PALACKÝ UNIVERSITY IN OLOMOUC
FACULTY OF SCIENCE

DISSERTATION THESIS

Economic applications of statistical analysis of compositional data

Department of Mathematical Analysis and Applications of Mathematics
Supervisor: **Doc. RNDr. Karel Hron, Ph.D.**
Author: **Mgr. Klára Hrůzová**
Study program: P1104 Applied Mathematics
Field of study: Applied Mathematics
Form of study: full-time
The year of submission: 2016

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Mgr. Klára Hrůzová

Název práce: Ekonomické aplikace statistické analýzy kompozičních dat

Typ práce: Dizertační práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: Doc. RNDr. Karel Hron, Ph.D.

Rok obhajoby práce: 2016

Abstrakt: Logratio analýza kompozičních dat, mnohorozměrných pozorování nesoucích relativní informaci, je již hojně využívána v přírodních vědních disciplínách, jako je geologie nebo chemie, avšak ve vědách společenských - ekonomie, psychologie a další, ještě není příliš známá. Tato práce se zabývá adaptací známých statistických metod pro kompoziční data s ekonomickými aplikacemi. Ukazuje se, že pokud se bere v úvahu relativní charakter dat, modely poskytují relevantní výsledky. Práce obsahuje kromě metod pro redukci dimenze (metoda hlavních komponent, PARAFAC) zejména regresní analýzu, která je v ekonomických aplikacích velmi oblíbená. V jejím rámci se pak zabývá zejména situací, kdy je kompoziční závisle i nezávisle proměnná, speciálně když regresi uvažujeme mezi složkami kompozice. V takovém případě je potřeba použít pro odhady parametrů ortogonální regresi, což je typ regrese s chybami v proměnných, namísto obvyklé metody nejmenších čtverců. Nakonec práce popisuje funkcionální obdodu metody hlavních komponent, která je aplikována na hustoty, neboli funkcionální kompozice.

Klíčová slova: kompoziční data; metoda hlavních komponent; regresní analýza; ortogonální regrese; funkcionální data; hustoty

Počet stran: 99

Počet příloh: 0

Jazyk: anglický

BIBLIOGRAPHICAL IDENTIFICATION

Author: Mgr. Klára Hrůzová

Title: Economic applications of statistical analysis of compositional data

Type of thesis: Dissertation thesis

Department: Department of Mathematical Analysis and Applications of Mathematics

Supervisor: Doc. RNDr. Karel Hron, Ph.D.

The year of presentation: 2016

Abstract: Logratio analysis of compositional data, multivariate observations carrying relative information, is nowadays widely used in nature sciences, such as geology or chemistry, however, it is not widespread in social sciences like economy, psychology, etc. The thesis deals with adaptations of known statistical methods for compositional data with economic applications. It reveals that by taking the relative nature of data into account the models provide relevant results. Besides the dimension reduction methods (principal component analysis, PARAFAC), the thesis particularly includes the regression analysis which is very popular in economic applications. Within regression analysis, the thesis mainly deals with the situation where both the dependent and independent variables are compositional, especially when the regression between the parts of a composition is considered. In such a case, orthogonal regression, a kind of errors-in-variable models, needs to be applied for parameter estimation instead of ordinary least squares method. Finally, functional analogy to principal component analysis is applied for the density functions, i.e. functional compositions.

Key words: compositional data; principal component analysis; linear regression; orthogonal regression; functional data; density functions

Number of pages: 99

Number of appendices: 0

Language: English

Statement of originality

I hereby declare that this dissertation thesis has been completed independently, under the supervision of Doc. RNDr. Karel Hron, 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

I would like to thank my supervisor Doc. RNDr. Karel Hron, Ph.D. for helpfulness, guidance and patience during the preparation of the scientific papers and this thesis. I would also like to thank RNDr. Jitka Machalová, Ph.D. for computing the B-spline coefficients for the final chapter. And, of course, I want to thank my parents for their support, enthusiasm and patience during the studies.

Introduction

Compositional data (or compositions for short) are known as column vectors with positive components that carry relative information, in other words, the only relevant information is contained in ratios between components [1]. Mostly, compositions sum to a constant, like 1 in case of proportions or 100 for percentages, however, it is just a proper representation in the equivalence class of proportional vectors, forming the sample space of compositional data. Accordingly, possible choice of constant sum constraint should not influence results of statistical analysis due to scale invariance property of compositions [87, 89].

The standard Euclidean geometry defined in real space is not appropriate for compositional data. It is caused by relative character of compositions, since Euclidean geometry deals with absolute values of components [89]. Hence, the Aitchison geometry with Euclidean vector space properties was developed which captures the relative nature of compositions [6, 88].

Nevertheless, almost all statistical methods rely on the Euclidean geometry in real space [21]. Accordingly, it is not appropriate to apply them directly to compositions. Instead, the logratio methodology [1, 28, 89] is used to express compositional data in real space using appropriate coordinates and, if necessary, to transform the results back to the original sample space [80, 87]. It is of particular importance to choose such coordinates that lead to interpretable and meaningful results.

The analysis of compositional data is nowadays popular in fields such as geology or chemometrics [12, 87], however, in social sciences like economy, psychology or sociology, compositional data are not widespread yet. Up to rare applications of the logratio methodology in economics [5, 43], despite of compositional nature of data [7, 19, 107], the analysis does not reflect this fact. Therefore, this thesis is aimed to present popular statistical tools adapted to compositional data and applied to economic data.

The thesis is divided into four chapters. The first chapter introduces the compositional data analysis - the main definitions, geometry, coordinate representation and descriptive statistics. The second chapter is focused on dimension reduction methods, namely principal component analysis together with construction and interpretation of the compositional biplot, and PARAFAC method for analysis of three-way compositional data are employed. In the first part of this chapter, the variance structure of three-part composition is construed in order to provide better insight into principal components [55]. Moreover, the chapter also contains an application of dimension reduction methods in analysis of trade flows structure [56]. The third chapter is aimed to linear regression in both classical and robust versions. Four possible cases are contained. The first two sections, where regression with compositional response [23] and compositional covariates [52], respectively, is described, are introduced as a basis for other sections. The third section contains a simple regression model for the case, when both the response and explanatory variables are compositional, and the main part of this chapter, the fourth section, is intended to regreesion between compositional parts [57]. Here the orthogonal regression with statistical inferences obtained by bootstrap sampling is applied together with the robust counterpart. The final chapter focuses on functional data analysis, particularly on functional principal component analysis (FPCA) applied on density functions, i.e. functional compositions [53]. A brief introduction to functional data analysis with B-spline representation of functions and description of density functions forms the first section. Consequently, FPCA and simplicial FPCA are described and applied to data containing salary distributions in regions of Austria.

This dissertation thesis is based on the following papers that were published, accepted or submitted during my Ph.D. study:

- **Hrůzová K.**, Hron K., Rypka M., Fišerová E. (2013). Covariance structure of principal components for three-part compositional data. *Acta Universitatis Palackianae Olomucensis, Facultas Rerum Naturalium. Mathematica* **52** (2), 61–69.

- Monti G.S., Migliorati S., Hron K., **Hrůzová K.**, Fišerová E. (2014). Logratio approach in curve fitting for concentration-response experiments. *Environmental and Ecological Statistics* **21** (2), 275–295.

- Hron K., Menafoglio A., Templ M., **Hrůzová K.**, Filzmoser P. (2016). Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics and Data Analysis* **94**, 330–350.

- **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..

- **Hrůzová K.**, Rypka M., Hron K. (2016). Compositional analysis of trade flows structure. *submitted*

All the computations and graphical outputs were performed by the statistical software R [93] using the basic packages and, unless otherwise stated, packages `robCompositions` [101], `compositions` [10], `ThreeWay` [47] and `fda` [92].

Chapter 1

Compositional data

Compositional data [1, 89] are strictly positive multivariate observations that carry only relative information. By the relative information it is meant that absolute values are no longer important for the analysis, instead, ratios between parts of a composition capture the only relevant information. The sample space of representations of compositional data within the equivalence class of proportional vectors is the simplex [1, 87, 89], which is defined as a set of strictly positive real numbers that sum up to a constant,

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

where κ is any positive real number, e.g. 1 in case of proportions or 100 for percentages.

A composition is not necessarily characterized by a constant sum of its components, e.g. every country has different gross domestic product (when its relative structure is of interest), different population or an area, however, it is possible to scale the data using the closure operation defined as

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

Note that the relative information between components remains unchanged, this is a consequence of scale invariance - crucial property of compositional data described bellow.

Since this work is focused on economic data, the basic difference between real and compositional data will be shown on a simple example of household expenditures. Every family disposes with the family budget which is distributed into a few basic expenditure categories, e.g. housing, bank payments, food, clothes, health, savings, etc.. The absolute values of such expenditures are different between families and thus they are not informative if the relative structure of expenditures is of main interest. In such a case, by expressing the expenditures in percentages, the comparison becomes more meaningful. Another feature is the relative scale of compositions. Suppose that housing expenditures increased from 500 EUR to 600 EUR and the savings from 100 EUR to 200 EUR. The difference from the Euclidean (standard) perspective is the same - 100 EUR, but taking ratios of these values into account the change in the first case is 1.2 times while in the second case it is twice as much.

Properties of compositional data can be formalized by principles of compositional data analysis [22, 89]. Among them, *scale invariance* property and *subcompositional coherence* seem to be the most important when analyzing compositional data. The first one means that the information conveyed by a composition does not depend on the units in which a composition is measured, i.e. characteristics of compositions should be invariant under a change of scale. According to the second one, the information contained in a composition of D parts should not be in a conflict with that coming from a subcomposition containing d parts, where $d \leq D$. The last principle is called *permutation invariance* - reordering parts of a composition does not affect the included information. Specific nature of compositional data as described above is captured by the Aitchison geometry on the simplex [87, 89].

1.1. Aitchison geometry

The Aitchison geometry with Euclidean vector space structure follows closely the above stated principles of compositional data analysis [87]. Basic operations substituting sum of two real vectors and multiplication of a vector by a scalar are called perturbation and power transformation, respectively. Their definition for $\mathbf{x} \in \mathcal{S}^D$, $\mathbf{y} \in \mathcal{S}^D$ and $\alpha \in \mathbb{R}$ follows,

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 \cdot y_1, \dots, x_D \cdot y_D)', \;\; \alpha \odot \mathbf{x} = \mathcal{C}(x_1^{\alpha}, \dots, x_D^{\alpha})'.$$

The triple $(\mathcal{S}^D, \oplus, \odot)$ forms vector space structure [89]. It means that operations of perturbation and power transformation follow the same properties as sum and scalar multiplication in the Euclidean geometry:

1. commutative property: $\mathbf{x} \oplus \mathbf{y} = \mathbf{y} \oplus \mathbf{x}$;
2. associative property: $(\mathbf{x} \oplus \mathbf{y}) \oplus \mathbf{z} = \mathbf{x} \oplus (\mathbf{y} \oplus \mathbf{z})$; $\alpha \odot (\beta \odot \mathbf{x}) = (\alpha \cdot \beta) \odot \mathbf{x}$;
3. distributive property 1: $\alpha \odot (\mathbf{x} \oplus \mathbf{y}) = (\alpha \odot \mathbf{x}) \oplus (\alpha \odot \mathbf{y})$;
4. distributive property 2: $(\alpha + \beta) \odot \mathbf{x} = (\alpha \odot \mathbf{x}) \oplus (\beta \odot \mathbf{x})$;
5. neutral element: $\mathbf{n} = \mathcal{C}(1, \dots, 1)'$; $1 \odot \mathbf{x} = \mathbf{x}$; where $\mathbf{n}$ is the barycenter of the simplex, note that neutral element is unique;

To obtain Euclidean vector space, inner product and the corresponding norm and distance are defined as well:

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

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

Note that $\mathbf{x} \ominus \mathbf{y}$ is also called perturbation-subtraction of compositions $\mathbf{x}$ and $\mathbf{y}$.

1.2. Coordinate representation

Since almost all standard statistical methods are defined in real space, it is not appropriate to apply them directly to compositions. In order to perform statistical processing using standard multivariate tools, it is necessary to express compositions first in proper real coordinates [89].

A challenging question is how to define coordinates. Any real vector, $\mathbf{x} \in \mathbb{R}^D$, can be expressed using coordinates of the canonical basis,

$$\mathbf{x} = x_1(1, 0, \ldots, 0)' + x_2(0, 1, 0, \ldots, 0)' + \cdots + x_D(0, \ldots, 0, 1)' = \sum_{i=1}^{D} x_i \mathbf{e}_i.$$

The main problem here is that the basis $\mathbf{e}_i$ does not respect the vector space structure of $\mathcal{S}^D$. To find an appropriate orthonormal basis, which seems to be preferable from the geometrical perspective, we can firstly find a generating system with respect to the Aitchison geometry. Taking $\mathbf{w}_i = \mathcal{C}(\exp \mathbf{e}_i) = \mathcal{C}(1, 1, \ldots, e, \ldots, 1)'$, $i = 1, \ldots, D$, where e is placed in the i-th place, we can express a composition $\mathbf{x} \in \mathcal{S}^D$ as

$$\mathbf{x} = \bigoplus_{i=1}^{D} \ln x_i \oplus \mathbf{w}_i = \ln x_1 \odot (e, 1, \ldots, 1)' \oplus \cdots \oplus \ln x_D \odot (1, \ldots, 1, e)'.$$

Due to fact that coefficients with respect to generating system are not unique, we can use the following expression

$$\mathbf{x} = \bigoplus_{i=1}^{D} \ln \frac{x_i}{g(\mathbf{x})} \odot \mathbf{w}_i,$$

where $g(\mathbf{x}) = \left(\prod_{i=1}^{D} x_i\right)^{1/D} = \exp\left(\frac{1}{D}\sum_{i=1}^{D} \ln x_i\right)$ is the geometric mean of a composition. These coefficients are known as centred logratio (clr) coordinates that are defined as

$$\operatorname{clr}(\mathbf{x}) = \left(\ln \frac{x_1}{g(\mathbf{x})}, \ldots, \ln \frac{x_D}{g(\mathbf{x})}\right)'. \tag{1.3}$$

Although the clr coordinates are symmetric in the components, the sum of the coefficients is zero and this leads to singular covariance matrix. Nevertheless, they are still used in the practice because they translate operations and metrics from the simplex endowed with the Aitchison geometry into real space. Particularly, for compositions $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ and real constants α, β it holds that

$$\operatorname{clr}(\alpha \odot \mathbf{x} \oplus \beta \odot \mathbf{y}) = \alpha \cdot \operatorname{clr}(\mathbf{x}) + \alpha \cdot \operatorname{clr}(\mathbf{y});$$

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle \operatorname{clr}(\mathbf{x}), \operatorname{clr}(\mathbf{y}) \rangle ;$$

$$\|\mathbf{x}\|_a = \|\operatorname{clr}(\mathbf{x})\| ; \ d_a(\mathbf{x}, \mathbf{y}) = d(\operatorname{clr}(\mathbf{x}), \operatorname{clr}(\mathbf{y})).$$

Since the Aitchison geometry has dimension one less than the number of components $(D-1)$, the clr coefficients are not coordinates with respect to a basis of the simplex. At the early stage of the logratio methodology, the additive logratio (alr) coordinates [1] were used as well. In this case, each part of the composition is divided by one chosen part, e.g. the last part x_D, to obtain the respective logratio. This leads to a vector of alr coordinates which is of dimension $D-1$:

$$\operatorname{alr}(\mathbf{x}) = \left(\ln \frac{x_1}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D} \right)'.$$

It is obvious that alr coordinates are not symmetrical in the components and, unlike clr coordinates, they do not preserve distances. Thus they can be used only for modeling purposes. The reason is that alr coordinates do not correspond to an orthonormal basis of the simplex. To find them, the Gram-Schmidt procedure can be used.

In general suppose $\mathbf{e}_i, \ i = 1, \ldots, D-1$, form an orthonormal basis of $\mathcal{S}^D$ and $\boldsymbol{\psi}$ is the $(D-1, D)$-matrix whose rows are $\operatorname{clr}(\mathbf{e}_i)$. Orthonormal basis satisfies condition that $\langle \mathbf{e}_i, \mathbf{e}_j \rangle_a = 1$ for $i = j$ and zero otherwise. This implies that $\boldsymbol{\psi}\boldsymbol{\psi}' = \mathbf{I}_{D-1}$, where $\mathbf{I}_{D-1}$ is the identity matrix of dimension $D-1$.

With a particular choice of the orthonormal basis, the composition $\mathbf{x} \in \mathcal{S}^D$ can be written as

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

where $\mathbf{x}^* = (x_1^*, \ldots, x_{D-1}^*)'$ is the vector of coordinates of $\mathbf{x}$ with respect to this basis. The resulting coordinates are called isometric logratio (ilr) coordinates

[28]. The corresponding mapping is isometric isomorphism between $\mathcal{S}^D$ and $\mathbb{R}^{D-1}$ and thus it preserves distances and translates operations similarly as for the clr coordinates.

Now we introduce ilr coordinates that are used in this work [52, 57]. A set of D orthonormal coordinate systems is considered, namely $\mathbf{z}^{(l)} = (z_1^{(l)}, \ldots, z_{D-1}^{(l)})'$, $l = 1, \ldots, D$,

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}}, \; i = 1, \ldots, D-1. \tag{1.4}$$

Here $(x_1^{(l)}, x_2^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})'$ stands for such a permutation of the parts $(x_1, \ldots, x_D)'$, that always the l-th compositional part fills the first position, $(x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)'$. In such a configuration first ilr variable $z_1^{(l)}$ explains all the relative information (logratios) about the original compositional part x_l (it is nothing else than a scaled aggregation of all logratios with x_l), the coordinates $z_2^{(l)}, \ldots, z_{D-1}^{(l)}$ then explain the remaining logratios in the composition [39]. Note that the only important position is that of $x_1^{(l)}$ (that is interpretable through $z_1^{(l)}$), the other parts can be chosen arbitrarily because different ilr coordinates are orthogonal rotations of each other [28]. Of course, $z_1^{(l)}$ cannot be identified with compositional part x_l, as the other parts are also naturally involved through the corresponding logratios. Its interpretation is thus limited due to the specific structure of the Aitchison geometry. We can also see that this coordinate is formed by a logratio between the part x_l and an “average part”, resulting from the geometric mean of the remaining parts in the composition. Therefore, values of $z_1^{(l)}$ represent a measure of dominance of the part x_l with respect to the other parts.

The simplest case is when a two-part composition $\mathbf{x} = (x, \kappa - x)'$ is considered. This type is used for the univariate data expressed in percentages or parts of a whole. The constant κ stands either for the unit constraint (in case of proportions, $\kappa = 1$) or, generally, for a chosen positive number, and represents just a proper scale representation of compositions. For further details concerning the particular case of two-part compositions, see, e.g., [36, 54]. The ilr coordinate is then defined as

$$x^* = \frac{1}{\sqrt{2}} \ln \frac{x}{\kappa - x}. \tag{1.5}$$

It is easy to see that this coordinate is proportional to well-known logit transformation.

1.3. Descriptive statistics

To explore the basic characteristics of any dataset, the standard descriptive statistics are used. However, they are not appropriate for compositional data since they do not follow the Aitchison geometry. Instead of the arithmetic mean and variance (covariance matrix), new measures called center, variation matrix and total variance were introduced [89].

Center is a measure of central tendency for compositional dataset, $\mathbf{X}_{n\times D}$, with compositions $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD})'$ for $i = 1, \ldots, n$ in its rows. It is defined as

$$\operatorname{cen}(\mathbf{X}) = \mathcal{C}(g_1, \ldots, g_D)', \tag{1.6}$$

where $g_i = \left(\prod_{j=1}^{n} x_{ij}\right)^{1/n}$, $i = 1, \ldots, D$ is the geometric mean. The (closed) center corresponds to barycenter of the simplex.

The dispersion is described by the variation matrix,

$$\mathbf{T} = \begin{pmatrix} t_{11} & t_{12} & \ldots & t_{1D} \\ t_{21} & t_{22} & \ldots & t_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ t_{D1} & t_{D2} & \ldots & t_{DD} \end{pmatrix}, \; t_{ij} = \operatorname{var}\left(\ln \frac{x_i}{x_j}\right). \tag{1.7}$$

It is obvious that the main diagonal is formed by zeros since $t_{ii} = \operatorname{var}\left(\ln \frac{x_i}{x_i}\right) = 0$. From $t_{ij},\ i, j = 1, \ldots, D$ we can conclude about proportionality of x_i and x_j. If t_{ij} is zero, or nearly so, then x_i and x_j are proportional, or nearly so.

To measure the global dispersion, the total variance is given by

$$\operatorname{totvar}(\mathbf{X}) = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} t_{ij}. \tag{1.8}$$

It summarises elements of the variation matrix and can also be computed using clr and ilr coordinates. For more, see the Section 2.3.

Chapter 2

Dimension reduction methods

In economic world, one can be interested in analyzing the multivariate structure of a dataset. One of the most popular methods for this purpose is principal component analysis. It leads to dimension reduction based on linear combination of the original data which depletes most of the variability. The results can then be displayed in biplot [45] which is a scatterplot of the first two principal components where scores are displayed as points and loadings as rays.

In what follows, we briefly describe the basic idea of principal component analysis and the construction and interpretation of compositional biplot. Next, the covariance structure of principal components for three-part composition together with an illustrative example is considered. For dimension reduction of three-way compositional data, we also briefly introduce parallel factor analysis and finally, all the mentioned methods are applied to real-world data on trade flows structure.

2.1. Principal component analysis

Given the mean-centered real data matrix $\mathbf{X}_{(n\times D)}$, principal components (PCs) are defined as linear combinations of original data such as $\mathbf{U} = \mathbf{XB}$, where $\mathbf{U}_{(n\times D)}$ is the score matrix, whose columns $(\mathbf{u}_1, \ldots, \mathbf{u}_D)$ are called principal components, and matrix $\mathbf{B}_{(D\times D)}$ is the loading matrix [50]. For the first PC ($\mathbf{u}_1 = \mathbf{b}_1\mathbf{x}_1$) we require maximal variance which is achieved by determining the vector $\mathbf{b}_1$

with the condition $\|\mathbf{b}_1\| = 1$ [58]. The vector $\mathbf{b}_2$ is also determined by the requirement of maximum variance of the second PC, in addition $\|\mathbf{b}_2\| = 1$ and $\mathbf{b}_1'\mathbf{b}_2 = 0$. Generally, j-th PC is defined as $\mathbf{u}_j = \mathbf{b}_j\mathbf{x}_j$ for $2 < j \leq D$, where $\mathbf{b}_j$ is chosen to maximize the variance of $\mathbf{u}_j$ under the conditions $\|\mathbf{b}_j\| = 1$ and $\mathbf{b}_j'\mathbf{b}_k = 0$ for $1 \leq k < j$. Thus the loading vectors are normalised and orthogonal to each other (the orthogonality condition holds also for the principal components) [61].

The question is how we can obtain such matrix $\mathbf{B}$ which fulfills these conditions. There are several options but one of the most popular is the eigenvalue decomposition of the covariance matrix [61]. Let's denote the covariance matrix of $\mathbf{X}$ by $\mathbf{\Sigma}$. It can be decomposed as follows,

$$\mathbf{\Sigma} = \mathbf{B}\mathbf{\Lambda}\mathbf{B}' = \sum_{i=1}^{D} \lambda_i \mathbf{b}_i \mathbf{b}_i',$$

where $\mathbf{\Lambda} = \mathrm{Diag}\{\lambda_1, \ldots, \lambda_D\}$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D \geq 0$ are the eigenvalues of $\mathbf{\Sigma}$ and $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_D)$ is the matrix of the orthonormal eigenvectors that form the loading vectors in PCA. An alternative approach to PCA is developed in the next section.

2.2. Compositional biplot

Since we are interested in relative structure of data at hand, compositional biplot [2, 70] seems to be their appropriate graphical display, if dimension reduction is an inherent requirement. Below we explain how it is constructed and interpreted. Suppose we have a compositional data matrix $\mathbf{X}_{n\times D}$. As principal components are applied to centered data, it is necessary to center the data matrix first using perturbation as $\mathbf{x}_i \oplus \mathrm{cen}^{-1}(\mathbf{X}) = \mathbf{x}_i \ominus \mathrm{cen}(\mathbf{X})$ for $i = 1, \ldots, n$. This operation shifts the center of the data into neutral element $\mathbf{n}$ (we also refer to centering in the Aitchison sense). Next step is to express the dataset in clr coordinates, $\mathbf{Z} = \mathrm{clr}(\mathbf{X})$ [89]. Since the clr coordinates preserve distances, we can apply standard singular value decomposition such as

$$\mathbf{Z} = \mathbf{LKM}',$$

where $\mathbf{L}$ and $\mathbf{M}$ are the orthonormal matrices of eigenvectors of $\mathbf{ZZ}'$ and $\mathbf{Z}'\mathbf{Z}$, respectively. The matrix $\mathbf{K} = \mathrm{Diag}(k_1, \ldots, k_s)$, where $k_i = \sqrt{\lambda_i}$ are singular values that are square roots of the s positive eigenvalues of either $\mathbf{ZZ}'$ or $\mathbf{Z}'\mathbf{Z}$. Biplot is usually formed using the first two principal components, in our case we take the first two singular values and corresponding eigenvectors. Then $\mathbf{Z}$ has to be written as a product of two matrices $\mathbf{GH}'$, where $\mathbf{G}_{n\times 2}$ and $\mathbf{H}_{D\times 2}$. The biplot is just a representation of vectors $\mathbf{g}_i$, rows of $\mathbf{G}$, and $\mathbf{h}_j$, rows of $\mathbf{H}$, in a plane. The vectors $\mathbf{g}_i$ are termed row markers of $\mathbf{Z}$ and correspond to projections of the n samples on the plane defined by the first two eigenvectors of $\mathbf{ZZ}'$. Vectors $\mathbf{h}_j$ are called column markers and correspond to projections of D clr-coefficients on the plane defined by the first two eigenvetors of $\mathbf{Z}'\mathbf{Z}$.

Due to construction of the biplot in clr coordinates, it is necessary to adapt its interpretation accordingly [70, 89]. The basic terms are ray, which joins the origin to a vertex $\mathbf{h}_j$, and link, which joins two vertices $\mathbf{h}_j$ and $\mathbf{h}_k$. Links and rays provide information about the relative variability in a compositional dataset: length of a link between $\mathbf{h}_j$ and $\mathbf{h}_k$ approximates standard deviation of the logratio between j-th and k-th compositional parts and length of a ray approximates standard deviation of the respective clr coefficient. Consequently, if the vertices coincide, then the variance of corresponding logratio is approximately zero and this means that the corresponding two parts are proportional. Links also provide information about correlation of two pairwise logratios: suppose two links $\overline{jk}$ and $\overline{il}$ intersect in M, then

$$\cos(jMi) \approx \mathrm{corr}\left(\ln\frac{x_j}{x_k}, \ln\frac{x_i}{x_l}\right).$$

2.3. Covariance structure of principal components for three-part compositions

Special attention deserve three-part compositions, $\mathbf{x} = (x_1, x_2, x_3)'$, due to the possibility of representing them graphically in ternary diagram. Ternary diagram [89] is an equilateral triangle consisting of vertices X_1, X_2 and X_3, where a composition $\mathbf{x}$ is plotted at a distance x_1 from the opposite side of vertex X_1, at a distance x_2 from the opposite side of vertex X_2 and at a distance x_3 from the opposite side of X_3.

There are three possible choices of the ilr coordinates according to (1.4) (up to orientation of coordinates), that differ only in permutation of the parts x_1, x_2, x_3,

$$z_{11} = \frac{\sqrt{2}}{\sqrt{3}} \ln \frac{x_1}{\sqrt{x_2 x_3}}, \; z_{12} = \frac{1}{\sqrt{2}} \ln \frac{x_2}{x_3}, \tag{2.1}$$

$$z_{21} = \frac{\sqrt{2}}{\sqrt{3}} \ln \frac{x_2}{\sqrt{x_1 x_3}}, \; z_{22} = \frac{1}{\sqrt{2}} \ln \frac{x_1}{x_3}, \tag{2.2}$$

$$z_{31} = \frac{\sqrt{2}}{\sqrt{3}} \ln \frac{x_3}{\sqrt{x_1 x_2}}, \; z_{32} = \frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2}. \tag{2.3}$$

The interpretation of orthonormal coordinates can be obtained from their covariance structure, expressed using variances of log-ratios [39, 40]. In case of (2.1), the variances of z_{11} and z_{12} are given by

$$\mathrm{var}(z_{11}) = \frac{1}{3}\mathrm{var}\left(\ln \frac{x_1}{x_2}\right) + \frac{1}{3}\mathrm{var}\left(\ln \frac{x_1}{x_3}\right) - \frac{1}{6}\mathrm{var}\left(\ln \frac{x_2}{x_3}\right), \; \mathrm{var}(z_{12}) = \frac{1}{2}\mathrm{var}\left(\ln \frac{x_2}{x_3}\right). \tag{2.4}$$

Thus the first coordinate captures all the relative information about the first compositional part (expressed by log-ratios between x_1, x_2 and x_1, x_3, respectively). The second coordinate stands for the remaining log-ratio between x_2, x_3. The variance of z_{11} consists of variances of the first two mentioned log-ratios including x_1 in the nominator and it is reduced by the variance of log-ratio of remaining two compositional parts. This is a consequence of the fact that each ilr variable forms a log-contrast, i.e. term of the form $\mathbf{h}'\ln\mathbf{x}$, where $\mathbf{h}'\mathbf{1} = h_1 + h_2 + h_3 = 0$. Furthermore, the total variance, which represents the sum of variances of both coordinates, results in

$$\mathrm{totvar}(\mathbf{x}) = \mathrm{var}(z_{11}) + \mathrm{var}(z_{12}) = \frac{1}{3}\left[\mathrm{var}\left(\ln\frac{x_1}{x_2}\right) + \mathrm{var}\left(\ln\frac{x_1}{x_3}\right) + \mathrm{var}\left(\ln\frac{x_2}{x_3}\right)\right]. \tag{2.5}$$

Analogous relations would be obtained also for (2.2) and (2.3) by permutation of parts of the original composition.

The main goal of this section is to analyze the variance structure of the well-known principal components as a popular tool for dimension reduction and its impact to interpretation of these orthonormal coordinates [55]. Note that principal components are obtained from such rotation of the original variables which maximizes variance of the resulting coordinates. Although in case of standard real data the covariance structure of principal components can be also expressed using elements of the original covariance matrix [59], we will follow an alternative way of its derivation that enables a deeper insight into covariance structure of three-part compositional data.

2.3.1. Variance structure of principal components

At the beginning of this part we introduce a general constrained problem of finding stationary values [48] that will be used consequently to derive the main theorem concerning covariance structure of principal components for three-part compositional data, denoted in the following as z_1^*, z_2^*. Taking the main idea of principal component analysis into account, we search for maximal difference between variances of both variables.

Let $\mathbf{A}$ be a real symmetric matrix of order D and $\mathbf{c}$ a given real vector that fulfills the condition $\mathbf{c}'\mathbf{c} = 1$. The goal is to find the stationary values of $\mathbf{h}'\mathbf{A}\mathbf{h}$, taking constraints $\mathbf{h}'\mathbf{h} = 1$, $\mathbf{c}'\mathbf{h} = 0$ into account. Denote

$$\varphi(\mathbf{h}, \nu, \mu) = \mathbf{h}'\mathbf{A}\mathbf{h} - \nu(\mathbf{h}'\mathbf{h} - 1) + 2\mu\mathbf{h}'\mathbf{c}, \tag{2.6}$$

where ν, μ are Lagrange multipliers. Differentiating (2.6) with respect to $\mathbf{h}$ leads to

$$\mathbf{Ah} - \nu\mathbf{h} + \mu\mathbf{c} = \mathbf{0}. \tag{2.7}$$

Multiplying (2.7) from left by $\mathbf{c}'$ and using the condition $\mathbf{c}'\mathbf{c} = 1$, we have

$$\mu = -\mathbf{c}'\mathbf{Ah}. \tag{2.8}$$

Then substituting (2.8) into (2.7) we obtain

$$\mathbf{PAh} = \nu\mathbf{h}, \tag{2.9}$$

where $\mathbf{P} = \mathbf{I} - \mathbf{cc}'$. Although $\mathbf{P}$ and $\mathbf{A}$ are symmetric, $\mathbf{PA}$ is not necessarily so. Note that $\mathbf{P}^2 = \mathbf{P}$, so that $\mathbf{P}$ is a projection matrix.

It is well-known that for two arbitrary square matrices $\mathbf{E}$ and $\mathbf{F}$, the eigenvalues of $\mathbf{EF}$ equal the eigenvalues of $\mathbf{FE}$. Thus we can write

$$\lambda(\mathbf{PA}) = \lambda(\mathbf{P}^2\mathbf{A}) = \lambda(\mathbf{PAP}),$$

where λ corresponds to any (fixed) eigenvalue of the matrix in brackets.

The matrix $\mathbf{PAP}$ is symmetric and hence one can use the standard algorithms for finding its eigenvalues. Then if we denote $\mathbf{K} = \mathbf{PAP}$ and if $\mathbf{Kz}_i = \lambda_i\mathbf{z}_i$, it follows that $\mathbf{h}_i = \mathbf{Pz}_i$, where $\mathbf{h}_i$ is the eigenvector which satisfies (2.9) and also the initial problem. At least one eigenvalue of $\mathbf{K}$ will be equal to zero, and $\mathbf{c}$ will be an eigenvector associated with a zero eigenvalue.

The following lemma (see [1, p. 93]) establishes a relation between log-contrasts, corresponding to orthonormal coordinates and their covariance structure.

Lemma 1. *Variances and covariances for log-contrasts* $\mathbf{h}_1' \ln \mathbf{x}$ *and* $\mathbf{h}_2' \ln \mathbf{x}$ *of a D-part composition* $\mathbf{x}$ *are*

$$\operatorname{var}(\mathbf{h}_1' \ln \mathbf{x}) = -\frac{1}{2}\mathbf{h}_1'\mathbf{T}\mathbf{h}_1, \quad \operatorname{var}(\mathbf{h}_2' \ln \mathbf{x}) = -\frac{1}{2}\mathbf{h}_2'\mathbf{T}\mathbf{h}_2, \tag{2.10}$$

$$\operatorname{cov}(\mathbf{h}_1' \ln \mathbf{x}, \mathbf{h}_2' \ln \mathbf{x}) = -\frac{1}{2}\mathbf{h}_1'\mathbf{T}\mathbf{h}_2, \tag{2.11}$$

where $\mathbf{T}$ *is the variation matrix defined by*

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

Theorem 2. *The covariance structure of principal components (orthonormal coordinates)* z_1^*, z_2^* *for three-part composition* $\mathbf{x} = (x_1, x_2, x_3)'$ *can be expressed as*

$$\begin{aligned} \mathrm{var}(z_1^*) &= \frac{a+b+c}{6} + \frac{\sqrt{(a-b)^2 + (b-c)^2 + (c-a)^2}}{3\sqrt{2}}, \\ \mathrm{var}(z_2^*) &= \frac{a+b+c}{6} - \frac{\sqrt{(a-b)^2 + (b-c)^2 + (c-a)^2}}{3\sqrt{2}}, \end{aligned} \tag{2.12}$$

where $a,\ b,\ c$ *correspond to* $\mathrm{var}\left(\ln \frac{x_1}{x_2}\right),\ \mathrm{var}\left(\ln \frac{x_1}{x_3}\right),\ \mathrm{var}\left(\ln \frac{x_2}{x_3}\right)$*, respectively.*

Proof. Taking properties of the variation matrix into account [1], the general problem of finding stationary values can be replaced by maximizing $\mathbf{h}'\mathbf{T}\mathbf{h}$ with respect to constraints $\mathbf{h}'\mathbf{c} = 0,\ \mathbf{h}'\mathbf{h} = 1$. Here $\mathbf{c} = \frac{1}{\sqrt{3}}(1,1,1)'$ and

$$\mathbf{T} = -\frac{1}{2} \begin{pmatrix} 0 & a & b \\ a & 0 & c \\ b & c & 0 \end{pmatrix}.$$

Consequently, by solving the equation $\mathbf{K}\mathbf{h} = \lambda\mathbf{h}$ ($\mathbf{K} = \mathbf{P}\mathbf{T}\mathbf{P}$, $\mathbf{P} = \mathbf{I} - \mathbf{c}'\mathbf{c}$), the resulting non-zero eigenvalues correspond to variances of principal components and eigenvectors to their log-contrasts. □

Note that in context of compositional data analysis, the matrix $\mathbf{K}$ represents covariance matrix of clr coordinates of compositions [1]. It is easy to see that principal components and their variances, resulting as log-contrasts of eigenvectors and (non-zero) eigenvalues of the clr covariance matrix, respectively, correspond to those coming from ilr coordinates [34]. Log-contrasts, corresponding to coordinates z_1^*, z_2^*, thus can be expressed as

$$\mathbf{h}_1 = \left(-\frac{a-c+S}{2(b-c)\sqrt{S^2+(a-c)(a-b)S}}, \frac{a-b+S}{2(b-c)\sqrt{S^2+(a-c)(a-b)S}}, \frac{1}{2\sqrt{S^2+(a-c)(a-b)S}}\right)'$$

and

$$\mathbf{h}_2 = \left(-\frac{a-c-S}{2(b-c)\sqrt{S^2-(a-c)(a-b)S}}, \frac{a-b-S}{2(b-c)\sqrt{S^2-(a-c)(a-b)S}}, \frac{1}{2\sqrt{S^2-(a-c)(a-b)S}}\right)',$$

respectively, where $S = \sqrt{\frac{1}{2}[(a-b)^2+(b-c)^2+(c-a)^2]}$. Because z_1^*, z_2^* are orthonormal coordinates, $\mathbf{h}_1$, $\mathbf{h}_2$ are standard and orthogonal log-contrasts, i.e. $\mathbf{h}_1'\mathbf{h}_1 = \mathbf{h}_2'\mathbf{h}_2 = 1$, $\mathbf{h}_1'\mathbf{h}_2 = 0$ (see [1, p. 85] for details). The latter property as well as zero covariance between z_1^* and z_2^* results from construction of principal components [50].

Note that big differences between variances of logratios contribute for maximization of the first principal component at the expense of the second one. This is obvious from the second part of (2.12) - in variance of z_1^* we add square root of the sum of squared differences of these variances while in $\mathrm{var}(z_2^*)$ we subtract it. Furthermore, it is not necessary to consider the covariance because principal components are uncorrelated [50]. Obviously, the interpretation of principal components seems to be not straightforward even with the above decomposition of the covariance structure using variance of log-ratios of compositional parts. It will strongly depend on the analyzed problem. On the other hand, some features of variances of these coordinates are now easily detectable. As already mentioned, the first part of both variances is formed by half of the total variance. Particularly, for higher difference between variances of both principal components high differences between variances of logratios are crucial (see the term contained in the square root).

2.3.2. Gross value added in Germany

The theory described above will be shown on an example of gross value added (GVA) in German regions (recent data are available from the year 2009), see [71] for details.

GVA is a measure of the value of goods and services produced in an area, industry or sector of an economy. It represents the output minus intermediate consumption and it is linked to gross domestic product (GDP) in the following sense:

$$\text{GVA} + \text{taxes on products} - \text{subsidies on products} = \text{GDP}.$$

GVA is used mainly for measuring gross regional domestic product and other measures of the output of entities smaller than the whole economy.

The analyzed dataset consists of the gross value added structure (2009) [71] which is divided into agriculture, production and services of 411 German regions. The values are expressed in percentages, thus we have three-part composition represented with a constant sum constraint 100%.

The three-part compositions can be displayed in ternary diagram. As we can see in Figure 2.1 (left) the data are clustered on the side between Production and Services; this means that the Agriculture part contains mostly small positive values. For better visualization we centered the compositions (in the Aitchison sense) and result is plotted in Figure 2.1 (right).

In the next step the ilr coordinates are constructed according to (2.1)–(2.3). Their scatterplots together with the scatterplot of the first two principal components of the ilr coordinate $\mathbf{z}_1 = (z_{11}, z_{12})'$ are displayed in Figure 2.2. Note that these coordinates are rotations of each other. The upper left plot corresponds to coordinates resulting from (2.1). We can observe that the main data cloud contains higher negative values of z_{11} and rather negative values of z_{12}. It means that the third part (services), which is in the denominator of z_{12}, is dominating in the composition, followed by production and agriculture parts. It is also easy to see that the first coordinate captures more variability of the data

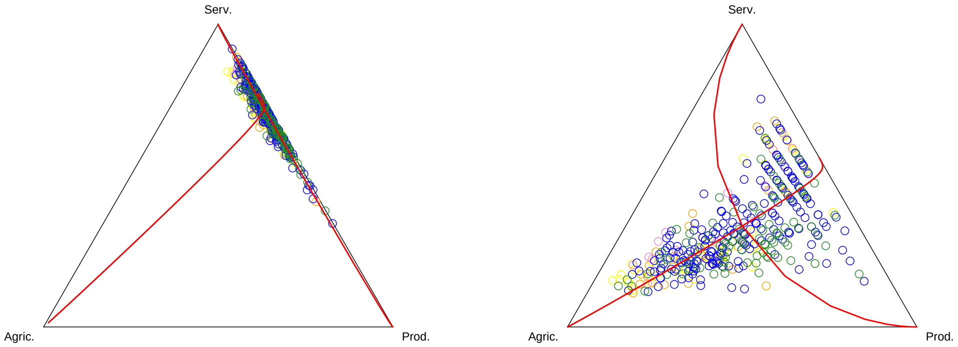

Figure 2.1: Ternary diagrams - non-centered (left) and centered (right) GVA structure data together with the first and second principal components (red lines).

set which is confirmed by Table 1. From this figure we can also see that the main data cloud in the other two coordinates systems is located in the second and fourth quadrant. This means that ratios x_1/x_3 and x_1/x_2 are mostly below zero. This confirms the fact that the part x_1 contributes at least to the GVA structure. Analogous interpretation can be derived also for the remaining two coordinate systems. Consequently, it is not suprising that the scatterplot for principal components is quite close just to coordinates (2.1).

Note that different colours distinguish the federal states of Germany in this sense (that correspond to natural geographical regions):

- yellow for Mecklenburg-Vorpommern, Sachsen and Thüringen,
- orange for Sachsen-Anhalt and Brandenburg,
- violet for Berlin, Schleswig-Holstein, Hamburg and Bremen,
- blue for Niedersachsen, Nordrhein-Westfalen, Hessen and Bayern,
- green for Rheinland-Pfalz, Baden-Württemberg and Saarland.

From the variation matrix (2.13) is evident that the largest variability is contained in logratio between the first and third variable and a bit smaller between

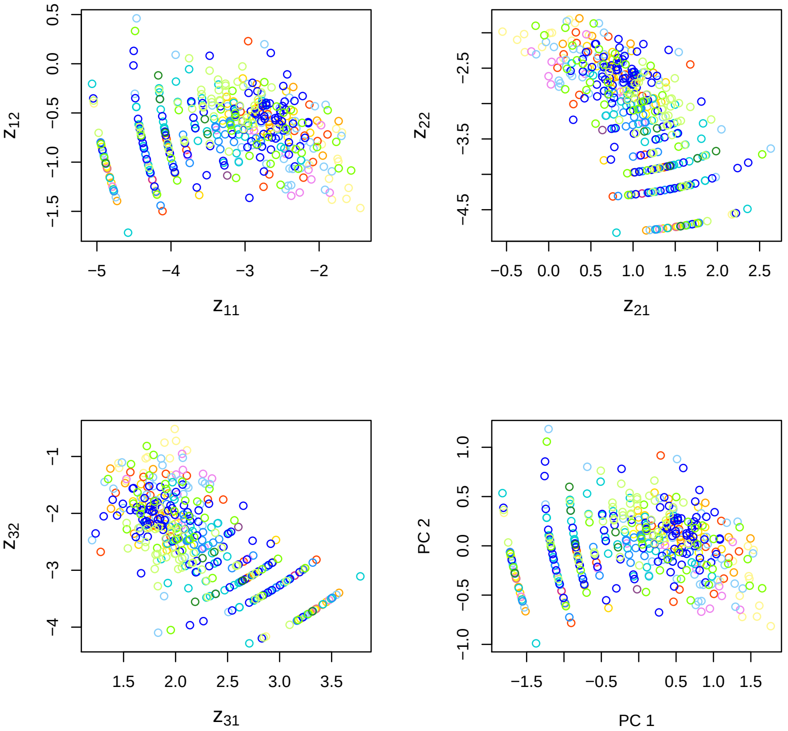

Figure 2.2: GVA structure data, plots of orthonormal coordinates. The upper left plot corresponds to formula (2.1), upper right plot to (2.2), lower left plot to (2.3) and lower right plot to principal components.

first and second variable, while the smallest variance has logratio of the second and third part. This was evident also in Figure 2.2.

$$\mathbf{T} = \begin{pmatrix} 0 & 1.153 & 1.206 \\ 1.153 & 0 & 0.225 \\ 1.206 & 0.225 & 0 \end{pmatrix} \tag{2.13}$$

From the variation matrix (2.13) the variances of principal components using (2.12) can be easily computed, $\mathrm{var}(z_1^*) = 0.749,\ \mathrm{var}(z_2^*) = 0.112$, where the first part of both variance terms, half of the total variance $\mathrm{totvar}(\mathbf{x})$, equals 0.431. Difference between both variances results from the sum of squared differences between variances of log-ratios. Variances of log-ratios with x_1 differ substantially from variance of $\ln(x_2/x_3)$ that once more confirm the exceptional role of the

i	$\mathrm{var}(z_{i1})$	$\mathrm{var}(z_{i2})$
1	0.749	0.112
2	0.258	0.603
3	0.285	0.576

Table 2.1: Variances of ilr coordinates.

services part for the overall variability of the compositional data set.

2.4. Parallel factor analysis (PARAFAC)

By adding another mode into the analysis, bilinear PCA is no longer appropriate. For example, time mode is frequently considered as well, i.e. the samples are observed for given variables in several time slots. For dimension reduction of such data (we refer also to three-way data), the PARAFAC or CANDECOMP method (abbreviation comes from CANonical DECOMPosition) is preferred instead of standard PCA. The input data are decomposed into trilinear components where each component consists of one score vector and unlike PCA two loading vectors; usually we refer simply to three loading vectors. A PARAFAC model of three-way array [11] is given by three loading matrices $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ with elements a_{if}, b_{jf} and c_{kf} that minimize the sum of squares of the residuals e_{ijk},

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \tag{2.14}$$

for $i = 1, \ldots, I,\ j = 1, \ldots, J$ and $k = 1, \ldots, K$.

The advantage of the PARAFAC model is the uniqueness of the solution, when a proper number of components is chosen. The meaning of uniqueness is that the estimated PARAFAC model cannot be rotated without a loss of fit. In [72] have been shown that unique solutions can be expected, if the loading vectors are linearly independent in two of the modes and in the third mode no two loading vectors are linearly dependent. In [68, 69] it was proved that the PARAFAC solution is unique iff $\mathrm{rank}(\mathbf{A}) + \mathrm{rank}(\mathbf{B}) + \mathrm{rank}(\mathbf{C}) \geq 2F + 2$, where F is the number of components.

The solution of the PARAFAC model (estimations of matrices $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$) can be found using the alternating least squares (ALS) method by assuming the loading vectors of two modes known and then estimating the unknown set of parameters of the last mode [11, 66]. Let's consider the case of $F = 1$ first. If an estimate of loading vectors $\mathbf{b}$ and $\mathbf{c}$ is given, $\mathbf{a}$ can be determined by the least-squares solution to the model $\mathbf{X} = \mathbf{a}(\mathbf{b} \otimes \mathbf{c}) + \mathbf{E}$, where $\mathbf{X}$ is unfolded array of size $I \times JK$ (i.e. with respect to the first mode), $(\mathbf{b} \otimes \mathbf{c})$ is the tensor product of the vectors $\mathbf{b}$ and $\mathbf{c}$ and $\mathbf{E}$ is a matrix of errors. We can denote the tensor product $(\mathbf{b} \otimes \mathbf{c})$ by $\mathbf{z}$ (or $\mathbf{Z}$ in case of more components). Then we can define the model as $\mathbf{X} = \mathbf{A}\mathbf{Z} + \mathbf{E}$. The conditional least squares estimate of $\mathbf{A}$ is then given by $\mathbf{A} = \mathbf{X}\mathbf{Z}'(\mathbf{Z}\mathbf{Z}')^{-1}$.

General ALS algorithm [67] can be written in the following way:

1. Choose the number of components, F;
2. Initialize $\mathbf{B}$ and $\mathbf{C}$;
3. Estimate $\mathbf{A}$ from unfolded $\mathbf{X}$, $\mathbf{B}$ and $\mathbf{C}$ by least squares regression;
4. Estimate $\mathbf{B}$ and $\mathbf{C}$ likewise;
5. Continue from 3 until convergence.

Accordingly, $\mathbf{A}$ is $I \times F$ matrix containing the fth loading vector in its fth column. The matrices $\mathbf{B}$ and $\mathbf{C}$ are defined in the same way. The ALS algorithm is improving the fit of the model in each iteration. If the algorithm converges to the global minimum, then the solution is found. Although the ALS method enables to find a solution, it is time consuming, especially for large number of variables.

2.5. Compositional analysis of trade flows structure

The aim of this section is to apply the above introduced PCA and PARAFAC models to real-world dataset that contains structure of export and import in the end-use categories. The motivation comes from the initial question, how to compare export and import of different countries. In the standard case one would compute simply differences between components. However, each country has different area, different size of population, different GDP and different structure of the economy. This means that if we would just subtract import from export values, the results could be completely misleading. The problem can be solved using the perturbation-subtraction (Section 1.1, [3]), i.e. by taking the ratios of export and import for every end-use category.

The motivation for this section comes from the fact that in today's globalised world, export and import play an important role in the country's economic situation. Globalisation causes growth of international trade in goods and services and two structural changes in trade patterns: the increasing importance of emerging economies and rapid growth of trade in intermediate goods as a result of vertical specialisation, meaning that each country is specialised in one or more innovation and production processes and thus it is common for the value chain of a particular final product to span several countries. Trade in intermediate goods currently represents about 56 % of total global trade in goods [82] and therefore we intend to explore trade flows breaking down by end-use categories to better monitor international trade patterns.

2.5.1. Dataset

The dataset called The OECD STAN Bilateral Trade by Industry and End-use [109] was first released in 2011 to highlight the increasing influence of export and import of intermediate goods. The values of import and export of goods are broken down by industrial sectors and, simultaneously, by end-use categories. Estimates are expressed in nominal terms, in current US dollars, and are collected from more than a hundred reporters and partners, including all 34 members of OECD and a wide range of non-members. Note here for the purpose of standard statistical analysis, without consider the relative nature of data, we would have to convert the current US dollars into constant US dollars in order to consider time. However, we are dealing with compositional data which means that we are working only with the ratios between the parts and thus multiplication by any constant does not affect results of the analysis. Following this idea it is not necessary to convert the currency prior to further statistical processing using the logratio methodology.

Breaking down trade in goods according to their end-use [86] adds a new dimension to the traditional commodity-based trade statistics and provides a link to National Accounts Input-Output Tables, in which flows of goods and services are reported according to end-users. Using the basic kinds of domestic end-use categories from the System of National Accounts and the detailed classification systems of trade in goods, bilateral flows of exports and imports can be classified into intermediate goods, household consumption goods and capital goods. However, some kinds of products can be either for intermediate demand and household consumption, or for capital goods in industry and household consumption. Thus it was introduced mixed end-use category which contains personal computers, passenger cars, personal phones, packed medicines and precious goods. The last category is miscellaneous which includes commodities that don't belong to any other categories. To keep the presented study simple we will not consider this category for further calculations.

2.5.2. Statistical analysis

Patterns in the relative structure of export and import of goods cannot be revealed by applying standard multivariate techniques to the raw data as the relevant information is contained exclusively in ratios between the respective components. Nevertheless, for the sake of comparison, principal component analysis was applied both to the original data and to clr coordinates for the year 2012, the most recent completed one in the database.

Obviously, when dealing with economies of different size of trade (with different population, share of trade in economy), straightforward application of PCA (see Figure 2.3) becomes useless. From these biplots it is hard to recognize any

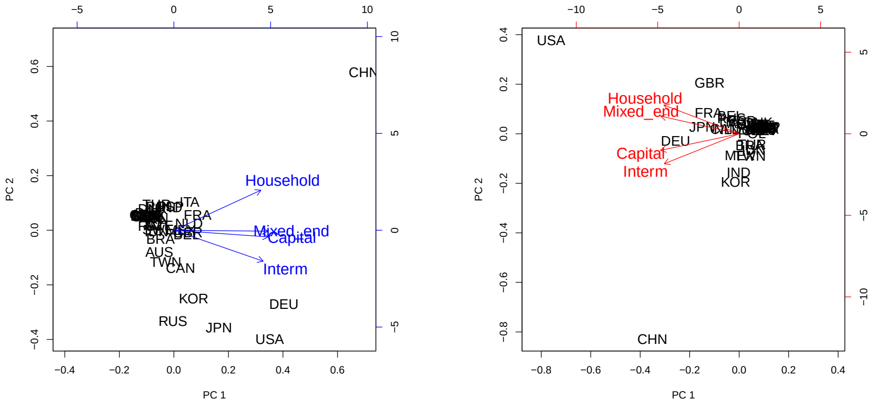

Figure 2.3: Biplots of export and import, applied to the original data.

structure in the dataset, it also seems that all variables are highly correlated.

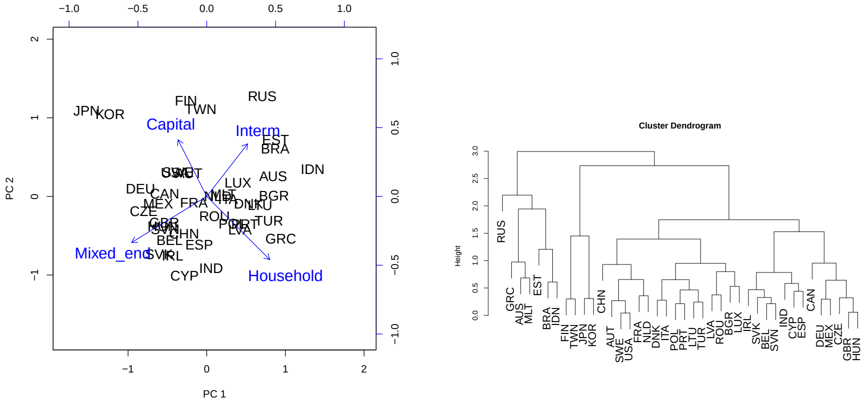

Figure 2.4: Compositional biplot of export (on the left), dendrogram of clr coordinates of export (on the right).

In contrast, when relative contributions of the components, conveyed by clr coordinates (here of end-use categories), are considered instead, PCA and biplot diagrams are much easier interpretable (see Figure 2.4, left). The countries exporting relatively more intermediate goods (Russia, Australia, Brazil), household (Greece, Turkey, India), mixed end-use (middle Europe countries), capital goods (Japan, Korea, Finland) can be well distinguished, no matter of their size. These clusters are also evident from the dendrogram of export (Figure 2.4, right), where the well-known hierarchical clustering with complete linkage [60] was applied on

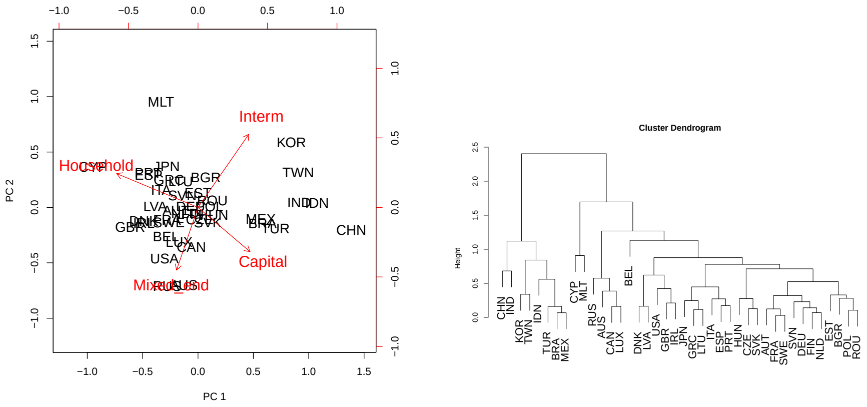

Figure 2.5: Compositional biplot of import (on the left) and dendrogram of clr coordinates of import (on the right).

clr coordinates.

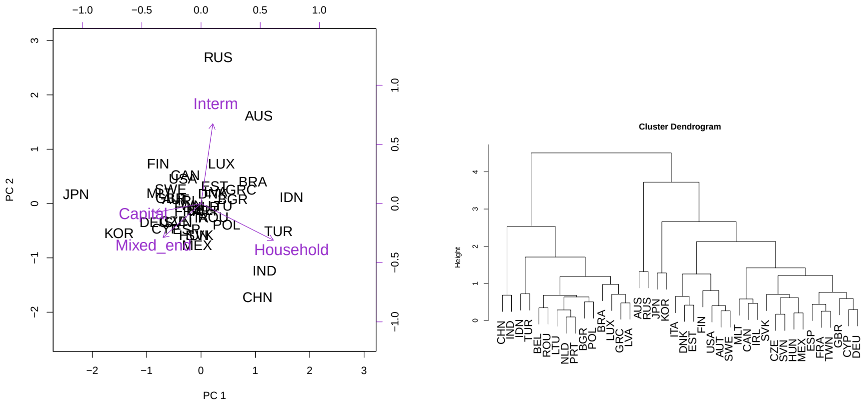

Figure 2.6: Compositional biplot (on the left) and dendrogram (on the right) of clr coordinates of differences between export and import.

Similarly, in Figure 2.5 on the left, the compositional biplot of import is displayed. It is evident that Asian countries such as Korea, Taiwan, India and China import mainly intermediate and capital goods. On the other hand, mixed end-use goods are imported into large countries, namely Russia, Australia, USA and Canada. Middle Europe countries are spread around the origin and Cyprus imports mostly the household consumption goods. This corresponds well to the general perspective of international trade structure of that year [102]. The clusters of countries, recognized in the biplot, can be seen again also from the dendrogram in clr coordinates.

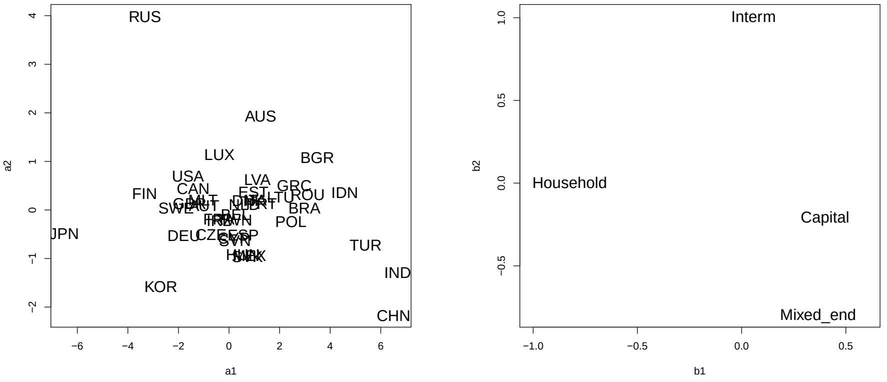

Figure 2.7: Results of the PARAFAC method for differences between exports and imports, mode A (on the left) and mode B (on the right), using clr coordinates.

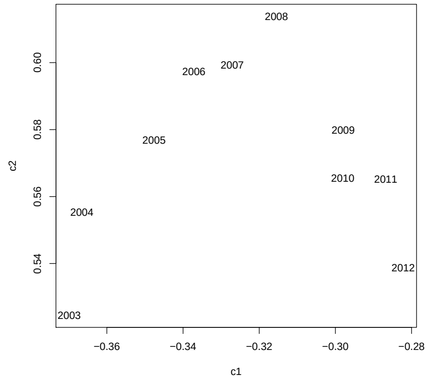

Figure 2.8: Results of the PARAFAC method for differences between exports and imports, mode C, using clr coordinates.

The perturbation operation can be now used to capture relative differences between export and import structure through ratios between the respective components. Consequently, large values of the (log-)ratios will indicate discrepancy between both international flow aspects. From the respective link in Figure 2.6 it is visible that the variance of pairwise logratio between export/import ratios of capital goods and mixed end-use goods, respectively, are very small. Thus the ratios between exports and imports of these end-use categories are very similar. The cluster of China, India, Indonesia and Turkey lies near the household goods variable, thus these countries have the largest difference between export and import for this variable. Russia and Australia have largest difference between export and import of intermediate goods, while Korea and Japan in capital goods.

In order to include also time variable and to get a complete picture about the development in a larger time scale, also PARAFAC modeling was applied to the perturbed data, i.e. to the ratio of export and import components (after expressing them in clr coordinates) for years 2003–2012. Similar results as for the previous figures were obtained that confirms a certain stability of the export/import structure comparing to the only year 2012, considered above. In the mode A (Figure 2.7, left), corresponding to samples, cluster of China, India, Turkey and Indonesia can be seen, as well as cluster of Japan and Korea. In the middle of the plot there is a group of middle European countries and it also seems that Russia differs significantly from the other countries. Mode B (Figure 2.7, right) confirms the result that components Capital and Mixed end-use goods are very similar, when considering ratios of export and import for the years 2003–2012. And finally, mode C displayed in Figure 2.8 shows the development in time, where a clear time pattern with a change point in 2008 is observed, interpretable in terms of global integration. Accordingly, this loading plot well reflects the global crisis in 2008–2009 that has temporarily brought the long-run trend of rising global integration through trade to a halt.

Chapter 3

Regression analysis

Linear regression is a very popular statistical tool in economic world. When we are dealing with compositional data, four main cases might occur. In the latter we are going to describe these cases together with appropriate examples.

This chapter is organized as follows - firstly we briefly remind basics of linear regression analysis and introduce the robust approach to regression modeling. The following sections describe approaches to regression with compositional response, compositional covariates and a special case of regression model with compositional response and covariates, respectively, together with economic examples. Note that the first two cases are introduced since they are necessary for development of regression models described afterwards. The final section is aimed to the case when the relation between parts of a composition is analyzed.

3.1. Linear regression

Firstly, we briefly remind the basics of linear regression. Linear regression or linear model, is used to model the relationship between response, dependent, variables and explanatory, independent, variables, also called covariates [50, 60]. The relation between the response and the explanatory variables is given in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.1}$$

where $\mathbf{y}_{n\times 1}$ is the response vector, $\mathbf{X}_{n\times p}$ is the matrix of covariates, $\boldsymbol{\beta}_{p\times 1}$ is a vector of unknown parameters and $\boldsymbol{\varepsilon}_{n\times 1}$ is the error term which is assumed to have zero expectation and variance $\sigma^2\mathbf{I}$, where $\mathbf{I}$ denotes identity matrix of order n. The parameters $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ are estimated by the least squares method, where sum of residuals is minimized,

$$S_e = (\mathbf{y} - \widehat{\mathbf{y}})'(\mathbf{y} - \widehat{\mathbf{y}}) = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}), \tag{3.2}$$

$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ being the least square estimate of $\boldsymbol{\beta}$ [83]. In many cases, first column of explanatory matrix contains vector of ones, then we define the parameter β_0 which is called the intercept, or the absolute term, and the number of parameters changes to $p + 1$.

To compare how well the model fits the data at hand, the adjusted coefficient of determination, R^2_{adj}, is computed

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1}\frac{S_e}{S_T}, \tag{3.3}$$

where p is the number of parameters and $S_T = (\mathbf{y} - \overline{y}\mathbf{1})'(\mathbf{y} - \overline{y}\mathbf{1})$, $\overline{y} = 1/n\sum_{i=1}^{n} y_i$ is the total sum of squares [83] and $\mathbf{1}$ is an $n \times 1$ vector of ones. When the coefficient is close to one, the regression fits the data well.

Assuming normal distribution of the error term, we can test whether the response depends on the ith column of covariates. Denote v_{ij} elements of the matrix $(\mathbf{X}'\mathbf{X})^{-1}$ and $s^2 = \frac{S_e}{n-p-1}$, then the test statistic for the hypothesis $\beta_i = \beta_i^0$ is

$$T_i = \frac{b_i - \beta_i^0}{\sqrt{s^2 v_{ii}}} \sim \; t_{n-p-1}, \tag{3.4}$$

where t_{n-p-1} denotes Student's t-distribution with $n - p - 1$ degrees of freedom [83].

3.2. Introduction to robust regression

All statistical methods are based on certain assumptions on the data that are supposed to be analyzed. These assumptions improve the efficiency of the methods and can lead to nice mathematical properties of the results. Among the basic assumptions is that all observations belong to the same distribution. However, this property is not often fulfilled in practical situations which can lead to inappropriate results. As a way out, robust methods were designed in order to resist to irregular or outlying observations [42]. A raw approach for obtaining a robust estimation is the identification of irregular observations [37, 42], e.g. using Mahalanobis distances, in case of compositional data [33], their removal and the subsequent application of standard statistical tools to the remaining dataset. However, it is not always possible to find the outliers or to eliminate them from the analysis. In this case the robust estimators are used. A well chosen robust estimator will provide a reliable fit for the whole range spanned by the data points without being influenced by deviating points. Most robust methods can be described as classical methods where the data are weighted, with weights depending on the data. The majority of the data will receive a uniform weight while the more atypical individual cases are, the lower the weight they will get.

Every robust estimator should fulfill some properties to proceed with statistical analysis reasonably [42]. The first property is related to influence function (IF). It measures the influence which a negligible amount of contamination has to an estimator regarding its position in space. Evaluating the IF at the points of a data set reveals how each data point changes the estimator´s behaviour. The IF should be bounded and smooth to keep the robustness. The next property, maxbias curve, measures the bias which an estimator has with respect to the percentage of the worst possible type of contamination. From this curve it is possible to observe that for each estimator, there exists a point where the bias tends to infinity. This point is called breakdown point. It indicates which percentage of the data may be replaced with outliers before the estimator yields aberrant results. It usually ranges from 0 % to 50 % which means that asymptotically half of the data can be contaminated arbitrarily without obtaining completely arbitrary results. And the last basic property concerns statistical efficiency which results in minimal variance. It is well-known that robust estimators have higher variance than classical estimators. Thus it is necessary, when designing the robust estimators, not only to investigate the robustness properties, but also the efficiency properties.

To avoid influence of the outlying observations to regression estimates, the classical regression should be replaced by robust one [37]. Consequently, properties of the regression model are based on properties of the particular robust method. The linear regression model (3.1) can be rewritten as

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i,\ \ i = 1, \dots, n,$$

where $\mathbf{x}_i$ is the ith row of the covariate matrix $\mathbf{X}_{n\times p}$, y_i is an element of the response vector $\mathbf{y}_{n\times 1}$, $\boldsymbol{\beta}_{p\times 1}$ is the vector of unknown regression parameters and ε_i is the error term. Denote the ith residual for a given estimator $\mathbf{b}$ (not necessarily the least squares estimator) as $r_i = r_i(\mathbf{b}) = y_i - \mathbf{x}_i'\mathbf{b}$. The estimation of regression parameters is mostly based on minimization of the size of residuals. For the purposes of this thesis, the least trimmed squares (LTS) method was applied [20]. It is defined as

$$\mathbf{b} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{h} r_{(i)}^2(\boldsymbol{\beta}), \tag{3.5}$$

where $r_{(i)}$ denotes the ordered absolute residuals, $|r_{(1)}| \leq \cdots \leq |r_{(n)}|$, and h is called trimming constant which has to satisfy condition $\frac{n}{2} < h \leq n$. This constant determines the breakdown point of LTS estimator, since $n - h$ observations with the largest residuals do not affect the estimator. For $h = \lceil n/2 \rceil$ (rounded to the nearest integer) the breakdown point is 50% and for $h = n$ the ordinary least squares estimator with breakdown point 0% is achieved. From other properties, $\sqrt{n}$-consistency and asymptotic normality [95] in the location-scale model can be mentioned, however the efficiency of estimator is rather lower.

3.3. Regression analysis with compositional response

It frequently happens that the response variable has compositional character. It implies that we should follow the Aitchison geometry when estimating the regression parameters [23, 89]. Suppose that compositional response variable has n observations and is denoted as $\mathbf{y}_i = (y_{i1}, \ldots, y_{iD})' \in \mathcal{S}^D$, then the regression model is stated as follows,

$$\mathbf{y}_i = \boldsymbol{\beta}_0 \oplus \bigoplus_{j=1}^{p} (x_{ij} \odot \boldsymbol{\beta}_j) \oplus \boldsymbol{\varepsilon}_i, \ i = 1, \ldots, n,$$

where $\mathbf{x}_i = (x_0, x_{i1}, \ldots, x_{ip})'$ is a vector of real covariates with $x_0 = 1$ and $\boldsymbol{\beta}_j$ for $j = 0, \ldots, p$ are compositional regression parameters. The least squares method is often used for fitting the model but it is necessary to apply it in sense of the Aitchison geometry. Thus we minimize the sum of square-norms of the error $\text{SSE} = \sum_{i=1}^{n} \|\boldsymbol{\varepsilon}_i\|_a^2$; generalization of S_e as introduced in (3.2). Due to dimensionality of compositions, the number of coefficients to be estimated in this model is $(p + 1) \times (D - 1)$.

Instead of solving this computationally difficult problem, we can express compositional responses in coordinates with respect to orthonormal basis of simplex [23, 28, 89]. Let $h(\cdot)$ is a coordinate function for the particular orthonormal basis, denote $\mathbf{y}_i^* = h(\mathbf{y}_i), \boldsymbol{\varepsilon}_i^* = h(\boldsymbol{\varepsilon}_i)$ for $i = 1, \ldots, n$ and $\boldsymbol{\beta}_j^* = h(\boldsymbol{\beta}_j), \ j = 0, \ldots, p$. Taking such coordinates, the transformed model is

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

and also SSE can be expressed as

$$\text{SSE} = \sum_{i=1}^{n} \sum_{k=1}^{D-1} (\boldsymbol{\varepsilon}_{ik}^*)^2.$$

Estimation in the regression model thus reduces to $D - 1$ ordinary least squares problems, in other words, we are dealing with multiple regression. After estimating the parameters, we can use inverse ilr mapping to show the results in the original space.

For the purpose of coordinate representation we can choose ilr coordinates (1.4), where always the l-th part of the original composition fills the first position. Then the coordinate $z_1^{(l)}$ expresses all the relative information about the part x_l regarding the other parts. Thus the response is now denoted as $\mathbf{z}^{(l)} = (z_1^{(l)}, \ldots, z_{D-1}^{(l)})', \ l = 1, \ldots, D$. But we need to realize that only the coordinate $z_1^{(l)}$ covers the relative information about the part x_l, the coordinates $z_2^{(l)}, \ldots, z_{D-1}^{(l)}$ include information about the other parts through their logratios. In order to find the model corresponding to each part of the original composition, only the first coordinate is relevant. Accordingly, we obtain D regression models, where real explanatory variables remain the same and the response changes according to which part fills the first position in the composition to construct the coordinate $z_1^{(l)}$.

For illustration of the described theory, the dataset consisting of household expenditures of 27 EU countries (available on http://ec.europa.eu/eurostat or in the package `robCompositions` in software R) and gross domestic product (GDP) per capita from the year 2008 (available on http://data.worldbank.org/indicator/NY.GDP.PCAP.CD) are analyzed. The goal is to find the relation between expenditures on food as a part of composition describing the expenditures structure and GDP per capita. The household expenditures consist of expenditures on food, alcohol, clothing, housing, furnishings, health, transport, communications, recreation, education, restaurants and other. It is clear that expenditures can be considered as composition with 12 parts and to express the relative information on food with respect to all the other parts, the coordinate $z_1^{(1)}$ from (1.4) needs to be computed. This coordinate forms the response variable and GDP per capita is the real explanatory variable. Note that we could also analyze the other coordinates $z_1^{(l)}$ for $l = 2, \ldots, 12$ in the same way.

	GDP		log(GDP)	
parameter	estimate	p-value	estimate	p-value
classical	R^2_{adj}=0.5435		R^2_{adj}=0.765	
b_0	1.721	5.28e-14	7.48822	5.07e-11
b_1	-1.486e-05	6.96e-06	-0.61165	1.50e-09
robust	R^2_{adj}=0.6197		R^2_{adj}=0.812	
b_0	1.781	6.77e-14	7.74831	8.04e-12
b_1	-1.803e-05	1.84e-06	-0.63868	2.11e-10

Table 3.1: Summary of regression outputs for compositinal response.

The regression model is then built as

$$z_{1i}^{(1)} = \beta_0^* + \beta_1^* x_i + \varepsilon_i,\ i = 1, \ldots, n, \tag{3.6}$$

where β_0^*, β_1^* are the regression parameters and x_i contains value of GDP per capita in i-th country. Results are summarized in Table 3.1 and displayed graphically in Figure 3.1. Due to multiplicative scale of positive variables [79], the logarithmic transformation of GDP was applied as well (see the right plot). Note that using logarithmic transformation is very common for economic indicators. From Figure 3.1 (right plot), it is evident that logarithmic transformation of GDP leads also to better fit by the regression model (3.6) comparing to the untransformed version.

We can see that all estimated parameters are significant on the level < 0.0001. From the adjusted coefficient of determination we can observe that logarithmic transformation of GDP fits data better and that there is only a slight difference between the classical and robust regressions. The regression line shows decreasing trend for relative expenditures on food with increasing value of GDP per capita. Thus we can conclude that households in countries with higher GDP spend relatively less money on food; the main reason might be higher relative expenditures on services in these countries.

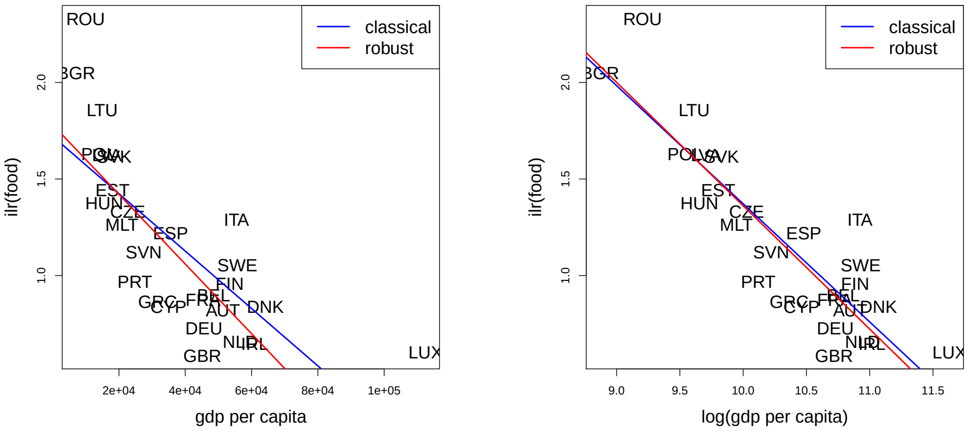

Figure 3.1: Plots of fitted regression lines using original (left) and logarithmized (right) GDP per capita.

3.4. Regression analysis with compositional covariates

Similar approach is used when the covariates are compositional, $\mathbf{x}_i \in \mathcal{S}^D,\ i = 1, \ldots, n$, and the response is observed as a real vector [52, 89]. Then we estimate a surface on $\mathcal{S}^D \times \mathbb{R}$ with the equation

$$y_i = \beta_0 + \langle \boldsymbol{\beta}, \mathbf{x}_i \rangle_a + \varepsilon_i,\ i = 1, \ldots, n,$$

where $\boldsymbol{\beta} \in \mathcal{S}^D$ is the gradient of $\mathbf{y}$ with respect to compositions $\mathbf{x}_i$ and β_0 is a real intercept [89]. Because the response is a real vector, the classical least squares fit can be applied,

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \beta_0 - \langle \boldsymbol{\beta}, \mathbf{x}_i \rangle_a)^2.$$

To avoid computing of the Aitchison inner product, the ilr coordinates of $\mathbf{x}_i$ can be used instead of the original composition. The SSE becomes

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \beta_0 - \langle \boldsymbol{\beta}^*, \mathbf{x}_i^* \rangle)^2.$$

Thus we can fit a linear regression to the response $\mathbf{y}$ as a linear function of $\mathbf{x}^*$. The parameters $\boldsymbol{\beta}^*$ contain the slope coefficients. Although the clr coordinates can be used as well as the ilr coordinates for the purpose of regression modeling, it is not appropriate here due to singular covariance matrix.

We can again use ilr coordinates (1.4) to express the explanatory compositions in real space. Thus the regression model has form

$$y_i = \beta_0 + \beta_1^{(l)} z_{1i}^{(l)} + \cdots + \beta_{D-1}^{(l)} z_{D-1,i}^{(l)} + \varepsilon_i,\ i = 1, \ldots, n,\ l = 1, \ldots, D. \tag{3.7}$$

Due to orthogonality of different ilr bases, the intercept β_0 remains unchanged. Since the coordinate $z_1^{(l)}$ explains all the relative information about the original part x_l, the coefficient $\beta_1^{(l)}$ can be assigned to this part. The remaining coefficients are not straightforward to interpret since the assigned regressor variables do not fully represent one particular part [52]. Thus, the only way how to interpret the role of each compositional part for explaining the response $\mathbf{y}$ is to consider D different regression models and to interpret the coefficients $\beta_1^{(l)}$, representing the part x_l.

To show an economic application of the model, we analyze the relation between GDP per capita (available on http://data.worldbank.org/indicator/NY.GDP.PCAP.CD) and relative structure of gross value added (GVA) from the year 2010. The GVA consists of 4 parts, namely manufacturing, agriculture, industry and services (the dataset is described in a more detail in Section 3.6.5). As it was mentioned earlier, only the coordinates $z_1^{(l)}$ for $l = 1, 2, 3, 4$ contain the relative information about part x_l with respect to all the other parts, accordingly, only the corresponding parameters are important for interpretation of the model. Following this idea, four multivariate regression models were estimated using classical and robust methods. The results are summarized in Table 3.2. Again, the logarithm of GDP per capita was considered as in the previous section.

The logarithmic transformation of GDP leads again to better fit of the model - this is evident both from the adjusted coefficient of determination (Table 3.2) and from the plots of the response vs. fitted values in Figure 3.2. From the

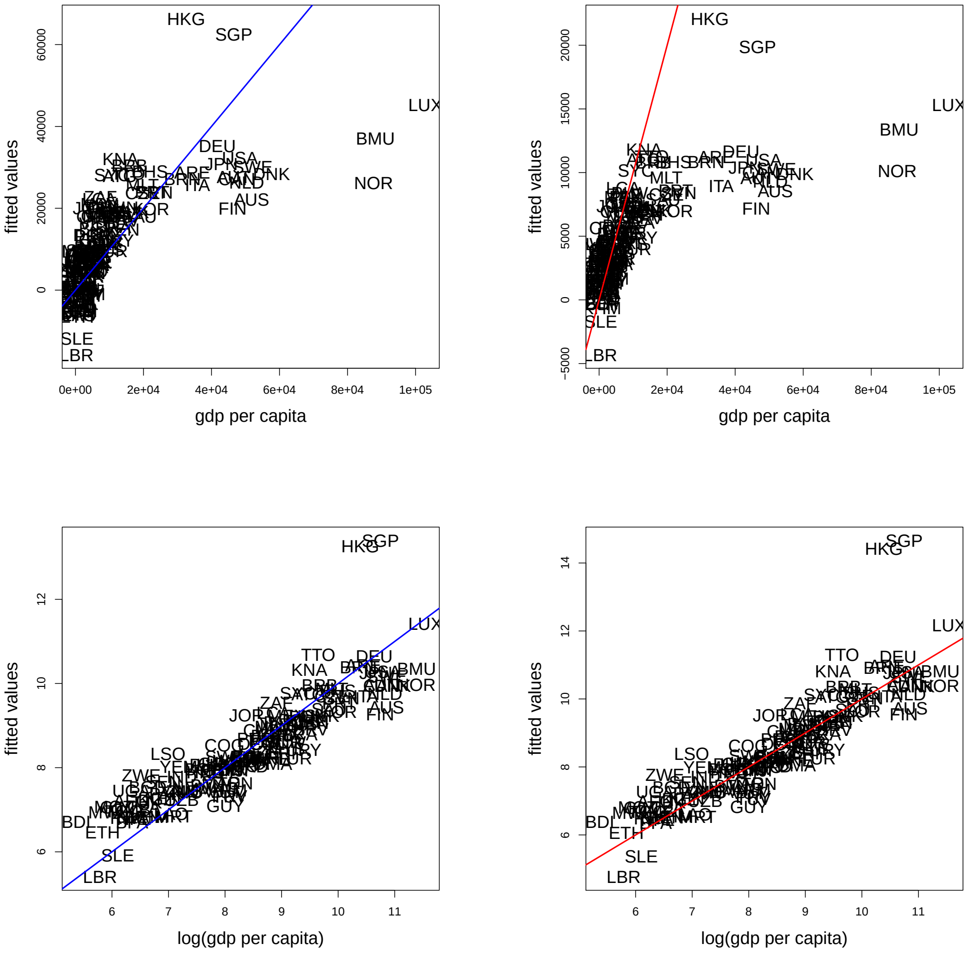

Figure 3.2: Plots of the response variable (GDP per capita) vs. its predicted values with the reference line $y = x$ using classical (blue line) and robust (red line) approaches for the original (upper panel) and log-transformed GDP (lower panel) values.

	GDP		log(GDP)	
parameter	estimate	p-value	estimate	p-value
classical	R^2_{adj}=0.5089		R^2_{adj}=0.7755	
b_0	-8803.2	0.0083	6.91786	<2e-16
$b_1^{(1)}$	-479.1	0.7889	0.13019	0.1707
$b_1^{(2)}$	-11016.3	<2e-16	-1.09364	<2e-16
$b_1^{(3)}$	-189.0	0.9050	0.19092	0.0241
$b_1^{(4)}$	11684.4	2.24e-07	0.77253	2.97e-10
robust	R^2_{adj}=0.7249		R^2_{adj}=0.8713	
b_0	-924.0	0.0825	6.67117	<2e-16
$b_1^{(1)}$	-675.1	0.0192	0.19435	0.01153
$b_1^{(2)}$	-3395.8	<2e-16	-1.35128	<2e-16
$b_1^{(3)}$	479.2	0.0481	0.23173	0.00073
$b_1^{(4)}$	3413.8	1.65e-15	0.92519	<2e-16

Table 3.2: Summary of regression outputs for compositional covariates.

classically estimated parameters, summarized in Table 3.2, we can conclude that relative contributions of manufacturing and industry within the GVA composition do not significantly influence the GDP. However, by considering the log of GDP as the preferable choice, industry becomes significant. Furthermore, it is easy to see that services have the largest positive relative influence on GDP, while higher dominance of agriculture within the GVA composition has negative relative influence on GDP. On the other side, robust regression yields different results, all of the parameters are significant on the usual level 0.05 (except of the intercept). Again, the largest positive relative influence on GDP show services and the largest negative relative influence has the agriculture sector. Different regression outputs using the classical and robust approaches indicate that the classical (least squares) estimation was strongly influenced by outlying observations.

3.5. Regression analysis with compositional response and explanatory variables

The third possibility is that both the response and explanatory variables have compositional character. This part is motivated by the problem of modeling the ecotoxicological experiments [30, 106], where the responses of subjects to different stimuli (e.g. drug, poison, etc.) are observed in a quantitative way. In the typical ecotoxicological experiment, a series of concentrations of a toxic agent $x_1, \ldots, x_r$ are chosen in increasing order of magnitude, and the proportions p_i of responding subjects are recorded. In these experiments, only the proportions p_i are known, while the number of subjects tested for each concentration and corresponding number of respondents are not given.

The general model has following form,

$$p_i = f(x_i, \boldsymbol{\beta}) + \varepsilon_i,\ i = 1, \ldots, n,$$

where p_i is the response measured at concentration x_i, the function $f(x_i, \boldsymbol{\beta})$ represents the mean of the response given the concentration x_i, $\boldsymbol{\beta}$ stands for the vector of unknown parameters and ε_i is the measurement error. Standard choices of the regression function f are logit, generalized logit, probit and weibull models [38, 106]. Instead of working with these difficult models we can consider the proportions p_i and concentrations x_i as two-part compositional data in the form $\mathbf{x} = (x, \kappa - x)'$ [84]. For proportions p_i and the respective two-part compositions $\mathbf{p}_i$ the constant $\kappa = 1$, while for concentrations x_i (measured in mg/l) $\kappa = 10^6$. Model that takes the compositional character of variables into account is defined as follows

$$\mathbf{p}_i = \boldsymbol{\beta}_0 \oplus \beta_1 \odot \mathbf{x}_i \oplus \boldsymbol{\varepsilon}_i,\ i = 1, \ldots, n. \tag{3.8}$$

To estimate the regression parameters we need to express the model in real space. For this purpose, the isometric logratio coordinates (1.5) are applied, the resulting variables are marked with an asterisk. Thus the regression line is obtained,

$$p_i^* = \beta_0^* + \beta_1 x_i^* + \varepsilon_i^*,\ i = 1, \ldots, n, \tag{3.9}$$

parameter	estimate	ilr^{-1}(estimate)	p-value
classical	R^2_{adj}=0.412		
b_0	-1.40322	12.08453	<2e-16
b_1	0.58434	69.55889	4.01e-07
robust	R^2_{adj}=0.6233		
b_0	-1.44082	11.53089	<2e-16
b_1	0.77419	74.92977	7.01e-11

Table 3.3: Summary of regression outputs.

where the parameters β_0^*, β_1 are estimated using ordinary least-squares method [84]. Fitted values for the original proportions p_i are obtained using inverse ilr mapping

$$\widehat{p}_i = h^{-1}(\widehat{p}_i^*) = \frac{\kappa \exp(\sqrt{2}\widehat{p}_i^*)}{1+\exp(\sqrt{2}\widehat{p}_i^*)}, \tag{3.10}$$

for $\kappa = 1$.

It is easy to see that here the ilr coordinates are proportional up to constant $\frac{1}{\sqrt{2}}$ to well-known logit transformation. Nevertheless, within the logratio methodology it represents a coordinate that enables to apply standard methods for statistical analysis of compositional data (including regression methods [84]). Note that the ilr coordinates also allows a straightforward generalization of the regression model to a multiple-compositional-response case.

The presented model was applied to real-world data that contain percentage shares of manufacturing value added (mva) in GDP [103] and percentage shares of import of intermediate goods [109]. Data were collected for 49 countries of the world from the year 2009. The aim of the analysis is to analyze whether the import of intermediate goods has an influence on the mva.

The results are summarized in Table 3.3 and it is easy to see that both regression parameters are significant and with increasing percentage of imported intermediate goods, the percentage of mva increases. This result reflects the trend of international trade - it is cheaper to import intermediate goods than to manufacture the whole product. Accordingly, countries can product more goods

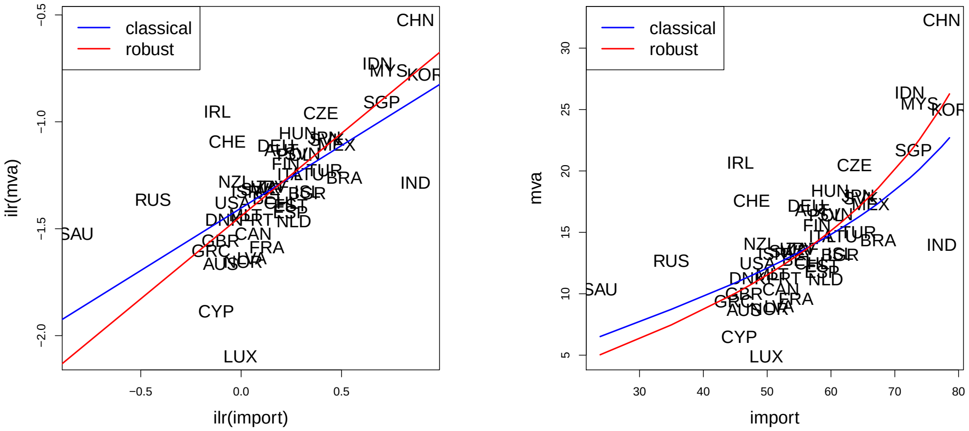

Figure 3.3: Plots of fitted regression in ilr space (on the left) and in original space (on the right).

and the mva grows. Furthermore, there is only a slight difference between classical and robust approach, thus there are no influential outliers in the dataset. In Figure 3.3 the estimated regression lines, both in real space and within the Aitchison geometry, are plotted. The regression results obtained in ilr coordinates are transformed back to the original space using (3.10).

3.6. Regression within a composition

Most of the economic indicators, such as gross domestic product, value added, export, import, etc., consist of many variables. For example GDP, in the income approach, is computed as a sum of compensation of employees, gross operating surplus, gross mixed income and taxes less subsidies on production and imports. Generally, we are interested in analyzing GDP but we can be also interested in analyzing the relation between the variables that form the GDP composition. For this purpose, orthogonal regression in proper coordinates seems to be the preferable option [57].

A particular challenge for the choice of coordinates comes from the fact that at least two parts in the composition are of simultaneous interest, the response part and covariate part(s). The first idea could be to use the centered logratio (clr) coordinates (1.3). Because of the relation $\ln\frac{x_l}{g(\mathbf{x})}=\sqrt{\frac{D-1}{D}}z_1^{(l)}$, each clr coordinate can be interpreted analogously as the corresponding ilr coordinates with respect to the original compositional parts. Nevertheless, their use for regression purposes is not appropriate due to the inherent constraint

$$\ln\frac{x_1}{g(\mathbf{x})}+\cdots+\ln\frac{x_D}{g(\mathbf{x})}=0$$

that leads to a singular covariance matrix of $\mathrm{clr}(\mathbf{x})$ and would thus result in misleading conclusions from the regression model.

Consequently, the question arises, how to use coordinates (1.4) for the case of regression of one of the compositional parts to the remaining parts. Similar problem was analyzed in the previous sections, where the compositional response or compositional covariates were considered in the regression model. There, in order to analyze the influence of a single compositional part on the explanatory variables, D multiple regression models according to the coordinate representations (1.4) were constructed. In our case, x_l plays the role of the response variable that should be represented by a coordinate as well. Since the main task is to analyze the influence of the other parts on x_l, it seems reasonable that also the corresponding coordinate will contain information on the relation of x_l to all remaining parts in the composition. Thus, in the above notation, $z_1^{(l)}$ plays the role of such a coordinate. Consequently, we can proceed with the coordinate representation of the explanatory subcomposition $(x_1,\ldots,x_{l-1},x_{l+1},\ldots,x_D)'$. For this purpose, the coordinates $z_2^{(k)},\ldots,z_{D-1}^{(k)}$ according to the reordered subcomposition $(x_k,x_2,\ldots,x_i,\ldots,x_D)'$, $i\neq\{k,l\}$, $k=2,\ldots,D$, can be used. Similarly as before, the coordinate $z_2^{(k)}$ explains all the relative information about part x_k in the resulting subcomposition. Considering the range of k, we arrive at $D-1$ regression models

$$z_1^{(l)}=\beta_0^{(k)}+\beta_1^{(k)}z_2^{(k)}+\ldots+\beta_{D-2}^{(k)}z_{D-1}^{(k)}+\varepsilon \tag{3.11}$$

(in theoretical form, ε stands for an error term), assigned to single explanatory compositional parts. The interpretation of these models is similar to the case of regression with compositional covariates [52], i.e. in each model just the absolute term parameter and the parameter corresponding to the coordinate $z_2^{(k)}$ are used for further interpretation and to perform statistical inference (confidence intervals, hypotheses testing).

Since both response and explanatory variables arise from one composition, it cannot be assumed that the covariates represent errorless variables like in the case of a real-valued response [52]. Consequently, the use of an ordinary multiple regression model is inappropriate and can even lead to biased results. Therefore, we apply an orthogonal regression model (or, equivalently, a total least squares model) for this purpose, which is a specific type of errors-in-variable (EIV) model [44].

3.6.1. Orthogonal regression

For simplification of the notation, we denote the matrix of n realizations of the vector $(z_2^{(k)}, \dots, z_{D-1}^{(k)})'$, for a chosen $k \in \{1, \dots, D\}$, $k \neq l$, as $\mathbf{X} \in \mathbb{R}^{n \times D-2}$ (centered data are assumed), and $\mathbf{y} \in \mathbb{R}^n$ the observation vector of the response coordinate $z_1^{(l)}$. The total least-squares (TLS) method was originally introduced to solve overdetermined systems of equations $\mathbf{X}\boldsymbol{\beta} \approx \mathbf{y}$, where $\mathbf{X}$ and $\mathbf{y}$ are given data (here compositions expressed in orthonormal coordinates), and $\boldsymbol{\beta} \in \mathbb{R}^{D-2}$ is the vector of unknown parameters. There is no exact solution; particularly in the case of $n > D-2$, we are seeking for an approximation.

In the classical TLS problem [76] we are looking for the minimal errors $\boldsymbol{\varepsilon}_X, \boldsymbol{\varepsilon}_y$ (in the sense of the Frobenius norm, denoted by the subscript F in the following) on the given data $\mathbf{X}, \mathbf{y}$ that make the system of equations $\widehat{\mathbf{X}}\mathbf{b} = \widehat{\mathbf{y}}$, $\widehat{\mathbf{X}} = \mathbf{X} + \boldsymbol{\varepsilon}_X$, $\widehat{\mathbf{y}} = \mathbf{y} + \boldsymbol{\varepsilon}_y$ solvable, i.e.

$$\{\widehat{\mathbf{X}}, \widehat{\mathbf{y}},\ \boldsymbol{\varepsilon}_X,\ \boldsymbol{\varepsilon}_y\} := \operatorname{argmin}_{\boldsymbol{\varepsilon}_X, \boldsymbol{\varepsilon}_y} \|[\boldsymbol{\varepsilon}_X,\ \boldsymbol{\varepsilon}_y]\|_F\,, \tag{3.12}$$

subject to $(\mathbf{X} + \boldsymbol{\varepsilon}_X)\boldsymbol{\beta} = \mathbf{y} + \boldsymbol{\varepsilon}_y$, resulting in the estimate $\mathbf{b}$ of the regression parameters. The Frobenius norm of $n \times p$ matrix $\mathbf{A}$ with elements a_{ij} is defined as $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{p} a_{ij}^2}$. The solution is a maximum likelihood estimator $\mathbf{b}$

in the optimally corrected EIV model $\widehat{\mathbf{X}}\mathbf{b} = \widehat{\mathbf{y}}$, $\widehat{\mathbf{X}} = \mathbf{X} + \boldsymbol{\varepsilon}_X$, $\widehat{\mathbf{y}} = \mathbf{y} + \boldsymbol{\varepsilon}_y$, if the usual assumptions are fulfilled, namely that $\text{vec}[\boldsymbol{\varepsilon}_X, \boldsymbol{\varepsilon}_y]$ ("vec" forms one vector, composed of the columns of the matrix in the argument) has zero mean, and is a normally distributed random vector with a covariance matrix that is a multiple of the identity.

From the methodological point of view, the singular value decomposition is applied to $\mathbf{Z} = [\mathbf{X}, \mathbf{y}] = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}'$, where $\boldsymbol{\Lambda} = \text{Diag}(\lambda_1, \ldots, \lambda_{D-1})$ and $\lambda_1 \geq \cdots \geq \lambda_{D-1} \geq 0$ are the singular values of $\mathbf{Z}$, and $\mathbf{U}$ and $\mathbf{V}$ are the corresponding orthonormal matrices. Let us define the partitions

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{v}_{12} \\ \mathbf{v}_{21} & v_{22} \end{bmatrix}, \quad \boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_1 & \mathbf{0} \\ \mathbf{0} & \lambda_D \end{bmatrix},$$

where the matrices $\mathbf{V}_{11}$ and $\boldsymbol{\Lambda}_1$ are of dimension $(D-2) \times (D-2)$. Then a TLS solution exists iff v_{22} is non-zero; moreover, it is unique iff $\lambda_{D-2} \neq \lambda_{D-1}$. In this case it is given by

$$\mathbf{b} = -\mathbf{v}_{12}/v_{22} \tag{3.13}$$

and the corresponding TLS error matrix equals $\boldsymbol{\varepsilon}_Z = [\boldsymbol{\varepsilon}_X, \boldsymbol{\varepsilon}_y] = -\mathbf{U}\text{Diag}(\mathbf{0}, \lambda_{D-1})\mathbf{V}'$ [76], with $\mathbf{0}$ being a vector with $D-2$ zeros. Thus, when a unique solution $\mathbf{b}$ exists, it is computed from the scaled right singular vector corresponding to the smallest singular value. It is important to note that as different ilr coordinate systems are just rotation of each other [28], TLS estimates transform accordingly.

It is well-known that the matrices $\boldsymbol{\Lambda}$ and $\mathbf{V}$ from an SVD on the centered explanatory and response variables correspond to outputs of an eigenvalue decomposition on the (estimated) covariance matrix $\boldsymbol{\Sigma}$, performed within principal component analysis (PCA). Thus, except for the intercept term in the orthogonal regression model (that is discussed in the next section), the same results as above in (3.13) can be obtained also using the smallest eigenvalue and the corresponding eigenvector (loading vector) of the covariance matrix. We will follow this approach further in the next section.

Geometrical motivation

As mentioned in the previous section, the TLS (orthogonal regression) estimates of the parameters can be obtained by means of principal component analysis. We apply the proposed procedure directly to the case of four-part compositional data where a geometrical illustration of the problem is still possible. Moreover, this approach also shows how to proceed with non-centered data. For this purpose, we assume to have a random vector $\mathbf{z} = (z_1, z_2, z_3)'$ (an orthonormal coordinate representation of the composition) and the task is to find a relationship between the response variable z_1 and the covariates z_2, z_3, expressed in the form $z_1 = \beta_0 + \beta_1 z_2 + \beta_2 z_3 + \varepsilon$, with the regression parameters $\beta_0, \beta_1, \beta_2$.

From the geometrical point of view, the basic idea is to fit a plane to the data using PCA. The loadings of the first two principal components define a basis of the plane. As the third principal component is orthogonal to the previous ones, its loadings define the normal vector to the plane, $\mathbf{n} = (n_1, n_2, n_3)'$, forming the last column of the matrix $\mathbf{V}$ in terms of the previous section. The plane passes through the point $\mathbf{t}$, representing the location estimate of the $n \times 3$ data matrix $\mathbf{Z}$ (the arithmetic mean in the classical case), and its perpendicular distance from the origin is $\mathbf{t}'\mathbf{n}$. The perpendicular distance from each point in $\mathbf{Z}$ to the plane (the norm of the residuals) is the inner product of each centered point and the normal vector to the plane. The fitted plane minimizes the sum of squared errors.

Consequently, the estimated regression parameters are obtained using the elements of the normal vector, namely

$$b_0 = \frac{\mathbf{t}'\mathbf{n}}{n_3},\ b_1 = -\frac{n_1}{n_3},\ b_2 = -\frac{n_2}{n_3}.$$

We can also consider the general case, where a vector of orthonormal coordinates has $D-1$ components, $\mathbf{z} = (z_1, z_2, \ldots, z_{D-1})'$. As in the previous case, the response variable is z_1 and covariates $z_2, \ldots, z_{D-1}$. Then the regression relation is expressed in the form $z_1 = \beta_0 + \beta_1 z_2 + \beta_2 z_3 + \cdots + \beta_{D-2} z_{D-1} + \varepsilon$ for a vector of regression parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_{D-2})'$. Denote the loading vector corresponding to the smallest eigenvalue as $\mathbf{n} = (n_1, n_2, \ldots, n_{D-1})'$. Then the estimated parameters $\mathbf{b}$ are obtained using values of the loading vector as follows,

$$b_0 = \frac{\mathbf{t}'\mathbf{n}}{n_{D-1}},\ b_1 = -\frac{n_1}{n_{D-1}},\ b_2 = -\frac{n_2}{n_{D-1}}, \ldots,\ b_{D-2} = -\frac{n_{D-2}}{n_{D-1}},$$

where $\mathbf{t}$ is the mean vector of $\mathbf{Z}$.

3.6.2. Nonparametric bootstrap sampling

In order to support the interpretation of the outcome of orthogonal regression, it is desirable to obtain confidence intervals for the regression parameters, and p-values for tests about these parameters. This statistical inference is only possible with strict distributional assumptions but even then it would be challenging to derive the exact distribution of the parameters in the robust case. A better strategy is to derive the inference by resampling methods. In order to relax the assumptions about the distribution of the input data, the nonparametric bootstrap [41] was chosen for this purpose.

Generally, bootstrapping is based on building a sampling distribution for a statistic by resampling from the data at hand. Consequently, the nonparametric bootstrap allows us to estimate the sampling distribution of a statistic empirically without making assumptions about the form of the population, and without deriving the sampling distribution explicitly. The basic idea is that, after drawing a sample of size n from $\mathsf{S} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ with replacement (without loss of generality, we fix the concrete choice of orthonormal coordinates again), we are treating the sample S as an estimate of the whole population which means that each element $\mathbf{z}_i$ of S is selected with probability $1/n$ to mimic the original selection of the sample S. This procedure is repeated R times, where R is a large number, to obtain a sufficient number of bootstrap samples.

The r-th bootstrap sample is denoted as $\mathsf{S}^r = \{\mathbf{z}_{r_1}, \ldots, \mathbf{z}_{r_n}\}$, $r = 1, \ldots, R$. In the next step we compute the regression estimates b_i for each bootstrap sample to get b_i^{r*}, $i = 1, \ldots, D-1$. Then the distribution of b_i^{r*} around the original estimate b_i is analogous to the sampling distribution of the estimator b_i around the population parameter β_i. In context of orthogonal regression, the bootstrap distribution of b_i can be directly used to derive sample p-values for significance testing of the regression parameters. For this purpose, the p-value p_i for the regression parameter β_i for a two-sided alternative is derived by comparing the values of the bootstrap parameter estimates with zero. By denoting l_i and h_i as the number of estimated values lower and higher than zero, respectively, we get $p_i = 2 \cdot \min\{l_i, h_i\}/R$ [15].

Furthermore, we can proceed also to construct bootstrap confidence intervals. For this purpose, several approaches are available. A natural choice is to take bootstrap percentile intervals that are free of any distributional assumptions [15, 41]. The bootstrap percentile interval uses the empirical quantiles of b_i^{r*} (computed from S^r, $r = 1, \ldots, R$) to form a confidence interval for β_i, $(b^*_{i(l)}, b^*_{i(u)})$, $i = 1, \ldots, D-1$. Concretely, from the ordered bootstrap replicates of the statistic b_i, i.e. $b^*_{i(1)}, b^*_{i(2)}, \ldots, b^*_{i(R)}$, and for a given $\alpha \in (0, 1)$ we set $l = \lceil (R+1)\alpha/2 \rceil, u = \lceil (R+1)(1-\alpha/2) \rceil$ (rounded to the nearest integer).

3.6.3. Robust orthogonal regression

Regression estimators which are based on classical SVD or PCA are sensitive to outliers that naturally occur in most real-world data sets. Therefore, we also considered a robust version of the orthogonal regression. Although robust versions of SVD are available (e.g. [14]), it is simpler and computationally more attractive to use robust PCA, which is obtained through a robust estimation of the covariance matrix (e.g. [32]). Among other possibilities like those in [13, 31, 77], MM-estimators [94] are employed for this purpose in the following.

MM-estimators

MM-estimators were chosen because they are highly efficient when the errors have a normal distribution, their breakdown point is 0.5 and they have bounded influence function.

Multivariate MM-estimators are extensions of S-estimators [78]. They are based on two loss functions ρ_0 and ρ_1 that satisfy the conditions: (a) ρ is symmetric and twice continuously differentiable, with $\rho(0) = 0$; (b) ρ is strictly increasing on an interval $[0, k]$ and constant on $[k, \infty]$ for some finite constant k. Given the matrix with the observations in any chosen orthonormal coordinates (like those from the beginning of this section), $\mathbf{Z} = [\mathbf{X}, \mathbf{y}] = (\mathbf{z}_1, \ldots, \mathbf{z}_n)' \in \mathbb{R}^{D-1}$, the MM-estimators for location and covariance are defined in two steps:

1. Let $(\widetilde{\boldsymbol{\mu}}_n, \widetilde{\boldsymbol{\Sigma}}_n)$ be S-estimators of location and covariance, respectively, that is $(\widetilde{\boldsymbol{\mu}}_n, \widetilde{\boldsymbol{\Sigma}}_n)$ minimize $|\mathbf{C}|$ subject to

$$\frac{1}{n}\sum_{i=1}^{n} \rho_0\left([(\mathbf{z}_i - \mathbf{t})'\mathbf{C}^{-1}(\mathbf{z}_i - \mathbf{t})]^{1/2}\right) = b,$$

   among all $(\mathbf{t}, \mathbf{C}) \in \mathbb{R}^{D-1}$. Denote $\widehat{s} = |\widetilde{\boldsymbol{\Sigma}}_n|^{1/[2(D-1)]}$.

2. The MM-estimator for location and shape $(\widehat{\boldsymbol{\mu}}_n,\ \widehat{\boldsymbol{\Gamma}}_n)$ minimizes

$$\frac{1}{n}\sum_{i=1}^{n} \rho_1\left(\left[(\mathbf{z}_i - \mathbf{t})'\mathbf{S}^{-1}(\mathbf{z}_i - \mathbf{t})\right]^{1/2} / \widehat{s}\right)$$

   among all $\mathbf{t}$ and all symmetric positive definite $\mathbf{S}$ with $|\mathbf{S}| = 1$. The MM-estimator of the covariance matrix is then $\widehat{\boldsymbol{\Sigma}}_n = \widehat{s}^2\widehat{\boldsymbol{\Gamma}}_n$.

The idea is to estimate the scale by means of a very robust S-estimator and then to estimate the location and shape using different ρ functions that yields a better efficiency. Once the location and covariance are obtained using the MM-estimator, they can be used to compute the robust orthogonal regression estimates as described above.

3.6.4. Fast and robust bootstrap

The available theory for robust estimators is limited to asymptotic results. Although bootstrap is a very useful tool, in case of robust estimators there are two problems: computational complexity of robust estimators and the instability of the bootstrap in case of outliers. Thus we used fast and robust bootstrap

[99, 104] which is based on the fact that the robust estimators (namely S- and MM-estimators) can be represented by smooth fixed point equations that allow to calculate only a fast approximation of the estimates in each bootstrap sample. For the case of MM-estimators, the fixed point equations are as follows,

$$\widehat{\boldsymbol{\mu}}_n = \left(\sum_{i=1}^{n} \frac{\rho_1'(d_i/|\widetilde{\boldsymbol{\Sigma}}_n|^{1/[2(D-1)]})}{d_i} \right)^{-1} \left(\sum_{i=1}^{n} \frac{\rho_1'(d_i/|\widetilde{\boldsymbol{\Sigma}}_n|^{1/[2(D-1)]})}{d_i} \mathbf{z}_i \right); \qquad (3.14)$$

$$\widehat{\boldsymbol{\Gamma}}_n = G\left(\sum_{i=1}^{n} \frac{\rho_1'(d_i/|\widetilde{\boldsymbol{\Sigma}}_n|^{1/[2(D-1)]})}{d_i} (\mathbf{z}_i - \widehat{\boldsymbol{\mu}}_n)(\mathbf{z}_i - \widehat{\boldsymbol{\mu}}_n)' \right); \qquad (3.15)$$

$$\widetilde{\boldsymbol{\Sigma}}_n = \frac{1}{nb} \left(\sum_{i=1}^{n} (D-1) \frac{\rho_0'(\widetilde{d}_i)}{\widetilde{d}_i} (\mathbf{z}_i - \widetilde{\boldsymbol{\mu}}_n)(\mathbf{z}_i - \widetilde{\boldsymbol{\mu}}_n)' + \left(\sum_{i=1}^{n} \widetilde{w}_i \right) \widetilde{\boldsymbol{\Sigma}}_n \right); \qquad (3.16)$$

$$\widetilde{\boldsymbol{\mu}}_n = \left(\sum_{i=1}^{n} \frac{\rho_0'(\widetilde{d}_i)}{\widetilde{d}_i} \right)^{-1} \left(\sum_{i=1}^{n} \frac{\rho_0'(\widetilde{d}_i)}{\widetilde{d}_i} \mathbf{z}_i \right); \qquad (3.17)$$

where we denote $G(\mathbf{A}) = |\mathbf{A}|^{-1/(D-1)}\mathbf{A}$ for a $(D-1) \times (D-1)$ matrix $\mathbf{A}$, $d_i = [(\mathbf{z}_i - \widehat{\boldsymbol{\mu}}_n)'\widehat{\boldsymbol{\Gamma}}_n^{-1}(\mathbf{z}_i - \widehat{\boldsymbol{\mu}}_n)]^{1/2}$, $\widetilde{d}_i = [(\mathbf{z}_i - \widetilde{\boldsymbol{\mu}}_n)'\widetilde{\boldsymbol{\Sigma}}_n^{-1}(\mathbf{z}_i - \widetilde{\boldsymbol{\mu}}_n)]^{1/2}$ and $\widetilde{w}_i = \rho_0(\widetilde{d}_i) - \rho_0'(\widetilde{d}_i)\widetilde{d}_i$. Generally, denote the equations (3.14) - (3.17) by means of a function $\mathbf{f} : \mathbb{R}^{2[(D-1)+(D-1)^2]} \to \mathbb{R}^{2[(D-1)+(D-1)^2]}$ such that $\mathbf{f}(\widehat{\Theta}_n) = \widehat{\Theta}_n$, where $\widehat{\Theta}_n$ contains all estimates in the vectorized form and can be represented as a solution of fixed-point equations. For example, for MM-estimators we have $\widehat{\Theta}_n := \Big((\widehat{\boldsymbol{\mu}}_n)', \mathrm{vec}(\widehat{\boldsymbol{\Gamma}}_n)', \mathrm{vec}(\widetilde{\boldsymbol{\Sigma}}_n)', (\widetilde{\boldsymbol{\mu}}_n)'\Big)'$. Instead of recalculating the estimates $\widehat{\Theta}_n^*$ for each bootstrap sample we can calculate its one-step approximation starting from the initial value $\widehat{\Theta}_n$,

$$\widehat{\Theta}_n^{1*} = \mathbf{f}(\widehat{\Theta}_n). \qquad (3.18)$$

Unfortunately, this approximation underestimates the variability of $\widehat{\Theta}_n$ because the initial value in the approximation remains the same. To remedy this we can apply a linear correction [98] as follows. Given the smoothness of $\mathbf{f}$ we can calculate a Taylor expansion about the limiting value of $\widehat{\Theta}_n$

$$\widehat{\Theta}_n = \mathbf{f}(\Theta) + \nabla\mathbf{f}(\Theta)(\widehat{\Theta}_n - \Theta) + R_n, \tag{3.19}$$

where $\Theta = (\boldsymbol{\mu}', \mathrm{vec}(\boldsymbol{\Gamma})', \mathrm{vec}(\boldsymbol{\Sigma})', \boldsymbol{\mu}')'$, R_n is the remainder term and $\nabla\mathbf{f}(\cdot)$ is the matrix of partial derivatives. If the remainder term is sufficiently small, we can rewrite (3.19) as

$$\sqrt{n}(\widehat{\Theta} - \Theta) \approx [I - \nabla\mathbf{f}(\Theta)]^{-1}\sqrt{n}(\mathbf{f}(\Theta) - \Theta). \tag{3.20}$$

Since both sides of this equation are asymptotically equivalent, the distribution of the bootstrapped statistics will also converge to the same limit. Finally, we can define the linearly corrected version of the one-step approximation (3.18) as

$$\widehat{\Theta}_n^{R*} := \widehat{\Theta}_n + [I - \nabla\mathbf{f}(\widehat{\Theta}_n)]^{-1}(\widehat{\Theta}_n^{1*} - \widehat{\Theta}_n). \tag{3.21}$$

Note that the estimating equations involve weighted least squares estimates and covariances. The weights will be small or even zero for observations detected as outliers. This guarantees that $\widehat{\Theta}_n^{R*}$ is as robust as $\widehat{\Theta}_n$. The idea is to draw bootstrap samples as usual, but instead of computing the actual estimator in each bootstrap sample, a fast approximation is computed based on the estimating equations of the estimator.

The above theoretical developments were implemented into new R package `oreg` [57, Section 6]. It provides functions for classical and robust versions of orthogonal regression that are possible to use for both standard multivariate and compositional data - the package also includes computation of ilr coordinates. This package was used for all the computations and graphical outputs in the following section.

3.6.5. Activities of gross value added

The example focuses on the relation between different activities of gross value added. The data set comes from the World Bank database (http://data.worldbank.org) and includes observations for 131 countries in 2010 at constant 2005 USD.

Gross value added (GVA) is the most important measure of productivity of the economy of a country or region, representing the difference between production output and intermediate consumption, i.e. the monetary value of the amount of goods and services that have been produced, less the cost of all inputs and raw materials that are directly attributable to that production. Gross value added is less than GDP because it excludes value-added tax (VAT) and other product taxes.

GVA can be decomposed into the following economic activities:

- agriculture (consisting of agriculture, forestry, hunting and fishing);
- manufacturing[1];
- other industry (consisting of mining and quarrying; electricity, gas, steam and air conditioning supply; water supply; sewerage, waste management and remediation activities; construction);
- services (consisting of education, health and other personal services; public administration and defense).

Thus, GVA can be expressed as the sum of these four activities. The goal of the study is to analyze the relation between manufacturing and the rest of the activities by considering relative contributions of the mentioned activities to the overall GVA.

Although the original data are expressed in monetary units (USD), and no constant sum constraint is present (like it is the case of proportions or percentages), from the relative structure of GVA we can conclude that these four economic activities form a composition $\mathbf{x} = (x_1, x_2, x_3, x_4)'$, where x_1 corresponds to manufacturing, x_2 to agriculture, x_3 to other industry and x_4 to services. In such case, using an arbitrary regression technique either for the original observations or any constrained form of them, would lead to biased results [52]. Figure 3.4

[1] Manufacturing is defined as the physical or chemical transformation of materials of components into new products.

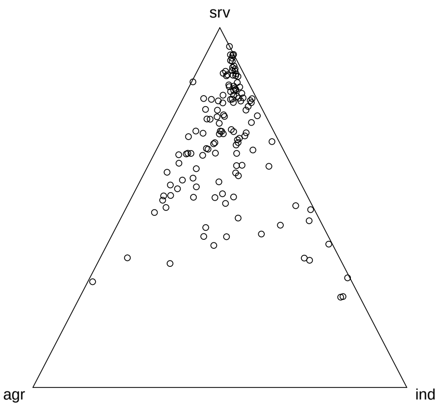

Figure 3.4: Ternary diagram of the explanatory variables `agr` (agriculture), `ind` (other industry), and `srv` (services).

displays a ternary diagram of the explanatory variables x_2, x_3, x_4. It can be observed that the part `srv` (services) contains the largest relative contribution and `agr` (agriculture) the smallest one in this subcomposition. This corresponds to the fact that the points are concentrated mainly along the segment between `srv` and `ind` (other industry), rather closer to the vertex `srv`.

For further statistical processing, the compositional response and the explanatory variables are expressed in *ilr* coordinates (1.4). Following the previous considerations, the response coordinate is defined as $z_1^{(1)} = \sqrt{\frac{3}{4}} \ln \frac{x_1}{\sqrt[3]{x_2 x_3 x_4}}$, i.e., it explains all the relative information about manufacturing with respect to the other three parts in the composition through an aggregation of the corresponding logratios. Permutation of the remaining three activities results in three regression models, where always the respective coordinate $z_2^{(k)}$ for $k = 2, 3, 4$ includes the most interesting information - (scaled) aggregation of logratios of x_i with the remaining explanatory parts. The resulting regression models that favor one of the explanatory compositional parts x_2, x_3, x_4 thus contain the following coordinates (in addition to $z_1^{(1)}$),

$$\begin{aligned} z_2^{(2)} &= \sqrt{\tfrac{2}{3}}\ln\tfrac{x_2}{\sqrt{x_3x_4}}, \quad & z_3^{(2)} &= \sqrt{\tfrac{1}{2}}\ln\tfrac{x_3}{x_4}; \\ z_2^{(3)} &= \sqrt{\tfrac{2}{3}}\ln\tfrac{x_3}{\sqrt{x_2x_4}}, \quad & z_3^{(3)} &= \sqrt{\tfrac{1}{2}}\ln\tfrac{x_2}{x_4}; \\ z_2^{(4)} &= \sqrt{\tfrac{2}{3}}\ln\tfrac{x_4}{\sqrt{x_2x_3}}, \quad & z_3^{(4)} &= \sqrt{\tfrac{1}{2}}\ln\tfrac{x_2}{x_3}, \end{aligned}$$

respectively.

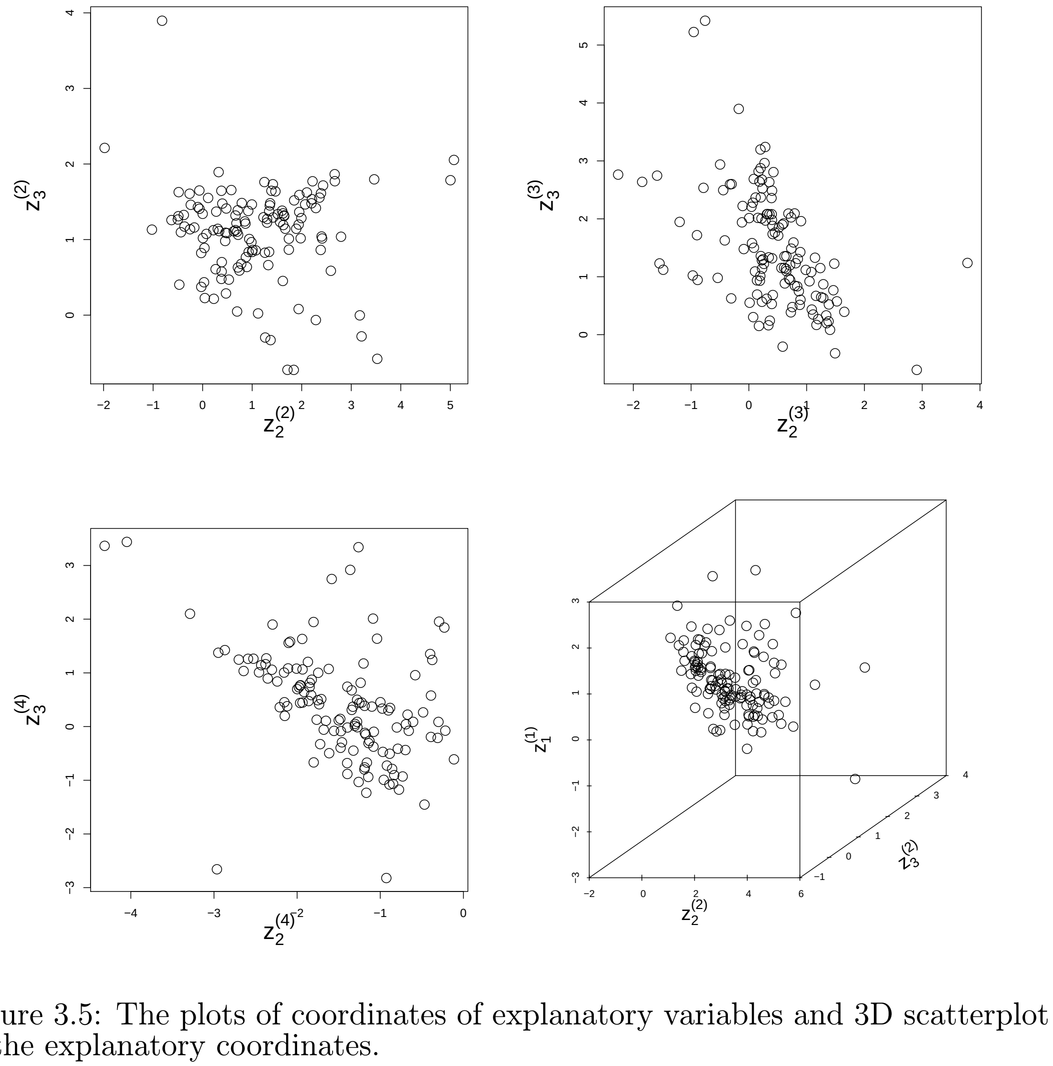

Figure 3.5: The plots of coordinates of explanatory variables and 3D scatterplot of the explanatory coordinates.

In Figure 3.5, scatterplots of the explanatory coordinates are displayed, where the part of interest corresponds to x_2 (upper left), x_3 (upper right) and x_4 (lower left). Particularly, it can be seen that the x-coordinates of the upper left and upper right plots, $z_2^{(2)}$ and $z_2^{(3)}$, are mainly negative which means that the relative

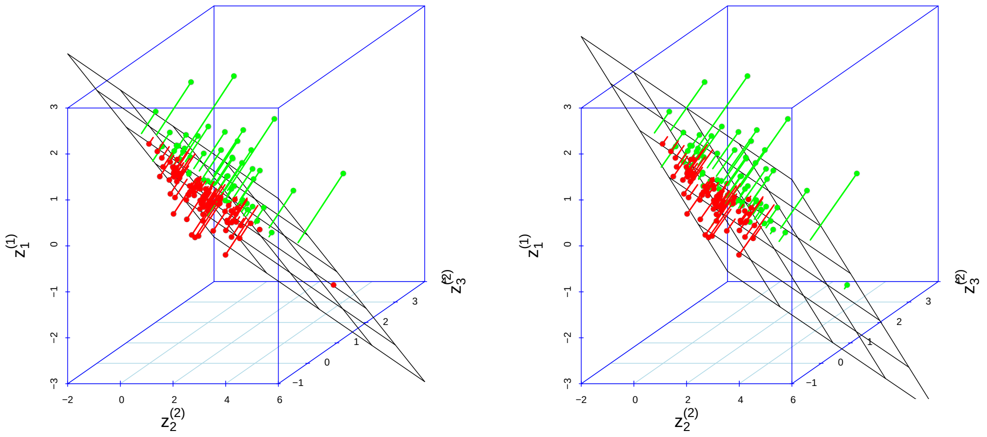

Figure 3.6: The 3D plots of estimated regression plane for coordinates $z_2^{(2)}, z_3^{(2)}, z_1^{(1)}$ following classical (on the left) and robust (on the right) approach.

contributions of agriculture and other industry are lower than the mean contribution of the other parts. On the other hand, the coordinate $z_2^{(4)}$ clearly shows the relative dominance of services. The 3D scatterplot in Figure 3.5 (lower right) contains all three coordinates $z_2^{(2)}, z_3^{(2)}, z_1^{(1)}$ (in this order) to see the relation between the covariates and the response variable. Although a certain linear relationship can be observed from this scatterplot, orthogonal regression modeling needs to be performed in order to specify the possible influence of covariates.

	par. estimate	perc. CI	p-value
intercept	-2.151	(-4.464, -1.559)	0.002
$b_1^{(2)}$	-0.394	(-0.584, -0.115)	0.020
$b_1^{(3)}$	-0.878	(-2.745, -0.498)	0.000
$b_1^{(4)}$	1.272	(0.858, 2.978)	0.002

Table 3.4: Summary of regression outputs using classical orthogonal regression for all defined models.

The results of classical orthogonal regression in coordinates are summarized in Table 3.4. Note that the intercept for all regression models is identical (similarly as for LS regression [52]), which is a consequence of the orthogonal relation between the different *ilr* coordinate systems. Therefore, although the basic model consists of three regression parameters (corresponding to intercept and two orthonormal coordinates), for the interpretation purposes it is enough to summarize just the intercept and the parameters corresponding to the coordinates $z_2^{(2)}, z_2^{(3)}, z_2^{(4)}$ from all three models. Nonparametric bootstrap (with $R = 1000$) was used to derive the corresponding statistical inference (confidence intervals, p-values for significance testing). Note that it would be possible to compute the regression estimates for all remaining models just from the estimates in one particular model using orthogonal transformations, similar as in the standard case of LS (PLS) regression [63].

According to Table 3.4, all regression parameters are significant on the usual level $\alpha = 0.05$, although $b_1^{(2)}$ is closer to zero. Moreover, the estimated parameters $b_1^{(2)}$ and $b_1^{(3)}$ are negative which means that "agriculture" and "other industry" have small negative relative influence on manufacturing. On the other hand, "services" (resulting from the estimation in the last model) have strong positive relative influence on "manufacturing". This is explained by the fact that the growth of the manufacturing sector is inevitably induced by the growth of the services sector, necessary to support it. Transportation, communication, financial and business services are required by the manufacturing and thus there is no increase in manufacturing without (relative) growth of these services. To illustrate the regression results geometrically (see Section 3.6.1 to recall the geometric motivation), Figure 3.6 displays the 3D plot of the estimated regression plane for coordinates $z_1^{(1)}$ (response), and $z_2^{(2)}, z_3^{(2)}$ (covariates) with all the points projected on the plane.

To restrict possible influence of outlying observations on the estimates, a robust version of orthogonal regression using MM-estimators was applied as well. The summary of the regression outputs (including confidence intervals and p-values computed by fast and robust bootstrap) are displayed in Table 3.5. The results are similar to those from Table 3.4. In contrast to the classical analysis, here the regression parameter $b_1^{(2)}$ is not significant. Consequently, the difference

	par. estimate	perc. CI	p-value
intercept	-2.311	(-6.391, -1.666)	0.006
$b_2^{(2)}$	-0.389	(-0.605, 0.180)	0.116
$b_2^{(3)}$	-1.075	(-4.994, -0.556)	0.002
$b_2^{(4)}$	1.464	(0.996, 5.184)	0.002

Table 3.5: Summary of regression outputs using robust orthogonal regression for all defined models.

for the inference on $b_1^{(2)}$ can be attributed to the outliers, underlining the need for a robust analysis. Of course, there are differences among countries, and the relation between agriculture and industry is a long debated topic, see for example [96] and [97] for a detailed analysis of these linkages in India using the input-output framework.

Chapter 4

Functional principal component analysis for density functions

Nowadays, an increasing number of studies are based on complex data, such as curves, surfaces or images. As a direct consequence, the importance of functional data analysis (FDA), e.g. [90], has recently strongly increased. In recent years, a large body of literature has been developed in this field, e.g. [51, 91], however, still little attention has been paid to the problem of dealing with functional data that are probability density functions [17, 18, 81, 85, 108]. Even though it might seem that density functions are just a special case of functional data – with a constant-integral constraint equal to one – standard FDA methods appear to be inappropriate for their statistical treatment, as they do not consider the particular constrained nature of the data. In this context, probability density functions have recently been interpreted as functional compositional data, i.e., functional data carrying only relative information. To handle this kind of data, the Aitchison geometry has been extended to the so called Bayes spaces: a Hilbert space structure for σ-finite measures, including probability measures, has been worked out in [9], based on the pioneering work of [24] and the subsequent developments of [8] and [29].

This chapter is organized as follows. Firstly we introduce functional data analysis and B-spline smoothing. Next, we highlight the main differences between standard functional data and density functions, functional compositions. Then we describe the case of functional principal component analysis (FPCA) and simplicial FPCA. Finally, the methodological outputs are applied to a data set with salary distributions in Austria.

4.1. Functional data analysis

Functional data [91], as the title prompts, consist of functions. The basic philosophy is to think of observed data functions as single entities. Functional data are usually observed and recorded discretely as n pairs (t_j, y_j), and y_j is a snapshot of the function x at t_j. By t_j we can consider time, frequency, weight, interval and so on [90]. Functional data can be obtained from original observations that are interpolated, we can also take large number of independent observations whose estimated probability densities are the functional data, or images and curves appear as functional data.

Aspects of functional data:

- functional data are continuously defined;
- the individual datum is a whole function;
- smoothness or other regularity is a key aspect of other analysis.

As mentioned, we usually require that function x is smooth, which means that a pair of adjacent data values y_j and y_{j+1} are necessarily linked together to some extent and unlikely to be too different form each other. However, the observed data may not be at all smooth due to the presence of measurement error. Thus in the following sections we will show one way of how to obtain smooth functions.

The aims of the analysis of functional data [91] are the same as those for conventional data: to represent the data in ways that aid further analysis; to develop ways of presenting the data that highlight interesting and important features; to investigate variability as well as mean characteristics; to build models for the data observed, including those that allow for dependence of one observation or variable on another, and so on.

4.1.1. Summary statistics

For functional data, the summary statistics [91] are similar to the discrete data, however, it is important to note that now we obtain functions not just constant values. The mean function is the average of functions $x_i(t),\ i = 1, \dots, N$ from the sample point-wise across replications,

$$\overline{x}(t) = N^{-1} \sum_{i=1}^{N} x_i(t), \tag{4.1}$$

and the variance function follows

$$\mathrm{var}_X(t) = (N-1)^{-1} \sum_{i=1}^{N} [x_i(t) - \overline{x}(t)]^2. \tag{4.2}$$

The covariance function summarizes the dependence of records across different argument values,

$$\mathrm{cov}_X(t_1, t_2) = (N-1)^{-1} \sum_{i=1}^{N} [x_i(t_1) - \overline{x}(t_1)][x_i(t_2) - \overline{x}(t_2)], \tag{4.3}$$

for all t_1, t_2 and the correlation function is then computed as

$$\mathrm{corr}_X(t_1, t_2) = \frac{\mathrm{cov}_X(t_1, t_2)}{\sqrt{\mathrm{var}_X(t_1)\mathrm{var}_X(t_2)}}. \tag{4.4}$$

The cross-covariance and cross-correlation are obtained similarly.

4.1.2. Smoothing functions

Assuming that a functional datum for replication i arrives as a set of discrete measured values, $y_{i1}, \dots, y_{in}$, the first task is to convert these values to a function x_i with values $x_i(t)$ computable for any desired argument t. If the discrete values are assumed to be errorless, then the process is called interpolation. However, if they have some observational errors that need removing, then the conversion from discrete data to functions may involve smoothing [91].

The smoothing is based on representing the functions by basis functions. A basis function system is a set of known functions ϕ_k that are independent and that enable to approximate any function by taking linear combination of K of these functions. Thus we can express a function x as

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t). \tag{4.5}$$

The interpolation is achieved for $K = n$ in the sense that we can choose the coefficients c_k to yield $x(t_j) = y_j$ for each j. Therefore the degree of smoothing is determined by the number of basis functions, K, which is considered as a parameter corresponding to the characteristics of the data.

B-spline representation

Spline functions are the most common choice of approximation system for non-periodic functional data.

The interpolation spline is a function which is a piecewise polynomial of lower degree whose parts are smoothly joined [75, 91]. Let $\Delta\lambda$ is an increasing sequence of $g+2$ knots, i.e. $a = \lambda_0 < \lambda_1 < \cdots < \lambda_g < b = \lambda_{g+1}$, and $s_k(t)$ is the polynomial spline defined on finite interval $[a, b]$ with characteristics

- $s_k(t)$ is the polynomial of maximum degree k on each interval $[\lambda_i, \lambda_{i+1}]$ for $i = 0, 1, \ldots, g-1$,
- $s_k(t)$ has continuous derivatives up to order $k-1$ in knots λ_i.

Moreover, let $\mathcal{S}_k^{\Delta\lambda}[a, b]$ denotes the vector space of polynomial splines of degree $k > 0$, defined on a finite interval $[a, b]$ with the sequence of knots $\Delta\lambda$. The dimension of $\mathcal{S}_k^{\Delta\lambda}[a, b]$ equals $g + k + 1$. Then every spline $s_k(t)$ can be uniquely expressed as linear combination of $g + k + 1$ basis functions. However, when working with B-spline basis, only $g + k - 1$ linearly independent B-splines can be constructed on $\Delta\lambda$. Therefore we need to add $2k$ knots that fulfill condition

$$\lambda_{-k} \leq \lambda_{-k+1} \leq \cdots \leq \lambda_0, \;\; \lambda_{g+1} \leq \lambda_{g+2} \leq \cdots \leq \lambda_{g+k+1}.$$

In the following, we work with such extended sequence of knots, where additional knots equal λ_0 and λ_{g+1}, i.e.

$$\lambda_{-k} = \lambda_{-k+1} = \cdots = \lambda_0, \ \lambda_{g+1} = \lambda_{g+2} = \cdots = \lambda_{g+k+1}.$$

Then every spline $s_k(t) \in \mathcal{S}_k^{\Delta\lambda}[a,b]$ can be expressed as

$$s_k(t) = \sum_{i=-k}^{g} b_i B_i^{k+1}(t), \tag{4.6}$$

where $\mathbf{b} = (b_{-k}, \ldots, b_g)'$ is a vector of B-spline coefficients and $B_i^{k+1}(t), \ i = -k, \ldots, g$ are B-splines of degree k and form basis in $\mathcal{S}_k^{\Delta\lambda}[a,b]$. Spline can be also defined using collocation matrix as

$$s_k(t) = \mathbf{B}_{k+1}(t)\mathbf{b}, \tag{4.7}$$

where $\mathbf{b}$ is a vector of B-spline coefficients and $\mathbf{B}_{k+1}(t)$ is the mentioned collocation matrix, which is defined for given $\mathbf{t} = (t_1, \ldots, t_n)'$ and B-spline basis $B_i^{k+1}(t), \ i = -k, \ldots, g$ as

$$\mathbf{B}_{k+1}(\mathbf{t}) = \begin{pmatrix} B_{-k}^{k+1}(t_1) & \ldots & B_g^{k+1}(t_1) \\ \vdots & \ddots & \vdots \\ B_{-k}^{k+1}(t_n) & \ldots & B_g^{k+1}(t_n) \end{pmatrix} \in \mathbb{R}^{n,g+k+1}. \tag{4.8}$$

For $l \in \{1.\ldots, k-1\}$ the derivative of order l of the spline $s_k(t) \in \mathcal{S}_k^{\Delta\lambda}[a,b]$ is a spline $s_{k-l}(t) \in \mathcal{S}_{k-l}^{\Delta\lambda}[a,b]$ with the same knots. The l-th spline derivative can be written as

$$s_k^{(l)}(t) = \mathbf{B}_{k+1-l}(t)\mathbf{b}^{(l)}, \tag{4.9}$$

where $\mathbf{b}^{(l)} \in \mathbb{R}^{g+k+1-l}$ is given by

$$\mathbf{b}^{(l)} = \mathbf{D}_l\mathbf{L}_l\mathbf{b}^{(l-1)} = \mathbf{D}_l\mathbf{L}_l \ldots \mathbf{D}_1\mathbf{L}_1\mathbf{b} = \mathbf{S}_l\mathbf{b}$$

and $\mathbf{b}^{(0)} = \mathbf{b}$. $\mathbf{S}_l \in \mathbb{R}^{g+k+1-l,g+k+1}$ is upper triangular matrix with full rank, $\mathbf{D}_j = (k+1-j)\mathrm{Diag}(d_{-k+j}, \ldots, d_g) \in \mathbb{R}^{g+k+1-j,g+k+1-j}$ with $d_i = \frac{1}{\lambda_{i+k+1-j} - \lambda_i}$ $\forall i =$

$-k+j,\dots,g$ and

$$\mathbf{L}_j := \begin{pmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix} \in \mathbb{R}^{g+k+1-j,g+k+2-j}.$$

The spline smoothing is the methodology of data aproximation, which is a compromise between the spline interpolation and least-squares approximation.

Given a raw datum $x(t_i)$, observed at the time points $t_i \in [a,b]$, $i=1,\dots,n$, we seek for the coefficients of the smoothing spline [16] that minimizes the functional

$$J_l(s_k) = \sum_{i=1}^{n} w_i \left[x(t_i) - s_k(t_i)\right]^2 + \psi \int_I \left[s_k^{(l)}(t)\right]^2 \mathrm{d}t, \tag{4.10}$$

where $w_i \geq 0$, $i=1,\dots,n$, $n \geq g+1$, are given weights and $\psi \geq 0$ a given parameter. The parameter ψ controls the impact of the differential penalization appearing in (4.10) and is thus associated with the smoothness of the resulting approximation. We set the weights as well as the value of the smoothing parameter ψ to one, following the default setting of [75]. For possible sensible determination of ψ using, e.g. cross-validation, we refer to [64] and [73]. In general, for the purpose of setting the parameters all the techniques which are in use in FDA can be employed in this case as well.

Let $\mathbf{b}^* = (b^*_{-k},\dots,b^*_g)'$ is the resulting vector of B-spline coefficients [75], then spline

$$s^*_k(t) = \sum_{i=-k}^{g} b^*_i B_i^{k+1}(t) \tag{4.11}$$

is the best approximation of $J_l(s_k)$ in the sense of least squares.

4.1.3. Density functions

As it was mentioned above, density functions can be considered as functional compositional data with the integral constraint equal to one. Since compositional data are discrete (in sense that they have just a finite number of parts), the generalization of the Aitchison geometry to functions led to introduce Bayes spaces [8, 9, 24, 29].

We denote by μ an absolutely continuous measure with respect to the Lebesgue measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, with compact support $I \subset \mathbb{R}$ and density f. To keep the notation simple, hereafter we refer to the properties of μ through the properties of f, even though [8, 9, 29] develop the theory of Bayes spaces in a complete generality. We say that two density functions f, g are equivalent if they are proportional, and we denote this by $f =_{\mathcal{B}} g$. As such, the integral constraint $\int_I f(x)\,\mathrm{d}x = 1$ of a probability density function singles out a representative within an equivalence class of functional compositions that provide the same set of information. Indeed, any other representative $\widetilde{f}$ within the same class (and characterized by a constraint $\int_I \widetilde{f}(x)\,\mathrm{d}x = c$ for $c > 0$) carries the same relative information on the contribution of any Borel subset of $\mathbb{R}$ to the measure of the support. In this setting – as noted by [29] – the probability of a given event has not a meaning *per se*, but should be compared with the probability of the entire sample space, which is conventionally set to 1, but could be equivalently set to another positive constant c. This property is known as *scale invariance*.

A second important feature of functional compositions is the *relative scale* property: the relative increase of a probability over a Borel set from 0.05 to 0.1 (2 multiple) differs from the increase 0.5 to 0.55 (1.1 multiple), although the absolute differences are the same in both cases. This property reflects the relative nature of functional compositions, and further motivates the use of the log-ratio approach – already extensively employed in compositional data analysis – to deal with density functions.

In fact, both the scale invariance and the relative scale properties are completely ignored when considering probability density functions just like unconstrained functional data. In particular, the usual notions of sum and product by a constant appear inappropriate when applied to compositions, since the space of functional compositions endowed with those operations is not a vector space (e.g., the point-wise sum of two compositions is not necessarily a composition).

Instead, the geometry of the Bayes Hilbert space of [9], which is described below, enables one to capture and properly incorporate these properties. In the following, we restrict our attention to density functions with compact support, as in [18]. Both theoretical and practical reasons motivate this choice. Indeed, when the support is the whole real line, the Lebesgue measure cannot be used as reference probability measure, leading to highly technical issues. Moreover, in most real datasets, finite values for the inferior and superior extremes of the support can be determined without a substantial loss of generality.

We call $\mathcal{B}^2(I)$ the Bayes space of (equivalence classes of) positive functional compositions f on I with square-integrable logarithm. In particular, we here consider continuous (hence bounded) functional compositions on the compact support I. Hereafter, the representative of an equivalence class will be its element integrating to 1. Moreover, the symbol I will denote an interval $[a, b]$ but any subset of $\mathbb{R}$ with finite measure could be dealt with analogously. Given two absolutely integrable density functions $f, g \in \mathcal{B}^2(I)$ and a real number $\alpha \in \mathbb{R}$ we indicate with $f \oplus g$ and $\alpha \odot f$ the perturbation and powering operations, respectively, defined as [9, 24]:

$$(f \oplus g)(t) = \frac{f(t)g(t)}{\int_I f(s)g(s)\,\mathrm{d}s}, \quad (\alpha \odot f)(t) = \frac{f(t)^\alpha}{\int_I f(s)^\alpha\,\mathrm{d}s}, \quad t \in I.$$

The resulting functions are readily seen to be probability density functions. [24] prove that $\mathcal{B}^2(I)$ endowed with the operations $(\oplus, \odot)$ is a vector space. Note that the neutral elements of perturbation and powering are $e(t) = 1/\eta$, with $\eta = b - a$ (i.e., the uniform density), and 1, respectively. Moreover, the difference between two elements $f, g \in \mathcal{B}^2(I)$, denoted by $f \ominus g$, is obtained as perturbation of f with the reciprocal of g, i.e., $(f \ominus g)(t) = (f \oplus [(-1) \odot g])(t), t \in I$.

To endow $\mathcal{B}^2(I)$ with a Hilbert space structure, [24] define the inner product

$$\langle f, g \rangle_{\mathcal{B}} = \frac{1}{2\eta} \int_I \int_I \ln \frac{f(t)}{f(s)} \ln \frac{g(t)}{g(s)}\,\mathrm{d}t\,\mathrm{d}s, \quad f, g \in \mathcal{B}^2(I), \tag{4.12}$$

with $\eta = b - a$, which induces the following norm,

$$||f||_{\mathcal{B}} = \left[\frac{1}{2\eta}\int_I\int_I \ln^2\frac{f(t)}{f(s)}\,\mathrm{d}t\,\mathrm{d}s\right]^{1/2}.$$

The space $\mathcal{B}^2(I)$, endowed with the inner product (4.12), is proved to be a separable Hilbert space in [24]. As such, it is isomorphic to the Hilbert space $L^2(I)$ of (equivalence classes of) square-integrable real functions on I. An isometric isomorphism between $\mathcal{B}^2(I)$ and $L^2(I)$ is defined by the *centred log-ratio* (clr) transformation [9, 81], which is defined, for $f \in \mathcal{B}^2(I)$, as

$$\mathrm{clr}(f)(t) = f_c(t) = \ln f(t) - \frac{1}{\eta}\int_I \ln f(s)\,\mathrm{d}s. \tag{4.13}$$

We remark that such an isometry allows to compute operations and inner products among the elements in $\mathcal{B}^2(I)$ in terms of their counterpart in $L^2(I)$ among the clr-transforms, i.e.

$$\mathrm{clr}(f \oplus g)(t) = f_c(t) + g_c(t), \quad \mathrm{clr}(\alpha \odot f)(t) = \alpha \cdot f_c(t),$$

$$\langle f, g\rangle_{\mathcal{B}} = \langle f_c, g_c\rangle_2 = \int_I f_c(t)g_c(t)\,\mathrm{d}t.$$

However, note that, by construction, the constraint

$$\int_I \mathrm{clr}(f)(t)\mathrm{d}t = \int_I \ln f(t)\,\mathrm{d}t - \int_I \frac{1}{\eta}\int_I \ln f(s)\,\mathrm{d}s\,\mathrm{d}t = 0$$

occurs. This additional condition needs to be taken into account for computation and analysis on clr-transformed density functions, as we shall show in Section 4.3. Accordingly, by considering this condition when smoothing discretized clr transformed densities as described in Section 4.1.2, the optimal coefficients are obtained as

$$\bar{\mathbf{b}}^* = \mathbf{DK}\left[(\mathbf{B}_{k+1}(\mathbf{t})\mathbf{DK})^\top \mathbf{W}\mathbf{B}_{k+1}(\mathbf{t})\mathbf{DK} + \psi\,(\mathbf{DK})^\top \mathbf{N}_{kl}\mathbf{DK}\right]^+ \mathbf{K}^\top\mathbf{D}^\top\mathbf{B}_{k+1}^\top(\mathbf{t})\mathbf{W}\mathbf{y},$$

where $\mathbf{W} = \mathrm{Diag}(\mathbf{w})$, $\mathbf{A}^+$ denotes the Moore-Penrose pseudoinverse of a matrix $\mathbf{A}$,

$$\mathbf{D} = (k+1)\,\mathrm{Diag}\left(\frac{1}{\lambda_1 - \lambda_{-k}}, \ldots, \frac{1}{\lambda_{g+k+1} - \lambda_g}\right) \in \mathbb{R}^{g+k+1,g+k+1},$$

$$\mathbf{K} = \begin{pmatrix} 1 & 0 & 0 & \cdots & -1 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \end{pmatrix} \in \mathbb{R}^{g+k+1,g+k+1}$$

and $\mathbf{N}_{kl} = \mathbf{S}_l^\top \mathbf{M}_{kl} \mathbf{S}_l$ is a positive semidefinite matrix with

$$\mathbf{M}_{kl} = \begin{pmatrix} \langle B_{-k+l}^{k+1-l}, B_{-k+l}^{k+1-l} \rangle_2 & \dots & \langle B_g^{k+1-l}, B_{-k+l}^{k+1-l} \rangle_2 \\ \vdots & & \vdots \\ \langle B_{-k+l}^{k+1-l}, B_g^{k+1-l} \rangle_2 & \dots & \langle B_g^{k+1-l}, B_g^{k+1-l} \rangle_2 \end{pmatrix} \in \mathbb{R}^{g+k+1-l,g+k+1-l}.$$

4.2. Principal component analysis for functional data

Principal component analysis (PCA) is a widely used multivariate statistical technique aiming to capture the main modes of variability of the data by means of a small number of linear combinations of the original variables. In the functional context, the same aim is reached by functional principal component analysis (FPCA). Here, we briefly recall FPCA, referring the reader, e.g. to [90, Chapter 8], [51, Chapter 3] and [100], for further details on this topic.

Let us consider a functional random sample $X_1, ..., X_N$ in $L^2(I)$, and indicate with $\langle x, y \rangle_2 = \int_I x(t)y(t)dt$ the inner product between two elements x, y in $L^2(I)$ and with $\|x\|_2 = (\int_I |x(t)|^2 dt)^{1/2}$ the induced norm. For ease of notation and without loss of generality, we assume the samples to be centred. FPCA firstly looks for the main mode of variability, i.e., for the element ξ_1 in $L^2(I)$ – called first functional principal component (FPC) – maximizing over $\xi \in L^2(I)$

$$\frac{1}{N} \sum_{i=1}^{N} \langle X_i, \xi \rangle_2^2 \text{ subject to } \|\xi\|_2 = 1. \tag{4.14}$$

The remaining FPCs, $\{\xi_j\}_{j \geq 2}$, capture the remaining modes of variability subject to be mutually orthogonal, and are thus obtained by solving problem (4.14) with the additional orthogonality constraint $\langle \xi_k, \xi \rangle_2 = 0, k < j$.

Analogously to the multivariate case, the FPCs $\{\xi_j\}_{j\geq 1}$ coincide with the eigenfunctions of the sample covariance operator $V : L^2(I) \to L^2(I)$, e.g. [51, p. 26], acting on $x \in L^2(I)$ as

$$Vx = \frac{1}{N}\sum_{i=1}^{N} \langle X_i, x\rangle_2 X_i,$$

or, equivalently, as

$$Vx = \int_I v(\cdot, t)x(t)dt,$$

the kernel $v : I \times I \to \mathbb{R}$ being the sample covariance function

$$v(s,t) = \frac{1}{N}\sum_{i=1}^{N} x_i(s)x_i(t), \quad s, t \in I.$$

Therefore, the j-th FPC ξ_j and the associated scores $\Psi_{ij} = \langle X_i, \xi_j\rangle_2$, $i = 1, ..., N$, are obtained by solving the eigenvalue equation

$$V\xi_j = \rho_j \xi_j, \tag{4.15}$$

where ρ_j denotes the j-th eigenvalue, with $\rho_1 \geq \rho_2 \geq ...$. As in multivariate PCA, for each j, the term $\rho_j / \sum_j \rho_j$ is associated with the proportion of total variability explained by the FPC ξ_j.

Several computational methods can be utilized to solve equation (4.15), e.g. [62, 65, 91]. [91, Chapter 8.4] suggest to express each datum X_i, $i = 1, ..., N$, as a linear combination of K known basis functions $\phi_1, ..., \phi_K$ and to solve the eigenproblem (4.15) through an appropriate matrix coefficient. Indeed, suppose that each datum X_i, $i = 1, ..., N$, admits the basis expansion

$$X_i(\cdot) = \sum_{k=1}^{K} c_{ik}\phi_k(\cdot), \tag{4.16}$$

or, in matrix notation, $\mathbf{X}(\cdot) = \mathbf{C}\boldsymbol{\phi}(\cdot)$, with $\mathbf{C} = (c_{ik}) \in \mathbb{R}^{N,K}$, $\mathbf{X}(\cdot) = (X_i(\cdot))$, and $\boldsymbol{\phi}(\cdot) = (\phi_i(\cdot))$. Then the variance-covariance function takes the form $v(s,t) =$

$N^{-1}\boldsymbol{\phi}(s)'\mathbf{C}'\mathbf{C}\boldsymbol{\phi}(t)$, $s,t \in I$. Suppose further that the eigenfunction ξ_j, $j \geq 1$, admits the expansion $\xi_j(\cdot) = \sum_{k=1}^{K} a_{jk}\phi_k(\cdot)$, or in matrix notation $\xi_j(\cdot) = \boldsymbol{\phi}(\cdot)'\mathbf{a}_j$. This yields $V\xi_j(\cdot) = \boldsymbol{\phi}(\cdot)'N^{-1}\mathbf{C}'\mathbf{C}\mathbf{M}\mathbf{a}_j$, where $\mathbf{M}_{kl} = \langle \phi_k, \phi_l \rangle_2$. Therefore the eigenvalue equation (4.15) reduces to

$$N^{-1}\mathbf{C}'\mathbf{C}\mathbf{M}\mathbf{a}_j = \rho_i \mathbf{a}_j, \tag{4.17}$$

and $\mathbf{a}_j$ is obtained as solution of the linear system (4.17). Note that in case of basis orthonormality $\mathbf{M} = \mathbf{I}$ the FPCA problem reduces to standard multivariate PCA of the coefficient matrix $\mathbf{C}$. Otherwise, [91, Chapter 8.4] show that problem (4.17) is equivalent to the eigenproblem

$$\frac{1}{N}\mathbf{M}^{1/2}\mathbf{C}'\mathbf{C}\mathbf{M}^{1/2}\mathbf{u}_j = \rho_i \mathbf{u}_j$$

with $\mathbf{u}_j = \mathbf{M}^{1/2}\mathbf{a}_j$, i.e., FPCA reduces to a multivariate PCA of the transformed coefficient matrix $\mathbf{C}\mathbf{M}^{1/2}$ followed by the transformation $\mathbf{a} = \mathbf{M}^{-1/2}\mathbf{u}$.

4.3. Simplicial functional principal component analysis

As functional compositions, probability density functions are featured by specific properties, such as the scale invariance and relative scale properties. The latter would be neglected, if one applied the functional principal component analysis described in Section 4.2 to density functions. Aim of this Section is to derive a simplicial version of FPCA, named SFPCA, by following the same scheme that led to the formulation of FPCs in Section 4.2, but in agreement with the Bayes Hilbert space geometry introduced in Section 4.1.3.

Let $\widetilde{X}_1, ..., \widetilde{X}_N$ be a sample in $\mathcal{B}^2(I)$, and denote by $X_1, ..., X_N$ the corresponding centred sample, i.e., for $i = 1, ..., N$, $X_i = \widetilde{X}_i \ominus \overline{X}$, where $\overline{X}$ stands for the sample mean $\overline{X} = \frac{1}{N} \odot \bigoplus_{i=1}^{N} \widetilde{X}_i$. We consider the problem of finding the simplicial functional principal components (SFPCs) in $\mathcal{B}^2(I)$, i.e. the elements

$\{\zeta_j\}_{j\geq 1}$, $\zeta_j \in \mathcal{B}^2(I)$, maximizing the following objective function over $\zeta \in \mathcal{B}^2(I)$:

$$\frac{1}{N}\sum_{i=1}^{N}\langle X_i, \zeta\rangle_{\mathcal{B}}^2 \text{ subject to } \|\zeta\|_{\mathcal{B}} = 1;\ \langle \zeta, \zeta_k\rangle_{\mathcal{B}} = 0,\ k < j, \tag{4.18}$$

where the orthogonality condition $\langle \zeta, \zeta_k\rangle_{\mathcal{B}} = 0$, for $k < j$, holds only for $j \geq 2$.

Because $\mathcal{B}^2(I)$ is a separable Hilbert space, the minimization problem (4.18) is well posed [51, Theorem 3.2, p. 38]. Thus, the solution of (4.18) exists and is unique. Indeed, analogously to the $L^2(I)$ case previously discussed, the j-th SFPC solves the eigenvalue equation

$$V\zeta_j = \delta_j \odot \zeta_j, \tag{4.19}$$

(δ_j, ζ_j) being the j-th eigenpairs of the sample covariance operator $V : \mathcal{B}^2(I) \to \mathcal{B}^2(I)$, acting on $x \in \mathcal{B}^2(I)$ as

$$Vx = \frac{1}{N} \odot \bigoplus_{i=1}^{N} \langle X_i, x\rangle_B \odot X_i.$$

In order to proceed with (4.18) in practice, i.e. to express densities in the standard L^2 space, we apply the isometric isomorphism between $\mathcal{B}^2(I)$ and $L^2(I)$ defined by the clr-transform (4.13) that allows to rewrite the original problem (4.18) as a maximization of the term

$$\frac{1}{N}\sum_{i=1}^{N}\langle \mathrm{clr}(X_i), \mathrm{clr}(\zeta)\rangle_2^2 \text{ subject to } \|\mathrm{clr}(\zeta)\|_2 = 1;\ \langle \mathrm{clr}(\zeta), \mathrm{clr}(\zeta_k)\rangle_2 = 0,\ k < j$$

over $\zeta \in \mathcal{B}^2(I)$. Accordingly, for $j \geq 1$ the maximization problem (4.18) can be equivalently restated as finding $\nu \in L^2$ which maximizes

$$\frac{1}{N}\sum_{i=1}^{N}\langle \mathrm{clr}(X_i), \nu\rangle_2^2 \text{ subject to } \|\nu\|_2 = 1;\ \langle \nu, \nu_k\rangle_2 = 0,\ k < j;\ \int_I \nu = 0, \tag{4.20}$$

where the orthogonality constraint is meaningful only for $j \geq 2$ and the zero-integral constraint incorporates the corresponding clr-transform property.

We now show that (4.20) is solved by the eigenfunctions $\{\xi_j\}_{j\geq 1}$ of the sample covariance operator $V_{\mathrm{clr}} : L^2(I) \to L^2(I)$ of the transformed sample $\mathrm{clr}(X_1), ...,$ $\mathrm{clr}(X_N)$, acting on $x \in L^2(I)$ as

$$V_{\mathrm{clr}} x = \frac{1}{N} \sum_{i=1}^{N} \langle \mathrm{clr}(X_i), x \rangle_2 \, \mathrm{clr}(X_i).$$

We first notice that the eigenfunctions $\{\xi_j\}_{j\geq 1}$ would have solved problem (4.20), if it had been stated without the zero-integral condition $\int_I \nu = 0$, since in that case (4.20) would have been equivalent to (4.14) (with the orthogonality constraints for $j \geq 2$). Therefore, to prove that $\nu = \xi_j$ maximizes (4.20) it suffices to show that ξ_j fulfills also the constraint $\int_I \xi_j = 0$, for all $j \geq 1$. To this end, we note that the zero-integral property of the clr-transformed sample $\mathrm{clr}(X_1), ..., \mathrm{clr}(X_N)$ implies that V_{clr} admits a zero eigenvalue with associated eigenfunction $\xi_0 \equiv 1/\sqrt{b-a}$:

$$V_{\mathrm{clr}}\, \xi_0 = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\sqrt{b-a}} \left[\int_I \mathrm{clr}(X_i) \right] \mathrm{clr}(X_i) \equiv 0.$$

Since the eigenfunctions $\{\xi_j\}$ corresponding to the remaining non-null eigenvalues $\{\rho_j\}$ are to be orthogonal to the eigenfunction ξ_0, the ξ_j's need to satisfy the zero-integral condition $\int_I \xi_j = 0$, as $\langle \xi_j, \xi_0 \rangle_2 = 1/\sqrt{b-a} \int_I \xi_j$. Another way to see this is to notice that (a) the image of the sample covariance operator V_{clr} is the span of the clr-transformed observations, (and the constant function $\xi_0 \equiv \frac{1}{\sqrt{b-a}}$ belongs to its kernel), and (b) the eigenfunctions corresponding to the non-null eigenvalues form a basis of the image of V_{clr}. As such, each eigenfunction ξ_j can be written as a unique linear combination of the functions $\mathrm{clr}(X_1), ..., \mathrm{clr}(X_N)$. Therefore, the zero-integral condition is fulfilled since it holds, by construction, for each of the functions $\mathrm{clr}(X_i)$, $i = 1, ..., N$. Thus, problem (4.18) can be restated in terms of clr-transforms as (4.20) and the SFPCs can be obtained by transforming the eigenfunctions $\{\xi_j\}_{j\geq 1}$ associated to the non-null eigenvalues $\{\rho_j\}_{j\geq 1}$ of V_{clr} through the inverse of the function clr, namely $\zeta_j = \mathrm{clr}^{-1}(\xi_j) =_{\mathcal{B}} \exp(\xi_j)$, with $j \geq 1$. Note that, as in classical PCA, the eigenfunctions ξ_j are determined up to sign changes. Accordingly, the SFPCs are determined up to a powering by ±1 (i.e., if ζ_j solves problem (4.18), $-1 \odot \zeta_j$ is a solution as well).

To compute the eigenfunctions ξ_j we resort to a method based on a B-spline basis expansion. Following [75], we consider for $\mathrm{clr}(X_1), ..., \mathrm{clr}(X_N)$ and ξ_j, $j \geq 1$, a B-spline basis fulfilling the zero-integral constraint,

$$\mathrm{clr}(X_i)(\cdot) = \sum_{k=1}^{K} c_{ik}\phi_k(\cdot), \quad \xi_j(\cdot) = \sum_{k=1}^{K} a_{jk}\phi_k(\cdot). \tag{4.21}$$

To compute the B-spline coefficients the usual parametrization of smoothing splines applies, and the additional constraint is incorporated in the estimation algorithm as described in [75]. Hence, with the same arguments used in Section 4.2, $\mathbf{a}_j = (a_{jk})$ is obtained as solution of the eigenproblem

$$N^{-1}\mathbf{C}'\mathbf{C}\mathbf{M}\mathbf{a}_j = \rho_i \mathbf{a}_j,$$

with analogue orthogonality arguments as those previously introduced, the zero integral constraint is inherently kept in the PCA algorithm, and thus does not need to be explicitly imposed.

For the purpose of dimensionality reduction, the choice of the number of SFPCs to be retained can follow the same strategies as those used in FPCA: one may fix a threshold in the amount of variability explained by the retained SFPCs, or look for an elbow in the scree plot. Even the interpretation of the results of SFPCA may follow the main lines used in the $L^2(I)$ case, since the SFPCs represent the main modes of variability of the observations around the global mean function, but in the space $\mathcal{B}^2(I)$ endowed with its own geometry. Finally, a useful tool to visualize and interpret the results of SFPCs is also the scores plan graph.

4.4. Analysis of salary distributions in Austria

To demonstrate usefulness of SFPCA in economic applications, an example of hourly wages in Austria is presented. The data set is a subset of synthetically generated real Austrian SES (Structural Earnings Survey) data, contained in the R-package `laeken` [4]. The total of 15691 data was divided into 10 salary intervals, according to Sturges rule, for non-zero wages up to 44.5 EUR. Note that only data below 99%-quantile were used to eliminate extreme values. The wages were also divided according to location (eastern, southern and western Austria) and five age intervals. In the next step, the absolute values were transformed to proportions to obtain density functions and then the clr transformation (4.13), in its descrete form (1.3), was applied.

The resulting values were smoothed using B-splines as described in Section 4.1.2 on the support $I = [0, 44.5]$. The cubic smoothing splines with knots 2.22752, 10, 20, 30 and 42.25 were used to obtain the B-spline coefficients that were used afterwards for representing the clr transformed densities. The B-spline cubic basis is displayed in Figure 4.1. The resulting smoothed densities are displayed in Figure 4.2, where clr transformed densities are on the left and the original data on the right. Note that different colours distinguish the location - green for western Austria, blue for eastern Austria and red for southern Austria. To simplify the notation, age intervals were denoted as 1 for $\langle 15, 29)$, 2 for $\langle 29, 39)$, 3 for $\langle 39, 49)$, 4 for $\langle 49, 59)$ and 5 for $\langle 59, 120)$.

From Figure 4.3 the covariance structure is rescaled. By considering the original densities (Figure 4.3, right), most of the variability is captured by the left part of densities, however the covariance structure of clr densities (Figure 4.3, left) shows that the variability is distributed between the left and right tails of the distribution.

In the next step, functional principal component analysis was applied to clr densities and to the original density functions. According to scree plot (Figure 4.4), two or maximum three SFPCs need to be taken. In Figure 4.5 scatter plots of scores for the first two SFPCs and FPCs (the case of analyzing the original data), respectively, are displayed. The left plot shows that age structure dominates the location - cluster of younger people (15–29 years) is on the right side and cluster of the oldest people (59 and more years) is on the lower left part

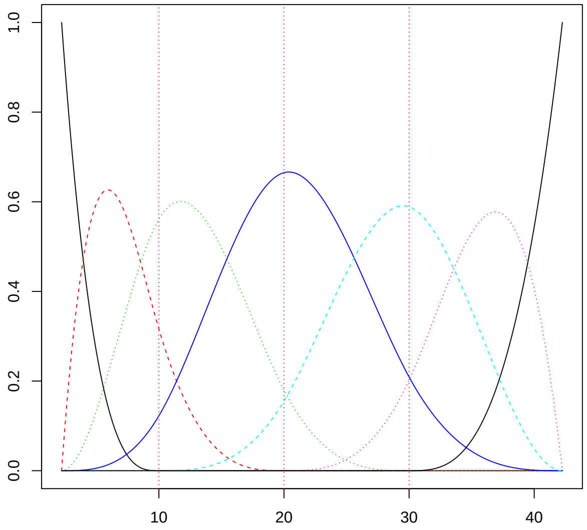

Figure 4.1: The B-spline basis.

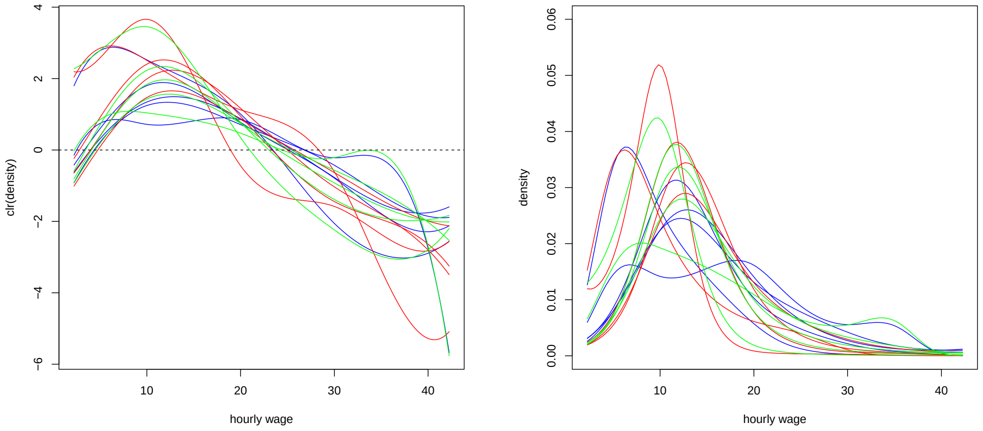

Figure 4.2: Density of clr-transformed densities (left) and original densities (right) of hourly wages in Austria.

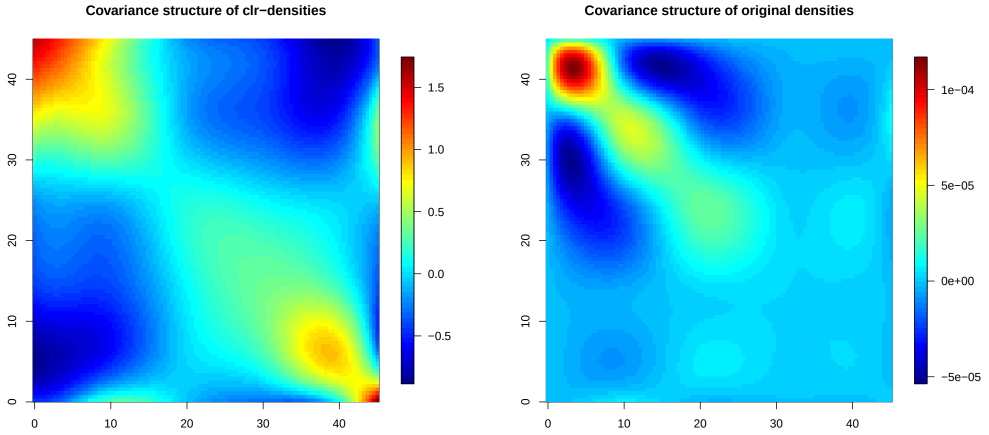

Figure 4.3: The covariance structure of clr-densities (left) and original densities (right).

of the plot (except of the oldest group of people from southern Austria). In the second quadrant of the scatter plot most of the other observations are located, surprisingly still clustered according to age groups. By looking at data structure in Figure 4.5 (left), it can be concluded that the first SFPC can be linked to height of salary. Particularly, it nicely reflects the general fact that in early age salaries are not very high and with increasing experiences (followed by increasing age) the salary increases as well. The only exception seems to be southern Austria for the oldest age group that might indicate some specific structure of the labour market there. Moreover, it seems that in eastern Austria, people have in general higher salaries - blue observations are always located in the left part of each age group. It might correspond to the fact that in eastern Austria also Vienna, capital of the country, is located. In the right plot of Figure 4.5, grouping according to age is still preserved, but clusters are far not so well separated as before and also the role of the oldest age group is no more so clear. Regional effects are completely lost.

In Figure 4.6, the first three functional principal components (harmonics) are visualized. In the upper panel, the SFPCs are displayed - the first component characterizes the variability of the right tail of the distribution, the main contri-

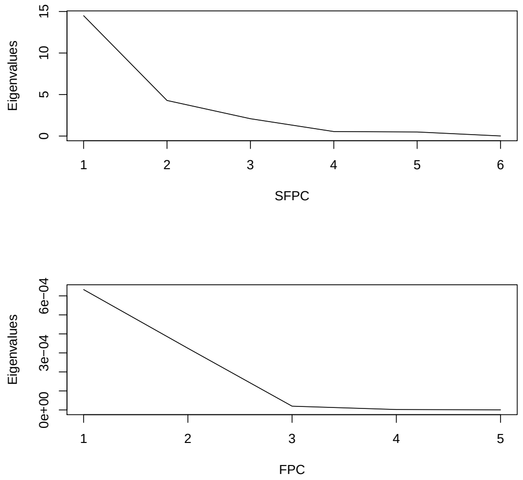

Figure 4.4: Scree plot SFPCA (upper plot) and FPCA (lower plot).

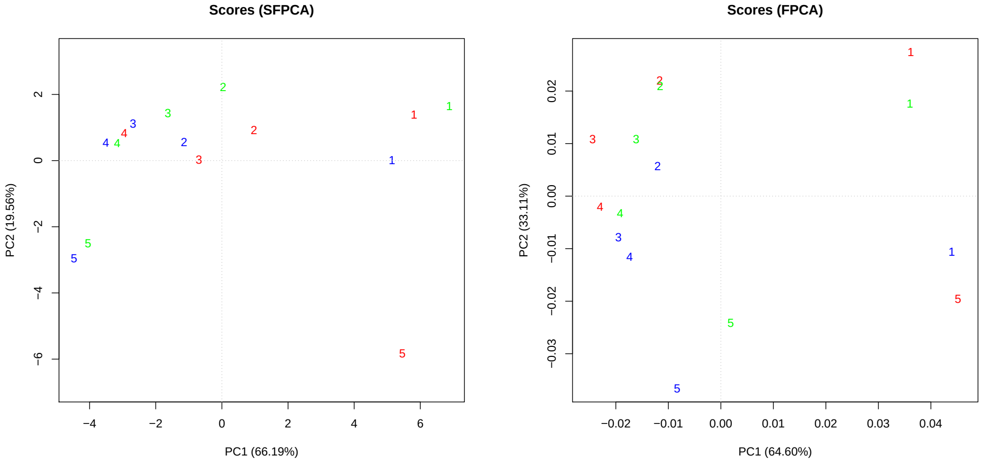

Figure 4.5: Scatter plot of scores of SFPCA (left) and FPCA (right).

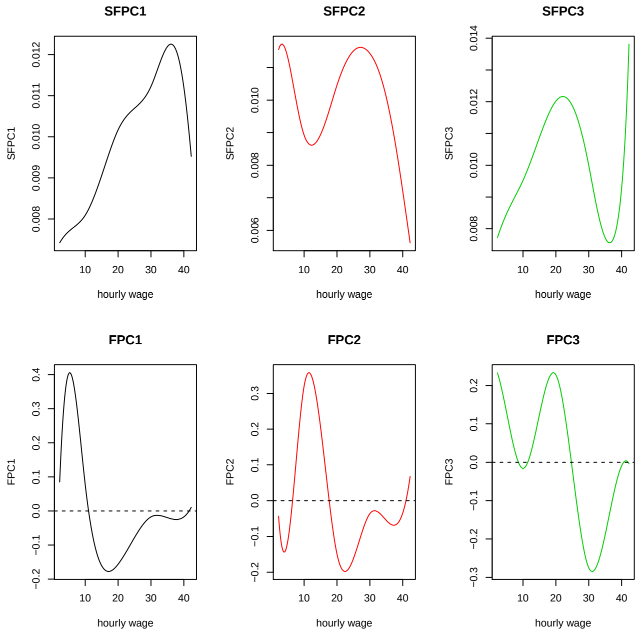

Figure 4.6: Plots of harmonics of SFPCA (upper panel) and FPCA (lower panel).

bution is provided by wages higher than 30 EUR. The second SFPC highlights lower incomes, similarly as the third SFPC. The lower panel of Figure 4.6 provides a completely different picture. The first two harmonics display the largest variability on the left tail of the distribution, the third FPC is hardly interpretable. Variability in higher wages is thus poorly represented by analyzing the original densities.

In Figure 4.7 projections of densities using SFPC1 and SFPC2 (left) and FPC1 and FPC2 (right) are displayed. We can see that projection using the results of SFPCA comparing to the right plot of Figure 4.2 is satisfactory - the main modes of variability are pretty captured. However, the right plot of Figure 4.7 is hardly comparable with the original functions - some curves are even below zero, unacceptable for density functions, and the curves have also different shape.

From the results or SFPCA we can conclude that taking the relative nature of density functions, i.e. functional compositions, into account leads to better, more meaningful and interpretable results.

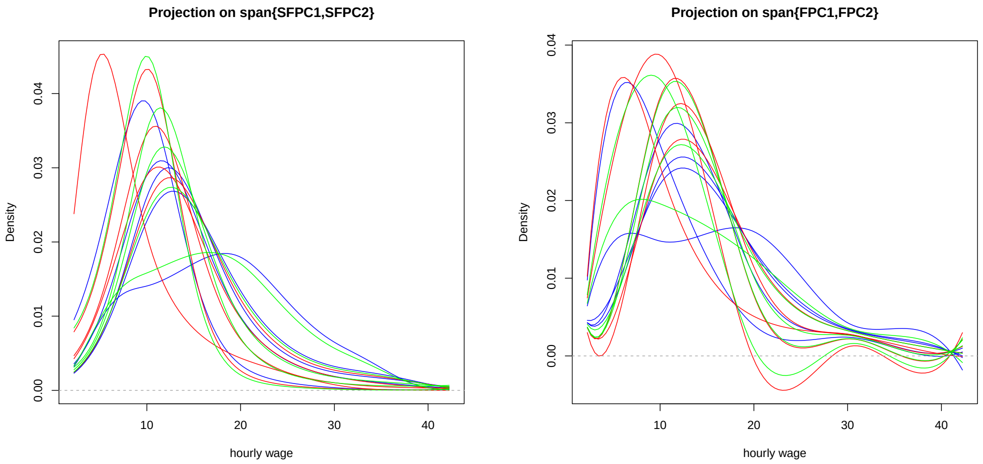

Figure 4.7: Approximated densities.

Conclusions

The main goal of this thesis was to show how the compositional analysis can be useful and meaningful in economic applications. Although not for all papers that form the base of the thesis we finally succeeded to include an economic application, it was shown in previous chapters that the logratio methodology has a great potential also in this field.

Firstly, the definition of compositional data, together with their geometry, coordinate representation and basic descriptive statistics, was necessary to introduce. Next parts of this thesis were focused on the popular statistical methods in compositional context.

Dimension reduction methods belong to the most favourite statistical tools when analyzing the structure of data. Firstly, the principal component analysis was introduced. Its results are often displayed using biplot - for the purposes of compositional data analysis its interpretation needs to be adapted due to clr coordinate representation. When dealing with three-part compositions, only two principal components are obtained that enables to analyze deeply their interpretation in sense of pairwise logratios of the original parts. If the input compositional data are enriched with a third mode, parallel factor analysis can be applied to reveal main patterns in the data structure. The dimension reduction methods were applied to trade flows structure data set to see the advantages of taking the relative nature of compositional data into account for statistical processing.

In regression analysis, several possibilities when analyzing compositional data can occur. In the basic setting, either the response or explanatory variables are compositional. Introducing these two regression models was necessary for further developments presented in the thesis. Particularly, for regression with compositional response and explanatory variables, where only a special case of two-part compositions is considered. Using standard regression for such data is completely inappropriate, but even for the compositional model it is always necessary to check first, whether data follow the trend proposed by the model. From our experience, this is very important step particularly for economic data.

A different methodological approach must be considered, when the response and also explanatory variables come from one composition. Accordingly, in addition to proper coordinate representation they are both naturally burdened by measurement errors which leads to errors-in-variable models, concretely to orthogonal regression based on principal component analysis. In order to proceed with the corresponding statistical inference, bootstrap sampling was used to construct confidence intervals for regression parameters and to test their significance. Due to outlying observations that might occur in real compositional data sets also robust version of the regression model was considered.

The final part of this thesis aimed to present more advanced field of statistical analysis - functional data analysis that is still not very common in economic applications. The chapter introduces the basics of functional data analysis, descriptive statistics, smoothing of functions and explains specifics of density functions as functional compositions. The general methodology was presented for the special case of functional principal component analysis, performed for clr transformed densities. Finally, the resulting simplicial functional principal component analysis was applied to synthetic data with salary distributions in Austria and compared with the standard approach.

The most difficult part of this thesis was to present different methods and their economic applications in a coincise form. The hope is that it could serve as a “guide” or, better, as a list of (necessarily incomplete) options how to deal with economic data that have compositional character. There are, of course, further important methods in economic context, not listed here, among them for example compositional time series. However, due to volatility of financial time series, that form one potential economic application, we are faced with multivariate ARCH and GARCH models that belong to the most difficult ones even in their univariate cases. Hence, this is one of many other possibilities how to further develop this topic in the future.

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] Aitchison J., Ng K.W. (2005). The role of perturbation in compositional data analysis. *Statistical Modelling* **5**: 173–185.

[4] Alfons A., Templ M. (2013). Estimation of social exclusion indicators from complex surveys: The R package laeken. *Journal of Statistical Software* **54** (15): 1–25.

[5] Barceló-Vidal C., Aguilar L., Martín-Fernández J.A. (2011). Compositional VARIMA time series. In Pawlowsky-Glahn V., Buccianti A. (eds.), *Compositional data analysis: Theory and applications*, 89–103. Chichester: Wiley.

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

[7] 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.

[8] Van den Boogaart K.G., Egozcue J.J., Pawlowsky-Glahn V. (2010). Bayes linear spaces. *Statistics and Operations Research Transactions* **34** (2): 201–222.

[9] Van den Boogaart K.G., Egozcue J.J., Pawlowsky-Glahn V. (2014). Bayes Hilbert spaces. *Australian & New Zealand Journal of Statistics* **56** (2): 171–194.

[10] Van den Boogaart K.G., Tolosana-Delgado R. (2008). compositions: A unified R package to analyze compositional data. *Computers & Geosciences* **34** (4): 320–338.

[11] Bro, R. (1997). PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems* **38**: 149–171

[12] Buccianti A., Mateu-Figueras G., Pawlowsky-Glahn V., eds. (2006). *Compositional Data Analysis in the Geosciences: From Theory to Practice*, London: Geological Society.

[13] Croux C., Fekri M., Ruiz-Gazen A. (2010). Fast and robust estimation of the multivariate errors in variables model. *Test* **19**: 286–303.

[14] Croux C., Filzmoser P., Pison G., Rousseeuw P.(2003). Fitting multiplicative models by robust alternating regressions. *Statistics and Computing* **13** (1): 23–36.

[15] Davison A.C., Hinkley D.V. (1997). *Bootstrap methods and their application.* Cambridge: Cambridge University Press.

[16] de Boor C. (2001). *A Practical Guide to Splines*. New York: Springer.

[17] Delicado P. (2007). Functional k-sample problem when data are density functions. *Computational Statistics* **22**: 391–410.

[18] Delicado P. (2011). Dimensionality reduction when data are density functions. *Computational Statistics and Data Analysis* 5**5**: 401–420.

[19] Devarajan S., Swaroop V., Zou H. (1996). The composition of public expenditure and economic growth. *Journal of Monetary Economics* **37**: 313–344.

[20] Doornik J.A. (2011). Robust estimation using least trimmed squares. Available online at
http://econ.au.dk/fileadmin/site_files/filer_oekonomi/subsites/creates/Seminar_Papers/2011/ELTS.pdf

[21] Eaton M.L. (1983). *Multivariate Statistics. A Vector Space Approach.* New York: Wiley.

[22] Egozcue J.J. (2009). Reply to "On the Harker Variation Diagrams; ..." by J.A. Cortés. *Mathematical Geosciences* **41** (7): 829–834.

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

[24] Egozcue J.J., Díaz-Barrero J.L., Pawlowsky-Glahn V. (2006). Hilbert space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica, English Series* **22** (4): 1175–1182.

[25] Egozcue J.J., Pawlowsky-Glahn V. (2011). Basic concepts and procedures. In: Pawlowsky-Glahn V, Buccianti A, editors. *Compositional data analysis: Theory and applications.* Chichester: Wiley, 12–28.

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

[27] Egozcue J.J., Pawlowsky-Glahn V. (2006). Simplicial geometry for compositional data. In Buccianti A., Mateu-Figueras G., Pawlowsky-Glahn V. (eds) *Compositional Data Analysis in the Geosciences: From Theory to Practice.* London: Geological Society, Special Publications 264, 145–160.

[28] 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.

[29] Egozcue J.J., Pawlowsky-Glahn V., Tolosana-Delgado R., Ortego M.I., van den Boogaart K.G. (2013). Bayes spaces: use of improper distributions and exponential families. *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas* **107**: 475–486.

[30] Environment Canada (2005). Guidance document on statistical methods for environmental toxicity tests. Report EPS /RM/46. Ottawa, ON.

[31] Fekri M., Ruiz-Gazen A. (2004). Robust weighted orthogonal regression in the errors-in-variables model. *Journal of Multivariate Analysis* **88**: 89–108.

[32] Filzmoser P. (1999). Robust principal components and factor analysis in the geostatistical treatment of environmental data. *Environmetrics* **10**: 363–375.

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

[34] Filzmoser P., Hron K. (2013). Robustness for compositional data. In Becker C., Fried R., Kuhnt S. (eds) *Robustness and Complex Data Structures.* Heidelberg: Springer, 117–131.

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

[36] Filzmoser P., Hron K., Reimann C. (2009). Univariate statistical analysis of environmental (compositional) data: Problems and possibilities. *Science of the Total Environment* **407** (23): 6100–6108.

[37] Filzmoser P., Serneels S., Maronna R., Van Espen P.J. (2007). *Robust multivariate methods in chemometrics.* Vienna University of Technology.

[38] Finney D.J. (1979). Bioassay and the practice of statistical inference. *International Statistical Review / Revue Internationale de Statistique* **47**: 1–12.

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

[40] 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 - Theory and Methods* **41**: 2367–2385.

[41] Fox, J. (2002). Bootstrapping Regression Models. Appendix to an R and S-PLUS Companion to Applied Regression.

[42] Fritz H. (2011). *Robust clustering and dimension reduction: methods, algorithms and implementation.* Vienna University of Technology (unpublished dissertation thesis).

[43] Fry T. (2011). Applications in Economics. In *Compositional data analysis: Theory and applications*, edited by Pawlowsky-Glahn V., Buccianti A., 318–326. Chichester: Wiley.

[44] Fuller, W.A. (1987). *Measurement Error Models.* New York: Wiley.

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

[46] Gallo M. (2013). Log-ratio and Parallel Factor Analysis: An Approach to Analyze Three-way Compositional Data. In *Advanced Dynamic Modeling of Economic and Social Systems*, edited by Proto A.N., Squillante M., Kacprzyk J., 209–221. Heidelberg: Springer.

[47] Giordani P., Kiers H., Del Ferraro M. (2014). Three-way component analysis using the R package ThreeWay. *Journal of Statistical Software*, **57** (7): 1–23.

[48] Golub G.H. (1973). Modified matrix eigenvalue problems. *SIAM Review* **15**: 318–334.

[49] Gower J.C., Hand D.J. (1996). *Biplots.* London: Chapman & Hall.

[50] Härdle W.K., Simar L. (2012). *Applied Multivariate Statistical Analysis.* Heidelberg: Springer.

[51] Horváth L., Kokoszka P. (2012). *Inference for Functional Data with Applications.* Heidelberg: Springer.

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

[53] Hron K., Menafoglio A., Templ M., Hrůzová K., Filzmoser P. (2016). Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics and Data Analysis* **94**: 330–350.

[54] Hron K., Templ M., Filzmoser P. (2013). Estimation of a proportion in survey sampling using the logratio approach. *Metrika* **76**: 799–818.

[55] Hrůzová K., Hron K., Rypka M., Fišerová E. (2013). Covariance structure of principal components for three-part compositional data. *Acta Universitatis Palackianae Olomucensis. Facultas Rerum Naturalium. Mathematica* 52 (2): 61–69.

[56] Hrůzová K., Rypka M., Hron K. (2016). Compositional analysis of trade flows structure. *submitted*

[57] 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.

[58] Jackson J.E. (1991). *A User's Guide to Principal Components.* New York: Wiley & Sons.

[59] 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**: 7–14.

[60] Johnson R.A., Wichern D.W. (2002). *Applied Multivariate Statistical Analysis.* London: Prentice Hall, fifth edition.

[61] Jolliffe I.T. (2002). *Principal Component Analysis, 2nd ed.*. New York: Springer.

[62] Jones M.C., Rice J.A. (1992). Displaying the important features of large collections of similar curves. *The American Statistician* **46** (2): 140–145.

[63] 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**: 21–28.

[64] Kneip A., Sickles R.C., Song W. (2012). A new panel data treatment for heterogeneity in time trends. *Econometric Theory* **28**: 590-–628.

[65] Kneip A., Utikal K. (2001). Inference for density families using functional principal component analysis. *Journal of the American Statistical Association* **96**: 519–542.

[66] Kroonenberg P.M. (1983). *Three-mode Principal Component Analysis. Theory and applications.* Leiden: DSWO Press.

[67] Kroonenberg P.M., de Leeuw J. (1980). Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika* **45** (1): 69–97.

[68] Kruskal J.B. (1976). More factors than subjects, test and treatments: An indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika* **41** (3): 281–293.

[69] Kruskal J.B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications* **18** (2): 95–138.

[70] Kynčlová P., Filzmoser P., Hron K. (2016). Compositional biplots including external noncompositional variables. *Statistics* : DOI: 10.1080/02331888.2015.1135155.

[71] Landesbetrieb für Statistik und Kommunikationstechnologie Niedersachsen (2012). *Kreiszahlen: Ausgewählte Regional Daten für Deutschland*. Hannover: Landesbetrieb für Statistik und Kommunikationstechnologie Niedersachsen.

[72] Leurgans S.E., Ross R.T., Abel R.B. (1993). A Decomposition for Three-Way Arrays. *SIAM Journal on Matrix Analysis an Applications* **14** (4): 1064–1083.

[73] Liebl D. (2013). Modeling and forecasting electricity spot prices: A functional data perspective. *Annals of Applied Statistics* **7**: 1562-–1592.

[74] Long J.S. (1997). *Regression Models for Categorical and Limited Dependent Variables.* London: Sage Publications, Inc..

[75] Machalová J., Hron K., Monti G.S. (2015). Preprocessing of centred logratio transformed density functions using smoothing splines. *Journal of Applied Statistics*: DOI: 10.1080/02664763.2015.1103706.

[76] Markovsky I., Van Huffel S. (2007). Overview of total least-squares methods. *Signal Processing* **87**: 2283–2302.

[77] Maronna R.A. (2005). Principal Components and Orthogonal Regression Based on Robust Scales. *Technometrics* **47**: 264–273.

[78] Maronna R., Martin R.D., Yohai V.J. (2006). *Robust Statistics: Theory and Methods.* New York: John Wiley.

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

[80] Mateu-Figueras G., Pawlowsky-Glahn V., Egozcue J.J. (2011). The principle of working on coordinates. In Pawlowsky-Glahn V., Buccianti A. (Eds.), *Compositional Data Analysis: Theory and Applications*, 31–42. Chichester: Wiley.

[81] Menafoglio A., Guadagnini A., Secchi P. (2014). A kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers. *Stochastic Environmental Research and Risk Assessment* **28** (7): 1835–1851.

[82] Miroudot S., Lanz R., Ragoussis A. (2009). Trade in intermediate goods and services. *OECD Trade Policy Working Paper* **93**.

[83] Montgomery D.C., Peck E.A., Vining G.G. (2001). *Introduction to Linear Regression Analysis, 4th ed.*. Hoboken: John Wiley & Sons.

[84] Monti G.S., Migliorati S., Hron K., Hrůzová K., Fišerová E. (2014). Log-ratio approach in curve fitting for concentration-response experiments. *Environmental and Ecological Statistics* **21** (2): 275–295.

[85] Nerini D., Ghattas B. (2007). Classifying densities using functional regression trees: applications in oceanology. *Computational Statistics and Data Analysis* **51**: 4984–4993.

[86] OECD Directorate for Science, Technology and Industry. Division for Economic Analysis and Statistics (2014). *OECD Bilateral Trade Database by Industry and End-use Category.* OECD Publishing.

[87] Pawlowsky-Glahn V., Buccianti A., eds. (2011). *Compositional Data Analysis: Theory and Applications.* Chichester: Wiley.

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

[89] Pawlowsky-Glahn V., Egozcue J.J., Tolosana-Delgado R. (2015). *Modeling and Analysis of Compositional Data.* Chichester: Wiley.

[90] Ramsay J., Silverman B.W. (2002). *Applied Functional Data Analysis: Methods and Case Studies.* New York: Springer-Verlag.

[91] Ramsay J., Silverman B.W. (2005). *Functional Data Analysis, 2nd ed.*. New York: Springer.

[92] Ramsay J.O., Wickham H., Graves S. (2015). *fda: Functional data analysis.* R package version 2.4.4.

[93] R Core Team (2012). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna.

[94] Rousseeuw P., Hubert M. (2013). High-breakdown estimators of multivariate location and scatter. In Becker C., Fried R., Kuhnt S. (eds.), *Robustness and complex data structures.* Heidelberg: Springer, 49–66.

[95] Rousseeuw P.J., Leroy A.M. (1987). *Robust regression and outlier detection.* New York: John Wiley & Sons.

[96] Saikia D. (2009). Agriculture-Industry Interlinkages: Some Theoretical and Methodological Issues in the Indian Context, available at: http://mpra.ub.uni-muenchen.de/27820, last accessed on 1 December 2014.

[97] Saikia D. (2011). Analyzing inter-sectoral linkages in India. *African Journal of Agricultural Research* **6**: 6766–6775.

[98] Salibian-Barrera M., Zamar R.H. (2002). Bootstrapping robust estimates of regression. *Annals of Statistics* **30**: 556–582.

[99] Salibian-Barrera M., Van Aelst S., Willems G. (2006). PCA based on multivariate MM-estimators with fast and robust bootstrap. *Journal of the American Statistical Association* **101**: 1198–1211.

[100] Shang H.L. (2014). A survey of functional principal component analysis. *Advances in Statistical Analysis* **98**: 121–142.

[101] Templ M., Hron K., Filzmoser P. (2011). *robCompositions: an R-package for robust statistical analysis of compositional data.*

[102] UN (2012). *World Economic Situation and Prospects.* New York: United Nations.

[103] United Nations Industrial Development Organization Vienna (2013). *International yearbook of industrial statistics 2013.* Cheltenham: Edward Elgar Publishing.

[104] Van Aelst S., Willems G. (2013). Fast and robust bootstrap for multivariate inference: The R package FRB. *Journal of Statistical Software* **53** (3).

[105] Venables W.N., Ripley B.D. (2002). *Modern Applied Statistics with S.* New York: Springer.

[106] Vighi M., Migliorati S., Monti G.S. (2009). Toxicity on the luminescent bacterium Vibrio fischeri (Beijerinck). I: QSAR equation for narcotics and polar narcotics. *Ecotoxicology and Environmental Safety* **72**: 154–161.

[107] Wierts P., Van Kerkhoff H., De Haan J. (2014). Composition of exports and export performance of Eurozone countries. *JCMS: Journal of Common Market Studies* **52** (4): 928–941.

[108] Zhang Z., Müller H.G. (2011). Functional density synchronization. *Computational Statistics and Data Analysis* **55**: 2234–2249.

[109] Zhu S., Yamano N., Cimper A. (2011). Compilation of bilateral trade database by industry and end-use category. *OECD Science, Technology and Industry Working Papers*, OECD Publishing.

PALACKÝ UNIVERSITY OLOMOUC
FACULTY OF SCIENCE

DISSERTATION THESIS SUMMARY

Economic applications of statistical analysis of compositional data

Department of Mathematical Analysis and Applications of Mathematics
Supervisor: **Doc. RNDr. Karel Hron, Ph.D.**
Author: **Mgr. Klára Hrůzová**
Study programme: P1104 Applied Mathematics
Field of study: Applied Mathematics
Form of study: full-time
The year of submission: 2016

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. Klára Hrůzová**
Department of Mathematical Analysis and Applications
of Mathematics
Faculty of Science
Palacký University Olomouc

Supervisor: **Doc. RNDr. Karel Hron, Ph.D.**
Department of Mathematical Analysis and Applications
of Mathematics
Faculty of Science
Palacký University Olomouc

Reviewers: **Prof. Dr. Vera Pawlowsky-Glahn**
Department of Computer Science, Applied Mathematics and Statistics
University of Girona

Doc. Matthias Templ, Ph.D.
Department of Statistics and Probability Theory
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.

Contents

1. Abstract

Logratio analysis of compositional data, multivariate observations carrying relative information, is nowadays widely used in nature sciences, such as geology or chemistry, however, it is not widespread in social sciences like economy, psychology, etc. The thesis deals with adaptations of known statistical methods for compositional data with economic applications. It reveals that by taking the relative nature of data into account the models provide relevant results. Besides the dimension reduction methods (principal component analysis, PARAFAC), the thesis particularly includes the regression analysis which is very popular in economic applications. Within regression analysis, the thesis mainly deals with the situation where both the dependent and independent variables are compositional, especially when the regression between the parts of a composition is considered. In such a case, orthogonal regression, a kind of errors-in-variable models, needs to be applied for parameter estimation instead of ordinary least squares method. Finally, functional analogy to principal component analysis is applied for the density functions, i.e. functional compositions.

Key words: compositional data; principal component analysis; linear regression; orthogonal regression; functional data; density functions

2. Abstrakt v českém jazyce

Logratio analýza kompozičních dat, mnohorozměrných pozorování nesoucích relativní informaci, je již hojně využívána v přírodních vědních disciplínách, jako je geologie nebo chemie, avšak ve vědách společenských - ekonomie, psychologie a další, ještě není příliš známá. Tato práce se zabývá adaptací známých statistických metod pro kompoziční data s ekonomickými aplikacemi. Ukazuje se, že pokud se bere v úvahu relativní charakter dat, modely poskytují relevantní výsledky. Práce obsahuje kromě metod pro redukci dimenze (metoda hlavních komponent, PARAFAC) zejména regresní analýzu, která je v ekonomických aplikacích velmi oblíbená. V jejím rámci se pak zabývá zejména situací, kdy je kompoziční závisle i nezávisle proměnná, speciálně když regresi uvažujeme mezi složkami kompozice. V takovém případě je potřeba použít pro odhady parametrů ortogonální regresi, což je typ regrese s chybami v proměnných, namísto obvyklé metody nejmenších čtverců. Nakonec práce popisuje funkcionální obdodu metody hlavních komponent, která je aplikována na hustoty, neboli funkcionální kompozice.

Klíčová slova: kompoziční data; metoda hlavních komponent; regresní analýza; ortogonální regrese; funkcionální data; hustoty

3. Introduction

Compositional data (or compositions for short) are known as column vectors with positive components that carry relative information, in other words, the only relevant information is contained in ratios between components [1]. Mostly, compositions sum to a constant, like 1 in case of proportions or 100 for percentages, however, it is just a proper representation in the equivalence class of proportional vectors, forming the sample space of compositional data. Accordingly, possible choice of constant sum constraint should not influence results of statistical analysis due to scale invariance property of compositions [46, 48].

The standard Euclidean geometry, defined in real space is not appropriate for compositional data. It is caused by relative character of compositions, since Euclidean geometry deals with absolute values of components [48]. Hence, the Aitchison geometry with Euclidean vector space properties was developed which captures the relative nature of compositions [3, 47].

Nevertheless, almost all statistical methods rely on the Euclidean geometry in real space [15]. Accordingly, it is not appropriate to apply them directly to compositions. Instead, the logratio methodology [1, 19, 48] is used to express compositional data in real space using appropriate coordinates and, if necessary, to transform the results back to the original sample space [42, 46]. It is of particular importance to choose such coordinates that lead to interpretable and meaningful results.

The analysis of compositional data is nowadays popular in fields such as geology or chemometrics [8, 46], however, in social sciences like economy, psychology or sociology, compositional data are not widespread yet. Up to rare applications of the logratio methodology in economics [2, 27], despite of compositional nature of data [4, 14, 55], the analysis does not reflect this fact. Therefore, this thesis is aimed to present popular statistical tools adapted to compositional data and applied to economic data.

4. Recent state summary

4.1. Compositional data

Compositional data [1,48] are strictly positive multivariate observations that carry only relative information. By the relative information it is meant that absolute values are no longer important for the analysis, instead, ratios between parts of a composition capture the only relevant information. The sample space of representations of compositional data within the equivalence class of proportional vectors is simplex [1,46,48], which is defined as a set of strictly positive real numbers that sum up to a constant,

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

where κ is any positive real number, e.g. 1 in case of proportions or 100 for percentages.

Properties of compositional data that distinguish compositional data from standard multivariate observations can be formalized by principles of compositional data analysis [16,48]. Among them, *scale invariance* property and *subcompositional coherence* seem to be the most important when analyzing compositional data. The first one means that the information conveyed by a composition does not depend on the units in which a composition is measured, i.e. characteristics of compositions should be invariant under a change of scale. According to the second one, the information contained in a composition of D parts should not be in a conflict with that coming from a subcomposition containing d parts, where $d \leq D$. The last principle is called *permutation invariance* - reordering parts of a composition does not affect the included information.

The Aitchison geometry with Euclidean vector space structure follows closely the above stated principles of compositional data analysis [46]. Basic operations substituting sum of two real vectors and multiplication of a vector by a scalar are called perturbation and power transformation, respectively. Their definition for $\mathbf{x} \in \mathcal{S}^D$, $\mathbf{y} \in \mathcal{S}^D$ and $\alpha \in \mathbb{R}$ follows,

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 \cdot y_1, \ldots, x_D \cdot y_D)', \ \alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, \ldots, x_D^\alpha)'.$$

The triple $(\mathcal{S}^D, \oplus, \odot)$ forms vector space structure [48] and to obtain Euclidean vector space, inner product and the corresponding norm and distance are defined as well:

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

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

Since almost all standard statistical methods are defined in real space, it is not appropriate to apply them directly to compositions. In order to perform statistical processing using standard multivariate tools, it is necessary to express compositions first in proper real coordinates [48].

One of the possibilities are centred logratio (clr) coordinates that are coordinates with respect to generating system of simplex. They are defined as

$$\mathrm{clr}(\mathbf{x}) = \left(\ln \frac{x_1}{g(\mathbf{x})}, \ldots, \ln \frac{x_D}{g(\mathbf{x})} \right)', \tag{2}$$

where $g(\mathbf{x})$ stands for geometric mean of $\mathbf{x}$. Although the clr coordinates are symmetric in the components, the sum of the coefficients is zero and this leads to singular covariance matrix. Nevertheless, they are still used in the practice because they translate operations and metrics from the simplex endowed with the Aitchison geometry into real space. Particularly, for compositions $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ and real constants α, β it holds that

$$\mathrm{clr}(\alpha \odot \mathbf{x} \oplus \beta \odot \mathbf{y}) = \alpha \cdot \mathrm{clr}(\mathbf{x}) + \alpha \cdot \mathrm{clr}(\mathbf{y});$$

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle \mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{y}) \rangle;$$

$$\|\mathbf{x}\|_a = \|\mathrm{clr}(\mathbf{x})\|; \;\; d_a(\mathbf{x}, \mathbf{y}) = d(\mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{y})).$$

Since the Aitchison geometry has dimension one less than the number of components $(D-1)$, the clr coefficients are not coordinates with respect to a basis of the simplex.

At the early stage of the logratio methodology, the additive logratio (alr) coordinates [1] were used as well. In this case, each part of the composition is divided by one chosen part, e.g. the last part x_D, to obtain the respective logratio. This leads to a vector of alr coordinates which is of dimension $D-1$:

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

It is obvious that alr coordinates are not symmetrical in the components and, unlike clr coordinates, they do not preserve distances. Thus they can be used only for modeling purposes. The reason is that alr coordinates do not correspond to an orthonormal basis of the simplex.

To avoid singular covariance matrix, coordinates with respect to an orthonormal basis of the simplex could be another option. With a particular choice of the orthonormal basis, the composition $\mathbf{x} \in \mathcal{S}^D$ can be written as

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

where $\mathbf{x}^* = (x_1^*, \ldots, x_{D-1}^*)'$ is the vector of coordinates of $\mathbf{x}$ with respect to this basis. The resulting coordinates are called isometric logratio (ilr) coordinates [19]. The corresponding mapping is isometric isomorphism between $\mathcal{S}^D$ and $\mathbb{R}^{D-1}$ and thus it preserves distances and translates operations similarly as for the clr coordinates.

The ilr coordinates that are used in the thesis [32, 35] are defined as a set of D orthonormal coordinate systems, namely $\mathbf{z}^{(l)} = (z_1^{(l)}, \ldots, z_{D-1}^{(l)})'$, $l = 1, \ldots, D$,

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}}\,\ln\frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}},\ \ i = 1, \ldots, D-1. \tag{3}$$

Here $(x_1^{(l)}, x_2^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})'$ stands for such a permutation of the parts $(x_1, \ldots, x_D)'$, that always the l-th compositional part fills the first position, $(x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)'$. In such a configuration, the first ilr variable $z_1^{(l)}$ explains all the relative information (logratios) about the original compositional part x_l (it is nothing else than a scaled aggregation of all logratios with x_l), the coordinates $z_2^{(l)}, \dots, z_{D-1}^{(l)}$ then explain the remaining logratios in the composition [24]. Note that the only important position is that of $x_1^{(l)}$ (that is interpretable through $z_1^{(l)}$), the other parts can be chosen arbitrarily, because different ilr coordinates are orthogonal rotations of each other [19]. Of course, $z_1^{(l)}$ cannot be identified with compositional part x_l, as the other parts are also naturally involved through the corresponding logratios. Its interpretation is thus limited due to the specific structure of the Aitchison geometry. We can also see that this coordinate is formed by a logratio between the part x_l and an "average part", resulting from the geometric mean of the remaining parts in the composition. Therefore, values of $z_1^{(l)}$ represent a measure of dominance of the part x_l with respect to the other parts.

4.2. Economic applications of compositional data

Although there is a big potential in economic applications of compositional data analysis, logratio methodology has not been widespread in this field yet. Up to now, just few papers with an economic application are available, e.g. [2,7,26,27,39]. For example, in [27] the application of compositional data analysis to consumer demand systems is described. Alternative approaches are mentioned as well and possible applications of the logratio methodology in economics are proposed there.

The compositional VARIMA time series were introduced in [2]. For an example of expenditure shares in the UK was shown that the model does not depend on the chosen logratio coordinates, however, the clr coordinates should be avoided due to singular covariance matrix. Accordingly, the choice of coordinates depends mainly on the interpretability of the resulting model.

5. Thesis objectives

The main purpose of the thesis is to adapt popular statistical tools within the logratio methodology and to develop new methods that are of primary interest in economic applications of compositional data. When the (relative) structure of a dataset is of main interest, dimension reduction methods need to be applied. Among them, principal component analysis is one of the most popular one and by considering three-mode data, parallel factor analysis (PARAFAC) as well. On the other hand, when relations between variables are analyzed, which is very common in economics, the regression analysis is applied. And finally, when mass data are collected, that form (approximately continuous) density functions, functional data analysis tools are applied by considering the relative nature of densities using the Bayes space methodology. When the interest is devoted to dimension reduction of density functions, an adapted version of functional principal component analysis is used.

6. Theoretical framework and applied methods

6.1. Dimension reduction methods

In economic world, one can be interested in analyzing the multivariate structure of a dataset. One of the most popular methods for this purpose is principal component analysis (PCA). It leads to dimension reduction based on linear combination of the original data which depletes most of the variability (for more see, e.g. [30, 37]). For compositional data, clr coordinates (2) are used instead of the original data [23].

The results of PCA can be displayed in biplot [29] which is a scatterplot of the first two principal components, where scores are displayed as points and loadings as rays. However, when the PCA is applied to centred data in clr coordinates, the interpretation needs to be adjusted [38,48]. The basic terms are ray, which joins the origin to a vertex $\mathbf{h}_j$ (formed by the first two components of the respective loading vector), and link, which joins two vertices $\mathbf{h}_j$ and $\mathbf{h}_k$. Links and rays provide information about the relative variability in a compositional dataset: length of a link between $\mathbf{h}_j$ and $\mathbf{h}_k$ approximates standard deviation of the logratio between j-th and k-th compositional parts and length of a ray approximates standard deviation of the respective clr coefficient. Consequently, if the vertices coincide, then the variance of corresponding logratio is approximately zero and this means that the corresponding two parts are proportional. Links also provide information about correlation of two pairwise logratios: suppose two links $\overline{jk}$ and $\overline{il}$ intersect in M, then

$$\cos(jMi) \approx \mathrm{corr}\left(\ln\frac{x_j}{x_k}, \ln\frac{x_i}{x_l}\right).$$

For three-part composition, $\mathbf{x} = (x_1, x_2, x_3)'$, we can deeply analyze the variance structure of the principal components and its impact to interpretation of the resulting orthonormal coordinates (3). Although in case of standard real data the covariance structure of principal components can be also expressed using elements of the original covariance matrix [36], we will follow an alternative way of its derivation that enables a deeper insight into covariance structure of three-part compositional data. The covariance structure is described in the following theorem, for more information see [34].

Theorem 1. *The covariance structure of principal components (orthonormal coordinates) z_1^*, z_2^* for three-part composition $\mathbf{x} = (x_1, x_2, x_3)'$ can be expressed as*

$$\begin{aligned} \mathrm{var}(z_1^*) &= \frac{a+b+c}{6} + \frac{\sqrt{(a-b)^2+(b-c)^2+(c-a)^2}}{3\sqrt{2}}, \\ \mathrm{var}(z_2^*) &= \frac{a+b+c}{6} - \frac{\sqrt{(a-b)^2+(b-c)^2+(c-a)^2}}{3\sqrt{2}}, \end{aligned} \tag{4}$$

where $a,\ b,\ c$ correspond to $\mathrm{var}\left(\ln\frac{x_1}{x_2}\right),\ \mathrm{var}\left(\ln\frac{x_1}{x_3}\right),\ \mathrm{var}\left(\ln\frac{x_2}{x_3}\right)$*, respectively.*

Note that big differences between variances of logratios contribute for maximization of the first principal component at the expense of the second one. This is obvious from the second part of (4) - in variance of z_1^* we add square root of the sum of squared differences of these variances while in $\mathrm{var}(z_2^*)$ we subtract it. Furthermore, it is not necessary to consider the covariance because principal components are uncorrelated [30]. The above theorem was applied to interpret principal components of gross value added compositions in German regions [34].

6.2. Linear regression analysis

Linear regression is a very popular statistical tool in economic world. When we are dealing with compositional data, four main cases might occur. The first two deal with either compositional response or explanatory variables; they are described, e.g., in [17,32,48]. The third case joins the first two together which means that we work with both compositional response and explanatory variables [44]. The last case considers analyzing the relation between parts of a composition [35] - this will be described in a more detail here.

Most of the economic indicators, such as gross domestic product, value added, export, import, etc., consist of many variables. For example GDP, in the income approach, is computed as a sum of compensation of employees, gross operating surplus, gross mixed income and taxes less subsidies on production and imports. Generally, we are interested in analyzing GDP, but we can be also interested in analyzing the relation between the variables that form the GDP composition. For this purpose, orthogonal regression in proper coordinates seems to be the preferable option [35].

A particular challenge for the choice of coordinates comes from the fact that at least two parts in the composition are of simultaneous interest, the response part and covariate part(s). Consequently, the question arises, how to use coordinates (3) for the case of regression of one of the compositional parts to the remaining parts. In order to analyze the influence of a single compositional part on the explanatory variables, D multiple regression models according to the coordinate representations (3) were constructed. Let x_l plays the role of the response variable that should be represented by a coordinate as well. Since the main task is to analyze the influence of the other parts on x_l, it seems reasonable that also the corresponding coordinate will contain information on the relation of x_l to all remaining parts in the composition. Thus, in the notation of (3), $z_1^{(l)}$ plays the role of such a coordinate. Consequently, we can proceed with the coordinate representation of the explanatory subcomposition $(x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)'$. For this purpose, the coordinates $z_2^{(k)}, \ldots, z_{D-1}^{(k)}$ according to the reordered subcomposition $(x_k, x_2, \ldots, x_i, \ldots, x_D)'$, $i \neq \{k, l\}$, $k = 2, \ldots, D$, can be used. Similarly as before, the coordinate $z_2^{(k)}$ explains all the relative information about part x_k in the resulting subcomposition. Considering the range of k, we arrive at $D-1$ regression models

$$z_1^{(l)} = \beta_0^{(k)} + \beta_1^{(k)} z_2^{(k)} + \ldots + \beta_{D-2}^{(k)} z_{D-1}^{(k)} + \varepsilon \tag{5}$$

(in theoretical form, ε stands for an error term), assigned to single explanatory compositional parts. The interpretation of these models is similar to the case of regression with compositional covariates [32], i.e. in each model just the absolute term parameter and the parameter corresponding to the coordinate $z_2^{(k)}$ are used for further interpretation and to perform statistical inference (confidence intervals, hypotheses testing).

Since both response and explanatory variables arise from one composition, it cannot be assumed that the covariates represent errorless variables like in the case of a real-valued response [32]. Consequently, the use of an ordinary multiple regression model is inappropriate and can even lead to biased results. Therefore, we apply an orthogonal regression model (or, equivalently, a total least squares model) for this purpose, which is a specific type of errors-in-variable (EIV) model [28].

The estimation of regression parameters will be described (following the geometrical motivation) for the case of the four-part composition, $\mathbf{x} = (x_1, x_2, x_3, x_4)'$, where x_1 was chosen for the response and the other parts form explanatory variables. For this purpose, we assume to have a random vector $\mathbf{z} = (z_1, z_2, z_3)'$ (an orthonormal coordinate representation of the composition following (3), where $z_1 \equiv z_1^{(1)}$ and $z_i \equiv z_i^{(2)}$ for $i = 2, 3$) and the task is to find a relationship between the response variable z_1 and the covariates z_2, z_3, expressed in the form $z_1 = \beta_0 + \beta_1 z_2 + \beta_2 z_3 + \varepsilon$, with the regression parameters $\beta_0, \beta_1, \beta_2$.

From the geometrical point of view the basic idea is to fit a plane to the data using PCA. The loadings of the first two principal components define a basis of the plane. As the third principal component is orthogonal to the previous ones, its loadings define the normal vector to the plane, $\mathbf{n} = (n_1, n_2, n_3)'$. The plane passes through the point $\mathbf{t}$, representing the location estimate of the corresponding $n \times 3$ data matrix $\mathbf{Z}$ (the arithmetic mean in the classical case), and its perpendicular distance from the origin is $\mathbf{t}'\mathbf{n}$. The perpendicular distance from each point in $\mathbf{Z}$ to the plane (the norm of the residuals) is the inner product of each centered point and the normal vector to the plane. The fitted plane minimizes the sum of squared errors.

Consequently, the estimated regression parameters are obtained using the elements of the normal vector, namely

$$b_0 = \frac{\mathbf{t}'\mathbf{n}}{n_3},\ b_1 = -\frac{n_1}{n_3},\ b_2 = -\frac{n_2}{n_3}.$$

We can also consider the general case, where a vector of orthonormal coordinates has $D-1$ components, $\mathbf{z} = (z_1, z_2, \ldots, z_{D-1})'$. As in the previous case, the response variable is z_1 and covariates $z_2, \ldots, z_{D-1}$. Then the regression relation is expressed in the form $z_1 = \beta_0 + \beta_1 z_2 + \beta_2 z_3 + \cdots + \beta_{D-2} z_{D-1} + \varepsilon$ for a vector of regression parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_{D-2})'$. Denote the loading vector corresponding to the smallest eigenvalue as $\mathbf{n} = (n_1, n_2, \ldots, n_{D-1})'$. Then the estimated parameters $\mathbf{b}$ are obtained using values of the loading vector as follows,

$$b_0 = \frac{\mathbf{t}'\mathbf{n}}{n_{D-1}},\ b_1 = -\frac{n_1}{n_{D-1}},\ b_2 = -\frac{n_2}{n_{D-1}}, \ldots,\ b_{D-2} = -\frac{n_{D-2}}{n_{D-1}},$$

where $\mathbf{t}$ is the mean vector of $\mathbf{Z}$.

In order to support the interpretation of the outcome of orthogonal regression, it is desirable to obtain confidence intervals for the regression parameters, and p-values for tests about these parameters. This statistical inference is only possible with strict distributional assumptions, a better strategy is to derive the inference by resampling methods. In order to relax the assumptions about the distribution of the input data, the nonparametric bootstrap [11, 25] can be applied.

Regression estimators which are based on classical SVD or PCA are sensitive to outliers that naturally occur in most real-world data sets. Therefore, we also considered a robust version of the orthogonal regression. Although robust versions of SVD are available (e.g. [10]), it is simpler and computationally more attractive to use robust PCA, which is obtained through a robust estimation of the covariance matrix (e.g. [22]).

Among other possibilities like [9,21,41], MM-estimators [51] are employed, because they are highly efficient when the errors have a normal distribution, their breakdown point is 0.5 and they have a bounded influence function.

The available theory for robust estimators is limited to asymptotic results. Although bootstrap is a very useful tool, in case of robust estimators there are two problems: computational complexity of robust estimators and the instability of the bootstrap in case of outliers. Thus we used fast and robust bootstrap [52,54] which is based on the fact that the robust estimators (concretely S- and MM-estimators) can be represented by smooth fixed point equations which allow to calculate only a fast approximation of the estimates in each bootstrap sample.

In [35], the above model was used to analyze the relation between manufacturing and other parts of gross value added.

6.3. Functional principal component analysis applied on density functions

Nowadays, an increasing number of studies are based on complex data, such as curves, surfaces or images. As a direct consequence, the importance of functional data analysis (FDA), e.g. [49], has recently strongly increased. In recent years, a large body of literature has been developed in this field, e.g. [31,50], however, still little attention has been paid to the problem of dealing with functional data that are probability density functions [12,13,43,45,56]. Even though it might seem that density functions are just a special case of functional data – with a constant-integral constraint equal to one – standard FDA methods appear to be inappropriate for their statistical treatment, as they do not consider the particular constrained nature of the data. In this context, probability density functions have recently been interpreted as functional compositional data, i.e., functional data carrying only relative information. To handle this kind of data, the Aitchison geometry has been extended to the so called Bayes spaces: a Hilbert space structure for σ-finite measures, including probability measures, has been worked out in [6], based on the pioneering work of [18] and the subsequent developments of [5] and [20].

We call $\mathcal{B}^2(I)$ the Bayes space of (equivalence classes of) positive functional compositions f on I with square-integrable logarithm. In particular, we here consider continuous (hence bounded) functional compositions on the compact support $I = [a, b]$. Given two absolutely integrable density functions $f, g \in \mathcal{B}^2(I)$ and a real number $\alpha \in \mathbb{R}$ we indicate with $f \oplus g$ and $\alpha \odot f$ the perturbation and powering operations, respectively, defined as [6,18]:

$$(f \oplus g)(t) = \frac{f(t)g(t)}{\int_I f(s)g(s)\,\mathrm{d}s}, \quad (\alpha \odot f)(t) = \frac{f(t)^\alpha}{\int_I f(s)^\alpha\,\mathrm{d}s}, \quad t \in I.$$

The resulting functions are readily seen to be probability density functions. [18] prove that $\mathcal{B}^2(I)$ endowed with the operations $(\oplus, \odot)$ is a vector space. Note that the neutral elements of perturbation and powering are $e(t) = 1/\eta$, with $\eta = b - a$ (i.e., the uniform density), and 1, respectively. Moreover, the difference between two elements $f, g \in \mathcal{B}^2(I)$, denoted by $f \ominus g$, is obtained as perturbation of f with the reciprocal of g, i.e., $(f \ominus g)(t) = (f \oplus [(-1) \odot g])(t), t \in I$. To endow $\mathcal{B}^2(I)$ with a Hilbert space structure, [18] define the inner product

$$\langle f, g \rangle_{\mathcal{B}} = \frac{1}{2\eta} \int_I \int_I \ln \frac{f(t)}{f(s)} \ln \frac{g(t)}{g(s)}\,\mathrm{d}t\,\mathrm{d}s, \quad f, g \in \mathcal{B}^2(I), \tag{6}$$

with $\eta = b - a$, which induces the following norm,

$$||f||_{\mathcal{B}} = \left[\frac{1}{2\eta} \int_I \int_I \ln^2 \frac{f(t)}{f(s)}\,\mathrm{d}t\,\mathrm{d}s\right]^{1/2}.$$

The space $\mathcal{B}^2(I)$, endowed with the inner product (6), is proved to be a separable Hilbert space in [18].

As functional compositions, probability density functions are featured by specific properties, such as the scale invariance and relative scale properties. The latter would be neglected, if one applied the functional principal component analysis (FPCA) to density functions; for more see [49, Chapter 8], [31, Chapter 3] and [53]. The simplicial version of FPCA, named SFPCA, is derived by following the Bayes space methodology.

Let $\widetilde{X}_1, ..., \widetilde{X}_N$ be a sample in $\mathcal{B}^2(I)$, and denote by $X_1, ..., X_N$ the corresponding centred sample, i.e., for $i = 1, ..., N$, $X_i = \widetilde{X}_i \ominus \overline{X}$, where $\overline{X}$ stands for the sample mean $\overline{X} = \frac{1}{N} \odot \bigoplus_{i=1}^{N} \widetilde{X}_i$. We consider the problem of finding the simplicial functional principal components (SFPCs) in $\mathcal{B}^2(I)$, i.e., the elements $\{\zeta_j\}_{j\geq 1}$, $\zeta_j \in \mathcal{B}^2(I)$, maximizing the following objective function over $\zeta \in \mathcal{B}^2(I)$:

$$\frac{1}{N} \sum_{i=1}^{N} \langle X_i, \zeta \rangle_{\mathcal{B}}^2 \text{ subject to } \|\zeta\|_{\mathcal{B}} = 1; \; \langle \zeta, \zeta_k \rangle_{\mathcal{B}} = 0, \; k < j, \tag{7}$$

where the orthogonality condition $\langle \zeta, \zeta_k \rangle_{\mathcal{B}} = 0$, for $k < j$, holds only for $j \geq 2$.

Because $\mathcal{B}^2(I)$ is a separable Hilbert space, the minimization problem (7) is well posed [31, Theorem 3.2, p. 38]. Thus, the solution of (7) exists and is unique. Accordingly, the j-th SFPC solves the eigenvalue equation

$$V\zeta_j = \delta_j \odot \zeta_j, \tag{8}$$

(δ_j, ζ_j) being the j-th eigenpairs of the sample covariance operator $V : \mathcal{B}^2(I) \to \mathcal{B}^2(I)$, acting on $x \in \mathcal{B}^2(I)$ as

$$Vx = \frac{1}{N} \odot \bigoplus_{i=1}^{N} \langle X_i, x \rangle_B \odot X_i.$$

In order to proceed with (7) in practice, i.e. to express densities in the standard L^2 space, we apply the isometric isomorphism between $\mathcal{B}^2(I)$ and $L^2(I)$ defined by the clr-transform

$$\operatorname{clr}(f)(t) = f_c(t) = \ln f(t) - \frac{1}{\eta} \int_I \ln f(s) \, \mathrm{d}s. \tag{9}$$

However, note that, by construction, the constraint

$$\int_I \operatorname{clr}(f)(t) \mathrm{d}t = \int_I \ln f(t) \, \mathrm{d}t - \int_I \frac{1}{\eta} \int_I \ln f(s) \, \mathrm{d}s \, \mathrm{d}t = 0$$

occurs. This additional condition needs to be taken into account for computation and analysis on clr-transformed density functions.

The clr-transform allows to rewrite the original problem (7) as a maximization of the term

$$\frac{1}{N}\sum_{i=1}^{N}\langle \mathrm{clr}(X_i), \mathrm{clr}(\zeta)\rangle_2^2 \text{ subject to } \|\mathrm{clr}(\zeta)\|_2 = 1;\ \langle \mathrm{clr}(\zeta), \mathrm{clr}(\zeta_k)\rangle_2 = 0,\ k < j$$

over $\zeta \in \mathcal{B}^2(I)$. Accordingly, for $j \geq 1$ the maximization problem (7) can be equivalently restated as finding $\nu \in L^2$ which maximizes

$$\frac{1}{N}\sum_{i=1}^{N}\langle \mathrm{clr}(X_i), \nu\rangle_2^2 \text{ subject to } \|\nu\|_2 = 1;\ \langle \nu, \nu_k\rangle_2 = 0,\ k < j;\ \int_I \nu = 0, \tag{10}$$

where the orthogonality constraint is meaningful only for $j \geq 2$ and the zero-integral constraint incorporates the corresponding clr-transform property.

We now show that (10) is solved by the eigenfunctions $\{\xi_j\}_{j\geq 1}$ of the sample covariance operator $V_{\mathrm{clr}} : L^2(I) \to L^2(I)$ of the transformed sample $\mathrm{clr}(X_1), ..., \mathrm{clr}(X_N)$, acting on $x \in L^2(I)$ as

$$V_{\mathrm{clr}}x = \frac{1}{N}\sum_{i=1}^{N}\langle \mathrm{clr}(X_i), x\rangle_2 \, \mathrm{clr}(X_i).$$

We first notice that the eigenfunctions $\{\xi_j\}_{j\geq 1}$ would have solved problem (10), if it had been stated without the zero-integral condition $\int_I \nu = 0$. Therefore, to prove that $\nu = \xi_j$ maximizes (10) it suffices to show that ξ_j fulfills also the constraint $\int_I \xi_j = 0$, for all $j \geq 1$. To this end, we note that the zero-integral property of the clr-transformed sample $\mathrm{clr}(X_1), ..., \mathrm{clr}(X_N)$ implies that V_{clr} admits a zero eigenvalue with associated eigenfunction $\xi_0 \equiv 1/\sqrt{b-a}$:

$$V_{\mathrm{clr}}\,\xi_0 = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sqrt{b-a}}\left[\int_I \mathrm{clr}(X_i)\right] \mathrm{clr}(X_i) \equiv 0.$$

Since the eigenfunctions $\{\xi_j\}$ corresponding to the remaining non-null eigenvalues $\{\rho_j\}$ are to be orthogonal to the eigenfunction ξ_0, the ξ_j's need to satisfy the zero-integral condition $\int_I \xi_j = 0$, as $\langle \xi_j, \xi_0\rangle_2 = 1/\sqrt{b-a}\int_I \xi_j$. Another way to see this is to notice that

(a) the image of the sample covariance operator V_{clr} is the span of the clr-transformed observations, (and the constant function $\xi_0 \equiv \frac{1}{\sqrt{b-a}}$ belongs to its kernel), and (b) the eigenfunctions corresponding to the non-null eigenvalues form a basis of the image of V_{clr}. As such, each eigenfunction ξ_j can be written as a unique linear combination of the functions $\text{clr}(X_1), ..., \text{clr}(X_N)$. Therefore, the zero-integral condition is fulfilled since it holds, by construction, for each of the functions $\text{clr}(X_i)$, $i = 1, ..., N$. Thus, problem (7) can be restated in terms of clr-transforms as (10) and the SFPCs can be obtained by transforming the eigenfunctions $\{\xi_j\}_{j\geq 1}$ associated to the non-null eigenvalues $\{\rho_j\}_{j\geq 1}$ of V_{clr} through the inverse of the function clr, namely $\zeta_j = \text{clr}^{-1}(\xi_j) =_{\mathcal{B}} \exp(\xi_j)$, with $j \geq 1$. Note that, as in classical PCA, the eigenfunctions ξ_j are determined up to sign changes. Accordingly, the SFPCs are determined up to a powering by ±1 (i.e., if ζ_j solves problem (7), $-1 \odot \zeta_j$ is a solution as well).

To compute the eigenfunctions ξ_j we resort to a method based on a B-spline basis expansion. Following [40], we consider for $\text{clr}(X_1), ..., \text{clr}(X_N)$ and ξ_j, $j \geq 1$, a B-spline basis fulfilling the zero-integral constraint,

$$\text{clr}(X_i)(\cdot) = \sum_{k=1}^{K} c_{ik}\phi_k(\cdot), \quad \xi_j(\cdot) = \sum_{k=1}^{K} a_{jk}\phi_k(\cdot). \tag{11}$$

To compute the B-spline coefficients the usual parametrization of smoothing splines applies, and the additional constraint is incorporated in the estimation algorithm as described in [40]. Hence, $\mathbf{a}_j = (a_{jk})$ is obtained as solution of the eigenproblem

$$N^{-1}\mathbf{C}'\mathbf{C}\mathbf{M}\mathbf{a}_j = \rho_i \mathbf{a}_j,$$

where $\mathbf{C} = c_{ik}$. With analogue orthogonality arguments as those previously introduced, the zero integral constraint is inherently kept in the PCA algorithm, and thus does not need to be explicitly imposed.

The above introduced methodology was applied to dimension reduction of salary distributions in Austria regions.

7. Original results and summary

The main goal of this thesis was to show how the compositional analysis can be useful and meaningful in economic applications. Although not for all papers that form the base of the thesis we finally succeeded to include an economic application, it was shown that the logratio methodology has a great potential also in this field.

The dissertation thesis contributes to three basic methods useful for the statistical analysis of economic data. The first part consists of dimension reduction methods. Here, the covariance structure of principal components for three-part compositions brings a deeper insight into variance structure of orthonormal coordinates. Other extension consists in application of principal component analysis and PARAFAC to a compositional dataset of trade flows.

The second part of the thesis aims to regression analysis, where two basic cases (regression with compositional either response or explanatory variables) are extended by the case, where both the response and covariates are compositional [44] and by analyzing the relation between parts of a composition [35]. In the latter regression model, the first important task was how to define interpretable orthonormal coordinates, since one part forms the response and other parts of the same composition form explanatory variables. Second important task was to define such a regression model which is able to deal with measurement errors contained also in the independent variables. Accordingly, the orthogonal regression based on the outputs of principal component analysis was applied. In order to obtain statistical inference, to support the regression results, nonparametric boots̃trap was used. Furthermore, the classical regression models are sensitive to outlying observations, thus the robust counterpart was developed and fast and robust bootstrap was applied to obtain the statistical inference also for the robust parameter estimates. Both classical and robust regression models were applied to an economic example.

The final part of the thesis deals with functional compositions - density functions. Here reformulation of the standard functional principal component analysis (FPCA) led to development of simplicial functional principal component analysis, where the FPCA is adapted to the space of clr transformed densities. Again, the theoretical considerations are applied to an economic example - salary distributions in regions of Austria.

The most difficult part of this thesis was to present different methods and their economic applications in a coincise form. The hope is that the thesis could serve as a list of (necessarily incomplete) options how to deal with economic data that have compositional character. There are, of course, further important methods in economic context, not listed here, among them for example compositional time series. However, due to volatility of financial time series, that form one potential economic application, we are faced with multivariate ARCH and GARCH models that belong to the most difficult ones even in their univariate cases. Hence, this is one of many other possibilities how to further develop this topic in the future.

List of publications

Research papers

- **Hrůzová K.**, Hron K., Rypka M., Fišerová E. (2013). Covariance structure of principal components for three-part compositional data. *Acta Universitatis Palackianae Olomucensis, Facultas Rerum Naturalium. Mathematica* **52** (2), 61–69.

- Monti G.S., Migliorati S., Hron K., **Hrůzová K.**, Fišerová E (2014). Log-ratio approach in curve fitting for concentration-response experiments. *Environmental and Ecological Statistics* **22** (2), 275–295.

- Hron K., Menafoglio A., Templ M., **Hrůzová K.**, Filzmoser P. (2016). Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics and Data Analysis* **94**, 330–350.

- **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.

- **Hrůzová K.**, Rypka M., Hron K. (2016). Compositional analysis of trade flows structure. *submitted*

Proceedings papers

- **Hrůzová K.**, Hron K. (2013). Compositional data analysis of German national accounts. In Hron K., Filzmoser P., Templ M. (eds.) *Proceedings of The 5th International Workshop on Compositional Data*, 64–70.

- **Hrůzová K.**, Hron K., Filzmoser P., Todorov V. (2014). Regression among parts of compositional data with an economic application. *MME 2014 Conference proceedings*, 337–342. Olomouc: Palacký University.

List of conferences

- ROBUST, 9.-14.9.2012, Němčičky (CZ): Bilance a bilanční dendrogram kompozičních dat (poster, in Czech)
- CoDaWork 2013, 3.-.7.6.2013, Vorau (AT): Compositional data analysis of German gross value added (poster)
- ODAM 2013, 12.-14.6.2013, Olomouc (CZ): Log-ratio approach in curve fitting for concentration - response experiments (presentation)
- ERCIM, 14.-16.12.2013, London (UK): Classical and robust principal component analysis for density functions using Bayes spaces (presentation)
- ROBUST, 19.-24.1.2014, Jetřichovice (CZ): Ekonomická aplikace kompozičního regresního modelu (poster+presentation, in Czech)
- StatGeo 2014, 17.-20.6.2014, Olomouc (CZ): Klasická a robustní regrese mezi složkami kompozice (presentation, in Czech)
- LinStat, 24.-28.8.2014, Linköping (SW) - K. Hrůzová, K. Hron, P. Filzmoser, V. Todorov: Orthogonal regression among parts of compositional data (presentation)
- Mathematical Methods in Economics 2014, 10.-12.9.2014, Olomouc (CZ): Orthogonal regression among parts of compositional data with an economic application (presentation)
- ODAM, 20.-22.5.2015, Olomouc (CZ): Compositional analysis of trade flows structure (presentation)
- CoDaWork 2015, 1.-5.6.2015, L´Escala (ES): Orthogonal regression among parts of compositional data with an economic application (presentation)

References

[1] Aitchison J. (1986). *The Statistical Analysis of Compositional Data.* London: Chapman and Hall.

[2] Barceló-Vidal C., Aguilar L., Martín-Fernández J.A. (2011). Compositional VARIMA time series. In Pawlowsky-Glahn V., Buccianti A. (eds.), *Compositional data analysis: Theory and applications*, 89–103. Chichester: Wiley.

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

[4] 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.

[5] Van den Boogaart K.G., Egozcue J.J., Pawlowsky-Glahn V. (2010). Bayes linear spaces. *Statistics and Operations Research Transactions* **34** (2): 201–222.

[6] Van den Boogaart K.G., Egozcue J.J., Pawlowsky-Glahn V. (2014). Bayes Hilbert spaces. *Australian & New Zealand Journal of Statistics* **56** (2): 171–194.

[7] Brunsdon T.M., Smith T.M.F. (1998). The time series analysis of compositional data. *Journal of Official Statistics* **14** (3), 237-–253.

[8] Buccianti A., Mateu-Figueras G., Pawlowsky-Glahn V., eds. (2006). *Compositional Data Analysis in the Geosciences: From Theory to Practice*, London: Geological Society.

[9] Croux C., Fekri M., Ruiz-Gazen A. (2010). Fast and robust estimation of the multivariate errors in variables model. *Test* **19**: 286–303.

[10] Croux C., Filzmoser P., Pison G., Rousseeuw P.(2003). Fitting multiplicative models by robust alternating regressions. *Statistics and Computing* **13** (1): 23–36.

[11] Davison A.C., Hinkley D.V. (1997). *Bootstrap methods and their application.* Cambridge: Cambridge University Press.

[12] Delicado P. (2007). Functional k-sample problem when data are density functions. *Computational Statistics* **22**: 391–410.

[13] Delicado P. (2011). Dimensionality reduction when data are density functions. *Computational Statistics and Data Analysis* 5**5**: 401–420.

[14] Devarajan S., Swaroop V., Zou H. (1996). The composition of public expenditure and economic growth. *Journal of Monetary Economics* **37**: 313–344.

[15] Eaton M.L. (1983). *Multivariate Statistics. A Vector Space Approach.* New York: Wiley.

[16] Egozcue J.J. (2009). Reply to "On the Harker Variation Diagrams; ..." by J.A. Cortés. *Mathematical Geosciences* **41** (7): 829–834.

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

[18] Egozcue J.J., Díaz-Barrero J.L., Pawlowsky-Glahn V. (2006). Hilbert space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica, English Series* **22** (4): 1175–1182.

[19] 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.

[20] Egozcue J.J., Pawlowsky-Glahn V., Tolosana-Delgado R., Ortego M.I., van den Boogaart K.G. (2013). Bayes spaces: use of improper distributions and exponential families. *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas* **107**: 475–486.

[21] Fekri M., Ruiz-Gazen A. (2004). Robust weighted orthogonal regression in the errors-in-variables model. *Journal of Multivariate Analysis* **88**: 89–108.

[22] Filzmoser P. (1999). Robust principal components and factor analysis in the geostatistical treatment of environmental data. *Environmetrics* **10**: 363–375.

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

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

[25] Fox, J. (2002). Bootstrapping Regression Models. Appendix to an R and S-PLUS Companion to Applied Regression.

[26] Fry J.M., Fry T.R.L., McLaren K.R. (1996). The stochastic specification of demand share equations: Restricting budget shares to the unit simplex. *Journal of Econometrics* **73** (2), 377-–385.

[27] Fry T. (2011). Applications in Economics. In *Compositional data analysis: Theory and applications*, edited by Pawlowsky-Glahn V., Buccianti A., 318–326. Chichester: Wiley.

[28] Fuller, W.A. (1987). *Measurement Error Models.* New York: Wiley.

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

[30] Härdle W.K., Simar L. (2012). *Applied Multivariate Statistical Analysis.* Heidelberg: Springer.

[31] Horváth L., Kokoszka P. (2012). *Inference for Functional Data with Applications.* Heidelberg: Springer.

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

[33] Hron K., Menafoglio A., Templ M., Hrůzová K., Filzmoser P. (2016). Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics and Data Analysis* **94**: 330–350.

[34] Hrůzová K., Hron K., Rypka M., Fišerová E. (2013). Covariance structure of principal components for three-part compositional data. *Acta Universitatis Palackianae Olomucensis. Facultas Rerum Naturalium. Mathematica* 52 (2): 61–69.

[35] 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.

[36] 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**: 7–14.

[37] Jolliffe I.T. (2002). *Principal Component Analysis, 2nd ed.*. New York: Springer.

[38] Kynčlová P., Filzmoser P., Hron K. (2016). Compositional biplots including external noncompositional variables. *Statistics* : DOI: 10.1080/02331888.2015.1135155.

[39] Larrossa J. (2003). A compositional statistical analysis of capital stock. In Thió-Henestrosa S., Martín-Fernández J.A. (eds.), *Proceedings of CoDaWork'03, The 1st Compositional Data Analysis Workshop.* Universitat de Girona, ISBN 84-8458-111-X, http://ima.udg.es/Activitats/CoDaWork03/, Girona (E). CD-ROM.

[40] Machalová J., Hron K., Monti G.S. (2015). Preprocessing of centred logratio transformed density functions using smoothing splines. *Journal of Applied Statistics*: DOI: 10.1080/02664763.2015.1103706.

[41] Maronna R.A. (2005). Principal Components and Orthogonal Regression Based on Robust Scales. *Technometrics* **47**: 264–273.

[42] Mateu-Figueras G., Pawlowsky-Glahn V., Egozcue J.J. (2011). The principle of working on coordinates. In Pawlowsky-Glahn V., Buccianti A. (Eds.), *Compositional Data Analysis: Theory and Applications*, 31–42. Chichester: Wiley.

[43] Menafoglio A., Guadagnini A., Secchi P. (2014). A kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers. *Stochastic Environmental Research and Risk Assessment* **28** (7): 1835–1851.

[44] Monti G.S., Migliorati S., Hron K., Hrůzová K., Fišerová E. (2014). Log-ratio approach in curve fitting for concentration-response experiments. *Environmental and Ecological Statistics* **21** (2): 275–295.

[45] Nerini D., Ghattas B. (2007). Classifying densities using functional regression trees: applications in oceanology. *Computational Statistics and Data Analysis* **51**: 4984–4993.

[46] Pawlowsky-Glahn V., Buccianti A., eds. (2011). *Compositional Data Analysis: Theory and Applications*. Chichester: Wiley.

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

[48] Pawlowsky-Glahn V., Egozcue J.J., Tolosana-Delgado R. (2015). *Modeling and Analysis of Compositional Data.* Chichester: Wiley.

[49] Ramsay J., Silverman B.W. (2002). *Applied Functional Data Analysis: Methods and Case Studies*. New York: Springer-Verlag.

[50] Ramsay J., Silverman B.W. (2005). *Functional Data Analysis, 2nd ed.*. New York: Springer.

[51] Rousseeuw P., Hubert M. (2013). High-breakdown estimators of multivariate location and scatter. In Becker C., Fried R., Kuhnt S. (eds.), *Robustness and complex data structures.* Heidelberg: Springer, 49–66.

[52] Salibian-Barrera M., Van Aelst S., Willems G. (2006). PCA based on multivariate MM-estimators with fast and robust bootstrap. *Journal of the American Statistical Association* **101**: 1198–1211.

[53] Shang H.L. (2014). A survey of functional principal component analysis. *Advances in Statistical Analysis* **98**: 121–142.

[54] Van Aelst S., Willems G. (2013). Fast and robust bootstrap for multivariate inference: The R package FRB. *Journal of Statistical Software* **53** (3).

[55] Wierts P., Van Kerkhoff H., De Haan J. (2014). Composition of exports and export performance of Eurozone countries. *JCMS: Journal of Common Market Studies* **52** (4): 928–941.

[56] Zhang Z., Müller H.G. (2011). Functional density synchronization. *Computational Statistics and Data Analysis* **55**: 2234–2249.