Diplomarbeit

zum Thema

Classical and Robust Regression for Compositional Data Analysis

ausgeführt am
Institut für Statistik und Wahrscheinlichkeitstheorie
der Technischen Universität Wien

unter der Anleitung von
Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Peter Filzmoser

durch
Judith Zehetgruber, BSc.
Schwadorferstraße 5
3100 St. Pölten

Wien, 15. Mai 2013 _____

Danksagung

Zu Beginn möchte ich mich bei meinem Diplomarbeitsbetreuer, Herrn Ao. Univ.-Prof. Dipl.-Ing. Dr.techn. Peter Filzmoser bedanken, dass er immer gewillt war mir bei Problemen zu helfen und mich während des Schreibens sehr unterstützt hat. Danke für die aufgebrachte Geduld mit mir und der Arbeit.

Ein besonderes Danke möchte ich natürlich auch meinen Eltern aussprechen, dafür, dass sie es mir möglich gemacht haben dieses Studium zu absolvieren and mir die nötige Zeit dazu gelassen haben, meinem Papa, dass er, solange er da war, mir seine Geduld geschenkt hat und mir in schweren Zeiten wieder Kraft gegeben hat. Und meiner Mama, Mag. Maria Zehetgruber, möchte ich danken, dass sie für mich da war, wenn ich sie gebraucht habe, sei es um einfach nur da zu sein, oder um zu telefonieren oder mich voran zu treiben.

Ein großes Danke gilt auch meiner Schwester, Magdalena Zehetgruber, BA, für die vielen, vielen Stunden die wir während unserer gemeinsamen Studienzeit miteinander verbracht haben und die mich sehr gestärkt haben. Danke, dass du mich oft aufgebaut hast und mir immer hilfreich zur Seite gestanden bist!

Weiters möchte ich mich bei meinen Freunden bedanken, die mich sowohl beim Schreiben meiner Arbeit, aber vor allem auch während des ganzen Studiums begleitet und unterstützt haben und mit denen ich eine sehr schöne Zeit verbringen durfte. Im Besonderen danke ich Dipl.-Ing. Karoline Geißler und Dipl.-Ing. Tamara Steinlechner.

Schlussendlich möchte ich mich auch bei meinem Freund, Florian Spechtenhauser, BSc bedanken für das Verständnis und die Geduld, deine Hilfe und Unterstützung während des Schreibens meiner Diplomarbeit, für das Korrekturlesen und dafür, dass du für mich da bist.

Abstract

In this diploma thesis, classical and robust multivariate regression for compositions is developed. Therefore, proper transformations from the simplex to the usual Euclidean space have to be applied on the compositional variables for being able to interpret the results in terms of coordinates. Consequently, special kinds of balances are proposed to obtain reasonable results, that are easy to interpret. The regression analysis is divided into three parts. A model with just a compositional response is considered, as well as a model with compositional explanatory variables and finally a model with both, compositional response and compositional explanatory variables is taken into account. Special attention has to be paid to the ilr transformations of the original variables as well as to the resulting models given in coordinates. Further, classical and robust regression analysis can be applied and coefficients are computed. In the robust case, the multivariate least-trimmed squares estimator is used and a fast mlts algorithm has been used for the computations. Geochemical data was available to present the results. Inference statistics on the one hand and diagnostic plots on the other hand are used to display the properties of the data and the models that have been observed.

Contents

Chapter 1

Introduction

1.1 General remarks

The aim of the thesis is to develop classical regression and robust (multivariate) regression analysis for compositional data which is a special kind of data defined on the simplex. Moreover, these methods are applied on certain data using the statistics software R. This introduction should imbed the work into a context and give a short overwiev of the parts of the thesis.
Recent work on this topic has been done by Filzmoser et al. (2012) and Filzmoser and Hron (2012), who started to work out robust regression analysis for compositions. General information about compositional data and the simplex, as well as the introduction of the alr and clr transformation can be found in Aitchison (1986), which was the first attempt to treat compositions differently. The lecture notes of Pawlowsky-Glahn et al. (2007) give an interesting overview of the topic and for more details see Pawlowsky-Glahn and Buccianti (2011). Further work about log-ratio transformations, especially about the ilr transformation is given in Egozcue et al. (2003). To gain insight about robust multivariate methods, Hubert et al. (2008) and Rousseeuw et al. (2004) are proposed to read.

1.2 Overview of the contents of the Diploma thesis

In chapter 2, the main aspects of compositional data are explained. One will learn about the special geometry that is used for compositions. Moreover, the chapter will give an overview of how to handle compositions and it also gives a detailed description of the three possible transformations that can be

applied on this kind of data. Special coordinates of the ilr transformation, called balances, will be considered intensely, and finally the relationships between the transformations will be stated. In the end some elements of simplicial statistics are added.

In chapter 3, linear regression analysis will be introduced for compositional data. First a compositional response and non-compositional explanatory variables will be considered. The main focus lies on the transformation of the compositions and on the estimation of the parameters in the regression model. Moreover, some inference statistics such as the coefficient of determination, a t-test and a F-test will be described.
A model with compositional explanatory variables and a non-compositional response is developed and, again the parameters will be estimated by means of the least-squares estimation. Once again, a t-test as well as a F-test and some other inference statistics will be shown in that context.
Finally, a model with both, compositional response and compositional explanatory variables, is taken into account. This model is rather difficult to handle due to transformations of the response as well as of the explanatory variables. These transformations will be done separately. However, when one works with coordinates inference statistics can be applied similarly to that of the other cases mentioned before.

In chapter 4, robust multiple and multivariate linear regression will be explained. Therefore, two methods are introduced - the minimum covariance determinant (MCD) regression based on robust estimation of location and scatter, and the (multivariate) least trimmed squares (MLTS) regression. For the MLTS regression a subset of h observations should be found whose covariance matrix of its residuals from a least squares fit is minimal. One will find out, that the MLTS estimator is more general and therefore it is preferred to use it for further applications.

Chapter 5 will shortly discuss the three regression models when one applies a robust method. The robust estimations are just applied on the model with coordinates and therefore, the analysis is very similar to that in chapters 3 and 4.

Finally, in chapter 6 some examples will be presented by using a large data set from geochemistry. For example, one is interested to find out if there is a relationship between the parts of Iron in the soil and the magnetic characteristics there. Balances will be used to apply regression analysis on these data. Examples for all different models are given and diagnostic plots, that should detect outliers, are performed.

In chapter 7, the results will be discussed and the most important facts will be pointed out.

Chapter 2

Compositional data

In this chapter we introduce basic concepts and some properties of compositional data. The main reason that data analysis (e.g. regression analysis) also was adapted for compositional data follows from the fact, that we often face the situation, where we are not interested in the size of data alternatively the amount of a certain variable, but we want to obtain information about the ratios. For instance, this is the case if the data are given in percentages.

2.1 Some properties of compositional data

First we start with some definitions.
A row vector, $\mathbf{x} = [x_1, x_2, \ldots, x_D]$, is defined as a *D-part composition* when all its components are strictly positive real numbers and when they carry only relative information. Compositions can be considered as representatives of equivalence classes of real vectors with positive components.
The sample space of compositional data is called the *simplex*, defined as

$$\mathcal{S}^D = \{\mathbf{x} = [x_1, x_2, \ldots, x_D] | x_i > 0, i = 1, 2, \ldots, D; \sum_{i=1}^{D} x_i = \kappa\}$$

The goal is now to work on the simplex (e.g. to apply regression analysis) and therefore we have to introduce proper mathematical tools for this sample space.
For any vector of D real positive compositions we will define a *closure function*, that sums all the components up to a constant value κ. This closure function just rescales the vector. Let $\mathbf{z} = [z_1, z_2, \ldots, z_D] \in \mathbb{R}_+^D$ where $z_i > 0$ for all $i = 1, 2, \ldots, D$.

The closure of $\mathbf{z}$ is defined as

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

That means, that closure is nothing else but a projection of any point in the positive orthant of the D-dimensional real space onto the simplex.
Two vectors of D positive real components $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^D$ $(x_i, y_i \leq 0 \ \forall i = 1, 2, \ldots, D)$ are *compositionally equivalent*, if there exists a positive scalar $\lambda \in \mathbb{R}^+$ such that $\mathbf{x} = \lambda \cdot \mathbf{y}$ and, equivalently, $\mathcal{C}(\mathbf{x}) = \mathcal{C}(\mathbf{y})$.

There are three conditions that should be fulfilled by any statistical method to be applied on compositions: scale invariance, permutation invariance and subcompositional coherence (Aitchison, 1986). These conditions are necessary to receive correct results.
A function $f(.)$ is *scale invariant* if for any positive real value $\lambda \in \mathbb{R}^+$ and for any compostion $\mathbf{x} \in \mathcal{S}^D$, the function satisfies $f(\lambda \mathbf{x}) = f(\mathbf{x})$, i.e. it yields the same result for all vectors that are compositionally equivalent. This property can only be achieved if $f(.)$ is a function only of log-ratios of the parts in $\mathbf{x}$.
A function is *permutation invariant*, if it yields equivalent results when we change the ordering of our parts in the composition.
Subcompositional coherence means that subcompositions should behave as orthogonal projections do in conventional real analysis. The size of a projected segment is less than or equal to the size of the segment itself (cf. Pawlowsky-Glahn et al., 2007).

2.2 The Aitchison geometry

As already mentioned before, for compositional data the Euclidean geometry does not form a proper geometry. This is the reason why we have to work in a different geometry, called the Aitchison geometry. There are two operations that give the simplex a vector space structure. The first one is called perturbation operation.
Perturbation of a composition $\mathbf{x} \in \mathcal{S}^D$ by a composition $\mathbf{y} \in \mathcal{S}^D$ is defined as

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

This operation is analogous to the addition in real space. The second one is analogous to multiplication by a scalar in real space and is called power transformation.
Power transformation of a composition $\mathbf{x} \in \mathcal{S}^D$ by a constant $\alpha \in \mathbb{R}$ is

defined as

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

The simplex $(\mathcal{S}^D, \oplus, \odot)$ with the pertubation and power transformation is a vector space (Pawlowsky-Glahn et al., 2007). To obtain the vector space structure, we take the following inner product:
The *Aitchison inner product* of $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ is given by

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

Hence, we can also define the associated norm:
We define the *Aitchison norm* of $\mathbf{x} \in \mathcal{S}^D$ as the following:

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

And furthermore, the *Aitchison distance* between two vectors $\mathbf{x}$ and $\mathbf{y} \in \mathcal{S}^D$ is defined as:

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

2.3 Important transformations

The aim now is to find suitable transformations to represent our data in coordinates so that they can also be easily interpreted. Since size is irrelevant for compositional data, Aitchison (1986) introduced two transformations based on ratios, for the ratios being the most important information we can gain from this kind of data. First, we need to find an appropriate basis so that any vector $\mathbf{x} \in \mathcal{S}^D$ can be expressed in terms of this basis. Therefore, one has to use a generating system, such as the following used by Aitchison (1986),

$$\mathbf{w}_i = \mathcal{C}[\exp(\mathbf{e}_i)] = \mathcal{C}[1, 1, \ldots, e, \ldots, 1], \; i = 1, 2, \ldots, D,$$

where $\forall \mathbf{w}_i$ Euler's number e is at the i-th position since $\mathbf{e}_i$ is just the i-th unit vector.
Thus, any vector $\mathbf{x} \in \mathcal{S}^D$ can be written as

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

2.3.1 alr transformation

The transformation $\text{alr} : \mathcal{S}^{\text{D}} \to \mathbb{R}^{\text{D}-1}$ assigns the real $(n-1)$-tuple

$$\text{alr}(\mathbf{x}) = \log\left(\frac{x_1}{x_D}, \frac{x_2}{x_D}, \ldots, \frac{x_{D-1}}{x_D}\right)$$

to the composition $\mathbf{x} \in \mathcal{S}^{\mathbf{D}}$. (The part x_D in the denominator could be replaced by any of the other parts.)
It is used to investigate the dimension of the vector space and actually called *additive log-ratio transformation* (alr) (Aitchison, 1986).
There exist coordinates that correspond to that transformation. They are called additive log-ratio coordinates. Thus we need a basis again, where we will choose the one described above: $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{D-1}\}$. Now any vector $\mathbf{x} \in \mathcal{S}^D$ can be written as

$$\mathbf{x} = \bigoplus_{i=1}^{D-1} \ln \frac{x_i}{x_D} \odot \mathbf{w}_i.$$

This is the reason why the perturbation and the power transformation of a composition $\mathbf{x} \in \mathcal{S}^D$ have been introduced previously (see section 2.2).

There are some important properties of the alr transformation:
Firstly, the transformation $\text{alr} : \mathcal{S}^D \to \mathbb{R}^{D-1}$ is one-to-one. If $\mathbf{x}^* \in \mathbb{R}^{D-1}$, then the inverse alr transformation is

$$\text{alr}^{-1}(\mathbf{x}^*) = \mathcal{C}[\exp(x_1^*, x_2^*, \ldots, x_{D-1}^*, 0)].$$

Moreover, the alr transformation is an isomorphism of vector spaces, but it is not symmetrical in the components. Anyway, the essential problem with alr coordinates is the non-isometric characteristic of this transformation. That means, there are coordinates in an oblique basis, something that affects distances if the usual Euclidean distance is computed from the alr coordinates (Pawlowsky-Glahn et al., 2007). For this reason, Aitchison (1986) used alr coordinates for modeling when applying statistical analysis on compositional data, but he could not apply techniques based on a metric. Therefore, he proposed another transformation.

2.3.2 clr transformation

The *centred log-ratio transformation* (clr) gives the expression of a composition in the centred log-ratio coefficients, which are given by

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

where

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

is the component-wise geometric mean of the composition. Since the denominator can be replaced by any constant, the transformation is not unique and therefore the clr transformation is consistent with the concept of compositions as equivalence classes (cf. Barceló-Vidal et al., 2001).
We can also define the inverse transformation which gives us the coefficients in the canonical basis of the real space:

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

where $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_D]$.
The following part will describe the most important properties of the clr transformation and its coefficients. The clr transformation, clr: $\mathcal{S}^D \to U \subset \mathbb{R}^D$ is an isomorphism of $(D-1)$-dimensional vector spaces. Moreover, it is symmetrical in the components. This fact implies a new constraint on the transformed sample. The sum of the components has to be zero:

$$\sum_{i=1}^{D} \xi_i = 0,$$

where the i-th clr coefficient is given by $\xi_i = \frac{\ln x_i}{g(\mathbf{x})}$. That fact also means that the covariance matrix of $\boldsymbol{\xi}$ is singular. Hence, some problems while analysing data may follow. Furthermore, clr coefficients are not subcompositionally coherent (cf. section 2.1). Hence, it may happen that the measured distance between two full compositions is smaller than the distance between them when considering subcompositions. That characteristic would not be reasonable. Moreover, clr coefficients are not coordinates with respect to a basis of the simplex. Anyway, there are also some positive important properties in connection with the Aitchison inner product, norm and distance.

Let $\mathbf{x}, \mathbf{y}$ be compositions in $\mathcal{S}^D$ and $\mathrm{clr}(\mathbf{x})$, $\mathrm{clr}(\mathbf{y})$ their respective clr transformations. Then

$$\begin{aligned} <\mathbf{x}, \mathbf{y}>_a &= <\mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{y})> \\ \|\mathbf{x}\|_a &= \|\mathrm{clr}(\mathbf{x})\| \\ d_a(\mathbf{x}, \mathbf{y}) &= d(\mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{y})) \end{aligned}$$

where $<.,.>$ denotes the inner product, $\|.\|$ the standard norm and $d(.,.)$ the distance in $\mathbb{R}^D$.

Furthermore, the clr transformation is an isometry between $(D-1)$-dimensional Euclidean spaces.

The fact, that neither the alr- nor the clr transformation can be directly associated with an orthogonal coordinate system in the simplex lead Egozcue et al. (2003) to define a new transformation, called isometric log-ratio transformation. It is an isometry between $\mathcal{S}^D$ and $\mathbb{R}^{D-1}$, thus avoiding the drawbacks of both, the alr- and the clr transformation (Pawlowsky-Glahn et al., 2007).

2.3.3 ilr transformation

The *isometric log-ratio transformation* (ilr) was introduced by Egozcue et al. (2003). The aim was to find a transformation that can be directly associated with an orthogonal coordinate system in the simplex. Thus, it should be both isometric and an isomorphism to improve the situation that we are confronted with in the situation of the alr- as well as the clr transformation. To obtain an orthonormal basis, we omit the last element $\mathbf{w}_D$ of the generating system in section 2.3. The resulting system $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{D-1}\}$ is a basis, but it is not orthonormal yet. Once an independent set of $D-1$ compositions in $\mathcal{S}^D$ is given, an orthonormal basis with respect to the inner product can be found using the Gram-Schmidt procedure, since this is known to be so for any vector space with an inner product (related to the Aitchison geometry). Consequently, the only drawback results in the non-uniqueness of the basis. Finally we will define the isometric log-ratio transformation as the following: Let $\mathbf{e}_i, i = 1, 2, \ldots, D-1$, be an orthonormal basis in $\mathcal{S}^D$. The coordinate function assigning coordinates with respect to $\mathbf{e}_i, i = 1, 2, \ldots, D-1$, to a composition $\mathbf{x} \in \mathcal{S}^D$ is called isometric log-ratio transformation, if

$$\begin{aligned} &\text{ilr} : \mathcal{S}^D \to \mathbb{R}^{D-1} \\ &\text{ilr}(\mathbf{x}) = (< \mathbf{x}, \mathbf{e}_1 >_a, < \mathbf{x}, \mathbf{e}_2 >_a, \ldots, < \mathbf{x}, \mathbf{e}_{D-1} >_a). \end{aligned}$$

Any $\mathbf{x} \in \mathcal{S}^D$ can be expressed as

$$\mathbf{x} = \text{ilr}^{-1}(\mathbf{x}^*) = \bigoplus_{i=1}^{D-1} x_i^* \odot \mathbf{e}_i \ , \ x_i^* =< \mathbf{x}, \mathbf{e}_i >_a$$

where $\mathbf{x}^* = [x_1^*, x_2^*, \ldots, x_{D-1}^*]$ is the vector of coordinates. An important property of the ilr transformation is that it is an isometric isomorphism of vector spaces, i.e. if $\delta \in \mathbb{R}, \mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ and $\mathbf{x}^* = \text{ilr}(\mathbf{x}), \mathbf{y}^* = \text{ilr}(\mathbf{y})$, then

$$\text{ilr}(\mathbf{x} \oplus \mathbf{y}) = \mathbf{x}^* + \mathbf{y}^*, \quad \text{ilr}(\delta \odot \mathbf{x}) = \delta \mathbf{x}^*.$$

Moreover, there are some more characteristics of the ilr transformation: Consider $\mathbf{x}_k \in \mathcal{S}^D$ and real constants α, β, then:

$$\begin{aligned} \text{ilr}(\alpha \odot \mathbf{x}_1 \oplus \beta \odot \mathbf{x}_2) &= \alpha \cdot \text{ilr}(\mathbf{x}_1) + \beta \cdot \text{ilr}(\mathbf{x}_2) = \alpha \cdot \mathbf{x}_1^* + \beta \cdot \mathbf{x}_2^* \\ < \mathbf{x}_1, \mathbf{x}_2 >_a &= < \text{ilr}(\mathbf{x}_1), \text{ilr}(\mathbf{x}_2) > = < \mathbf{x}_1^*, \mathbf{x}_2^* > \\ \|\mathbf{x}_1\|_a &= \|\text{ilr}(\mathbf{x}_1)\| = \|\mathbf{x}_1^*\| \\ d_a(\mathbf{x}_1, \mathbf{x}_2) &= d(\text{ilr}(\mathbf{x}_1), \text{ilr}(\mathbf{x}_2)) = d(\mathbf{x}_1^*, \mathbf{x}_2^*) \end{aligned}$$

Referring to Pawlowsky-Glahn et al. (2007), the main difference between the properties of clr and ilr is that the clr refers to vectors of coefficients in $\mathbb{R}^D$, whereas the latter deals with vectors of coordinates in $\mathbb{R}^{D-1}$ and is thus matching the actual dimension of $\mathcal{S}^D$.
There is one matrix (given in Chapter 4.4 of Pawlowsky-Glahn et al., 2007) that has some special properties, and it is useful when considering the relationship between the transformations. We will call it $\mathbf{\Psi}$. Let $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$ again be a generic orthonormal basis of the simplex. The $(D-1, D)$-matrix $\mathbf{\Psi}$ contains $\text{clr}(\mathbf{e}_i)$ in its rows. Moreover, any orthonormal basis satisfies $< \mathbf{e}_i, \mathbf{e}_j >_a = \delta_{ij}$, where δ_{ij} is the Kronecker-delta. Hence, formally it can be written down as $< \mathbf{e}_i, \mathbf{e}_j >_a = < \text{clr}(\mathbf{e}_i), \text{clr}(\mathbf{e}_j) > = \delta_{ij}$. Therefore, the matrix $\mathbf{\Psi}$ satisfies $\mathbf{\Psi}\mathbf{\Psi}' = \mathbf{I}_{D-1}$, being $\mathbf{I}_{D-1}$ the identity matrix of dimension $D-1$. To recover the compositions of the basis from $\mathbf{\Psi}$ it is now just necessary to apply clr^{-1} in each row of the matrix. Another important fact is that these rows of $\mathbf{\Psi}$ add up to 0 because they are clr coefficients. We will return to this issue later.

2.3.4 Balances

For defining balances, we will apply a sequential binary partition that was developed by Egozcue and Pawlowsky-Glahn (2005) on compositional vectors. At first we part our composition into two groups and mark the r components of one with +1 and the s of the other group with -1. Then each group is again split into two and this procedure continues until all groups have a single part. The i-th step symbolizes the i-th order partition and all groups that are not split in the i-th partition are signed with 0.
Then the balance is defined as the normalised log-ratio of the geometric mean of each group of parts:

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

This means that, for the i-th balance, the parts receive a weight of either

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

a_0 is for those parts, that are not involved in the splitting.
Then we can write the i-th balance as

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

where a_{ij} equals a_+ if the code in the i-th order partition is $+1$ for the j-th part and a_- if the code is -1 and a_0 if the code is zero. A very interesting fact is that this matrix with entries a_{ij} is just the matrix $\mathbf{\Psi}$, that we have defined in section 2.3.3.
Since the geometric mean is used in the nominator as well as in the denominator, its ratio measures the relative weight of each group. The logarithm is used to provide the appropriate scale. A positive balance means, that in (geometric) mean, the group of parts in the numerator has more weight in the composition than the group in the denominator (and conversely for negative balances) (Pawlowsky-Glahn et al., 2007). Furthermore, balances are useful because it is possible to obtain information within the groups, and sometimes it is even more interesting to know something about that than just the relation between the two groups. Summarising, we can state that balances project compositions onto special subspaces just by retaining some balances and making other ones null and this is a useful and important task while doing applications with compositional data.

It is not just valid for balances, but for all orthogonal bases, that when we are performing analysis of compositional data, results that could be obtained using compositions and the Aitchison geometry are exactly the same as those obtained using the coordinates of the compostions and the ordinary Euclidean geometry. That fact helps to facilitate working with compositions but it is very crucial to select the basis carefully due to interpretative reasons.

2.3.5 Relationships between transformations

Again we consider the $(D-1, D)$-matrix $\mathbf{\Psi}$ whose rows are $\text{clr}(\mathbf{e}_i)$, associated with an orthonormal basis of the simplex $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$. There are some relations between the ilr(.), alr(.) and the clr(.) transformation. First we will define $\mathbf{\Xi}' = [\mathbf{I}_{D-1} : -\mathbf{1}'_{D-1}]$, where $\mathbf{I}_{D-1}$ is the identity matrix of dimension

$(D-1)$ and $\mathbf{1}_{D-1}$ is a $(D-1)$ row vector of units. Furthermore $\boldsymbol{\Upsilon}$ is the Moore-Penrose generalised inverse of $\boldsymbol{\Xi}$, that is

$$\boldsymbol{\Upsilon} = \frac{1}{D}\begin{pmatrix} D-1 & -1 & -1 & \cdots & -1 & -1 \\ -1 & D-1 & -1 & \cdots & -1 & -1 \\ -1 & -1 & D-1 & \cdots & -1 & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -1 & -1 & -1 & \cdots & D-1 & -1 \end{pmatrix}.$$

The resulting relations are stated below:

$$\begin{aligned} \mathbf{x}^* = \operatorname{ilr}(\mathbf{x}) &= \operatorname{clr}(\mathbf{x}) \cdot \boldsymbol{\Psi}' = \operatorname{alr}(\mathbf{x}) \cdot \boldsymbol{\Upsilon} \cdot \boldsymbol{\Psi}' \\ \operatorname{clr}(\mathbf{x}) &= \operatorname{alr}(\mathbf{x}) \cdot \boldsymbol{\Xi} \\ \operatorname{clr}(\mathbf{x}) &= \operatorname{ilr}(\mathbf{x}) \cdot \boldsymbol{\Psi} \\ \operatorname{alr}(\mathbf{x}) &= \operatorname{ilr}(\mathbf{x}) \cdot \boldsymbol{\Psi} \cdot \boldsymbol{\Xi} \end{aligned}$$

2.4 Elements of simplicial statistics

Standard methods in descriptive statistics, e.g. the arithmetic mean, are not very informative when applying them on compositions. Therefore, it is necessary to introduce alternatives.
At first we will define the so called *centre*, the closed geometric mean or also called simplex-average, that is a measure of central tendency for compositional data. For a data set $\mathbf{X} = [x_{ij}]$ with n observations (rows) and D parts (columns) we can write the centre $\mathbf{g}$ as

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

with $g_j = (\prod_{i=1}^{n} x_{ij})^{\frac{1}{n}}, j = 1, 2, \ldots, D$. Hence, the geometric mean is considered column-wise.

There are two ways to describe dispersion in a compositional data set. Either one uses the *variation matrix*, originally defined by Aitchison (1986), given by

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

Var(.) stands for the variance of the log-ratio of parts i and j.
Or, dispersion can also be described by means of the *normalised variation matrix*, where the entries of the matrix $\mathbf{T}^*$ are given by $t^*_{ij} = \mathrm{Var}(\frac{1}{\sqrt{2}} \ln \frac{x_i}{x_j})$, so t^*_{ij} is the variance of the normalised log-ratio of parts i and j, and hence the log-ratio is a balance.

A measure of global dispersion is the *total variance* of a random composition given by

$$\mathrm{TotVar}[\mathbf{x}] = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \mathrm{Var}\left(\ln \frac{x_i}{x_j} \right) = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} t_{ij} = \frac{1}{D} \sum_{i=1}^{D} \sum_{j=1}^{D} t^*_{ij}.$$

Due to our work with ratios, all the upper measures are independent of the constant κ, consequently rescaling has no effect. Moreover, the variation matrix, in both versions, explains how the total variation is split among the parts (among all log-ratios).

Chapter 3

Regression with compositional data

Regression analysis is one of the most important tools in statistical analysis. In real space, a linear regression model can be written in terms of a conditional expected value as

$$\mathbb{E}(Y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_D x_D \tag{3.1}$$

with unknown parameters $\beta_0, \ldots, \beta_D$ that need to be estimated. That regression model is just reasonable for data carrying absolute information, but in case of compositional data, where components of the explanatory variables x_i carry only relative information, such a model is inappropriate. Note that also the response can be compositional, but then we are in the context of multivariate linear regression.
Since most of the standard statistical methods are designed for the usual Euclidean geometry, but not for the Aitchison one, the family of log-ratio transformations from the simplex with the Aitchison geometry to the Euclidean real space was introduced (see Filzmoser et al., 2012).
The isometric log-ratio transformation is preferable because it expresses the original compositions in $D-1$ orthonormal coordinates with respect to the Aitchison geometry. The proper choice of the orthonormal coordinates is crucial since we want to find a good interpretation of the result in the real space.

3.1 Regression with compositional response

This chapter relies mainly on three papers (Egozcue et al., 2011; Pawlowsky-Glahn et al., 2007; Filzmoser et al., 2012). We consider a linear regression

model with compositional response. Evidently, the response has to be multivariate then. Hence we are just able to distinguish between two cases: univariate predictors, multivariate predictors. The goal is to estimate the parameters while using suitable methods for analysing such special data.

3.1.1 Multivariate response, multivariate predictors

Let $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iD})', i = 1, \ldots, n$ represent a data set in which the i-th observation is a composition. Furthermore, let $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ be the values of p (non-compositional) covariates. A prediction in the simplex $\mathcal{S}^D$ consists of a deterministic function of the covariates $\mathbf{p}(\mathbf{x}) \in \mathcal{S}^D$ and a perturbation-additive residual $\mathbf{e} \in \mathcal{S}^D$. $\mathbf{B} = (\mathbf{b}_0', \mathbf{b}_1', \ldots, \mathbf{b}_p')'$ is the $(p+1) \times D$ - matrix including the unknown compositional regression parameters to be estimated. A linear predictor in the simplex is defined as

$$\mathbf{p}(\mathbf{x}) = \mathbf{b}_0 \oplus \bigoplus_{j=1}^{p} (\mathbf{x}_j \odot \mathbf{b}_j) \ , \ \mathbf{b}_j \in \mathcal{S}^D.$$

For modeling we use least squares regression. Therefore, we have to find estimates $\hat{\mathbf{b}}_j$ of the compositional coefficients $\mathbf{b}_j, j = 0, 1, \ldots, p$ in our model:

$$\begin{aligned} \mathbf{y}_i &= \mathbf{p}(\mathbf{x}_i) \oplus \mathbf{e}_i \\ \mathbf{y}_i &= \mathbf{b}_0 \oplus \bigoplus_{j=1}^{p} (x_{ij} \odot \mathbf{b}_j) \oplus \mathbf{e}_i \ , \ i = 1, 2, \ldots, n \end{aligned}$$

The estimation is performed by minimizing the sum of square-norms of the error

$$\text{SSE} = \sum_{i=1}^{n} \|\mathbf{e}_i\|_a^2 = \sum_{i=1}^{n} \|\mathbf{p}(\mathbf{x}_i) \ominus \mathbf{y}_i\|_a^2.$$

Since we are still working in the Aitchison geometry, all expressions are referred to it. Thus, we cannot say anything about the result yet, and that is the reason why we will transform the compositions and express them in orthonormal coordinates of the simplex (see section 2.3.3). Let the transformed compositions be marked by asterisk ($\text{ilr}(\mathbf{y}_i) = \mathbf{y}_i^* = (y_{i1}^*, y_{i2}^*, \ldots, y_{i,D-1}^*)'$ and $\text{ilr}(\mathbf{b}_0) = \mathbf{b}_0^*, \text{ilr}(\mathbf{b}_j) = \mathbf{b}_j^*, j = 1 \ldots, D-1$). Thus the model can be rewritten in coordinates:

$$\mathbf{y}_i^* = \mathbf{b}_0^* + \sum_{j=1}^{p} (x_{ij} \cdot \mathbf{b}_j^*) + \mathbf{e}_i^* \ , \ i = 1, 2, \ldots, n$$

and hence

$$\text{SSE} = \sum_{i=1}^{n} \|\mathbf{e}_i^*\|^2$$

which is a consequence of the isometric character of ilr(.).
Let us use matrix notation now. $\mathbf{B}^* = (\mathbf{b}_0^{*\prime}, \mathbf{b}_1^{*\prime}, \ldots, \mathbf{b}_p^{*\prime})'$ is the $(p+1)\times(D-1)$ matrix of regression parameters in coordinates and $\mathbf{Y}^* = (\mathbf{y}_1^{*\prime}, \mathbf{y}_2^{*\prime}, \ldots, \mathbf{y}_n^{*\prime})'$ the $n \times (D-1)$ matrix of response compositions in coordinates and $\mathbf{X} = ((1, \mathbf{x}_1')', (1, \mathbf{x}_2')', \ldots, (1, \mathbf{x}_n')')'$ the $n \times (p+1)$ design matrix.
Thus the (usual) least-squares estimate of $\mathbf{B}^*$ equals

$$\hat{\mathbf{B}}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}^*.$$

One can show that the least-squares regression problem in the simplex is equivalent to $D-1$ ordinary least-squares problems for the coordinates (Egozcue et al., 2011). Furthermore, the result of the problem in coordinates is independent of the selected orthonormal basis and can be solved independently. Finally, the estimated covariance matrix $\mathbf{\Sigma}_{\mathbf{e}^*}$ of the (mutually uncorrelated) errors $\mathbf{e}_i^*$ is

$$\mathbf{\Sigma}_{\mathbf{e}^*} = \frac{1}{n-p}(\mathbf{Y}^* - \mathbf{X}\hat{\mathbf{B}}^*)'(\mathbf{Y}^* - \mathbf{X}\hat{\mathbf{B}}^*).$$

Usually, the regression parameters are back-transformed after estimation and the results are interpreted directly on the simplex. On the other hand, inference concerning regression parameters (significance testing etc.) can only be performed (and interpreted) in the orthonormal coordinates (Filzmoser et al., 2012).

Finally, we can summarize the steps in the procedure to estimate the regression coefficients in a model with a compositional response (cf. Egozcue et al., 2011):

- an orthonormal basis has to be selected, a good choice would be a sequential binary partition of the compositional response vector,
- representation of the compositional response by means of its orthonormal coordiantes that are possibly balance-coordinates,
- perform least-squares estimation of the regression coefficients and the sums of squares for each coordinate of the response using the available covariates,
- reconstruct the compositional coefficients, predictors and residuals.

These steps correspond to the principle of working on coordinates (Mateu-Figueras et al., 2011).

Coefficient of determination

To derive the coefficient of determination in that model, we need to define the decomposition of the total sum of squares in the simplex:

$$\widehat{SST} = \sum_{i=1}^{n} \|\mathbf{y}_i \ominus \mathbf{g}(\mathbf{Y})\|_a^2$$

where $\mathbf{g}(\mathbf{Y})$ is the geometric mean of the sample response (for definition see section 2.4), which is the natural estimator of the centre of the (random) composition $\mathbf{y}$: $\text{Cen}[\mathbf{y}] = \text{ilr}^{-1}\mathbf{E}[\text{ilr}(\mathbf{y})]$.

The decomposition of $\widehat{SST}$ is then defined by

$$\widehat{SST} = \widehat{SSR} + \widehat{SSE},$$

where $\widehat{SSR} = \sum_{i=1}^{D} \|\widehat{\mathbf{p}}(\mathbf{x}_{ij}) \ominus \mathbf{g}(\mathbf{Y})\|_a^2$. Finally, we define the coefficient of determination coefficient as:

$$R^2 = \frac{\widehat{SSR}}{\widehat{SST}} = 1 - \frac{\widehat{SSE}}{\widehat{SST}}$$

The received value R^2 is the per unit of metric (or total)-variance of the compositional response explained by the regression.

The coefficient of determination can also be expressed in terms of the sums of squares of the regression for the coordinates:

$$R^2 = \frac{\sum_{j=1}^{D-1} \widehat{SSR_j}}{\widehat{SST}} = \frac{\sum_{j=1}^{D-1} \widehat{SST_j} \cdot R_j^2}{\widehat{SST}}$$

where $R_j^2 = \widehat{SSR_j}/\widehat{SST_j}$ and $\widehat{SSR} = \sum_{j=1}^{D-1} \widehat{SSR_j}$.

Inference statistics

To obtain inference statistics or rather conclusions about certain hypotheses, we assume the normality of the random errors $\mathbf{e}_i, i = 1, \ldots, n$. The first question we want to find an answer to is, if the whole matrix of estimated coefficients $\hat{\mathbf{B}}^*$ is equal to zero. This result would lead to the conclusion, that the set of independent explanatory variables does not describe the multivariate response coordinates at all. The hypotheses for this issue are given by:

$$\begin{aligned} H_0 &: \hat{\mathbf{B}}^* = 0 \\ H_1 &: \hat{\mathbf{B}}^* \neq 0 \end{aligned}$$

The corresponding teststatistic

$$\Lambda = \frac{|\mathbf{Y}^{*\prime}\mathbf{Y}^* - \hat{\mathbf{B}}^{*\prime}\mathbf{X}'\mathbf{X}\hat{\mathbf{B}}^*|}{|\mathbf{Y}^{*\prime}\mathbf{Y}^* - n\bar{\mathbf{y}}^*\bar{\mathbf{y}}^{*\prime}|}$$

is called Wilks Λ with $\bar{\mathbf{y}}^* = (\bar{y}_1, \ldots, \bar{y}_D)$ and under the null hypothesis it follows: $\Lambda \sim \Lambda(D, n-p-1, p)$ (cf. Fahrmeir et al., 1996). Unfortunately, distributions of this test statistic are only asymptotic (Filzmoser and Hron, 2012) and therefore we will consider other tests.
The hypothesis of interest is given by the question: does the j-th explanatory variable ($j = 1, \ldots, p$) have a siginificant influence on the response variables? Formally the null hypothesis is expressed as

$$\begin{aligned} H_0 &: \mathbf{h}_j'\hat{\mathbf{B}}^* = \mathbf{0} \\ H_1 &: \mathbf{h}_j'\hat{\mathbf{B}}^* \neq \mathbf{0}, \end{aligned}$$

where $\mathbf{h}_j$ equals a p-part column vector with zero entries with the exception of an entry of 1 at position $(j+1)$. $\hat{\mathbf{B}}^*$ is the matrix of estimated regression parameters in coordinates.
Under the null hypothesis the following test statistic holds,

$$F_j = \frac{\mathbf{h}_j'\hat{\mathbf{B}}^*(\mathbf{Y}^*\mathbf{M}_X\mathbf{Y}^*)^{-1}\hat{\mathbf{B}}^{*\prime}\mathbf{h}_j}{\mathbf{h}_j(\mathbf{X}'\mathbf{X})^{-1}\mathbf{h}_j}\frac{n-p-D}{D} \sim F_{D,n-p-D}, \tag{3.2}$$

where $\mathbf{M}_X = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ ($\mathbf{I}_n$ stands for the identity matrix of order n). If the realization of the statistic (3.2) exceeds the quantile $F_{D,n-p-D;1-\alpha}$, the null hypothesis is rejected on a significance level α (cf. Filzmoser and Hron, 2012).
Moreover another hypothesis can be formulated: the j-th explanatory variable does not have a significant influence on the k-th response variable ($k = 1, \ldots, D$). In that case the $(j+1, k)$-th entry of the matrix $\hat{\mathbf{B}}^*$ equals zero. Formally the null hypothesis as well as the corresponding alternative hypothesis are expressed as:

$$\begin{aligned} H_0 &: \mathbf{h}_j'\hat{\mathbf{B}}^*\mathbf{m}_k = 0 \\ H_1 &: \mathbf{h}_j'\hat{\mathbf{B}}^*\mathbf{m}_k \neq 0. \end{aligned}$$

The vector $\mathbf{m}_k$ defines a D-part vector with zero entries except a 1 at position k. Under the null hypothesis the following test statistic

$$F_{jk} = \frac{\mathbf{h}_j'\hat{\mathbf{B}}^*\mathbf{m}_k(\mathbf{m}_k'\mathbf{Y}^*\mathbf{M}_X\mathbf{Y}^*\mathbf{m}_k)^{-1}\mathbf{m}_k'\hat{\mathbf{B}}^{*\prime}\mathbf{h}_j}{\mathbf{h}_j(\mathbf{X}'\mathbf{X})^{-1}\mathbf{h}_j}\frac{n-p-1}{1} \tag{3.3}$$

is $F_{1,n-p-1;1-\alpha}$ distributed at a significance level α.
The interpretation lies in the choice of the transformation of the compositions in the model.

3.1.2 Multivariate response, univariate predictors

The idea will be written down just shortly, the main difference to the general case is that the covariates are just one-dimensional. Hence, our model can be described as follows:

$$\begin{aligned} \mathbf{y}_i &= (y_{i1}, y_{i2}, \ldots, y_{iD})', \quad i = 1, \ldots, n \\ \mathbf{x}_i &= (1, x_{i1})', \quad i = 1, \ldots, n \text{ and } x_{i0} = 1 \\ \mathbf{b}_j &= (b_{j1}, b_{j2}, \ldots, b_{jD})' \in \mathcal{S}^D, \quad j = 0, 1 \\ \mathbf{e}_i &= (e_{i1}, e_{i2}, \ldots, e_{iD})' \in \mathcal{S}^D, \quad i = 1, \ldots, n \end{aligned}$$

The model to find the appropriate coefficients is then given by

$$\mathbf{y}_i = \mathbf{b}_0 \oplus (x_{i1} \odot \mathbf{b}_1) \oplus \mathbf{e}_i \ , \ i = 1, 2, \ldots, n,$$

and the resulting model in coordinates can be written as

$$\mathbf{y}_i^* = \mathbf{b}_0^* + x_{i1} \cdot \mathbf{b}_1^* + \mathbf{e}_i^*, i = 1, \ldots, n.$$

It is now possible to use all the methods which are mentioned above for this model.

3.2 Regression analysis with compositional explanatory variables

The aim of this section is to estimate parameters from a linear regression model, when a multivariate (non-compositional) response is predicted by compositional explanatory variables. First attempts have been done by using the clr(.) transformation, which results for the multiple case in a model

$$\mathbb{E}(Y|\mathbf{x}) = \beta_0 + \sum_{i=1}^{D} \beta_i \xi_i.$$

The problem is, that the resulting clr variables are singular like it was already mentioned in section 2.3.2, and thus, regression parameters should have been estimated using the theory of singular linear models. Moreover, the clr variables as a whole explain some ratios more than once. The subcompositional incoherence indicates problems as well: any subset would alter each clr variable because all the parts in the currently used subset are contained in the denominator (the geometric mean).
That is the main reason why these computations are usually performed in

orthogonal coordinates. Therefore, we will apply an ilr transformation on the composition $\mathbf{x}$, which seems to be the only method to achieve a regression model without constraints and with a meaningful interpretation of the unknown parameters.
A good choice of an orthonormal basis is resulting in a $(D-1)$-dimensional real vector $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, z_{D-1}^*)$, where the components are defined as

$$x_i^* = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}}, \; i = 1, \ldots, D-1$$

These coordinates are a special form of balances, which were described in section 2.3.4, namely the partition is always one element and all the others.
The inverse transformation of $\mathbf{x}^*$ to the original composition $\mathbf{x}$ is then given, before closure, by

$$\begin{aligned}
x_1 &= \exp\left(\frac{\sqrt{D-1}}{\sqrt{D}} x_1^*\right), \\
x_i &= \exp\left(-\sum_{j=1}^{i-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} x_j^* + \frac{\sqrt{D-i}}{\sqrt{D-i+1}} x_i^*\right), \; i = 2, \ldots, D-1 \\
x_D &= \exp\left(-\sum_{j=1}^{D-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} x_j^*\right).
\end{aligned}$$

When using this form of balances, the variable x_1^* represents all the relevant information about the compositional part x_1 and moreover, it is invariant against permutation of the parts $x_2, \ldots, x_D$. Obviously, the coordinate x_2^* does not explain all the relative information about x_2, because the part x_1 is not contained therein.
From that fact follows, that we will probably construct another orthonormal basis where the first ilr coordinate explains the compositional part x_2 and then another, until we have finally $D-1$ coordinates, that explain all the relative information about their compositional part x_i, respectively. Explicitely, we can write down this construction in the following way:
For $l = 1, 2, \ldots, D$, the D-tuple $(x_1, x_2, \ldots, x_D)$ is replaced by

$$(x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D) := (x_1^{(l)}, x_2^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)}).$$

The corresponding ilr-transformation is

$$x_i^{*(l)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}}, \; i = 1, \ldots, D-1 \tag{3.4}$$

and we have $x_i^{*(1)} = x_i^*$ for $i = 1, 2, \ldots, D-1$.

3.2.1 Univariate response, multivariate compositional explanatory variables

When we have a sample of n observations of the response $y_i \in \mathbb{R}, i = 1, \ldots, n$ and of the compositional explanatory variables $\mathbf{x}_i \in \mathcal{S}^D, i = 1, \ldots, n$ we obtain a standard multiple linear regression of y_i on the ilr-transformed explanatory variables $\mathbf{x}_i^* \in \mathbb{R}^{D-1}$ with the regression coefficients $c_j, j = 0, \ldots, D-1$:

$$\mathbb{E}(y_i|\mathbf{x}_i^*) = c_0 + \sum_{j=1}^{D-1} x_{ij}^* c_j$$

The regression coefficients $c_j, j = 1, \ldots, D-1$ can be estimated by the least squares method. The intercept term is directly related to the response y_i and therefore it is not depending to the choice of the orthonormal basis on the simplex. Since the other coefficients are directly connected to the ilr coordinates, their interpretation is difficult and therefore we will now consider the l-th ilr basis, for $l = 1, \ldots, D$ (cf. equation 3.4), which will lead to the following regression model:

$$\mathbb{E}(y_i|\mathbf{x}_i^*) = c_0 + c_1^{(l)} x_1^{*(l)} + \cdots + c_{D-1}^{(l)} x_{D-1}^{*(l)} \tag{3.5}$$

Due to the orthogonality of different ilr bases, the intercept term c_0 as well as the model fit remains unchanged (Hron et al., 2012). Since $x_1^{*(l)}$ explains all the relative information about part $x_1^{(l)}$, the coefficient $c_1^{(l)}$ can be assigned to this part. Since we cannot interpret the other regression coefficients in a reasonable way we have to consider D different regression models according to model (3.5) by taking $l \in \{1, \ldots, D\}$ and hence interpret the coefficient $c_1^{(l)}$ which represents part $x_1^{(l)}$. It has to be noted that a regression model with the ilr variables $x_1^{*(1)}, \ldots, x_1^{*(D)}$ would not be appropriate because it results in singularity (Hron et al., 2012).

Finally, the regression model can be written down as:

$$y_i = c_0 + c_1 z_{i1} + \cdots + c_{D-1} z_{i,D-1} + e_i, \; i = 1, \ldots, n$$

The random variables $\mathbf{e} = (e_1, \ldots, e_n)'$ are assumed to be uncorrelated and with the same variance σ^2.
Further explanation and parameter estimation will be done in the next section for the general case of a multivariate response and multivariate covariates.

3.2.2 Inference statistics in multiple linear regression models with compositional explanatory variables

Now that we have developed a reasonable model, the estimated parameters should be tested. One interesting aspect to test are hypotheses on the parameters $\mathbf{c}_0, \mathbf{c}_1, \ldots, \mathbf{c}_{D-1}$. Therefore, the error terms $\mathbf{e}$ require certain assumptions, namley $\mathbf{e}$ should be multivariate normally distributed with mean vector $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$.

The test statistic associated with the significance of the individual regression parameters is:

$$\begin{aligned} T_0 &= \frac{\hat{c}_0}{\sqrt{\frac{(\mathbf{y}-\mathbf{X}^*\hat{\mathbf{c}})'(\mathbf{y}-\mathbf{X}^*\hat{\mathbf{c}})}{n-D}\{(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\}_{0,0}}} \\ T_i &= \frac{\hat{c}_0}{\sqrt{\frac{(\mathbf{y}-\mathbf{X}^*\hat{\mathbf{c}})'(\mathbf{y}-\mathbf{X}^*\hat{\mathbf{c}})}{n-D}\{(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\}_{i,i}}}, \; i = 1, \ldots, D-1 \end{aligned}$$

The hypotheses for T_0 are defined as:

$$\begin{aligned} H_0 &: \; c_0 = 0 \\ H_1 &: \; c_0 \neq 0. \end{aligned}$$

and the corresponding hypotheses for T_1 are defined as:

$$\begin{aligned} H_0 &: \; c_i = 0 \\ H_1 &: \; c_i \neq 0, \end{aligned}$$

for $i = 1, \ldots, D-1$. Assuming the validity of the null hypotheses, T_0 and T_i follow a Student t-distribution with $n - D$ degrees of freedom.
The reason why we actually test just for the first two parameters lies in the ilr transformation. We cannot properly interpret the other parameters. But, more general, we can test $c_1^{(l)}$ for $l = 1, \ldots, D$. The goal of this test is to find out, if a subcomposition of the given compositional covariate can replace the original composition in the regression model (cf. Hron et al., 2012). Another important task for inference in regression analysis is whether the values of $\mathbf{Y}$ at all depend on values of the ilr coordinates $x_1^*, \ldots, x_{D-1}^*$. That means we want to test whether all the parameters c_i, for $i = 1, \ldots, D-1$ are equal to 0. Formally the hypothesis is defined as:

$$\begin{aligned} H_0 &: \; c_i = 0 \; \forall i \\ H_1 &: \; c_i = 0 \text{ for at least one } i \end{aligned}$$

The following test-statistic is used:

$$F = \frac{1}{(D-1)S^2}\hat{\mathbf{c}}_*'\{(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\}_{(-1,-1)}\hat{\mathbf{c}}_*$$

where $\hat{\mathbf{c}}_* = (\hat{c}_1, \ldots, \hat{c}_{D-1})'$ and $\{(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\}_{(-1,-1)}$ denotes that the first row and the first column were excluded from the matrix $(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}$. Furthermore, S^2 denotes an unbiased estimator of the residual variance σ^2: $S^2 = (\mathbf{y} - \mathbf{X}^*\hat{\mathbf{c}})'(\mathbf{y} - \mathbf{X}^*\hat{\mathbf{c}})/(n-D)$. If the null hypothesis holds, this test statistic follows the Fisher F-distribution with $D-1$ and $n-D$ degrees of freedom. An important property of this test statistic is that F is invariant with respect to a change of the order of $x_1^{(l)}, \ldots, x_D^{(l)}$ in equation (3.4).
For compositional data analysis there also exists the coefficient of determination R^2, given as

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i + \bar{y})^2}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ and $(\hat{y_1}, \ldots, \hat{y_n})' = \mathbf{X}^*\hat{\mathbf{c}}$ are predicted values of the response variable. Values close to one indicate a strong relation between the explanatory variables to the response.

3.2.3 Multivariate response, multivariate compositional explanatory variables

In this section we consider a model with a non-compositional multivariate response $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iq})$ and covariates forming a composition $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$:

$$\mathbf{y}_i = \mathbf{c}_0 + \sum_{j=1}^{D-1}(x_{ij}^* \cdot \mathbf{c}_j) + \mathbf{e}_i,\ i = 1, 2, \ldots, n \tag{3.6}$$

with $x_i^* = (x_{i1}^*, x_{i2}^*, \ldots, x_{i,D-1}^*)'$ as the coordinates of the ilr-transformation and errors $\mathbf{e}_i$. We can rewrite the model in matrix notation:

$$\mathbf{Y} = \mathbf{X}^*\mathbf{C} + \mathbf{E}$$

where $\mathbf{Y}$ is the $(n \times q)$-matrix of responses, $\mathbf{X}^* = ((1, \mathbf{x}_1')', (1, \mathbf{x}_2')', \ldots, (1, \mathbf{x}_n')')'$ is the $(n \times D)$-matrix of ilr coefficients with an intercept vector of ones included, and $\mathbf{C} = (\mathbf{c_0}, \ldots, \mathbf{c}_{D-1})'$ is the $(D \times q)$-matrix of the regression parameters. Again, the sum of squares SSE will be minimized to obtain the least-squares solution which is given by

$$\hat{\mathbf{C}} = [(\mathbf{X}^*)'\mathbf{X}^*]^{-1}(\mathbf{X}^*)'\mathbf{Y}.$$

Further, the estimated covariance matrix of the errors e_i is

$$\hat{\mathbf{\Sigma}}_{\mathbf{e}} = \frac{1}{n-D}(\mathbf{Y} - \mathbf{X}^*\hat{\mathbf{C}})'(\mathbf{Y} - \mathbf{X}^*\hat{\mathbf{C}}).$$

Since we want to analyse the effect of the original compositional parts on the response, one has to think about how to do that. We cannot use all the ilr variables $x_1^{*(1)}, \ldots, x_1^{*(D)}$, because then the regression model would result in singularity again (one has to remember, that these coordinates are just multiples of the clr coordinates). So therefore, the model should be done with that permutation of the compositions, so that the l-th coordinate $x_1^{*(l)}$, which we want to analyse in particular is on position one and after obtaining the parameters, all the relative information about that compositional part can be interpreted (cf. 3.2.1).

Inference statistics

Hypotheses on the model coefficients and test statistics in case of a multivariate regression model can be adapted from section 3.1.1. Instead of $\mathbf{X}$ coordinates $\mathbf{X}^*$ have to be considered. On the other side, due to our assumptions of the model $\mathbf{Y}$ and $\hat{\mathbf{C}}$ are not coordinates.

3.3 Regression with compositional response and compositional explanatory variables

In this case we are confronted with a situation, where $\mathbf{y}_i$ as well as $\mathbf{x}_i$ form compositions. Hence, we consider a multivariate response and multivariate predictors to define our regression model.
The situation is different compared to those mentioned in the section above, since both $\mathbf{y}_i$ and $\mathbf{x}_i$ have to be ilr-transformed. So there is one important question that turns up when thinking about applying the ilr transformation, that is: should the variables $\mathbf{y}_i$ and $\mathbf{x}_i$ be transformed all together or separately? A joint transformation would result in one big matrix $\mathbf{Z}$ including all ilr-transformed variables of $\mathbf{X}$ and $\mathbf{Y}$. Alternatively, when applying ilr transformation on both $\mathbf{X}$ and $\mathbf{Y}$ separately we obtain individual coordinates $\mathbf{Y}^*$ and $\mathbf{X}^*$.
There are good reasons to choose a separate transformation, because otherwise the information contained in $\mathbf{X}$ and $\mathbf{Y}$ would get mixed in $\mathbf{Z}$.
Let us write down all the necessary model variables and parameters now.

Therefore, we consider a number of observations of n:

$$\begin{aligned} \mathbf{y}_i &= (y_{i1}, \ldots, y_{iD}) \in \mathcal{S}^D, i = 1, \ldots, n \\ \mathbf{x}_i &= (x_{i1}, \ldots, x_{iP}) \in \mathcal{S}^P, i = 1, \ldots, n \end{aligned}$$

where $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)'$ is the $(n \times D)$-matrix of compositional responses, $\mathbf{X} = ((1, \mathbf{x}_1'), \ldots, (1, \mathbf{x}_n'))'$ is the $(n \times (P+1))$-matrix of predictor variables (including a vector $\mathbf{1}$ of ones for the intercept) and $\mathbf{D}$ is the $((P+1) \times D)$-dimensional matrix of the regression coefficients and the vectors $\mathbf{d}_j$ are compositions.
The ilr-transformed variables are again marked by asterisk: $\text{ilr}(\mathbf{y}_i) = \mathbf{y}_i^*$) and $\text{ilr}(\mathbf{x}_i) = \mathbf{x}_i^*$, although it is not necessary that the same ilr transformation is used for both, $\mathbf{y}_i$ and $\mathbf{x}_i$. The model in terms of coordinates is given by:

$$\mathbf{y}_i^* = \mathbf{d}_0^* + \sum_{j=1}^{P-1} (x_{ij}^* \cdot \mathbf{d}_j^*) + \mathbf{e}_i^*, \ i = 1, \ldots, n$$

It is important to note here that the coefficients $\mathbf{d}_j^*$ cannot be directly associated with $\mathbf{d}_j$. Nevertheless, it is correct that $\mathbf{y}_i^* = \text{ilr}(\mathbf{y}_i)$ and in the same way for $\mathbf{x}$, respectively. Defining the relationship between $\mathbf{d}_j$ and $\mathbf{d}_j^*$ exceeds the range of this work and needs to be thought about intensively. Anyway, the least squares estimations can be done, since we consider $\mathbf{D}^*$ just as the coefficients in the model in terms of coordinates, and therefore, the dimensions coincide. It is a $(P \times (D-1))$-matrix.
To estimate the parameters the sum of squares $\text{SSE} = \sum_{i=1}^n \|\mathbf{e}_i^*\|^2$ will be minimized again. The least-squares estimator of the coefficients $\hat{\mathbf{D}}^*$ can be calculated as

$$\hat{\mathbf{D}}^* = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{Y}^*$$

The estimated covariance matrix $\mathbf{\Sigma}_{\mathbf{e}^*}$ is

$$\mathbf{\Sigma}_{\mathbf{e}^*} = \frac{1}{n-P-1}(\mathbf{Y}^* - \mathbf{X}^*\hat{\mathbf{D}}^*)'(\mathbf{Y}^* - \mathbf{X}^*\hat{\mathbf{D}}^*)$$

Coefficient of determination

The coefficient of determination is explained similar to that mentioned in the section before (cf. section 3.1.1). But then the value of R^2 is derived by means of $\hat{y}^*$. This is to say coordinates have to be used due to our assumption, that $\mathbf{y}_i$ as well as $\mathbf{x}_i$ are compositions.

Inference statistics

For inference statistics we consider balances which we introduced in the section before (cf. equation (3.4)). The coefficients $d_{1k}, k = 1, \ldots, D$ explain the influence of the relative information corresponding to the part x_1 on the k-th coordinate of the response variable. For this reason, the response composition $\mathbf{y} = (y_1, \ldots, y_D)'$ also needs to be transformed with the special choice of orthonormal coordinates using equation (3.4), so that d_{11} evaluates the strength of the described infuence on y_1 (or, more precisely, to its relative contribution in the whole composition $\mathbf{y}$ as expressed using the coordinate y_1. Thus, to evaluate all possible combinations of the response and explanatory compositional variables we need to construct $D \cdot P$ regression models (Filzmoser et al., 2012).
For the proper test statistics see section 3.1.1. Obviously the parameters have to be adapted, as the coefficients are called $\mathbf{D}^*$. Furthermore, both, $\mathbf{X}^*$ and $\mathbf{Y}^*$ are ilr-transformed compositions and therefore coordinates.

Chapter 4

Robust multiple and multivariate linear regression

By applying (multivariate) regression we want to explain relations between a multivariate response and one or more explanatory variables. In that issue one is often confronted with the problem of outliers. Outliers can bias the results and will lead to inappropriate regression parameters. For example, the very common least-squares method is very sensitive to outliers. That is the reason why robust multiple and multivariate regression have been introduced. Robust methods can deal with a certain fraction of outlying observations in the data (concept of breakdown point) and therefore they will lead to better results.
This chapter is mainly based on the paper by Hubert et al. (2008). General methods for robust regression will be explained here.

4.1 Basic concepts

When using now least-squares regression to estimate the parameters of a certain model, outliers can have a big influence on the results. There are two types of outliers that could occur. *Leverage points* are observations $(\mathbf{x}_i, y_i)$ whose $\mathbf{x}_i$ are outlying. That means that $\mathbf{x}_i$ deviates from the majority in x-space (Hubert et al., 2008). If such a point follows the linear trend of the majority, we call it a good leverage point, otherwise, if it does not follow a linear trend, it is called a bad leverage point. *Regression outliers* are those observations, that deviate in the y-space. Hence, if we do not look at the data carefully, there are mainly two things that may happen when applying statistical methods. Firstly, the multivariate estimates differ from the "cor-

rect" answer, which is defined as the estimate that would have been obtained without the outliers. And secondly, the resulting fitted model does not allow to detect the outliers by means of their residuals, Mahalanobis distances or "leave-one-out" diagnostics (Hubert et al., 2008). Therefore, we have to find all outliers that matter, which is equivalent to finding a robust fit.
In order to quantify the robustness of a method one usually uses the concepts of breakdown point and influence function. The breakdown point of an estimator is the smallest proportion of arbitrary observations (outliers, other deviating points) that an estimator can handle before giving a non-sense result. The influence function gives an idea of how an estimator behaves under small amounts of data contamination. For a robust estimator, the influence of an infinitesimal contamination is bounded, while it is unbounded for a common non-robust estimator (Filzmoser and Hron, 2011).
A very important property for our goal to find robust estimates for compositional data is that the estimates should be affine equivariant. That means that they should behave properly under affine transformations of the data: for the data $\mathbf{X}$, the location estimator $\hat{\boldsymbol{\mu}}$ and the covariance estimator $\hat{\boldsymbol{\Sigma}}$ as well as for any nonsingular $(D-1)\times(D-1)$ matrix $\mathbf{A}$ and for any vector $\mathbf{b}\in\mathbb{R}^{D-1}$ the following two conditions should be fulfilled:

$$\begin{aligned}\hat{\boldsymbol{\mu}}(\mathbf{XA}+\mathbf{1}_n\mathbf{b}') &= \hat{\boldsymbol{\mu}}(\mathbf{X})\mathbf{A}+\mathbf{b}\\ \hat{\boldsymbol{\Sigma}}(\mathbf{XA}+\mathbf{1}_n\mathbf{b}') &= \mathbf{A}'\hat{\boldsymbol{\Sigma}}(\mathbf{X})\mathbf{A}\end{aligned}$$

where $\mathbf{1}_n$ is a vector with n elements with ones.

4.2 Robust multiple linear regression

First, let us define the multiple linear regression model, where predictor variables $\mathbf{x}_i$ and a response y_i are measured: $(\mathbf{x}_i, y_i), i=1,\ldots,n$

$$y_i=\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}+e_i$$

where the errors e_i are assumed to be normally distributed with zero mean and constant variance σ^2. We call $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)$ the slope and β_0 the intercept. The rather strict model assumptions may be violated for the robust regression case.
A very convenient method to estimate $\boldsymbol{\beta}$ in a robust way is to use the least-trimmed squares firstly introduced by Rousseeuw (1984). Contrary to the ordinary least-squares estimation, the sum consists just of $h<n$ summands. Let us define this idea formally:
The residuals are defined as

$$r_i=y_i-\hat{y}_i \quad \text{with} \quad \hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\ldots+\hat{\beta}_p x_{ip}.$$

They describe the error between the response and the fitted model. Ordinary least-squares regression is then defined as:

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} r_i^2(\boldsymbol{\beta})$$

The problem with this estimator is the big chance of being biased due to outliers in x- or y-space. Therefore, the LTS (least-trimmed squares) estimator has been introduced and is given by

$$\hat{\boldsymbol{\beta}}_{\text{LTS}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{h} r_{(i)}^2(\boldsymbol{\beta})$$

with $r_{(1)}^2 \leq r_{(2)}^2 \leq \ldots \leq r_{(h)}^2 \leq \ldots \leq r_{(n)}^2$, which are the ordered residual squares. h has to be fixed beforehand and is chosen the following way: $\lfloor \frac{n}{2} \rfloor \leq h \leq n$. Hence, the LTS estimate is the least-squares fit to these h points. This method has a breakdown point up to 50%, which is the maximum possible and reasonable value. Refering to Hubert et al. (2008), a usual choice of h is $h \approx 0.75n$ that yields a breakdown point about 25%, since a maximal breakdown point does not mean maximal efficiency.
Hence, the standard deviation of the errors can be estimated by

$$\hat{\sigma} = c_{h,n} \sqrt{\frac{1}{h} \sum_{i=1}^{h} r_{(i)}^2}$$

where r_i are the residuals from the LTS fit and $c_{h,n}$ makes $\hat{\sigma}$ consistent and unbiased at Gaussian error distributions (Pison et al., 2002).

4.3 Robust multivariate regression

4.3.1 Multivariate regression based on robust covariance estimation

Now we consider the multivariate regression model, where we have p-variate predictors as well as q-variate responses. The model is then given by

$$\begin{aligned} \mathbf{y}_i &= \boldsymbol{\beta}_0 + \mathbf{x}_i' \mathbf{B} + \boldsymbol{e}_i \\ \mathbf{y}_i &= (y_{i1}, \ldots, y_{iq})', \quad \mathbf{x}_i = (x_{i1}, \ldots, x_{ip})' \text{ and } \boldsymbol{e}_i = (e_{i1}, \ldots, e_{1q})'. \end{aligned}$$

$\mathbf{B}$ is the $p \times q$ slope matrix and $\boldsymbol{\beta}_0$ is the q-dimensional intercept vector. The errors are independently and identically distributed with zero mean and

$\mathrm{Cov}(\boldsymbol{e}) = \boldsymbol{\Sigma}_e$ is a positive definite matrix of size $q \times q$.
An important task for multivariate statistics is the estimation of location and covariance. The usual way to estimate is to compute the arithmetic mean and the sample covarince matrix for the data. Both estimators are very sensitive to outliers, they even have a breakdown point of 0%. That means, that just one outlier or deviating point could lead to an arbitrarly large value for the mean, for example. One good proposal for a robust estimator is the MCD (Minimum Covariance Determinant) estimator, which will be described below.
First, let us write down the well known least-squares solution for our model. The empirical mean and covariance matrix of the joint $(\mathbf{x}, \mathbf{y})$ variables are: $\hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\boldsymbol{\mu}}_x \\ \hat{\boldsymbol{\mu}}_y \end{pmatrix}$ and $\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{xx} & \hat{\boldsymbol{\Sigma}}_{xy} \\ \hat{\boldsymbol{\Sigma}}_{yx} & \hat{\boldsymbol{\Sigma}}_{yy} \end{pmatrix}$. Hence we get:

$$\begin{aligned} \hat{\boldsymbol{B}} &= \hat{\boldsymbol{\Sigma}}_{xx}^{-1} \hat{\boldsymbol{\Sigma}}_{xy} \\ \hat{\boldsymbol{\beta}}_0 &= \hat{\boldsymbol{\mu}}_y - \hat{\mathbf{B}}' \hat{\boldsymbol{\mu}}_x \\ \hat{\boldsymbol{\Sigma}}_e &= \hat{\boldsymbol{\Sigma}}_{yy} - \hat{\mathbf{B}}' \hat{\boldsymbol{\Sigma}}_{xx} \hat{\mathbf{B}} \end{aligned}$$

First we will now introduce a robust estimator for location and scatter.

Minimum covariance determinant regression

The idea of the method is to look for the h observations whose classical covariance matrix has the lowest possible determinant. The MCD location estimate is then defined as the average of these h points whereas the MCD estimate of scatter is a multiple of their covariance matrix (Rousseeuw, 1985). The property of affine equivariance is fulfilled.
The MCD estimator yields robust estimates of location and covariance with a maximum breakdown point of 50%. However, it has been noted that the MCD can have a low efficiency. Hence, it is sometimes better to accept a lower breakdown point, and therefore reaching a higher statistical efficiency. Typically one chooses a breakdown point of 25%, which is still sufficiently robust for most applications and is more efficient (Hubert et al., 2008). Moreover, MCD estimators do have a bounded influence function and they are asymptotically normal.
Consider a data set $\mathbf{Z}_n = \{\mathbf{z}_i : i = 1, \ldots, n\} \in \mathbb{R}^{p+q}$. In that data set the response $\mathbf{y}_i$ and the covariates $\mathbf{x}_i$ are merged. Then the MCD method looks for the subset $\{\mathbf{z}_{i_1}, \ldots, \mathbf{z}_{i_h}\}$ of size h whose covariance matrix has the smallest determinant , where $\lceil \frac{n}{2} \rceil \leq h \leq n$ and $\gamma = (n-h)/n, 0 \leq \gamma \leq 0.5$.

The estimate for the center and the covariance are then defined as:

$$\begin{aligned} \mathbf{t}_n &= \frac{1}{h}\sum_{j=1}^{h}\mathbf{z}_{i_j} \\ \mathbf{C}_n &= c_n c_\gamma \frac{1}{h}\sum_{j=1}^{h}(\mathbf{z}_{i_j}-\mathbf{t}_n)(\mathbf{z}_{i_j}-\mathbf{t}_n)', \end{aligned}$$

where c_γ is a consistency factor and c_n is a small sample correction factor (see Pison et al., 2002). The breakdown point of the MCD estimator is approximately γ.
Hence, observations that lie far from the center only have a small effect on the MCD estimate and therefore both, leverage points and regression outliers only have a small impact on the estimates. As already stated before, the efficieny is rather low when gaining a breakdown point of about 50%. To improve the situation, it is a good choice to apply a reweighting algorithm (reweighting multivariate regression). For further details, see Rousseeuw et al. (2004).

4.3.2 Multivariate least-trimmed squares regression

This section is mainly based on one paper called "The multivariate least-trimmed squares estimator" by Agulló et al. (2008).
The multivariate least-trimmed squares regression is an extension of the multiple LTS-regression mentioned above. The idea of this approach is similar to that of the MCD estimator. A subset of h observations should be found, where these h observations have a minimal determinant of the covariance matrix of its residuals from a LS-fit. Hence, the correlation between the different components of the error term is taken into account.
The approach is equivalent to the selection of the value of $\mathbf{B}$ which minimizes the determinant of the robust MCD scatter matrix of the residuals.
For the sake of simplicity, we will use another notation in the model of multivariate regression:

$$\mathbf{y}_i = \mathbf{x}_i'\mathbf{B} + \boldsymbol{e}_i$$

where $i = 1, \ldots, n$, and $\mathbf{B}$ already includes the intercept term $\mathbf{b}_0$. Moreover, the first element of x_i is 1 for the intercept extimation. The ordinary least-squares estimator is then given by $\hat{\boldsymbol{B}}_{\mathrm{LS}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$. If we consider a data set $\mathbf{Z}_n = \{(\mathbf{x}_i, \mathbf{y}_i) : i = 1, \ldots, n\}$, we are able to define the residuals the following way:

$$\mathbf{r}_i(\mathbf{B}) = \mathbf{y}_i - \mathbf{x}_i'\mathbf{B} \tag{4.1}$$

Furthermore, the set $\mathcal{H}$ is given by $\mathcal{H} = \{H \in \{1, \ldots, n\} \mid \#H = h\}$. Then $\hat{\boldsymbol{B}}_{\mathrm{LS}}(H)$ is the LS- estimator based solely on the observations $(\mathbf{x}_j, \mathbf{y}_j) : j \in H$. The covariance of the residuals depending on H and $\mathbf{B}$ is then given by:

$$\mathrm{Cov}(H, \mathbf{B}) := \frac{1}{h}(\mathbf{r}_j(\mathbf{B}) - \overline{\mathbf{r}}_H(\mathbf{B}))(\mathbf{r}_j(\mathbf{B}) - \overline{\mathbf{r}}_H(\mathbf{B}))'$$

where $\overline{\mathbf{r}}_H(\mathbf{B}) = \frac{1}{h}\sum_{j \in H} \mathbf{r}_j(\mathbf{B})$. Finally, the multivariate least-trimmed squares estimator is derived as the least-squares estimator but using just those h elements of the dataset that results in the minimum determinant. Formally it can be written as:

$$\hat{\boldsymbol{B}}_{\mathrm{MLTS}}(\mathbf{x}_i, \mathbf{y}_i) = \hat{\boldsymbol{B}}_{\mathrm{LS}}(\hat{H}) \quad \text{where} \quad \hat{H} \in \operatorname*{argmin}_{H \in \mathcal{H}} \det \hat{\boldsymbol{\Sigma}}_{\mathrm{LS}}(H)$$

with $\hat{\boldsymbol{\Sigma}}_{\mathrm{LS}}(H) = \mathrm{Cov}(H, \hat{\mathbf{B}}_{\mathrm{LS}}(H))$ for any $H \in \mathcal{H}$.
An estimator for the covariance of the errors is then given by

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{LS}}(\mathbf{Z}_n) = c_\gamma \hat{\boldsymbol{\Sigma}}_{\mathrm{LS}}(\hat{H})$$

with a consistency factor c_γ.
If there is more than one solution in the minimization problem, one of them is selected arbitrarly. If h=n we are in the situation of the classical LS-estimator.
To deepen the knowledge about this estimator, see Agulló et al. (2008).

Finally we can conclude that the MCD-based procedure focuses on random designs and the MLTS approach is more general, since it is based only on the covariance matrix of the residuals, instead of on the covariance matrix of the joint distribution. Therefore, we will prefer the MLTS estimator, although for robust estimation of location and scatter the MCD estimator will be used later as well.

Chapter 5

Robust linear regression with compostitional data

The development of robust statistical methods for compositional data is still in the beginning. A very important step for that task was done by introducing the ilr transformation together with the concepts of balances, since the ilr variables of a D-part composition represent coefficients of an orthonormal basis on the simplex. Nowadays, the clr transformation is only important for the construction of a compositional biplot because it makes its interpretation possible in terms of the original compositional parts (Filzmoser and Hron, 2011).
The most frequently used methods for multivariate robust regression analysis are outlier detection (see Filzmoser and Hron 2008), principal component analysis (see Filzmoser et al. 2009a), factor analysis (see Filzmoser et al. 2009c), discriminant analysis, or the estimation of missing values. Here we focus on robust regression for compositional data.

5.1 Robust regression with compositional response

Now we use the robust methods described in chapter 4 when our data includes compositions. In this section we consider the case that just the response $\mathbf{y}_i$ is compositional and therefore multivariate. As it was shown in section 3.1, an ilr transformation will be applied on the dependent variable $\mathbf{y}_i$ and the result is the same model that has been mentioned there:

$$\mathbf{y}_i^* = \mathbf{b}_0^* + \sum_{j=1}^{p}(x_{ij} \cdot \mathbf{b}_j^*) + \mathbf{e}_i^* \ , \ i = 1, 2, \ldots, n \tag{5.1}$$

The difference is, that the regression parameters will be estimated in a robust way (cf. chapter 4). Therefore, it is possible to consider a MCD regression by estimating the location and scatter of the joint $(\mathbf{x}, \mathbf{y})$ and just using the h observations that give the minimum determinant of the covariance matrix (cf. section 4.3.1). Otherwise we can apply a MLTS regression (cf. section 4.3.2) on the model given in equation (5.1). The procedure is the same as it has been described in section 4.3.2. The model just differs a little due to the ilr-transformed variables $\mathbf{y}_i^*, \mathbf{b}_j^*$ and $\mathbf{e}_i^*$.
Hence, we use robust estimators that yield more reliable results in the regression analysis of real compositional data.
To obtain inference statistics or rather conclusions about certain hypotheses, we assume the normality of the random errors $\mathbf{e}_i, i = 1, \ldots, n$. The associated test statistics are explained and specified in section 3.1.1. Instead of the classical least-squares estimator for $\hat{\mathbf{B}}^*$ its robust counterpart is chosen to perform the executions. In summary, we obtain new results for our regression parameters and these coefficients can be interpreted according to the choice of balances. It has to be noted that for the computations of the F-statistics and the associated value R^2 we assume that a modification of the formulas would lead to results which are more accurate. The problem follows from the usage of $\mathbf{X}$ and $\mathbf{Y}$ in the calculations because these variables still contain leverage points, although $\mathbf{B}$ has been estimated in a robust way. An idea of adaption is to multiply these variables with a vector including ones on those positions which are in the best subset (which is evaluated for the robust estimation of center and covariance), and zeros on those postions that are not. Nevertheless, even in literature researchers are discordant about the correctness of this adaption.

5.2 Robust regression with compositional explanatory variables

Also here we just refer to section 3.1 and chapter 4. Model (3.6) is taken into account and robust multivariate methods of regression analysis are applied.
For the multivariate least-trimmed squares estimations we will give some more details:
The very common least-squares estimator for the coefficients $\mathbf{c}_j, j = 0, \ldots, D-1$ is given by $\hat{\mathbf{C}} = [(\mathbf{X}^*)'\mathbf{X}^*]^{-1}(\mathbf{X}^*)'\mathbf{Y}$. The data set $\mathbf{Z}_n = \{(\mathbf{x}_i^*, \mathbf{y}_i) : i = 1, \ldots, n\}$ is similar to that in section 4.3.2 and also the residuals $\mathbf{r}_i$ are calculated the same way as in equation (4.1), except for the coordinates. Again we consider the set $\mathcal{H} = \{H \in \{1, \ldots, n\} \mid \#H = h\}$. Then $\hat{\boldsymbol{B}}_{\mathrm{LS}}(H)$ is the LS-

estimator based solely on the observations $(\mathbf{x}_j^*, \mathbf{y}_j) : j \in H$. The multivariate least-trimmed squares estimator is then given by

$$\hat{\boldsymbol{C}}_{\mathrm{MLTS}}(\mathbf{x}_i^*, \mathbf{y}_i) = \hat{\boldsymbol{C}}_{\mathrm{LS}}(\hat{H}) \quad \text{where} \quad \hat{H} \in \operatorname*{argmin}_{H \in \mathcal{H}} \det \hat{\boldsymbol{\Sigma}}_{\mathrm{LS}}(H)$$

with $\hat{\boldsymbol{\Sigma}}_{\mathrm{LS}}(H) = \mathrm{Cov}(H, \hat{\mathbf{C}}_{\mathrm{LS}}(H))$ for any $H \in \mathcal{H}$ (cf. section 4.3.2). All the formulas, above all the corresponding F-test, that have not been mentioned in this part again can be looked up in the referred sections and have to be adapted respectively.

5.3 Robust regression with compositional explanatory variables and response

Finally, robust regression should be applied on a model with both compositional response and compositional explanatory variables. We consider the model from section 3.3. Similar to the two other cases before we just estimate the coefficients in a robust way contrary to the way we did it in chapter 3. Since we are confronted with a multivariate regression model again, the two F-statistics (3.2) and (3.3) are the proper values to make decisions on the estimated coefficients. In particular, the statistics F_{11} can be used to test whether the relative information (the ratios) on a chosen compositional explanatory variable has a significant influence on the corresponding counterpart concerning a certain part of the response composition, if both compositions were expressed in coordinates (Filzmoser and Hron, 2012).

Chapter 6

Examples with R

The aim of this section is to apply all the techniques described in the theoretical part above on compositional data. Therefore, classical and robust linear regression will be performed as well as the resulting inference statistics. Outputs will be interpreted and analysed. For the execution we used a special data set which we will describe in detail in the next section.

6.1 The data set

For the analysis of compositional data we used a data set that has been collected for a project called GEMAS (The EuroGeoSurveys geochemical mapping of agricultural and grazing land soils project) in order to help industries, dealing with natural ressources, to get to know the bioavailablity of metals and other chemical elements in soil. Moreover, they want to obtain information about the long-term fate of metals and other chemical elements added to soil. Another important goal is to recognize toxic concentrations in soil that may influence plant and animal productivity as well as human health. Therefore, samples were taken from places all over Europe to measure the soil quality at the European scale. The GEMAS project delivers good quality and comparable exposure data of metals in agricultural and grazing land soil. A Geochemical Atlas of Europe has been produced (for detailed maps and informations see http://www.gtk.fi/publ/foregsatlas) which demonstrates that low-sample density geochemical mapping at the European scale is possible for a variety of sample materials, including surface water, stream and floodplain sediments and soil (Reimann et al., 2009). Another important part of the project is to establish an archive of samples, that would be invaluable in case of catastrophic events. For more information about the topic, Reimann et al. (2009) is interesting to read.

6.1.1 Description of the variables

The data set contains a set of 142 variables, basically elements which are components of the soil, with all in all 2108 observations. The variable names are:

```
  [1] "ID"          "COUNTRY"     "C_ID"
  [4] "TYPE"        "TYPE2"       "XCOO"
  [7] "YCOO"        "XLAEA"       "YLAEA"
 [10] "ALT"         "CIA"         "sand"
 [13] "silt"        "clay"        "sand_norm"
 [16] "silt_norm"   "clay_norm"   "soiltype"
 [19] "soilclass"   "climate"     "MeanTemp"
 [22] "AnnPrec"     "PM"          "CEC"
 [25] "pH_CaCl2"    "TOC"         "Ag"
 [28] "Al"          "As"          "Au"
 [31] "B"           "Ba"          "Be"
 [34] "Bi"          "Ca"          "Cd"
 [37] "Ce"          "Co"          "Cr"
 [40] "Cs"          "Cu"          "Fe"
 [43] "Ga"          "Ge"          "Hf"
 [46] "Hg"          "In"          "K"
 [49] "La"          "Li"          "Mg"
 [52] "Mn"          "Mo"          "Na"
 [55] "Nb"          "Ni"          "P"
 [58] "Pb"          "Pd"          "Pt"
 [61] "Rb"          "Re"          "S"
 [64] "Sb"          "Sc"          "Se"
 [67] "Sn"          "Sr"          "Ta"
 [70] "Te"          "Th"          "Ti"
 [73] "Tl"          "U"           "V"
 [76] "W"           "Y"           "Zn"
 [79] "Zr"          "C_tot"       "S_tot"
 [82] "SiO2"        "Si_XRF"      "TiO2"
 [85] "Ti_XRF"      "Al2O3"       "Al_XRF"
 [88] "Fe2O3"       "Fe_XRF"      "MnO"
 [91] "Mn_XRF"      "MgO"         "Mg_XRF"
 [94] "CaO"         "Ca_XRF"      "Na2O"
 [97] "Na_XRF"      "K2O"         "K_XRF"
[100] "P2O5"        "P_XRF"       "SO3"
[103] "S_XRF"       "Cl_XRF"      "F_XRF"
[106] "LOI"         "As_XRF"      "Ba_XRF"
```

```
[109] "Bi_XRF"                "Ce_XRF"                "Co_XRF"
[112] "Cr_XRF"                "Cs_XRF"                "Cu_XRF"
[115] "Ga_XRF"                "Hf_XRF"                "La_XRF"
[118] "Mo_XRF"                "Nb_XRF"                "Ni_XRF"
[121] "Pb_XRF"                "Rb_XRF"                "Sb_XRF"
[124] "Sc_XRF"                "Sn_XRF"                "Sr_XRF"
[127] "Ta_XRF"                "Th_XRF"                "U_XRF"
[130] "V_XRF"                 "W_XRF"                 "Y_XRF"
[133] "Zn_XRF"                "Zr_XRF"                "X208_207"
[136] "X207_208"              "X208_206"              "X206_208"
[139] "X206_207"              "X207_206"              "SUSCEPTIBILITY"
[142] "SUSCEPTIBILITY.FE2O3"
```

As a fist setp we deleted those observations that included missing values (NAs) which resultet in a working set with finally 2061 observations. With the aid of Dr. Clemens Reimann, project coordinator of the GEMAS project, we selected some groups of variables to observe during the regression analysis. At first some non-compositional variables are chosen:

$\mathbf{y}_1$ ... MeanTemp (mean temperature)
$\mathbf{y}_2$ ... log(AnnPrec) (annual precipitation)
$\mathbf{y}_3$ ... ALT (altitude)
$\mathbf{y}_4$ ... log(SUSECPTIBILITY) (magnetic characteristics).

Due to the skewness of the variables `AnnPrec` and `SUSCEPTIBILITY` a log-transformation has been applied on them. $\mathbf{Y} = (y_1, y_2)$ characterizes in some sense the climate.
As next step we defined some sets of compositions. First we chose the three parts `sand_norm, silt_norm` and `clay_norm`, which already have been normed to sum up to 100%. In the following this data set is stored in $\mathbf{X}_1$. Moreover, $\mathbf{X}_2$ represents the `XRF`(X-ray fluorescence method) variables and another one, namely $\mathbf{X}_3$, the oxides. Finally all the raw elements form set $\mathbf{X}_4$ of compositions. Variables that haven't had good data quality have been removed so that the regression performance is improved.
The four sets are given by:

```
X_1=(sand_norm silt_norm clay_norm)
X_2=(Si_XRF Ti_XRF Al_XRF Fe_XRF Mn_XRF Mg_XRF Ca_XRF Na_XRF K_XRF
P_XRF Ba_XRF Cr_XRF Nb_XRF Pb_XRF Rb_XRF Sr_XRF V_XRF Y_XRF Zn_XRF
Zr_XRF)
```

```
X_3=(SiO2 TiO2 Al2O3 Fe2O3 MnO MgO CaO Na2O K2O P2O5)
X_4=(Ag Al As B Ba Be Bi Cd Ce Co Cr Cs Cu Fe Ga K La Li Mg Mn Mo
Na Nb Ni P Pb Rb S Sb Sc Se Sn Sr Th Ti Tl U V W Y Zn Zr)
```

6.2 Linear regression analysis and inference statistics with compositions

6.2.1 Classical and robust linear regression with compositional explanatory variables

We consider the following regression model. We have a univariate non-compositional response and (obviously) multivariate compositional explanatory variables (cf. model 3.1).
Functions for these analyses have already been implemented in *R*. We used the function `lmCoDaX`, but we adapted it for our own purpose to apply classical and robust regression with compositional explanatory variables. Within the function, the variables get ilr-transformed by means of a sequential binary partition (cf. Egozcue and Pawlowsky-Glahn, 2005).
In this context, we used the dependent variable $\mathbf{y}_4$ and as the independent explanatory variables the set $\mathbf{X}_3$. Consequently, we want to gain information about the relationship of the magnetic characteristics in the soil and the parts of oxides. Since $\mathbf{y}_4$ is log-transformed and the variable `SUSCEPTIBILITY` had one negative value, it was necessary to exclude one more data point so that the total size of observations is 2060 in this model.
For both, the classical and the robust estimations, the coefficients are derived in the same way (according to model 3.5). Hence, it has to be noted that the output shows the estimates for c_0, which is the same for all models and furthermore it consists of $c_1^{(1)}$, $c_2^{(2)}$, etc. These coefficients can not be used to describe one model, but D different models where all the relative information about the first part of $\mathbf{X}$ is explained. For a better understanding, imagine $c_2^{(2)}$ is the first coefficient of the model, where $\mathbf{x}_2$ is the first explanatory variable. In Section 3.2.1 it is written as $x_1^{(2)}$. To keep the output short, below we stated just the first coefficients. See Appendix A for the function details
.
The *R*-Input is given by

```
> ### univarite Response, multivariate compositional explanatory variables
>
```

```
> ### Model 1: y4~ "O"-elements
> ### classical and robust estimations
>
> ### classical estimation with lm(.) and robust estimation with lts(.)
> ### use function lmCoDaX(.) from package robCompositions
>
> model1_class<-lmCoDaX_adapted(y=y4[-446],X=X3[-446,],method="classical")
> model1_rob<-lmCoDaX_adapted(y4[-446],X3[-446,],method="robust")
>
> ### the model also prints the estimations for not-transformed compositions
```

In the case of a classical estimation we obtained table 6.1.

```
Call:
lm(formula = y ~ ., data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2139 -0.6520 -0.0598  0.6001  3.4388

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.80470    0.40259   6.967 4.36e-12 ***
X.SiO2      -1.00931    0.05457 -18.497  < 2e-16 ***
X.TiO2       0.11860    0.08454   1.403  0.16081
X.Al2O3     -0.16430    0.13403  -1.226  0.22041
X.Fe2O3      1.01764    0.10732   9.482  < 2e-16 ***
X.MnO        0.33458    0.04883   6.852 9.58e-12 ***
X.MgO       -0.42940    0.05200  -8.257 2.64e-16 ***
X.CaO       -0.06112    0.02201  -2.777  0.00554 **
X.Na2O       0.04291    0.03052   1.406  0.15986
X.K2O        0.02705    0.07932   0.341  0.73313
X.P2O5       0.12336    0.04629   2.665  0.00776 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8999 on 2049 degrees of freedom
Multiple R-squared:  0.3581,        Adjusted R-squared:  0.3553
F-statistic: 127.1 on 9 and 2050 DF,  p-value: < 2.2e-16
```

Table 6.1: Classical regression with ilr-transformed explanatory variables

Unfortunately, R^2 is rather low with 35.81%. So the susceptibility probably also depends on other things apart from our variables. Anyway, we got some interesting results. There are variables that are highly significant, which are `SiO2, Fe2O3, MgO, MnO`, but also `CaO` and `P2O5` show a big influence on the response. If we have a look on the coefficients we can see that Iron(III)-oxide for example has a very positive effect on the response, whereas Silicon dioxide (quartz) has a siginificant negative effect and that was something that we expected. If the part of Silicon dioxide increases, then we are confronted with a so called thinning effect because then the sample is not or less magnetic and hence, the susceptibility can not be measured any more.
By means of the F-statistic we can say that the variables describe the response somehow, even if R^2 is not that high. However, the p-value is small, therefore we have to refuse the hypothesis that all coefficients are equal to 0. Since it is also possible to show the results for common least-squares regression for the original data (and not for the coordinates), we will also have a look on that output to compare the situation (see table 6.2).

```
Call:
lm(formula = y ~ ., data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6565 -0.6792 -0.0679  0.6119  3.5850

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.526380   0.340041 -13.311  < 2e-16 ***
X.SiO2       0.016125   0.003540   4.555 5.54e-06 ***
X.TiO2       0.320401   0.109168   2.935  0.00337 **
X.Al2O3      0.045555   0.013834   3.293  0.00101 **
X.Fe2O3      0.214803   0.026135   8.219 3.60e-16 ***
X.MnO        0.733538   0.287986   2.547  0.01093 *
X.MgO       -0.021428   0.020322  -1.054  0.29182
X.CaO        0.040037   0.006867   5.830 6.41e-09 ***
X.Na2O      -0.014871   0.028317  -0.525  0.59953
X.K2O       -0.017729   0.037159  -0.477  0.63334
X.P2O5       1.991945   0.201931   9.864  < 2e-16 ***
---
Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9023 on 2049 degrees of freedom
Multiple R-squared:  0.3549, Adjusted R-squared:  0.3518
F-statistic: 112.7 on 10 and 2049 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = y ~ ., data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2139 -0.6520 -0.0598  0.6001  3.4388

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.80470    0.40259   6.967 4.36e-12 ***
X.SiO2      -1.00931    0.05457 -18.497  < 2e-16 ***
X.TiO2       0.11860    0.08454   1.403  0.16081
X.Al2O3     -0.16430    0.13403  -1.226  0.22041
X.Fe2O3      1.01764    0.10732   9.482  < 2e-16 ***
X.MnO        0.33458    0.04883   6.852 9.58e-12 ***
X.MgO       -0.42940    0.05200  -8.257 2.64e-16 ***
X.CaO       -0.06112    0.02201  -2.777  0.00554 **
X.Na2O       0.04291    0.03052   1.406  0.15986
X.K2O        0.02705    0.07932   0.341  0.73313
X.P2O5       0.12336    0.04629   2.665  0.00776 **
---
Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8999 on 2049 degrees of freedom
Multiple R-squared:  0.3581, Adjusted R-squared:  0.3553
F-statistic: 127.1 on 9 and 2050 DF,  p-value: < 2.2e-16
```

Table 6.2: Comparison of common least-squares regression without transformation of the compositions (left) and regression with ilr-transformed explanatory variables (right)

An interesting result is that R^2 is quite the same in the model with the original variables and that with ilr-transformed variables, so the quality fit seems to be similar. The big difference, however, is the interpretation of the inference statistics. The results based on the original data are misleading because the data are not represented in the usual Euclidean space. Moreover, table 6.2 shows that significant variables are not exactly the same in the case

of original variables and coordinates. MgO, for example, is highly significant in the model using coordinates, but it is not describing the response significantly when using original data (compositions). That is the reason why one should be careful with compositional data and their interpretation.
In table 6.3 the output of the robust estimation is shown.

```
Call:
ltsReg.formula(formula = y ~ ., data = d)

Residuals (from reweighted LS):
    Min      1Q  Median      3Q     Max
-1.9887 -0.5421  0.0000  0.5892  1.9770

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
Intercept  2.29681    0.37113   6.189 7.36e-10 ***
X.SiO2    -0.78620    0.05079 -15.479  < 2e-16 ***
X.TiO2     0.05779    0.07887   0.733  0.46380
X.Al2O3   -0.37420    0.12386  -3.021  0.00255 **
X.Fe2O3    0.90863    0.09888   9.189  < 2e-16 ***
X.MnO      0.44748    0.04472  10.006  < 2e-16 ***
X.MgO     -0.38515    0.04831  -7.973 2.60e-15 ***
X.CaO     -0.02807    0.02003  -1.401  0.16126
X.Na2O     0.04043    0.02830   1.428  0.15331
X.K2O      0.10619    0.07346   1.446  0.14847
X.P2O5    -0.01931    0.04303  -0.449  0.65363
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7985 on 1984 degrees of freedom
Multiple R-Squared: 0.3363,        Adjusted R-squared: 0.3332
F-statistic: 111.1 on 9 and 1973 DF,  p-value: < 2.2e-16
```

Table 6.3: Robust regression with ilr-transformed explanatory variables

Also here R^2, with a value of about 33.63%, is quite low, but again the F-test defines it as a reasonable and valid model. The conclusion is that a lot of inputs are probably missing for a better description of the susceptibility. On the other hand, we can not set all estimated coefficients equal to 0. In this model, surprisingly, some other chemical substances than in the classical

regression model have a significant effect when their part in the soil is bigger. The most sigificant substances are `SiO2, Fe2O3, MnO, MgO` and we can also add `Al2O3`, where again `Fe2O3` and also `MnO` influence the response by means of a positive coefficient. The other significant elements do have negative effects.
Finally, we made a regression diagnostic plot. In that matter the robust distance computed by a MCD estimator is plotted against the standardized LTS residuals. The robust distance is defined as:

$$\mathrm{RD}(\mathbf{x}_i, \mathbf{X}) = \sqrt{(\mathbf{x}_i - \mathbf{t}_n(\mathbf{X}))' \mathbf{C}_n(\mathbf{X})^{-1} (\mathbf{x}_i - \mathbf{t}_n(\mathbf{X}))},$$

which is the robustified Mahalanobis distance and uses the robust MCD estimates described in section 4.3.1 for the mean and the covariance estimation. In figure 6.1 the regression diagnostic plot for our model is shown. The cut-off values separating outliers from regular observations are taken as ± 2.5 for the standardized residuals, and $\sqrt{\chi^2_d}$ for the robust distance, where d is the number of columns of $\mathbf{X}$.

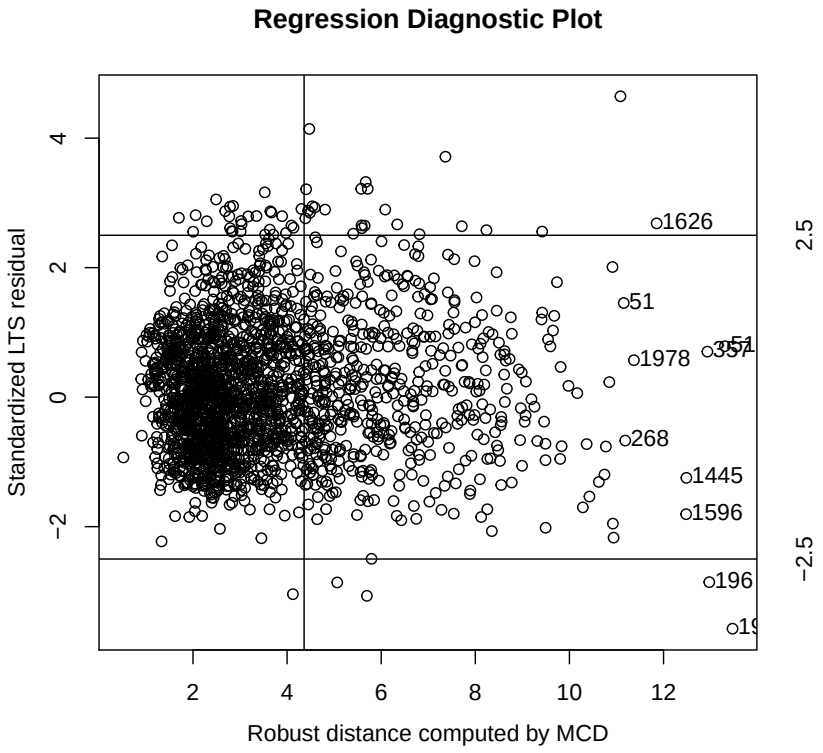

Figure 6.1: Regression diagnostic plot for model $\mathbf{y}_4 \sim \mathbf{X}_3$

The data between the horizontal cutoffs on the left side are the regular observations and on the right side there are the so called good leverage points because they follow the linear pattern, but they deviate in x-direction. In the bottom and topright part of the plot we can see some bad leverage points and to the left we see some vertical or regression-outliers.
To obtain another graph for comparison, we calculate the classical Mahalanobis distances and use it on the x-axis instead of the robust counterpart. On the y-axis the standardized least-squares residuals are plotted. The corresponding image is given in figure 6.2.
The classical Mahalanobis distance is a measure to detect outliers by computing the distance from the center of the cloud to the observations. In that case, classical estimates for the center $\boldsymbol{\mu}$ and the covariance $\boldsymbol{\Sigma}$ are used. It is given by:

$$\mathrm{MD}(\mathbf{x}_i, \mathbf{X}) = \sqrt{(\mathbf{x}_i - \hat{\boldsymbol{\mu}})' \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})}$$

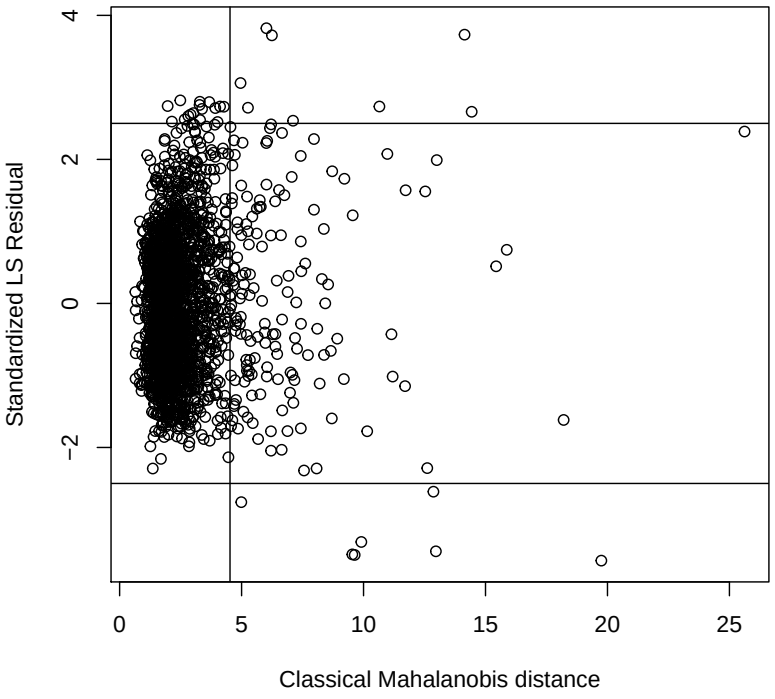

Figure 6.2: Regression diagnostic plot for model $\mathbf{y}_4 \sim \mathbf{X}_3$ using the classical Mahalanobis distance

As a consequence of figure 6.2 we agree with the assumption that robust estimation is reasonable as well as necessary to obtain proper results. In figure 6.2 we see that less points are classified as outliers, but nevertheless some points have a Mahalanobis distance that exceed the maximum robust distance by far.

Next, we consider the same regression models but we use different explanatory variables to describe the response y_4. Now the relationship between susceptibility and the parts of X_2 will be figured out.

```
>   options(width=60)
> ### use now the XRF data instead of the "O" data
> model1a_class<-lmCoDaX_adapted(y=y4[-446],X=X2[-446,],method="classical")
> model1a_rob<-lmCoDaX_adapted(y=y4[-446],X=X2[-446,],method="robust")
```

The output table has to be interpreted the same way as before. That means the estimated coefficients origine from D different regression models, where D is the number of variables that describe our model. The first table 6.4 shows the output for the classical least-squares regression and the second for the robust case (table 6.5).

```
Call:
lm(formula = y ~ ., data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5645 -0.6411 -0.0601  0.5649  3.7322

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.30189    1.06142  -2.169   0.0302 *
X.Si_XRF    -1.01942    0.06684 -15.251  < 2e-16 ***
X.Ti_XRF     0.03977    0.13571   0.293   0.7695
X.Al_XRF     0.28251    0.15103   1.871   0.0615 .
X.Fe_XRF     1.14947    0.11226  10.239  < 2e-16 ***
X.Mn_XRF     0.39193    0.04926   7.957 2.91e-15 ***
X.Mg_XRF    -0.31235    0.05782  -5.402 7.37e-08 ***
X.Ca_XRF     0.03219    0.03217   1.001   0.3172
X.Na_XRF    -0.04052    0.03634  -1.115   0.2650
X.K_XRF     -0.07637    0.14650  -0.521   0.6022
X.P_XRF      0.12764    0.05017   2.544   0.0110 *
X.Ba_XRF     0.11299    0.08958   1.261   0.2073
X.Cr_XRF    -0.07941    0.04704  -1.688   0.0916 .
X.Nb_XRF    -0.01915    0.09746  -0.197   0.8442
X.Pb_XRF     0.27468    0.05195   5.287 1.37e-07 ***
X.Rb_XRF    -0.13066    0.11336  -1.153   0.2492
X.Sr_XRF    -0.14416    0.05922  -2.435   0.0150 *
X.V_XRF      0.02654    0.10831   0.245   0.8064
X.Y_XRF     -0.90983    0.09646  -9.432  < 2e-16 ***
X.Zn_XRF    -0.17288    0.08055  -2.146   0.0320 *
X.Zr_XRF     0.46704    0.08744   5.341 1.03e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8709 on 2039 degrees of freedom
Multiple R-squared:  0.4017,        Adjusted R-squared:  0.3961
F-statistic: 72.09 on 19 and 2040 DF,  p-value: < 2.2e-16
```

Table 6.4: Classical ilr regression of $\mathbf{y}_4 \sim \mathbf{X}_2$

In this model ($\mathbf{y}_4 \sim \mathbf{X}_2$), R^2 is higher than in the one including the data set $\mathbf{X}_3$, so these variables obviously describe the susceptibility somehow better

(This fact could also result from the higher number of explanatory variables.) Nevertheless 40.17% are still less than 50%. The hypothesis that all coefficients are equal to zero can again be refused. There are a lot of highly significant variables, let us point out Iron (`Fe_XRF`) again. There is a highly positive effect on the response when the part of Iron in the soil increases whereas the part of Silicon in the sample, for example, influences with a very negative effect. Therefore, a plausible explication has already been stated in the last model. Other highly significant elements are `Mn_XRF, Mg_XRF, Pb_XRF, Y_XRF, Zr_XRF`. It is not just the significance of the variables that is interesting but also the sign of their coefficients. Therefore we should have a look if they have a positive or negative impact on the response.

```
Call:
ltsReg.formula(formula = y ~ ., data = d)

Residuals (from reweighted LS):
    Min      1Q  Median      3Q     Max
-1.8897 -0.4996  0.0000  0.5228  1.9166

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
Intercept -3.047018   0.990807  -3.075 0.002132 **
X.Si_XRF  -1.016951   0.061346 -16.577  < 2e-16 ***
X.Ti_XRF  -0.111997   0.127299  -0.880 0.379078
X.Al_XRF   0.314397   0.138016   2.278 0.022836 *
X.Fe_XRF   1.219539   0.105655  11.543  < 2e-16 ***
X.Mn_XRF   0.484628   0.044891  10.796  < 2e-16 ***
X.Mg_XRF  -0.341711   0.052834  -6.468 1.25e-10 ***
X.Ca_XRF   0.122427   0.028782   4.254 2.20e-05 ***
X.Na_XRF  -0.097024   0.033038  -2.937 0.003356 **
X.K_XRF    0.017551   0.133630   0.131 0.895519
X.P_XRF    0.107138   0.046787   2.290 0.022132 *
X.Ba_XRF   0.336312   0.082515   4.076 4.77e-05 ***
X.Cr_XRF  -0.256777   0.044729  -5.741 1.09e-08 ***
X.Nb_XRF  -0.323015   0.093030  -3.472 0.000528 ***
X.Pb_XRF   0.155300   0.047557   3.266 0.001111 **
X.Rb_XRF  -0.431899   0.109603  -3.941 8.41e-05 ***
X.Sr_XRF  -0.345202   0.054697  -6.311 3.41e-10 ***
X.V_XRF    0.001357   0.103059   0.013 0.989499
X.Y_XRF   -0.617372   0.088995  -6.937 5.41e-12 ***
X.Zn_XRF  -0.043829   0.076659  -0.572 0.567558
X.Zr_XRF   0.619511   0.079303   7.812 9.11e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7713 on 1972 degrees of freedom
Multiple R-Squared: 0.4322,        Adjusted R-squared: 0.4267
F-statistic: 78.51 on 19 and 1960 DF,  p-value: < 2.2e-16
```

Table 6.5: Robust ilr regression of $\mathbf{y}_4 \sim \mathbf{X}_2$

There is quite a big difference between the classical and the robust model. Other, and all above more variables show influence (either positive or nega-

tive) when their parts in the soil increase or decrease. Here it is interesting to point out that actually almost all variables are necessary to describe the response in a good way. Just a few elements do not have a siginificant effect, which are `Ti_XRF, K_XRF, V_XRF, Zn_XRF`. However, it is again Iron which shows a big positive influence and Silicon which has a high negative effect on the response.
Now it is necessary to point out that the regression coefficients of the model, where the compositions have not been ilr-transformed, do not suit at all to them of the correct model. Silicon, for example, has a significantly high positive effect, which is, according to Dr. Reimann, a specialist for geochemistry, not reasonable at all. Moreover, almost all variables are significant in that matter.

Also here we will look at diagnostic plots (see figure 6.3) for both, classical and robust estimations.

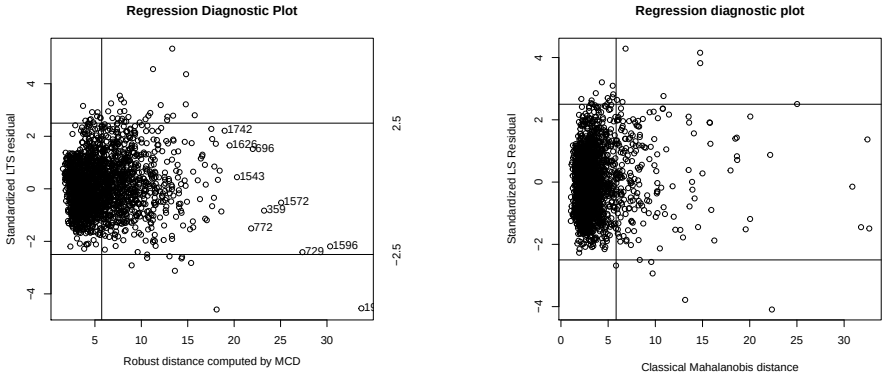

Figure 6.3: Regression diagnostic plot for Model $\mathbf{y}_4 \sim \mathbf{X}_2$ when using robust distances and LTS residuals (left) and classical Mahalanobis distances and LS residuals (right)

The robust estimation associated with an LTS regression detects again much more points that do not follow the pattern and therefore they are classified as outliers. In addition, the classical Mahalanobis distances are more scattered. That means, some distances are very big compared to the others. In contrast, the robust distances are larger in average, but there are not so many points that deviate that much.

So the final conclusion after this executions is that robust methods yield different results, they detect more outliers and all regression analyses that are applied on data with outliers should be done like that.

6.2.2 Multivariate response and compositional explanatory variables

We consider a model where the response $\mathbf{Y}$ somehow represents the climate $(\mathbf{y}_1, \mathbf{y}_2)$ and where the explanatory variables are the different aqua regia elements $\mathbf{X}_4$. The variables with bad quality have been removed. Moreover, the response variable $\mathbf{y}_2$ including the mean precipitation has been log-transformed due to a strong skewness.
The algorithm is mainly taken from Joossens (2008), but it has been adapted. It is a fast MLTS algorithm that has been introduced in Hubert et al. (2008). Additionally, the compositions have been transformed by means of an ilr transformation. A special kind of balances that has been described in section 3 is used so that we gain all the information about the first variable relatively to all the others. For the exact implementation see Appendix A, function `mlts`. γ is chosen as 0.75. This value is a good trade-off between robustness and efficiency.
In table 6.6 we see the output of the model. The first matrix contains the estimated regression coefficients, but just the first row of coefficients is straightforward to interpret. The others are just parts relatively to the rest of the variables and therefore, due to the choice of the balances, not easy to interpret. The second matrix includes the covariance matrix of the residuals $\mathbf{\Sigma}$. In addition the coefficient of determination R^2 has been derived. It is about 46% and thus higher than for the models we have studied before. Nevertheless, one has to be careful with the interpretation of R^2 because we suppose it has to be modified for our robust regression model. To be specific, the estimates $\mathbf{B}$ and $\mathbf{\Sigma}$ are robust, but $\mathbf{X}$ is unadjusted of outliers and therefore, they still contain leverage points.

```
> model2b<-mlts(x=isomLR(X4),y=cbind(y1,y2),gamma=0.75,ns=3000,nc=20,delta=0.01)
```

```
First 6 rows of matrix B containing the estimated coefficients:

              y1          y2
[1,]  1.03029642  0.07231715
[2,] -0.36543233 -0.27125751
[3,] -1.21361199  0.06356850
[4,]  0.08639065  0.03728086
[5,] -0.08608759  0.10319597
[6,] -0.16355394  0.01016086

Estimated covariance matrix of the residuals:

            y1          y2
y1  1.92735975 -0.04492821
y2 -0.04492821  0.02828361

The coefficient of determination R_squared is 0.4616483
```

Table 6.6: Model output for $\mathbf{Y} \sim \mathbf{X}_4$

In figure 6.4 we can see a regression outlier map. On the x-axis there are the robust distances of $\mathbf{X}_3$ estimated by the MCD method, and on the y-axis there are the robust distances of residuals estimated by MLTS. The latter robust distances are defined as

$$d(\mathbf{r}_i(\hat{\boldsymbol{\mu}}_{\text{MCD}})) = \sqrt{\mathbf{r}_i(\hat{\boldsymbol{\mu}}_{\text{MCD}})'(\hat{\boldsymbol{\Sigma}}_e)^{-1}\mathbf{r}_i(\hat{\boldsymbol{\mu}}_{\text{MCD}})}, \tag{6.1}$$

where $\hat{\boldsymbol{\mu}}_{\text{MCD}}$ and $\hat{\boldsymbol{\Sigma}}_e$ are the MCD regression estimates.
For both distances, a quantile (e.g. 0.975) of the χ^2-distribution with the corresponding dimensions as degrees of freedom can be used as cutoff values.

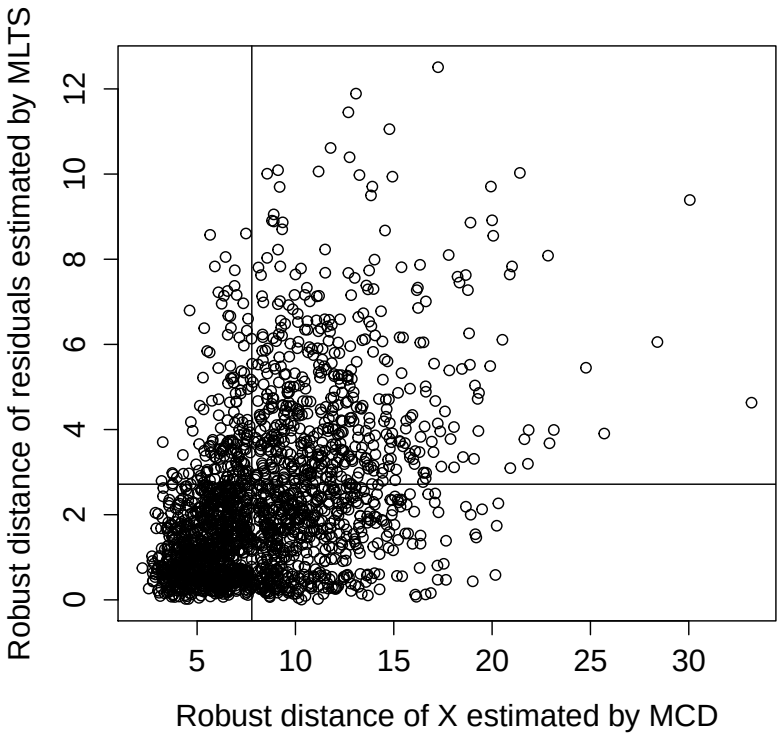

Figure 6.4: Regression outlier map

The observations on the bottom left are the regular ones, those to the top left are vertical outliers. Their residuals are outlying, but their x-values are not. Observations on the top right can be classified as bad leverage points. According to the plot, there are a lot of outliers, vertical outliers as well as leverage points.
To verify the validity of a robust model, we want to show another figure having on the x-axis the classical Mahalanobis distance and on the y-axis the distances of the residuals coming from an LS fit. Figure 6.5 shows the results.

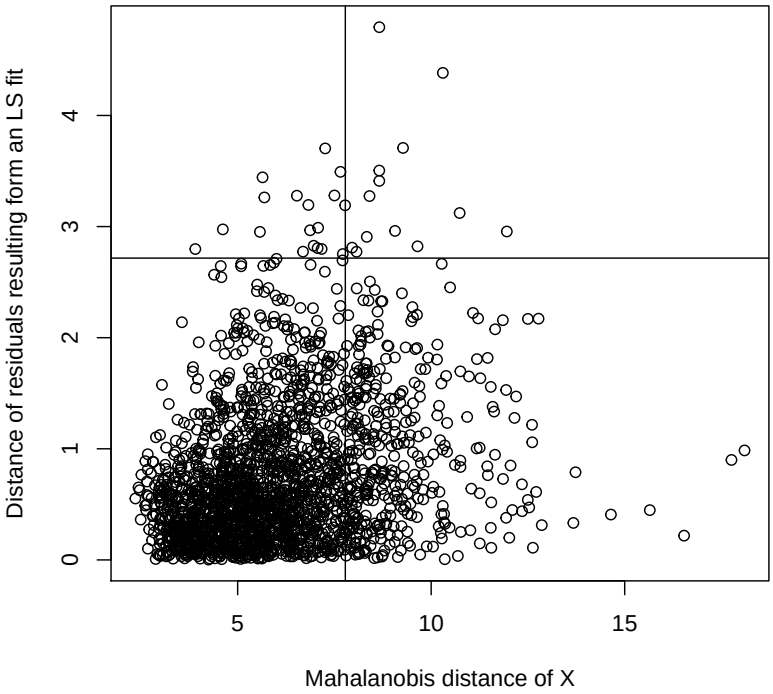

Figure 6.5: Regression outlier map of the model $(\mathbf{y}_1, \mathbf{y}_2) \sim \mathbf{X}_4$

The next step is to compare classical Mahalanobis distances with their robust counterparts. Figure 6.6 shows the results. The data on the bottom left are the regular points. On the top left there are those points that are classified as outliers when robust methods are applied, but which are not seen as outliers when the estimates of center and covariance are computed by classical methods. On the top right those points are situated which are classified as outliers with both methods.

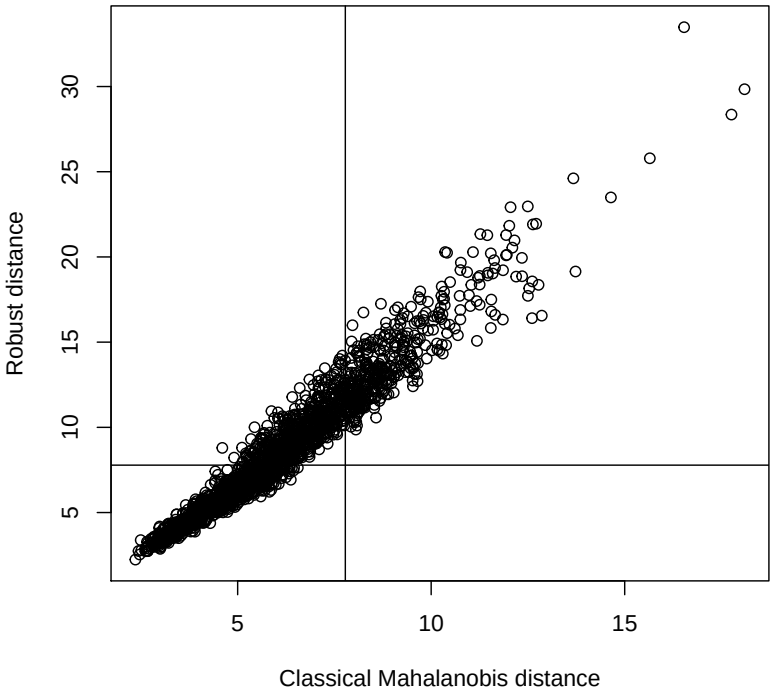

Figure 6.6: Classical Mahalanobis distance vs robust distance of the data $\mathbf{X}_3$

To sum up, we can say that there are a lot of points which can not be identified as outliers if classical estimates are used. Therefore, it makes sense to apply MCD regression.
Furthermore, we want to compare the results using ilr-transformed variables and original variables. In figure 6.7 we see distance-distance plots that compare outlier detection for the two cases.

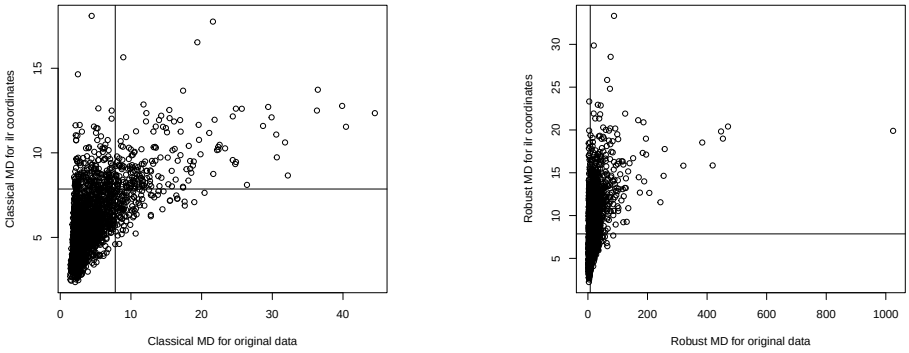

Figure 6.7: Distance-distance plots comparing outlier detection for the original and ilr-transformed data, based on classical (left) and robust (right) estimates

From figure 6.7 we want to get to know if the transformation (the opening) of the data to the Euclidean space is relevant to outlier detection or whether the same results would appear without any transformation. Classical as well as robust (MCD) estimates were used. The horizontal axes represent the Mahalanobis distances using the untransformed original data, and the vertical axes are for the ilr-transformed data. The lines indicate the cut-off values. Since there are points in all four quadrants we can conclude that we would not have obtained the same results when using the original data. Moreover, robust regression shows an even more different picture. There is one extraordinary outlier that would not have been detected without robust estimates.
In table 6.7 we can see the robust estimated coefficients of the regression model where again the regression coefficients are the first elements of 42 models, respectively.

```
coordinate: Ag   F-statistic: 137.226    coefficients:  1.804756   0.0350523
coordinate: Al   F-statistic: 171.0029   coefficients: -1.082672  -0.4973153
coordinate: As   F-statistic: 19.15011   coefficients: -0.7073182 -0.03008587
coordinate: B    F-statistic: 8.898786   coefficients:  0.203309   0.03606078
coordinate: Ba   F-statistic: 71.61085   coefficients: -0.2971173  0.1705481
coordinate: Be   F-statistic: 4.295982   coefficients: -0.2644376  0.03447392
coordinate: Bi   F-statistic: 4.747388   coefficients:  0.1867888 -0.04461269
coordinate: Cd   F-statistic: 7.761615   coefficients: -0.4673227 -0.04310271
coordinate: Ce   F-statistic: 17.57624   coefficients:  0.4302734  0.3172852
coordinate: Co   F-statistic: 2.12691    coefficients: -0.06903757 0.05157953
coordinate: Cr   F-statistic: 5.745188   coefficients: -0.7294582  0.0273239
coordinate: Cs   F-statistic: 18.21438   coefficients: -0.4975111 -0.1056292
coordinate: Cu   F-statistic: 32.88867   coefficients: -0.9063549 -0.0546561
coordinate: Fe   F-statistic: 95.43264   coefficients:  0.7207731 -0.3400886
coordinate: Ga   F-statistic: 74.94313   coefficients:  2.129636   0.3713608
coordinate: K    F-statistic: 3.53295    coefficients: -0.4874833 -0.02025005
coordinate: La   F-statistic: 34.02161   coefficients:  3.005086   0.3183177
coordinate: Li   F-statistic: 19.87053   coefficients: -0.9264747 -0.0789191
coordinate: Mg   F-statistic: 32.26881   coefficients:  0.2520048 -0.108691
coordinate: Mn   F-statistic: 10.90119   coefficients:  0.2784748 -0.06038848
coordinate: Mo   F-statistic: 17.41589   coefficients:  0.6194729  0.02747035
coordinate: Na   F-statistic: 4.060501   coefficients: -0.2073859  0.02438701
coordinate: Nb   F-statistic: 4.186128   coefficients:  0.4094687 -0.007552187
coordinate: Ni   F-statistic: 1.701966   coefficients:  0.2130337  0.04046002
coordinate: P    F-statistic: 44.97306   coefficients:  1.25025   -0.07842878
coordinate: Pb   F-statistic: 36.42795   coefficients: -1.603703  -0.02139326
coordinate: Rb   F-statistic: 3.849817   coefficients:  0.1123132  0.07384382
coordinate: S    F-statistic: 28.70643   coefficients:  0.5250282 -0.03202781
coordinate: Sb   F-statistic: 27.04833   coefficients: -0.900579  -0.03989462
coordinate: Sc   F-statistic: 7.595004   coefficients:  1.062606   0.02095384
coordinate: Se   F-statistic: 2.566379   coefficients: -0.249249  -0.009471893
coordinate: Sn   F-statistic: 35.17172   coefficients: -0.9139126 -0.08494874
coordinate: Sr   F-statistic: 32.22416   coefficients: -0.6366728  0.0238663
coordinate: Th   F-statistic: 4.265048   coefficients: -0.1943988  0.03889979
coordinate: Ti   F-statistic: 6.91777    coefficients:  0.3672628  0.01836594
coordinate: Tl   F-statistic: 9.051492   coefficients:  0.7228025  0.04108
coordinate: U    F-statistic: 13.68895   coefficients:  0.3863221 -0.06502947
coordinate: V    F-statistic: 3.779437   coefficients:  0.4848099 -0.03940839
coordinate: W    F-statistic: 8.176248   coefficients: -0.2772011  0.02357514
coordinate: Y    F-statistic: 19.2789    coefficients: -1.150997  -0.1010479
coordinate: Zn   F-statistic: 10.9982    coefficients:  1.030445   0.05020505
coordinate: Zr   F-statistic: 14.29625   coefficients:  0.3963563  0.02501617

The corresponding 95%-quantile of the F-distribution is 3.000184
```

Table 6.7: Robust coefficients and F-statistics of the model $(\mathbf{y}_1, \mathbf{y}_2) \sim \mathbf{X}_4$

There are just a few elements whose coordinates do not influence the response. But sometimes it is even interesting to know exactly these, so we want to enumerate them: `Co, Ni, Se`. According to an expert of geochemistry, it is very surprising that Selen does not have an impact on the climate. From that fact follows that we have to be careful with results and interpretations and besides, it increases the uncertainty of the correctness of both, the F-statistics and the coefficient of determination R^2 in case of a robust regression. All the other elements do have either a significant positive or

negative effect on the mean temperature and annual precipitation.

Furthermore we want to obtain information about how the `XRF` data influence the climate $(\mathbf{y}_1, \mathbf{y}_2)$.

```
> model2c<-mlts(x=isomLR(X2),y=cbind(y1,y2),gamma=0.75,ns=3000,nc=20,delta=0.01)
```

```
estimated covariance matrix of residulas:

             y1          y2
y1  2.51913661 -0.07683724
y2 -0.07683724  0.04015629

The coefficient of determination equals  0.2394992
```

Table 6.8: Estimated matrix $\mathbf{\Sigma}$ and coefficient of determination R^2 in the model $(\mathbf{y}_1, \mathbf{y}_2) \sim \mathbf{X}_2$

The coefficient of determination is rather low and therefore, we have to suppose that there are much more inputs that describe the climate (which is somehow obvious) or that the evaluation of the F-statistics is not coherent with the robust multivariate regression model. Nevertheless, we can conclude that the oxides have much more influence on the climate. At least this is a meaningful result.
In table 6.9 we see the estimated coefficients as well as the corresponding F-statistics for the coordinates. Unfortunately, all the coordinates are significant, except for one that is `Zn_XRF`. According to the output, all the others do show either a postive or a negative influence on the climate or rather on the mean temperature and annual precipitation.

```
coordinate: Si_XRF   F-statistic: 64.01032   coefficients: -2.551768  -0.04536396
coordinate: Ti_XRF   F-statistic: 24.80653   coefficients: -1.048606  -0.291589
coordinate: Al_XRF   F-statistic: 51.13712   coefficients: -1.697115  -0.4478831
coordinate: Fe_XRF   F-statistic: 9.614859   coefficients:  1.342485   0.1082234
coordinate: Mn_XRF   F-statistic: 25.50529   coefficients:  0.4899351 -0.1017633
coordinate: Mg_XRF   F-statistic: 40.73675   coefficients:  1.474431   0.1171863
coordinate: Ca_XRF   F-statistic: 16.12205   coefficients: -0.5968735  0.01671626
coordinate: Na_XRF   F-statistic: 37.60736   coefficients: -0.004207717 -0.09741506
coordinate: K_XRF    F-statistic: 146.9889   coefficients:  5.571251  -0.5745473
coordinate: P_XRF    F-statistic: 42.57841   coefficients:  0.3500521 -0.1460284
coordinate: Ba_XRF   F-statistic: 396.2044   coefficients:  2.365908   0.7958941
coordinate: Cr_XRF   F-statistic: 51.83016   coefficients: -0.8691962  0.1279036
coordinate: Nb_XRF   F-statistic: 10.36878   coefficients:  0.07018358 0.1431354
coordinate: Pb_XRF   F-statistic: 56.52273   coefficients: -1.927884  -0.05923645
coordinate: Rb_XRF   F-statistic: 50.41385   coefficients: -1.546535   0.3099586
coordinate: Sr_XRF   F-statistic: 13.86738   coefficients:  0.101475  -0.1000597
coordinate: V_XRF    F-statistic: 10.57578   coefficients:  0.1582521 -0.1484785
coordinate: Y_XRF    F-statistic: 37.52931   coefficients:  2.378658   0.1596321
coordinate: Zn_XRF   F-statistic: 0.4736852  coefficients:  0.1922837  0.02018438
coordinate: Zr_XRF   F-statistic: 13.42876   coefficients:  1.40903    0.08318843

The corresponding 95%-quantile of the F-distribution is 3.000136
```

Table 6.9: F statistics and estimated parameters of model $(\mathbf{y}_1, \mathbf{y}_2) \sim \mathbf{X}_2$

6.2.3 Linear regression with compositional response and non-compositional explanatory variables

Now we consider a regression model where we want to explain the parts of sand, silt and clay by the non-compositional explanatory variables that describe the climate. Formally we can write it down as:

$$\boldsymbol{X}_1 = \boldsymbol{Y}\boldsymbol{B} + \mathbf{E}$$

The code to derive the coefficients by means of the robust regression with compositional variables is given in Appendix A (function `mlts`). Again a fast MLTS algorithm is used to calculate the coefficients as well as the residual variance and the distance of the robust residuals. To describe the output, a regression outlier map is plotted in figure 6.8. In table 6.10 the estimated coefficients for beta and the matrix $\boldsymbol{\Sigma}$ are given.

```
> yn<-isomLR(X1)
> k<-which(is.infinite(yn)[,1])
> y<-yn[-k,]
> x<-cbind(y1,y2)[-k,]
> model3<-mlts(x=x,y=y,gamma=0.75,ns=3000,nc=20)
```

```
> model3$beta

           [,1]       [,2]
y1  0.01449668  0.0425192
y2 -0.04881073 -0.1271665

> model3$sigma

            [,1]        [,2]
[1,]  0.36199968 -0.03620171
[2,] -0.03620171  0.08823674
```

Table 6.10: First estimates for $\mathbf{B}$ and estimator for $\mathbf{\Sigma}$ in model $\mathbf{X}_1 \sim (\mathbf{y}_1, \mathbf{y}_2)$

To derive inference statistics, one has to compute $D \cdot 2$ regression models and calculate the F-statistic (see section 5.1) for each of the estimations separately. Then information about how the climate influences parts of the soil can be given.

```
### calculations of the coefficients
> for (i in 1:2){
+ yn<-isomLR(X1)
+ k<-which(is.infinite(yn)[,1])
+ y<-yn[-k,]
+ x<-cbind(y1,y2)[-k,]
+ model3<-mlts(x=x,y=y,gamma=0.75,ns=500,nc=10)
+ M<-diag(model3$n)-x%*%solve(t(x)%*%x)%*%t(x)
+ h<-c(1,rep(0,model3$p-1))
+ nom<-model3$beta[1,]%*%solve(t(y)%*%M%*%y)%*%t(model3$beta)%*%h
+ den<-t(h)%*%solve(t(x)%*%x)%*%h
+ quod<-(model3$n-model3$p-model3$q) /model3$q
+ F[i]<-nom/den*quod
+ sse<-sum((model3$res)^2)
+ }
```

```
### calculation of r.squared
> for (k in 1:model3$n)
+ {
+   for (j in 1:(model3$q-1))
+   {
+     ssr=ssr+((x%*%model3$beta)[k,j]-mean(y[,j]))^2
+   }
```

```
+ }
> R<-ssr/(ssr+sse)
```

```
coordinate: y1  F-statistic: 116.2002  coefficients: 0.01449668 0.0425192
coordinate: y2  F-statistic: 116.4164  coefficients: 0.01477982 0.04256854

The corresponding 95%-quantile of the F-distribution is 3.000102

The coefficient of determination equals 0.1727367
```

Table 6.11: Regression parameters of model $\mathbf{X}_1 \sim (\mathbf{y}_1, \mathbf{y}_2)$

In table 6.11 we can find the estimated F-statistics as well as the coefficients for the coordinates. Moreover the 95%-quantile of the F-distribution is denoted. The coefficient of determination is very low. That leads us to the conclusion that the non-compositional explanatory variables $\mathbf{y}_1$ und $\mathbf{y}_2$ do not contribute well to describe the coordinates of data set $\mathbf{X}_1$.

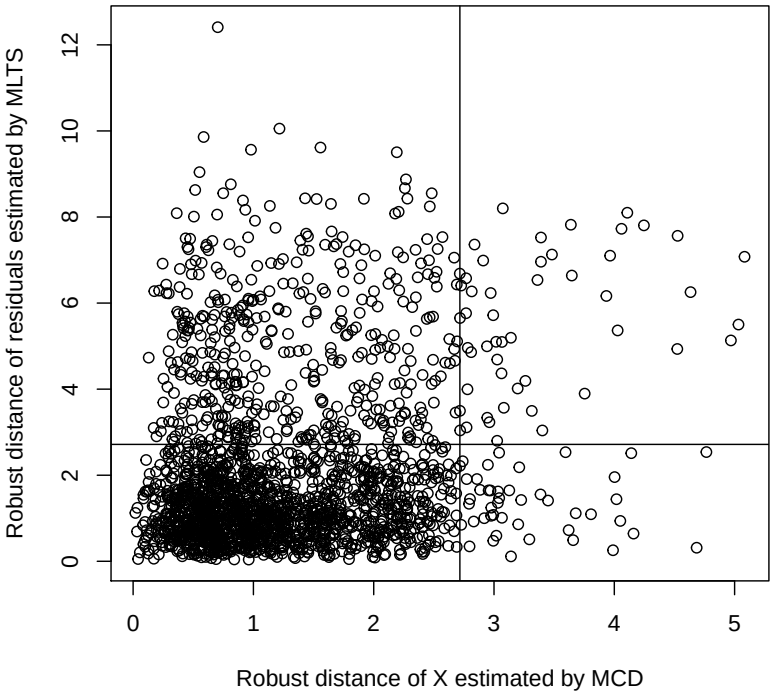

Figure 6.8: Regression outlier map for model $\mathbf{X}_1 \sim \mathbf{Y}$

In figure 6.8 we see a big amount of vertival outliers, but significantly less leverage points.

Figure 6.9: Regression outlier map for common multivariate least-squares regression associated with the classical Mahalanobis distance

6.2.4 Linear regression with compositional response and compositional explanatory variables

Finally, we consider the model, where we are confronted with compositional explanatory variables as well as a compositional response. An ilr transformation has to be applied on both separately and afterwards, the regression analysis can be started.
Here, we like to have a look on the relationship between the variable $\mathbf{X}_1$ and the variable $\mathbf{X}_3$. Both contain elements of the soil and sum up to about one so that they can be considered as compositions. Again an MLTS estimator will be used to derive the coefficients and the robust distances of the coordinates.
To define proper inference statistics one has to apply $D \cdot P$ estimations where D is the dimension of $\mathbf{X}_1$ and P is the dimension of $\mathbf{X}_3$ (cf. section 3.3). In figure 6.10 the regression outlier map is shown for robust estimates as well

as for classical estimates.

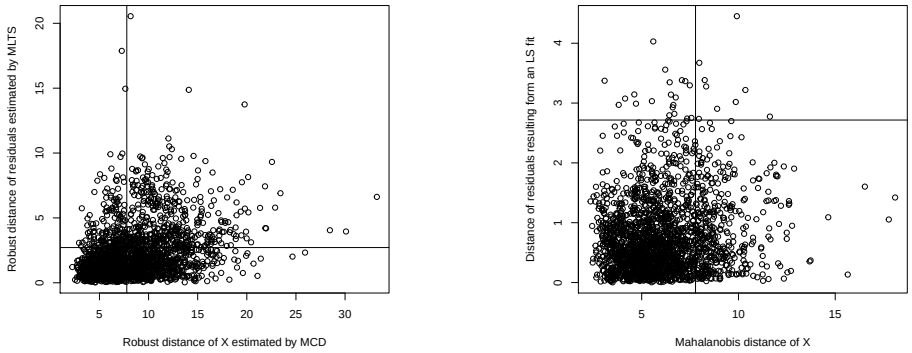

Figure 6.10: Regression outlier map of the model $\mathbf{X}_1 \sim \mathbf{X}_3$ with robust esimates (left) and classical estimates (right)

Figure 6.7 represents the associated plots to compare the Mahalanobis distances with the original data as well as the ilr-transformed variables in case of a classical and a robust estimation of center and covariance. Since we use the same explanatory variables, we do not have to repeat calculations and plots.
Finally, we want to find out which coordinates are the significant ones to describe the parts of $\mathbf{X}_1$. Therefore, first we use formula (3.2), which is the F-statistic to get to know if some variables do not help to describe the response at all. We calculate F_1 to determine the impact of part 1 on the response and repeat the computations for all variables.

```
> model4<-mlts_model4(x=x,y=y,gamma=0.75,ns=3000,nc=20,delta=0.01)
```

```
coordinate: Ag  F-statistic: 19.49802   coefficients: -0.1612934   0.03285746
coordinate: Al  F-statistic: 54.21285   coefficients: -0.6341268   0.1467685
coordinate: As  F-statistic: 10.90387   coefficients: -0.112539   -0.05246823
coordinate: B   F-statistic: 18.88954   coefficients: -0.1103841   0.06341695
coordinate: Ba  F-statistic: 2.6927     coefficients: -0.07798289 -0.01040735
coordinate: Be  F-statistic: 24.17039   coefficients:  0.1986603  -0.1298603
coordinate: Bi  F-statistic: 13.31353   coefficients:  0.1554583  -0.09720959
coordinate: Cd  F-statistic: 17.52207   coefficients:  0.1295278  -0.1352273
coordinate: Ce  F-statistic: 8.115012   coefficients: -0.4482347   0.1844639
coordinate: Co  F-statistic: 6.81569    coefficients: -0.1919943  -0.07551753
coordinate: Cr  F-statistic: 4.937498   coefficients: -0.1803643   0.0005673683
coordinate: Cs  F-statistic: 2.278934   coefficients: -0.03587578  0.06610231
coordinate: Cu  F-statistic: 0.3754524  coefficients:  0.02596895 -0.002259261
coordinate: Fe  F-statistic: 15.3314    coefficients:  0.3112942  -0.09983005
coordinate: Ga  F-statistic: 10.30004   coefficients:  0.3107987   0.1213082
coordinate: K   F-statistic: 11.19385   coefficients:  0.02478652 -0.1652272
coordinate: La  F-statistic: 6.006913   coefficients:  0.2926289  -0.2323421
coordinate: Li  F-statistic: 0.004125703  coefficients: -0.003279296 -0.001582096
coordinate: Mg  F-statistic: 12.07416   coefficients:  0.03308041 -0.1220261
coordinate: Mn  F-statistic: 12.49982   coefficients:  0.006591762 0.1310796
coordinate: Mo  F-statistic: 11.31325   coefficients: -0.05793445  0.08744876
coordinate: Na  F-statistic: 2.398094   coefficients: -0.01802845 -0.04364006
coordinate: Nb  F-statistic: 12.82898   coefficients: -0.1796762  -0.005182686
coordinate: Ni  F-statistic: 1.336236   coefficients: -0.06490433  0.05063552
coordinate: P   F-statistic: 1.905363   coefficients: -0.04239332  0.04943112
coordinate: Pb  F-statistic: 6.843804   coefficients:  0.1408551   0.06457379
coordinate: Rb  F-statistic: 5.693206   coefficients:  0.1862649  -0.07778464
coordinate: S   F-statistic: 243.2536   coefficients:  0.4156775   0.1151362
coordinate: Sb  F-statistic: 1.842761   coefficients:  0.05274823  0.01944933
coordinate: Sc  F-statistic: 15.32511   coefficients: -0.293194   -0.1558724
coordinate: Se  F-statistic: 7.133014   coefficients:  0.01751019  0.07677634
coordinate: Sn  F-statistic: 3.855366   coefficients:  0.02440537 -0.06917625
coordinate: Sr  F-statistic: 3.514989   coefficients: -0.0496704  -0.01799025
coordinate: Th  F-statistic: 3.996911   coefficients: -0.09741875 -0.02228071
coordinate: Ti  F-statistic: 38.16463   coefficients:  0.1382673   0.1275837
coordinate: Tl  F-statistic: 5.341979   coefficients: -0.1450275   0.02358795
coordinate: U   F-statistic: 0.9095538  coefficients:  0.04715913  0.00545617
coordinate: V   F-statistic: 29.50146   coefficients:  0.4301445   0.04849563
coordinate: W   F-statistic: 1.191041   coefficients: -0.03353508 -0.006012791
coordinate: Y   F-statistic: 9.35543    coefficients:  0.04430097  0.1777767
coordinate: Zn  F-statistic: 1.937728   coefficients:  0.02261359 -0.08325237
coordinate: Zr  F-statistic: 6.425577   coefficients: -0.07088554  0.01223383
```

Table 6.12: F-statistics and regression coefficients for the model $\mathbf{X}_1 \sim \mathbf{X}_4$

The output gives us the following conclusions. There are just a few coordinates which do not help to describe the model. The corresponding elements are `Ba, Cs, Cu, Li, Na, Ni, P, Sb, U, W, Zn`. Hence, if the part in the soil of these elements is increasing or decreasing it does not have a big effect on the response (on the parts of sand, silt or clay in the soil). However, one has to be careful with interpretations since the coefficients are always derived for the elements as a part of the rest of the others. So they describe the effect of one part relatively to all the others. The estimated coefficients as well as the values of the F-statistics are shown in table 6.12. Hence, we can identify in which way each coordinate influences the coordinates of the

response, which is either a positive or a negative effect.

```
The coefficient of determination equals  0.3310495
```

This value for the coefficient of determination is again rather low, but since we assume that the associated calcualtions have to be modified, we should not pay too much attention on it here.

Finally, experts are interested in how the three parts `sand_norm`, `silt_norm` and `clay_norm` do describe the aqua regia compositions $\mathbf{X}_4$. Thus, we consider the following model in coordinates:

$$\mathbf{X}_4^* = \mathbf{X}_1^* \mathbf{D}^* + \mathbf{E}^*$$

Special attention should be paid on `clay_norm` which is supposed to have the biggest impact on the response.

```
coordinate: sand_norm  F-statistic: 51.761  coefficients: -0.1292351
0.2280944 -0.04827934 -0.04255178 -0.03812094 -0.2292504 -0.1749652
-0.02653513 0.07722091 -0.182723 -0.09893409 -0.1314296 -0.1106287
0.2565445 -0.09313518 0.0979192 0.03925649 -0.06903242 0.1306204
0.1377602 0.03849332 0.06822224 -0.04619676 -0.2354581 0.421024
0.2076411 0.07989358 0.4071942 -0.05701899 -0.1568177 -0.0790466
0.06057745 -0.06715394 -0.2236793 0.4067046 -0.06341329 0.1944305
0.09009965 0.2568383 0.1353969 0.5686421
coordinate: silt_norm  F-statistic: 104.1527  coefficients: -6.328288
8.114455 -0.6969301 -1.060469 2.437355 -3.085465 -4.709611 -4.704712
1.327212 -0.3718612 0.7493356 -2.693116 0.1659322 8.326003 -1.123514
6.012682 0.9908091 0.5269292 6.939519 5.215803 -3.055548 2.451072
-2.530963 1.399134 5.553065 1.539647 1.863546 4.764256 -2.889851
-0.1077786 -2.54748 -2.027077 1.948487 0.2001244 4.135395 -3.45367
-1.872938 2.218568 -4.993134 -0.2926793 2.023815
coordinate: clay_norm  F-statistic: 459.6819  coefficients: 6.362659
-8.208366 0.7616963 1.103356 -2.350418 3.288606 4.828801 4.686759
-1.391697 0.5564396 -0.6364328 2.780329 -0.04822509 -8.443961 1.205417
-6.021147 -1.02412 -0.4441121 -6.964346 -5.239063 2.96821 -2.504109
2.510506 -1.126744 -5.881061 -1.690053 -1.916853 -5.094725 2.94726
0.2728092 2.605574 1.968038 -1.850967 0.02299315 -4.528684 3.466756 1.623705
-2.271234 4.668933 0.1588271 -2.531733

The corresponding quantile of the F-distribution is 1.395007
```

Table 6.13: Estimated coefficients and F-statistics of model $\mathbf{X}_4 \sim \mathbf{X}_1$

Table 6.13 showes the estimated coefficients and F-statistics of the model. The computations were very difficult because during the calculations there occured problems of singularity. So, we cannot be sure that the algorithm has converged. Nevertheless, the hypothesis that `clay_norm` may have to biggest impact on the response cannot be refused, since the F-statistic is big.

The coefficient of determination is given by 0.4852269 and is therefore the highest ever obtained during this research.

Chapter 7

Conclusions

Multivariate classical and robust analysis of compositional data is a recent topic. In this thesis, the goal was to gain information between the difference of classical and robust regression applied on compositions. Several methods, like regression outlier plots, coefficients estimations or computing test statistics, have been used to make proper interpretations possible.
First of all, compositions have to be transformed because otherwise they cannot be interpreted in terms of coordinates of the Euclidean space. For the transformations three different methods were proposed - additive log-ratio transformation (alr), centered log-ratio transformation (clr) and isometric log-ratio transformation (ilr). The ilr transformation is chosen to be used for applications since it has some important properties. It is both, isometric and an isomorphism. Moreover, it yields coordinates in $\mathbb{R}^{D-1}$ which is the actual dimension of the simplex and therefore reasonable. Special ilr transformations, called balances, have been introduced. Balances make coordinates easier to interpret. The most important balances describe all the information about one coordinate relative to all the others. By means of these balances, classical and robust regression have been introduced, so that there is also an easy way for interpretation.
For the different models with compositions and non-compositional parts, regression coefficients were derived and the related test statistics were formulated.
Finally, the most important part was to implement the regression analysis in R and to try to obtain interesting results and conclusions about the topic. To sum up, we first had a look on a multiple regression model with compositional explanatory variables. We could see that robust regression does make a difference in the estimations. More outliers were detected by this method. Moreover, we can conclude, that also the transformation (the opening of the data) is necessary to receive correct results. Otherwise, the results and con-

clusions of the estimated coefficients differed and aside from that there is no way to interpret the coefficients in the simplex.
However, the coefficient of determination is almost the same when we compare a model with original (compositional) data and one with coordinates. That means that the adaption is similar in both models, but again, the problem occurs when we want to interpret the results. Therefore, the proper selection of the balances (ilr transformation) is crucial. Here we always chose the balances where one element is represented as part of all the others. Nonetheless, if we want to obtain information about one group relative to another group, a balance expressing this relationship can be used.
Further, one has to be careful with interpreting inference statistics, when both, the response as well as the independent variables form compositions. In that case, the response variables, which will be described in terms of estimated coefficients and coordinates, are coordinates as well and therefore, represent also parts of variables, relative to the rest. Besides, the dimension of the response decreases as a result from an ilr transformation.
Referring to our data set, we want to sum up the most important conclusions that we got from the study. It is reasonable that Iron(III)-oxide has a big positive impact on the response susceptibiliy. The magnetic character is obviously depending on Iron, besides quartz was supposed to infect the response negative. This result could also be seen in our regression. The magnetic character is lost if there is more quartz in the soil.
Accoring to specialist Dr. Reimann, the results of the model where the climate is described by the aqua regia data are surprising. For example, Selen should be influenced by the climate, but in our model it is one of those few variables which do not show any significance. Moreover, the values of R^2 are very low on average. Hence, we presume there is missing a lot of information and probably a modification of the coefficient of determination in case of robust multivariate regression analysis is necessary and would lead to results that are more accurate.
To sum up, we have to admit that the theoretical problem of compositions defined on the simplex is straight-forward to introduce, whereas the conclusions of applications with compositions are rather difficult to interpret as well as to obtain. However, a lot of new insight in the topic has been achieved and the interesting effect of a transformation from the simplex to the Euclidean space has been examined.

Appendix A

R-Codes

This code is mainly taken from the *R* package `robCompositions` and just a little bit adapted to my idea. It applies classical and robust regression analysis on the ilr- transformed explanatory compositions and returns the associated model parameters.

```
lmCoDaX_adapted<-function (y, X, method = "robust")
{
  ilrregression <- function(X, y) {
    d <- data.frame(y = y, X = X)
    lmcla <- lm(y ~ ., data = d)
    lmcla.sum <- summary(lmcla)
    require(robCompositions)
    ilr.sum <- lmcla.sum
    for (j in 1:ncol(X)) {
      Zj <- -robCompositions::isomLR(cbind(X[, j], X[, -j]))
      dj <- data.frame(y = y, Z = Zj)
      res <- lm(y ~ ., data = dj)
      res.sum <- summary(res)
      if (j == 1) {
        ilr.sum$coefficients[1:2, ] <- res.sum$coefficients[1:2, ]
        ilr.sum$residuals <- res.sum$residuals
        ilr.sum$sigma <- res.sum$sigma
        ilr.sum$r.squared <- res.sum$r.squared
        ilr.sum$adj.r.squared <- res.sum$adj.r.squared
        ilr.sum$fstatistic <- res.sum$fstatistic
                    }
      else {
        ilr.sum$coefficients[j + 1, ] <- res.sum$coefficients[2, ]
           }
```

```
    }
    list(lm = lmcla, lm1 = lmcla.sum, ilr = ilr.sum)
  }
  robilrregression <- function(X, y) {
    require(robustbase)
    d <- data.frame(y = y, X = X)
    lmcla <- ltsReg(y ~ ., data = d)
    lmcla.sum <- summary(lmcla)
    require(robCompositions)
    ilr.sum <- lmcla.sum
    for (j in 1:ncol(X)) {
      Zj <- -robCompositions::isomLR(cbind(X[, j], X[, -j]))
      dj <- data.frame(y = y, Z = Zj)
      res <- ltsReg(y ~ ., data = dj)
      res.sum <- summary(res)
      if (j == 1) {
        plot(res,which="rdiag",id.n=10)
        ilr.sum$coefficients[1:2, ] <- res.sum$coefficients[1:2, ]
        ilr.sum$residuals <- res.sum$residuals
        ilr.sum$sigma <- res.sum$sigma
        ilr.sum$r.squared <- res.sum$r.squared
        ilr.sum$adj.r.squared <- res.sum$adj.r.squared
        ilr.sum$fstatistic <- res.sum$fstatistic
                  }
      else {
        ilr.sum$coefficients[j + 1, ] <- res.sum$coefficients[2, ]
      }
    }
    list(lm = lmcla, lm1 = lmcla.sum, ilr = ilr.sum)
  }
  if (method == "classical") {
    reg <- ilrregression(X, y)
  }
  else if (method == "robust") {
    reg <- robilrregression(X, y)
  }
  return(reg)
}
```

```
mlts <- function(x,y,gamma,ns=500,nc=10,delta=0.01)
{
  d <- dim(x); n <- d[1]; p <- d[2]
  q <- ncol(y)
  h <- floor(n*(1-gamma))+1
  obj0 <- 1e10
  for (i in 1:ns)
  { set.seed(i)
    sorted <- sort(runif(n),na.last = NA,index.return=TRUE)
    istart <- sorted$ix[1:(p+q)]
    xstart <- as.matrix(x[istart,])
    ystart <- as.matrix(y[istart,])
    bstart <- solve(t(xstart)%*%xstart,t(xstart)%*%ystart)
    sigmastart <- (t(ystart-xstart%*%bstart))%*%(ystart-xstart%*%bstart)/q
    for (j in 1:nc)
    { res  <-  y - x %*% bstart
      tres <- t(res)
      dist2 <- colMeans(solve(sigmastart,tres)*tres)
      sdist2 <- sort(dist2,na.last = NA,index.return = TRUE)
      idist2 <- sdist2$ix[1:h]
      xstart <- as.matrix(x[idist2,])
      ystart <- as.matrix(y[idist2,])
      bstart <- solve(t(xstart)%*%xstart,t(xstart)%*%ystart)
      sigmastart <- (t(ystart-xstart%*%bstart))%*%(ystart-xstart%*%bstart)/(h-p)
    }
    obj <- det(sigmastart)
    if (obj < obj0)
    { result.beta <- bstart
      result.sigma <- sigmastart
      obj0 <- obj
               }
  }
  cgamma <- (1-gamma)/pchisq(qchisq(1-gamma,q),q+2)
  result.sigma <- cgamma * result.sigma
  res <- y - x %*% result.beta
  tres<-t(res)
  result.dres <- colSums(solve(result.sigma,tres)*tres)
  result.dres <- sqrt(result.dres)
  rd_temp<-covMcd(x)
  rd<-sqrt(mahalanobis(x,rd_temp$center,rd_temp$cov))
  plot(rd,result.dres,xlab="Robust distance of X estimated by MCD",
  ylab="Robust distance of residuals estimated by MLTS")
  abline(h=sqrt(qchisq(0.975,ncol(y))))
```

```
  abline(v=sqrt(qchisq(0.975,ncol(x))))
  list(beta=result.beta,sigma=result.sigma,dres=result.dres,rd=rd,p=p,q=q,n=n)
}




### fast mlts algorithm

mlts_model4<-function(x,y,gamma,ns=500,nc=10,delta=0.01)
{

  d <- dim(x); n <- d[1]; p <- d[2]
  q <- ncol(y); n2<-nrow(y)
  h <- floor(n*(1-gamma))+1
  obj0 <- 1e100

  for (i in 1:ns)
  { set.seed(i)
    sorted <- sort(runif(n),na.last = NA,index.return=TRUE)
    istart <- sorted$ix[1:(p+q)]
    xstart <- as.matrix(x[istart,])
    ystart <- as.matrix(y[istart,])
    bstart <- solve(t(xstart)%*%xstart,t(xstart)%*%ystart)
    sigmastart <- (t(ystart-xstart%*%bstart))%*%(ystart-xstart%*%bstart)/q
    for (j in 1:nc)
    { res  <-  y - x %*% bstart
      tres <- t(res)
      dist2 <- colMeans(solve(sigmastart,tres)*tres)
      sdist2 <- sort(dist2,na.last = NA,index.return = TRUE)
      idist2 <- sdist2$ix[1:h]
      xstart <- as.matrix(x[idist2,])
      ystart <- as.matrix(y[idist2,])
      bstart <- solve(t(xstart)%*%xstart,t(xstart)%*%ystart)
      sigmastart <- (t(ystart-xstart%*%bstart))%*%(ystart-xstart%*%bstart)/(h-p)
    }
    obj <- det(sigmastart)
    if (obj < obj0)
    { result.beta <- bstart
      result.sigma <- sigmastart
      obj0 <- obj
    }
  }
```

```
  cgamma <- (1-gamma)/pchisq(qchisq(1-gamma,q),q+2)
  result.sigma <- cgamma * result.sigma
  res <- y - x %*% result.beta
  tres<-t(res)
  result.dres <- colSums(solve(result.sigma,tres)*tres)
  result.dres <- sqrt(result.dres)
  rd_temp<-covMcd(x)
  rd<-sqrt(mahalanobis(x,rd_temp$center,rd_temp$cov))
  plot(rd,result.dres,xlab="Robust distance of X estimated by MCD",
       ylab="Robust distance of residuals estimated by MLTS")
  abline(h=sqrt(qchisq(0.975,ncol(y))))
  abline(v=sqrt(qchisq(0.975,ncol(x))))

  M<-diag(n)-x%*%solve(t(x)%*%x)%*%t(x)
  h<-c(1,rep(0,p-1))
  nom<-result.beta[1,]%*%solve(t(y)%*%M%*%y)%*%t(result.beta)%*%h
  den<-t(h)%*%solve(t(x)%*%x)%*%h
  quod<-(n-p-q) /q
  F<-nom/den*quod

  list(beta=result.beta,sigma=result.sigma,dres=result.dres,F=F,q=q,p=p,n=n)
}
```

List of Figures

List of Tables

Bibliography

J. Agulló, C. Croux, and S. Van Aelst. The multivariate least-trimmed squares estimator. *Journal of Multivariate Analysis*, 99(3):311–338, 2008.

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

C. Barceló-Vidal, J. A. Martín-Fernández, and V. Pawlowsky-Glahn. Mathematical foundations of compositional data analysis. In *Proceedings of IAMG*, volume 1, 2001.

J. J. Egozcue, J. Daunis-i-Estadella, V. Pawlowsky-Glahn, K. Hron, and P. Filzmoser. Simplicial regression. the normal model. *Journal of Applied Probability and Statistics*, 6(1 & 2):87–108, 2011.

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

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

Ludwig Fahrmeir, Hans Wolfgang Brachinger, Alfred Hamerle, and Gerhard Tutz. *Multivariate statistische Verfahren*. Walter de Gruyter, 1996.

P. Filzmoser. Robuste Statistik, 2009.

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

P. Filzmoser and K. Hron. Multivariate outlier detection with compositional data. In *Proceedings of the Ninth International Conference on Computer Data Analysis and Modeling, volume 1*, pages 45–52, 2010a.

P. Filzmoser and K. Hron. Robust methods for compositional data. In *Proceedings of COMPSTAT'2010*, pages 79–88. Springer, 2010b.

P. Filzmoser and K. Hron. Robust statistical analysis. In *Compositional Data Analysis: Theory and Applications*, pages 59–70, Chichester, UK, 2011. Wiley.

P. Filzmoser and K. Hron. Robust compositional regression (i): general concepts and models with compositional response, 2012. Unpublished paper.

P. Filzmoser, K. Hron, and M. Templ. Robust regression of compositional data with applications, 2012. Unpublished paper.

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

M. Hubert, P. J. Rousseeuw, and S. Van Aelst. High-breakdown robust multivariate methods. *Statistical Science*, pages 92–119, 2008.

K. Joossens. MLTS algorithm. 2008. http://www.econ.kuleuven.be/ public/NDBAE06/programs/mlts/mlts.r.txt.

G. Mateu-Figueras, V. Pawlowsky-Glahn, and J. J. Egozcue. The principle of working on coordinates. In *Compositional Data Analysis: Theory and Applications*, pages 31–42, Chichester, UK, 2011. Wiley.

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

V. Pawlowsky-Glahn and J. J. Egozcue. Exploring compositional data with the coda-dendrogram. *Austrian Journal of Statistics*, 40(1-2):103–113, 2011.

V. Pawlowsky-Glahn, J. J. Egozcue, and R. Tolosana-Delgado. Lecture notes on compositional data analysis, 2007. http://hdl.handle.net/10256/297.

G. Pison, S. Van Aelst, and G. Willems. Small sample corrections for LTS and MCD. *Metrika*, pages 11–123, 2002.

C. Reimann, A. Demetriades, O. A. Eggen, P. Filzmoser, and the EuroGeoSurveys Geochemistry expert group. The EuroGeoSurveys geochemical mapping of agricultural and grazing land soils project (GEMAS) - Evaluation of quality control results of aqua regia extraction analysis. *Norges geologiske undersøkelse*, 2009.

P. J. Rousseeuw. Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, 8:283–297, 1985.

P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. Wiley & Sons, New York, 1987.

P. J. Rousseeuw, S. Van Aelst, K. Van Driessen, and J.A. Agulló. Robust multivariate regression. *Technometrics*, 46(3), 2004.