PALACKÝ UNIVERSITY OLOMOUC
FACULTY OF SCIENCE
DEPARTMENT OF MATHEMATICAL ANALYSIS AND MATHEMATICAL APPLICATIONS

Dissertation Thesis

Regression analysis for compositional data

Supervisor:
doc. RNDr. Eva Fišerová, Ph.D.
Year of submission: 2017

Author:
Mgr. Sandra Donevska
Applied mathematics

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Mgr. Sandra Donevska

Název práce: Regresní analýza pro kompoziční data

Týp práce: Dizertační práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: Doc. RNDr. Eva Fišerová Ph.D.

Rok obhajoby práce: 2017

Abstrakt: Práce se zabývá regresní analýzou pro kompoziční data. Relativní charakter kompozičních dat, který je odlišuje od standardních mnohorozměrných dat, vyžaduje speciální zacházení. Jedním ze základních přístupů ke statistické analýze kompozičních dat, který je použit i v této práci, je vyjádření kompozičních dat ve vhodném souřadnicovém systému.

Nejprve je pozornost soustředěna na problematiku regresního modelu s kompoziční vysvětlovanou proměnnou. Pro kompoziční data vyjádřená v ortonormálních souřadnicích je v práci vytvořen mnohorozměrný regresní model a uvedeny explicitní vzorce pro odhady neznámých regresních parametrů a testové statistiky pro ověření jejich statistické významnosti. Dále je navržena jiná souřadnicová reprezentace kompozičních dat, která umožňuje zjednodušit výpočty pro odhady regresních parametrů a testové statistiky a vyhodnocena kvalita predikce v různých souřadnicových systémech.

Druhá část této práce je věnována kalibračnímu problému pro kompoziční data. V práci je použit přístup založený na lineárním modelu s podmínkami typu II. Je zde dokázána ekvivalence mezi lineárními modely s podmínkami typu II a ortogonální regresí. Dále je zde navržena procedura pro kalibraci kompozičních měření a prezentovány testy pro shodu dvou měřicích přístrojů (metod).

V poslední části této práce je navržena procedura pro výběr kompozičních složek, která zaručuje, že výsledná redukce dimenze kompozice nezpůsobí podstatnou ztrátu informace o mnohorozměrné variabilitě datové struktuře.

Všechny teoretické výsledky jsou aplikovány při řešení reálných úloh.

Klíčová slova: kompoziční data; regrese s kompoziční vysvětlovanou proměnnou; kalibrace; ortogonální regrese; lineární regresní model s podmínkami typu II; výběr proměnných

Počet stran: 98

Počet příloh: 0

Jazyk: anglický

BIBLIOGRAPHICAL IDENTIFICATION

Author: Mgr. Sandra Donevska

Title: Regression analysis for compositional data

Type of thesis: Dissertation thesis

Department: Department of Mathematical Analysis and Applications of Mathematics

Supervisor: Doc. RNDr. Eva Fišerová Ph.D.

The year of presentation: 2017

Abstract: This thesis is focused on regression analysis for compositional data. Relative nature of compositional data that distinguishes them from the standard multivariate data call for a special treatment. Since for the most of the statistical techniques there is still not developed stay-in the simplex approach, the log-ratio methodology presents a proper statistical approach that enable to express the data in a coordinate system.

Firstly, a regression model with compositional response variable is studied. A multivariate regression model is built for the compositional data expressed in orthonormal coordinates. The explicit formulas for the estimators of regression parameters and as well test statistics for the verification of their significance are provided. Further, new coordinate representation of the compositional data allowing to simplify the computation concerning regression parameters estimation and hypothesis testing is proposed and as well, the quality of prediction in different coordinate system is evaluated.

The second part of this thesis is devoted to the calibration problem for compositions. Here the calibration approach based on linear models with the type-II constraints is used. The equivalence between the linear model with type-II constraints and the total least squares regression is proved. A procedure for calibration of compositional measurements is proposed and tests for conformity of two measuring devices (methods) are presented.

In the last part of the thesis, a variable selection procedure for compositions that guarantees that a reduction of the original composition to a subcomposition causes only negligible change of the information is presented.

All theoretical results are applied to real-world examples.

Key words: compositional data; regression with compositional response; calibration; total least squares; linear model with type-II constraints; variable selection

Number of pages: 98

Number of appendices: 0

Language: English

Declaration of Authorship

I hereby declare that this dissertation thesis has been completed independently, under the supervision of doc. RNDr. Eva Fišerová Ph.D. All the materials and resources are cited with regard to the scientific ethics, copyrights and the laws protecting intellectual property. This thesis or its parts were not submitted to obtain any other or the same academic title.

In Olomouc, ...

signature

Contents

Acknowledgement

Firstly, I would like to express my sincere gratitude to my supervisor doc. RNDr. Eva Fišerová, Ph.D. for the continuous support of my Ph.D study and related research, for her patience, motivation, and immense knowledge. My sincere thanks also goes to doc. RNDr. Karel Hron, Ph.D.. Without his precious support it would not be possible to conduct this research. Last but not the least, I would like to thank my family for supporting me spiritually throughout writing this thesis.

Background & Goals of the Study

Background

Regression is a common statistical method for modelling and analysing the relationship between the response and predictor variable(s). In the frame of the parametric approach of the regression analysis, the linear relationship between the variables is only investigated.

The regression techniques discussed in this thesis are performed on a special kind of multivariate data known as compositional data, or compositions for short. The definition for D-part composition as quantitative descriptions of the parts of some whole, conveying relative information, dates from the 1986 and it is given by Aitchison. This strictly positive data that quite often sum up into an arbitrary constant, have the simplex $\mathcal{S}^D$ with the Aitchison geometry, to be their sample space. As it is well known, the simplex lacks the Euclidean vector space structure. It is the underlying reason that the standard statistical methods, like the regression analysis in our case, can not be applied directly on the compositions. Hence, this fact led to the necessity to translate the compositions from the simplex to the real space in order to apply standard statistical methods. The log-ratio methodology presents a proper statistical approach that enable to express the data isometrically in the real Euclidean space [4]. This approach permits one to release the fixed constant sum constraint and follow natural principles of compositions. These principles consist of the scale invariance, permutation invariance and subcompositional coherence [15, 66, 68]. The scale invariance provides the same results of a statistical analysis irrespective of particular representation of the positive vector whose parts carry relative contributions on a whole. Subcompositional coherence means that results for subcompositions are not in contradictions

with those for compositions. Permutation invariance means that reordering parts of a composition does not affect the included information.

Centered log-ratio (clr) coordinates represent historically the first isometric mapping between the Aitchison geometry and the real space endowed with the Euclidean geometry [4]. Clr coordinates are characterized by a zero sum of the variables and, consequently, by a singular covariance matrix. Often it is preferable to have orthonormal coordinates that avoid singularity of the covariance matrix, therefore the isometric log-ratio (ilr) transformation was proposed [21]. The composition after the ilr transformation results in a vector of orthonormal coordinates in the R^{D-1}. Naturally, there are inverse transformations back to simplex $\mathcal{S}^D$. Hence, the results can be interpreted either in the coordinates or on the simplex.

First meaningful studies about the regression models for compositions [1, 2] come from the 1980's together with the log-ratio approach. Here are behind the others, studied models where only the response random variables are compositional while the predictors are real fixed variables. Two types of regression models depending on the distribution of the residuals are compared, namely the logistic-normal and the Dirichlet regression models. The Dirichlet regression model [36, 38] has showed to have too strong restrictive independence properties. Additionally it is shown that the Dirichlet distribution [15, 68] can be approximated by some distributions in the logistic-normal family. Further, some tests about the significance of the model and its parameters are developed and discussed in [4].

Remarkable invention in the field of the regression analysis for compositional data came in the paper of J.J. Egozcue et al. (2012). The expression of the regression model in the orthonormal coordinates offers opportunity to use the least squares (LS) method for obtaining the estimates of the unknown regression parameters [18].

The LS problem is presented on both the simplex $\mathcal{S}^D$ and in coordinates on the R^{D-1}.

Important advantage of this approach is that one can solve independently $D-1$ least squares problems. In addition the results from these $D-1$ univariate regressions are the same as those from the multivariate approach. This result from the fact that the response variables are not correlated. Obtained univariate results are independent on the chosen orthonormal basis, i.e. the compositional parameters and residuals after inverse ilr transformation do not depend on the chosen orthonormal basis. Procedure for the LS estimation that is suggested there should be consistent to the principle of working on coordinates [66]. Likewise it appears in [24] that the sum of squares (total, explained and residual) and the coefficient of determination of the multivariate model of the orthonormal coordinates can be also partitioned: summing up the sum of squares and the coefficient of determination from the $D-1$ univariate regression models gives the same result as from the multivariate one.

In order to perform standard statistical inference like hypothesis testing, or constructing confidence and prediction intervals and bounds for the unknown regression parameters, etc., it is necessary to assume normal distribution of the residuals from the model in coordinates. The normal distribution on the R^{D-1} is equivalent to the normal or logistic normal distribution on the $\mathcal{S}^D$ [2, 56, 58]. Simultaneously, J.J. Egozcue et al. (2012) highlights the potentiality to perform regression tests that are based on marginal normality of the residuals. Also it is referred that the marginal tests depend on the chosen orthonormal basis on the simplex. However, one will have to confront with the problem of the multiple testing task. There is a need to develop more complex testing theory in the compositional regression model. One particular aim of this dissertation thesis is to study in depth the regression model with the compositional response. Namely, here we present the benefits of the theory of the multivariate linear models. As well we show on the different coordinate representation of such a multivariate model, and on opportunity to choose the one that will enhance the interpretation of the outcomes [see Section 4.2 for more].

Another type of the regression technique for compositions that is developed

is the one based on the total least squares. This technique is also known as the orthogonal regression, regression with errors in variables, or as a calibration problem. A model is established just for the three-part compositions after the ilr transformation and can be used for modelling the relationship between the parts of compositions. Primary contribution to this quite new regression technique for compositional data can be find in the papers [29, 31]. Authors there overcome the standard TLS by the linear regression models with the type II constraints [52]. An important requirement to build such a model is the assumption of independence and homoscedasticity of the orthonormal coordinates. Otherwise, when this is violated, then it is not satisfied the invariance of the results on the simplex under the orthogonal rotation of the orthonormal coordinates. Namely, when transforming the results of the analysis back on the simplex they will differ from these obtained in the ilr space. A real world example in [29] shows that the proposed method works for heteroscedastic orhtonormal coordinates too. The linear model approach is favourable for finite samples, unlike, the TLS which is an asymptotic approach. Moreover, the linear model approach enables to perform the standard statistical inference, being difficult or sometimes impossible in the frame of the TLS approach.

An iterative algorithm is proposed for the estimation of the calibration line [29], [see Section 2.2 for more]. Advantages of this iteration procedure is that it converges very quickly, and in addition, stable values of the estimates are achieved in the first few iterations. Problems with numerical stability of the proposed algorithm may occur if the angle between the calibration line and the axis represented by the first orthonormal coordinate tends to be 90°. Thus the calibration line is estimated.

Furthermore, it is checked the model and its assumptions for adequacy and validity for e.g. testing significance of calibration line's parameters, overall test of significance, verifying the quality of the estimated calibration line or deciding whether the additional observation can be explained by the estimated line [31] etc. Moreover, under the assumption of normality, there are suggested confidence

regions for the calibration line parameters, i.e., the confidence interval for each of the calibration line parameters, confidence ellipse for the vector of the calibration parameters, pointwise and simultaneous confidence bounds for the calibration line [31]. As well, there are constructed confidence ellipses for the location of the unknown errorless results of measurement [29]. However, all these results are done for 3-part compositions only.

In this thesis we will develop the calibration problem for D-part compositions based on the linear model with the type II - constraints. The calibration is a process whereby the scale of a measuring device or method is determined on a basis of an experiment. There are two stages in the calibration process. In the first stage, the calibration curve is specified. It describes a relationship between the quantity values with measurement uncertainties provided by a measurement standard (a measuring device or method with assigned correctness) and a calibrated one. The second stage concerns the prediction of values for measurement standard based on measurements by calibrated device. The values of measurement standard are considered either fixed (non-random), or random. In the former case we speak about controlled calibration, in the latter about random or natural calibration. For more details see, e.g., [14, 62]. In this thesis we focus on determination of a calibration line with random values of the standard.

It is quite surprising that the calibration problem is not reflected (except for some introductory issues [29, 31]) in literature on compositional data analysis. Different approach to compositional calibration is proposed in [80]. Here, the calibration model is fitted by means of the generalized linear model with the multi-Poisson distribution.

Very recent study in [44, 45] shows generalization of the TLS problem on modeling the linear relationship between all D compositional parts. Again here authors suggest to consider the special choice of the orthonormal coordinates [30, 44, 45] which improves the interpretation of the model and the statistical inference. The unknown regression parameters of the TLS problem are estimated by means of the singular value decomposition (SVD) or the principal component

analysis (PCA). The comparison of the estimates resulting from the both techniques is done as well. Also robust modification is offered of the PCA estimation technique, i.e. MM-estimators [35] of the location and shape. This technique is emphasized under the robust version of the SVD [13] mainly because it is simpler and less computationally complex. For performing statistical inferences the authors suggest, to use bootstrap methods.

Through the history of regression modelling of the compositions also other types of models are improved. Multiple regression models with the real response variable and predictors that are compositional parts are studied firstly in sense of the experiments with mixtures [73, 74]. The model in this point describes the response as a linear or canonical polynomial function of the mixture variables, related with the compositional parts. Problems in such models are connected with the constant sum constraint of the compositions which reflects in singularity. Attempt to overcome this difficulty is done with involving higher order polynomial regression function than the primary considered linear regression function. Estimates of the unknown regression parameters are obtained using the LS method. However this method usually fails, that is affected by the approximate linear dependence of the design matrix which leads to necessity to address for, e.g. the ridge regression [40]. Another trouble being caused by the constant sum constraint is the interpretation of the estimated parameters from the mixture model: it is impossible to alter one proportion without altering at least one of the other proportions. Aitchison contribute to the field of the mixture models for compositional data by implementation of the log-contrast, being the linear combination of the logarithm of the parts and the regression parameters (except for the intercept), with the requirement on the regression parameters to sum up into zero [3, 4]. Worth mentioning circumstance is if one of the regression parameters in the log-contrast model is equal to 1 and some other equal to -1, while the rest of them are zero, than the model considers the log-ratio of the two parts associated with the parameters being equal to 1 and -1. Regrettably, mentioned unwanted consequences of the analysis are not completely removed.

Little bit later, further approval in this field is done by Hron et al. (2012) where the model with compositional predictors and non-compositional response is developed. The authors propose to use the orthonormal coordinates as predictors, and consequently the LS for estimation of the unknown regression parameters.

The problem with interpretation of the resulting parameters can be solved if reasonable orthonormal basis for the ilr transformation is considered [30].

Special orthonormal coordinates in this case can be constructed on a such a way that are directly connected with certain compositional part. Accordingly the first orthonormal coordinate explains all the relative information about the first part of the composition. The interpretation of the remaining $D-2$ orthonormal coordinates is definitely not straightforward, because they are not directly connected to just one compositional part. Thus just the regression parameters connected with the first predicor or the first orthonormal coordinate is crucial. So, in order to interpret how the remaining $D-1$ parts contribute to the response variable $D-1$ number of regression models are obtained. Let us shortly describe how other $D-1$ systems of orthonormal coordinates have to be construct [see Section 1.2 for more]. On the beginning one needs to permute the parts of the composition on a such a way that the first compositional part will be replaced with the second compositional part. In this way permuted composition is ilr transformed with the respect to another orthonormal basis, that is just an orthogonal rotation of the previous orthonormal basis [21]. Once more we have a model where only the first predictor is of concern, which this time carries all the relative information about the second part. Briefly said, later in every further step firstly, the permuted composition is created, such that the part on the first place is replaced with the next one that remains to be interpreted. Again it is ilr transformed with respect to such an orthonormal basis that will magnify the interpretability of the desired part of the composition. The resulting orthonormal coordinates are being predictors for the new model. Additionally, standard statistical inference like the t-test about the significance of the regression parameters and the overall F-test can be done. The invariance of this test under the

choice of the orthonormal basis is also discussed in [41]. One of the latest in the area of regression analysis of the compositional data concerns the developments of the models with compositional random response and compositional non-random predictor variables [86] directly on the simplex $\mathcal{S}^D$. The algorithmic procedure consisting of 4 steps is presented for building a linear regression model. Starting point in the algorithm is to formulate a regression model in the Aitchison geometry. One step concerns the application of the LS method on the simplex that gives the estimates of the regression parameters. Another step is about evaluating the regression model, where the observed squared correlation is taken as a measure of goodness of fit. Goodness of prediction of the model is evaluated by the cross-validated squared correlation coefficients based on the leave-one-out method [77].

Furthermore, it is proved that results on the $\mathcal{S}^D$ are equivalent with those on the R^{D-1}.

Statistical models for discrete compositions are firstly disccused in [2] and first full analysis of such models can be find in [12]. Resent study of models for discrete compostions by Bacon-Shone (2008) concerns overcoming of the restriction of the total sum of the parts. An overview of log-linear models approach for contingency tables can be find in [25].

Goals of the Study

Regression analysis for the compositional data started to expand in the early 80's. Great progress in this field has been done till now, but there are still some topics that deserve special attention. Four types of regression models, depending on the type of the response and predictors variables can be distinguished: a regression model with compositional response and non-compositional predictors, a regression model with non-compositional response and compositional predictors, a regression between parts of compositions, and a regression model with compositional response and predictor variables.

The motivation for writing this thesis lies in satisfying the current needs for further development in regression analysis for compositional data. Because the branch is quite wide, the thesis is mainly focused on regression models with compositional response, the calibration problem for compositions, and the simplification of regression models with compositional data in terms of reducing dimensionality of the compositions. The calibration problem is related to a regression between parts and the TLS problem. The particular goals of the thesis are the following:

- Formulate a multivariate regression model with a compositional response and find explicit formulas for the estimators of the regression parameters and proper test statistics. Find coordinate representation of compositional data allowing to simplify the computation concerning regression parameters estimation and hypothesis testing. Evaluate the quality of prediction in different coordinate systems.

- Prove the equivalence between the TLS approach and the linear model

with the type-II constraints. Propose a procedure for calibration of compositional measurements and suggest tests for conformity of two measuring devices (methods).

- Propose a variable selection procedure for compositions that guarantees that a reduction of the original composition to a subcomposition causes only negligible change of the information.

- Apply theoretical results to real-world examples.

This dissertation thesis is based on the following published papers:

- **Donevska S.**, Fišerová E., Hron K. (2011). On the equivalence between orthogonal regression and linear model with type-II constraints. *Acta Universitatis Palackianae Olomucensis. Facultas Rerum Naturalium. Mathematica* **50** (2), 19–27.

- Hron K., Filzmoser P., **Donevska S.**, Fišerová E. (2013). Covariance-based variable selection for compositional data. *Mathematical Geosciences* **45** (4), 487–498.

- **Donevska S.**, Fišerová E., Hron K. (2016). Calibration of compositional measurements. *Communications in Statistics - Theory and Methods* **45** (22), 6773–6788.

- Fišerová E., **Donevska S.**, Hron K., Bábek O., Váňkátová K. (2016). Practical aspects of log-ratio coordinate representations in regression with compositional Response. *Measurement Science Review* **16** (5), 235–243.

Statistical software that was used to preform the simulation study and as well, for demonstrating the theoretical considerations on real world examples, is the free available software R [70]. The known compositions packages `robCompositions` [66], `compositions` [81], and as well, packages `mvoutlier` [28], `StatDA` [27] were used.

Part I

Research methods

1. Compositional data

Compositional data are commonly represented in percentages or proportions. These representations induce a constant sum constraint on the compositional parts, i.e., 100 in the case of percentages and 1 when proportions are used. As will be shown later, any representation of compositions does not alter information carried by the data. These data can be find in many scientific area, e.g. in biochemistry when it is examined a chemical composition of blood plasma, urine and renal calculi [4, 20], in geochemistry when it is studied the composition of the sedimentary rocks [43] or the composiiton of reservoir sediments [33], etc.

Statistical society was discussing for a long time whether it is appropriate to perform standard statistical analysis on R^D on the data carrying only relative information. Attention was divert when Pearson identified the problem of spurious correlation [65, 69]. It is about the correlation of the ratios of the variables, he showed that if y_1, y_2 and y_3 are uncorrelated , than $\frac{y_1}{y_3}$ and $\frac{y_2}{y_3}$ are correlated.

Effort was done through the years to overcome the problem of the constrained data when finally Aitchison in 1980's revolutionize the field of statistics with the statistical methodology respecting the nature of the compositions. He has point out on the importance of the relative nature of the compositions, being data that contain the only relevant information in the ratios between their parts [4, 65, 68]. Moreover, he has developed the theory of the simplex being a vector space convenient for the compositions, this will be disscused in the following section. Simultaneously, he gave the idea of the log-ratio transformations that enable usage of standard multivariate statistical methods on the real space [4]. The last section in this part presents the tools for exploratory analysis for compositions.

1.1 Aitchison geometry

Compositional data [4] are strictly positive multivariate observations that carry only relative information.
Compositions, denoted as $\mathbf{y} = (y_1, y_2, \ldots, y_D)'$, have their own sample space the simplex $\mathcal{S}^D$ defined as

$$\mathcal{S}^D = \left\{ \mathbf{y} = (y_1, y_2, \ldots, y_D)' | y_i > 0, i = 1, 2, \ldots, D; \sum_{i=1}^{D} y_i = k \right\}.$$

Crucial in this framework is the operation of closure for $\mathbf{y} = (y_1, y_2, \ldots, y_D)' \in \mathbb{R}_+^D$, given by

$$\mathcal{C}(\mathbf{y}) = \left(\frac{ky_1}{\sum_{i=1}^{D} y_i}, \frac{ky_2}{\sum_{i=1}^{D} y_i}, ..., \frac{ky_D}{\sum_{i=1}^{D} y_i} \right)',$$

with which we can express the compositions as a non-negative vectors summing up into an arbitrary constant $k > 0$. Basically, information contained in the composition remains same it is just matter of change of the units. Such compositions are compositionally equivalent, hence it is does not depend on the choice of k.

The vector space structure of the simplex $\mathcal{S}^D$ is obtained with the following two operations defined on it [15, 66]:

- perturbation of $\mathbf{y} \in \mathcal{S}^D$ by $\mathbf{w} \in \mathcal{S}^D$, analogous to addition in the real space:

$$\mathbf{y} \oplus \mathbf{w} = \mathcal{C}\left(y_1 w_1, y_2 w_2, ..., y_D w_D\right)',$$

- power transformation or powering of $\mathbf{y} \in \mathcal{S}^D$ by a constant $\alpha \in R$, analogous to scalar multiplication in the real space:

$$\alpha \odot \mathbf{y} = \mathcal{C}\left(y_1^\alpha, y_2^\alpha, ..., y_D^\alpha\right)'.$$

Further it is desired to work with the compositions on the simplex on comparable way as we do we the standard multivariate data on the real space. We demand to compute the length of a composition, to determine angles between compositional vectors or to find the distance between them. To obtain the Euclidean vector space, the inner product, norm and distance are defined in the following way:

- the Aitchison inner product of compositions $\mathbf{y}, \mathbf{w} \in \mathcal{S}^D$,

$$\langle \mathbf{y}, \mathbf{w} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{y_i}{y_j} \ln \frac{w_i}{w_j},$$

- the Aitchison norm of $\mathbf{x} \in \mathcal{S}^D$

$$\|\mathbf{y}\|_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{y_i}{y_j} \right)^2} = \sqrt{\langle \mathbf{y}, \mathbf{y} \rangle_a},$$

- the Aitchison distance between $\mathbf{w}$ and $\mathbf{y} \in \mathcal{S}^D$

$$d_a(\mathbf{y}, \mathbf{w}) = \|\mathbf{y} \ominus \mathbf{w}\|_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{y_i}{y_j} - \frac{w_i}{w_j} \right)^2}. \tag{1.1}$$

1.2 Coordinate representation

The primary source of information in compositions is contained in (log-)ratio between parts. Therefore, the representation by the log-ratio of compositional parts seems to be convenient. As already mentioned there are certain log-ratio transformations which translates the composition from the simplex into a coordinate vector on the real space.

Firstly the centred log-ratio (clr) transformation was invented, which is mapping between the simplex $\mathcal{S}^D$ and the Euclidean space $\mathbb{R}^D$, defined by,

$$\mathrm{clr}(\mathbf{y}) = \left(\ln \frac{y_1}{\mathrm{g}(\mathbf{y})}, \ln \frac{y_2}{\mathrm{g}(\mathbf{y})}, \ldots, \ln \frac{y_D}{\mathrm{g}(\mathbf{y})} \right)' = \mathbf{h}, \ \mathbf{y} \in \mathcal{S}^D, \ \mathbf{h} \in \mathbb{R}^D. \tag{1.2}$$

where $\mathrm{g}(\mathbf{y}) = \sqrt[D]{\prod_{j=1}^{D} y_j}$ is the geometric mean of the parts of the composition.

Obviously, there is a possibility of inverse clr transformation that back transforms the real vector $\mathbf{h} \in \mathbb{R}^D$ into a composition $\mathbf{y} \in \mathcal{S}^D$ on the simplex, that is reached by,

$$\mathrm{clr}^{-1}(\mathbf{y}) = \mathcal{C}(\exp(h_1), \exp(h_2), \ldots, \exp(h_D))'.$$

Clr transformation actually, expresses the composition $\mathbf{y} \in \mathcal{S}^D$ in coordinates with respect to the generating system on the simplex $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_D\}$, where

$$\mathbf{v}_j = \mathcal{C}(\exp(\mathbf{e}_j))' = \mathcal{C}(1, 1, \ldots, e, \ldots, 1)', \ j = 1, 2, \ldots, D.$$

Consequently, a composition $\mathbf{y} \in \mathcal{S}^D$ relying on the generating system $\{\mathbf{v}_1, \ldots, \mathbf{v}_D\}$ is expressed in terms of [4, 15, 66]

$$\begin{aligned}\mathbf{y} &= \bigoplus_{j=1}^{D} h_j \odot \mathbf{v}_j = \bigoplus_{j=1}^{D} \frac{y_j}{\mathrm{g}(\mathbf{y})} \odot \mathbf{v}_j = \\ &= \frac{y_1}{\mathrm{g}(\mathbf{y})} \odot (e, 1, \ldots, 1)' \oplus \cdots \oplus \frac{y_D}{\mathrm{g}(\mathbf{y})} \odot (1, 1, \ldots, e)'.\end{aligned}$$

The geometric mean in the denominator can be substituted by any constant, which is possible because of the compositional equivalence.

The isometric property makes this transformation applicable for techniques based on distances. This property also reflects the straightforward interpretation of the clr transformed composition. Unfortunately, one disadvantage property of this transformation that comes from the symmetry of the components of the vector of the clr coordinates is that leads to singular covariance matrix which causes computational issues. Another disadvantage property that the clr transformation dispose is that the clr coefficients do not satisfy the principle of subcompositional coherence. This principle is of crucial importance whose meaning is that the information carried in the composition should not be contradictory with the one carried in the subcomposition. Every method before applied to the compositional

data should meet this requirement. Here the geometric mean of a subcompostion does not necessary have to be the same with the one we have for the full D-part composition.

Despite of the disadvantage properties, the clr coordinates are still frequently used because of an intuitive interpretation. For example, the compositional biplot of the clr coordinates [7] can be constructed, that is an important visualization tool for investigation of the compositional data structure. Here, the single clr coordinates are usually interpreted in terms of the original compositional parts [26, 79].

To avoid disadvantages of the clr coordinates, orthonormal coordinates with respect to an orthonormal basis on the simplex were proposed [21]. The transformation is called as the isometric log-ratio (ilr) transformation.

There exist many ways to obtain an orthonormal basis of the simplex. Unfortunately, there is no canonical basis on the simplex, where by the interpretation of the orthonormal coordinates is not that straightforward. The choice of the method for construction of the basis may improve the interpretation of the resulting coordinates. Behind the commonly used methods belong the Gram - Schmidt procedure [21] and the very intuitive - sequential binary partition (SBP) [22]. The resulting coordinates coming from the SBP, called balances, give interpretation in sense of grouped parts of the composition. In each of the $D-1$ consecutive steps of the SBP, partitioning of the parts into two non-overlapping, distinguished groups is done. Groups of compositional parts are formed according to expert knowledge, or can be formed blindly, without any preliminary knowledge about the grouping of the parts.

Often used orthonormal basis leads to the $(D-1) \times D$ matrix $\mathbf{V}$, such that $\mathbf{V}\mathbf{V}' = \mathbb{I}_{(D-1)}$, with the rows vectors [33]

$$\mathbf{v}_i = \sqrt{\frac{D-i}{D-i+1}} \left(0, \ldots, 0, 1, -\frac{1}{D-i}, \ldots, -\frac{1}{D-i}\right), \; i = 1, 2, \ldots, D. \quad (1.3)$$

This basis relates with the orthonormal coordinates [30],

$$\operatorname{ilr}(\mathbf{y})_i = \mathrm{z}_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{y_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} y_j}}, i = 1, 2, \ldots, D-1. \tag{1.4}$$

There exist unique relationship between the ilr and the clr coordindates [66], given by

$$\mathbf{z} = \mathbf{h}\mathbf{V}', \tag{1.5}$$

where $\mathbf{h} \in \mathbb{R}^D$ is the clr transformed composition $\mathbf{y} \in \mathcal{S}^D$. Moreover, for the first coordinates of both systems it holds [66],

$$h_1 = \sqrt{\frac{D-1}{D}} z_1.$$

In this case, the first orthonormal coordinate z_1 explains all the relative information about the first compositional part y_1 within the first given composition [30]. Unfortunately, the remaining orthonormal coordinates do not have such straightforward interpretation.

In order to obtain the interpretation for the remaining orthonormal coordinates, we just need to make permutation of the compositional parts,

$$\begin{aligned} \mathbf{y}^{(l)} &= (y_l, y_1, \ldots, y_{l-1}, y_{l+1}, \ldots, y_D) = \\ &= \left(y_1^{(l)}, y_2^{(l)}, \ldots, y_{l-1}^{(l)}, y_{l+1}^{(l)}, \ldots, y_D^{(l)}\right), \ l = 1, 2, \ldots D, \end{aligned}$$

and subsequently apply the formula in (1.4) to the permuted compositions $\mathbf{y}^{(l)}$, $l = 1, 2, \ldots, D$. Thus the first orthonormal coordinate obtained for permuted composition $y^{(l)}, l = 1, 2, \ldots, D$, contains all the relative information about the l-th compositional part y_l, $l = 1, 2, \ldots, D$ and, consequently

$$h_l = \sqrt{\frac{D-1}{D}} z_1^{(l)}, \ l = 1, 2, \ldots, D. \tag{1.6}$$

Behind all, it is also possible to back transform the ilr transformed composition, by the following formulas [29],

$$y_1^{(l)} = \exp\left(\sqrt{\frac{D-1}{D}} z_1^{(l)}\right),$$

$$y_i^{(l)} = \exp\left(-\sum_{j=1}^{i-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} z_j^{(l)} + \sqrt{\frac{D-i}{D-i+1}} z_i^{(l)}\right),\ i = 2, \ldots, D-1,$$

$$y_D^{(l)} = \exp\left(-\sum_{j=1}^{D-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} z_j^{(l)}\right).$$

At the conclusion of this section it remains to be emphasized that the log-ratio approach for the compositions is of fundamental importance since without it the mass of the well known statistical techniques could not be applied on the compositions.

1.3 Exploratory analysis

Standard descriptive statistics like the arithmetic mean, the variance or the standard deviation are not compatible with the Aitchison geometry for compositions, and therefore, should be substituted by another proper descriptive statistics for compositions.

The basic measure of variability of a random composition $\mathbf{y} \in \mathcal{S}^D$ is the variation matrix [4], defined as

$$\mathbf{T} = \left\{ \mathrm{var}\left(\ln \frac{y_i}{y_j}\right) \right\}_{i,j=1}^{D}. \tag{1.7}$$

The elements of the variation matrix describe the variability of the random log-ratio $\ln \frac{y_i}{y_j}$: the smaller the value of this variance, the more the log-ratio tends to be a constant. The (normed) sum of the elements of the variation matrix is called total variance,

$$\mathrm{totvar}(\mathbf{y}) = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \mathrm{var}\left(\ln \frac{y_i}{y_j}\right), \tag{1.8}$$

expressing the total variability of the compositional data set. Note that

$$\text{totvar}(\mathbf{y}) = \sum_{i=1}^{D} \text{var}(h_i) = \sum_{i=1}^{D-1} \text{var}(z_i^{(l)}),\ l = 1, 2, \ldots, D, \tag{1.9}$$

i.e. the total variance can also be computed using the variability of the clr coordinates or the orthonormal coordinates, respectively [63].

Further, what worth to be mentioned for the purposes of the thesis is the compositional variation array, defined as the simplest and minimum way of summarizing the patterns of location and variability within a compositional data set [5],

$$\mathbf{V} = \begin{pmatrix} 0 & \text{var}\left(\ln \frac{y_1}{y_2}\right) & \text{var}\left(\ln \frac{y_1}{y_3}\right) & \cdots & \text{var}\left(\ln \frac{y_1}{y_D}\right) \\ \text{E}\left(\ln \frac{y_2}{y_1}\right) & 0 & \text{var}\left(\ln \frac{y_2}{y_3}\right) & \cdots & \text{var}\left(\ln \frac{y_2}{y_D}\right) \\ \cdots & \cdots & \cdots & \cdot & \cdots \\ \text{E}\left(\ln \frac{y_D}{y_1}\right) & \text{E}\left(\ln \frac{y_D}{y_2}\right) & \text{E}\left(\ln \frac{y_D}{y_3}\right) & \cdots & 0 \end{pmatrix},$$

where in the upper triangle of the array the log-ratio variances and in the lower triangle the log-ratio expectations are displayed. Moreover, if basic properties of logarithm as well as those of expectation and variance of log-ratio are taken into account, the following properties hold. It is sufficient to know variances of log-ratios $\ln(y_i/y_j)$, $i < j$, because $\text{var}\left(\ln \frac{y_i}{y_j}\right) = \text{var}\left(-\ln \frac{y_j}{y_i}\right)$, hence the covariance structure of a D-part composition is entirely determined by the $\frac{D(D-1)}{2}$ log-ratio variances. For the log-ratio expectations the triangular equality holds, $\text{E}\left(\ln \frac{y_i}{y_m}\right) + \text{E}\left(\ln \frac{y_m}{y_j}\right) = \text{E}\left(\ln \frac{y_i}{y_j}\right)$.

Finally, it is possible to express the covariance structure of balances using linear combinations of variances of log-ratios [30]. Taking the balances $z_i = z_i^{(1)}$ given in (1.4) we obtain [30]

$$\text{var}(z_i) = \frac{1}{D-i+1} \sum_{p=i+1}^{D} \text{var}\left(\ln \frac{y_i}{y_p}\right) - \frac{1}{2(D-i)(D-i+1)} \sum_{p=i+1}^{D} \sum_{q=i+1}^{D} \text{var}\left(\ln \frac{y_p}{y_q}\right). \tag{1.10}$$

2. Regression models

Two regression techniques for compositions are deeply examined in this thesis: the regression with the compositional response and the calibration problem for compositions. For each of these techniques, an overview of known results is presented in the following sections. Firstly, some basics of the multivariate linear regression models together with statistical inference on regression parameters are recalled. The second section in this chapter is devoted to the total least squares (TLS) problem for compositions. Here two possible approaches for the estimation are presented, namely the maximum likelihood approach and the linear model approach.

2.1 Multivariate regression with non-compositional response

In this section we would like to point out that only here $\mathbf{y}$ will not denote a composition but it will stand for a standard random vector. This notation is used in order to follow the usual one in the statistical literature.

A multivariate regression model presents a regression model where multiple response variables appear simultaneously. Consider we have q random variables $y_1, y_2, \ldots, y_q$ and for each of these we have n observations. Let us denote by $\mathbf{y}_j = (y_{1j}, y_{2j}, \ldots, y_{nj})'$, $j = 1, 2, \ldots, q$, the observation vector that corresponds to the random variable y_j. For every vector $\mathbf{y}_j$ we assume the following linear model [47, 54]

$$\mathbf{y}_j = \mathbf{X}\mathbf{b}_j + \boldsymbol{\varepsilon}_j, \ j = 1, 2, \ldots, q, \tag{2.1}$$

and, simultaneously, for all vectors $\mathbf{y}_j$ we assume the multivariate linear model

$$\underline{\mathbf{y}} = \mathbf{XB} + \underline{\boldsymbol{\varepsilon}}, \tag{2.2}$$

where $\underline{\mathbf{y}} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_q)$ is the $(n \times q)$ dimensional matrix of response vectors, $\mathbf{X}$ is the $(n \times k)$ dimensional design matrix which has full column rank, $\boldsymbol{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_q)$ is the $(k \times q)$ dimensional matrix of the unknown regression parameters, $\mathbf{b}_j = (b_{1j}, b_{2j}, \ldots, b_{kj})'$, $j = 1, 2, \ldots, q$ and $\underline{\boldsymbol{\varepsilon}} = (\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_q)$ is the $(n \times q)$ dimensional matrix of the random errors. Further, let us assume that the multivariate responses $\mathbf{y}_{i\cdot} = (y_{i1}, y_{i2}, \ldots, y_{iq})'$, $i = 1, 2, \ldots, n$, are independent with the same unknown variance-covariance matrix $\boldsymbol{\Sigma}$, i.e.

$$\begin{aligned} \operatorname{cov}(\mathbf{y}_{i\cdot}, \mathbf{y}_{j\cdot}) &= \mathbf{0}, \ i \neq j, \\ \operatorname{var}(\mathbf{y}_{i\cdot}) &= \boldsymbol{\Sigma}, \ i = 1, 2, \ldots, n. \end{aligned}$$

In order to derive the estimator of $\mathbf{B}$, to construct confidence intervals and confidence regions for unknown regression parameters, or to do some tests for significance of the regression coefficient, etc., the model (2.2) can be rewritten in the following vectorized form [54]

$$\operatorname{vec}\left(\underline{\mathbf{y}}\right) = (\mathbf{I}_q \otimes \mathbf{X}) \operatorname{vec}(\mathbf{B}) + \operatorname{vec}(\boldsymbol{\varepsilon}), \ \operatorname{var}[\operatorname{vec}(\underline{\mathbf{y}})] = \boldsymbol{\Sigma} \otimes \mathbf{I}_n,$$

where $\operatorname{vec}(\underline{\mathbf{y}}) = (\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_q')'$ and the symbol $\otimes$ denotes the Kronecker product. Thus, the least squares estimator of $\operatorname{vec}(\mathbf{B})$ is obtained by minimizing the square of the Mahalanobis distance of the residuals [54]

$$\arg \min_{\operatorname{vec}(B)} \left\| \operatorname{vec}\left(\underline{\mathbf{y}}\right) - (\mathbf{I}_q \otimes \mathbf{X}) \operatorname{vec}(\boldsymbol{B}) \right\|^2_{(\Sigma \otimes \mathbf{I}_n)}. \tag{2.3}$$

The solution of the minimization problem (2.3), after de-vectorization, is

$$\widehat{\boldsymbol{B}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \underline{\mathbf{y}}. \tag{2.4}$$

The estimator $\widehat{\boldsymbol{B}}$ is the best linear unbiased estimator (BLUE) of the parameter matrix $\boldsymbol{B}$ [54].

One can notice that this estimator is invariant with respect to the variance-covariance matrix of vec $(\underline{\mathbf{y}})$.

However, the variance-covariance matrix of the vector $\text{vec}(\widehat{\mathbf{B}}) = (\widehat{\mathbf{b}}_1', \widehat{\mathbf{b}}_2', \ldots, \widehat{\mathbf{b}}_q')'$

$$\text{var}\left[\text{vec}(\widehat{\mathbf{B}})\right] = \boldsymbol{\Sigma} \otimes (\mathbf{X}'\mathbf{X})^{-1} \tag{2.5}$$

depends on $\boldsymbol{\Sigma}$. Since the variance-covariance matrix $\boldsymbol{\Sigma}$ is unknown, it is necessary to estimate it. The unbiased estimator of $\boldsymbol{\Sigma}$ is given by [54] $\widehat{\boldsymbol{\Sigma}} = \underline{\mathbf{y}}'\mathbf{M}_X\underline{\mathbf{y}}/(n-k)$, where $\mathbf{M}_X = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is a projector on the orthogonal complement of the vector space $\mathcal{M}(\mathbf{X})$ generated by the columns of the design matrix $\mathbf{X}$, i.e. $\mathcal{M}(\mathbf{X}) = \{\mathbf{X}\mathbf{u} : \mathbf{u} \in \mathbb{R}^k\}$. Under normality, the estimators $\widehat{\mathbf{B}}$ and $\widehat{\boldsymbol{\Sigma}}$ are statistically independent. Moreover, if $n-k > q$, then $(n-k)\widehat{\boldsymbol{\Sigma}}$ has the Wishart distribution $W_q[n-k, \boldsymbol{\Sigma}]$.

Let us note that the univariate approach (2.1) leads to the same estimators of the regression parameters $\mathbf{b}_j$ and of the variances $\sigma^{jj} = \{\boldsymbol{\Sigma}\}_{jj}$, $j = 1, 2, \ldots, q$. Specifically,

$$\widehat{\mathbf{b}}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_i,$$

$$\text{var}(\widehat{\mathbf{b}}_i) = \sigma^{ii}(\mathbf{X}'\mathbf{X})^{-1},$$

$$\widehat{\sigma}^{ii} = \mathbf{y}_i'\mathbf{M}_X\mathbf{y}_i/(n-k).$$

The theory of multivariate linear regression models [54] provide a range of tests, that are easy to compute due to explicit formulas. Usually three basic issues of hypotheses testing in a multivariate regression context are considered: significance of covariates for the responses $\mathbf{y}_j$, $j = 1, 2, \ldots, q$, point wise and simultaneously, and verification that the predictor $\mathbf{x}_i$, $i = 1, 2, \ldots, k$, contributes to the explanation of the overall variability in $\underline{\mathbf{y}}$.

It is easy to see that hypothesis testing on single regression parameters as well as on the whole vector parameter $\mathbf{b}_j$, $j = 1, 2, \ldots, q$, that conveys contributions of all covariates to the j-th response simultaneously can be performed within univariate multiple regressions using standard $t-$ and $F-$ test statistics,

respectively. Particularly, the test statistics for the null hypothesis $\mathbf{b}_j = \mathbf{0}$ can be expressed as

$$F_j = \frac{(n-k)\,\widehat{\mathbf{b}}_j'\mathbf{X}'\mathbf{X}\widehat{\mathbf{b}}_j}{k\widehat{\sigma}^{jj}}, \tag{2.6}$$

which has F-distribution with k and $n-k$ degrees of freedom under the null hypothesis.

The case of significance testing of the i-th predictor, $i = 1, 2, \ldots, k$, requires already the multivariate setting. Symbolically, the null hypothesis about the i-th predictor is expressed as $H_{0i} : \mathbf{B}_{i.} = (\mathbf{b}_{i1}, \mathbf{b}_{i2}, \ldots, \mathbf{b}_{iq}) = \mathbf{0}$. The corresponding test statistic is given by

$$F_i = \frac{(n-q-k+1)\,\widehat{\mathbf{B}}_{i.}\left(\underline{\mathbf{y}}'\mathbf{M}_X\underline{\mathbf{y}}\right)^{-1}\widehat{\mathbf{B}}_{i.}'}{q\left\{(\mathbf{X}'\mathbf{X})^{-1}\right\}_{ii}}, \tag{2.7}$$

which is distributed as $F_{q,n-q-k+1}$ under the null hypothesis H_{0i}.

Finally, in some cases even significance of the whole matrix of regression parameters $\mathbf{B}$, or a more general hypothesis $H_0 : \mathbf{AB} = \mathbf{C}$, where $\mathbf{A}$ is a $r \times k$ hypothesis matrix having full-row rank $r \leq k$, and $\mathbf{C}$ is a $r \times q$ matrix, are of interest. For this purpose a battery of tests is available, mostly used tests are Pillai-Bartlett trace, Wilks's Lambda, Hotelling-Lawley trace and Roy's largest root [47]. All of them are based directly or indirectly on $p = \min(r, q)$ non-zero eigenvalues λ_j of the product matrix $\mathbf{HE}^{-1}$, where $\mathbf{H}$ is the matrix for the hypothesis sums of squares and cross products, and $\mathbf{E}$ is the residual matrix, i.e.

$$\mathbf{E} = (\underline{\mathbf{y}} - \mathbf{X}\widehat{\mathbf{B}})'(\underline{\mathbf{y}} - \mathbf{X}\widehat{\mathbf{B}}) \tag{2.8}$$

$$\mathbf{H} = (\mathbf{A}\widehat{\mathbf{B}} - \mathbf{C})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\widehat{\mathbf{B}} - \mathbf{C}). \tag{2.9}$$

The multivariate models enable to describe more complex designs, thus concerning the association between the outcomes. Definitely, they are more efficient tool for modelling convoluted designs than the univariate ones.

Further, testing in the multivariate models avoid problems with the multiple testing. The tests for the univariate models are not simultaneous tests for all the

regressions and they do not consider the influence of the correlations among the responses, which can result in less powerful tests. Consequently, the univariate tests cannot evaluate joint influence on all outcomes. Among the difficulties when one uses the multivariate linear model approach is the necessity of disposing with large number of observations and complex interpretation of the results.

2.2 Regression between parts of 3-part compositions

This section is focused on modelling the relationship within the composition, i.e. between the compositional parts. The fundamental concept of the total least squares (TLS) problem for 3-part compositions, proposed in [29, 31], is summarized in the following.

Here the TLS procedure will be applied on the ilr transformed 3-part compositions.

The TLS in the simplest form attempts to fit a line that explains the set of n two-dimensional data points in such way that the sum of the orthogonal squares distances from the data points to the estimated line is minimal. The 3-part compositional data can be expressed as an $(n \times 2)$-dimensional data matrix $(\mathbf{z}_1, \mathbf{z}_2)$ of the corresponding orthonormal coordinates. The orthonormal coordinates follow the model [19],

$$z_{1i} = \mu_i + \varepsilon_{1i},\ z_{2i} = \nu_i + \varepsilon_{2i},\ i = 1, 2, \cdots, n, \tag{2.10}$$

where μ_i and ν_i are the unobserved errorless recordings (true values) of z_{1i} and z_{2i}, $i = 1, 2, \ldots, n$, respectively, ε_{1i} and ε_{2i} are independent random errors with zero mean and with variance equal to $\text{var}(\varepsilon_{1i}) = \sigma_1^2$ and $\text{var}(\varepsilon_{2i}) = \sigma_2^2$. In this thesis we take $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Moreover, we assume that the errorless recordings satisfy

$$\nu_i = a + b\mu_i,\ i = 1, 2, \ldots, n, \tag{2.11}$$

where a is the unknown intercept and b is the unknown slope of the orthogonal

regression line.

The TLS problem represents the following minimization problem

$$\min_{a,b} \frac{\sum_{i=1}^{n}(a + bz_{1i} - z_{2i})^2}{b^2 + 1}.$$

Standard calculus gives the minimum and we find estimators [16],

$$\widehat{a} = \overline{z}_2 - \widehat{b}\overline{z}_1, \tag{2.12}$$

$$\widehat{b} = \frac{s_{z_2}^2 - s_{z_1}^2 + \sqrt{(s_{z_2}^2 - s_{z_1}^2)^2 + 4s_{z_1,z_2}^2}}{2s_{z_1,z_2}}, \tag{2.13}$$

where $\overline{z}_1$, $\overline{z}_2$ are sample means, $s_{z_1}^2$, $s_{z_2}^2$ are sample variances, and s_{z_1,z_2} is a sample covariance. Consequently, under the normality, the maximum likelihood method gives the same estimators [8, 16, 50].

We have to point out that the maximum likelihood estimators and TLS solution are the same only in a considered special case, i.e., z_{1i} and z_{2i} are independent normally distributed random variables with the same variance. If the variances σ_1^2 and σ_2^2 are different such that $\sigma_1^2 = \lambda\sigma_2^2$, where $\lambda > 0$ is known, the maximum likelihood estimators are given by the expressions [16],

$$\widehat{a} = \overline{z}_2 - \widehat{b}\overline{z}_1,$$

$$\widehat{b} = \frac{\lambda s_{z_2}^2 - s_{z_1}^2 + \sqrt{(\lambda s_{z_2}^2 - s_{z_1}^2)^2 + 4\lambda s_{z_2z_1}^2}}{2\lambda s_{z_2z_1}}.$$

It is shown in [50] that the estimators (2.12) and (2.13) are weakly consistent. Conditions for strong consistency can be found in [83, 84]. General results on consistency see, e.g., in [16, 34, 50].

In addition, if we consider that the variance σ^2 is unknown, then its maximum likelihood estimator results in [8, 50]

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left[\left(z_{2i} - \widehat{a} - \widehat{b}\widehat{\mu}_i\right)^2 + (z_{1i} - \widehat{\mu}_i)^2\right]}{2n},$$

where the estimator $\widehat{\mu}_i$ of the true value μ_i, $i = 1, 2, \ldots, n$, is of the form

$$\widehat{\mu}_i = \frac{z_{1i} + \widehat{b} z_{2i} - \widehat{a}\widehat{b}}{1 + \widehat{b}^2}. \tag{2.14}$$

The estimator $\widehat{\sigma}^2$ converges in probability to $\sigma^2/2$. This particular inconsistency causes no difficulty, the consistent estimator of σ^2 is simply $2n\widehat{\sigma}^2/(n-2)$. Further, the estimator $\widehat{\mu}_i$ is also inconsistent. Finally, the estimator of the errorless recordings ν_i, $i = 1, 2, \ldots, n$, is

$$\widehat{\nu}_i = \widehat{a} + \widehat{b}_2 \widehat{\mu}_i. \tag{2.15}$$

Hence, we have shown how to obtain the predicted values $(\widehat{\mu}_i, \widehat{\nu}_i)'$ when the observed values are $(z_{1i}, z_{2i})'$, $i = 1, 2, \ldots, n$.

One of the possible disadvantages of the maximum likelihood estimators, given in this section, are their asymptotic properties; therefore they are not satisfactory when making statistical inference with finite samples. Nevertheless, some approximate procedures can be found in e.g. [46, 48, 50, 61].

Linear models with type-II constrains [32], based on the calibration line approach [53, 87, 88], overcome the difficulties of the TLS approach. The regression model is of the form of

$$\begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\nu} \end{pmatrix} + \boldsymbol{\varepsilon},$$

where the unknown regression parameters a, b, and the vector of the errorless recordings $\boldsymbol{\mu}, \boldsymbol{\nu}$ satisfy

$$\boldsymbol{\nu} = a\mathbf{1}_n + b\boldsymbol{\mu}, \quad \operatorname{var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_{2n}. \tag{2.16}$$

Such constraints on regression parameters involving other unknown parameters a and b are called type-II constraints. Evidently, this is a non-linear function of the unknown parameters b and $\boldsymbol{\mu}$. Using linearization by the Taylor series locally

at $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\nu}^{(0)}$, $a^{(0)}$ and $b^{(0)}$, when the second and higher derivatives are neglected, the locally BLU estimators of $\boldsymbol{\mu}$, $\boldsymbol{\nu}$, a and b can be derived [29]:

$$\widehat{\boldsymbol{\mu}} = \boldsymbol{z}_1 + \frac{b^{(0)}}{\left[b^{(0)}\right]^2 + 1} \boldsymbol{M}^{(0)} \left[\boldsymbol{z}_2 - \boldsymbol{\nu}^{(0)} - b^{(0)} \left(\boldsymbol{z}_1 - \boldsymbol{\mu}^{(0)}\right)\right], \tag{2.17}$$

$$\widehat{\boldsymbol{\nu}} = \boldsymbol{z}_2 - \frac{1}{\left[b^{(0)}\right]^2 + 1} \boldsymbol{M}^{(0)} \left[\boldsymbol{z}_2 - \boldsymbol{\nu}^{(0)} - b^{(0)} \left(\boldsymbol{z}_1 - \boldsymbol{\mu}^{(0)}\right)\right], \tag{2.18}$$

$$\begin{aligned} \begin{pmatrix} \widehat{a} \\ \widehat{b} \end{pmatrix} = \begin{pmatrix} a^{(0)} \\ b^{(0)} \end{pmatrix} + \begin{pmatrix} n, & \mathbf{1}'\boldsymbol{\mu}^{(0)} \\ \left[\boldsymbol{\mu}^{(0)}\right]' \mathbf{1}, & \left[\boldsymbol{\mu}^{(0)}\right]' \boldsymbol{\mu}^{(0)} \end{pmatrix}^{-1} \\ \times \begin{pmatrix} \mathbf{1}' \left[\boldsymbol{z}_2 - \boldsymbol{\nu}^{(0)} - b^{(0)} \left(\boldsymbol{z}_1 - \boldsymbol{\mu}^{(0)}\right)\right] \\ \left[\boldsymbol{\mu}^{(0)}\right]' \left[\boldsymbol{z}_2 - \boldsymbol{\nu}^{(0)} - b^{(0)} \left(\boldsymbol{z}_1 - \boldsymbol{\mu}^{(0)}\right)\right] \end{pmatrix}, \end{aligned} \tag{2.19}$$

where

$$\boldsymbol{M}^{(0)} = \boldsymbol{I}_n - \left(\mathbf{1}, \boldsymbol{\mu}^{(0)}\right) \begin{pmatrix} n, & \mathbf{1}'\boldsymbol{\mu}^{(0)} \\ \left[\boldsymbol{\mu}^{(0)}\right]' \mathbf{1}, & \left[\boldsymbol{\mu}^{(0)}\right]' \boldsymbol{\mu}^{(0)} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{1}' \\ \left[\boldsymbol{\mu}^{(0)}\right]' \end{pmatrix}. \tag{2.20}$$

The accuracy characteristics of these estimators are the following. The covariance matrices of the estimators $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\nu}}$ are

$$\operatorname{var}\left[\widehat{\boldsymbol{\mu}}\right] = \sigma^2 \boldsymbol{I}_n - \frac{\left[b^{(0)}\right]^2 \sigma^2}{\left[b^{(0)}\right]^2 + 1} \boldsymbol{M}^{(0)}, \tag{2.21}$$

$$\operatorname{var}\left[\widehat{\boldsymbol{\nu}}\right] = \sigma^2 \boldsymbol{I}_n - \frac{\sigma^2}{\left[b^{(0)}\right]^2 + 1} \boldsymbol{M}^{(0)}. \tag{2.22}$$

The cross-covariance matrix of the estimators $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\nu}}$ is

$$\operatorname{cov}\left[\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\nu}}\right] = \frac{b^{(0)} \sigma^2}{\left[b^{(0)}\right]^2 + 1} \boldsymbol{M}^{(0)}. \tag{2.23}$$

The covariance matrix of the estimator $(\widehat{a}, \widehat{b})'$ is

$$\operatorname{var}\left[\begin{pmatrix} \widehat{a} \\ \widehat{b} \end{pmatrix}\right] = \sigma^2 \left(\left[b^{(0)}\right]^2 + 1\right) \begin{pmatrix} n, & \mathbf{1}'\boldsymbol{\mu}^{(0)} \\ \left[\boldsymbol{\mu}^{(0)}\right]' \mathbf{1}, & \left[\boldsymbol{\mu}^{(0)}\right]' \boldsymbol{\mu}^{(0)} \end{pmatrix}^{-1} \tag{2.24}$$

and the cross-covariance matrix of the estimators $(\widehat{\boldsymbol{\mu}}', \widehat{\boldsymbol{\nu}}')'$ and $(\widehat{a}, \widehat{b})'$ is

$$\operatorname{cov}\left[\begin{pmatrix} \widehat{a} \\ \widehat{b} \end{pmatrix}, \begin{pmatrix} \widehat{\boldsymbol{\mu}} \\ \widehat{\boldsymbol{\nu}} \end{pmatrix}\right] = -\sigma^2 \begin{pmatrix} n, & \mathbf{1}'\boldsymbol{\mu}^{(0)} \\ \left[\boldsymbol{\mu}^{(0)}\right]' \mathbf{1}, & \left[\boldsymbol{\mu}^{(0)}\right]' \boldsymbol{\mu}^{(0)} \end{pmatrix}^{-1} \begin{pmatrix} b^{(0)} \mathbf{1}', & -\mathbf{1}' \\ b^{(0)} \left[\boldsymbol{\mu}^{(0)}\right]', & -\left[\boldsymbol{\mu}^{(0)}\right]' \end{pmatrix}.$$

One can see that the estimators $\widehat{\boldsymbol{\mu}}$, $\widehat{\boldsymbol{\nu}}$, $\widehat{a}$ and $\widehat{b}$ depend on the unknown approximate values $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\nu}^{(0)}, a^{(0)}$ and $b^{(0)}$, therefore it is necessary to solve them on an iterative manner. Accuracy characteristics can be evaluated using the resulting estimates of the iterative procedure.

The variance σ^2 is usually unknown and can be unbiasedly estimated by [53]

$$\widehat{\sigma}^2 = \frac{(\mathbf{z}_1 - \widehat{\boldsymbol{\mu}})'(\mathbf{z}_2 - \widehat{\boldsymbol{\mu}}) + (\mathbf{z}_1 - \widehat{\boldsymbol{\nu}})'(\mathbf{z}_2 - \widehat{\boldsymbol{\nu}})}{n-2}. \tag{2.25}$$

In the following we outline the standard iterative algorithm for estimating the calibration line, described in four main steps [29]. The first step consists of determining initial values of the intercept a and the slope b of the calibration line and the errorless recordings $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$. In case a specific prior information on the true values of these parameters occurs, we should take it into account, otherwise the choice should satisfy the relation (2.16), for example

$$a^{(0)} = \frac{z_{1j}z_{2i} - z_{1i}z_{2j}}{z_{1j} - z_{1i}},\ b^{(0)} = \frac{z_{2j} - z_{2i}}{z_{1j} - z_{1i}},$$

$$\boldsymbol{\mu}^{(0)} = \mathbf{z}_1,\ \boldsymbol{\nu}^{(0)} = a^{(0)}\mathbf{1}_n + b^{(0)}\boldsymbol{\mu}^{(0)},$$

where $z_{1i} = \min\{z_{1k} : k = 1, 2, \ldots, n\}$, $z_{2j} = \max\{z_{2k} : k = 1, 2, \ldots, n\}$ and z_{2i}, z_{2j} are the corresponding z_2 coordinates. The choice of the initial values does not have any impact on convergence of this algorithm. In the second step, we calculate $\widehat{a}$, $\widehat{b}$, $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\nu}}$ for every data point $(z_{1i}, z_{2i})'$, $i = 1, 2, \ldots, n$, using the equations (2.17)-(2.20). Further, in the the third step we need to update the initial values by the scheme

$$\boldsymbol{\nu}^{(0)} = \widehat{\boldsymbol{\nu}} + (\widehat{b} - b^{(0)})(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^{(0)}), \quad \boldsymbol{\mu}^{(0)} = \widehat{\boldsymbol{\mu}}, \quad a^{(0)} = \widehat{a}, \quad b^{(0)} = \widehat{b}. \tag{2.26}$$

We repeat steps 2 and 3 until estimates converge, i.e. until changes in estimates at each iteration are less than some pre-set tolerance. Estimates obtained from this iterative algorithm converge usually very quickly, and they also preserve the prescribed condition (2.16). Thus, linear models approach represents an alternative technique for the TLS; however, the solution is only approximative

due to linearization step. On the other hand, this technique enables to provide further statistical inference. This means that under the assumption of normality we can construct, e.g. approximative confidence domains or statistical tests.

Part II

Results

3. Covariance - based variable selection

Dimension reduction of data is often desired in order to simplify the multivariate statistical analysis and as well, to simplify the interpretation of the results. Usually when we want to know which variables to exclude, we ask the experts. Regrettably, their choice may lead to major changes of the multivariate statistical analysis results.

The first section in this chapter introduces some usefull features of the variation matrix. In the following section, we propose the covariance-based stepwise procedure for variable selection for compositions that guarantees that the loss of the information when moving from a composition to a subcomposition is rather negligible. Subsequent two sections involve application of this procedure on real world data from geochemistry. This chapter is mainly based on the article [43].

Obviously this stepwise variable selection is useful in regression analysis with compositions as well. The reduction of dimension of compostions simplify the regression analysis, which enable the analysis to be even faster and the final results easier to interpret.

3.1 Properties of the variation matrix

The basic idea of the compositional variable reduction algorithm is to obtain such a subcomposition that will give the same information about the multivariate data structure like the initial composition. The information on compositional variability is included in the variation matrix given by the expression (1.7). More-

over, the total variability of the compositional data set is captured in the total variance, which is equal to the sum of all elements of the variation matrix like given in (1.8). Equivalently the total variance is equal to the sum of the variances of the clr (orthonormal) coordinates as it is stated in (1.9).

Let us consider D-part composition $\mathbf{y}$ and its clr coordinates $\mathbf{h}$ given by (1.2) and D-systems of orthonormal coordinates $\mathbf{z}^{(i)}$, $i = 1, 2, \ldots, D$, given by (1.4). As a consequence of (1.6), the variance of the clr coordinate h_i corresponds (up to a constant) to the variance of $z_1^{(i)}$. Also the covariance structure of the clr coordinates can be analyzed, which has been done thoroughly in [4].

From (1.10), multiplied by $(D-1)/D$ to obtain clr variances, we can also expect quite a strong relation between $\mathrm{var}(h_i)$ and the sum of the i-th row (column) of the corresponding variation matrix $\mathbf{T}$. This finding induces a useful property, mentioned in the next theorem [43]. Particularly, it shows that ordered variances of different clr coordinates (or, alternatively, of the first orthonormal coordinates from (1.4) correspond to the same order of the sums in the variation matrix connected with the related compositional parts.

Theorem 3.1. *Consider the clr coordinates h_i and h_j, $i \neq j$, $i, j \in \{1, 2, \ldots, D\}$ (or, equivalently, balances $z_1^{(i)}$ and $z_1^{(j)}$ from (1.4), corresponding to two different orthonormal bases). Then* $\mathrm{var}(h_i) \geq \mathrm{var}(h_j)$, *if and only if*

$$\sum_{p=1}^{D} \mathrm{var}\left(\ln \frac{y_i}{y_p}\right) \geq \sum_{p=1}^{D} \mathrm{var}\left(\ln \frac{y_j}{y_p}\right).$$

Proof. Let $\mathrm{var}(h_i) \geq \mathrm{var}(h_j)$, $i, j = 1, 2, \ldots, D$. According to (1.10), this is equivalent to

$$\frac{D-1}{D^2} \sum_{p=1}^{D} \mathrm{var}\left(\ln \frac{y_i}{y_p}\right) - \frac{1}{2D^2} \sum_{\substack{p=1 \\ p \neq i}}^{D} \sum_{\substack{s=1 \\ s \neq i}}^{D} \mathrm{var}\left(\ln \frac{y_p}{y_s}\right) \geq$$

$$\frac{D-1}{D^2} \sum_{p=1}^{D} \mathrm{var}\left(\ln \frac{y_j}{y_p}\right) - \frac{1}{2D^2} \sum_{\substack{p=1 \\ p \neq j}}^{D} \sum_{\substack{s=1 \\ s \neq j}}^{D} \mathrm{var}\left(\ln \frac{y_p}{y_s}\right).$$

Extending the left-hand side of this inequality by the term $\pm\left(\frac{1}{D^2}\right)\sum_{p=1}^{D}\text{var}\left(\ln\frac{y_i}{y_p}\right)$ and using the relationship $\text{var}\left(\ln\frac{y_p}{y_s}\right) = \text{var}\left(\ln\frac{y_s}{y_p}\right)$, the left-hand side can be rewritten in the form

$$\frac{1}{D}\sum_{p=1}^{D}\text{var}\left(\ln\frac{y_i}{y_p}\right) - \frac{1}{2D^2}\sum_{p=1}^{D}\sum_{s=1}^{D}\text{var}\left(\ln\frac{y_p}{y_s}\right).$$

Similarly we can adjust the right-hand side of the inequality, and thus $\text{var}(h_i) \geq \text{var}(h_j)$ if and only if

$$\sum_{p=1}^{D}\text{var}\left(\ln\frac{y_i}{y_p}\right) \geq \sum_{p=1}^{D}\text{var}\left(\ln\frac{y_j}{y_p}\right).$$

□

This theorem can be used to identify compositional parts ("markers") that are responsible for larger clr variances. In particular, it allows to detect possible sources of changes in the multivariate analysis of compositional data, like those resulting from the compositional biplot [7]. Theorem 3.1 makes it possible to identify the ordered contribution of the single compositional parts to the overall variance with the corresponding clr coordinates. Using this fact, a stepwise algorithm is introduced in the following that helps to derive a subcomposition with a minimal loss concerning the total variance of the original composition.

3.2 Stepwise procedure

Let us consider a composition $\mathbf{y} = (y_1, y_2, \ldots, y_D)'$. Without loss of generality, let

$$\text{var}(h_1) \geq \cdots \geq \text{var}(h_D),$$

which is, according to Theorem 3.1, equivalent to

$$\sum_{p=1}^{D}\text{var}\left(\ln\frac{y_1}{y_p}\right) \geq \sum_{p=1}^{D}\text{var}\left(\ln\frac{y_2}{y_p}\right) \geq \cdots \geq \sum_{p=1}^{D}\text{var}\left(\ln\frac{y_D}{y_p}\right).$$

Since h_D has the smallest variance, its contribution to the overall variance of the compositional data set, totvar($\mathbf{y}$), is minimal. This is equivalent to the statement that the aggregated variances of the log-ratios with the part y_D have the smallest contribution to the overall variance. Consequently, the part y_D is not determining the multivariate data structure and it can be omitted from the composition. Hence, we arrive at a subcomposition $\mathbf{y}_1 = (y_1, y_2, \ldots, y_{D-1})'$. In the next step we perform a clr transformation on $\mathbf{y}_1$, calculate variances of the clr transformed variables and again omit the part corresponding to the clr coordinate with the smallest variance. So we continue until a certain number of parts is obtained, and we stop at latest after $D-2$ steps.

The order of the variances of the clr coordinates is generally not maintained after omitting the part of composition $\mathbf{y}$ corresponding to the clr coordinate with the smallest variance. In fact, as a simple consequence of Theorem 3.1, the order of the clr variances when moving from a D-part to a $(D-1)$-part composition is maintained only under the assumption

$$\operatorname{var}\left(\ln\frac{y_1}{y_D}\right) \geq \operatorname{var}\left(\ln\frac{y_2}{y_D}\right) \geq \cdots \geq \operatorname{var}\left(\ln\frac{y_{D-1}}{y_D}\right).$$

Nevertheless, from simulations using real geochemical data (see next sections) it follows that the ordering of the clr coordinates of the original composition according to their variances is a relatively accurate indicator whether the corresponding part of the original composition will be included in the final subcomposition or not.

The prescribed number of parts of the target subcomposition is usually not provided. Thus, an important question is when the selection of compositional parts should be stopped. It is easy to see that the main idea of the above algorithm is to select a subcomposition such that the loss in total variance of the composition from the previous step is minimal. This inspires to find such a criterion that compares the total variance of the subcomposition, obtained in the i-th step of the algorithm, $i = 1, 2, \ldots, D-2$, with the total variance of the composition from the previous step. In more detail, denote $\widehat{\text{totvar}}(\mathbf{y}_i)$ the

total variance of the i-th subcomposition, estimated from the data. We want to test whether its difference to the total variance $\mathrm{totvar}(\mathbf{y}_{i-1})$ can be considered as negligible (i.e., the null hypothesis is $\mathrm{totvar}(\mathbf{y}_i) = \mathrm{totvar}(\mathbf{y}_{i-1})$) or rather as a result of a systematic pattern (alternative hypothesis $\mathrm{totvar}(\mathbf{y}_i) < \mathrm{totvar}(\mathbf{y}_{i-1})$). Obviously, $\mathrm{totvar}(\mathbf{y}_{i-1})$ is not known and also needs to be estimated from the data (in the previous step of the algorithm), as it is the case with $\widehat{\mathrm{totvar}}(\mathbf{y}_i)$. Here we assume that $\mathrm{totvar}(\mathbf{y}_{i-1})$ is fixed from the previous step of the algorithm, and thus it can be considered as a given (non-random) number in the current step. The following test statistic from [42],

$$U_i^+ = \frac{\widehat{\mathrm{totvar}}(\mathbf{y}_i) - \mathrm{totvar}(\mathbf{y}_{i-1})}{\sqrt{\frac{2}{n-1}\mathrm{tr}(\widehat{\mathbf{\Sigma}}_i^2)}}, \tag{3.1}$$

is used for this purpose; the matrix $\widehat{\mathbf{\Sigma}}_i$ stands for the sample covariance matrix of the composition $\mathbf{y}_i$ in (arbitrarily chosen) orthonormal coordinates. Small values of U_i^+ favor the alternative, so we reject the null hypothesis, if U_i^+ realizes in the critical region $\mathsf{W} = (-\infty, u_\alpha\rangle$, where u_α denotes the α-quantile (preferably $\alpha = 0{,}05$) of the standard normal distribution (being inspired by the asymptotic distribution of U_i^+, see [42] for details). Thus, in each step of the algorithm we compute the statistic U_i^+, and the procedure is stopped when U_i^+ realizes for the first time in W.

Practical properties of the proposed iterative procedure will be demonstrated on real-world examples in the next sections.

3.3 Illustrative example: Kola

Firstly, we will demonstrate the results of the proposed stepwise algorithm at the well-known Kola data [71]. This data set is the result of a large geochemical mapping project, carried out from 1992 to 1998 by the Geological Surveys of Finland and Norway, and the Central Kola Expedition, Russia. An area covering 188 000 km^2 at the peninsula Kola in northern Europe was sampled. In

total, around 600 samples of soil were taken in 4 different layers (moss, humus, B-horizon, C-horizon), and subsequently analyzed by a number of different techniques for more than 50 chemical elements. The project was primarily designed to reveal the environmental conditions in the area. The data are available in the package `StatDA` [27] of the software environment R [70].

Four experiments with this data set are realized, each of them having a purpose to show on the benefits of the proposed stepwise procedure. The tasks for each experiment are the following:

- observe the reduction in total variance when the procedure is applied on a subset of the moss layer data set consisting of 15 out of 31 compositional parts;
- investigate whether the test statistic is able to select appropriate compositional parts when applied on a subset of the moss layer data set consisting of 15 out of 31 compositional parts;
- check how the stepwise procedure behaves when we have different sizes of the start composition;
- examine the results of the procedure after being applied on the whole moss layer data set.

Below is described in detail how each of the tasks is solved.

First experiment

In the first experiment we are interested in observing the reduction in total variance by the stepwise procedure. For this purpose we use all the 31 elements of the moss layer and select randomly 15 variables. Then the stepwise algorithm is applied until a two-part subcomposition is reached (i.e. here we are not using the proposed stopping criterion). After each step the total variance is computed. The whole procedure is repeated 1000 times, and the results are shown in Figure 3.1.

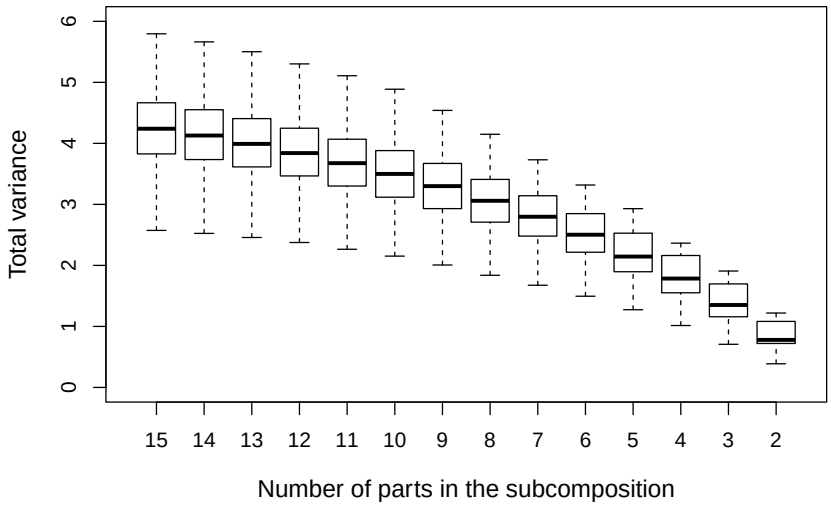

Figure 3.1: Total variances of subcompositions obtained from the stepwise algorithm.

Each boxplot summarizes the total variances achieved by the given size of the subcomposition. A decreasing sequence of the total variance (and its variability among subcompositions of the given size, see whiskers of the boxplots) is clearly visible. Subsequent steps of the algorithm result in increasing relative differences of the median total variances. This is important for obtaining significance at a certain step using the proposed test statistic.

Note that this feature as well as the below mentioned properties are characteristic for the stepwise procedure even in general, independent on the concrete data set chosen.

Second experiment

In the second experiment we want to check if proposed test statistic is able to select appropriate compositional parts. For this we again select randomly 15 parts of the Kola moss data as a starting composition. Then

the stepwise procedure is applied until the test statistic suggests to stop the process. Repeating this experiment 1000 times results in a distribution of the number of remaining parts, which is visualized by a barplot in Figure 3.2 (left).

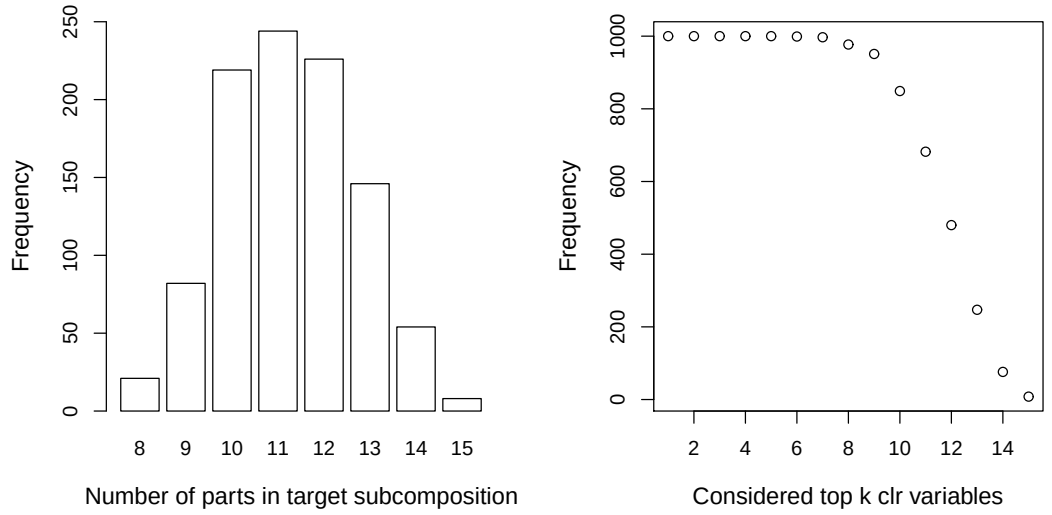

Figure 3.2: Barplot of the number of parts of the subcomposition resulting from the stepwise procedure using the stopping-criterion (left); clr coordinates of the initial composition, sorted according to decreasing variance, versus number of times the corresponding compositional parts were included in the resulting subcomposition (right).

The algorithm arrives typically at subcompositions with 10 to 12 parts, i.e. around two thirds of the starting number of parts. The important question is whether the resulting target compositions are indeed consisting of parts with large clr variances of the initial compositions. Therefore, we sort the parts of all 1000 initial subcompositions according to decreasing values of their clr variances, and count how often the top k clr coordinates were included in the target compositions, where $k = 1, 2, \ldots, 15$. Figure 3.2 (right) shows the result. The counts have to decrease for larger values of k because of the possibly smaller total number of parts in the target compositions, see Figure 3.2 (left). We can see that the initial clr coordinate with largest variance ($k = 1$) was selected in all 1000 cases. This also holds for k up

to 5, i.e., the top 5 clr coordinates were always selected. Then the counts drop, partly because the resulting subcompositions were smaller than k, and partly because not all considered k clr coordinates were selected. The figure, however, clearly indicates that the important clr coordinates were included in the target subcompositions, although Theorem 3.1 provides here no theoretical guarantee.

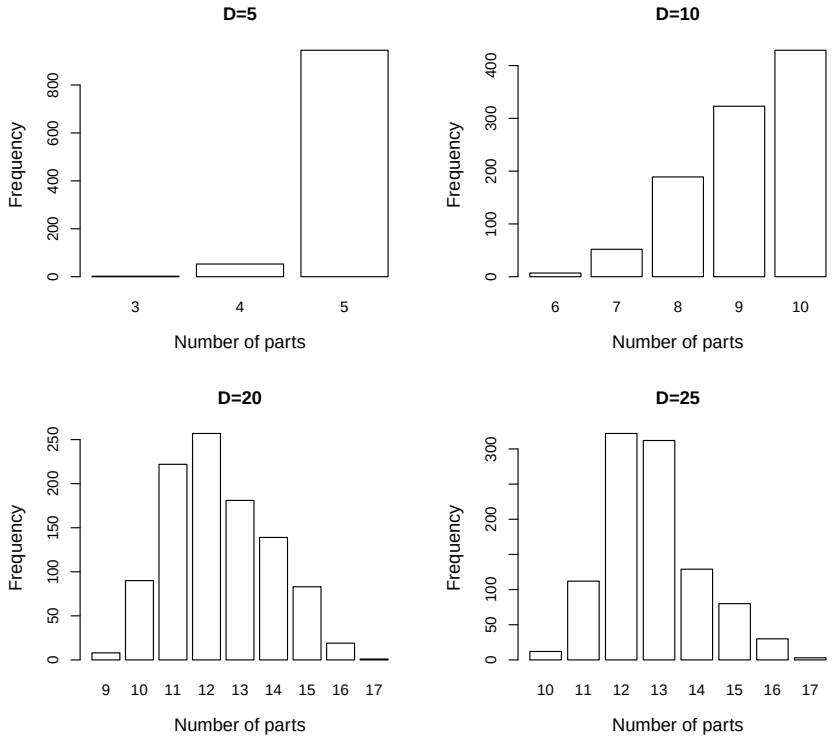

Figure 3.3: Barplots of the number of parts of the subcomposition resulting from the stepwise procedure using the stop-criterion with 5-part (upper left), 10-part (upper right), 20-part (lower left) and 25-part (lower right) original compositions.

Third experiment

In the third simulation experiment we analyze the behaviour of the stepwise procedure for different sizes of the starting composition. We use the same simulation setting as before, but select as initial composition 5, 10, 20, and 25 parts of the Kola moss data, respectively. For each case 1000 simulations are performed, and the distributions of the resulting numbers of parts in

the target compositions are shown in Figure 3.3. If the starting composition has only $D = 5$ parts (upper left), the procedure usually arrives at a target composition again with 5 parts. On the other hand, if one starts with $D = 25$ parts (lower right), the number of parts will be reduced to about a half. This behaviour of shrinking larger compositions more and more is very desirable for practice.

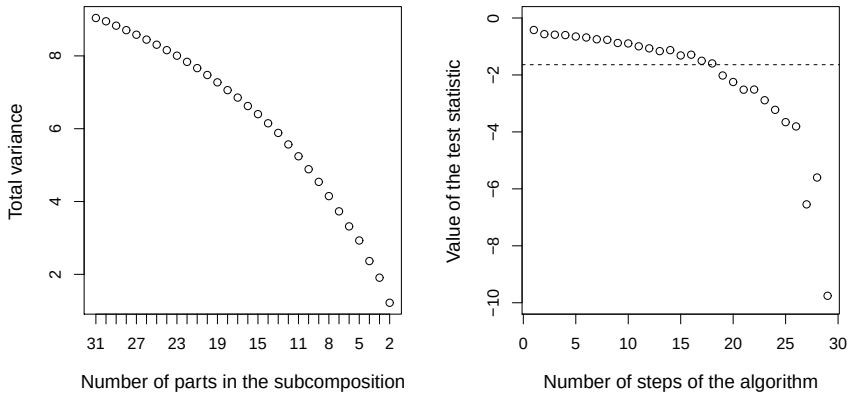

Figure 3.4: Total variances of subcompositions obtained from the stepwise algorithm for the whole moss layer data set (left), corresponding values of the test statistic U_i^+ in (3.1) together with the cut-off value (right).

Forth experiment

In the last illustration we apply the stepwise procedure to the whole moss layer data set consisting of 31 compositional parts. The total variances of the resulting subcompositions are plotted in Figure 3.4. They quite nicely correspond to the trend as indicated in Figure 3.1 (left). The right picture shows the values of the test statistic U_i^+ given in (3.1), for each of the $i = 1, 2, \ldots, 29$ possible steps, and they reflect the same trend. The algorithm will stop after the value of the test statistic falls below $u_{0.05} = -1.64$ (horizontal line). This happens at step 19 of the algorithm, and thus a 13-part subcompositions is remaining.

3.4 Illustrative example: Baltic soil survey

In this section, we demonstrate the effect of the stepwise procedure using the compositional biplot [7] as a visualization tool. Here we employ the Baltic Soil Survey (BSS) data [72], which originate from a large-scale geochemistry project carried out in northern Europe, in an area of about 1 800 000 km^2. On an irregular grid, 769 samples of agricultural soils have been collected. The samples came from two different layers, the top layer (0-25 cm) and the bottom layer (50-75 cm). All samples were analyzed for the concentration of more than 40 chemical compounds. The data sets of the top and bottom layer are available in the R package `mvoutlier` [28]. Here we use the major elements (Al_2O_3, Fe_2O_3, K_2O, MgO, MnO, CaO, TiO_2, Na_2O, P_2O_5 and SiO_2), plus LOI (Loss on ignition) of the top layer, i.e. an 11-part composition. Note that the same elements were used also in [26], where classical and robust biplots of both log- and clr-transformed compositions were compared.

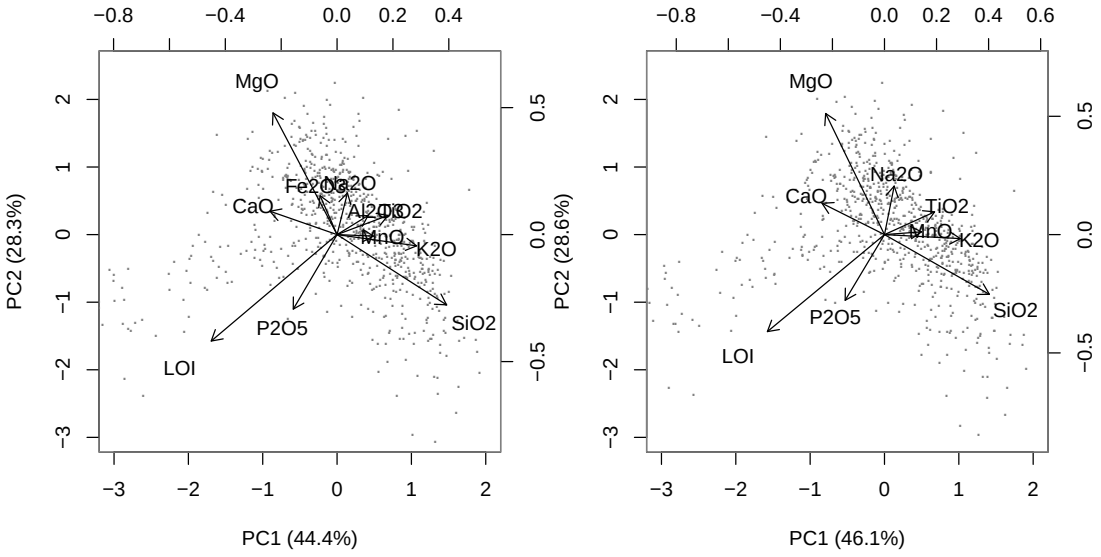

Figure 3.5: Biplots of the BSS data with all major elements (left) and after exclusion of Al_2O_3 and Fe_2O_3 (right).

Figure 3.5 (left) shows the classical compositional biplot of the initial 11-part composition. If we apply the stepwise procedure, we arrive at a 9-part

subcomposition. The elements Al_2O_3 and Fe_2O_3 were subsequently excluded with the corresponding values of the test statistics $U_1^+ = -0.7185$ and $U_2^+ = -1.4753$, respectively. The next step with an exclusion of TiO_2 would already lead to significance with a value of $U_3^+ = -2.2712$. The resulting biplot (Figure 3.5, right) shows that there is nearly no difference visible in the relations among the remaining compositional parts (arrows in the biplot) compared to the original biplot.

Thus, the multivariate data structure is widely preserved and the information of the excluded elements is still contained in the remaining subcomposition.

4. Regression with compositional response

Regression analysis with compositional response is of great potential interest in geochemistry [43] and also in medical applications [49], e.g., in human metabolomics, where concentrations of metabolites are frequently influenced by external factors (temperature, age of patients, etc.).

Despite this intensive care, there are still some practical aspects concerning linear regression with compositional response that deserve to be further analysed. The first one concerns special orthonormal coordinate systems that enable interpretation in terms of the original compositional parts (with respect to the other parts in the actual composition) and were applied in a number of applications including regression modelling [30, 41, 49]. Although it is theoretically sound to work exclusively in orthonormal coordinates, this particular choice of coordinates seems to be also a bit impractical as for a D-part composition D coordinate systems are needed. It can be shown that due to the relation between these particular orthonormal coordinates and centred log-ratio coordinates (1.6) that are formed by coefficients with respect to a generating system, it is possible to get easily the same numerical outputs (or possibly up to a constant multiple) concerning regression parameters estimation and hypotheses testing in multivariate regression models by using just one coordinate system. The second aspect concerns the relation between the mean square error (MSE) and the coefficient of determination, obtained from a regression model in orthonormal coordinates, or after applying a log-transformation to the original compositional data (in units

that do not clearly indicate relative structure of components, like proportions or percentages). This fact may be useful for further methodological developments, similarly as it was the case of inequality between Euclidean distance in orthonormal log-ratio coordinates (or, equivalently, the Aitchison distance [6] between the original compositions) and Euclidean distance between log-transformed compositions [66]. For example, the mentioned relation was successfully used for the case of compositional data with an informative total (sum of parts), characterized by so called T-spaces [67], where a log-transformation plays an important role of a possible coordinate representation as an alternative to orthonormal log-ratio coordinates plus a variable representing the total. Each of these aspects is examined in the next sections. In the last section in this chapter we present an illustrative example from geochemistry. The chapter is assembled from the article [33].

4.1 Multivariate regression model with compositional response

Regression with a D-part compositional response leads to a multivariate linear model with a $(D-1)$-dimensional response variable formed by orthonormal coordinates. Although by using orthonormal coordinates, it is possible to decompose the multivariate model into $D-1$ multiple regressions [24], in general, the multivariate approach has several advantages in comparison with a series of univariate models.

Let us consider D-part compositional responses $\mathbf{y}_i$, $i = 1, 2, \ldots, n$. Let us denote $\underline{\mathbf{z}}_{i\cdot} = (z_{i1}, z_{i2} \ldots, z_{iD-1})'$, $i = 1, 2, \ldots, n$, the corresponding orthonormal coordinates given by (1.3) such that the first orthonormal coordinate z_{i1} explains all the relative information about the first compositional part y_{i1}. In order to obtain coordinates with similar interpretation for each of the compositional parts y_{il}, $l = 1, 2, \ldots, D$, D different orthonormal coordinate systems are needed. These are obtained by means of permutation of the first part of the composition $\mathbf{y}$. Moreover, there exists unique relationship between these coordinate systems,

therefore, without loss of generality, we can focus just on the coordinates $\underline{\mathbf{z}}_{i\cdot}$.

According to (2.1) and (2.2), the multivariate linear model can be expressed as

$$(\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_{D-1}) = \mathbf{X}(\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{D-1}) + (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{D-1}),$$

or, equivalently, in the matrix form

$$\underline{\mathbf{z}} = \mathbf{X}\mathbf{B} + \underline{\varepsilon}.$$

Here it is assumed that $\mathbf{X}$ is an $(n \times k)$ dimensional design matrix of full column rank, $\mathbf{b}_j$, $j = 1, 2, \ldots, D-1$, is a k dimensional vector of unknown regression parameters and $\underline{\varepsilon}$ is an $[n \times (D-1)]$ dimensional matrix of the random errors. The multivariate responses $\underline{\mathbf{z}}_{i\cdot} = (z_{i1}, z_{i2}, \ldots, z_{iD-1})'$, $i = 1, 2, \ldots, n$, are assumed to be independent with the same unknown variance-covariance matrix $\mathbf{\Sigma}$.

According to (2.4), the BLUE of the parameter matrix $\mathbf{B}$ is

$$\widehat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{D-1}).$$

The estimator of $\mathbf{B}$ is invariant under a change of the variance-covariance matrix $\mathbf{\Sigma}$. With respect to (2.5), the variance-covariance matrix of $\mathrm{vec}(\widehat{\mathbf{B}})$ is

$$\mathrm{var}\left[\mathrm{vec}(\widehat{\mathbf{B}})\right] = \mathbf{\Sigma} \otimes (\mathbf{X}'\mathbf{X})^{-1},$$

furthermore the unbiased estimator of the variance-covariance matrix $\mathbf{\Sigma}$ is given by $\widehat{\mathbf{\Sigma}} = \underline{\mathbf{z}}'\mathbf{M}_X\underline{\mathbf{z}}/(n-k)$.

Under the assumption of normally distributed coordinate representation $\underline{\mathbf{z}}_{i\cdot}$ of the compositional response [57], hypotheses testing can be performed. In the following we will present some tests introduced in Section 2.1 adjusted for the study of multivariate regression model with compositional response.

To verifying the significance of the covariates for the ilr coordinate z_j, $j = 1, 2, \ldots, D-1$, point wise and simultaneously, i.e. testing the null hypothesises $\mathbf{b}_j = \mathbf{0}$, $j = 1, 2, \ldots, D-1$, the test statistic (2.6) can be used. In the following,

this statistic will be denoted as F_j^{ilr} in order to point out that the multivariate model in orthonormal coordinates is considered.

Another test that can be taken into account is the test for the significance of the i-th predictor, $i = 1, 2, \ldots, k$, i.e. test of the hypothesis $\mathbf{B}_{i.} = \left(b_{i1}, \ldots, b_{i(D-1)}\right) = \mathbf{0}$. The test statistic for this case is derived from the relation (2.7) which results in the following expression for the test statistic

$$F_{pred,i}^{ilr} = \frac{(n - D - k + 2)\, \widehat{\mathbf{B}}_{i.} \left(\underline{\mathbf{z}}' \mathbf{M}_X \underline{\mathbf{z}}\right)^{-1} \widehat{\mathbf{B}}_{i.}'}{(D-1) \left\{\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right\}_{ii}},$$

which is distributed as $F_{D-1,n-D-k+2}$ under the null hypothesis H_{0i}.

Lastly sometimes it is of interest to verify the significance of the whole matrix of regression parameters $\mathbf{B}$, or in other words to test the hypothesis $\mathbf{AB} = \mathbf{C}$, where $\mathbf{A}$ is a $q \times k$ hypothesis matrix having full-row rank $q \leq k$, and $\mathbf{C}$ is a $q \times D - 1$ matrix. Therefore, we will use the well-known Pillai-Barttlet trace, Wilk's Lambda, Hotelling-Lawley trace and Roy's largest root that rely on the $p = \min(q, D-1)$ non-zero eigenvalues λ_j of $\mathbf{HE}^{-1}$ where the matrices $\mathbf{H}$ and $\mathbf{E}$ are given in (2.8) and (2.9).

The behaviour of these matrices in different coordinate systems is thus crucial for statistical properties of the above test statistics. Obviously, all of them are invariant under a change of a basis thus they follow the behaviour of the sample covariance matrix under affine transformations [55].

4.2 Regression with compositional response in different coordinate systems

Due to (1.6) that describes the relationship between single clr coefficients and the first orthonormal coordinates from (1.4) it seems to be quite intuitive possibility to replace orthonormal coordinates in the response simply by their clr counterparts and then proceed with the regression analysis.

Nevertheless, due to singularity of the covariance matrix of clr coordinates it is not possible to decompose the multivariate model into univariate ones as it was

the case for orthonormal coordinates. Though, as it is shown below, even taking multivariate regression in clr coordinates would yield the same results of the respective test statistics as one would obtain by considering single orthonormal coordinates, coming from D regression models.

In the following, the relation between clr and ilr coordinate systems (1.5) is extensively used. Then the multivariate model can be also written in the form

$$\underline{\mathbf{h}} = \mathbf{X}\mathbf{B}^{clr} + \underline{\varepsilon}_{clr}, \tag{4.1}$$

where $\underline{\mathbf{h}} = (\mathbf{h}_1, (\mathbf{h}_2, \ldots, \mathbf{h}_D)$ is the $(n \times D)$ dimensional matrix of response vectors that stand for the clr coordinates given in (1.2). The variance - covariance matrix of independent D-variate responses $\underline{\mathbf{h}}_{i\cdot}$ is $\text{var}(\underline{\mathbf{h}}_{i\cdot}) = \Sigma_{clr} = \mathbf{V}'\Sigma_{ilr}\mathbf{V}$, $i = 1, 2, \ldots, n$ where $\mathbf{V}$ is the $[(D-1) \times D]$ matrix, such that satisfies $\mathbf{V}\mathbf{V}' = \mathbb{I}_{(D-1)}$, having the rows vectors given by the relation (1.3). The variance - covariance matrix Σ_{clr} is a $D \times D$ positive semi-definite matrix with the rank $D-1$ unlike Σ_{ilr}, which is a full rank $(D-1) \times (D-1)$ positive definite matrix. Obviously, $\Sigma_{ilr} = \mathbf{V}\Sigma_{clr}\mathbf{V}'$. The relationships between the parameter matrices and multivariate responses are the following

$$\begin{aligned} \mathbf{B}^{clr} &= \mathbf{B}^{ilr}\mathbf{V}, \\ \mathbf{B}^{ilr} &= \mathbf{B}^{clr}\mathbf{V}', \\ \underline{\mathbf{h}} &= \underline{\mathbf{z}}\mathbf{V}, \\ \underline{\mathbf{y}} &= \underline{\mathbf{h}}\mathbf{V}'. \end{aligned} \tag{4.2}$$

Theorem 4.1. *(i) The test statistics for the hypotheses* $\mathbf{B}^{ilr}_{i\cdot} = \mathbf{0}$ *and* $\mathbf{B}^{clr}_{i\cdot} = \mathbf{0}$ *are the same for an arbitrary* $i = 1, 2, \ldots, k$, *i.e.* $F^{ilr}_{pred,i} = F^{clr}_{pred,i}$.

(ii) Let us denote $\beta_{clr,l}$ *the* l*-th column vector of the parameter matrix* $\mathbf{B}^{clr}$ *in the model with clr coordinates responses, and* $\beta^{(l)}_{ilr,1}$ *the first column vector of the parameter matrix* $\mathbf{B}^{ilr}$ *in the* l*-th model with orthonormal coordinates* $\mathbf{y}^{(l)}$ *considered as multivariate responses. Then the test statistics for the null hypotheses* $\mathbf{b}^{(l)}_{ilr,1} = \mathbf{0}$ *and* $\mathbf{b}_{clr,l} = \mathbf{0}$ *for an arbitrary* $l = 1, 2, \ldots, D$, *are the same, i.e.* $F^{ilr,(l)}_1 = F^{clr}_l$.

Proof. Let us consider the first statement. According to the relations (4.2) and (4.2), as well as from the fact that the matrix $\widehat{\Sigma}_{clr}$ is singular with the rank $D-1$, the test statistic $F^{clr}_{pred,i}$ that arises from a general formula in [54] can be rewritten as

$$\begin{aligned} F^{clr}_{pred,i} &= \frac{(n - \mathrm{r}(\mathbf{X}) - \mathrm{r}(\widehat{\Sigma}_{clr}) + 1)\widehat{\mathbf{B}}^{clr}_{i\cdot}\widehat{\Sigma}^{-}_{clr}(\widehat{\mathbf{B}}^{clr}_{i\cdot})'}{\mathrm{r}(\widehat{\Sigma}_{clr})\{(\mathbf{X}'\mathbf{X})^{-1}\}_{ii}} \\ &= \frac{(n - D - k + 2)\widehat{\mathbf{B}}^{ilr}_{i\cdot}\mathbf{V}(\mathbf{V}'\widehat{\Sigma}_{ilr}\mathbf{V})^{-}\mathbf{V}'(\widehat{B}^{ilr}_{i\cdot})'}{(D-1)\{(\mathbf{X}'\mathbf{X})^{-1}\}_{ii}} \\ &= \frac{(n - D - k + 2)\widehat{\mathbf{B}}^{ilr}_{i\cdot}\mathbf{V}\mathbf{V}^{-}_{L}\widehat{\Sigma}^{-1}_{ilr}\left(\mathbf{V}'_{R}\right)^{-}\mathbf{V}'(\widehat{\mathbf{B}}^{ilr}_{i\cdot})'}{(D-1)\{(\mathbf{X}'\mathbf{X})^{-1}\}_{ii}}, \end{aligned}$$

where the matrix $\mathbf{V}'_R$ is the right inverse of $\mathbf{V}'$ and the matrix $\mathbf{V}_L$ is the left inverse of $\mathbf{V}$, i.e.,

$$\left(\mathbf{V}'_{R}\right)^{-} = \widehat{\Sigma}_{ilr}\mathbf{V}\left(\mathbf{V}'\widehat{\Sigma}_{ilr}\mathbf{V}\right)^{-} \quad \text{and} \quad \mathbf{V}'\left(\mathbf{V}'_{R}\right)^{-} = \mathbf{I}_{D},$$

$$\mathbf{V}^{-}_{L} = \left(\mathbf{V}'\widehat{\Sigma}_{ilr}\mathbf{V}\right)^{-}\mathbf{V}'\widehat{\Sigma}_{ilr} \quad \text{and} \quad \mathbf{V}^{-}_{L}\mathbf{V}' = \mathbf{I}_{D-1}$$

and $\mathbf{A}^{-}$ denotes a generalized inverse of a matrix $\mathbf{A}$, i.e., a matrix fulfilling the property $\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}$.

The desired equality $F^{ilr}_{pred,i} = F^{clr}_{pred,i}$ is gained by pre-multiplying and post-multiplying the matrix $\widehat{\Sigma}_{ilr}$ by $\mathbf{V}\mathbf{V}' = \mathbf{I}_{D-1}$.

The statement (ii) is a direct consequence of (1.6). □

Theorem 4.2. *The test statistics for the null hypotheses* $\mathbf{B}^{ilr} = \mathbf{0}$ *and* $\mathbf{B}^{clr} = \mathbf{0}$*, as listed in Section 4.1, are the same.*

Proof. The statement follows from invariance under a change of a basis of the matrices $\mathbf{E}$ and $\mathbf{H}$ given by (2.8) and (2.9), $\mathbf{E}_{ilr} = \mathbf{V}\mathbf{E}_{clr}\mathbf{V}'$, $\mathbf{E}_{clr} = \mathbf{V}'\mathbf{E}_{ilr}\mathbf{V}$, $\mathbf{H}_{ilr} = \mathbf{V}\mathbf{H}_{clr}\mathbf{V}'$, $\mathbf{H}_{clr} = \mathbf{V}'\mathbf{H}_{ilr}\mathbf{V}$, and the fact that the matrices $\mathbf{H}_{clr}\mathbf{E}^{-}_{clr}$ and $\mathbf{H}_{ilr}\mathbf{E}^{-1}_{ilr}$ have the same non-zero eigenvalues. □

The above findings can be used to perform parameter estimation and significance testing in clr coordinates instead of taking D orthonormal coordinate systems of type (1.4), when the interpretation in sense of the original compositional parts (with respect to the others) is required. Although methodically working in orthonormal coordinates is preferred in any case, numerical outputs are the same (test statistics) or differ just up to a constant resulting from (1.6).

Finally, note that the interpretation of the regression parameters can be enhanced by considering orthogonal coordinates, resulting from suppressing scaling constants in orthonormal coordinates. Concretely, they are formed from (1.4) by omitting scaling constants and replacing the natural logarithm by its binary counterpart (or any other interpretable base of logarithm), i.e.

$$z_i^* = \log_2 \frac{y_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} y_j}}, \ i = 1, 2, \ldots, D-1$$

[60]. By considering regression in clr coordinates, the parameters of the resulting regression model in orthogonal coordinates, adapted to favour the l-th compositional part (denoted as $b_1^{*(l)}$), would be related through

$$b_1^{*(l)} = \log_2(\mathrm{e})\sqrt{\frac{D}{D-1}} b_{ilr,1}^{(l)} = \log_2(\mathrm{e})\frac{D}{D-1} b_{clr,l}.$$

Consequently, by taking the j-th element of $b_1^{*(l)}$, i.e. $b_{1;j}^{*(l)}$, for $j = 1, 2, \ldots, k$, then for a unit additive change in the j-th explanatory variable (by constant values of the other covariates), the ratio of x_l to the mean relative contributions of the other parts grows (decreases) $\delta = 2^{b_{1;j}^{*(l)}}$ times.

4.3 Quality of prediction in log-ratio coordinates versus log-transformed data

In practice, the simple log-transformation, $z_i = \log(y_i)$, $i = 1, 2, \ldots, D$, is often used in geochemistry, chemometrics and related fields for modelling data

with strictly positive parts. Nevertheless, it has important consequences also in the compositional context. If not just the relative structure of compositional parts is of interest, but also their absolute abundances in the original units like mg/l, cps, or monetary units [67], the log-transformation serves for an appropriate coordinate representation of the data at hand. Namely, compositional data with informative absolute values of parts induce an Euclidean vector space structure again (we refer to T-space) that should be taken into account for the construction of any relevant real coordinates.

An obvious consequence in the case of positive data (i.e., compositional data with an informative total) is lack of scale invariance, but relative scale of compositions (not absolute differences, but ratios form the source of dissimilarity between compositional vectors) is still taken into account for statistical processing. Interestingly, it is easy to see that the standard Euclidean distance of log-transformed data is always greater or equal to the Aitchison distance between two compositions $\mathbf{y}$ and $\mathbf{w}$ [6], defined as (1.1). To compare log-ratio and log-transformed regression models one has to analyse, whether something similar holds also in the regression context. Such a finding would be an important step to understand the behaviour of regression models in different coordinate systems. For this purpose, the matrix of sums of residual squares is taken for both the cases of orthonormal coordinates and log-transformed compositions,

$$\begin{aligned}\mathbf{E}_{ilr} &= (\underline{\mathbf{z}} - \mathbf{X}\widehat{\mathbf{B}})'(\underline{\mathbf{z}} - \mathbf{X}\widehat{\mathbf{B}}) = \underline{\mathbf{z}}'\mathbf{M}_X\underline{\mathbf{z}},\\ \mathbf{E}_{log} &= [\log(\underline{\mathbf{y}})]'\mathbf{M}_X \log(\underline{\mathbf{y}}),\end{aligned}$$

respectively. Here the symbol $\underline{\mathbf{y}}$ denotes an $n \times D$ matrix with D-part compositions in rows. The overall variability in data corresponds to the matrices of total sum of squares

$$\mathbf{T}_{ilr} = \underline{\mathbf{z}}'\mathbf{M}_{X_1}\underline{\mathbf{z}} = \mathbf{V}\mathbf{T}_{log}\mathbf{V}', \quad \mathbf{T}_{log} = [\log(\underline{\mathbf{y}})]'\mathbf{M}_{X_1}\log(\underline{\mathbf{y}}).$$

The matrix $\mathbf{E}$ is commonly used to measure the discrepancy between the data and a fitted model in case of multivariate regression [47]. Although also an alternative

exists, based directly on the norm between the observed and predicted response [24, 86], using directly $\mathbf{E}$ seems to be more coherent with the current regression methodology. Particularly, the trace of $\mathbf{E}$ is of primary importance, because it aggregates residual sums of squares of single response variables and leads to the multivariate analogy of the residual sum of squares (RSS). The inequalities between the traces of matrices $\mathbf{E}$ and $\mathbf{T}$ for compositions in orthonormal coordinates and by taking log-transformation are stated in the following theorem.

Theorem 4.3. *The traces of the matrices* $\mathbf{E}_{ilr}$ *(sums of residual squares) and* $\mathbf{T}_{ilr}$ *(total sum of squares) for compositions represented in orthonormal coordinates are always less or equal than the traces of the matrices* $\mathbf{E}_{log}$ *and* $\mathbf{T}_{log}$ *for log-transformed compositions, i.e. the following inequalities hold*

$$0 \leq \mathrm{tr}(\mathbf{E}_{ilr}) \leq \mathrm{tr}(\mathbf{E}_{log}), \quad 0 \leq \mathrm{tr}(\mathbf{T}_{ilr}) \leq \mathrm{tr}(\mathbf{T}_{log}).$$

Proof. The relationships between the ilr, clr coordinates and log-transformations [4, 21] can be expressed as

$$\underline{\mathbf{z}} = \underline{\mathbf{h}}\mathbf{V}', \quad \underline{\mathbf{h}} = \mathbf{M}_{X_1} \log(\underline{\mathbf{y}}),$$

where $\mathbf{V}$ contains in its rows orthonormal basis in clr coordinates, i.e. it is a $(D-1) \times D$ matrix with the property $\mathbf{V}\mathbf{V}' = \mathbf{I}_{D-1}$, and $\mathbf{M}_{X_1}$ is a projection matrix on the orthogonal complement of the vector space $\mathcal{M}(\mathbf{1}) \subset \mathbb{R}^D$ generated by the vector $\mathbf{1}$ of n ones, i.e., on the hyperplane formed by clr coordinates. Using these equalities, the matrix $\mathbf{E}_{ilr}$ can be rewritten as

$$\begin{aligned} \mathbf{E}_{ilr} &= \mathbf{V}[\underline{\mathbf{h}}]'\mathbf{M}_X\underline{\mathbf{h}}\mathbf{V}' \\ &= \mathbf{V}\mathbf{M}_{X_1}[\log(\underline{\mathbf{y}})]'\mathbf{M}_X \log(\underline{\mathbf{y}})\mathbf{M}_{X_1}\mathbf{V}'. \end{aligned}$$

The matrix $\mathbf{V}$ contains basis of the vector space that is orthogonal to the vector space $\mathcal{M}(\mathbf{1})$, and thus $\mathbf{M}_{X_1}\mathbf{V}' = \mathbf{V}'$. Hence

$$\mathbf{E}_{ilr} = \mathbf{V}[\log(\underline{\mathbf{y}})]'\mathbf{M}_X \log(\underline{\mathbf{y}})\mathbf{V}' = \mathbf{V}\mathbf{E}_{log}\mathbf{V}',$$

and the trace of the matrix $\mathbf{E}_{ilr}$ is $\mathrm{tr}(\mathbf{E}_{ilr}) = \mathrm{tr}(\mathbf{E}_{log}\mathbf{V}'\mathbf{V})$. The matrix $\mathbf{E}_{log}$ is positive semidefinite, $\mathbf{V}'\mathbf{V}$ is symmetric, and thus, the upper and lower bounds for $\mathrm{tr}(\mathbf{E}_{ilr})$ are [51]

$$\lambda_{\min}(\mathbf{V}'\mathbf{V})\mathrm{tr}(\mathbf{E}_{log}) \leq \mathrm{tr}(\mathbf{E}_{ilr}) \leq \lambda_{\max}(\mathbf{V}'\mathbf{V})\mathrm{tr}(\mathbf{E}_{log}),$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and largest eigenvalues of the matrix $\mathbf{V}'\mathbf{V}$. Since the matrix $\mathbf{V}'\mathbf{V}$ is idempotent with the rank $D-1$, it has $D-2$ eigenvalues $\lambda_{max} = 1$ and one eigenvalue $\lambda_{\min} = 0$ [37]. Thus we have

$$0 \leq \mathrm{tr}(\mathbf{E}_{ilr}) \leq \mathrm{tr}(\mathbf{E}_{log}).$$

Similarly we can prove the inequality for the trace of matrices of total sum of squares. □

Theorem 4.3 states that the trace of the matrix $\mathbf{E}$ obtained for orthonormal coordinates is less or equal to that one for log-transformed compositions. Thus, the mean squared error (MSE) for orthonormal coordinates is less or equal to the MSE for log-transformed data. Since the same inequality holds also for the trace of the matrix $\mathbf{T}$, the relationship between the coefficients of determination R^2_{ilr} and R^2_{log} does not exist in general. These measures of goodness of fit, defined as

$$R^2_{ilr} = 1 - \frac{\mathrm{tr}(\mathbf{E}_{ilr})}{\mathrm{tr}(\mathbf{T}_{ilr})}, \quad R^2_{log} = 1 - \frac{\mathrm{tr}(\mathbf{E}_{log})}{\mathrm{tr}(\mathbf{T}_{log})},$$

thus reflect structural changes that arise by avoiding the scale invariance property of compositions, i.e. when log-transformation is applied instead of taking the orthonormal coordinates.

It is not difficult to demonstrate that there is no relation in general between both coefficients R^2_{ilr} and R^2_{log}. For this purpose, let us consider two matrices of response compositions,

$$\underline{\mathbf{y}}_1 = \begin{pmatrix} 1 & 5 & 1 \\ 9 & 2 & 2 \\ 1 & 8 & 3 \\ \mathbf{1} & 2 & 5 \end{pmatrix}, \quad \underline{\mathbf{y}}_2 = \begin{pmatrix} 1 & 5 & 1 \\ 9 & 2 & 2 \\ 1 & 8 & 3 \\ \mathbf{10} & 2 & 5 \end{pmatrix},$$

observed for the values $x = 1, 2, 3, 4$ of the explanatory variable. Note that both matrices differ just by the entry on the position (4,1). Though, by taking linear regression with an absolute term, the first case results in $R^2_{ilr} = 0.707 < 0.736 = R^2_{log}$, while in the second one $R^2_{ilr} = 0.788 > 0.674 = R^2_{log}$ is obtained.

Finally, it is worth to mention that the trace of any covariance matrix (residual or total) is equal to the mean of the distances between the samples and the centre on the simplex. This fact can be used to reformulate Theorem 4.3 in terms of distances in the respective spaces, if appropriate.

4.4 Illustrative example: reservoir sediments in the Czech Republic

The findings from the above sections are briefly illustrated using a geological data set from lacustrine sediments of the Nové Mlýny reservoir in the Czech Republic (underwater core NM1, WGS-84: $48°53'8.771''$N, $16°31'52.966''E$) [75].

Thirty-four samples from the core were air dried, manually ground in agate mortar and subjected to element composition analysis using Energy Dispersive X-ray Fluorescence (EDXRF) spectrometry. A PANalytical MiniPal 4.0 EDXRF spectrometer with a Peltier-cooled silicon drift energy dispersive detector (Institute of Anorganic chemistry in Řež, Prague) was used. Signals of Al and Si were acquired at 4 kV/200 μA with Kapton filter 151 under He flush; zn, Mn and Fe at 20 kV/100 μA with Al filter in air 152 and Rb and Pb at 30 kV/200 μA with Ag filter in air [59]. The EDXRF results are provided in counts per second (cps).

Fifteen elements Al, Si, P, Ti, K, Ca, Fe, Cr, Mn, Ni, Cu, zn, zr, Rb and Pb were selected for further statistical processing using regression analysis. The elements represent common lithophile elements, which are used for geochemical description of common parameters of sediments and sedimentary rocks, such as the grain size (Al, Si and Ti), degree of weathering (K, Al and Rb), heavy-mineral composition (zr, Ti, Fe), organic production (P, Ca, Cu, zn), redox state (Mn, Ni, Cu, zn) and anthropogenic impact by toxic compounds (Cr, Ni, zn, Pb).

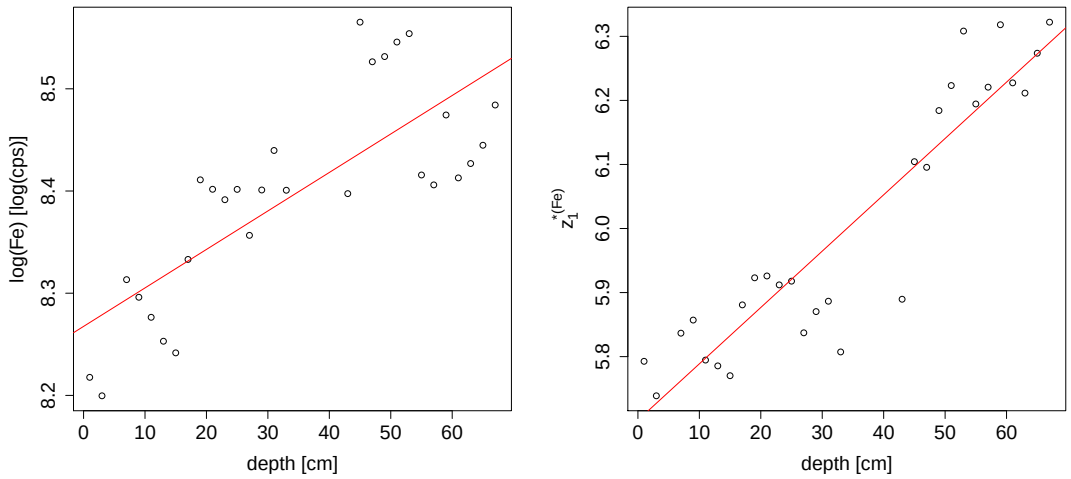

Figure 4.1: Regression model for iron using log-transformation (left) and using orthogonal coordinates (right).

In this concrete case, both absolute and relative information were of simultaneous interest; the total concentrations of the elements in the Nové Mlýny reservoir have been recently interpreted in [9]. Accordingly, in the following both log-ratio coordinates and log-transformed compositions were employed.

In addition to other site-specific geological tasks the aim was to investigate whether the distribution of these 15 elements in the core is random or organized. For this purpose linear regression models with the polynomial trend (up to the 4th-degree) in depth, and with the response composition in clr coordinates and log-transformed variables were taken. Particularly, the models 4.1 and

$$\log(\underline{\mathbf{y}}) = \mathbf{X}\mathbf{B}^{log} + \underline{\varepsilon}_{log},$$

where $\mathbf{B}^{log} = (b_{log,1}, \ldots, b_{log,15})$, were analyzed. The j-th row of the design matrix $\mathbf{X}$ was considered in the following forms

$$(1, \mathrm{depth}_j), \ \ldots, (1, \mathrm{depth}_j, \ldots, \mathrm{depth}_j^4).$$

In all cases the simplest possible model that was consistent with data was chosen.

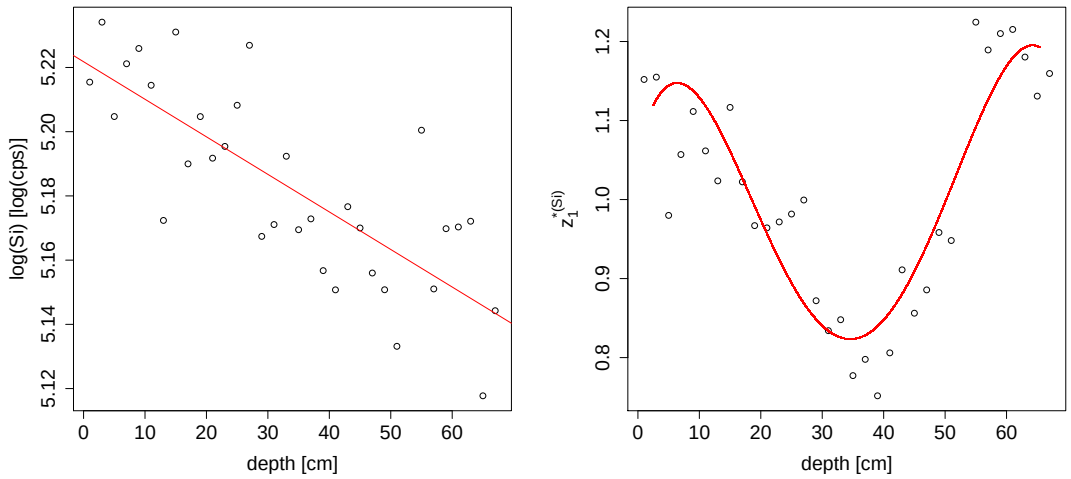

Figure 4.2: Regression model for iron using log-transformation (left) and using orthogonal coordinates (right).

By considering the regression outputs (realizations of test statistics F_l^{clr} and F-statistics to verify significance of the whole vector parameter $\mathbf{b}_{clr,l}$ and $\mathbf{b}_{log,l}$, respectively, T-statistics for significance testing of single regression parameters, p-values, coefficients of determination and visualization of data together with the corresponding regression functions), only zirconium (zr) didn't show any systematic pattern (i.e. does not change with changing depth) either for log-transformation or clr coordinate of the response. A systematic increase/decrease was observed in a majority of the elements but their clr coordinates usually indicate a more complex (polynomial) underlying pattern.

A typical example is Fe (Figure 4.1) in which an increasing trend was observed. For an easier interpretation, the response was expressed in orthogonal coordinates. From regression outputs (the slope parameter estimate was $8.796 \cdot 10^{-3}$ with standard error $0.710 \cdot 10^{-3}$ and p-value $\ll 0.001$, $MSE_{ilr} = 0.049$, $R^2_{ilr} = 0.845$) it can be concluded that by the increase of depth by 1 cm the ratio of Fe to the geometric mean of the other 14 elements increases approximately once ($\delta = 1.009146$ times, it means 1%); similarly, by considering log-transformed

response (the slope parameter estimate was $3.766 \cdot 10^{-3}$ with standard error $0.580 \cdot 10^{-3}$ and p-value $\ll 0.001$, $MSE_{log} = 0.060$, $R^2_{log} = 0.609$), the increase of depth by 1 cm means that the absolute amount of Fe (in cps) grows approximately once, $\exp\{3.766 \cdot 10^{-3}\} = 1.003773$. From the lower value of MSE_{ilr} than MSE_{log} as indicated by Theorem 4.3 one can in general conclude that for given scales the MSE values show always better fit in the ilr space (that would be no longer the case for scaling-free R^2 values). On the other hand, data configuration for the clr representation suggests that the linear trend could be enhanced by a more complex regression function, here polynomial of degree four. An extreme case of this general feature is Si (Figure 4.2), in which the linear trend for the response in clr coordinates is replaced by the polynomial one of degree four.

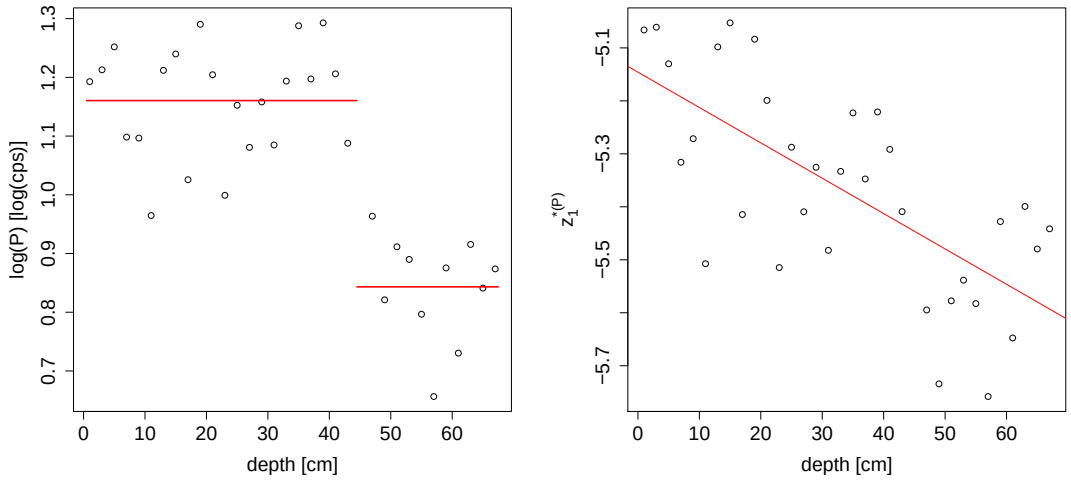

Figure 4.3: Regression model for iron using log-transformation (left) and using orthogonal coordinates (right).

It is important to mention that the depth range from 45 to 55 cm in the NM1 core is a transitional zone between lower pre-dam fluvial sediments and upper, fully dam-reservoir ones [76]. This layer, strongly enriched in organic carbon, has critical effect on the depth distribution of various elements, including P (sensitive

to organic productivity), Si (sensitive to grain size) and Fe (sensitive to redox conditions) (Figures 4.1, 4.2, 4.3). Distribution of these elements shows breaks at the base or on top of this layer, i.e. at 55 or 45 cm depth, which can be explained by their different geochemical behaviour. In particular, Si break is related to decrease of grain size at a break in sedimentation style from fluvial to lacustrine, Fe peak between 45 and 55 cm depth is likely related to diagenetic sulphide precipitation under dysoxic/anoxic conditions (high organic carbon) and P is related to increased organic productivity in water column of the lake.

Consequently, linear (= continuous in depth) regression trends are less likely than those represented by a polynomial function (= discontinuous in depth). In this respect, the clr data provide a better representation of the core stratigraphy. Mathematically, this effect can be easily explained by the remaining elements in the composition, which are incorporated in the denominator of the centred log-ratio. This facilitates identifying geochemical patterns related to the geochemical matrix in which the particular element is contained. On the other hand, there are also some exceptions, like for phosphorus (P, Figures 4.3), where this change seems to be better reflected by the log-transformed response (accordingly, even two constant lines instead of one regression line were taken). In this case, the piecewise constant model with the j-th row of the design matrix given as $(1, I[\text{depth}_j \geq 45 \text{ cm}])$, where the symbol $I[\text{depth}_j \geq 45 \text{ cm}]$ denotes a dummy variable coded 1 for the j-th measurements in the depth 45 cm and more, and 0 otherwise, fitted best the data.

Based on the purpose of the analysis, one can consider purely relative information, or to take also absolute abundances of positive data into account. Nevertheless, like here, such decision of the analyst should always follow also previous expert knowledge on possible underlying processes in data.

5. Calibration problem for compositional data

The last chapter of this dissertation thesis is devoted to the calibration problem for compositions. On the very beginning we show that indeed the TLS and the linear model with type-II constraints lead to the same estimates. The next section handles a calibration for compositional measurements. The calibration is usually used to express a linear relationship between errorless measurements obtained by two methods (or, alternatively, by two measuring devices). In the subsequent section, we derive the analogy between the compositional variation array and the matrices of the predicted values and residual variances from univariate calibrations, which are useful in descriptive statistics for compositions. Consequently, tests for verification of conformity between two methods of measurement are proposed. The quality of the above mentioned tests is verified by means of simulations. Lastly, the theoretical results are applied to a real-world example from biochemistry. This section is based on the papers [19, 20].

5.1 Equivalence between total least squares regression and linear model with type-II constraints

Section 2.2 contents three possible approaches for estimation of the unknown parameters in a model where the linear relationship between the compositional parts is analysed. On the beginning we have presented the standard approaches

to estimation, namely the TLS and the maximum likelihood method. Asymptotic character of these approaches being their main disadvantage property has been overcome when a calibration approach based on the linear model with type-II constraints is used instead. In the following theorem we will point out on the equivalence between the two mentioned approaches to TLS regression estimation, namely the TLS approach and the linear model approach. This subsection relies on the article [19]. As before, we assume that random errors have the same variance σ^2.

Theorem 5.1. *Let us consider the TLS regression model that is given by (2.10) and (2.11) where z_{1i} and z_{2i}, $i = 1, 2, \ldots, n$, are independent random variables with the same variance σ^2. The estimates of the calibration line coefficients, obtained from the iterative algorithm (2.19), converge to the TLS estimates given by relations (2.12) and (2.13). Moreover, under the assumption of normality, the estimates from the iterative algorithm converge to the maximum likelihood estimates.*

Proof. The estimators obtained from the iterative algorithm are the BLUEs in the linearized model (2.19). The nonlinear model (2.16) can also be expressed as

$$\begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu} \\ a\mathbf{1}_n + b\boldsymbol{\mu} \end{pmatrix} + \boldsymbol{\varepsilon}, \qquad \operatorname{var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I},$$

or, simply as

$$\mathbf{Z} = \mathbf{f}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon},$$

where $\mathbf{Z} = (\mathbf{z}_1', \mathbf{z}_2')'$, $\boldsymbol{\theta} = (a, b, \mu_1, \cdots, \mu_n)'$, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1', \boldsymbol{\varepsilon}_2')'$, and finally $\mathbf{f}(\boldsymbol{\theta})$ is a nonlinear function of the unknown parameter $\boldsymbol{\theta}$. Hence, the least squares minimization function is

$$\sum_{i=1}^{2n} \left(Z_i - f_i(\boldsymbol{\theta})\right)^2 = \sum_{i=1}^{n} \left[(z_{1i} - \mu_i)^2 + (z_{2i} - a - b\mu_i)^2\right]. \tag{5.1}$$

Since the model is nonlinear, we linearize the function $\mathbf{f}(\boldsymbol{\theta})$ by the Taylor series

locally at $\boldsymbol{\theta}^{(0)}$, when the second and higher derivatives are neglected. Thus, the resulting linearized model is given by

$$\mathbf{Z} = \mathbf{f}\left(\boldsymbol{\theta}^{(0)}\right) + \boldsymbol{\varphi}^{(0)} \triangle\boldsymbol{\theta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varphi}^{(0)} = \dfrac{\partial \mathbf{f}\,(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)}}$ and $\triangle\boldsymbol{\theta} = \boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}$. Now the BLUE of $\triangle\boldsymbol{\theta}$ can be derived by the least squares method as

$$\widehat{\triangle\boldsymbol{\theta}} = \left(\left[\boldsymbol{\varphi}^{(0)}\right]' \boldsymbol{\varphi}^{(0)}\right)^{-1} \left[\boldsymbol{\varphi}^{(0)}\right]' \left[\mathbf{Z} - \mathbf{f}\left(\boldsymbol{\theta}^{(0)}\right)\right].$$

Hence, $\widehat{\boldsymbol{\theta}} = \widehat{\triangle\boldsymbol{\theta}} + \boldsymbol{\theta}^{(0)}$. If $\widehat{\triangle\boldsymbol{\theta}}^{(k)}$ is calculated in the kth iteration from the iterative algorithm, the values of $\boldsymbol{\theta}^{(0)}$ are determined according to (2.26) when the estimated values of $\boldsymbol{\theta}$ from the $(k-1)$th iteration are used. Thus, the estimate in the kth iteration is

$$\widehat{\boldsymbol{\theta}}^{(k)} = \widehat{\triangle\boldsymbol{\theta}}^{(k)} + \boldsymbol{\theta}^{(0)}.$$

If the starting point $\boldsymbol{\theta}^{(0)}$ is sufficiently good chosen, then the iterative algorithm converges, i.e., $\widehat{\triangle\boldsymbol{\theta}}^{(k)}$ converges to zero and $\widehat{\boldsymbol{\theta}}^{(k)}$ converges to a point that minimizes (5.1).

The TLS estimators minimize, over all a and b, the quantity

$$\sum_{i=1}^{n} \left[(z_{1i} - \widehat{\mu}_i)^2 + (z_{2i} - \widehat{\nu}_i)^2\right] = \sum_{i=1}^{n} \left[(z_{1i} - \mu_{1i})^2 + (z_{1i} - a - b\mu_i)^2\right], \qquad (5.2)$$

where $(\widehat{z}_{1i}, \widehat{z}_{2i})$ given by (2.14) and (2.15) is the closest point to an observed point (z_{1i}, z_{2i}) on the calibration line $\nu_i = a + b\mu_i$.

The functions (5.1) and (5.2) minimize the same problem and, thus if the iterative algorithm converges, obtained estimates of the calibration line coefficients converge to the TLS estimates. The rest of the proof follows from the fact that under normality the maximum likelihood and the TLS estimators are the same.

□

5.2 Calibration problem

A calibration line describes the linear relationship between the errorless measurements obtained by measuring the same object by two different methods. Algorithms for fitting a calibration line and statistical inference in case of three-part compositions are derived in [29, 31] and these are presented in Section 2.2. For D-part compositions, the calibration problem can be partitioned into $D(D-1)/2$ partial calibration problems, performed on log-ratios of compositional parts. In other words, the calibration is performed for the corresponding coordinate of all possible two-part subcompositions separately.

Let there be n different compositions that have D parts which are measured using two methods A and B with the same precision. Let $\mathcal{R} = \{r = 1, 2, \ldots, D-1,\ s = r+1,\ r+2, \ldots, D\}$ be the set of subscripts. For two-part subcompositions (y_r, y_s) and (w_r, w_s), corresponding to the measurements obtained by methods A and B, respectively, the log-ratios are formed and arranged in data matrices

$$(\mathbf{Z}^A_{rs}, \mathbf{Z}^B_{rs}) = \begin{pmatrix} \ln \frac{y_{1r}}{y_{1s}} & \ln \frac{w_{1r}}{w_{1s}} \\ \ln \frac{y_{2r}}{y_{2s}} & \ln \frac{w_{2r}}{w_{2s}} \\ \vdots & \vdots \\ \ln \frac{y_{nr}}{y_{ns}} & \ln \frac{w_{nr}}{w_{ns}} \end{pmatrix}, \tag{5.3}$$

where $(r, s) \in \mathcal{R}$ (note that multiplying the log-ratios by $1/\sqrt{2}$ ilr coordinate would be formed). Let us assume that $\mathbf{Z}^A_{rs}$ and $\mathbf{Z}^B_{rs}$ represent a realization of a normally distributed n-dimensional random vector $\mathbf{z}^A_{rs} \sim N_n(\boldsymbol{\mu}_{rs}, \sigma^2\mathbf{I}_n)$ and $\mathbf{z}^B_{rs} \sim N_n(\boldsymbol{\nu}_{rs}, \sigma^2\mathbf{I}_n)$, respectively. Let $\boldsymbol{\mu}_{rs} = (\mu_{1rs}, \mu_{2rs}, \ldots, \mu_{nrs})'$ denote the vector of errorless measurement results of $\mathbf{z}^A_{rs}$, and $\boldsymbol{\nu}_{rs} = (\nu_{1rs}, \nu_{2rs}, \ldots, \nu_{nrs})'$ the vector of errorless measurement results of $\mathbf{z}^B_{rs}$, where $(r, s) \in \mathcal{R}$. Moreover, these measurement results are taken to be mutually independent. Thus, the calibration line (2.16) can be expressed as

$$\boldsymbol{\nu}_{rs} = a_{rs}\mathbf{1}_n + b_{rs}\boldsymbol{\mu}_{rs}, \tag{5.4}$$

where $(r, s) \in \mathcal{R}$, and $\mathbf{1}_n$ stands for the vector of n ones. The parameter a_{rs}

represents a systematic deviation of log-ratios between parts r and s obtained by measurement methods A and B, and b_{rs} denotes the scaling factor between them.

The BLUE of $\boldsymbol{\mu}_{rs}$, $\boldsymbol{\nu}_{rs}$, a_{rs} and b_{rs}, denoted as $\widehat{\boldsymbol{\mu}}_{rs}$, $\widehat{\boldsymbol{\nu}}_{rs}$, $\widehat{a}_{rs}$ and $\widehat{b}_{rs}$, are obtained by the iterative procedure described in Section 2.2 [29], applied to single log-ratios. For the estimation of the unknown model parameters and of the variance - covariance matrices we use the formulas (2.17) - (2.24). Of course, they can be used after appropriate adjustment i.e. after substituting $\mathbf{z}_1$ and $\mathbf{z}_2$ by $\mathbf{z}_{rs}^A$ and $\mathbf{z}_{rs}^B$ respectively, a by a_{rs}, b by b_{rs}, $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ by $\boldsymbol{\mu}_{rs}$ and $\boldsymbol{\nu}_{rs}$ respectively and as well substituting their initial values $\boldsymbol{\mu}^{(0)}$ and $\boldsymbol{\nu}^{(0)}$ by $\boldsymbol{\mu}_{rs}^{(0)}$ and $\boldsymbol{\nu}_{rs}^{(0)}$ respectively.

5.3 Exploratory analysis of calibration results

From the theoretical point of view, it is interesting that the fitted calibration lines can be also used to predict the values of the method B by the method A and vice versa. For this purpose, let us define the matrices of the predicted averages $\mathbf{M}^{(j)}$, $j = 1, 2$, as

$$\mathbf{M}^{(j)} = \begin{pmatrix} 0 & \widehat{m}_{12}^{(j)} & \widehat{m}_{13}^{(j)} & \cdots & \widehat{m}_{1D}^{(j)} \\ \widehat{m}_{21}^{(j)} & 0 & \widehat{m}_{23}^{(j)} & \cdots & \widehat{m}_{2D}^{(j)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \widehat{m}_{D1}^{(j)} & \widehat{m}_{D2}^{(j)} & \widehat{m}_{D3}^{(j)} & \cdots & 0 \end{pmatrix},$$

where for $j = 1$ the elements of $\mathbf{M}^{(1)}$ are the averages of method B as predicted by the averages of method A. Particularly, elements of $\mathbf{M}^{(1)}$ are defined as predicted averages using the fitted calibration line (5.4) when the parameters are marked with superscript (1),

$$\widehat{m}_{rs}^{(1)} = \widehat{a}_{rs}^{(1)} + \widehat{b}_{rs}^{(1)} \overline{z_{rs}^A}, \quad \overline{z_{rs}^A} = \frac{1}{n} \sum_{i=1}^{n} \ln \frac{x_{ir}}{x_{is}}. \tag{5.5}$$

Conversely, the elements of $\mathbf{M}^{(2)}$ are the averages of method A as predicted by the averages of method B, i.e., the elements of $\mathbf{M}^{(2)}$ are defined as predictions using the fitted calibration line

$$\boldsymbol{\mu}_{rs} = a_{rs}^{(2)} \mathbf{1}_n + b_{rs}^{(2)} \boldsymbol{\nu}_{rs},$$

specifically

$$\widehat{m}_{rs}^{(2)} = \widehat{a}_{rs}^{(2)} + \widehat{b}_{rs}^{(2)}\overline{z_{rs}^{B}}, \quad \overline{z_{rs}^{B}} = \frac{1}{n}\sum_{i=1}^{n} \ln \frac{y_{ir}}{y_{is}}. \tag{5.6}$$

Further, the matrix of residual variances is defined as

$$\mathbf{T}_* = \begin{pmatrix} 0 & \widehat{\sigma}_{12}^2 & \widehat{\sigma}_{13}^2 & \cdots & \widehat{\sigma}_{1D}^2 \\ \widehat{\sigma}_{21}^2 & 0 & \widehat{\sigma}_{23}^2 & \cdots & \widehat{\sigma}_{2D}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \widehat{\sigma}_{D1}^2 & \widehat{\sigma}_{D2}^2 & \widehat{\sigma}_{D3}^2 & \cdots & 0 \end{pmatrix},$$

where $\widehat{\sigma}_{rs}^2$ is the estimate of the residual variance (2.25) for $r, s = 1, 2, \ldots, D$, $r \neq s$.

In the following, some properties of matrices $\mathbf{M}^{(1)}$, $\mathbf{M}^{(2)}$ and $\mathbf{T}_*$ are established that will reveal their close connection to the variation array, described in Section 1.3. These matrices are useful for exploratory analysis of calibration results and for testing conformity of two methods as well, which will be discussed in more detail in the next section.

Lemma 5.1. *Elements $\widehat{m}_{rs}^{(1)}$ and $\widehat{m}_{rs}^{(2)}$ are sample means of the log-ratios of the measurements of the parts (r, s) obtained by the measuring devices B and A, respectively, i.e., $\widehat{m}_{rs}^{(1)} = \overline{z_{rs}^{B}}$ and $\widehat{m}_{rs}^{(2)} = \overline{z_{rs}^{A}}$.*

Proof. By substituting the relation of $\widehat{a}_{rs}^{(1)}$ given by (2.12) into (5.5) we obtain the statement for the elements of the matrix $\mathbf{M}^{(1)}$. Validity for the matrix $\mathbf{M}^{(2)}$ can be shown in the same way. □

As a direct consequence of Lemma 5.1 the following properties are fulfilled:

i) For the elements of the matrices $\mathbf{M}^{(j)}$, $j = 1, 2$, the triangular equality holds, i.e.,

$$\widehat{m}_{rs}^{(j)} = \widehat{m}_{rl}^{(j)} + \widehat{m}_{ls}^{(j)}, \; r, s, l = 1, 2, \ldots, D.$$

ii) Matrices $\mathbf{M}^{(j)}$, $j = 1, 2$, are antisymmetric, i.e., $\widehat{m}_{rs}^{(j)} = -\widehat{m}_{sr}^{(j)}$ and $\widehat{m}_{rr}^{(j)} = 0$, $r, s = 1, 2, \ldots, D$.

Similarly, it is a direct consequence of the logarithm properties that the matrix of residual variances $\mathbf{T}_*$ is symmetric. Thus we can conclude that the elements of the above matrices have the same properties as the elements of the variation array.

The above findings can be used for descriptive statistics based on the results of the calibration problem. In particular, we can compare elements of matrices $\mathbf{M}^{(1)}$ and $\mathbf{M}^{(2)}$. If the elements are the same, or nearly so, then both methods yield on average the same values for observations with the given compositional parts (recall the fact that all relevant information in a composition is contained in log-ratios of the parts). Taking just the i-th row/column of $\mathbf{M}^{(1)} - \mathbf{M}^{(2)}$ (or its sum) into account, we can observe this effect for the i-th compositional part. In addition, the smaller the values in $\mathbf{T}_*$, the stronger the systematic relation between the corresponding log-ratios of the two methods. Summing up the rows/columns of $\mathbf{T}_*$, this pattern can be compared for the single compositional parts. Finally, in the next section some tests for conformity of two measurement methods will be introduced.

5.4 Tests for conformity of two measurement methods

When matrices $\mathbf{M}^{(1)}$, $\mathbf{M}^{(2)}$ and $\mathbf{T}_*$ are estimated, it is natural to use them to construct tests, related either to log-ratios or to the original compositional parts, that handle the calibration problem. In the following we will assume that the sample of ilr transformed compositions follow a normal distribution. As a simple consequence, each log-ratio of the composition follows a normal distribution as well. Basic tasks in a calibration problem are to test whether two measurement methods give the same results and whether they measure with some predetermined precision. Here we propose five different tests that can be splitted into three families. The first family includes tests F_{rs}, T^1_{rs} and T^2_{rs}; the second one corresponds to T_{rs}; and the third one to C_{rs}. The first two families

are suitable for identification of a significant systematic difference between results obtained by two methods; the third one for verification that the methods follow the same prescribed precision.

Two measurement methods give the same result if and only if the calibration line passes through the origin at angle of 45 degrees. As we have explained in the previous section, for D-part compositions we fit $D(D-1)/2$ partial calibrations lines. Hence, the problem is to test the null hypothesis that all intercepts $a_{rs}^{(1)} = 0$ and all slopes $b_{rs}^{(1)} = 1$, $(r,s) \in \mathcal{R}$, simultaneously. The test statistic for each hypothesis $H_{0rs} : a_{rs}^{(1)} = 0, b_{rs}^{(1)} = 1$ individually, according to [31], is given as

$$F_{rs} = \left[\begin{pmatrix} \widehat{a}_{rs}^{(1)} \\ \widehat{b}_{rs}^{(1)} \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right]' \left[\widehat{\text{var}} \begin{pmatrix} \widehat{a}_{rs}^{(1)} \\ \widehat{b}_{rs}^{(1)} \end{pmatrix} \right]^{-1} \left[\begin{pmatrix} \widehat{a}_{rs}^{(1)} \\ \widehat{b}_{rs}^{(1)} \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right]. \tag{5.7}$$

The symbol $\widehat{\text{var}}[(\widehat{a}_{rs}^{(1)}, \widehat{b}_{rs}^{(1)})']$ stands for the covariance matrix of the estimator $(\widehat{a}_{rs}^{(1)}, \widehat{b}_{rs}^{(1)})'$, where the estimated value of dispersion σ_{rs}^2 (2.25) is plugged into the formula (2.24). Under the null hypothesis, the statistic F_{rs} is distributed as $F_{2,n-2}$. For testing the whole set of hypotheses H_{0rs}, $(r,s) \in \mathcal{R}$, simultaneously, it is necessary to use some techniques for multiple comparisons. In order to retain a prescribed significance level α for all tests simultaneously, the significance level for each test must be less than α. The Bonferroni-adjusted α-level of significance $\alpha_{adj} = \frac{2\alpha}{D(D-1)}$ for each test is one of the most commonly used approaches. Applying the Bonferroni correction, we reject H_{0rs} when $f_{rs} \geq F_{2,n-2}\,(1 - \alpha_{adj})$, where f_{rs} is a realization of the test statistic F_{rs}. Equivalently, the hypothesis H_{0rs} is rejected if p-value $\leq \alpha_{adj}$.

In the case we reject some of H_{0rs} we want to identify which of the components $a_{rs}^{(1)}$ or $b_{rs}^{(1)}$ is responsible for the rejection. This is done by performing multiple tests of intercepts and slopes separately. Therefore, in order to test whether the intercepts are significantly different from zero, i.e., to test $a_{rs}^{(1)} = 0$, $(r,s) \in \mathcal{R}$,

simultaneously, the test statistic [31]

$$T_{rs}^1 = \frac{\widehat{a}_{rs}^{(1)} \sqrt{n \left[\widehat{\boldsymbol{\mu}}_{rs}\right]' \widehat{\boldsymbol{\mu}}_{rs} - \left[\mathbf{1}'\widehat{\boldsymbol{\mu}}_{rs}\right]^2}}{\widehat{\sigma}_{rs} \sqrt{\left(\widehat{b}_{rs}^{(1)} + 1\right) \left[\widehat{\boldsymbol{\mu}}_{rs}\right]' \widehat{\boldsymbol{\mu}}_{rs}}}, \tag{5.8}$$

can be used. Under the null hypothesis T_{rs}^1 has the t_{n-2} distribution. The hypothesis $a_{rs}^{(1)} = 0$ is rejected if $|t_{rs}^1| \geq t_{n-2}\left(1 - \frac{\alpha_{adj}}{2}\right)$, where t_{rs}^1 is a realization of T_{rs}^1, or else if p-value$\leq \alpha_{adj}$. The test that the slopes are significantly different from 1, i.e., to test $b_{rs}^{(1)} = 1$, $(r,s) \in \mathcal{R}$, simultaneously, is done by using the test statistic [31]

$$T_{rs}^2 = \frac{\left(\widehat{b}_{rs}^{(1)} - 1\right) \sqrt{n \left[\widehat{\boldsymbol{\mu}}_{rs}\right]' \widehat{\boldsymbol{\mu}}_{rs} - \left[\mathbf{1}'\widehat{\boldsymbol{\mu}}_{rs}\right]^2}}{\widehat{\sigma}_{rs} \sqrt{n \left(\widehat{b}_{rs}^{(1)} + 1\right)}}, \tag{5.9}$$

which is distributed as t_{n-2}, if the hypothesis is true. Conditions for rejection of hypotheses are the same as for the intercept.

Another way of hypothesis testing that the two measurement methods give the same results is based on matrices of predicted averages $\mathbf{M}^{(1)}$ and $\mathbf{M}^{(2)}$. Both methods give the same results if and only if the matrices $\mathbf{M}^{(1)}$ and $\mathbf{M}^{(2)}$ are the same. Hence, the problem is to test the hypothesis that all elements of matrices $\mathbf{M}^{(1)}$ and $\mathbf{M}^{(2)}$ are the same, i.e., to test $\mathrm{H}_{0rs}^{'}$: $\overline{\mu}_{rs} = \overline{\nu}_{rs}$ for all $(r,s) \in \mathcal{R}$, simultaneously. Thus, according to Lemma 5.1, test statistics are of the form

$$T_{rs} = \frac{\widehat{m}_{rs}^{(1)} - \widehat{m}_{rs}^{(2)}}{\widehat{\sigma}_{rs}} \sqrt{n/2}, \tag{5.10}$$

$(r,s) \in \mathcal{R}$. Under the null hypothesis the test statistic T_{rs} follows the t_{n-2} distribution. We reject $\mathrm{H}_{0rs}^{'}$ if $|t_{rs}| \geq \mathrm{t}_{n-2}\left(1 - \frac{\alpha_{adj}}{2}\right)$, where t_{rs} is a realization of the test statistic T_{rs}.

The testing procedures, introduced above, can be split into two natural families. The first family is formed by tests F_{rs}, T_{rs}^1, T_{rs}^2, the second by tests T_{rs}. The second family with the test T_{rs} represents a new proposed alternative to the

first family with the test F_{rs}. For both proposed families of tests, if we reject the null hypothesis, then there exist significant systematic differences between the methods. By proposed multiple comparisons, we directly obtain for which log-ratios of parts the methods give different values. Moreover, for the test F_{rs} we also employ the fitted calibration line, i.e., the line that describes a linear relationship between log-ratios of parts for both methods. On the contrary, the advantage of tests T_{rs} is their simplicity. They only require calculation of sample means and sample deviations of log-ratios of parts. However, under significant systematic differences between log-ratios of some parts for both methods, these tests do not provide information on the linear relationship expressed by the corresponding calibration lines. Quality of both the proposed test F_{rs} and T_{rs} will be compared in the following section.

Finally, with the third family of tests, we will discuss a problem of precision of two measurement devices (methods). In order to verify that the methods follow a prescribed precision (the same for both methods), we provide multiple tests on elements of matrix $\mathbf{T}_*$, $\mathrm{H}_{0rs}^{''}$: $\sigma_{rs}^2 = \sigma_{rs0}^2$ versus $\mathrm{H}_{1rs}^{''}$: $\sigma_{rs}^2 \neq \sigma_{rs0}^2$. Here we use χ^2-tests on variance,

$$C_{rs} = \widehat{\sigma}_{rs}^2 \frac{n-2}{\sigma_{rs0}^2}. \tag{5.11}$$

In the case that $\mathrm{H}_{0rs}^{''}$ is true the test statistic C_{rs} is distributed as χ_{n-2}^2. Hypothesis $\mathrm{H}_{0rs}^{''}$ is rejected if $c_{rs} \leq \chi_{n-2}^2\left(\frac{\alpha_{adj}}{2}\right)$ or $c_{rs} \geq \chi_{n-2}^2\left(1-\frac{\alpha_{adj}}{2}\right)$ for c_{rs} taken to be a realization of C_{rs}.

All the tests proposed in this section are uniformly more powerful if one uses, e.g., the modified Bonferroni adjustment, known as Holm-Bonferroni adjustment [39], which consists of a three step algorithm. In the first step p-values are computed and arranged in ascending order, thus constituting an increasing sequence of p-values, i.e., $p_{(1)} < p_{(2)} < \cdots < p_{(k)}$, where k is the number of comparisons made (here $D(D-1)/2$). Subsequently they are compared with the corresponding adjusted α-levels of significance calculated as $\frac{\alpha}{k-j+1}$, j is the position in the sequence of p-values. We are starting with comparing the first p-value with the

appropriate adjusted α-level of significance. The algorithm stops when it finds such a p-value that exceeds the adjusted α-level of significance, i.e., when it finds minimum j such that $p_{(j)} > \frac{\alpha}{k-j+1}$. Finally, in the last step a conclusion about the acceptance or rejection of the hypotheses is done, i.e., reject hypotheses corresponding with p-values $p_{(1)}, \ldots, p_{(j-1)}$ and do not reject the remaining hypotheses. For other possible methods for addressing multiple testing see, e.g., [11, 78].

5.5 Simulation study

In this section we compare the quality of the both proposed tests F_{rs} and T_{rs} by simulations. We will explore probability of rejecting the conformity between the two methods with respect to chosen difference between the results of measurement. In the first study, the differences are added to the log-ratios between parts to identify properties of the proposed tests in general. In the second study, the differences come from perturbation of compositions with three, five and ten parts to verify the proposed methodology for calibration of compositional measurements.

We will consider 30 samples of three-part compositions that are obtained by two different methods A and B. The data matrices $\mathbf{z}_{rs}^A$ and $\mathbf{z}_{rs}^B$ were generated in a natural way, a normally distributed error term was added to the true mean. Observations were considered independent and each having the same precision $\sigma = 0.1$ and $\sigma = 0.5$. The true mean values $\boldsymbol{\mu}_{rs}$ of log-ratios between parts r and s for the method A were considered such that $\boldsymbol{\mu}_{12}$ takes on values from the interval $(2.1, 2.8)$, $\boldsymbol{\mu}_{13} = \boldsymbol{\mu}_{12} + 1$ and $\boldsymbol{\mu}_{23} = \boldsymbol{\mu}_{12} - 1$. The true mean values $\boldsymbol{\nu}_{rs}$ for the method B were considered either the same as for the method A, i.e., $\boldsymbol{\nu}_{rs} = \boldsymbol{\mu}_{rs}$, or as linear functions of $\boldsymbol{\mu}_{rs}$. Coefficients of linear functions were chosen in the interval $\langle 2, 4 \rangle$.

5000 simulations were done for each case and hypotheses $H_{0rs} : a_{rs} = 0$, $b_{rs} = 1$ (test F_{rs}) and $H'_{0rs} : \bar{\mu}_{rs} = \bar{\nu}_{rs}$ (test T_{rs}) for $r = 1, 2$, $s = 2, 3$, $r < s$,

simultaneously, were tested on the significance level 0.05. Obtained empirical probabilities of rejection of the null hypotheses for the case when mean values of log-ratios of parts were considered the same for both methods are presented in Table 5.1. In the other cases, mean values $\boldsymbol{\nu}_{rs}$ of the log-ratios of parts for method B were considered as linear functions of the mean values $\boldsymbol{\mu}_{rs}$ for method A and hypotheses were rejected in all provided tests regardless of adjustment and precision ($0.05/3 = 0.017$ for the Bonferroni-adjusted 0.05-level of significance; $0.05/3 = 0.017$, $0.05/2 = 0.025$ and 0.05 for the Holm-Bonferroni adjustments).

	Bonferroni adjustment				Holm-Bonferroni adjustment			
	$\sigma = 0.1$		$\sigma = 0.5$		$\sigma = 0.1$		$\sigma = 0.5$	
(r, s)	T_{rs}	F_{rs}	T_{rs}	F_{rs}	T_{rs}	F_{rs}	T_{rs}	F_{rs}
$(1, 2)$	0.0190	0.0260	0.0228	0.4030	0.0190	0.0260	0.0228	0.5198
$(1, 3)$	0.0190	0.0260	0.0228	0.4030	0.0228	0.0360	0.0336	0.3136
$(2, 3)$	0.0140	0.0250	0.0248	0.3950	0.0140	0.0280	0.0262	0.4228

Table 5.1: Empirical probabilities of rejecting the hypothesis $H_{0rs} : a_{rs} = 0$, $b_{rs} = 1$ (test F_{rs}) and $H'_{0rs} : \bar{\mu}_{rs} = \bar{\nu}_{rs}$ (test T_{rs}) for $r = 1, 2$, $s = 2, 3$, $r < s$, simultaneously, on significance level 0.05 for data simulated with the same mean values for both methods.

In Table 5.1 we can see that the results differ for different accuracy of observations. For relatively precise observations (parameters are estimated with sufficient precision), both proposed tests give similar results with the Bonferroni and Holm-Bonferroni adjustments. Conformity of the two methods between log-ratios of two parts is rejected with probability less than 0.023 and 0.036 for tests T_{rs} and F_{rs}, respectively. The test T_{rs} also gives a small empirical probability of rejecting conformity for less precise observations ($\sigma = 0.5$), less than 0.034 for each partial test. This is due to the fact that this statistic only tests the conformity of averages. On the contrary, the statistic F_{rs} tests whether the calibration line passes through the origin at an angle of 45 degrees. Simulation for the mean values from the short interval with greater variance shows that although the data were simulated with the same mean for both methods, the greater variance results in a significant difference in both estimated mean values and calibration line. The test T_{rs} does not detect this, i.e., mean value averages are not significantly

different.

Summarizing obtained results, both proposed tests detect with certainty the situation in which methods disagree. For relatively precise observations, both tests give similar results and can be used for verification of conformity of methods. However, for cases with less accuracy, the test T_{rs} does not detect nonconformity sufficiently sensitively. In this case the test F_{rs} is more sensitive.

Now, we will investigate the probability of rejecting the conformity between the two methods with respect to differences $c > 0$ added to compositional measurements. Thirty samples of three-, five-, and ten-part compositions obtained by the two methods A and B were generated in the following way. In the case of three-part compositions, in the first step, two data matrices of the order 2×30 were generated from a two-dimensional normal distribution with the mean equal to equidistant 30 points from the interval $\langle -3, 3\rangle$ and with the variance matrix $0.01\mathbf{I}$. Next, the inverse ilr transformation was applied to obtain two samples with sample size 30 of three-part compositions (denoted as $\mathbf{y}_A$ and $\mathbf{y}_B$). Finally, the compositions $\mathbf{y}_B$ were perturbated by $(c, 1, 1)$ with a constant c from the interval $(0, 10)$. If $c = 1$, compositions $\mathbf{y}_B$ do not change after the perturbation, and thus the compositional measurements obtained by methods A and B are the same. If $c \neq 1$, the first part of compositions $\mathbf{y}_B$ is multiplied by c and other parts remain unchanged, i.e., compositional measurements obtained by the methods A and B are different. Then the proposed procedure for the calibration was used. Analogously we proceeded with five- and ten-part compositions.

Similarly as in the first study, 5000 simulations were done and conformity of the two methods were tested on significance level 0.05. Recall that only log-ratios with the part y_1 were affected by the constant c; remaining log-ratios are unchanged for both methods. Thus, if log-ratios of parts include y_1, differences between log-ratios of the corresponding compositional measurements for methods A and B are increasing for c going to zero or infinity; otherwise, they are stable. The boxplots of log-ratios of parts including and not including y_1 with respect to the constant c are demonstrated in Figure 5.1. Resulting empirical probabilities

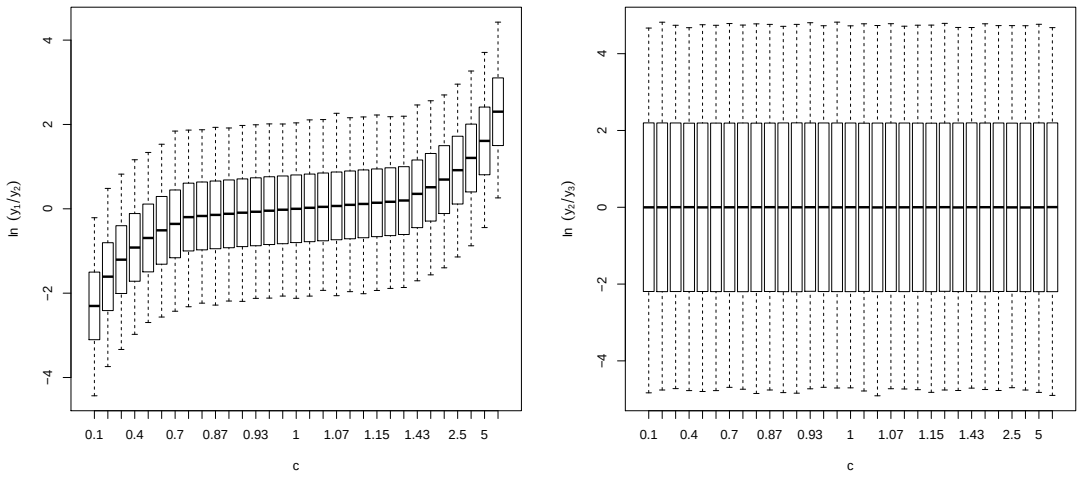

Figure 5.1: Boxplots of log-ratios of parts y_1 and y_2 (top), and y_2 and y_3 (bottom) with respect to constant c chosen for perturbation of composition (y_1, y_2, y_3) by $(c, 1, 1)$.

of rejecting conformity of three-part compositional measurements with respect to the constant c are presented in Figures 5.2 and 5.3.

In Figure 5.2, results for tests F_{rs} and T_{rs} on a single log-ratio between parts r and s on the 0.05/3=0.017 Bonferroni-adjusted 0.05-level of significance are demonstrated together with multiple tests F_{rs} and T_{rs} (tests on all log-ratios simultaneously). The power of tests for Holm-Bonferroni correction is the same for three-part compositions and therefore the results are not presented. We can see that probabilities of rejection remain stable for log-ratio of parts x_2 and x_3. The empirical probability of type I error for single tests F_{23}, T_{23} is not greater than 0.020 and 0.053 for the Bonferroni-adjusted and Holm-Bonferroni-adjusted level, respectively. The overall probability of type I error is 0.043 and 0.051 for the tests T_{rs} and F_{rs}, respectively. Thus the test T_{rs} is a bit conservative. Further, we can see that the tests F_{12}, F_{13} and T_{12}, T_{13} detect with high probability the situation in which log-ratios between parts y_1, y_2, and y_1, y_3, of the two methods differ. In the case of compositions with higher number of parts, more powerful corrections seem to be necessary.

In Figure 5.3 we can see that the power of both overall tests systematically

decrease with increasing number of parts. Furthermore, the overall probability of the type I error slightly increases; in the case of ten-part, it is less than 0.068 and 0.056 for test F_{rs} and T_{rs}, respectively. It is also easy to see that the test T_{rs} is more powerful. Because of symmetry of results for $c \in (0, 1)$ and $c > 1$ and easier comparability, only the case of $c \in (0.7, 1\rangle$ is plotted. Thus,the above results confirm that the proposed tests work well up to a moderate number of parts, nevertheless, for compositions with more parts an alternative approach seems to be necessary.

5.6 Illustrative example: blood plasma

Proposed approaches from the previous sections are used to analyze a real-world data set from clinical biochemistry that consists of 10 samples of blood plasma with concentrations (in mmol/l) of four selected amino-acids, alanine (part 1), glycine (part 2), leucine (part 3), and isoleucine (part 4), containing an aliphatic chain. These samples are obtained by two different methods: external standard (method A) and internal standard (method B) (see Table 5.2). The task is to analyze whether there is a significant difference between measurement results from the external and internal standards and whether both methods measure with the same given reasonable precision of 0.1mmol/l (a common standard error for all log-ratios of the composition). Note that with these data, not the absolute values of concentrations but rather their relative contributions to the overall composition of blood plasma are of interest. Thus, although the observations are not represented by a constant sum constraint (like 1 or 100), they are typical compositional data.

At the beginning, data matrices (5.3) of log-ratios were computed; in addition, their elements were multiplied by $1/\sqrt{2}$. Calibration lines for log-ratios of parts of internal standard (method B) subject to external standard (method A) are estimated by the iterative algorithm described in Section 2.2 with accuracy more than $\epsilon = 10^{-9}$. Note that the calibration lines are determined with different precisions (see Table 5.3).

External standard method

n	1	2	3	4
1	0.158	0.108	0.041	0.020
2	0.244	0.086	0.050	0.029
3	0.243	0.182	0.078	0.039
4	0.264	0.094	0.077	0.041
5	0.186	0.192	0.061	0.039
6	0.137	0.134	0.053	0.036
7	0.143	0.089	0.043	0.026
8	0.258	0.156	0.074	0.040
9	0.211	0.117	0.067	0.042
10	0.441	0.195	0.220	0.084

Internal standard method

n	1	2	3	4
1	0.449	0.235	0.168	0.044
2	0.560	0.151	0.165	0.050
3	0.379	0.217	0.175	0.047
4	0.455	0.124	0.190	0.053
5	0.324	0.256	0.152	0.052
6	0.237	0.178	0.131	0.047
7	0.329	0.157	0.142	0.045
8	0.508	0.236	0.211	0.059
9	0.355	0.145	0.163	0.054
10	0.649	0.194	0.307	0.097

Table 5.2: Measurement results of 10 samples of blood plasma with concentrations (in mmol/l) of four selected amino-acids, alanine (part 1), glycine (part 2), leucine (part 3), and isoleucine (part 4), from external standard method (left) and internal standard method (right).

Elements of the matrices $\mathbf{M}^{(j)}$, $j = 1, 2$, are calculated using (5.5) for $j = 1$ and (5.6) for $j = 2$. Elements of the matrix $\mathbf{T}_*$ are obtained by (2.25),

$$\mathbf{M}^{(1)} = \begin{pmatrix} 0 & 0.7947 & 0.8447 & 2.0350 \\ -0.7947 & 0 & 0.0500 & 1.2403 \\ -0.8447 & -0.0500 & 0 & 1.1903 \\ -2.0350 & -1.2403 & -1.1903 & 0 \end{pmatrix},$$

$$\mathbf{M}^{(2)} = \begin{pmatrix} 0 & 0.5119 & 1.1660 & 1.7629 \\ -0.5119 & 0 & 0.6542 & 1.2510 \\ -1.1660 & -1.5595 & 0 & 0.5969 \\ -1.7629 & -1.2510 & -0.5969 & 0 \end{pmatrix},$$

$$\mathbf{T}_* = \begin{pmatrix} 0 & 0.0007 & 0.0068 & 0.0001 \\ 0.0007 & 0 & 0.0032 & 0.0012 \\ 0.0068 & 0.0032 & 0 & 0.0080 \\ 0.0001 & 0.0012 & 0.0080 & 0 \end{pmatrix}.$$

By computing the difference between matrices $\mathbf{M}^{(1)}$ and $\mathbf{M}^{(2)}$,

$$\mathbf{M}^{(1)} - \mathbf{M}^{(2)} = \begin{pmatrix} 0 & 0.2828 & -0.3213 & 0.2721 \\ -0.2828 & 0 & -0.6042 & -0.0107 \\ 0.3213 & 1.5095 & 0 & 0.5934 \\ -0.2721 & 0.0107 & -0.5934 & 0 \end{pmatrix},$$

parts (r,s)	calibration line standard errors of $(\widehat{a}_{rs},\ \widehat{b}_{rs})$	iterations
$(1,2)$	$\mathbf{z}_{12}^B = 0.2629 + 1.0391\mathbf{z}_{12}^A$ $(0.0225,\ 0.0362)$	7
$(1,3)$	$\mathbf{z}_{13}^B = 0.0600 + 0.6729\mathbf{z}_{13}^A$ $(0.1710,\ 0.1441)$	15
$(1,4)$	$\mathbf{z}_{14}^B = 0.2601 + 1.0068\mathbf{z}_{14}^A$ $(0.0424,\ 0.0239)$	7
$(2,3)$	$\mathbf{z}_{23}^B = -0.5040 + 0.8469\mathbf{z}_{23}^A$ $(0.0486,\ 0.0651)$	9
$(2,4)$	$\mathbf{z}_{24}^B = -0.1072 + 1.0771\mathbf{z}_{24}^A$ $(0.0731,\ 0.0569)$	9
$(3,4)$	$\mathbf{z}_{34}^B = 0.9158 + 0.4599\mathbf{z}_{34}^A$ $(0.1229,\ 0.1992)$	25

Table 5.3: Estimates of the calibration line parameters for the amino-acids data.

and summing up absolute values of elements in rows (or columns) of $\mathbf{M}^{(1)} - \mathbf{M}^{(2)}$ (0.8762, 0.8977, 2.4242, 0.8762), we observe the highest differences between log-ratios with leucine (part 3). From the row sums of the variation matrix $\mathbf{T}_*$ (0.0076, 0.0051, 0.0180, 0.0093) we can tentatively conclude that the strongest relation between log-ratios of the two measurement methods corresponds to those with glycine (part 2).

Furthermore, we can interpret the elements $\widehat{m}_{rs}^{(1)}$ and $\widehat{m}_{rs}^{(2)}$ (predictions from the estimated calibration lines) as averages of the log-ratio between the parts r and s obtained by external standard and internal standard methods, respectively (see Lemma 5.1), e.g. $\widehat{m}_{12}^{(1)} = 0.7947$ represents the average of log-ratios between alanine and glycine obtained from external standard. The precision of both methods when considering the log-ratios between the parts r and s is given by $\widehat{\sigma}_{rs}$, e.g., $\widehat{\sigma}_{12} = \sqrt{0.0007} = 0.0265$ stands for the precision of the methods for log-ratio between alanine and glycine.

In order to get the required information about conformity of both methods, the proposed multiple tests from the previous section are applied. At first, it is necessary to verify normality for log-ratios of compositional data. The Shapiro-Wilk normality test with p-values of 0.6000, 0.6550, 0.9230, 0.4658, 0.6248, 0.4562

(external standard) and 0.4250, 0.2504, 0.9728, 0.6319, 0.8179, 0.6549 (internal standard) ensures that normality of transformed compositional samples cannot be rejected, although the sample size is not very large (as is common in medical applications). Next we can proceed to multiple tests for identification of systematic differences between both methods. The resulting values of the test statistics T_{rs} and F_{rs}, calculated according to (5.7) and (5.10), respectively, are displayed in Table 5.4.

It is easy to see that p-values of the test statistics T_{rs} and F_{rs}, except for the log-ratio between glycine and isoleucine, are less than 0.0001. Each of these p-values is thus less than the Bonferroni adjusted 0.05-level of significance, $0.0083 = \frac{0.05}{6}$, and also smaller than the Holm-Bonferroni adjusted 0.05-levels of significance (0.0083, 0.0100, 0.0125, 0.0167, 0.0250, 0.0500). Hence, there is a statistically significant difference between the results from the external and internal standards; it means that the two methods do not give us the same results when all log-ratios of parts are measured except, for the pair $(2, 4)$.

The differences within log-ratios (considering both methods) are modeled using the calibration lines given in Table 5.3. It remains to verify whether the calibration line can be simplified, i.e., to reveal the source of differences between the methods. For this purpose we use the multiple t-tests using the test statistics T^1_{rs} and T^2_{rs}, given by (5.8) and (5.9), respectively; results are collected in Table 5.4. The hypothesis that the slope of calibration line is equal to 1 cannot be rejected (all p-values of the test statistics T^2_{rs} are higher than the Bonferroni-adjusted, or the Holm-Bonferroni-adjusted 0.05-level of significance); however the intercept for most of the log-ratios is significantly different from zero (p-values of $T^1_{12}, T^1_{14}, T^1_{23}$ and T^1_{34} are less than 0.0083 - the Bonferroni adjustment, and also they are less than 0.0083, 0.0100, 0.0125 and 0.0167, respectively - the Holm-Bonferroni adjustment). In particular, the log-ratios between alanine and glycine, alanine and isoleucine, and leucine and isoleucine, respectively, are systematically higher for the internal standard than those obtained by the external standard; the estimated differences are equal to the estimates of intercept in proper calibration line (see

Table 3) $\widehat{a}_{12} = 0.2629$, $\widehat{a}_{14} = 0.2601$, $\widehat{a}_{34} = 0.9158$ corresponding to the above mentioned log-ratios. On the other hand, the log-ratios of glycine and leucine for internal standard method are systematically smaller than those obtained by the external standard method; the estimated difference is $\widehat{a}_{23} = -0.5040$.

parts (r,s)	T_{rs} p-value	F_{rs} p-value	T^1_{rs} p-value	T^2_{rs} p-value	C_{rs} p-value
$(1,2)$	22.5300 $<< 0.0001$	244.6636 $<< 0.0001$	11.7921 $<< 0.0001$	1.0904 0.3072	0.6305 0.0006
$(1,3)$	-8.7043 $<< 0.0001$	54.7259 $<< 0.0001$	0.3271 0.7520	-2.1150 0.0673	5.4527 0.5834
$(1,4)$	50.6504 $<< 0.0001$	1274.0529 $<< 0.0001$	6.1389 0.003	0.2863 0.7820	0.1155 $<< 0.0001$
$(2,3)$	-23.9906 $<< 0.0001$	337.9413 $<< 0.0001$	-10.0091 $<< 0.0001$	-2.2686 0.0530	2.5374 0.0801
$(2,4)$	-0.6820 0.5145	1.1317 0.7383	-1.4953 0.1732	1.3806 0.2048	0.9908 0.0034
$(3,4)$	14.8754 $<< 0.0001$	186.3208 $<< 0.0001$	6.7900 0.0001	-2.4700 0.0387	6.3674 0.7877

Table 5.4: Results of the test statistics and their corresponding p-values for the amino-acids data.

Finally, we will analyze the (common) accuracy of both methods. To verify that the methods measure with prescribed precision 0.1 we use multiple χ^2-tests with the test statistics C_{rs}. From Table 5.4 we can see that the p-values corresponding to the test statistics C_{12}, C_{14}, C_{24} are smaller than the Bonferroni-adjusted 0.05-level of significance ($\alpha = 0.0083$). Taking the Holm-Bonferroni adjustment into account, the conclusion is the same. Hence, the accuracy of measurement is significantly higher for the log-ratios of alanine and glycine, alanine and isoleucine, and glycine and isoleucine, respectively, with estimated accuracies obtained from the matrix $\mathbf{T}_*$ equal to $\widehat{\sigma}_{12} = \sqrt{0.0007} = 0.0265$, $\widehat{\sigma}_{14} = \sqrt{0.0001} = 0.01$, and $\widehat{\sigma}_{24} = \sqrt{0.0012} = 0.0346$.

The above results using tests, performed on the log-ratios, have also a direct consequence for parts of the original composition. In particular, multiple tests using T_{rs} and F_{rs} statistics revealed significant differences in measuring the compositional parts using the two methods (also by taking into account the exceptional

behavior of the log-ratio between glycine and isoleucine). For each of the compositional parts, two of the corresponding log-ratios show a significant shift using statistics T_{rs}^1 - this thus seems to be the main reason for the difference between the external and internal standards for measurement of the amino-acids compounds (and corresponds to the theoretical principle of both methods). Finally, by comparing the C_{rs} statistics with the entries of matrix $\mathbf{T}_*$, we can conclude that leucine was determined with the lowest precision using both methods. This fact might be related to a bigger molecule of leucine, which affects the precision of the measurement. In general, the more log-ratios containing a certain compositional part show a significant difference between both methods, the stronger relative difference between the methods can be assigned to that compositional part itself. Although we do not provide the corresponding statistical tests here, a graphical visualization of rejections (for all the mentioned tests) could provide a good insight into calibration behavior of the original compositional parts, in particular when compositions with more parts are involved in the analysis.

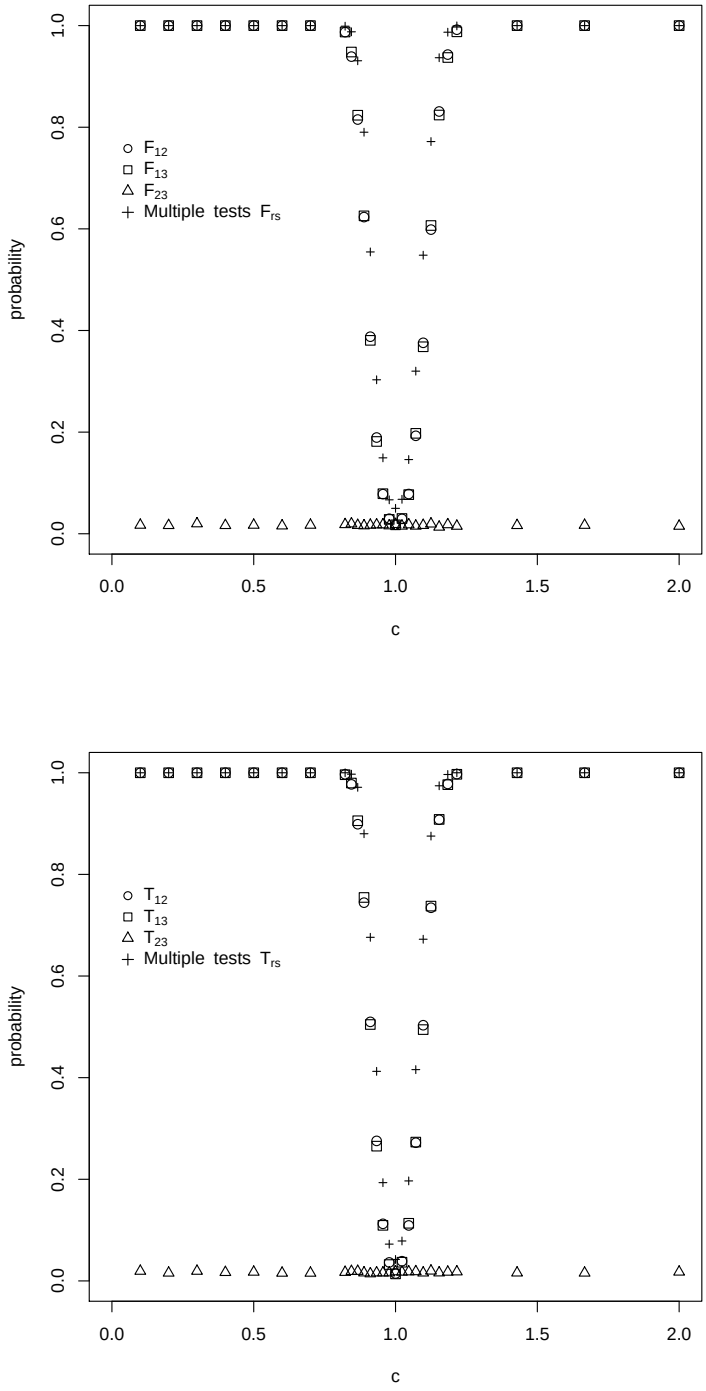

Figure 5.2: Probabilities of rejecting the conformity of three-part compositional measurements obtained by two methods with respect to constant c on the 0.017 Bonferroni-adjusted 0.05-level of significance. Results from tests F_{rs} (top) and T_{rs} (bottom) on a single log-ratio between parts r, s together with multiple tests.

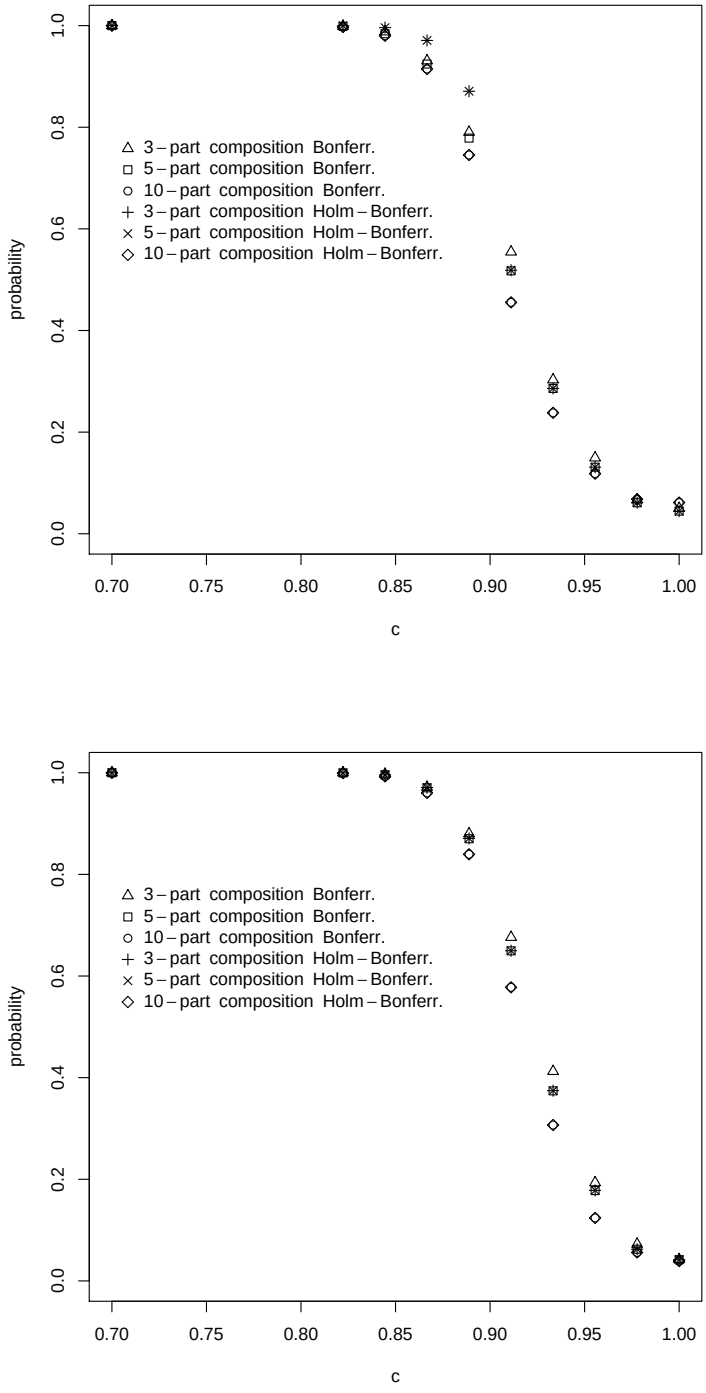

Figure 5.3: Overall probabilities of rejecting the conformity of three-, five-, and ten-part compositional measurements obtained by two methods with respect to constant c for multiple tests F_{rs} (top) and T_{rs} (bottom) on significance level 0.05 with the Bonferroni and the Holm-Bonferroni adjustments.

Conclusion

The thesis is focused on regression analysis with compositional data. In practice, for example in geosciences, compositions usually have high number of parts. Therefore, it is convenient to reduce the number of compositional parts before performing regression analysis for compositions. The proposed procedure for variable selection reduces the dimension of the compositions, and, consequently, it simplifies the statistical analysis and the interpretation of the results is easier to understand. This procedure provides only negligible loss of information about the multivariate data structure.

An intuitive selection of parts based on expert knowledge of subject matter specialists may lead to major changes of the multivariate statistical analysis results. For example, experts are interested in the analyses of certain geochemical processes and select elements for the statistical analysis which are somehow related to these processes. In this selection they may miss variables that are responsible for the substantial information about the multivariate data structure, and their omission changes the statement about the resulting subcomposition. Note that such an approach to variable selection differs from the known problem of subcompositional incoherence of the original composition with a (prescribed) unit-sum representation of the compositional vector [4, 17], that is against the general definition of compositional data as multivariate observations where the only relevant information is contained in the ratios between the parts [23].

The proposed stepwise procedure for excluding compositional parts allows to arrive at a subcomposition that still retains the important information contained in the multivariate data structure. The goal of this procedure is to retain the total variance from one step to the next, and it is stopped before a significant reduction would occur. The larger the original composition, the more reduction of the number of parts is made. Examples have demonstrated that indeed those "marker" variables are selected, and an omission of these variables would have resulted in substantial changes of multivariate statistical analyses of compositional data.

Between the targets stated on the beginning of this thesis it is to contribute

in the field of regression analysis with compositional response. Although regression analysis with compositional response represents one of the most tasks of compositional data analysis, there are still some aspects that deserve to be analysed in more detail. One of the aspects elaborated here, concerns the particular coordinate representation useful for the estimation and interpretation of regression parameters. Further aspect discussed in this thesis deals with the quality of prediction by considering (or not) also absolute abundances instead of purely relative information conveyed by compositional data. They both have in common that even clr coordinates and the simple log-transformation are nowadays rather suppressed in compositional data analysis, they might be useful for some specific tasks and also help to understand differences between various methodological viewpoints. Particularly, clr coordinates simplify the computation of the regression coefficients instead of considering D different regression models with orthonormal coordinates, just the principal difference between both options arising from a singularity of a covariance matrix for clr coordinates needs to be taken into account. Clr coordinates cannot be considered separately due to their zero sum constraint, while this is not the case for orthonormal coordinates. The theory was endowed with a real data example from sedimentology, where interesting patterns were revealed. From this perspective, we believe that the presented methodological outputs are useful steps for a practical analysis of compositional data.

The last aim of this thesis concerns the calibration problem of compositional measurements. The calibration problem (sometimes also referred to the orthogonal regression, the total least squares, or the regression with errors-in-variables) belongs to basic tasks in statistical analysis. In the thesis, an alternative approach to the orthogonal regression by means of linear models with the type-II constraints was used. The equivalence between this approach and the standard orthogonal regression estimation was proved under assumptions of independent random errors with equal variances leads. Because all the relevant information in a composition is contained in log-ratios, the multivariate problem can be con-

verted into univariate calibration of single log-ratios that are easy to handle and interpret. It means, for D-part compositions, the calibration problem can be partitioned into $D(D-1)/2$ partial univariate calibration problems, performed on log-ratios of compositional parts. Hence, the calibration line is fitted to the corresponding coordinate of all possible two-part subcompositions separately. As a result of calibration, an analogy between the compositional variation array and the matrices of the predicted values and residual variances from univariate calibrations was derived, which is a popular tool in descriptive statistics of compositional data. Further, tests for conformity of two measurement methods were proposed. Particularly, tests for the identification of a significant systematic difference between results obtained by two methods and for the verification that the methods follow the same prescribed precision. All proposed tests are univariate, and, thus the multiple comparison approach was used to summarize results into a multivariate decision. Theoretical results were applied to a real example from biochemistry.

Bibliography

[1] Aitchison J., Shen S. M. (1980). Logistic-normal distributions. Some properties and uses. *Biometrika* **67** (2), 261-272.

[2] Aitchison J. (1982). The statistical analysis of compositional data (with discussion). *Journal of the Royal Society. Series B (Statistical Methodology)* **44** (2), 139–177.

[3] Aitchison J., Bacon-Stone J. (1984). Log contrast models for experiments with mixtures. *Biometrika* **71** (2), 323–330.

[4] Aitchison J. (1986) *The statistical analysis of compositional data*. London: Chapman and Hall.

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

[6] Aitchison J., Barcel-Vidal C., Martn-Fernndez J.A., Pawlowsky-Glahn V. (2000). Logratio analysis and compositional distance. *Mathematical Geology* **32** (3), 271–275.

[7] Aitchison J., Greenacre M. (2002). Biplots of compositional data. *Applied Statistics* **51**, 375–392.

[8] Anděl J. (2007). *Statistical methods* (in Czech). Praha: Matfyzpress.

[9] Bábek, O., Matys Grygar T., Faměra M., Hron K., Nováková T., Sedláček J. (2015). Geochemical background in polluted river sediments: How to separate the effects of sediment provenance and grain size with statistical rigour? *Catena* 135, 240–253.

[10] Bacon-Scone J. (2008). Discrete and continuous compositions. In *Proceedings of CoDaWork'08. The 3rd Compositional Data Analysis Workshop* (ed Daunis-i-Estadella J. and Martín-Fernández J.A.). University of Girona, Girona.

[11] Benjamini Y., Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Serie B (Methodological)* **57** (1), 289–300.

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

[13] Blejer M.I., Fernandez R.B. (1980). Effects of unanticipated money growth on prices and on output and its composition in a fixed-exchange-rate open economy. *Canadian Journal of Economy* **13**, 82–95.

[14] Brown P. J. (1993). *Measurement, regression and calibration.* Oxford: Clarendon Press.

[15] Buccianti A., Mateu-Figueras G., Pawlowsky-Glahn V., eds. (2006). *Compositional data analysis in the geosciences: from theory to practice*, London: Geological Society.

[16] Casella G., Berger R. L. (2002). *Statistical inference, 2nd ed.* Pacific Grove: Duxbury Press.

[17] Chayes F. (1960). On correlation between variables of constant sum. *Journal of Geophysical Research* **65**, 4185-4193.

[18] Daunis-i-Estadella J., Egozcue J. J., Pawlowsky-Glahn V. (2002). Least squares regression in the simplex. In *Proceedings of IAMG02. The eigth annual conference of the International Association for Mathematical Geology, Bayer, U., Burger, H.,and Skala, W., editors, volume I and II, International Association for Mathematical Geology, Selbstverlag der Alfred-Wegener-Stiftung, Berlin* 411–416.

[19] Donevska S., Fišerová E., Hron K. (2011). On the equivalence between orthogonal regression and linear model with type-II constraints. *Acta Universitatis Palackianae Olomucensis. Facultas Rerum Naturalium. Mathematica* **50** (2), 19–27.

[20] Donevska S., Fišerová E., Hron K. (2016). Calibration of compositional measurements. *Communications in Statistics - Theory and Methods* **45** (22), 6773–6788.

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

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

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

[24] Egozcue J.J., Pawlowsky-Glahn V., Hron K., Filzmoser P. (2012). Simplicial regression. The normal model. *Journal of Applied Probability and Statistics* **6**, 87–106.

[25] Fačevicová K. (2016). Complex structures of compositional data. Dissertation, Palacký University in Olomouc, Faculty of Science, Czech Republic.

[26] Filzmoser P., Hron, K., Reimann, C. (2009). Principal component analysis for compositional data with outliers. *Environmetrics* **20** (6), 621–632.

[27] Filzmoser P., Steiger B. (2011). *StatDA: statistical analysis for environmental data.*

[28] Filzmoser P., Gschwandtner M.(2012). *Multivariate outlier detection based on robust methods.*

[29] Fišerová E., Hron K. (2010). Total least squares solution for compositional data using linear models. *Journal of Applied Statistics* **37**, 1137–1152.

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

[31] Fišerová E., Hron K. (2012). Statistical inference in orthogonal regression for three-part compositional data using a linear model with type-II constraints. *Communications in Statistics* **41**, 2367–2385.

[32] Fišerová E., Kubáček L. and Kunderová P. (2007). *Linear statistical models. Regularity and singularities*. Praha: Academia.

[33] Fišerová E., Donevska S., Hron K., Bábek O., Váňkátová K. (2016). Practical aspects of log-ratio coordinate representations in regression with compositional response. *Measurement Science Review* **16** (5), 235–243.

[34] Fuller W.A. (1987). *Measurement error models*. New York: John Wiley & Sons.

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

[36] Gueorguieva R., Rosenheck R., Zelterman D. (2008). Dirichlet component regression and its applications to psychiatric data. *Computational Statistics and Data Analysis* **52**, 5344–5355.

[37] Harville D.A. (1997). *Matrix algebra from a statistician's perspective*. New York: Springer.

[38] Hijazi R.H., Jernigan R.W. (2009). Modelling compositional data using Dirichlet regression models. *Journal of Applied Probability and Statistics* **4**, (1) 77–91.

[39] Holm S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal Of Statistics* **2**, 65–70.

[40] Hoerl A. E., Kennard R. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.

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

[42] Hron K., Kubáček L. (2011). Statistical properties of the total variation estimator for compositional data. *Metrika* 74, 221–230.

[43] Hron K., Filzmoser P., Donevska S., Fišerová E. (2013). Covariance-based variable selection for compositional data. *Mathematical geosciences* **45** (4), 6773–6788.

[44] Hrůzová K., Todorov V., Hron K., Filzmoser P. (2016). Classical and robust orthogonal regression between parts of compositional data. *Statistics*, DOI: 10.1080/02331888.2016.1162164..

[45] Hrůzová K. (2016). Economic applications of statistical analysis of compositional data. Dissertation, Palacký University in Olomouc, Faculty of Science, Czech Republic.

[46] Jackson J. D., Dunlevy J. A. (1988). Orthogonal least squares and the interchangeability of alternative proxy variables in the social sciences. *Journal of the Royal Statistical Society Series D (The Statistician)* **37** (1), 7–14.

[47] Johnson R.A., Wichern D.W. (2007). *Applied Multivariate Statistical Analysis, 6th ed.* Pearson.

[48] Jolicoeur P. (1968). Interval estimation of the slope of the major axis of a bivariate normal distribution in the case of a small sample. *Biometrics* **24**, 679–682.

[49] Kalivodová A., Hron K., Filzmoser P., Najdekr L., Janečková H., Adam T. (2015). PLS-DA for compositional data with application to metabolomics. *Journal of Chemometrics* **29** (1), 21–28.

[50] Kendall M.G., Stuart A. (1967). *The advanced theory of statistics, 2nd vol.* London: Charles Griffin.

[51] Kleinman D.L., Athans M. (1968). The design of suboptimal linear time-varying systems. *IEEE Transactions on Automatic Control*, AC-13, 150–159.

[52] Kubáček K., Kubáčková L. (1995). *Statistical models with linear structures*. Bratislava: Veda.

[53] Kubáček, L., Kubáčková L. (1997). One of the calibration problems. *Acta Universitatis Palackianae Olomucensis. Facultas Rerum Naturalium. Mathematica* **36**, 117–130.

[54] Kubáček L. (2008). *Multivariate statistical models revisited*. Olomouc: Palacký University.

[55] Martín-Fernández J.A., Daunis-i-Estadella J., Mateu-Figueras G. (2015). On the interpretation of differences between groups for compositional data. *Statistics and Operations Research Transactions* **39**, 231–252.

[56] Mateu-Figueras G., Pawlowsky-Glahn V., Barceló-Vidal C. (2003). *Distributions on the simplex*. In *Proceedings of the Compositional Data Analysis Workshop CoDaWork'03*, S. Thió-Henestrosa and J.A. Martín-Fernández, eds, Universitat de Girona, Girona.

[57] Mateu-Figueras G., Pawlowsky-Glahn V. (2008). Critical approach to probability laws in geochemistry. *Mathematical Geosciences* **40** (5), 489–502.

[58] Mateu-Figueras G., Pawlowsky-Glahn V., Egozcue J.J. (2013). The normal distribution in some constrained sample spaces. *Statistics and Operations Research Transactions* **37** (1), 29–55.

[59] Matys Grygar T., Elznicová J., Bábek O., Hošek M., Engel Z., Kiss T. (2014). Obtaining isochrones from pollution signals in a fluvial sediment record: a case study in a uranium-polluted floodplain of the Ploučnice River, Czech Republic. *Applied Geochemistry* **48**, 1–15.

[60] Müller I., Hron K., Fišerová E., Šmahaj J., Cakirpaloglu P., Vančáková J. (2016). Time budget analysis using logratio methods. *arXiv:1609.07887* [math.ST].

[61] Nestares O., Fleet D. J., Heeder D. J.(2000). Likellihood functions and confidence bands for total-least squares problems. *In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'00)* **1**, 1523–1530.

[62] Osborne C. (1991). Statistical calibration: a review. *International Statistical Review*, **59**, 309–336.

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

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

[65] Pawlowsky-Glahn V., Buccianti A., Mateu-Figueras G. (2006). *Compositional data analysis in the geosciences: from theory to practice*. London: The Geological Society of London.

[66] Pawlowsky-Glahn V., Buccianti A. (2011). *Compositional data analysis: theory and applications*. Chichester: Wiley.

[67] Pawlowsky-Glahn V., Egozcue J.J., Lovell D. (2015). Tools for compositional data with a total. *Statistical Modelling*, **15** (2), 175–190.

[68] Pawlowsky-Glahn V., Egozcue J.J., Tolsana-Delgado R. (2015). *Modeling and analysis of compositional data*. Chichester: Wiley.

[69] Pearson K. (1897). Mathematical contributions to the theory of evolution-On a form of spurious correlation which may arise when indices are used in the measurement of organs. In *Proceedings of the Royal Society of London* **60**, 489–497.

[70] R development core team (2012). *R: A language and environment for statistical computing*. R Foundation for statistical computing, Vienna.

[71] Reimann C., Äyräs M., Chekushin V., Bogatyrev I., Boyd R., Caritat P., Dutter R., Finne T., Halleraker J., Jæger O., Kashulina G., Lehto O., Niskavaara H., Pavlov V., Räisänen M., Strand T., Volden T. (1998). Environmental geochemical atlas of the Central Barents Region. *Geological Survey of Norway (NGU), Geological Survey of Finland (GTK), and Central Kola Expedition (CKE), Special Publication, Trondheim, Espoo, Monchegorsk.*

[72] Reimann C., Siewers U., Tarvainen T., Bityukova L., Eriksson J., Gilucis A., Gregorauskiene V., Lukashev V.K., Matinian N.N., Pasieczna A. (2003). *Agricultural soils in northern Europe: A geochemical atlas*. Schweizerbartśche Verlagsbuchhandlung, Stuttgart: Geologisches Jahrbuch.

[73] Scheffé H. (1958). Experiments with mixtures. *Journal of the Royal Statistical Society* **B20**, 344–360.

[74] Scheffé H.(1963). The simplex-centroid design for experiments with mixtures. *Journal of the Royal Statistical Society* **B25**, 235–263.

[75] Sedláček J., Bábek O., Kielar O. (2016). Sediment accumulation rates and high-resolution stratigraphy of recent fluvial suspension deposits in various fluvial settings, Morava River catchment area, Czech Republic. *Geomorphology* **254**, 73–87.

[76] Sedláček J., Bábek O., Nováková T. (2016). Sedimentary record and anthropogenic pollution of a step-wise filled, multiple source fed dam reservoir: An example from Nové Mlýny reservoir, Czech Republic. *Science of the Total Environment*.

[77] Shi L. M., Fang H., Tong W., Wu J., Perkins R., Blair R. M., Branham W.S., Dial S.I., Moland C.I., Sheehan D.M. (2001). QSAR models using a large diverse set of estrogens. *Journal of Chemical Information and Modeling* **41** (1), 186–195.

[78] Storey J.D., Tibshirani R. (2003). Statistical significance for genomewide studies. *PNAS* **100** (16), 9440–9445.

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

[80] Tolosana-Delgado R., van den Boogaart K.G., Fišerová E., Hron K., Dunkl I. (2015). Joint compositional calibration: a geochronological example. In *Proceedings of CoDaWork'15. The 6th international Workshop on Compositional Data analysis, L'Escala* 268–284.

[81] van den Boogaart K. G., Tolsana R. (2008). "compositions": A unified R package to analyze compositional data. *Computers & Geosciences* **34** (4), 320–338.

[82] van den Boogaart K.G., Tolosana-Delgado R. (2013). *Analyzing compositional data with R*. Verlag Berlin Heidelberg: Springer.

[83] Van Huffel S., Lemmerling P., eds. (2002). *Total least squares and errors-in-variables modelling: analysis, algorithms and applications*. Dordrecht: Kluwer.

[84] Van Huffel S., Vandevalle J. (1991). *The total least squares problem: computational aspects and analysis*. Philadelphia: SIAM.

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

[86] Wang H., Shangguan L., Wu J., Guan R. (2013). Multiple linear regression modelling for compositional data. *Neurocomputing* **122**, 490–500.

[87] Wimmer G., Witkovský V. (2007). Univariate linear calibration via replicated errors-in-variables model models. *Journal of Statistical Computation and Simulation* **77**, 213–227.

[88] Wimmer G., Witkovský V. (2011). Note on a calibration problem: selected results and extensions of professor Kubáček's Research. *Acta Universitatis Palackianae Olomucensis. Facultas Rerum Naturalium. Mathematica* **50**, 121–128.

PALACKÝ UNIVERSITY OLOMOUC
FACULTY OF SCIENCE
DEPARTMENT OF MATHEMATICAL ANALYSIS AND MATHEMATICAL APPLICATIONS

DISSERTATION THESIS SUMMARY

Regression analysis for compositional data

Supervisor:
doc. RNDr. Eva Fišerová, Ph.D.
Year of submission: 2017

Author:
Mgr. Sandra Donevska
Applied mathematics

The dissertation thesis was carried out under the full-time postgradual programme Applied Mathematics, field Applied Mathematics, in the Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University Olomouc.

Applicant: **Mgr. Sandra Donevska**
Department of Mathematical Analysis and
Applications of Mathematics
Faculty of Science
Palacký University Olomouc

Supervisor: **doc. RNDr. Eva Fišerová, Ph.D.**
Department of Mathematical Analysis and
Applications of Mathematics
Faculty of Science
Palacký University Olomouc

Reviewers: **prof. RNDr. Gejza Wimmer, DrSc.**
Department of Mathematics and Statistics
Faculty of Science
Masaryk University, Brno

Ao. Univ.-Prof. Dipl.-Ing. Peter Filzmoser, Dr.techn.
Institute of Statistics and Mathematical Methods
in Economics
Faculty of Mathematics and Geoinformation
Vienna University of Technology

Dissertation thesis summary was sent to distribution on ..... Oral defence of dissertation thesis will be performed on ..... at Department of Mathematical Analysis and Applications of Mathematics in front of the committee for Ph.D. study programme Applied Mathematics, Faculty of Science, Palacký University Olomouc, room ....., 17. listopadu 12, Olomouc. Full text of the dissertation thesis is available at Study Department of Faculty of Science, Palacký University Olomouc.

1. Abstract

This thesis is focused on regression analysis for compositional data. Relative nature of compositional data that distinguishes them from the standard multivariate data call for a special treatment. Since for the most of the statistical techniques there is still not developed stay-in the simplex approach, the log-ratio methodology presents a proper statistical approach that enable to express the data in a coordinate system.

Firstly, a regression model with compositional response variable is studied. A multivariate regression model is built for the compositional data expressed in orthonormal coordinates. The explicit formulas for the estimators of regression parameters and as well test statistics for the verification of their significance are provided. Further, new coordinate representation of the compositional data allowing to simplify the computation concerning regression parameters estimation and hypothesis testing is proposed and as well, the quality of prediction in different coordinate system is evaluated.

The second part of this thesis is devoted to the calibration problem for compositions. Here the calibration approach based on linear models with the type-II constraints is used. The equivalence between the linear model with type-II constraints and the total least squares regression is proved. A procedure for calibration of compositional measurements is proposed and tests for conformity of two measuring devices (methods) are presented.

In the last part of the thesis, a variable selection procedure for compositions that guarantees that a reduction of the original composition to a subcomposition causes only negligible change of the information is presented.

All theoretical results are applied to real-world examples.

Key words: compositional data; regression with compositional response; calibration; total least squares; linear model with type-II constraints; variable selection

2. Abstract v českém jazyce

Práce se zabývá regresní analýzou pro kompoziční data. Relativní charakter kompozičních dat, který je odlišuje od standardních mnohorozměrných dat, vyžaduje speciální zacházení. Jedním ze základních přístupů ke statistické analýze kompozičních dat, který je použit i v této práci, je vyjádření kompozičních dat ve vhodném souřadnicovém systému.

Nejprve je pozornost soustředěna na problematiku regresního modelu s kompoziční vysvětlovanou proměnnou. Pro kompoziční data vyjádřená v ortonormálních souřadnicích je v práci vytvořen mnohorozměrný regresní model a uvedeny explicitní vzorce pro odhady neznámých regresních parametrů a testové statistiky pro ověření jejich statistické významnosti. Dále je navržena jiná souřadnicová reprezentace kompozičních dat, která umožňuje zjednodušit výpočty pro odhady regresních parametrů a testové statistiky a vyhodnocena kvalita predikce v různých souřadnicových systémech.

Druhá část této práce je věnována kalibračnímu problému pro kompoziční data. V práci je použit přístup založený na lineárním modelu s podmínkami typu II. Je zde dokázána ekvivalence mezi lineárními modely s podmínkami typu II a ortogonální regresí. Dále je zde navržena procedura pro kalibraci kompozičních měření a prezentovány testy pro shodu dvou měřicích přístrojů (metod).

V poslední části této práce je navržena procedura pro výběr kompozičních složek, která zaručuje, že výsledná redukce dimenze kompozice nezpůsobí podstatnou ztrátu informace o mnohorozměrné variabilitě datové struktuře.

Všechny teoretické výsledky jsou aplikovány při řešení reálných úloh.

Klíčová slova: kompoziční data; regrese s kompoziční vysvětlovanou proměnnou; kalibrace; ortogonální regrese; lineární regresní model s podmínkami typu II; výběr proměnných

3. Introduction

Regression is a common statistical method for modelling and analysing the relationship between the response and predictor variable(s). In the frame of the parametric approach of the regression analysis, the linear relationship between the variables is only investigated.

The regression techniques discussed in this thesis are performed on a special kind of multivariate data known as compositional data, or compositions for short. The definition for D-part composition as quantitative descriptions of the parts of some whole, conveying relative information, dates from the 1986 and it is given by Aitchison. This strictly positive data that quite often sum up into an arbitrary constant, have the simplex $\mathcal{S}^D$ with the Aitchison geometry, to be their sample space. As it is well known, the simplex lacks the Euclidean vector space structure. The log-ratio methodology presents a proper statistical approach that enable us to express the data isometrically in the real Euclidean space [1].

Regression analysis for the compositional data started to expand in the early 80's. Great progress in this field has been done till now, but there are still some topics that deserve special attention. Four types of regression models, depending on the type of the response and predictors variables can be distinguished: a regression model with compositional response and non-compositional predictors, a regression model with non-compositional response and compositional predictors, a regression between parts of compositions, and a regression model with compositional response and predictor variables.

The motivation for writing this thesis lies in satisfying the current needs for further development in regression analysis for compositional data. Because the branch is quite wide, the thesis is mainly focused on regression models with compositional response, the calibration problem for compositions, and the simplification of regression models with compositional data in terms of reducing dimensionality of the compositions. The calibration problem is related to a regression between parts and the TLS problem.

4. Recent state summary

4.1 Compositional data

Compositional data, or compositions are strictly positive multivariate observations that carry only relative information.
Compositions, denoted as $\mathbf{y} = (y_1, y_2, \ldots, y_D)'$, have their own sample space the simplex $\mathcal{S}^D$ defined as

$$\mathcal{S}^D = \left\{ \mathbf{y} = (y_1, y_2, \ldots, y_D)' | y_i > 0, i = 1, 2, \ldots, D; \sum_{i=1}^{D} y_i = k \right\}.$$

Crucial in this framework is the operation of closure for $\mathbf{y} = (y_1, y_2, \ldots, y_D)' \in \mathbb{R}_+^D$, given by

$$\mathcal{C}(\mathbf{y}) = \left(\frac{ky_1}{\sum_{i=1}^{D} y_i}, \frac{ky_2}{\sum_{i=1}^{D} y_i}, ..., \frac{ky_D}{\sum_{i=1}^{D} y_i} \right)',$$

with which we can express the compositions as a non-negative vectors summing up into an arbitrary constant $k > 0$. Basically, information contained in the composition remains same it is just matter of change of the units. Such compositions are compositionally equivalent, hence it is does not depend on the choice of k.

The vector space structure of the simplex $\mathcal{S}^D$ is obtained with the following two operations defined on it [3, 22]:

- perturbation of $\mathbf{y} \in \mathcal{S}^D$ by $\mathbf{w} \in \mathcal{S}^D$, analogous to addition in the real space:

$$\mathbf{y} \oplus \mathbf{w} = \mathcal{C}\left(y_1 w_1, y_2 w_2, ..., y_D w_D\right)',$$

- power transformation or powering of $\mathbf{y} \in \mathcal{S}^D$ by a constant $\alpha \in R$, analogous to scalar multiplication in the real space:

$$\alpha \odot \mathbf{y} = \mathcal{C}\left(y_1^\alpha, y_2^\alpha, ..., y_D^\alpha\right)'.$$

Further it is desired to work with the compositions on the simplex on comparable way as we do we the standard multivariate data on the real space. We demand to

compute the length of a composition, to determine angles between compositional vectors or to find the distance between them. The Aitchison inner product, norm and distance were invented to satisfy these purposes [3, 22, 23]. Moreover, these functions form the Aitchison geometry on the simplex.

There are certain log-ratio transformations which translates the composition from the simplex into coordinate vector on the the real space.

Firstly the centred log-ratio (clr) transformation was invented, which is mapping between the simplex $\mathcal{S}^D$ and the Euclidean space $\mathbb{R}^D$, defined by,

$$\operatorname{clr}(\mathbf{y}) = \left(\ln \frac{y_1}{\mathrm{g}(\mathbf{y})}, \ln \frac{y_2}{\mathrm{g}(\mathbf{y})}, \ldots, \ln \frac{y_D}{\mathrm{g}(\mathbf{y})} \right)' = \mathbf{h},\ \mathbf{y} \in \mathcal{S}^D,\ \mathbf{h} \in \mathbb{R}^D. \tag{4.1}$$

where $\mathrm{g}(\mathbf{y}) = \sqrt[D]{\prod_{j=1}^{D} y_j}$ is the geometric mean of the parts of the composition.

Clr transformation actually, expresses the composition $\mathbf{y} \in \mathcal{S}^D$ in coordinates with respect to the generating system on the simplex.

The isometric property makes this transformation applicable for techniques based on distances. This property also reflects the straightforward interpretation of the clr transformed composition. Unfortunately, one disadvantage property of this transformation that comes from the symmetry of the components of the vector of the clr coordinates is that leads to singular covariance matrix which causes computational issues. Another disadvantage property that the clr transformation dispose is that the clr coefficients do not satisfy the principle of subcompositional coherence. This principle is of crucial importance whose meaning is that the information carried in the composition should not be contradictory with the one carried in the subcomposition. Every method before applied to the compositional data should meet this requirement. Here the geometric mean of a subcompostion does not necessary have to be the same with the one we have for the full D-part composition.

Despite of the disadvantage properties, the clr coordinates are still frequently used because of an intuitive interpretation. For example, the compositional biplot of the clr coordinates [2] can be constructed, that is an important visualization tool for investigation of the compositional data structure. Here, the single clr coordinates are usually interpreted in terms of the original compositional parts [8, 24].

To avoid disadvantages of the clr coordinates, orthonormal coordinates with respect to an orthonormal basis on the simplex were proposed [5]. The transformation is called as isometric log-ratio (ilr) transformation.

There exist many ways to obtain an orthonormal basis of the simplex. Unfortunately, there is no canonical basis on the simplex, where by the interpretation of the orthonormal coordinates is not that straightforward. The choice of the method for construction of the basis may improve the interpretation of the resulting coordinates. Behind the commonly used methods belong the Gram - Schmidt procedure [5] and the very intuitive - sequential binary partition (SBP) [6]. The resulting coordinates coming from the SBP, called balances, give interpretation in sense of grouped parts of the composition. In each of the $D-1$ consecutive steps of the SBP, partitioning of the parts into two non-overlapping, distinguished groups is done. Groups of compositional parts are formed according to expert knowledge, or can be formed blindly, without any preliminary knowledge about the grouping of the parts.

Often used orthonormal basis leads to the $(D-1)\times D$ matrix $\mathbf{V}$, such that $\mathbf{V}\mathbf{V}' = \mathbb{I}_{(D-1)}$, with the rows vectors [11]

$$\mathbf{v}_i = \sqrt{\frac{D-i}{D-i+1}}\left(0,\ldots,0,1,-\frac{1}{D-i},\ldots,-\frac{1}{D-i}\right),\ i=1,2,\ldots,D. \tag{4.2}$$

This basis relates with the orthonormal coordinates [11],

$$\operatorname{ilr}(\mathbf{y})_i = \mathrm{z}_i = \sqrt{\frac{D-i}{D-i+1}}\ln\frac{y_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} y_j}}, i=1,2,\ldots,D-1. \tag{4.3}$$

There exist unique relationship between the ilr and the clr coordindates [22], given by

$$\mathbf{z} = \mathbf{h}\mathbf{V}',$$

where $\mathbf{h}\in\mathbb{R}^D$ is the clr transformed composition $\mathbf{y}\in\mathcal{S}^D$. Moreover, for the first coordinates of both systems it holds [22],

$$h_1 = \sqrt{\frac{D-1}{D}}z_1.$$

In this case, the first orthonormal coordinate z_1 explains all the relative information about the first compositional part y_1 within the first given composition [11]. Unfortunately, the remaining orthonormal coordinates do not have such straightforward interpretation.

In order to obtain the interpretation for the remaining orthonormal coordinates, we just need to make permutation of the compositional parts,

$$\mathbf{y}^{(l)} = (y_l, y_1, \ldots, y_{l-1}, y_{l+1}, \ldots, y_D) =$$
$$= \left(y_1^{(l)}, y_2^{(l)}, \ldots, y_{l-1}^{(l)}, y_{l+1}^{(l)}, \ldots, y_D^{(l)}\right), \ l = 1, 2, \ldots D,$$

and subsequently apply the formula in (4.3) to the permuted compositions $\mathbf{y}^{(l)}$, $l = 1, 2, \ldots, D$. Thus the first orthonormal coordinate obtained for permuted composition $y^{(l)}, l = 1, 2, \ldots, D$, contains all the relative information about the l-th compositional part y_l, $l = 1, 2, \ldots, D$ and, consequently

$$h_l = \sqrt{\frac{D-1}{D}} z_1^{(l)}, \ l = 1, 2, \ldots, D. \tag{4.4}$$

4.2 Regression analysis for compositional data

In the thesis we firstly devote to regression model with compositional response.

Remarkable invention in the field of the regression analysis for compositional data came in the paper of J.J. Egozcue et al. (2012). Here the regression model is expressed in the orthonormal coordinates which offers opportunity to use the least squares (LS) method for obtaining the estimates of the unknown regression parameters [4, 7].

The LS problem is presented on both the simplex $\mathcal{S}^D$ and in coordinates on the R^{D-1}.

Secondly we deal with the calibration problem for compositions. The calibration is a process whereby the scale of a measuring device or method is determined on a basis of an experiment. There are two stages in the calibration process. In the first stage, the calibration curve is specified. It describes a relationship between the quantity values with measurement uncertainties provided by a measurement standard (a measuring device or method with assigned correctness) and a calibrated one. The second stage concerns the prediction of values for measurement standard based on measurements by calibrated device. The values of measurement standard are considered either fixed (non-random), or random. In the former case we speak about controlled calibration, in the latter about random or natural calibration. In this thesis we focused on determination of a calibration line with random values of the standard. The calibration problem is handled with linear model approach namely with linear model with type two constraints. The linear model approach as an alternative approach of the total least squares problem for 3-part composition

is studied already in [10, 11]. Generalization of the total least squares for D-part composition can be find in [14, 15].

5. Thesis objectives

The particular goals of the thesis are the following:

- Formulate a multivariate regression model with a compositional response and find explicit formulas for the estimators of the regression parameters and proper test statistics. Find coordinate representation of compositional data allowing to simplify the computation concerning regression parameters estimation and hypothesis testing. Evaluate the quality of prediction in different coordinate systems.
- Prove the equivalence between the TLS approach and the linear model with the type-II constraints. Propose a procedure for calibration of compositional measurements and suggest tests for conformity of two measuring devices (methods).
- Propose a variable selection procedure for compositions that guarantees that a reduction of the original composition to a subcomposition causes only negligible change of the information.
- Apply theoretical results to real-world examples.

6. Theoretical framework and applied methods

6.1 Covariance - based variable selection for compositoinal data

The basic measure of variability of a random composition $\mathbf{y} \in \mathcal{S}^D$ is the variation matrix [1], defined as

$$\mathbf{T} = \left\{ \mathrm{var}\left(\ln \frac{y_i}{y_j} \right) \right\}_{i,j=1}^{D}.$$

The elements of the variation matrix describe the variability of the random log-ratio $\ln \frac{y_i}{y_j}$: the smaller the value of this variance, the more the log-ratio tends to be a constant. The (normed) sum of the elements of the variation matrix is called total variance,

$$\mathrm{totvar}(\mathbf{y}) = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \mathrm{var}\left(\ln \frac{y_i}{y_j} \right), \tag{6.1}$$

expressing the total variability of the compositional data set. Note that

$$\mathrm{totvar}(\mathbf{y}) = \sum_{i=1}^{D} \mathrm{var}(h_i) = \sum_{i=1}^{D-1} \mathrm{var}(z_i^{(l)}),\ l = 1, 2, \ldots, D, \tag{6.2}$$

i.e. the total variance can also be computed using the variability of the clr coordinates or the orthonormal coordinates, respectively [21].
Further, what worth to be mentioned for the purposes of the thesis is the compositional variation array, defined as the simplest and minimum way of summarizing the patterns of location and variability within a compositional data set [1],

$$\mathbf{V} = \begin{pmatrix} 0 & \mathrm{var}\left(\ln \frac{y_1}{y_2}\right) & \mathrm{var}\left(\ln \frac{y_1}{y_3}\right) & \cdots & \mathrm{var}\left(\ln \frac{y_1}{y_D}\right) \\ \mathrm{E}\left(\ln \frac{y_2}{y_1}\right) & 0 & \mathrm{var}\left(\ln \frac{y_2}{y_3}\right) & \cdots & \mathrm{var}\left(\ln \frac{y_2}{y_D}\right) \\ \cdots & \cdots & \cdots & \cdot & \cdots \\ \mathrm{E}\left(\ln \frac{y_D}{y_1}\right) & \mathrm{E}\left(\ln \frac{y_D}{y_2}\right) & \mathrm{E}\left(\ln \frac{y_D}{y_3}\right) & \cdots & 0 \end{pmatrix}, \tag{6.3}$$

where in the upper triangle of the array the log-ratio variances and in the lower triangle the log-ratio expectations are displayed.

The variance of the orthonormal coordinates $z_i,\ i = 1, \ldots, D$ in (4.3) is given by the relation [11],

$$\mathrm{var}(z_i) = \frac{1}{D-i+1} \sum_{p=i+1}^{D} \mathrm{var}\left(\ln \frac{y_i}{y_p}\right) - \frac{1}{2(D-i)(D-i+1)} \sum_{p=i+1}^{D} \sum_{q=i+1}^{D} \mathrm{var}\left(\ln \frac{y_p}{y_q}\right). \quad (6.4)$$

As a consequence of (4.4), the variance of the clr coordinate h_i corresponds (up to a constant) to the variance of $z_1^{(i)}$.

From (6.4), multiplied by $(D-1)/D$ to obtain clr variances, we can also expect quite a strong relation between $\mathrm{var}(y_i)$ and the sum of the i-th row (column) of the corresponding variation matrix $\mathbf{T}$. This finding induces a useful property:

$$[\mathrm{var}(h_i) \geq \mathrm{var}(h_j)] \ \Leftrightarrow \ \left[\sum_{p=1}^{D} \mathrm{var}\left(\ln \frac{y_i}{y_p}\right) \geq \sum_{p=1}^{D} \mathrm{var}\left(\ln \frac{y_j}{y_p}\right)\right],$$

$$i \neq j,\ i, j = 1, 2, \ldots, D,$$

where h_i are the clr coordinates given in (4.1). Particularly, it shows that ordered variances of different clr coordinates (or, alternatively, of the first orthonormal coordinates from (4.3)) correspond to the same order of the sums in the variation matrix connected with the related compositional parts. Thus, can be used to identify compositional parts ("markers") that are responsible for larger clr variances. It is possible to identify the ordered contribution of the single compositional parts to the overall variance with the corresponding clr coordinates. Using this fact, a stepwise algorithm was introduced in the thesis that helps to derive a subcomposition with a minimal loss concerning the total variance of the original composition. In each step of the algorithm we can omit one part of the composition that has the smallest contribution to the overall variability (6.1),(6.2). The algorithm will stop once we will reach 2-part composition or we can use a stop criteria for the algorithm that will recommend us where the reduction of parts should stop. We use the following stop criterion

$$U_i^+ = \frac{\widehat{\text{totvar}}(\mathbf{y}_i) - \text{totvar}(\mathbf{y}_{i-1})}{\sqrt{\frac{2}{n-1}\text{tr}(\widehat{\mathbf{\Sigma}}_i^2)}},$$

the matrix $\widehat{\mathbf{\Sigma}}_i$ stands for the sample covariance matrix of the composition $\mathbf{y}_i$ in (arbitrarily chosen) orthonormal coordinates. Small values of U_i^+ favor the alternative, so we reject the null hypothesis, if U_i^+ realizes in the critical region $\mathsf{W} = (-\infty, u_\alpha\rangle$, where u_α denotes the α-quantile (preferably $\alpha = 0{,}05$) of the standard normal distribution.

6.2 Multivariate regression with compositional response

Let us consider in the following $\mathbf{y}$ will not stand for a composition but will stand for a standard random vector. This notation is used in order to follow the usual one in the statistical literature.

A multivariate regression model presents a regression model where multiple response variables appear simultaneously. Consider we have q random variables $y_1, y_2, \ldots, y_q$ and for each of these we have n observations. Let us denote by $\mathbf{y}_j = (y_{1j}, y_{2j}, \ldots, y_{nj})'$, $j = 1, 2, \ldots, q$, the observation vector that corresponds to the random variable y_j. For every vector $\mathbf{y}_j$ we assume the following linear model [16, 19]

$$\mathbf{y}_j = \mathbf{X}\mathbf{b}_j + \boldsymbol{\varepsilon}_j, \ j = 1, 2, \ldots, q, \tag{6.5}$$

and, simultaneously, for all vectors $\mathbf{y}_j$ we assume the multivariate linear model

$$\underline{\mathbf{y}} = \mathbf{X}\mathbf{B} + \underline{\boldsymbol{\varepsilon}}, \tag{6.6}$$

where $\underline{\mathbf{y}} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_q)$ is the $(n \times q)$ dimensional matrix of response vectors, $\mathbf{X}$ is the $(n \times k)$ dimensional design matrix which has full column rank, $\boldsymbol{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_q)$ is the $(k \times q)$ dimensional matrix of the unknown regression parameters, $\mathbf{b}_j = (b_{1j}, b_{2j}, \ldots, b_{kj})'$, $j = 1, 2, \ldots, q$ and $\underline{\boldsymbol{\varepsilon}} = (\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_q)$ is the $(n \times q)$ dimensional matrix of the random errors. Further, let us assume that the multivariate responses $\mathbf{y}_{i\cdot} = (y_{i1}, \ldots, y_{iq})'$, $i = 1, 2, \ldots, n$, are independent with the same unknown variance-covariance matrix $\mathbf{\Sigma}$, i.e.

$$\text{cov}(\mathbf{y}_{i\cdot}, \mathbf{y}_{j\cdot}) = \mathbf{0}, \ i \neq j,$$
$$\text{var}(\mathbf{y}_{i\cdot}) = \mathbf{\Sigma}, \ i = 1, 2, \ldots, n.$$

In order to derive the estimator of $\mathbf{B}$, to construct confidence intervals and confidence regions for unknown regression parameters, or to do some tests for significance of the regression coefficient, etc., the model (6.6) can be rewritten in the following vectorized form [19]

$$\operatorname{vec}\left(\underline{\mathbf{y}}\right) = \left(\mathbf{I}_q \otimes \mathbf{X}\right) \operatorname{vec}\left(\mathbf{B}\right) + \operatorname{vec}\left(\boldsymbol{\varepsilon}\right), \ \operatorname{var}[\operatorname{vec}(\underline{\mathbf{y}})] = \boldsymbol{\Sigma} \otimes \mathbf{I}_n,$$

where $\operatorname{vec}(\underline{\mathbf{y}}) = (\mathbf{y}_1', \ldots, \mathbf{y}_q')'$ and the symbol $\otimes$ denotes the Kronecker product. Thus, the least squares estimator of $\operatorname{vec}(\mathbf{B})$ is obtained by minimizing the square of the Mahalanobis distance of the residuals [19]

$$\arg \min_{\operatorname{vec}(B)} \left\| \operatorname{vec}\left(\underline{\mathbf{y}}\right) - \left(\mathbf{I}_q \otimes \mathbf{X}\right) \operatorname{vec}\left(\boldsymbol{B}\right) \right\|^2_{(\Sigma \otimes \mathbf{I}_n)}. \tag{6.7}$$

The solution of the minimization problem (6.7), after de-vectorization, is

$$\widehat{\boldsymbol{B}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\underline{\mathbf{y}}. \tag{6.8}$$

The estimator $\widehat{\boldsymbol{B}}$ is the best linear unbiased estimator (BLUE) of the parameter matrix $\boldsymbol{B}$ [19].

One can notice that this estimator is invariant with respect to the variance-covariance matrix of $\operatorname{vec}\left(\underline{\mathbf{y}}\right)$.

However, the variance-covariance matrix of the vector $\operatorname{vec}(\widehat{\mathbf{B}}) = (\widehat{\mathbf{b}}_1', \widehat{\mathbf{b}}_2', \ldots, \widehat{\mathbf{b}}_q')'$

$$\operatorname{var}\left[\operatorname{vec}(\widehat{\mathbf{B}})\right] = \boldsymbol{\Sigma} \otimes \left(\mathbf{X}'\mathbf{X}\right)^{-1} \tag{6.9}$$

depends on $\boldsymbol{\Sigma}$. Since the variance-covariance matrix $\boldsymbol{\Sigma}$ is unknown, it is necessary to estimate it. The unbiased estimator of $\boldsymbol{\Sigma}$ is $\widehat{\boldsymbol{\Sigma}} = \underline{\mathbf{y}}'\mathbf{M}_X\underline{\mathbf{y}}/(n-k)$, where $\mathbf{M}_X = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is a projector on the orthogonal complement of the vector space $\mathcal{M}(\mathbf{X})$ generated by the columns of the design matrix $\mathbf{X}$, i.e. $\mathcal{M}(\mathbf{X}) = \{\mathbf{X}\mathbf{u} : \mathbf{u} \in \mathbb{R}^k\}$. Under normality, the estimators $\widehat{\mathbf{B}}$ and $\widehat{\boldsymbol{\Sigma}}$ are statistically independent. Moreover, if $n - k > q$, then $(n-k)\widehat{\boldsymbol{\Sigma}}$ has the Wishart distribution $W_q[n-k, \boldsymbol{\Sigma}]$.

Let us note that the univariate approach (6.5) leads to the same estimators of the regression parameters $\mathbf{b}_j$ and of the variances $\sigma^{jj} = \{\boldsymbol{\Sigma}\}_{jj}$, $j = 1, 2, \ldots, q$. Specifically,

$$\widehat{\mathbf{b}}_i = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{y}_i,$$

$$\operatorname{var}(\widehat{\mathbf{b}}_i) = \sigma^{ii} \left(\mathbf{X}'\mathbf{X}\right)^{-1},$$

$$\widehat{\sigma}^{ii} = \mathbf{y}_i' \mathbf{M}_X \mathbf{y}_i / (n-k).$$

The theory of multivariate linear regression models [19] provide a range of tests, that are easy to compute due to explicit formulas. Usually three basic issues of hypotheses testing in a multivariate regression context are considered: significance of covariates for the responses $\mathbf{y}_j$, $j = 1, 2, \ldots, q$, point wise and simultaneously, and verification that the predictor $\mathbf{x}_i$, $i = 1, 2, \ldots, k$, contributes to the explanation of the overall variability in $\underline{\mathbf{y}}$.

The multivariate models enable to describe more complex designs, thus concerning the association between the outcomes. Definitely, they are more efficient tool for modelling convoluted designs than the univariate ones.

Further, testing in the multivariate models avoid problems with the multiple testing. The tests for the univariate models are not simultaneous tests for all the regressions and they do not consider the influence of the correlations among the responses, which can result in less powerful tests. Consequently, the univariate tests cannot evaluate joint influence on all outcomes. Among the difficulties when one uses the multivariate linear model approach is the necessity of disposing with large number of observations and complex interpretation of the results.

Regression with a D-part compositional response leads to a multivariate linear model with a $(D-1)$-dimensional response variable formed by orthonormal coordinates z_i, $i = 1, 2, \ldots, D-1$, given by 4.3). Although by using orthonormal coordinates, it is possible to decompose the (multivariate model into $D-1$ multiple regressions [7], in general, the multivariate approach has several advantages in comparison with a series of univariate models.

According to (6.5) and (6.6), the multivariate linear model can be expressed as

$$(\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_{D-1}) = \mathbf{X}(\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{D-1}) + (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{D-1}),$$

or, equivalently, in the matrix form

$$\underline{\mathbf{z}} = \mathbf{X}\mathbf{B} + \underline{\varepsilon}.$$

Here it is assumed that $\mathbf{X}$ is an $(n \times k)$ dimensional design matrix of full column rank, $\mathbf{b}_j$, $j = 1, 2, \ldots, D-1$, is a k dimensional vector of unknown regression parameters and $\underline{\varepsilon}$ is an $[n \times (D-1)]$ dimensional matrix of the random errors. The multivariate responses $\underline{\mathbf{z}}_{i\cdot} = (z_{i1}, z_{i2}, \ldots, z_{iD-1})'$, $i = 1, 2, \ldots, n$, are assumed to be independent with the same unknown variance-covariance matrix $\mathbf{\Sigma}$. Regression with a D-part compositional response leads to a multivariate linear model with a $(D-1)$-dimensional response variable formed by orthonormal coordinates.

Although by using orthonormal coordinates, it is possible to decompose the multivariate model into $D-1$ multiple regressions [7], in general, the multivariate approach has several advantages in comparison with a series of univariate models.

According to (6.8), the BLUE of the parameter matrix $\mathbf{B}$ is

$$\widehat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{D-1}).$$

The estimator of $\mathbf{B}$ is invariant under a change of the variance-covariance matrix $\boldsymbol{\Sigma}$. The variance-covariance matrix of $\mathrm{vec}(\widehat{\mathbf{B}})$ is given in (6.9).

To verifying the significance of the covariates for the ilr coordinate z_j, $j = 1, 2, \ldots, D-1$, point wise and simultaneously, i.e. testing the null hypothesises $\mathbf{b}_j = \mathbf{0}$, $j = 1, 2, \ldots, D-1$, the following test statistic can be used

$$F_j^{ilr} = \frac{(n-k)\,\widehat{\mathbf{b}}_j'\mathbf{X}'\mathbf{X}\widehat{\mathbf{b}}_j}{k\widehat{\sigma}^{jj}}, \tag{6.10}$$

which has F-distribution with k and $n-k$ degrees of freedom under the null hypothesis.

Another test that can be taken into account is the test for the significance of the i-th predictor, $i = 1, 2, \ldots, k$, i.e. test of the hypothesis $\mathbf{B}_{i.} = \left(b_{i1}, b_{i2}, \ldots, b_{i(D-1)}\right) = \mathbf{0}$. The test statistic for this case is

$$F_{pred,i}^{ilr} = \frac{(n-D-k+2)\,\widehat{\mathbf{B}}_{i.}\,(\underline{\mathbf{z}}'\mathbf{M}_X\underline{\mathbf{z}})^{-1}\,\widehat{\mathbf{B}}_{i.}'}{(D-1)\left\{(\mathbf{X}'\mathbf{X})^{-1}\right\}_{ii}}, \tag{6.11}$$

which is distributed as $F_{D-1,n-D-k+2}$ under the null hypothesis H_{0i}.

Lastly sometimes it is of interest to verify the significance of the whole matrix of regression parameters $\mathbf{B}$, or in other words to test the hypothesis $\mathbf{AB} = \mathbf{C}$, where $\mathbf{A}$ is a $q \times k$ hypothesis matrix having full-row rank $q \leq k$, and $\mathbf{C}$ is a $q \times D-1$ matrix. Therefore we will use the well-known Pillai-Barttlet trace, Wilk's Lambda, Hotelling-Lawley trace and Roy's largest root that rely on the $p = \min(q, D-1)$ non-zero eigenvalues λ_j of $\mathbf{HE}^{-1}$ where the matrices $\mathbf{H}$ and $\mathbf{E}$ are

$$\mathbf{E} = (\underline{\mathbf{y}} - \mathbf{X}\widehat{\mathbf{B}})'(\underline{\mathbf{y}} - \mathbf{X}\widehat{\mathbf{B}})$$

$$\mathbf{H} = (\mathbf{A}\widehat{\mathbf{B}} - \mathbf{C})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\widehat{\mathbf{B}} - \mathbf{C}).$$

Due to (4.4) that describes the relationship between single clr coefficients and the first orthonormal coordinates from (4.3) it seems to be quite intuitive possibility to replace orthonormal coordinates in the response simply by their clr counterparts and then proceed with the regression analysis.

Nevertheless, due to singularity of the covariance matrix of clr coordinates it is not possible to decompose the multivariate model into univariate ones as it was the case for orthonormal coordinates. Though, as it is shown below, even taking multivariate regression in clr coordinates would yield the same results of the respective test statistics as one would obtain by considering single orthonormal coordinates, coming from D regression models.

Then the multivariate model can be also written in the form

$$\underline{\mathbf{h}} = \mathbf{X}\mathbf{B}^{clr} + \underline{\varepsilon}_{clr}, \tag{6.12}$$

where $\underline{\mathbf{h}} = (\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_D)$ is the $(n \times D)$ dimensional matrix of response vectors that stand for the clr coordinates given in (4.1). The variance-covariance matrix of independent D-variate responses $\underline{\mathbf{h}}_{i.}$ is $\text{var}(\underline{\mathbf{h}}_{i.}) = \Sigma_{clr} = \mathbf{V}'\Sigma_{ilr}\mathbf{V}$, $i = 1, 2, \ldots, n$ where $\mathbf{V}$ is the $[(D-1) \times D]$ matrix, such that satisfies $\mathbf{V}\mathbf{V}' = \mathbb{I}_{(D-1)}$, having the rows vectors given by the relation (4.2). The variance-covariance matrix Σ_{clr} is a $D \times D$ positive semi-definite matrix with the rank $D-1$ unlike Σ_{ilr}, which is a full rank $(D-1) \times (D-1)$ positive definite matrix. Obviously, $\Sigma_{ilr} = \mathbf{V}\Sigma_{clr}\mathbf{V}'$. The relationships between the parameter matrices and multivariate responses are the following

$$\begin{aligned}
\mathbf{B}^{clr} &= \mathbf{B}^{ilr}\mathbf{V}, \\
\mathbf{B}^{ilr} &= \mathbf{B}^{clr}\mathbf{V}', \\
\underline{\mathbf{h}} &= \underline{\mathbf{z}}\mathbf{V}, \\
\underline{\mathbf{y}} &= \underline{\mathbf{h}}\mathbf{V}'.
\end{aligned}$$

In the thesis is shown that if we consider the model (6.12), the test statistics given by the relation (6.10) and (6.11) can be used for testing hypothesis significance of covariates for the clr coordinate $\mathbf{h}_j$, $j = 1, 2, \ldots, D$ point wise and simultaneously, and verification that the predictor $\mathbf{x}_i, \imath = 1, 2, \ldots, k$, contributes to the explanation of the overall variability. Follows that it is possible to perform parameter estimation and significance testing in clr coordinates instead of taking D orthonormal coordinate systems of type (4.3), when the interpretation in sense of the original compositional parts (with respect to the others) is required. Although methodically working in orthonormal coordinates is preferred in any case, numerical outputs are the same (test statistics) or differ just up to a constant resulting from (4.4).

Finally, note that the interpretation of the regression parameters can be enhanced by considering orthogonal coordinates, resulting from suppressing scaling

constants in orthonormal coordinates. Concretely, they are formed from (4.3) by omitting scaling constants and replacing the natural logarithm by its binary counterpart (or any other interpretable base of logarithm), i.e.

$$z_i^* = \log_2 \frac{y_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} y_j}}, \ i = 1, 2, \ldots, D-1$$

[20]. By considering regression in clr coordinates, the parameters of the resulting regression model in orthogonal coordinates, adapted to favour the l-th compositional part (denoted as $b_1^{*(l)}$), would be related through

$$b_1^{*(l)} = \log_2(\mathrm{e})\sqrt{\frac{D}{D-1}} b_{ilr,1}^{(l)} = \log_2(\mathrm{e})\frac{D}{D-1} b_{clr,l}.$$

Consequently, by taking the j-th element of $b_1^{*(l)}$, i.e. $b_{1;j}^{*(l)}$, for $j = 1, 2, \ldots, k$, then for a unit additive change in the j-th explanatory variable (by constant values of the other covariates), the ratio of y_l to the mean relative contributions of the other parts grows (decreases) $\delta = 2^{b_{1;j}^{*(l)}}$ times.

To compare log-ratio and log-transformed regression models one has to analyse, whether something similar holds also in the regression context. Such a finding would be an important step to understand the behaviour of regression models in different coordinate systems. For this purpose, the matrix of sums of residual squares is taken for both the cases of orthonormal coordinates and log-transformed compositions,

$$\begin{aligned} \mathbf{E}_{ilr} &= (\underline{\mathbf{z}} - \mathbf{X}\widehat{\mathbf{B}})'(\underline{\mathbf{z}} - \mathbf{X}\widehat{\mathbf{B}}) = \underline{\mathbf{z}}'\mathbf{M}_X\underline{\mathbf{z}}, \\ \mathbf{E}_{log} &= [\log(\underline{\mathbf{y}})]'\mathbf{M}_X \log(\underline{\mathbf{y}}), \end{aligned}$$

respectively. Here the symbol $\underline{\mathbf{y}}$ denotes an $n \times D$ matrix with D-part compositions in rows. The overall variability in data corresponds to the matrices of total sum of squares

$$\mathbf{T}_{ilr} = \underline{\mathbf{z}}'\mathbf{M}_{X_1}\underline{\mathbf{z}} = \mathbf{V}\mathbf{T}_{log}\mathbf{V}', \quad \mathbf{T}_{log} = [\log(\underline{\mathbf{y}})]'\mathbf{M}_{X_1} \log(\underline{\mathbf{y}}).$$

The matrix $\mathbf{E}$ is commonly used to measure the discrepancy between the data and a fitted model in case of multivariate regression [16]. Although also an alternative exists, based directly on the norm between the observed and predicted response [7], using directly $\mathbf{E}$ seems to be more coherent with the current regression methodology. Particularly, the trace of $\mathbf{E}$ is of primary importance, because it aggregates

residual sums of squares of single response variables and leads to the multivariate analogy of the residual sum of squares (RSS). The following relation between the traces of matrices $\mathbf{E}$ and $\mathbf{T}$ for compositions in orthonormal coordinates and by taking log-transformation holds:

$$0 \leq \mathrm{tr}(\mathbf{E}_{ilr}) \leq \mathrm{tr}(\mathbf{E}_{log}), \quad 0 \leq \mathrm{tr}(\mathbf{T}_{ilr}) \leq \mathrm{tr}(\mathbf{T}_{log}).$$

Thus the trace of the matrix $\mathbf{E}$ obtained for orthonormal coordinates is less or equal to that one for log-transformed compositions. Thus, the mean squared error (MSE) for orthonormal coordinates is less or equal to the MSE for log-transformed data. Since the same inequality holds also for the trace of the matrix $\mathbf{T}$, the relationship between the coefficients of determination R^2_{ilr} and R^2_{log} does not exist in general. These measures of goodness of fit, defined as

$$R^2_{ilr} = 1 - \frac{\mathrm{tr}(\mathbf{E}_{ilr})}{\mathrm{tr}(\mathbf{T}_{ilr})}, \quad R^2_{log} = 1 - \frac{\mathrm{tr}(\mathbf{E}_{log})}{\mathrm{tr}(\mathbf{T}_{log})},$$

thus reflect structural changes that arise by avoiding the scale invariance property of compositions, i.e. when log-transformation is applied instead of taking the orthonormal coordinates.

6.3 Calibration problem for compositional data

The total least squares (TLS) presents regression technique that deals with modelling the relationship within the composition, i.e. between the compositional parts. This technique is also known as the orthogonal regression, regression with errors in variables, or as a calibration problem. A model is established just for the three-part compositions after the ilr transformation and can be used for modelling the relationship between the parts of compositions. Primary contribution to this quite new regression technique compositional data can be find in the papers [10, 12]. Authors there overcome the standard TLS by the linear regression models with the type II constraints [17].

An important requirement to build such a model is the assumption of independence and homoscedasticity of the orthonormal coordinates. Otherwise, when this is violated, then it is not satisfied the invariance of the results on the simplex under the orthogonal rotation of the orthonormal coordinates. Namely, when transforming the results of the analysis back on the simplex they will differ from these obtained in the ilr space. Linear models with type-II constrains [9], based on the

calibration line approach [18, 25], overcome the difficulties of the TLS approach. The linear regression model is of the form of

$$\begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\nu} \end{pmatrix} + \boldsymbol{\varepsilon},$$

where $\mathbf{z}_1$ and $\mathbf{z}_2$ are the orthonormal coordinates, and where the unknown regression parameters a, b, and the vector of the errorless recordings $\boldsymbol{\mu}, \boldsymbol{\nu}$ satisfy

$$\boldsymbol{\nu} = a\mathbf{1}_n + b\boldsymbol{\mu}, \quad \text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_{2n}. \tag{6.13}$$

Such constraints on regression parameters involving other unknown parameters a and b are called type-II constraints. Evidently, this is a non-linear function of the unknown parameters b and $\boldsymbol{\mu}$. Using linearization by the Taylor series locally at $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\nu}^{(0)}$, $a^{(0)}$ and $b^{(0)}$, when the second and higher derivatives are neglected, the locally BLU estimators of $\boldsymbol{\mu}$, $\boldsymbol{\nu}$, a and b are derived in [10].

The variance σ^2 is usually unknown and can be unbiasedly estimated by [18]

$$\widehat{\sigma}^2 = \frac{(\mathbf{z}_1 - \widehat{\boldsymbol{\mu}})'(\mathbf{z}_2 - \widehat{\boldsymbol{\mu}}) + (\mathbf{z}_1 - \widehat{\boldsymbol{\nu}})'(\mathbf{z}_2 - \widehat{\boldsymbol{\nu}})}{n-2}. \tag{6.14}$$

The linear model approach is favourable for finite samples, unlike, the TLS which is an asymptotic approach. Moreover, the linear model approach enables to perform the standard statistical inference, being difficult or sometimes impossible in the frame of the TLS approach.

An iterative algorithm is proposed for the estimation of the calibration line [10]. Advantages of this iteration procedure is that it converges very quickly, and in addition, stable values of the estimates are achieved in the first few iterations. Problems with numerical stability of the proposed algorithm may occur if the angle between the calibration line and the axis represented by the first orthonormal coordinate tends to be 90°. Thus the calibration line is estimated.

Moreover, we have shown in the thesis that indeed the TLS and the linear model with type-II constraints lead to the same estimates.

In the thesis is discussed the calibration problem for compositions. For D-part compositions, the calibration problem can be partitioned into $D(D-1)/2$ partial calibration problems, performed on log-ratios of compositional parts. In other words, the calibration is performed for the corresponding coordinate of all possible two-part subcompositions separately.

Let there be n different compositions that have D parts which are measured using two methods A and B with the same precision. Let $\mathcal{R} = \{r = 1, 2, \ldots, D -$

$1,\ s = r+1,\ r+2, \ldots, D\}$ be the set of subscripts. For two-part subcompositions (y_r, y_s) and (w_r, w_s), corresponding to the measurements obtained by methods A and B, respectively, the log-ratios are formed and arranged in data matrices

$$(\mathbf{Z}^A_{rs}, \mathbf{Z}^B_{rs}) = \begin{pmatrix} \ln \frac{y_{1r}}{y_{1s}} & \ln \frac{w_{1r}}{w_{1s}} \\ \ln \frac{y_{2r}}{y_{2s}} & \ln \frac{w_{2r}}{w_{2s}} \\ \vdots & \vdots \\ \ln \frac{y_{nr}}{y_{ns}} & \ln \frac{w_{nr}}{w_{ns}} \end{pmatrix},$$

where $(r,s) \in \mathcal{R}$ (note that multiplying the log-ratios by $1/\sqrt{2}$ ilr coordinate would be formed). Let us assume that $\mathbf{Z}^A_{rs}$ and $\mathbf{Z}^B_{rs}$ represent a realization of a normally distributed n-dimensional random vector $\mathbf{z}^A_{rs} \sim N_n(\boldsymbol{\mu}_{rs}, \sigma^2\mathbf{I}_n)$ and $\mathbf{z}^B_{rs} \sim N_n(\boldsymbol{\nu}_{rs}, \sigma^2\mathbf{I}_n)$, respectively. Let $\boldsymbol{\mu}_{rs} = (\mu_{1rs}, \mu_{2rs}, \ldots, \mu_{nrs})'$ denote the vector of errorless measurement results of $\mathbf{z}^A_{rs}$, and $\boldsymbol{\nu}_{rs} = (\nu_{1rs}, \nu_{2rs}, \ldots, \nu_{nrs})'$ the vector of errorless measurement results of $\mathbf{z}^B_{rs}$, where $(r,s) \in \mathcal{R}$. Moreover, these measurement results are taken to be mutually independent. Thus, the calibration line (6.13) can be expressed as

$$\boldsymbol{\nu}_{rs} = a_{rs}\mathbf{1}_n + b_{rs}\boldsymbol{\mu}_{rs}, \tag{6.15}$$

where $(r,s) \in \mathcal{R}$, and $\mathbf{1}_n$ stands for the vector of n ones. The parameter a_{rs} represents a systematic deviation of log-ratios between parts r and s obtained by measurement methods A and B, and b_{rs} denotes the scaling factor between them. The formulas for the BLUE's of the unknown model parameters and as well instruction for the iterative procedure for estimation can be find in [10].

From the theoretical point of view, it is interesting that the fitted calibration lines can be also used to predict the values of the method B by the method A and vice versa. For this purpose, let us define the matrices of the predicted averages $\mathbf{M}^{(j)}$, $j = 1, 2$, as

$$\mathbf{M}^{(j)} = \begin{pmatrix} 0 & \widehat{m}^{(j)}_{12} & \widehat{m}^{(j)}_{13} & \cdots & \widehat{m}^{(j)}_{1D} \\ \widehat{m}^{(j)}_{21} & 0 & \widehat{m}^{(j)}_{23} & \cdots & \widehat{m}^{(j)}_{2D} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \widehat{m}^{(j)}_{D1} & \widehat{m}^{(j)}_{D2} & \widehat{m}^{(j)}_{D3} & \cdots & 0 \end{pmatrix},$$

where for $j = 1$ the elements of $\mathbf{M}^{(1)}$ are the averages of method B as predicted by the averages of method A. Particularly, elements of $\mathbf{M}^{(1)}$ are defined as predicted averages using the fitted calibration line (6.15) when the parameters are marked

with superscript (1),

$$\widehat{m}_{rs}^{(1)} = \widehat{a}_{rs}^{(1)} + \widehat{b}_{rs}^{(1)}\overline{z_{rs}^{A}}, \quad \overline{z_{rs}^{A}} = \frac{1}{n}\sum_{i=1}^{n} \ln \frac{x_{ir}}{x_{is}}.$$

Conversely, the elements of $\mathbf{M}^{(2)}$ are the averages of method A as predicted by the averages of method B, i.e., the elements of $\mathbf{M}^{(2)}$ are defined as predictions using the fitted calibration line

$$\boldsymbol{\mu}_{rs} = a_{rs}^{(2)}\mathbf{1}_n + b_{rs}^{(2)}\boldsymbol{\nu}_{rs},$$

specifically

$$\widehat{m}_{rs}^{(2)} = \widehat{a}_{rs}^{(2)} + \widehat{b}_{rs}^{(2)}\overline{z_{rs}^{B}}, \quad \overline{z_{rs}^{B}} = \frac{1}{n}\sum_{i=1}^{n} \ln \frac{y_{ir}}{y_{is}}.$$

Further, the matrix of residual variances is defined as

$$\mathbf{T}_* = \begin{pmatrix} 0 & \widehat{\sigma}_{12}^2 & \widehat{\sigma}_{13}^2 & \cdots & \widehat{\sigma}_{1D}^2 \\ \widehat{\sigma}_{21}^2 & 0 & \widehat{\sigma}_{23}^2 & \cdots & \widehat{\sigma}_{2D}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \widehat{\sigma}_{D1}^2 & \widehat{\sigma}_{D2}^2 & \widehat{\sigma}_{D3}^2 & \cdots & 0 \end{pmatrix},$$

where $\widehat{\sigma}_{rs}^2$ is the estimate of the residual variance (6.14) for $r, s = 1, 2, \ldots, D$, $r \neq s$.

The the matrices $\mathbf{M}^{(j)}$, $j = 1, 2$ and $\mathbf{T}_*$ fulfil the following properties:

i) For the elements of the matrices $\mathbf{M}^{(j)}$, $j = 1, 2$, the triangular equality holds, i.e.,

$$\widehat{m}_{rs}^{(j)} = \widehat{m}_{rl}^{(j)} + \widehat{m}_{ls}^{(j)}, \ r, s, l = 1, 2, \ldots, D.$$

ii) Matrices $\mathbf{M}^{(j)}$, $j = 1, 2$, are antisymmetric, i.e., $\widehat{m}_{rs}^{(j)} = -\widehat{m}_{sr}^{(j)}$ and $\widehat{m}_{rr}^{(j)} = 0$, $r, s = 1, 2, \ldots, D$.

Similarly, it is a direct consequence of the logarithm properties that the matrix of residual variances $\mathbf{T}_*$ is symmetric. Thus we can conclude that the elements of the above matrices have the same properties as the elements of the variation array given by (6.3).

The above findings can be used for descriptive statistics based on the results of the calibration problem. Consequently, some tests for conformity of two measurement methods can be introduced.

The test statistic for each hypothesis $H_{0rs} : a_{rs}^{(1)} = 0, b_{rs}^{(1)} = 1$ individually, according to [12], is given as

$$F_{rs} = \left[\begin{pmatrix} \widehat{a}_{rs}^{(1)} \\ \widehat{b}_{rs}^{(1)} \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right]' \left[\widehat{\text{var}} \begin{pmatrix} \widehat{a}_{rs}^{(1)} \\ \widehat{b}_{rs}^{(1)} \end{pmatrix} \right]^{-1} \left[\begin{pmatrix} \widehat{a}_{rs}^{(1)} \\ \widehat{b}_{rs}^{(1)} \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right].$$

The symbol $\widehat{\text{var}}[(\widehat{a}_{rs}^{(1)}, \widehat{b}_{rs}^{(1)})']$ stands for the covariance matrix of the estimator $(\widehat{a}_{rs}^{(1)}, \widehat{b}_{rs}^{(1)})'$ [10]. Under the null hypothesis, the statistic F_{rs} is distributed as $F_{2,n-2}$. For testing the whole set of hypotheses H_{0rs}, $(r,s) \in \mathcal{R}$, simultaneously, it is necessary to use some techniques for multiple comparisons. In order to retain a prescribed significance level α for all tests simultaneously, the significance level for each test must be less than α. The Bonferroni-adjusted α-level of significance $\alpha_{adj} = \frac{2\alpha}{D(D-1)}$ for each test is one of the most commonly used approaches. Applying the Bonferroni correction, we reject H_{0rs} when $f_{rs} \geq F_{2,n-2}\,(1 - \alpha_{adj})$, where f_{rs} is a realization of the test statistic F_{rs}. Equivalently, the hypothesis H_{0rs} is rejected if p-value $\leq \alpha_{adj}$.

In the case we reject some of H_{0rs} we want to identify which of the components $a_{rs}^{(1)}$ or $b_{rs}^{(1)}$ is responsible for the rejection. This is done by performing multiple tests of intercepts and slopes separately. Therefore, in order to test whether the intercepts are significantly different from zero, i.e., to test $a_{rs}^{(1)} = 0$, $(r,s) \in \mathcal{R}$, simultaneously, the test statistic [12]

$$T_{rs}^{1} = \frac{\widehat{a}_{rs}^{(1)} \sqrt{n\,[\widehat{\boldsymbol{\mu}}_{rs}]'\,\widehat{\boldsymbol{\mu}}_{rs} - [\mathbf{1}'\widehat{\boldsymbol{\mu}}_{rs}]^2}}{\widehat{\sigma}_{rs} \sqrt{\left(\widehat{b}_{rs}^{(1)} + 1\right) [\widehat{\boldsymbol{\mu}}_{rs}]'\,\widehat{\boldsymbol{\mu}}_{rs}}},$$

can be used. Under the null hypothesis T_{rs}^{1} has the t_{n-2} distribution. The hypothesis $a_{rs}^{(1)} = 0$ is rejected if $|t_{rs}^{1}| \geq t_{n-2}\left(1 - \frac{\alpha_{adj}}{2}\right)$, where t_{rs}^{1} is a realization of T_{rs}^{1}, or else if p-value$\leq \alpha_{adj}$. The test that the slopes are significantly different from 1, i.e., to test $b_{rs}^{(1)} = 1$, $(r,s) \in \mathcal{R}$, simultaneously, is done by using the test statistic [12]

$$T_{rs}^{2} = \frac{\left(\widehat{b}_{rs}^{(1)} - 1\right) \sqrt{n\,[\widehat{\boldsymbol{\mu}}_{rs}]'\,\widehat{\boldsymbol{\mu}}_{rs} - [\mathbf{1}'\widehat{\boldsymbol{\mu}}_{rs}]^2}}{\widehat{\sigma}_{rs} \sqrt{n\left(\widehat{b}_{rs}^{(1)} + 1\right)}},$$

which is distributed as t_{n-2}, if the hypothesis is true. Conditions for rejection of hypotheses are the same as for the intercept.

Another way of hypothesis testing that the two measurement methods give the same results is based on matrices of predicted averages $\mathbf{M}^{(1)}$ and $\mathbf{M}^{(2)}$. Both methods give the same results if and only if the matrices $\mathbf{M}^{(1)}$ and $\mathbf{M}^{(2)}$ are the same. Hence, the problem is to test the hypothesis that all elements of matrices $\mathbf{M}^{(1)}$ and $\mathbf{M}^{(2)}$ are the same, i.e., to test $\mathrm{H}_{0rs}^{'}$: $\overline{\mu}_{rs} = \overline{\nu}_{rs}$ for all $(r,s) \in \mathcal{R}$, simultaneously. The test statistics for this hypothesis are of the form

$$T_{rs} = \frac{\widehat{m}_{rs}^{(1)} - \widehat{m}_{rs}^{(2)}}{\widehat{\sigma}_{rs}} \sqrt{n/2},$$

$(r,s) \in \mathcal{R}$. Under the null hypothesis the test statistic T_{rs} follows the t_{n-2} distribution. We reject $\mathrm{H}_{0rs}^{'}$ if $|t_{rs}| \geq \mathrm{t}_{n-2}\left(1 - \frac{\alpha_{adj}}{2}\right)$, where t_{rs} is a realization of the test statistic T_{rs}.

In order to verify that the methods follow a prescribed precision (the same for both methods), we provide multiple tests on elements of matrix $\mathbf{T}_*$, $\mathrm{H}_{0rs}^{''}$: $\sigma_{rs}^2 = \sigma_{rs0}^2$ versus $\mathrm{H}_{1rs}^{''}$: $\sigma_{rs}^2 \neq \sigma_{rs0}^2$. Here we use χ^2-tests on variance,

$$C_{rs} = \widehat{\sigma}_{rs}^2 \frac{n-2}{\sigma_{rs0}^2}.$$

In the case that $\mathrm{H}_{0rs}^{''}$ is true the test statistic C_{rs} is distributed as χ_{n-2}^2. Hypothesis $\mathrm{H}_{0rs}^{''}$ is rejected if $c_{rs} \leq \chi_{n-2}^2\left(\frac{\alpha_{adj}}{2}\right)$ or $c_{rs} \geq \chi_{n-2}^2\left(1 - \frac{\alpha_{adj}}{2}\right)$ for c_{rs} taken to be a realization of C_{rs}.

All the tests proposed in this section are uniformly more powerful if one uses, e.g., the modified Bonferroni adjustment, known as Holm-Bonferroni adjustment [13], which consists of a three step algorithm. In the first step p-values are computed and arranged in ascending order, thus constituting an increasing sequence of p-values, i.e., $p_{(1)} < p_{(2)} < \cdots < p_{(k)}$, where k is the number of comparisons made (here $D(D-1)/2$). Subsequently they are compared with the corresponding adjusted α-levels of significance calculated as $\frac{\alpha}{k-j+1}$, j is the position in the sequence of p-values. We are starting with comparing the first p-value with the appropriate adjusted α-level of significance. The algorithm stops when it finds such a p-value that exceeds the adjusted α-level of significance, i.e., when it finds minimum j such that $p_{(j)} > \frac{\alpha}{k-j+1}$. Finally, in the last step a conclusion about the acceptance or rejection of the hypotheses is done, i.e., reject hypotheses corresponding with p-values $p_{(1)}, \ldots, p_{(j-1)}$ and do not reject the remaining hypotheses.

7. Original results and summary

The thesis is focused on regression analysis with compositional data. In practice, for example in geosciences, compositions usually have high number of parts. Therefore, it is convenient to reduce the number of compositional parts before performing regression analysis for compositions. The proposed procedure for variable selection reduces the dimension of the compositions, and, consequently, it simplifies the statistical analysis and the interpretation of the results is easier to understand. This procedure provides only negligible loss of information about the multivariate data structure.

An intuitive selection of parts based on expert knowledge of subject matter specialists may lead to major changes of the multivariate statistical analysis results. For example, experts are interested in the analyses of certain geochemical processes and select elements for the statistical analysis which are somehow related to these processes. In this selection they may miss variables that are responsible for the substantial information about the multivariate data structure, and their omission changes the statement about the resulting subcomposition.

The proposed stepwise procedure for excluding compositional parts allows to arrive at a subcomposition that still retains the important information contained in the multivariate data structure. The goal of this procedure is to retain the total variance from one step to the next, and it is stopped before a significant reduction would occur. The larger the original composition, the more reduction of the number of parts is made.

Although regression analysis with compositional response represents one of the most tasks of compositional data analysis, there are still some aspects that deserve to be analysed in more detail. One of the aspects elaborated here, concerns the particular coordinate representation useful for the estimation and interpretation of regression parameters. Further aspect discussed in this thesis deals with the quality of prediction by considering (or not) also absolute abundances instead of purely relative information conveyed by compositional data. They both have in common that even clr coordinates and the simple log-transformation are nowadays rather suppressed in compositional data analysis, they might be useful for some specific tasks and also help to understand differences between various methodological viewpoints. Particularly, clr coordinates simplify the computation of the regression coefficients instead of considering D different regression models with orthonormal coordinates, just the principal difference between both options arising

from a singularity of a covariance matrix for clr coordinates needs to be taken into account. Clr coordinates cannot be considered separately due to their zero sum constraint, while this is not the case for orthonormal coordinates. From this perspective, we believe that the presented methodological outputs are useful steps for a practical analysis of compositional data.

The last aim of this thesis concerned the calibration problem of compositional measurements. The calibration problem belongs to basic tasks in statistical analysis. In the thesis, an alternative approach to the total least squares by means of linear models with the type-II constraints was used. The equivalence between this approach and the standard total least squares estimation was proved under assumptions of independent random errors with equal variances leads. Because all the relevant information in a composition is contained in log-ratios, the multivariate problem can be converted into univariate calibration of single log-ratios that are easy to handle and interpret. It means, for D-part compositions, the calibration problem can be partitioned into $D(D-1)/2$ partial univariate calibration problems, performed on log-ratios of compositional parts. Hence, the calibration line is fitted to the corresponding coordinate of all possible two-part subcompositions separately. As a result of calibration, an analogy between the compositional variation array and the matrices of the predicted values and residual variances from univariate calibrations was derived, which is a popular tool in descriptive statistics of compositional data. Further, tests for conformity of two measurement methods were proposed. Particularly, tests for the identification of a significant systematic difference between results obtained by two methods and for the verification that the methods follow the same prescribed precision. All proposed tests are univariate, and, thus the multiple comparison approach was used to summarize results into a multivariate decision.

List of publications

- **Donevska S.**, Fišerová E., Hron K. (2011). On the equivalence between orthogonal regression and linear model with type-II constraints. *Acta Universitatis Palackianae Olomucensis. Facultas Rerum Naturalium. Mathematica* **50** (2), 19–27.

- Hron K., Filzmoser P., **Donevska S.**, Fišerová E. (2013). Covariance-based variable selection for compositional data. *Mathematical Geosciences* **45** (4): 487–498.

- Potěšil J., Kopřiva F., Džubák P., **Donevska S.**, Hajdúch M. (2013). Systémovy vliv průduškového astmatu na expersi proteinu vícečetné lékové rezidence 1 v lymfocytech periferní krve u dětí. *Alergie* **15** (3): 171–174.

- Potěšil J., Kopřiva F., Džubák P., **Donevska S.**, Srovnal J., Michal V., Hajdúch M. (submitted). Overexpression of multidrug resistance protein 1 in peripheral blood lymphocytes in subjects on cetirizine of levocetirizine antihistamines. *Pediatric Rheumatology.*

- Kadlecová A., Nagy M., Mertins B., **Donevska S.** (2014). Identifikace konstituentů u předškolních dětí. *Naše řeč* **97** (3): 146–163.

- Procházka R., Řehan V., **Donevska S.** (2014). Problematika alexithymie a disociace u zavislosti na alkoholu. *Československá psychologie* **58** (2): 168–178.

- Horaková D., Azeem K., Benešová R., Pastuha D., Horák V., Dumbrovská L., Martínek A., Novotný D., Švagera Z., Hobzová M., Galuszková D., Janout V., **Donevska S.**, Vrbková J., Kollárová H. (2015). Total and High Molecular Weight Adiponectin Levels and Prediction of Cardiovascular Risk in Diabetic Patients. *International Journal of Endocrinology* **2015**, doi:10.1155/2015/545068.

- **Donevska S.**, Fišerová E., Hron K. (2016). Calibration of compositional measurements. *Communications in Statistics - Theory and Methods* **45** (22): 6773–6788.

- Fišerová E., **Donevska S.**, Hron K., Bábek O., Váňkátová K. (2016). Practical aspects of log-ratio coordinate representations in regression with compositional Response. *Measurement Science Review* **16** (5): 235–243.

List of conferences

- ODAM 2011, 26. - 28. 01. 2011, Olomouc (CZ) Compositional aspects of orthogonal regression (poster)
- CoDaWork 2011, 10. - 13. 05. 2011, Sant Feliu de Guixols (SP) Interpretation of orthonormal coordinates in case of three-part compositions applied to orthogonal regression for compositional data (poster)
- LINSTAT 2012, 14. - 20. 07. 2012, Bedlewo (PL): Calibration between log-ratios of parts of compositional data (presentation)
- ROBUST 2012, 9. - 14. 09. 2012, Němčičky (CZ): Calibration between log-ratios of parts of compositional data using linear models (poster)
- CoDaWork 2013, 3. - 7. 6. 2013, Vorau (AT): Covariance-based variable selection for compositional data (presentation)
- ODAM 2013, 12. - 16. 6. 2013, Olomouc (CZ): Calibration line problem for compositions (presentation)
- ROBUST 2014, 19. - 24. 1. 2014, Jetřichovice (CZ): Výběr proměnných v kompozičních datech (presentation)
- LINSTAT 2014, 24. - 28. 8. 2014, Linköping (SW): Variable selection stepwise procedure for compositional data (presentation)

Bibliography

[1] Aitchison J. (1986) *The statistical analysis of compositional data*. London: Chapman and Hall.

[2] Aitchison J., Greenacre M. (2002). Biplots of compositional data. *Applied Statistics* **51**, 375–392.

[3] Buccianti A., Mateu-Figueras G., Pawlowsky-Glahn V., eds. (2006). *Compositional data analysis in the geosciences: From Theory to Practice*, London: Geological Society.

[4] Daunis-i-Estadella J., Egozcue J. J., Pawlowsky-Glahn V. (2002). Least squares regression in the simplex. In *Proceedings of IAMG02. The eigth annual conference of the International Association for Mathematical Geology, Bayer, U., Burger, H.,and Skala, W., editors, volume I and II, International Association for Mathematical Geology, Selbstverlag der Alfred-Wegener-Stiftung, Berlin* 411–416.

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

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

[7] Egozcue J.J., Pawlowsky-Glahn V., Hron K., Filzmoser P. (2012). Simplicial regression. The normal model. *Journal of Applied Probability and Statistics* **6**, 87–106.

[8] Filzmoser P., Hron, K., Reimann, C. (2009). Principal component analysis for compositional data with outliers. *Environmetrics* **20**(6), 621–632.

[9] Fišerová E., Kubáček L. and Kunderová P. (2007). *Linear statistical models. Regularity and singularities*. Praha: Academia.

[10] Fišerová E., Hron K. (2010). Total least squares solution for compositional data using linear models. *Journal of Applied Statistics* **37**: 1137–1152.

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

[12] Fišerová E., Hron K. (2012). Statistical inference in orthogonal regression for three-part compositional data using a Linear model with Type-II constraints. *Communications in Statistics* **41**, 2367–2385.

[13] Holm S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal Of Statistics* **2**, 65–70.

[14] Hrůzová K., Todorov V., Hron K., Filzmoser P. (2016). Classical and robust orthogonal regression between parts of compositional data. *Statistics*, DOI: 10.1080/02331888.2016.1162164..

[15] Hrůzová K. (2016). Economic applications of statistical analysis of compositional data. Dissertation, Palacký University in Olomouc, Faculty of Science, Czech Republic.

[16] Johnson R.A., Wichern D.W. (2007). *Applied multivariate statistical analysis, 6th ed.* Pearson.

[17] Kubáček K., Kubáčková L. (1995). *Statistical models with linear structures.* Bratislava: Veda.

[18] Kubáček, L., Kubáčková L. (1997). One of the calibration problems. *Acta Universitatis Palackianae Olomucensis. Facultas Rerum Naturalium. Mathematica***36**, 117–130.

[19] Kubáček L. (2008). *Multivariate statistical models revisited.* Olomouc: Palacký University.

[20] Müller I., Hron K., Fišerová E., Šmahaj J., Cakirpaloglu P., Vančáková J. (2016). Time budget analysis using logratio methods. *arXiv:1609.07887* [math.ST].

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

[22] Pawlowsky-Glahn V., Buccianti A. (2011). *Compositional data analysis: theory and applications.* Chichester: Wiley.

[23] Pawlowsky-Glahn V., Egozcue J.J., Tolsana-Delgado R. (2015). *Modeling and analysis of compositional data.* Chichester: Wiley.

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

[25] Wimmer G., Witkovský V. (2007). Univariate linear calibration via replicated errors-in-variables model models. *Journal of Statistical Computation and Simulation* **77**, 213–227.