Die approbierte Originalversion dieser Dissertation ist in der Hauptbibliothek der Technischen Universität Wien aufgestellt und zugänglich.
http://www.ub.tuwien.ac.at

The approved original version of this thesis is available at the main library of the Vienna University of Technology.
http://www.ub.tuwien.ac.at/eng

TECHNISCHE
UNIVERSITÄT
WIEN
Vienna | Austria

DISSERTATION

Theoretical and Practical Aspects in 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

Dipl.-Ing. Mehmet Can Mert
Matrikelnummer 0630095

Diese Dissertation haben begutachtet:

Ao. Univ.-Prof. Dipl.-Ing. Dr.techn. Peter Filzmoser

Doc. RNDr. Karel Hron Ph.D.

Wien, 1. September 2016

Dipl.-Ing. Mehmet Can Mert

Erklärung zur Verfassung der Arbeit

Dipl.-Ing. Mehmet Can Mert
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, 1. September 2016

Dipl.-Ing. Mehmet Can Mert

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Prof. Peter Filzmoser for the continuous and endless support and motivation, and for sharing his academic experience. I am grateful for his trust by giving me that opportunity and believing in me that i will succeed. Besides my advisor, i would like to thank Doc. Karel Hron for his insightful comments, but also for theoretical discussions which encouraged me to widen my research from various perspectives. I am thankful to all my colleges in our department and especially to those i shared the office with for the theoretical productive discussions and also for their encouragement when i encountered difficulties in my research.

Above all, i want to thank especially my family and my friends for supporting me in every situation and difficult moments, and also for sharing the happy moments.

Abstract

Compositional data represent the relative information between variables that are parts of some whole. The relevant information is contained only in the ratios between the measured variables, and not in the absolute values. A common procedure how to analyze this relative information is to use the so-called log-ratio approach, proposed by John Aitchison in the 1980s. From a geometrical point of view, the compositions live in the simplex sample space, and the log-ratio approach enables a representation in terms of coordinates in the usual Euclidean geometry. The well known coordinates are the additive log-ratio (alr), the centered log-ratio, and the isometric log-ratio (ilr) coordinates. The clr and ilr coordinates are preferred, since the ilr representation constructs orthonormal coordinates and the clr representation allows for an interpretation in terms of the original variables.

We focus on different aspects of compositional data: One field of interest are high-dimensional compositional data, where the interpretation of the resulting coordinates can become a complex task. Another concern is the propagation of measurement errors in the construction of the orthonormal coordinates. Applications in geochemistry, but also in epidemiology, which is a new field for this kind of analysis, underline the usefulness of this approach.

Kurzfassung

Kompositionsdaten stellen die relative Information zwischen den Variablen, die Teile eines Ganzen sind, dar. Die relevante Information is nicht in den absoluten Werten enthalten, sondern in den Verhältnissen der kompositionellen Variablen. Diese relative Information kann mit dem in den 1980er Jahren vorgeschlagenen Log-Ratio Ansatz analysiert werden.

Aus geometrischer Sicht befinden sich die Kompositionsdaten im Simplex, einer Teilmenge des euklidischen Raums. Der Log-Ratio Ansatz representiert die Daten in Form von Koordinaten in der üblichen euklidischen Geometrie. Die bekannten Koordinaten sind die additive Log-Ratio (alr), die centered Log-Ratio (clr), und die Isometric Log-Ratio Koordinaten (ilr). Die clr und ilr Koordinaten werden bevorzugt, da die ilr Darstellung orthonormale Koordinaten erstellt und die clr Darstellung eine Interpretation in Bezug auf die ursprünglichen Variablen ermöglicht.

Diese Dissertation legt den Schwerpuntk auf die verschiedenen Aspekte der Kompositionsdaten: Ein Interessensgebiet sind die hochdimensionalen Kompositionsdaten, wobei die Interpretation der extrahierten Koordinaten eine komplexe Aufgabe sein kann. Ein weiteres Anliegen ist die Ausbreitung von Messfehlern bei der Darstellung der Kompositionen in Form von orthonormalen Koordinaten. Anwendungen in der Geochemie sowie in der Epidemiologie, ein neues Feld für diese Art von Analyse, heben die Brauchbarkeit des Log-Ratio Ansatzes hervor.

Contents

Abstract **vii**

Kurzfassung **ix**

Contents **xi**

1 Introduction **1**
- 1.1 Introduction and History . . . 1
- 1.2 Basic Concepts . . . 2
- 1.3 Principles of Compositional Data . . . 3
- 1.4 Geometric Structure of the Simplex . . . 4
- 1.5 Log-ratio Methodology . . . 5
- 1.6 Balances . . . 7
- 1.7 Irregularities in Compositional Data . . . 11
- 1.8 Exploring the Compositional Data . . . 14
- 1.9 Robustness in Compositional Data Analysis . . . 18
- 1.10 Robust PCA and Biplots of Compositional Data . . . 21
- 1.11 Outline . . . 22

2 Sparse Principal Balances **25**
- 2.1 Introduction . . . 25
- 2.2 Existing algorithms for constructing principal balances . . . 27
- 2.3 A new algorithm for constructing sparse principal balances . . . 29
- 2.4 Simulations . . . 31
- 2.5 Application to metabolomics data . . . 34
- 2.6 Conclusions . . . 38

3 Error Propagation in Isometric Log-ratio Coordinates for Compositional Data: Theoretical and Practical Considerations **41**
- 3.1 Introduction . . . 41
- 3.2 Error Propagation in the Standard Euclidean Geometry . . . 45
- 3.3 Error Propagation on the Simplex . . . 45
- 3.4 Simulation-Based Investigations of Error Propagation . . . 48
- 3.5 Discussion and Conclusions . . . 58

4 Compositional Data Analysis in Epidemiology 63
4.1 Introduction 63
4.2 The log-ratio approach 65
4.3 Analyzing single epidemiological variables 67
4.4 Joint analysis of epidemiological variables 70
4.5 Regression modeling 76
4.6 Summary and conclusions 79

List of Figures 83

List of Tables 85

Bibliography 87

Index 93

Curriculum Vitae 95

CHAPTER 1

Introduction

1.1 Introduction and History

Compositional data contain relative information consisting of non-negative values which are parts of some whole. Such data are usually called compositions and presented in the form of percentages, concentrations, or frequencies. The relevant information lies in the ratios between the parts of the compositions. However, a common approach of the practitioners is to consider such data as absolute values and apply standard statistical methods which can lead to biased results. To illustrate this we consider a dataset, the number of premature births in Austria, collected from 1984 to 2014, covering 30 years, and published by Statistics Austria (sta, 2016). Depending on how premature the birth is, the data are categorized into 4 groups: late preterm (34 to ≤36 weeks), moderate preterm (32th or 33th week), very preterm (28 to 32 weeks), and extremely preterm (less than 28 weeks). The data are normalized by dividing each number of premature births by the total size in a country to obtain percentages for each preterm group resulting in so-called closed data. The relation between the premature birth groups is analyzed by the Pearson correlation coefficients between the groups. First, the correlation matrix of the full composition is analyzed. Next, the group late preterm is omitted from the original data and normalized again yielding a subcomposition. The correlation matrices of the full composition and the subcomposition are given in Tables 1.1 and 1.2.

	extreme preterm	very preterm	moderate preterm	late preterm
extreme preterm	1.00	0.40	0.43	-0.73
very preterm	0.40	1.00	0.41	-0.74
moderate preterm	0.43	0.41	1.00	-0.86
late preterm	-0.73	-0.74	-0.86	1.00

Table 1.1: Pearson correlations between the proportions of premature births in Austria: four part composition: extremely, very, moderate, late preterm.

	extreme preterm	very preterm	moderate preterm
extreme preterm	1.00	-0.38	-0.50
very preterm	-0.38	1.00	-0.61
moderate preterm	-0.50	-0.61	1.00

Table 1.2: Pearson correlations between the proportions of premature births in Austria: three part composition: extremely, very, moderate preterm.

Table 1.1 shows that all groups except late preterm correlate positively in the full composition. Late preterm correlates negatively with all other groups. In Table 1.2, we see that when we analyze the subcomposition (where the group late preterm has been omitted), the relationships between the three groups change drastically. In his paper (Pearson, 1897), Karl Pearson deals with this phenomenon and coins a term for it: *spurious correlation.* Pearson draws attention to the improperness of applying standard statistical methods on closed data such as proportions. At the beginning of the 1960s, the problem with the constant sum constraint is analyzed further by Felix Chayes. He argues that the correlation is negatively biased when standard methods are applied to data with constant sum constraint (Chayes, 1960). This led to a greater awareness of the issues and problems associated with closed data, especially among geologists and biologists.

The concept of log-ratio transformations introduced by Aitchison in the 1980s was the beginning of a new perspective on how to analyze such data. He proposed the concept of log-ratio transformations for compositions by arguing that compositions provide relative information between the parts rather than absolute information. The resulting one-to-one mapping onto the real space made it possible to represent the compositions in an unconstrained space and allowed the use of standard methods.

1.2 Basic Concepts

A composition defines a vector $\boldsymbol{x} = (x_1, \ldots, x_D)$ with D parts, where all values are strictly positive and the relevant information is given by the ratios between those parts (Pawlowsky-Glahn et al., 2015). Consequently, we can say that the absolute values in a composition are non-informative. A composition can exist in the form of closed data such as proportions or percentages with a constant sum constraint. A composition can also be expressed in units such as ppm (parts per million) or ppb (parts per billion) that occur in geology or chemometrics. Data with concentration units such as mg/L or meg/L can be a composition, where no constant sum has to be given for the observations. One can even linearly switch between different units; however, the ratios between the parts remain the same. For the purpose of interpretability and visualization a composition can be expressed as closed data with the help of the closure operator $\mathcal{C}$. It is a rescaling of the composition so that its parts sum up to a chosen constant value κ,

$$\mathcal{C}(\boldsymbol{x}) = \left(\frac{\kappa x_1}{\sum_{j=1}^{D} x_j}, \cdots, \frac{\kappa x_D}{\sum_{j=1}^{D} x_j} \right). \tag{1.1}$$

The sample space of compositional data is a subset of the usual Euclidean space, called simplex, with the so called Aitchison geometry (Aitchison, 1986). For a composition $\boldsymbol{x} = (x_1, \cdots, x_D)$ with D parts, the simplex sample space is defined as

$$\mathcal{S}^D = \{\boldsymbol{x} = (x_1, \ldots, x_D) \text{ such that } x_j > 0 \ \forall j, \sum_{j=1}^{D} x_j = \kappa\},$$

where κ is an arbitrary constant.

In practice, a full composition may not be always available, or the focus can be only on a specific subset of the composition. For example, in epidemiology the cause of deaths by diseases can be considered as a composition, and the epidemiologist can be interested only in the chronic diseases. In such a case, a subcomposition should be analyzed rather than the full composition. A subcomposition can be obtained by selecting a subset of the parts of a full composition.

1.3 Principles of Compositional Data

Aitchison (1986) proposed three important principles that compositional data analysis should fulfill. Those principles are scale invariance, the equivalence class of compositions, subcompositional coherence, the independence of the results from a chosen subset of the composition, and permutation invariance, the independence of the results from the different permutations of the parts.

1.3.1 Scale Invariance

Compositional data analysis considers the ratios between the parts as the relevant information. Consequently, the statistical inference remains the same independent of how or whether the composition is scaled. For instance, recalling the example dealing with premature births, one might express the data in form of percentages or rates per 10000 births, as it is common in epidemiology. For scaling purposes, the *closure* operator (see Equation (1.1)) can be used. Considering the premature births in 2014, the composition with four preterm groups can be given as absolute frequencies (320, 698, 875, 4587), as probabilities (0.049, 0.108, 0.135, 0.708) or as rates per 10000 births (490, 1080, 1350, 7080). Considering the ratios between the corresponding parts, all three expressions yield the same ratios.

1.3.2 Subcompositional Coherence

Subcompositional coherence states that the conclusions made for the common parts of two subcompositions must agree with each other. In literature, two common criteria stand for the coherence: The first one is that the scale invariance principle holds for any possible subcomposition, and the second one is that the distances between two full compositions are greater or equal than the distances between their subcompositions, known as subcompositional dominance (Pawlowsky-Glahn and Buccianti, 2011; Pawlowsky-Glahn et al., 2015).

1.3.3 Permutation Invariance

This principle emphasizes that the results of the compositional data analysis should not depend on the order of the compositional parts. For instance, the parts of a composition can be arranged alphabetically or even according to their importance in ascending or descending order. However, the composition remains the same, and due to that the ratios between the parts do not change, no matter which component is ordered first and which component is ordered last.

1.4 Geometric Structure of the Simplex

A D-dimensional simplex is a $(D-1)$-dimensional subset of the real space $\mathbb{R}^D$. However, the simplex does not possess the geometrical properties of the usual Euclidean geometry. Hence, the traditional vector operations such as addition or multiplication do not work properly, and the Euclidean distance, which most of the statistical analysis relies on, do not have a proper meaning in the simplex. Therefore, a proper geometry with a vector space structure is needed to be established to work with compositional data in simplex. This geometry is known as Aitchison geometry and defined by special operations of perturbation, power transformation and Aitchison inner product:

The Aitchison geometry forms a vector space under two operations: The first one is the perturbation operator, which corresponds to addition in Euclidean geometry, and the second one is the powering operator, which corresponds to multiplication in Euclidean geometry. Let us consider two D-part compositions, $\boldsymbol{x} = (x_1, \ldots, x_D)$ and $\boldsymbol{y} = (y_1, \ldots, y_D)$. The perturbation operator is defined as

$$\boldsymbol{x} \oplus \boldsymbol{y} = \mathcal{C}(x_1 y_1, x_2 y_2, \cdots, x_D y_D).$$

Powering by a scalar $\alpha \in \mathbb{R}$ is defined as

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

Using the perturbation operator, the neutral element is defined as $\boldsymbol{n} = \mathcal{C}(1, 1, \ldots, 1) = \left(\frac{1}{D}, \frac{1}{D}, \ldots, \frac{1}{D}\right)$ satisfying $\boldsymbol{x} \oplus \boldsymbol{n} = \boldsymbol{x}$. Note that $\boldsymbol{n}$ is unique and is called the barycenter of the simplex. Using the powering operator, the opposite element of a composition $\boldsymbol{x}$,

$\boldsymbol{x}^{-1}$, can be defined as $\boldsymbol{x}^{-1} = \mathcal{C}[x_1^{-1}, x_2^{-1}, \ldots, x_D^{-1}]$. It is clear that the perturbation of $\boldsymbol{x}$ and $\boldsymbol{x}^{-1}$ results in the neutral element $\boldsymbol{n}$.

Furthermore, a Euclidean vector space structure can be constructed on the simplex, $(S^D, \oplus, \odot)$, by defining the Aitchison inner product, the Aitchison norm and the Aitchison distance based on the logarithmic ratios between the compositional parts (Pawlowsky-Glahn et al., 2015). The Aitchison inner product provides geometrical benefits such as the projection of the compositions onto specific directions and the identification of angles between the compositional vectors. For $\boldsymbol{x}$ and $\boldsymbol{y}$, the inner product is defined as

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_A = \frac{1}{2D} \sum_{j=1}^{D} \sum_{k=1}^{D} \ln \frac{x_j}{x_k} \ln \frac{y_j}{y_k}.$$

The Aitchison norm, providing the length of a composition, and the Aitchison distance between compositions are

$$||\boldsymbol{x}||_A = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_A}, \quad d_A(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} \oplus (-1) \odot \boldsymbol{y}||_A,$$

respectively.

1.5 Log-ratio Methodology

The Euclidean space is the geometry that we work on and we are familiar with. The algebraic structure allows us to define a vector $\boldsymbol{x}^D = (x_1, \ldots, x_D) \in \mathbb{R}^D$ as linear combination of any chosen orthonormal basis $(\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_D)$ as follows,

$$\boldsymbol{x} = x_1 \boldsymbol{e}_1 + x_2 \boldsymbol{e}_2 + \cdots + x_D \boldsymbol{e}_D, \tag{1.2}$$

where $\boldsymbol{e}_i = [0, 0, \ldots, 1, 0, \ldots, 0]$ and 1 is the i-th component of the vector. Let us now assume that $\boldsymbol{x} \in S^D$ is a composition defined in the simplex. By intuition we would assume that the composition $\boldsymbol{x}$ can be expressed by an orthonormal basis as above. However, the algebraic-geometric structure of the simplex is different from the Euclidean space and does not possess the same characteristics. As a matter of fact, the sum and product operators are not a closed operation and Equation (1.2) is not a linear combination anymore (Pawlowsky-Glahn and Buccianti, 2011). Hence, even the linear combination of the coefficients $(x_1, \ldots, x_D)$ can result in negative or zero values which are not defined in the simplex. Therefore, a need emerges to define a generating system $(\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_D)$ of the simplex within the Aitchison geometry, where $\boldsymbol{w}_i = (1, \ldots, e, \ldots, 1)$ and e is the i-th component of $\boldsymbol{w}_i$. Using the perturbation and powering operators ($\oplus$, $\odot$), the composition $\boldsymbol{x}$ can be expressed as follows:

$$\boldsymbol{x} = (\ln x_1 \odot \boldsymbol{w}_1) \oplus (\ln x_2 \odot \boldsymbol{w}_2) \oplus \cdots \oplus (\ln x_D \odot \boldsymbol{w}_D) \tag{1.3}$$

Perturbation and powering operators include closure and are scale invariant. Thus, adding any constant or scaling by any constant does not alter the composition. Dividing the coefficients $(x_1, \dots, x_D)$ by the geometric mean of $\boldsymbol{x}$, denoted by $g(\boldsymbol{x}) = \sqrt[D]{\prod_{j=1}^{D} x_j}$, results in the equivalent representation:

$$\boldsymbol{x} = \left(\ln \frac{x_1}{g(\boldsymbol{x})} \odot \boldsymbol{w}_1\right) \oplus \left(\ln \frac{x_2}{g(\boldsymbol{x})} \odot \boldsymbol{w}_2\right) \oplus \cdots \oplus \left(\ln \frac{x_D}{g(\boldsymbol{x})} \odot \boldsymbol{w}_D\right). \tag{1.4}$$

Equation (1.4) defines a generating system for the simplex. The resulting coefficients define the centered log-ratio (clr) transformation (Aitchison, 1986), a one-to-one mapping from S^D to $\mathbb{R}^D$:

$$clr(\boldsymbol{x}) = \boldsymbol{y} = (y_1, \dots, y_D) = \left(\ln \frac{x_1}{g(\boldsymbol{x})}, \dots, \ln \frac{x_D}{g(\boldsymbol{x})}\right). \tag{1.5}$$

The clr transformation divides each component symmetrically by the geometric mean of all the components, which simplifies the interpretation of the components. Therefore, the original variables of the composition can be represented directly by their corresponding clr coefficients. However, the clr transformation has some disadvantageous properties. Due to the zero sum constraint of the coefficients, $\sum_{j=1}^{D} y_j = 0$, the transformation results in a singularity. As a consequence, some statistical methods such as robust procedures can have numerical problems when applied on these coefficients.

As an alternative to clr coefficients, a new basis can be extracted from the generating system shown in Equation (1.2) by taking any $D-1$ vectors from $(\boldsymbol{w}_1, \dots, \boldsymbol{w}_D)$. In this way, the composition $\boldsymbol{x}$ can be expressed as follows:

$$\boldsymbol{x} = \left(\ln \frac{x_1}{x_j} \odot \boldsymbol{w}_1\right) \oplus \cdots \oplus \left(\ln \frac{x_{j-1}}{x_j} \odot \boldsymbol{w}_{j-1}\right) \oplus \left(\ln \frac{x_{j+1}}{x_j} \odot \boldsymbol{w}_{j+1}\right) \oplus \cdots \oplus \left(\ln \frac{x_D}{x_j} \odot \boldsymbol{w}_D\right) \tag{1.6}$$

The index $j \in 1, \dots, D$ refers to the variable chosen as a denominator. The resulting coefficients of Equation (1.6) correspond to the additive-log-ratio (alr) transformation (Aitchison, 1986), a mapping from S^D to $\mathbb{R}^{D-1}$,

$$alr(\boldsymbol{x}) = \left(\ln \frac{x_1}{x_j}, \dots, \ln \frac{x_{j-1}}{x_j}, \ln \frac{x_{j+1}}{x_j}, \dots, \ln \frac{x_D}{x_j}\right). \tag{1.7}$$

However, the corresponding basis $(\boldsymbol{w}_1, \dots, \boldsymbol{w}_{j-1}, \boldsymbol{w}_{j+1}, \dots, \boldsymbol{w}_D)$ of the alr transformation is not orthogonal, which results in a non-isometric transformation (Egozcue and Pawlowsky-Glahn, 2006). Furthermore, the alr transformation is not symmetrical as the results vary according to the denominator chosen.

The disadvantages of alr and clr transformations are solved by the isometric log-ratio (ilr) transformation introduced by Egozcue et al. (2003b). The fundamental idea of the ilr transformation is to choose an orthonormal bases in the simplex. The simplex possesses the Euclidean space structure which enables to generate an orthonormal basis by employing the Gram-Schmidt orthonormalisation process to any basis (Pawlowsky-Glahn and Buccianti, 2011). Note that infinitely many orthonormal bases can be obtained by this procedure. Therefore, one must choose a basis that is relevant for a given task. A particular basis is proposed by Egozcue et al. (2003b),

$$\boldsymbol{v}_j = \sqrt{\frac{j}{j+1}}\left(\frac{1}{j}, \ldots, \frac{1}{j}, -1, 0, \ldots, 0\right)' \text{ for } j = 1, \ldots, D-1, \tag{1.8}$$

resulting in ilr coordinates

$$\boldsymbol{z} = (z_1, \ldots, z_{D-1}) \text{ with } z_j = \sqrt{\frac{j}{j+1}} \ln \frac{\sqrt[j]{\prod_{k=1}^{j} x_k}}{x_{j+1}} \text{ for } j = 1, \ldots, D-1. \tag{1.9}$$

Since the ilr coordinates are related to the orthonormal basis on the hyperplane $\mathcal{H}: y_1 + \cdots + y_D = 0$ in $\mathbb{R}^D$, formed by the clr transformation, there is a clear relationship between the clr coefficients and ilr coordinates,

$$\boldsymbol{y} = \boldsymbol{V}\boldsymbol{z}$$

where $\boldsymbol{V}$ is a $D \times (D-1)$ matrix consisting of the columns $\boldsymbol{v}_j$ in Equation (1.8), defining the vectors of orthonormal basis on hyperplane $\mathcal{H}$ (Filzmoser et al., 2009b). A specific choice of orthonormal coordinates is suggested by Fišerová and Hron (2011).

$$z_j = ilr_j(\boldsymbol{x}) = \sqrt{\frac{D-j}{D-j+1}} \ln \left(\frac{x_j}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_k}} \right), j = 1, \ldots, D-1, \tag{1.10}$$

where the first ilr coordinate z_1 explains all the relative information concerning the corresponding original part x_1 of the considered composition. The coordinates $z_2, \ldots, z_{D-1}$ explain the log-ratios of the remaining parts in the composition and it is not straightforward to interprete them. Considering the Equations (1.5) and (1.10) we see that the first clr coefficient and the first ilr coordinates are proportional. This proportion can be shown by the following linear relationship:

$$y_1 = \sqrt{\frac{D-1}{D}} z_1.$$

Moreover, the relative information of part x_k to the remaining parts $(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_D)$ can be easily extracted by putting part x_k on the first position and constructing the ilr coordinates based on the rearranged order (Fišerová and Hron, 2011).

1.6 Balances

Working with the ilr coordinates based on an orthonormal basis has various geometric advantages. However, in most cases it is challenging to interpret the ilr coordinates easily, especially with a high number of parts in a composition. To circumvent this problem, balances can be used. They are coordinates of an orthonormal basis expressing the relative information between groups of parts in a composition. Hence, specific orthonormal coordinates can be chosen to represent specific information according to the problem to be analyzed.

In order to construct orthonormal balances, a procedure called sequential binary partition (SBP) can be applied to a composition (Egozcue and Pawlowsky-Glahn, 2005). SBP is an iterative procedure which generates a hierarchy of the parts of a composition. In the first step, two groups of parts are defined, usually based on expert knowledge. In the following steps, the groups generated are split into smaller groups until a single part remains in each group. Note that the generated groups refer to the subcompositions. Let us assume that the i-th step of SBP results in two subcompositions $\boldsymbol{x}_+$ with r parts and $\boldsymbol{x}_-$ with s parts of a D-part composition $\boldsymbol{x}$, where $i < D - 1$ and $r + s \leq D$. A balance between $\boldsymbol{x}_+$ and $\boldsymbol{x}_-$ can be expressed as a log-ratio of the geometric mean of the two subcompositions (Egozcue and Pawlowsky-Glahn, 2005):

$$b_i = \sqrt{\frac{rs}{r+s}} \ln \frac{g_m(\boldsymbol{x}_+)}{g_m(\boldsymbol{x}_-)}. \tag{1.11}$$

The balancing element of such a balance is then given as $\mathcal{C}(a_1, a_2, \ldots, a_D) \in S^D$, where $\mathcal{C}$ is the closure operator and a_i gets the value $\exp\left(\frac{1}{r}\sqrt{\frac{rs}{r+s}}\right)$ for the part i in $\boldsymbol{x}_+$, and $\exp\left(-\frac{1}{s}\sqrt{\frac{rs}{r+s}}\right)$ for the part i in $\boldsymbol{x}_-$ (Pawlowsky-Glahn et al., 2011).

1.6.1 Example: Childbirth data

An illustration of SBP can be shown by using a dataset, expressing the number of live childbirths in the countries of the European Union by the mother's age. The corresponding data matrix has 28 observations (countries), and 7 parts describing the age groups of mother: age_{10-19}, age_{20-24}, age_{25-29}, age_{30-34}, age_{35-39}, age_{40-44}, $\text{age}_{>45}$. The data are shown in detail in Table 1.3. The idea of the balances is that they are easy to interpret. Therefore, in most cases a-priori information is needed to split the parts into groups, so that the generated balance represents the sought information.

Here, the first balance (b_1) is chosen as the relative information between the live childbirths at normal age phase and early-late age phase of mother at birth. This balance can be used to identify the countries with higher rate of childbirths at the early-late phase compared to the normal phase. The first balance is given by the quantitative information expressed in the following equation,

$$b_1 = \sqrt{\frac{4*3}{4+3}} \ln \frac{g_m(\text{age}_{10-19}, \text{age}_{20-24}, \text{age}_{40-44}, \text{age}_{>45})}{g_m(\text{age}_{25-29}, \text{age}_{30-34}, \text{age}_{35-39})}.$$

The sign of the balance indicates which country is dominated by which age group and the quantity of the balance shows the intensity of that dominance. A second balance (b_2) can be considered by taking the parts in the previous group with the negative sign and

country	age$_{10-19}$	age$_{20-24}$	age$_{25-29}$	age$_{30-34}$	age$_{35-39}$	age$_{40-44}$	age$_{>45}$
BE	2130.00	15225	41977	42177	18189	3908	236.00
BG	6655.00	14454	20770	16204	7978	1401	116.00
CZ	2734.00	13339	32643	38620	19449	2935	140.00
DK	632.00	6254	18155	19414	10192	2117	106.00
DE	15467.00	80364	202696	253567	133254	27484	1574.00
EE	460.00	2156	4510	3803	2084	514	24.00
IE	1226.00	5954	13123	24659	18188	3887	248.00
GR	2272.00	7905	21682	34800	20541	4347	602.00
ES	8552.00	30752	77856	153787	124271	28819	2039.00
FR	19520.00	109500	258919	267351	129480	32602	1956.00
HR	1222.00	5750	12376	13087	5912	1161	53.00
IT	7819.00	46029	112818	167806	128358	36654	3112.00
CY	128.00	906	2932	3477	1438	316	52.00
LV	866.00	3993	7357	5754	3044	705	26.00
LT	1158.00	5430	10762	8476	3703	796	38.00
LU	92.00	558	1490	2228	1419	257	21.00
HU	6040.00	13308	23444	28846	18177	3268	101.00
MT	152.00	565	1265	1408	689	105	7.00
NL	1796.00	16731	54087	66460	30118	5706	283.00
AT	1686.00	11305	24498	27622	13424	2996	191.00
PL	13287.00	60053	129245	117017	46891	8321	346.00
PT	2491.00	8772	19040	28645	19156	4034	229.00
RO	19381.00	39176	61486	46319	22422	4092	227.00
SI	233.00	2256	7151	7554	3413	529	29.00
SK	3470.00	8938	16544	16906	7891	1242	42.00
FI	1101.00	8259	17158	19013	9659	1897	145.00
SE	1300.00	14554	35425	38676	20003	4663	286.00
GB	29244.00	125377	219258	241237	128610	29950	2172.00

Table 1.3: Number of live births according to age groups of mother at birth in the countries of the European Union.

representing the ratio of the childbirths between the early and late phase,

$$b_2 = \sqrt{\frac{2*2}{2+2}} \ln \frac{g_m(\text{age}_{10-19}, \text{age}_{20-24})}{g_m(\text{age}_{40-44}, \text{age}_{>45})}.$$

	age_{10-19}	age_{20-24}	age_{25-29}	age_{30-34}	age_{35-39}	age_{40-44}	$\text{age}_{>45}$	r	s
1	+1	+1	-1	-1	-1	+1	+1	4	3
2	+1	+1	0	0	0	-1	-1	2	2
3	+1	-1	0	0	0	0	0	1	1
4	0	0	0	0	0	+1	-1	1	1
5	0	0	+1	-1	-1	0	0	1	2
6	0	0	0	+1	-1	0	0	1	1

Table 1.4: An example of sequential binary partition to construct an orthonormal basis based on the live birth data.

The constructed balances are depicted in a scatter plot in Figure 1.1. The horizontal axis represents the first balance (b_1), while the second balance (b_2) is shown on the vertical axis. The points are labeled by abbreviations of the European countries. In countries like Romania (RO), Bulgaria (BG), Slovakia (SK), and Hungary (HU) the rate of childbirths at early phase is higher than in other countries compared to the late phase. On the other hand, in southern countries such as Greece (GR), Spain (ES), and Italy (IT) the early-late age phase at childbirth is more dominant than the normal age phase at childbirth compared to other countries. By considering the first balance, it is not possible to make any statement about the second balance since both reveal completely different information as they are geometrically orthogonal. This is an important fact to consider about the concept of balances.

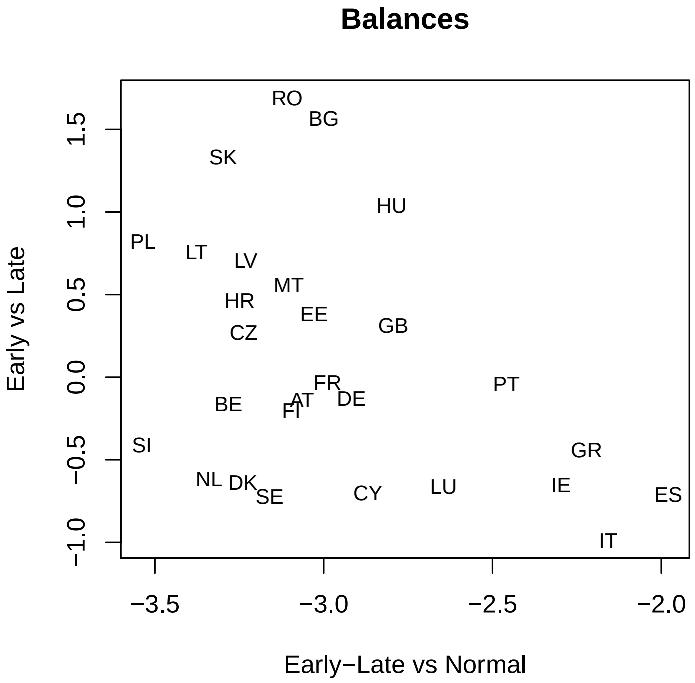

Figure 1.1: Scatter plot of the constructed balances based on the live birth data.

An important drawback of balances is that the extracted balances do not represent the maximum information unconditionally as the primary focus is on the interpretation aspect. That issue is treated with the concept of principal balances, balances representing the maximum information of the data. Regarding that three algorithms are suggested by (Pawlowsky-Glahn et al., 2011). Moreover, those algorithms work for low dimensions. For this reason, a new tool called sparse principal balance is proposed in Chapter 2 to build an interpretable orthonormal basis for high-dimensional data with a minimum loss of information.

1.7 Irregularities in Compositional Data

In contrast to the simulated data, issues with data quality can occur when the focus is on real data. Especially for compositional data, zero values and missing values cause problems when applying the log-ratio methodology since the logarithm of a zero or missing value is not defined in Euclidean geometry. From a geometrical point of view, the closure operator would not work properly when a part presents zero or missing values as the total sum is either biased or can not be calculated. Having such irregular values in most of the observations of a component could even lead us to disregard that component which means that we would then work with a subcomposition.

In the fields of geology, chemometrics, and metabolomics, compositional data are gathered by analyzing samples with measurement tools. In some cases, the values are so small that those measurement tools can not detect the value. Therefore, the value is rounded to zero or defined as a missing value. Usually, such cases are treated with imputation methods or by substituting that value with a small value, such as 2/3 of its detection limit, the smallest value that can be measured. Zeros can also occur conditionally due to the specific structure of a population or when a specific factor is analyzed as a compositional component that distinguishes two or more populations (Aitchison, 1986). For instance, the proportion of alcohol in non-alcoholic cocktails or proportion of causes of deaths from a tropic disease in a country in which nobody was infected from that disease can be given as examples. A zero value can also report an absence of a part rather than a data irregularity, which is known by the term *counting zero problem* (Pawlowsky-Glahn et al., 2015). Such cases usually occur when the composition is deduced from count data. This type of zeros can strongly associate with the considered sample size of the data. For example, considering the number of deaths caused by diseases in a rural area compared to an urban area: The counts of the diseases in a rural area will not be fully equivalent to the counts of the same disease in an urban area, as the number of the population may be higher in the urban area than in the rural area.

Different strategies have been developed to overcome the zero problem in compositional data. One popular and practical approach is to replace the zero values with a *convenient value*, for example a smaller positive value less than the detection limit (Martín-Fernández et al., 2003). The replacement can be a fixed value or a random value. The latter case is more appropriate when the amount of the values to be replaced is high in order to retain the variability. For a small number of zeros the former case can be more expedient (Pawlowsky-Glahn et al., 2015). Some important replacement procedures are additive replacement (Aitchison, 1986), multiplicative replacement (Martín-Fernández et al., 2003), parametric models (Palarea-Albaladejo et al., 2007; Palarea-Albaladejo and Martín-Fernández, 2008), Box-Cox-based replacement (Greenacre, 2010, 2011a,b) and robust procedures (Hron et al., 2010).

Next, we illustrate an imputation example using the childbirth dataset. We mimic the count zero problem by artificially defining zero values for five randomly selected countries (Germany, Ireland, United Kingdom, Portugal, Romania) of the last age group $\text{age}_{>45}$. To impute the zero values we use the k-means imputation method proposed in Hron et al. (2010). This method is implemented in the R-package 'robCompositions' (Templ et al., 2016) and can be carried out to the childbirth dataset, *birth*, with the following R-function.

```r
birthImp <- impKNNa(birth)$xImp
```

The function uses the Aitchison distances to identify the k-nearest neighbors to an observation with missing parts or zero values. For robustness aspects, the median is used for the estimation of the k-nearest neighbors.

The output object *birthImp* includes the following information, which can be viewed easily:

```
names(birthImp)
##  [1] "xOrig" "xImp" "criteria" "iter" "w"
##  "wind" "metric"
```

The original data is stored in *xOrig*, the imputed data in *xImp*, *w* includes the number of missing values, and indices of the imputed values are listed in the element *wind*. Based on this information, one can create summaries and diagnostic plots to validate the imputation.

The first ilr coordinate of the imputed data *birthImp* is illustrated in Figure 1.2. The imputed observations are marked by green color, while the original values that were artificially defined as ”zero values” are marked by red color.

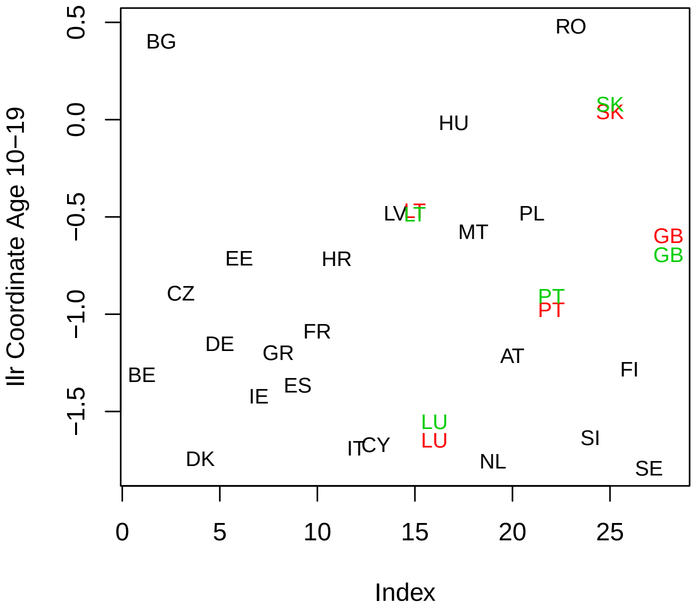

Figure 1.2: First ilr-coordinate of *birth* data. The observations marked by red color were defined as zero values. Observations after imputation are marked by green color.

1.8 Exploring the Compositional Data

Exploratory data analysis is usually the first step to get an initial overview about the data that will be subjected to further analysis. It helps data analysts to understand the data and to look for specific problems and structures in the data. Exploratory data analysis also includes descriptive statistics defining the center and dispersion of the data, and visualization tools such as scatter plots and biplots, which can be used in univariate and multivariate manner to discover patterns and data irregularities, such as errors and outliers.

1.8.1 Center and Dispersion

For data defined in the Euclidean space, calculating standard descriptive statistics such as arithmetic mean and variance or standard deviation is straightforward. However, when analyzing compositional data the nature of the Aitchison geometry needs to be taken into account. Consider a compositional data set with n observations and D parts, represented in the $n \times D$ data matrix $\boldsymbol{X}$, with the matrix elements x_{ij}. The center of $\boldsymbol{X}$ can be estimated by the closed geometric mean of the individual compositional parts as

$$cen(\boldsymbol{X}) = \mathcal{C}\left(g_1, \ldots, g_D\right), \tag{1.12}$$

where $g_j = \left(\prod_{i=1}^{n} x_{ij}\right)^{1/n}$ for $j = 1, \ldots, D$. In the simplex, the center of a composition can be interpreted as the multivariate mean of the composition (Pawlowsky-Glahn and Buccianti, 2011). Alternatively, the center can be estimated by expressing the compositions in ilr coordinates, see Equation (1.10),

$$z_{ij} = \sqrt{\frac{D-j}{D-j+1}} \ln \left(\frac{x_{ij}}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_{ik}}} \right), \tag{1.13}$$

for $j = 1, \ldots, D-1$ and $i = 1, \ldots, n$. Now the arithmetic mean can be used to estimate the center,

$$\overline{\boldsymbol{z}} = \left(\overline{z}_1, \ldots, \overline{z}_{D-1}\right)', \text{ with } \overline{z}_j = \frac{1}{n} \sum_{i=1}^{n} z_{ij}, \text{ for } j = 1, \ldots, D-1.$$

Finally, $\overline{\boldsymbol{z}}$ needs to be back-transformed to the simplex by the inverse ilr transformation. The inverse ilr transformation $\boldsymbol{x} = ilr^{-1}(\boldsymbol{z})$ is given as

$$
\begin{aligned}
x_1 &= exp\left(\sqrt{\frac{D-1}{D}}z_1\right), \\
x_j &= exp\left(-\sum_{k=1}^{j-1}\frac{1}{\sqrt{(D-k+1)(D-k)}}z_k + \sqrt{\frac{D-j}{D-j+1}}z_j\right), j = 2, \dots, D-1 \\
x_D &= exp\left(-\sum_{j=1}^{D-1}\frac{1}{\sqrt{(D-j+1)(D-j)}}z_j\right).
\end{aligned}
$$

The dispersion between two parts in a compositional data set can be expressed by the variance of log-ratios of those parts. The matrix form of log-ratios of all compositional parts is known as the variation matrix (Aitchison, 1986) and is denoted by $\mathbf{T}$.

$$
\mathbf{T} = \begin{bmatrix} t_{11} & x_{12} & t_{13} & \dots & t_{1D} \\ t_{21} & x_{22} & t_{23} & \dots & t_{2D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_{D1} & t_{D2} & t_{D3} & \dots & t_{DD} \end{bmatrix}
$$

where $t_{jk} = \mathrm{Var}\left[\ln\frac{x_j}{x_k}\right]$ for $j \in \{1, \dots, D\}$ and $k \in \{1, \dots, D\}$. ”Var” stands for the variance, and for a concrete compositional sample, the empirical variance can be used as an estimator. Using the variation matrix, a total variance which defines the global dispersion of a compositional data set can be expressed as the sum of variances of all pairwise log-ratios,

$$
\mathrm{totalvar}[\boldsymbol{X}] = \frac{1}{D}\sum_{j=1}^{D-1}\sum_{k=j+1}^{D} t_{jk}. \tag{1.14}
$$

Alternatively, the total variance can be decomposed into the variance of the ilr coordinates of the composition as follows,

$$
\mathrm{totalvar}[\boldsymbol{X}] = \sum_{j=1}^{D-1} Var[z_{1j}, \dots, z_{nj}]. \tag{1.15}
$$

This expression reveals, which coordinates have a high contribution to the variance, in other words which coordinates carry more information, as the ilr coordinates are orthogonal to each other. Typically, the variability and center of a composition can be presented by a variation array (Aitchison, 1986).

Table 1.5 shows these descriptive statistics for the childbirth data introduced previously. The sample mean of the pairwise log-ratios (numerator by column, denominator by row) is given in the lower triangle of Table 1.5. Positive values refer to the dominance of the part in the numerator. For instance, the number of births in the age group age_{30-34} is clearly

		numerator						
		age_{10-19}	age_{20-24}	age_{25-29}	age_{30-34}	age_{35-39}	age_{40-44}	$\text{age}_{>45}$
denominator	age_{10-19}		0.21	0.39	0.53	0.59	0.70	1.06
	age_{20-24}	-1.59		0.04	0.16	0.25	0.32	0.63
	age_{25-29}	-2.41	-0.82		0.06	0.16	0.22	0.48
	age_{30-34}	-2.54	-0.94	-0.12		0.03	0.08	0.27
	age_{35-39}	-1.89	-0.30	0.52	0.64		0.03	0.19
	age_{40-44}	-0.29	1.30	2.12	2.24	1.60		0.14
	$\text{age}_{>45}$	2.52	4.11	4.93	5.06	4.41	2.81	

Table 1.5: Variation array of childbirth data. The upper triangle shows the variance of the log-ratios, while the lower triangle includes the log-ratio means.

dominating the other age categories on average, especially the groups $\text{age}_{10-19}$and $\text{age}_{>45}$. On the other hand, the age group age_{10-19} is dominated by all other groups on average, except by $\text{age}_{>45}$. The variability of the pairwise log-ratios across the countries can be observed in the upper triangle of Table 1.5. Higher values indicate higher variability of the log-ratios. For instance, the variability of the births of age group $\text{age}_{>45}$ is clearly higher compared to the other age groups. On the other hand the small variance of the log-ratio of age group age_{30-34} over age_{35-39} (0.03) indicates proportional birth rates of both age groups across the countries.

1.8.2 Visualization of Compositional Data

Ternary Diagram

A popular visualization tool for a composition is the ternary diagram, presenting a three part (sub-)composition in a two-dimensional scatter plot. It is an equilateral triangle with vertices $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$. A composition with three parts located in the neighborhood of a vertex indicates a high proportion of the part represented by that vertex. The center of the ternary diagram is called the *barycenter*, with equal proportions to all vertices. Such a representation of compositional data is often used by geoscientists and petrologists.

An example of a ternary diagram is shown in Figure 1.3 for the childbirth data. Here, an amalgamated version of the data is considered by summarizing the age groups into three main groups: age_{10-19}, age_{20-24} as early, age_{25-29}, age_{30-34}, age_{35-39} as normal, age_{40-44}, $\text{age}_{>45}$ as late.

Almost all European countries yield a high proportion of birth in the normal age phase of the mother. Romania (RO) and Bulgaria (BG) deviate from other countries with higher rate of births at early age. When a composition is dominated by one part (in our case the composition is dominated by the age group ”normal”), it is not straightforward to have a detailed idea about the proportions of the other groups. One possibility is rescaling the observations of the data before visualization. This can be done by shifting the center of the composition to the barycenter of the simplex. The idea is to apply the perturbation operator on each observation by the inverse of its geometric center. The centered version of the ternary diagram is shown on the right panel of Figure 1.3. Now the two countries Italy (IT) and Spain (ES) in the direction of age group ”late” are more visible.

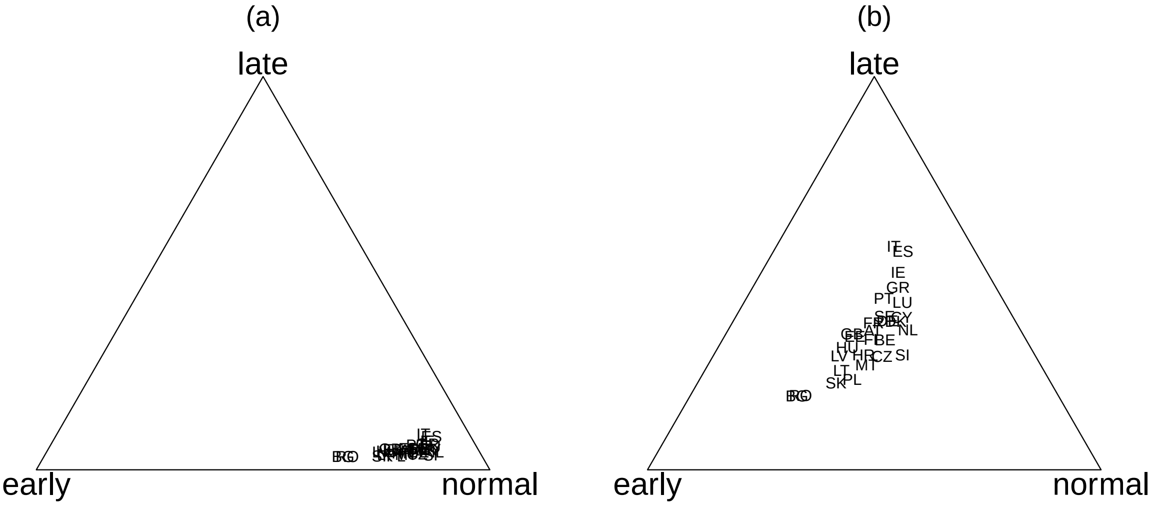

Figure 1.3: Childbirth data before (a) and after (b) centering.

The ternary diagram can be constructed by using the *ternaryDiag()* function of the R-package *robCompositions*, where $\boldsymbol{x}$ represents the amalgamated 3-part composition.

```
ternaryDiag(x)
## Centered Ternary Diagramm
d.amg.center <- acomp(d.amg) + acomp(1/apply(d.amg, 2, gm))
ternaryDiag(d.amg.center)
```

Absolute Versus Relative Information

The usage of ternary diagrams can be restrictive, depending on which compositional information needs to be visualized. The focus of the visualization can be a multivariate task or even plotting an individual part of the composition. In such cases, special attention should be devoted to choose the form of the information to be visualized. Working with the absolute values, instead of the ratios of the components, would not incorporate the rest of the information in the composition. For instance, the composition could be dominated by a specific part. In such a case, considering the absolute information of a part with small values may lead the practitioner to a conclusion that the analyzed part is of less importance in the data, which must not be the case unconditionally.

The above discussed aspects are now demonstrated by analyzing the spatial distribution of amalgamated childbirth data with the groups early, normal and late age phase of the mother, as used in the example for the ternary diagrams. Figure 1.4 shows the regional distribution of the proportion of childbirths at normal age of mother expressed as absolute information (left) and as relative information (right), given by the corresponding clr coordinate for the variable normal age phase, on the European map. In both maps, a scale according to the quantiles 0.05, 0.25, 0.50, 0.75 and 0.95 of the distribution is used. The absolute and relative information reveal different regional patterns. Considering the absolute information, the countries Netherlands, Denmark, Slovenia followed by Spain, Greece, and Ireland reveal the highest proportion of childbirths at normal age phase. A different picture is provided by taking the clr coordinate, where the dominance of normal age phase is shifted more to central European countries. Here, the dominance refers to a higher relative contribution of the normal age group.

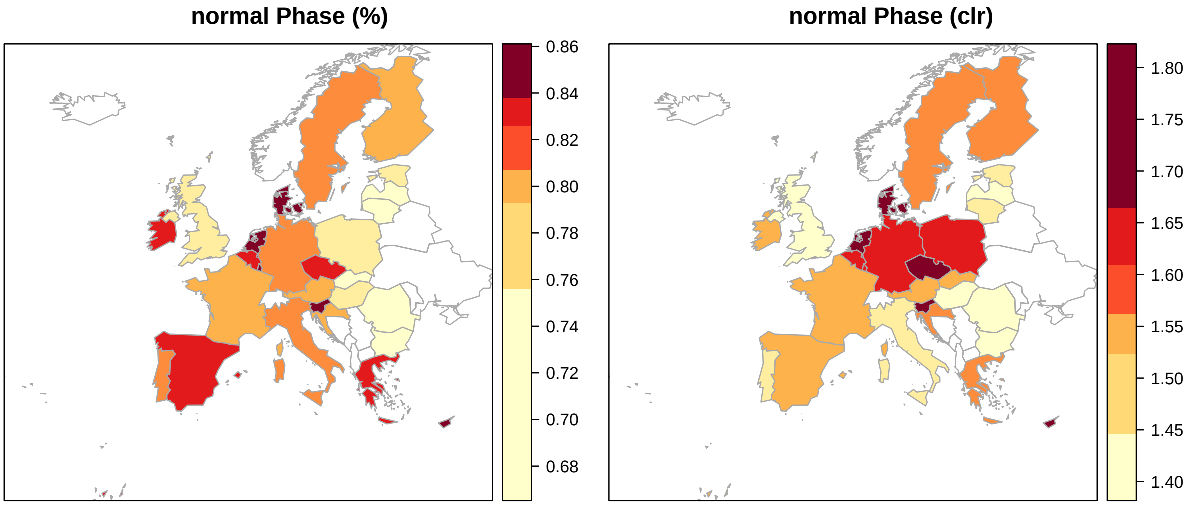

Figure 1.4: The proportion of childbirths at normal age of mother expressed as absolute information (left) and relative information in form of the clr coordinate (right). The breaks of the color scale are defined by the quantiles 0.05, 0.25, 0.50, 0.75 and 0.95, yielding six classes.

1.9 Robustness in Compositional Data Analysis

One of the important tasks of exploratory data analysis is to investigate the irregularities in the data. The irregular observations can act as outliers yielding distorted results and biased models, especially in the context of multivariate data analysis. In such cases, robust methods are more appropriate to analyze the data. This also holds for the case of compositional data.

In order to apply the robust procedures on compositional data, the sample space must be a proper one. For that purpose, the compositional data should be represented in the usual Euclidean space. Among the known log-ratio methods, the alr transformation is not isometric and the clr transformation results in singularity. The proposal of an isometric transformation from the simplex to the Euclidean space allows to use robust methods and avoids numerical problems (Filzmoser and Hron, 2008; Filzmoser et al., 2009b; Hron et al., 2010; Filzmoser et al., 2012). The Mahalanobis distance is commonly used to investigate the outlyingness of an observation in a multivariate data set. A compositional data matrix $\boldsymbol{X}$ with D parts and n observations can be represented in ilr coordinates by using the Equation (1.13), resulting in a matrix $\boldsymbol{Z}$ with $D-1$ coordinates and n observations. Now the Mahalanobis distance for each observation $\boldsymbol{z}_i$, for $i = 1, \ldots, n$, can be given as follows

$$\mathrm{MD}(\boldsymbol{z}_i) = \sqrt{(\boldsymbol{z}_i - T(\boldsymbol{Z}))' \, C(\boldsymbol{Z})^{-1} \, (\boldsymbol{z}_i - T(\boldsymbol{Z}))}, \tag{1.16}$$

where T stands for the location estimator and $\mathcal{C}$ for the covariance estimator, respectively. The usual, non-robust, way of estimating location and covariance of a data matrix is to use the arithmetic mean and sample covariance matrix. In order to downweight the effects of outliers on those sensible estimators, robust counterparts, such as Minimum Covariance Determinant (MCD) (Rousseeuw and Van Driessen, 1999) or MM-estimator (Yohai, 1987) can be used. The MCD estimator seeks h observations whose empirical covariance matrix has the smallest possible determinant. The location and covariance are then estimated from those observations. The MCD estimator reaches the maximum breakdown point, namely 50%, when $h \approx \frac{n}{2}$ observations are chosen. Besides having the maximum breakdown point, MCD is an affine equivariant estimator which is an important property for robust multivariate methods for compositional data (Pawlowsky-Glahn and Buccianti, 2011). The estimators T and C are affine equivariant if they satisfy the conditions

$$T(\boldsymbol{A}\boldsymbol{z}_1 + \boldsymbol{b}, \ldots, \boldsymbol{A}\boldsymbol{z}_n + \boldsymbol{b}) = \boldsymbol{A}T(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) + \boldsymbol{b}$$

$$C(\boldsymbol{A}\boldsymbol{z}_1 + \boldsymbol{b}, \ldots, \boldsymbol{A}\boldsymbol{z}_n + \boldsymbol{b}) = \boldsymbol{A}C(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)\boldsymbol{A}',$$

where $\boldsymbol{A}$ is a $(D-1) \times (D-1)$ non-singular matrix and $\boldsymbol{b} \in \mathbb{R}^{D-1}$ is a shift vector. In case of multivariate normal distribution, the squared Mahalanobis distances follow approximately a χ^2-distribution. Based on that, a cutoff value can be determined to identify the outliers. Usually, the 0.975 quantile of the χ^2-distribution with $D-1$ degrees of freedom is used. Observations with squared Mahalanobis distances larger than $\chi^2_{0.975,D-1}$ are identified as outliers.

An example is shown by using the childbirth data. Figures 1.1 and 1.3 have already revealed some certain structures in the data. The focus is on a comparision of the robust Mahalanobis distances, calculated from the raw data given in proportions as absolute information, and the ilr coordinates as relative information. For this purpose, the robust covariance and location estimators of both datasets are calculated by using the *covMcd* function implemented in R-package *robustbase* (version 0.92-5). Since each row in the raw data matrix sums up to 1, the robust covariance matrix and location estimations of the raw data encounter singularity problems. Therefore, the estimations are given as a numerical approximation by the function. In contrast, the ilr coordinates do not suffer from the singularity problem.

The calculated Mahalanobis distances are illustrated in Figure 1.5. The horizontal axis represents the robust Mahalanobis distance of the raw data given in proportions. The Mahalanobis distances of the ilr coordinates are shown on the vertical axis. The horizontal and vertical lines represent the cutoff values for both datasets, 0.975 quantiles of the χ^2-distributions with corresponding degrees of freedom. None of the observations are detected as outliers based on the Mahalanobis distances of the raw data. On the other hand, some potential outliers are identified by considering the ilr coordinates. Those are the countries Bulgaria (BG), Malta (MT), Greece (GR), Romania (RO), Czech Republic (CZ), Slovenia (SI), Cyprus (CY), and Finland (FI).

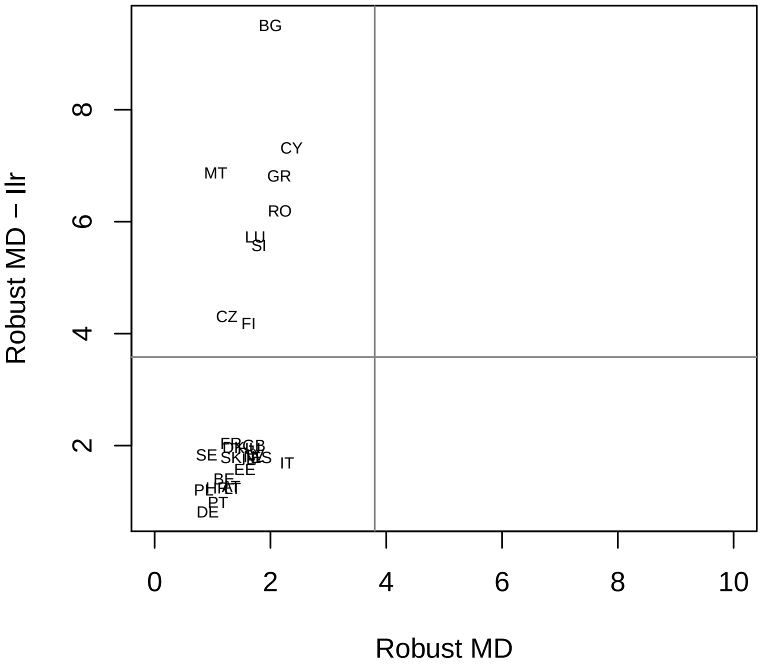

Figure 1.5: Multivariate outlier detection of childbirth data given in proportions. The horizontal axis represents the robust Mahalanobis distances of the raw data. The vertical axis represents the robust Mahalanobis distances based on the ilr coordinates.

1.10 Robust PCA and Biplots of Compositional Data

Principal component analysis (PCA) is a statistical procedure developed by Hotelling (1933) that aims at dimension reduction. This is achieved by representing the data as linear combinations of the usually correlated original variables. These linear combinations are called principal components (PCs) and they define a projection into a lower dimensional space. It uses an orthonogonal transformation with an objective function of minimum information loss (Jolliffe, 2014). The minimum information loss is guaranteed by extracting a first component with maximum variance among all linear combinations of variables, then a second component with the maximum of the remaining variation, and so on. The different components are supposed to be orthogonal to each other. The resulting PCs can be used for visualization, or to detect outliers or clusters. PCA is one of the most important methods in multivariate statistics and applied almost in every field.

PCA is also a useful tool for compositional data since in fields of chemometrics, proteomics or genomics, the data occur in high dimensions with hundreds or even thousands of variables. In such cases, PCA can be applied to achieve a reduction of dimension. However, as mentioned earlier, the geometry of the compositional data is not appropriate for PCA, as PCA is defined for the Euclidean geometry. The ilr transformation can be applied to represent the data in coordinates.

Consider a $n \times D$ compositional data matrix $\boldsymbol{X}$ with n compositions and D parts, and $\boldsymbol{Z}$ is the matrix of the corresponding ilr coordinates of $\boldsymbol{X}$ calculated by Equation (1.13). The PCs can be obtained by eigenvalue decomposition of the covariance matrix $C(\boldsymbol{Z})$ of ilr-transformed data or singular value decomposition of $\boldsymbol{Z}$. The decomposition of $C(\boldsymbol{Z})$ in terms of eigenvectors and eigenvalues can be expressed as $C(\boldsymbol{Z}) = \boldsymbol{\Gamma}\boldsymbol{A}\boldsymbol{\Gamma}^\top$. The eigenvectors of $C(\boldsymbol{Z})$ are represented by the orthogonal $(D-1) \times (D-1)$ matrix $\boldsymbol{\Gamma}$ and the eigenvalues of $C(\boldsymbol{Z})$ by the diagonal elements of $\boldsymbol{A}$. The PCA transformation can be defined as follows,

$$\boldsymbol{Z}^* = (\boldsymbol{Z} - \mathbf{1}T(\boldsymbol{Z})^\top)\boldsymbol{\Gamma}. \tag{1.17}$$

as stated in Filzmoser et al. (2009b). Here, $\mathbf{1}$ is a vector with n ones and $\boldsymbol{Z}^*$ is called the score matrix with colums $i = 1, \ldots, D-1$ denoting the scores of the i-th PC. In classical PCA, $T(\boldsymbol{Z})$ and $C(\boldsymbol{Z})$ are the sample mean and sample covariance matrix of $\boldsymbol{Z}$, which are sensitive to outliers. The existence of outliers can cause a distortion of the covariance structure and consequently distorted directions of principal components. In such a case, a robust estimation of $T(\boldsymbol{Z})$ and $C(\boldsymbol{Z})$ should be preferred, like the MCD or MM-estimator (Maronna et al., 2006). However, the resulting robust scores of the PCs and loadings are not interpretable in terms of the original compositional parts, as they are based on the ilr coordinates. That can be circumvented by back-transformation of the results to the clr coordinates where all parts are symmetrically divided by the geometric mean. The back-transformation is given as

$$\boldsymbol{Y}^* = \boldsymbol{Z}^*\boldsymbol{V}^\top. \tag{1.18}$$

The matrix $\boldsymbol{V} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_{D-1})$ is defined in Equation (1.8). The robust loading matrix of the clr variables can be obtained using the same relation as above.

$$C(\boldsymbol{Y}) = C(\boldsymbol{Z}\boldsymbol{V}^\top) = \boldsymbol{V}C(\boldsymbol{Z})\boldsymbol{V}^\top = \boldsymbol{V}\boldsymbol{\Gamma}^*\boldsymbol{A}\boldsymbol{\Gamma}^{*\top}\boldsymbol{V}^\top$$

Due to the affine equivariance property of the robust estimators the new eigenvectors $\boldsymbol{\Gamma}^*$of $C(\boldsymbol{Y})$ are equal to $\boldsymbol{V}\boldsymbol{\Gamma}$ and the eigenvalues of $C(\boldsymbol{Y})$ remain the same as for $C(\boldsymbol{Z})$.

Consequently, the transformed loadings and scores lead to a *compositional biplot* providing a two-dimensional visualization of a rank two approximation of the observations and variables of the composition. In the biplot representation, the variables (clr coefficients) are represented by rays and the observations by points. The distances between vertices of the rays (links) describe the variability of the log-ratios of the involved parts. The more the vertices lay apart from each other, the higher is the variability of the log-ratio of those parts. The variability of a clr coefficient is represented by the length of its ray. Nevertheless, a large ray does not indicate a large variability of the corresponding original part unconditionally as the variability of that log-ratio also depends on the full composition due to the geometric mean.

In the following, an example is given by using the childbirth data. A compositional biplot is constructed by applying a robust PCA on ilr-transformed data. For comparative purposes, also a standard biplot is constructed by considering the childbirth data as raw data. The resulting biplots are shown in Figure 1.6. The compositional (left) and standard (right) biplots yield at first sight similar results. On the standard biplot, Bulgaria (BG) and Romania (RO) are placed in direction of part age_{10-19}, indicating childbirths at very early age. Central European countries are located between age_{25-29} and age_{30-34}. Slovakia (SK) is located with respect to part age_{20-24}, referring to a high proportion of that part in the raw data.

Considering the compositional biplot, Slovakia (SK) is now dominated by the corresponding clr variable of part age_{10-19}, indicating a high log-ratio of that part to all other parts. This holds also for the countries Bulgaria (BG), Romania (RO), and Hungary (HU). The northern European countries Denmark (DK), Sweden (SE) and Netherlands (NL) are located close to the ray of the age group age_{30-34}, indicating a dominance of part age_{30-34}. The compositional biplot shows more heterogeneous groups of countries than the standard biplot. The southern countries are located on the upper-right side of the biplot, the northern countries on the lower side and the eastern countries on the left side of the biplot. The central European countries are located in the center of the biplot.

1.11 Outline

This thesis treats topics that provide insight into theoretical and practical concepts of compositional data analysis. Specific issues related to compositional data are analyzed and methods are developed. With practical examples using real data and simulation data, the advantages and benefits of the proposed methods are shown. The first chapter

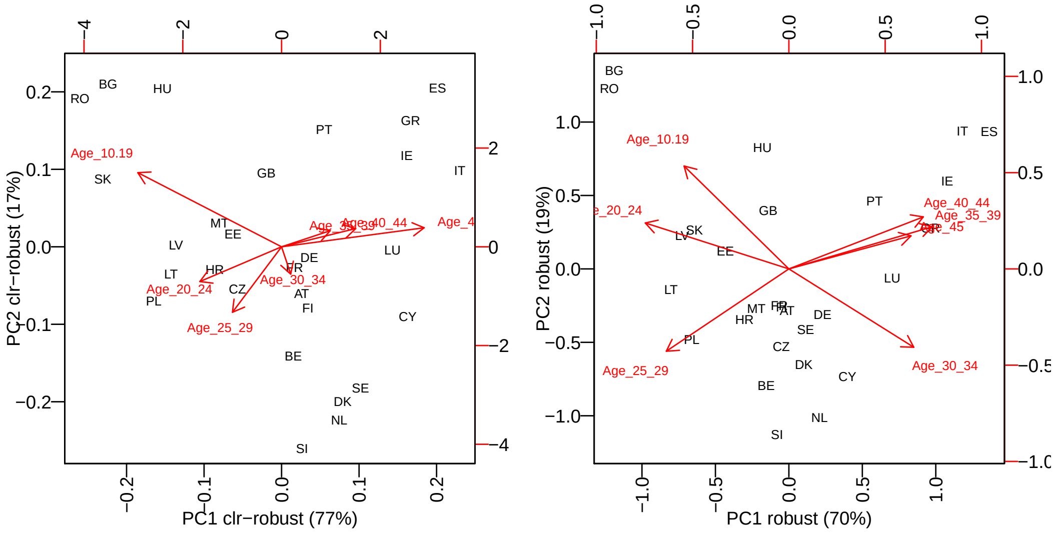

Figure 1.6: Compositional (left) and standard (right) biplot of the first two robust PCs based on the childbirth data.

introduces the historical background and the theoretical developments of compositional data analysis, followed by univariate and multivariate practical applications emphasizing the advantageous properties over standard statistical methods. The second chapter introduces a new approach related to constructing principal balances for high-dimensional compositional data including the concept of sparsity for dimension reduction. The third chapter is devoted to the problem of how the representation of compositional data in Euclidean space is affected by measurement errors and detection limit problems. The last chapter concludes with a discussion of differences in approaches of compositional data analysis and standard statistical methods in epidemiology. It is shown that the compositional data analysis approach leads to new and interesting insights.

Chapter 2 proposes a new method for constructing balances for high dimensional compositional data where the new balances allow for interpretability of the new coordinates. Pawlowsky-Glahn and Buccianti (2011) proposed an approach called *principal balances* that works for compositional data with low dimensions. The method uses the concept of sparsity to perform a dimension reduction and increase interpretability with involving only the relevant parts. The resulting directions, *sparse principal balances*, build an orthonormal basis in simplex and are easy to interpret.

Mert, C.M., Filzmoser, P., Hron, K. (2014). Sparse principal balances. *Statistical Modeling*, pp. 173–176.

Chapter 3 investigates the effect of measurement errors or detection limit problems on the representation of the compositions in terms of ilr coordinates in Euclidean space. The concept of error propagation is used to analyze the problem theoretically. Through simulation experiments, certain types of contamination is applied to real data sets and the behavior of the errors after the ilr transformation is analyzed. Recommendations on the amount of the error and the expected distortion of the results are provided for practitioners dealing with such problems.

Mert, C.M., Filzmoser, P., Hron, K. (2016). Error propagation in isometric log-ratio coordinates for compositional data: theoretical and practical considerations. *Mathematical Geosciences*, pp. 1–21.

Chapter 4 discusses the role of compositional data analysis in epidemiology. The differences of compositional data analysis, representing the relative information in terms of coordinates, and the standard case, analyzing the absolute information, for epidemiological data are outlined. Compositional data analysis considers the multivariate information by incorporating other available information than the analyzed variable. In contrast, the data from epidemiology are usually considered as absolute information and therefore the variables are isolated from each other. Univariate and multivariate analyses are applied on real health care data sets considering relative and absolute information. It is demonstrated that compositional data analysis leads to new and interesting insights when used for epidemiological data.

Mert, C.M., Filzmoser, P., Endel, G., Wilbacher, I. (2016). Compositional data analysis in epidemiology. *Statistical Methods in Medical Research*, Accepted.

CHAPTER 2

Sparse Principal Balances

Abstract: Compositional data analysis deals with situations where the relevant information is contained only in the ratios between the measured variables, and not in the reported values. This paper focuses on high-dimensional compositional data (in sense of hundreds or even thousands of variables), as they appear in chemometrics (e.g. mass spectral data), proteomics, or genomics. The goal of this contribution is to perform a dimension reduction of such data, where the new directions should allow for interpretability. An approach named *principal balances* turned out to be successful for low dimensions. Here, the concept of sparse principal component analysis is proposed for constructing principal directions, so-called sparse principal balances. They are sparse (contain many zeros), build an orthonormal basis in the sample space of the compositional data, are efficient for dimension reduction, and are applicable to high-dimensional data.

Key words: principal component analysis; compositional data; isometric log-ratio transformation; sparseness

2.1 Introduction

Dimension reduction is often a first step for analyzing multivariate data. The representation of the data in a lower-dimensional space can be used in an exploratory way, e.g. to get an impression of the data structure, also as a precursor to other statistical methods, like regression or discriminant analysis. The latter is important in particular for high-dimensional data, where the number of variables is higher than the number of observations. The most meaningful method in terms of dimension reduction is principal component analysis (PCA), since it is the goal of PCA to maximize the explained variance. However, PCA cannot directly be applied to compositional data, because they live in a different geometry.

A composition $\boldsymbol{x} = (x_1, \ldots, x_D)^t$ consists of $D > 1$ strictly positive components, the so-called compositional parts, and it is represented in the D-part simplex

$$\mathcal{S}^D = \{\boldsymbol{x} = (x_1, \ldots, x_D)^t, x_j > 0, \sum_{j=1}^{D} x_j = \kappa\}.$$

The constant κ can be chosen arbitrarily (e.g., 100 in case of percentages), without changing the multivariate information (Aitchison, 1986; Egozcue, 2009; Pawlowsky-Glahn and Buccianti, 2011). The simplex sample space is a $(D-1)$-dimensional subset of the D-dimensional real Euclidean space. Typical examples for compositional data are income components, chemical element concentrations in some material, or spectral data where only the intensity is informative and not the absolute number of the intensity (e.g. data from mass spectrometry). The latter example is a case for a high-dimensional composition, i.e. where D is in the order of hundreds.

PCA for compositional data has been studied in many papers, starting with AITCHISON (1983); Aitchison (1984). The main idea is to transform the compositional data appropriately to the real Euclidean space $\mathbb{R}^D$, before PCA is carried out. For this purpose, the centered log-ratio (clr) transformation turned out to be useful (Aitchison and Greenacre, 2002). The clr transformation is a transformation from $\mathcal{S}^D$ to $\mathbb{R}^D$, and the results for a composition $\boldsymbol{x} \in \mathcal{S}^D$ are the transformed data $\boldsymbol{y} \in \mathbb{R}^D$ with

$$\boldsymbol{y} = (y_1, \ldots, y_D)^t = \left(\log \frac{x_1}{\sqrt[D]{\prod_{j=1}^{D} x_j}}, \ldots, \log \frac{x_D}{\sqrt[D]{\prod_{j=1}^{D} x_j}} \right)^t. \tag{2.1}$$

This transformation results, however, in singularity, because $\sum_{j=1}^{D} y_j = 0$. Thus, PCA applied to the clr-transformed data will result in $D-1$ principal components (PCs) that express the complete data information. A further option is to use the isometric log-ratio (ilr) transformation (Egozcue et al., 2003b), which is based on the choice of an orthonormal basis on the hyperplane $\mathcal{H} : y_1 + \cdots + y_D = 0$ in $\mathbb{R}^D$ that is formed by the clr transformation. The resulting ilr-transformed data are expressed in $\mathbb{R}^{D-1}$, and the collinearity problem is avoided. Note that the first $D-1$ PCs from the clr-transformed data are a special case of an ilr basis.

The coordinates of ilr-transformed data, or, equivalently, the PCA scores of clr-transformed data are in general difficult to interpret. PCs are linear combinations of ilr- (or clr-) variables, but in general the information of the original compositional parts is contained in not one, but several of the transformed variables. For example, each clr-variable contains information of all compositional parts due to the geometric mean in the denominator (see Equation (2.1)). This problem also remains when using the ilr transformation proposed in Egozcue et al. (2003b). It can be circumvented by the approach of Egozcue and Pawlowsky-Glahn (2005), where so-called *balances* are constructed, representing coordinates of an orthonormal basis in the simplex, which describe all the relative information of groups of compositional parts. If prior knowledge is available, balances can summarize the information of groups of meaningfully related components. Otherwise, orthonormal balances can be easily constructed by a procedure called sequential binary partitioning (Egozcue and Pawlowsky-Glahn, 2005).

Unlike PCs, balances are not constructed in order to maximize the explained variance. The idea of *principal balances* (PBs) (Pawlowsky-Glahn et al., 2011) is an attempt to satisfy both needs: to form an orthonormal basis on the simplex, and to maximize the explained variance with the constraint of allowing for a simpler interpretability. In other words, PBs are balances, describing all the relative information of groups of parts, but they should be close to the PCs of clr-transformed data in order to express much of the data information.

Pawlowsky-Glahn et al. (2011) suggested three algorithms to construct PBs. Two of them are computationally intensive and are thus limited to low dimension. The third is based on hierarchical clustering; it is fast to compute also for high-dimensional data, but the explained variance of the first few PBs can be quite low. In this paper, we propose a fourth algorithm based on sparse PCA (Witten et al., 2009) for computing interpretable orthonormal directions with high explained variance. The resulting balances are called *sparse principal balances* (SPB).

The paper is organized as follows: Section 2 reviews the algorithms of Pawlowsky-Glahn et al. (2011) to construct PBs. Section 3 introduces an algorithm for constructing sparse principal balances. Simulation studies in Section 4 compare the performance of the four different algorithms, and Section 5 presents an application with metabolomics data. The final Section 6 concludes with a discussion about the main features of the proposed algorithm and an outlook for future work.

2.2 Existing algorithms for constructing principal balances

Denote by $\boldsymbol{x}_+$ and $\boldsymbol{x}_-$ subcompositions of two disjoint groups of parts from a D-part composition $\boldsymbol{x} = (x_1, \ldots, x_D)$, where the number of parts in $\boldsymbol{x}_+$ and $\boldsymbol{x}_-$ is r and s, respectively, with $r + s \leq D$. Further, denote by $g_m(\cdot)$ the geometric mean of the arguments. Then a balance is defined as

$$b = \sqrt{\frac{rs}{r+s}} \ln \frac{g_m(\boldsymbol{x}_+)}{g_m(\boldsymbol{x}_-)} \tag{2.2}$$

(Egozcue and Pawlowsky-Glahn, 2005), and represents a coordinate of the composition with respect to a (compositional) basis vector $\boldsymbol{e}$ of unit length, formed in the clr-space. This basis vector $\boldsymbol{e}$, also called *balancing element*, is defined in the clr-space as $\text{clr}(\boldsymbol{e}) = (a_1, \ldots, a_D)$, with components

$$a_j = \begin{cases} 0 & \text{if } x_j \notin \boldsymbol{x}_+ \text{ and } x_j \notin \boldsymbol{x}_- \\ \sqrt{\frac{s}{r(r+s)}} & \text{if } x_j \in \boldsymbol{x}_+ \\ -\sqrt{\frac{r}{s(r+s)}} & \text{if } x_j \in \boldsymbol{x}_- \end{cases} \tag{2.3}$$

for $j = 1, \dots, D$. The possibility of including just disjoint subgroups of compositional parts into the construction of balances will be useful later on for achieving sparsity.

Orthonormal balances can be defined using a sequential binary partition (Egozcue and Pawlowsky-Glahn, 2005). The basic idea is to recursively partition each of the groups $\boldsymbol{x}_+$ and $\boldsymbol{x}_-$ in two further disjoint groups and to compute the balance and the corresponding balancing element according to (2.2) and (2.3), until a complete set of $D-1$ orthonormal basis vectors has been formed. In each step of the procedure, there are many possible disjoint partitions: with increasing dimension D the number of binary partitions grows exponentially (Pawlowsky-Glahn et al., 2011). The resulting orthonormal bases are just rotations of each other.

With PBs one wants to identify a set of $D-1$ orthonormal balances which successively maximize the explained variance. An exhaustive search for PBs is only feasible if D is very small. Pawlowsky-Glahn et al. (2011) thus introduced the following three approximate algorithms to compute PBs. Some of these algorithms use the resulting PCs of the clr-transformed data; these PCs will be called CoDa-PCs (CoDa stands for Compositional Data):

AP (angular proximity to principal components): In the first step of this recursive algorithm, all possible binary partitions of the full D-part composition are created. The balancing element with the smallest geometric angle with one of the CoDa-PCs is stored and removed from the set of possible directions. The procedure is then applied to each group of the previously identified balance separately, where the geometric angle to one of the remaining CoDa-PCs is minimized. This is repeated step-by-step, until $D-1$ balances are extracted, i.e., until a complete sequential binary partition is achieved.

HC (hierarchical clustering of components): The set of orthonormal balances is constructed by hierarchical cluster analysis. The variance of the balance between two groups (2.2) is used as a criterion to link two clusters. It turns out that this corresponds to a Ward clustering (Everitt, 1993) based on the variation matrix $t_{jk} = \mathrm{Var}[\log(x_j/x_k)]$, $j, k = 1, \dots, D$, where "Var" denotes the variance (Aitchison, 1986).

MV (maximum explained variance hierarchical balances): This sequential algorithm starts with the first CoDa-PC, and uses two groups with the signs of the loadings for constructing a balance. Let r denote the number of positive signs and s the number of negative signs. Then it is checked whether a change of one positive sign to the other group increases the explained variance. This check is also carried out for all combinations of $2, \dots, r-1$ positive signs. The balance with the maximum explained variance is stored, a new CoDa-PCA is performed with the larger group, and so on.

For more details on the algorithms we refer readers to Pawlowsky-Glahn et al. (2011). The computation time of AP and MV explodes quickly with increasing dimension because of the exponentially growing number of possible combinations that are used as candidates. Creating all possible combinations also leads to memory allocation problems for larger D. The HC algorithm is just based on a $D \times D$ distance matrix, which is unproblematic even for larger dimension.

2.3 A new algorithm for constructing sparse principal balances

Sparse principal balances (SPBs) can be defined as balances according to Equation (2.2), that make a tradeoff between maximizing explained variance and the number of involved components $r + s \ll D$. The latter condition simplifies the interpretation, particularly for large D. A SPB should describe the information of only a few compositional parts with zero contribution from the other (majority of) parts. This is similar to the aim of sparse PCA, where many of the entries of the loading matrix are forced to be zero (Zou et al., 2006). Here we use an algorithm proposed by Witten et al. (2009), implemented as function `SPC` in the R package `PMA` (Witten D and B, 2011). The idea is to perform a penalized matrix decomposition to obtain a rank-K approximation of the original data matrix.

Suppose we have given n compositions collected into the data matrix $\boldsymbol{X}$ of dimension $n \times D$. The clr-transformed matrix is denoted by $\boldsymbol{Y}$, and it has the same dimension. The rank-K approximation of $\boldsymbol{Y}$ is

$$\hat{\boldsymbol{Y}} = \sum_{l=1}^{K} d_l \boldsymbol{u}_l \boldsymbol{v}_l^t, \tag{2.4}$$

where d_l (scalar), $\boldsymbol{u}_l$ (vector of length n) and $\boldsymbol{v}_l$ (vector of length D) minimize the squared Frobenius norm of $\boldsymbol{Y} - \hat{\boldsymbol{Y}}$, subject to penalties on $\boldsymbol{u}_l$ and $\boldsymbol{v}_l$. In this paper, L_1 penalties are considered which lead to sparse solutions for $\boldsymbol{u}_l$ and $\boldsymbol{v}_l$. For $K = 1$, the problem can be formulated as

$$\max_{\boldsymbol{u}_1, \boldsymbol{v}_1} \boldsymbol{u}_1^t Y \boldsymbol{v}_1 \text{ subject to } \|\boldsymbol{u}_1\|_2^2 \leq 1,\ \|\boldsymbol{v}_1\|_2^2 \leq 1,\ \|\boldsymbol{u}_1\|_1 \leq c_1,\ \|\boldsymbol{v}_1\|_1 \leq c_2, \tag{2.5}$$

where $\|\cdot\|_2$ is the Euclidean norm, $\|\cdot\|_1$ is the L_1 norm (Witten et al., 2009), and c_1 and c_2 are tuning parameters.

For $K > 1$, Witten et al. (2009) propose the following approximation procedure:

1. Let $\boldsymbol{Y}^1 \leftarrow \boldsymbol{Y}$.
2. For $l \in 1, \ldots, K$:
    (a) Find $\boldsymbol{u}_l$, $\boldsymbol{v}_l$, and d_l by applying the above rank-1 approximation to the data $\boldsymbol{Y}^l$.
    (b) $\boldsymbol{Y}^{l+1} \leftarrow \boldsymbol{Y}^l - d_l \boldsymbol{u}_l \boldsymbol{v}_l^t$.

Without the L_1 constraints, this procedure leads to a rank-K singular value decomposition of $\boldsymbol{Y}$ with orthogonal vectors, i.e. $\boldsymbol{u}_l^t \boldsymbol{u}_m = 0$ and $\boldsymbol{v}_l^t \boldsymbol{v}_m = 0$, for $l \neq m$ and $l, m \in \{1, \ldots, K\}$. In particular, $\boldsymbol{v}_l$ corresponds to the l-th principal component direction of $\boldsymbol{Y}$. Orthogonality is also desired in the case with L_1 constraints, because we are interested that few PCs, explaining as much of the total variance as possible, are also easily interpretable. Witten et al. (2009) outline a procedure that allows for orthogonality also for the L_1 case.

SPBs can now be constructed as follows. As with the algorithms in Section 2, we start with a sparse PCA of the clr-transformed data matrix $\boldsymbol{Y}$, by extracting a fixed number K $(1 \leq K < D)$ of components. The resulting sparse loadings matrix is denoted by the $D \times K$ matrix $\boldsymbol{V} = [v_{jk}]$. Usually, K will be small (e.g., $K \leq 5$), since we are only interested in a few balances that explain an essential part of the variability. The resulting loadings, however, need to be modified for SPBs, because for a simplification of the interpretation we want to ensure that different balances relate to different parts. In other words, we want to force a disjoint pattern of non-zero entries in different balances. The proposed algorithm is as follows:

1) Force a disjoint non-zero pattern:
   For $j \in \{1, \ldots, D\}$:

   > Find the smallest k for which $v_{jk} \neq 0$, and set all entries v_{jl} to zero, for $l > k$ (if they are non-zero). If $v_{jk} = 0$ for all k, store the index j in the index set $\mathcal{J}_0$.

   $\mathcal{J}_0$ contains the indices of all zero-lines in $\boldsymbol{V}$. The number of elements in $\mathcal{J}_0$ is denoted by $|\mathcal{J}_0|$.

2) Denote by $\mathcal{K}_p$ and $\mathcal{K}_n$ the index sets containing the column numbers $k \in \{1, \ldots, K\}$ of $\boldsymbol{V}$, where the k-th column has no positive or negative entry, respectively. For example, if $v_{jk} \leq 0$ for all j, the index k will be stored in $\mathcal{K}_p$. The sizes of these sets are $|\mathcal{K}_p|$ and $|\mathcal{K}_n|$.

3) While $|\mathcal{J}_0| < |\mathcal{K}_p| + |\mathcal{K}_n|$:

   - Identify the entry (j, k) of $\boldsymbol{V}$ with the smallest positive absolute value, and set this v_{jk} to zero.
   - Update $\mathcal{J}_0$, $\mathcal{K}_p$ and $\mathcal{K}_n$.

   As a result we have at least as many zero-lines in $\boldsymbol{V}$ as "problematic" columns, i.e. columns without positive *and* negative sign.

4) Guarantee positive *and* negative sign in each column of $\boldsymbol{V}$:

   - Repeat for all $k_p \in \mathcal{K}_p$:
     For any $j_0 \in \mathcal{J}_0$: set $v_{j_0 k_p}$ to any positive value and update $\mathcal{J}_0 := \mathcal{J}_0 \backslash \{j_0\}$.

   - Repeat for all $k_n \in \mathcal{K}_n$:
     For any $j_0 \in \mathcal{J}_0$: set $v_{j_0 k_n}$ to any negative value and update $\mathcal{J}_0 := \mathcal{J}_0 \backslash \{j_0\}$.

5) The resulting columns of the (modified) matrix $\boldsymbol{V}$ are in general not formed by vectors in the clr space, i.e. their elements do not sum up to zero. In order not to change the zero elements of $\boldsymbol{V}$, we project the $d_l \leq D$ non-zero elements of each column $\boldsymbol{v}_l$, denoted as $\boldsymbol{v}_l^*$, to $\boldsymbol{w}_l^* = \boldsymbol{G}_l \boldsymbol{v}_l^*$, $l = 1, \ldots, K$, using the $d_l \times d_l$ matrix $\boldsymbol{G}_l = \boldsymbol{I} - (1/d_l)\boldsymbol{J}$ from Aitchison (1986). Here $\boldsymbol{I}$ and $\boldsymbol{J}$ stand for the identity matrix and matrix of ones, respectively (both of order d_l). The resulting vector $\boldsymbol{w}_l$ with non-zero elements corresponding to $\boldsymbol{w}_l^*$ thus form a "suboptimal projection" (resulting from the sparsity constraint) of $\boldsymbol{v}_l$ to the clr hyperplane. This projection is necessary to retain sparse PCA loadings in the clr space as it is the case for classical PCA.

6) As a final modification of $\boldsymbol{V}$, the nearest balances (with respect to the Euclidean distance) to the resulting clr vectors are constructed (Egozcue and Pawlowsky-Glahn, 2005). For this purpose, we simply compute the arithmetic mean from the positive/negative entries of $\boldsymbol{w}_l$, and by rescaling to unit norm the balancing element is obtained, see Equation (2.2). The set of extracted balances is orthogonal by construction.

Execution time comparisons on a standard computer show immediately, which of the four algorithms is feasible for the computation with high-dimensional data. Figure 2.1 shows the average computation time (average over 10 trials) for computing the first balance with the different algorithms. The number of observations is always $n = 100$, but the number of parts D varies from 4 to 50 (left plot). The computation times for the algorithms AP and MV explodes for quite moderate values of D, and thus these algorithms will not be considered later on for the simulations in high-dimensional settings. On the other hand, the time for HC and SPB is still very low. In the right plot, the same setting is used, but D is taken much higher (100, 500, 1000, 2000). Here we only compare HC and SPB which are still feasible to compute. Although the computation time for HC is still quite moderate, the one for SPB is far below in all runs.

2.4 Simulations

In this section we compare the methods HC and SPB which are still feasible to compute for high-dimensional compositional data. By simulating data of varying dimension, we focus on the explained variance of the resulting balances, and compare with the explained variance achieved by CoDa-PCA. Also the resulting sparseness will be of interest in the evaluation.

The simulation setting is as follows:

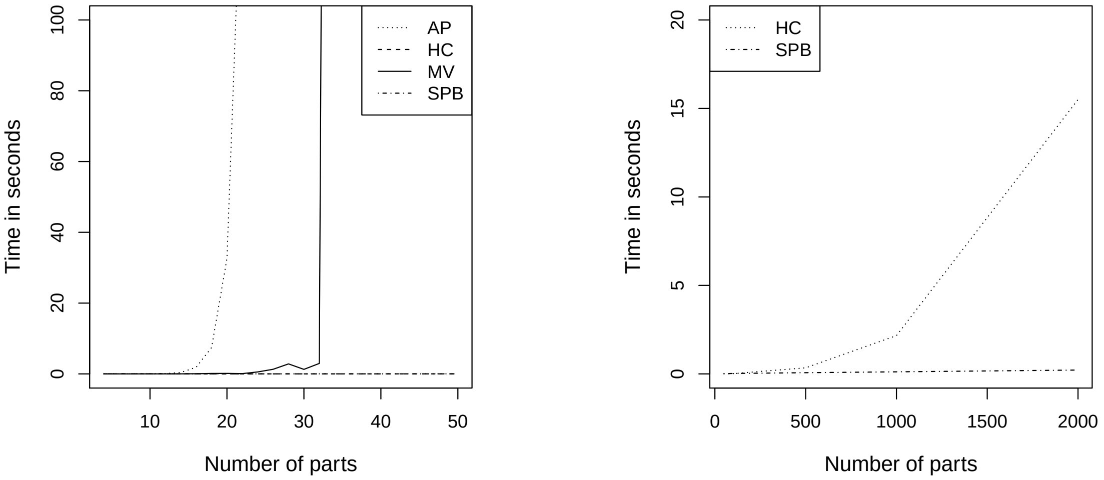

Figure 2.1: Time comparison to compute the first balance according to the algorithms AP, HC, MV, and SPB. Shown are average computation times for compositional data with $n = 100$ observations and varying number D of parts. The right plot compares only the algorithms HC and SPB in high-dimensional settings.

1. A matrix $\boldsymbol{L}_{ilr}$ of size $(D-1) \times (D-1)$ is generated with uniformly distributed values in $[-1, 1]$. The columns are then standardized to unit length vectors (loadings matrix in ilr-space).
2. A matrix $\boldsymbol{Z}_{ilr}$ of size $n \times (D-1)$ is generated according to a $(D-1)$-dimensional normal distribution with mean vector $(0, \dots, 0)^t$ and a diagonal covariance matrix $\boldsymbol{C}_Z = \mathrm{diag}(0.9^1, 0.9^2, \dots, 0.9^k, 0.01, \dots, 0.01)$. So, the PCA scores are uncorrelated and their variances are decreasing exponentially.
3. The data matrix is reconstructed in the ilr-space by $\boldsymbol{X}_{ilr} = \boldsymbol{Z}_{ilr}\boldsymbol{L}_{ilr}^t$.
4. $\boldsymbol{X}_{ilr}$ is transformed to the simplex by the inverse ilr transformation

$$\begin{aligned}
\boldsymbol{x}_s^1 &= \exp\left(\frac{\sqrt{D-1}}{\sqrt{D}}\boldsymbol{x}_{ilr}^1\right), \\
\boldsymbol{x}_s^j &= \exp\left(-\sum_{k=1}^{j-1}\frac{1}{\sqrt{(D-k+1)(D-k)}}\boldsymbol{x}_{ilr}^k + \frac{\sqrt{D-j}}{\sqrt{D-j+1}}\boldsymbol{x}_{ilr}^j\right), \\
&\quad \text{for } i = 2, \dots, D-1, \\
\boldsymbol{x}_s^D &= \exp\left(-\sum_{j=1}^{D-1}\frac{1}{\sqrt{(D-j+1)(D-j)}}\boldsymbol{x}_{ilr}^j\right).
\end{aligned}$$

Here, $\boldsymbol{x}_{ilr}^1, \dots, \boldsymbol{x}_{ilr}^{D-1}$ are the columns of $\boldsymbol{X}_{ilr}$, and the matrix in the simplex, $\boldsymbol{X}_s$, is formed by the columns $\boldsymbol{x}_s^1, \dots, \boldsymbol{x}_s^D$ (see Martín-Fernández et al., 2012).

5. The clr transformation (Equation (2.1)) is applied to $\boldsymbol{X}_s$, resulting in $\boldsymbol{X}_{clr}$.

CoDa-PCA is applied to $\boldsymbol{X}_{clr}$. Note that the total variance corresponds to the sum of the diagonal elements of $\boldsymbol{C}_Z$. HC is applied to $\boldsymbol{X}_s$, and SPB is applied to $\boldsymbol{X}_{clr}$. In all simulation runs we fix the number of observations with $n = 100$ and the number of replications with 1000.

2.4.1 Experiments with $K = 5$

The number of extracted PCs is $K = 5$, and the data dimension is taken as $D \in \{50, 100, 500, 1000\}$. For the SPB method, the tuning parameters c_1 and c_2 need to be specified, see (2.5). This is done for each considered D separately, at the basis of one simulated data set. The tuning parameters that lead to the highest cumulative explained variance in SPB for $K = 5$ are selected.

Figure 2.2 compares the results for the cumulative explained variances of the three methods. Each boxplot summarizes the results of the 1000 simulations. In lower dimension ($D = 50$ and $D = 100$) the methods SPB and HC result in a quite similar explained variance, but with increasing dimension, SPB clearly gets better. There is still a clear gap to the explained variance of CoDa-PCA, but remember that SPBs are based on sparse PCA which forms a compromise between explained variance and sparsity of the loadings matrix.

The method SPB results in balances which involve non-overlapping groups of variables. The balances thus have a certain number of non-zeros, and Figure 2.3 displays for the above simulation results the proportion of non-zeros in each balance, cumulated by subsequent balances. This information is useful for evaluating the level of sparseness of the algorithm. For example, for $D = 50$ (upper left plot), about 30% of the entries of the first balance are non-zero, for the second balance we obtain slightly less than 30% non-zeros, which gives in total somewhat less than 60% of non-zeros for the first two balances. With the first 5 balances, about 95% of the variables have entered the sparse principal balances. Also in higher dimension, these percentages are comparable. Note that the data are not simulated with inherent sparseness structure.

2.4.2 Experiments with $K = 2$

Here we are only interested in the first $K = 2$ components. Similar as before, the tuning parameters for SPB are selected according to maximize the explained variance for $K = 2$. Now we divide the cumulative explained variances for $K = 2$ obtained from SPB by those from HC after each simulation run. The results are shown by boxplots in Figure 2.4. It turns out that only for $D = 10$, HC leads to a higher median value, for $D = 50$ and $D = 100$ the method SPB achieves higher values in the vast majority of runs, and for even higher values of D, SPB is clearly improving over HC.

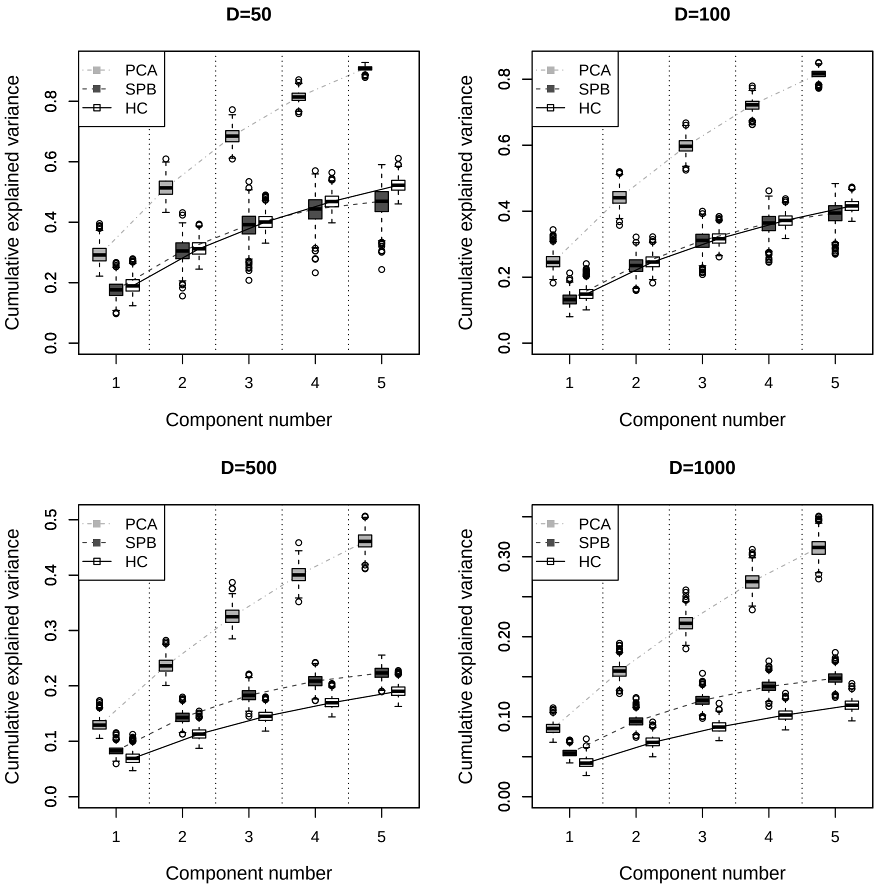

Figure 2.2: Cumulative explained variance of the first 5 components for CoDa-PCA, SPB, and HC for simulated data with different dimension D. With increasing dimension, SBP results in higher explained variance compared to HC.

2.5 Application to metabolomics data

The data set considered here contains NMR metabolomic spectra from urine samples of 18 mice, each belonging to one of two treatment groups. Each spectrum has 189 spectral bins, and they are measured in parts per million (ppm). This already clearly emphasizes the compositional nature of the data. The data set is described in detail in Nyamundanda et al. (2010), and it is available in the R package `MetabolAnalyze` as data set `UrineSpectra` (Gift et al., 2010).

Our goal is a two-dimensional representation of the spectral information. We compare the CoDa-PCA method with the hierarchical clustering (HC) and the sparse principal balances (SPB) approaches. Table 2.1 shows the explained variances for the three methods when using two components (balances). There is of course an information loss compared

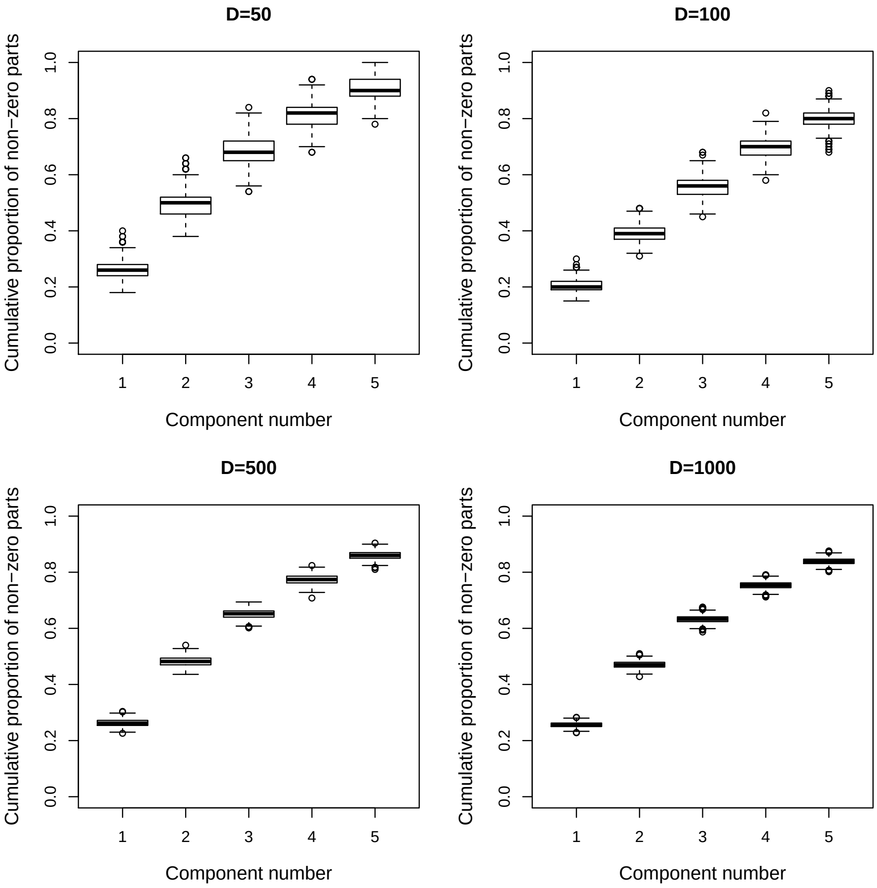

Figure 2.3: Cumulative proportions of non-zero entries in the balances of SPB.

to the clr-based CoDa-PCA approach. The explained variance of the first SPB component is much more than that of HC, and the cumulative explained variance of HC and SPB for two components is comparable, although SPB aims at a simpler interpretation than HC.

Figure 2.5 shows the original spectral data, overlaid with the information of the signs of the first two balances from HC. Specifically, the vertical lines refer to the information of the sign: if the sign of the balance for a part is positive, we see a light gray vertical line (red, in the online-version) at the corresponding position in the plot, and if the sign is negative, the vertical line is shown in dark gray (blue). It can be seen that for the first balance, there are few negative signs, and all remaining signs are positive. For the second balance we have the reverse behavior, but here we also get zeros at positions where the signs of the first balance are negative (no vertical lines). Overall, an interpretation of the balances is not obvious.

In comparison to Figure 2.5 for HC, Figure 2.6 shows the information of the first two

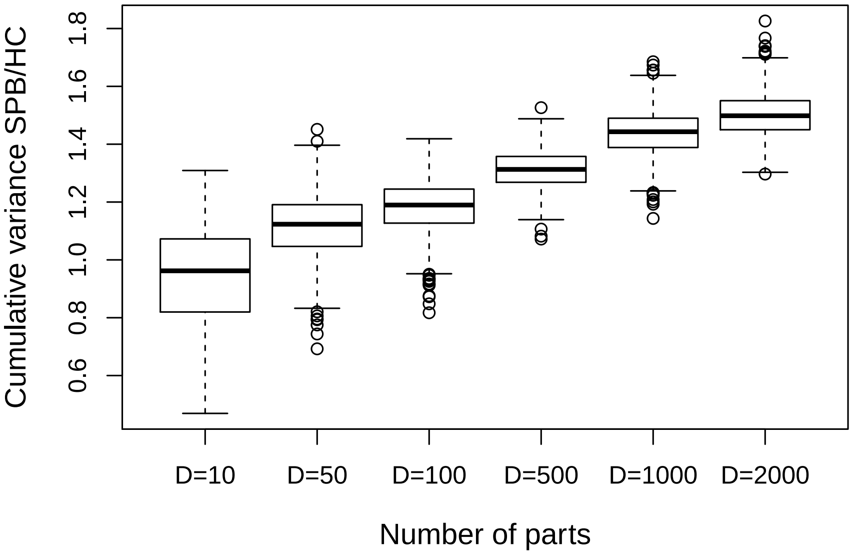

Figure 2.4: Cumulative explained variance for $K = 2$ components. Displayed is the ratio between SPB and HC. The explained variance clearly gets larger for SPB compared to HC with increasing dimension.

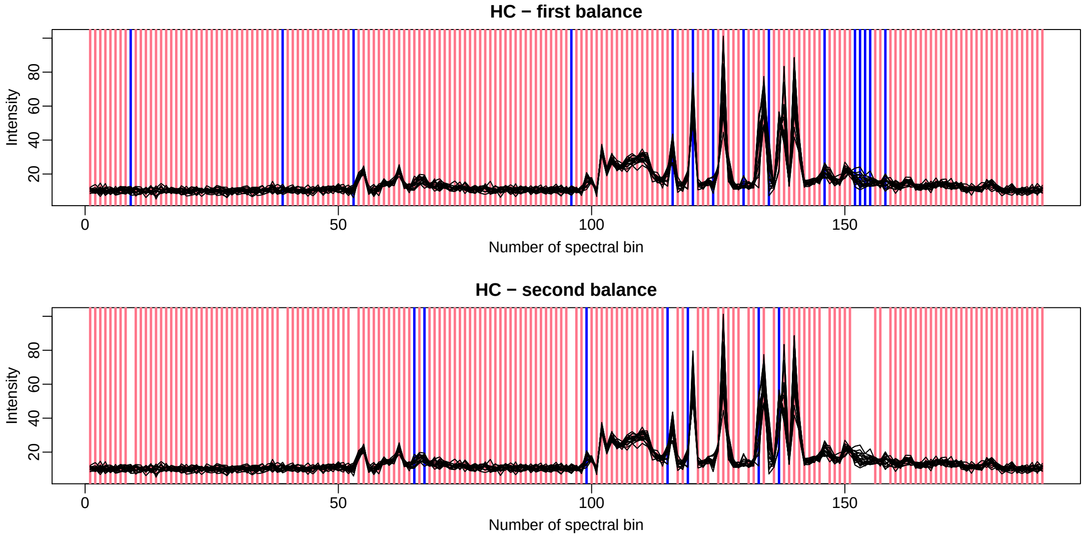

Figure 2.5: First two balances of HC applied to the urine data. Shown are the original data (black), and the position of the positive (light gray vertical line, or red in the online-version) and negative (dark gray vertical lines, or blue) signs of the balances.

Table 2.1: Cumulative explained variances for CoDa-PCA, hierarchical clustering (HC) and sparse principal balances (SPB) for the urine data set.

	Cumulative explained variance [%]	
Method	one component	two components
CoDa-PCA	28.1	38.5
HC	8.9	16.7
SPB	13.9	15.6

balances from SPB. One can see that only few non-zero entries exist (light and dark gray lines), allowing for a severe simplification of the interpretation. The upper plot for the first balance shows non-zero entries essentially at the peaks of the signal. Non-zeros in the second balance are at smaller peaks or in ranges of larger variability of the signals.

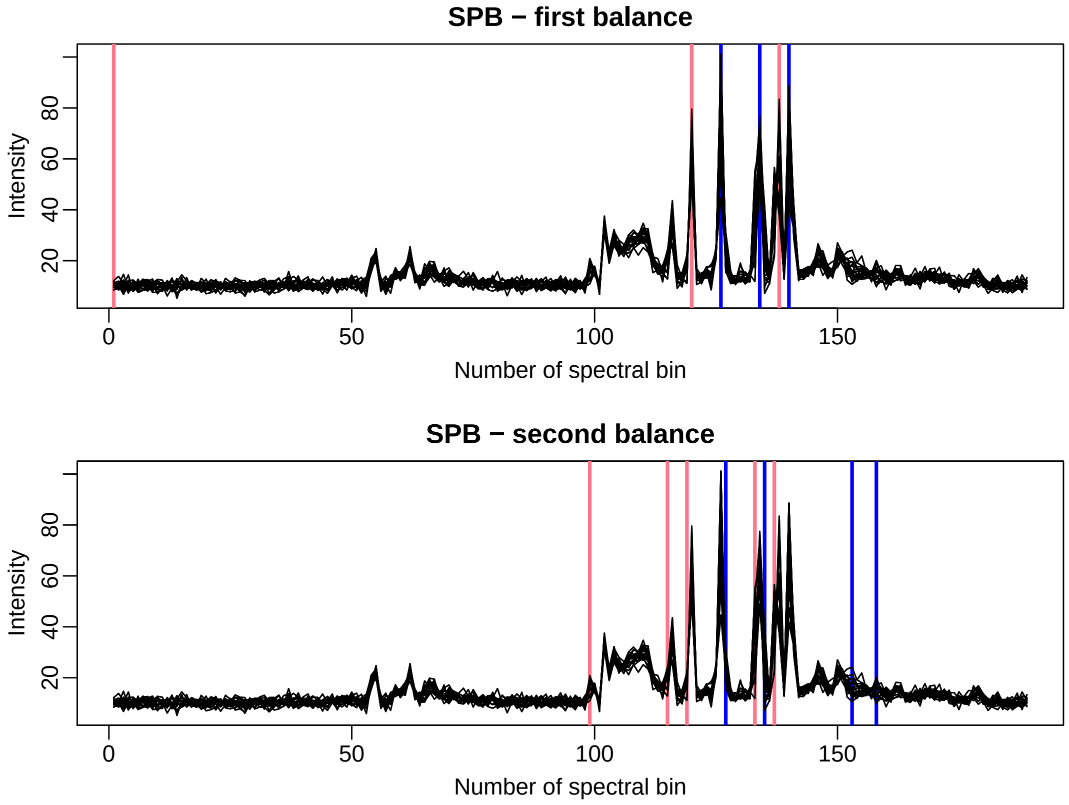

Figure 2.6: First two balances of SPB applied to the urine data. Shown are the original data (black), and the position of the positive (light gray vertical line, or red in the online-version) and negative (dark gray vertical lines, or blue) signs of the balances.

Finally, Figure 2.7 shows the projection of the data in the space of the first two PCs (balances). More specifically, the left plot shows the first two PCA scores of CoDa-PCA, the middle plot shows the first two coordinates derived from HC, and the right plot presents the first two coordinates from SPB. The plot symbols are according to the grouping information (control and treatment group). The groups show a clear pattern in all three plots, which means that the between-group variability is reasonably high in order to be visible in the first two main directions. For CoDa-PCA, the group separation is visible only along the first PC, while for HC and SPB both components are informative for distinguishing the groups. In particular for SPB, the groups have an easier interpretation, since only very few variables contribute to the resulting components, see Figure 2.6.

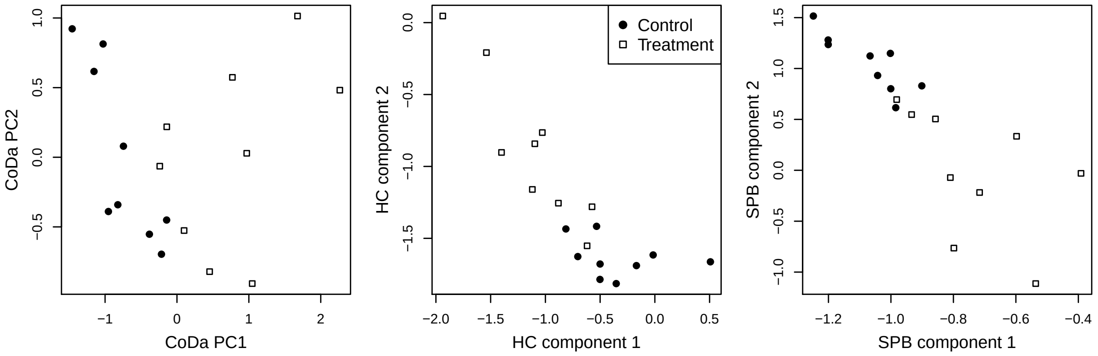

Figure 2.7: First two components of CoDa-PCA, HC and SPB for the urine data.

2.6 Conclusions

In many applied fields like analytical chemistry, metabolomics, or genomics, the experimental measurement processes yield high-dimensional data where the reported data values are not of direct interest, but rather the ratios between the variables. This type of data is called compositional data, and the log-ratio approach for their analysis allows for a more reliable insight into the real data structure. The main aim of this paper is to show that the compositional data analysis method of principal balances leads to reasonable and easily interpretable results, if *sparse* principal component analysis is employed for dimensionality reduction. Hereat, the necessity of working in the clr-space, leading to keep the zero-sum constraint and thus to limited possibilities of data manipulation, affected also the resulting algorithm. Despite that, both simulations and the real data example suggest that sparsity provides a good trade-off between maximization of the explained total variance and interpretability.

Especially for spectral data, where the order of the variables has a meaning, this work could be extended in the direction of sparse fused PCA (Guo et al., 2010). This method yields non-zero loadings for neighboring spectra, resulting in an even clearer interpretation.

Also in the context of compositional data, non-zero loadings for neighboring spectra would lead to the interpretation that a whole range of spectra, related to an inherent property of the measurement, is relevant for a sparse principal balance.

Another possible extension is to use a robust sparse PCA method for the clr-transformed data as a starting point for the SPB algorithm. Robust sparse principal components are less sensitive to data outliers (Croux et al., 2013).

In conclusion, we feel that sparsity constraints on compositional data analysis methods offer a practical way to gain insight into high-dimensional compositions.

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 PrF_2013_013 Mathematical Models of the Internal Grant Agency of the Palacký University in Olomouc.

CHAPTER 3

Error Propagation in Isometric Log-ratio Coordinates for Compositional Data: Theoretical and Practical Considerations

Abstract: Compositional data, as they typically appear in geochemistry in terms of concentrations of chemical elements in soil samples, need to be expressed in log-ratio coordinates before applying the traditional statistical tools if the relative structure of the data is of primary interest. There are different possibilities for this purpose, like centered log-ratio coefficients, or isometric log-ratio coordinates. In both approaches, geometric means of the compositional parts are involved, and it is unclear how measurement errors or detection limit problems affect their presentation in coordinates. This problem is investigated theoretically by making use of the theory of error propagation. Due to certain limitations of this approach, the effect of error propagation is also studied by means of simulations. This allows to provide recommendations for practitioners on the amount of error and on the expected distortion of the results, depending on the purpose of the analysis.

Key words: Aitchison geometry; Orthonormal coordinates; Taylor approximation; Compositional differential calculus; Detection limit

3.1 Introduction

Compositional data analysis is concerned with analyzing the relative information between the variables, the so-called compositional parts, of a multivariate data set. Here, relative information refers to the log-ratio methodology (Aitchison, 1986), so in fact to an analysis of logarithms of ratios between the compositional parts. It has been demonstrated that the sample space of compositions is not the usual Euclidean space, but the simplex with the so-called Aitchison geometry (Pawlowsky-Glahn et al., 2015). For a composition $\boldsymbol{x} = (x_1, \cdots, x_D)$ with D parts, the simplex sample space is defined as

$$\mathcal{S}^D = \{\boldsymbol{x} = (x_1, \ldots, x_D) \text{ such that } x_j > 0 \ \forall j, \sum_{j=1}^{D} x_j = \kappa\}$$

for an arbitrary constant κ. Nevertheless, according to recent developments, the sample space of compositional data is even more general (Pawlowsky-Glahn et al., 2015): A vector $\boldsymbol{x}$ is a D-part composition when all its components are strictly positive real numbers and carry only relative information. Note that the term relative information is equivalent to information lies in the ratios between the components, not in the absolute values. As a consequence, the actual sample space of compositional data is formed by equivalence classes of proportional positive vectors. Therefore, any constant sum constraint is just a proper representation of compositions that honors the scale invariance principle of compositions: the information in a composition does not depend on the particular units in which the composition is expressed (Egozcue, 2009). In practical terms, the choice of the constant κ is irrelevant, since it does not alter the results from a log-ratio based analysis. In that sense, a discussion on whether the values of an observation sum up to the same constant is needless, this would not make any difference for the analysis considered in this paper. Though, for the purpose of better interpretability or visualization, one could also express compositions with the closure operator

$$\mathcal{C}(\boldsymbol{x}) = \left(\frac{\kappa x_1}{\sum_{j=1}^{D} x_j}, \cdots, \frac{\kappa x_D}{\sum_{j=1}^{D} x_j} \right),$$

which then sum up to the constant κ.

The Aitchison geometry defines a vector space structure of the simplex by the basic operations of perturbation and powering. Given two compositions $\boldsymbol{x} = (x_1, \cdots, x_D)$ and $\boldsymbol{y} = (y_1, \cdots, y_D)$ in $\mathcal{S}^D$, perturbation refers to vector addition, and is defined as

$$\boldsymbol{x} \oplus \boldsymbol{y} = \mathcal{C}(x_1 y_1, \cdots, x_D y_D).$$

Powering refers to a multiplication of a composition $\boldsymbol{x} = (x_1, \cdots, x_D) \in \mathcal{S}^D$ by a scalar $\alpha \in \mathbb{R}$, and is defined as

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

Further, the Aitchison inner product, the Aitchison norm, and the Aitchison distance have been defined, and they lead to a Euclidean vector space structure (Pawlowsky-Glahn et al., 2015). All these definitions employ log-ratios between the compositional parts; for instance, the Aitchison inner product between the compositions $\boldsymbol{x}$ and $\boldsymbol{y}$ is given as

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_A = \frac{1}{2D} \sum_{j=1}^{D} \sum_{k=1}^{D} \ln \frac{x_j}{x_k} \ln \frac{y_j}{y_k},$$

that leads to the Aitchison norm and distance

$$||\boldsymbol{x}||_A = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_A}, \quad d_A(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} \oplus (-1) \odot \boldsymbol{y}||_A$$

respectively. Working directly in the simplex sample space is not straightforward. Rather, it is common to express compositional data in the usual Euclidean geometry. In the literature, one frequently refers to transformations; here it is prefered to use the terminology of expressing the compositions in appropriate coordinates with respect to the Aitchison geometry (Pawlowsky-Glahn and Egozcue, 2001) that allows to analyze compositions in the usual Euclidean geometry.

The focus in this paper is on isometric log-ratio (ilr) coordinates (Egozcue et al., 2003b), which allow to express a composition $\boldsymbol{x} \in \mathcal{S}^D$ in the real space $\mathbb{R}^{D-1}$. A particular choice for ilr coordinates is

$$z_j = ilr_j(\boldsymbol{x}) = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{x_j}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_k}}, \; j = 1, \ldots, D-1, \tag{3.1}$$

and the coordinates $\boldsymbol{z} = (z_1, \ldots, z_{D-1})$ indeed correspond to an orthonormal basis in $\mathbb{R}^{D-1}$ (Egozcue et al., 2003b). The particular choice of the ilr coordinates in (4.4) allows for an interpretation of the first coordinate z_1 as that one expressing all relative information about part x_1, since x_1 is not included in any other ilr coordinate.

The definition of ilr coordinates (4.4) reveals that geometric means of (subsets of) the parts are involved. Note that the geometric mean of $\boldsymbol{x}$ can also be expressed as

$$g_m(\boldsymbol{x}) = \left(\prod_{j=1}^{D} x_j \right)^{1/D} = \exp \left(\frac{1}{D} \sum_{j=1}^{D} \ln x_j \right)$$

involving the arithmetic mean of the log-transformed values. It is well known that the arithmetic mean is sensitive to data outliers (Maronna et al., 2006). Consequently, also data imprecision in one or some compositional parts (that are usually measured without respecting the relative nature of compositional data), or detection limit problems, may act like outliers and lead to a distortion of the geometric mean. The resulting ilr coordinates will suffer from data quality problems, and subsequent analyses based on these coordinates can be biased.

This unwanted effect is investigated here under the terminology of error propagation, where the effect of the errors on the output of a function is analyzed. Propagation of error can be performed by a calculus-based approach, or by simulation studies. A calculus-based approach makes use of the Taylor series expansion and calculates the first two statistical moments of the error of output, the mean and the variance, under the assumption that the errors are statistically independent (Ku, 1966). With few exceptions, almost all analyses of error propagation with the calculus-based approach use the first-order Taylor approximation, and neglect the higher order terms (Birge, 1939).

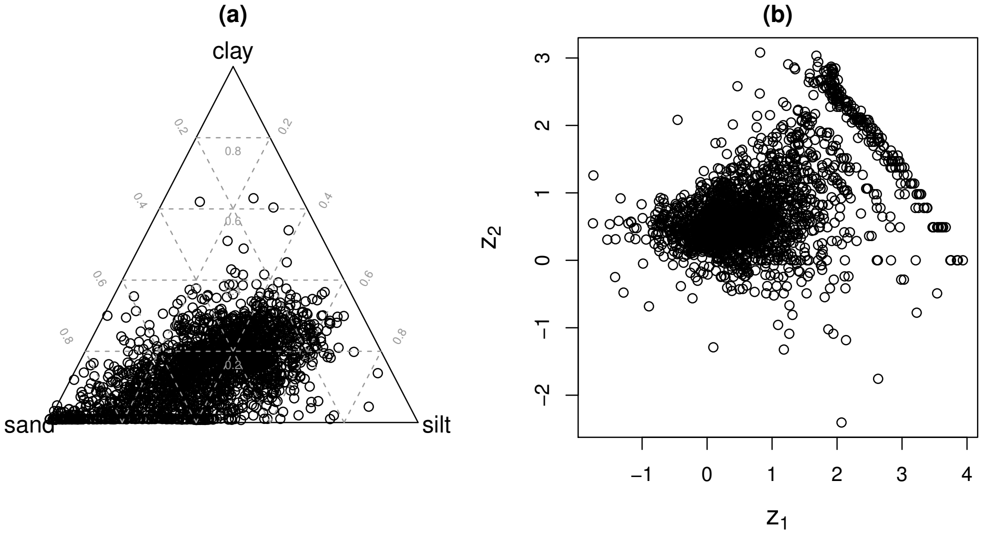

Figure 3.1: Composition of sand, silt, and clay in agricultural soils of Europe. Ternary diagram (a), representation in ilr coordinates (b)

This approach is briefly reviewed in Section 3.2. Section 3.3 starts with a motivating example about the effect of the errors on ilr coordinates and applies the concept of Taylor approximation to error propagation in the simplex. While this is done in a general form for any function (transformation), particular emphasis is given to error propagation for ilr coordinates that causes one source of distortion of outputs in practical geochemical problems (Filzmoser et al., 2009c).

Determining error propagation only for the first two moments is unsatisfactory, because it would also be interesting how the data structure is changed in case of data problems like detection limits or imprecision of the measurements. Thus, simulation-based methods for error propagation are considered as well. The Monte Carlo method is adaptable and simple for the propagation of errors (Feller and Blaich, 2001; Cox and Siebert, 2006), and various applications of this method can be found (Liu, 2008). The simulation-based approach in Section 3.4 makes use of a practical data set and shows the effect of imprecision and detection limit effects on the ilr coordinates. The interest lies particularly in error propagation on the first ilr coordinate, because this contains all relative information about the first compositional part, and on error propagation on all ilr coordinated jointly, because they contain the full multivariate information. The final Section 3.5 discusses the findings and concludes.

3.2 Error Propagation in the Standard Euclidean Geometry

Consider a p-dimensional random variable $\boldsymbol{x} = (x_1, \ldots, x_p)$, and a function $f : \mathbb{R}^p \to \mathbb{R}$ that gives the output y as a result of $y = f(\boldsymbol{x})$. The propagation of the errors of each variable through the function f on the output can be derived by using Taylor approximation (Ku, 1966). This yields a linear approximation of the function f by the tangent plane where the slopes in $x_1, \ldots, x_p$ are described by the partial derivatives $\frac{\partial y}{\partial x_1}, \ldots, \frac{\partial y}{\partial x_p}$ at a single point. One can express the random variables $(x_1, \ldots, x_p)$ as the sum of their expected values $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$ and random deviations from the expected value $\boldsymbol{\epsilon} = (\varepsilon_1, \ldots, \varepsilon_p)$, so that $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, assuming that the errors have mean zero. Taking the first-order Taylor approximation of $f(\boldsymbol{x})$ results in

$$\begin{aligned} y &= f(x_1, \cdots, x_p) = f(\mu_1 + \epsilon_1, \cdots, \mu_p + \epsilon_p) \\ &\approx f(\mu_1, \cdots, \mu_p) + \left[\frac{\partial f}{\partial x_1}(\mu_1)\right]\epsilon_1 + \cdots + \left[\frac{\partial f}{\partial x_p}(\mu_p)\right]\epsilon_p. \end{aligned} \tag{3.2}$$

In the framework of error propagation it is common to assume that $(x_1, \ldots, x_p)$ follow a known distribution, in most cases a multivariate normal distribution (Ku, 1966). If the distribution is known, the partial derivatives are evaluated at the true means, if not, the sample averages are used for the estimation. The approximation in Equation (3.2) can now be used to calculate mean and variance of y, which both depend on the function f. The second central moment, the variance $\text{Var}(y)$, describes the uncertainty, which is mainly used to investigate the effect of error propagation and is given as

$$\text{Var}(y) \approx \sum_{j=1}^{p} \left(\frac{\partial f}{\partial x_j}(\mu_j)\right)^2 E({\epsilon_j}^2) + \sum\sum_{j \neq k} \left(\frac{\partial f}{\partial x_j}(\mu_j)\right)\left(\frac{\partial f}{\partial x_k}(\mu_k)\right) E(\epsilon_j \epsilon_k). \tag{3.3}$$

Equation (3.3) reveals how the variability of the output y depends on the errors and on the function f.

3.3 Error Propagation on the Simplex

As a motivating example the composition of sand, silt, and clay in agricultural soils in Europe is considered. The data are reported in Reimann et al. (2014). From the ternary diagram (Figure 3.1(a)) it can be seen that the clay concentrations can be very small, but data artifacts are not immediately visible. The resulting ilr coordinates z_1 and z_2 are shown in Figure 3.1(b). Here the small clay values are visible in form of a band that deviates clearly from the joint data structure. In fact, small values of clay have been rounded in the laboratory, which causes already a distortion of the multivariate data structure. Thus, the imprecision here is visible as a rounding effect in the part clay. Variables with values below a detection limit can result in similar artifacts, since usually the values below detection are set to some constant, like 2/3 times the values of the detection limit (Martín-Fernández et al., 2003). This is still the usual practice in geosciences rather than employing more sophisticated algorithms for their imputation (Martín-Fernández et al., 2012).

Similar as in Section 3.2, error propagation is derived for a general function using first-order Taylor approximation. However, since this is directly done on the simplex, also the Taylor approximation needs to be done on the simplex. The theoretical background for the differential calculus on the simplex can be found in Barceló-Vidal and Martín-Fernández (2002) and Barceló-Vidal et al. (2011). Here the tools necessary to carry out the Taylor approximation are recalled.

Let $f : U \to \mathbb{R}^m$ be a vector-valued function defined on a subset $U \subset \mathbb{R}^D_+$. Let $\underline{U} = \{\mathcal{C}(\boldsymbol{w}), \boldsymbol{w} \in U\}$, the compositional closure of U, be a subset of $\mathcal{S}^D$. If f is scale invariant, that is $f(\boldsymbol{w}) = f(k\boldsymbol{w})$ for any $k > 0$, it induces a vector-valued function $\underline{f} : \underline{U} \to \mathbb{R}^m$. It suffices to define

$$\underline{f}(\boldsymbol{x}) = f(\boldsymbol{w}) \ ,\forall \boldsymbol{w} \in U,$$

where $\mathcal{C}(\boldsymbol{w}) = \boldsymbol{x}$ (Barceló-Vidal et al., 2011). The function $\underline{f}$ is $\mathcal{C}$-differentiable at $\boldsymbol{x} \in \underline{U}$, if there exists an $m \times D$ matrix $\boldsymbol{A} = (a_{ij})$, satisfying $\boldsymbol{A}\mathbf{1}_D = \mathbf{0}_m$ (defining a linear transformation from $\mathbb{R}^D$ to $\mathbb{R}^m$), such that

$$\lim_{\boldsymbol{u} \xrightarrow{\mathcal{C}} \boldsymbol{n}} \frac{\|\underline{f}(\boldsymbol{x} \oplus \boldsymbol{u}) - \underline{f}(\boldsymbol{x}) - \boldsymbol{A} \ \mathrm{ln}\boldsymbol{u}\|}{\|\boldsymbol{u}\|_A} = 0$$

for $\boldsymbol{u} \in \underline{U}$, where $\mathbf{1}_D = (1, \cdots, 1)$ with length D, and $\mathbf{0}_m = (0, \cdots, 0)$ with length m. Note that $\boldsymbol{n} = \mathcal{C}(1, \cdots, 1)$ is the neutral element of $(\mathcal{S}^D, \oplus)$ and $\boldsymbol{u} \xrightarrow{\mathcal{C}} \boldsymbol{n}$ denotes that $\boldsymbol{u}$ converges to $\boldsymbol{n}$ on the simplex. From the definitions above, the first-order Taylor approximation of a real-valued function $\underline{f}$ can be written as

$$\underline{f}(\boldsymbol{x} \oplus \boldsymbol{u}) \approx \underline{f}(\boldsymbol{x}) + \sum_{j=1}^{D} \ln(u_j) \left[\frac{\partial_{\mathcal{C}} \underline{f}}{\partial x_j}(\boldsymbol{x}) \right], \tag{3.4}$$

where the $\mathcal{C}$-derivative of $\underline{f}$ exists and is equal to

$$\frac{\partial_{\mathcal{C}} \underline{f}}{\partial x_j}(\boldsymbol{x}) = x_j \left(\frac{\partial \underline{f}}{\partial x_j}(\boldsymbol{x}) - \sum_{i=1}^{D} x_i \frac{\partial \underline{f}}{\partial x_i}(\boldsymbol{x}) \right) \text{ for } j = 1, \ldots, D. \tag{3.5}$$

Given a D-part composition $\boldsymbol{x} = (x_1, \cdots, x_D) \in S^D$, which can be expressed as a perturbation of its center $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_D)$ (Pawlowsky-Glahn and Egozcue, 2002) and random deviations $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_D)$ from the center, so that $\boldsymbol{x} = \boldsymbol{\mu} \oplus \boldsymbol{\epsilon}$, then (3.4) can be rewritten as

$$\underline{f}(\boldsymbol{\mu} \oplus \boldsymbol{\epsilon}) \approx \underline{f}(\boldsymbol{\mu}) + \sum_{j=1}^{D} \ln(\epsilon_j) \left[\frac{\partial_{\mathcal{C}} \underline{f}}{\partial \mu_j}(\boldsymbol{\mu}) \right]. \tag{3.6}$$

One can proceed as in Section 3.2 to derive the variance of the components of $\underline{f}(\boldsymbol{\mu} \oplus \boldsymbol{\epsilon})$. Similar as for the Taylor expansion (3.2) from Section 3.2, also here the approximation is valid just for small perturbations. Moreover, in contrast to the previous case, the error is now multiplicative. Although this fits well with the nature of compositional data, particularly with their scale invariance, in practice error terms are often additive (van den Boogaart et al., 2015). This fact should be taken into account for an error propagation analysis of compositional data.

In case of ilr coordinates, however, the investigation of the error propagation simplifies. By considering (3.6) with ilr coordinate $ilr_i(\boldsymbol{x})$ as i-th component of $\underline{f}$

$$\frac{\partial_c ilr_i}{\partial \mu_j} = \begin{cases} 0 & \text{if } j < i, \\ \sqrt{\frac{D-i}{D-i+1}} & \text{if } j = i, \\ -\sqrt{\frac{D-i}{D-i+1}}\frac{1}{D-i} & \text{if } j > i, \end{cases} \tag{3.7}$$

where $i = 1, \ldots, D-1$. This corresponds exactly to a logcontrast (Aitchison, 1986) of the the i-th ilr coordinate of the compositional error $\boldsymbol{\epsilon}$, and consequently

$$ilr_i(\boldsymbol{x}) = ilr_i(\boldsymbol{\mu} \oplus \boldsymbol{\epsilon}) = ilr_i(\boldsymbol{\mu}) + ilr_i(\boldsymbol{\epsilon}),\ i = 1, \ldots, D-1.$$

In the context of error propagation this shows that the ilr coordinates are additive with respect to multiplicative errors. On the other hand, for other forms of errors a non-linear behavior can be expected. This issue is further investigated within the simulation study in Section 3.4.

In addition, this leads to an alternative verification of the linearity of ilr coordinates

$$\boldsymbol{z} = ilr(\boldsymbol{x}) = ilr(\boldsymbol{\mu} \oplus \boldsymbol{\epsilon}) = ilr(\boldsymbol{\mu}) + ilr(\boldsymbol{\epsilon}),$$

that is commonly shown directly with the definitions from Section 4.1. Even more, ilr coordinates represent an isometry, which means that all metric concepts in the simplex are maintained after taking the ilr coordinates (Pawlowsky-Glahn et al., 2015). The variance can now be considered component-wise, for example for the j-th component z_j of $\boldsymbol{z}$ one obtains

$$\mathrm{Var}(z_j) = \mathrm{Var}(ilr_j(\boldsymbol{x})) = \mathrm{Var}(ilr_j(\boldsymbol{\epsilon})).$$

This variance can be expressed by log-ratios of the compositional parts as shown in Fišerová and Hron (2011) as

$$\begin{aligned} \mathrm{Var}(z_j) &= A - B \quad \text{with} \\ A &= \frac{1}{D-j+1}\sum_{k=j+1}^{D} \mathrm{Var}\left(\ln\frac{\epsilon_j}{\epsilon_k}\right), \\ B &= \frac{1}{2(D-j)(D-j+1)}\sum_{k=j+1}^{D}\sum_{l=j+1}^{D} \mathrm{Var}\left(\ln\frac{\epsilon_k}{\epsilon_l}\right). \end{aligned} \tag{3.8}$$

The contributions of log-ratio variances in this linear combination are clearly higher for terms in A that include ϵ_j, and lower for terms in B where ϵ_j is not involved, and their magnitude depends on the number of parts D. In particular, if D is large and contamination (imprecision, detection limit problem) is expected only in one compositional part, the effect on the variance of z_j will be small. Note, however, that for a multivariate analysis the focus is in all coordinates $z_1, \ldots, z_{D-1}$ simultaneously, and thus it is not so straightforward to investigate the effect, since there may also be dependencies among the error terms. There is a simple exception: suppose that an error is to be expected only in log-ratios with one compositional part. From a practical perspective, it would then appear that only one compositional part is erroneous. If this part is taken as the first one, the ilr coordinates from Equation (4.4) will allow to assign this error exclusively to z_1, but not to the other coordinates.

Besides investigating the variance of the coordinates, it is also important to know how the errors affect distances between different compositions, that is between observations of a compositional data set, and how the multivariate data structure is affected. All these aspects will be investigated in more detail by simulations in the next section.

3.4 Simulation-Based Investigations of Error Propagation

For a simulation-based analysis of error propagation a real data set is used, namely the GEMAS data mentioned in Section 4.1, described in Reimann et al. (2014). More than 2,000 samples of agricultural soils have been analyzed in an area covering 5.6 million km^2 of Europe across 33 countries, and for the simulations the concentrations of the elements Al, Ba, Ca, Cr, Fe, K, Mg, Mn, Na, Nb, P, Pb, Rb, Si, Sr, Ti, V, Y, Zn and Zr are considered. Precision or detection limit problems of these elements are rather small or even not existing (Reimann et al., 2014), and thus these elements form a good base for carrying out simulations where contamination is artificially introduced in the form of imprecision and detection limit problems.

Denote the resulting compositional data matrix by $\boldsymbol{X}$, where the observations are forming the rows and the above mentioned compositional parts the columns. The number of observations is $n = 2107$, and the number of parts is $D = 20$. The cells of the matrix $\boldsymbol{X}$ are denoted as x_{ij}, for $i = 1, \ldots, n$ and $j = 1, \ldots, D$.

In the simulations problems with detection limit and imprecision are reproduced as follows:

- Detection limit (DL): Set all observations x_{ij} of the j-th part to the value

$$x_{ij}^{*} = \begin{cases} \frac{2}{3}\mathrm{DL}_j & \text{if } x_{ij} \leq \mathrm{DL}_j \\ x_{ij} & \text{otherwise ,} \end{cases} \tag{3.9}$$

  where $i = 1, \ldots, n$, and DL_j is taken as some quantile of that part.

- Imprecision rate (IR): A noise term ϵ_{ij} is added to each observation x_{ij}, where the noise depends on the magnitude of the observation and follows a uniform distribution. Thus, the values x_{ij}, $i = 1, \ldots, n$, are set to

  $$x^*_{ij} = x_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{U}(-\alpha_j x_{ij}, \alpha_j x_{ij}), \tag{3.10}$$

  where $\alpha_j > 0$ defines the imprecision rate of the j-th part, and the resulting simulated value x^*_{ij} must be positive. Note that this contamination is not additive but multiplicative, since

  $$x^*_{ij} = x_{ij}(1 + \gamma_j), \quad \gamma_j \sim \mathcal{U}(-\alpha_j, \alpha_j).$$

  Thus, this contamination scheme corresponds to the error model of the previous section, while contamination by a detection limit introduces a non-linear effect.

As mentioned previously, the main interest is the investigation of error propagation for ilr coordinates. If the i-th row of $\boldsymbol{X}$ is denoted by $\boldsymbol{x}_i$, then the ilr coordinates are obtained by Equation (4.4), leading to the values $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{i,D-1})$. The complete $n \times (D - 1)$ matrix of coordinates is denoted by $\boldsymbol{Z}$, with cells z_{ij}.

As an illustrative example the last 10 parts of the composition are picked, and contaminated with errors. A detection limit problem is imitated, by choosing DL_j as the 0.25-quantile in each of these components, and setting the values in these parts according to Equation (3.9). The results are shown in the left panels of Figure 3.2: the upper panel shows the first ilr coordinate of the original versus the contaminated data. One can see clear distortions in form of deviations from the main structure, but also in form of nonlinearities. For a clearer picture of the multivariate data structure the Mahalanobis distances of all ilr coordinates for the original and contaminated data are presented in the lower panel of Figure 3.2. The Mahalanobis distance (MD) for the i-th composition expressed in coordinates is

$$\mathrm{MD}(\boldsymbol{z}_i) = \sqrt{(\boldsymbol{z}_i - \boldsymbol{t}_{\boldsymbol{z}})' \boldsymbol{C}_{\boldsymbol{z}}^{-1} (\boldsymbol{z}_i - \boldsymbol{t}_{\boldsymbol{z}})}, \text{ for } i = 1, \ldots, n, \tag{3.11}$$

where $\boldsymbol{t}_{\boldsymbol{z}}$ and $\boldsymbol{C}_{\boldsymbol{z}}$ are robust estimators of location and covariance of the ilr coordinates $\boldsymbol{Z}$, respectively. For reasons of comparability the Mahalanobis distances for the contaminated data are computed with the estimators $\boldsymbol{t}_{\boldsymbol{z}}$ and $\boldsymbol{C}_{\boldsymbol{z}}$ based on the uncontaminated data. Plugging in robust estimators is essential since they guarantee that the Mahalanobis distance estimation is not spoiled by single outliers but based on the data majority. For this purpose the minimum covariance determinant (MCD) estimator is used (Rousseeuw and Van Driessen, 1999) .

The right panel of Figure 3.2 shows the results of a simulated precision problem. Again, the last ten parts are contaminated, α_j is set to 0.25 for these parts, and Equation (3.10) is applied. The upper panel compares the first ilr coordinates for the original and distorted data. Since the contamination is symmetric in each part, the outcome is also

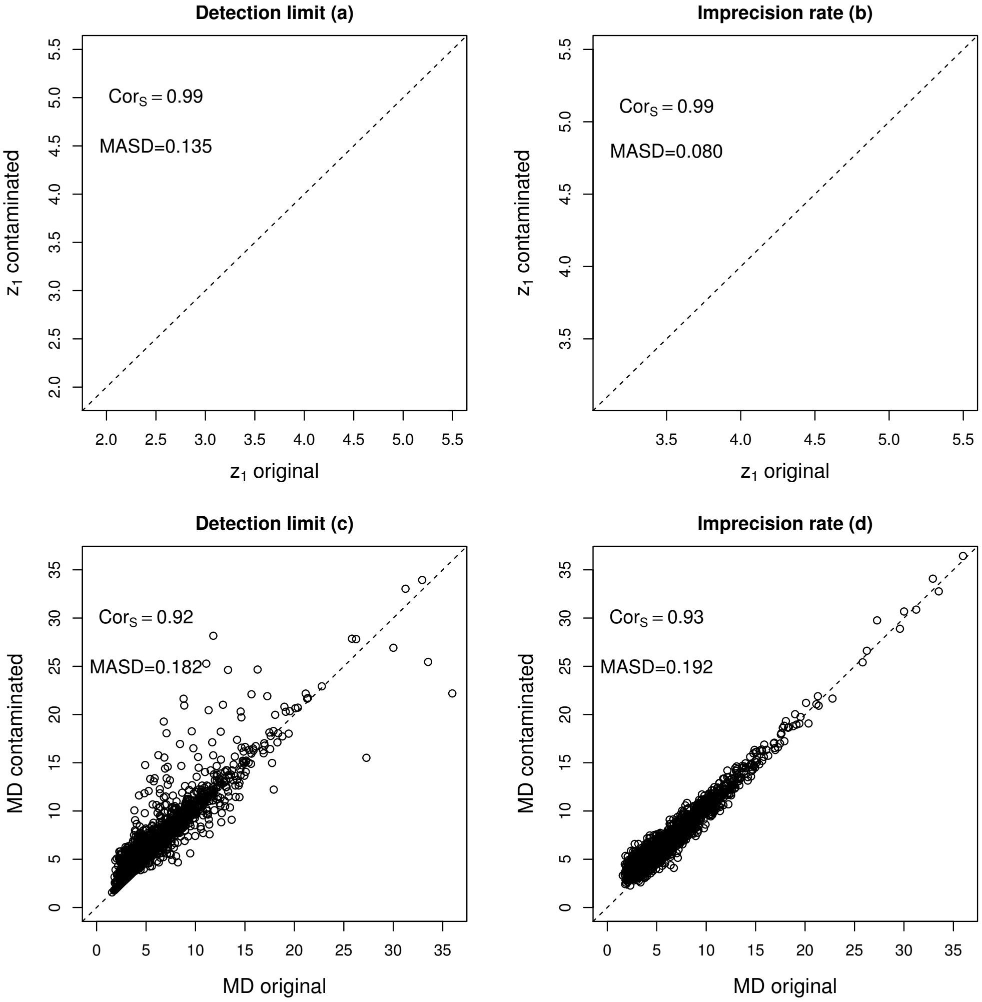

Figure 3.2: Effect of the DL and IR contamination on the first ilr coordinate ((a) and (b)), and on all ilr coordinates jointly ((c) and (d))

relatively symmetric around the line of 45 degrees. The comparison of the Mahalanobis distance shows that those distances for the contaminated data increase in general.

The above example already provides an idea about possible choices of measures for quantifying the resulting error. The focus is on the first ilr coordinate as well as on all coordinates jointly in terms of Mahalanobis distances, and the original data will be compared with the contaminated data.

Denote the values of the first ilr coordinate by $\boldsymbol{z}_{.1} = (z_{11}, \ldots, z_{n1})$, and the corresponding contaminated version by $\boldsymbol{z}^*_{.1} = (z^*_{11}, \ldots, z^*_{n1})$. The two vectors are compared by:

- Spearman rank correlation, expressed as

$$\mathrm{Cor}_S(\boldsymbol{z}_{.1}, \boldsymbol{z}^*_{.1}) = \frac{\mathrm{Cov}(R(\boldsymbol{z}_{.1}), R(\boldsymbol{z}^*_{.1}))}{\sqrt{\mathrm{Var}(R(\boldsymbol{z}_{.1}))}\sqrt{\mathrm{Var}(R(\boldsymbol{z}^*_{.1}))}}, \tag{3.12}$$

  where $R(\cdot)$ gives the ranks of its argument vector.

- Mean absolute scaled deviation (MASD), defined as

$$\mathrm{MASD}(\boldsymbol{z}_{.1}, \boldsymbol{z}^*_{.1}) = \frac{1}{n}\sum_{i=1}^{n} \frac{|z_{i1} - z^*_{i1}|}{\sqrt{\mathrm{Var}(\boldsymbol{z}_{.1})}}. \tag{3.13}$$

The Spearman rank correlation coefficient measures the monotone relation between the uncontaminated and contaminated coordinates; a value of one would refer to the same ordering of the values of the coordinates. On the other hand, MASD is more strict and evaluates the error in reproducing the values of the coordinate. Note that the scaling in MASD by the variance is used to allow for a comparison of the corresponding first ilr coordinates if the parts in the data matrix are permuted.

Similar measures for comparison are proposed in the multivariate case. Denote by $\mathrm{MD}(\boldsymbol{Z})$ the vector of the Mahalanobis distances $\mathrm{MD}(\boldsymbol{z}_i)$, for $i = 1, \ldots, n$, see Equation (3.11), and by $\mathrm{MD}(\boldsymbol{Z}^*)$ the corresponding contaminated version, with entries $\mathrm{MD}(\boldsymbol{z}^*_i)$. Then the Spearman rank correlation coefficient $\mathrm{Cor}_S(\mathrm{MD}(\boldsymbol{Z}), \mathrm{MD}(\boldsymbol{Z}^*))$ investigates if the overall ordering in the multivariate data structure, represented in coordinates, is maintained. A mean absolute scaled deviation (MASD) measure relates to the Mahalanobis distances

$$\mathrm{MASD}(\mathrm{MD}(\boldsymbol{Z}), \mathrm{MD}(\boldsymbol{Z}^*)) = \frac{1}{n}\sum_{i=1}^{n} \frac{|\mathrm{MD}(\boldsymbol{z}_i) - \mathrm{MD}(\boldsymbol{z}^*_i)|}{Q_{0.5}(\mathrm{MD}(\boldsymbol{Z}))}. \tag{3.14}$$

The scaling is done by the 0.5-quantile (median) of the Mahalanobis distances of $\boldsymbol{Z}$ in order to allow for comparability of subcompositions with different numbers of parts. This measure thus indicates the error in reproducing the multivariate data structure. As mentioned previously, the Mahalanobis distances $\mathrm{MD}(\boldsymbol{Z}^*)$ are based on the estimates of location $\boldsymbol{t}_{\boldsymbol{z}}$ and covariance $\boldsymbol{C}_{\boldsymbol{z}}$ of the matrix $\boldsymbol{Z}$, see Equation (3.11), leading to a MASD value of zero for observations which have not been changed.

These measures have been computed for the example shown in Figure 3.2 in order to get an idea about the meaning of the magnitude of these values. The Spearman rank correlation is in all cases clearly above 0.9, in spite of the deviations of some points. The scaled distances MASD for the first ilr coordinates are lower that those for all coordinates jointly (Mahalanobis distances).

3.4.1 Simulation 1: One Uncontaminated, 1 to 19 Contaminated Parts

Start with the first column $\boldsymbol{x}_{.1}$ of the composition $\boldsymbol{X}$, and add step-by-step another column. After the $(k-1)$-st step one ends up with the subcomposition $\boldsymbol{X}_k = (\boldsymbol{x}_{.1}, \boldsymbol{x}_{.2}, \ldots, \boldsymbol{x}_{.k})$, where $k = 2, \ldots, 20$. A contaminated version is generated by contaminating all parts except the first one; this yields $\boldsymbol{X}_k^* = (\boldsymbol{x}_{.1}, \boldsymbol{x}_{.2}^*, \ldots, \boldsymbol{x}_{.k}^*)$. Then the ilr coordinates are computed from $\boldsymbol{X}_k$ and $\boldsymbol{X}_k^*$, and the measures Cor_S and MASD are calculated for the first coordinates and for all coordinates jointly in terms of Mahalanobis distances.

The number of simulation replications is 100. In each replication, the parts of the original composition are permuted. In that way, the first (uncontaminated) part changes, but also the sequence of the parts that are added changes. All simulations are done for the contamination in form of detection limit (DL) and for imprecision (IR). In the first case, the value DL_j of the detection limit is taken as the 0.25-quantile, see Equation (3.9), while in the latter case the imprecision rate is taken as $\alpha_j = 0.25$, see Equation (3.10).

The results are presented by boxplots in Figure 3.3. The left panels show the outcome for the detection limit simulations, the right panels show the results of the imprecision simulations. The upper figures show the comparison of original versus contaminated versions in terms of Spearman correlations, while the lower figures compare in terms of MASD. The grey boxplots compare the first ilr coordinates, while the white boxplots summarize the Mahalanobis distances of all joint coordinates. The plots allow to compare the impact of an increasing number of contaminated parts (horizontal axes). Although the amount of contamination is quite high, the correlations reveal that the covariance structure of the multivariate data is basically preserved. In particular, the comparison of the first ilr coordinates leads to a remarkably high correlation, which is quite stable with an increasing number of parts (for DL), and even improving in case of IR. This means that additional parts coupled with a symmetric contamination scheme, as in case of IR, still provide important and useful information that stabilizes the first ilr coordinate. The MASD results for the first ilr coordinate are quite stable in case of DL, while in the IR case with increasing number of parts an improvement is observed.

The picture is somewhat different when comparing all ilr coordinates jointly. The Spearman correlation is clearly lower, and it gets more stable with an increasing number of parts. In case of DL, the MASD measure is nearly constant with an increasing number of parts, while for IR first a decline is observed, but then a clear increase. It is, however, surprising that the Mahalanobis distances do not change more drastically, given that the amount of contamination is relatively high.

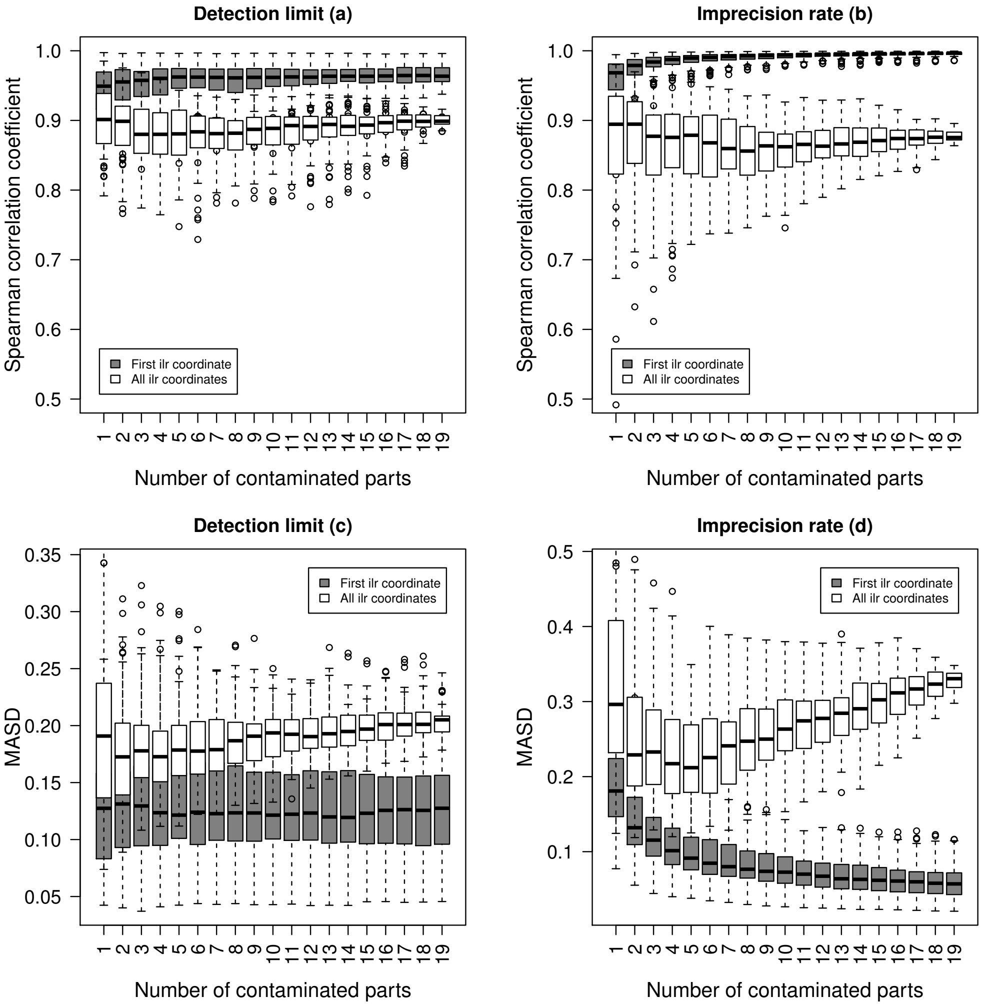

Figure 3.3: One uncontaminated part, and 1 to 19 contaminated parts added. Univariate and multivariate structural changes between the original and contaminated ilr coordinates with increasing number of contaminated parts in case of DL ((a) and (c)) and IR ((b) and (d))

3.4.2 Simulation 2: 10 Uncontaminated, 1 to 10 Contaminated Parts

In a further simulation experiment a block of 10 compositional parts is fixed and left uncontaminated. Step by step a contaminated part is added, until all 10 remaining (contaminated) parts have been included. The comparison is done in the same way as before. The simulation is repeated 100 times, and the parts are randomly permuted for each replication. Thus, the uncontaminated block changes, but also the contaminated parts differ from simulation to simulation. The results are shown in Figure 3.4.

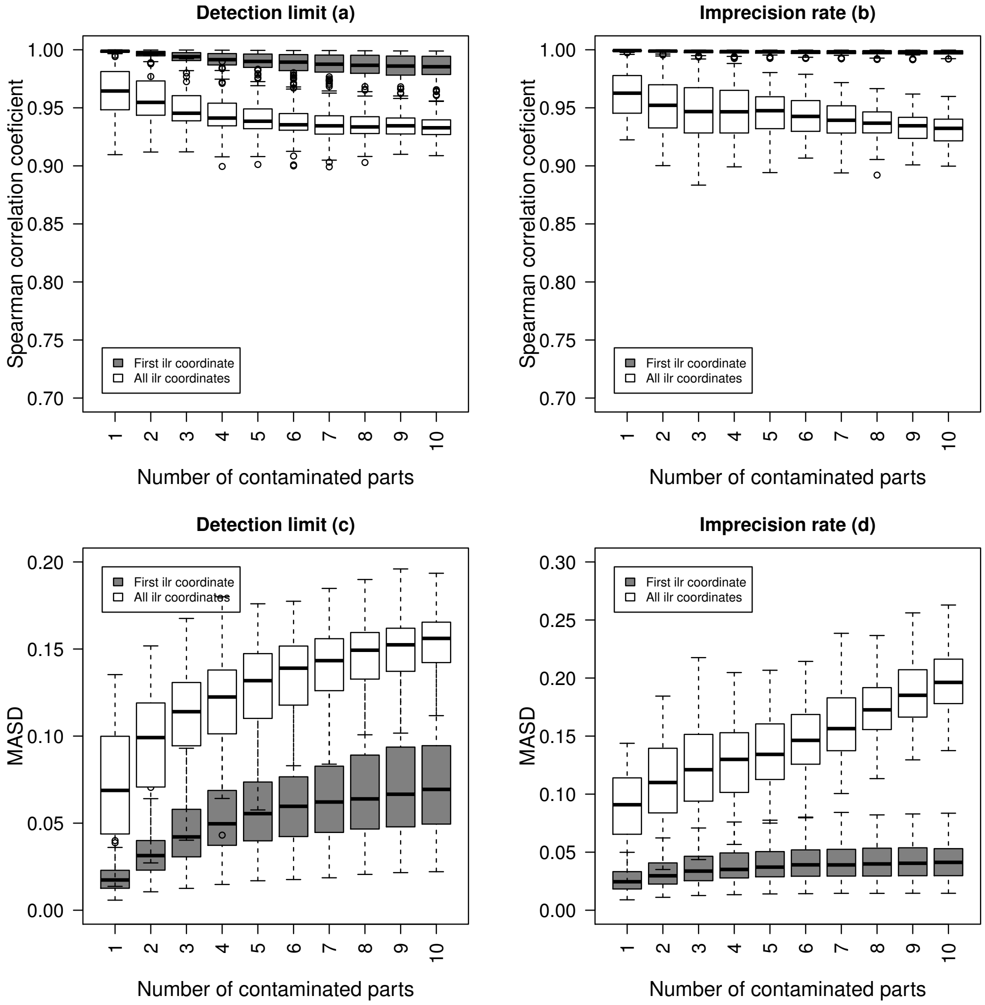

Figure 3.4: Ten uncontaminated parts, and 1 to 10 contaminated parts added. Univariate and multivariate structural changes between the original and contaminated ilr coordinates with increasing number of contaminated parts in case of DL ((a) and (c)) and IR ((b) and (d))

Basically, a similar impression can be observed as in Figure 3.3. For the first ilr coordinates, the correlations are now very close to one, and the values of MASD, although increasing slightly with increasing number of contaminated parts, are close to zero. So, having good data quality for a major part of the data set is a good protection against poor data quality in additional parts - at least for the first ilr coordinate. The multivariate data structure is well maintained in terms of ordering, expressed by the Spearman rank correlations, which are still clearly above 0.9. The MASD values for the Mahalanobis distances now increase for DL as well as for IR, with an increasing number of parts, but they are lower than in the previous simulation.

3.4.3 Simulation 3: Changing the Amount of Contamination

In the previous simulations the amount of contamination is fixed. Here the effect of changing the amount of contamination is investigated. For that purpose, 10 parts are selected randomly to leave them uncontaminated, while the remaining 10 parts are contaminated by the same amount: in case of DL contamination, the value DL_j is varied from the 0.05-quantile to the 0.95-quantile; for IR contamination, the imprecision rate α_j is varied from 0.05 to 0.95. Note that the imprecision in real studies can be much higher, in particular for small concentrations (Reimann et al., 2014). Figure 3.5 summarizes the outcome of the simulations, where again 100 replications were performed.

The resistance against contamination of the first ilr coordinate is remarkable. Both, the correlation and the MASD report relatively small deviations, even for very high amounts of contamination. Contamination according to DL has more effect than that based on imprecision. This is different when looking at the multivariate data structure, expressed by the joint coordinates. The correlations get severely low, and also the MASD increases rapidly. The effect for IR contamination is more severe than that for DL. A MASD value of one means that the average change of the Mahalanobis distances before and after contamination is as large as the median Mahalanobis distance, and thus this would correspond to a substantial change in the multivariate data structure.

3.4.4 Simulation 4: Changing the Number of Observations

In a final simulation the effect of the number of observations in the data set, which has been fixed before with all available observations (i.e., more than 2,000) is analyzed. As before, 10 parts are randomly selected and not modified, and the remaining 10 parts are contaminated at a level of 25%, that is for DL contamination 25% of values below detection limit in each of these parts, and for IR contamination $\alpha_j = 0.25$ for these parts. The results in Figure 3.6 for the 100 simulations show that there is no visible effect for the first ilr coordinate. However, the multivariate structure suffers severely if the number of observations is smaller than 100.

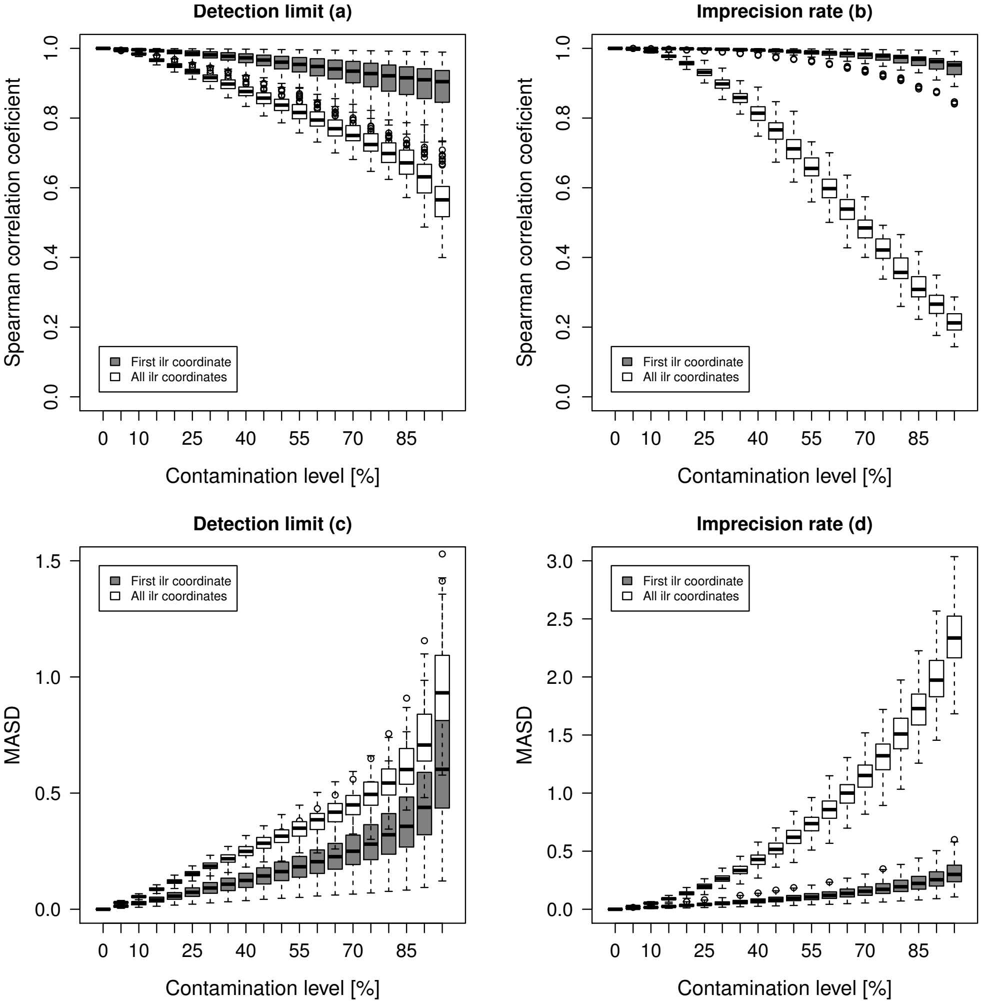

Figure 3.5: Ten uncontaminated parts, and 10 contaminated parts added. Univariate and multivariate structural changes between the original and contaminated ilr coordinates with increasing amount of contamination in case of DL ((a) and (c)) and IR ((b) and (d))

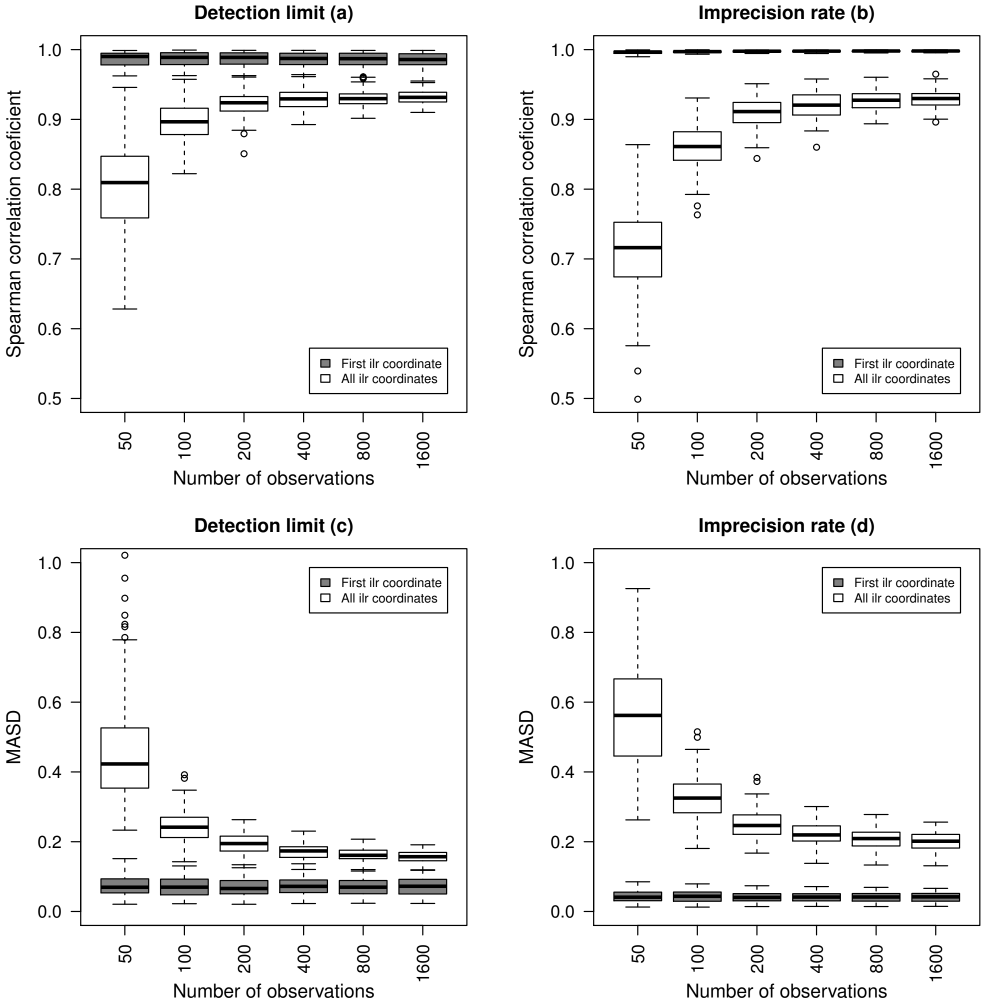

Figure 3.6: Ten uncontaminated parts, and 10 contaminated parts added. Univariate and multivariate structural changes between the original and contaminated ilr coordinates with varying number of observations in the data set; the contamination level is fixed with 25%; DL ((a) and (c)) and IR ((b) and (d))

3.5 Discussion and Conclusions

To many practitioners it looks almost obvious that geometric means, as they are used in log-ratio approaches, may cause instabilities due to the involved products of the data values. Even worse, measurement errors could be propagated by the use of geometric means. This problem is investigated in more detail, by focusing on the most important log-ratio approach based on ilr coordinates (Pawlowsky-Glahn and Buccianti, 2011; Pawlowsky-Glahn et al., 2015).

In a first attempt, the classical theory of error propagation has been formulated on the simplex, the sample space of compositional data. While this gets complex if any transformation function would be considered, the results are straightforward when using ilr coordinates because of their linearity. It has been shown that the variance of an ilr coordinate is just the variance of the same ilr coordinate of the random deviations from the center. Using Equation (3.8) it can be seen which terms contribute by which magnitude to this variance. For non-linear contamination schemes, these variance contributions cannot be computed from the random errors, but they have to be computed directly from the ilr coordinate. This has been done for the simulation scheme outlined in Section 3.4.1 for the first ilr coordinates z_1 of the uncontaminated data, the data contaminated by a detection limit, and contaminated by the imprecision rate. The resulting variance contributions are shown in Figure 3.7 in the form of ratios A/B according to Equation (3.8) as non-colored boxplots. With increasing number of parts, the term B (which does not involve variance contributions with log-ratios to x_1) gets more dominant. This can be seen in the uncontaminated case, as well as in the contaminated cases due to the inherent variability contained within the log-ratios of the remaining parts. Interestingly, detection limit contamination has almost no effect on the variance contributions A and B when compared to the uncontaminated case. This is also shown by the dark boxplots which represent the ratios of A-contaminated to A-uncontaminated. Only for contamination by the imprecision rate, the variance contributions are clearly higher compared to the uncontaminated case if the number of contaminated parts is low. For higher numbers of contaminated parts, the variance contributions are about the same.

Further investigations have been carried out through simulation experiments. The contamination is studied in terms of mimicking a detection limit problem, and in terms of imprecision in form of a multiplicative factor. In all experiments it turned out that the structure of the first ilr coordinate can almost not be destroyed with poor data quality, except in case of extremely high amounts of contamination. This is an interesting outcome, since due to the proposed formula (4.4) to derive the ilr coordinates, the first coordinates describes all relative information about the first compositional part (Fišerová and Hron, 2011). Clearly, if the main interest is not in the first but in another part, then this part is simply put to the first position. Note that the first coordinate is proportional to the corresponding centered log-ratio (clr) coefficient (Aitchison, 1986) for this part (Fišerová and Hron, 2011). Practitioners often explore just the structure of the resulting clr coefficients. For example, one can study the clr coefficients for the different chemical elements in maps, which is the compositional alternative to the traditional maps based

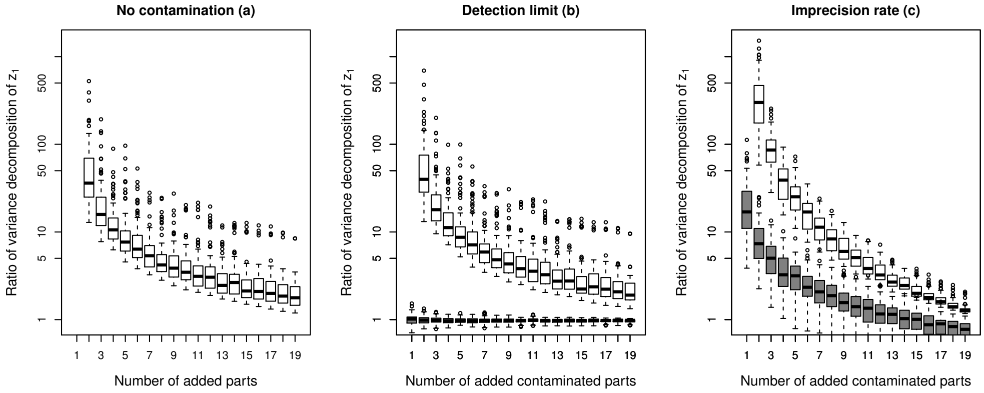

Figure 3.7: Variance decomposition results: The light boxplots show the ratios A/B according to Equation (3.8) for the simulation scheme in Section 3.4.1. The dark boxplots compare the ratio of A-contaminated (either DL or IR) to A-uncontaminated. No contamination (a), detection limit (b), imprecision rate (c)

on the absolute concentrations. Examples are shown in Reimann et al. (2014).

It is not studied, how the contamination of the first part effects the first ilr coordinate (z_1), because it is clear that the contamination would be immediately reflected in the first ilr coordinate, and any additional contamination in other parts would make things worse. Hence, variations of z_1 are only due to variations of $(x_2, ..., x_D)$. It is therefore quite logical that the impact of DL or IR on z_1 remains limited, and that its growth decreases as D increases, due to compensation effects when computing $g_m(x_2, ...x_D)$.

Especially when applying multivariate statistical methods, like principal component analysis or discriminant analysis, all ilr coordinates have to be analyzed jointly. Therefore, the effect of errors on the multivariate data structure are also investigated in the simulations. It depends very much on the setting if the multivariate data structure is destroyed by the contamination or not. If dimension increases, the effects of the contamination generally increase. It depends a lot on the contamination level if the multivariate data structure after contamination is still closely related to that before, but this also depends on the sample size of the data: higher numbers of observation (e.g., at least 100 in the data set used here) stabilize the results.

Consider again the example shown in Figure 3.2, where 10 parts out of 20 have been contaminated at a level of 25%. Here the DL contamination scheme is considered. Figure 3.8 shows the biplot for the first two principal components (PCs): left panels for the uncontaminated data, right panels for the contaminated data. A comparison is also done with robust PCs (Filzmoser et al., 2009b), which are shown at the lower panels. While there is almost no difference visible between the uncontaminated and contaminated versions, there is a clear difference in the outcome of classical and robust principal component analysis. This shows that, although the MASD is around 0.18 (Figure 3.2(c)), the outliers that are present in the data have a much stronger effect than the artificial contamination used here.

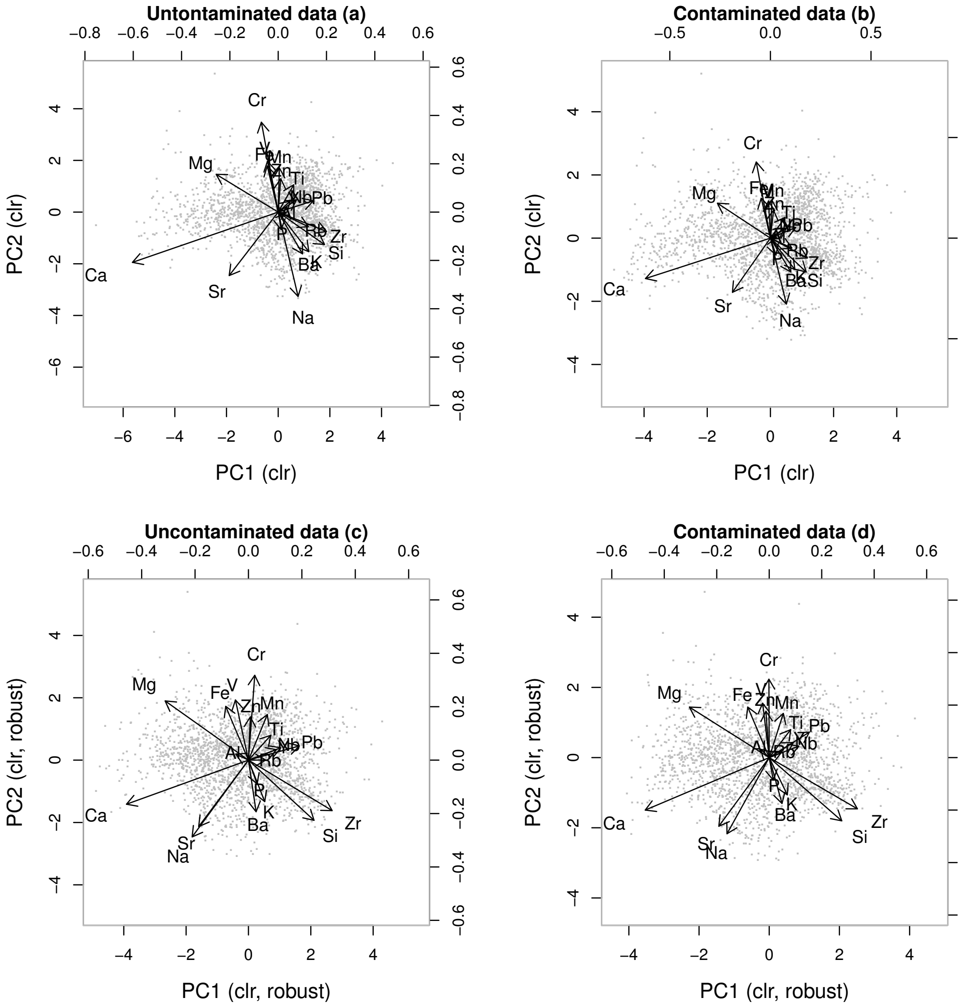

Figure 3.8: Biplots of the first two PCs based on the data shown in Figure 3.2. Classical PCs of uncontaminated data (a), classical PCs of contaminated data (b), robust PCs of uncontaminated data (c), robust PCs of contaminated data (d)

The overall conclusion of this paper is not that one does not have to care anymore about data quality issues. In contrary, good data quality is the basis of any sound statistical analysis. Rather, it should provide an answer to researchers who have a data set available, and who carefully think about which compositional parts to include in the analysis. Often it is known which parts have precision problems, and sometimes even the level of imprecision is known. Also, the amount of values below detection is known. Including such parts with moderate quality in the analysis will in general not have a major effect on a single (the first) ilr variable, and the effects will also be limited in general for the multivariate data structure.

The point why one should consider including as much information as possible in the analysis is because the reliable values of such parts with moderate data quality also contribute to the log-ratio analysis and they might contain important and relevant information.

Acknowledgments

This work has been partly funded by the Austrian Science Fund (FWF), project I 1910-N26, by the grant COST Action CRoNoS IC1408, and by the K-project DEXHELPP through COMET - Competence Centers for Excellent Technologies, supported by BMVIT, BMWFI and the province Vienna. The COMET program is administrated by FFG.

CHAPTER 4

Compositional Data Analysis in Epidemiology

Abstract: Compositional data analysis refers to analyzing relative information, based on ratios between the variables in a data set. Data from epidemiology are usually treated as absolute information in an analysis. We outline the differences in both approaches for univariate and multivariate statistical analyses, using illustrative data sets from Austrian districts. Not only the results of the analyses can differ, but in particular the interpretation differs. It is demonstrated that the compositional data analysis approach leads to new and interesting insights.

Key words: Log-ratio approach; multivariate statistics; Euclidean geometry; compositional data; isometric log-ratio coordinates

4.1 Introduction

Health care research is one of the most important research fields in epidemiology with high benefits to the society. This research covers a variety of topics, such as disease trends, risk factors, or structural and regional changes in public health. Health care data are generally perceived as being ”information rich” but still ”knowledge poor”. (Lincoln and Builder, 1999) Nevertheless, statistical methods are used more and more often to discover patterns and trends of such data and to support decision makers. In most cases, health care data are measured as count data, for example the numbers of deaths caused by a disease, or the numbers of patients affected by a disease. Often epidemiologic information is represented in form of incidence, prevalence, morbidity, or analyzed as raw data.

Let us consider a data matrix $\boldsymbol{X}$ with n rows and D columns, where n denotes the number of observations (e.g. the number of districts), and D stands for the number of variables being investigated. The variables could refer to specific causes of death or to certain diseases. The i-th row of the matrix is represented by $\boldsymbol{x}_i$ and the cells of the matrix are denoted as x_{ij}, where $i = 1, \ldots, n$ and $j = 1, \ldots, D$. Here we expect that the data are already standardized with regard to geographic and sociodemographic (e.g. age, population size) attributes, in order to allow for a comparison of data subgroups.(Hennekens and Buring, 1987)

Depending on the focus of the analysis, the raw data $\boldsymbol{X}$ can be used, or this information needs to be preprocessed or aggregated. However, the raw data are not always expedient to analyze when the focus is on comparing observations or groups of observations based on different scales. For example, the number of diseases could be generally higher in one region than in another region, and thus a specific disease is not comparable across the regions, e.g. due to different population size. In this case, the data can be normalized by dividing each observation by the corresponding row sum. More formally, for a given observation $\boldsymbol{x}_i$ the proportion of the j-th variable to the sum of all variables is denoted by p_{ij} and defined as

$$p_{ij} = \frac{x_{ij}}{\sum_{j=1}^{D} x_{ij}}. \tag{4.1}$$

Clearly, $p_{i1} + \ldots + p_{iD} = 1$, for $i = 1, \ldots, n$, which allows for a numerical comparison of the observations. For a fixed j, the multivariate information is thus expressed as a univariate quantity. However, the denominator might suppress and hide relevant information. For instance, the data can be dominated by one or more variables with high values, which leads to small proportions for all observations and to a loss of possible patterns and data structures.

There are yet other problems with proportional data. It is often not so clear in a study, which variables can and should be included in a data set. For example, the set of diseases to be considered in an epidemiological study is usually not clearly defined. If one is interested in investigating the relation between two variables, expressed as proportions, this relation can be driven by the total sum, which is just depending on the considered variables in the study. This has been noted already by Pearson Pearson (1897), who talked about "spurious" correlations in this context, i.e., correlations which are depending on the choice of the variables in the analysis.

An alternative to working with proportional data is to consider ratios of variable pairs, or even better, logarithms of ratios (log-ratios), $\ln \frac{x_{ij}}{x_{ik}}$, $i = 1, \ldots, n$ and $j, k \in \{1, \ldots, D\}$. A positive value of the log-ratio between the j-th and the k-th variable refers to the dominance of the j-th variable (nominator) over the k-th variable (denominator). By taking pairwise log-ratios, the variance of the resulting new variable does not depend on which variable is used for the nominator (denominator). The result does also not depend on which variables are included in the whole data set – this would just lead to a different number of possible pairwise log-ratios. A further advantage is that the pairwise log-ratio does not depend on the units in which the variables are expressed: the multiplication of a row of the original data matrix by an arbitrary positive number does not alter the pairwise log-ratios. This principle is called *scale invariance*, and the approach of working with log-ratios has been introduced in detail in Aitchison's work.Aitchison (1986)

The goal of this paper is to analyze the possibilities of the log-ratio approach in the context of epidemiology. We will make use of the original ideas introduced in the book of AitchisonAitchison (1986), but focus more on the recent research in this field.(Pawlowsky-Glahn et al., 2015) Generally, the log-ratio approach turned out to be suitable for analyzing so-called *compositional data*, which are representing parts of some "whole". This definition frequently leads to the wrong assumption that the "whole" must refer to a constant sum of the data, such as the sum of the individual causes of deaths needs to sum up to the total number of deaths. However, as mentioned above, the log-ratio approach is scale invariant, and by considering only a subset of the variables (subcomposition), one will get coherent results compared to those from the full composition. (Pawlowsky-Glahn et al., 2015) The log-ratio approach in the context of epidemiology is rarely used. Exceptions are the articles by Leite(Leite, 2014), where its usefulness in the area of nutritional epidemiology has been demonstrated, and by Tsilimigras and Fodor(Tsilimigras and Fodor, 2016), where the advantages and challenges of using the log-ratio approach in microbiome studies has been represented.

This paper is structured as follows. In Section 4.2 we will introduce some basic concepts of the log-ratio approach. Section 4.3 illustrates for a specific data set how the log-ratio approach can be used for analyzing single quantities of interest. Section 4.4 uses another data set to show how a joint analysis of the variables can be made, following the log-ratio approach. Section 4.5 demonstrates another approach from compositional data analysis, namely regression modeling. For illustrating this approach, both data sets from the previous sections are used. In all our analyses we put emphasis on the use of robust statistical methods, since data from epidemiological practice frequently do not meet classical requirements such as normal distribution, or may contain outliers. The final Section 4.6 summarizes and concludes.

4.2 The log-ratio approach

The basic sources of information for the log-ratio approach are pairwise log-ratios, as introduced in the previous section. If the interest of the analysis is in one particular variable, one could think of computing all pairwise log-ratios of this variable to each of the remaining variables, and to analyze each of the resulting quantities separately. For small datasets that might be an intuitive approach, however, with an increasing number of variables the effort increases linearly and the general view of the data structure can get lost. For the purpose of simplicity one can consider the arithmetic mean of all pairwise log-ratios for the j-th variable of observation $\boldsymbol{x}_i$,

$$\frac{1}{D}\left(\ln\frac{x_{ij}}{x_{i1}}+\ldots+\ln\frac{x_{ij}}{x_{ij}}+\ldots+\ln\frac{x_{ij}}{x_{iD}}\right)=\ln\frac{x_{ij}}{\sqrt[D]{\prod_{k=1}^{D}x_{ik}}}, \tag{4.2}$$

where $i = 1, \ldots, n$ and $j \in \{1, \ldots, D\}$. Clearly, the term $\ln \frac{x_{ij}}{x_{ij}}$ is zero. The right-hand side of Equation (4.2) is found in the usual definition of the so-called *centered log-ratio* (clr) transformation (Aitchison, 1986), defined as

$$\boldsymbol{y}_i = clr(\boldsymbol{x}_i) = (y_{i1}, y_{i2}, \ldots, y_{iD})' = \left(\ln \frac{x_{i1}}{\sqrt[D]{\prod_{k=1}^{D} x_{ik}}}, \ldots, \ln \frac{x_{iD}}{\sqrt[D]{\prod_{k=1}^{D} x_{ik}}} \right)'. \tag{4.3}$$

The clr transformation is an isometric mapping between the D-part simplex sample space $\mathcal{S}^D$, the sample space of compositional data with D compositional parts (variables), and the D-dimensional real space $\mathbb{R}^D$. From Equation (4.2) it can be seen that the j-th clr coefficient y_{ij}, for $j \in \{1, \ldots, D\}$, contains all relative information of the j-th variable to the remaining variables, in terms of averaged log-ratios. This allows the analyst to consider a different viewpoint of the problem: The question might not be whether the data at hand “are compositional data”, where an appropriate transformation needs to be performed before the statistical analysis is carried out, but rather if it is desirable to analyze “relative information”. As mentioned in Section 4.1, pairwise log-ratios as the key ingredients of the clr coefficients have more attractive properties (depending on the application) than simple proportions, and thus they can avoid undesirable effects like spurious correlations (Pearson, 1897) or dependence on scale.

The interpretation of the clr coefficients is straightforward, since they represent all relative information of one compositional part to the others. However, the sum of those coefficients is zero, $y_{i1} + y_{i2} + \cdots + y_{iD} = 0$, for $i = 1, \ldots, n$, resulting in data collinearity due to the constant sum constraint. For some statistical methods and estimators, especially with robust procedures (Maronna et al., 2006), this data collinearity leads to numerical and sometimes conceptual problems. Robust statistical methods give reliable results even in presence of outliers or deviations from strict model assumptions such as normal distribution, and thus they are considered as very useful for health care data and studies in epidemiology.(Beaumont et al., 2006)

The collinearity problem of clr coefficients can be avoided by the *isometric log-ratio* (ilr) transformation, an isometric mapping from $\mathcal{S}^D$ to the real space $\mathbb{R}^{D-1}$.(Egozcue et al., 2003a) The idea is to construct an orthonormal basis, so-called ilr coordinates, in the $D-1$-dimensional hyperplane formed by the clr coefficients. Clearly, only $D-1$ coordinates are necessary, and there are infinitely many possibilities to construct this coordinate system. A specific choice is (Fišerová and Hron, 2011)

$$z_{ij} = ilr_j(\boldsymbol{x}_i) = \sqrt{\frac{D-j}{D-j+1}} \ln \left(\frac{x_{ij}}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_{ik}}} \right), j = 1, \ldots, D-1, \tag{4.4}$$

which is the j-th ilr coordinate of observation $\boldsymbol{x}_i$. Then $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{i,D-1})'$ is the representation of $\boldsymbol{x}_i$ in this orthonormal basis.

There is a reason why Equation (4.4) has been proposed to construct the ilr coordinates: It can be shown that

$$y_{i1} = \frac{D-1}{D} z_{i1}, \tag{4.5}$$

for $i = 1, \ldots, n$ (Fišerová and Hron, 2011), and thus the first ilr coordinate is proportional to the first clr coefficient. Therefore, y_{i1} and z_{i1} have the same interpretation. However, different to the clr case, none of $z_{i2}, \ldots, z_{i,D-1}$ contains any information about the first part. Thus, this choice of ilr coordinates makes it possible to fully isolate all the relative information in the composition or subcomposition considered about the first part into the first ilr coordinate. If one is interested in relative information about another compositional part, one simply needs to put this part on the first position and construct the ilr coordinates from Equation (4.4).(Fišerová and Hron, 2011)

For epidemiological data it may be possible to use expert knowledge for defining groups of compositional parts that relate to certain diseases, for instance. In this case one can use so-called *balances*. This alternative approach to construct coordinates of an orthonormal basis, expresses the relative information between the groups of parts in a composition. Balances can be constructed by a procedure called *sequential binary partition* (SBP).(Egozcue and Pawlowsky-Glahn, 2005; Pawlowsky-Glahn and Buccianti, 2011)

4.3 Analyzing single epidemiological variables

In this section, the concepts of compositional data analysis are illustrated with a data set from Austria, describing six causes of death, see Table 4.1. The data were recorded in 2007, obtained in terms of absolute frequencies from Statistics Austria, the Austrian institution for official statistics, and aggregated (summed up) to the 99 Austrian districts. Note that the districts of Vienna are considered together as one main district. Since the data depend mainly on the regional age structure of the population, and on the high variation of the regional population size, they were standardized by the overall age distribution of the world standard population given by the World Health Organization (WHO) as underlying distribution(Hennekens and Buring, 1987), and by dividing by the number of residents in the corresponding district. Following this, the rates are calculated per 10.000 population, resulting in an absolute *risk rate* per 10.000. This is our data matrix $\boldsymbol{X}$ with $n = 99$ rows and $D = 6$ columns. It should be noted that the data are collected from reports of medical doctors and hospitals, and that the cause of death of a person can not always be assigned uniquely to one of the codes reported in Table 4.1, or that the assignment procedure might not be unique among all hospitals. Thus, a certain degree of robustness of the analysis will be appreciated.

We are interested in analyzing the spatial distribution of single death causes. The main purpose is not to provide deep reasonings and rigorous interpretations of regional differences, but rather to compare different methods, including compositional approaches, in an explanatory way. We focus here on the risk of deaths caused by circulatory systems, see Table 4.1. Figure 4.1 shows the regional distribution of the preprocessed raw data information x_{ij} (upper plot) and proportional information p_{ij}, see Equation (4.1) (lower plot). In all these map presentations we use a scaling according to the quantiles $0.05, 0.25, 0.50, 0.75$ and 0.95 of the distribution. The upper figure shows a clear east-west

ICD10	Group
C00-C97	Neoplasms
I00-I99	Circulatory systems
J00-J99	Respiratory systems
K00-K93	Digestive systems
V01-Y89	Injuries
A00-B99, D01-H95,L00-R99	Other diseases

Table 4.1: Causes of deaths by aggregated disease groups.

trend, with higher risk rate caused by circulatory systems diseases in the east and lower rate in the west of Austria. This can be valuable information for decision makers, with a possible consequence of trying to improve the situation of circulatory diseases in eastern Austria. A different picture is provided by the proportional information in the lower figure, where the east-west gradient has disappeared. We observe some districts such as Bludenz (BZ) in the west, and Waidhofen an der Thaya (WT) in the north, with a high proportion of the death rate caused by circulatory systems, despite their low absolute risk rate. Obviously, the differences in both presentations are caused by the sum of all death rates, which also varies across the regions. So, even if the risk is small for circulatory systems, it can be relatively high as proportion on all risk rates, if the sum is also small. The latter happens if all risk rates are small compared to the other districts. This gives a different view, and it now seems to make sense to invest in regions with high proportional risk rate, rather than with high absolute risk rate.

Figure 4.2 shows analyses based on the log-ratio approach. The upper plot presents the pairwise log-ratio of the risk rates regarding circulatory systems to respiratory systems. Since a value of zero would correspond to equal risk, we conclude that the risk rate for circulatory systems is generally higher and thus dominating. Some regional patterns are visible, such as the cluster of high values in the north-east of Austria. This analysis is not yet very informative, since the log-ratios to the other causes of death need to be analyzed in order to get a complete picture of the risk rates for circulatory systems. The lower plot of Figure 4.2 shows this information in terms of the ilr coordinate for the variable circulatory systems. For this purpose, the variable circulatory systems has to be put to the first position before Equation (4.4) is applied. So, this plot summarizes all relative information of cause of death by circulatory systems to the remaining causes of death considered here. Since the values are mostly positive, cause of death by circulatory systems is dominating, but varying in the area. Some regional patterns appear, such as low values in all districts of Tyrol (with the exception of Reutte (RE)). Since all pairwise log-ratios to the variable circulatory systems are taken into account in this analysis, the source of information is of high granularity, and the resulting picture should be most informative. Although the interest is basically in just one variable, the whole multivariate information by considering all remaining variables has been used and analyzed. One can thus claim that this information should be most relevant for decision makers. It should,

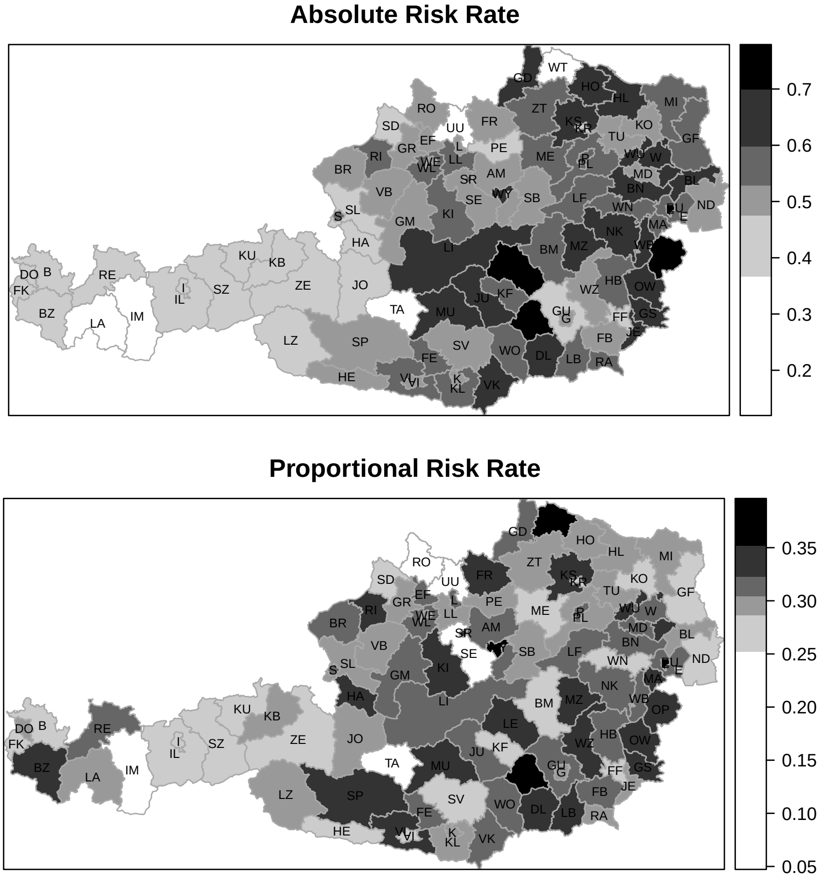

Figure 4.1: The upper figure shows the absolute risk rate of death per 10.000 population (age standardized and divided by number of residents) caused by circulatory systems. The lower figure shows the proportional risk rate, proportion of absolute risk rate of deaths caused by circulatory systems to the total absolute risk rate of deaths. The breaks of the gray scale are defined by the quantiles 0.05, 0.25, 0.50, 0.75 and 0.95, yielding six classes.

however, be noted that the scale of the new ilr coordinate is not interpretable in terms of a risk rate. It can just be interpreted in terms of different levels of dominance of the variable under investigation, and the resulting pattern in the geographical map can serve as a guidance for setting local activities.

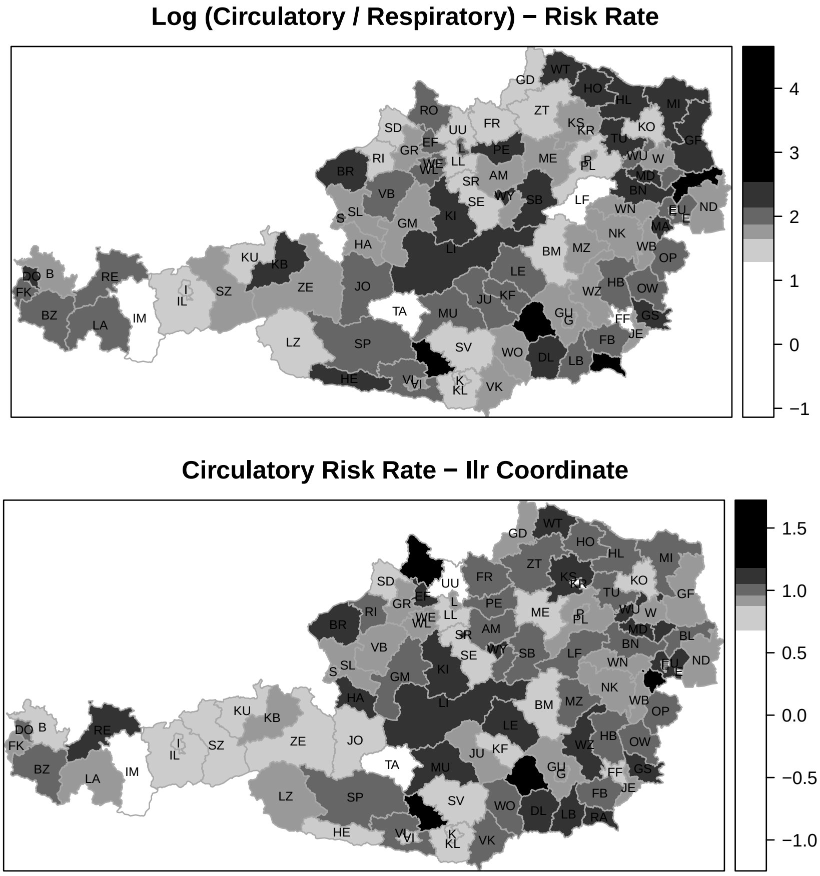
Log (Circulatory / Respiratory) − Risk Rate
4
3
2
1
0
−1
Circulatory Risk Rate − Ilr Coordinate
1.5
1.0
0.5
0.0
−0.5
−1.0

Figure 4.2: The upper figure shows the log-ratio of the absolute risk rate of circulatory systems to respiratory systems. The lower figure represents the regional distribution of the first ilr coordinate of absolute risk rate of deaths caused by circulatory systems. The breaks of the gray scale are defined by the quantiles 0.05, 0.25, 0.50, 0.75 and 0.95, yielding six classes.

4.4 Joint analysis of epidemiological variables

While in the previous section we were interested in presenting information about just one variable, we will analyze now the information of all variables jointly. A variety of statistical methods for multivariate data analysis is available, and here we focus just on two procedures: principal component analysis to investigate the main data structure, and multivariate outlier detection to discover atypical observations.

The methods are illustrated by hospital discharge data. The data have been provided by the Austrian social insurance institutions. The number of hospital discharges in Austria has been recorded for the year 2007. We use the data for the disease groups shown in Table 4.2. The data were aggregated to the 99 Austrian districts, standardized for age, and divided by the population size of the districts, as mentioned in Section 4.3. The disease groups form our compositional parts, and here we are interested in a joint analysis of all parts.

ICD10	Plot symbol	Group
A00-B99	AB	Certain infectious and parasitic diseases
C00-C99	C	Neoplasms
E10-E15	E10	Diabetes mellitus
E66	E66	Overweight and obesity
E78	E78	Lipid and lipoprotein abnormalities
I05-I59	I05	Circulatory Systems
I60-I69	I60	Cerebrovascular disease
I70-I72	I70	Atherosclerosis
I73-I74	I73	Peripheral vascular diseases
Others	others	All other diseases

Table 4.2: Selected disease groups for hospital discharge data.

4.4.1 Principal component analysis

Principal component analysis (PCA) is a popular multivariate method for reducing the dimensionality of the data while aiming at minimal loss of information. (Jolliffe, 2014) The data are projected into a lower dimensional space with orthogonal directions, called principal components (PC), which are linear combinations of the original variables. The coefficients defining the linear combinations are called *loadings*, and they allow for an interpretation of the PCs. Loadings and scores from a PCA can be used to construct a biplot, which provides a joint interpretation of variables and observations. (Gabriel, 1971)

Standard PCA is defined in the Euclidean geometry and can lead to biased results when the relevant information is contained in the ratios between the compositional parts. (Filzmoser et al., 2009a) Since the $D-1$ ilr coordinates capture all relative information of the D compositional parts, the space spanned by these coordinates is appropriate for PCA. Using ilr coordinates also avoids the mentioned collinearity problem of clr coefficients, and it is thus possible to perform robust PCA which downweights the effect of outlying observations. (Filzmoser et al., 2009a) For the interpretation of the results it is common to transform loadings and scores to the clr space, which leads to the *compositional biplot*. (Aitchison and Greenacre, 2002) In the biplot representation, the information related to the variables (disease groups) is shown by rays, and the information related to the observations by the abbreviations of the districts at the appropriate positions. The most important properties are:Kynčlová et al. (2016)

- The lengths of the rays are proportional to the standard deviation of the corresponding variables expressed as clr coefficients.
- The links between the vertices of the rays represent the standard deviation of the log-ratios between the involved compositional parts.
- The orthogonal projections of the observations on the rays represent the clr coefficients.

Figure 4.3 (left) shows the clr biplot resulting from a robust PCA on the hospital discharge data. The presented first two PCs explain 75% of the variance. Along the first PC with 60% explained variance, a clear grouping of the districts is visible. This PC is mainly determined by relative information of I70 and I73 (see Table 4.2 for the abbreviations). In more detail, high scores on the first PC refer to a dominance of atherosclerosis (I70-I72), and/or to an absence of peripheral vascular diseases (I73-I74). Figure 4.4 shows the scores in a map, and the high scores in dark gray concentrate mainly in the districts of Styria in the south-east of Austria. It may be difficult to interpret the reason for this behavior; one of the possible explanations would be that in Styria the system for coding this information is different to other regions – at least for the coding of I70-I72 and I73-I74.

The right plot of Figure 4.3 shows the results of robust PCA applied on the original (robustly scaled) data, i.e. on the preprocessed raw data information. The explained variance is somewhat lower (67%). Since the rays are arranged in just the left part of the plot, one would conclude that the variables tend to be positively correlated. This is an artifact of the geometry when it is assumed that the geometry of the sample space is the standard Euclidean geometry in $\mathbb{R}^D$, and a consequence of the spurious correlation, and it is a typical phenomenon for PCA applied to compositions. (Reimann et al., 2008) In this biplot, low values on the first PC point at districts with high hospital discharges in most considered disease groups. The second PC has similarities to the first PC from the compositional approach, but referring to the raw data information. The relation between I73 and C is rather weak according to the log-ratio approach, while it seems to be very strong when analyzing raw data information.

4.4.2 Outlier Detection

The biplot in Figure 4.3 (left) reveals some districts such as RA (Bad Radkersburg) that are deviating from the data majority. These districts could be denoted as outliers. However, the biplots only show a two-dimensional projection of the data and may thus

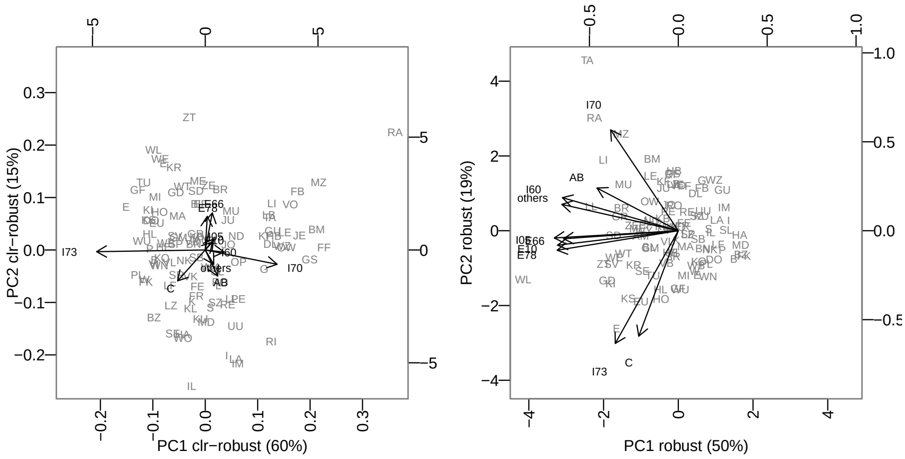

Figure 4.3: Biplots of first two robustly estimated principal components, expressed in clr coefficients (left plot) and robust PCA for raw data information (right plot).

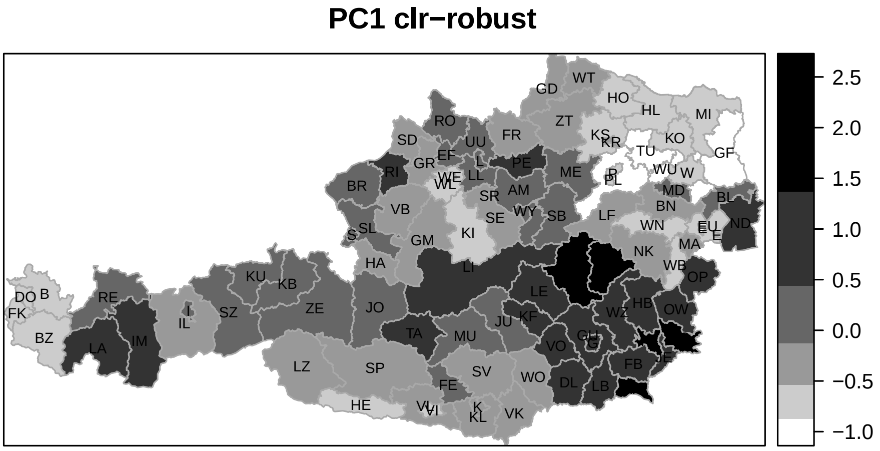

Figure 4.4: The first robustly estimated principal component for clr coefficients. The breaks of the gray scale are defined by the quantiles 0.05, 0.25, 0.50, 0.75 and 0.95, yielding six classes.

hide the position of other outliers in the remaining data structure. Moreover, there is no threshold value available that allows to distinguish outliers from regular observations.

Multivariate outlier detection is usually based on Mahalanobis distances (MD), a distance measure that accounts for the joint covariance structure of the data. (Mahalanobis, 1936) For a given data matrix $\boldsymbol{X}$, the MD of observation $\boldsymbol{x}_i$, for $i = 1, \ldots, n$, to the data center $T(\boldsymbol{X})$ with respect to the covariance $C(\boldsymbol{X})$ is defined as

$$\mathrm{MD}(\boldsymbol{x}_i) = \sqrt{(\boldsymbol{x}_i - T(\boldsymbol{X}))' \, C(\boldsymbol{X})^{-1} \, (\boldsymbol{x}_i - T(\boldsymbol{X}))}, \tag{4.6}$$

where T denotes the location estimator and C stands for the covariance estimator of $\boldsymbol{X}$. Note that if $C(\boldsymbol{X})$ is equal to the identity matrix (uncorrelated variables with variance one), the MD reduces to the Euclidean distance of an observation to the center. A large MD indicates that the observation is far away from the center with respect to the joint covariance structure, and thus the observation is a candidate for a univariate or a multivariate outlier. It is essential that both T and C are robust estimators in order to avoid the influence of the outliers on the estimation of location and covariance. Here, the MM-estimator for location and covariance is used (Yohai, 1987), leading to highly robust and efficient estimators.

A cut-off value for the identification of outliers can be determined by the distribution of the squared MDs, which are approximately χ^2 distributed with D degrees of freedom if the data are multivariate normally distributed. (Rousseeuw and van Zomeren, 1990) It is common to use the quantile 0.975 of that distribution as a cut-off value, i.e., observations with larger squared MD than the 0.975 quantile of the χ^2_D are declared as outliers.

The discussed methodology for outlier detection needs to be modified if the interest is in relative rather than in absolute information. In this case, the data need to be expressed in ilr coordinates, as it has been done for PCA. Equation (4.4) can be used for this purpose, resulting in the observations $\boldsymbol{z}_i$, for $i = 1, \ldots, n$, which form the rows of the matrix $\boldsymbol{Z}$. Equation (4.6) then modifies to

$$\mathrm{MD}(\boldsymbol{z}_i) = \sqrt{(\boldsymbol{z}_i - T(\boldsymbol{Z}))' \, C(\boldsymbol{Z})^{-1} \, (\boldsymbol{z}_i - T(\boldsymbol{Z}))}, \tag{4.7}$$

where $T(\boldsymbol{Z})$ and $C(\boldsymbol{Z})$ are robust estimators of location and covariance, respectively, for the ilr coordinates. The outlier cut-off value is the 0.975 quantile of the χ^2_{D-1} distribution, since the number of columns in $\boldsymbol{Z}$ is only $D - 1$. It has been shownFilzmoser and Hron (2008) that the MDs from Equation (4.7) remain the same, independent of the choice of the ilr coordinates, as long as T and C are affine equivariant estimators, such as MM-estimators. Note that using clr coefficients rather than ilr coordinates would cause numerical difficulties for MM-estimators due to their collinearity.

The discussed methodology for outlier detection is applied in the following to the hospital discharge data described earlier in this section. Our aim is to compare the results for outlier detection for the conventional approach based on the original data information, and for the log-ratio approach. The former MDs are shown on the horizontal axis of

Figure 4.5, while the latter MDs are presented on the vertical axis. The corresponding outlier cut-off values are shown as solid lines, and they divide the plot into four quadrants: (1) regular observations with respect to both methods in the lower left quadrant, (2) outliers for both methods in the upper right quadrant, (3) outliers from the raw-data point of view, but regular observations for the log-ratio approach in the lower right quadrant, and (4) outliers for ilr coordinates but regular observations for the absolute information in the upper left quadrant. These four cases are distinguished by different gray scales in the regional presentation of the information in Figure 4.6. Overall, there are no clear spatial patterns visible in the map. From Figure 4.5 one would also conclude that both sources of information lead to a joint trend for the MDs, and differences concerning outlyingness are only in few districts. However, the interpretation of outlyingness is different.

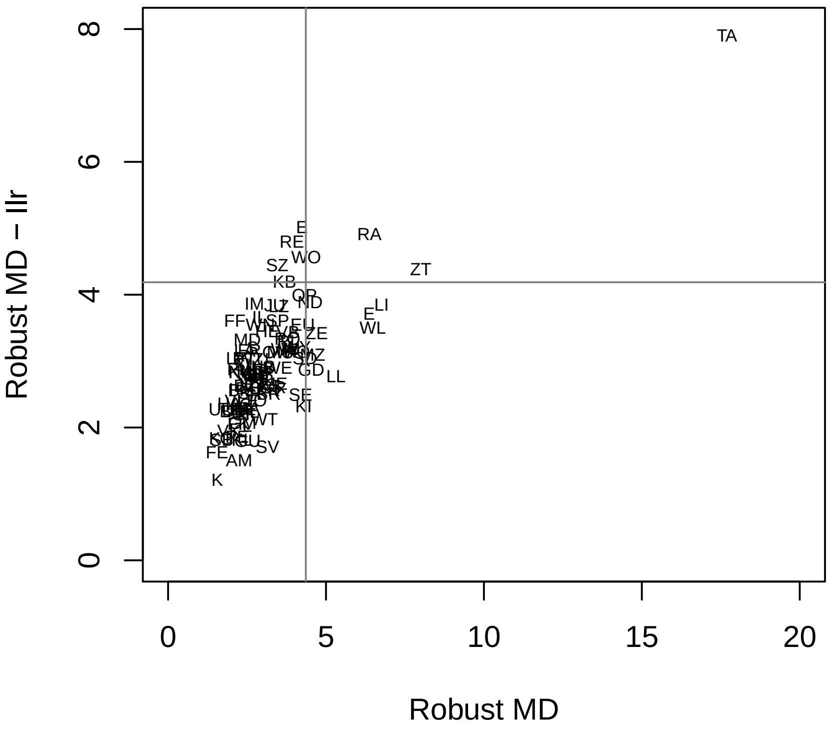

Figure 4.5: Multivariate outlier detection for the hospital discharge data based on robust Mahalanobis distances using raw data information (horizontal axis) and relative information expressed in ilr coordinates (vertical axis). Outlier cut-off values are shown as solid lines.

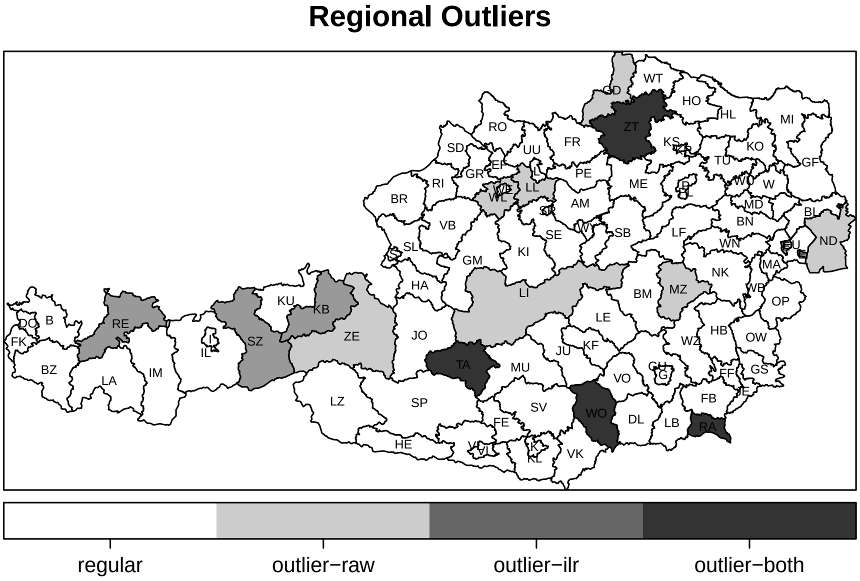

Figure 4.6: Outlying districts in the hospital discharge data according to the results shown in Figure 4.5 for absolute (raw) and relative (ilr) information.

This interpretation is supported by the plots in Figure 4.7 and 4.8. The single graphics show univariate scatterplots (the horizontal axes are randomly generated values) of the variables involved in the computation of the MDs. Accordingly, Figure 4.7 shows univariate scatterplots of the columns of $\boldsymbol{X}$ for the raw data information, and Figure 4.8 represents the relative information in terms of ilr coordinates constructed for each compositional part (see end of Section 4.2). In these presentations, dark gray symbols for the district abbreviations refer to outliers according to the corresponding MDs. Figure 4.7 shows that for variable I70 the district TA (Tamsweg) is an upper outlier: TA has a much higher number of hospital discharges for I70-I72 than all other districts (at least twice as high). Figure 4.8 also reveals TA for I70 as extreme value, but the interpretation is that for TA, the discharges for I70-I72 are dominating, relative to the other discharge groups. Figure 4.7 reveals WL (Wels-Land) on the upper extreme for several discharge groups, thus this district has exceptionally high discharges in those groups, compared to the other districts. Compared to the other discharge groups, however, and thus looking at relative information, WL is no longer extreme, see Figure 4.8. From the plots it can also be seen that the multivariate outliers are not necessarily outliers in the single dimensions.

4.5 Regression modeling

While the previous sections focused more on exploratory tools to investigate differences in the conventional and the log-ratio approach, this section shows differences for regression modeling, since regression is a widely used tool in statistics. We make use of both data

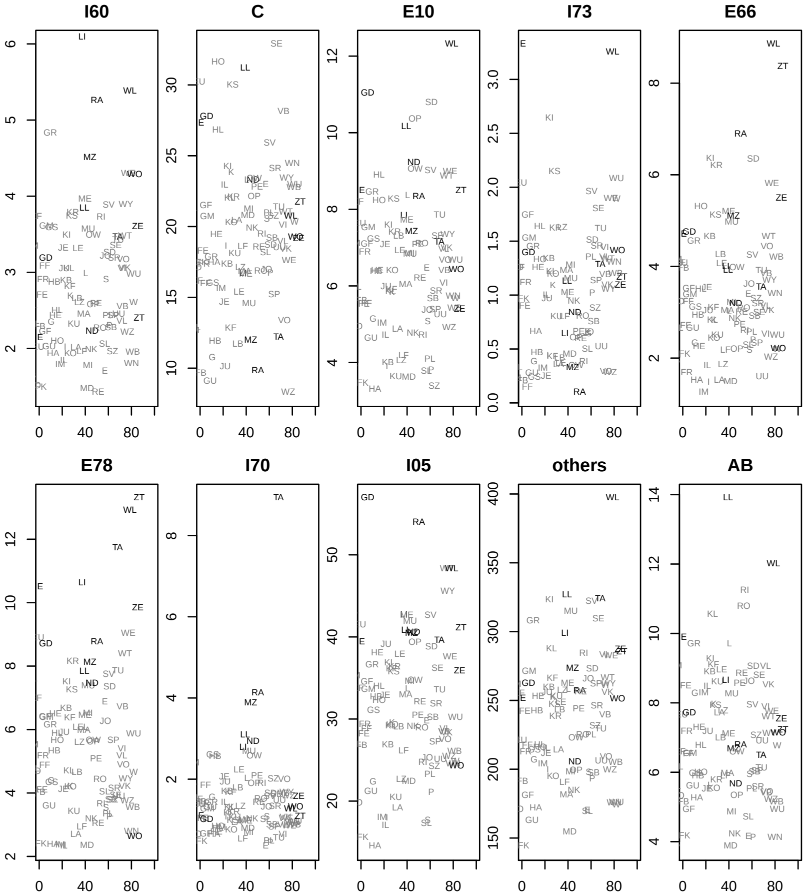

Figure 4.7: Univariate scatterplots of the hospital discharge data as raw data; outliers are coded in dark gray.

sets introduced previously, and try to explain deaths caused by circulatory systems, see Table 4.1, by the hospital discharge data, see Table 4.2. The response variable is treated as absolute information, given in form of the (standardized) risk rate. In the first case, the explanatory variables, contained in the matrix $\boldsymbol{X}$, are (standardized) frequencies of hospital discharges. In the log-ratio approach, the ilr variables contained in the matrix $\boldsymbol{Z}$ are used as regressor variables. In both cases, robust regression based on the MM-estimator is applied (Yohai, 1987), and robust inference as implemented in the R package `robustbase` is carried out. (Rousseeuw et al., 2016) In the log-ratio approach, robust inference is applied to the single ilr variablesHron et al. (2012), as implemented in the R package `robCompositions`. (Templ et al., 2016) This means that specific ilr coordinates are constructed, using Equation (4.4), by putting the variable of interest to the first position (see end of Section 4.2). Thus, for D compositional parts, D ilr bases are constructed, and for each basis, regression is applied. The regression inference (test

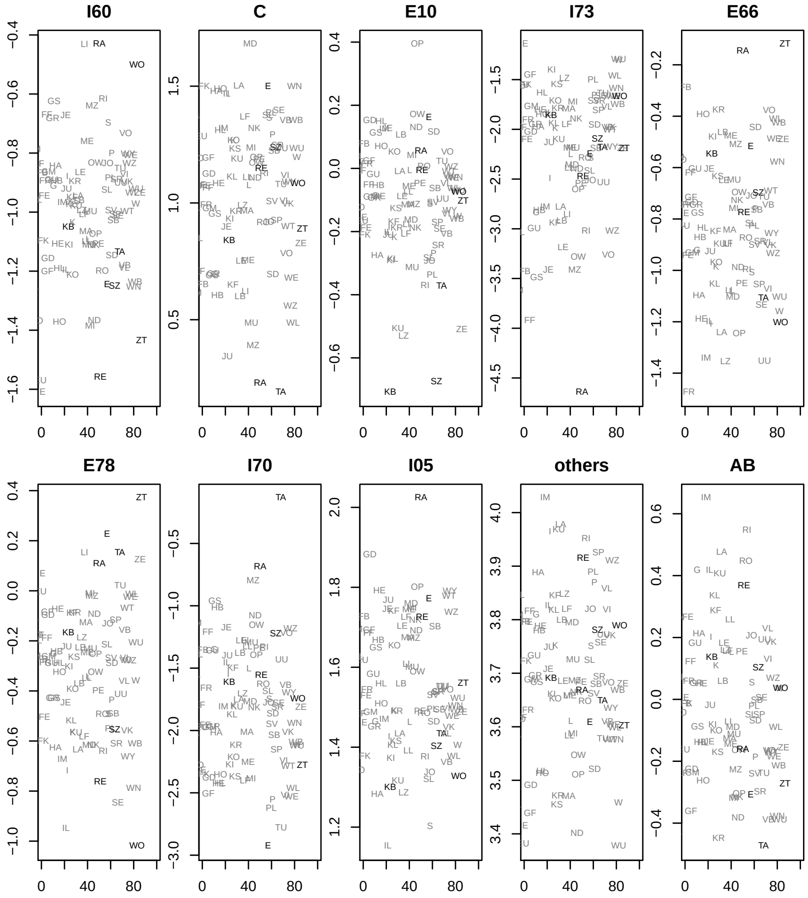

Figure 4.8: Univariate scatterplots of ilr coordinates of the hospital discharge data; outliers are coded in dark gray.

if the regression parameter is zero) is then just reported for this first ilr variable.

The results are shown in Table 4.3 (absolute information) and Table 4.4 (log-ratio approach). In both cases, the model fit in terms of the multiple R-square is rather low. On the other hand, it might have been surprising if death rates could be well explained just by hospital discharges, and thus this example should rather be considered as an illustration of differences between two approaches. Looking at the significance of the variables (last column in the tables), for the approach based on absolute information, the terms `intercept`, E78 and I05 are significant, based on a significance level of 5%. In the log-ratio approach, the significant terms are E78, I05, and the remaining diseases (others). E78 (lipid and lipoprotein abnormalities) are regarded as a modifiable risk factor for cardiovascular disease due to their influence on atherosclerosis, and I05-I59 refer to diseases of the circulatory systems. Both are thus logical candidates for being significant in the regression. The essential difference in the outcome is thus in the variable “others”.

```
               Estimate   Std. Err.  t value  Pr(>|t|)
(Intercept)   -0.8741096  0.0888319  -9.840   0.000
I60            0.0290419  0.0327317   0.887   0.377
C             -0.0005565  0.0039548  -0.141   0.888
E10            0.0276863  0.0208307   1.329   0.187
I73            0.0247177  0.0538451   0.459   0.647
E66            0.0106079  0.0231088   0.459   0.647
E78           -0.0304454  0.0140143  -2.172   0.032
I70           -0.0185205  0.0191972  -0.965   0.337
I05            0.0133381  0.0045693   2.919   0.004
AB            -0.0004597  0.0151218  -0.030   0.975
others        -0.0012793  0.0008483  -1.508   0.135

Robust residual standard error: 0.1642
Multiple R-squared: 0.35, Adjusted R-squared: 0.28
```

Table 4.3: Results of the robust regression analysis for modeling the absolute death rate caused by circulatory systems by raw hospital disease data.

The negative sign of the estimated coefficient means that the under-dominance of the remaining diseases leads to higher death rates of circulatory systems, and dominance of “others” causes lower circulatory system death rates. Note that “others” refers to the corresponding ilr variable, expressing all relative information of “others” to the remaining compositional parts, including E78 and I05, see Equation (4.4). Also E78 and I05 are ilr variables in the log-ratio approach, thus expressing all relative information about these compositional parts. Since the absolute frequencies in “others” are high, compared to the remaining variables (see scale in the plots in Figure 4.7), it can be expected that a change in this variable will have some consequence on the other variables, i.e. more hospital discharges for other diseases will affect the number of discharges of the considered disease groups. Therefore, considering (log-) ratios of the variables as regressors might make more sense than using the variables in the model, isolated from each other. The log-ratio approach accounts for the interplay between the variables in a natural way, and the interpretation of the resulting model in terms of relative contributions rather than absolute ones is more intuitive.

4.6 Summary and conclusions

We have demonstrated different possibilities for analyzing health care information, such as frequencies of different diseases or death rates. There are big conceptual differences if the reported (appropriately normalized) absolute numbers are analyzed, or if ratios between the different variables (disease groups) form the source of information for the analysis. The latter case leads to analyzing relative information, with consequences for

```
                Estimate   Std. Err. t value Pr(>|t|)
(Intercept)      1.05114    0.63858   1.646  0.103
ilr(I60)         0.14114    0.09423   1.498  0.137
ilr(C)          -0.10876    0.07717  -1.409  0.162
ilr(E10)         0.19479    0.13227   1.473  0.144
ilr(I73)         0.04556    0.05421   0.840  0.402
ilr(E66)         0.11214    0.08876   1.263  0.209
ilr(E78)        -0.20675    0.07023  -2.944  0.004
ilr(I70)        -0.02432    0.05849  -0.416  0.678
ilr(I05)         0.40531    0.12981   3.122  0.002
ilr(AB)          0.02246    0.12318   0.182  0.855
ilr(others)     -0.58318    0.18265  -3.193  0.001

Robust residual standard error: 0.1642
Multiple R-squared: 0.36, Adjusted R-squared: 0.30
```

Table 4.4: Results of the robust regression analysis for modeling the absolute death rate caused by circulatory systems by ilr coordinates of the hospital disease data.

the methodology and for the interpretation. The well-established log-ratio approach not only leads to desirable properties such as scale invariance, it also allows to work in the usual Euclidean geometry, where the standard statistical methods are designed for. It is unpredictable if the numerical results of an analysis of absolute or relative information differ, but it is clear that the interpretation of the results is different. When absolute information is in focus, one can interpret the involved variables isolated from each other, and think in terms of the measured unit. In contrast, variables involved in a log-ratio analysis always need to be interpreted in a relative sense, typically in terms of dominance of one variable with respect to the other variables. In this way, one variable is never isolated from the rest, but always considered in a multivariate context by incorporating other available information. In many cases, this might be more appropriate for analyzing epidemiological data. For example, the log-ratio approach allows to analyze and compare the whole “risk profile” of certain diseases among different regions even if the total sum differs (different population sizes, different health conditions, different health care systems, etc.). Normalization with respect to such effects is not necessary with this approach since the resulting ratios would not change, but it would be crucial for an analysis of the absolute information.

Summing up, the log-ratio approach is recommended in epidemiology if certain variables need to be directly related to each other. Since all pairwise relationships are taken into account, the analyzed information is of high granularity, leading to a more comprehensive picture. In some cases even just one ratio of a part to another part can vary in neighboring districts indicating regional differences. This information can be a promising indicator for decision makers. However for a correct conclusion a detailed analysis of all log-ratios to that part, or the corresponding clr variable need to be considered. A “traditional” approach which investigates the variables as proportions to a total, for instance, is not wrong in a strict sense, but does not make use of the granularity of the multivariate information. Using exclusively this proportional information for investing into the infrastructure of a region with a high proportion on a certain disease thus ignores information that might be relevant for developing more general strategies. Log-ratios, on the other hand, ignore the absolute values (e.g. if a proportion on a disease is high in a district), and thus a combination of both sources of information may be the optimal support for decision makers.

All methods for compositional data analysis considered here can be carried out with the R package `robCompositions`. (Templ et al., 2016)

Acknowledgments

This work has been partly funded by the K-project DEXHELPP through COMET - Competence Centers for Excellent Technologies, supported by BMVIT, BMWFI and the province Vienna. The COMET program is administrated by FFG.

List of Figures

1.1 Scatter plot of the constructed balances based on the live birth data. 11

1.2 First ilr-coordinate of *birth* data. The observations marked by red color were defined as zero values. Observations after imputation are marked by green color. 13

1.3 Childbirth data before (a) and after (b) centering. 17

1.4 The proportion of childbirths at normal age of mother expressed as absolute information (left) and relative information in form of the clr coordinate (right). The breaks of the color scale are defined by the quantiles 0.05, 0.25, 0.50, 0.75 and 0.95, yielding six classes. 18

1.5 Multivariate outlier detection of childbirth data given in proportions. The horizontal axis represents the robust Mahalanobis distances of the raw data. The vertical axis represents the robust Mahalanobis distances based on the ilr coordinates. 20

1.6 Compositional (left) and standard (right) biplot of the first two robust PCs based on the childbirth data. 23

2.1 Time comparison to compute the first balance according to the algorithms AP, HC, MV, and SPB. Shown are average computation times for compositional data with $n = 100$ observations and varying number D of parts. The right plot compares only the algorithms HC and SPB in high-dimensional settings. 32

2.2 Cumulative explained variance of the first 5 components for CoDa-PCA, SPB, and HC for simulated data with different dimension D. With increasing dimension, SBP results in higher explained variance compared to HC. 34

2.3 Cumulative proportions of non-zero entries in the balances of SPB. 35

2.4 Cumulative explained variance for $K = 2$ components. Displayed is the ratio between SPB and HC. The explained variance clearly gets larger for SPB compared to HC with increasing dimension. 36

2.5 First two balances of HC applied to the urine data. Shown are the original data (black), and the position of the positive (light gray vertical line, or red in the online-version) and negative (dark gray vertical lines, or blue) signs of the balances. 36

2.6 First two balances of SPB applied to the urine data. Shown are the original data (black), and the position of the positive (light gray vertical line, or red in the online-version) and negative (dark gray vertical lines, or blue) signs of the balances. 37

2.7 First two components of CoDa-PCA, HC and SPB for the urine data. 38

3.1 Composition of sand, silt, and clay in agricultural soils of Europe. Ternary diagram (a), representation in ilr coordinates (b) 44

3.2 Effect of the DL and IR contamination on the first ilr coordinate ((a) and (b)), and on all ilr coordinates jointly ((c) and (d)) 50

3.3 One uncontaminated part, and 1 to 19 contaminated parts added. Univariate and multivariate structural changes between the original and contaminated ilr coordinates with increasing number of contaminated parts in case of DL ((a) and (c)) and IR ((b) and (d)) 53

3.4 Ten uncontaminated parts, and 1 to 10 contaminated parts added. Univariate and multivariate structural changes between the original and contaminated ilr coordinates with increasing number of contaminated parts in case of DL ((a) and (c)) and IR ((b) and (d)) 54

3.5 Ten uncontaminated parts, and 10 contaminated parts added. Univariate and multivariate structural changes between the original and contaminated ilr coordinates with increasing amount of contamination in case of DL ((a) and (c)) and IR ((b) and (d)) 56

3.6 Ten uncontaminated parts, and 10 contaminated parts added. Univariate and multivariate structural changes between the original and contaminated ilr coordinates with varying number of observations in the data set; the contamination level is fixed with 25%; DL ((a) and (c)) and IR ((b) and (d)) 57

3.7 Variance decomposition results: The light boxplots show the ratios A/B according to Equation (3.8) for the simulation scheme in Section 3.4.1. The dark boxplots compare the ratio of A-contaminated (either DL or IR) to A-uncontaminated. No contamination (a), detection limit (b), imprecision rate (c) 59

3.8 Biplots of the first two PCs based on the data shown in Figure 3.2. Classical PCs of uncontaminated data (a), classical PCs of contaminated data (b), robust PCs of uncontaminated data (c), robust PCs of contaminated data (d) 60

4.1 The upper figure shows the absolute risk rate of death per 10.000 population (age standardized and divided by number of residents) caused by circulatory systems. The lower figure shows the proportional risk rate, proportion of absolute risk rate of deaths caused by circulatory systems to the total absolute risk rate of deaths. The breaks of the gray scale are defined by the quantiles 0.05, 0.25, 0.50, 0.75 and 0.95, yielding six classes. 69

4.2 The upper figure shows the log-ratio of the absolute risk rate of circulatory systems to respiratory systems. The lower figure represents the regional distribution of the first ilr coordinate of absolute risk rate of deaths caused by circulatory systems. The breaks of the gray scale are defined by the quantiles 0.05, 0.25, 0.50, 0.75 and 0.95, yielding six classes. 70
4.3 Biplots of first two robustly estimated principal components, expressed in clr coefficients (left plot) and robust PCA for raw data information (right plot). 73
4.4 The first robustly estimated principal component for clr coefficients. The breaks of the gray scale are defined by the quantiles 0.05, 0.25, 0.50, 0.75 and 0.95, yielding six classes. 73
4.5 Multivariate outlier detection for the hospital discharge data based on robust Mahalanobis distances using raw data information (horizontal axis) and relative information expressed in ilr coordinates (vertical axis). Outlier cut-off values are shown as solid lines. 75
4.6 Outlying districts in the hospital discharge data according to the results shown in Figure 4.5 for absolute (raw) and relative (ilr) information. 76
4.7 Univariate scatterplots of the hospital discharge data as raw data; outliers are coded in dark gray. 77
4.8 Univariate scatterplots of ilr coordinates of the hospital discharge data; outliers are coded in dark gray. 78

List of Tables

1.1 Pearson correlations between the proportions of premature births in Austria: four part composition: extremely, very, moderate, late preterm. 1
1.2 Pearson correlations between the proportions of premature births in Austria: three part composition: extremely, very, moderate preterm. 2
1.3 Number of live births according to age groups of mother at birth in the countries of the European Union. 9
1.4 An example of sequential binary partition to construct an orthonormal basis based on the live birth data. 10
1.5 Variation array of childbirth data. The upper triangle shows the variance of the log-ratios, while the lower triangle includes the log-ratio means. 16

2.1 Cumulative explained variances for CoDa-PCA, hierarchical clustering (HC) and sparse principal balances (SPB) for the urine data set. 37

4.1 Causes of deaths by aggregated disease groups. 68
4.2 Selected disease groups for hospital discharge data. 71
4.3 Results of the robust regression analysis for modeling the absolute death rate caused by circulatory systems by raw hospital disease data. 79
4.4 Results of the robust regression analysis for modeling the absolute death rate caused by circulatory systems by ilr coordinates of the hospital disease data. 80

Bibliography

Statistik Austria, 2016. URL `http://www.statistik.at/web_en/statistics/PeopleSociety/population/births/028950.html`.

J. AITCHISON. Principal component analysis of compositional data. *Biometrika*, 70(1):57–65, 1983.

J. Aitchison. Reducing the dimensionality of compositional data sets. *Journal of the International Association for Mathematical Geology*, 16(6):617–635, 1984.

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

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

C. Barceló-Vidal and J. A. Martín-Fernández. Differential calculus on the simplex. *Terra Nostra*, 3:393–398, 2002.

C. Barceló-Vidal, J. Martín-Fernández, and G. Mateu-Figueras. Compositional differential calculus on the simplex. In V. Pawlowsky-Glahn and A. Buccianti, editors, *Compositional Data Analysis: Theory and Applications*, pages 176–190. Wiley, Chichester, 2011.

J. Beaumont, L. Lix, K. Yost, and E.A.Hahn. Application of robust statistical methods for sensitivity analysis of health-related quality of life outcomes. *Quality of Life Research*, 15:349–356, 2006.

R. T. Birge. The propagation of errors. *American Journal of Physics*, 7(6):351–357, 1939.

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

M. G. Cox and B. R. L. Siebert. The use of a Monte Carlo method for evaluating uncertainty and expanded uncertainty. *Metrologia*, 43(4):S178, 2006.

C. Croux, P. Filzmoser, and H. Fritz. Robust sparse principal component analysis. *Technometrics*, 55(2):202–214, 2013.

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

J. J. Egozcue and V. Pawlowsky-Glahn. Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37:795–820, 2005.

J. J. Egozcue and V. Pawlowsky-Glahn. Simplicial geometry for compositional data. *Geological Society, London, Special Publications*, 264(1):145–159, 2006.

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

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, 2003b.

B. Everitt. *Cluster Analysis*. Camdridge, 1993.

S. E. Feller and C. F. Blaich. Error estimates for fitted parameters: Application to hcl/dcl vibrational-rotational spectroscopy. *Journal of Chemical Education*, 78(3):409, 2001.

P. Filzmoser and K. Hron. Outlier detection for compositional data using robust methods. *Mathematical Geosciences*, 40(3):233–248, 2008.

P. Filzmoser, K. Hron, and C. Reimann. Principal component analysis for compositional data with outliers. *Environmetrics*, 20(6), 2009a.

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

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, 2009c.

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 the interpretation of orthonormal coordinates for compositional data. *Mathematical Geosciences*, 43(4):455–468, 2011.

E. Fišerová and K. Hron. On interpretation of orthonormal coordinates for compositional data. *Mathematical Geosciences*, 43(4):455–468, 2011.

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

N. Gift, I. C. Gormley, and L. Brennan. *MetabolAnalyze: probabilistic principal components analysis for metabolomic data*, 2010. R package version 1.3.

M. Greenacre. Log-ratio analysis is a limiting case of correspondence analysis. *Mathematical Geosciences*, 42(1):129–134, 2010.

M. Greenacre. Measuring subcompositional incoherence. *Mathematical Geosciences*, 43 (6):681–693, 2011a.

M. Greenacre. *Compositional Data and Correspondence Analysis*, pages 104–113. John Wiley & Sons, Ltd, 2011b.

J. Guo, G. James, E. Levina, G. Michailidis, and J. Zhu. Principal component analysis with sparse fused loadings. *Journal of Computational and Graphical Statistics*, 19(4): 930–946, 2010.

C. Hennekens and J. Buring. *Epidemiology in Medicine.* Lippincott, Williams & Wilkins, Philadelphia, PA, USA, 1987.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933.

K. Hron, M. Templ, and P. Filzmoser. Imputation of missing values for compositional data using classical and robust methods. *Computational Statistics & Data Analysis*, 54(12):3095 – 3107, 2010.

K. Hron, P. Filzmoser, and K. Thompson. Linear regression with compositional explanatory variables. *Journal of Applied Statistics*, 39(5):1115–1128, 2012.

I. Jolliffe. *Principal Component Analysis.* John Wiley & Sons, London, 2014.

H. H. Ku. Notes on the use of propagation of error formulas. *Journal of Research of the National Bureau of Standards*, 70, 1966.

P. Kynčlová, P. Filzmoser, and K. Hron. Compositional biplots including external non-compositional variables. *Statistics*, ( ): , 2016. To appear.

M. C. Leite. Applying compositional data methodology to nutritional epidemiology. *Statistical Methods in Medical Research*, 24:43–70, 2014.

T. L. Lincoln and C. Builder. Global healthcare and the flux of technology. *International journal of medical informatics*, 53(2-3):213—224, 1999.

J. S. Liu. *Monte Carlo Strategies in Scientific Computing.* Springer, 2008.

P. Mahalanobis. On the generalised distance in statistics. *Proceedings of the National Institute of Science of India*, A2:49–55, 1936.

R. Maronna, D. Martin, and V. Yohai. *Robust Statistics: Theory and Methods*. John Wiley & Sons Canada Ltd., Toronto, ON, 2006.

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

J. A. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser, and J. Palarea-Albaladejo. Model-based replacement of rounded zeros in compositional data: Classical and robust approaches. *Computational Statistics and Data Analysis*, 56(9):2688–2704, 2012.

G. Nyamundanda, L. Brennan, and I. C. Gormley. Probabilistic principal component analysis for metabolomic data. *BMC Bioinformatics*, 11(1):1–11, 2010.

J. Palarea-Albaladejo and J. Martín-Fernández. A modified {EM} alr-algorithm for replacing rounded zeros in compositional data sets. *Computers & Geosciences*, 34(8): 902 – 917, 2008.

J. Palarea-Albaladejo, J. A. Martín-Fernández, and J. Gómez-García. A parametric approach for dealing with compositional rounded zeros. *Mathematical Geology*, 39(7): 625–645, 2007.

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 and J. J. Egozcue. Blu estimators and compositional data. *Mathematical Geology*, 34(3):259–274, 2002.

V. Pawlowsky-Glahn, J. J. Egozcue, and R. Tolosana-Delgado. Principal balances. In R. Tolosana-Delgado and M. Ortego, editors, *Proceedings of the 4th International Workshop on Compositional Data Analysis*, pages 1–10, Spain, 2011. Girona.

V. Pawlowsky-Glahn, J. 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, 1897.

C. Reimann, P. Filzmoser, R. Garrett, and R. Dutter. *Statistical Data Analysis Explained: Applied Environmental Statistics with R*. Wiley, Chichester, UK, 2008.

C. Reimann, M. Birke, A. Demetriades, P. Filzmoser, and P. O'Connor. *Chemistry of Europe's agricultural soils - Part A: Methodology and interpretation of the GEMAS data set*. Schweizerbarth, Hannover, 2014.

P. J. Rousseeuw and K. Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):pp. 212–223, 1999.

P. J. Rousseeuw and B. C. van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–639, 1990.

P. J. Rousseeuw, C. Croux, V. Todorov, A. Ruckstuhl, M. Salibian-Barrera, T. Verbeke, and M. Maechler. ***robustbase**: Basic Robust Statistics*, 2016. URL `http://CRAN.R-project.org/package=robustbase`. R package version 0.92-5.

M. Templ, K. Hron, and P. Filzmoser. ***robCompositions**: Robust Estimation for Compositional Data.*, 2016. URL `http://CRAN.R-project.org/package=robCompositions`. R package version 2.0.2.

M. C. Tsilimigras and A. A. Fodor. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Annals of Epidemiology*, 26(5):330 – 335, 2016.

K. van den Boogaart, R. Tolosana-Delgado, and M. Templ. Regression with compositional response having unobserved components or below detection limit values. *Statistical Modelling*, 15(2):191–213, 2015.

D. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 7 2009.

G. S. Witten D, Tibshirani R and N. B. *PMA: Penalized Multivariate Analysis*, 2011. R package version 1.0.8.

V. J. Yohai. High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 15(2):642–656, 1987.

H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

Index

- Aitchison, 3
  - distance, 5, 42
  - geometry, 4, 42
  - inner product, 4, 5, 42
  - norm, 5, 42
- alr, *see* additive log-ratio

- balances, 8, 26, 67
  - orthonormal, 8, 28
  - principal, 27
  - sparse principal, 27
- barycenter, 4, 16
- biplot, 59
  - compositional, 22, 72

- center, 14
- closure, 2, 42
- clr, *see* centered log-ratio
- collinearity, 66, 71
- composition, 2
- correlation
  - Pearson, 1
  - Spearman, 51
  - spurious, 2, 64

- detection limit, 12, 43, 45, 48
- DL, *see* detection limit

- error propagation, 43, 47, 48

- generating system, 5

- ilr, *see* isometric log-ratio
- imprecision rate, 49
- information
  - absolute, 78
  - multivariate, 64, 68
  - proportional, 68
  - relative, 66, 68, 79
- IR, *see* imprecision rate

- log-ratio transformations
  - additive, 6
  - centered, 6, 7, 22, 26, 66
  - isometric, 6, 22, 26, 66

- Mahalanobis distance, 19, 49, 74
  - robust, 19
- MD, *see* Mahalanobis distance
- mean absolute scaled deviation, 51
- metabolomic spectra, 34

- orthonormal basis, 5, 7, 66
- orthonormal coordinates, 7

- PBs, *see* principal balances
- PCA, *see* principal component analysis
- penalized matrix decomposition, 29
- permutation invariance, 3
- perturbation, 4, 5, 42
- powering, 5, 42
- principal component analysis, 21, 25, 71

- regression, 76

- `robCompositions`, 12, 81

- SBP, *see* sequential binary partition
- scale invariance, 3, 65
- sequential binary partition, 8, 27, 67

simplex, 3, 26
singular value decomposition, 21
sparse principal balances, 29, 34
SPB, *see* sparse principal balances
subcomposition, 3, 8
subcompositional coherence, 3

Taylor approximation, 43, 46
ternary diagram, 16, 17
total variance, 15, 30

variation array, 15
variation matrix, 15

zero problem, 12

Curriculum Vitae

Personal Data

First Name:	Mehmet Can
Last Name:	Mert
Date of Birth:	16.08.1985
Nationality:	Turkey
Home Adress:	Neustift am Walde 78/1 1190 Vienna, Austria
Mobile:	+43 660 580 4696
Email:	mehmet.mert@tuwien.ac.at

Academic Experience

2012 - 2016	PhD programme , Vienna University of Technology Institute of Statistics and Mathematical Methods in Economics *Thesis: Theoretical and practical aspects in compositional data analysis*
2009 - 2012	Master's programme, Vienna University of Technology Institute of Statistics and Mathematical Methods in Economics *Thesis: An empirical comparison of stochastic option pricing models*
2006 - 2009	Bachelor's programme, Graz University of Technology Technical Mathematics *Thesis: Empirical analysis of air pollutants*

Professional Experience

2012 - 2014	Junior Statistician, Vienna University of Technology *Institute of Statistics and Mathematical Methods in Economics* *Development of new statistical methods for compositional data analysis* *Statistical analysis of research projects with social insurance data of Austria*
2014 - 2016	Junior Statistician, K-Project: DEXHELPP *Decision Support for Health Policy and Planning: Developing statistical methods, models and technologies based on existing health care data*

Computer Skills

Programming:	R, SAS, SPSS, Matlab, C, LaTeX, Microsoft Office
Database:	Postgres, MSSQL, MySQL, Oracle

Language Skills

Turkish:	Mother tongue
German:	Fluent
English:	Fluent