TECHNISCHE UNIVERSITÄT WIEN
Vienna | Austria

DISSERTATION

Methodological Contributions to Compositional Data Analysis

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Doktors der technischen Wissenschaften unter der Leitung von

Ao. Univ.-Prof. Dipl.-Ing. Dr.techn. Peter Filzmoser, Institut für Stochastik und Wirtschaftsmathematik (E105)

eingereicht an der Technischen Universität Wien an der Fakultät für Mathematik and Geoinformation

von

Mgr. Petra Kynčlová
Matrikelnummer 1228699

Diese Dissertation haben begutachtet:

Ao. Univ.-Prof. Dipl.-Ing. Dr.techn. Peter Filzmoser	Doc. RNDr. Karel Hron Ph.D.

Wien, 10. Dezember 2015

Mgr. Petra Kynčlová

Erklärung zur Verfassung der Arbeit

Mgr. Petra Kynčlová
Wiedner Gürtel 28/7
1040 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 10. Dezember 2015

Mgr. Petra Kynčlová

Acknowledgements

Initially, I would like to thank two most important persons who made my studies even possible. In the first place, my supervisor Prof. Peter Filzmoser for continuous and endless support and motivation, and for sharing his academic experience. Secondly, to Doc. Karel Hron for giving me an opportunity to start my studies in Vienna and also for his ideas and constant contribution. I am beyond grateful to all my collegues, I had a chance to meet and work with during my studies, especially to those I shared the office with. Further, I would like to express my appreciation for their help and understanding. I hope to continue cooperation with all of them in the future.

Last but not least, I want to thank my family and friends for always being there for me and understanding all my mood swings caused by the stress and the lack of time.

Abstract

The world is full of a large amount of data that can be analyzed to understand the behavior, structure, inner patterns or mutual relations between various variables, explaining how the things around us principally work. Some of the data could, however, be specific and require a special treatment when applying standard statistical analysis. This is the case with so called compositional data.

Compositional data represent a special type of multivariate data carrying exclusively relative information, and they can frequently be found in various experimental fields. The main information of interest is then given by the respective ratios between the compositional parts, and the data are often expressed as proportions or percentages, i.e. as data with a constant sum constraint. All this makes the corresponding statistical analysis difficult, because compositional data do not follow the standard Euclidean geometry, which is required for applying the usual statistical procedures. Despite the fact that a lot of progress has been made with defining a new geometry followed by the log-ratio methodology, there are still several open issues in the field of compositional data analysis.

This thesis is exclusively dedicated to compositional data analysis and its specific methods and tools developed and investigated based on the current needs of practitioners. The aim of the thesis is to present a comprehensive concept of the statistical analysis for compositional data and to introduce new methodological contributions in the field of time series analysis, correlation analysis, and an extension of compositional biplots. All new concepts are demonstrated on data examples to see the practical impact of using the appropriate geometry and methods for compositional data.

Kurzfassung

In der heutigen Zeit beobachten wir mehr und mehr Daten, die analysiert werden können, um das Verhalten, die Struktur, innere Muster, oder wechselseitige Beziehungen in einer Vielzahl von Variablen zu verstehen, und eine Erklärung zu erhalten, wie die Dinge um uns generell zusammenhängen. Einige dieser Daten könnten jedoch von spezieller Natur sein und eine besondere Vorbehandlung benötigen, bevor Standardmethoden der Statistik angewandt werden. Dies ist der Fall bei sogenannten Kompositionsdaten.

Kompositionsdaten repräsentieren einen Spezialfall von multivariaten Daten, die ausschließlich relative Information beinhalten, und sie können häufig in einer Vielzahl von Anwendungen gefunden werden. Die interessierende Information liegt in den entsprechenden Verhältnissen der kompositionellen Variablen, und die Daten sind oft ausgedrückt in Anteilen oder Prozenten, d.h. als Daten mit konstanter Summe. Das macht die statistische Analyse aufwändiger, weil Kompositionsdaten nicht der üblichen euklidischen Geometrie folgen, die den meisten statistischen Prozeduren zugrunde liegt. Obwohl viele Fortschritte zur Entwicklung einer neuen Geometrie gemacht wurden, gefolgt von der sogenannten *log-ratio* Methodik, gibt es noch immer viele offene Themen im Bereich der Analyse von Kompositionsdaten.

Diese Dissertation ist ausschließlich der Analyse von Kompositionsdaten gewidmet, mit ihren spezifischen Methoden und Werkzeugen, entwickelt und untersucht anhand der derzeitigen Bedürfnisse der Praxis. Das Ziel der Dissertation ist es, ein umfangreiches Konzept der statistischen Analyse von Kompositionsdaten zu präsentieren, und neue methodische Beiträge in den Bereichen Zeitreihenanalyse und Korrelationsanalyse zu liefern, sowie eine Erweiterung von kompositionellen Biplots. Alle neuen Konzepte werden auf Datenbeispiele angewendet, um den praktischen Nutzen der geeigneten Geometrie und der Methoden für Kompositionsdaten aufzuzeigen.

Contents

CHAPTER 1

Introduction

1.1 Introduction and history

Compositional data (CoDa) can be found in all experimental fields. They represent a special type of multivariate data that provide relative information between their components, i.e. they describe the parts of some given whole. Considering the fact that only the ratios are informative, compositional data occur frequently in geochemistry and biosciences, but also in fields such as economics or political sciences. The scientists are then more interested in the data structure provided by the ratios than in the absolute mass of compounds depending on the sample size. These observations are thus called compositional data, or compositions, and they mostly appear as proportions, percentages or frequencies.

A simple example can be given to understand the basis of compositional data and to distinguish the different interest of the data analysis and methods. Let us consider the famous cocktail called Margarita. One cocktail of Margarita consists of mixing 3.5 cl of tequila, 2 cl of Cointreau (triple sec) and 1.5 cl of freshly squeezed lime juice. In mathematical terms, we obtain a composition $x = (3.5, 2, 1.5)$ with the sum $\kappa = 7$. To prepare two cocktails of Margarita, we can simply double the amounts of ingredients to $x^* = (7, 4, 3)$. The total sum has changed to $\kappa^* = 14$, but the ratios between the compounds are unchanged. If we change the structure of ingredients, the resulting cocktail would not be the famous Margarita, but we would shake another drink. Such easy example shows why the relative information is important when dealing with compositional data. The existence of this type of data was discovered by the failure of standard statistical methods applied on them.

The starting point for the statistical analysis of compositional data can be dated back to a paper by Pearson (1897), where the problem of so called *spurious correlation* and its interpretation was pointed out. The paper describes the difficulties obtained by applying

standard correlation analysis to data with constant sum constraint. The criticism of the application of standard multivariate analysis to compositional data continued by Chayes (1960) in the interpretation of the product-moment correlation between components of a geochemical composition, with negative bias as the distorting factor from the viewpoint of any sensible interpretation. Nevertheless, the findings still did not lead to start building a new appropriate methodology for working with compositional data. The main step towards a proper compositional approach was done by Aitchison (1986), who decided to define compositions in terms of ratios between parts and stated that the information carried by compositional data is relative. It was also Aitchison (1986), who described the specific principles for compositional data and came up with an idea that a log-ratio (logarithm of a ratio) transformation provides a one-to-one mapping onto a real space and started to build a methodology based on a variety of log-ratio coordinates. The main advantage of using log-ratio coordinates enables to use standard unconstrained multivariate statistics applied to transformed data with a possibility of coming back to the simplex.

In the following years, it turned out that compositional data are not restricted entirely to observations with a constant sum constraint, proportions or percentages, but the concept covers all observations carrying relative information with a possibility of being described with any prescribed sum constraint without altering the ratios between the parts (Pawlowsky-Glahn and Egozcue, 2001). Moreover, the Aitchison geometry with the Euclidean vector space structure was introduced (Pawlowsky-Glahn and Egozcue, 2001) in order to express compositions in proper log-ratio coordinates followed by applying standard statistical methods. Recently, compositional data analysis has consisted in representing the data in a log-ratio type of coordinates, applying the standard statistical tools on the coordinates treated as real random variables followed by the final interpretation. The interpretation of the resulting models can then carried out either in coordinates, or in terms of the original units, called *the principle of working in coordinates* (Mateu-Figueras et al., 2011).

In recent years, various statistical methods have been developed for or adjusted to the field of compositional data with respect to their special geometric properties. This dissertation consists of a collection of papers describing exclusively tools and methods for compositional data.

1.2 Geometrical aspects of compositional data

The geometry of compositional data is a very important issue for enhancing statistical methods when working with this type of data. Compositions are naturally represented as *closed data*, where the constant sum constraint influences their sample space. Nevertheless, data with some constraint do not follow the well-known Euclidean geometry traditionally used for any kind of statistical analysis. The natural sample space for compositional data is the *simplex*. For this reason, the proper geometry on the simplex has to be established.

1.2.1 Principles

Compositional data are characterized as multivariate observations containing relative contributions of parts on a whole. In mathematical terms, a D-part composition is a vector $\mathbf{x} = (x_1, \ldots, x_D)$ with the simplex as the sample space

$$\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, \ldots, x_D) : x_i > 0 \quad (i = 1, \ldots, D), \sum_{i=1}^{D} x_i = \kappa \right\}, \tag{1.1}$$

where κ is a given constant. Frequently in practice, the constant κ is chosen to be one, as a unit sum constraint. As it was described in the introduction, an arbitrary prescribed sum constraint does not change the information contained in the ratios between the parts.

The compositions are usually represented as vectors of two or more components, thus omitting one component should not change the resulting values. For this reason, compositional data analysis is meaningful only when three important conditions are fulfilled: scale invariance, subcompositional coherence and permutation invariance (Aitchison, 1986).

Scale invariance

This principle assumes that vectors with proportional components represent the same composition. The vectors of proportional positive components form an equivalence class. The selection of representatives of the equivalent class can be achieved by applying the closure operation. The closure for a composition $\mathbf{x} = (x_1, \ldots, x_D)$ is defined as

$$\mathcal{C}(\mathbf{x}) = \left(\frac{\kappa x_1}{\sum_{i=1}^{D} x_i}, \cdots, \frac{\kappa x_D}{\sum_{i=1}^{D} x_i} \right). \tag{1.2}$$

The set of vectors with D positive components summing up to the constant κ form the D-part simplex, $\mathcal{S}^D$. The size and the total weight of the whole is irrelevant, when considering the compositional data analysis. By applying the closure operation, the data from samples with different sizes can be compared. The statistical analysis is then scaling invariant, when it provides the same answer independent from the value κ.

Considering the former example with Margarita and applying a closure operation on the cocktail composition with $\kappa = 100$, percentages are obtained and the total sum of ingredients has been filtered out. The new closed composition is then $x^* = (50, 28.57, 21.43)$ containing exactly the same information. Moreover, the percentages tell us how to prepare an arbitrary amount of drinks.

Subcompositional coherence

A subcomposition is a subset of the initial parts of the original composition that is formed by reclosing the vector of the chosen components. Explicitly, a subcomposition $\mathbf{x}_S$ with

$s < D$ parts from a composition $\mathbf{x}$ is obtained by applying a closure operation to the subvector $(x_{i_1}, x_{i_2}, \ldots, x_{i_s})$ of $\mathbf{x}$, where the set of subscripts $S = \{i_1, \ldots, i_s\}$ indicates which parts are selected in the subcomposition (Pawlowsky-Glahn and Buccianti, 2011). It is obvious that analyses concerning a subset of parts must not depend on other non-involved parts. Moreover, if we add a new random component and work with the resulting $(D + 1)$-part composition, the result should not change.

There are two additional principles related to subcompositional coherence. First, the principle of scale invariance should hold for any of the possible subcompositions, keeping the ratios of parts. The second one is called *subcompositional dominance* which states if the distance or divergence is used to compare compositions, this distance or divergence should be greater or equal to that obtained comparing the corresponding subcompositions.

Permutation invariance

Last but not least, the conclusions and results of a compositional analysis should not depend on the order of the parts given in the data set. Based on the cocktail example, this means that the order of the added ingredients should not affect the taste of the resulting drink.

1.2.2 Aitchison geometry

Working with compositional data requires establishing a proper geometry on the simplex $\mathcal{S}^D$ to fulfill the conditions from function 1.2.1. The basic operations used on the vector space structure of the simplex are based on the closure operation and they represent a parallel operation to addition and multiplication by a constant in the real space. Consider the compositions $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$. The perturbation of $\mathbf{x}$ with $\mathbf{y}$ is defined as the composition

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, x_2 y_2, \ldots, x_D y_D), \tag{1.3}$$

and powering of $\mathbf{x}$ by a real number α as the composition

$$\alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha). \tag{1.4}$$

Then it can be shown that $\mathbf{x} \oplus \mathbf{n} = \mathbf{x}$ for $\mathbf{n} = \mathcal{C}(1, 1, \ldots, 1)$, thus the composition with equal parts is the neutral element of perturbation. Using the opposite element of $\mathbf{y}$, $\mathbf{y}^{-1} = \mathcal{C}(y_1^{-1}, y_2^{-1}, \ldots, y_D^{-1})$, the inverse perturbation $\ominus$ can be defined as

$$\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus \mathbf{y}^{-1}. \tag{1.5}$$

The simplex with operations perturbation and powering, $(\mathcal{S}^D, \oplus, \odot)$, represents a vector space. Furthermore, the following inner product with its associated norm and distance

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}, \tag{1.6}$$

$$\|\mathbf{x}\|_a^2 = \langle \mathbf{x}, \mathbf{x} \rangle_a, \tag{1.7}$$

$$d_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_a, \tag{1.8}$$

can be used to obtain a finite $(D-1)$-dimensional Hilbert space structure. It is important to point out that this is Euclidean vector space structure on the simplex (Billheimer et al., 2001; Pawlowsky-Glahn and Egozcue, 2001). Then the properties of $(\mathcal{S}^D, \oplus, \odot)$ refer to the so called Aitchison geometry on the simplex, as well as to the Aitchison distance, norm and inner product, denoted by the subscript a.

1.2.3 Log-ratio methodology

We already know that the simplex $\mathcal{S}^D$ is a $(D-1)$-dimensional subset of D-dimensional real space. Therefore, compositions from $\mathcal{S}^D$ are usually expressed in terms of a canonical basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_D\}$ of $\mathbb{R}^D$. Then any composition $\mathbf{x} \in \mathcal{S}^D$ can be written as

$$\mathbf{x} = x_1 \cdot (1, 0, \ldots, 0) + x_2 \cdot (0, 1, \ldots, 0) + \cdots + x_D \cdot (0, \ldots, 0, 1). \tag{1.9}$$

Nevertheless, the set of vectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_D\}$ is neither a basis nor a generating system on $\mathcal{S}^D$ with respect to the given vector space structure. Despite that Equation 1.9 represents a convex linear combination on $\mathbb{R}^D$, it is not a linear combination on $\mathcal{S}^D$, because the sum and the product does not stand for a closed operation on the simplex to be able to capture the property of scale invariance.

Building the basis within the Aitchison geometry requires finding a generating system of $\mathcal{S}^D$. Such a generating system can be obtained as $\{\mathbf{w}_1, \ldots, \mathbf{w}_D\}$ with

$$\mathbf{w}_i = \mathcal{C}(1, 1, \ldots, e, \ldots, 1, 1) \qquad \text{for} \qquad i = 1, \ldots, D, \tag{1.10}$$

where e is the i-th component. Then, any composition $\mathbf{x} \in \mathcal{S}^D$ can be expressed as

$$\mathbf{x} = (\ln x_1 \odot \mathbf{w}_1) \oplus (\ln x_2 \odot \mathbf{w}_2) \oplus \cdots \oplus (\ln x_D \odot \mathbf{w}_D). \tag{1.11}$$

Since the closure operation included in perturbation and powering suggests the scale invariance, adding an arbitrary constant does not change $\mathbf{x}$. Thus the following equivalent expression can be considered

$$\mathbf{x} = \left(\ln \frac{x_1}{g(\mathbf{x})} \odot \mathbf{w}_1\right) \oplus \left(\ln \frac{x_2}{g(\mathbf{x})} \odot \mathbf{w}_2\right) \oplus \cdots \oplus \left(\ln \frac{x_D}{g(\mathbf{x})} \odot \mathbf{w}_D\right), \tag{1.12}$$

where $g(\mathbf{x})$ stands for the geometric mean of all components. The coefficients coming from Equation (1.12) represent the centered log-ratio (clr) transformation introduced by Aitchison (1986),

$$\mathbf{y} = (y_1, \ldots, y_D) = \mathrm{clr}(\mathbf{x}) = \left(\ln \frac{x_1}{g(\mathbf{x})}, \ln \frac{x_2}{g(\mathbf{x})}, \ldots, \ln \frac{x_D}{g(\mathbf{x})}\right). \tag{1.13}$$

The clr coordinates represent a one-to-one mapping from $\mathcal{S}^D$ to $\mathbb{R}^D$, so it is possible to use the original variable names for the interpretation of statistical results based on clr

transformed data. However, the resulting data are collinear due to the new constraint $y_1 + \cdots + y_D = 0$, and the corresponding covariance matrix is singular.

Considering the same generating system $\{\mathbf{w}_1, \ldots, \mathbf{w}_D\}$, another basis can be obtained by taking any $(D-1)$ vectors, for instance $\{\mathbf{w}_1, \ldots, \mathbf{w}_{D-1}\}$. Then any compositional vector $\mathbf{x} \in \mathcal{S}^D$ can be expressed as

$$\mathbf{x} = \left(\ln \frac{x_1}{x_D} \odot \mathbf{w}_1 \right) \oplus \left(\ln \frac{x_2}{x_D} \odot \mathbf{w}_2 \right) \oplus \cdots \oplus \left(\ln \frac{x_{D-1}}{x_D} \odot \mathbf{w}_D \right), \tag{1.14}$$

where the coefficients belong to the additive log-ratio (alr) transformation defined by Aitchison (1986)

$$\mathrm{alr}(\mathbf{x}) = \left(\ln \frac{x_1}{x_D}, \ln \frac{x_2}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D} \right). \tag{1.15}$$

The basis $\{\mathbf{w}_1, \ldots, \mathbf{w}_{D-1}\}$ is not orthogonal, which can be shown by computing the inner product and the norm (Egozcue and Pawlowsky-Glahn, 2005). The alr transformation is then not symmetrical in the components. Moreover, the essential problem with alr coordinates is the non-isometric character of this transformation.

The Euclidean vector space structure of the simplex assures the existence of an orthonormal basis with respect to the inner product by applying the usual Gram-Schmidt orthonormalization to any given basis. This assumption formed the background for introducing the isometric log-ratio (ilr) transformation by Egozcue et al. (2003). This transformation results in orthonormal coordinates $\mathbf{z} = (z_1, \ldots, z_{D-1})$ with respect to the Aitchison geometry, and it also leads to an orthonormal basis of the hyperplane $\mathcal{H} : y_1 + \cdots + y_D = 0$, formed by the clr transformation. Moreover, there exists a linear relationship between the clr variables and the orthonormal coordinates,

$$\mathbf{y} = \mathcal{V}\mathbf{z}. \tag{1.16}$$

The columns of the $D \times (D-1)$ matrix $\mathcal{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_{D-1})$ are orthonormal basis vectors on the hyperplane $\mathcal{H}$,

$$\mathbf{v}_{D-i} = \sqrt{\frac{i}{i+1}} \left(0, \ldots, 0, 1, -\frac{1}{i}, \ldots, -\frac{1}{i} \right)^\top, \qquad i = 1, \ldots, D-1, \tag{1.17}$$

resulting in the ilr coordinates $\mathbf{z}$.

There are infinitely many possibilities to construct an orthonormal basis. A special choice of orthonormal coordinates that allows to interpret them in terms of the contributions of the single compositional parts is as follows (Filzmoser et al., 2012). Consider the compositions $(x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)$, which are re-arranged such that the l-th part is at the first position. We will use the notation $\mathbf{x}^{(l)} = (x_1^{(l)}, x_2^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})$, where each part with index $l = 1, \ldots, D$ could be placed on the first position, and the sequence of the other parts remains unchanged. The ilr transformation of $\mathbf{x}^{(l)}$ results in

$\mathbf{z}^{(l)} = (z_1^{(l)}, \ldots, z_{D-1}^{(l)})$, where the components are defined by

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}}, \qquad i = 1, \ldots, D-1. \tag{1.18}$$

Then, the first ilr variable $z_1^{(l)}$ explains all the relative information (log-ratios) about the original compositional part x_l. The coordinates $z_2^{(l)}, \ldots, z_{D-1}^{(l)}$ explain the remaining log-ratios in the composition (Fišerová and Hron, 2011). Note that the only important position is that of $x_1^{(l)}$, because it can be fully explained by $z_1^{(l)}$. The other parts can be chosen arbitrarily, because different ilr coordinates are orthogonal rotations of each other (Egozcue et al., 2003). Note that the relation

$$y_l = \sqrt{\frac{D-1}{D}} z_1^{(l)}, \qquad l = 1, \ldots, D, \tag{1.19}$$

confirms our preliminary requirement on interpretability of the resulting coordinates, for $D \to \infty$ both variables approach the same values. On the other hand, both y_l and $z_1^{(l)}$ thus share also interpretational doubts, mentioned by defining the clr variables.

The advantage of obtaining an interpretation for each compositional part is redeemed by the necessity of constructing D coordinate systems, where always just one variable is of primary interest (at the first position). It is obvious that always the first coordinate $z_1^{(l)}$ in each given system corresponds to the clr coordinate y_l, for $l = 1, \ldots, D$, differing by the constant $\sqrt{\frac{D}{D-1}}$.

Another strategy, how to obtain an orthonormal basis of the simplex and their respective ilr coordinates with useful interpretation properties, is called *sequential binary partitioning* (Egozcue and Pawlowsky-Glahn, 2005). It enables to build a special orthonormal basis with coordinates called *balances*. This procedure is explained in detail in Chapter 4.

1.2.4 Center and variability

As the compositional data follow the Aitchison geometry on the simplex $\mathcal{S}^D$, the standard descriptive statistics is not very informative in this case. Central tendency and dispersion, the alternatives of the arithmetic mean and variance or standard deviation should be properly defined, since they are described in the framework of the Euclidean geometry in real space. For this reason, the concept of the center (Aitchison, 1997), the variation matrix and the total variance (Aitchison, 1986) was introduced.

Following Pawlowsky-Glahn and Egozcue (2001), the center of a random composition $\mathbf{x} \in \mathcal{S}^D$ is defined as $\mathrm{cen}(\mathbf{x})$ that minimizes the expression $\mathrm{E}[d_a^2(\mathbf{x}, \mathrm{cen}(\mathbf{x}))]$, thus

$$\mathrm{cen}(\mathbf{x}) = \mathcal{C}(\exp(\mathrm{E}(\ln(\mathbf{x})))). \tag{1.20}$$

The definition of the center of $\mathbf{x}$ can then be expressed as a closed geometric mean (Aitchison, 1997) and represents a mean of the simplex as a sample space.

A measure of global dispersion of a compositional sample is called the total variance that can be defined as

$$\mathrm{TotVar}(\mathbf{x}) = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\mathrm{var}\Big(\ln\frac{x_i}{x_j}\Big), \tag{1.21}$$

where var denotes the variance. Furthermore, it is also possible to estimate the center and the total variance using the ilr components and their properties corresponding to the estimators of the mean and the variance-covariance in real sample spaces (Pawlowsky-Glahn and Egozcue, 2001).

The dispersion of a compositional sample can also be described by the variation matrix (Aitchison, 1986). The variation matrix of a D-part composition is a symmetric matrix of order D, defined as

$$\mathbf{T} = [t_{ij}] = \left[\mathrm{var}\Big(\ln\frac{x_i}{x_j}\Big)\right], \quad i,j = 1,\ldots,D, \tag{1.22}$$

with zeros on the diagonal. When the elements of $\mathbf{T}$ are close to zero, the ratio x_i/x_j is nearly constant, i.e. the two parts x_i and x_j are almost proportional. On the contrary, high variability of the log-ratio indicates very different ratios of two parts among all the observations.

The log-ratios in (1.22) can also be rescaled according to (1.18) so that they correspond, up to orientation, to the normed coordinate of the two-part composition (x_i, x_j). The resulting normalized variation matrix (Pawlowsky-Glahn et al., 2015) is defined as

$$\mathbf{T}^* = [t^*_{ij}] = \left[\mathrm{var}\Big(\frac{1}{\sqrt{2}}\ln\frac{x_i}{x_j}\Big)\right], \quad i,j = 1,\ldots,D, \tag{1.23}$$

where t^*_{ij} stands for the usual (sample) variance of the normalized log-ratio of parts i and j (balance). Subsequently, the relation between $\mathbf{T}$ and $\mathbf{T}^*$ is given as

$$\mathbf{T} = \frac{1}{2}\mathbf{T}^*. \tag{1.24}$$

The measure of variability could be normalized to the range (0,1] as

$$\tau_{ij} = \exp(-\mathrm{var}(t^*_{ij})) \tag{1.25}$$

for $1 \le i,j \le D, i \ne j$ (Buccianti and Pawlowsky-Glahn, 2005; Filzmoser et al., 2010). High variability of the log-ratio then tends to a result approaching zero and, conversely, small variability is reflected by values of τ_{ij} close to one with the limiting case of perfect proportionality.

1.3 Compositional data analysis in practice

Compositional data occur frequently in practice, because most measured features around us are expressed in terms of contributions to some given whole. To demonstrate different properties of compositional data, a practical example is considered in the following.

1.3.1 Example: Employment data

Consider a data set including the number of employed people in the countries of the European Union; the data come from EUROSTAT (Eurostat, the statistical office of the European Union, 2013) and they play an important role for an application also presented in Chapter 3. The six-part composition describes the number of employed people in different fields of economic activity: agriculture, forestry and fishing (*agriculture*); industry and construction (*industry*); financial and insurance activities (*finance*); real estate activities (*real estate*); public administration, defense, education, human health and social work activities (*public*); arts, entertainment, recreation and other service activities (*arts*). The data are shown in Table 1.1.

	agri	finance	real estate	public	arts	industry
BE	53.00	157.00	26.10	1474.10	217.20	1340.70
BG	189.00	52.70	7.80	563.60	92.30	1061.90
CZ	149.20	136.70	45.70	957.80	194.50	1513.30
DK	69.60	82.10	26.80	896.90	140.50	767.00
DE	620.30	1313.40	277.30	10292.90	1945.60	11532.70
EE	29.10	10.50	10.70	139.40	26.60	209.20
IE	583.00	85.80	91.00	10.30	487.10	99.30
GR	490.00	112.70	5.90	857.30	177.20	1336.90
ES	753.20	424.80	96.20	3860.50	1380.30	6121.60
FR	773.80	857.30	318.20	7978.20	1710.10	7595.10
HR	198.10	35.30	2.70	276.20	52.80	460.90
IT	849.10	642.90	141.80	4641.60	1750.30	7470.10
CY	11.30	23.30	2.20	70.00	42.20	158.30
LV	73.30	24.20	23.00	202.30	41.40	293.00
LT	112.20	18.20	13.30	293.20	52.70	438.60
LU	3.10	29.50	1.60	72.50	25.90	51.40
HU	201.10	93.90	23.00	895.80	156.80	1217.20
MT	1.80	7.70	0.80	47.20	8.20	60.20
NL	208.30	220.80	64.80	2464.60	353.90	2323.60
AT	204.60	148.40	37.20	933.50	196.80	1482.40
PL	1960.20	393.70	146.90	3121.10	451.80	4768.10
PT	486.00	97.80	23.60	1038.80	281.40	1484.80
RO	2682.30	140.10	15.80	1228.40	237.80	2519.80
SI	77.10	31.30	2.70	195.00	31.10	257.80
SK	75.40	51.90	16.00	503.50	69.30	784.70
FI	102.80	51.10	23.30	696.70	142.30	705.40
SE	95.40	95.40	66.90	1508.60	231.70	1254.10
UK	347.20	1201.50	332.20	8741.00	1601.70	9098.20

Table 1.1: Number of employed people (in thousands) in the member states of the European Union in 2013.

The data represent the structure of employment based on economic activity. The data set contains the absolute values which differ a lot among the different countries. This fact is given by the size of total employed population in different states. However, we are interested in the relative information contained in the ratios between variables. For instance, we can look at two particular countries like Austria and France. It is obvious that the number of people working in France is much higher than in Austria due to the number of inhabitants. But looking exclusively on the ratios, they yield very similar results of the relative behavior.

1.3.2 Graphical representation

For a graphical representation of three-part compositions ($D = 3$), the ternary diagram is widely used, as it is built as a two-dimensional plot. This graphical tool is widely known for example in geology or petrology. The ternary diagram is an equitorial triangle, where its vertices represent three parts of the composition. Each vertex is associated with one part. The interpretation of data points in the ternary diagram is pretty simple. When the given observation lies close to the vertex, it means that the proportion of the corresponding variable represented by the vertex is high. If the point is situated directly on the link between vertices, the proportion of the variable represented by the opposite vertex is zero. A second possible interpretation is that all points on a straight line through one of the corners have equal relative portions of the two remaining components. The data point lying in the center of the ternary diagram, called *baricenter*, then has equal proportions on all observed variables.

Figure 1.1 shows the ternary diagram for the employment data example. A three-part subcomposition, with parts *agri*, *finance* and *arts*, is chosen to be depicted in the ternary diagram and in the scatter plot of coordinates.

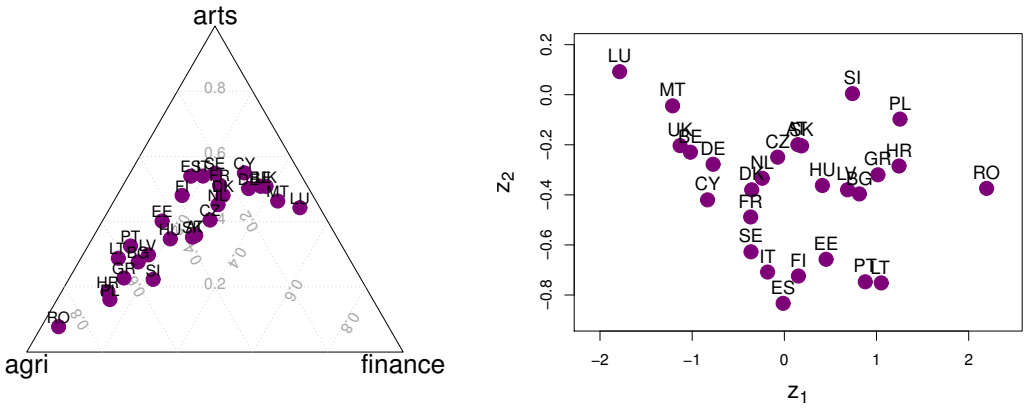


Figure 1.1: Ternary diagram and ilr coordinates for employment data

Comparing the ternary diagram with the scatter plot of ilr coordinates nicely indicates the different behavior of data points in the graphs. The ternary diagram clearly shows the proportional structure of the compositional parts. For instance, Romania yields a high proportion of people employed in agriculture, while variables *arts* and finance are negligible. On the other hand, Luxembourg shows low proportion of agricultural employment against financial activities and activities corresponding to the variable *arts*.

In the case of the scatter plot, following ilr coordinates based on Equation (1.18) were considered,

$$z_1 = \sqrt{\frac{2}{3}} \ln \frac{x_1}{\sqrt{x_2 x_3}}, \tag{1.26}$$

$$z_2 = \sqrt{\frac{1}{2}} \ln \frac{x_2}{x_3}, \tag{1.27}$$

in order to describe all relative information about the variable *agri* by the coordinate z_1. Romania is then located on the right, showing a high value of the coordinate z_1, followed by Croatia and Poland. On the other side with small values of z_1, it is possible to find Luxembourg and Malta. However, the ilr coordinates can be chosen in a different way with respect to another variable. The use of coordinates enables us to apply standard statistical methods, because we transformed the compositions into the Euclidean vector space.

1.3.3 Absolute vs. relative information

The main idea of working with compositional data consists of realizing that the information of interest contained in the data is relative. It is not a coincidence that the problems of standard statistical analyses applied on compositional data was pointed out by geochemists. In geochemistry, all elements present in a sample are rarely analyzed jointly and it is very common to work only with subcompositions.

Figure 1.2 shows how important the aspect of relative versus absolute information is for the analysis. The data coming from our employment example are again investigated. To see the structure, the data were recalculated into percentages with respect to the total employment in the given country. For illustration, the variable *industry* is taken into account and plotted with focus on the absolute information (left) and the relative information given by the clr transformation of the values (right). Comparing these two European maps, both approaches show quite different patterns across the countries. For instance, looking at the absolute values (percentages), the countries with the highest proportion of industrial employment are Austria, Bulgaria, Czech Republic and Slovakia. Applying the clr transformation, the structure of employment in industry has changed a lot and the relative contribution of industry is more significant in countries like Croatia and Greece. The completely different behavior is visible also for the Baltic countries, and for example for Romania. Romania reveals quite a small percentage of industrial employment, but from a relative point of view the performance is much higher. The opposite pattern is then shown by Austria, where the contribution to employment in the

industrial sector is relatively smaller then it is observed by the map covering absolute information. The difference between these two maps is given by the dominance of the agricultural economic activity (*agri*) over the sample, that is clearly visible in Figure 1.3. After using the clr transformation, the data are no longer dominated by the compositional part *agri* and the true data structure with the relations between the variables are shown.

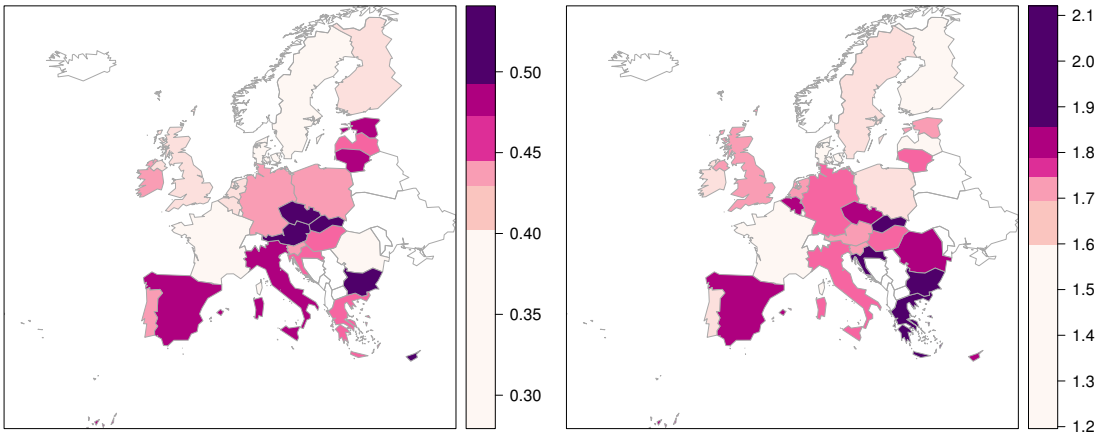


Figure 1.2: The variable *industry* of the employment data: absolute information expressed by percentages (left) and relative information expressed by clr coordinates (right). The color scale is according to regular quantiles of the distribution.

1.4 Principal component analysis

Principal component analysis (PCA) represents a statistical method to find a set of orthogonal directions in a data set, which maximize the variance of the data projected on them (Jolliffe, 2013). These directions provide an uncorrelated representation of the original data structure and they are called principal components. The number of principal components is less than or equal to the number of original variables. Then the first principal component has maximum variance among all linear combinations and thus it explains as much variation in the data as possible. Accordingly, the second principal component is the linear combination describing a maximum of the remaining variation with the constraint that the correlation between the first and second principal component is zero, and so on. The resulting vectors form an uncorrelated orthogonal basis set. It is apparent that PCA can not only reveal patterns in the data structure, but it is also a useful tool for dimension reduction. Principal component analysis belongs to the most important methods of multivariate statistics and it represents also the essential method for various multivariate procedures. For this reason, PCA for compositional data is of primary interest and should be investigated.

The compositional approach to principal component analysis dates back to Aitchison

(1983, 1986) in the sense of finding a useful transformation. The main idea is to transform the compositional data to the real Euclidean space $\mathbb{R}^D$, where standard PCA can be applied. A useful possibility seems to use the clr coordinates, because they represent a one-to-one mapping from the simplex $\mathcal{S}^D$ to the real space $\mathbb{R}^D$. This transformation treats all components symmetrically by dividing by the geometric mean, and therefore it allows to use the original variable names for the interpretation. However, these coordinates result in singularity, because $\sum_{i=1}^{D} y_i = 0$. Then PCA applied on the clr-transformed data will result in $D-1$ principal components. This problem can be avoided by using the ilr coordinates, which can be derived from the clr transformation by using their mutual relationship (1.16).

For the purpose of the principal component analysis, the matrix expression of the clr coefficients is given as

$$\mathbf{y} = \mathbf{F}\log(\mathbf{x}), \tag{1.28}$$

where

$$\mathbf{F} = \begin{pmatrix} 1 & & -1 \\ & \ddots & \vdots \\ & 1 & -1 \end{pmatrix}, \tag{1.29}$$

where the undisplayed elements of the matrix are zero.

Now, we can assume an $n \times D$ data matrix $\mathbf{X}$ with n compositions $\mathbf{x}_i$, $i = 1, \ldots, n$, in the rows and apply (1.28) to each row to obtain a matrix of clr coefficients

$$\mathbf{Y} = \log(\mathbf{X})\mathbf{F}^\top, \tag{1.30}$$

and transform it into the $n \times (D-1)$ matrix $\mathbf{Z}$ of corresponding ilr coordinates

$$\mathbf{Z} = \mathbf{Y}\mathbf{V}, \tag{1.31}$$

where $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_{D-1})$ (with $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_{D-1}$), that is the $D \times (D-1)$ matrix containing orthonormal basis vectors from (1.17) on the hyperplane $\mathcal{H}$. Denote $T(\mathbf{Z})$ a location estimator and $C(\mathbf{Z})$ a covariance estimator for the ilr transformed data. After singular value decomposition of $C(\mathbf{Z}) = \mathbf{G}_\mathbf{z}\mathbf{L}_\mathbf{z}\mathbf{G}_\mathbf{z}^\top$ with the diagonal matrix $\mathbf{L}_\mathbf{z}$ of eigenvalues and the matrix $\mathbf{G}_\mathbf{z}$ of eigenvectors of $C(\mathbf{Z})$, the PCA transformation can be defined as

$$\mathbf{Z}^* = [\mathbf{Z} - \mathbf{1}T(\mathbf{Z})^\top]\mathbf{G}_\mathbf{z}, \tag{1.32}$$

with $\mathbf{1}$ standing for a vector of n ones (Filzmoser et al., 2009a). If the original data matrix $\mathbf{X}$ has full rank D, then $\mathbf{Z}$ will have full rank $D-1$ with scores $\mathbf{Z}^*$ and loadings $\mathbf{G}_\mathbf{z}$, represented by the columns of the respective matrices. The main problem arises from the interpretation, since scores and loading obtained for ilr coordinates are usually not interpretable. The solution is to go simply back to the clr space by back-transforming the results. Then the scores are

$$\mathbf{Y}^* = \mathbf{Z}^*\mathbf{V}^\top, \tag{1.33}$$

and for an arbitrary affine equivariant covariance estimator we obtain

$$C(\mathbf{Y}) = \mathbf{V}C(\mathbf{Z})\mathbf{V}^\top = \mathbf{V}\mathbf{G}_\mathbf{z}\mathbf{L}_\mathbf{z}\mathbf{G}_\mathbf{z}^\top\mathbf{V}^\top. \tag{1.34}$$

The resulting back-transformed loading matrix is

$$\mathbf{G}_\mathbf{y} = \mathbf{V}\mathbf{G}_\mathbf{z}. \tag{1.35}$$

$\mathbf{G}_\mathbf{y}$ represents the matrix of eigenvectors to the nonzero eigenvalues of $C(\mathbf{Y})$, which are the same as for $C(\mathbf{Z})$, and moreover, the explained variance corresponding to the chosen number of principal components remains unchanged.

The convenience of such introduced notation is given by the possible extension with properties of robustness by replacing the classical estimators of location $T(\mathbf{Z})$ and covariance $C(\mathbf{Z})$ by robust versions.

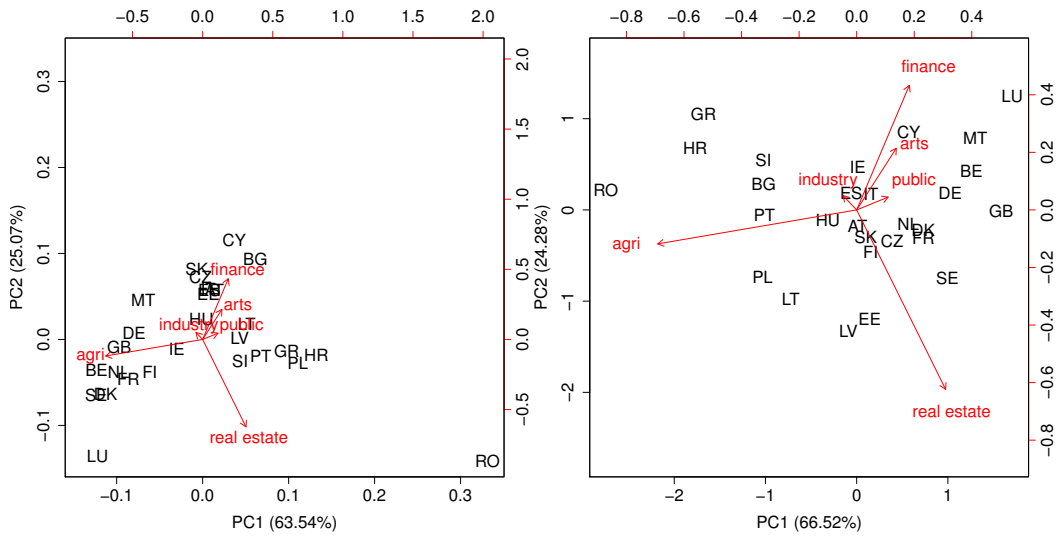


Figure 1.3: Standard (left) and compositional (right) biplot for the employment data.

Principal component analysis is widely used for constructing biplots, the most popular method of visualizing the resulting loadings and scores of a dimension reduction technique applied to multivariate data. It is possible to compare PCA results for the standard and compositional approach on the given example with employment data. Figure 1.3 shows biplots constructed in two different ways. The left plot represents a biplot based on the percentage data of total employment, where standard PCA was used. The right

one is called *compositional biplot*, where PCA is applied on clr coordinates and thus the original values in thousands of employees can be used for this analysis. Although the total variance explained by the first two components is very similar, we can see quite different properties of such constructed biplots. On the standard biplot, Romania lies far away from the rest of the countries and also its location does not really correspond with the real structure of employment by economic activity in this country. Romania is well known for being an agricultural country, so one would expect Romania to be situated with respect to the variable *agri* as it is on the compositional biplot. A similar inappropriate behavior is observed by the countries Luxembourg and Bulgaria, where Bulgaria is placed in the direction of the variable *finance*, and on the other hand, Luxembourg is in between *real estate* and *agri*. Moreover, Romania located far away from the main group of observations may suggest a possibility that this country can be considered as an outlier. Accordingly, a robust extension of the biplot construction should be taken into account in this case to investigate possible outlying observations (Hron and Filzmoser, 2014).

The compositional biplot then shows nice geographic patterns like the Baltic countries together with Poland or central European states concentrated in the middle of the biplot. There is also a group of countries like Cyprus, Luxembourg and Malta, located close to the ray of the variable *finance*, that corresponds to the fact that the gross domestic product of these countries is dominated by the financial sector.

The construction of a biplot will be fully described in Chapter 3. One can be also interested in the influence of some other variables on the structure of employment in these countries. For this reason, a new tool called *ilr biplot* is introduced in Chapter 3 extended also by how to incorporate non-compositional external variables to investigate their mutual relations.

1.5 Correlation analysis

Correlation analysis represents one of the crucial statistical problems when applying the standard statistical analysis on compositional data. Although correlation analysis still represents a widely used tool to express the strength of a linear relation between compositional parts in a quantitative way, it is not directly applicable on them (Pearson, 1897). Standard correlation measures are based on variances and covariances that are defined for the Euclidean space and not for the simplex. The definition of the center and variability for compositional data was already introduced in Section 1.2.4.

The term *spurious correlation* arises quite often in the case of compositional data and it was already mentioned by Pearson (1897) in his seminal paper. The main idea of spurious correlation is shown by the fact that the relation between compositional parts may completely change, when only a subcomposition, coming out of the whole given composition, is considered. The problem of spurious correlation can be described by using the employment data example.

We can consider different attempts to analyze the mutual relations in the employment

data. A first approach consists in computing usual Pearson correlation coefficients for the raw data of absolute numbers in thousands of employees (Table 1.1). Correlations obtained in this way are obviously influenced by the size of the total employment in the respective country, which is the main factor dominating the resulting correlation coefficients. Thus the corresponding values are all positive, mostly larger than 0.9. Some of these correlations are shown in Table 1.2. The size of the total employment can be filtered out by dividing each number of employees by the total size in a country to obtain percentages for each category of economic activities. The correlation coefficients are displayed in the second row of Table 1.2 and the correlation matrix is completely different from the correlation matrix for absolute values of employees. The high positive values of correlations totally disappeared in this case, that supports the impression of high correlations being caused by a size effect. However, expressing our data in percentages induces also some negative correlations. This fact is caused by analyzing correlations of *closed data*, which are vectors with positive components adding up to a constant, in this case 100. This effect is called *negative bias* and it was described by Aitchison (1986) and is represented by the relation

$$\mathrm{cov}(x_1, x_2) + cov(x_1, x_3) + \cdots + \mathrm{cov}(x_1, x_D) = -\mathrm{var}(x_1). \tag{1.36}$$

The presence of negative bias in the covariance structure on the simplex $\mathcal{S}^D$ is very crucial when attempting to apply standard correlation analysis on compositional data. For this reason, using standard correlation analysis for compositional data produces results that do not follow the properties of scale independence and subcompositional coherence, which are essential for working with compositions.

One would intuitively expect that conclusions obtained by one analyst should be compatible with results obtained by another one, for instance by taking some subcomposition of our sample. Taking into account only subcompositions is common in practice. For example, in geochemistry, all elements present in a sample are rarely analyzed and it is preferred to work only with subcompositions. With respect to our example, another approach to correlation analysis is considered by taking only a subcomposition of *industry*, *finance*, *real estate* and *public*, and investigating the mutual relations between these components. The corresponding correlation coefficients are then available in Table 1.2. Nevertheless, one would expect to see coherent results when analyzing different subcompositions of the original data, which is not the case occurred for the employment data set. This phenomenon is called *spurious correlation* and causes inconsistencies when applying standard multivariate methods based on covariances on the compositional data. For this reason, it would be convenient to find a new alternative approach to correlation analysis for compositional data which allows to avoid the problems mentioned above.

One possible solution to measuring an association between compositional parts is the variation matrix (1.22) introduced by Aitchison (1986). When the elements of the variation matrix are close to zero, it can be stated that these two variables, parts, are proportional. The opposite is then given by the high values of ratios, which indicate a high relative variability among all the observations.

	industry finance	*industry real estate*	*industry public*	*industry arts*	*finance real estate*	*finance public*
absolute	0.974	0.924	0.973	0.970	0.958	0.991
percentages	-0.088	-0.022	-0.428	-0.212	-0.137	0.197
subcomposition	-0.502	-0.208	-0.911		-0.177	0.116
	finance arts	*public real estate*	*public arts*	*real estate arts*		
absolute	0.936	0.969	0.943	0.897		
percentages	0.647	0.303	0.120	-0.013		
subcomposition		0.199				

Table 1.2: Spurious correlation: Pearson correlation coefficients for the employment data.

In Table 1.3, the variation matrix for the employment data example is illustrated. For instance, the corresponding element of the variation matrix for *public* and *industry* is 0.08, close to zero, which demonstrates the fact that these two parts are almost proportional among all the observations given in the data set. The same can be observed for the variables *arts* and *finance*, or *industry* and *arts*. On the other hand, *agri* and *real estate* reveal quite low proportionality among all the countries in the EU.

	industry	*agri*	*finance*	*real estate*	*public*	*arts*
industry	0.00	0.68	0.24	0.48	0.08	0.14
agri	0.68	0.00	1.32	1.63	1.02	1.12
finance	0.24	1.32	0.00	0.64	0.18	0.12
real estate	0.48	1.63	0.64	0.00	0.36	0.47
public	0.08	1.02	0.18	0.36	0.00	0.12
arts	0.14	1.12	0.12	0.47	0.12	0.00

Table 1.3: Variation matrix for the employment data.

Focusing only on the bivariate case, the two-part subcomposition (x_1, x_2), the ilr coefficients (1.18) can be simplified to

$$z = \frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2}. \tag{1.37}$$

Then the ilr variable z is only univariate, but it contains all the relevant information between x_1 and x_2 covered by the log-ratio. The variation matrix could then be normalized by defining

$$\text{corr}(x_1, x_2) = \exp(-\text{var}(z)), \tag{1.38}$$

to the interval $[0, 1]$ (Buccianti and Pawlowsky-Glahn, 2005; Filzmoser et al., 2010). Large variability is indicated by values close to zero and small variability by values approaching one. Nevertheless, this alternative does not fulfill the properties of the usual correlation, especially the interpretation of positive and negative association between parts. Another

problem related to correlation analysis for compositional data is to enable some statistical inference, such as significance testing.

	industry	*agri*	*finance*	*real estate*	*public*	*arts*
industry	1.00	0.71	0.89	0.79	0.96	0.93
agri	0.71	1.00	0.52	0.44	0.60	0.57
finance	0.89	0.52	1.00	0.72	0.92	0.94
real estate	0.79	0.44	0.72	1.00	0.84	0.79
public	0.96	0.60	0.92	0.84	1.00	0.94
arts	0.93	0.57	0.94	0.79	0.94	1.00

Table 1.4: $\exp(-\mathrm{var}(z))$ for the employment data.

Correlation analysis of compositional data is meaningful only when it is applied on orthonormal coordinates. The main approach to correlation analysis for compositional data is based on using balances constructed mostly by sequential binary partitioning (Egozcue and Pawlowsky-Glahn, 2005). The procedure of constructing balances is described in detail in Chapter 4. However, all the previously mentioned approaches are based on the ratios between two parts and the influence of other parts is completely ignored. It would be convenient to build such balances containing also the rest of the parts, because the association can be affected by the remaining components. Moreover, it is necessary to symmetrize the orthonormal coordinates with respect to the investigated parts x_1 and x_2 (without loss of generality). Accordingly, such built coordinates can be constructed from two different coordinate systems resulting from permuting the parts in a given composition and then focusing on the role of x_1 or x_2. The procedure of constructing symmetric orthonormal coordinates is described in Section 4.3.

Consequently, correlations of symmetric balances can be computed for the employment data example and compared to the previous approaches. Table 1.5 shows the resulting Pearson correlation coefficients. Using orthonormal coordinates enables us to obtain the correlation coefficient in the common sense of positive and negative values. Nevertheless, special care should be devoted to the interpretation of the resulting correlation, because it is expressed in terms of dominance of both parts to the average behavior of the rest. The correlation coefficients presented in Table 1.5 are no longer comparable to the previous results, because here the remaining parts are considered as well.

The problems of applying correlation analysis on compositional data are discussed in Filzmoser and Hron (2009) and enriched by a new approach based on symmetrical balances introduced in Chapter 4.

1.6 $\mathcal{T}$ spaces: incorporating a total

Compositional data were previously introduced as a special type of multivariate data carrying relative information contained in the ratios between compositional parts. This means that compositional data analysis usually deals with data sets where the total is

	industry	*agri*	*finance*	*real estate*	*public*	*arts*
industry	1.00	0.61	0.09	-0.07	0.23	0.03
agri	0.61	1.00	-0.39	-0.37	-0.28	-0.35
finance	0.09	-0.39	1.00	0.00	0.49	0.70
real estate	-0.07	-0.37	0.00	1.00	0.45	0.20
public	0.23	-0.28	0.49	0.45	1.00	0.39
arts	0.03	-0.35	0.70	0.20	0.39	1.00

Table 1.5: Correlations computed for symmetric balances for the employment data.

unknown or uninformative. The standard practice, the log-ratio approach for working with compositional data is to express the observations in coordinates. Thus, the data are projected into the simplex $\mathcal{S}^D$, a $(D-1)$-dimensional subset of the real space $\mathbb{R}^D$, by applying the closure operation (1.2), where κ is frequently chosen as one. However, the data are then prepared for performing further compositional data analysis, but all information about the total amount is ignored.

Nevertheless, there are some situations in practice, where both relative and absolute information are of interest and should be taken into account to provide reasonable results of the corresponding data analysis. For instance, this is the case in multivariate time series analysis adjusted for compositional data, called *compositional time series* (Barceló-Vidal et al., 2011). Compositional time series should be treated carefully from the perspective of relative information contained in the ratios in investigated time points by using the log-ratio approach. One of the typical goals of time series analysis is to produce predictions of the future based on past and present data, and the analysis of trends. As it was already described above, information about the total abundance is completely ignored by using log-ratio transformations and it is not available for further data analysis as required by time series procedures.

In general, to investigate any compositional vector $\mathbf{x}$ preserving also information about the total, two alternative approaches arise frequently in practice. The first alternative suggests to take the logarithms of each component to form $\ln(\mathbf{x})$. The second one corresponds to treating a log-transformed total together with a composition in one joint analysis. For those purposes, the new product space $\mathcal{T} = \mathbb{R}^+ \times \mathcal{S}^D$ has been introduced (Pawlowsky-Glahn et al., 2014).

1.6.1 Space structure of $\mathcal{T} = \mathbb{R}^+ \times \mathcal{S}^D$

Incorporating the total information into the analysis, the space structure of the product space $\mathcal{T} = \mathbb{R}^+ \times \mathcal{S}^D$ should be investigated. Firstly, the space structure of the positive orthant $\mathbb{R}^D_+$ needs to be analyzed by taking the logarithmic transformation of each component for $\mathbf{x} \in \mathbb{R}^D_+$ in order to establish operations for further analysis in the product space $\mathcal{T}$. Consider a vector with D strictly positive components, $\mathbf{x} \in \mathbb{R}^D_+$. According to Pawlowsky-Glahn and Egozcue (2001), the logarithmic transformation applied to

each component in $\mathbb{R}_+^D$ induces Euclidean space structure over $\mathbb{R}$. Then it is possible to define corresponding operations on $\mathbb{R}_+^D$. Concretely, the Abelian group operation is called *plus-perturbation*, and the external multiplication is *plus-powering*. In mathematical terms, they are defined for $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^D$ and $\alpha \in \mathbb{R}$ as

$$\mathbf{x} \oplus_+ \mathbf{y} = [x_1 \cdot y_1, \ldots, x_D \cdot y_D], \tag{1.39}$$

$$\alpha \odot_+ \mathbf{x} = [x_1^\alpha, \ldots, x_D^\alpha], \tag{1.40}$$

and the respective inner product in $\mathbb{R}_+^D$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^D$ is called *plus-inner-product* defined as

$$\langle \mathbf{x}, \mathbf{y} \rangle_+ = \langle \ln \mathbf{x}, \ln \mathbf{y} \rangle, \tag{1.41}$$

where $\langle , \rangle$ stands for the usual Euclidean inner product in $\mathbb{R}^D$. Assuming the above mentioned operations and the inner product, $\mathbb{R}_+^D$ is a D-dimensional real Euclidean vector space. Additionally, the associated distance and norm are defined as

$$d_+(\mathbf{x}, \mathbf{y}) = d(\ln \mathbf{x}, \ln \mathbf{y}), \qquad \|\mathbf{x}\|_+ = \|\ln \mathbf{x}\|, \tag{1.42}$$

where d and $\|\cdot\|$ represent the Euclidean distance and norm in $\mathbb{R}^D$.

With such introduced definitions and alternative Euclidean structure concerning different group operation $\oplus_+$ and alternative metrics in $\mathbb{R}_+^D$, one can proceed to introduce the space structure of the new product space $\mathcal{T} = \mathbb{R}^+ \times \mathcal{S}^D$.

Consider a vector with D strictly positive components, $\mathbf{x} \in \mathbb{R}_+^D$, which results after the closure operation in a D-part composition $\mathcal{C}(\mathbf{x}) \in \mathcal{S}^D$. The absolute information $t(\mathbf{x}) \in \mathbb{R}_+$ can then be defined by some appropriate function $t(\cdot)$, which can stand as the total sum, the product, the arithmetic or geometric mean, or any other value related to the respective problem.

Thus, the extended vector $\tilde{\mathbf{x}}$,

$$\tilde{\mathbf{x}} = [t(\mathbf{x}), \mathcal{C}(\mathbf{x})] = [t(x), \tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_D], \tag{1.43}$$

represents an element of the product space $\mathcal{T} = \mathbb{R}^+ \times \mathcal{S}^D$, the set of possible values of $t(\mathbf{x})$ and $\mathcal{C}(\mathbf{x})$. As for the Aitchison geometry in Section 1.2.2, it is necessary to adopt Euclidean space structure in $\mathcal{T}$ by defining appropriate operations corresponding to the general principles for working with compositional data.

The Abelian inner group operation and the external multiplication in the product space $\mathcal{T}$ are called $\mathcal{T}$*-perturbation* and $\mathcal{T}$*-powering* (Pawlowsky-Glahn et al., 2014). For $\tilde{\mathbf{x}}, \tilde{\mathbf{y}} \in \mathcal{T}$ and $\alpha \in \mathbb{R}$, they correspond to

$$\tilde{\mathbf{x}} \oplus_T \tilde{\mathbf{y}} = [t(\mathbf{x}) \oplus_+ t(\mathbf{y}), \mathbf{x} \oplus_a \mathbf{y}] = [t(\mathbf{x}) \cdot t(\mathbf{y}), \mathcal{C}(\tilde{x}_1 \tilde{y}_1, \ldots, \tilde{x}_D \tilde{y}_D)], \tag{1.44}$$

$$\alpha \odot_T \tilde{\mathbf{x}} = [\alpha \odot_+ t(\mathbf{x}), \alpha \odot_a \mathbf{x}] = [(t(\mathbf{x}))^\alpha, \mathcal{C}(\tilde{x}_1^\alpha, \ldots, \tilde{x}_D^\alpha)], \tag{1.45}$$

where $\oplus_+$ and $\odot_+$ represent perturbation and powering in $\mathbb{R}_+$ and $\oplus_a$ and $\odot_a$ perturbation and powering in $\mathcal{S}^D$ as introduced in Section 1.2.2. The inner product in $\mathcal{T}$ is called *$\mathcal{T}$-inner-product* and for $\tilde{\mathbf{x}}, \tilde{\mathbf{y}} \in \mathcal{T}$, it is defined as

$$\langle \tilde{\mathbf{x}}, \tilde{\mathbf{y}} \rangle_T = \langle t(\mathbf{x}), t(\mathbf{y}) \rangle_+ + \langle \mathcal{C}(\mathbf{x}), \mathcal{C}(\mathbf{y}) \rangle_a, \tag{1.46}$$

where $\langle , \rangle_+$ stands for the inner product in $\mathbb{R}_+$ and $\langle , \rangle_a$ for the Aitchison inner product in $\mathcal{S}^D$ (Pawlowsky-Glahn and Egozcue, 2001).

As $(\oplus_+, \odot_+, \langle , \rangle_+)$ and $(\oplus_a, \odot_a, \langle , \rangle_a)$ define the operations and metrics in $\mathbb{R}_+$ and $\mathcal{S}^D$, respectively the same then holds for $(\oplus_T, \odot_T$ and $\langle , \rangle_T)$ in $\mathcal{T}$. Consequently, the product space $\mathcal{T} = \mathbb{R}^+ \times \mathcal{S}^D$ with $\mathcal{T}$-perturbation $(\oplus_T)$, $\mathcal{T}$-powering $(\odot_T)$ and $\mathcal{T}$-inner product $(\langle , \rangle_T)$ is a D-dimensional Euclidean vector space on $\mathbb{R}$. The definition of the Euclidean vector space allows us to define the corresponding square distance in $\mathcal{T}$ as

$$d_T^2(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = d_+^2(t(\mathbf{x}), t(\mathbf{y})) + d_a^2(\mathcal{C}(\mathbf{x}), \mathcal{C}(\mathbf{y})) = \ln^2 \frac{t(\mathbf{x})}{t(\mathbf{y})} + d_a^2(\mathcal{C}(\mathbf{x}), \mathcal{C}(\mathbf{y})). \tag{1.47}$$

1.6.2 Practical consequences

In practical situations, it is important to consider if the total abundance is of interest and if it should be part of the compositional data analysis. The next step is then to decide, which is the relevant total function $t(\cdot)$ to use. The main task is to find an isometry between $\mathbb{R}_+^D$ and $\mathcal{T} = \mathbb{R}^+ \times \mathcal{S}^D$, because then the mathematical properties between these two spaces are equivalent. In general, an arbitrary value $t(\cdot)$ related to the specific problem can be appointed. However, it is obvious that a special treatment is required to establish the conditions of compatibility on the total function $t(\cdot)$, so that the statistics or calculus applied on $\mathcal{C}(\mathbf{x})$ are compatible with those performed on $\mathcal{T} = \mathbb{R}^+ \times \mathcal{S}^D$. Two special cases, the total sum and the product total, with all their properties are discussed in detail in Pawlowsky-Glahn et al. (2014), as well as their characteristics of centers and metric variances.

Figure 1.4 shows compositional biplots constructed for the employment data example including the total abundance information in the way described above. Compositional parts are treated in the usual way for constructing a biplot, therefore the clr coordinates are used. The total is then considered in the logarithmized and scaled form in order to investigate the mutual relations between compositional parts and the total. It is clear that incorporating the total either as the total sum, or the product total, yield quite similar results, only with few distinctions, in this case. However, this does not represent a general rule. We can observe that the total information has a significant effect in the whole data set. The observations located in the direction of the ray standing for the total are countries with a high number of inhabitants and also with the total employed population. The importance of the total is also pointed out by the length of the link, which is very dominating in the biplot together with the variable *agri*. In addition, the total variance explained by the first two principal components is much lower than in the case of considering the variables without a total (see Figure 1.3).

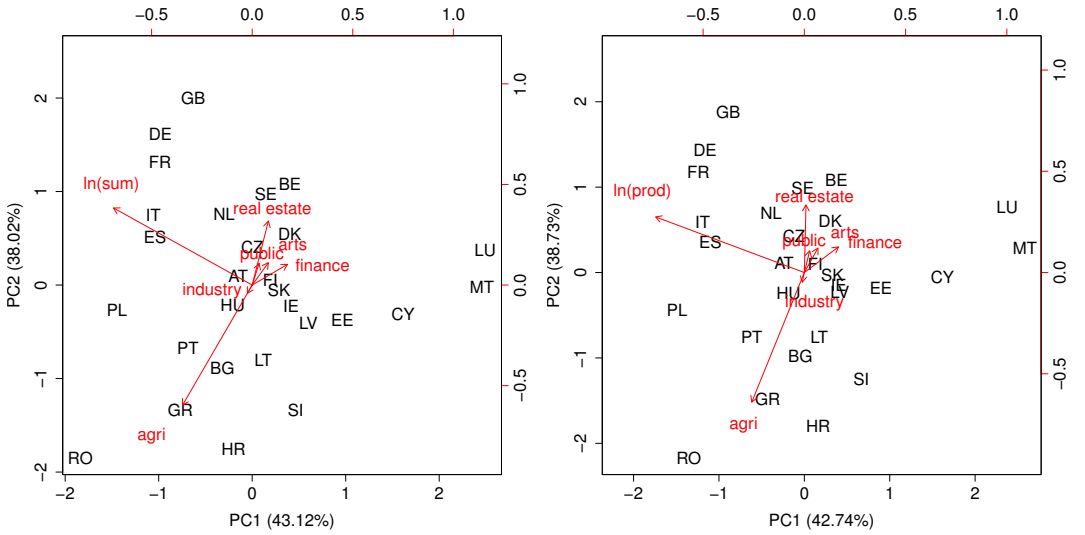


Figure 1.4: Biplot for the employment data including a total: the total sum (left) and the product total (right).

A particular application of $\mathcal{T}$ spaces will be demonstrated in Chapter 2, where multivariate time series for compositional data are investigated and the total abundance is incorporated as a total sum treated jointly with compositional time series.

1.7 Implementation in R

The first implementation of compositional data analysis in R (R Core Team, 2015) was introduced by van den Boogaart et al. (2010) in the package *compositions*. The package provides functions for a consistent analysis of compositional data and positive numbers in the way proposed originally by John Aitchison. The implemented methods offer a statistical analysis of four different scales of amount data: compositional data with relative geometry (Aitchison simplex), compositional data in absolute geometry (classical simplex), positive data with relative geometry (log-scale analysis) and positive data with absolute geometry ($\mathbb{R}_+^D$).

The second package for compositional data analysis and methods is called *robCompositions* implemented by Templ et al. (2011). In addition to classical statistical procedures in *compositions*, the package *robCompositions* considers also robust statistical tools for

compositional data together with corresponding graphics. To express compositions in coordinates, three different possibilities are available: additive, centered and isometric log-ratio transformations. Their implementation differs from the package *compositions* by preserving variable names and absolute values. Considering the ilr transformation, the special choice of orthonormal coordinates based on Fišerová and Hron (2011) is used due to its convenient interpretation properties. The package provides methods for robust principal component analysis, multivariate outlier detection, discriminant analysis and robust imputation of missing values, to name a few.

1.8 Outline of the thesis

This thesis is principally dedicated to compositional data analysis and methods. New tools and methods are developed and compared with the results of standard statistical procedures to show corresponding problems and main advantages of using compositional data analysis. The first introductory chapter presents the history and the main concept of working with compositions and gives an overview on differences and awareness that should be taking into account by using methods developed for compositional data. The second chapter introduces compositional time series and an application of $\mathcal{T}$ spaces in their modeling. The third chapter focuses on the construction of biplots based on the special choice of orthonormal coordinates with respect to enhancing useful interpretation properties followed by incorporating also external non-compositional variables into such constructed biplots. The last chapter discusses potential measures of association between compositional parts focused on building new symmetrical balances and their particular correlation analysis. All statistical procedures and graphics were performed in R, an environment for statistical computing and graphics (R Core Team, 2015).

Chapter 2 introduces compositional time series as multivariate time series describing relative contributions to some total. Vector autoregressive models are used to compare the standard and compositional methods. The standard approach based on raw data is then compared with the compositional one applied on transformed data. The theory of $\mathcal{T}$ spaces is customized for the application of the time series concept. The chapter provides also a concise methodology for an interpretation of the coordinates in the transformed space together with the corresponding statistical inference (like hypotheses testing).

Kynčlová, P., Filzmoser, P., Hron, K. (2015). Modeling compositional time series with vector autoregressive models. *Journal of Forecasting 34 (4)*, pp. 303–314.

Chapter 3 examines biplots as a widely used statistical tool for visualizing the resulting loadings and scores of a dimension reduction technique applied to multivariate data. In the case of compositions, the data have to be pre-processed with a log-ratio transformation before the dimension reduction is carried out. The chapter shows the properties of the compositional biplot and introduces an alternative called the *ilr biplot* as a new tool based on a special choice of orthonormal coordinates resulting from an isometric log-ratio (ilr) transformation. The methodology is demonstrated on real data sets.

Kynčlová, P., Filzmoser, P., Hron, K. (2015). Compositional biplots including external non-compositional variables. Submitted to Statistics.

Chapter 4 discusses different existing and potential measures of association between compositional parts. Correlation coefficients are most popular in statistical practice for measuring pairwise variable associations, but for identifying the association between two compositional parts, standard correlation analysis is not suitable. This chapter introduces an approach of symmetrical balances that capture all relative information in form of aggregated log-ratios of both compositional parts of interest. The balances form orthonormal coordinates, which enables to use standard correlation measures relying on the Euclidean geometry. The idea is supported by simulation studies and an example providing deeper insight into the proposed approach to compare it with alternative measures.

Kynčlová, P., Hron, K., Filzmoser, P. (2015). Correlation between compositional parts based on symmetric balances. Submitted to Mathematical Geosciences.

CHAPTER 2

Modeling compositional time series with vector autoregressive models

Abstract: Multivariate time series describing relative contributions to a total (like proportional data) are called compositional time series. They need to be transformed first to the usual Euclidean geometry before a time series model is fitted. It is shown how an appropriate transformation can be chosen, resulting in coordinates with respect to the Aitchison geometry of compositional data. Using vector autoregressive models, the standard approach based on raw data is compared with the compositional approach based on transformed data. The results from the compositional approach are consistent with the relative nature of the observations, while the analysis of the raw data leads to several inconsistencies and artifacts. The compositional approach is extended to the case when also the total of the compositional parts is of interest. Moreover, a concise methodology for an interpretation of the coordinates in the transformed space together with the corresponding statistical inference (like hypotheses testing) is provided.

Key words: VAR model; Compositional data; Isometric log-ratio transformation; Granger causality

2.1 Introduction

Compositional data represent a special type of multivariate data that generally describe parts of a given whole (Aitchison, 1986; Pawlowsky-Glahn and Buccianti, 2011). A D-part composition is defined as a vector $\mathbf{x} = (x_1, \ldots, x_D)^\top$ with strictly positive real components. They carry only relative information, which is given by the ratios between the components (parts). Most standard statistical methods assume that the analyzed data come from the real Euclidean space with the Euclidean geometry, whereas the natural sample space of compositions is the simplex (Aitchison, 1986). Thus, using classical statistical tools for modeling compositional data may lead to inadequate results.

Compositional time series (CTS) represent multivariate time series of compositions, often characterized by a constant sum constraint representation, at each time point t. Thus a CTS can be defined as the series $\{\mathbf{x}_t : t = 1, \ldots, n\}$, where $\mathbf{x}_t = (x_{1t}, \ldots, x_{Dt})^\top$ are elements of the simplex $\mathcal{S}^D$, the sample space of representations of compositional data to a chosen constant sum constraint κ. CTS are thus characterized by positive components $x_{1t}, \ldots, x_{Dt}$ with a constant sum at each time t (frequently the constant is taken as 1). This constraint forms in practice the crucial problem when modeling compositional time series by standard multivariate time series methods. From the methodological point of view, the problem with a statistical analysis of CTS using standard methods is caused by the specific geometry of compositional data, the Aitchison geometry on the simplex (Egozcue and Pawlowsky-Glahn, 2006), that accounts for inherent properties of compositional data (Egozcue, 2009).

Several approaches for modeling CTS have been introduced. The principal strategy is based on using log-ratio transformations. This procedure consists of transforming given CTS to the space of coordinates – the real vector space with the Euclidean structure – to leave the Aitchison geometry and, practically, break the unit sum constraint of the original time series. Subsequently, standard multivariate time series methods can be applied to the transformed time series.

In the context of CTS, the most frequently used transformations have been additive log-ratio (alr) transformations (Aitchison, 1986; Mills, 2010; Barceló-Vidal et al., 2011), although they lead to oblique coordinates with respect to the Aitchison geometry. The reasonable alternative is represented by the isometric log-ratio (ilr) transformations (Egozcue et al., 2003) that result in orthonormal coordinates. In Bergmann (2008) a particular choice of ilr coordinates was used in order to facilitate the interpretation of the results; nevertheless, due to the apparent complexity of the interpretation of the ilr coordinates, their systematic use for the analysis of CTS is still not fully accepted (Barceló-Vidal et al., 2011). Consequently, although several different approaches for analyzing CTS have been proposed (see Larrosa, 2005), even with a compositional VARIMA model on the simplex (Barceló-Vidal et al., 2011), compositional time series modeling does not appear to be extensively known.

According to Barceló-Vidal et al. (2011), the full compositional VARIMA model and the estimation of the parameters do not depend on the specific log-ratio transformation used.

However, restricted models, that are applied to facilitate the interpretation of parameters, lead to different compositional ARIMA models depending on the transformation applied to the data.

This paper is based on using a special choice of an ilr transformation in order to facilitate a concise approach for the interpretation in coordinates. We focus on vector autoregressive (VAR) models, but an extension to more general models would be possible. Section 2.2 provides a general introduction to the geometry of compositional data, and Section 2.3 refers to special transformations in this context. Section 2.4 explains how the VAR model can be used for compositional data, and it also shows that the resulting final model and the predictions do not depend on the particular choice of the transformation. Further extensions concerning modeling both the relative and absolute information in the context of time series are contained in Section 2.5. Practical examples in Section 2.6 highlight major differences of time series modeling and hypothesis testing when using untransformed or appropriately transformed data. The final Section 2.7 concludes.

2.2 The simplex $\mathcal{S}^D$ as a compositional space

The sample space of representations of D-part compositions to a chosen constant sum constraint is given by the simplex

$$\mathcal{S}^D = \Big\{(x_1, \dots, x_D)^\top : x_i > 0, i = 1, \dots, D; \sum_{i=1}^{D} x_i = \kappa\Big\}, \tag{2.1}$$

where κ is a positive constant. Due to the relative character of compositional data, the specific choice of κ is not relevant; the information contained in the ratios between the compositional parts remains the same. The $(D-1)$-dimensional vector space structure on the simplex $\mathcal{S}^D$ is induced by the operations perturbation and power transformation, defined for compositions $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ and $\alpha \in \mathbb{R}$ as

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, x_2 y_2, \dots, x_D y_D)^\top, \qquad \alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \dots, x_D^\alpha)^\top, \tag{2.2}$$

respectively. Here $\mathcal{C}(\cdot)$ denotes the closure operation that converts each compositional vector from $\mathbb{R}_+^D$ into its representation in $\mathcal{S}^D$. Using the opposite element of $\mathbf{y}$, $\mathbf{y}^{-1} = \mathcal{C}(y_1^{-1}, y_2^{-1}, \dots, y_D^{-1})^\top$, the inverse perturbation $\ominus$ can be defined as

$$\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus \mathbf{y}^{-1}. \tag{2.3}$$

Additionally, the Aitchison inner product is defined for two compositions $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ as

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}, \tag{2.4}$$

which induces the Euclidean vector space structure of the simplex $\mathcal{S}^D$. The inner product can be used to construct a norm and a distance in the simplex

$$\|\mathbf{x}\|_a^2 = \langle \mathbf{x}, \mathbf{x} \rangle_a, \qquad d_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_a. \tag{2.5}$$

The distance is known as the Aitchison distance and it holds important properties associated with compositional data, like invariance under perturbation, invariance under permutation of parts and subcompositional coherence (Pawlowsky-Glahn and Buccianti, 2011).

Compositions in $\mathcal{S}^D$ can be expressed as perturbation-linear combination of compositional vectors, forming a basis or a generating system of $\mathcal{S}^D$ (with respect to the Aitchison geometry). The corresponding coordinates of compositions (real vectors) thus result from transformations of $\mathcal{S}^D$ onto $\mathbb{R}^{D-1}$ or a hyperplane of $\mathbb{R}^D$. Because the coordinates are formed by logarithms of ratios (log-ratios), we refer to log-ratio transformations. The preferable representation of compositions is formed by their coordinates with respect to an orthonormal basis, leading to a one-to-one isometric mapping of the Aitchison geometry on the simplex $\mathcal{S}^D$ to the Euclidean geometry in the real space $\mathbb{R}^{D-1}$. A brief review of the frequently used log-ratio transformations is provided in the next section.

2.3 Log-ratio transformations of compositions and their interpretation

Consider a composition $\mathbf{x} = (x_1, \dots, x_D)^\top \in \mathcal{S}^D$. The additive log-ratio (alr) transformation is a mapping from the simplex $\mathcal{S}^D$ to the real space $\mathbb{R}^{D-1}$, and it depends on the choice of the denominator in the log-ratios, forming the coordinates. Accordingly, the alr transformations are defined as

$$\mathbf{y}^{(k)} = \mathrm{alr}_k(\mathbf{x}) = \left(\ln \frac{x_1}{x_k}, \dots, \ln \frac{x_{k-1}}{x_k}, \ln \frac{x_{k+1}}{x_k}, \dots, \ln \frac{x_D}{x_k} \right)^\top, \quad k = 1, \dots, D. \tag{2.6}$$

Although the alr transformations seem to be easily interpretable, they are not isometric, because their corresponding bases on the simplex are not orthonormal with respect to the Aitchison geometry (Egozcue and Pawlowsky-Glahn, 2006).

The centered log-ratio (clr) transformation of $\mathbf{x} \in \mathcal{S}^D$ is defined as

$$\mathbf{z} = (z_1, \dots, z_D)^\top = \mathrm{clr}(\mathbf{x}) = \left(\ln \frac{x_1}{g(\mathbf{x})}, \dots, \ln \frac{x_D}{g(\mathbf{x})} \right)^\top, \tag{2.7}$$

where $g(\mathbf{x})$ is the geometric mean of the parts of $\mathbf{x}$. This transformation is isometric and maps $\mathcal{S}^D$ into the subspace $V = \{\mathbf{z} \in \mathbb{R}^D : z_1 + \cdots + z_D = 0\}$ of $\mathbb{R}^D$. Thus, the transformed composition lies on a hyperplane through the origin of $\mathbb{R}^D$, which is orthogonal to the vector of units $\mathbf{1}_D$. The clr transformation is closely connected with the isometric log-ratio transformation. Assuming that the inverse clr transformation is an isometry of V onto $\mathcal{S}^D$, then an orthonormal basis in $\mathcal{S}^D$ can be derived from an orthonormal basis in V.

Let $\{\mathbf{v}_1, \dots, \mathbf{v}_{D-1}\}$ be an arbitrary of the space $V \subset \mathbb{R}^D$, then the vectors $\mathbf{e}_i = \mathrm{clr}^{-1}(\mathbf{v}_i)$, $i = 1, \dots, D-1$, represent an orthonormal basis in the simplex $\mathcal{S}^D$. According to this

apparent finding, we can define the isometric log-ratio (ilr) transformations as one-to-one mappings, assigning for a composition $\mathbf{x} \in \mathcal{S}^D$ coordinates with respect to a basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_{D-1}\}$ on the simplex, i.e.

$$\mathbf{u} = \text{ilr}(\mathbf{x}) = (\langle \mathbf{x}, \mathbf{e}_1 \rangle_a, \ldots, \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a)^\top. \tag{2.8}$$

The ilr transformations represent an isometric isomorphism of vector spaces. Thus, for $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ and $\alpha, \beta \in \mathbb{R}$,

$$\text{ilr}(\alpha \odot \mathbf{x} \oplus \beta \odot \mathbf{y}) = \alpha \cdot \text{ilr}(\mathbf{x}) + \beta \cdot \text{ilr}(\mathbf{y}) \tag{2.9}$$

and also

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle \text{ilr}(\mathbf{x}), \text{ilr}(\mathbf{y}) \rangle, \quad d_a(\mathbf{x}, \mathbf{y}) = d(\text{ilr}(\mathbf{x}), \text{ilr}(\mathbf{y})), \quad \|\mathbf{x}\|_a = \|\text{ilr}(\mathbf{x})\| = \|\mathbf{u}\|. \tag{2.10}$$

The ilr coordinates can also be expressed as linear combinations of logarithms of parts whose coefficients add to zero. Considering the $D \times (D-1)$ matrix $\mathbf{V}$ with columns $\mathbf{v}_i = \text{clr}(\mathbf{e}_i)$, the vector of ilr coordinates associated to the matrix $\mathbf{V}$ of a composition $\mathbf{x} \in \mathcal{S}^D$ with respect to $\mathbf{e}_i$, $i = 1, \ldots, D-1$, is

$$\mathbf{u}_\mathbf{V} = \text{ilr}_\mathbf{V}(\mathbf{x}) = \mathbf{V}^\top \text{clr}(\mathbf{x}) = \mathbf{V}^\top \log(\mathbf{x}), \tag{2.11}$$

where the matrix $\mathbf{V}$ is called contrast-matrix associated with the orthonormal basis $\mathbf{e}_i$, $i = 1, \ldots, D-1$ (Egozcue et al., 2003).

Due to the relative character of compositional data and the dimension of the simplex (one less than the number of parts in a composition), a problem of interpretation of the orthogonal coordinates (also called balances) arises in the sense of their relation to the original compositional parts. This problem was solved by introducing the sequential binary partition procedure (Egozcue and Pawlowsky-Glahn, 2005) that consists in splitting parts of a composition into separated groups so that balances representing the groups and the relations between the groups are constructed. A special choice of balances leads to coordinates

$$\text{ilr}(\mathbf{x}) = (z_1, \ldots, z_{D-1})^\top, \quad z_j = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{x_j}{\sqrt[D-j]{\prod_{l=j+1}^{D} x_l}}, \quad j = 1, \ldots, D-1. \tag{2.12}$$

Here, all the relative information (ratios) of part x_1 to the parts $x_2, \ldots, x_D$ is contained in the balance z_1 (Fišerová and Hron, 2011; Filzmoser et al., 2012). Parts of the remaining subcomposition are represented by $z_2, \ldots, z_{D-1}$, nevertheless, already without a similar interpretation as for z_1. This can be simply achieved by perturbing parts of the original composition in (2.12) and considering the particular role of z_1. Finally, the inverse ilr

transformation $\mathbf{x} = \text{ilr}^{-1}(\mathbf{z})$, where

$$x_1 = \exp\Big(\sqrt{\frac{D-1}{D}} z_1\Big), \tag{2.13}$$

$$x_i = \exp\Big(-\sum_{j=1}^{i-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} z_j + \sqrt{\frac{D-i}{D-i+1}} z_i\Big), \qquad i = 2, \ldots, D-1, \tag{2.14}$$

$$x_D = \exp\Big(-\sum_{j=1}^{D-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} z_j\Big). \tag{2.15}$$

is used to express the coordinates back on the simplex.

Coordinate representations given by different log-ratio transformations are related by linear relationships, because vectors $\text{alr}_k(\mathbf{x})$, $\text{clr}(\mathbf{x})$ and $\text{ilr}_{\mathbf{V}}(\mathbf{x})$ represent coordinates of the same composition $\mathbf{x}$ with respect to different bases of the Euclidean vector space $(\mathcal{S}^D, \oplus, \odot)$. Specially, consider two different orthonormal bases of V, $\{\mathbf{v}_1, \ldots, \mathbf{v}_{D-1}\}$ and $\{\mathbf{v}_1^*, \ldots, \mathbf{v}_{D-1}^*\}$, and the corresponding matrices $\mathbf{V}$ and $\mathbf{V}^*$,

$$\mathbf{V} = [\mathbf{v}_1 : \cdots : \mathbf{v}_{D-1}], \qquad \mathbf{V}^* = [\mathbf{v}_1^* : \cdots : \mathbf{v}_{D-1}^*]. \tag{2.16}$$

Then a linear relationship between two ilr transformations of a composition $\mathbf{x} \in \mathcal{S}^D$ with respect to the different bases can be defined as

$$\text{ilr}_{\mathbf{V}}(\mathbf{x}) = \mathbf{V}^\top \mathbf{V}^* \text{ilr}_{\mathbf{V}^*}(\mathbf{x}). \tag{2.17}$$

Other relations between log-ratio transformations can be found in Egozcue et al. (2003); Barceló-Vidal et al. (2011).

Finally, let us introduce a (perturbation) matrix product in the simplex, defined for $\mathbf{A} \in \mathbb{R}_{D\times D}$ and $\mathbf{x} \in \mathcal{S}^D$ as

$$\mathbf{A} \boxdot \mathbf{x} = \mathcal{C}\Big(\prod_{j=1}^{D} x_j^{a_{1j}}, \ldots, \prod_{j=1}^{D} x_j^{a_{Dj}}\Big)^\top. \tag{2.18}$$

This operation forms a linear transformation with respect to the Aitchison geometry, but only if the rows of $\mathbf{A}$ add up to zero, i.e. $\mathbf{A}\mathbf{1}_D = \mathbf{0}_D$. Otherwise, the matrix product on the simplex is not scale invariant, i.e. $\mathbf{A} \boxdot \mathbf{x} \neq \mathbf{A} \boxdot (k\mathbf{x})$ for $k > 0$. Assuming the same restriction for the columns of $\mathbf{A}$, $\mathbf{A}^\top \mathbf{1}_D = \mathbf{0}_D$, then the function $\mathbf{x} \to \mathbf{A} \boxdot \mathbf{x}$ represents an endomorphism on the simplex $\mathcal{S}^D$. The matrix associated with the identity endomorphism is the so-called centering matrix $\mathbf{G}_D = \mathbf{I}_D - D^{-1}\mathbf{1}_D\mathbf{1}_D^\top$.

Let $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ and $\mathbf{y} = \mathbf{A} \boxdot \mathbf{x}$ be an endomorphism on $\mathcal{S}^D$. It is easy to see that this endomorphism can be expressed as

$$\text{clr}(\mathbf{y}) = \mathbf{A} \cdot \text{clr}(\mathbf{x}) \tag{2.19}$$

in the space of clr coordinates (hyperplane of $\mathbb{R}^D$). By using the relationship between clr and ilr transformations, this results in

$$\text{ilr}_{\mathbf{V}}(\mathbf{y}) = \mathbf{A}_{\mathbf{V}} \cdot \text{ilr}_{\mathbf{V}}(\mathbf{x}), \tag{2.20}$$

where the matrix $\mathbf{A}_{\mathbf{V}}$ is obtained from $\mathbf{A}$ as $\mathbf{A}_{\mathbf{V}} = \mathbf{V}^\top \mathbf{A}\mathbf{V}$. It can be shown that $\mathbf{A}$ is not the only matrix that corresponds to $\mathbf{A}_{\mathbf{V}}$ in this transformation. In fact, $\mathbf{A}_{\mathbf{V}}$ can also be expressed as $\mathbf{A}_{\mathbf{V}} = \mathbf{V}^\top \mathbf{A}_0 \mathbf{V}$, where $\mathbf{A}_0 = \mathbf{V}\mathbf{A}_{\mathbf{V}}\mathbf{V}^\top = \mathbf{V}\mathbf{V}^\top \mathbf{A}\mathbf{V}\mathbf{V}^\top = \mathbf{G}_D \mathbf{A} \mathbf{G}_D$ (see Pawlowsky-Glahn and Buccianti, 2011). Accordingly, $\mathbf{A}$ and $\mathbf{A}_0$ represent the same linear transformation on the simplex $\mathcal{S}^D$, i.e. $\mathbf{A} \boxdot \mathbf{x} = \mathbf{A}_0 \boxdot \mathbf{x}$.

2.4 VAR model for compositional time series

2.4.1 The vector autoregressive (VAR) model

Let $\mathbf{x}_t = (x_{1t}, \ldots, x_{Dt})^\top$ be a compositional vector measured at time t, $t = 1, \ldots, n$. Then $\mathbf{z}_t = (z_{1t}, \ldots, z_{D-1,t})^\top$ represents coordinates of $\mathbf{x}_t$ obtained by using an ilr transformation (determined by a contrast-matrix $\mathbf{V}$).

Here we consider a vector autoregressive (VAR) model in reduced form with p lags, denoted as VAR(p) model, that is defined as

$$\mathbf{z}_t = \mathbf{c}_{\mathbf{V}} + \mathbf{A}_{\mathbf{V}}^{(1)} \mathbf{z}_{t-1} + \mathbf{A}_{\mathbf{V}}^{(2)} \mathbf{z}_{t-2} + \cdots + \mathbf{A}_{\mathbf{V}}^{(p)} \mathbf{z}_{t-p} + \boldsymbol{\epsilon}_t, \tag{2.21}$$

where $\mathbf{c}_{\mathbf{V}}$ is a real vector, $\mathbf{A}_{\mathbf{V}}^{(i)}$ $(i = 1, \ldots, p)$ are parameter matrices, and $\boldsymbol{\epsilon}_t$ is the error component (see, e.g., Lütkepohl, 2007). The error process is considered to be a zero mean white noise process with covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. This means that the transformed observation $\mathbf{z}_t$ is modeled based on the p earlier observations $\mathbf{z}_{t-1}, \ldots, \mathbf{z}_{t-p}$. The VAR($p$) model can be equivalently expressed directly on the simplex as

$$\mathbf{x}_t = \mathbf{b} \oplus \left(\mathbf{A}^{(1)} \boxdot \mathbf{x}_{t-1}\right) \oplus \left(\mathbf{A}^{(2)} \boxdot \mathbf{x}_{t-2}\right) \oplus \cdots \oplus \left(\mathbf{A}^{(p)} \boxdot \mathbf{x}_{t-p}\right) \oplus \mathbf{w}_t, \tag{2.22}$$

where $\mathbf{b}$ represents the compositional counterpart to $\mathbf{c}_{\mathbf{V}}$ and $\{\mathbf{w}_t\}$ is the white noise process on the simplex (see Barceló-Vidal et al., 2011).

Let us consider two different ilr transformed coordinates $\mathbf{z}_t$ and $\mathbf{z}_t^*$ $(t = 1, \ldots, n)$ for given compositional time series $\{\mathbf{x}_t : t = 1, \ldots, n\}$. Let $\mathbf{z}_t = \text{ilr}_{\mathbf{V}}(\mathbf{x}_t)$ represent ilr coordinates of the composition $\mathbf{x}_t \in \mathcal{S}^D$ associated with the matrix $\mathbf{V}$, and $\mathbf{z}_t^* = \text{ilr}_{\mathbf{V}^*}(\mathbf{x}_t)$ represent ilr coordinates associated with the matrix $\mathbf{V}^*$. Using the following relations,

$$\mathbf{z}_t^* = \mathbf{V}^{*\top} \mathbf{V} \mathbf{z}_t, \quad \mathbf{A}_{\mathbf{V}^*}^{(i)} = \mathbf{V}^{*\top} \mathbf{V} \mathbf{A}_{\mathbf{V}}^{(i)} \mathbf{V}^\top \mathbf{V}^*, \quad i = 1, \ldots, p, \tag{2.23}$$

it can be shown that a VAR model for compositional time series does not depend on the concrete choice of the ilr transformation. In this case we say that two VAR(p) models, resulting from two different ilr transformations

$$\mathbf{z}_t = \mathbf{c}_{\mathbf{V}} + \mathbf{A}_{\mathbf{V}}^{(1)} \mathbf{z}_{t-1} + \mathbf{A}_{\mathbf{V}}^{(2)} \mathbf{z}_{t-2} + \ldots \mathbf{A}_{\mathbf{V}}^{(p)} \mathbf{z}_{t-p}, \tag{2.24}$$

$$\mathbf{z}_t^* = \mathbf{c}_{\mathbf{V}^*} + \mathbf{A}_{\mathbf{V}^*}^{(1)}\mathbf{z}_{t-1}^* + \mathbf{A}_{\mathbf{V}^*}^{(2)}\mathbf{z}_{t-2}^* + \dots \mathbf{A}_{\mathbf{V}^*}^{(p)}\mathbf{z}_{t-p}^*, \tag{2.25}$$

are compositionally equivalent. This means that the final model on the simplex (2.22), obtained from using the inverse ilr transformation, is invariant to the choice of the ilr transformation, and the same predictions are thus obtained (Barceló-Vidal et al., 2011) (the equivalent properties also holds for alr and clr transformations). Moreover, within the log-ratio methodology, the obtained predictions can be always rescaled to a prescribed constant sum constraint without loss of information.

While for prediction purposes, any of the above mentioned log-ratio transformations can be applied due to the compositional equivalence of VAR models, the role of an appropriate coordinate representation becomes crucial if also statistical inference (like hypotheses testing) is considered. In the following sections we show how ilr coordinates (2.12) can be used for this purpose.

2.4.2 Model specification

The order of a VAR model, i.e., the number of lags p of VAR(p), is unknown in practice, but it can be chosen by using selection criteria. The general approach is to fit VAR(p) models for $p = 0, \dots, p_{max}$ and then choose that number of lags which minimizes the corresponding function of the given selection criterion.

In this paper, the Akaike information criterion (AIC), the Hannan–Quin criterion (HQ), the Schwarz criterion (SC) and the final prediction error (FPE) are computed to choose the value p. They are defined as

$$\mathrm{AIC}(p) = \ln|\hat{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}(p)| + \frac{2}{n}pK^2, \tag{2.26}$$

$$\mathrm{HQ}(p) = \ln|\hat{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}(p)| + \frac{2\ln\ln n}{n}pK^2, \tag{2.27}$$

$$\mathrm{SC}(p) = \ln|\hat{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}(p)| + \frac{\ln n}{n}pK^2, \tag{2.28}$$

$$\mathrm{FPE}(p) = \left[\frac{n+Kp+1}{n-Kp-1}\right]^K |\hat{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}(p)|, \tag{2.29}$$

where $K = D - 1$ is the dimension of $\mathbf{z}_t$, n is the length of the time series, and $\hat{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}(p)$ is the maximum likelihood estimator of the residual covariance matrix (see, e.g., Lütkepohl, 2007). All the above criteria for model specification are invariant to the choice of the ilr transformation, since the value of the determinant of $\hat{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}(p)$ remains unchanged.

The above criteria have different properties: The AIC criterion tends to asymptotically overestimate the real order with a positive probability. On the contrary, HQ and SC yield consistent estimates of the order p, and under general conditions the estimated order converges in probability, if the true VAR order p is less than or equal to p_{max} (see, e.g., Lütkepohl, 2007). In many cases, the choice of the order depends on the objective of the analysis. For instance, if forecasting is the main aim, then the correct order of the VAR

model is not needed. In this case it is reasonable to find a suitable model for prediction by choosing such an order that minimizes a measure of forecast precision. Note that the order of a VAR model can also be specified by sequential testing procedures, namely by testing zero restrictions on parameter matrices (see Lütkepohl, 2007).

2.4.3 Estimation of VAR models

The stationary VAR(p) model (2.21) can be written in the form of a matrix equation

$$\mathbf{Y} = \mathbf{Z}\mathbf{B} + \mathbf{E}, \tag{2.30}$$

where $\mathbf{Y} = (\mathbf{z}_1, \dots, \mathbf{z}_n)^\top$, the $t-$th row of the $n \times [(D-1)p+1]$ matrix $\mathbf{Z}$ equals $\mathbf{Z}_t = (1, \mathbf{z}_{t-1}^\top, \dots, \mathbf{z}_{t-p}^\top)^\top$ and $\mathbf{B} = [\mathbf{c}, \mathbf{A}^{(1)}, \dots, \mathbf{A}^{(p)}]^\top$ contains the parameters. Assuming a sample of size n, $\mathbf{z}_1, \dots, \mathbf{z}_n$, and p presample values, $\mathbf{z}_{-p+1}, \dots, \mathbf{z}_0$, the parameters $\mathbf{B}$ can be estimated separately for each equation (formed by the columns of $\mathbf{Y}$) by the ordinary least squares (OLS) method. If the regressors in all equations are the same (no restriction for parameters are imposed), the estimator is identical to the generalized least squares (GLS) estimator. This estimator is also identical to the maximum likelihood (ML) estimator (conditional on the size of a given initial presample), if the VAR(p) process $\mathbf{z}_t$ is normally distributed and $\boldsymbol{\epsilon}_t$ (rows of the $n \times p$ error matrix $\mathbf{E}$) represent a white noise process, thus $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}), t = 1, \dots, n$.

Such an estimator has the convenient asymptotic properties of standard estimators, it is consistent and asymptotically efficient. If the VAR(p) process is not stationary, or if restrictions are imposed on the parameters, the GLS estimator may be more beneficial (Lütkepohl, 2007).

2.4.4 Hypotheses testing

A frequent task in the context of multivariate time series analysis is testing for causality. For that reason, the Granger causality was introduced (see, e.g., Lütkepohl, 2007), which represents a statistical concept that is based on prediction. In other words, we are interested in testing whether one variable could help to improve predictions of the remaining observed variables. Considering our special choice (2.12) of the ilr transformation, our aim is to test whether variability of z_1, that carries all relative information of the chosen compositional part x_1 to all parts $x_2, \dots, x_D$ (up to a permutation of these parts), has an effect on the coordinates $z_2, \dots, z_{D-1}$, representing the remaining subcomposition.

Consider our VAR(p) model (2.21) with no restrictions on the parameters. Using the operation of vectorization for the estimated parameters $\hat{\mathbf{B}}$, we obtain

$$\hat{\boldsymbol{\beta}} = \mathrm{vec}(\hat{\mathbf{B}}) = \mathrm{vec} \begin{pmatrix} \hat{\boldsymbol{c}}^\top \\ (\hat{\mathbf{A}}^{(1)})^\top \\ \vdots \\ (\hat{\mathbf{A}}^{(p)})^\top \end{pmatrix}. \tag{2.31}$$

Under general assumptions for VAR models, $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normally distributed with asymptotic covariance matrix

$$\widehat{avar}(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} \otimes (\mathbf{Z}^\top \mathbf{Z})^{-1} \tag{2.32}$$

with

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \frac{1}{n-(D-1)p-1} \sum_{t=1}^{n} (\mathbf{z}_t - \mathbf{Z}\hat{\mathbf{B}})(\mathbf{z}_t - \mathbf{Z}\hat{\mathbf{B}})^\top. \tag{2.33}$$

Considering the mentioned properties, we may test linear hypotheses for parameters of the general form $\mathbf{R}\hat{\boldsymbol{\beta}} = \mathbf{r}$ by using the Wald statistics

$$(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})^\top \left\{ \mathbf{R}[\widehat{avar}(\hat{\boldsymbol{\beta}})]\mathbf{R}^\top \right\}^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) \sim F(q, n-(D-1)p-1), \tag{2.34}$$

where q is the number of parameters tested and $\mathbf{R}$ is a $q \times (D-1)^2 p + D - 1$ restricting matrix. The structure of the restricting matrix $\mathbf{R}$ depends on the particular tested null hypothesis. The elements of $\mathbf{R}$ are 0, when the corresponding parameter is not tested, and they are 1, when the significance of a parameter is tested (see, e.g., Lütkepohl, 2007).

Without loss of generality, the main interest consist in testing Granger causality from z_1 to the remaining coordinates $z_2, \ldots, z_{D-1}$. However, the hypothesis can also be tested reversely, i.e. if the coordinates $z_2, \ldots, z_{D-1}$ have a significant effect on z_1. The particular null hypothesis about non Granger causality generally depends on the concrete objective of the analysis and the structure of data.

2.5 $\mathcal{T}$ spaces in the time series context

In this section we introduce an extension of the above considerations to the case, when both relative and absolute information are of interest. The concept of compositional data analysis as it was introduced in sections 2.2 and 2.3 was based on the assumption that compositional data carry exclusively relative information, which is contained in the ratios between their parts. Nevertheless, in practice both relative and absolute information is often necessary to be taken into account in order to provide a reasonable output of a data analysis. The latter information is expressed by modeling absolute values of the sum of the original compositional parts (if this is not trivial, like for proportional representations of compositions summing up to one). Although a careful treatment of relative information is required, provided by the log-ratio approach, the absolute values predictions may represent the final objective of multivariate time series analysis in the case of forecasting.

Formally, consider a vector with D strictly positive components, $\mathbf{x} \in \mathbb{R}_+^D$. For the log-ratio approach, the data can be expressed as closed observations with a constant sum κ, frequently with $\kappa = 1$, without loss of information (Aitchison, 1986). Thus, all information about the total amount is ignored. This information can be incorporated by defining an extended vector space $\mathcal{T} = \mathbb{R}_+ \times \mathcal{S}^D$, that allows to model the relative

structure of the values and the absolute total sum as an additional variable jointly in one model (see Pawlowsky-Glahn et al., 2013).

The vector $\tilde{\mathbf{x}} = [t(\mathbf{x}), \mathcal{C}(\mathbf{x})] = [t_x, \tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_D]$ is the element of $\mathcal{T} = \mathbb{R}_+ \times \mathcal{S}^D$, where $t(\mathbf{x})$ stands for the total sum, i.e. $t(\mathbf{x}) = \sum_{i=1}^{D} x_i$. In the time series context, often the logarithm of the total $t(\mathbf{x})$ will be taken. The compositions $\mathbf{x} = (x_1, \ldots, x_D)^\top$ are modeled by employing a log-ratio transformation; in our case the suggested ilr transformation. Subsequently, the statistical analysis can be performed (including Granger causality with the total variable). In the case of time series analysis, also the back-transformation to the original values is often required for prediction purposes. Hence the log-total needs to be back-transformed by the exponential transformation, and the forecasted compositions in proportions are multiplied by these values. An example in the next section will show the usefulness of this procedure.

2.6 Illustrative examples

2.6.1 First example

We consider a data set that contains compositional time series of monthly measured paper production shares in Europe from May 2004 to December 2009, The data set was kindly provided by Statistics Austria. The time series is represented as proportions of the overall paper production per month in Austria (x_1), the eurozone countries without Austria (x_2), and EU countries not in the eurozone plus the remaining countries in Europe (x_3). Therefore, for each month, the values of the three categories sum up to one.

The aim of this section is to compare the standard approach, when the VAR model is applied directly to the original time series, and the compositional approach based on using the ilr transformation (2.12) and applying the VAR model to the coordinates. Figure 2.1 shows plots of the raw (left) and ilr transformed (right) data. The ilr-variable z_1 (solid line) represents th relative information of the Austrian paper production to the other two parts, while z_2 describes the relative information between x_2 and x_3.

The process of modeling time series consists of model specification, parameter estimation, and diagnostic checking of an assumed VAR(p) model. In this case, a seasonal VAR model is considered because of the monthly observed data. The number of lags is chosen based on using the model selection criteria mentioned in section 2.4.2. In our example, just the proportional data are available and thus a singularity problem occurs when analyzing the raw data with the standard approach due to the unit sum constraint. This problem is usually "circumvented" by omitting one variable, and applying the VAR model only to the remaining two parts; the values for the omitted part is supposed to be calculated afterwards as the complement to one. Nevertheless, the information contained in the last variable is dropped out as the unit sum constraint is just a proper representation of compositional data, not their inherent property. Thus to demonstrate the results of the standard approach, three models were built for the case of omitting subsequently variable x_1, x_2, and x_3, respectively, from the original time series. For the compositional

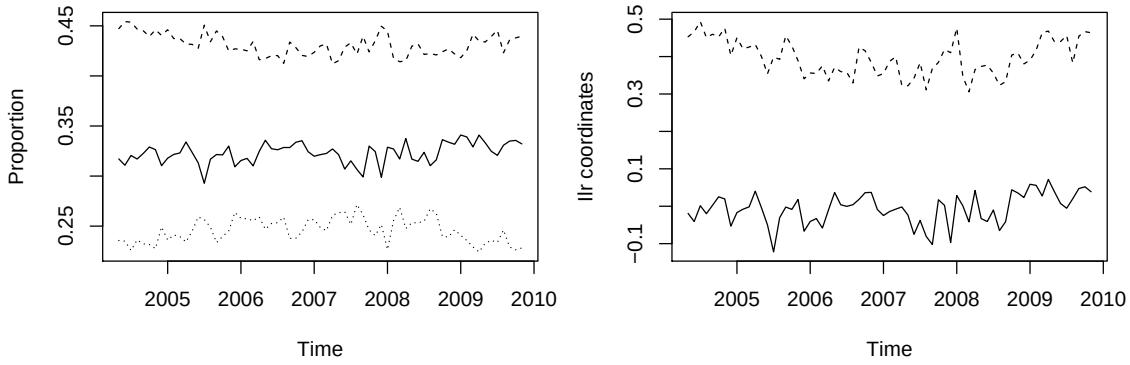


Figure 2.1: Left: Raw untransformed paper production time series (solid line: Austria (x_1), dashed line: eurozone countries without Austria (x_2), dotted line: EU states not in the eurozone plus other countries in Europe (x_3)); right: ilr-transformed time series (solid line z_1, dashed line z_2).

approach, all three time series can be used. Using the mentioned model selection criteria, the resulting numbers of lags p for the original and the ilr-transformed time series are shown in Table 2.1. The same numbers of lags are obtained here for all three cases of omitting one variable, nevertheless, with no guarantee that the results will not differ in general. A VAR(1) model is selected for the standard and the compositional approach in order to allow a subsequent comparison of the results.

Table 2.1: Resulting numbers of lags, using different model selection criteria, for the untransformed and the ilr-transformed data.

Approach	AIC(p)	HQ(p)	SC(p)	FPE(p)
untransformed	4	1	1	4
ilr-transformed	4	1	1	4

The undesirable effect of omitting one compositional part in our VAR(1) models for the original time series can be seen, e.g., in case of forecasting future values. For this purpose we use data from May 2004 to November 2009 for forecasting the (relative) values for December 2009. We apply all three possible VAR models for the original data as well as the ilr approach and compare the corresponding forecasts also with the true December 2009 composition. The results differ for the standard approach in each case of omitting one variable from the model, whereas for the compositional approach the predictions are always the same due to the invariance of using log-ratio transformations. The elimination of a variable from the standard analysis and its subsequent calculation as the complement to one might even cause that the obtained predictions can be negative or equal to zero, or they could be greater than one. This can not occur when using log-ratio transformations.

The relationships between the variables (original parts or the ilr coordinates, respectively)

can be further analyzed by Granger causality analysis. In our example, the hypothesis tested using the ilr approach is $H_0 : a_{21}^{(1)} = 0$, which represents Granger non-causality between the coordinates defined by (2.12). With this transformation, the aim is to test the influence of the relative amount of paper production in Austria to the other European countries, and also conversely. In other words, we are interested in testing, if coordinate z_1 has no effect on z_2, i.e., whether relative information on x_1 (production of paper in Austria) does not influence the ratio to x_2 and x_3. In our case, the Granger non-causality of z_1 to z_2 is not rejected on the significance level 0.05. This means that past values of z_1 probably does not contain information that is useful for predicting z_2. The relative information contained in the paper production of Austria is then not useful in forecasting the ratio to the other countries. Conversely, we can test, whether $H_0 : a_{12}^{(1)} = 0$ (the ratio between x_2 and x_3 has no effect on the variable x_1). In this case, the null hypothesis is also not rejected on the significance level 0.05.

Similar tests can also be carried out for the standard approach. However, the crucial problem consists in omitting one variable as the initial step of the analysis. Granger causality can be investigated only between two variables. The null hypothesis cannot be rejected in each case on the significance level 0.05. The resulting p-values for reasonable combinations are summarized in Table 2.2.

Apparently, tests for Granger causality for the standard and compositional case are not comparable, because they produce results with completely different interpretation with respect to the original time series. Using the standard approach, one effect is always excluded from the complex analysis. According to the obtained results, one cannot state any information about the excluded variable and its influence to the remaining observations. On the contrary, the log-ratio approach offers the overall analysis of all components. The testing is performed in the space of ilr coordinates, which enhances interpretability of the results. Thus, although the null hypothesis was not rejected in all mentioned cases, the ilr results are more reasonable due to modeling effects of all variables.

Table 2.2: Testing Granger causality.

Null hypothesis	p-value
z_1 does not Granger cause z_2	0.741
z_2 does not Granger cause z_1	0.203
x_1 does not Granger cause x_2	0.426
x_1 does not Granger cause x_3	0.291
x_2 does not Granger cause x_1	0.387
x_3 does not Granger cause x_1	0.387

2.6.2 Second example

The second example consists of modeling a four-part compositional time series data set. The data represent gross bonuses of metal production in Austria (in thousands of euros) considering white-collar workers (x_1), blue-collar workers (x_2), commercial apprentices (x_3) and industrial apprentices (x_4). The data are available monthly from January 2004 till November 2010, see Figure 2.2. Our goal is to model the relative structure (relative contributions of the parts on the total metal production) of the compositional time series. However, one can also be interested in predictions of the original absolute values (in thousands of euros), based on the multivariate (relative) structure of the compositional data and the total of the compositional parts. For interpretation purposes, we can define the total of the metal production as the sum of all compositional parts in (original) absolute values as

$$X_t = x_1 + x_2 + x_3 + x_4$$

and then investigate this total as an additional variable in coordinate representation of the compositional time series. Consequently, the predicted total is used to compute predictions in absolute numbers. The compositional approach using the methodology of $\mathcal{T}$ spaces (Section 2.5) for modeling multivariate time series of compositions with a total is compared to the standard approach for the original data.

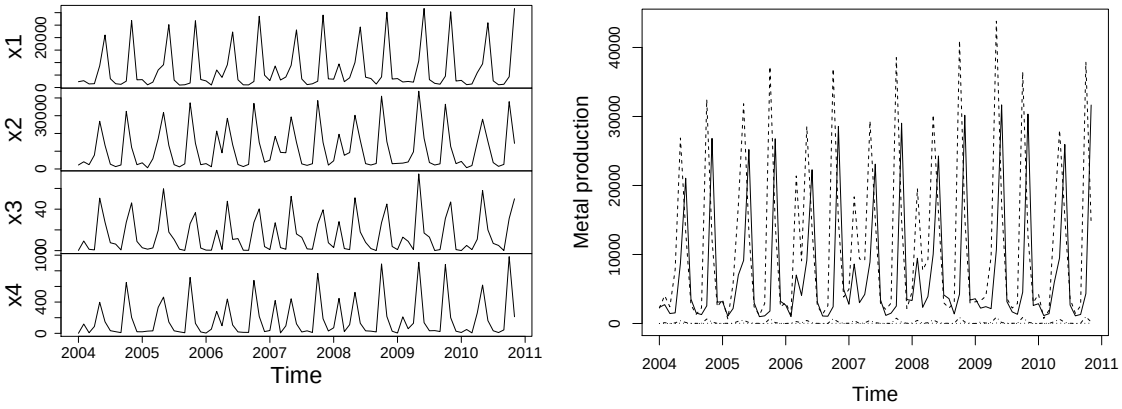


Figure 2.2: Time series of metal production in absolute numbers plotted separately (left) and jointly (right); x_1 is represented by a solid, x_2 by a dashed, x_3 by a dotted and x_4 by a dashed-dotted line, respectively.

Initially, the VAR model is applied to the variables x_1, x_2, x_3, x_4 in the standard way. The metal production data from Austria are represented as monthly measured data, thus the seasonal effect is considered by including dummy variables in the model. The suggested numbers of lags according to the mentioned model selection criteria are summarized in Table 2.3, and a lag of one is selected. In contrast to the first example, the time series

observations were collected in absolute values (without any constant sum constraint), thus omitting a variable for the standard approach is not necessary.

Considering the compositional structure of the data, the theory of $\mathcal{T}$ spaces can be involved to perform a reasonable analysis taking the Aitchison geometry of compositional data into account. The isometric log-ratio transformation (2.12) is applied to the observed variables, and the total sum is taken as an additional variable in the model. The ilr transformed data values and the log-transformed total sum X_t are displayed in Figure 2.3. Subsequently, the time series are investigated also by taking into account the monthly seasonality to choose the corresponding number of lags p (summarized in Table 2.3). Finally, the VAR model with $p = 1$ lag is selected for both the standard and compositional analyses.

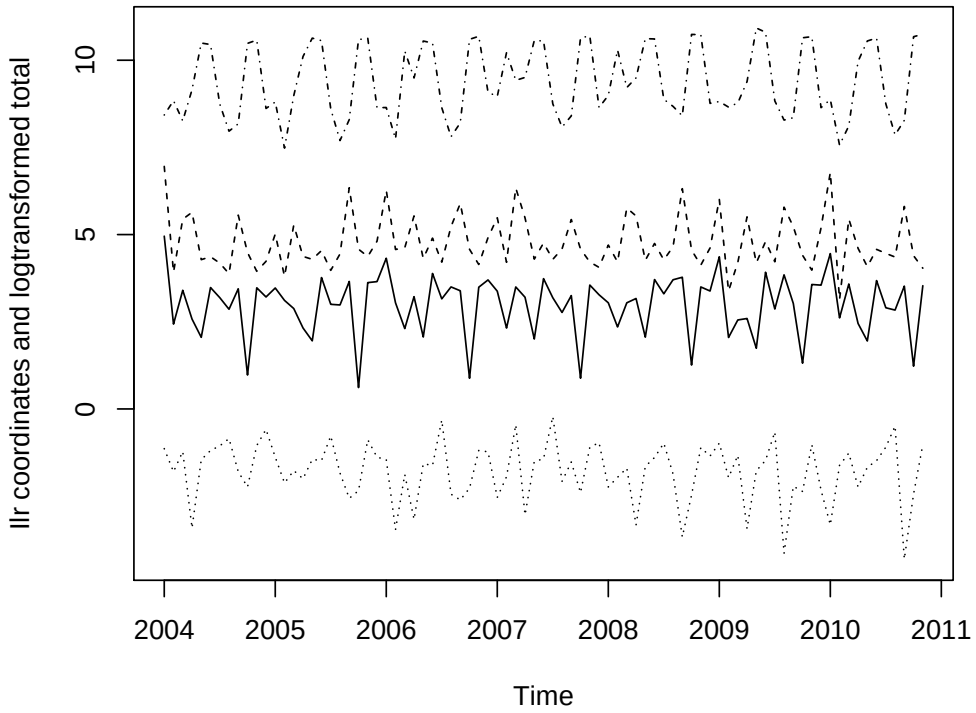


Figure 2.3: Ilr transformed time series (z_1 solid, z_2 dashed and z_3 dotted line) and log transformed total X_t of metal production (dashed-dotted line).

As in the previous example, the time series can be tested for Granger causality in order to detect significant effects between variables, and between variables and the total, respectively. In the case of applying the compositional approach, hypothesis testing is carried out in the space of orthonormal coordinates. Consequently, we consider the ilr coordinates z_1, z_2, z_3 and log-transformed total X_t, where z_1 explains all relative information (ratios) about the original compositional part x_1 with respect to the remaining parts x_2, x_3 and x_4. The reasonable null hypothesis seems to be whether z_1 does not

Table 2.3: Resulting numbers of lags, using different model selection criteria, for the standard approach and the compositional approach using $\mathcal{T}$ spaces.

Approach	AIC(p)	HQ(p)	SC(p)	FPE(p)
standard	10	1	1	2
compositional	10	1	1	1

Granger cause z_2, z_3 and $\log(X_t)$. The resulting p-value is 0.526, which indicates that Granger non-causality cannot be rejected at the significance level 0.05. In other words, the relative information of x_1, here gross bonuses of white-collar workers in the metal production in Austria, to the other compositional parts (included in the coordinate z_1) does not have a significant effect for predicting the coordinates z_2, z_3 and $\log(X_t)$.

The previous null hypothesis focused only on all the relative information of the part x_1 to the remaining parts. Nevertheless, one might also be interested in testing the relative information concerning x_2 (and x_3, x_4, respectively). This can be achieved by exchanging x_1 with another part of interest in the ilr-transformation (2.12), and again focusing on the coordinate z_1. In that way we obtain significance for z_1, which accounts for all the relative information of x_4 to x_1, x_2, x_3, on the coordinates z_2, z_3 and the log-transformed total X_t. The corresponding p-value of 0.043 indicates that the relative information contained in x_4, representing industrial apprentices, has a significant effect on predicting time series of the other variables x_1, x_2, x_3 and the total. In other words, the development of gross bonuses of industrial apprentices influences the data structure of the other workers employed in the metal industry in Austria. Nevertheless, note that by omitting the total variable from the above test, the influence of z_1 on the (purely) relative structure of the other parts (represented by the variables z_2, z_3) is not significant.

The alternative possibility of investigating causality consists in testing, whether the additional total variable affects the observed variables x_1, x_2, x_3 and x_4. In the space of coordinates, we thus test whether $\log(X_t)$ does not Granger cause z_1, z_2, z_3. This particular null hypothesis is not rejected on the significance level 0.05, where the resulting p-value of 0.115 obviously does not depend on a permutation of the compositional parts in (2.12), resulting in different interpretations of the coordinate z_1.

Testing Granger causality with the standard approach is definitely not comparable with the compositional one. The testing by applying compositional techniques is performed in the space of coordinates and the total is considered as an additional variable involved in the investigated model (in its log-transformed form). Nevertheless, the null hypothesis that x_4 does not Granger cause x_1, x_2, x_3 using the original values is rejected on the significance level 0.05 (based on the p-value $6.16 \cdot 10^{-6}$) as well. The similar hypotheses, whether one concrete variable has no influence on the other ones, cannot be rejected in all the remaining cases. We can conclude that past values of x_4, gross bonuses of industrial apprentices, could be very useful in predicting the future values of the remaining variables.

Finally, it might be interesting to compare the accuracy of the predictions obtained by the different approaches. In the compositional case, a back-transformation to the simplex is required to attain the final predictions. As a measure of prediction accuracy we consider the root mean squared error of prediction (RMSEP),

$$\text{RMSEP} = \sqrt{\frac{1}{m}\sum_{t=1}^{m} \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|^2},$$

where m is the number of predicted values ahead. In our case, predictions for 24 subsequent months are made (from December 2009 to November 2011) to compare the accuracy of forecasting. The results show that the RMSEP for the compositional case is 167.02, whereas when using the standard approach the RMSEP is 170.94. According to these results, the compositional approach results in slightly better predictive ability than the conventional one.

2.7 Concluding remarks

Compositional time series are by definition multivariate, mostly represented with a constant sum constraint, and they carry only relative information. Since their sample space is the simplex rather than the real space with the usual Euclidean geometry, they need to be expressed in appropriate (preferably orthonormal) coordinates with respect to the Aitchison geometry before VAR models are employed. We have proposed a specific ilr-transformation to represent the compositional parts in orthonormal coordinates, that is preferable among other log-ratio transformations, because it allows for a meaningful interpretation of the results in terms of the original compositional parts. Moreover, the particular choice of the ilr transformation does not change the resulting predictions of the original compositional values.

Applying VAR models directly to the raw untransformed data may lead to inappropriate models that do not respect the compositional nature of the data. One may get artifact like singularity of the data due to their constant sum constraint, or predictions outside the data range of proportional data (negative or larger than one). Omitting a variable in case of time series with constant sum can even lead to different models and prediction, depending on which variable is omitted. Also for compositional time series without a constant sum constraint, the analysis of the raw data will not focus on the relative information inherent in the data, and may be driven by large values rather than by small ones, which might be of equal importance in a relative sense.

Compositional time series can also be represented as compositional data with a total. The main idea of modeling these time series is based on the theory of $\mathcal{T}$ spaces that enables modeling the relative structure of variables with the absolute total sum together in one model. This approach is thus especially useful when both relative and absolute information is of interest for time series analysis. Furthermore, using the $\mathcal{T}$ spaces approach, the total is included in the model as a separate information from the relative

one (represented by orthonormal coordinates), so that Granger causality also between the coordinates and the total can be investigated. This fact can be of interest in many real-world problems.

Acknowledgments

The paper is supported by the Operational Program Education for Competitiveness-European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic), and the grant IGA PrF_2014_028 Mathematical Models of the Internal Grant Agency of the Palacký University in Olomouc.

CHAPTER 3

Compositional biplots including external non-compositional variables

Abstract: Biplots represent a widely used statistical tool for visualizing the resulting loadings and scores of a dimension reduction technique applied to multivariate data. If the underlying data carry only relative information (i.e. compositional data expressed in proportions, mg/kg, etc.) they have to be pre-processed with a log-ratio transformation before the dimension reduction is carried out. In the context of principal component analysis, the resulting biplot is called compositional biplot. We introduce an alternative, the *ilr biplot*, which is based on a special choice of orthonormal coordinates resulting from an isometric log-ratio (ilr) transformation. This allows to incorporate also external non-compositional variables, and to study the relations to the compositional variables. The methodology is demonstrated on real data sets.

Key words: Compositional data; Log-ratio transformations; Principal component analysis; Singular value decomposition; Compositional biplot; Ilr biplot

3.1 Introduction

Compositional data represent multivariate observations where the relevant information is contained in the ratios between the variables. Usually, already the measurement unit of such data (proportions, percentages, mg/kg, ppm, etc.) reflects their relative character.

Since the interest is only in the ratios, the chosen unit is irrelevant, and it forms just a proper representation of the variables, called compositional parts Aitchison (1986). Geometrically, compositional data follow the Aitchison geometry on the simplex Egozcue and Pawlowsky-Glahn (2006). Consequently, standard statistical methods that rely on the standard Euclidean geometry in real space usually fail when they attempt to capture the multivariate structure of compositional data.

In the last two decades, several papers related to the proper statistical treatment of compositional data have appeared, employing the log-ratio methodology to compositional data analysis Aitchison (1986); Buccianti (2013); Buccianti et al. (2006); Egozcue et al. (2003); Egozcue and Pawlowsky-Glahn (2005); Pawlowsky-Glahn and Buccianti (2011). This is also the case in the context of principal component analysis (PCA) for compositional data Aitchison and Greenacre (2002); i Estadella et al. (2011); Filzmoser et al. (2009a); Filzmoser and Hron (2013); Hron and Filzmoser (2014). Nevertheless, the recent developments concern just the case of PCA working only with compositional parts Aitchison and Greenacre (2002); Filzmoser et al. (2009a); Hron and Filzmoser (2014) or when supplementary variables are projected into a PCA biplot of compositional data i Estadella et al. (2011). A concise methodology on how to incorporate also additional non-compositional variables into one PCA is still not available, despite the fact that these cases frequently occur in practice. Examples are chemical concentration data of air quality measurements with external information like wind-speed or solar radiation, or election data with external information characterizing the districts or regions.

The goal of this paper is to introduce an approach, based on the isometric log-ratio transformation for compositional data Egozcue et al. (2003), for exploring the relations between compositional parts and external non-compositional variables using biplots of principal components. In the next section, some basics on biplots are recalled (Section 3.2). Section 3.3 treats biplots from a compositional data analysis point of view. Section 3.4 provides a detailed description of the methodology to include additional variables to compositional data in this context. Its usefulness for practical applications is demonstrated on two examples (Sections 3.5): for a data set from the German federal election, and for employment data in the European Union. The final Section 3.6 discusses possible problems and extension of the new analytical tool.

3.2 The PCA biplot: construction and interpretation

Consider a given data matrix $\mathbf{X}$ of dimension $n \times D$. The n rows are formed by the observation vectors $\mathbf{x}_{i.}$, for $i = 1, \ldots, n$, and the D columns by the variable vectors $\mathbf{x}_j$, for $j = 1, \ldots, D$. Throughout the manuscript, a "." in the index of a vector will refer to the corresponding row of a matrix, and a vector will always be a column-vector. Thus, $\mathbf{X} = (\mathbf{x}_{1.}^\top, \ldots, \mathbf{x}_{n.}^\top)^\top = (\mathbf{x}_1, \ldots, \mathbf{x}_D)$. We further assume that $\mathbf{X}$ is mean-centered, i.e. the column-wise arithmetic mean is subtracted from each column.

A PCA biplot can be constructed using singular value decomposition (SVD) of $\mathbf{X}$, given

by

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top, \tag{3.1}$$

where $\mathbf{U} \in \mathbb{R}^{n\times n}$ and $\mathbf{V} \in \mathbb{R}^{D\times D}$ represent orthogonal matrices and $\mathbf{D} \in \mathbb{R}^{n\times D}$ is a (rectangular) diagonal matrix, where the diagonal consists of non-negative values, the *s*ingular values, which are arranged in descending order ($d_{11} \geq d_{22} \geq \cdots \geq d_{kk} \geq 0$). Here, $k \leq \min(n, D)$ denotes the rank of $\mathbf{X}$. With this decomposition, $\mathbf{X}$ can be expressed as

$$\mathbf{X} = \sum_{i=1}^{k} d_{ii}\mathbf{u}_i{\mathbf{v}_i}^\top, \tag{3.2}$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$, respectively, represent the i-th column of the matrix $\mathbf{U}$ and $\mathbf{V}$, respectively. Due to the orthogonality of $\mathbf{U}$ and $\mathbf{V}$ the following equations hold:

$$\mathbf{X}\mathbf{X}^\top\mathbf{u}_i = d_{ii}^2\mathbf{u}_i, \tag{3.3}$$

$$\mathbf{X}^\top\mathbf{X}\mathbf{v}_i = d_{ii}^2\mathbf{v}_i. \tag{3.4}$$

Thus, $\mathbf{u}_i$ is the i-th eigenvector of $\mathbf{X}\mathbf{X}^\top$ to the eigenvalue d_{ii}^2, and $\mathbf{v}_i$ is the i-th eigenvector of $\mathbf{X}^\top\mathbf{X}$ to the same eigenvalue d_{ii}^2. From the latter equation it is immediate that $\mathbf{v}_i$ is also an eigenvector of the sample covariance matrix

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}^\top\mathbf{X}, \tag{3.5}$$

which thus corresponds to the i-th loading vector of a classical PCA. Accordingly, the PCA scores information is contained in the matrix $\mathbf{V}$ Jackson (1991).

The goal of the biplot is to plot information of the observations (PCA scores) as well as information of the variables (PCA loadings) in one plot Gabriel (1971). For this purpose we define the decomposition $\mathbf{X} = \mathbf{G}\mathbf{H}^\top$, where the rows of the matrix

$$\underset{n\times k}{\mathbf{G}} = (\mathbf{g}_{1\cdot}, \ldots, \mathbf{g}_{n\cdot})^\top = \sqrt{n-1}\mathbf{U} \tag{3.6}$$

contain the information of the observations, and the rows of the matrix

$$\underset{D\times k}{\mathbf{H}} = (\mathbf{h}_{1\cdot}, \ldots, \mathbf{h}_{D\cdot})^\top = \frac{1}{\sqrt{n-1}}\mathbf{V}\mathbf{D}, \tag{3.7}$$

contain information of the variables. The scores information is usually shown by points in the biplot, while the loadings information is drawn by rays. Since a biplot is usually two-dimensional, the information contained in $\mathbf{X}$ is exactly reproduced if the rank k of $\mathbf{X}$ is two (or less). Otherwise, the descriptive ability of the biplot relies on the amount of variability explained by the first two principal components, and we only obtain $\mathbf{G}\mathbf{H}^\top \approx \mathbf{X}$.

With the above choices of the matrices $\mathbf{G}$ and $\mathbf{H}$, the following properties are obtained Gabriel (1971) and visually explained in Figure 3.1:

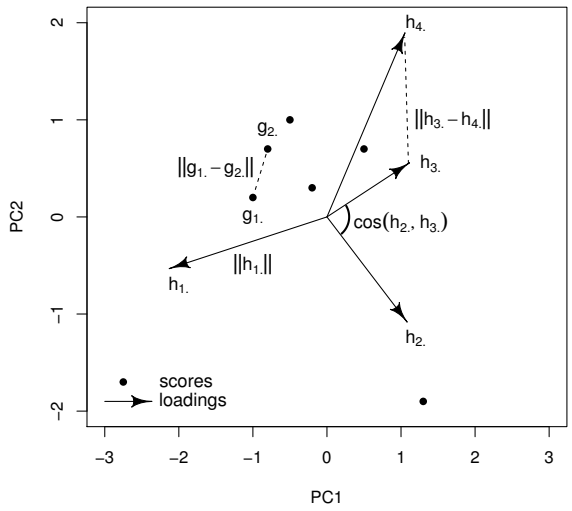


Figure 3.1: Graphical illustration of standard biplot properties.

- The inner product between the rows of $\mathbf{G}$ and the rows of $\mathbf{H}$ estimates the original matrix of observations $\mathbf{X}$, i.e. $\mathbf{g}_{i\cdot}^{\top}\mathbf{h}_{j\cdot} \approx x_{ij}$.

- Since $\mathbf{H}\mathbf{H}^{\top} \approx \frac{1}{n-1}\mathbf{X}^{\top}\mathbf{X} = \mathbf{S}$, a biplot constructed in this way is called *covariance biplot*.

- The length of a ray estimates the standard deviation of the respective variable, $\|\mathbf{h}_{j\cdot}\|^2 = \mathbf{h}_{j\cdot}^{\top}\mathbf{h}_{j\cdot} \approx \frac{1}{n-1}\mathbf{x}_j^{\top}\mathbf{x}_j$.

- Consequently, the cosine of the angle between two rays expresses the approximated correlation coefficients between the corresponding variables, $\cos(\mathbf{h}_{i\cdot}, \mathbf{h}_{j\cdot}) = \frac{\mathbf{h}_{i\cdot}^{\top}\mathbf{h}_{j\cdot}}{\|\mathbf{h}_{i\cdot}\|\|\mathbf{h}_{j\cdot}\|} \approx \frac{\mathbf{x}_i^{\top}\mathbf{x}_j}{\|\mathbf{x}_i\|\|\mathbf{x}_j\|}$.

- The squared distances between the rows of $\mathbf{H}$ approximate the mean squared difference between the variables, $\|\mathbf{h}_{i\cdot} - \mathbf{h}_{j\cdot}\|^2 \approx \frac{1}{n-1}\|\mathbf{x}_i - \mathbf{x}_j\|^2$.

- The squared distances between the rows of $\mathbf{G}$ approximate the squared Mahalanobis distance between the observations, $\|\mathbf{g}_{i\cdot} - \mathbf{g}_{j\cdot}\|^2 \approx (\mathbf{x}_{i\cdot} - \mathbf{x}_{j\cdot})^{\top}\mathbf{S}^{-1}(\mathbf{x}_{i\cdot} - \mathbf{x}_{j\cdot})$.

The above well-known properties for the covariance biplot will be explored in the following for compositional data.

3.3 Biplots for compositional data

3.3.1 The clr transformation and corresponding biplot properties

Compositional data follow the Aitchison geometry on the simplex. Before applying PCA and constructing a biplot, the data need to be transformed to the usual Euclidean geometry. A popular transformation for this purpose Aitchison and Greenacre (2002) is the centered log-ratio (clr) transformation Aitchison (1986), defined for a D-part composition $\mathbf{x} = (x_1, \ldots, x_D)^\top$ as

$$\mathbf{y} = (y_1, \ldots, y_D)^\top = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right)^\top . \tag{3.8}$$

The expression in the denominator, $\sqrt[D]{\prod_{i=1}^{D} x_i}$, represents the geometric mean of the given composition $\mathbf{x}$, denoted as $g(\mathbf{x})$.

Let us assume the $n \times D$ matrix $\mathbf{Y}$ as a matrix of clr coefficients of $\mathbf{X}$, the original uncentered compositional data matrix. The elements of $\mathbf{Y}$ are denoted by y_{ij}, the rows by $\mathbf{y}_{i\cdot}$, and the columns by $\mathbf{y}_j$. Since the clr transformation preserves the distances between the objects Egozcue et al. (2003), the standard procedures can be applied for the newly constructed matrix $\mathbf{Y}$. For the sake of convenience, we will use the same notation as in the last section. Accordingly, in analogy to (3.1), the SVD decomposition of $\mathbf{Y}$ is given by

$$\mathbf{Y} = \mathbf{U}\mathbf{D}\mathbf{V}^\top. \tag{3.9}$$

Further, the matrices $\mathbf{G}$ and $\mathbf{H}$ are defined according to (3.6) and (3.7), respectively. Using only the first two components of these matrices for the biplot construction, the relation $\mathbf{G}\mathbf{H}^\top = \mathbf{Y}$ holds if the rank of $\mathbf{Y}$ is not larger than two–otherwise this relation is only approximately valid, and the quality of the approximation depends on the rank of $\mathbf{Y}$. The rows of the matrix $\mathbf{G}$ contain the object information, and the rows of the matrix $\mathbf{H}$ contain the information of the clr variables. Both sources of information are used to construct the so-called *c*ompositional biplot Aitchison and Greenacre (2002).

The essential difference between the standard and the compositional biplot is that $\mathbf{H}$ does not directly represent the original variables but transformed versions thereof. It is possible to interpret the single clr variables as those capturing all the relative information (ratios) about the corresponding compositional parts (in the numerator of the log-ratio) Fišerová and Hron (2011). Nevertheless, from a numerical perspective, one should be aware of the fact that the geometric mean in the denominator can be driven by possible distortion (like rounding errors) of the involved parts. For this reason, the interpretation of clr variables in the sense of the original compositional parts (in terms of (sub)dominance of the part of interest to the "mean" part in the composition) requires a careful selection of parts, included in the parent composition. As a consequence, the interpretation of the relations in the compositional biplot has to be adapted (Figure 3.1):

- Similar to the standard biplot, the inner product between the rows of $\mathbf{G}$ and the rows of $\mathbf{H}$ estimates the matrix of clr coefficients $\mathbf{Y}$,

$$\mathbf{g}_{i\cdot}^{\top}\mathbf{h}_{j\cdot} = \sqrt{n-1}\mathbf{u}_{i\cdot}^{\top}\frac{1}{\sqrt{n-1}}(\mathbf{v}_{j\cdot}\mathbf{D}) = \mathbf{u}_{i\cdot}^{\top}\mathbf{D}\mathbf{v}_{j\cdot} \approx y_{ij} = \ln\frac{x_{ij}}{g(\mathbf{x})}, \tag{3.10}$$

  where $\mathbf{u}_{i\cdot}$ and $\mathbf{v}_{j\cdot}$ are i-th and j-th row of $\mathbf{U}$ and $\mathbf{V}$, respectively.

- The lengths of the rays estimate the standard deviations of clr transformed variables (clr coefficients),

$$\|\mathbf{h}_{j\cdot}\|^2 = \mathbf{h}_{j\cdot}^{\top}\mathbf{h}_{j\cdot} = \frac{1}{n-1}(\mathbf{v}_{j\cdot}\mathbf{D})^{\top}(\mathbf{v}_{j\cdot}\mathbf{D}) \approx \frac{1}{n-1}\mathbf{y}_j^{\top}\mathbf{y}_j = \operatorname{var}\left(\ln\frac{\mathbf{x}_j}{g(\mathbf{x})}\right). \tag{3.11}$$

- The links between the vertices of the rays estimate the standard deviation of the log-ratio between the corresponding compositional parts, hence

$$\begin{aligned}\|\mathbf{h}_{i\cdot} - \mathbf{h}_{j\cdot}\|^2 &\approx \frac{1}{n-1}(\mathbf{y}_i - \mathbf{y}_j)^{\top}(\mathbf{y}_i - \mathbf{y}_j) = \frac{1}{n-1}\sum_{l=1}^{n}(\mathbf{y}_{li} - \mathbf{y}_{lj})^2 \\ &= \frac{1}{n-1}\sum_{l=1}^{n}\left(\ln\frac{x_{li}}{g(\mathbf{x})} - \ln\frac{x_{lj}}{g(\mathbf{x})}\right)^2 = \frac{1}{n-1}\sum_{l=1}^{n}\left(\ln\frac{x_{li}}{x_{lj}}\right)^2 \\ &= \operatorname{var}\left(\ln\frac{\mathbf{x}_i}{\mathbf{x}_j}\right).\end{aligned} \tag{3.12}$$

- The projection of a score onto a link represents an approximate difference between the two clr coordinates y_{ij} and y_{ik}, which is the log-ratio between the original values x_{ij} and x_{ik},

$$\begin{aligned}\mathbf{g}_{i\cdot}{}^{\top}(\mathbf{h}_{j\cdot} - \mathbf{h}_{k\cdot}) &= \sqrt{n-1}\mathbf{u}_{i\cdot}{}^{\top}\frac{1}{\sqrt{n-1}}(\mathbf{v}_{j\cdot} - \mathbf{v}_{k\cdot})\mathbf{D} \\ &\approx y_{ij} - y_{ik} = \ln\frac{x_{ij}}{g(\mathbf{x})} - \ln\frac{x_{ik}}{g(\mathbf{x})} = \ln\frac{x_{ij}}{x_{ik}}.\end{aligned} \tag{3.13}$$

- The Euclidean distance between the rows of $\mathbf{G}$ approximates the Mahalanobis distance between the clr coefficients in the full space with the estimated covariance matrix $\mathbf{S}_{\mathbf{Y}}$ of the clr-transformed variables,

$$\begin{aligned}\|\mathbf{g}_{i\cdot} - \mathbf{g}_{j\cdot}\|^2 &= (\mathbf{g}_{i\cdot} - \mathbf{g}_{j\cdot})^{\top}(\mathbf{g}_{i\cdot} - \mathbf{g}_{j\cdot}) = (n-1)(\mathbf{u}_{i\cdot} - \mathbf{u}_{j\cdot})^{\top}(\mathbf{u}_{i\cdot} - \mathbf{u}_{j\cdot}) \\ &\approx (\mathbf{y}_{i\cdot} - \mathbf{y}_{j\cdot})^{\top}\mathbf{S}_{\mathbf{Y}}^{-1}(\mathbf{y}_{i\cdot} - \mathbf{y}_{j\cdot}).\end{aligned} \tag{3.14}$$

Several further properties of the compositional biplot are listed in Aitchison and Greenacre (2002). Although these are important for interpreting the relations among the compositional parts, they cannot be explored for relating compositional variables with external information.

The important difference between the standard and the compositional biplot is in the interpretation of the rays and of the links between the vertices of the rays. While in the standard biplot, rays and links represent variability among the variables, they represent *r*elative variability in the compositional biplot. Specifically, the correlation measure expressed by the cosine of the angle between two rays (standard biplot) is replaced by the variance of a log-ratio, expressed as the (squared) length of a link in the compositional biplot Aitchison (1986). Accordingly, when the vertices coincide, or nearly so, then the variance $\mathrm{var}(\ln \frac{x_i}{x_j})$ is approximately equal to zero. Thus, the ratio between x_i and x_j is constant, or nearly so, and it could be stated that variables x_i and x_j are interchangeable.

In many situations, the clr coordinates themselves are not appropriate for a statistical analysis, because due to the constraint $y_1 + \cdots + y_D = 0$, resulting from the fact that clr variables represent coordinates with respect to a generating system, the corresponding covariance matrix is singular. A correlation coefficient between clr variables would thus result in biased values. The reason is that for the covariance structure of clr variables the following relations hold: $\sum_{i \neq j} \mathrm{cov}(y_i, y_j) = -\mathrm{var}(y_i)$, $i = 1, \ldots, D$. Consequently, the corresponding correlation coefficients loose their predicative value, because they cannot vary freely between -1 and 1. From this perspective, also for combining the clr variables with external non-compositional ones, the singularity constraint would result in problematic issues. For example, any clr variable cannot be principally taken separately without considering its relation to the other variables, expressed by the zero sum constraint. It thus complicates intepretability of the biplot in the sense of relative information on single compositional parts, discussed in the following. To sum up, this all makes the use of clr variables for the purpose of PCA and the compositional biplot with additional non-compositional variables not recommendable.

3.3.2 The ilr transformation and biplot construction

The isometric log-ratio (ilr) transformation results in orthonormal coordinates $\mathbf{z} = (z_1, \ldots, z_{D-1})^\top$ with respect to the Aitchison geometry, and it also leads to an orthonormal basis of the hyperplane $\mathcal{H} : y_1 + \cdots + y_D = 0$, formed by the clr transformation Egozcue et al. (2003). Consequently, there exists a linear relation between the clr variables and the orthonormal coordinates Egozcue et al. (2003),

$$\mathbf{y} = \mathcal{V}\mathbf{z}. \tag{3.15}$$

The columns of the $D \times (D-1)$ matrix $\mathcal{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_{D-1})$ are orthonormal basis vectors on the hyperplane $\mathcal{H}$,

$$\mathbf{v}_{D-i} = \sqrt{\frac{i}{i+1}}\left(0, \ldots, 0, 1, -\frac{1}{i}, \ldots, -\frac{1}{i}\right)^\top, \qquad i = 1, \ldots, D-1, \tag{3.16}$$

resulting in the ilr coordinates $\mathbf{z}$. In particular, this means that PCA results in the same principal component scores with non-zero variances (the last principal component is formed by the normal vector on $\mathcal{H}$, thus having zero variability).

There are infinitely many possibilities to construct an orthonormal basis. A special choice of orthonormal coordinates that allows to interpret them in terms of the contributions of the single compositional parts is as follows Filzmoser et al. (2012). Consider the compositions $(x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)$, which are re-arranged such that the l-th part is in the first position. We will use the notation $\mathbf{x}^{(l)} = (x_1^{(l)}, x_2^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})$, where each part with index $l = 1, \ldots, D$ could be placed on the first position, and the sequence of the other parts remains unchanged. The ilr transformation of $\mathbf{x}^{(l)}$ results in $\mathbf{z}^{(l)} = (z_1^{(l)}, \ldots, z_{D-1}^{(l)})^\top$, where the components are defined by

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}}, \qquad i = 1, \ldots, D-1. \tag{3.17}$$

Then, the first ilr variable $z_1^{(l)}$ explains all the relative information (log-ratios) about the original compositional part x_l. The coordinates $z_2^{(l)}, \ldots, z_{D-1}^{(l)}$ explain the remaining log-ratios in the composition Fišerová and Hron (2011). Note that the only important position is that of $x_1^{(l)}$, because it can be fully explained by $z_1^{(l)}$. The other parts can be chosen arbitrarily, because different ilr transformations are orthogonal rotations of each other Egozcue et al. (2003). Note that the relation

$$y_l = \sqrt{\frac{D-1}{D}} z_1^{(l)}, \; l = 1, \ldots, D \tag{3.18}$$

confirms our preliminary requirement on interpretability of the resulting coordinates, for $D \to \infty$ both variables approach the same values. On the other hand, both y_l and $z_1^{(l)}$ thus share also interpretational doubts, mentioned by defining the clr variables.

The advantage of obtaining an interpretation for each compositional part is redeemed by the necessity of constructing D coordinate systems, where always just one variable is of primary interest (at the first position). It is obvious that always the first coordinate $z_1^{(l)}$ in each given system corresponds to the clr coordinate y_l, for $l = 1, \ldots, D$, differing by the constant $\sqrt{\frac{D}{D-1}}$.

Consider now an $n \times (D-1)$ matrix $\mathbf{Z}^{(l)}$ with ilr coefficients due to (3.17), for each of the n observations. Assuming D different coordinate systems, then D singular value decompositions are required to obtain scores and loadings for the biplot construction. For $l = 1, \ldots, D$, an SVD gives

$$\mathbf{Z}^{(l)} = \mathbf{U}^{(l)} \mathbf{D} \mathbf{V}^{(l)\top}. \tag{3.19}$$

As it has been shown in Filzmoser et al. (2009a), the diagonal matrix $\mathbf{D}$ is the same as in (3.9) for the clr-transformed data. Moreover, all matrices $\mathbf{U}^{(l)}$ are equal, and they correspond to the matrix $\mathbf{U}$ in (3.9). This means that the scores in the clr space are identical to the scores of the ilr space, apart from the last column of the clr score matrix that contains zeros. Due to the relationship (3.15) between clr and ilr coordinates by

a matrix with orthonormal columns, and the fact that different ilr-transformations are orthogonally related, we get

$$\mathbf{V} = \mathcal{V}^{(l)}\mathbf{V}^{(l)}, \quad \text{for } l = 1, \ldots, D, \tag{3.20}$$

where $\mathbf{V}$ are the loadings from an SVD of $\mathbf{Y}$, and the matrix $\mathcal{V}^{(l)}$ stands for corresponding permutations of the orthonormal basis matrix $\mathcal{V}$, see Egozcue et al. (2003) and Filzmoser et al. (2009a). Considering relation (3.18) it is immediate that the l-th row of $\mathbf{V}$ is equivalent to the first row of $\mathbf{V}^{(l)}$, differing only by the constant $\sqrt{\frac{D}{D-1}}$.

For constructing the biplot, a decomposition of the form

$$\mathbf{Z}^{(l)} = \mathbf{G}^{(l)}{\mathbf{H}^{(l)}}^{\top}, \qquad l = 1, \ldots, D, \tag{3.21}$$

is required. With the above statements, and in analogy to the clr biplot, it is clear that

$$\mathbf{G}^{(l)} = \mathbf{G} = \sqrt{n-1}\mathbf{U}, \tag{3.22}$$

and

$$\mathbf{H}^{(l)} = \frac{1}{\sqrt{n-1}}\mathbf{V}^{(l)}\mathbf{D}. \tag{3.23}$$

Due to the relation between the matrices $\mathbf{V}$ and $\mathbf{V}^{(l)}$, the first row $\mathbf{h}_{1\cdot}^{(l)}$ of the ilr loadings information $\mathbf{H}^{(l)}$ is related to the l-th row $\mathbf{h}_{l\cdot}$ of the clr loadings information $\mathbf{H}$ by

$$\mathbf{h}_{1\cdot}^{(l)} = \sqrt{\frac{D}{D-1}}\mathbf{h}_{l\cdot}, \qquad l = 1, \ldots, D. \tag{3.24}$$

The relationships between the loadings of ilr and clr coefficients are leading to similar properties and to an interpretability as in the compositional biplot. It was shown that the loadings, corresponding to the D first coordinates $z_1^{(l)}$ of the coordinate systems (3.17), only differ by a constant from those coming from the D clr variables. The important aspect is that now the loadings come from different *orthonormal* coordinate systems, those that we are used to deal with in practice Eaton (1983). Methodologically, this is crucial to employ further (non-compositional) variables and to study relationships between them and the coordinate representations of the composition. Doing that by ignoring the zero constant sum constraint of clr variables with singular covariance matrix has no theoretical justification. The properties of the ilr biplot, formed by merging the loading information (coming from D coordinate systems) and scores of a D-part composition, are illustrated in Figure 3.2.

- The inner product between the rows of $\mathbf{G}$ and the rows of $\mathbf{H}^{(l)}$ gives

$$\begin{aligned}\mathbf{g}_{i\cdot}^{\top}\mathbf{h}_{1\cdot}^{(l)} &= \sqrt{\frac{D}{D-1}}\mathbf{g}_{i\cdot}^{\top}\mathbf{h}_{l\cdot} = \sqrt{\frac{D}{D-1}}\mathbf{u}_{i\cdot}^{\top}\mathbf{D}\mathbf{v}_{l\cdot} \\ &\approx \sqrt{\frac{D}{D-1}}y_{il} = \sqrt{\frac{D}{D-1}}\ln\frac{x_{il}}{g(\mathbf{x})}.\end{aligned} \tag{3.25}$$

- The lengths of the rays represent

$$\|\mathbf{h}_{1\cdot}^{(l)}\|^2 = \frac{D}{D-1}\mathbf{h}_{l\cdot}^\top \mathbf{h}_{l\cdot} \approx \frac{D}{D-1}\frac{1}{n-1}\mathbf{y}_l^\top \mathbf{y}_l = \frac{D}{D-1}\mathrm{var}\left(\ln\frac{\mathbf{x}_l}{g(\mathbf{x})}\right). \tag{3.26}$$

- The links between the vertices are

$$\begin{aligned}\|\mathbf{h}_{1\cdot}^{(i)} - \mathbf{h}_{1\cdot}^{(j)}\|^2 &= \frac{D}{D-1}\|\mathbf{h}_{i\cdot} - \mathbf{h}_{j\cdot}\|^2 \\ &\approx \frac{D}{D-1}\frac{1}{n-1}(\mathbf{y}_i - \mathbf{y}_j)^\top(\mathbf{y}_i - \mathbf{y}_j) = \frac{D}{D-1}\mathrm{var}\left(\ln\frac{\mathbf{x}_i}{\mathbf{x}_j}\right).\end{aligned} \tag{3.27}$$

- The projection of a score to the link yields

$$\begin{aligned}\mathbf{g}_{i\cdot}{}^\top(\mathbf{h}_{1\cdot}^{(j)} - \mathbf{h}_{1\cdot}^{(k)}) &= \sqrt{\frac{D}{D-1}}\mathbf{g}_{i\cdot}{}^\top(\mathbf{h}_{j\cdot} - \mathbf{h}_{k\cdot}) \\ &\approx \sqrt{\frac{D}{D-1}}(y_{ij} - y_{ik}) = \sqrt{\frac{D}{D-1}}\ln\frac{x_{ij}}{x_{ik}}.\end{aligned} \tag{3.28}$$

- As for the clr biplot, the Euclidean distance between the rows of $\mathbf{G}$ gives

$$\|\mathbf{g}_{i\cdot} - \mathbf{g}_{j\cdot}\|^2 \approx (\mathbf{y}_{i\cdot} - \mathbf{y}_{j\cdot})^\top \mathbf{S}_{\mathbf{Y}}^{-1}(\mathbf{y}_{i\cdot} - \mathbf{y}_{j\cdot}). \tag{3.29}$$

- The angles between ilr coordinates and clr coefficients remain the same, despite the fact that they are not used for interpreting a correlation structure of a compositional biplot.

$$\cos(\mathbf{h}_{1\cdot}^{(i)}, \mathbf{h}_{1\cdot}^{(j)}) = \frac{\frac{D}{D-1}\mathbf{h}_{i\cdot}^\top \mathbf{h}_{j\cdot}}{\frac{D}{D-1}\|\mathbf{h}_{i\cdot}\|\|\mathbf{h}_{j\cdot}\|} \approx \frac{\mathbf{y}_i^\top \mathbf{y}_j}{\|\mathbf{y}_i\|\|\mathbf{y}_j\|}. \tag{3.30}$$

In the following, the biplot constructed by merging information from loadings of D orthonormal coordinate systems together into one planar graph, as described above, will be called *ilr biplot*. In order to avoid possible confusion, we should note that the ilr biplot as defined here thus corresponds to a scaled compositional biplot of clr variables; they both differ just in the interpretation of the loadings, coming from the employed orthonormal coordinate systems in the ilr biplot. This helps to consider the (scaled) clr variables separately (consequently also within the compositional biplot), and not as an inherent part of the coordinates with respect to a generating system. On the other hand, a *biplot of ilr coordinates* for an interpretable choice of balances Egozcue and Pawlowsky-Glahn (2005), following the properties of a standard biplot, can be constructed as well. In the next step we describe how the ilr biplot can be extended by additional non-compositional variables.

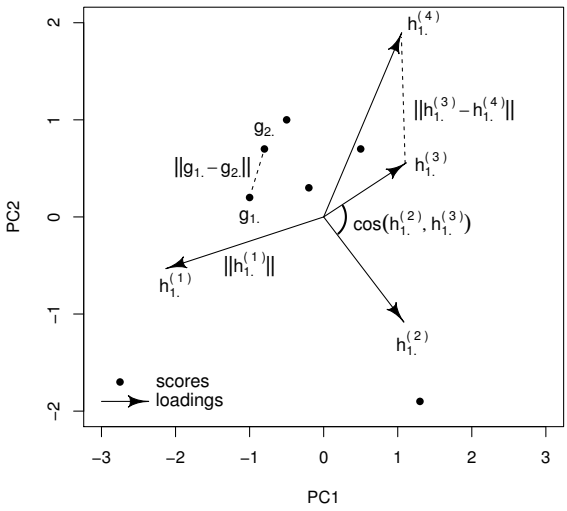


Figure 3.2: Graphical illustration of ilr biplot properties.

3.4 Compositional biplots with additional variables

The next step to construct a meaningful biplot for both compositional data and external non-compositional variables is to analyze, whether the use of a clr transformation or ilr coordinate systems (3.17) for the compositional part of the data would yield the same results (up to a scaling constant) as in the previous section. Consider q additional non-compositional variables $\mathbf{X}^* = (\mathbf{x}_1^*, \ldots, \mathbf{x}_q^*)^\top$, which have already been preprocessed accordingly (e.g. scaled). In the following we have to distinguish different cases how to combine external and compositional variables.

Initially, let us assume only one composition and external variables. We could consider two joint matrices $(\mathbf{Y} \vdots \mathbf{X}^*) \in \mathrm{R}^{n \times (D+q)}$ and $(\mathbf{Z}^{(l)} \vdots \mathbf{X}^*) \in \mathrm{R}^{n \times (D+q-1)}$, where $\mathbf{Y}$ represents clr coordinates and $\mathbf{Z}^{(l)}, l = 1, \ldots, D$, are ilr coefficients for D different coordinate systems. Subsequently, it is required to apply the SVD for both matrices to compare scores and loadings of a compositional and ilr biplot, respectively. For $l = 1, \ldots, D$ the SVD gives

$$(\mathbf{Y} \vdots \mathbf{X}^*) = \mathbf{U}^* \mathbf{D}^* \mathbf{V}^{*\top} = \mathbf{G}^* \mathbf{H}^{*\top}, \tag{3.31}$$

$$(\mathbf{Z}^{(l)} \vdots \mathbf{X}^*) = \mathbf{U}^{*(l)} \mathbf{D}^{*(l)} \mathbf{V}^{*(l)\top} = \mathbf{G}^{*(l)} \mathbf{H}^{*(l)\top}. \tag{3.32}$$

The diagonal matrices $\mathbf{D}^{*(l)}, l = 1, \ldots, D$, are the same and they are equal to $\mathbf{D}^*$ (up to its last zero row/column) corresponding to the SVD for clr coordinates with external variables. Similarly, it is straightforward to show that the scores for the compositional

and ilr biplot, respectively, are identical,

$$\mathbf{G}^{*(l)} = \mathbf{G}^* = \sqrt{n-1}\mathbf{U}^*, \qquad l = 1, \ldots, D. \tag{3.33}$$

The loadings of the SVD of (3.9) and (3.32) are related according to a linear relation between the clr and the ilr transformation (3.15) as

$$\mathbf{V}^* = \mathcal{V}^{(l)}\mathbf{V}^{*(l)}, \qquad l = 1, \ldots, D, \tag{3.34}$$

where the matrix $\mathcal{V}^{(l)}$ represents the corresponding permutation of the orthonormal basis matrix $\mathcal{V}$ (3.16). Accordingly, a relation between the loadings using the ilr transformation and the clr transformation to construct a biplot including external non-compositional variables is obtained as

$$\mathbf{h}_{1.}^{*\,(l)} = \sqrt{\frac{D}{D-1}}\mathbf{h}_{l.}^* \quad \text{for} \quad l = 1, \ldots, D. \tag{3.35}$$

Since we have stated the same relation for loadings without external variables (3.24), it is obvious that incorporating new non-compositional variables to the construction of a biplot does not influence the resulting loadings and scores of the compositional parts.

Consequently, a meaningful interpretation between compositional parts and external variables can be investigated. The representation of the relations among the compositional variables has been introduced in Section 3.3 and in the case of external variables the important role is played by the angles showing the approximate correlation coefficient between two external variables as in the standard biplot. Similarly, for the purpose of interpreting the relations between both types of variables only angles can be considered. Thus the angles can also approximate the correlation structure between the chosen external variable x_i^* $i = 1, \ldots, q$ and an arbitrary compositional part x_l $(l = 1, \ldots, D)$, since the compositional variable is expressed (in the above sense) using coordinate $z_1^{(l)}$, $l = 1, \ldots, D$, being a standard real variable.

Furthermore, let us assume two different compositional variables to investigate their mutual relations among parts in a biplot (external variables are not considered for simplicity). Let $\mathbf{X}_1 = (\mathbf{x}_{11}, \ldots, \mathbf{x}_{1D_1})^\top$ and $\mathbf{X}_2 = (\mathbf{x}_{21}, \ldots, \mathbf{x}_{2D_2})^\top$ be two different compositions with D_1 respectively D_2 parts. To compare loadings and scores it is necessary to construct the SVD for the merged matrices of the clr and ilr coordinates for both compositional variables as follows

$$(\mathbf{Y}_1 \vdots \mathbf{Y}_2) = \tilde{\mathbf{U}}\tilde{\mathbf{D}}\tilde{\mathbf{V}}^\top = \tilde{\mathbf{G}}\tilde{\mathbf{H}}^\top, \tag{3.36}$$

$$(\mathbf{Z}_1^{(l)} \vdots \mathbf{Z}_2^{(k)}) = \tilde{\mathbf{U}}^{(lk)}\tilde{\mathbf{D}}^{(lk)}\tilde{\mathbf{V}}^{(lk)} = \tilde{\mathbf{G}}^{(lk)}\tilde{\mathbf{H}}^{(lk)}, \tag{3.37}$$

where $l = 1, \ldots, D_1$ and $k = 1, \ldots, D_2$. Here $\mathbf{Y}_1$ and $\mathbf{Y}_2$, respectively, represent clr coefficients of $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively, and $\mathbf{Z}_1^{(l)}, \mathbf{Z}_2^{(k)}$ stand for their ilr coordinates according to (3.17). The relationships between scores and loadings for the compositional

biplot and the ilr biplot correspond directly to the simple case of one composition in Section 3.2. Since, by omitting the last two rows and columns of $\tilde{\mathbf{D}}$, the diagonal matrices $\tilde{\mathbf{D}}^{(lk)}$ and $\tilde{\mathbf{D}}$ are the same, for $l = 1, \ldots, D$, the corresponding scores are equal,

$$\tilde{\mathbf{G}}^{(lk)} = \tilde{\mathbf{G}} = \sqrt{n-1}\tilde{\mathbf{U}}. \tag{3.38}$$

We can derive an analogous relation also for the loadings,

$$\tilde{\mathbf{h}}_{1\cdot}^{(lk)} = \sqrt{\frac{D_1}{D_1 - 1}}\tilde{\mathbf{h}}_{l\cdot}, \qquad l = 1, \ldots, D_1, \tag{3.39}$$

thus the ilr loadings concerning the first composition differ only by a constant $\sqrt{\frac{D_1}{D_1-1}}$, where D_1 is the number of parts of the first composition. A similar relation can also be derived for the loadings highlighting parts of the latter composition, i.e.

$$\tilde{\mathbf{h}}_{D_1\cdot}^{(lk)} = \sqrt{\frac{D_2}{D_2 - 1}}\tilde{\mathbf{h}}_{(D_1+k)\cdot}, \qquad k = 1, \ldots, D_2. \tag{3.40}$$

Taking into account the mentioned relations between scores and loadings, an appropriate interpretation of the properties can be incorporated for the case of a biplot constructed for two different compositional variables. Because the ilr coordinates represent standard real variables, their relation for those coordinates resulting from different compositions can be analyzed using angles of the corresponding rays like in the standard biplot. Of course, for measuring the strength of the relative relation between the parts within one composition, the links between the rays still represent the preferred option.

Generally, it is feasible to construct a meaningful biplot for more compositions and external non-compositional variables simultaneously as a simple extension of two previously described cases. The main idea consists in applying a special choice of ilr coordinates (3.17) for each composition and preprocessing external non-compositional variables by the corresponding transformations. Consequently, the transformed variables are merged into one joint matrix followed by SVD to obtain scores and loadings for a biplot construction. Such a biplot representation reflects all possible combinations of the previously mentioned cases.

The main convenience is given by a simple relationship between the resulting SVD for clr and ilr coordinates. Obviously, it is not necessary to construct D coordinate systems when scores are always the same and loadings differ from using the clr transformation only by a scaling constant. It is possible to apply the clr transformation for the compositional parts followed by the same interpretation of the biplot as for a special choice of ilr coordinates (3.17). It is apparent that an appropriate interpretation of scores and loadings always depends also on the characteristic structure of the examined data. Selected cases are demonstrated on real-world data examples in Section 3.5.

3.5 Applications

3.5.1 Election data

The first example describes the results of a federal election in Germany in different federal states (Table 3.1) in September 2013 (data come from German Federal Statistical Office). The aim is to analyze the relations between the votes for the political parties in the elections (compositional variables), and their relation to the unemployment rate and the average monthly income (external non-compositional variables). We consider the votes for the Christian Democratic Union and Christian Social Union of Bavaria, also called The Union (CDU/CSU), Social Democratic Party (SDP), The Left (DIE LINKE), Alliance '90/The Greens (GRÜNE), Free Democratic Party (FDP) and the rest of the parties participated in the elections (other parties). The votes are examined in absolute values (number of valid votes). The unemployment in the federal states is reported in percentages, and the average monthly income in euros.

As mentioned formerly, we are interested in relative information (ratios between the votes for the parties) contained in the data and also the influence of some additional effects. Initially, it is necessary to use appropriate transformations for all variables to obtain a meaningful biplot structure. For the numbers of valid votes, the ilr transformation (3.17) is used. The unemployment information, provided in percentages, is logit-transformed in order to change the relative scale of percentages (as a special case of compositional data) into the absolute one Filzmoser et al. (2009a), and the average monthly income is scaled using its mean and its standard deviation. Subsequently, PCA is performed on these joint data to obtain scores and loadings for constructing the biplot.

Figure 3.3 (left) shows the resulting biplot. In order to avoid possible confusion, names of the original compositional parts are displayed using a function $ilr(.)$. It should stress the fact that the relative information, conveyed by the corresponding coordinates $z_1^{(l)}$ from (3.17), is considered instead of the parts themselves. The explained variance is high, with 92.8%. It is obvious that the federal states are split into two groups. The right located group of states corresponds exactly to the states of former East Germany, except Berlin. The rest of them, left located, are states of former West Germany.

The lengths of the rays of the compositional variables represent the variability of respective ilr coordinates, and the lengths of the rays of the external variables stand for their own variability. The longest ray of the compositional variables represents the standard deviation of ilr variable DIE LINKE, which explains all the relative information of DIE LINKE to the rest of the considered parties. This means that the relative variability of the obtained votes differs a lot among all observed states. On the other hand, SDP and other parties show the smallest relative variability.

The important role in the interpretation of compositional variables in biplots is played by links between vertices of the rays. As the links stand for standard deviations of log-ratios, they can provide the information about relative variability of compositional parts. When the variance of the log-ratios $\mathrm{var}(\ln \frac{x_i}{x_j})$ is approximately zero or nearly so, we can say

that the proportion of the variables is stable, thus x_i and x_j are interchangeable. This is the case for the pair GRÜNE and SDP, and to some extent also for the pair CDU/CSU and other parties. It means their proportion is almost equal among all observations. On the other hand, GRÜNE and DIE LINKE, FDP and DIE LINKE show the highest proportional variability.

The relation between external non-compositional variables can be examined as in the standard biplot. Accordingly, since the rays for income and unemployment are almost orthogonal, these variables seem to be nearly uncorrelated. The angles of the rays are also informative for investigating the relations between external variables and compositional ones, since the latter are ilr coordinates $z_1^{(l)}$ which explain all relative information about the original part x_l. Accordingly, the parties GRÜNE and SDP are strongly positively related to average monthly income. In contrast, the income variable is uncorrelated with voters of FDP and DIE LINKE, there is no essential relationship between income and votes for these political parties. The variable unemployment is strongly negatively related to FDP and CDU/CSU. The opposite relation seems to exist between unemployment and DIE LINKE, i.e. the rate of unemployment influences the proportional structure of people voting DIE LINKE.

Also the federal states can now be associated with the variables: The division of the states into the western and the eastern group is based on differences in the income (higher in the west) and unemployment (higher in the east), but also in the voting behavior. For example, in the eastern states DIE LINKE is much more dominant, and FDP is stronger in the west.

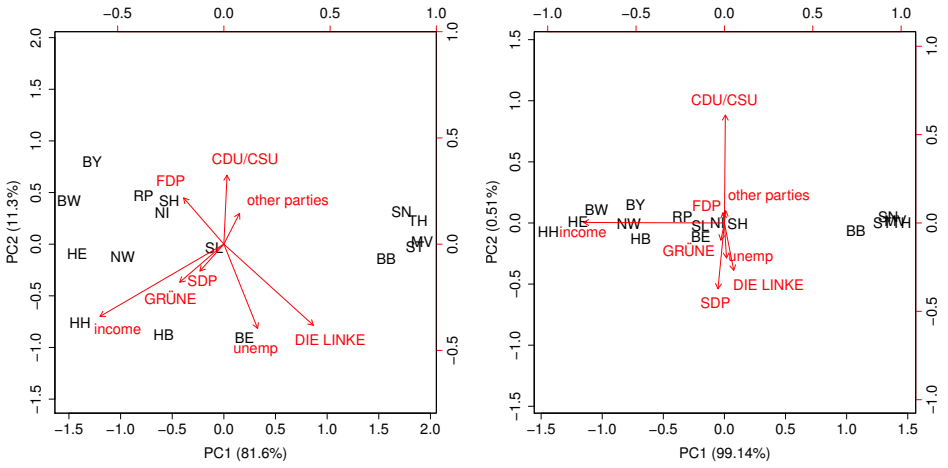


Figure 3.3: Biplots for the German federal elections including unemployment and average monthly income: ilr biplot (left) and standard biplot (right).

We also want to compare the results obtained using the ilr biplot with the case, where the compositional nature of the election data is not accounted for. Therefore, these raw percentage data are combined with the external variables unemployment (in percent) and income (in absolute numbers, scaled). Then, an SVD is carried out to the combined data, and the results are shown in a standard biplot in Figure 3.3 (right). Despite the high explained proportion of variance (99.65%), it is obvious that the resulting biplot differs a lot from the previous solution. We still have the separation of the states into the two groups, which are the result of different income. However, all other variables are essentially uncorrelated to this main direction. Also, this second PCA direction expresses not even 1% of the variability, and is thus rather irrelevant for an interpretation.

We also tried to use a logit-transformation for each of the compositional variables, and join this information with the external variables, i.e. with logit-transformed unemployment and scaled income. The resulting biplot is quite similar to the ilr biplot. There is, however, no guarantee for this phenomenon, as it will be shown in the next example.

3.5.2 Employment data

The aim of the second example is to show how it is possible to construct and interpret a biplot for two different compositions with external non-compositional variables. We consider a data set consisting of the number of employed people in the countries of the European Union (except of Ireland); the data come from EUROSTAT. The first composition describes the number of employed people in different fields of economic activity: agriculture, forestry and fishing (*agri*); industry and construction (*industry*); financial and insurance activities (*finance*); real estate activities (*real estate*); public administration, defense, education, human health and social work activities (*public*); arts, entertainment, recreation and other service activities (*arts*). The second composition illustrates employment in various age categories: from 15 to 24 years (15-24); from 25 to 64 years (25-64); and from 65 years and over (65+). The external variables are: shares of young people living with their parents (*young*), and people at the risk of poverty or social exclusion (*poverty*); both are given in percentages.

Each compositional data set is ilr-transformed with D coordinate systems (3.17) (instead, for simplicity, just the clr transformation can be taken), and afterwards joined together with the external variables. Figure 3.4 shows the resulting ilr biplot. The proportion of explained variance for these first two components is 79.1%. The notation *ilr1* and *ilr2* is used to stress that the respective relative information on compositional parts is related to two particular compositions. It is visible that many observations which are close to each other are also geographically in a neighborhood, for instance the Baltic states (Estonia, Latvia, Lithuania) or the Scandinavian countries (Denmark, Finland, Sweden). Close groups of observations have a similar proportional behavior of the considered variables. In general, richer countries are concentrated in the left part of the biplot, whereas less economically strong ones are in the right part. This is also supported by the external variables *p*overty, pointing to the right side, and *r*eal estate, pointing to the richer countries. On the top of graph we can recognize a group of countries whose

gross domestic product (GDP) consists particularly from activities of the financial sector (Luxembourg, Malta and also Cyprus).

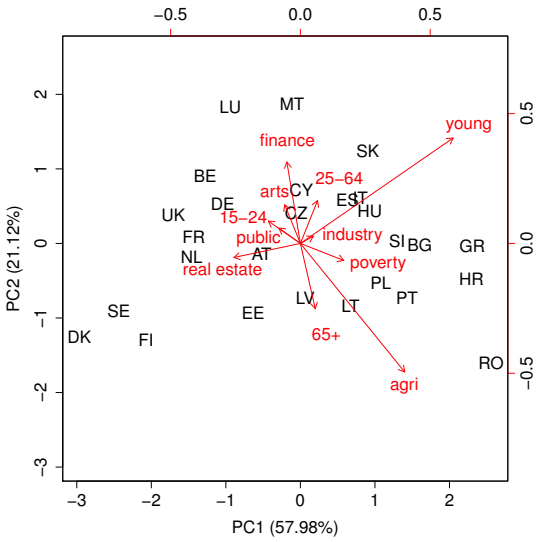


Figure 3.4: Ilr biplot of employed people by economic activity and age, including the risk of poverty or social exclusion and the share of young people living with their parents.

Initially, let us consider only the first composition (economic activity) and relations within these variables. The most significant ray is apparent for the agricultural sector, expressing a large standard deviation of all relative information of the variable *agri* to the remaining sectors. The links suggest that variables *public* and *industry* are proportionally almost equal, the proportion remains almost the same among all the observations. The same behavior can also be observed for the pairs *finance* and *arts*, *industry* and *arts*, *public* and *arts*, *finance* and *public*. On the contrary, the highest proportional variability is evident between agriculture and real estate activities, resp. financial activities. Within the second composition, the links seem to be very similar for all given parts.

Subsequently, we can also investigate relations between both compositions. Since the same ilr transformation (3.17) was used for both of them, the resulting conclusions (concerning the biplot interpretation) are made in the same way as dealing with standard real variables. We can see that the rays for *public* and young employees (15-24) nearly coincide, thus the behavior of these two variables within their parent compositions is positively correlated. The analogous relation can also be identified between 15-24 to *arts* and *finance*, then between 25-64 to *industry* and *finance*. Oppositely, the dominance of

agriculture and young workers in their respective compositions is negatively correlated. It means that the agricultural sector is more important in countries with lower relative representation of young workers (this corresponds also to its positive correlation with employees over 65 years).

Considering now the external variables, we see that the percentage of young people living with their parents is uncorrelated to the proportion of employed people in the agricultural sector. The same conclusion can be stated also for the arts sector. On the contrary, the variable *young* is strongly related to the relative information of the industrial sector. On the other hand, the *young* is strongly negatively correlated with real estate activities since these variables lay approximately on the same line. The risk of poverty appears uncorrelated with employed people between 25 and 64 years. Moreover, the variable *poverty* is strongly negatively correlated with the relative amount of people employed in the public administration. Additionally, the risk of poverty seems to be related also to the variables *agri* and *industry*.

It is interesting to compare the compositional biplot also with a biplot constructed in the standard way, i.e. by ignoring the compositional nature of both compositions. Figure 3.5 shows two standard biplots with different data preprocessing transformations. The left graph represents data without scaling, since the data are expressed in percentages, scaling seems to be unreasonable in this case. Regardless, the resulting biplot does not look very meaningful for the purpose of interpretation. The non-compositional variables reflect significantly higher variability than other observed variables (much longer rays). For this reason, the logit transformation was used for all non-compositional variables and the biplot is shown in Figure 3.5 (right). The explained variance is much lower in this case (72.18%) and there are some significant differences to the ilr biplot. For instance, the age structure of the employed people is completely different. The ray of employed people in age of 25 to 64 is slightly visible and the variance of young people (15-24) has changed its direction. In the ilr biplot the ray coincides with the *public* variable, whereas here it seems to be correlated with real estate activities.

In conclusion, the construction of the ilr biplot enhances the applicability of the compositional biplot, whereas they visualize the same scores and loadings (up to a scaling constant). Frequently, standard biplots result in misleading representations and their construction does not consider the natural geometric structure of compositional data. As it was shown in the examples, the ilr biplot usually yields more reasonable results.

3.6 Discussion

The multivariate data structure of compositional data can be analyzed with the clr biplot, i.e. a biplot based on singular value decomposition of the clr-transformed data. Instead of the clr transformation, we considered a specific ilr transformation for each compositional variable. Variable-wise, this yields the same information as clr, up to a scaling constant. However, from an interpretation point of view, the ilr version is more convenient, since

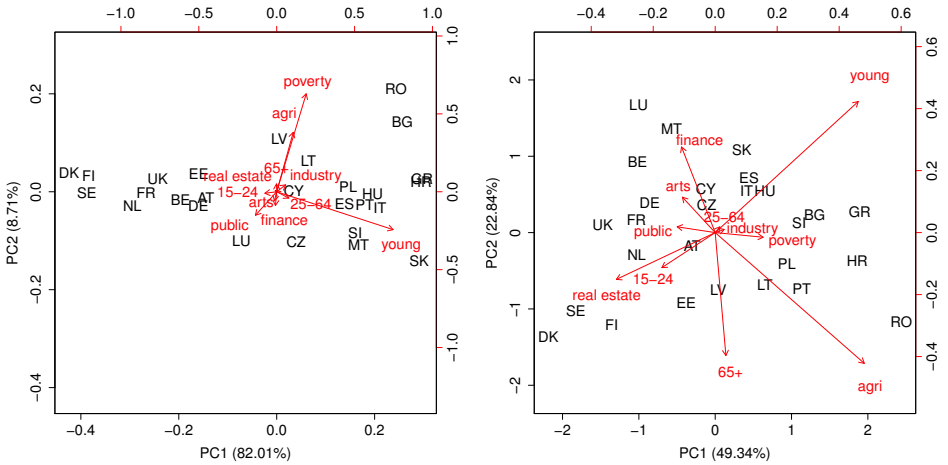


Figure 3.5: Standard biplot of employed people by economic activity and age with external non-compositional variables: no transformation used (left), using logit transformation (right).

each ray in the plot represents an individual orthonormal coordinate with a meaningful interpretation.

The ilr version of the biplot has the additional advantage that it is possible to reasonably combine compositional data with other compositions, and/or with non-compositional (external) variables. The idea is that each composition is ilr-transformed, the results are combined, and then external variables merged. We have shown how the relations between the variables of different compositions, relations to external variables, and relations to the observations can be interpreted.

As in the non-compositional case, a proper preprocessing of external variables should be considered. It has been shown on real examples that the most convenient transformations are logit transformation for percentage data and simple scaling for variables containing absolute values. Possibly also the log-transformation can be applied, when the effect of relative scale of the original variable needs to be suppressed Mateu-Figueras and Pawlowsky-Glahn (2008). A scaling of the compositions is not necessary since the log-ratio transformations are invariant with respect to scaling.

It has been shown on practical real-world examples that the ilr biplot provides a more reasonable representation of the data structure than standard biplots since it captures the different geometrical features of compositional data. As in the usual case, a proper interpretation depends also on the explained proportion of variance. The higher variance,

the better the ilr biplot reveals the real multivariate data structure. It is of course possible to show an ilr biplot not only for the first two components, but also for higher-order pairs.

In further research, a robust version of the ilr biplot can be considered and constituted, based on robust PCA for compositional data Filzmoser et al. (2009a). A robustified version will be less sensitive to outlying observations.

Acknowledgment

This work was supported by the Operational Program Education for Competitiveness - European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic).

Abbreviations

Abbreviation	State
BB	Brandenburg
BE	Berlin
BY	Bavaria
BW	Baden-Württemberg
HB	Bremen
HE	Hesse
HH	Hamburg
MV	Mecklenburg-Vorpommern
NI	Lower Saxony
NW	North Rhine-Westphalia
RP	Rhineland-Palatinate
SH	Schleswig-Holstein
SL	Saarland
SN	Saxony
ST	Saxony-Anhalt
TH	Thuringia

Table 3.1: Codes representing names of German states

Abbreviation	Country
AT	Austria
BE	Belgium
BG	Bulgaria
CY	Cyprus
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GR	Greece
HR	Croatia
HU	Hungary
IT	Italy
LT	Lithuania
LU	Luxembourg
LV	Latvia
MT	Malta
NL	Netherlands
PL	Poland
PT	Portugal
RO	Romania
SE	Sweden
SI	Slovenia
SK	Slovakia
UK	United Kingdom

Table 3.2: Codes representing names of European countries

CHAPTER 4

Correlation between compositional parts based on symmetric balances

Abstract: Correlation coefficients are most popular in statistical practice for measuring pairwise variable associations. Compositional data, carrying only relative information, require a different treatment in correlation analysis. For identifying the association between two compositional parts, symmetrical balances are constructed that capture all relative information in form of aggregated log-ratios of both compositional parts of interest. The balances form orthonormal coordinates, and thus standard correlation measures relying on the Euclidean geometry can be used to measure the association. Simulation studies and an example provide deeper insight into the proposed approach, and allow for comparisons with alternative measures.

Key words: Correlation analysis; Compositional data; Sequential binary partitioning; Symmetrical balances; Log-ratio transformations

4.1 Introduction

Compositional data are characterized by observations on compositional parts that contribute to some whole. Typical examples are the number of votes for political parties in a regional election with a given population, or concentrations of chemical elements in some material with defined weight. An analysis of the associations between the compositional parts (political parties, chemical elements) based on the underlying data is often a first

step to understand the multivariate data structure. However, applying correlation analysis to compositional data can lead to so-called spurious correlations. The problem of spurious correlations dates back to the seminal paper by Pearson (1897) where difficulties obtained by applying standard correlation analysis to data with a constant sum constraint are described. There was a long way (with one important milestone (Chayes, 1960)) to realize that any such reasonable measure cannot be based on the original compositional parts, but exclusively on (log)-ratios forming the only relevant information in compositions Aitchison (1986). In the following years, it turned out that compositional data are not restricted entirely to observations with a constant sum constraint (like proportions or percentages), but the concept covers all observations carrying relative information, with a possibility of being expressed with any prescribed sum constraint without altering the ratios between the parts (Pawlowsky-Glahn et al., 2015). The specific principles of compositional data (scale invariance, permutation invariance and subcompositional coherence) induce the Aitchison geometry (Pawlowsky-Glahn and Egozcue, 2001) with the Euclidean vector space structure that enables to express compositions in proper log-ratio coordinates and continue with statistical processing using standard multivariate statistical tools.

Aitchison (1986) proposed to change completely the point of view on association between compositional parts by introducing the variation matrix. Accordingly, the association between two parts, expressed by the variance of the corresponding log-ratio, is stronger, when the ratio between them tends to be constant. Although this concept turned out to be successful in a range of applications during the last 30 years (Pawlowsky-Glahn and Buccianti, 2011), there are still certain limitations of the approach that inhibit its wider acceptance by the geochemical community (Filzmoser et al., 2010; Reimann et al., 2012). They result mainly from the lack of possibilities of distinguishing *positive* and *negative* association, an essential feature in case of the correlation coefficient. In order to get an impression about such a behavior between geochemical variables, many researchers in the field tend to return back to improper preprocessing tools like the log-transformation that violates the scale invariance principle of compositional data.

This paper proposes to measure the strength of association between compositional parts through the correlation coefficient between a particular choice of orthonormal coordinates with respect to the Aitchison geometry. The orthonormal coordinates are based on log-ratios, formed always by a part of interest and the remaining variables, aggregated in terms of a weighted geometric mean. Methodologically, it follows the idea of having log-ratio coordinates that express all relative information about the parts of interest (Filzmoser et al., 2009b). Two such coordinates need to be constructed simultaneously in a coordinate system, each corresponding to one of the parts. After a brief review of recent possibilities concerning association between compositional parts in the next section, these coordinates are derived in Section 4.3. A detailed discussion of the new correlation measure together with some possible alternatives is provided in Section 4.4. Sections 4.5 and 4.6 employ a geochemical data set in simulations and comparisons to provide deeper insight into the properties of the proposed association measure. The final

Section 4.7 concludes and provides some outlook.

4.2 Measures of compositional association

4.2.1 Correlation analysis for compositional data

The most popular way of measuring association (relation) between variables in practice is by using a correlation measure. Nevertheless, its application on compositional data does not get so straightforward. Let us recall that a D-part composition is represented as a vector $\mathbf{x} = (x_1, \ldots, x_D)^\top$, where all components are positive real numbers that carry only relative information (Aitchison, 1986; Pawlowsky-Glahn et al., 2015). This means that only the ratios between the parts are informative and they form the basis of a reasonable (statistical) processing. Moreover, one should follow the principles of compositional data (Egozcue, 2009) in order to have a guarantee of a reliable analysis. Particularly, the representation of a compositional vector with any sum of components (proportions, percentages, mg/kg, ...) should yield the same results according to the scale invariance principle. These essential assumptions constitute the source of the problems to apply standard correlation analysis on compositional data.

Let us consider compositional data with a fixed prescribed constant sum constraint (the case of proportions), that still occur sometimes in the literature. In this case, correlation analysis is influenced also by the presence of negative bias in the covariance structure. It is represented by the relations

$$\mathrm{cov}(x_i, x_1) + \mathrm{cov}(x_i, x_2) + \cdots + \mathrm{cov}(x_i, x_{i-1}) + \mathrm{cov}(x_i, x_{i+1}) + \mathrm{cov}(x_i, x_D) = -\mathrm{var}(x_i), \tag{4.1}$$

(for $i = 1, \ldots, D$), that make the interpretation of the correlation coefficient meaningless (its value cannot freely vary between -1 and 1). Consequently, this leads to the effect known as *subcompositional incoherence* (Aitchison, 1986), i.e. the correlation between parts of a composition with D parts can be completely in contradiction with the correlation resulting from a subcomposition containing d parts, $d \leq D$. Nevertheless, this is just an illustration of the fact that standard statistical analysis of the original compositional data (that are driven by the Aitchison geometry) cannot be recommended in general.

The Euclidean vector space structure of the Aitchison geometry enables to get a coordinate representation of compositions in the real space, where standard statistical methods can be applied. The resulting centered log-ratio (clr) (Aitchison, 1986) and isometric log-ratio (ilr) coordinates (Egozcue et al., 2003), which seem to be recently the most popular in practice, correspond to coordinates with respect to a generating system and an orthonormal basis, respectively.

Accordingly, the clr coordinates are defined as

$$\mathbf{y} = (y_1, \ldots, y_D)^\top = \left(\ln \frac{x_1}{\sqrt{\prod_{i=1}^{D} x_i}}, \ldots, \ln \frac{x_D}{\sqrt{\prod_{i=1}^{D} x_i}} \right)^\top, \tag{4.2}$$

imposing the zero sum constraint of the new variables, $y_1+\cdots+y_D = 0$. Although it seems to be attractive to assign each single original compositional part to a clr coefficient (and then even continue with correlation analysis), this effort has no geometrical background and should be avoided. Particularly, similar relations as those in (4.3),

$$\mathrm{cov}(y_i, y_1)+\mathrm{cov}(y_i, y_2)+\cdots+\mathrm{cov}(y_i, y_{i-1})+\mathrm{cov}(y_i, y_{i+1})+\mathrm{cov}(y_i, y_D) = -\mathrm{var}(y_i), \quad (4.3)$$

(for $i = 1, \ldots, D$), that show a distortion of the covariance structure, support the argumentation.

Following general theoretical assumptions (Eaton, 1983), correlation analysis of compositional data in the usual sense is meaningful exclusively in log-ratio coordinates with respect to a basis, preferably to an orthonormal one, that guarantees isometry between the Aitchison geometry and the real space. Nevertheless, only $D-1$ such coordinates exist and it is not possible to assign a coordinate to each part simultaneously like in the case of a canonical basis for standard observations. Searching for interpretable orthonormal (ilr) coordinates led to the concept of balances (Egozcue and Pawlowsky-Glahn, 2005) as coordinates with a specific interpretation in terms of balances between groups of compositional parts. These new coordinates are constructed using a procedure called *sequential binary partitioning* (SBP), where the original parts are separated sequentially into non-overlapping groups of parts (Egozcue and Pawlowsky-Glahn, 2005). Concretely,

order	x_1	x_2	x_3	x_4	x_5	x_6	x_7	r	s
1	+	+	+	+	−	−	−	4	3
2	+	+	−	−	0	0	0	2	2
3	+	−	0	0	0	0	0	1	1
4	0	0	+	−	0	0	0	1	1
4	0	0	0	0	+	−	−	1	2
5	0	0	0	0	0	+	−	1	1

Table 4.1: Example of SBP of a seven-part composition.

the main idea of SBP consists of dividing a given group of parts into two subgroups in each order of partition until $D-1$ steps are performed. At the beginning, all parts of a composition are separated into two groups. In each step, the groups formed in the previous step are split again into two subgroups: the first group labeled by +, the second one labeled by −. The zero entries represent parts which are not involved in a partition of a given order. The process ends when all groups consist of a single part. The procedure is usually accompanied with a table, where the resulting partitioning scheme is depicted; see Table 4.1 for an example. The corresponding balances are computed as follows,

$$z_i = \sqrt{\frac{rs}{r+s}} \ln \frac{\left(\prod_+ x_j\right)^{1/r}}{\left(\prod_- x_k\right)^{1/s}}, \qquad i = 1, \ldots, D-1. \quad (4.4)$$

Here. the products $\prod_+$ and $\prod_-$ include parts labeled as $+$ or $-$, and r and s stand for the number of positive and negative signs in the i-th partition, respectively (see Table 4.1). Formula (4.4) indeed supports the naming *balances*: each coordinate represents a log-ratio between two groups of parts, given by their respective geometric means, or alternatively, it contains all the information about the ratios between parts coded as $+$ and parts coded as – (Fišerová and Hron, 2011). Although correlation analysis of balances is now possible (Filzmoser and Hron, 2009), due to (4.4) its interpretation is not straightforward without a deeper prior (expert) knowledge of how the SBP should be constructed.

Consequently, an alternative approach was introduced coming from the idea of having "automated" coordinates that would better stress the role of single compositional parts (Filzmoser et al., 2009b; Fišerová and Hron, 2011). A particular form of SBP (see Table 4.2) leads to coordinates

$$\mathbf{z} = (z_1, \ldots, z_{D-1})^\top, \quad z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}}, \quad i = 1, \ldots, D-1. \tag{4.5}$$

It is obvious that the balance z_1, being proportional to y_1, contains all the relative information of the part x_1 with respect to the remaining parts of the composition, since this part is not contained in any other coordinate of (4.5). The variable z_1 can be interpreted in terms of dominance of x_1 to the other parts, represented by their geometric mean, thus to their *average behavior*. Unfortunately, the same interpretation in sense

order	x_1	x_2	x_3	$\ldots$	x_{D-2}	x_{D-1}	x_D	r	s
1	$+$	$-$	$-$	$\ldots$	$-$	$-$	$-$	1	$D-1$
2	0	$+$	$-$	$\ldots$	$-$	$-$	$-$	2	$D-2$
	.	.	.	.	..	..	.		
$D-2$	0	0	0	0	$+$	$-$	$-$	1	2
$D-1$	0	0	0	0	0	$+$	$-$	1	1

Table 4.2: SBP corresponding to coordinates (4.5).

of explaining *all* relative information cannot be assigned to z_2 and x_2, because this balance already does not contain the first part. Nevertheless, a good candidate for the correlation between relative contributions of x_1 and x_2 in a given composition would be a symmetrical form of z_1 and z_2 because of the exclusive position of the parts of interest (x_1, x_2) in the respective coordinates. This task will be further developed in Section 3. Obviously, the role of x_1 and x_2 can also be interchanged, and a similar construction for different parts can be obtained by permuting the parts in (4.5). Without loss of generality, just the case of x_1 and x_2 will be considered in the following.

4.2.2 Variation matrix as a measure of stability

A main tool of measuring compositional association between two compositional parts has been the variation matrix as a measure of stability (Aitchison, 1986). The variation matrix of a D-part composition is a symmetric matrix of order D, defined as

$$\mathbf{T} = [t_{ij}] = \left[\mathrm{var}\Big(\ln \frac{x_i}{x_j}\Big)\right], \quad i,j = 1,\ldots,D, \tag{4.6}$$

with zero diagonal elements. When the elements of $\mathbf{T}$ are close to zero, the ratio of x_i/x_j is nearly constant, i.e. the two parts x_i and x_j are almost proportional. On the contrary, high variability of the log-ratio indicates very different ratios of two parts among all the observations.

The log-ratios in (4.6) can also be rescaled according to (4.5) so that they correspond, up to orientation, to the normed coordinate of the two-part composition $(x_i, x_j)^\top$. The resulting normalized variation matrix (Pawlowsky-Glahn et al., 2015) is defined as

$$\mathbf{T}^* = [t^*_{ij}] = \left[\mathrm{var}\Big(\frac{1}{\sqrt{2}} \ln \frac{x_i}{x_j}\Big)\right], \quad i,j = 1,\ldots,D, \tag{4.7}$$

where t^*_{ij} stands for the usual (sample) variance of the normalized log-ratio of parts i and j (balance). Subsequently, the relation between $\mathbf{T}$ and $\mathbf{T}^*$ is given as $\mathbf{T} = \frac{1}{2}\mathbf{T}^*$. The measure of variability could be normalized to the range (0,1] as $\tau_{ij} = \exp(-\mathrm{var}(t^*_{ij}))$ for $1 \leq i,j \leq D, i \neq j$ (Buccianti and Pawlowsky-Glahn, 2005; Filzmoser et al., 2010). High variability of the log-ratio then tends to a result approaching zero and, conversely, small variability is reflected by values of τ_{ij} close to one with the limiting case of perfect proportionality. However, this is still just a proper normalization of the elements of the variation matrix and not a correlation measure in the common sense. Particularly, the concept of proportionality does not allow to think in terms of positive and negative association, as it is known from the correlation coefficient.

A different approach to normalization, proposed by Egozcue et al. (2013), is based on the idea that "completely non-proportional" components of a D-part composition correspond to consistent values off the diagonal. Hence, each element out of the diagonal would have the value $2D \cdot \mathrm{TotVar}/D(D-1) = 2\mathrm{TotVar}/(D-1)$, where

$$\mathrm{TotVar}(\mathbf{x}) = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \mathrm{var}\Big(\ln \frac{x_i}{x_j}\Big) \tag{4.8}$$

stands for a measure of global dispersion, called the (Pawlowsky-Glahn and Egozcue, 2001). If we divide each non-diagonal element by this value, we obtain a matrix for complete non-proportionality. This normalization thus leads to the matrix $\tilde{\mathbf{T}}$,

$$\tilde{\mathbf{T}} = \frac{D-1}{2\mathrm{TotVar}}\mathbf{T}. \tag{4.9}$$

If the values of $\tilde{\mathbf{T}}$ are greater than one, the corresponding pair of parts is less proportional than the log-ratio variance that would be observed in a complete non-proportional composition. Values less than one show association between the parts. The smaller the value is, the more associated these two parts are. One can ask if the possible association is significant and statistical hypothesis testing can be carried out. For this purpose, the following two balances,

$$z^v_{1(i,j)} = \frac{1}{\sqrt{2}} \ln \frac{x_i}{x_j}, \quad z^v_{2(i,j)} = \frac{\sqrt{2(D-2)}}{\sqrt{D}} \ln \frac{x_i x_j}{\sqrt[D-2]{\prod_{k \neq \{i,j\}} x_k}}, \tag{4.10}$$

accompanied by the complementary $D-2$ orthonormal coordinates, were employed (Egozcue et al., 2013). The first of them corresponds to an element of the normalized variation matrix, and the latter one links this log-ratio (capturing relative information on the subcomposition $(x_i, x_j)^\top$) with the remaining parts in the given composition. These coordinates allow to test for significance of the elements in the variation matrix indirectly through a regression model, so that deviations from the exact association would be explained within the whole composition. Nevertheless, the interpretation of the elements of the variation matrix themselves is not further enhanced by using this approach.

4.3 Constructing symmetric balances

All the introduced approaches to measuring association between compositional parts are based, directly or indirectly, on working with orthonormal coordinates. However, constructing interpretable balances with SBP (4.4) for correlation analysis needs some prior expertise. It is also important to note that the normalized variation matrix considers only associations between two parts of a given composition. Although this seems to be a clear advantage, one should be aware that any part in the compositional vector is defined by ratios with all other parts in the composition. Consequently, the association based on the simple log-ratio ignores that both parts are unavoidably influenced also by the remaining components. This fact should be taken into account for considering any reasonable coordinates that would allow for a correlation analysis between relative contributions conveyed by both parts. As mentioned in the previous section, one possible setting of coordinates would be (4.5). Nevertheless, it is necessary to symmetrize with respect to parts x_1 and x_2 (without loss of generality).

Accordingly, we consider two coordinate systems $\mathbf{z}$ and $\mathbf{z}^*$ resulting from the permutation of the parts in (4.5) and focus on the role of x_1 and x_2, respectively. It is obvious from Table 4.3 with the corresponding SBPs that the first two coordinates from each system (4.11)(4.12) fully describe the subcomposition $(x_1, x_2)^\top$ within the given composition.

$$z_1 = \sqrt{\frac{D-1}{D}} \ln \frac{x_1}{\sqrt[D-1]{\prod_{i=2}^{D} x_i}}, \; z_2 = \sqrt{\frac{D-2}{D-1}} \ln \frac{x_2}{\sqrt[D-2]{\prod_{i=3}^{D} x_i}}, \tag{4.11}$$

	x_1	x_2	x_3	x_4	...	x_{D-1}	x_D
z_1	+	−	−	−	...	−	−
z_2	0	+	−	−	...	−	−
z_3	0	0	+	−	...	−	−
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
z_{D-1}	0	0	0	0	...	+	−

	x_1	x_2	x_3	x_4	...	x_{D-1}	x_D
z_1^*	−	+	−	−	...	−	−
z_2^*	+	0	−	−	...	−	−
z_3	0	0	+	−	...	−	−
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
z_{D-1}	0	0	0	0	...	+	−

Table 4.3: Construction of balances $\mathbf{z}$ (left) and $\mathbf{z}^*$ (right).

$$z_1^* = \sqrt{\frac{D-1}{D}} \ln \frac{x_2}{\sqrt[D-1]{x_1 \prod_{i=3}^{D} x_i}}, \; z_2^* = \sqrt{\frac{D-2}{D-1}} \ln \frac{x_1}{\sqrt[D-2]{\prod_{i=3}^{D} x_i}}, \tag{4.12}$$

Using the SBPs from Table 4.3, it is now possible to build matrices of clr representations of orthonormal basis vectors corresponding to the first two balances of $\mathbf{z}$ and $\mathbf{z}^*$ (Egozcue et al., 2003) as

$$\underset{D\times 2}{\mathbf{V}_{\mathbf{z}}} = \begin{pmatrix} \sqrt{\frac{D-1}{D}} & 0 \\ -\frac{1}{D-1}\sqrt{\frac{D-1}{D}} & \sqrt{\frac{D-2}{D-1}} \\ -\frac{1}{D-1}\sqrt{\frac{D-1}{D}} & -\frac{1}{D-2}\sqrt{\frac{D-2}{D-1}} \\ \vdots & \vdots \\ -\frac{1}{D-1}\sqrt{\frac{D-1}{D}} & -\frac{1}{D-2}\sqrt{\frac{D-2}{D-1}} \end{pmatrix} \quad \underset{D\times 2}{\mathbf{V}_{\mathbf{z}^*}} = \begin{pmatrix} -\frac{1}{D-1}\sqrt{\frac{D-1}{D}} & \sqrt{\frac{D-2}{D-1}} \\ \sqrt{\frac{D-1}{D}} & 0 \\ -\frac{1}{D-1}\sqrt{\frac{D-1}{D}} & -\frac{1}{D-2}\sqrt{\frac{D-2}{D-1}} \\ \vdots & \vdots \\ -\frac{1}{D-1}\sqrt{\frac{D-1}{D}} & -\frac{1}{D-2}\sqrt{\frac{D-2}{D-1}} \end{pmatrix},$$

where $\mathbf{V}_{\mathbf{z}^*}$ results from a permutation of the first two rows of $\mathbf{V}_{\mathbf{z}}$. Consequently, the first two balances of $\mathbf{z}$ and $\mathbf{z}^*$ are related through an orthogonal transformation as

$$\mathbf{z}^* = \mathbf{V}_{\mathbf{z}}{}^{\top} \mathbf{V}_{\mathbf{z}^*} \mathbf{z}, \tag{4.13}$$

where the orthogonal matrix $\mathbf{V}_{\mathbf{z}}{}^{\top}\mathbf{V}_{\mathbf{z}^*}$ has the form

$$\mathbf{V}_{\mathbf{z}}{}^{\top} \mathbf{V}_{\mathbf{z}^*} = \begin{pmatrix} -\frac{1}{D-1} & \sqrt{\frac{D-2}{D}}\frac{D}{D-1} \\ \sqrt{\frac{D-2}{D}}\frac{D}{D-1} & \frac{1}{D-1} \end{pmatrix}. \tag{4.14}$$

Note that both matrices $\mathbf{V}_{\mathbf{z}}$ and $\mathbf{V}_{\mathbf{z}^*}$ are closely connected to the respective coordinates. Namely, their columns $\mathbf{v}_1 = (v_{11}, \ldots, v_{D1})^{\top}, \mathbf{v}_2 = (v_{12}, \ldots, v_{D2})^{\top}$ and $\mathbf{v}_1^* = (v_{11}^*, \ldots, v_{D1}^*)^{\top}$, $\mathbf{v}_2^* = (v_{12}^*, \ldots, v_{D2}^*)^{\top}$ with zero sums of their elements represent logcontrast coefficients of z_1, z_2 and z_1^*, z_2^*, respectively (Aitchison, 1986), i.e.

$$z_1 = \sum_{i=1}^{D} v_{i1} \ln x_i, \; z_2 = \sum_{i=1}^{D} v_{i2} \ln x_i, \; z_1^* = \sum_{i=1}^{D} v_{i1}^* \ln x_i, \; z_2^* = \sum_{i=1}^{D} v_{i2}^* \ln x_i. \tag{4.15}$$

Because of the roles of the above mentioned coordinates with respect to the single parts x_1 and x_2, one can construct new symmetric balances capturing their relative contributions expressed through log-ratios to other parts in the composition. Let x_1 be the first part of interest; the case of x_2 can be processed accordingly. Based on basic geometry, a symmetric coordinate z_1^s capturing relative information about x_1 corresponds to an angle bisector of $\mathbf{v}_1$ and $\mathbf{v}_2^*$. Similarly, the coordinate z_2^s (that stands for x_2) would correspond to an angle bisector of $\mathbf{v}_2$ and $\mathbf{v}_1^*$. See Figure 4.1 for an illustration.

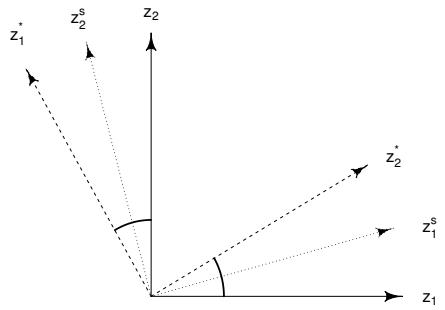


Figure 4.1: Graphical illustration of the symmetric balances.

Particularly, the new symmetric orthonormal coordinate is computed using the respective logcontrast coefficients as

$$z_1^s = \frac{1}{\|\mathbf{v}_1 + \mathbf{v}_2^*\|}(\mathbf{v}_1 + \mathbf{v}_2^*)^\top \ln \mathbf{x}. \tag{4.16}$$

The sum of $\mathbf{v}_1$ and $\mathbf{v}_2^*$ results in a vector with elements

$$\begin{aligned}\mathbf{v}_1 + \mathbf{v}_2^* = \Bigg(& \frac{D-1+\sqrt{D(D-2)}}{\sqrt{D(D-1)}}, -\frac{1}{\sqrt{D(D-1)}}, \\ & -\frac{\sqrt{D-2}+\sqrt{D}}{\sqrt{D(D-1)(D-2)}}, \ldots, -\frac{\sqrt{D-2}+\sqrt{D}}{\sqrt{D(D-1)(D-2)}}\Bigg)^\top,\end{aligned} \tag{4.17}$$

and norm

$$\|\mathbf{v}_1 + \mathbf{v}_2^*\| = \sqrt{\frac{2 \cdot (D-1+\sqrt{D(D-2)})}{D-1}}. \tag{4.18}$$

Subsequently, logcontrast coefficients of the symmetric coordinate z_1^s are given as

$$\frac{\mathbf{v}_1+\mathbf{v}_2^*}{\|\mathbf{v}_1+\mathbf{v}_2^*\|} = \Bigg(\frac{\sqrt{D-1+\sqrt{D(D-2)}}}{\sqrt{2D}}, -\frac{1}{\sqrt{2D(D-1+\sqrt{D(D-2)})}}, \\ -\frac{\sqrt{D-2}+\sqrt{D}}{\sqrt{2D(D-2)(D-1+\sqrt{D(D-2)})}}, \dots, \\ -\frac{\sqrt{D-2}+\sqrt{D}}{\sqrt{2D(D-2)(D-1+\sqrt{D(D-2)})}} \Bigg)^\top, \tag{4.19}$$

followed by the resulting coordinate,

$$z_1^s = \sqrt{\frac{D-1+\sqrt{D(D-2)}}{2D}} \ln \frac{x_1}{x_2^{\frac{1}{D-1+\sqrt{D(D-2)}}} \Big(x_3 x_4 \cdots x_D \Big)^{\frac{\sqrt{D-2}+\sqrt{D}}{\sqrt{D-2}(D-1+\sqrt{D(D-2)})}}}. \tag{4.20}$$

The same procedure is applied to the coordinates z_1^* and z_2, describing information about the compositional part x_2, in order to obtain the second symmetric coordinate z_2^s. Thus

$$z_2^s = \sqrt{\frac{D-1+\sqrt{D(D-2)}}{2D}} \ln \frac{x_2}{x_1^{\frac{1}{D-1+\sqrt{D(D-2)}}} \Big(x_3 x_4 \cdots x_D \Big)^{\frac{\sqrt{D-2}+\sqrt{D}}{\sqrt{D-2}(D-1+\sqrt{D(D-2)})}}}. \tag{4.21}$$

From the above construction it is clear that $z_1^s, z_2^s, z_3, \dots, z_{D-1}$, or alternatively $z_1^s, z_2^s, z_3^*, \dots, z_{D-1}^*$, form orthonormal coordinates of the composition $\mathbf{x}$. The interpretation of the resulting symmetric balances is indeed as expected, they both capture dominance of x_1 and x_2, respectively, with respect to the other components in a symmetric manner. Although the coefficients in the denominator of (4.20) and (4.21) seem to be quite complicated, one does not need to take care about them in practice, because they result just from the normalization needed to achieve orthonormality of the coordinates. More important is weighting of x_2 in z_1^s (and x_1 in z_2^s) that is different for the remaining parts, which reflects the compromise resulting from symmetrizing the input coordinates (4.11) and (4.12). Nevertheless, it is visible that the ratio of both weights, i.e.

$$\frac{\frac{1}{D-1+\sqrt{D(D-2)}}}{\frac{\sqrt{D-2}+\sqrt{D}}{\sqrt{D-2}(D-1+\sqrt{D(D-2)})}} = \frac{\sqrt{D-2}}{\sqrt{D-2}+\sqrt{D}} \tag{4.22}$$

(see Figure 4.2), is stabilized quite soon with an increasing number of parts to approximately one half in favor of the remaining parts.

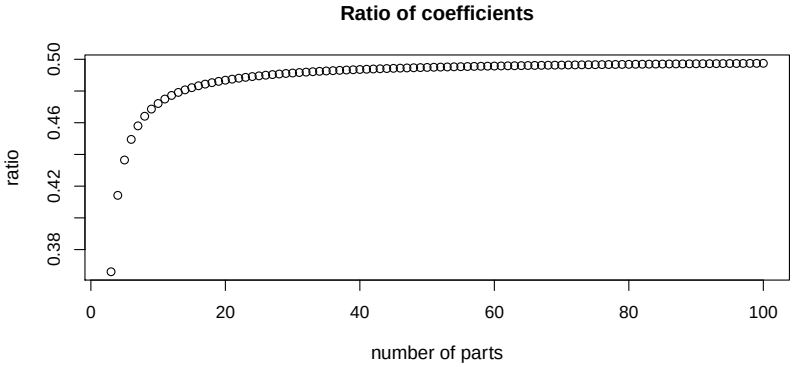


Figure 4.2: Ratio between weights in symmetric balances.

4.4 Correlation analysis with symmetric balances

The symmetric balances, as constructed in the previous section, allow to perform correlation analysis between coordinates which express one part of interest with respect to the other parts in the composition. For this purpose, the Pearson correlation coefficient can be taken,

$$\rho(z_1^s, z_2^s) = \frac{\mathrm{cov}(z_1^s, z_2^s)}{\sqrt{\mathrm{var}(z_1^s)\mathrm{var}(z_2^s)}}, \tag{4.23}$$

or any other alternative correlation measure. The interpretation in the sense of positive and negative association (known from the correlation coefficient) is possible and statistical inference like significance testing can be performed as usual. It is just important to emphasize that it is not a correlation between the original components, but between coordinates assigned to them. Their specific interpretation consists in terms of dominance of both parts to the average behavior of the rest as described in detail above. Hence, the remaining parts can influence the value of the correlation coefficient as well, which fully corresponds to the relative nature of compositional data. Similarly, the correlation for any other pair of parts in $\mathbf{x}$ can be calculated by permuting the parts in (4.20) and (4.21). By summarizing all corresponding correlation coefficients in one matrix, the *compositional correlation matrix* $\mathbf{R}_C(\mathbf{x})$ of dimension $D \times D$ is obtained. It is symmetric with unit diagonal as the standard correlation matrix. Moreover, any scaling and shifting in the compositional sense, i.e. by perturbing $\mathbf{x}$ with a non-random composition $\mathbf{b} = (b_1, \ldots, b_D)^\top$ and powering with a real constant a in order to get a composition $a \odot \mathbf{x} \oplus \mathbf{b} = (x_1^a b_1, \ldots, x_D^a b_D)$ (up to an arbitrary scaling constant, see Pawlowsky-Glahn et al., 2015, for details), yields the same result, $\mathbf{R}_C(a \odot \mathbf{x} \oplus \mathbf{b}) = \mathbf{R}_C(\mathbf{x})$. Although practical experience shows some further interesting properties (like positive definiteness), it is crucial to realize that the elements of $\mathbf{R}_C(\mathbf{x})$ are formed by using $D(D-1)/2$ different coordinate systems, and thus the matrix cannot be processed as a whole, e.g. by computing principal components.

Constructing symmetric balances seems to be the most relevant way how to perform correlation analysis between relative contributions of compositional parts. Nevertheless, the form of the coordinates $\mathbf{z}$ and $\mathbf{z}^*$ inspires to consider also other possibilities that will be briefly mentioned. The first option consists in taking correlation coefficients between the coordinates (4.11) and (4.12), respectively,

$$\rho(z_1, z_2) = \frac{\mathrm{cov}(z_1, z_2)}{\sqrt{\mathrm{var}(z_1)\mathrm{var}(z_2)}}, \qquad \rho(z_1^*, z_2^*) = \frac{\mathrm{cov}(z_1^*, z_2^*)}{\sqrt{\mathrm{var}(z_1^*)\mathrm{var}(z_2^*)}}, \tag{4.24}$$

and then compute their average as follows

$$\rho_{ave}(\mathbf{z}, \mathbf{z}^*) = \frac{\rho(z_1, z_2) + \rho(z_1^*, z_2^*)}{2}. \tag{4.25}$$

Another idea to construct a correlation coefficient with similar interpretation as for the symmetric balances follows the approach from linear discriminant analysis (Johnson and Wichern, 2007) based on calculating the so-called *pooled covariance matrix* from

$$\Sigma_{\mathbf{z}} = \begin{pmatrix} \mathrm{var}(z_1) & \mathrm{cov}(z_1, z_2) \\ \mathrm{cov}(z_2, z_1) & \mathrm{var}(z_2) \end{pmatrix} \quad \text{and} \quad \Sigma_{\mathbf{z}^*} = \begin{pmatrix} \mathrm{var}(z_1^*) & \mathrm{cov}(z_1^*, z_2^*) \\ \mathrm{cov}(z_2^*, z_1^*) & \mathrm{var}(z_2^*) \end{pmatrix}. \tag{4.26}$$

The pooled covariance matrix represents here an average of the covariance matrices $\Sigma_{\mathbf{z}}$ and $\Sigma_{\mathbf{z}^*}$,

$$\Sigma_{\mathrm{p}}(\mathbf{z}, \mathbf{z}^*) = \frac{\Sigma_{\mathbf{z}} + \Sigma_{\mathbf{z}^*}}{2} = \begin{pmatrix} \Sigma_{\mathrm{p}_{11}} & \Sigma_{\mathrm{p}_{12}} \\ \Sigma_{\mathrm{p}_{21}} & \Sigma_{\mathrm{p}_{22}} \end{pmatrix}, \tag{4.27}$$

and the elements are taken to get the resulting correlation coefficient,

$$\rho_{\mathrm{pool}}(\mathbf{z}, \mathbf{z}^*) = \frac{\Sigma_{\mathrm{p}_{12}}}{\sqrt{\Sigma_{\mathrm{p}_{11}}\Sigma_{\mathrm{p}_{22}}}}. \tag{4.28}$$

The next section will be devoted to thorough simulation studies to investigate, whether one would benefit from employing these alternative approaches in addition to the main proposal formed by correlation analysis of symmetric balances.

4.5 Simulation studies

The main aim of the following simulation studies is to investigate the properties of the different correlation coefficients as introduced in the previous section, and to compare also with some other approaches that are used in the literature. In this section we use the data obtained from the moss layer in the Kola Project (Reimann et al., 1998). The data are available in the R-package `mvoutlier` as data set `moss` (R Core Team, 2015), and they contain concentrations of 31 chemical elements in more than 600 soil samples measured in the moss layer.

4.5.1 Simulation 1: Dependence on the number of parts

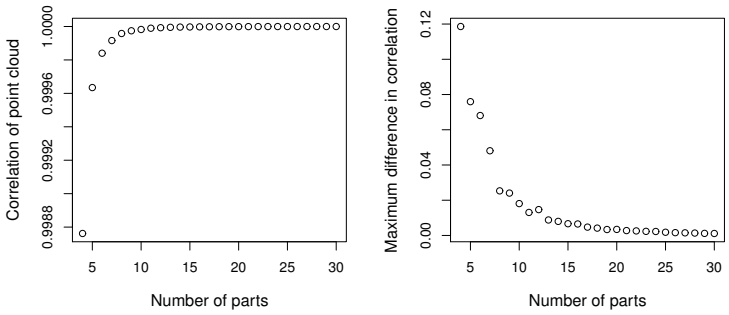
The first simulation setting compares the different approaches for correlation analysis for a varying number of parts involved in the computation of the correlation coefficients. We select randomly k parts ($4 \leq k \leq 30$), and compute the correlation between the first two parts. For each fixed k, the random selection is done 10.000 times, resulting in 10.000 correlation values for each method. When comparing two methods, we compare the outcomes of all results for fixed k in terms of the Pearson correlation. A value close to one would indicate approximately the same outcome of both methods. The left panels in Figure 4.3 show these pairwise comparisons of the different correlation measures, with the considered number of parts on the horizontal axes, and the resulting correlations between the point clouds of the 10.000 outcomes on the vertical axes. The right panels show again pairwise comparisons of correlation measures, but this time the maximum difference of the 10.000 results is computed.

It can be seen that the relation between correlations of symmetric balances and average correlations or correlations based on the pooled covariance matrix, respectively, are very close regardless of the number of parts, with some few exceptions for lower numbers of parts in the compositional data set (Figure 4.3 (a,b)). It is then clear that the difference between correlations for symmetric balances and correlations based on either the coordinates z_1, z_2 or the coordinates z_1^*, z_2^* are also small. On the other hand, the correlations between the coordinates z_1, z_2 and z_1^*, z_2^*, respectively, reveal quite a different behavior (Figure 4.3 (c)). Particularly, for a lower numbers of parts, the correlation coefficients can differ quite substantially, though they converge readily with increasing dimension of the data. Although using simply coordinates (4.5) to capture the relation between compositional parts seems to be attractive (Buccianti et al., 2014), one should be aware that the asymmetry matters. Just for the sake of curiosity, the same simulation has also been done for correlations between symmetric balances and the respective clr variables, with a similar result as before (Figure 4.3 (d)). This fact supports the conclusion that by suppressing alternative correlation measures from the previous section in favor of correlation between symmetric balances, almost no new information is lost. On contrary, it could be quite dangerous to compute correlations from non-symmetric coordinates or negatively-biased clr coefficients. Particularly for lower numbers of parts, the difference between them and the more relevant symmetrized coordinate approach can be substantial.

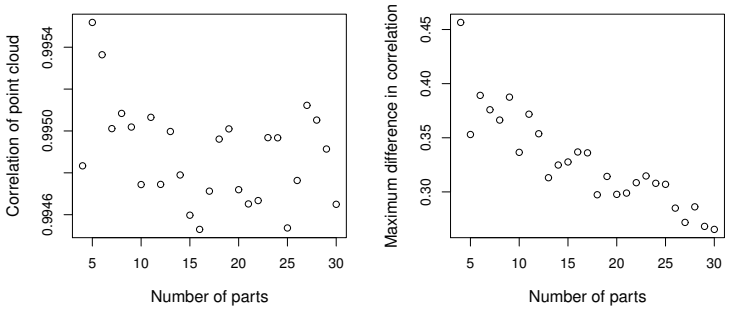
4.5.2 Simulation 2: Permutation tests

Here we consider only the approach of symmetrical balances. It is possible to test the respective correlation coefficient by applying permutation tests for correlations, where the random permutations are drawn without replacement among the observations in the data set. The goal of this simulation study is to investigate the behavior of the correlation coefficient between symmetrical balances $\rho(z_1^s, z_2^s)$ while permuting the observations, and thus to perform hypothesis testing using the permutation test (Good, 2000).

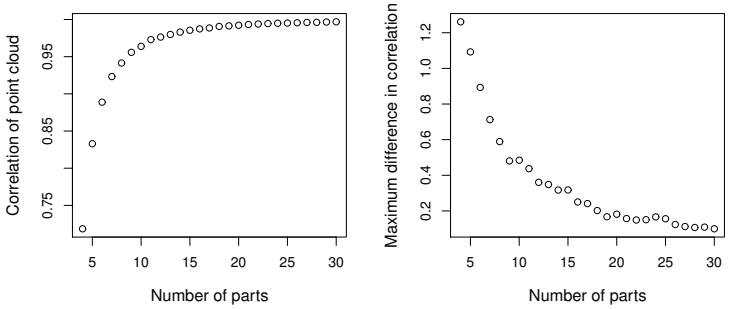
(a) Comparison of correlations between symmetric balances and average correlation coefficient.



(b) Comparison of correlations between symmetric balances and pooled covariance matrix approach.



(c) Comparison of correlations between coordinates $\mathbf{z}$ and $\mathbf{z}^*$.



(d) Comparison of correlations between symmetric balances and clr coefficients.

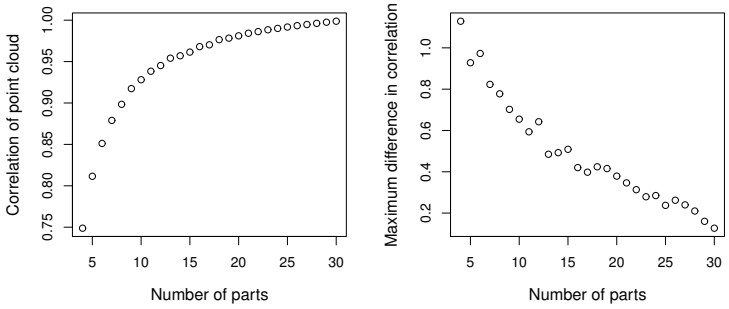


Figure 4.3: Pairwise comparisons of different correlation measures based on 10.000 random selections of data sets with k parts ($4 \leq k \leq 30$); left: Pearson correlations of the resulting point clouds; right: maximum difference between all results.

We can test the null hypothesis about no association between relative contributions to x_1 and x_2, represented by $H_0 : \rho(z_1^s, z_2^s) = 0$, against the alternative hypothesis, $H_1 : \rho(z_1^s, z_2^s) \neq 0$, in three different cases:

(P1) permuting the observations of the second part x_2,

(P2) permuting the observations of both parts x_1 and x_2 independently,

(P3) permuting the observations of the remaining variables $x_3, \ldots, x_D$.

The correlation coefficient r_0 is computed as correlation between our symmetrical balances corresponding to the parts x_1 and x_2, i.e. $r_0 = \hat{\rho}(z_1^s, z_2^s)$. A permutation of the observations according to one of the above schemes (P1)-(P3) leads to modified correlation coefficients $r_i = \hat{\rho}(z_{1(i)}^s, z_{2(i)}^s)$, $i = 1, \ldots, M$, where M stands for the number of permutations. Then we can calculate the resulting p-value as

$$p = \frac{\#(|r_0| < |r_i|, \forall i)}{M}, \qquad i = 1, \ldots, M. \tag{4.29}$$

If $p < 0.05$, the null hypothesis can be rejected at the significance level 0.05 and it can be stated that the correlation between two parts x_1 and x_2 is significant. Note that this procedure is strictly defined for non-compositional variables. In case of compositions we can only expect the same test behavior if the observations in both parts x_1 and x_2 are permuted independently.

The test was performed for all possible combinations of pairs of compositional parts in the data set `moss`, and all schemes (P1)-(P3) of permutation testing were considered. The number of replicates M was set to 1000. The main interest of the tests consists in investigating the behavior of the p-values depending on the used scheme (P1)-(P3), and depending on the size of the original correlation coefficient r_0. The results are shown in Figure 4.4 and 4.5.

Figure 4.4 shows, how the resulting p-values are depending on the original correlation coefficients r_0. The cases, where the null hypothesis is rejected, are depicted in black and non-rejecting situations are in gray. The null hypothesis is not rejected when the original correlation is approximately zero, i.e. relative contributions to x_1 and x_2 are uncorrelated. The relative frequency of non-reject events is then displayed on Figure 4.5. The majority of non-rejects are cumulated around a correlation of zero. The results for the situations where only observations in part x_2 are permuted, and where observations in both x_1 and x_2 are permuted independently are quite the same. A different outcome is observed when only the observations in the remaining parts $x_3, \ldots, x_D$ are permuted. Here it is seen that even for original correlations further away from zero, the test does not report significance. This demonstrates that the correlation between x_1 and x_2, represented by the balances z_1^s and z_2^s, is also heavily depending on the behavior of the remaining parts in the composition. Since the `moss` data set has $D = 31$ parts, the remainder can inherently contain important relative information for the association. As noted above, the

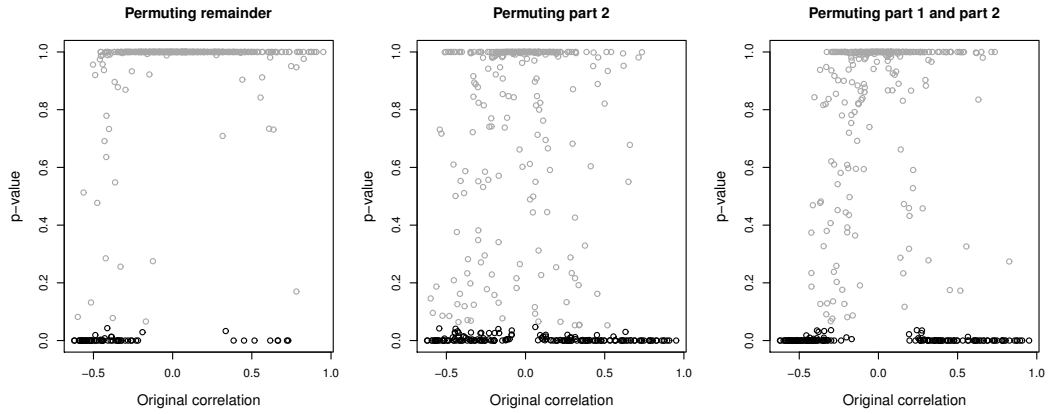


Figure 4.4: p-values of permutation tests for correlations between symmetrical balances, in relation to the original correlation coefficient r_0; left: only observations in the remainder are permuted (P3); middle: observations in part x_2 is permuted (P1); right: observations in parts x_1 and x_2 are permuted independently (P2).

classical permutation test for uncorrelatedness is not properly defined in this situation, but it is still interesting to see the effects. Overall, the permutation tests provide useful information about the importance of the remaining variables when processing correlation analysis for compositional data.

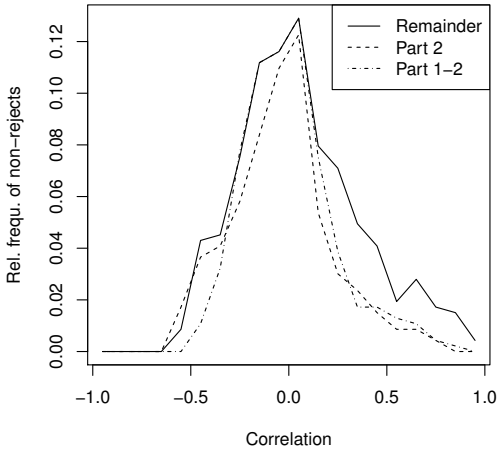


Figure 4.5: Relative frequency of non-rejects in permutation tests for the correlation coefficient of symmetrical balances.

4.6 Example

As in the previous section, we consider the *moss* data set by comparing different association measures. The resulting pairwise correlation coefficients are presented by so-called heat maps, see Figure 4.6, where the values are simply color coded. In addition, the variables are grouped in order to identify patterns in the matrix of pairwise correlations. In Figure 4.6 we compare the heatmaps for associations based on the variation matrix coefficients (upper left), and further correlations for log-transformed data (upper right), for symmetrical balances (lower left), and for clr coordinates (lower right). Due to the individual grouping in each heatmap, the order of the rows and columns changes and makes a direct comparison difficult. However, in this representation one can clearly see the difference in patterns. The variation matrix approach leads to a very different structure due to the non-negative association measures. Also the heatmap for log-transformed data, still very commonly applied in geochemistry, reveal a different structure compared to that for symmetrical balances. In particular, only few negative correlations, but mainly positive ones are obtained. Finally, the heatmaps for symmetrical balances and for clr coordinates are very similar. This is to be expected from the simulation results, see Figure 4.3(d), since for larger numbers of parts the two approaches for computing correlations get very similar.

4.7 Discussion

Correlation analysis of the original compositional parts fails to provide interpretable results, if a fixed constant sum constraint is employed. This is due to the relative nature of compositions represented particularly by scale invariance, and it leads to a negative bias of the correlation structure. The only safe way to perform correlation analysis of compositional data is to express them in orthonormal log-ratio coordinates. Although sequential binary partitioning and the resulting balances can be very useful, when prior knowledge about geochemical processes in the data are available, automated and interpretable orthonormal coordinates that capture relative information about single compositional parts can help to reveal hidden geochemical patterns, when such information is not available.

For the purpose of interpretable correlation analysis in orthonormal log-ratio coordinates, so-called symmetric balances were introduced by using a special choice of balance coordinates. They allow to treat two compositional parts in a symmetric way in one coordinate system and to compute the correlation coefficient. Although the symmetric balances cannot be simply identified with the original compositional parts, because they capture just relative contributions of the parts within a given composition, it seems to be the first successful attempt to have correlation analysis of compositional data interpretable in terms of a pair of compositional parts. Particularly, the possibility of analyzing negative and positive associations as often required in practice (and not available using the variation matrix approach) can help to eliminate inappropriate data processing, for instance using the popular (but scale dependent) log-transformation. Moreover, one should be aware that also other parts are naturally involved into the correlation between

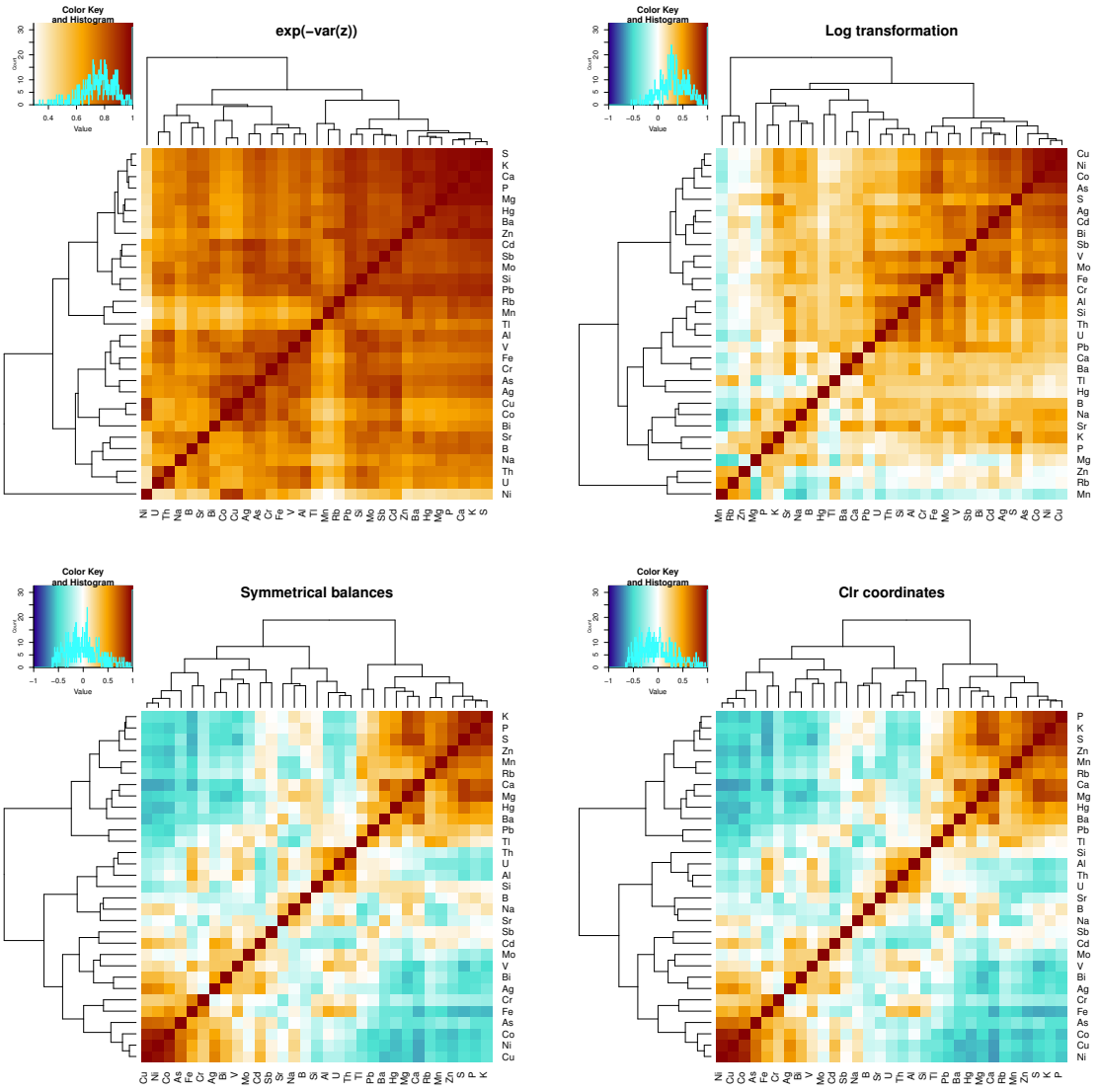


Figure 4.6: Heatmaps of correlations for the moss data set based on the variation matrix coefficients (upper left), log-transformed data (upper right), symmetrical balances (lower left), and clr coordinates (lower right).

two given components by constructing symmetric balances. Nevertheless, it follows closely the definition of compositional data, that none of the parts can be analyzed without considering relations (ratios) to the other parts. This, however, has the consequence that measurement errors in some parts may affect the resulting correlation coefficients of symmetric balances A possible way out seems to be appropriate weighting of the parts according to their relevance, as proposed recently in (Egozcue and Pawlowsky-Glahn, 2015; Filzmoser and Hron, 2015). This will be further investigated in subsequent work.

Correlation coefficients can be seen as summarizing the information of the variable relations shown in scatter plots. With the concept of symmetrical balances we also have an appropriate graphical representation of two compositional parts in terms of orthonormal coordinates. This can serve as a new way of investigating pairwise relations.

Finally, here only the Pearson correlation was used to measure association. Clearly, one can also employ alternative correlation estimators, like the Spearman correlation for identifying non-linear relations, or robust correlation estimators for downweighting the influence of outlying observations.

Acknowledgements

The paper was supported by the grant COST Action CRoNoS IC1408 and the grant IGA_PrF_2015_013 Mathematical Models of the Internal Grant Agency of the Palacký University in Olomouc.

List of Figures

List of Tables

Bibliography

J. Aitchison. Principal component analysis of compositional data. *Biometrika*, 70(1):57–65, 1983. doi: 10.1093/biomet/70.1.57. URL http://biomet.oxfordjournals.org/content/70/1/57.abstract.

J. Aitchison. *The Statistical Analysis of Compositional Data*. Chapman and Hall, London, UK, 1986.

J. Aitchison. The one-hour course in compositional data analysis or compositional data analysis is simple. In V. Pawlowsky-Glahn, editor, *The third annual conference of the International Association for Mathematical Geology – IAMG'97, Proceedings, International Center for Numerical Methods in Engineering (CIMNE)*, 1997.

J. Aitchison and M. Greenacre. Biplots of compositional data. *Applied Statistics*, 51: 375–392, 2002.

C. Barceló-Vidal, L. Aguilar, and J. A. Martín-Fernández. *Compositional VARIMA Time Series*, pages 87–103. John Wiley & Sons, Ltd, 2011. ISBN 9781119976462. doi: 10.1002/9781119976462.ch7. URL http://dx.doi.org/10.1002/9781119976462.ch7.

J. Bergmann. Compositional time series: An application. In J. D. i Estadella and J. A. Martín-Fernandez, editors, *Proceedings of CoDaWork'08, The 3rd Compositional Data Analysis Workshop*, Girona, Spain, 2008.

D. Billheimer, P. Guttorp, and W. F. Fagan. Statistical interpretation of species composition. *Journal of the American Statistical Association*, 96(456):1205–1214, 2001. ISSN 01621459. URL http://www.jstor.org/stable/3085883.

A. Buccianti. Is compositional data analysis a way to see beyond the illusion? *Computers & Geosciences*, 50:165–173, 2013.

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

A. Buccianti, G. Mateu-Figueras, and V. Pawlowsky-Glahn, editors. *Compositional Data Analysis in the Geosciences: From Theory to Practice*. Number 264. Geological Society, special publications edition, 2006.

A. Buccianti, J. Egozcue, and V. Pawlowsky-Glahn. Variation diagrams to statistically model the behavior of geochemical variables: Theory and applications. *Journal of Hydrology*, 519:988–998, 2014.

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

M. Eaton. *Multivariate Statistics. A Vector Space Approach.* John Wiley & Sons, New York, 1983.

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

J. Egozcue and V. Pawlowsky-Glahn. Proceedings of the 6th international workshop on compositional data analysis. In S. Thió-Henestrosa and J. Martín Fernández, editors, *Changing the reference measure in the simplex and its weighting effects*, pages 1–10. University of Girona, Girona, 2015.

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

J. J. Egozcue and V. Pawlowsky-Glahn. Simplicial geometry for compositional data. In A. Buccianti, G. Mateu-Figueras, and V. Pawlowsky-Glahn, editors, *Compositional Data Analysis in the Geosciences: From Theory to Practice*, pages 67–77. Geological Society Publishing House, London, UK, 2006.

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

J. J. Egozcue, D. Lovell, and V. Pawlowsky-Glahn. Testing compositional association. In K. Hron, P. Filzmoser, and M. Templ, editors, *Proceedings of the 5th International Workshop on Compositional Data Analysis*, Vorau, Austria, 2013.

Eurostat, the statistical office of the European Union. Employment by type of disability, sex, age and economic activity, 2013.

P. Filzmoser and K. Hron. Correlation analysis for compositional data. *Mathematical Geosciences*, 41(8):905–919, 2009.

P. Filzmoser and K. Hron. Robustness for compositional data. In C. Becker, R. Fried, and S. Kuhnt, editors, *Robustness and Complex Data Structures*, pages 117–131. Springer, Heidelberg, DE, 2013.

P. Filzmoser and K. Hron. Robust coordinates for compositional data using weighted balances. In K. Nordhausen and S. Taskinen, editors, *Modern Nonparametric, Robust and Multivariate Methods*. Springer, Heidelberg, 2015.

P. Filzmoser, K. Hron, and C. Reimann. Principal component analysis for compositional data with outliers. *Environmetrics*, 20:621–632, 2009a.

P. Filzmoser, K. Hron, and C. Reimann. Univariate statistical analysis of environmental (compositional) data: Problems and possibilities. *Science of the Total Environment*, 407:6100–6108, 2009b.

P. Filzmoser, K. Hron, and C. Reimann. The bivariate statistical analysis of enviromental (compositional) data. *Science of the Total Enviroment*, 408(19):4230–4238, 2010.

P. Filzmoser, K. Hron, and C. Reimann. Interpretation of multivariate outliers for compositional data. *Computers & Geosciences*, 39:77–85, 2012.

E. Fišerová and K. Hron. On interpretation of orthonormal coordinates for compositional data. *Mathematical Geosciences*, 43:455–468, 2011.

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

P. Good. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer-Verlag, New York, 2nd edition, 2000.

K. Hron and P. Filzmoser. Exploring compositional data with the robust compositional biplot. In M. Carpita, E. Brentari, and E. Qannari, editors, *Advances in Latent Variables*, pages 219–226. Springer, Heidelberg, 2014.

J. D. i Estadella, S. Thió-Henestrosa, and G. Mateu-Figueras. Including supplementary elements in a compositional biplot. *Computers & Geosciences*, 37:696–701, 2011.

J. E. Jackson. *A User's Guide to Principal Components*. Wiley & Sons, New York, 1991.

R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, International, 2007. sixth edition.

I. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer New York, 2013. ISBN 9781475719048. URL `https://books.google.at/books?id=-ongBwAAQBAJ`.

J. Larrosa. Compositional time series: Past and present. Econometrics, EconWPA, 2005.

H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, Berlin, 1st ed. 2006. corr. 2nd printing edition, 2007. ISBN 3540262393.

G. Mateu-Figueras and V. Pawlowsky-Glahn. A critical approach to probability laws in geochemistry. *Mathematical Geosciences*, 40(5):489–502, 2008.

G. Mateu-Figueras, V. Pawlowsky-Glahn, and J. J. Egozcue. *The Principle of Working on Coordinates*, pages 29–42. John Wiley & Sons, Ltd, 2011.

T. Mills. Forecasting compositional time series. *Quality & Quantity*, 44(4):673–690, 2010. ISSN 0033-5177. doi: 10.1007/s11135-009-9229-8. URL http://dx.doi.org/10.1007/s11135-009-9229-8.

V. Pawlowsky-Glahn and A. Buccianti. *Compositional Data Analysis: Theory and Applications*. Wiley, Chichester, 2011.

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

V. Pawlowsky-Glahn, J. Egozcue, and D. Lovell. The product space $\mathcal{T}$ (tools for compositional data with a total). In P. F. K. Hron and M. Templ, editors, *Proceedings of CoDaWork'13, The 5th Compositional Data Analysis Workshop*, Vorau, Austria, 2013.

V. Pawlowsky-Glahn, J. J. Egozcue, and D. Lovell. Tools for compositional data with a total. *Statistical Modelling*, 2014. doi: 10.1177/1471082X14535526. URL http://smj.sagepub.com/content/early/2014/11/25/1471082X14535526.abstract.

V. Pawlowsky-Glahn, J. Egozcue, and R. Tolosana-Delgado. *Modeling and Analysis of Compositional Data*. Wiley, Chichester, 2015.

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

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL http://www.R-project.org/.

C. Reimann, M. Äyräs, and V. C. et al. *Environmental Geochemical Atlas of the Central Barents Region*. NGU-GTK-CKE Special publication, Geological Survey of Norway, Trondheim, Norway, 1998. ISBN 9783510652631.

C. Reimann, P. Filzmoser, K. Fabian, K. Hron, M. Birke, A. Demetriades, E. Dinelli, A. Ladenberger, and The GEMAS Project Team. The concept of compositional data analysis in practice – Total major element concentrations in agricultural and grazing land soils of Europe. *Science of the Total Environment*, 426:196–210, 2012.

M. Templ, K. Hron, and P. Filzmoser. *robCompositions: an R-package for robust statistical analysis of compositional data*. John Wiley and Sons, 2011. ISBN 978-0-470-71135-4.

K. G. van den Boogaart, R. Tolosana, and M. Bren. *compositions: Compositional Data Analysis*, 2010. URL http://CRAN.R-project.org/package=compositions. R package version 1.40-1.

Index

Curriculum Vitae

Personal Data

Name:	Petra Kynčlová
Date of birth:	11.03.1988
Place of birth:	Olomouc, Czech Republic
Nationality:	Czech

Education

since 2012	PhD in Technical Mathematics, Vienna University of Technology, Austria
2010 – 2012	Master in Applications of Mathematics in Economy, Palacký University Olomouc, Czech Republic
2007 – 2010	Bachelor in Mathematics-Economics of Insurance Systems, Palacký University Olomouc, Czech Republic
1999 – 2007	Slovanské gymnázium Olomouc, Czech Republic

Work Experience

2012 – 2015	Junior Statistician, StatGISTeam, Faculty of Geoinformatics, Palacký University Olomouc, Czech Republic
Jan – Mar 2015	Junior Statistical Consultant, United Nations Industrial Development Organization (UNIDO), Vienna, Austria

Awards

Oct 2015	Co-winner of the 2015 YSM Data Analysis Competition hosted by UNIDO: Industrial development in least developed countries.