Spherical regression*, Ann. Statist. **14** (1986), no. 3, 907–924.
 - [DaJ74] J.N. Darroch and I.R. James, *F-independence and null correlations of bounded sum positive variables*, J. R. Statist. Soc. B **36** (1974), 247–52.
 - [DaR70] J. N. Darroch and D. Ratcliff, *Null correlations for proportions II*, J. Math. Geol. **2** (1974), 307–312.
 - [DaR78] J. N. Darroch and D. Ratcliff, *No association of proportions*, J. Math. Geol. **10** (1978), 361–368.
 - [EPM03] J.J. Egozcue, V. Pawlowsky-Glahn and G. Mateu-Figueras, *Isometric Logratio Transformations for Compositional Data Analysis*, Mathematical Geology **35** (2003), 279–300.
 - [EP05] J.J. Egozcue and V. Pawlowsky-Glahn, *Groups of parts and their balances in compositional data analysis*, Mathematical Geology **37** (2005), no. 7, 795–828.
 - [EP06] J.J. Egozcue and V. Pawlowsky-Glahn, *Compositional data in the geosciences*, vol. 264, ch. Simplicial geometry for compositional data, pp. 145–159, Geological Society, London, 2006.
 - [EB11] J.J. Egozcue, C. Barceló-Vidal, J.A. Martín-Fernández, E. Jarauta-Bragulat, J.L. Díaz-Barrero and G. Mateu-Figueras, *Compositional Data Analysis: Theory and Applications*, ch. Elements of simplicial linear algebra and geometry, pp. 141–157, John Wiley & Sons, Ltd, Chichester, UK, 2011. DOI:10.1002/9781119976462.ch 11
 - [EP11] J.J. Egozcue and V. Pawlowsky-Glahn, *Compositional data analysis: Theory and applications*, ch. Basic concepts and procedures, pp. 12–27, John Wiley & Sons, Ltd, Chichester (UK), 2011.
 - [FH11] E. Fišerová and K. Hron, *On the interpretation of orthonormal coordinates for compositional data*, Mathematical Geosciences **43** (2011), no. 4, 455–468.
 - [Kru62] C. Krumbein, *Open and closed number systems in stratigraphic mapping*, AAPG Bulletin **46** (1962), no. 12, 2229–2245.

-
- [Mar19] J.A. Martín-Fernández, *Comments on: Compositional data: the sample space and its structure by Egozcue, J.J. and Pawlowsky-Glahn, V.*, TEST **28** (2019), no. 3, 653–657.
- [Mos62] J.E. Mosimann, *On the compound multinomial distribution, the multivariate β -distribution, and correlations among proportions*, Biometrika **49** (1962), no. 1–2, 65–82.
- [Mos63] J.E. Mosimann, *On the compound negative binomial distribution and correlations among inversely sampled pollen counts*, Biometrika **50** (1963), no. 1, 47–54.
- [Pea97] K. Pearson, *Mathematical contributions to the theory of evolution. on a form of spurious correlation which may arise when indices are used in the measurements of organs*, Proc. R. Soc. **60** (1897), 489–498.
- [SV59] O.V. Sarmanov and A.B. Vistelius, *On the correlation of percentage values*, Doklady Akademii Nauk. SSSR **126** (1959), 22–25.

Exploratory analysis

Contents

- 2.1 Centre of a compositional data set
 - 2.1.1 Centre
 - 2.1.2 Centring a compositional data set
- 2.2 Covariance structure of a compositional data set
 - 2.2.1 Variation matrix and variation array
 - 2.2.2 Centred logratio covariance matrix
 - 2.2.3 Additive logratio covariance matrix
 - 2.2.4 *olr*-covariance matrix
 - 2.2.5 Total variance
 - 2.2.6 Scaling a compositional data set
- 2.3 CoDa-dendrogram
- 2.4 Reduced-dimensionality representation of a compositional data set: clr-biplot
 - 2.4.1 Singular value decomposition
 - 2.4.2 Logcontrast principal components analysis
 - 2.4.3 Compositional biplot
 - 2.4.4 Interpretation of a compositional biplot
 - 2.4.5 Subcompositional analysis
- 2.5 Principal balances
- 2.6 Distributions on the simplex
 - 2.6.1 Most relevant distributions
 - 2.6.2 Normality tests
 - 2.6.2.1 Radius tests
 - 2.6.2.2 Marginal tests

Objectives

- ✓ To present the assumptions, principles, and techniques necessary to gain insight into CoDa via exploratory data analysis (EDA).
- ✓ To analyse the peculiarities of the reduced-dimensionality representation of a CoDa set.
- ✓ To show a procedure for creating an SBP according the criterion of maximizing the proportion of total variability retained by the balances.
- ✓ To introduce the most important probability distributions models on the simplex.

2.1. Centre of a compositional data set [†]

Standard descriptive statistics are not very informative in the case of CoDa. In particular, the arithmetic mean and the variance or standard deviation of individual components do not fit in with the compositional geometry as measures of central tendency and dispersion. They were defined as such in the framework of Euclidean geometry in real space, which is not a sensible geometry for CoDa. Therefore, it is necessary to introduce alternatives, which we find in the concepts of *centre*, *variation matrix* and *total variance*.

Let

$$(2.1) \quad \mathbf{X} = \{\mathbf{x}_i = [x_{i1}, \dots, x_{iD}] \in \mathcal{S}^D : i = 1, \dots, n\}$$

be a CoDa set of size n . The n rows $\mathbf{x}_1, \dots, \mathbf{x}_n$ of the matrix \mathbf{x} correspond to samples, and the D columns X_1, \dots, X_D correspond to parts of CoDa.

2.1.1. Centre. A measure of central tendency for the CoDa set \mathbf{X} is the closed geometric mean which is called the *centre* and is defined as

$$(2.2) \quad \mathbf{g} = \mathcal{C}[g_1, \dots, g_D], \text{ with } g_j = \left(\prod_{i=1}^n x_{ij} \right)^{1/n}, \quad j = 1, \dots, D.$$

where \mathcal{C} is the closure operator to a constant κ .

Note that in the definition of the centre of a compositional data set the geometric mean is considered column-wise (i.e. by variables), whereas in the *clr* transformation,

$$\text{clr } \mathbf{x} = \left[\ln \frac{x_1}{g(\mathbf{x})}, \dots, \ln \frac{x_D}{g(\mathbf{x})} \right],$$

the geometric mean $g(\mathbf{x}) = \left(\prod_{j=1}^D x_j \right)^{1/D}$ is considered row-wise (i.e. by samples).

It is easy to prove that the centre \mathbf{g} can be calculated from the arithmetic mean of the *clr*-scores set $\mathbf{Z} = \text{clr } \mathbf{x}$, where *clr* is applied row-wise. More precisely,

$$(2.3) \quad \mathbf{g} = \text{clr}^{-1}[\bar{Z}_1, \dots, \bar{Z}_D] = \mathcal{C}[\exp \bar{Z}_1, \dots, \exp \bar{Z}_D],$$

[†]This section is an adaptation of [DBB06, p. 161–163], [Ait86, Sections 4.1–4.9, p. 64–83]. Further information in [PET15, Sections 5.2–5.3, p. 66–69].

with

$$\bar{Z}_j = \frac{1}{n} \sum_{i=1}^n \ln \frac{x_{ij}}{g(\mathbf{x}_i)}, \quad j = 1, \dots, D.$$

Similarly, \mathbf{g} can be calculated from the arithmetic mean of the row-wise *alr*-transformed data set $\mathbf{Y} = \text{alr } \mathbf{X}$, that is,

$$(2.4) \quad \mathbf{g} = \text{alr}^{-1}[\bar{Y}_1, \dots, \bar{Y}_{D-1}] = \mathcal{C}[\exp \bar{Y}_1, \dots, \exp \bar{Y}_{D-1}, 1],$$

with

$$\bar{Y}_j = \frac{1}{n} \sum_{i=1}^n \ln \frac{x_{ij}}{x_D}, \quad j = 1, \dots, D-1.$$

In addition, \mathbf{g} can also be calculated from the arithmetic mean of the row-wise *olr*-transformed data set $\mathbf{X}^* = \text{olr } \mathbf{X}$, that is, $\mathbf{g} = \text{olr}^{-1}(\bar{\mathbf{X}}^*)$, where $\bar{\mathbf{X}}^*$ is the arithmetic mean vector of the *olr*-coordinates set.

Indeed, let $\mathcal{B} = \{\mathbf{e}_1, \dots, \mathbf{e}_{D-1}\}$ be an *olr*-basis of the vector space \mathcal{S}^D , and Φ be the $(D-1) \times D$ -matrix whose i -th row is the vector $\text{clr } \mathbf{e}_i$, for $i = 1, \dots, D-1$.

The centre \mathbf{g} of \mathbf{X} can also be calculated from the arithmetic mean of the *olr*-transformed data (with respect to basis \mathcal{B}) $\mathbf{X}^* = \text{olr } \mathbf{X}$ since it holds that

$$\mathbf{g} = \text{olr}^{-1}[\bar{X}_1^*, \dots, \bar{X}_{D-1}^*] = \mathcal{C}(\exp([\bar{X}_1^*, \dots, \bar{X}_{D-1}^*] \cdot \Phi)) ,$$

with

$$\bar{X}_j^* = \frac{1}{n} \sum_{i=1}^n x_{ij}^*, \quad j = 1, \dots, D-1.$$

Samples $\mathbf{x}_i = [x_{i1}, \dots, x_{iD}]$, $i = 1, \dots, n$ can be originally closed to a scale constant κ (i.e. $\sum_{j=1}^D x_{ij} = \kappa$, $i = 1, \dots, n$) or they can be non-closed. Note that the geometric mean vector $[g_1, \dots, g_D]$ is not closed and it is expressed in terms of original units of compositional set \mathbf{X} [Mar+20]. Once one applies the closure to the geometric mean vector, the centre $\mathbf{g} = \mathcal{C}[g_1, \dots, g_D]$ of the compositional set \mathbf{X} expresses information about the ratios of parts of the centre. The centre \mathbf{g} is in sharp contrast to what is almost universally quoted in raw CoDa analysis, namely the arithmetic mean vector of the CoDa set

$$\mathbf{m} = [m_1, \dots, m_D], \quad \text{with } m_j = \sum_{i=1}^n x_{ij}, \quad j = 1, \dots, D.$$

The arithmetic average \mathbf{m} of compositions is plotted in Figure 2.1, and, compared to centre \mathbf{g} , is more like an outlier than a central characteristic of a data set.

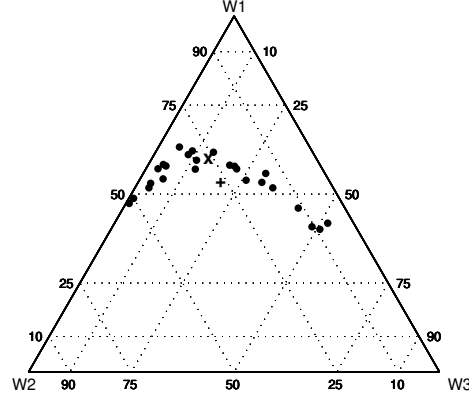


Figure 2.1. Centre of a CoDa set in S^3 : symbol “ \times ” ($= [0.606, 0.272, 0.122]$) indicates the position of the (geometric) centre \mathbf{g} ; symbol “ $+$ ” ($= [0.443, 0.229, 0.148]$) indicates the position of the typical centre \mathbf{m} (arithmetic average of the compositions).

Example 1.

The **statisticiantimebudget** CoDa set \mathbf{X} (see [Appendix](#)) records the daily time (in hours) devoted to six (D) daily activities undertaken by an academic statistician: teaching (T); consultation (C); administration (A); research (R); other wakeful activities (O); and sleep (S). The data were recorded on each of 20 (n) days, selected randomly from working days.

The table shows the centre of the CoDa set (\mathbf{X}) expressed in raw units (\mathbf{g}), in *alr*-coordinates ($\hat{\boldsymbol{\mu}}_{\text{alr}}$), in *clr*-scores ($\hat{\boldsymbol{\mu}}_{\text{clr}}$), and in *olr*-coordinates ($\hat{\boldsymbol{\mu}}_{\text{olr}}$).

Raw data (%)	T	C	A	R	O	S
\mathbf{g}	14.75	10.49	12.43	11.29	23.16	27.88
<i>alr</i>	$\ln(T/S)$	$\ln(C/S)$	$\ln(A/S)$	$\ln(R/S)$	$\ln(O/S)$	
$\hat{\boldsymbol{\mu}}_{\text{alr}}$	-0.6367	-0.9779	-0.8078	-0.9042	-0.1856	
<i>clr</i>	clr_1	clr_2	clr_3	clr_4	clr_5	clr_6
$\hat{\boldsymbol{\mu}}_{\text{clr}}$	-0.0513	-0.3925	-0.2225	-0.3188	0.3997	0.5854
<i>olr</i>	x_1^*	x_2^*	x_3^*	x_4^*	x_5^*	
$\hat{\boldsymbol{\mu}}_{\text{olr}}$	-0.8531	0.1224	0.1891	-0.1203	-0.1313	

The centre expressed in raw units (%) indicates that the largest proportions are in the ‘leisure’ (O and S). Note that the reference is $\frac{1}{D} = 1/6 \approx 16.66\%$. The *alr*-coordinates of the centre are all negative because the largest part (sleep) is used as

denominator. This is not the case of the *clr*-scores, where the scores of the ‘leisure’ are positive. The SBP used for the *olr*-coordinates is

i	T	C	A	R	O	S	p	n
1	-1	-1	-1	-1	+1	+1	2	4
2	+1	-1	-1	+1	0	0	2	2
3	+1	0	0	-1	0	0	1	1
4	0	+1	-1	0	0	0	1	1
5	0	0	0	0	+1	-1	1	1

The sign ‘-’ of x_1^* in $\hat{\boldsymbol{\mu}}_{\text{olr}}$ indicates that, in average, the time dedicated to each ‘leisure’ activity (O, S) is greater than the time spent in each ‘work’ activity (T, C, A, R). The other *olr*-coordinates can be analogously interpreted. For example, the sign ‘+’ of x_2^* suggests that, in average, the time for the activities T or R are greater than the time in the activities C or A .

2.1.2. Centring a compositional data set. A standard way to visualise data in a ternary diagram is to rescale the observations in such a way that their range is approximately the same. This involves nothing more than applying a perturbation to the data set, a perturbation which is usually chosen by trial and error. To overcome this somewhat arbitrary approach, note that for a composition \mathbf{x} and its opposite $\ominus\mathbf{x}$, it holds that $\mathbf{x} \oplus (\ominus\mathbf{x}) = \mathbf{u} = [1/D, \dots, 1/D]$. This means that we can move by perturbation any composition to the barycentre \mathbf{u} of the simplex, in the same way as we move real data in real space to the origin by translation. This property, together with the definition of centre, allows us to design a strategy to move the set of samples in such a way, that its structure is better visualised and all the pairwise ratios between parts are preserved. To do that, we just need to compute the centre \mathbf{g} of our data set \mathbf{X} and perturb each sample of \mathbf{X} by the opposite $\ominus\mathbf{g}$. That is, the centred data set \mathbf{X}^C is formed by the compositions $\mathbf{x}_i^C = \mathbf{x}_i \oplus (\ominus\mathbf{g})$, for $i = 1, \dots, n$. This has the effect of moving the centre of the data set (\mathbf{g}) to the barycentre of the simplex (\mathbf{u}), and the set of samples will gravitate around \mathbf{u} , giving us an automatic and optimal way of visualising our data.

This strategy was first introduced in order to improve the visualisation of CoDa sets in ternary diagrams [MB01]. An extensive discussion can be found in [EPE02], 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 nonlinear distortion. See Fig. 2.2 for an example of a data set before (Fig. 2.2a) and after (Fig. 2.2b) centering and the effect on the gridlines.

Example 2.

Figure 2.2a shows the 3-part subcomposition \mathbf{X}_s of non-academic tasks $[A, O, S]$ for the *statisticiantimebudget* CoDa set (see Appendix). The centre (in %) of this subcomposition is $\mathbf{g}_s = [19.58, 36.49, 43.93]$.

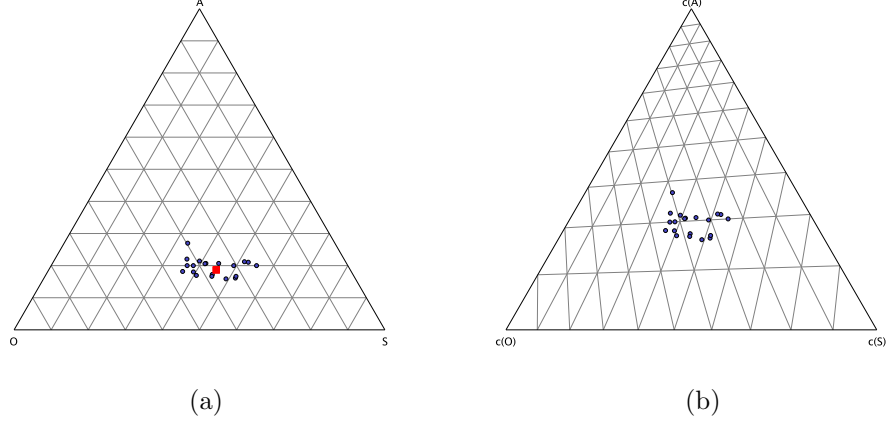


Figure 2.2. Centering of the 3-part subcomposition $[A, O, S]$ of `statisticiantimebudget` data set: (a) original subcompositions \mathbf{X}_s (blue dots) and the centre \mathbf{g}_s (red square); (b) centred subcompositions \mathbf{X}_s^C (blue dots). The vertices $c(A)$, $c(O)$, and $c(S)$ represent centred parts.

Figure 2.2b shows the samples after its perturbation by $\ominus \mathbf{g}_s$. They are distributed around $\mathbf{u} = [1/3, 1/3, 1/3]$, the barycenter of the ternary diagram ($1/3 \approx 33.33\%$).

Activities for Section 2.1

[Click here to get the activities of this section](#)

2.2. Covariance structure of a compositional data set [§]

We can describe the covariance structure of a CoDa set in different ways.

2.2.1. Variation matrix and variation array. A way to describe dispersion in a CoDa set \mathbf{x} is by the *variation matrix* defined as

$$\mathbf{T} = [\tau_{ij}] = \begin{bmatrix} \text{var} \left(\ln \frac{X_1}{X_1} \right) & \text{var} \left(\ln \frac{X_1}{X_2} \right) & \cdots & \text{var} \left(\ln \frac{X_1}{X_D} \right) \\ \text{var} \left(\ln \frac{X_2}{X_1} \right) & \text{var} \left(\ln \frac{X_2}{X_2} \right) & \cdots & \text{var} \left(\ln \frac{X_2}{X_D} \right) \\ \vdots & \vdots & \ddots & \vdots \\ \text{var} \left(\ln \frac{X_D}{X_1} \right) & \text{var} \left(\ln \frac{X_D}{X_2} \right) & \cdots & \text{var} \left(\ln \frac{X_D}{X_D} \right) \end{bmatrix},$$

[§]This section is an adaptation of [DBB06, p. 161–163], [Ait86, Sections 4.1–4.9, p. 64–83]. Further information in [PET15, Section 5.2, p. 66–68].

where $\tau_{ij} = \text{var}(\ln(X_i/X_j))$ stands for the usual variance of the logratio of parts i and j .

Note that, by definition, \mathbf{T} is symmetric and its diagonal contains only zeros. Indeed, it holds $\text{var}(\ln \frac{X_i}{X_j}) = \text{var}(\ln \frac{X_j}{X_i})$, for $i, j : 1, \dots, D$ and $\text{var}(\ln \frac{X_i}{X_i}) = \text{var}(\ln \mathbf{1}) = 0$, where $\mathbf{1}$ is the all-ones column vector. The variation matrix \mathbf{T} can be generalised to the *variation array* \mathbf{T}^* replacing the lower diagonal elements of the matrix $\text{var}(\ln \frac{X_i}{X_j})$, for $i > j : 1, \dots, D$ by the log-ratio expectations $E(\ln \frac{X_i}{X_j})$. That is, the array \mathbf{T}^* is

$$\mathbf{T}^* = \begin{bmatrix} 0 & \text{var}(\ln \frac{X_1}{X_2}) & \cdots & \text{var}(\ln \frac{X_1}{X_D}) \\ E(\ln \frac{X_2}{X_1}) & 0 & \cdots & \text{var}(\ln \frac{X_2}{X_D}) \\ \vdots & \vdots & \ddots & \vdots \\ E(\ln \frac{X_D}{X_1}) & E(\ln \frac{X_D}{X_2}) & \cdots & 0 \end{bmatrix}.$$

Considering all the pairwise logratios, the array \mathbf{T}^* simultaneously shows information about *location* ($E(\cdot)$) and *spread* ($\text{var}(\cdot)$) of the compositional set \mathbf{x} . Furthermore, note that any single entry in \mathbf{T} and \mathbf{T}^* does not depend on the scale constant κ associated with the sample space \mathcal{S}^D , as constants cancel out when taking ratios. Consequently, rescaling has no effect.

2.2.2. Centred logratio covariance matrix. Another way to describe the relative variability in a CoDa set \mathbf{x} is by means of the *clr-covariance matrix* $\mathbf{\Gamma}$, that is, the covariance matrix of $\mathbf{Z} = \text{clr } \mathbf{x}$, where the *clr*-scores are calculated row-wise. The matrix $\mathbf{\Gamma}$ is:

$$\mathbf{\Gamma} = [\gamma_{ij}] = \begin{bmatrix} \text{var}(Z_1) & \text{cov}(Z_1, Z_2) & \cdots & \text{cov}(Z_1, Z_D) \\ \text{cov}(Z_2, Z_1) & \text{var}(Z_2) & \cdots & \text{cov}(Z_2, Z_D) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(Z_D, Z_1) & \text{cov}(Z_D, Z_2) & \cdots & \text{var}(Z_D) \end{bmatrix},$$

with Z_1, \dots, Z_D the column-variables of \mathbf{Z} . The covariance matrix $\mathbf{\Gamma}$ is singular because $\sum_{j=1}^D \gamma_{ij} = 0$, for $i = 1, 2, \dots, D$. In addition, the value of a particular γ_{ij} , for $i, j = 1, 2, \dots, D$ should be not interpreted because depends on the subcomposition. For example, when $\mathbf{Z} = \text{clr } \mathbf{x}$ is calculated using the full composition, for $i \neq j$, one can get $\text{cov}(Z_i, Z_j) > 0$, whereas for a particular subcomposition this value γ_{ij} can be negative. That is, matrix $\mathbf{\Gamma}$ is not subcompositionally coherent.

2.2.3. Additive logratio covariance matrix. We can describe the relative variability in a CoDa set \mathbf{x} by means of a *alr-covariance matrix* $\mathbf{\Sigma}$, that is, the covariance matrix of $\mathbf{Y} = \text{alr } \mathbf{x}$, where the *alr*-coordinates are calculated row-wise. The matrix $\mathbf{\Sigma}$ is:

$$\mathbf{\Sigma} = [\sigma_{ij}] = \begin{bmatrix} \text{var}(Y_1) & \text{cov}(Y_1, Y_2) & \cdots & \text{cov}(Y_1, Y_{D-1}) \\ \text{cov}(Y_2, Y_1) & \text{var}(Y_2) & \cdots & \text{cov}(Y_2, Y_{D-1}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(Y_{D-1}, Y_1) & \text{cov}(Y_{D-1}, Y_2) & \cdots & \text{var}(Y_{D-1}) \end{bmatrix},$$

with Y_1, \dots, Y_{D-1} the column-variables of \mathbf{Y} . The covariance matrix Σ is not unique because it depends of the part used as denominator in the *olr*-coordinates.

2.2.4. *olr*-covariance matrix. We can also describe the relative variability in a CoDa set \mathbf{x} by means of a *olr-covariance matrix* Ω , that is, the covariance matrix of $\mathbf{x}^* = \text{olr } \mathbf{x}$, where the *olr*-coordinates are calculated row-wise. The matrix Ω is:

$$\Omega = [\omega_{ij}] = \begin{bmatrix} \text{var}(X_1^*) & \text{cov}(X_1^*, X_2^*) & \cdots & \text{cov}(X_1^*, X_{D-1}^*) \\ \text{cov}(X_2^*, X_1^*) & \text{var}(X_2^*) & \cdots & \text{cov}(X_2^*, X_{D-1}^*) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(X_{D-1}^*, X_1^*) & \text{cov}(X_{D-1}^*, X_2^*) & \cdots & \text{var}(X_{D-1}^*) \end{bmatrix},$$

with X_1^*, \dots, X_{D-1}^* the column-variables of \mathbf{x}^* . Note that *olr*-coordinates are calculated once a *olr*-basis is selected. Consequently, the covariance matrix Ω is not unique because the values of ω_{ij} , for $i, j : 1, \dots, D-1$, depend of the *olr*-basis used.

We can use any of the matrices \mathbf{T} , $\mathbf{\Gamma}$, Σ , and Ω to describe the covariance structure of a CoDa set. We can move readily from one matrix representation to another (see [Ait86, PET15]). The four representations have advantages and disadvantages: \mathbf{T} in its confinement to 2-part variation is simple, treats parts symmetrically but is not a covariance matrix; $\mathbf{\Gamma}$ is a covariance matrix, treats parts symmetrically but is singular and not subcompositionally coherent; Σ and Ω are both covariance matrices, generally non-singular but are not unique, that is, they are asymmetric in its treatment of parts.

Example 3.

Let \mathbf{X} be the CoDa set of the `statisticiantimebudget.cdp` file (see Appendix). The variation array \mathbf{T}^* is

$$\mathbf{T}^* = \begin{bmatrix} \text{E}(\cdot) | \text{var}(\cdot) & T & C & A & R & O & S \\ T & & 0.0608 & 0.0826 & 0.0325 & 0.0418 & 0.1023 \\ C & -0.3412 & & 0.0415 & 0.0394 & 0.0458 & 0.0708 \\ A & -0.1712 & 0.1701 & & 0.0406 & 0.0473 & 0.0538 \\ R & -0.2675 & 0.0737 & -0.0963 & & 0.0502 & 0.0701 \\ O & 0.4511 & 0.7923 & 0.6222 & 0.7185 & & 0.0967 \\ S & 0.6367 & 0.9779 & 0.8078 & 0.9042 & 0.1856 & \end{bmatrix}.$$

The lowest pairwise log-ratio variance is $\text{var}(\ln \frac{R}{T}) = 0.0325$. If we assume that this value is close to zero then (R, T) are activities with a proportional time. In this case, the estimate of pairwise log-ratio expectation is $\text{E}(\ln \frac{R}{T}) \approx -0.2675$, where, in average,

$$\ln \frac{R}{T} \approx -0.2675 \rightarrow \frac{R}{T} \approx 0.7653 \rightarrow R \approx 0.7653 T,$$

the time dedicated to R is approximately 76.53% time spent in T .

The matrix $\mathbf{\Gamma}$, the covariance matrix of *clr*-scores set, is:

$$\mathbf{\Gamma} = \begin{bmatrix} & \text{clr}_T & \text{clr}_C & \text{clr}_A & \text{clr}_R & \text{clr}_O & \text{clr}_S \\ \text{clr}_T & 0.0290 & -0.0066 & -0.0168 & 0.0055 & 0.0049 & -0.0160 \\ \text{clr}_C & -0.0066 & 0.0187 & -0.0014 & -0.0031 & -0.0022 & -0.0054 \\ \text{clr}_A & -0.0168 & -0.0014 & 0.0200 & -0.0031 & -0.0023 & 0.0037 \\ \text{clr}_R & 0.0055 & -0.0031 & -0.0031 & 0.0145 & -0.0066 & -0.0072 \\ \text{clr}_O & 0.0049 & -0.0022 & -0.0023 & -0.0066 & 0.0226 & -0.0164 \\ \text{clr}_S & -0.0160 & -0.0054 & 0.0037 & -0.0072 & -0.0164 & 0.0413 \end{bmatrix},$$

where the most relevant information are the *clr*-variances recorded in the diagonal. The part *R* has lowest variance (0.0145) suggesting that the proportion of time dedicated to research activities has a low relative variation along the 20 days. In contrast, the part *S* has the largest contribution (0.0413) to the variability of the data set.

The covariance matrix Σ of *alr*-coordinates when the part sleep (*S*) is used as denominator is

$$\Sigma = \begin{bmatrix} \ln \frac{T}{S} & \ln \frac{C}{S} & \ln \frac{A}{S} & \ln \frac{R}{S} & \ln \frac{O}{S} \\ \ln \frac{T}{S} & 0.1023 & 0.0561 & 0.0368 & 0.0700 & 0.0786 \\ \ln \frac{C}{S} & 0.0561 & 0.0708 & 0.0416 & 0.0507 & 0.0608 \\ \ln \frac{A}{S} & 0.0368 & 0.0416 & 0.0538 & 0.0417 & 0.0516 \\ \ln \frac{R}{S} & 0.0700 & 0.0507 & 0.0417 & 0.0701 & 0.0583 \\ \ln \frac{O}{S} & 0.0786 & 0.0608 & 0.0516 & 0.0583 & 0.0967 \end{bmatrix},$$

where the elements of the diagonal are equal to the last column (right side) of the variation array \mathbf{T}^* . Indicating that each column of a part X_j in the variation matrix \mathbf{T} is equal to the diagonal of the matrix Σ obtained when this part X_j is used as denominator for the *alr*-coordinates. The positive value of all covariances (off-diagonal elements in Σ) suggest some *isotemporal substitution*. That is, when a pairwise ratio increases because the proportion of time dedicated to *S* decreases then all the other daily activities increase its proportion.

Using the same *olr*-basis than in Example 1, the covariance matrix Ω is

$$\Omega = \begin{bmatrix} x_1^* & x_2^* & x_3^* & x_4^* & x_5^* \\ x_1^* & 0.0233 & 0.0081 & -0.0016 & 0.0055 & 0.0114 \\ x_2^* & 0.0081 & 0.0374 & 0.0112 & 0.0041 & 0.0086 \\ x_3^* & -0.0016 & 0.0112 & 0.0162 & 0.0052 & 0.0102 \\ x_4^* & 0.0055 & 0.0041 & 0.0052 & 0.0207 & 0.0046 \\ x_5^* & 0.0114 & 0.0086 & 0.0102 & 0.0046 & 0.0484 \end{bmatrix},$$

Because the *olr*-coordinates x_3^* , x_4^* , and x_5^* are proportional (factor 1/2) to pairwise logratios $\ln(T/R)$, $\ln(C/A)$ and $\ln(O/S)$ respectively, then its values in the matrix Ω can be easily deduced from the counterparts in matrices \mathbf{T} and Σ . Using the variances of balances x_1^* (0.0233) and x_2^* (0.0374) and its covariance (0.0081) one can obtain that the Pearson correlation coefficient is 0.2727. If one assumes that it is significant, it suggests that when the ratio between the average of ‘working’ and ‘leisure’ activities increases then the ratio between the average of (*T*, *R*) and (*C*, *A*) also increases.

2.2.5. Total variance. A measure of total relative variability of a CoDa set \mathbf{X} was historically defined by

$$(2.5) \quad \text{totvar}(\mathbf{X}) = \frac{1}{2D} \sum_{i=1}^D \sum_{j=1}^D \text{var} \left(\ln \frac{X_i}{X_j} \right),$$

where *totvar* signifies the *total variance* of the data set, that is, the sum of all the elements of the variation matrix \mathbf{T} divided by $2D$. The total variance of \mathbf{X} coincides with the *trace* (tr) of the *clr*-covariance matrix $\mathbf{\Gamma}$ and of any *olr*-covariance matrix

$\mathbf{\Omega}$:

$$(2.6) \quad \text{totvar}(\mathbf{X}) = \text{tr}(\mathbf{\Gamma}) = \sum_{j=1}^D \text{var}(Z_j) = \text{tr}(\mathbf{\Omega}) = \sum_{j=1}^{D-1} \text{var}(X_j^*),$$

where $\text{tr}(\cdot)$ of a matrix is equal to the sum of elements on the main diagonal (from the upper left to the lower right). Total variance inherits the properties of the trace of a matrix. In particular, it is invariant under a change of basis, that is, $\text{totvar}(\mathbf{X})$ is invariant as regards the *olr*-basis selected. In addition, because total variance is a sum of variances, it is invariant by perturbation. Consequently, a CoDa set \mathbf{X} and the centred data set \mathbf{X}^C have the same total variance.

From their definition, it is clear that total variance summarizes the variation matrix in a single quantity, 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 variance for a pressure-temperature random vector, for instance). Conversely, the variation matrix explains how the total variation is split among the parts (or more precisely, among all logratios). Indeed, given a part X_j , expressing $\text{var}(Z_j) / \text{totvar}(\mathbf{X})$ as a percentage, one can interpret the amount of total variance accounted by the part. In addition, once an *olr*-coordinates are calculated based on an SBP, the ratio $\text{var}(X_j^*) / \text{totvar}(\mathbf{X})$ indicates the part of total variance retained by a particular balance which is interesting for the analyst.

Example 4.

Let \mathbf{X} be the CoDa set recorded in the file `statisticiantimebudget.cdp` (see [Appendix](#)). The total variance is $\text{totvar}(\mathbf{X}) = 0.1460$, which is equal to the trace of matrix $\mathbf{\Gamma}$:

$$\text{tr}(\mathbf{\Gamma}) = 0.0290 + 0.0187 + 0.020 + 0.0145 + 0.0226 + 0.0413 = 0.1460.$$

The time dedicated to ‘sleep’ has the largest *clr*-variability (0.0413), which contribution to the total variance is 28.28% ($=0.0413/0.1460$). Because $\text{totvar}(\mathbf{X})$ is also equal to $\text{tr}(\mathbf{\Omega})$, the total variance also decomposes as a sum of the *olr*-variances. The largest *olr*-variance (diagonal elements of $\mathbf{\Omega}$) is $\text{var}(x_5^*) = 0.0484$ which contribution is 33.15% to the total variance. Importantly, x_5^* is proportional to the logratio $\ln(O/S)$, suggesting that the ratio between activities (O, S) are the responsible of large amount of the variability in the CoDa set.

2.2.6. Scaling a compositional data set. In the real space, a centred random variable can be scaled to unit variance dividing it by its standard deviation. Consequently, scaling a real data set consists of scaling each variable (column) separately. On the other hand, a (centred) compositional set \mathbf{X} can be scaled by row-wise powering it with the factor $(\text{totvar}(\mathbf{X}))^{-1/2}$. That is, the compositional set $(\text{totvar}(\mathbf{X}))^{-1/2} \odot \mathbf{X}$ has unit total variance (i.e., $\text{totvar}((\text{totvar}(\mathbf{X}))^{-1/2} \odot \mathbf{X}) = 1$). Importantly, each pairwise logratio in the variation matrix of the scaled CoDa set has the same relative contribution to the total variance. This is a relevant difference with conventional standardisation in the real space: with real vectors, the relative contribution is an artifact of the units of each single variable, and usually it should be ignored. In contrast, in compositional vectors, all parts share the

same ‘units’, and their relative contribution to total variance is a rich source of information.

Example 5.

Figure 2.3a shows the *centred* 3-part subcomposition \mathbf{X}_s^C of non-academic tasks $[A, O, S]$ for the `statisticiantimebudget` data set (Fig. 2.2). The total variance of the data set is $\text{totvar}(\mathbf{X}) = \text{totvar}(\mathbf{X})^C = 0.0659$.

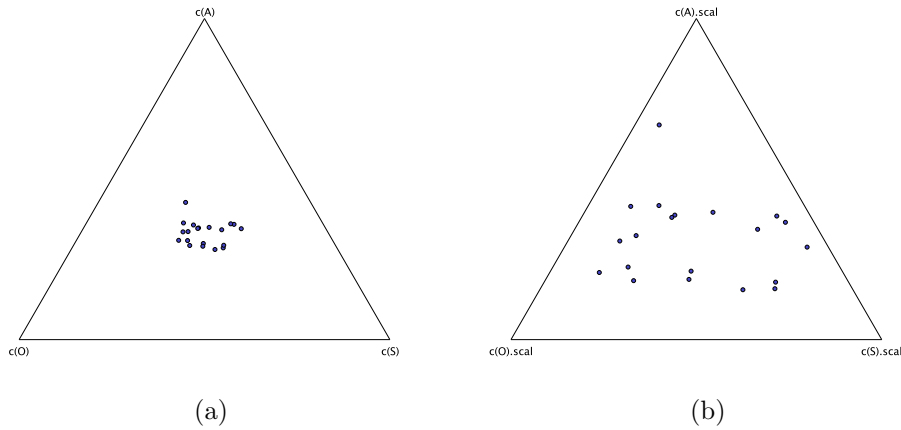


Figure 2.3. Scaling of a *centred* 3-part subcomposition of `statisticiantimebudget` data set: (a) original centred subcompositions \mathbf{X}_s^C : $[c(A), c(O), c(S)]$; (b) scaled subcompositions. The vertices $c(A).scal$, $c(O).scal$, and $c(S).scal$ represent scaled parts.

Figure 2.3b shows the samples after its powering by $1/\sqrt{0.0659} \approx 3.895$. The scaled data set has total variance equal to one (1), it is more spread than the original data set.

Activities for Section 2.2

🔗 [Click here to get the activities of this section](#)

2.3. CoDa-dendrogram

A *balance-dendrogram* or CoDa-dendrogram available in CoDaPack is the joint representation of: a) the SBP (in the form of a tree structure); b) the sample mean and variance of each balance (or *olr*-coordinate); and c) a box plot summarising the order statistic of each balance. The tree structure created is similar to a typical

dendrogram used in hierarchical cluster analysis. In the CoDa-dendrogram, each *olr*-coordinate (balance) is represented on a horizontal axis, whose limits correspond to a certain range (the same for every coordinate). The length of the vertical bar going up from each one of these coordinate axes represents the variance of the *olr*-coordinate, and the contact point represents the value of the arithmetic mean of the *olr*-coordinate. The *leaves of the tree* (i.e., the parts) are ordered according to its inclusion in the numerator (right side of the tree) or the denominator (left side of the tree) of the balances. In this way, the part always included in the numerator of the SBP is the first leaf on the right hand side. On the other hand, the part always included in the denominator is the first leaf on the left. Moreover, CoDa-Pack adds to the current data frame the new columns with the *olr*-coordinates of the compositions with respect to the *olr*-basis associated to the SBP. Figure 2.4 shows a CoDa-dendrogram for the `statisticiantimebudget` CoDa set.

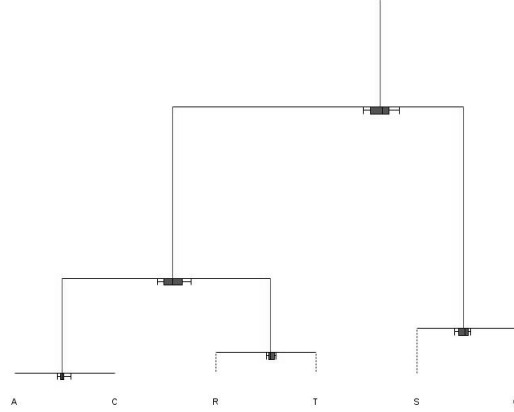


Figure 2.4. A balance dendrogram of the statistician's time budget (`statisticiantimebudget` CoDa set)

Example 6.

From Figure 2.4 one can deduce that the first partition splits the composition (A, C, R, T, S, O) into (S, O) and (A, C, R, T) included respectively in the numerator and denominator of the first balance. Following this procedure, the SBP derived from the CoDa-dendrogram is:

i	Sign matrix \mathbf{S} of the SBP						n	p
	A	C	R	T	S	O		
1	-1	-1	-1	-1	+1	+1	4	2
2	0	0	0	0	-1	+1	1	1
3	-1	-1	+1	+1	0	0	2	2
4	0	0	-1	+1	0	0	1	1
5	-1	+1	0	0	0	0	1	1


Given that the range of each coordinate is symmetric (in Figure 2.4 the range goes from -2 to $+2$), the box plots closer to one side indicate that the part (or the

geometric mean of group of parts) on this side is more abundant than the part (or the geometric mean of group of parts) corresponding to the opposite side. Thus, in Figure 2.4, the first box plot is placed entirely on the right half of the first horizontal axis. This indicates that the geometric mean of group $\{S, O\}$ (on the right side) is greater than the geometric mean of group $\{A, C, R, T\}$ (on the left side) in all samples (days). Similarly, the box plot relative to the fourth *olr*-coordinate is located almost entirely on the right side indicating that part T is greater than part R in almost all samples (days). No similar arguments can be made in relation to the parts involved in the second and third *olr*-coordinates because the zero is included in their box plots.

The longest vertical bar corresponds to the second *olr*-coordinate. This means that the logratio $\ln(O/S)$ has a high variability (in comparison with the other balances). That is, it has the largest contribution to the total variance of the data set ($\text{totvar}(\mathbf{X})$). In contrast, the shortest vertical bar corresponds to the fourth *olr*-coordinate indicating the small variability of the logratio $\ln(T/R)$ relative to the variability of the other balances.

Another easily read feature from a balance-dendrogram is the symmetry of balances. This can be assessed by comparing between the length of the two boxes, and also by comparing the length of the two whiskers. We can also compare the position of the median (located on the common border of the two quantile boxes) relative to that of the mean (the point where the vertical bar joins the coordinate axe). From Figure 2.4 the box plot for the third *olr*-coordinate (i.e., balance between the group $\{T, R\}$ and the group $\{C, A\}$) suggests a symmetric distribution when compared to the other balances.

Activities for Section 2.3

 [Click here to get the activities of this section](#)

2.4. Reduced-dimensionality representation of a compositional data set: clr-biplot ^{||}

Consider the CoDa matrix \mathbf{X} with n rows and D columns introduced in Eq. (2.1). Broadly speaking, the aim of dimensionality reduction techniques is to reduce the number of variables (D) in the CoDa set without having to lose much information. One approach, known as *variable selection*, consists of selecting S original parts forming a subcomposition (see Section 2.4.5) which retain a large amount of total variance ($\text{totvar}(\mathbf{X})$). Another popular approach consists of a transformation of data from a high-dimensional space into a low-dimensional space which creates a reduced set of new variables. The number of new variables (r) is usually $r = 2$ or 3 , being each new variable a combination of the original variables, containing basically the same information as the original variables. To create these new variables we use

^{||}This section is an adaptation of [DBB06, p. 163–165], [Ait03, sections 4.3–4.4, p. 83–90], [DTM11]. Further information in [PET15, Section 5.4, p. 70–76].

the *singular value decomposition* of the CoDa-matrix (Section 2.4.1) in the context of a *principal component analysis* (Section 2.4.2). To illustrate the dimensionality reduction of a CoDa matrix \mathbf{X} we will represent both samples and original variables in a plot known as a *compositional biplot*, a *CoDa-biplot* or a *clr-biplot* (Sections 2.4.3 and 2.4.4).

Example 7.

Figure 2.5 shows a CoDa-biplot for the `statisticiantimebudget` CoDa set (see Appendix). Let \mathbf{X} be the CoDa set recorded in the file. The matrix \mathbf{X} has 20 rows (n) and six columns (D). That is, the 20 6-part compositions are in the simplex \mathcal{S}^6 , a 5-dimensional sample space. Figure 2.5 is a 2-dimensional representation of the data set. The name *clr-biplot* indicates that the plot is created using the *clr*-scores of \mathbf{X} . The six parts, represented by the *clr*-variables (red *rays*) and the 20 samples, using its *clr*-scores (blue dots), are projected in a 2-dimensional space. This space is generated by the two first vectors of an *olr*-basis that are respectively represented by the horizontal and vertical axes (*ilr.1* and *ilr.2*). On the following pages, we explain how this *olr*-basis is created, the variables and samples are projected, how to evaluate the quality of such representation, and we give the basic guidelines to interpret the elements forming the biplot (axis, rays, etc).

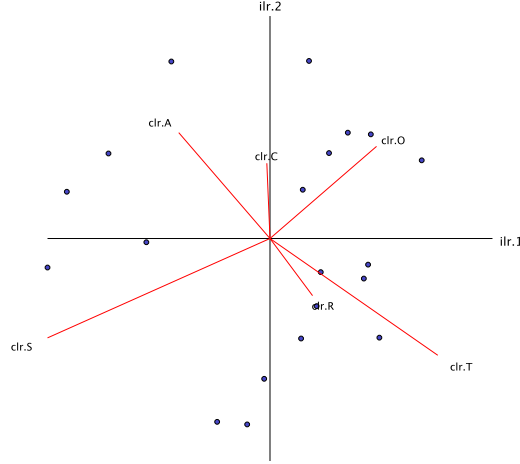


Figure 2.5. *clr*-biplot of the statistician's time budget (`statisticiantimebudget` data set).

2.4.1. Simplicial singular value decomposition. Let \mathbf{X} be a CoDa set with n samples and D parts which has been previously centred. Afterwards, we consider $\mathbf{Z} = \text{clr } \mathbf{X}$, the *clr*-scores set of the centred data set \mathbf{X} . In other words, let \mathbf{Z} be the double-centred data set obtained by centring by rows and columns the data set $\ln \mathbf{X}$. Note that \mathbf{Z} is of the same order as \mathbf{X} , that is, it has n rows and D columns. Since the *clr*-scores preserve distances, standard dimension reducing techniques can be applied to \mathbf{Z} , and in particular the singular value decomposition

(SVD). The SVD technique applied to a compositional set, simultaneously allows to reduce its dimensionality with minimum lose of information and to describe relations between variables and observations.

The SVD of the matrix \mathbf{Z} consists of the obtention of two type of elements: 1) \mathbf{U} and \mathbf{V} called eigenvectors and 2) $\lambda_1, \dots, \lambda_{D-1}$, called eigenvalues.

Thus, it holds that

$$(2.7) \quad \mathbf{Z} = \mathbf{U} \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_{D-1}} \end{bmatrix} \mathbf{V}^t,$$

where the matrices \mathbf{U} and \mathbf{V} are respectively of order $n \times D - 1$ and $D \times D - 1$. Let $\mathbf{u}_1, \dots, \mathbf{u}_{D-1}$ be the set of column vectors of matrix \mathbf{U} and $\mathbf{v}_1, \dots, \mathbf{v}_{D-1}$ those of \mathbf{V} . It holds that both sets are formed by orthonormal vectors. That is, the inner products $\langle \mathbf{u}_i, \mathbf{u}_j \rangle$ and $\langle \mathbf{v}_i, \mathbf{v}_j \rangle$ are equal to zero for $i \neq j$ and equal to one when $i = j$, for $i, j = 1, \dots, D - 1$. In other words, it holds $\mathbf{U}^t \mathbf{U} = \mathbf{I}_{D-1}$ and $\mathbf{V}^t \mathbf{V} = \mathbf{I}_{D-1}$, being \mathbf{I}_{D-1} the *identity* matrix (i.e., the square matrix with ones on the main diagonal and zeros elsewhere).

The expression (2.7) can be written equivalently as

$$(2.8) \quad \mathbf{Z} = [\mathbf{u}_1, \dots, \mathbf{u}_{D-1}] \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_{D-1}} \end{bmatrix} [\mathbf{v}_1, \dots, \mathbf{v}_{D-1}]^t.$$

The matrix \mathbf{U} of the SVD of a matrix \mathbf{Z} of order $n \times D$ is formed by the eigenvectors $\mathbf{u}_1, \dots, \mathbf{u}_{D-1}$ of matrix $\mathbf{Z}\mathbf{Z}^t$. Moreover, the columns of matrix \mathbf{V} are the eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_{D-1}$ of matrix $\mathbf{Z}^t\mathbf{Z}$. Finally, the diagonal $\sqrt{\lambda_1}, \dots, \sqrt{\lambda_{D-1}}$ are the square roots of the $D - 1$ positive eigenvalues $\lambda_1, \dots, \lambda_{D-1}$ of either $\mathbf{Z}\mathbf{Z}^t$ or $\mathbf{Z}^t\mathbf{Z}$, which are equal up to additional null eigenvalues. We assume that the eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{D-1}$ are in descending order of magnitude. Note that since the *clr*-covariance matrix $\mathbf{\Gamma}$ of \mathbf{x} coincides with $\frac{1}{n-1}\mathbf{Z}^t\mathbf{Z}$, it holds that the $\mathbf{Z}^t\mathbf{Z}$ and $\mathbf{\Gamma}$ have the same D eigenvectors and eigenvalues. At least one of the eigenvectors is equal to zero ($\lambda_D = 0$) because the matrix $\mathbf{\Gamma}$ is singular. Note that the matrix \mathbf{Z} is formed by the centred *clr*-scores, where its rows and columns add up to zero. Therefore, the rank of matrix \mathbf{Z} is lower or equal than $D - 1$, being the rank equal to $D - 1$ the most common case. Consequently, the eigenvector \mathbf{v}_D associated to the eigenvalue $\lambda_D = 0$ is always equal to $\mathbf{v}_D = \mathbf{1}$, the all-ones column vector, which is not considered in our SVD technique.

2.4.2. Logconstrast principal components analysis. Given that the eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{D-1}$ are sorted in descending order of magnitude, to obtain a reduced-dimensionality representation of \mathbf{X} , we can project \mathbf{Z} down into a reduced space defined by only the first r singular vectors $\mathbf{v}_1, \dots, \mathbf{v}_r$ of matrix \mathbf{V} , where r is usually 2 or 3.

Assuming the common case where the rank of matrix \mathbf{Z} is $D-1$, the interpretation of SVD in Equation (2.7) is straightforward:

- The rows $\mathbf{v}_1^t, \dots, \mathbf{v}_{D-1}^t$ of \mathbf{V}^t are the vectors of the *clr*-coordinates of an *olr*-basis in the simplex \mathcal{S}^D .
- Let $\mathbf{e}_1, \dots, \mathbf{e}_{D-1}$ be the *clr*-backtransformed eigenvectors, that is,

$$\mathbf{e}_j = \text{clr}^{-1} \mathbf{v}_j \quad (j = 1, \dots, D-1) .$$

Then $\mathcal{B} = \{\mathbf{e}_1, \dots, \mathbf{e}_{D-1}\}$ constitute an *olr*-basis of \mathcal{S}^D .

The direction $t \odot \mathbf{e}_j$ ($t \in \mathbb{R}$) in the simplex \mathcal{S}^D associated to \mathbf{e}_j is known as *jth-principal component* (PC) axis.

- The matrix product

$$\mathbf{U} \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_{D-1}}) = \mathbf{X}^*$$

is a $n \times (D-1)$ matrix \mathbf{X}^* whose n rows are the *olr*-coordinates of each composition data point of \mathbf{X} with respect to the *olr*-basis \mathcal{B} . By columns, we symbolize this matrix of *olr*-coordinates as $\mathbf{X}^* = [X_1^* \dots X_{D-1}^*]$.

Let $\mathbf{x}_i^* = [x_{i1}^*, \dots, x_{iD-1}^*]$ be the i^{th} -row of \mathbf{X}^* , for $i = 1, \dots, n$. Thus x_{ij}^* is the j^{th} -coordinate of the composition \mathbf{x}_i with respect to the *olr*-basis \mathcal{B} . That is, x_{ij}^* is the coordinate of \mathbf{x}_i on the j^{th} -PC axis. According to the expression of a logcontrast (see Chapter 1), since $x_{ij}^* = \langle \mathbf{x}_i, \mathbf{e}_j \rangle_a$, it holds that the coordinates x_{ij}^* can be written as logcontrasts

$$x_{ij}^* = v_{j1} \ln x_{i1} + \dots + v_{jD} \ln x_{iD} ,$$

where v_{j1}, \dots, v_{jD} are the components of the vector \mathbf{v}_j of matrix \mathbf{V} . Keep in mind that the vectors $\mathbf{v}_1, \dots, \mathbf{v}_{D-1}$ belong to the *clr*-subspace U (see Chapter 1) and, therefore, its components add up to zero.

- The eigenvalues $\lambda_1, \dots, \lambda_{D-1}$ are respectively equal to sample variances $\text{var}(X_1^*), \dots, \text{var}(X_{D-1}^*)$ of the *olr*-coordinates. Let $\mathbf{\Omega}$ be the *olr*-covariance matrix associated to the *olr*-basis \mathcal{B} . Thus, the sum $\text{tr}(\mathbf{\Omega}) = \lambda_1 + \dots + \lambda_{D-1}$ is equal to the total variance $\text{totvar}(\mathbf{X})$ (see Chapter 2).

Therefore, if we want a reduced-dimensionality representation of \mathbf{X} , we can project \mathbf{Z} down into the reduced space defined by only the first r eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_r$ of matrix \mathbf{V} . This is equivalent to selecting the first r columns X_1^*, \dots, X_r^* of \mathbf{X}^* . The first r *olr*-coordinates \mathbf{x}^* of a generic composition \mathbf{x} of \mathbf{X} in this reduced space are those given for the following logcontrasts:

$$\begin{aligned} x_1^* &= v_{11} \ln x_1 + \dots + v_{1D} \ln x_D , \\ x_2^* &= v_{21} \ln x_1 + \dots + v_{2D} \ln x_D , \\ &\vdots \\ x_r^* &= v_{r1} \ln x_1 + \dots + v_{rD} \ln x_D , \end{aligned}$$

The proportion of the total variability ($\text{totvar}(\mathbf{X})$) which is retained by, or contained in, the first r (logcontrast) PC is then $(\lambda_1 + \dots + \lambda_r) / (\lambda_1 + \dots + \lambda_{D-1})$.

2.4.3. Compositional biplot. [Gab71] introduced the biplot to represent simultaneously the rows and columns of any matrix by means of a 2-rank approximation. [Ait97] adapted it for CoDa and proved it to be a useful exploratory tool. Here we briefly describe first the philosophy and mathematics of this technique, and then its interpretation in depth. A very important reference is [GH96].

In order to reduce the dimension of the CoDa set, we can suppress some *olr*-coordinates, typically those with associated low variance. This can be thought as deletion of eigenvalues. Assume that we retain r eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r$ ($r \leq D - 1$). Then the proportion of retained variance is

$$\frac{\lambda_1 + \dots + \lambda_r}{\lambda_1 + \dots + \lambda_{D-1}}.$$

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

$$(2.9) \quad \mathbf{A} = \begin{bmatrix} u_{11} & u_{21} \\ \vdots & \vdots \\ u_{1n} & u_{2n} \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{bmatrix} \begin{bmatrix} v_{11} & \dots & v_{1D} \\ v_{21} & \dots & v_{2D} \end{bmatrix},$$

This expression can be written in a more abbreviated form as:

$$(2.10) \quad \mathbf{A} = [\mathbf{u}_1, \mathbf{u}_2] \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{bmatrix} [\mathbf{v}_1, \mathbf{v}_2]^t.$$

The proportion of variability retained by this approximation is $(\lambda_1 + \lambda_2) / (\sum_{j=1}^{D-1} \lambda_j)$.

To obtain a biplot, it is first necessary to write \mathbf{A} as the product of two matrices \mathbf{GH}^t , where \mathbf{G} is a $n \times 2$ matrix and \mathbf{H} is a $D \times 2$ matrix

$$(2.11) \quad \mathbf{A} = \begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_n \end{bmatrix} \begin{bmatrix} \mathbf{h}_1 & \dots & \mathbf{h}_D \end{bmatrix},$$

There are different possibilities to obtain such a factorisation of \mathbf{A} , one of which is

$$(2.12) \quad \mathbf{A} = \begin{bmatrix} \sqrt{n-1}(\sqrt{\lambda_1})^\alpha u_{11} & \sqrt{n-1}(\sqrt{\lambda_2})^\alpha u_{21} \\ \vdots & \vdots \\ \sqrt{n-1}(\sqrt{\lambda_1})^\alpha u_{1n} & \sqrt{n-1}(\sqrt{\lambda_2})^\alpha u_{2n} \end{bmatrix} \times \\ \times \begin{bmatrix} \frac{(\sqrt{\lambda_1})^{1-\alpha}}{\sqrt{n-1}} v_{11} & \dots & \frac{(\sqrt{\lambda_1})^{1-\alpha}}{\sqrt{n-1}} v_{1D} \\ \frac{(\sqrt{\lambda_2})^{1-\alpha}}{\sqrt{n-1}} v_{21} & \dots & \frac{(\sqrt{\lambda_2})^{1-\alpha}}{\sqrt{n-1}} v_{2D} \end{bmatrix} = \begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_n \end{bmatrix} \begin{bmatrix} \mathbf{h}_1 & \dots & \mathbf{h}_D \end{bmatrix},$$

for some constant α .

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

Note that the matrix of the square root of eigenvalues

$$(2.13) \quad \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{bmatrix}$$

now is included on \mathbf{G} (when $\alpha = 1$), or on \mathbf{H} (when $\alpha = 0$), or on both \mathbf{G} and \mathbf{H} (when $0 < \alpha < 1$). If the matrix of the square root of eigenvalues is included entirely on \mathbf{G} (for $\alpha = 1$), the biplot drawn is called *form biplot*. It favours the display of the individuals. If the matrix of the square root of eigenvalues is included entirely on \mathbf{H} (for $\alpha = 0$), the resulting biplot –called *covariance biplot*– favours the display of *clr*-variables. The two biplots differ only by scale changes along the horizontal and vertical axes of the display. If the matrix of the square root of eigenvalues is included half on \mathbf{G} and half on \mathbf{H} (for $\alpha = 0.5$), the resulting biplot is the classical PC analysis and is a compromise between the display of variables and individuals. In all the compositional biplots the variables are conventionally depicted by *rays* emanating from the origin, since both their lengths and directions are important to the interpretation.

Example 8.

Let \mathbf{X} be the CoDa set of `statisticiantimebudget.cdp` file (see [Appendix](#)). Figures 2.6a and 2.6b respectively show the covariance and form *clr*-biplots. The quality of the 3-dimensional representation is high because the proportion of variance retained is 82.41%. Following table shows PCs loadings and % of variance retained:

<i>PC</i>	<i>clr.T</i>	<i>clr.C</i>	<i>clr.A</i>	<i>clr.R</i>	<i>clr.O</i>	<i>clr.S</i>	Cum. Prop.
							Ret. (%)
PC1	0.5329	-0.0100	-0.2898	0.1352	0.3384	-0.7065	43.84
PC2	-0.5114	0.3296	0.4637	-0.2502	0.4035	-0.4352	66.89
PC3	0.1088	-0.5618	0.0106	-0.4611	0.6126	0.2909	82.41
PC4	-0.0503	-0.6357	0.5506	0.4764	-0.1200	-0.2210	95.74
PC5	0.5230	0.0697	0.4807	-0.5593	-0.4079	-0.1061	100.00

For example, the first vector (*ilr.1*) of the *olr*-basis linked to the PCs created by the biplot is the linear combination

$$\begin{aligned} ilr.1 = & 0.5329 \text{clr } X_T - 0.0100 \text{clr } X_C - 0.2898 \text{clr } X_A + 0.1352 \text{clr } X_R + \\ & + 0.3384 \text{clr } X_O - 0.7065 \text{clr } X_S, \end{aligned}$$

that is equivalent to the logcontrast

$$\begin{aligned} ilr.1 = & 0.5329 \ln X_T - 0.0100 \ln X_C - 0.2898 \ln X_A + 0.1352 \ln X_R + \\ & + 0.3384 \ln X_O - 0.7065 \ln X_S, \end{aligned}$$

which retains 43.84% of the total variance. In the expression above it is assumed that the CoDa set \mathbf{X} has been previously centred. When one wants to calculate the first *olr*-coordinate *ilr.1* using the original compositions (without centering) the logcontrast is

$$\begin{aligned} ilr.1 = & 0.5329 \ln \frac{X_T}{g_T} - 0.0100 \ln \frac{X_C}{g_C} - 0.2898 \ln \frac{X_A}{g_A} + 0.1352 \ln \frac{X_R}{g_R} \\ & + 0.3384 \ln \frac{X_O}{g_O} - 0.7065 \ln \frac{X_S}{g_S}, \end{aligned}$$

where $\mathbf{g} = [g_T, g_C, g_A, g_R, g_O, g_S] = [14.75, 10.49, 12.43, 11.29, 23.16, 27.88]$ is the center in %. Note that the *olr*-coordinates of \mathbf{g} in the *clr*-biplot are $[0, 0, \dots, 0]$.

When the logcontrast is expressed as a logratio, one obtains

$$ilr.1 = \ln \frac{X_T^{0.5329} \cdot X_R^{0.1352} \cdot X_O^{0.3384}}{X_C^{0.0100} \cdot X_A^{0.2898} \cdot X_S^{0.7065}}.$$

This expression is more complex than a *balance* of an SBP. To simplify the expression one can assume that only the largest coefficients in the logcontrast are relevant and create a balance using the other parts. That is, we consider that *ilr.1* is approximated by the balance

$$ilr.1 \approx \sqrt{\frac{2 \cdot 2}{2 + 2}} \ln \frac{X_T^{0.5} \cdot X_O^{0.5}}{X_A^{0.5} \cdot X_S^{0.5}} = \ln \frac{\sqrt{X_T \cdot X_O}}{\sqrt{X_A \cdot X_S}}.$$

This idea inspires the *Principal Balances* algorithm (see Section 2.5), designed for creating a *data driven* SBP.

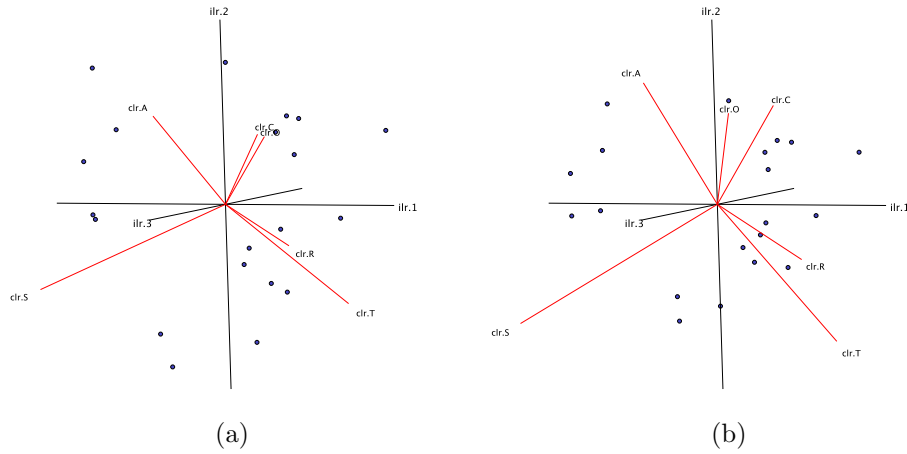


Figure 2.6. *clr*-biplot of `statisticiantimebudget` CoDa set: (a) covariance biplot; (b) form biplot. (Total variance retained: 82.41%).

2.4.4. Interpretation of a compositional biplot. A compositional biplot represent simultaneously the samples and *clr*-variables of any CoDa-matrix by means of a 2-rank approximation. A CoDa-biplot consists of (Fig. 2.7a):

- (1) an *origin* O which represents the centre of the CoDa set;
- (2) a *vertex* at position \mathbf{h}_j for each of the D *clr*-variable ($j = 1, \dots, D$); and
- (3) a *case marker* at position \mathbf{g}_i for each of the n samples or cases ($i = 1, \dots, n$).

We term the join of O to a vertex \mathbf{h}_j the *ray* $\overline{O\mathbf{h}_j}$, and the segment from a vertex \mathbf{h}_j to vertex \mathbf{h}_k , the *link* $\overline{\mathbf{h}_j\mathbf{h}_k}$.

These features constitute the basic characteristics of a CoDa-biplot with the following main properties for the interpretation of compositional variability (Fig. 2.7):

- (1) Links and rays (Fig. 2.7a) provide information on the relative variability in a CoDa-set, since

$$|\overline{\mathbf{h}_j\mathbf{h}_k}|^2 \approx \text{var} \left(\ln \frac{X_j}{X_k} \right) \quad \text{and} \quad |\overline{O\mathbf{h}_j}|^2 \approx \text{var}(\text{clr}_j \mathbf{X}) = \text{var}(Z_j) .$$

Nevertheless, one has to be careful in interpreting rays, which cannot be identified, neither with $\text{var}(X_j)$ nor with $\text{var}(\ln X_j)$, as the rays depend on the full compositions through $g(\mathbf{x}_1), \dots, g(\mathbf{x}_n)$ and vary when a subcomposition is considered.

- (2) Links provide information on the correlation of subcompositions (Fig. 2.7b). Thus, if links $\overline{\mathbf{h}_i\mathbf{h}_k}$ and $\overline{\mathbf{h}_j\mathbf{h}_l}$ intersect at M then

$$\cos(\widehat{\mathbf{h}_i M \mathbf{h}_j}) \approx \text{corr} \left(\ln \frac{X_i}{X_k}, \ln \frac{X_j}{X_l} \right) .$$

Furthermore, if the two links are at right angles, the $\cos(\widehat{\mathbf{h}_i M \mathbf{h}_j}) \approx 0$, and zero correlation of the two logratios can be expected. In particular, links $\overline{\mathbf{h}_i\mathbf{h}_l}$ and $\overline{\mathbf{h}_k\mathbf{h}_l}$ having a common vertex l (Fig. 2.7c) then

$$\cos(\widehat{\mathbf{h}_i\mathbf{h}_l\mathbf{h}_k}) \approx \text{corr}(\ln(X_i/X_l), \ln(X_k/X_l)) .$$

This is useful in the investigation of subcompositions for possible independence.

- (3) *Subcompositional analysis.* The centre O is the centroid (centre of gravity) of the D vertices $\mathbf{h}_1, \dots, \mathbf{h}_D$. Since ratios are preserved under formation of subcompositions, it follows that the biplot for any subcomposition S 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 O_S of these vertices (Fig. 2.7c).
- (4) *Coincident vertices.* If vertices \mathbf{h}_i and \mathbf{h}_k coincide, or nearly so, this means that $\text{var}(\ln(X_i/X_k))$ is zero, or nearly so, and that the ratio X_i/X_k is constant, or nearly so. Then, the two involved parts, x_i and x_k , can be assumed to be redundant (Fig. 2.7d). If the proportion of variance captured by the biplot is not very high, two coincident vertices suggest that $\ln(X_i/X_k)$ is orthogonal to the plane of the biplot, and this might be an indication of the possible independence of that logratio and the two directions of the axis for the CoDa-biplot.

- (5) *Collinear vertices.* If a subset of vertices \mathbf{h}_i , \mathbf{h}_j , and \mathbf{h}_l is approximately *collinear* (i.e., in a straight line), it might indicate that the associated subcomposition (X_i, X_j, X_l) has a biplot that is approximately one-dimensional, which means that the subcomposition has one-dimensional variability, that is, subcompositions (X_i, X_j, X_l) plot approximately along a compositional line (Fig. 2.7d).

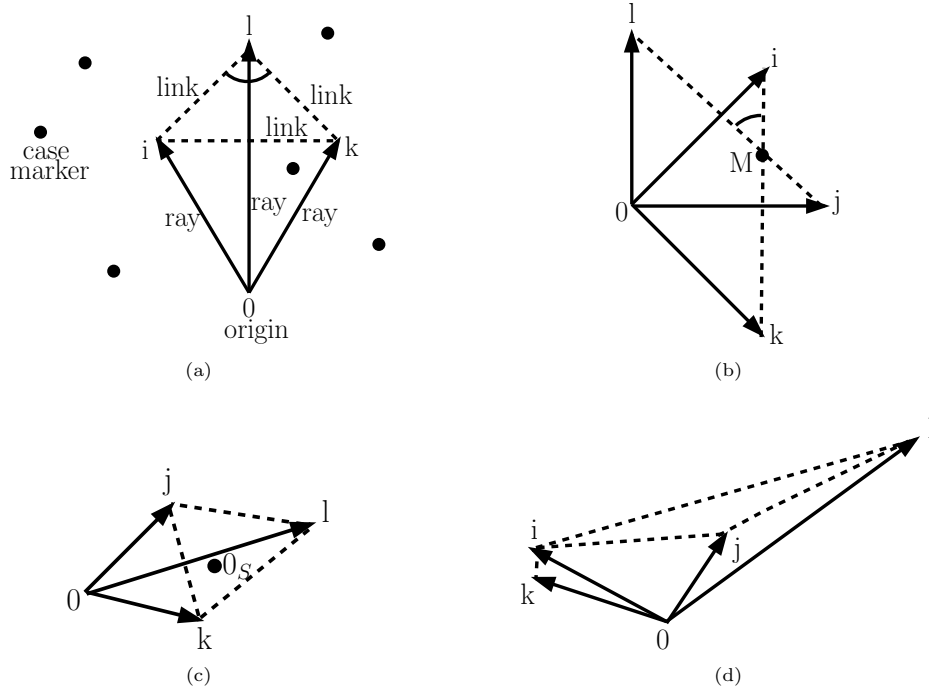


Figure 2.7. Elements for the interpretation of a compositional biplot.

From the aspects of interpretation set out above, it should be clear that the links are the fundamental elements of compositional biplots, rather than the rays which are the fundamental elements in the variation diagrams for unconstrained multivariate data. The lengths of links are (approximately) proportional to the square root of variance of pairwise logratios between single parts, as they appear in the variation matrix. The complete constellation of links provides information about the compositional covariance structure of pairwise logratios and provides hints on subcompositional variability and independence. The interpretation of a CoDa-biplot is concerned with its internal geometry and is unaffected by any rotation or mirror-imaging of the diagram.

Another fundamental difference between the practice of biplots for unconstrained and CoDa-biplot is in the use of data scaling. For an unconstrained data biplot, when there are substantial differences in the variances of the variables due to its scale, the biplot approximation may concentrate its efforts on capturing the nature of the variability of the variable with largest variability and it fails to provide any picture of the pattern of variability within the variables with lower variability.

Since such differences in variances may simply arise because of scales of measurement, a common technique in such biplot applications is to apply some form of individual scaling to the variables (standardization) of the unconstrained vectors. No such individual scaling is necessary for CoDa-biplot because we are interested in the analysis of the relative information (ratios). Indeed, since for any set of constants a_1, \dots, a_D , we have

$$\text{cov}(\ln(a_i X_i / a_j X_j), \ln(a_k X_k / a_l X_l)) = \text{cov}(\ln(X_i / X_j), \ln(X_k / X_l)) ,$$

it is obvious that the covariance structure and, therefore the compositional biplot, are unchanged by any differential scaling or perturbation of the compositions. This, of course, is simply an aspect of the perturbation invariance of measures of dispersion for CoDa.

Example 9.

Figure 2.8 shows the 2-dimensional covariance *clr*-biplot for the 6-part compositions $[T, C, A, R, O, S]$ of the `statisticiantimebudget` CoDa set (see Appendix). The quality is reasonably high because (66.89%) of total variance is retained. However, we should be cautious in our interpretations because it remains approximately 33% of non-accounted variance.

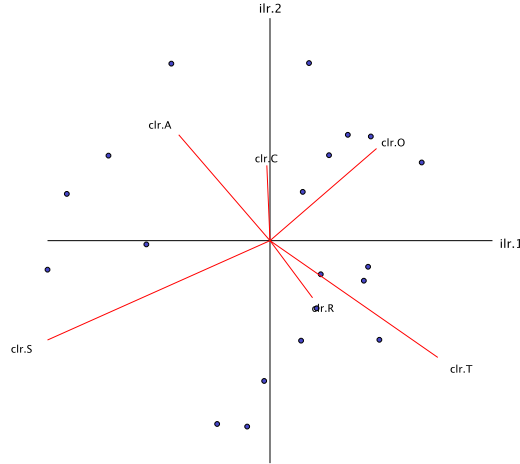


Figure 2.8. Covariance *clr*-biplot of the `statisticiantimebudget` CoDa set.
(Total variance retained: 66.89%)

The variable *clr.S* has largest ray. This is consistent with the diagonal elements of matrix Γ (see Example 3), where $\text{var}(Z_6) = 0.0413$ is the largest value among the *clr*-variables. Analogous interpretation can be done for the shortest rays (*clr.C* and *clr.R*) which respectively *clr*-variances are equal to 0.0187 and 0.0145. No coincident vertices are detected suggesting no redundant parts are present in the data set. Moreover, no collinear or *almost* collinear vertices are shown in the biplot. That is, there is no evidence of a 3-part subcomposition with one-dimensional variability.

There are several links that seems orthogonal. For example, the links $\overline{clr.A} \overline{clr.T}$ and $\overline{clr.O} \overline{clr.S}$ suggest that $\text{corr}(\ln(A/T), \ln(O/S)) \approx 0$. Using CoDaPack for calculating *clr*-coordinates we can obtain the pairwise logratios $\ln(A/T), \ln(O/S)$, whose correlation index is equal to -0.3017 suggesting a slight negative association (non-null). Other potential association can be analyzed following the same procedure. The largest rays in the positive part of the first PC-axis (ilr.1) are *clr.T* and *clr.O*. In the negative part are the rays of *clr.A* and *clr.S*. Consequently, samples with a large positive value in ilr.1 are samples with relative positive values in parts *T* and *O*, and relative small values in *A* and *S*. On the contrary, samples projected in the negative part of ilr.1 correspond to days where the time dedicated to *T* and *O* is relatively small whereas the time for *S* and *A* activities is relatively large. In this way, the closer the sample to the centre of biplot, the similar the sample is to the centre of the data set $\mathbf{g} = [14.75, 10.49, 12.43, 11.29, 23.16, 27.88]$.

2.4.5. Subcompositional analysis ^{††}. Let $S = \{i_1, \dots, i_C\}$ be a subset of the parts $1, \dots, D$ and $\text{sub}(\mathbf{x}; S)$ the subcomposition of the parts S of a composition \mathbf{x} of \mathcal{S}^D . Without loss of generality we may consider $S = \{1, \dots, C\}$, with $C \leq D$.

Let \mathbf{X} be the CoDa set defined in Equation (2.1). Let $\mathbf{X}_S = \{\text{sub}(\mathbf{x}; S) : \mathbf{x} \in \mathbf{X}\}$ be the CoDa set formed by the subcompositions $\text{sub}(\mathbf{x}; S)$ of all compositions of \mathbf{X} . Let \mathbf{T} and \mathbf{T}_S be the variation matrices of \mathbf{X} and \mathbf{X}_S respectively.

A common problem in compositional analysis appears to be marginal analysis in the sense of locating subcompositions of greatest or of least variability. For this purpose, the measure of total variation provides for any subcomposition S the estimate of the ratio

$$\frac{\text{totvar}(\mathbf{X}_S)}{\text{totvar}(\mathbf{X})}$$

as the proportion of the total variation accounted by the subcomposition. According to Equation (2.5) this proportion is equal to

$$(2.14) \quad \frac{D}{C} \times \frac{\sum_{i=1}^C \sum_{j=1}^C \text{var}\left(\ln \frac{X_i}{X_j}\right)}{\sum_{i=1}^D \sum_{j=1}^D \text{var}\left(\ln \frac{X_i}{X_j}\right)}.$$

The denominator of the second factor of Equation (2.14) is the sum of all the elements of \mathbf{T} , and the numerator is the sum of all the elements of \mathbf{T}_S , that is, those elements of \mathbf{T} which are in rows and columns associated with the parts of the subcomposition. Therefore, the proportion of the total variability of \mathbf{X} retained by a subcomposition \mathbf{X}_S is D/C times the ratio of the sum of the elements in the rows and columns of the variation matrix \mathbf{T} associated with the subcompositional parts to the sum of all the elements of \mathbf{T} . Consequently, when the interest is a subcomposition that retains large proportion of variability then one should look for large pairwise log-ratio variances in the variation matrix \mathbf{T} . In addition, because $\text{totvar}(\mathbf{X}) = \text{tr}(\mathbf{T})$ one can look for *clr*-variables with a large ray in the CoDa-biplot, which have large contribution into the decomposition of the total variance.

^{††}This section is an adaptation of [Ait86, Sections 8.5-8.6, p. 196-202], [Ait03, Section 4.8, p. 98-99].

However, we should be cautious because the length of a ray in a biplot depends on the composition (full or subcomposition) used for the *clr*-scores calculation.

Example 10.

Let \mathbf{X} be the CoDa set recorded in the `statisticiantimebudget.cdp` file: 20 6-part compositions $[A, C, O, R, S, T]$ (see [Appendix](#)). We are looking for the 3-parts subcomposition of \mathbf{X} that retains the largest proportion of $\text{totvar}(\mathbf{X}) = 0.1460$. Both [Fig. 2.6a](#) and [2.8](#), and also the variation array and covariance matrix $\mathbf{\Gamma}$ (Example 3) suggest the parts with largest contribution are $[A, O, S, T]$. The following table shows the four possible 3-parts subcompositions formed with $[A, O, S, T]$ and the proportion retained of $\text{totvar}(\mathbf{X})$.

Subcomposition	Total var.	Prop. ret. (%)
$[A, O, S]$	0.0659	45.14
$[A, O, T]$	0.0572	39.18
$[A, S, T]$	0.0796	54.52
$[O, S, T]$	0.0803	55.00

Observe that subcomposition $[O, S, T]$ retains 55% of the total variance but it still remains 45% non-accounted variability, suggesting that the rest of parts A, C and R have a large contribution to the variability of the CoDa set \mathbf{X} . If one wants to retain more variability a 4-part subcomposition should be considered. The subcomposition $[A, O, S, T]$ has a total variance 0.1061 which represents 72.67% of the variability in \mathbf{X} . According the Example 8, this percentage is larger than the proportion accounted by the 2-dimensional *clr*-biplot (66.89%, [Fig. 2.8](#)) but it is lower than the variance retained by the 3-dimensional biplot (82.41%, [Fig. 2.6a](#)).

Activities for Section 2.4

 [Click here to get the activities of this section](#)

2.5. Principal balances

CoDa analysis requires selecting an *olr*-basis with which to work on coordinates. In some cases, the selection of the *olr*-basis is based on a criterion of an expert (*expert driven*). In other cases, this selection is based on the information provided by the data set (*data driven*), for example, based on the CoDa-biplot (*principal component analysis* (PCA)). Compositional PCA provides bases that are, in general, logcontrasts of all the original parts, each with a different weight hindering their interpretation. For interpretative purposes, it would be better to have each basis component as a ratio (*balance*) of the geometric means of two groups of parts, leaving irrelevant parts with a zero weight. This is the role of *principal balances* (PBs), defined as a sequence of *olr*-balances which successively maximize the retained variance in a data set [[Mar+18](#)].

Given an SBP applied to a CoDa set \mathbf{X} and the coordinates (or balances) associated to it, $X_1^*, X_2^*, \dots, X_{D-1}^*$, we can order balances from the one with highest variance to the one with lowest variance: $X_{[1]}^*, X_{[2]}^*, \dots, X_{[D-1]}^*$ (note that such order

it is not necessary the same as the order of the SBP). The PBs are defined as the balances obtained from an SBP such that:

- For balance $X_{[1]}^*$, $\text{var}(X_{[1]}^*)$ is maximum and
- given $X_{[1]}^*, X_{[2]}^*, \dots, X_{[k-1]}^*$, for balance $X_{[k]}^*$, $\sum_{j=1}^k \text{var}(X_{[j]}^*)$ is maximum.

That is, the PBs set $\{X_{[1]}^*, \dots, X_{[D-1]}^*\}$ are a set of orthonormal balances where $\text{var}(X_{[1]}^*) > \text{var}(X_{[2]}^*) > \dots > \text{var}(X_{[D-1]}^*)$ and $\text{totvar}(\mathbf{X}) = \sum_{k=1}^{D-1} \text{var}(X_{[k]}^*)$.

In [Mar+18] three algorithms are presented for creating the PBs: *optimal*, *constrained PC*, and *Ward method*. The *optimal* algorithm to compute PBs requires an exhaustive and recursive search along all the possible sets of *olr*-balances. For example, to find the first PB $X_{[1]}^*$ the variance retained by each of the possible balances using 2, 3, \dots D parts is calculated. The balance that retains a largest proportion of variance is selected as $X_{[1]}^*$. This balance can be in any place of the CoDa-dendrogram from the top to the bottom. The procedure recursively continues *up&down* (*parent&child*) in the balance-dendrogram until the complete SBP is obtained. To reduce computational time of this recursive algorithm, the sets of possible partitions for up to 15 parts are provided. Figure 2.9 shows an example ($D = 8$) that illustrates the sequence followed by the algorithm to construct the PBs. First, the maximum variance associated with the partition $(+1, +1, +1, -1, -1, 0, 0, 0)$ was found and labelled as the *First PB*. According to the parts marked $+1$ or -1 , this partition is split into two partitions. The recursive algorithm applied to the parts marked with $+1$ found that the optimal child partition is $(+1, -1, 0, 0, 0, 0, 0, 0)$ (labelled 2), whose optimal parent partition is $(-1, -1, +1, 0, 0, 0, 0, 0)$ (labelled 3). The recursive algorithm applied to the parts marked with -1 in the First PB found that $(0, 0, 0, +1, -1, 0, 0, 0)$ is the optimal child partition (labelled 4). The list of consecutive optimal parent partitions of the First PB found is formed by the partition $(-1, -1, -1, -1, -1, +1, +1, 0)$ (labelled 5) and the top partition $(-1, -1, -1, -1, -1, -1, -1, +1)$ (labelled 6). When the parts marked with $+1$ in these parent partitions are recursively analyzed the optimal child partition was found to be $(0, 0, 0, 0, 0, +1, -1, -1, 0)$ (labelled 7). Once the SPB is completed the partitions are then sorted according to the variance of their corresponding *olr*-coordinates.

The *constrained PC* and *Ward method* algorithms are faster than the *optimal* but they are suboptimal. Both algorithms are based on approximations of the variance retained, respectively, using a new search for balances following a constrained PC approach and using the Ward method for hierarchical cluster analysis of parts. The Ward method is not a complicated procedure. Indeed, let \mathbf{X}^t the transposed data matrix of \mathbf{X} , that is, \mathbf{X}^t has D rows (the parts) and n columns (the samples). We apply the typical hierarchical cluster analysis with the Aitchison distance to \mathbf{X}^t for making a hierarchical structure of groups of parts. This structure is considered as the SBP for the PBs and the dendrogram is used as the CoDa-dendrogram.

The constrained PCs algorithm follows the *constrained PCs* approach introduced by [Chi2005]. Let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{D-1}$ be the directions of the PCs of the *clr*-scores set provided by the CoDa-biplot. Let $\alpha_1, \alpha_2, \dots, \alpha_{D-1}$ be the corresponding simplified PCs, that is the *clr*-vectors of the balancing elements to be

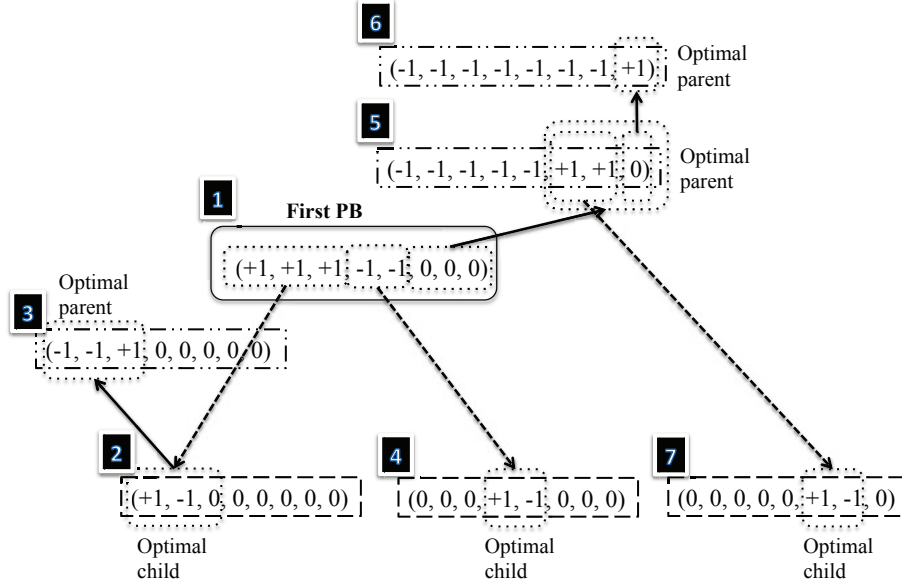


Figure 2.9. Example: optimal PBs algorithm. The label First PB indicates the PB that explains the maximum proportion of variance. Numbers in black rectangles indicate the sequence followed by the algorithm to construct the PBs (see text for more details).

constructed. Following [Chi2005], the components α_{ij} of the i -th vector α_i take only values $-c_1$, 0, and c_2 , such that $\sum_{j=1}^D \alpha_{ij}^2 = 1$ and $\sum_{j=1}^D \alpha_{ij} = 0$. [Chi2005] describe each vector α_i as “...a difference of the average of one set of variables and the average of another set of variables, called a *contrast*”. This is the characteristic of a *clr*-vector of a balancing element in the compositional case. Indeed, let $\mathbf{S} = [s_{ij}]$ be a $(D-1) \times D$ dimensional sign matrix with +1, -1 and 0 entries (0's correspond to parts not included in the partition). Following Chapter 1, the i -th coordinate x_i^* of a composition \mathbf{x} with respect to this basis is equal to the logcontrast

$$(2.15) \quad x_i^* = \sum_{j=1}^D \phi_{ij} \ln x_j.$$

The ϕ_{ij} is defined as follows:

$$(2.16) \quad \begin{aligned} \phi_{ij} &= 0, & \text{if } s_{ij} &= 0; \\ \phi_{ij} &= +\frac{1}{p_i} \sqrt{\frac{p_i \cdot n_i}{p_i + n_i}}, & \text{if } s_{ij} &> 0; \\ \phi_{ij} &= -\frac{1}{n_i} \sqrt{\frac{p_i \cdot n_i}{p_i + n_i}}, & \text{if } s_{ij} &< 0, \end{aligned}$$

where p_i and n_i are the number of parts in the i -th row of \mathbf{S} coded by +1 (*positive*) and -1 (*negative*). Consequently, for the constrained PCs algorithm we take $\alpha_{ij} = \phi_{ij}$, for $i = 1, \dots, D-1$ and $j = 1, \dots, D$.

Given the PC γ_i , the best simplification α_i minimises the angle $\arccos(\gamma_i \cdot \alpha_i^t)$. The idea is: the closer is the PB to the PC the larger variance retained by the PB. The search algorithm starts by identifying the largest positive and negative coefficients of γ_i and sets, respectively, the corresponding elements of α_i to $\pm\sqrt{2}/2$. All the other elements of α_i are forced to zero (Eq. 2.16). This procedure is repeated from three to D coefficients selected by absolute magnitude. Among these $D - 1$ possible balancing elements α_i the closest to γ_i is its best simplification. For example, for $\gamma_1 = (0.5329, -0.0100, -0.2898, 0.1352, 0.3384, -0.7065)$ the first PC of the data set in Example 8, the five candidates for simplification are the balances $(+1, 0, 0, 0, 0, -1)$, $(+1, 0, 0, 0, +1, -1)$, $(+1, 0, -1, 0, +1, -1)$, $(+1, 0, -1, +1, +1, -1)$, and $(+1, -1, -1, +1, +1, -1)$ accordingly normalised (Eq. 2.16). In this case, the closest balance, that is, the balance with the smallest angle, is $\alpha_1 = (\frac{1}{2}, 0, -\frac{1}{2}, 0, \frac{1}{2}, -\frac{1}{2})$ with an angle of 20.96 degrees. Both the calculation of the PCs and the posterior search algorithm of the constrained PCs are very straightforward. These techniques will replace the exhaustive search of the optimal algorithm to find approximate PBs. This step is mainly responsible for the consumption of computational time; thus, the reduction is very relevant.

None of these algorithms are available in the current version of CoDaPack (2.03.01, 2021 July). However, one can *imitate* these algorithms when creating a SBP based on the information provided by the *clr*-biplot.

Example 11.

Let \mathbf{X} be the data matrix of the 6-part compositions $[T, C, A, R, O, S]$ recorded in the `statisticiantimebudget.cdp` file (see Appendix) whose covariance *clr*-biplot is shown in Fig. 2.8. Let $\gamma_1 = (0.5329, -0.0100, -0.2898, 0.1352, 0.3384, -0.7065)$ be the first PC of the CoDa-biplot (see Example 8), which retains 43.84% of the total variance ($\text{totvar}(\mathbf{X}) = 0.1460$). The first option for a PB that approximates γ_1 is the balance $(\sqrt{2}/2, 0, 0, 0, 0, -\sqrt{2}/2)$ (Eq. 2.16). We calculate the *olr*-coordinates for this balance with CoDaPack ($\text{olr}.1 = \sqrt{2}/2(\ln T - \ln S)$) and, afterwards, calculate the variance of these coordinates (0.0512). This variance is 35.07% of $\text{totvar}(\mathbf{X})$. Another option is the balance $(+1, 0, -1, 0, +1, -1)$, that accordingly normalised is $(\frac{1}{2}, 0, -\frac{1}{2}, 0, \frac{1}{2}, -\frac{1}{2})$, that is, $\text{olr}.1 = \frac{1}{2}(\ln T - \ln A + \ln O - \ln S)$ (Eq. 2.16). This balance accounts 39.93% of the total variance ($\text{var}(\text{olr}.1) = 0.0583$) that is a reasonable approximation of the first PC (43.84%). The variance of all possible simplifications of γ_1 can be calculated using this technique for finding the balance retaining the largest proportion of $\text{totvar}(\mathbf{X})$. Assume, for example, that the best approximation is the balance $(+1, 0, -1, 0, +1, -1)$. To complete the SBP we have to find the best *child* and *parent* at each level as regards the variance retained. The balance that can be child are only $(+1, 0, 0, 0, -1, 0)$ and $(0, 0, +1, 0, 0, -1)$, conditional an irrelevant change of sign. The potential parents are $(-1, +1, -1, +1, -1, -1)$, $(-1, 0, -1, +1, -1, -1)$, or $(-1, +1, -1, 0, -1, -1)$. Once the best parent is selected, one can easily complete the SBP.

Sometimes, the analyst are not looking for dimensionality reduction as regards the proportion of variance retained but she/he wants to select an *olr*-basis according the CoDa-biplot. In this case, one can create an SBP inspired by the rays of the *clr*-variables. For example, because in the *clr*-biplot (Fig. 2.8) the rays associated to parts T , R and O are in the positive part of the first PC axis and the other in the negative then the first balance can be associated

to the partition $(+1, -1, -1, +1, +1, -1)$. Taking into account the second PC axis, the two child-balances can be respectively associated to $(-1, 0, 0, -1, +1, 0)$ and $(0, +1, +1, 0, 0, -1)$. To complete the *olr*-basis we can respectively add $(+1, 0, 0, -1, 0, 0)$ and $(0, +1, -1, 0, 0, 0)$ to the SBP. The following table shows the SBP of the PBs and the variance retained:

PB	T	C	A	R	O	S	Variance	Prop. (%)
								Var. Accounted
PB1	+1	-1	-1	+1	+1	-1	0.0492	33.70
PB3	-1	0	0	-1	+1	0	0.0253	17.33
PB5	+1	0	0	-1	0	0	0.0162	11.10
PB2	0	+1	+1	0	0	-1	0.0346	23.70
PB4	0	+1	-1	0	0	0	0.0207	14.18

As expected, the first PB retains lower variance (33.70%) than the first PC (43.84%, see Example 8). In addition, the accumulated variance accounted by the two first PBs $(33.70+23.70= 57.4\%)$ is lower than the variance retained in the *clr*-biplot (Example 8, 66.89%). This effect is the common case for the two first PB created from a CoDa-biplot. However, usually the last PBs account more variance than the last PCs. Note that, in both cases, the total accumulated variance must be equal $\text{totvar}(\mathbf{X}) = 0.1460$.

Activities for Section 2.5

 [Click here to get the activities of this section](#)

2.6. Distributions on the simplex^{§§}

2.6.1. Most relevant distributions.

The Dirichlet distribution is a common model for a random vector $\boldsymbol{\pi}$ of probabilities. From this point of view, a random vector \mathbf{x} , where $\sum_{j=1}^D x_j = 1$ follows a $\text{Dir}(D; \boldsymbol{\alpha})$, if its density function equals

$$f(\mathbf{x}) = \frac{\Gamma\left(\sum_{j=1}^D \alpha_j\right)}{\prod_{j=1}^D \Gamma(\alpha_j)} x_1^{\alpha_1-1} x_2^{\alpha_2-1} \dots x_D^{\alpha_D-1},$$

where Γ is the Gamma function and the real vector $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_D]$, $\alpha_j > 0$, is known as the *concentration* vector because the *closed* vector $\mathcal{C}(\boldsymbol{\alpha})$ informs about the expectation of the parts, that is, $E(x_j) = \frac{\alpha_j}{\sum_k \alpha_k}$. The variance is

$$\text{var}(x_j) = \frac{\alpha_j(\sum_k \alpha_k - \alpha_j)}{(\sum_k \alpha_k)^2(\sum_k \alpha_k + 1)},$$

and the correlation

^{§§}For further detailed information see [Ait86, Sections 3.4 (p. 58-61), 6.1-6.12 (p. 112-130) and 7.3 (p. 143-148)], [PET15, Section 6.3, p. 112-127].

$$\text{corr}(x_i, x_j) = - \frac{\alpha_i \alpha_j}{\sqrt{\alpha_i (\sum_k \alpha_k - \alpha_i) \alpha_j (\sum_k \alpha_k - \alpha_j)}} \quad (i \neq j).$$

Note that if one changes the *location* or concentration vector α then both the variance (*spread*) and the correlation are affected. This fact is a lack of flexibility when compared to other models, such as the Gaussian normal model. In addition, the correlation suggests a particular association between the variables. Unfortunately, this type of association is not present in many practical compositional studies [Ait86]. Figure 2.10 shows the isodensity lines of Dirichlet distribution for $\alpha = (2, 2, 2)$. It is a typical Dirichlet distribution on \mathcal{S}^3 when $\alpha = \alpha \cdot \mathbf{1}_D$, that is, $\alpha_i = \alpha$, for $i = 1, 2, \dots, D$. The resulting distribution is symmetric, that is, equal variance in all variables, and centred in the centre of the simplex. However, Figure 2.10b shows that these isodensity lines are not circles in the logratio space.

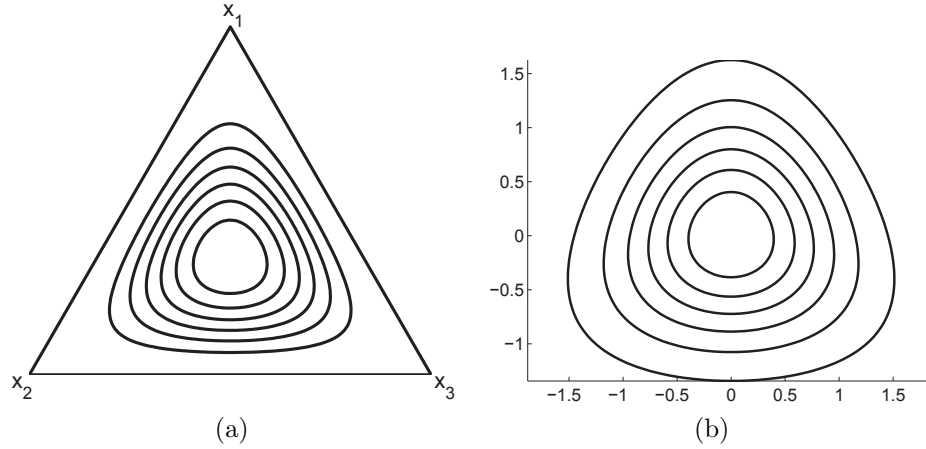


Figure 2.10. Dirichlet isodensity lines on \mathcal{S}^3 : (a) in ternary diagram; (b) *olr*-coordinate representation (default basis).

Furthermore, [Ait86, p. 60] stated that the most important difficulty is that “ $\text{Dir}(D; \alpha)$ has a very strong *implied independence structure*”. In other words, each ratio x_i/x_j is independent of any other ratio x_k/x_m . Once again, this property is not common in practical studies involving CoDa. These limitations of the Dirichlet model suggest that logratio methodology is a more general approach to the analysis of CoDa, when the relative, rather than the absolute, information is of interest.

Recall that most of the basic multivariate methods (e.g. linear discriminant analysis, manova, linear regression, etc) assume multivariate normality. Since CoDa have a particular sample space, a particular definition of normality is needed. Figure 2.11 shows a CoDa set in the ternary where isodensity ellipses of the classical normal distribution, centred on the arithmetical mean, were plotted. Although it is obvious that the model does not fit the data well, recall that the isodensity lines go beyond the sample space of the data. This fact suggests that the classical normal model is not a natural model in the simplex. Consequently, all multivariate methods that assume classical normality could provide erroneous inferences.

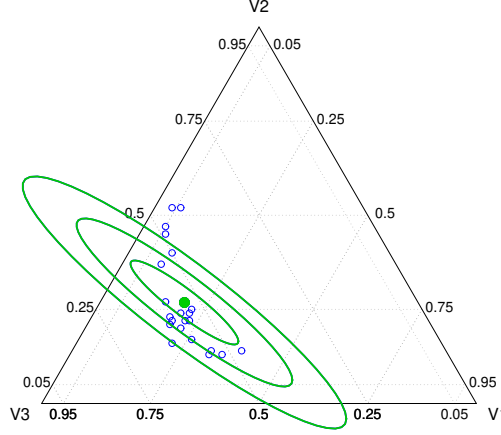


Figure 2.11. CoDa set (blue) in the ternary: isodensity lines (green) of typical multivariate normal.

Let \mathbf{x} be a compositional random vector. We say that \mathbf{x} follows a normal distribution on \mathcal{S}^D if, and only if, the vector of *olr*-coordinates \mathbf{x}^* follows a multivariate normal distribution on \mathbb{R}^{D-1} , that is,

$$\mathbf{x} \sim \mathcal{N}_{\mathcal{S}}(\boldsymbol{\mu}^*, \boldsymbol{\Omega}),$$

a random composition \mathbf{x} normally distributed on \mathcal{S}^D , with parameters $\boldsymbol{\mu}^*$ and $\boldsymbol{\Omega}$, has the density function

$$f_{\mathbf{x}}(\mathbf{x}) = (2\pi)^{-(D-1)/2} |\boldsymbol{\Omega}|^{-1/2} \exp \left[-\frac{1}{2} (\mathbf{x}^* - \boldsymbol{\mu}^*) \boldsymbol{\Omega}^{-1} (\mathbf{x}^* - \boldsymbol{\mu}^*)^t \right]$$

that is, the usual normal density applied to coordinates \mathbf{x}^* . In that case $\boldsymbol{\mu} = \text{olr}^{-1}(\boldsymbol{\mu}^*)$ can be interpreted as the centre $\text{cen}[\mathbf{x}]$ of the random vector \mathbf{x} in the simplex \mathcal{S}^D .

The standard mathematical statistics rely on real analysis, and real analysis is performed on the coefficients with respect to an orthonormal basis in a linear vector space. This approach is also justified from the standpoint of measure theory, but it would divert us too far from practical situations. The definition of normal distribution on the simplex, also known as the logratio-normal distribution or the logistic-normal distribution, is independent of the *olr*-basis chosen to express the coordinates. Moreover, the normal distribution on the simplex could be also defined using the *alr*-coordinates. In this case, the definition coincides with the definition in [Ait86] and we call it the Additive Logistic Normal distribution (ALN).

Let \mathbf{x}^* be the *olr*-coordinates of the random composition $\mathbf{x} \in \mathcal{S}^3$ provided by

$$\mathbf{x}^* = \left[\frac{1}{\sqrt{2}} \ln \left(\frac{x_1}{x_2} \right), \frac{1}{\sqrt{6}} \ln \left(\frac{x_1 x_2}{x_3 x_3} \right) \right].$$

Figure 2.12 shows the isodensity lines of two normal distributions on \mathcal{S}^3 where the centre are, respectively, $\boldsymbol{\mu}^* = (-0.5, -0.5)$ and $\boldsymbol{\mu}^* = (1.5, 1.5)$, and covariance

matrix $\mathbf{\Omega} = \mathbf{I}_2$, the identity matrix. In other words, the isodensity lines are log-ratio *circumferences*.

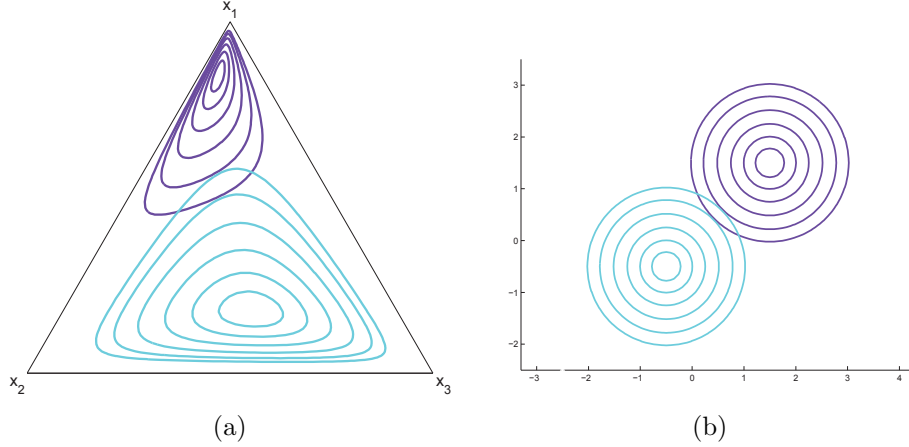


Figure 2.12. Normal isodensity lines on \mathcal{S}^3 : (a) in ternary diagram; (b) *olr*-coordinate representation (default basis).

Let \mathbf{X} be a CoDa set. With the assumption that the pattern of compositional variability is $\mathcal{N}_S(\boldsymbol{\mu}^*, \mathbf{\Omega})$, to estimate the distributional parameters:

- first, select an *olr*-basis and obtain the *olr*-coordinates of the data set \mathbf{X}^* ;
- and next, estimate the mean vector and covariance matrix of \mathbf{X}^* using the standard procedures in \mathbb{R}^{D-1} .

In some cases the multivariate normal distribution cannot properly fit the log-ratio coordinates because of remaining skewness. In these situations, the family of normal skewed distributions could be appropriate. This family, which includes the Gaussian one, can also be applied to CoDa [BMP06]. Indeed, a random D -composition \mathbf{x} follows a skew-normal distribution on \mathcal{S}^D when its coordinates \mathbf{x}^* follows a multivariate skew-normal distribution, $\mathbf{x}^* \sim \mathcal{SN}(\boldsymbol{\mu}^*, \mathbf{\Omega}, \boldsymbol{\lambda})$. Recall that the parameter $\boldsymbol{\lambda}$ controls the skewness. Figure 2.13 shows the isodensity lines of a typical skew-normal distribution on \mathcal{S}^3 . The two skew normal distributions are respectively centred in $\boldsymbol{\mu}^* = [-0.5, -0.5]$ and $\boldsymbol{\mu}^* = [1.5, 1.5]$, both with covariance matrix $\mathbf{\Omega} = \mathbf{I}_2$ and $\boldsymbol{\lambda} = [-1, 2]$. When one compares this figure with the Fig. 2.12 we detect the skewness of the lines, particularly in the coordinate space. Readers interested in knowing more about this distribution can find detailed information in [MP07].

2.6.2. Normality tests. We will work on *olr*-coordinates to test if a normal distribution on the simplex fits well a CoDa set. In other words, the test for normality on \mathcal{S}^D is

H_0 : the samples come from a normal distribution on \mathcal{S}^D

H_1 : the samples do not come from a normal distribution on \mathcal{S}^D

which is equivalent to test

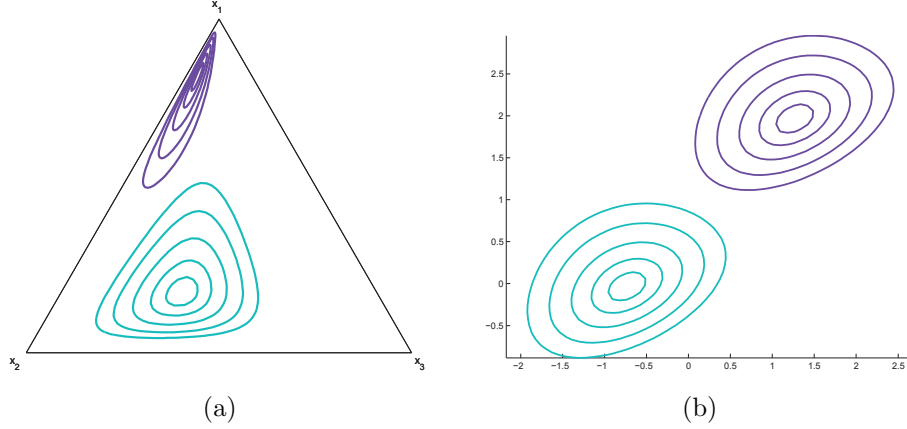


Figure 2.13. Skew-normal isodensity lines on \mathcal{S}^3 : (a) in ternary diagram; (b) olr -coordinate representation (default basis).

H_0 : the olr -coordinates come from a multivariate normal distribution on \mathbb{R}^{D-1}

H_1 : the olr -coordinates do not come from a normal distribution on \mathbb{R}^{D-1}

Among the numerous contrasts tools to test multivariate normality we focus here on the radius test.

2.6.2.1. Radius test. This test is based on the property that, under normality, the squared Mahalanobis distances (d_{MH}^2) from the samples of a data set \mathbf{X} ($n \times p$) to the arithmetic mean vector are chi-squared distributed with p degrees of freedom, where p is the number of variables. Let $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Omega}}$ respectively be the arithmetic mean vector and covariance matrix (i.e., the maximum likelihood estimates of center and variance). The squared Mahalanobis distance from a sample \mathbf{x}_i to the center is

$$d_{MH}^2(\mathbf{x}_i, \hat{\boldsymbol{\mu}}) = (\mathbf{x}_i - \hat{\boldsymbol{\mu}}) \hat{\boldsymbol{\Omega}}^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})^t \quad i = 1, \dots, n ;$$

The Mahalanobis distance from a sample to the centre is equivalent to the norm of the sample when the data set \mathbf{X} has been previously *spherised* $((\mathbf{X} - \hat{\boldsymbol{\mu}}) \hat{\boldsymbol{\Omega}}^{-1/2})$. That is, the samples are centred and its covariance matrix is equal to the identity matrix. Recall that when the covariance matrix is proportional to the identity matrix we have a spherical distribution. This case inspires the name *radii* to the distances from samples to the centre. For CoDa, the Mahalanobis distances are calculated using the olr -coordinates set \mathbf{X}^* . Importantly, the link of $d_{MH}^2(\mathbf{x}_i^*, \hat{\boldsymbol{\mu}}^*)$ with the spherical distribution indicates that Mahalanobis distances for CoDa are invariant under a change of olr -basis.

We are presenting a radius test based on the analysis of the empirical distribution function (EDF). To test normality on the simplex, the test is applied to a olr -coordinates set \mathbf{X}^* . Among all the possible EDF tests we have selected the tests based on Anderson-Darling (Q_a), Cramer-von Mises (Q_c) and Watson (Q_w) statistics, which are currently available in CoDaPack.

The procedure of a radius test has the following steps:

- (1) compute the maximum likelihood estimates $\hat{\boldsymbol{\mu}}^*$ and $\hat{\boldsymbol{\Omega}}$ (sample mean and covariance matrix of \mathbf{X}^*);
- (2) compute the squared Mahalanobis distances:

$$d_i^2 = (\mathbf{x}_i^* - \hat{\boldsymbol{\mu}}^*) \hat{\boldsymbol{\Omega}}^{-1} (\mathbf{x}_i^* - \hat{\boldsymbol{\mu}}^*)^t \quad i = 1, \dots, n;$$

- (3) obtain $p_i = F(d_i^2)$, $i = 1, \dots, n$ where F is the chi-squared distribution function, with $D - 1$ degrees of freedom (χ_{D-1}^2);
- (4) rearrange p -values in ascending order, $p_{(i)}$, $i = 1, \dots, n$;
- (5) calculate the EDF-statistics Q_a, Q_c, Q_w and modify them:
 - $Q_a^* = Q_a$;
 - $Q_c^* = (Q_c - \frac{0.4}{n} + \frac{0.6}{n^2}) (1 + \frac{1}{n})$;
 - $Q_w^* = (Q_w - \frac{0.1}{n} + \frac{0.1}{n^2}) (1 + \frac{0.8}{n})$;
- (6) compare values obtained with the corresponding critical values of the EDF-statistics for a different significance level (α):

mod. statistic	signif. level ($\alpha\%$)			
	10	5	2.5	1
Q_a^*	1.933	2.492	3.070	3.857
Q_c^*	0.347	0.461	0.581	0.743
Q_w^*	0.152	0.187	0.221	0.267

- (7) For a given significance level (α) and a test statistic (Anderson-Darling, Cramer-von Mises, or Watson), the null hypothesis H_0 will be rejected if the corresponding value Q^* obtained is larger than its critical value.

Obviously, the perfect situation when we analyse the normality would be that the three EDF tests agree in their decision. Unfortunately, this is not always the case. For those cases when they do not agree, the researcher is the one who has to take a *critical* and personal decision. Fortunately, these tools can be completed with the usual Q - Q plots, which provide a visual representation of each test. Another possibility is the generalisation of the Shapiro-Wilk test for multivariate normality. This test, available in R program, should be applied to the *olr*-coordinates data set \mathbf{X}^* .

2.6.2.2. Marginal tests. In some cases, when the null hypothesis H_0 is rejected for the normality on the simplex, the analyst could be interested in investigating if a particular *olr*-coordinate is in fact responsible for this lack of multivariate normality. In this scenario, some authors (e.g. [Ait86, p. 145]) also recommend analysing the normality of the *marginals*. That is, they recommend analysing if the distribution formed by each single marginal is normally distributed. Popular univariate normality tests are included in CoDaPack. The interested reader is referred to [Ait86] for further information.

Unlike the multivariate radius test, the univariate normality tests have one difficulty: they depend on the *olr*-basis used to calculate the coordinates. This feature, which could be annoying, could also be helpful when one is analysing which part contributes to the lack of normality of the coordinates.

Example 12.

Assuming a normal distribution on the simplex, Fig. 2.14 shows the predictive regions for 3-part subcomposition \mathbf{X}_s of non-academic tasks $[A, O, S]$ for the `statisticiantimebudget` CoDa set (see Appendix). The Gaussian predictive regions seem to fit reasonably well to the data. The results of radius and univariate marginal tests are shown in the table:

Normality tests						
	A^{2*}	p-value	W^{2*}	p-value	U^{2*}	p-value
Radius test						
\mathbf{X}^*	0.4738	>0.15	0.0584	>0.15	0.0653	>0.15
Marginal tests						
X_1^*	0.4083	>0.15	0.0587	>0.15	0.0587	>0.15
X_2^*	0.5799	[0.10, 0.15]	0.0779	>0.15	0.0692	>0.15

The three multivariate tests indicates *fail to reject* the null hypothesis because the p-values are > 0.15 . That is, the normal in the simplex model fits well the 3-part subcomposition. The univariate normality tests calculated for the *olr*-marginals (default partition) also suggests that both can be reasonably modeled by a normal distribution.

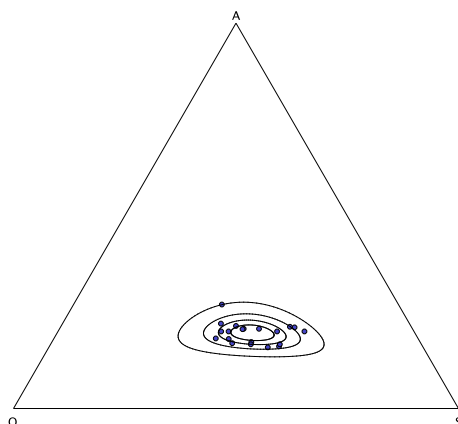


Figure 2.14. Gaussian predictive regions ($\alpha=0.25, 0.5, 0.75, 0.95$) in the ternary diagram of the 3-part subcomposition $[A, O, S]$ of the `statisticiantimebudget` CoDa set.

Activities for Section 2.6

 [Click here to get the activities of this section](#)

The chapter's key concepts

- ✓ The closed geometric mean is the measure of central tendency for a CoDa set.
- ✓ The covariance structure of a CoDa set can be described in different ways from the variation matrix or from the covariance matrices of the *clr*, *alr* and *ilr* transformed data set.
- ✓ The total variance is an overall measure of the total relative variability of a CoDa set.
- ✓ CoDa-dendrogram summarizes the coordinates created by an SBP.
- ✓ The biplot of the *clr*-transformed data set allows us to perform a first exploratory analysis of a CoDa set so as to discover significant statistical relationships between logratios of the parts and potential clusters of ‘similar’ compositions.
- ✓ Principal balances algorithms provide a particular type of SBP created by a data-driven procedure.
- ✓ Normal distribution on the simplex is a probability model that is consistent with simplicial geometry.

Specific references in this Chapter 2

- [Ait82] J. Aitchison, *The statistical analysis of compositional data (with discussion)*, J. R. Statist. Soc. B **44**, 1982, no. 2, 139–177.
- [Ait03] J. Aitchison, *A concise guide to compositional data analysis*, Available from <http://www.compositionaldata.com>, 2003.
- [Ait97] J. Aitchison, *The one-hour course in compositional data analysis or compositional data analysis is simple*, Proceedings of IAMG’97 — The third annual conference of the International Association for Mathematical Geology (V. Pawłowsky-Glahn, ed.), vol. I, II and addendum, 1997, pp. 3–35.
- [BMP06] A. Buccianti, G. Mateu-Figueras and V. Pawłowsky-Glahn, *Frequency distributions and natural laws in geochemistry*. In: Pawłowsky, V. and Buccianti, A. (eds) *Compositional Data Analysis: Theory and Applications*. Chichester (UK), John Wiley & Sons. Chapter 12, p. 175–189.
- [Chi2005] H.A. Chipman and H. Gu, *Interpretable dimension reduction*, Journal of Applied Statistics **32**, 2005, 969–987.
- [DBB06] J. Daunis-i-Estadella, C. Barceló-Vidal and A. Buccianti, *Compositional Data Analysis: Theory and Applications*, ch. Exploratory compositional data analysis, pp. 161–174, John Wiley & Sons, Ltd, Chichester, UK, 2011. DOI:10.1002/9781119976462.ch 11

- [DTM11] J. Daunis-i-Estadella, S. Thió-Henestrosa and G. Mateu-Figueras, *Including supplementary elements in a compositional biplot*, Computers & Geosciences **37**, 2011, 696–701.
- [EP06] J.J. Egozcue and V. Pawlowsky-Glahn, *Compositional data in the geosciences*, vol. 264, ch. Simplicial geometry for compositional data, pp. 145–159, Geological Society, London, 2006.
- [EB11] J.J. Egozcue, C. Barceló-Vidal, J.A. Martín-Fernández, E. Jarauta-Bragulat, J.L. Díaz-Barrero and G. Mateu-Figueras, *Compositional Data Analysis: Theory and Applications*, ch. Elements of simplicial linear algebra and geometry, pp. 141–157, John Wiley & Sons, Ltd, Chichester, UK, 2011. DOI:10.1002/9781119976462.ch 11
- [EPE02] H. von Eynatten, V. Pawlowsky-Glahn, and J.J. Egozcue, *Understanding perturbation on the simplex: a simple method to better visualise and interpret compositional data in ternary diagrams*, Mathematical Geology **34**, 2002, 249–257.
- [Gab71] K.R. Gabriel, K.R. *The biplot graphic display of matrices with application to principal component analysis*, Biometrika **58** (1971), no. 3, 453–467, 7. DOI:10.1093/biomet/58.3.453
- [GH96] J.C. Gower and D.J. Hand, *Biplots*, Chapman & Hall, London (UK), 1996, 277 p.
- [MB01] J. A. Martín-Fernández and M. Bren, *Some Practical Aspects on Multidimensional Scaling of Compositional Data*, Proceedings of IAMG’01 — The annual conference of the International Association for Mathematical Geology, Cancun (Mexico), CD-ROM, 16p.
- [Mar+18] J. A. Martín-Fernández, V. Pawlowsky-Glahn, J.J Egozcue, and R. Tolosona-Delgado, *Advances in Principal Balances for Compositional Data*, Mathematical Geosciences, **50** (2018), no. 3, 273–298.
- [Mar+20] J. A. Martín-Fernández, J. J. Egozcue, R. A. Olea, and V. Pawlowsky-Glahn, *Units recovery methods in compositional data analysis*, Natural Resources Research (2020). Doi: 10.1007/s11053-020-09659-7.
- [MP08] G. Mateu-Figueras and V. Pawlowsky-Glahn, *A critical approach to probability laws in geochemistry*, Mathematical Geosciences, **40** (2008), no. 5, 489–502.
- [MP07] G. Mateu-Figueras and V. Pawlowsky-Glahn, *The skew-normal distribution on the simplex*, Communications in Statistics-Theory and Methods, **36** (2007), no. 9, 1787–1802.
- [MMP11] G.S. Monti, G. Mateu-Figueras and V. Pawlowsky-Glahn, *Notes on the Scaled Dirichlet Distribution*, In: Pawlowsky-Glahn, V. and Buccianti (eds), *Compositional Data Analysis: Theory and Applications*, Chichester (UK), John Wiley & Sons, 128–138, 2011.
- [VG09] J.A. Villasenor-Alva and E. Gonzalez-Estrada, *A generalization of Shapiro-Wilk’s test for multivariate normality*, Communications in Statistics: Theory and Methods **38** (2009), no. 11, 1870–1883.

Data pre-processing: irregular data

Contents

- 3.1 Missing data
 - 3.1.1 The logratio EM algorithm for missing CoDa
- 3.2 Essential zeros
 - 3.2.1 Modelling essential zeros
 - 3.2.2 On analyzing the pattern of zeros
- 3.3 Count zeros
- 3.4 Censored data: rounded zeros
- 3.5 Dealing with missing values and zeros
- 3.6 Potential outliers

Objectives

- ✓ To deal with the most common irregular data in CoDa: missing data, values below detection limit and zeros.
- ✓ To distinguish the type of zeros and accordingly decide the procedure for dealing with them.
- ✓ To know how to detect potential outliers in CoDa.

In this chapter we present the techniques to deal with no-common data, that is, data that are unusual and have special features. We call these data *irregular data*. We consider several types of irregular data: non-available data (*missing data*), values below detection limit (*bdl*), zeros and outliers. The analysis and treatment of irregular data must be done before the statistical analysis (cluster, regression, MANOVA, etc) of the data. This step is known as *data pre-processing* or simply *pre-processing*. Because each type of irregular data requires its own specific pre-processing we introduce the particular techniques for doing so in following sub-sections. In any case, for this section and for the remainder of the chapter, we assume that groups in the data set do not exist. In other words, if there are groups in the data set then the techniques presented in this chapter should be applied separately to each group. In the next chapter we will introduce some basic multivariate techniques to deal with the groups of a data set.

3.1. Missing data

The first type of irregular data are the non-available data, usually labelled as “NA” as an entry in the incomplete data matrix. It is important to observe that there are different types of missing data because each type will need its own particular treatment. Let \mathbf{x} be an incomplete multivariate sample which can be split into two parts, *observed* and *missing*, that is, $\mathbf{x}=(\text{observed part}, \text{missing part})=(\mathbf{x}_{obs}, \mathbf{x}_{mis})$.

According to [LR87], there are three different types of missing data or missingness mechanism:

- Missing Completely At Random (**MCAR**): \mathbf{x}_{mis} are a simple random sample of all data values. Missingness does not depend on the data values.
Example: in a questionnaire, the accidental omission of an answer.
- Missing At Random (**MAR**): the probability that one value is missing depends on the \mathbf{x}_{obs} part but not on \mathbf{x}_{mis} .
Example: in a questionnaire, the probability of omitting an answer depends on the answer to other questions.
- Not Missing At Random (**NMAR**): the probability that one value is missing depends also on the \mathbf{x}_{mis} part.
Example: a question on a questionnaire has been deliberately skipped by the participant.

In the particular case of CoDa we can distinguish other simple cases (Table 3.1). In the next section we justify why the most usual NMAR value is the rounded zero value (e.g. Obs. 1). Furthermore, observe that the case when only one value is randomly missing in a closed composition where the sum of the observed part is less than κ (constant constraint sum) has an easy solution: impute the residual part to get the constraint sum value. Consequently, we deal with more complicated situations where the observed part \mathbf{x}_{obs} holds the constraint sum (e.g. Obs. 2) or we have more than one missing part (e.g. Obs. 3). Note that cases such as the Obs. 2 are equivalent to the cases where the samples are not closed. That is, when we have a non-closed sample with missing values and we apply the closure operation using the sum of \mathbf{x}_{obs} as the denominator, the result will be a vector similar to Obs. 2.

Table 3.1. Three *irregular* 5-part compositions

		x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	sum
Obs. 1	\mathbf{x}_1	0.0000	0.1250	0.1237	0.7253	0.0260	1
Obs. 2	\mathbf{x}_2	0.1304	0.3151	NA	0.2002	0.3543	1
Obs. 3	\mathbf{x}_3	0.1963	NA	NA	0.4819	0.0114	< 1
Obs. 4	\mathbf{x}_1	0.1430	0.2025	0.0736	0.1576	0.4233	1

There are three common strategies for handling NA:

- Listwise deletion: the entire composition is excluded from the analysis if any single value is NA. The analysis is only run on compositions which have a complete set of data.
- Pairwise deletion: compositions that contain some NA are used. For example, computed means are based on all available data for each part. Log-ratio variances are based on all data available for each pair of parts.
- Imputation: since deletion causes bias, we replace the NA of the data set \mathbf{X} to obtain a *complete data matrix* \mathbf{RX} .

Recall that our final purpose consists of performing a multivariate analysis which does not work when samples have missing values (e.g. cluster analysis). Consequently, at this stage (*preprocessing* data) our objective is to replace the missing values of the incomplete data set \mathbf{X} to obtain a complete data matrix \mathbf{RX} . Which imputation method should we use? It depends on how the NA is generated, that is, the missingness mechanism.

In the following we recommend a treatment of the missing part \mathbf{x}_{mis} depending on the scenario:

- (1) MCAR missingness and *few* missing values: *multiplicative* replacement.
Consider the case of a CoDa where the *few* missing values are the MCAR type. NB: the adjective *few* is a debatable issue and no unique opinion exists and no exact solution has been provided in the literature. For example, in the next section, when one is dealing with rounded zeros (NMAR values) the adjective *few* is empirical quantified as less than 10% of the data matrix entries. Let \mathbf{x}_i be an incomplete composition in \mathbf{X} . According to [MBP03], we replace \mathbf{x}_i with a composition $\mathbf{r}_i \in \mathcal{S}^D$ in data set \mathbf{RX} without missing values using the expression

$$(3.1) \quad r_{ij} = \begin{cases} m_{ij} & , \text{ if } x_{ij} \text{ is NA;} \\ x_{ij} \cdot \frac{(\kappa_i - \sum_{k|x_{ik} \text{ NA}} m_{ik})}{\sum_{k|x_{ik} \text{ non-missing}} x_{ik}} & , \text{ if } x_{ij} \text{ is non-missing,} \end{cases}$$

where the imputed value m_{ij} is a *free* choice made by the researcher and κ_i is the total sum desired for the composition \mathbf{r}_i . The total sum κ can be the same for all the compositions (*closed data*) or it can vary across the samples (*non closed data*). An example for the latter case is data in an election, where the total is the number of potential voters. The choice of m_{ij} could be based on

her/his knowledge about the part, the sample, the data set or external information related to the study (*cold deck*). Other common strategies consist of imputing values based on univariate statistics of the observed part (*hot deck*), e.g. the median or the geometric mean. Note in (3.1) that the *multiplicative* modification of non-missing values preserves the ratios between parts and the total sum of the vector, that is,

$$\frac{r_{ij}}{r_{ik}} = \frac{x_{ij}}{x_{ik}} ; \quad \sum_{j=1}^D r_{ij} = \kappa_i.$$

Example 1. Consider the samples from Table 3.1. We want to replace the missing values of Obs. 2 and Obs. 3 using the formula (3.1) where the imputed value is equal to the geometric mean of the observed value in the part.

	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	sum non-missing
\mathbf{x}_1	0.0000	0.1250	0.1237	0.7253	0.0260	1.0000
\mathbf{x}_2	0.1304	0.3151	NA	0.2002	0.3543	1.0000
\mathbf{x}_3	0.1963	NA	NA	0.4819	0.0114	0.6896
\mathbf{x}_4	0.1430	0.2025	0.0736	0.1576	0.4233	1.0000

Firstly, we simply *replace* missing values in \mathbf{X} by the geometric mean of the part. For example, assume that m_j is the geometric mean of the non-missing values x_{ij} in the j -th column of the full data set \mathbf{X} :

1st step		$m_2 =$ $g(x_{i2})$	$m_3 =$ $g(x_{i3})$				
	x_{i1}^*	x_{i2}^*	x_{i3}^*	x_{i4}^*	x_{i5}^*	sum	sum NA
\mathbf{x}_1^*	0.0000	0.1250	0.1237	0.7253	0.0260	1.0000	0.0000
\mathbf{x}_2^*	0.1304	0.3151	0.0954	0.2002	0.3543	1.0954	0.0954
\mathbf{x}_3^*	0.1963	0.1998	0.0954	0.4819	0.0114	0.9848	0.2952
\mathbf{x}_4^*	0.1430	0.2025	0.0736	0.1576	0.4233	1.0000	0.0000

Second, we *multiplicatively* modify the observed parts:

$$r_{ij} = x_{ij} \cdot \frac{(\kappa - \sum_{k|x_{ik} \text{ NA}} m_{ik})}{\sum_{k|x_{ik} \text{ non-missing}} x_k},$$

where $\kappa = 1$, a common closure constant.

2nd step						
	r_{i1}	r_{i2}	r_{i3}	r_{i4}	r_{i5}	sum
\mathbf{r}_1	0.0000	0.1250	0.1237	0.7253	0.0260	1.0000
\mathbf{r}_2	0.1180	0.2850	0.0954	0.1811	0.3205	1.0000
\mathbf{r}_3	0.2006	0.1985	0.0954	0.4925	0.0117	1.0000
\mathbf{r}_4	0.1430	0.2025	0.0736	0.1576	0.4233	1.0000

- (2) MCAR with *many* missing values or MAR: *logratio* EM-replacement.

When the number of missing values (MCAR or MAR) is large applying a more sophisticated method is recommended. The method should use the association of that part with the other. One of the most common methods is the Expectation and Maximization algorithm (EM) ([LR87]). This parametric method (multivariate normality is assumed) is an iterative method based on

the likelihood function. We will apply EM-algorithm to CoDa in a framework of the logratio methodology.

3.1.1. The logratio EM algorithm for missing CoDa. Here the situation can be split into two different scenarios:

- If there is one part without missing data:
 - (1) calculate *alr*-coordinates of the data: $\mathbf{Y} = \text{alr } \mathbf{X}$, using a part without NA as a denominator, where missing values in \mathbf{X} are transformed in NA in \mathbf{Y} ;
 - (2) replace *alr*-missing values in \mathbf{Y} with a typical EM-algorithm: $\mathbf{Y} \rightarrow \mathbf{RY}$;
 - (3) back-transform the data: $\mathbf{RX} = \text{alr}^{-1}(\mathbf{RY})$;
 - (4) use the values in \mathbf{RX} to impute the NA in \mathbf{X} . Making use of the fact that the relative ratios between components must be preserved, the estimated nondetect in the scale of the composition, r_{ij}^* , can be recovered as $r_{ij}^* = r_{ij} \cdot \frac{x_{ik}}{r_{ik}}$, where x_{ik} is the value of the originally observed component k .
- If all parts have missing values:
 - (1) initially, impute missing values with the multiplicative replacement: $\mathbf{X} \rightarrow \mathbf{RX}_0$;
 - (2) express the data using *olr*-coordinates: $\mathbf{RY}_0 = \text{olr}(\mathbf{RX}_0)$. Select the basis so as to minimize the distortion produced by the initial imputed values;
 - (3) replace *olr*-missing values with a typical EM-algorithm: $\mathbf{RY}_0 \rightarrow \mathbf{RY}$;
 - (4) back-transform the data: $\mathbf{RX} = \text{olr}^{-1}(\mathbf{RY})$.
 - (5) use the values in \mathbf{RX} to impute the NA in \mathbf{X} . Multiplicatively modify the imputed values to express the estimated nondetect in the original scale of the composition.

Because we can construct different *olr* basis, the best option is a basis so as to cause minimal distortion. Explicitly, each D -part composition $\mathbf{x} = [x_1, \dots, x_D]$ can be associated with another composition $\mathbf{x}^{(l)} = [x_1^{(l)}, \dots, x_l^{(l)}, x_{l+1}^{(l)}, \dots, x_D^{(l)}]$ resulting from a *decreasing* sorting of the parts x_1, \dots, x_D in relation to the number of missing values in the parts, that is, the part with more missing values is moved to the first position.

The *olr*-coordinates

$$y_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^D x_j^{(l)}}}, \quad i = 1, \dots, D-1,$$

express the compositions $\mathbf{x}^{(l)}$ as $(D-1)$ -dimensional real vectors $\mathbf{y}^{(l)} = [y_1^{(l)}, \dots, y_{D-1}^{(l)}]$, $l = 1, \dots, D$.

R-packages *Amelia* and *VIM* impute missing data using different methods, in particular, the EM-algorithm. R-package *zCompositions* and package *CoDaPack* include this algorithm for CoDa.

Under log-ratio normality, in its k th E-step, the EM-algorithm replaces the logratio missing values with the regression model

$$\hat{r}y_j^{(k)} = \hat{\beta}^{(k)} \cdot \mathbf{r}\mathbf{y}_{-j}^{(k)t}.$$

A robust-regression model could be applied as well, so as to minimize the effect of potential outliers. This option is supported by *zCompositions* and CoDaPack. Hereinafter *robust* refers to the methods that emulate typical statistical methods, but which are not unduly affected by potential outliers [MMY06].

Example 2. A simple evaluation of performance of logratio EM replacement. Let \mathbf{X} be a geochemical CoDa [PMO14] in \mathcal{S}^5 without missing values in its 229 samples:

- we forced (randomly) MCAR in \mathbf{X} , where 69.87% of samples have at least one NA;
- we replaced the missing data in \mathbf{X} using the log-ratio EM algorithm to obtain a complete data set \mathbf{RX} .

This presence/absence of missing values can be summarized in a “pattern” plot (Fig. 3.1) available in the submenu *ZPatterns plot* at the menu *Irregular data* in CoDaPack. The table shows the different pattern observed in a sample, where the blue cells indicate NA values. The marginal bar plots represent the profiles of the distributions of NA incidence. The distribution of missing values in each column is represented at the top. For example, the largest number of NA (22.71%) is collected in the part V. However, the distribution suggests a uniform distribution consistently with the type MCAR of missing values. The right marginal bar plot shows the distribution of the patterns. For example, the first bar indicates that 30.13% of samples have no NA.

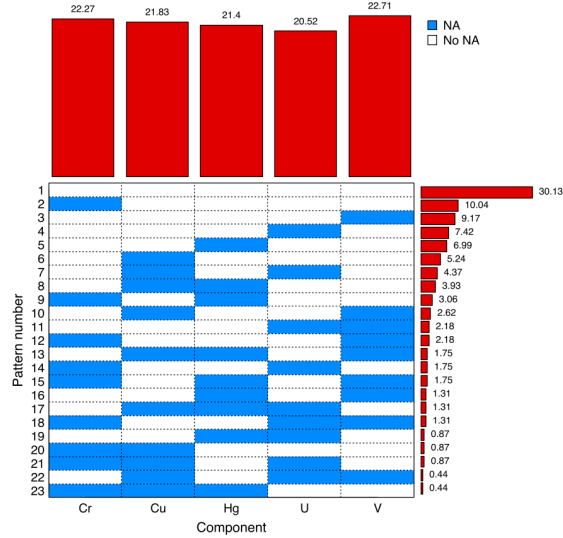


Figure 3.1. Pattern of forced missing values in the data set \mathbf{X} .

Figure 3.2 shows the performance of missing imputation using the EM algorithm with log-ratio methodology. The covariance *clr*-biplot in Figure 3.2a shows the original data set without missing data. There the blue circles are the samples in \mathbf{X} that were randomly selected to force some missing values. Figure 3.2b shows the resulting data set \mathbf{RX} after the imputation of the missing values. We state that the treatment worked reasonably well, in particular the distortion of the associations between the parts are minimal because the rays in both *clr*-biplots shows a similar pattern when an axes rotation is considered.

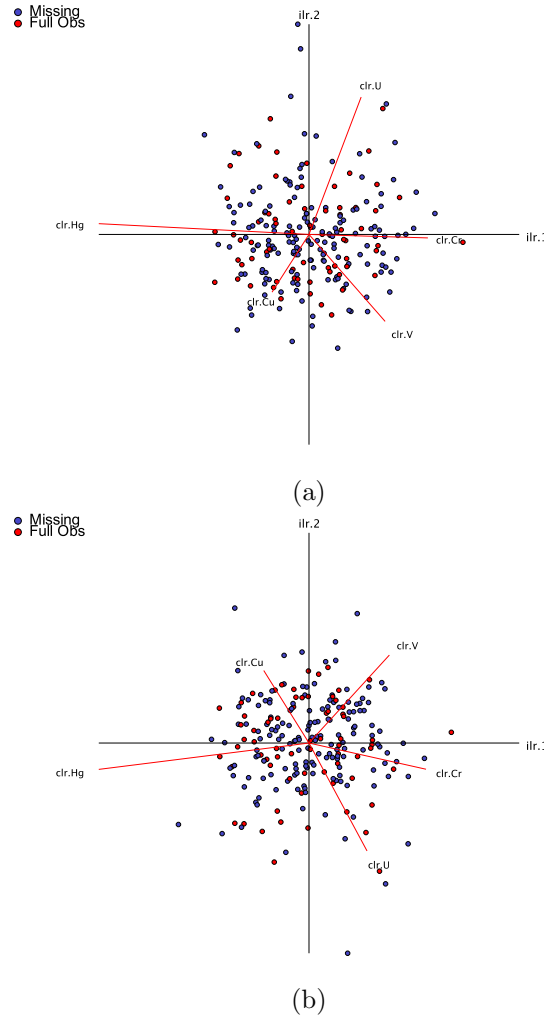



Figure 3.2. Covariance *clr*-biplot: effects of logratio EM replacement. (a) Original data set \mathbf{X} (75.95% variance retained); (b) Replaced data set \mathbf{RX} (79.08% variance retained).

Note that the second PC-axis in the biplot of Figure 3.2b has changed the sign as regards the second axis in Figure 3.2a, that is, there is a symmetry as regards the horizontal axis.

Activities for Section 3.1

 [Click here to get the activities of this section](#)

3.2. Essential zeros

In CoDa, ratios as well as logs require complete data matrix with nonzero entries. A zero and a nondetect value can be present for various reasons. In general, we consider three types of zeros:

- *essential* zeros: in continuous or discrete parts;
- *count* zeros: in discrete parts;
- *rounded* zeros: in continuous parts.

The techniques presented in this section are available in the R-package *zCompositions* and in the package CoDaPack.

An essential or absolute zero in a part of a sample is a *true* zero, that is, not a value recorded as zero because it is a value below a detection limit or because the time in an observational study was too short to observe this part. These absolute zeros are indicators of different subgroups within the data set. In consequence it does not make sense to replace these zeros by small values. We should admit that nowadays, there is not general methodology for dealing with essential zeros in a CoDa-analysis.

3.2.1. Modelling essential zeros. The more promising approach was proposed in [AK03] where a binomial conditional logratio normal model was suggested. The approach was developed in [BB16].

For example, let

$$\mathbf{X} = \begin{bmatrix} 0.25 & 0.25 & 0 & 0 & 0.5 \\ 0.1 & 0.1 & 0.3 & 0.1 & 0.4 \\ 0 & 0.5 & 0 & 0.15 & 0.35 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.3 & 0.1 & 0.1 & 0 & 0.5 \end{bmatrix}$$

be the data set with essential zeros. Let

$$\mathbf{U} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 0 & 1 \end{bmatrix}$$

be the *incidence* matrix, that is, a matrix where its entry equal to 1 means that the corresponding entry in \mathbf{X} is not zero. According to [AK03, BB16], in some conditions, it could be possible to make inferential calculations if we assume that \mathbf{x} is a normally distributed compositional random vector logratio and \mathbf{u} a random vector of Bernoulli distribution, for the *incidence* matrix \mathbf{U} . In this case, the likelihood function is

$$L(\theta, \xi, \Sigma) = \prod_{i=1}^n P[\mathbf{u}_i | \theta_{J(i)}] \cdot f(\mathbf{x}_{iJ(i)} | \xi_{J(i)}, \Sigma_{J(i)}),$$

where subindex i refers to the composition (*row*) and the corresponding subindex $J(i)$ represents the set of parts (*columns*) with non-zero values in the i th composition. In other words, the random vector \mathbf{u} is used to model the different zero patterns and, for each different pattern a log-ratio normal model is assumed.

To model log-ratio coordinates conditioned to distribution of zero patterns one has to estimate the parameters $\theta_{J(i)}, \xi_{J(i)}, \Sigma_{J(i)}$ from the samples. The parameters of the log-ratio normal distribution will be estimated from the compositional sample mean and covariance matrix. The parameters $\theta_{J(i)}$ of the incidence Bernoulli distribution will be estimated from the incidence matrix. This matrix can be summarized in a “pattern” plot (Fig. 3.3)

Figure 3.3 shows the distribution or “pattern” of essential zeros in a data set of a time use survey analyzed in [MDM15]. The table shows the 15 different patterns observed in a sample, where the blue cells indicate zero values in a part, that is, zero minutes in the corresponding daily activity. The marginal bar plots represent the profiles of the distributions of zero incidence. The distribution of zeros in each column is represented at the top. For example, the largest number of zeros (56.55%) is collected in the second part, the daily activity “Work&Stud”. The right marginal bar plot shows the distribution of the patterns. For example, the first bar indicates that 31.75% of individuals have no zeros, expending time in the five daily activities. The numbers in the cells are the value of the geometric mean (in %) for each pattern. For example, the subset of samples with a zero in the parts “House&Fam”, “SocActiv”, and “Commut&Others” (pattern #8) have the center (22.54, 77.46) for the subcomposition (*PCare&Sleep, Work&Stud*). The ratio between these two daily activities takes a very different value in the subcomposition (*PCare&Sleep, Work&Stud, SocActiv*) in pattern #6 where the center is (58.74, 22.04, 19.21). In some scenarios it could be worth exploring as well the variation array by pattern to analyze whether the value of basic statistics are affected by the pattern of zeros.

3.2.2. On analyzing the pattern of zeros. We introduce the analysis of the zeros pattern in the context of the essential zeros but this study can be done for

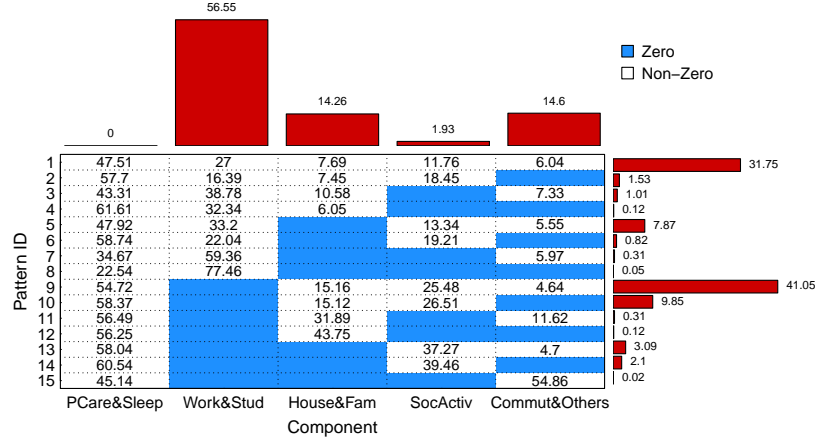


Figure 3.3. Pattern of zeros with marginal bar plots of column and row zero incidence distribution. Numbers in the cells are the value of the geometric mean for each pattern.

missing data, for the other type of zeros or for groups of samples defined by a grouping factor in fully observed zero-free data sets.

The R-package *zCompositions* supports the functions *lcTest*, *zVarArray*, *zVarArrayError*, and *zVarArrayTest* to do the analysis of zeros patterns. The function *zVarArray* returns overall and separate variation arrays for groups defined by zero patterns. That is, for each pattern of zero, log-ratio variances (upper triangle of variation matrix) and log-ratio means (lower triangle of variation matrix) are computed from the available data. The squared relative errors of variation arrays per group with respect to the overall variation array is provided by the function *zVarArrayError*. If one detects potential large errors, its significance can be tested with the function *zVarArrayTest*. This function performs a permutation test of the homogeneity of group-wise and overall variation arrays from all pair-wise log-ratios. This analysis performed for all pairwise logratios in the variation array can be extended for any log-contrast of interest. The function *lcTest* tests for homogeneity across groups of log-ratio means and log-ratio variances of user-defined log-contrasts.

Table 3.2 shows the overall variation array for time use survey data in Figure 3.3. The variation array for each zero pattern should be compared with this one. For example, Table 3.3) shows the variation array for subcomposition (*PCare&Sleep*, *Work&Stud*, *SocActiv*) in pattern #6. Note that the elements corresponding to the zero parts are recorded as “NA”. The weighted square error for each element to the overall value is recorded in parenthesis.

The squared relative errors, being accumulated across all patterns and pairwise logratios, are 0.226 for the log-ratio variances and 0.67 for the log-ratio means. The test returns p-values equal 0.001 both for the variances and for the means, indicating to reject the null hypothesis of homogeneity across the patterns of zeros. The samples in pattern #6 contribute to these errors with 20.57% for the variances

Table 3.2. Overall variation array for time use survey data in Figure 3.3. (PS: PCare&Sleep, WS: Work&Stud, HF: House&Fam, SA: SocActiv, CO: Commut&Others)


	PS	WS	HF	SA	CO
PS	—	0.60	0.93	0.42	0.63
WS	0.54	—	1.81	1.45	0.84
HF	1.50	1.24	—	1.59	1.54
SA	1.00	0.80	-0.50	—	1.03
CO	2.28	1.56	0.77	1.25	—

Table 3.3. Variation array for pattern #6 in Figure 3.3: subcomposition (*PCare&Sleep, Work&Stud, SocActiv*) with the weighted squared error in parenthesis. (PS: PCare&Sleep, WS: Work&Stud, HF: House&Fam, SA: SocActiv, CO: Commut&Others)

	PS	WS	HF	SA	CO
PS	—	1.18 (0.01)	NA	1.06 (0.02)	NA
WS	0.98 (0.01)	—	NA	3.65 (0.02)	NA
HF	NA	NA	—	NA	NA
SA	1.12 (0.00)	0.14 (0.01)	NA	—	NA
CO	NA	NA	NA	NA	—

and only 1.66% for the means, suggesting only large differences in the variances for this pattern.

Activities for Section 3.2

 [Click here to get the activities of this section](#)

3.3. Count zeros

This type of zero appears in sampling studies involving *counts* (i.e. elections). Usually, in the framework of a multinomial experiment, we detect that no “items” fall into the j -th category of counts and a zero value is recorded in the j -th part of a composition. Our assumption is that this j -th category could be unobserved due to the limited size of the sample. That is, if the time of the observation period or of the experiment was larger, then some item would *probably* fall into this category. In other words, the probability associated to this category in the multinomial experiment is not zero. Consequently, we can conclude that it makes sense to replace the count zero in the j -th part with a non-zero value. According to [MPO11], the imputation method based on a *Bayesian-multiplicative* (*BM*) replacement could be appropriate. The idea of these techniques consist of applying the closure to compositions ($\kappa = 1$) and, afterwards, to consider compositions as vectors of estimates

of probabilities in a multinomial probability model. Consequently, the zero values should be replaced by a small quantity.

Assume that outcomes in each of the n identical and independent trials can fall in any of the D mutually exclusive categories. Let $\boldsymbol{\pi} = [\pi_1, \dots, \pi_D]$ be the probabilities of outcome in categories ($\sum_j \pi_j = 1$). Let $\mathbf{c} = [c_1, \dots, c_D]$ be the vector of counts ($\sum_j c_j = n$). The *BM*-replacement is based on the property that states that the typical Dirichlet distribution $Dir(D; \boldsymbol{\alpha})$ is the conjugate prior to the multinomial probability distribution $M(n; \boldsymbol{\pi})$:

- The prior $Dir(D; \boldsymbol{\alpha})$ density function of $\boldsymbol{\pi}$ is $p(\boldsymbol{\pi}) \propto \prod_{j=1}^D \pi_j^{\alpha_j-1}$.
- For the multinomial distribution $M(n; \boldsymbol{\pi})$ the likelihood function is $L(\boldsymbol{\pi} | \mathbf{c}) \propto \prod_{j=1}^D \pi_j^{c_j}$.
- The posterior $Dir(D; \boldsymbol{\alpha} + \mathbf{c})$ density function of $\boldsymbol{\pi}$ is $p(\boldsymbol{\pi} | \mathbf{c}) = (L(\boldsymbol{\pi} | \mathbf{c}) \cdot p(\boldsymbol{\pi})) \propto \prod_{j=1}^D \pi_j^{c_j + \alpha_j - 1}$.

Note that the prior estimate of the expectation is $\hat{E}[\pi_j] = \alpha_j / \sum \alpha_k$. Therefore, given the observed data \mathbf{c} , the Bayesian posterior estimate $Dir(D; \boldsymbol{\alpha} + \mathbf{c})$ is

$$\hat{E}[\pi_j | \mathbf{c}] = \frac{c_j + \alpha_j}{n + \sum \alpha_k}.$$

Note that $n = \sum c_j$, the total of the original vector of counts. Using this approach, for a count zero in the j th part it holds that $c_j = 0$, and a candidate for the imputed value for the zero in the CoDa set is

$$\frac{\alpha_j}{n + \sum \alpha_k}.$$

To model $\boldsymbol{\alpha}$ we consider $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_D] = s \cdot \mathbf{t} = s \cdot [t_1, \dots, t_D]$. The parameter s is called the “strength” and the vector $\mathbf{t} = [t_1, \dots, t_D]$, where $\sum t_k = 1$, is the prior “estimate” of $\boldsymbol{\pi}$. Note that with this modelling, $p[\boldsymbol{\pi}] \propto \prod_{j=1}^D \pi_j^{s \cdot t_j - 1}$, the prior estimate is $E[\pi_j] = t_j$ and the posterior estimate for a count zero is

$$\frac{s \cdot t_j}{n + s}.$$

Commonly, one can take $\mathbf{t} = [1/D, \dots, 1/D]$, the non-informative prior, where we consider that the probabilities are the same in all parts. For example, when one considers the typical non-informative prior $t_j = \frac{1}{D}$ and $s = \frac{D}{2}$, then the classical *Jeffreys* posterior is obtained, that is, $\frac{1/2}{n + D/2}$ is the value that replaces the zero in the data sets. Since this non-informative assumption is not realistic in most situations, one could use the observed data \mathbf{c} to calculate a preliminary estimate. Following this approach, and using a leave-one-out strategy, for the i th composition we consider that $\hat{\alpha}_{ij} = \sum_{k=1, k \neq i}^N c_{kj}$ and that the prior estimate is $\frac{\hat{\alpha}_{ij}}{\sum_{k=1}^D \hat{\alpha}_{ik}}$. Let \hat{m}_{ij} be the prior estimate calculated using this leave-one-out procedure for the j th part in the i th composition, Table 3.4 shows the most common possibilities for the parameter s . [Ma+15] shows that when one takes the prior $t_{ij} = \hat{m}_{ij}$ for each sample then the strength $s_i = 1/g_i$ achieves the best performance in terms of minor distortion of the covariance structure.

Table 3.4. Dirichlet models and modified posterior Bayesian estimates (i.e. imputed value) when $t_{ij} = \hat{m}_{ij}$. The value g_i is the geometric mean of vector $\hat{\mathbf{m}}_i$ (SQ: square; GBM: geometric BM).

	Perks	Jeffreys	Bayes-Laplace	SQ	GBM
Strength s	1	$D/2$	D	\sqrt{n}	$1/g_i$
Imputed value	$\frac{\hat{m}_{ij}}{n+1}$	$\frac{\hat{m}_{ij} \cdot D}{2n+D}$	$\frac{\hat{m}_{ij} \cdot D}{n+D}$	$\frac{\hat{m}_{ij}}{\sqrt{n}+1}$	$\frac{\hat{m}_{ij}}{g_i \cdot n+1}$

Let \mathbf{x}_i , $i = 1, \dots, n$, be closed ($\kappa = 1$) D -part compositions with some count zero. We replace \mathbf{x}_i with a composition \mathbf{r}_i without zeros using the expression called the *BM*-replacement:

$$r_{ij} = \begin{cases} \frac{s_i \cdot t_{ij}}{n_i + s_i}, & \text{if } x_{ij} = 0, \\ x_{ij} \cdot \left(1 - \sum_{j|x_{ij}=0} \frac{s_i \cdot t_{ij}}{n_i + s_i}\right), & \text{if } x_{ij} > 0, \end{cases}$$

where n_i is the sum of the parts of the composition \mathbf{x}_i when it is expressed using the original counts, the parameter s_i is called the “strength” and the vector \mathbf{t}_i , the “prior”. The most common values for the strength s_i , are 1, $D/2$, D and $\sqrt{n_i}$. The value of \mathbf{t}_i are also decided by the analyst. For example, $t_{ij} = \frac{1}{D}$, $j = 1, \dots, D$, is the non-informative or “uniform” option. The multiplicative modification of non-zero values ($x_{ij} > 0$) is carried out to preserve the ratios between parts. Moreover, by carrying out the product $\mathbf{r}_i \cdot n_i$ an imputed pseudo-count is obtained if required.

Example 3.

The data set [PMS09] is formed by the individual male preferences of East African cichlids to court female colour polymorphic species of three colour morphs: plain type (P), white background (WB), and orange background (OB).

id. male	P	WB	OB	total courtships
1	0	20	8	28
1	30	14	28	72
\vdots	\vdots	\vdots	\vdots	\vdots
26	31	1	31	63
26	12	0	3	15

For example, in the first trial with male #1 the observed data was $[0, 20, 8]$. That is, the male made 20 courtship advances to female WB, 8 to OB. However, in its second trial the courtships were $[30, 14, 28]$, also making courtship advances also to the female P and supporting the assumption that the zeros are not essential zeros.

When we apply “Jeffreys prior” ($t_{ij} = \frac{1}{3}$ and $s_i = \frac{3}{2} \rightarrow s_i \cdot t_{ij} = \frac{1}{2}$) to replace zeros in the data of the first trial of male fish #1 we have to take the following steps:

Step	Composition	P	WB	OB	sum
1 st step	initial counts	0	20	8	28
2 nd step	proportions	0	0.7143	0.2857	1
3 rd step	replaced	0.0169	0.7143	0.2857	1.0169
4 th step	multiplicative modification	0.0169	0.7022	0.2809	1
5 th step-(optional)	pseudo-counts	0.4746	19.6610	7.8644	28

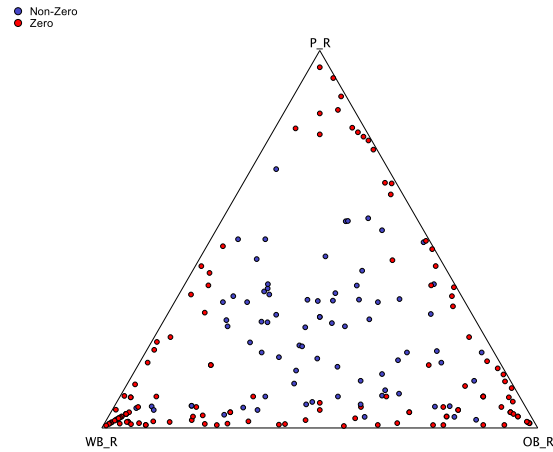



Figure 3.4. Data set [PMS09] in the ternary diagram after the replacement of count zeros (RX).

Figure 3.4 shows the CoDa set [PMS09] on the ternary diagram once the replacement method imputed the zeros using *Jeffreys* prior. Those samples with some zero are plotted as red circles, the compositions without zeros are blue circles. After the replacement we obtain a CoDa set without zeros.

Activities for Section 3.3

 [Click here to get the activities of this section](#)

3.4. Censored data: rounded zeros

The “rounded zeros problem” is a particular NMAR case of missing data. In this case, the data cannot be *observed* because their *true* value is below the maximum round-off error or a detection limit (DL). In the first case, when we represent the value using a finite string of digits, the true value is rounded to zero. In the second case, the true value is below the detection limit (BDL) of the experimental gadget and it records the true value as a *less-than* value or “< DL” value. This second case frequently appears in geochemical case studies. Hereafter, for simplicity, we will refer to both cases as “rounded zeros”. Certainly, it makes sense to replace zeros by *small* non-zero values. The algorithms have been formulated for left-censored

data as the most case in practice, but they could be easily extended to the case of right-censored data that has been recently considered in CoDa [MIK2020udl].

There are some types of imputation methods available in CoDaPack and in R-package *zCompositions*:

- univariate methods:
 - multiplicative replacement;
 - normal R_+ replacement: classical and robust;
- multivariate methods:
 - logratio EM replacement: classical and robust;
 - *data augmentation* algorithm.

By multivariate methods one refers to methods using information of other parts to replace the zero in one part, whereas univariate methods only use the information in the part to impute. By analogy to the missing data problem, when the data set has *few* rounded zeros the *multiplicative* replacement is appropriate. According to [MBP03], to have “few rounded zeros” means that the number of zeros is lower than 10% of the entries in the data matrix. In this case, a composition \mathbf{x}_i with some rounded zero is replaced by a composition $\mathbf{r}_i \in \mathcal{S}^D$ without zeros using the expression

$$r_{ij} = \begin{cases} \delta_{ij} & \text{if } x_{ij} = 0; \\ x_{ij} \cdot \frac{\kappa_i - \sum_{k|x_{ik}=0} \delta_{ik}}{\kappa_i} & \text{if } x_{ij} > 0, \end{cases}$$

where κ_i is the total sum of the non-zero values in the composition that one wants to preserve after the replacement. The value δ_{ij} must be “ $< \epsilon_{ij}$ ”, the DL or maximum rounding error. One reasonable option [MBP03] is $\delta_{ij} = 0.65\epsilon_{ij}$. In any case, the results obtained after the statistical analysis has been carried out with the replaced data set $\mathbf{R}\mathbf{X}$ should be checked by sensitivity analysis, where a range $\frac{\epsilon_{ij}}{10} \leq \delta_{ij} \leq \epsilon_{ij}$ is appropriate.

Example 4. Consider the compositions \mathbf{x}, \mathbf{x}^* :

	ϵ_1	ϵ_2	ϵ_3	
DL	0.1	0.05	0.08	
	x_1	x_2	x_3	κ_i
\mathbf{x}	0.0000	0.3333	0.6667	1
\mathbf{x}^*	0.0000	0.6400	0.3600	1

with one rounded zero each one. The first step consists of replacing rounded zeros by 65% of the DL= 0.1, that is, $r_{ij} = \delta_{ij} = 0.65 \cdot \epsilon_{ij}$:

	x_1	x_2	x_3	sum
\mathbf{r}	0.0650	0.3333	0.6667	1.065
\mathbf{r}^*	0.0650	0.6400	0.3600	1.065

Next, we perform a *multiplicative* modification of non-zero values $r_{ij} = x_{ij} \cdot (\kappa_i - \sum_{k|x_{ik}=0} \delta_{ik}) / \kappa_i$ to obtain the final compositions:

	x_1	x_2	x_3	sum
\mathbf{r}	0.0650	0.3117	0.6233	1
\mathbf{r}^*	0.0650	0.5984	0.3366	1

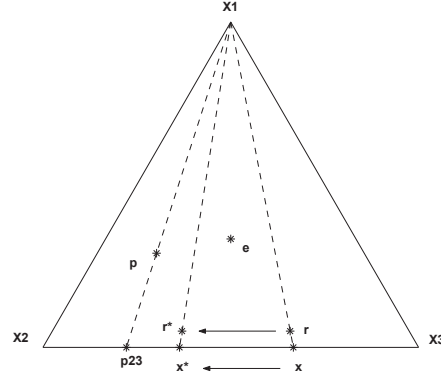


Figure 3.5. Ternary diagram: compositions \mathbf{x}, \mathbf{x}^* with rounded zeros and its replaced compositions \mathbf{r}, \mathbf{r}^* . The vectors \mathbf{p} and \mathbf{p}_{23} are the corresponding perturbation difference vectors. See text for more details.

Figure 3.5 shows that this replacement is fully consistent with the geometry of the simplex. Indeed, the replacement preserves the ratios between the non-zero values, making minimal distortion in the covariance structure. Furthermore, the replacement is coherent with the perturbation and subcomposition operations. Indeed, let $\mathbf{p} = \mathbf{r}^* \ominus \mathbf{r}$ be the perturbation vector difference between the replaced compositions \mathbf{r}^* and \mathbf{r} . Let \mathbf{p}_{23} be the 2-part subcomposition formed by the second and third parts. Thus, \mathbf{p}_{23} is equal to the perturbation vector difference between the corresponding subcompositions \mathbf{x}_{23} and \mathbf{x}_{23}^* respectively of \mathbf{x} and \mathbf{x}^* . That is, it holds $\mathbf{p}_{23} = \mathbf{x}_{23}^* \ominus \mathbf{x}_{23}$.

When a normal distribution representation on the own positive real line is assumed, the univariate technique of imputation is called *normal R_+ replacement* and its formula is

$$r_{ij} = \begin{cases} \delta_{ij} & \text{if } x_{ij} = 0; \\ x_{ij} \cdot \frac{\kappa_i - \sum_{k|x_{ik}=0} \delta_{ik}}{\kappa_i} & \text{if } x_{ij} > 0, \end{cases}$$

where κ_i is the sum of the non-zero values, the value $\delta_j = \exp(\hat{\mu}_j - \hat{\sigma}_j \cdot \hat{\lambda}_j)$, and $\hat{\lambda}_j = \frac{\phi((\ln \epsilon_j - \hat{\mu}_j)/\hat{\sigma}_j)}{\Phi((\ln \epsilon_j - \hat{\mu}_j)/\hat{\sigma}_j)}$

The estimates $\hat{\mu}_j$ and $\hat{\sigma}_j$ can be obtained in the log-space of coordinates using both ordinary or robust methods for normal variates. The value $\hat{\mu}_j - \hat{\sigma}_j \cdot \hat{\lambda}_j$ estimates the expected value of the mapped nondetects. Its back-transformation to the positive real line provides the optimal estimator of the expected value of the nondetects, actually their geometric mean [PM13, PM15].

When the number of rounded zeros is large (more than 10% of entries) [PM08] recommend the *multivariate modified EM log-ratio algorithm*. This algorithm is based on a parametric approach because log-ratio normality is assumed. In short, the method consists of modifying the EM-algorithm to include the information that (logratio) imputed values have to be lower than the (logratio) detection limit [PM08].

In other words,

- (1) log-ratio transform the CoDa set $\mathbf{X} \in \mathcal{S}^D$ which has rounded zeros to obtain the data set \mathbf{Y} . The log-ratio values of the zeros are NA values, that is, missing values of \mathbf{Y} ;
- (2) replace log-ratio missing values in \mathbf{Y} with the *modified* EM-algorithm to obtain the *complete* data set \mathbf{RY} . In the k th iteration the estimated value is calculated using the formula

$$\hat{r}y_j^{(k)} = \hat{\beta}^{(k)} \cdot \mathbf{ry}_{-j}^{(k)t} - \hat{\sigma}^{(k)} \frac{\phi\left(\frac{\psi_j^{(k)} - \hat{\beta}^{(k)} \cdot \mathbf{ry}_{-j}^{(k)t}}{\hat{\sigma}^{(k)}}\right)}{\Phi\left(\frac{\psi_j^{(k)} - \hat{\beta}^{(k)} \cdot \mathbf{ry}_{-j}^{(k)t}}{\hat{\sigma}^{(k)}}\right)},$$

where: ϕ and Φ are, respectively, the density and distribution functions of the standard normal distribution; $\hat{\sigma}^{(k)}$ is the estimated conditional standard deviation of $\mathbf{ry}_j^{(k)}$; and $\psi_j^{(k)}$ is the logratio transformed value of the detection limit.

- (3) back-transform the data set \mathbf{RY} to obtain $\mathbf{RX} \in \mathcal{S}^D$, a CoDa set without zeros.
- (4) use \mathbf{RX} to impute zeros in \mathbf{X} . If required, multiplicatively modify observed values in \mathbf{X} for preserving the total sum κ .

According to [Ma+12], robust techniques could be applied on the Expectation (regression) step of the EM algorithm to the *olr*-coordinates set. The CoDaPack and *zCompositions* R-package support this facility.

Another multivariate algorithm to replace zeros is the *Data Augmentation algorithm* (DA) [PM13]. This algorithm is a Markov Chain Monte Carlo (MCMC) algorithm originally designed for missing data problems. Indeed, the DA is remarkably similar to the EM algorithm, where the E and M steps are replaced by simulation-based I (*imputation*) and P (*posterior*) steps.

Under multivariate logratio normality, in the k th iteration of DA-algorithm:

- (1) I-step: given estimates $\hat{\theta}^{(k)}$, simulate from $P[\mathbf{y}_{mis} | \mathbf{y}_{obs}, \mathbf{y}_{non} < \psi, \hat{\theta}^{(k)}]$.
- (2) P-step: generate new estimates $\hat{\theta}^{(k+1)}$ by simulating from $P[\theta | \hat{\mathbf{y}}_{non}, \mathbf{y}_{obs}]$.

The above I-P sequence generates a Markov chain with the posterior predictive distribution of the censored values as stationary distribution. After a sufficient number of iterations, suitable random values can be drawn from the chain to replace nondetects. In addition, DA is easily used to implement a multiple imputation scheme for nondetects.

Example 5. Modified EM logratio algorithm

La Paloma stream CoDa set [Mo+10] consists of 96 samples of a 15-component geochemical composition coming from *La Paloma stream* which traverses the Cerro Pelado Fm., in north-western Venezuela. The observed composition, measured in $\mu\text{g/g}$, is [Cr, B, P, V, Cu, Ti, Ni, Y, Sr, La, Ce, Ba, Li, K, Rb]. The detection limit (DL) of each chemical component is respectively (2, 1, 0, 0, 2, 0, 6, 1, 0.6, 1, 1, 0, 0, 632, 10). In CoDaPack, we can use the submenu *Set detection limit* at the menu *Irregular data* to establish the DL for each chemical element.

To describe the distribution of zero values we can use the submenu *ZPatterns plot* at the menu *Irregular data*. Figure 3.6 shows the pattern of the presence of rounded zeros in the data set, both in numerical and graphical output format. A white cell means that the value was above the DL, and the blue cell means that the value is below DL (rounded zero). Observe that the right-hand column on the table corresponds to the marginal profile of the distribution of the different patterns. The profile on the top indicates, for example, that P has no zeros and 51.04% of the samples have a value below DL on Ni.

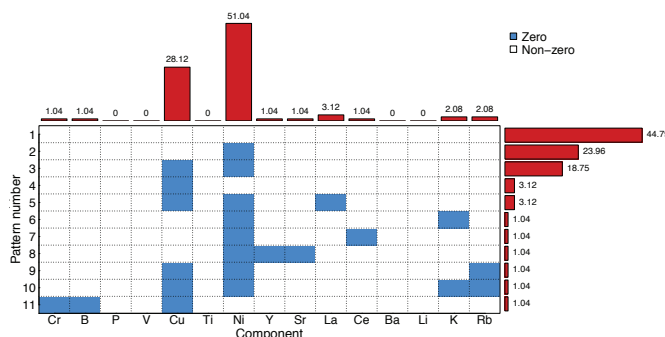


Figure 3.6. Pattern of values below detection limit for *La Paloma stream* CoDa set: blue color indicates nonobserved values.

The three first axes of the *clr*-biplot of this data set retain 51.89% of the original data variability after the modified EM log-ratio algorithm. Figure 3.7 shows the first two *clr*-biplot axes. The term *Imp* was added to the name of *clr*-variables to inform that rounded zeros were imputed. The blue circles are samples without zeros and the red circles are the samples with replaced rounded zeros. The largest ray is the ray of *clr.KImp* suggesting the chemical element that show larger variability is K. Note that this analyte is not a component with a larger number of rounded

zeros. In other words, the small values imputed do not dominate on the variability of the data set. Furthermore, we can state that, as expected, some samples with imputed zeros are far from the center because they take small values in some parts.

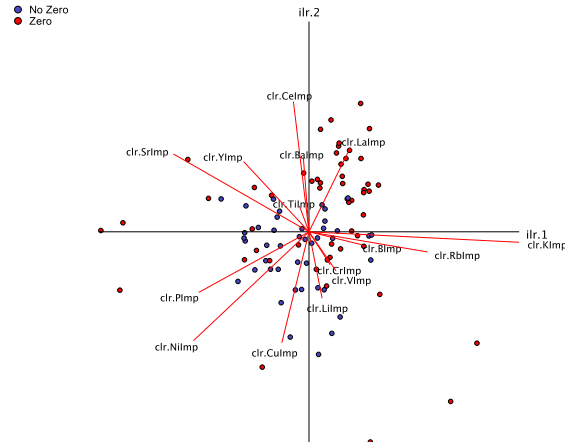


Figure 3.7. *clr*-biplot of *La Paloma stream* CoDa set after the zero replacement by the modified EM log-ratio algorithm

Activities for Section 3.4

[Click here to get the activities of this section](#)

3.5. Dealing with missing values and zeros

The connection between the nature of missing values and zeros is very strong. Moreover, in practice NA and zeros are simultaneously present in the data set. In these scenarios the recommended steps are:

- (1) Decide which problem (missing values or zeros) is the *minor* problem. Here *minor* indicates that the number of these cases are minor.
- (2) Non-parametric impute the irregular data of the minor problem.
- (3) Parametric impute the irregular data of the major problem.
- (4) Impute the minor problem with a parametric method.
- (5) Iterate (3) and (4) until convergence.

In addition, when there is a suspicion of potential outliers then apply robust algorithms. The convergence of this algorithm may be established in terms of the basic statistics means vector and covariance matrix. For example, the algorithm will stop

when the difference between two consecutive estimates of these statistics is small. CoDaPack, at the submenu *Logratio-EM Zero & missing replacement* at the menu *Irregular data*, and functions *multReplus* and *lrEMplus* in package *zCompositions* [PM15] support this algorithm for simultaneously dealing with NA and zeros.

Example 6. Modified EM log-ratio algorithm for NA and zeros

We consider again \mathbf{X} , the *La Paloma stream* CoDa set [Mo+10] with rounded zeros (Fig. 3.6):

- we forced (randomly) missing values MCAR in \mathbf{X} , where 35.42% of samples have at least one NA;
- we replaced the missing data and zeros in \mathbf{X} using the log-ratio EM algorithm to obtain a complete data set \mathbf{RX} .

Figure 3.8 shows the pattern of the presence of forced missing values in the data set \mathbf{X} , both in numerical and graphical output format. A white cell means that the value was observed, and the blue cell means that the value is NA. Observe that the right-hand column on the table corresponds to the marginal profile of the distribution of the different patterns, being 64.58% of samples without NA. The profile on the top indicates, for example, that Ni has no NA and 5.21% of the samples have a NA on Y.

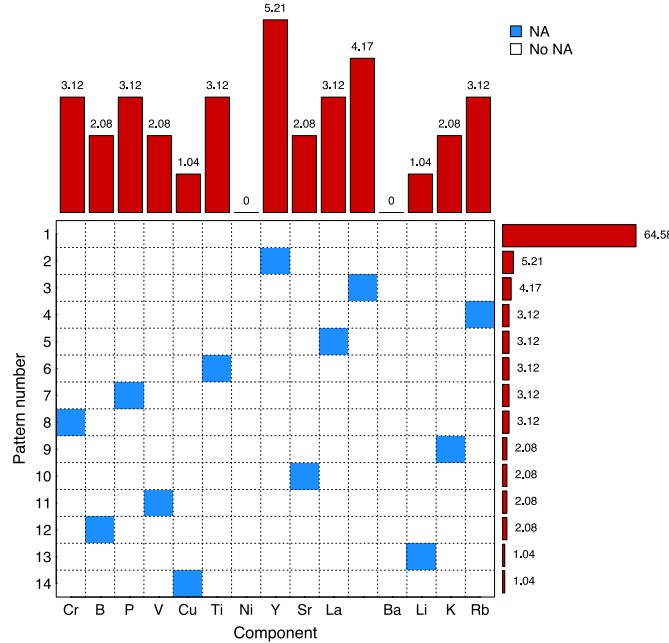


Figure 3.8. Pattern of forced missing values in *La Paloma stream* CoDa set: blue color indicates NA values.

Figure 3.9 shows the first two covariate *clr*-biplot axes for **RX**. The term *c* was added to the name of *clr*-variables to inform that zeros and NA values were imputed.

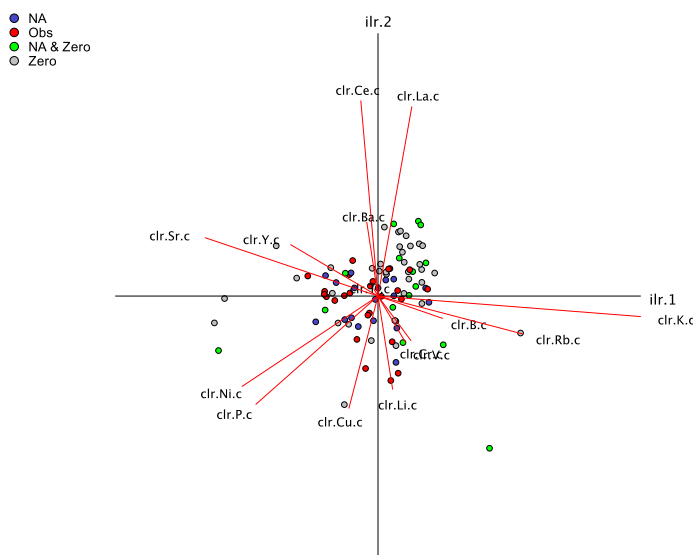


Figure 3.9. *clr*-biplot of La Paloma stream CoDa set **RX** after the NA and zero replacement by the EM log-ratio algorithm

Once compared to the Fig. 3.7 one may conclude that distortion by the NA replacement is not relevant. In this case, the three first axes of the *clr*-biplot of this data set retain 52.71% of the original data variability after the modified EM log-ratio algorithm. The red circles are samples fully observed, the gray circles are compositions with zeros, the blue circles represents samples with NA values, and green circles are the compositions with both zeros and NA values. Again the largest ray is the ray of *clr.K.c* suggesting the chemical element with largest variability is K. This analyte is not the component with the largest number of rounded zeros and NA values (Fig. 3.7 and 3.9). In other words, the values imputed do not dominate on the variability of the data set **RX**. Once again, as expected, some samples with imputed zeros (gray and green circles) are far from the center of the data set because they include imputed small values in some parts.

Activities for Section 3.5

 [Click here to get the activities of this section](#)

3.6. Potential outliers

Broadly speaking, an outlier is a datum which is *too different* from the rest of data. Our concern is to investigate whether this datum is *anomalous* (i.e., *erroneous*), to be removed from the data set; or, on the contrary, the datum is *special* because it is *different*, suggesting a particular analysis of the sample to conclude if there is a *reason* or *cause* that generated the datum. Statistics help us to identify *potential outliers* because they quantify *how different* a datum is from others, giving dissimilarities, measures of difference, divergences and distances. Although there are important mathematical properties to distinguish between these measures, hereafter we will call any these measures *distance*. When we conclude that there are outliers in our data set that may distort our classical statistical analysis then the use of *robust* statistical techniques is recommended.

Frequently, a multivariate outlier is a sample where we have found an *extreme* value (too large or too small) in one or several components. In other words, some values of the multivariate outlier are also univariate outliers. However, sometimes the multivariate outlier is characterised by its extreme values on the associations between the components [VDM14a, VDM14b, VDM15]. Figure 3.10 shows such a situation. The outlier “A”, in sky blue color, is clearly outside the “arch-shaped” distribution of the data set. However, when the three subcompositions of “A” (blue dashed lines) are compared to the maximum and minimum ratios from the distribution (gray dashed lines), one can conclude that the values are not extreme at all. On the other hand, consider now the conditional distribution of samples with the same x_1/x_3 ratio as “A” (red line in the data set). When we compare the ratio x_2/x_3 of “A” with the ratios of the conditional distribution (red line on the edge from \mathbf{x}_2 to \mathbf{x}_3) one can conclude that “A” is clearly outside the expected values.

Most tools for the detection of multivariate outliers are based on the location and the spread (*shape*) of the sample distribution. Indeed,

- **Location:** samples with high (low) values have large *distance* from the central location of the data set.
- **Spread or shape:** samples with middle values could be outliers.

From this point of view, the steps to identifying potential outliers in CoDa are:

- For each $\mathbf{x} \in \mathcal{S}^D$, work in *olr*-coordinates: $\mathcal{S}^D \rightarrow \mathbb{R}^{D-1}$;
- Calculate *typical* Mahalanobis distance d_M :

$$d_M(\text{olr } \mathbf{x}, \boldsymbol{\mu}) = (\text{olr } \mathbf{x} - \boldsymbol{\mu}) \boldsymbol{\Sigma}^{-1} (\text{olr } \mathbf{x} - \boldsymbol{\mu})^t,$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the expectation and covariance of the random composition. *Robust* statistics (centre and covariance) give a *robust* d_M .

- Because, under logratio normality, d_M^2 are approximately chi-square distributed with $D - 1$ degrees of freedom (χ_{D-1}^2), *potential outliers* are the samples having a large d_M^2 , for example, greater than the 97.5% quantile of the chi-squared distribution ($\chi_{(D-1);0.975}^2$).

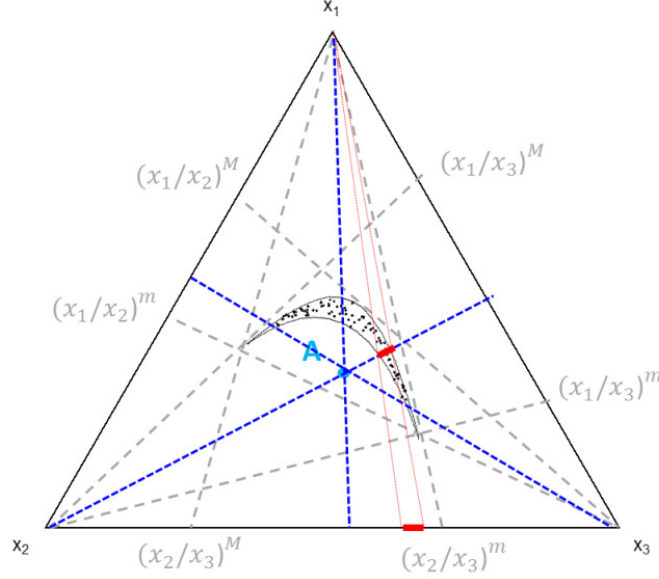


Figure 3.10. An outlier (“A”, sky blue color) in the ternary. Gray dashed lines are the maximum and minimum subcompositions of dataset. Dashed blue lines are the three subcompositions of “A”. Red lines are the conditional subcomposition.

Example 7. Outliers in a CoDa set

Using the *bloodMN* CoDa set we analyse the genotypes in the “MN” system of blood groups, a system of blood antigens also related to proteins of the red blood cell plasma membrane (see [Appendix](#) for more details).

Assuming log-ratio normality, Fig. 3.11a shows the Gaussian (90%, 95%, 99%) predictive regions of data set in the ternary diagram created using the submenu *Predictive Region* of the menu *Graphs* in CoDaPack. Having a look at this typical “arch” distribution, one can detect some samples as potential outlier at the tails of the distribution. Figure 3.11b, where the *olr*-transformed data set is plotted, corroborates this because it shows four samples (red circles) that have a large d_M^2 , for example, greater than the 97.5% quantile of the chi-squared distribution ($\chi_{(2);0.975}^2$). The *olr*-coordinates were created using the SBP shown in the table.

Signs in SBP for the *bloodMN* data set.

	<i>MN</i>	<i>MM</i>	<i>NN</i>
olr_1	1	-1	-1
olr_2	0	1	-1

Figure 3.11b shows that the variable clr_1 has the same direction that the first *olr*-coordinate according the well-known relation between a *clr*-variable and the *pivot* coordinates:


$$clr_1(\mathbf{x}) = \sqrt{\frac{D-1}{D}} \cdot olr_1 = \frac{D-1}{D} \cdot \ln \frac{x_1}{\left(\prod_{k=2}^D x_k\right)^{1/D-1}}.$$

As expected, by the Hardy-Weinberg principle, the first *olr*-coordinate (ilr_1) of the data set is approximately constant. The four samples at the tails of the vertical distribution (Fig. 3.11b) take small and large values in the second *olr*-coordinate (ilr_2) suggesting that the four compositions take a very relative extreme values in part *MM* as regards the part *NN*. This fact is suggested as well by the rays of variables clr_{MM} and clr_{NN} in Fig. 3.11b because the projection of these samples in the rays are respectively the largest values in both *clr*-variables.

To corroborate and complete the description of the four potential outliers we represent the data set in the ternary diagram (Fig. 3.11c). Indeed, the two compositions closest to the vertex *MM* take the largest values in this part and the smallest values in the part *NN*. On the other hand, the two samples closest to the vertex *NN* take the largest values in this part and the smallest values in the part *MM*. In addition, the four samples take the smallest values in the part *MN*.

Although these samples are classified as outliers they fits the distribution and they cannot be classified as anomalous, keeping the samples in the data set. In such a case, we recommend to use robust statistical techniques.

Activities for Section 3.6

 [Click here to get the activities of this section](#)

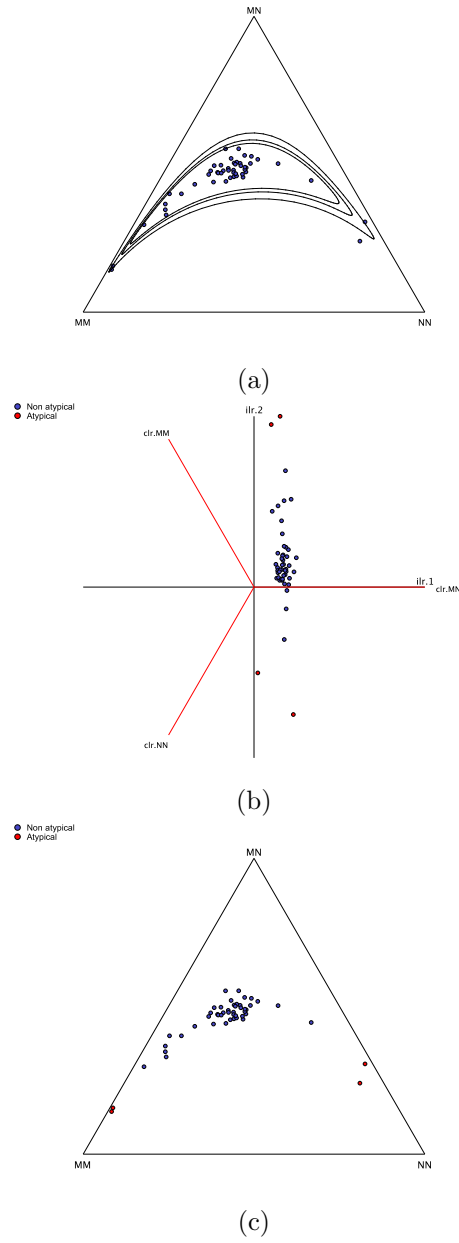


Figure 3.11. A typical Hardy Weinberg data set: (a) Gaussian predictive regions in the ternary diagram; (b) Atypical compositions in an *olr*-transformed space; (c) Atypical compositions in the ternary diagram. Red circles are samples detected as potential outliers.

The chapter's key concepts

- ✓ There are appropriate methods for dealing with irregular CoDa: missing data, outliers and zeros.
 - ✓ Different types of zeros require different techniques for dealing with them.
-

Specific references in Chapter 3

- [AK03] J. Aitchison and J. Kay, *Possible solution of some essential zero problems in compositional data analysis*. In: Proceedings of CoDaWork-03, The 1st Compositional Data Analysis Workshop, Thió-Henestrosa S. and Martín-Fernández J.A. (ed.) (2003), <http://ima.ud.es/Activitats/CoDaWork03/>, University of Girona, Girona (Spain).
- [BB16] J. Bear and D. Billheimer. *A Logistic Normal Mixture Model for Compositional Data Allowing Essential Zeros*. Austrian Journal of Statistics, **45**(4), 3-23, (2016)
- [LR87] R.J.A. Little and D.B. Rubin, *Statistical Analysis with Missing Data*, second ed.. Wiley, New Jersey, 1987.
- [MMY06] R.A. Maronna, R.D. Martin and V.J. Yohai, *Robust Statistics: Theory and Methods*, John Wiley & Sons, New York, 2006.
- [MBP03] J.A. Martín-Fernández, C. Barceló-Vidal and V. Pawlowsky-Glahn, *Dealing with Zeros and Missing Values in Compositional Data Sets*, Mathematical Geology **35** (2003), no. 3, 253–278.
- [MPO11] J.A. Martín-Fernández, J. Palarea-Albaladejo and R.A. Olea, *Dealing with Zeros*. In: Pawlowsky, V. and Buccianti, A. (eds), *Compositional Data Analysis: Theory and Applications*, Chichester (UK), John Wiley & Sons, 43–58, 2011.
- [Ma+12] J.A. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser and J. Palarea-Albaladejo, *Model-based replacement of rounded zeros in compositional data: classical and robust approach*, Computational Statistics & Data Analysis **56** (2012), 2688–2704.
- [Ma+15] J.A. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser and J. Palarea-Albaladejo. *Bayesian-multiplicative treatment of count zeros in compositional data sets*. StatModel (2015) (doi:10.1177/1471082X14535524).
- [MDM15] J.-A. Martín-Fernández, J. Daunis-i-Estadella, G. Mateu-Figueras *On the interpretation of differences between groups for compositional data*. SORT **39**

- (2015), no. 2, 231–252.
- [MIK2020udl] D. Miksová, P. Filzmoser, M. Middleton *Imputation of values above an upper detection limit in compositional data*. Computers & Geosciences **136** (2020), 104383. (doi:10.1016/j.cageo.2019.104383)
- [Mo+10] J.C. Montero-Serrano, J. Palarea-Albaladejo, J.A. Martín-Fernández, M. Martínez-Santana and J.V. Gutiérrez-Martín, *Sedimentary chemofacies characterization by means of multivariate analysis*, Sedimentary Geology **228** (2010), no. 3-4, 218–228.
- [PMS09] M.E.R. Pierotti, J.A. Martín-Fernández and O. Seehausen, *A Mapping individual variation in male mating preference space: multiple choice in a colour polymorphic cichlid fish*, Evolution **63** (2009), no. 9, 2372–2388.
- [PM08] J. Palarea-Albaladejo and J.A. Martín-Fernández, *A modified EM algorithm for replacing rounded zeros in compositional data sets*, Computers & Geosciences **34** (2008), no. 8, 902–917.
- [PM13] J. Palarea-Albaladejo and J.A. Martín-Fernández, *Values below detection limit in compositional chemical data*. Analytica Chimica Acta **764** (2013), 32–43.
- [PMO14] J. Palarea-Albaladejo, J.A. Martín-Fernández and R.A. Olea, *Bootstrap estimation of distributional statistics from compositional data with nondetects: a case study on coal ashes*. Journal of Chemometrics **28**(7) (2014), 585–599.
- [PM15] J. Palarea-Albaladejo and J.A. Martín-Fernández, *zCompositions - R package for multivariate imputation of left-censored data under compositional approach*. Chemometrics and Intelligent Laboratory Systems **143**, 85–96.
- [VDM14a] V. Vives-Mestres, J. Daunis-i Estadella and J.A. Martín-Fernández, *Individual T^2 Control Chart for Compositional Data*. Journal of Quality Technology **46**(2) (2014), 127–139.
- [VDM14b] V. Vives-Mestres, J. Daunis-i Estadella and J.A. Martín-Fernández, *Out-of-Control Signals in 3-part Compositional T^2 Control Chart*. Quality and Reliability Engineering International **30**(3) (2014), 337–346.
- [VDM15] V. Vives-Mestres, J. Daunis-i Estadella and J.A. Martín-Fernández, *Signal Interpretation in Hotelling's T^2 Control Chart for Compositional Data*. IIE Transactions **48**(7) (2016), 661–672.

Linear regression models (LRM)

Contents

- 4.1 LRM for a compositional response and scalar predictor
- 4.2 LRM for a scalar response and compositional predictor
- 4.3 [LRM extensions](#)
 - [4.3.1 Extensions of an LRM with a compositional predictor](#)
 - [4.3.2 Compositions as both predictor and response](#)

Objectives

- ✓ To estimate and interpret an LRM when the response is compositional.
- ✓ To estimate and interpret an LRM when the predictor is compositional.
- ✓ To introduce some extensions for an LRM

4.1. LRM for a compositional response and scalar predictor

In this section we are dealing with the Linear Regression Models (LRM) where the compositional variables are the response variables of the model [TB11].

Let \mathbf{X} be a data set in \mathcal{S}^D formed by n observations \mathbf{x}_i , for $i = 1, 2, \dots, n$. The i -th observation \mathbf{x}_i is associated with r external variables or covariates ($r \geq 1$) grouped in the real vector $\mathbf{t}_i = [t_{i0}, t_{i1}, \dots, t_{ir}]$, where $t_{i0} = 1$, for $i = 1, 2, \dots, n$.

The goal is to estimate the coefficients $\beta_0, \beta_1, \dots, \beta_r$ of a linear surface into \mathcal{S}^D whose equation is

$$\hat{\mathbf{x}}(\mathbf{t}) = \beta_0 \oplus (t_1 \odot \beta_1) \oplus \dots \oplus (t_r \odot \beta_r) = \bigoplus_{j=0}^r (t_j \odot \beta_j),$$

where $\mathbf{t} = [t_0, t_1, \dots, t_r]$ are real covariates and are identified as the parameters of the linear surface; the first parameter is defined as the constant $t_0 = 1$; and $\hat{\mathbf{x}}(\cdot)$ are the expected value of the CoDa-response variable. The compositional coefficients of the model, $\beta_j \in \mathcal{S}^D$, are to be estimated from the data. The most popular fitting method is the least-square deviation criterion which minimizes the sum of squared errors. Because this model is presented as a least-squares problem in the simplex, it could be formulated in terms of orthonormal log-ratio coordinates (olr). In other words,

- (1) we select a *olr*-basis in \mathcal{S}^D , for example, according to an SBP.
- (2) we represent the responses in coordinates: $\mathbf{x}_i^* = \text{olr}(\mathbf{x}_i) \in \mathbb{R}^{D-1}$.
- (3) we solve $D - 1$ ordinary-least-squares regression problems in coordinates to obtain the *olr*-coordinates β_j^* vectors of the β_j coefficients ($j = 1, 2, \dots, r$). That is, for the coordinates $k = 1, 2, \dots, D - 1$, find β_j^* minimizing the usual sum of squared errors:

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

where

$$\hat{x}_k^*(\mathbf{t}) = \beta_{0k}^* + \beta_{1k}^* t_1 + \dots + \beta_{rk}^* t_r, \text{ and}$$

- (4) back-transform the coefficients β_j^* to $\beta_j \in \mathcal{S}^D$ using $\beta_j = \text{olr}^{-1}(\beta_j^*)$.

Interpretation can thus alternatively be made in coordinates or in the simplex. Coefficients β_{jk}^* , $j = 1, \dots, r$ and $k = 1, \dots, D - 1$, can be interpreted as the effect of an increase in t_j by one unit (keeping the other t_j constant) on the *olr*-coordinate x_k^* . Thus, the values of a coefficient β_{jk}^* and its interpretation depend on the chosen *olr*-basis. The coefficient β_j is the perturbation vector which is applied to

the composition when the covariate t_j increases by one unit (keeping the other t_j constant). Note that this latter interpretation is invariant to the *olr*-basis.

Recall that sometimes it is interesting to transform (e.g. function `ln`) the real covariates in advance to estimate the $D - 1$ ordinary regression models in step (3). This decision is shared for all the LRM because it is an option due to the nature of the covariate (sample space, distribution, etc.). Moreover, in this step an outlier analysis could be done. In other words, robust regression models could be considered. For example, one can perform robust regression in R with the `rlm()` function in the MASS R-package.

Once the model is obtained one could ask for a global measure of goodness of fit beyond those provided for each of the $D - 1$ equations separately. A typical approach is based on the determination coefficient of the LRM (R^2) which compares the sum of squares explained by the regression model (SSR) with the sum of squared errors (SSE) and the total sum of squares (SST). This comparison is based on the decomposition:

$$\text{SST} = \text{SSR} + \text{SSE}.$$

In more formal terms, let \mathbf{g} be the centre of the CoDa set \mathbf{X} . The total sum of squares in the simplex is $\text{SST} = \sum_{i=1}^n \|\mathbf{x}_i \ominus \mathbf{g}\|_a^2$, the sum of squared errors is $\text{SSE} = \sum_{i=1}^n \|\mathbf{e}_i\|_a^2 = \sum_{i=1}^n \|\hat{\mathbf{x}}(\mathbf{t}_i) \ominus \mathbf{x}_i\|_a^2$ and the sum of squares explained is $\text{SSR} = \sum_{i=1}^n \|\hat{\mathbf{x}}(\mathbf{t}_i) \ominus \mathbf{g}\|_a^2$.

Using these elements, a determination coefficient of the LRM [Eg+12] can be defined as:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}},$$

where $0 \leq R^2 \leq 1$ and it is interpreted as the percentage of variance in \mathbf{X} explained by the LRM.

Inferential aspects on an LRM are typically based on the assumption of normality. Indeed, the residuals $\mathbf{e}_i = \hat{\mathbf{x}}(\mathbf{t}_i) \ominus \mathbf{x}_i$, for $i = 1, 2, \dots, n$, of a normal LRM are assumed to be normally distributed. Furthermore, the residuals \mathbf{e}_i are independent and homoskedastic across the samples, that is, for $i = 1, 2, \dots, n$. In other words, $\mathbf{e}_i \sim \mathcal{N}_S(\mathbf{0}, \mathbf{\Sigma})$ where \mathbf{e}_i and \mathbf{e}_j ($i \neq j$) are independent.

Example 1. Vulnerability of a dike

A dike is subjected to external actions, like 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 dike may be considered: θ_1 , service; θ_2 , damage; θ_3 , collapse. The dike can be designed for a design action, that is, wave-height, d , ranging $3 \leq d \leq 20$ (metres wave-height). Actions, parameterized by some wave-height of the storm, h , also ranging $3 \leq h \leq 20$ (metres wave-height). Vulnerability of the dike is described by the conditional probabilities

$$\hat{p}_k(d, h) = \text{P}[\theta_k | d, h] \quad , \quad k = 1, 2, 3 \quad , \quad \sum_{k=1}^3 p_k(d, h) = 1 \quad ,$$

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, $\mathbf{p}(d, h)$ is approximately known for only a limited number of values $\mathbf{p}(d_i, h_i)$, $i = 1, \dots, n$. The whole model of vulnerability can be expressed as a LRM

$$\mathbf{p}(d, h) = \beta_0 \oplus (d \odot \beta_1) \oplus (h \odot \beta_2),$$

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

Let \mathbf{X} be the CoDa set in the following table. The data set is formed by only nine ($n = 9$) samples where the compositional responses are the probabilities [**p.service**, **p.damage**, **p.collapse**], and **design** and **wave-height** are the covariates. These data were obtained from Monte Carlo simulations.

i	design (d)	wave-height (h)	$p.service$	$p.damage$	$p.collapse$
1	3.0	3.0	0.50	0.49	0.01
2	3.0	10.0	0.02	0.10	0.88
3	10.0	3.0	0.999	0.0009	0.0001
4	10.0	10.0	0.30	0.65	0.05
5	5.0	4.0	0.95	0.049	0.001
6	6.0	9.0	0.08	0.85	0.07
7	7.0	5.0	0.97	0.027	0.003
8	8.0	3.0	0.997	0.0028	0.0002
9	9.0	9.0	0.35	0.55	0.10

The *olr*-coordinates \mathbf{X}^* of \mathbf{X} were obtained as

$$x_1^* = \sqrt{\frac{2}{3}} \ln \left(\frac{\mathbf{p.service}}{(\mathbf{p.damage} \cdot \mathbf{p.collapse})^{1/2}} \right),$$

$$x_2^* = \sqrt{\frac{1}{2}} \ln \left(\frac{\mathbf{p.damage}}{\mathbf{p.collapse}} \right).$$

Once we have solved the two ordinary regression problems (Least Squares) in coordinates, we obtain the coordinates β_j^* :

	ilr.coef_1	p -value	ilr.coef_2	p -value
β_0^* (intercept)	4.0127	0.005	2.2767	0.132
β_1^* (design)	0.5069	0.003	0.1025	0.516
β_2^* (wave.height)	-0.8477	9E-05	-0.2325	0.117

Assuming the normal distribution, the p -values of significance of the coefficients on coordinates β_j^* (see table above) were calculated. Note that, in this case, only the regression coefficients of the first *olr*-coordinate are significant. The LRM in *olr*-coordinates is

$$(\hat{x}_1^*, \hat{x}_2^*) = (4.0127, 2.2767) + (0.5069, 0.1025) \cdot d + (-0.8477, -0.2325) \cdot h.$$

Back-transforming coefficients β_j^* to the simplex, the coefficients β_j are:

	reg_coef.service	reg_coef.damage	reg_coef.collapse
β_0 (intercept)	0.9632	0.0354	0.0014
β_1 (design)	0.4813	0.2781	0.2406
β_2 (wave.height)	0.1487	0.3563	0.4950

The LRM expressed in raw units is

$$[p.service, p.damage, p.collapse] = [0.9632, 0.0354, 0.0014] \oplus (d \odot [0.4813, 0.2781, 0.2406]) \\ \oplus (h \odot [0.1487, 0.3563, 0.4950]) .$$

When the compositional coefficients β_1 and β_2 of the model are compared with the neutral perturbation vector $[1/3, 1/3, 1/3]$ we can conclude that **p.service** increases with **design**, and that when **wave.height** increases we can expect **p.collapse** to increase.

The total sum of squares is equal to SST=0.8780, which decomposes on SSR=0.7582 and SSE=0.1198. In consequence, $R^2 = 86.34\%$, which suggests a high quality of the model.

The following table shows that normality can be assumed as a distribution of the compositional residuals. That is, the p -values suggest that a Gaussian LRM cannot be rejected.

	Anderson Darling		Cramér von Mises		Watson	
	A ² *	p -value	W ² *	p -value	U ² *	p -value
Radius test	0.4421	>0.15	0.0358	>0.15	0.0642	>0.15

The results show that increasing the **design** variable leads to a significant increase in the probability of service as compared to damage and collapse (x_1^*) and increasing **wave.height** leads to a significant decrease in the same coordinate. Figure 4.1a shows the regression lines in the ternary diagram obtained for several fixed values of variable **design** with **wave.height** as covariate. Figure 4.1b shows these lines in the *olr*-space. The green line is the regression line for the lowest value considered ($d = 3.5$), that is, the line

$$[p.service, p.damage, p.collapse] = ([0.9632, 0.0354, 0.0014] \oplus (3.5 \odot [0.4813, 0.2781, 0.2406])) \\ \oplus (h \odot [0.1487, 0.3563, 0.4950]) ,$$

where $d = 3.5$. Analogously, the blue line corresponds to the highest value considered ($d = 16.0$). Red lines are the lines for the intermediate values of d . All these lines are *parallel* because they share the gradient. Since the intercept of the LRM is $\beta_0 = [0.9632, 0.0354, 0.0014]$ (see table above) the regression lines depart from the vertex **p.service**. They arrive to the vertex **p.collapse** because the third component in β_2 takes the largest value (0.4950).

The univariate variation of the three parts in the composition [**p.service**, **p.damage**, **p.collapse**] is drawn in Fig. 4.2. In Fig. 4.2a we consider $d = 3.5$ whereas in Fig. 4.2b we take $d = 16.0$. In both figures, the covariate **wave.height** is in the horizontal axis.

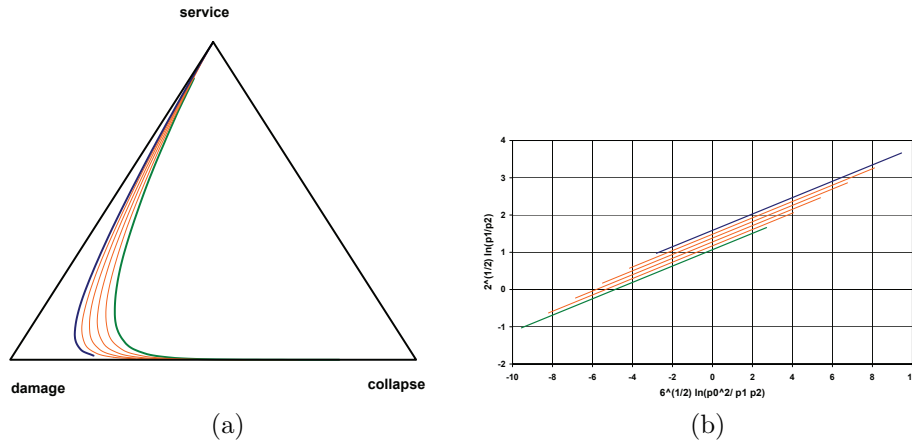


Figure 4.1. Regression compositional lines: (a) Ternary diagram; (b) on olr-coordinates. Green line corresponds to the lowest value of *design* ($d = 3.5$) and the blue line for the highest ($d = 16.0$). Red lines are the lines for the intermediate values.

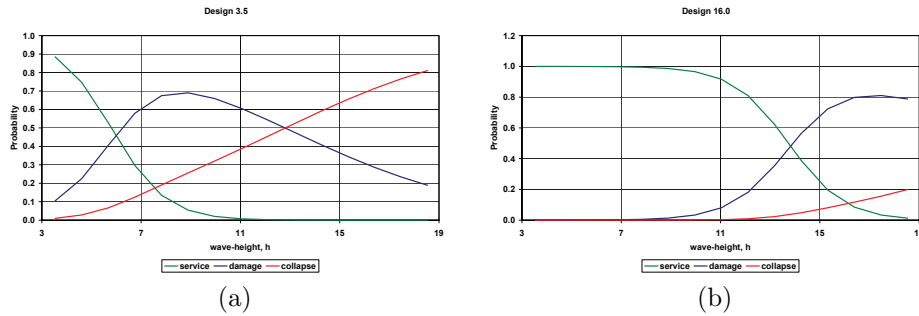



Figure 4.2. Univariate variation of the proportions $[p.\text{service}, p.\text{damage}, p.\text{collapse}]$ for different values of *wave-height*: (a) for design $d = 3.5$; (b) for design $d = 16.0$.

When one compares Figure 4.2a with 4.2b one observes that, accordingly β_j coefficients (see table above), $p.\text{service}$ (green line) increases with *design*. On the other hand, each figure shows that when *wave-height* increases then $p.\text{service}$ decreases and $p.\text{collapse}$ increases. The $p.\text{damage}$ (blue line) firstly increases but tends to decrease for large values of *wave-height*.

Activities for Section 4.1

 [Click here to get the activities of this section](#)

4.2. LRM for a scalar response and compositional predictor

We deal with an LRM where the compositional vector $\mathbf{x} \in \mathcal{S}^D$ are the explanatory variables of the model, which are used to explain a response variable y . This model is simpler than the previous one in many respects [CPG20]:

- no statistical assumption is made for the composition \mathbf{x} , but only for the residuals u of the variable y to be predicted,
- residual diagnostics are carried out exactly as in a standard LRM, and
- a single equation model is fitted, whose determination coefficient R^2 is interpretable directly.

The steps of the analysis are basically the same as before:

- (1) select an *olr*-basis in \mathcal{S}^D , for example using an SBP,
- (2) represent the predictor in *olr*-coordinates:

$$\mathbf{x}_i^* = \text{olr}(\mathbf{x}_i) \in \mathbb{R}^{D-1},$$

- (3) solve an ordinary regression problem in coordinates to obtain the β_j^* coefficients:

$$y_i = \beta_0 + \beta_1^* x_{i1}^* + \cdots + \beta_{D-1}^* x_{iD-1}^* + u_i,$$

- (4) back transform the coefficients (constant excluded) to obtain the *gradient* of the LRM:

$$\boldsymbol{\beta} = (\beta_1, \dots, \beta_D) = \text{olr}^{-1}(\beta_1^*, \dots, \beta_{D-1}^*).$$

The LRM model can be expressed in terms of the inner product of vectors $\boldsymbol{\beta}$ and \mathbf{x} . Indeed, it holds

$$\begin{aligned} y_i &= \beta_0 + \beta_1^* x_{i1}^* + \cdots + \beta_{D-1}^* x_{iD-1}^* + u_i = \\ &= \beta_0 + \langle \boldsymbol{\beta}^*, \mathbf{x}_i^* \rangle + u_i = \\ &= \beta_0 + \langle \boldsymbol{\beta}, \mathbf{x}_i \rangle_a + u_i. \end{aligned}$$

Note that the inner product $\langle \cdot, \cdot \rangle_a$ is invariant under change of *olr*-basis. Consequently, the predictions $\hat{y}_i = \beta_0 + \langle \boldsymbol{\beta}, \mathbf{x}_i \rangle_a$ are the same regardless the *olr*-coordinates used for the composition \mathbf{x} .

Global F tests and R^2 of the LRM are invariant under change of *olr*-basis. Individual t -tests provide the significance of individual coefficients. They are interpreted as the effect of increasing all parts in the numerator of the (logratio) coordinate by a common factor, while decreasing all components in the denominator by another common factor. This can lead to increases in the predicted value of the dependent variable (positive coefficient) or to decreases (negative coefficient). These individual coefficients and tests depend on the selected *olr*-basis and are as interpretable as coordinates themselves are [CPG20].

The gradient β is interpreted as the direction in the simplex along which the random composition \mathbf{x} must be perturbed to achieve the largest predicted increase in the dependent variable [TB11]. Alternative model formulations and interpretations are discussed in [CPG20].

Example 2. Employment distribution across industries.

According to the three-sector theory, as a country's economy develops, employment shifts from the primary sector (raw material extraction: farming, hunting, fishing, mining) to the secondary sector (industry, energy and construction) and finally to the tertiary sector (services). Thus, a country's employment distribution can be used as a predictor of several wealth indicators, such as the Gross Domestic Product (GDP).

The `eurostat_employment.cdp` file (see Appendix) contains EUROSTAT data on employment aggregated for both sexes, and all ages by economic activity (classification 1983-2008, NACE Rev. 1.1) in 2008 for the 29 EUROSTAT member countries, thus reflecting reality just before the 2008 financial crisis. Country codes in alphabetical order according to the country name in its own language are: Belgium (BE), Cyprus (CY), Czechia (CZ), Denmark (DK), Deutschland–Germany (DE), Eesti–Estonia (EE), Eire–Ireland (IE), España–Spain (ES), France (FR), Hellas–Greece (GR), Hrvatska–Croatia (HR), Iceland (IS), Italy (IT), Latvia (LV), Lithuania (LT), Luxembourg (LU), Macedonia (MK), Magyarország–Hungary (HU), Malta (MT), Netherlands (NL), Norway (NO), Österreich–Austria (AT), Portugal (PT), Romania (RO), Slovakia (SK), Suomi–Finland (FI), Switzerland (CH), Turkey (TR), United Kingdom (GB).

Our dependent variable is the logarithm (\ln) of gross domestic product per person in EUR at current prices ($\ln GDP$). Only for the purposes of exploratory data analyses it has also been categorised as a binary variable indicating values higher or lower than the median ($BinaryGDP$).

The employment composition ($D = 11$) is:

- *Primary sector*
 - x_1 : Primary_sector (agriculture, hunting, forestry, fishing, mining, quarrying)
- *Secondary sector*
 - x_2 : Manufacturing
 - x_3 : Energy (electricity, gas and water supply)
 - x_4 : Construction
- *Tertiary sector*
 - x_5 : Trade_repair_transport (wholesale and retail trade, repair, transport, storage, communications)
 - x_6 : Hotels_restaurants
 - x_7 : Financial_intermediation
 - x_8 : Real_estate (real estate, renting and business activities)
 - x_9 : Educ_admin_defense_soc_sec (education, public administration, defence, social security)
 - x_{10} : Health_social_work
 - x_{11} : Other_services (other community, social and personal service activities)

The biplot labelled by country and by the binary GDP variable (Figure 4.3) already suggests some meaningful associations between some tertiary industries and high GDP. Some of these industries played a role in the real state bubble leading to the 2008 financial crisis (Real_estate and Financial_intermediation). To define the

olr coordinates, we create a SBP. At the top of the SBP we separate the tertiary sector (all services) from the rest. Within services, we separate those involved in the pre-2008 real estate bubble (Financial.intermediation and Real.estate) from the rest of services, then these two from each other, then the pillars of the welfare state (Educ.admin.defense.soc.sec, Health.social.work) from the rest and from each other, then Other.services from the rest, and finally Trade.repair.transport from Hotels.restaurants. Finally, we separate the primary sector from the secondary sector. Within the secondary sector we separate Construction from the rest and then Energy from Manufacturing.

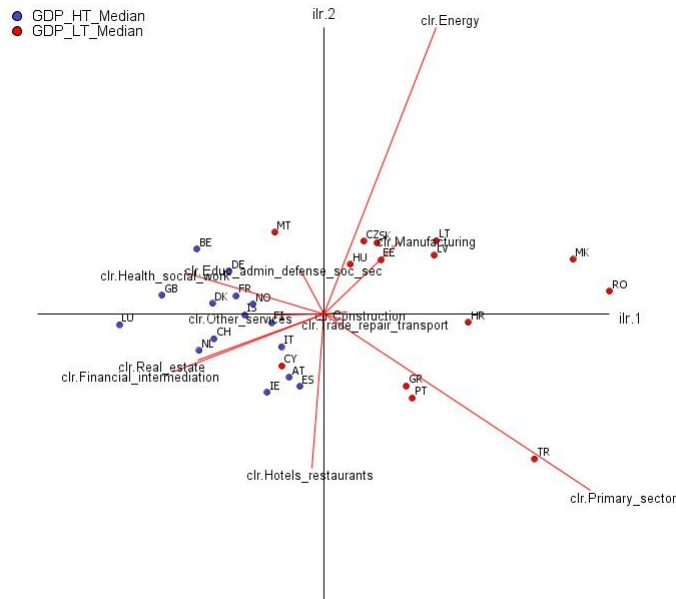


Figure 4.3. *clr*-biplot of employment distribution by GDP higher (blue dots: GDP_HT_Median) or lower (red dots: GDP_LT_Median) than the median.

The sign matrix for the partition is:

olr_k	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1^*	-1	-1	-1	-1	+1	+1	+1	+1	+1	+1	+1
x_2^*	0	0	0	0	-1	-1	+1	+1	-1	-1	-1
x_3^*	0	0	0	0	0	0	-1	+1	0	0	0
x_4^*	0	0	0	0	-1	-1	0	0	+1	+1	-1
x_5^*	0	0	0	0	0	0	0	0	-1	+1	0
x_6^*	0	0	0	0	-1	-1	0	0	0	0	+1
x_7^*	0	0	0	0	+1	-1	0	0	0	0	0
x_8^*	+1	-1	-1	-1	0	0	0	0	0	0	0
x_9^*	0	-1	-1	+1	0	0	0	0	0	0	0
x_{10}^*	0	-1	+1	0	0	0	0	0	0	0	0

The balance dendrogram separated by groups defined by having GDP per capita higher or lower than the median (Figure 4.4) seems to suggest that a higher employment in the service sector is associated to higher GDP, as does a higher weight of financial intermediation and real estate within services, or a higher ratio of Health_social_work over Educ_admin_defense_soc_sec. The largest variances are at the partitions involving the three sectors (tertiary from the rest $-x_1^*$ and primary from secondary $-x_8^*$).

We run the ordinary-least-squares regression to predict $\ln(\text{GDP})$ with the *olr*-coordinates built according to the same basis. The usual diagnostic plots (Figure 4.5) show residuals to be approximately normal and homoskedastic and no observations with a high Cook's distances emerge. The adjusted determination coefficient R^2 is very high at 93.2% (F -statistic = 39.39 on 10 and 18 d.f., p -value < 0.001). Employment composition as a whole is thus significantly related to $\ln(\text{GDP})$.

The significant regression coefficients at 5% (see table below) show that a higher employment in the service sector as opposed to all other sectors is associated to higher GDP (x_1^*). In more precise terms, increasing employment in all service industries by the same factor, while decreasing all primary and secondary industries by the same factor leads to a significant increase in $\ln(\text{GDP})$. In the same vein, a higher weight of financial intermediation and real estate services as opposed to the remaining services (x_2^*) leads to a higher predicted $\ln(\text{GDP})$.

	Estimate (β_j^*)	Std. Error	t -value	p -value
Intercept	10.196	0.791	12.893	<0.001
x_1^*	0.320	0.147	2.173	0.043
x_2^*	1.094	0.291	3.763	0.001
x_3^*	0.536	0.261	2.053	0.055
x_4^*	0.232	0.344	0.673	0.509
x_5^*	0.617	0.302	2.044	0.056
x_6^*	-0.436	0.417	-1.047	0.309
x_7^*	0.008	0.510	0.016	0.988
x_8^*	0.161	0.127	1.269	0.221
x_9^*	0.298	0.211	1.410	0.176
x_{10}^*	0.195	0.177	1.084	0.293

The inverse *olr*-transformation of the coefficient vector β^* is the perturbation vector β defining the direction in the simplex with a maximum increase of the dependent variable (see table below). In an 11-part composition, if the composition would be completely unrelated to the dependent variable, the perturbation vector would be the *neutral* perturbation $[1/11, 1/11, \dots, 1/11] = [0.0909, 0.0909, \dots, 0.0909]$. Once β is compared with the neutral perturbation we see that the increase in employment in Real_estate, Financial_intermediation, and Health_social_work, at the decrease of expenses in Manufacturing, Energy, Educ_admin_defense_soc_sec and Other_services lead to an increase of GDP per capita.

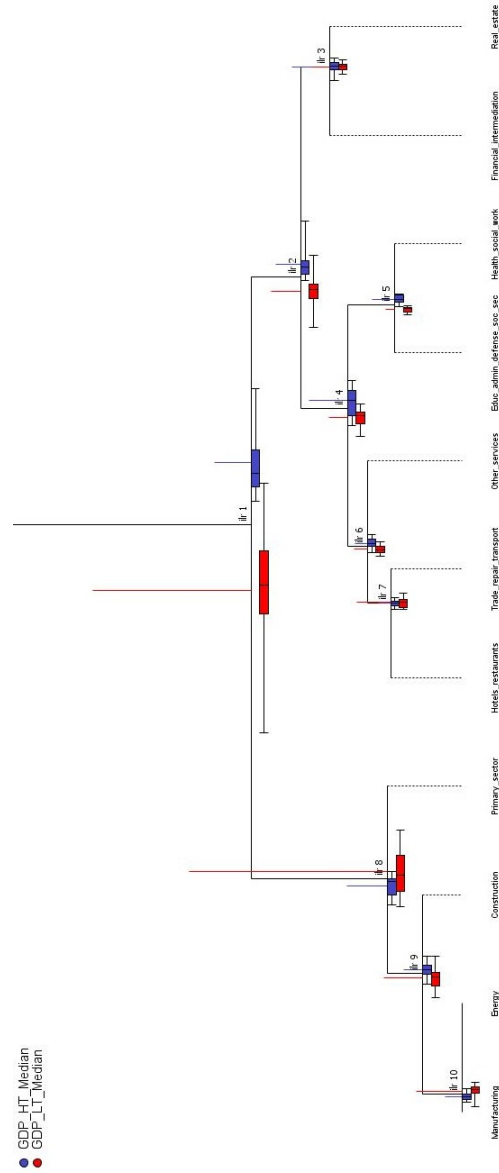


Figure 4.4. CoDa-dendrogram of employment distribution by GDP higher or lower than the median.

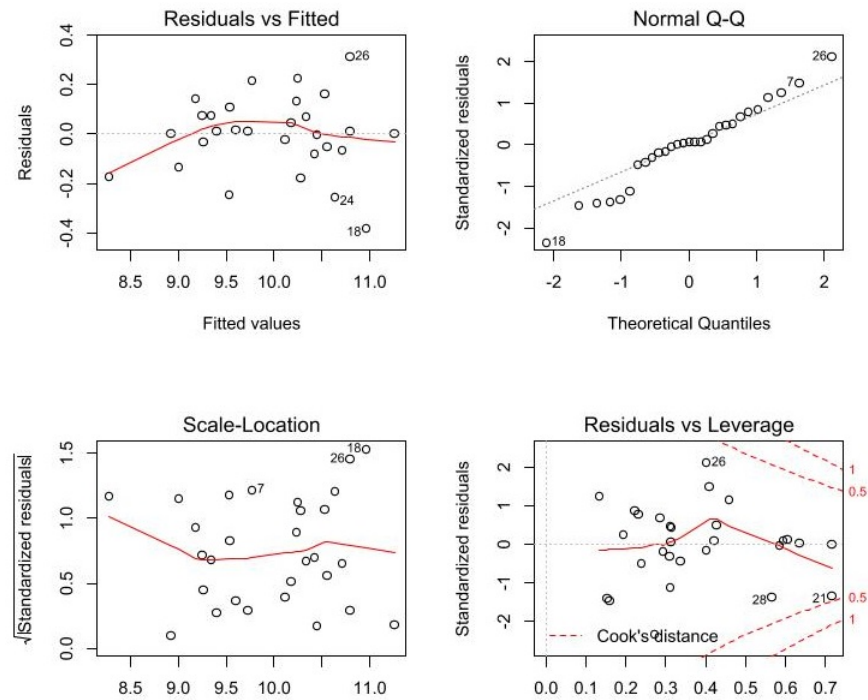


Figure 4.5. Residual plots. LRM of $\ln(\text{GDP})$ on employment distribution.

x_j	Estimate (β_j)
Primary_sector	0.0814
Manufacturing	0.0521
Energy	0.0686
Construction	0.0862
Trade_repair_transport	0.0735
Hotels_restaurants	0.0727
Financial_intermediation	0.1139
Real_estate	0.2430
Educ_admin_defense_soc_sec	0.0489
Health_social_work	0.1170
Other_services	0.0428

Activities for Section 4.2

[Click here to get the activities of this section](#)

4.3. LRM extensions

4.3.1. Extensions of an LRM with a compositional predictor. The extension from an LRM with an explanatory composition \mathbf{x} to a generalized linear model is straightforward [CMF17]. For instance, if the dependent variable y is a count, a *Poisson regression* can be specified, or if the dependent variable is ordinal or binary, an ordered or a binary *logit model* can be specified. Interpretation would then refer to the log expected count, to the logit, or to the appropriate expression in each case, taking the link function of the generalized linear model into account. Adding non-compositional predictors in the same model can also be done in a straightforward manner [CMF17] and nested models can be used to assess the predictive power of the compositional versus non-compositional predictors. The results of the non-compositional predictors are invariant to the selected *olr*-basis. The interpretation of the coefficients β of the compositional predictors \mathbf{x} , keeping the non-compositional predictors constant, is the same as outlined in the previous section. Non-compositional predictors may be numeric, dummy-coded factors or a combination of both. A very interesting particular case is including the information as regards total of parts as predictor, which [CMF17] recommend doing when the composition does not have a constant sum in their original units.

Example 3. Employment distribution (continued).

The `eurostat_employment.cdp` file (see Appendix) contains the additional factor variable `EU15_EU27_NOTEU` indicating if a country had joined the EU before or after 1995, or was not a EU member in 2008. This makes it possible to run three LRM: first, the composition-only model (see table above) where the predictor variable is the composition; second, the non-compositional-variables-only model with `EU15_EU27_NOTEU` as the unique predictor; and third, the complete model where both types of variables are predictors. When one compares the complete model with the composition-only model then one is testing the significance of variable `EU15_EU27_NOTEU` controlling by the employment distribution. Moreover, when one compares the complete model with the non-compositional-variables-only model then one is testing the significance of employment composition, controlling for the variable `EU15_EU27_NOTEU`. The results of both tests are shown in table below. The p -values suggest that employment composition is statistically significant whereas the factor `EU15_EU27_NOTEU` is not. Thus, we decide to keep the composition-only model (see table above) and discard the complete model.

	Variation in R^2	F	d.f.	p -value
Complete vs composition only	0.005	0.743	2 and 16	0.491
Complete vs <code>EU15_EU27_NOTEU</code> only	0.598	23.877	10 and 16	< 0.001

4.3.2. Compositions as both predictor and response. The set of predictors for a composition \mathbf{y} as dependent variable can include another composition \mathbf{x} as

explanatory variable. The β^* coefficients of \mathbf{x} refer to the effects of the increase of each explanatory *olr*-coordinate of \mathbf{x} on each dependent *olr*-coordinate of \mathbf{y} .

A particularly interesting case is to have the same composition measured at two points in time, that is, one has two CoDa sets \mathbf{X}_{t_1} and \mathbf{X}_{t_2} , where t_1 and t_2 are two different occasions. In this case, we can select $\mathbf{Y} = \mathbf{X}_{t_2}$ as the compositional response and $\mathbf{X} = \mathbf{X}_{t_1}$ as the predictor in the LRM. The same *olr*-basis should be selected for both compositions \mathbf{X}_{t_1} and \mathbf{X}_{t_2} . The key parameters are those showing to what extent each *olr*-coordinate in \mathbf{X}_{t_1} is a predictor of its own future value in \mathbf{X}_{t_2} .

When the LRM relates two different compositions \mathbf{y} and \mathbf{x} measured at the same point in time, parameters of the LRM are just as easy to interpret as the *olr*-bases are [BTD13]. When researchers are not able to build wholly interpretable *olr*-bases, they may rerun the model several times with several sets of *olr*-bases, both for the dependent \mathbf{y} and the explanatory \mathbf{x} compositions, in which particular parameters are of especial interest in each run [FHT18]. For example, if one component x_j of the explanatory composition \mathbf{x} is of especial interest, we define the *olr*-basis by an SBP where the first coordinate x_1^* balances the component x_j against the rest of the explanatory parts (i.e., pivot coordinates). Likewise, if one component y_k of the dependent composition \mathbf{y} is of especial interest, the first coordinate y_1^* of the dependent SPB can balance the component y_k against the rest of dependent parts. The coefficient β_{11}^* in the LRM relating both balances, if significant and positive, indicates that increasing x_j at the expense of decreasing all other parts in \mathbf{x} by a common factor leads to an increase in the relative importance of y_k within the composition \mathbf{y} . The LRM can be rerun as many times as combinations of parts in both \mathbf{y} and \mathbf{x} compositions are of interest to the researcher.

Activities for Section 4.3

 [Click here to get the activities of this section](#)

The chapter's key concepts

- ✓ The principle of working on coordinates applied to linear models allows us to fit compositional regression models by ordinary least squares.
- ✓ Coefficients can be interpreted in coordinates or back to the simplex.
- ✓ The composition may be the response or the predictor variable.

✓ The LRM can be extended to other general models.

Specific references in Chapter 4

- [BTD13] K.G. van den Boogaart and R. Tolosana-Delgado, *Analyzing compositional data with R*, Springer, Berlin, 2013.
- [CMF17] G. Coenders, J.A. Martín-Fernández and B. Ferrer-Rosell, *When relative and absolute information matter: compositional predictor with a total in generalized linear models*, Statistical Modelling **17**(6) (2017), 494–512.
- [CPG20] G. Coenders and V. Pawlowsky-Glahn, *On interpretations of tests and effect sizes in regression models with a compositional predictor*, SORT. Statistics and Operations Research Transactions **44**(1) (2020), 201–220.
- [Eg+12] J.J. Egozcue, J. Daunis-i-Estadella, V. Pawlowsky-Glahn, K. Hron and P. Filzmoser, *Simplicial regression. The normal model*, Journal of Applied Probability and Statistics **6** (2012), 87–108.
- [FHT18] P. Filzmoser, K. Hron and M. Templ *Applied compositional data analysis. With worked examples in R*. Springer Nature, Cham, 2018.
- [TB11] R. Tolosana-Delgado and K.G. van den Boogaart, *Linear Models with Compositions in R*. In: Compositional Data Analysis: Theory and Applications (eds V. Pawlowsky-Glahn and A. Buccianti), John Wiley & Sons, Ltd, Chichester, UK, 2011.

On the analysis of grouped data

Contents

- 5.1 Cluster analysis
- 5.2 Discriminant analysis
- 5.3 MANOVA

Objectives

- ✓ To learn how to form groups when the data set is compositional.
- ✓ To introduce how to calculate the linear discriminant function as a log-contrast.
- ✓ To properly analyse the difference between the centres of several groups using the MANOVA test.

5.1. Cluster analysis

The family of statistical *cluster analysis* methods focuses on the question: according to the values taken by the collected samples, can we make data groups? The answer to this question has application in many fields, including ecology, medicine, marketing, and social sciences. Broadly speaking, clustering methods are a wide range of multivariate methods to make *heterogeneous* groups or clusters composed of *homogeneous* samples. That is, we want samples which belong to the same cluster to be *similar*, and they should be *dissimilar* from samples in other groups. But, what does *similar* and *dissimilar* mean? Fig. 5.1 shows one possible strategy. The idea is to evaluate the homogeneity in each group when *comparing* samples with the corresponding mean of the group; and evaluate the heterogeneity among groups when *comparing* the group means with the overall mean. With this approach, one “simply” needs to decide how to make these *comparisons*.

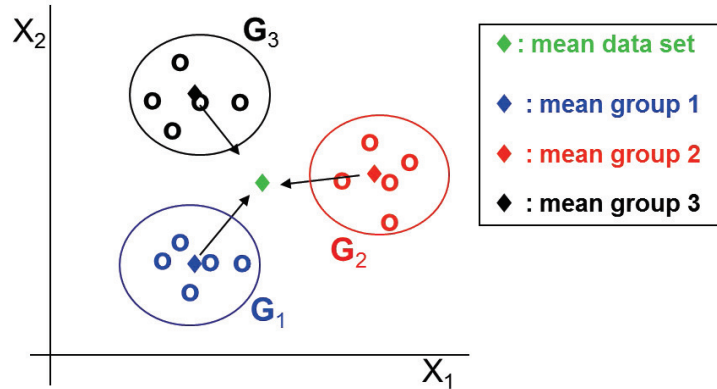


Figure 5.1. One strategy for evaluating heterogeneity among groups.

However, this strategy is not unique. A preliminary classification of clustering methods could be:

- *Parametric or model-based* methods: a probability distribution model is assumed for the data (e.g. normality of the *olr*-coordinates). The techniques to make groups are usually based on maximum likelihood methods [CMM16].

- *Non-parametric* methods: one needs to select an appropriate measure of dissimilarity for the data (e.g. compositional distance). The most common techniques are based on agglomerative hierarchical methods (e.g. Ward linkage method) and on k -means type methods.

Nowadays, there are many newer methods (e.g. fuzzy clustering [PMS12] and density-based clustering) that are described in short in approximately 2000 results under the *data clustering algorithms* category. In this course we are focusing on the non-parametric methods. The steps to performing a non-parametric cluster analysis in CoDa are:

- (1) Analysis of irregular data: outliers, missing and *zero* values.
- (2) Initial exploratory (logratio) data analysis: univariate, bivariate, PCA/biplot. Does any analysis suggests the existence of clusters?
- (3) Select a measure of dissimilarity: Aitchison distance, compositional Kullback-Leibler divergence, compositional L^p distances ($p \neq 2$), etc.
- (4) Select a clustering method: hierarchical, k -means, etc.
- (5) Decision regarding the number of clusters: how many clusters? (e.g. decision based on indices: Calinsky, Dunn and/or Silhouette).
- (6) Final exploratory data analysis using the cluster membership variable: biplot, comparison of group means, characterization of clusters in terms of original variables, etc.
- (7) Validation: cross validation, bootstrap, coephenetic correlation, etc.

After the universal first step, one has to look for any suggestion of groups. When some univariate or bivariate plot offers evidence about groups, then the variables involved will play an important role in the characterization of clusters (step #6). In any case, these simple plots have to be completed with some multivariate plot. The most widely used multivariate technique is the PCA/biplot. This method is based on a projection of the samples in a space of reduced dimension (see Chapter 2). In consequence, the distances between points in a form *clr*-biplot are approximations of the true Aitchison distances between samples. Usually, (but not always!) when the data set is formed by groups, a biplot representation gives evidence of clusters. However, in some cases, despite the groups being different, they appear mixed in the biplot.

To illustrate such problems we have considered the `statisticiantimebudget` CoDa set (see Appendix), which records the daily time (in hours) devoted to each activity (T, teaching; C, consultation; A, administration; R, research; O, other wakeful activities; S, sleep), recorded on each of 20 days. Next, we have generated a second group perturbing the 20 samples multiplying them component-wise by the vector $[1.2, 1, 1, 1, 1, 1]$, that is, increasing the first part (T) by 20%. Finally, we have created a third group perturbing the initial 20 samples by the vector $[1, 1, 1, 1.3, 1, 1]$ to increase the fourth component (R) by 30%. Figure 5.2 shows the *clr*-biplot of this data set. This representation is of a medium quality because the two first PCA axes only retain 61% of the variability. The samples in the first group are represented in blue. The compositions in the second group are shown in red and shifted slightly to the positive part of the *clr*-variable associated to the part T.

On the other hand, as expected, the green points representing the third group are shifted to the positive direction of the *clr*-part **R**. Using the MANOVA test (see Section 5.3), we obtain p -values below 0.00 indicating to us to reject the null hypothesis of equality of means. However, the samples from the different groups appear mixed in the *clr*-biplot, suggesting that there are no relevant differences between the groups artificially created and that there are other possible clusters within the data set.

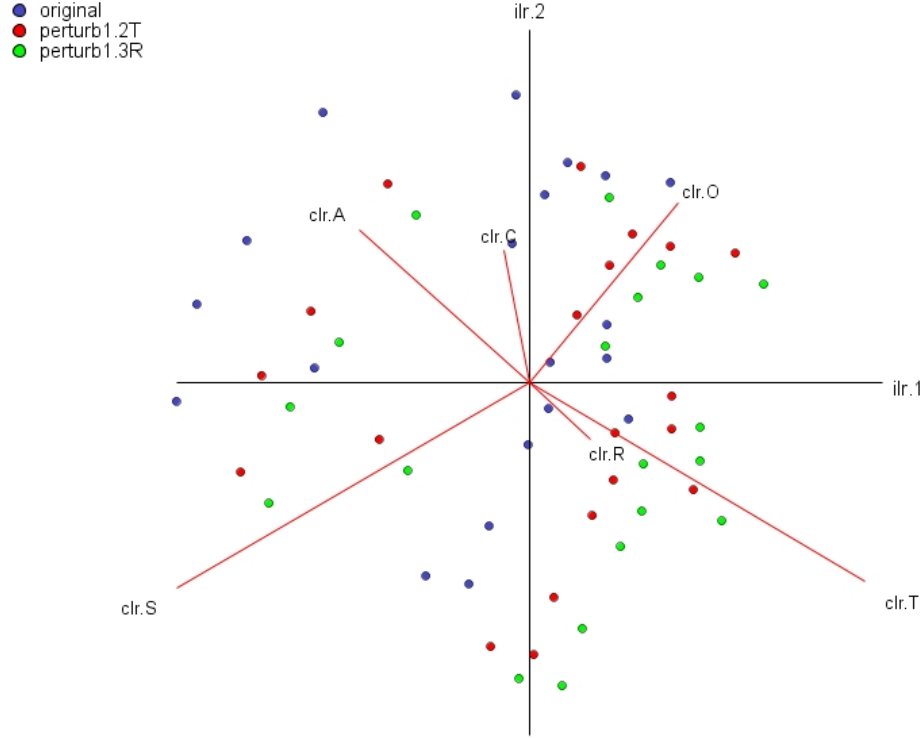


Figure 5.2. *clr*-biplot for **statisticiantimebudget** CoDa set and two artificial groups. Samples of the three groups are respectively represented in blue (original data set), red (data set perturbed by $[1.2, 1, 1, 1, 1, 1]$) and green (data set perturbed by $[1, 1, 1, 1.3, 1, 1]$).

A crucial point in a non-parametric cluster analysis consists of the selection of the measure of dissimilarity. It is obvious that the final structure of groups strongly depends on the measure applied. Although we are aware that, in some cases, when one uses a non natural measure one may obtain a reasonable classification, we advocate for selecting an appropriate measure of dissimilarity, coherent with the nature of the data.

For CoDa we have two well-known measures:

- *Aitchison distance*: Euclidean distance on coordinates (clr or olr) [ABMP00, MBP98]

$$d_a(\mathbf{x}, \mathbf{y}) = d(\text{clr } \mathbf{x}, \text{clr } \mathbf{y}) = d(\text{olr } \mathbf{x}, \text{olr } \mathbf{y});$$

- *compositional Kullback-Leibler (KL) divergence:*
 - traditional KL index: Given $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$, $d_{\text{KL}}(\mathbf{x}, \mathbf{y}) = \sum_k x_k \cdot \ln(x_k/y_k)$
 - CoDa-KL divergence:

$$d_{\text{M}}^2(\mathbf{x}, \mathbf{y}) = \frac{D}{2}(d_{\text{KL}}(\mathbf{u}, \mathbf{x} \ominus \mathbf{y}) + d_{\text{KL}}(\mathbf{u}, \mathbf{y} \ominus \mathbf{x})) = \frac{D}{2} \ln \left(\overline{\mathbf{x}/\mathbf{y}} \cdot \overline{\mathbf{y}/\mathbf{x}} \right) ,$$

where $\mathbf{u} = [1/D, 1/D, \dots, 1/D]$ is the neutral element of \mathcal{S}^D and $\overline{\mathbf{x}/\mathbf{y}}$ stands for the average of *ratio* vector $\mathbf{x}/\mathbf{y} = (x_1/y_1, x_2/y_2, \dots, x_D/y_D)$.

Both measures satisfy the following properties (d_* represents d_{a} and d_{M}):

$$d_*(\mathbf{x} \oplus \mathbf{p}, \mathbf{y} \oplus \mathbf{p}) = d_*(\mathbf{x}, \mathbf{y}) \quad \text{and} \quad d_*(\mathbf{x}_S, \mathbf{y}_S) \leq d_*(\mathbf{x}, \mathbf{y}) ,$$

for any $\mathbf{p} \in \mathcal{S}^D$ and for any subcomposition S . Moreover, d_{a} and d_{M} are approximately related by

$$d_{\text{c}}(\mathbf{x}, \mathbf{y}) \approx \sqrt{2} \cdot d_{\text{M}}(\mathbf{x}, \mathbf{y}) .$$

This property, based on a Taylor polynomial approximation, suggests that any monotone invariant cluster method will provide similar classifications using both measures d_{a} and d_{M} .

In the fourth step of a cluster analysis, one has to select the non-parametric method for clustering. A hierarchical agglomerative method starts by considering that each sample is a group. Then, in an iterative scheme, the technique hierarchically clusters samples and groups of samples until all the data are in a unique group. On analysing this hierarchical structure one can form the final classification.

On the other hand, k -means clustering seeks to partition the data set into a *specified number of groups*, k , by minimizing the *total within-group sum of squares*: $\text{trace}(\mathbf{W})$, where

$$\mathbf{W} = \sum_{g=1}^k \mathbf{W}_g , \quad \text{with } \mathbf{W}_g = (n_g - 1) \cdot \boldsymbol{\Sigma}_g ,$$

where $\boldsymbol{\Sigma}_g$ and n_g are the covariance matrix and the number of observations in the group g , respectively. Consequently, when k -means clustering is applied to log-ratio coordinates the measure of dissimilarity is, implicitly, the Aitchison distance [ABMP00, LCV18, MBP98]. In summary, the algorithm of the k -means clustering is:

- (1) decide k , number of groups (e.g. $k = 2$ in Fig. 5.3);
- (2) select k points (may be samples) as initial centres of the groups (e.g. two samples “*” connected by a line in Fig. 5.3(left));
- (3) for each sample \mathbf{x}_i , find the closest centre “*” and assign the sample to its group (e.g. clusters C_1^1 and C_2^1 in the Fig. 5.3(left));
- (4) recalculate the centres of the k groups (e.g. empty circles in the Fig. 5.3(left));
- (5) iterate steps 3 and 4 until convergence (e.g. clusters C_1^2 and C_2^2 in the Fig. 5.3(right) are the new groups and the filled circles its corresponding centres).

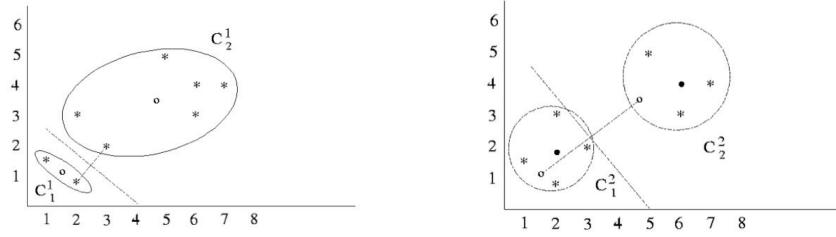


Figure 5.3. Two consecutive steps in the k -means algorithm for $k = 2$ groups: (left) initial step, where the two linked samples (*) are the initial centers and circles are the centers of the two formed groups; (right) second step, filled circles are the centers of the two formed groups.

For the convergence condition there are several possibilities, such as computing time or number of iterations. The most simple condition is stop the algorithm when the clusters obtained in one iteration are the same as in the previous one.

Once a clustering technique is applied and a classification is obtained, one has to decide if the number of groups is appropriate and how to characterise each group. These and the rest of the steps are illustrated via the following example.

Example 1. k -means for CoDa

The objective was to classify $n = 41$ districts of Catalonia (Fig. 5.4) according to the distribution of workers in $D = 8$ kinds of professions. Because the total number of workers in each district is not influent, a compositional approach may be useful.



Figure 5.4. Map of 41 districts in Catalonia

To make the compositions, the workers in each district were split into $D = 8$ kinds of professions (X_1 : technical; X_2 : staff; X_3 : administration; X_4 : commerce; X_5 : tourism; X_6 : agriculture; X_7 : industry; X_8 : armed forces).

After a preliminary analysis, no irregular data was detected. Figure 5.5 shows the clr-biplot of the data that retains 99.24% variability. This high quality indicates that the distances between points are very similar to the true distances. No evidences about well-separated groups is detected in the plot. However, some districts appear to be related to the Agriculture clr-axis and others to the Armed Forces direction.

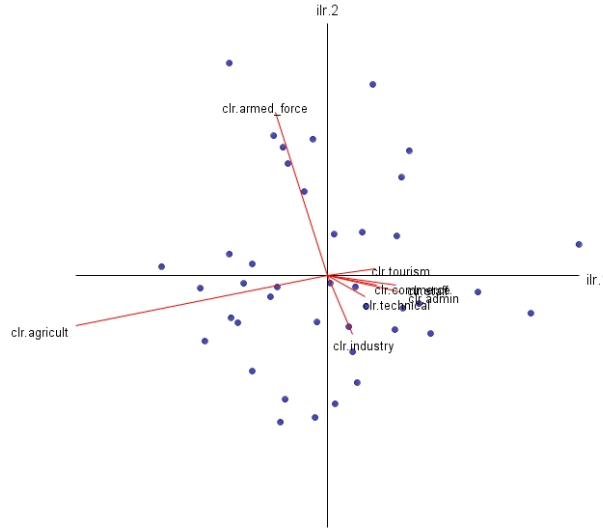


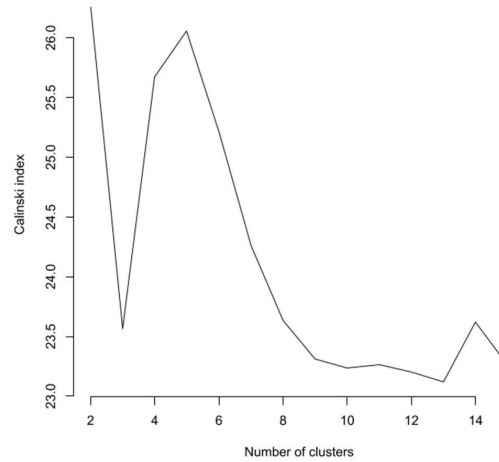
Figure 5.5. *clr*-biplot of labourers compositions in Catalonia.

The k -means algorithm can be applied once the number of clusters to be made is known. When the parameter k is unknown, a possible strategy is to apply the algorithm for several values of k . Later, these classifications in which the groups are more heterogeneous and the samples in each group are more homogeneous are selected as candidates for final classification. Since the algorithm minimises the total within-group sum of squares we can plot the values of $\text{trace}(\mathbf{W})$ for different values of k and look for the smallest values. Better still, we can plot the values of the *Calinski index*:

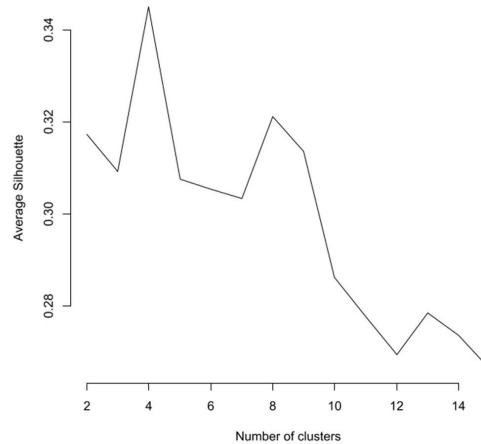
$$\text{Calinski}(k) = \frac{\text{trace}(\mathbf{B})/(k-1)}{\text{trace}(\mathbf{W})/(n-k)},$$

where matrix \mathbf{B} is the between-group sum of squares (see Section 5.3), that is, a measure of heterogeneity among clusters or the Average Silhouette [KR00] a measure of consistency which ranges from -1 to $+1$. When the Silhouette index for a particular sample approaches $+1$ indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. Consequently, the

further the Average Silhouette value is from +1, the worse the clusters' homogeneity. Assuming that the trace of a covariance matrix is a measure of total variability, note that the Calinski index is a ratio between “heterogeneity among groups” and “heterogeneity inside groups”. In consequence, we are interested in the values of k that provide the largest values of this index. Figure 5.6 shows the values of Calinski index and Average Silhouette until $k = 15$ groups. The Calinski index (Fig. 5.6a) takes its maximum for $k = 2$ clusters, followed by the cases of six and four groups. The plot of Average Silhouette (Fig. 5.6b) suggests a smoothed decreasing trend with maximums for $k = 4$ and $k = 8$.



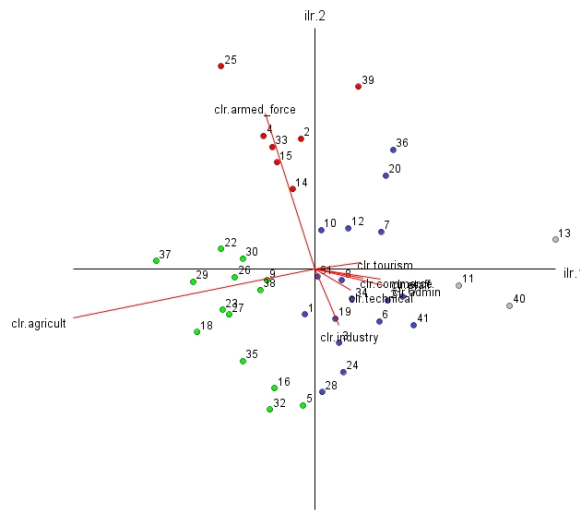
(a)



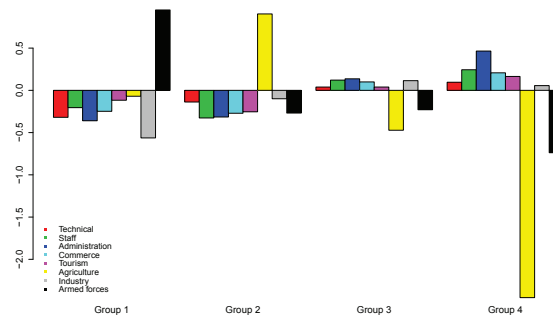
(b)

Figure 5.6. Heterogeneity plots for labourers compositions in Catalonia: (a) Calinski (b) Average Silhouette

After a preliminary examination of the groups, it was assumed that the best option is for four groups. A crucial question is “how to characterise each group in relation to the eight original parts?”. An initial possibility is represent the data in a *clr*-biplot, such as the plot in Fig. 5.7a. This figure suggests that the samples of group 1 (in red) are associated with relative high values in the *armed forces* part; the green group (group 2) is related to *agriculture*; the samples from the blue group (group 3) take centred values in the majority of the parts; and the grey group (group 4) is on the opposite side of the parts X_6 (agriculture) and X_8 (armed forces).



(a)



(b)

Figure 5.7. Plots to interpreting final classification for labourers compositions in Catalonia: (a) *clr*-biplot, where compositions are coloured by groups; (b) geometric-mean bar plot, where bars are coloured by parts. (X_1 : technical; X_2 : staff; X_3 : administration; X_4 : commerce; X_5 : tourism; X_6 : agriculture; X_7 : industry; X_8 : armed forces)

To analyse these patterns we should compare the average values of the samples with the overall average. To do this with CoDa, the geometric mean for each group and the overall geometric mean can be used. The geometric-mean bar plot (Fig. 5.7b) shows this comparison. To make this plot, we calculated the geometric mean vector for each group, and part wise, we made the *logratio* with the overall geometric mean. The bars with values near to zero suggest that the samples of the group take similar values to the overall geometric mean. The bars with large positive values indicate that the samples from the group take high values in the corresponding part. Large bars in the negative side of the plot correspond to the groups where the samples take values below the overall geometric mean. Figure 5.7b) suggests that samples in groups 1 (red group) and 2 (green group), take relative large values in parts X_8 (armed forces) and X_6 (agriculture), respectively. Group 4 (grey group) is characterised by relative small values in these two parts. Finally, the group 3 (blue group) is formed by samples that take relative averaged values.


There are several strategies to validate a final classification. For example, when the data set has a large number of samples one can use a k -fold scheme. This technique is a cross validation technique where the full data set is randomly split into k subsets and the clustering is separately applied to each one. Another possibility, also for validating the number of groups, is the use of a bootstrap algorithm. For each resampled data set one can plot the Calinski index and evaluate its sensitivity associated to the collected sample.

Both strategies can be used to evaluate the variability of the *coephenetic correlation*. This measure to evaluate whether the structure of the groups of samples is artificially produced by the clustering method or if it is natural, because of the groups that exist. The coephenetic correlation is the Pearson correlation coefficient between the original distances between samples and the distances associated to the cluster structure. The latter distances can be calculated in different ways according to the clustering method applied. For example, once the k -means provides the groups, one can decide that the distance between two samples regarding to the cluster structure is the distance between the corresponding group centres. According to this procedure, the value of coephenetic correlation for the compositions of laborers in Catalonia is 0.76, suggesting a high association between the original distances and the final clusters.

Cluster distance analysis: values shown in table below were obtained in R-program. The right column in the table shows the size of each group. In the upper triangle of the group-matrix, the distances between centres indicates that the centres of groups 1 and 3 are the closest. On the other hand, groups 2 and 4 are the more separated clusters. These patterns are coherent with the lower triangle values of the group-matrix. These values are the median of the distances between samples of corresponding groups. For example, 50% of the distances between samples in groups 2 and 4 are larger than 13.69. The values in the diagonal (boldface) are the maximum distance within the corresponding group. These values suggest that group 4 is most compact group whereas group 2 is the biggest cluster, twice as big as group 4.

	group 1	group 2	group 3	group 4	size
group 1	4.94	0.96	0.70	7.34	7
group 2	3.53	6.74	2.55	12.62	14
group 3	3.36	3.30	5.15	3.84	17
group 4	11.13	13.69	4.78	2.96	3

Activities for Section 5.1

 [Click here to get the activities of this section](#)

5.2. Discriminant analysis[†]

We take the word *discriminate* as a synonymous of *distinguish* or *differentiate*. That is, we want to be capable of recognising what makes a sample different or similar to other samples. To do this, we need an initial data set (*training data set*) where a collection of samples (i.e. compositions) is already classified in different groups. Using this sampling information, we will create some decision rules (*discriminant functions*) that will be applied to assign a new sample to one of the groups.

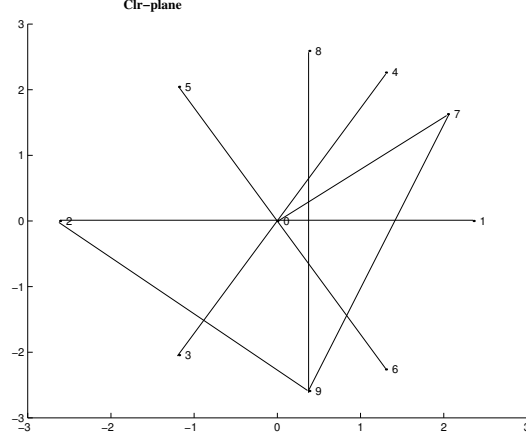
Let P be a population already divided into c classes or groups G_1, G_2, \dots, G_c . A discriminant analysis (DA) method aims to find a discrimination rule for assigning any individual $\mathbf{x} = [x_1, x_2, \dots, x_D]$ from P to a particular group G_j . From a *regression* perspective, we could assume that \mathbf{x} is the *explanatory* composition; the group G is the dependent variable, and we want to predict the value of G from the information collected in \mathbf{x} . Among the many different classification algorithms (Fisher's linear discriminant, logistic regression, naive Bayes classifier, support vector machines, quadratic classifiers, k -nearest neighbour, and neural networks) we will focus on the Fisher's or Linear DA (LDA) where the discrimination rule is based on compositional linear functions on \mathbf{x} . That is, the borders or discriminant curves will be simplicial linear manifolds or hyperplanes. These functions, which can be expressed in terms of a logcontrast (Chapter 1), take the simplicial patterns shown in Fig. 5.8a. When these compositional straight lines are expressed in log-ratio coordinates, one obtains the typical Euclidean straight lines drawn in Fig. 5.8b. In Fig. 5.8b we state that only the logcontrast with compositional direction to a vertex has the typical pattern of a Euclidean straight line.

Let \mathbf{X} be the training data set formed by n compositions $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{iD}]$ ($i = 1, 2, \dots, n$), where the group origin of each composition \mathbf{x}_i of \mathbf{X} is known. We assume *normality*, that is, compositions of a group G_j ($j = 1, \dots, c$), come from a log-ratio normal (or normal on the simplex) distribution $N_{SD}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. We also assume *homoscedasticity*, that is, the covariance matrices $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \dots, \boldsymbol{\Sigma}_c$ are all equal to $\boldsymbol{\Sigma}$. Usually, because of the parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}$ are unknown, they are estimated from compositions of \mathbf{X} .

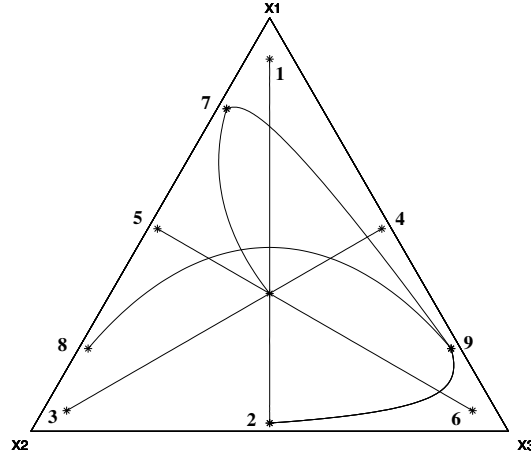
Fisher's discriminant rule is based on probabilities. That is, we will classify one composition \mathbf{x} to the group G_j with the largest probability $P[\mathbf{x} \in G_j]$. When these probabilities are calculated regardless of the information provided by compositions of the training data set \mathbf{X} , we call them *prior probabilities*. Indeed, let $\pi_j = P[\mathbf{x} \in G_j]$ be the initial or prior probability that an unclassified sample belongs to group G_j ($j = 1, 2, \dots, c$). There are two common options for these probabilities:

- *uniform*: $\pi_j = 1/c$, where c is the number of groups.

[†]Further information in [Ait86, Sections 7.11 (p. 176-180) and 12.6 (p. 293-297)] and [PET15, Section 8.4, p. 163-165].



(a)



(b)

Figure 5.8. To illustrate the change between simplicial and logratio coordinates. It can be observed straight lines connecting some compositions: (a) in log-ratio coordinates; (b) in the ternary diagram.

- *prior probabilities proportional* to sample size of groups $\pi_j = n_j/n$

Note that the uniform prior randomly assigns the group, whereas the latter prior gives the largest probability to the biggest group, which makes more sense for unbalanced sample size of groups.

When one wants to use the information provided by the training data set, then the discrimination rule has to maximise the *posterior* probabilities: $P[\mathbf{x} \in G_j | \mathbf{x}_{obs}]$. That is, given the observed random vector \mathbf{x}_{obs} , the probability that an unclassified sample \mathbf{x} belongs to group G_j ($j = 1, 2, \dots, c$). Note that the observed random vector \mathbf{x}_{obs} is the mathematical expression that refers to the information provided

by the training data set \mathbf{X} . In consequence, the classification rule classifies an unclassified sample \mathbf{x} to group j if

$$P[\mathbf{x} \in G_j | \mathbf{x}_{obs}] \geq P[\mathbf{x} \in G_k | \mathbf{x}_{obs}], \quad k = 1, 2, \dots, c$$

According to Bayes theorem

$$P[\mathbf{x} \in G_j | \mathbf{x}_{obs}] = \frac{P[\mathbf{x}_{obs} | \mathbf{x} \in G_j] \cdot \pi_j}{\sum_k P[\mathbf{x}_{obs} | \mathbf{x} \in G_k] \cdot \pi_k} = \frac{f_j(\mathbf{x}_{obs}) \cdot \Delta \mathbf{x}_{obs} \cdot \pi_j}{\sum_k f_k(\mathbf{x}_{obs}) \cdot \Delta \mathbf{x}_{obs} \cdot \pi_k} = \frac{f_j(\mathbf{x}_{obs}) \cdot \pi_j}{\sum_k f_k(\mathbf{x}_{obs}) \cdot \pi_k},$$

where the equality $P[\mathbf{x}_{obs} | \mathbf{x} \in G_j] = f_j(\mathbf{x}_{obs}) \cdot \Delta \mathbf{x}_{obs}$ is used to express a probability in terms of the corresponding density function. In consequence, because the denominator is the same for all the groups, the composition \mathbf{x} is classified to group j if

$$(5.1) \quad f_j(\mathbf{x}) \cdot \pi_j \geq f_k(\mathbf{x}) \cdot \pi_k, \quad k = 1, 2, \dots, c,$$

where, for the uniform prior, it reduces to $f_j(\mathbf{x}) \geq f_k(\mathbf{x})$, $k = 1, 2, \dots, c$, which indicates classifying \mathbf{x} to the group where it is most probable.

Moreover, when assuming normality and homoscedasticity:

$$f_j(\mathbf{x}) = \frac{1}{(2\pi)^{(D-1)/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (\text{olr } \mathbf{x} - \boldsymbol{\mu}_j) \Sigma^{-1} (\text{olr } \mathbf{x} - \boldsymbol{\mu}_j)^t \right\},$$

where $d_{Mah}(\mathbf{x}, G_j) = (\text{olr } \mathbf{x} - \boldsymbol{\mu}_j) \Sigma^{-1} (\text{olr } \mathbf{x} - \boldsymbol{\mu}_j)^t$ is the compositional Mahalanobis distance. In this case, the discriminant rule (Eq. (5.1)) is

$$d_{Mah}(\mathbf{x}, G_j) - \ln \pi_j \leq d_{Mah}(\mathbf{x}, G_k) - \ln \pi_k.$$

After some manipulation, it takes the form of a LDA function

$$(5.2) \quad \hat{L}_{jk}(\mathbf{x}) = (\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k) \hat{\Sigma}^{-1} (\text{olr } \mathbf{x})^t - \frac{1}{2} (\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k) \hat{\Sigma}^{-1} (\hat{\boldsymbol{\mu}}_j + \hat{\boldsymbol{\mu}}_k)^t - \ln(\pi_k / \pi_j).$$

Using these functions, the LDA rule classifies the composition \mathbf{x} to group G_j if $\hat{L}_{jk}(\mathbf{x}) \geq 0$ for any $k \neq j$. These functions can be expressed in terms of a logcontrast: $\hat{L}_{jk}(\mathbf{x}) = \alpha_0^{(jk)} + \alpha_1^{(jk)} \ln x_1 + \dots + \alpha_D^{(jk)} \ln x_D$, with $\sum_{i=1}^D \alpha_i^{(jk)} = 0$, where $\hat{L}_{jk}(\mathbf{x}) = 0$ defines the discrimination boundary between groups k and j , also known as *Bayes decision boundaries*. Importantly, note that another *olr*-basis or any other *alr*-coordinates produce the same LDA functions.

Moreover, using the functions $\hat{L}_{jk}(\mathbf{x})$ the posterior probabilities $P[\mathbf{x} \in G_j | \mathbf{x}_{obs}]$ can be calculated as

$$\begin{aligned} P[\mathbf{x} \in G_j | \mathbf{x}_{obs}] &= \frac{\exp(\hat{L}_{jc}(\mathbf{x}))}{1 + \sum_{k=1}^{c-1} \exp(\hat{L}_{kc}(\mathbf{x}))} \quad (j = 1, 2, \dots, c-1) \\ P[\mathbf{x} \in G_c | \mathbf{x}_{obs}] &= \frac{1}{1 + \sum_{k=1}^{c-1} \exp(\hat{L}_{kc}(\mathbf{x}))}. \end{aligned}$$

Note that $\sum_{j=1}^c P[\mathbf{x} \in G_j | \mathbf{x}_{obs}] = 1$ and that we assign the composition \mathbf{x} to the group G_j when $P[\mathbf{x} \in G_j | \mathbf{x}_{obs}] = \max \{P[\mathbf{x} \in G_k | \mathbf{x}_{obs}] : k = 1, 2, \dots, c\}$.

A crucial report in a DA is that of the evaluation of its performance. To make this evaluation we will use the information provided in the training data set \mathbf{X} . That is, we will apply the discrimination rule to the compositions in \mathbf{X} and evaluate the performance by making a comparison between the group assigned for the discrimination functions and the *true* group. A preliminary measure is the *misclassification rate*; that is, the proportion of wrongly classified compositions

after having applied the discriminant rule. However, because of the LDA functions used to classify a composition \mathbf{x}_i of \mathbf{X} has been calculated using the information included in \mathbf{x}_i , a biased estimation of the misclassification rate is obtained. To avoid this effect, the *cross-validation* techniques (e.g. leave-one-out or k -fold) are recommended.

In summary, to perform Fisher's LDA to CoDa we recommend the following steps:

- (1) Analysis of irregular data (outliers, missing and zero values) in the training set.
- (2) Initial exploratory data analysis using the group membership variable: univariate, bivariate, PCA/biplot.
Does any analysis suggests that the groups are well separated?
- (3) Decide the priori probabilities: uniform or proportional.
- (4) Estimate the LDA functions.
- (5) Classify the samples of the training set and estimate the misclassification rate. Use cross-validation techniques.
- (6) Check the assumptions: normality and homoscedasticity.
- (7) Classify new samples.

Example 2. Calc-alkaline vs. tholeiitic rocks

The `petrafm.cdp` file (see [Appendix](#)) contains the training CoDa set \mathbf{X} formed by $n = 100$ samples of 3-compositions of volcanic rocks from Ontario (Canada). The $D = 3$ parts are: $\mathbf{F} : FeO + 0.8998 \cdot Fe_2O_3$; $\mathbf{A} : Na_2O + K_2O$; and $\mathbf{M} : MgO$. The 100 samples are already classified in two groups: $n_c = 25$ from the *calc-alkaline* magma series, and $n_t = 75$ from the *tholeiitic* magma series.

After a preliminary analysis, no irregular data were detected. Figure 5.9 shows that rocks from the calc-alkaline magma series (in blue), can be well distinguished from rocks from the tholeiitic magma series (in red) on the *AFM* diagram. Note that although the two groups are well differentiated, a Euclidean straight line to separate these samples can not be drawn, because we are in the simplex and, moreover, such a line will not be able to separate these samples.

For this data set it is more reasonable to take the prior probabilities proportional to the group size because the tholeiitic group is thrice the size of the calc-alkaline group. However, we will show here the results obtained when uniform prior probabilities ($\pi_{calc} = \pi_{thol} = 0.5$) were assumed.

To work on coordinates we consider the data set $\text{olr } \mathbf{X} = \{\mathbf{u} = [u_1, u_2] = \text{olr } \mathbf{x} : \mathbf{x} \in \mathbf{X}\}$, where $\text{olr } \mathbf{x} = [\sqrt{2/3} \cdot \ln(\sqrt{x_1 x_2}/x_3), \sqrt{2}/2 \cdot \ln(x_1/x_2)]$. Using those coordinates, the parameters estimated for the normal distributions are

$$\hat{\Sigma} = \begin{bmatrix} 0.232 & -0.174 \\ -0.174 & 0.171 \end{bmatrix} \quad \hat{\mu}_{calc} = [0.433, 0.093] \quad \hat{\mu}_{thol} = [0.008, 1.046]$$

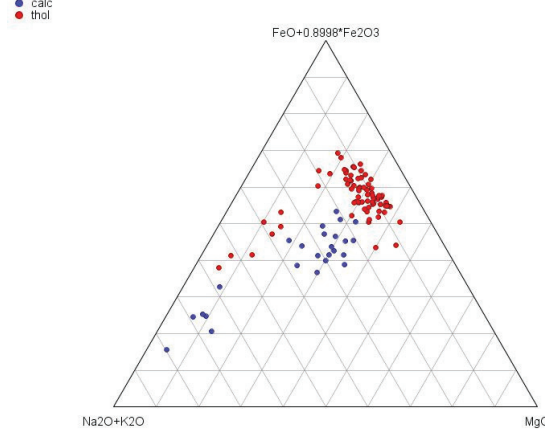


Figure 5.9. Petrafm CoDa set in the ternary diagram: calc-alkaline magma series (blue color) and tholeiitic magma series (red color)

Using these estimates and the prior probabilities in Eq. (5.2) we obtained the LDA function

$$\hat{L}_{calc,thol}(\mathbf{u}) = -9.68 u_1 - 15.39 u_2 + 10.90 = 0,$$

that after expressing the *olr*-coordinates in terms of the original variables is

$$-14.83 \ln F + 6.93 \ln A + 7.90 \ln M + 10.90 = 0.$$

Figure 5.10a shows the Bayes decision boundary in the log-ratio space. Note that there are some calc-alkaline data (in blue) and some tholeiitic data (in red) that are misclassified, they are on the *wrong* side of the LDA boundary. This LDA function is represented in the ternary diagram in Fig. 5.10b.

At the end of the LDA process, it is hoped that each group will have a normal distribution of discriminant scores (i.e. values for the discriminant function). The normality is due to the definition of the discriminant function as a linear combination of univariate random variables that are normally distributed. The degree of overlap between the discriminant score distributions can then be used as a measure of the success of the LDA technique. Figure 5.11 shows one example of two discriminant score distributions where the overlap degree is small and one expects a minimal misclassification.

When a preliminary evaluation of the performance was made, then the results shown in table (see below) were obtained. Note that these results are obtained without cross-validation, that is, they are biased. Only three calc-alkaline data were misclassified as tholeiitic. The same number of tholeiitic data was wrongly classified. In consequence, the misclassification rate (without cross-validation) is equal to $(3 + 3)/100 = 6\%$, which could be considered reasonable. Indeed, the *false calc-alkaline rate* (the fraction of the tholeiitic that are classified as calc-alkaline) is even better because is equal to $3/75 = 4\%$. However, the *false tholeiitic rate* (fraction of the calc-alkaline that are classified as tholeiitic) is much worse because is equal to $3/25 = 12\%$, which is not a reasonable misclassification level.

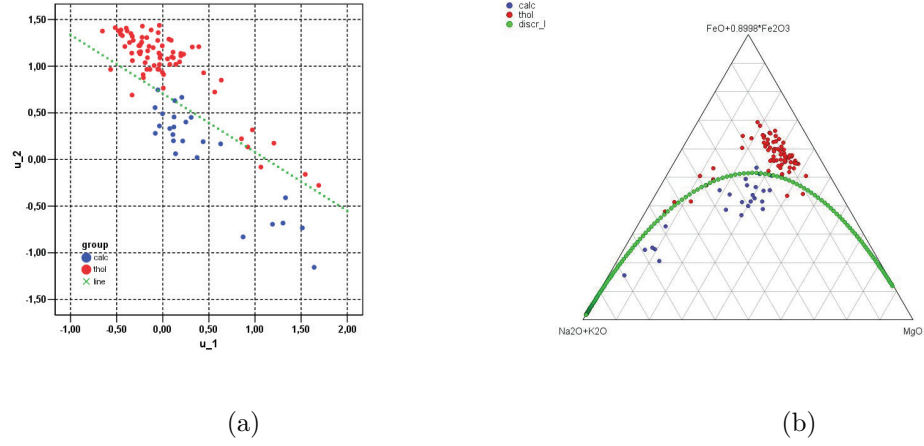


Figure 5.10. Bayes decision boundary for the **Petrafm** CoDa set: (a) in the logratio space; (b) in the ternary diagram. Calc-alkaline magma series (blue color) and tholeiitic magma series (red color).

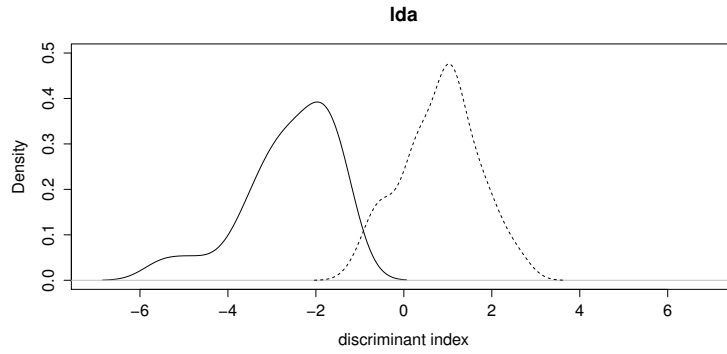


Figure 5.11. Densities of discriminant scores: degree of overlap represents the expected magnitude of misclassification.

Original group	Predicted group		
	calc-alk	thol	Total
calc-alk	22	3	25
thol	3	72	75
Total	25	75	100

To check the homoscedasticity we used the *Bartlett-Box test* where the null hypothesis is $H_0 : \Sigma_1 = \dots = \Sigma_c$ and the test statistic is

$$M = \frac{\prod_{k=1}^c |\hat{\Sigma}_k|^{(n_k-1)/2}}{|\hat{\Sigma}|^{(n-c)/2}},$$

that has a χ^2 distribution with $D \cdot (D - 1) \cdot (c - 1)/2$ degrees of freedom. Using the sample covariance matrices of each group

$$\hat{\Sigma}_{calc} = \begin{bmatrix} 0.292 & -0.260 \\ -0.260 & 0.277 \end{bmatrix} \quad \hat{\Sigma}_{thol} = \begin{bmatrix} 0.212 & -0.146 \\ -0.146 & 0.137 \end{bmatrix},$$

the result of the test is $M_{obs} = 6.37$. Because of the $p\text{-value}(\chi^2_3) = 0.896$ is clearly greater than $\alpha = 0.05$, it fails to reject H_0 , that is, we can assume that there is no difference between both covariance matrices.

There are many tools to check the multivariate normality. Two of them are the χ^2 -plot of the squared Mahalanobis distances and the *generalized Shapiro-Wilk* test [VG09]. In LDA, we apply these tools to the *residuals* (in coordinates), that is, the difference of each sample to the centre of its group. Under normality, the squared Mahalanobis distances of samples to the centre are χ^2 distributed. Figure 5.12 shows the chi-square plot where the OX axis contains the values of the χ^2 distribution and the OY axis represents the Mahalanobis distances. As for a typical Q-Q plot, to evaluate the normality we have to evaluate the linear trend of the cloud of points. Figure 5.12 shows an excellent linear pattern and suggests that the multivariate normal model fits the residuals well.

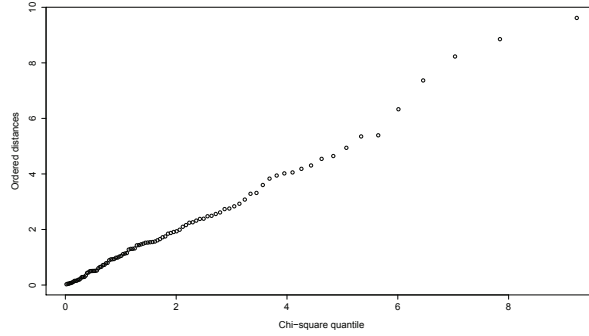


Figure 5.12. χ^2 -plot of the squared Mahalanobis distances for the `petrafm` data set.

To confirm the pattern observed in Figure 5.12, we applied the generalized Shapiro-Wilk test where the null hypothesis is H_0 : *multivariate normal model fits the data*. The value obtained for the test statistic was $MVW = 0.9898$, and its corresponding $p\text{-value} = 0.746$, therefore it indicates failure to reject H_0 .

Once the assumptions were checked, one can expect that the classification of new samples will be made with a misclassification rate equal to 6%.

Activities for Section 5.2

🔗 [Click here to get the activities of this section](#)

5.3. MANOVA

Broadly speaking, the multivariate analysis of variance technique (MANOVA) aims to test whether or not a categorical (independent) variable has an overall effect on a collection \mathbf{x} of continuous variables (dependent). That is, *testing* the statistical significance of the *mean differences* among *groups* (or *classes*) G_1, G_2, \dots, G_c . In consequence, the model for a random vector \mathbf{x} is $\mathbf{x}_j = \boldsymbol{\mu}_j + \boldsymbol{\varepsilon}$ ($j = 1, 2, \dots, c$), where $\boldsymbol{\mu}_j$ is the mean vector of group G_j and $\boldsymbol{\varepsilon}$ is a random vector of residuals. Under the usual *normality* and *homoscedasticity* assumptions, the distribution of the random vector \mathbf{x}_j is $\mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$, that is, the residuals $\boldsymbol{\varepsilon}$ are $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ distributed.

The MANOVA test has the null hypothesis $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \dots = \boldsymbol{\mu}_c$. In consequence, when it is rejected, is assumed that there is at least one group with a significant difference in its mean vector. Figure 5.13 shows a typical scenario where the null hypothesis should be rejected because there is evidence of differences among the mean vectors of three groups.

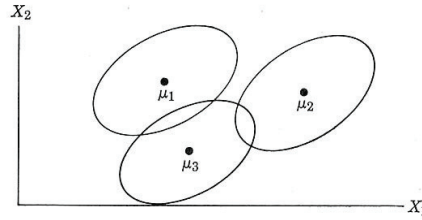


Figure 5.13. Typical scenario where the three mean vectors are significantly different from one another.

The crucial question is how to decide if there are no differences between the mean vectors of a collection of groups. The proposed approach in MANOVA is an extension of the strategy applied in the univariate case (ANOVA). In summary, one will measure *differences* between mean vectors in terms of variability, calculated using the sum of squares. Indeed, the property of decomposition of variability is the central concept in this analysis. This property states that the *total variability* is equal to the sum of *between variability* and *within variability*,

$$\mathbf{T} = \mathbf{B} + \mathbf{W},$$

where:

- *total variability* or Total Sum of Squares (TSS): $\mathbf{T} = (n - 1)\boldsymbol{\Sigma}$;
- *between variability* or Between-group Sum of Squares (BSS): $\mathbf{B} = (c - 1)\boldsymbol{\Sigma}_c$;
- *within variability* or Within-group Sum of Squares (WSS): $\mathbf{W} = \sum_{g=1}^c \mathbf{W}_g$,

where Σ is the overall covariance matrix; Σ_c is the pooled covariance matrix for the c mean vectors; and $\mathbf{W}_g = (n_g - 1)\Sigma_g$, where Σ_g is the covariance matrix of the g -th group.

Figure 5.14a shows that the *total variability* measures the overall differences when comparing each sample with the overall mean vector. To measure the *between variability* (Fig. 5.14b) the mean vector of each group is compared with the overall mean vector. This measurement informs about the degree of overlap among the groups. Figure 5.14c indicates that the *within variability* accounts for the accumulation of variability in each group.

Because \mathbf{B} measures the degree of overlap, one needs to evaluate the *relative* importance of \mathbf{B} , that is, if \mathbf{B} is large or not. Figure 5.15a shows the scenario in which the *between variability* is large and there are differences between the mean vectors. When there are no differences between groups (Fig. 5.15b) the importance of \mathbf{B} is small.

Importantly, the evaluation of \mathbf{B} must be *relative* to \mathbf{T} because we have to take into account the *total variability* of the data set. Indeed, one has to evaluate if the ratio between the size of \mathbf{B} and the size of \mathbf{T} is large or not. Because of $\mathbf{B} + \mathbf{W} = \mathbf{T}$, this evaluation may be done in terms of \mathbf{W} . This approach is in the definition of the test statistic Wilks' lambda $\Lambda = |\mathbf{W}|/|\mathbf{T}|$. Note that this statistic evaluates the size of each component through the determinant of a matrix, that is, the product of its eigenvalues. Rao's transformation of Λ , Ra , has approximately a *Fisher's F* distribution. A *large* value of the value observed of Ra , Ra_{obs} , means a p -value close to zero, that is, rejecting H_0 . Indeed,

- In Figure 5.15a: \mathbf{B} is large $\rightarrow \mathbf{T} \gg \mathbf{W} \rightarrow$ small $\Lambda_{obs} \rightarrow$ large $Ra_{obs} \rightarrow$ small p -value \rightarrow reject $H_0 : \mu_1 = \mu_2 = \dots = \mu_c$.
- In Figure 5.15b: $\mathbf{B} \approx \mathbf{0} \rightarrow \mathbf{T} \approx \mathbf{W} \rightarrow$ large $\Lambda_{obs} \rightarrow$ small $Ra_{obs} \rightarrow$ large p -value \rightarrow failure to reject $H_0 : \mu_1 = \mu_2 = \dots = \mu_c$.

The other common test statistics are Pillai's trace ($trace(\mathbf{B} \cdot \mathbf{T}^{-1})$), Lawley-Hotelling trace ($trace(\mathbf{W}^{-1} \cdot \mathbf{B})$), and Roy's largest root of matrix $\mathbf{W}^{-1} \cdot \mathbf{B}$. Nowadays, the discussion over the merits of each statistic continues and common software routines allow the four statistics to be calculated. In the case of two groups, the four statistics are equivalent and the MANOVA test reduces to Hotelling's T-square test.

Importantly, when using MANOVA with CoDa [PET15, Section 8.3, p. 160-163], the above tests are invariant under a change of log-ratio basis because the four statistics are invariant functions of the eigenvalues of matrix $\mathbf{W}^{-1} \cdot \mathbf{B}$. This fact facilitates the use of the tests because one can work with the *olr*-coordinates obtained from any SBP [MDM15].

To summarise, the recommended steps for performing a MANOVA for CoDa are:

- (1) Analysis of *irregular data*: outliers, missing and *zero* values.
- (2) Initial exploratory data analysis using the group membership variable: univariate, bivariate, PCA/biplot. Does any analysis suggests that the group means are different?

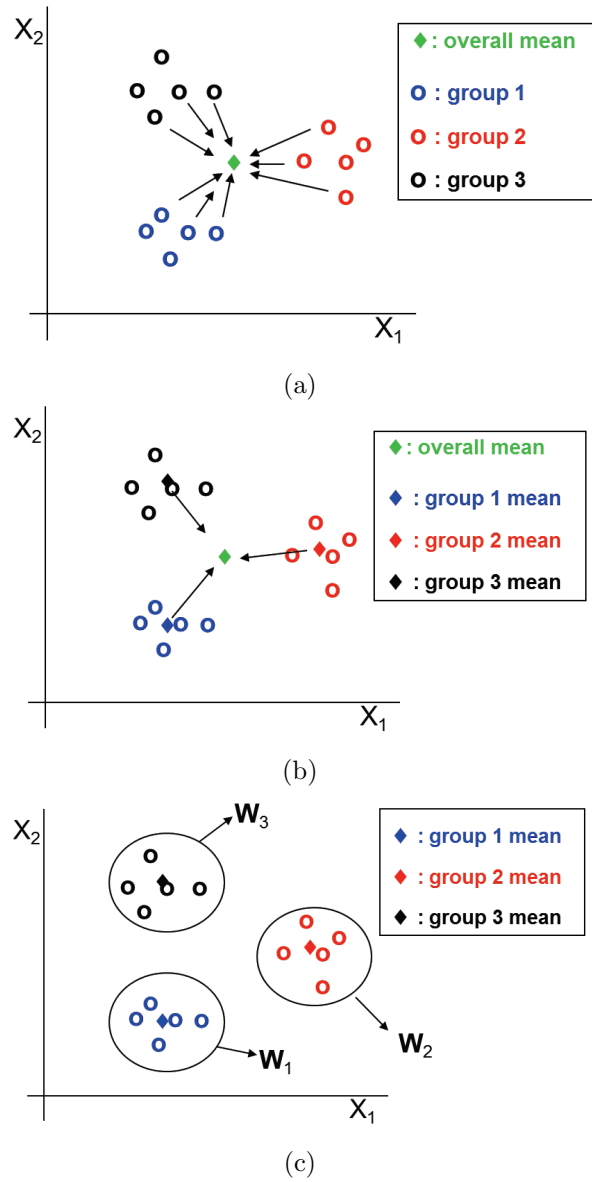


Figure 5.14. Variability decomposition: (a) Total variability; (b) Between variability; (c) Within variability.

- (3) Perform Wilks' lambda statistic. Reject or fail to reject H_0 ?
- (4) Check the assumptions: normality and homoscedasticity of residuals.
- (5) If H_0 is rejected, analyse the differences between group means: by group and by variable (*post-hoc* tests).

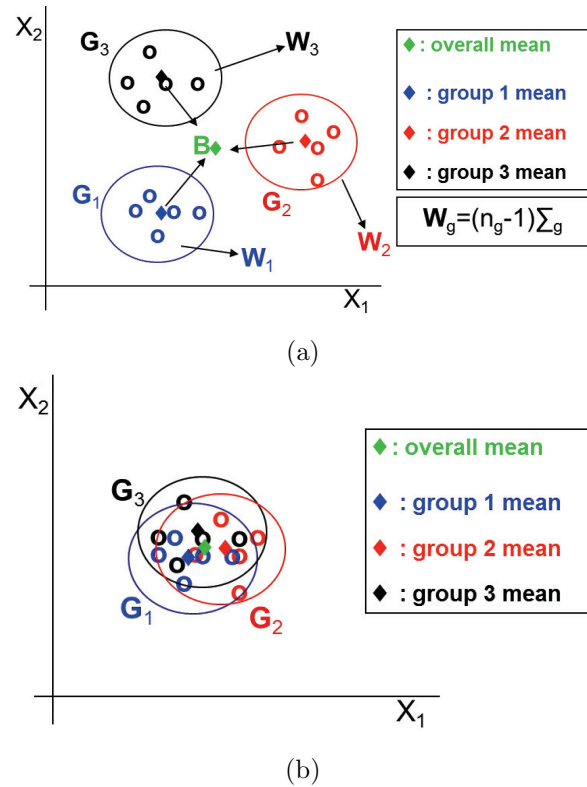


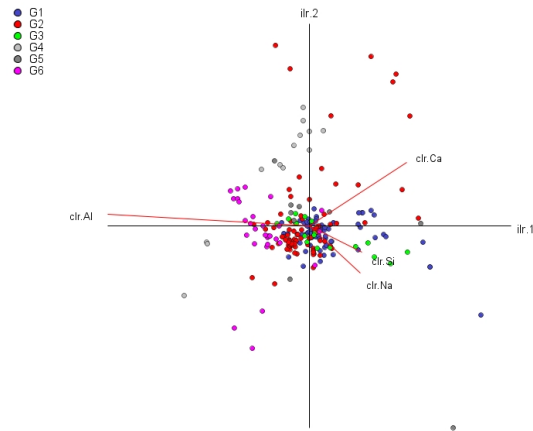
Figure 5.15. The between variability size: (a) large (reject H_0); (b) small (fail to reject H_0).

Example 3. Calc-alkaline vs tholeiitic rocks

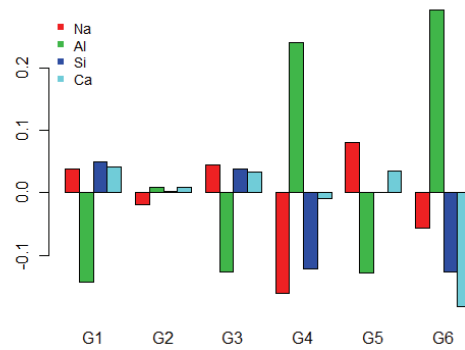
In this example we are going to analyse a subcomposition ($D = 4$) of major chemical elements. The CoDa set, called **forensic glass** data set, contains $n = 214$ samples classified in $c = 6$ different types of glasses. The full data set is one of the examples in the book “Pattern Recognition and Neural Networks”, Ripley, B.D., Cambridge University Press (1996). The data set is available at <http://www.stats.ox.ac.uk/~ripley/PRbook/>.

In a preliminary analysis, no irregular data (zeros) were detected. Figure 5.16a shows the *clr*-biplot of the data set. The groups are distinguished using colours. The quality of this representation is high because it retains 99% of the overall variability. Some groups, such as G_4 and G_6 are well separated, other groups, such as G_1 and G_3 , are mixed between them. The group G_2 is in the centre and mixed with the others. On the other hand, the data of G_5 is spread across the entire space. The profiles of the geometric centres (Fig. 5.16b) indicate that data in G_1 and G_3 are characterised by small values in the chemical element aluminium (Al). On the other hand, groups G_4 and G_6 have data that have large values in Al, and small values in silicon (Si). These groups are distinguished because data of G_4 have small values in

sodium (Na) but data from G_6 have small values in calcium (Ca). The largest values in the chemical element Na are in the G_5 data. The profile of the geometric mean of group G_2 is very similar to the overall profile. In summary, both plots suggest that there are differences between some mean vectors of the groups.



(a)



(b)

Figure 5.16. Subcomposition of Forensic data set: (a) *ckr*-biplot; (b) geometric means bar plot.

Table (see above) shows a summary of the results provided by Wilks' lambda test. The p -value for the test statistic ($\lambda = 0.378$) is smaller than $2.2 \cdot 10^{-6}$, which

leads to rejecting the null hypothesis and assuming that not all the mean vectors are equal.

	df	Wilks	approx. F	numer. df	denom. df	p -value
Type	5	0.378	16.031	15	569.08	$< 2.2e^{-16}$
Residuals	208					

df: degree freedom

The p -value obtained is a probability from an F -Fisher distribution that, under normality and homoscedasticity, fits the Ra transformed Wilks' lambda statistic. The homoscedasticity was checked using the statistics M associated to the Bartlett-Box test. The results were $M_{obs} = 255.428$, which for a χ^2_{30} gives a p -value > 0.99 , indicating failure to reject H_0 , that is, we can assume that the covariance matrices are equal. The residuals were expressed back to a raw composition by means of the inverse olr -transformation and the back-transformed residuals were submitted to log-ratio normality tests. The p -value for the radius test was below 0.01 for all available statistics (Anderson-Darling, Cramer-von Mises and Watson) and suggests that the normal distribution does not fit the residuals of the MANOVA model very well. Although many authors defend in the literature that MANOVA is a robust test for a lack of normality, in these situations it may be useful to perform a MANOVA non-parametric test based on distances.

Once we have assumed that at least one group is different, the differences between group means should be analysed. For $c = 6$ groups, there are $c \cdot (c - 1) / 2 = 15$ possible comparisons. To perform multiple tests some modification must be applied. One possibility is the Bonferroni correction, that is, to multiply by 15 the corresponding p -values and to compare them with the typical level $\alpha = 0.05$. The p -values set out in table (see below) indicate that there are no significant differences between groups G_1 and G_3 , or between G_2 and G_3 . The groups G_4 and G_5 are significantly different (0.046) but the p -value is larger than the p -values obtained when comparing these groups with the others. Group G_6 is significantly different from the rest of groups.

	G_1	G_2	G_3	G_4	G_5	G_6
G_1	1	< 0.01	> 0.99	< 0.01	< 0.01	< 0.01
G_2	< 0.01	1	0.34	< 0.01	< 0.01	< 0.01
G_3	> 0.99	0.34	1	< 0.01	0.02	< 0.01
G_4	< 0.01	< 0.01	< 0.01	1	0.046	< 0.01
G_5	< 0.01	< 0.01	0.02	0.046	1	0.01
G_6	< 0.01	< 0.01	< 0.01	< 0.01	0.01	1

When it is assumed that two groups are different, then it is recommended to analyse in which parts the differences are larger. Therefore, we have calculated the corresponding geometric means and compared the values for each part. To summarise this analysis we have calculated the logratio between the geometric mean vectors of each group and represented this vector in a bar plot. For example, table below shows the results for groups G_1 and G_6 . Samples of group G_1 have smaller values in the Na and Al chemical elements, and larger values in Si and Ca. In consequence, the log-ratio vector has negative values only in the two first components.

	Na	Al	Si	Ca
G_1	0.1382	0.0117	0.7584	0.0917
G_6	0.1474	0.0212	0.7454	0.0861
logratio	-0.0642	-0.5936	0.0173	0.0631

Because of the definition of a log-ratio vector, the closer the value to zero the more similar the parts are. Figure 5.17 suggests that the samples of groups G_1 and G_6 have similar values in all the parts except in the chemical element aluminium. For part Al the group G_6 samples have larger values than in G_1 . The horizontal lines in the plot inform about some of the ratios values, that is, the magnitude of the difference. Since the bar of part Al is larger than 1.5 we can interpret that, on average, G_6 samples have values more than 50% larger than in G_1 . For the Na and Ca chemical elements the differences are around 5%. No relevant differences are observed for silicon. The relevant differences between the groups G_1 and G_6 samples as regards aluminium are also shown in Figure 5.16. In the bar plot, the groups have the Al bar on opposite sides; and in the biplot, the samples have a different location along the direction of the axis associated to the aluminium.

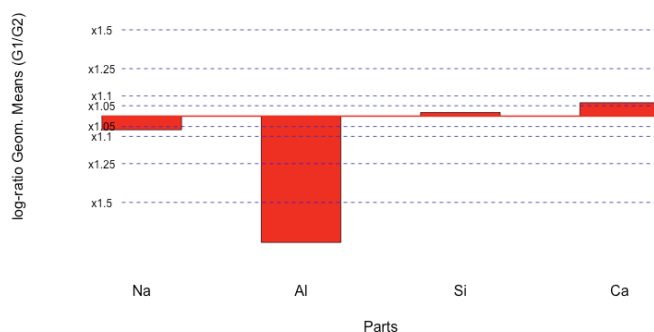


Figure 5.17. Bar plot of log-ratio vector of geometric means for groups G_1 and G_6 .

Activities for Section 5.3

[Click here to get the activities of this section](#)

The chapter's key concepts

- ✓ Multivariate methods can be applied to CoDa when the principle of working on coordinates is assumed.
- ✓ Non-parametric methods require Aitchison distance or Kullback-Leibler dissimilarity. The log-ratio normal distribution model is the most common model for the parametric methods on the simplex.
- ✓ There are specific techniques and plots to facilitate the interpretation of results provided by the multivariate methods.

Specific references in Chapter 5

- [ABMP00] J. Aitchison, C. Barceló-Vidal, J.A. Martín-Fernández and V. Pawlowsky-Glahn, *Logratio analysis and compositional distances*, Mathematical Geology, 32 (2000) no. 3, 271–275.
- [CMM16] M. Comas-Cufí, J.A. Martín-Fernández and G. Mateu-Figueras, *Logratio methods in mixture models for compositional data sets*, SORT-Statistics and Operations Research Transactions, 40 (2016), no. 2, 349–374
- [KR00] L. Kaufman and P.J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, New York, 1990.
- [LCV18] S. Linares-Mustarós, G. Coenders and M. Vives-Mestres, *Financial performance and distress profiles. From classification according to financial ratios to compositional classification*, Advances in Accounting, 40 (2028), 1–10.
- [MBP98] J.A. Martín-Fernández, C. Barceló-Vidal, and V. Pawlowsky-Glahn, *A critical approach to non-parametric classification of compositional data*. In: Advances in Data Science and Classification (eds A. Rizzi, M. Vichi and H.H. Bock), Springer, Berlin, 1998.
- [MDM15] J.A. Martín-Fernández, J. Daunis-i-Estadella and G. Mateu-Figueras *On the interpretation of differences between groups for compositional data*, SORT-Statistics and Operations Research Transactions, 39 (2015), no. 2, 231–252.
- [PMS12] J. Palarea-Albaladejo, J.A. Martín-Fernández and J.A. Soto, *Dealing with distances and transformations for fuzzy c-means clustering of compositional data*, Journal of Classification, 29 (2012), no. 2, 144–169.
- [VG09] J.A. Villasenor-Alva and E. Gonzalez-Estrada, *A generalization of Shapiro-Wilk's test for multivariate normality*, Communications in Statistics: Theory

and Methods, 38 (2009), no. 11, 1870–1883.

Appendix: Some typical compositional problems

Contents

- A.1 Chemical compositions of Roman-British pottery
- A.2 Arctic lake sediments at different depths
- A.3 Household budget patterns
- A.4 Milk composition study
- A.5 A statistician's time budget
- A.6 The MN blood system
- A.7 Mammal's milk
- A.8 Calc-alkaline and tholeiitic volcanic rocks
- A.9 Concentration of minor elements in carbon ashes
- A.10 Paleocological compositions
- A.11 Pollen composition in fossils
- A.12 Food consumption in European countries
- A.13 Household expenditures
- A.14 Serum proteins
- A.15 Physical activity and body mass index
- A.16 Hotel posts in social media
- A.17 The waste composition in Catalonia
- A.18 Employment distribution in EUROSTAT countries

We present the reader with a series of challenging problems in compositional data analysis, with typical data sets and the questions they pose. These come from a number of different disciplines and will be used to elicit the concepts and principles of compositional data analysis. For further detailed information see [Ait86, Sections 1.1-1.14, p. 1-20].

A.1. Chemical compositions of Romano-British pottery

In archaeology, the compositional analysis of raw materials (clays used to make pottery, lithic materials used to make stone tools, etc.) has become a key tool for examining trade and exchange in ancient economies. Different sources for such materials often have distinct chemical 'signatures' that can be identified in places far from their point of origin. One interpretative challenge is to take an often large, complex array of chemical assays and identify patterns that can be exploited in higher level interpretations.

The `pottery` data set consists of data pertaining to the chemical composition of 45 specimens of Romano-British pottery. The method used to generate these data is atomic absorption spectrophotometry, and readings for nine oxides (Al_2O_3 , Fe_2O_3 , MgO , CaO , Na_2O , K_2O , TiO_2 , MnO , and BaO) are provided. These samples come from five different kiln sites, and one of the issues we want to consider is the degree to which compositional data help distinguish pottery from these various kilns.

 [Back to Index](#)

A.2. Arctic lake sediments at different depths

In sedimentology, specimens of sediments are traditionally separated into three mutually exclusive and exhaustive constituents —sand, silt and clay— and the proportions of these parts by weight are quoted as [sand, silt, clay] compositions. The `arctic_lake` data set records the [sand, silt, clay] compositions of 39 sediment samples at different water depths in an Arctic lake. Again we recognise substantial variability between compositions. Questions of obvious interest here are the following. Is sediment composition dependent on water depth? If so, how can we quantify the extent of the dependence? If we regard sedimentation as a process, do these data provide any information on the nature of the process? Even at this stage of investigation we can see that this may be a question of compositional regression.

 [Back to Index](#)

A.3. Household budget patterns

An important aspect in the study of consumer demand is the analysis of household budget surveys, in which attention often focuses on the expenditures of a sample of households on a number of mutually exclusive and exhaustive commodity groups and their relation to total expenditure, income, type of housing, household composition and so on. In the investigation of such data the pattern or composition of expenditures, the proportions of total expenditure allocated to the commodity groups, can be shown to play a central role in a form of budget share approach to the analysis. Assurances of confidentiality and limitations of space preclude


the publication of individual budgets from an actual survey, but we can present a reduced version of the problem, which retains its key characteristics.

In a sample survey of single persons living alone in rented accommodation, twenty men and twenty women were randomly selected and asked to record over a period of one month their expenditures on the following four mutually exclusive and exhaustive commodity groups:

- **Hous**: Housing, including fuel and light.
- **Food**: Foodstuffs, including alcohol and tobacco.
- **Serv**: Services, including transport and vehicles.
- **Other**: Other goods, including clothing, footwear and durable goods.

The results are recorded in the `householdbudget` data set.

Interesting questions are readily formulated: to what extent does the pattern of the budget share of expenditures for men depend on the total amount spent? Are there any differences between men and women in their expenditure patterns? Are there any commodity groups which are given priority in the allocation of expenditure?

 [Back to Index](#)

A.4. Milk composition study

In an attempt to improve the quality of cow milk, milk from each of thirty cows was assessed by dietary composition before and after a strictly controlled dietary and hormonal regime over a period of eight weeks. Although seasonal variations in milk quality might have been regarded as negligible over this period, it was decided to have a control group of thirty cows kept under the same conditions but on a regular established regime. The sixty cows were of course allocated to control and treatment groups at random. The `milkcows` data set provides the complete set of before and after milk compositions for the sixty cows, showing the protein (`pr`), milk fat (`mf`), carbohydrate (`ch`), calcium (`Ca`), sodium (`Na`) and potassium (`K`) proportions by weight of total dietary content.

The purpose of the experiment was to determine whether the new regime had produced any significant change in the milk composition. It is, therefore, essential to have a clear idea of how change in compositional data is characterised by some meaningful operation. Thus, a key question here is how to formulate hypotheses of change of compositions, and indeed how we may investigate the full lattice of such hypotheses. Meanwhile we note that because of the before and after nature of the data within each experimental unit we have for compositional data the analogue of a paired comparison situation for real measurements where traditionally the differences in pairs of measurements are considered. Thus, we have to find the counterpart of difference for paired compositions.

 [Back to Index](#)

A.5. A statistician's time budget

Time budgets—how a day or a period of work is divided up into different activities—have become a popular source of data in psychology and sociology. To illustrate

such problems we consider six daily activities undertaken by an academic statistician: teaching (T); consultation (C); administration (A); research (R); other wakeful activities (O); and sleep (S).

The `statisticiantimebudget` data set records the daily time (in hours) devoted to each activity, recorded on each of 20 days, selected randomly from working days in alternate weeks so as to avoid possible carry-over effects such as a short-sleep day being compensated by make-up sleep on the succeeding day. The six activities may be divided into two categories: ‘work’ comprising activities T, C, A, and R, and ‘leisure’, comprising activities O and S. Our analysis may then be directed towards the work pattern consisting of the relative times spent in the four work activities, the leisure pattern, and the division of the day into work time and leisure time. Two obvious questions are as follows. To what extent, if any, do the patterns of work and of leisure depend on the times allocated to these major divisions of the day? Is the ratio of sleep to other wakeful activities dependent on the times spent in the various work activities?

 [Back to Index](#)

A.6. The MN blood system

In humans the main blood group systems are the ABO system, the Rh system and the MN system. The *MN blood system* is a system of blood antigens also related to proteins of the red blood cell plasma membrane. The inheritance pattern of the MN blood system is autosomal with codominance, a type of lack of dominance in which the heterozygous manifests a phenotype totally distinct from the homozygous. The possible phenotypical forms are three blood types: type M blood, type N blood and type MN blood. The frequencies of M, N and MN blood types vary widely depending on the ethnic population. However, the Hardy-Weinberg principle states that allele and genotype frequencies in a population will remain constant from generation to generation in the absence of other evolutionary influences. This implies that, in the long run, it holds that

$$(5.3) \quad \frac{x_{MM} \cdot x_{NN}}{(x_{MN})^2} = \frac{1}{4},$$

where x_{MM} and x_{NN} are the genotype relative frequencies of MM and NN homozygotes, respectively, and x_{MN} is the genotype relative frequency of MN heterozygotes. This principle was named after G.H. Hardy and W. Weinberg demonstrated it mathematically.

We will use the `bloodMN` data set to analyse how the relative frequencies of MM, NN and MN blood types are distributed, and to verify the Hardy-Weinberg principle. This data set records the information on the absolute frequencies of M, N, and MN blood types observed in samples coming from different ethnic groups around the world. This data set comes from [Boy50].

 [Back to Index](#)

A.7. Mammal's milk

The `mammalsmilk` data set contains the percentages of five constituents (W: water, P: protein, F: fat, L: lactose, and A: ash) of the milk of 24 mammals. The data are taken from [Har75]. We will analyse whether there are large differences between the compositions of milks and classify them into groups according to the similarity of their constituents.

 [Back to Index](#)

A.8. Calc-alkaline and tholeiitic volcanic rocks

This `petrafm` data set is formed by 100 classified volcanic rock samples from Ontario (Canada). The three parts are:

$$[A : Na_2O + K_2O; F : FeO + 0.8998 \cdot Fe_2O_3; M : MgO].$$

Rocks from the calc-alkaline magma series (25) can be well distinguished from samples from the tholeiitic magma series (75) on an AFM diagram. This data set is a typical example where a discriminant analysis based on the composition could be useful to classify new samples of volcanic rocks.

 [Back to Index](#)

A.9. Concentration of minor elements in carbon ashes

The `montana` data set consists of 229 samples of the concentration (in ppm) of minor elements [Cr, Cu, Hg, U, V] in carbon ashes from the Fort Union formation (Montana, USA), side of the Powder River Basin. The formation is mostly Palaeocene in age, and the coal is the result of deposition in conditions ranging from fluvial to lacustrine. All samples were taken from the same seam at different sites over an area of 430 km by 300 km, which implies that on average, the sampling spacing is 24 km. Using the spatial coordinates of the data, a semivariogram analysis was conducted for each chemical element in order to check for a potential spatial dependence structure in the data (not shown here). No spatial dependence patterns were observed for any component, which allowed us to assume an independence of the chemical samples at different locations.

The aforementioned chemical components actually represent a fully observed subcomposition of a much larger chemical composition. The five elements are not closed to a constant sum. Note that, as the samples are expressed in parts per million and all concentrations were originally measured, a residual element could be defined to fill up the gap to 10^6 . We use this data set to evaluate the algorithms for missing data.

 [Back to Index](#)

A.10. Paleocological compositions

The `foraminiferal` data set (Aitchison, 1986) is a typical example of paleocological data. It contains compositions of 4 different fossils (*Neogloboquadrina atlantica*, *Neogloboquadrina pachyderma*, *Globorotalia obesa*, and *Globigerinoides triloba*) at

30 different depths. Due to the rounded zeros present in the data set we will apply some zero replacement techniques to impute these values in advance. After data preprocessing, the analysis that should be undertaken is the association between the composition and the depth.

 [Back to Index](#)

A.11. Pollen composition in fossils

The `pollen` data set is formed by 30 fossil pollen samples from three different locations (recorded in variable `group`). The samples were analysed and the 3-part composition [`pinus`, `abies`, `quercus`] was measured. The aim was to determine whether the compositions differ significantly from one location to the other.

 [Back to Index](#)

A.12. Food consumption in European countries

The `alimentation` data set contains the percentages of consumption of several types of food in 25 European countries during the 80s. The categories are: red meat (pork, veal, beef), white meat (chicken), eggs, milk, fish, cereals, starch (potatoes), nuts, and fruits and vegetables. The file also contains a categorical variable that shows if the country is from the North or a Southern Mediterranean country. In addition, the countries are classified as Eastern European or as Western European. The aim is to analyse the similarities between countries as regards to their food consumption and to look for associations among the categorical variables.

 [Back to Index](#)

A.13. Household expenditures

From Eurostat (the European Union's statistical information service) the `houseexpend` data set records the composition on proportions of mean consumption expenditure of households expenditures on 12 domestic year costs in 27 states of the European Union. Some values in the data set are rounded zeros. In addition the data set contains the gross domestic product (`GDP05`) and (`GDP14`) in years 2005 and 2014, respectively. An interesting analysis is the potential association between expenditures compositions and GDP. Once a linear regression model is established, predictions can be provided.

 [Back to Index](#)

A.14. Serum proteins

The `serprot` data set records the percentages of the four serum proteins from the blood samples of 30 patients. Fourteen patients have one disease (1) and sixteen are known to have another different disease (2). The 4-compositions are formed by the proteins [`albumin`, `pre-albumin`, `globulin A`, `globulin B`]. The aim is to construct a diagnostic system based on these serum proteins so as to classify six new patients (0).

 [Back to Index](#)

A.15. Physical activity and body mass index

The `BMIPhisActi` data set records the proportion of daily time spent to sleep (`sleep`), sedentary behaviour (`sedent`), light physical activity (`Lpa`), moderate physical activity (`Mpa`) and vigorous physical activity (`Vpa`) measured on a small population of 393 children. Moreover the standardized body mass index (`zBMI`) of each child was also registered.

This data set was used in the example of the article [Dum19] to examine the expected differences in `zBMI` for reallocations of daily time between sleep, physical activity and sedentary behaviour. Because the original data is confidential, the data set `BMIPhisActi` includes simulated data that mimics the main features of the original data.

 [Back to Index](#)

A.16. Hotel posts in social media

The `weibo_hotels` data set aims at comparing the use of Weibo (Facebook equivalent in China) in hospitality e-marketing between small and medium accommodation establishments (private hostels, small hotels) and big and well-established business (such as international hotel chains or large hotels) in China. The 50 latest posts of the Weibo pages of each hotel ($n = 10$) are content-analyzed and coded regarding the count of posts featuring information on a 4-part composition [facilities, food, events, promotions]. Hotels were coded as large “L” or small “S” in the `hotel_size` categorical variable. As this small data set contains zeros we will use it to practice zero replacement methods for count zeros.

 [Back to Index](#)

A.17. The waste composition in Catalonia

The actual population residing in a municipality of Catalonia is composed by the census count and the so-called floating population (tourists, seasonal visitors, hostel students, short-time employees, and the like). Since actual population combines long and short term residents it is convenient to express it as equivalent full-time residents. Floating population may be positive if the municipality is receiving more short term residents than it is sending elsewhere, or negative if the opposite holds (expressed as a percentage above –if positive– or below –if negative– the census count). The `waste` data set includes this information in the variable `floating_population`. Floating population has a large impact on solid waste generation and thus waste can be used to predict floating population which is a hard to estimate demographic variable. This case study was presented in [Coe17].

Tourists and census population do not generate the same volume of waste and have different consumption and recycling patterns (waste composition). The Catalan Statistical Institute (IDESCAT) publishes official floating population data for all municipalities in Catalonia (Spain) above 5000 census habitants. The composition of urban solid waste is classified into $D = 5$ parts:

- x_1 : non recyclable (grey waste container in Catalonia),
- x_2 : glass (bottles and jars of any colour: green waste container),
- x_3 : light containers (plastic packaging, cans and tetra packs: yellow container),
- x_4 : paper and cardboard (blue container), and
- x_5 : biodegradable waste (brown container).

 [Back to Index](#)

A.18. Employment distribution in EUROSTAT countries

According to the three-sector theory, as a country's economy develops, employment shifts from the primary sector (raw material extraction: farming, hunting, fishing, mining) to the secondary sector (industry, energy and construction) and finally to the tertiary sector (services). Thus, a country's employment distribution can be used as a predictor of economic wealth.


The `eurostat_employment_2008` data set contains EUROSTAT data on employment aggregated for both sexes, and all ages distributed by economic activity (classification 1983-2008, NACE Rev. 1.1) in 2008 for the 29 EUROSTAT member countries, thus reflecting reality just before the 2008 financial crisis. Country codes in alphabetical order according to the country name in its own language are: Belgium (BE), Cyprus (CY), Czechia (CZ), Denmark (DK), Deutschland–Germany (DE), Eesti–Estonia (EE), Eire–Ireland (IE), España–Spain (ES), France (FR), Hellas–Greece (GR), Hrvatska–Croatia (HR), Iceland (IS), Italy (IT), Latvia (LV), Lithuania (LT), Luxembourg (LU), Macedonia (MK), Magyarország–Hungary (HU), Malta (MT), Netherlands (NL), Norway (NO), Österreich–Austria (AT), Portugal (PT), Romania (RO), Slovakia (SK), Suomi–Finland (FI), Switzerland (CH), Turkey (TR), United Kingdom (GB).

A key related variable is the logarithm of gross domestic product per person in EUR at current prices (“logGDP”). For the purposes of exploratory data analyses it has also been categorised as a binary variable indicating values higher or lower than the median (“Binary_GDP”). The employment composition ($D = 11$) is:

- Primary_sector (agriculture, hunting, forestry, fishing, mining, quarrying)
- Manufacturing
- Energy (electricity, gas and water supply)
- Construction
- Trade_repair_transport (wholesale and retail trade, repair, transport, storage, communications)
- Hotels_restaurants
- Financial_intermediation
- Real_estate (real estate, renting and business activities)
- Educ_admin_defense_soc_sec (education, public administration, defence, social security)
- Health_social_work

- Other_services (other community, social and personal service activities)

The aim is to construct a linear regression model to predict logGDP.

 [Back to Index](#)

Specific references in Appendix

- [Ait86] J. Aitchison, *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], 1986, 416 p.
- [Boy50] W.C. Boyd, *Genetics and the races of man: an introduction to modern physical anthropology*, Little, Brown & Co, 1950.
- [Coe17] G. Coenders, J.A. Martín-Fernández and B. Ferrer-Rosell, *When relative and absolute information matter: compositional predictor with a total in generalized linear models*, Statistical Modelling **17(6)** (2017), 494–512.
- [Dum19] D. Dumuid, Z. Pedisic, T.E. Stanford, J.A. Martín-Fernández, K. Hron, C. Maher, L.K. Lewis and T.S. Olds, *The Compositional Isotemporal Substitution Model: a Method for Estimating Changes in a Health Outcome for Reallocation of Time between Sleep, Sedentary Behaviour, and Physical Activity*, Statistical Methods in Medical Research **28(3)** (2019), 846–857, DOI:10.1177/0962280217737805.
- [Har75] J.A. Hartigan, *Clustering Algorithms*, John Wiley & Sons, 1975.

