PALACKÝ UNIVERSITY OLOMOUC
FACULTY OF SCIENCE

DISSERTATION THESIS

Multidimensional statistical methods for analysis of human metabolome

Department of Mathematical Analysis and Applications of Mathematics
Supervisor: **Doc. RNDr. Karel Hron, Ph.D.**
Author: **Mgr. Alžběta Gardlo (Kalivodová)**
Consultant: **prof. RNDr. Tomáš Adam, Ph.D.**
Study programme: P1104 Applied Mathematics
Field of study: Applied Mathematics
Form of study: full-time
The year of submission: 2016

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Mgr. Alžběta Gardlo (Kalivodová)

Název práce: Vícerozměrné statistické metody pro analýzu lidského metabolomu

Typ práce: Disertační práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

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

Rok obhajoby práce: 2016

Abstrakt: Metabolomika je poměrně novým oborem biochemie zabývající se studiem metabolitů, jejich dynamickými změnami, interakcemi a odpověďmi na podněty. Vzhledem k relativnímu charakteru metabolomických dat na ně může být pohlíženo jako na tzv. kompoziční data. Vektory takovýchto dat mají kladné složky; navíc nás nezajímají jejich absolutní hodnoty, ale podíly mezi nimi. Abychom mohli pracovat s kompozičními daty v klasickém euklidovském prostoru, musíme použít specifické souřadnicové systémy. Dále musíme při analýze metabolomických dat brát v úvahu materiál, který je použit pro měření, a v neposlední řadě i to, že máme k dispozici typicky řádově méně pozorování než proměnných, tedy hovoříme o tzv. vysoce-dimenzionálních datech. Pro analýzu takového souboru musí být použity speciální statistické metody. První částí statistické analýzy je předzpracování dat související s vyjádřením metabolomických (kompozičních) dat v tzv. logratio souřadnicích. V metabolomice také používáme tzv. kontroly kvality, které nám pomáhají v odstraňování chyb měření. Dalším problémem jsou nulové hodnoty. Většina v současnosti používaných statistických metod pro kompoziční data neumí pracovat s nulovými hodnotami, proto je musíme umět vhodně nahradit. Vlastní statistická analýza může být provedena pomocí celé řady postupů. První, nejpopulárnější, je metoda hlavních komponent. Ta je východiskem pro metodu částečných nejmenších čtverců či její ortogonální podobu. Pokud pracujeme s trojrozměrnými datovými tabulkami, můžeme analýzu provést také pomocí metody PARAFAC. Důležitou součástí této práce jsou také praktické příklady na reálných datových souborech z Laboratoře metabolomiky Univerzity Palackého Olomouc.

Klíčová slova: Kompoziční data, metabolomika, metoda částečných nejmenších čtverců, mnohorozměrná statistická analýza, praktická aplikace, nahrazování nul.

Počet stran: 130

Počet příloh: 0

Jazyk: anglický

BIBLIOGRAPHICAL IDENTIFICATION

Author: Mgr. Alžběta Gardlo (Kalivodová)

Title: Multidimensional statistical methods for analysis of human metabolome

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: The metabolomics is a quite new field of biochemistry which aims at studying metabolites, their dynamic changes, interactions and responses to stimuli. Because of relative character of metabolomic data, they can be considered as so called compositional data. They are characterized by positive entries, moreover, not their absolute values but ratios between them are of primary interest. In order to analyze statistically compositional data in standard Euclidean space, specific coordinate systems must be used. Furthermore, for the analysis of metabolomic data also the biochemical material must be considered, and finally, also te fact that substantially less observations than variables are available; we refer to so called high-dimensional compositional data. For statistical analysis of such data set, special statistical procedures must be applied. Prior to the statistical analysis itself, preprocessing of compositional data must be carried out, needed for further representation of logratio coordinates (quality control, zero values of compositional parts). So the statistical analysis itself can be performed using a wide range of proper methods. the most popular one is principal component analysis that can be accompanied by partial least squares method and its orthogonal modification. For the analysis of three-way metabolomic data, PARAFAC is recently preferred choice in chemometrics. Methodological outputs are demonstrated on real data from the Laboratory of Metabolomics, Palacký University Olomouc.

Key words: Compositional data, metabolomics, partial least squares regression, multivariate statistical analysis, practical application, imputation of zeros.

Number of pages: 130

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.

Olomouc, ...
signature

Contents

Acknowledgement

I would like to thank to my supervisor Doc. RNDr. Karel Hron, Ph.D. for helpfulness, guidance and patience during the preparation of the scientific papers and this thesis. I want to thank to all my colleagues and friends from the Laboratory of Metabolomics who provided a friendly environment for my research, especially to my consultant from the laboratory, prof. RNDr. Tomáš Adam, Ph.D. I am very grateful to my mother and my husband, who have always been supporting me in my studies.

List of abbreviations

Throughout the thesis, the following standard abbreviations are used. Other nonstandard abbreviations are introduced in the text as it is needed.

ADCS	average difference in covariance structure
alr coordinates	additive logratio coordinates
ALS	alternating least squares
AUC	area under the curve
CED	compositional error deviation
clr coorinates	centered logratio coordinates
CV	coefficient of variation
ilr coordinates	isometric logratio coordinates
LOESS	local regression
MCADD	medium chain acyl-CoA dehydrogenase deficiency
MSEP	mean squared error of prediction
m/z	mass to change ratio
NIPALS	nonliear iterative partial least squares
OPLS-DA	orthogonal partial least squares regression - discriminant analysis
PARAFAC	parallel factor analysis
PCA	principal component analysis
PLS-DA	partial least squares regression - discriminant analysis
PQN	probabilistic quotient normalization
PRESS	predicted error sum of squares
QC	quality control
SVD	singular value decomposition
VIP	variable importance in the projection

Aims of the thesis

This thesis aims to be a complex guide for the statistical processing of metabolomic (compositional) data sets. The main goal is to study the possibility of using advanced multivariate statistical methods for the statistical analysis of metabolomic data. Inputs from metabolomics have a specific structure - they have properties of so called compositional data. The thesis is focused on the analysis of this type of data. The second property of the metabolomic data is connected with the size of the data table, where typically much more metabolites than observations occur - these data have so called high-dimensional structure. Accordingly, the specific statistical approach must be used for their analysis. The thesis deals with popular methods for statistical processing of high-dimensional metabolomic data, like principal component analysis, partial least squares regression and parallel factor analysis (PARAFAC), which are adopted for the case of compositional data. The problem of the presence of zeros in the data table, that makes not possible to apply the logratio methodology to metabolomic (compositional) data, is also discussed. All algorithms are processed by appropriate software tool, the R software. The large part of the thesis is devoted to application of the logratio methodology to a wide range of data sets from metabolomics.

Introduction

Compositional data (or compositions for short) are multivariate observations with positive components, and they can be represented without loss of information as data with a constant sum constraint like proportions or percentages [1–3]. In such a case, the sum of the compounds (parts) is not important and the only relevant information is contained in the ratios between the parts. Hence, the constant sum is just a representation, not an inherent property of the data, describing quantitatively parts of a whole and following a relative scale. Compositional data occur in a wide range of applications involving geochemistry, analytical chemistry, and its related fields. Nevertheless, up to now just a few papers following the concept of compositional data were published in the field of metabolomics and proteomics [4–7].

Metabolomics aims at studying metabolites, their dynamic changes, interactions and responses to stimuli. It is applied to the metabolism of plants, bacteria, animals and humans. In humans all biological materials from biofluids (blood, urine) till tissues are analyzed. Although absolute values of biomarkers compared with reference ranges (data from the healthy population) is the most frequently used approach, ratios of metabolite data are frequently analyzed in the biochemical diagnostic practice. The reason for their use is, for example, to compensate for the "hydration" of an organism (correction for creatinine in case of urinary measurements), or for variability introduced by a sampling technique (dry blood spots). In diagnostic procedures that interpret data based on "profiling" (semiquantitative data [8, 9] on more variables in patient's biofluid by common techniques, e.g. organic acids in urine by gas chromatography - mass spectrometry), relative changes are more relevant/informative than absolute values. It suggests that metabolomic data can indeed be considered as observations carrying relative information, i.e. as compositional data [6].

Some authors correctly argue that the content of each biofluid is heavily influenced by endogenous sources (e.g. diet), thus concentration levels of metabolites can vary by orders of magnitude. In heavy insults (e.g. genetic enzyme defect, toxicity), the concentration of specific metabolite(s) can increase by orders of magnitudes. If this change represents a substantial part of a possible fixed constant sum constraint of compositional data (like 1 in case of proportions and 100 for percentages), it can lead to biased effects like spurious correlation (all correlation coefficients tend to be negative, thus, the covariance structure of the data is destroyed) and also the relative scale of compositional parts is completely ignored [1, 3]. This situation is pointed out in detail, e.g., by Sysi-Aho in [10]. Nevertheless, these doubts reflect exactly the case when compositional data, represented by a chosen constant sum constraint, are analyzed using standard statistical methods. On the other hand, this problem is correctly handled by the logratio approach to compositional data analysis that we expand throughout the thesis. The relative dominance of metabolites is then correctly reflected by the multivariate structure of the analyzed data.

Very important part of the statistical evaluation is the preprocessing of metabolomic data. The measuring instruments have some limitations and measuring errors can be present in data, for example pressure in the machine can diverge or the temperature in the room can change. To correct these errors special statistical methods must be used. Measurements of standard samples (so called quality control samples) must be stable in time. Then signal correction by LOESS method based on the quality control samples must be done before statistical processing itself is performed [11, 12].

Almost none of statistical methods is able to process data that contain measurement artifacts like missing values (pure absence of the measurement in some entries) or values below a detection limit (resulting as the effect of rounding errors, we also refer to rounded zeros). Especially, values below a detection limit occur frequently in natural sciences related to chemometric data or data from geochemistry. Their proper replacement must precede any further statistical analysis. Although for the case of standard multivariate data a comprehensive methodology exists [13], even applicable to high-dimensional data [14], it fails in case of compositional data. Due to their specific nature, each value to be imputed needs to be considered in a relative sense, as ratios with the other parts in a composition. Imputation methods are already developed for both, missing values [15] and rounded zeros [16–18]. However, these methods fail in case of high-dimensional compositional data sets.

It is widely common in chemometrics, and particularly in metabolomics, to normalize and scale the observations prior to further statistical analysis [19,20]. While most of the normalization techniques are heuristic ones, it is also possible to derive systematic approaches based on natural features of the underlying observations. The relative character of metabolite observations is reflected in the practice by many kinds of normalization techniques that are an integral part of research publications in the field worldwide and are exhaustively described in all chemometrics handbooks [19]. Let us mention, e.g., the well-known AUC normalization whose aim is to normalize a group of signals with peaks by standardizing the area under the curve (AUC) to the group median, mean or any other proper representation. Another approach is represented by rationing to landmarks, e.g. to normalization of urine end-product metabolites to creatinine, that is often used also in general in chemometrics. The choice of any such normalization is usually strongly data dependent in practice, which affects the objectivity and makes any further comparisons hardly attainable [21–26]. All these possibilities of normalization of the original biomarker values to dimensionless observations just reflect the fact that metabolomic observations are of relative nature, i.e. they are compositional data.

After the normalization step, data are popularly transformed using the log-transformation (popular in metabolomics), or alternatively (and preferably here) expressed as proper logratio coordinates that capture relative nature of metabolomic (compositional) data. Both transformations will be discussed in the following chapters.

The statistical analysis of two-way data starts typically with principal component analysis [27–29]. This method must be adapted to work with compositional data [3], i.e. a special coordinate system must be used for the analysis. This method will be introduced in this thesis with its imaging method called biplot. The application of biplot to compositional data is known [3, 29], but it belongs to basic processing methods of two-way data and must be shown for the better complexity of the whole procedure.

Concerning further statistical analysis, the problem occurs because more metabolites (in hundreds) than biological materials (only tens) are present in these data sets. Therefore, suitable methods must be applied for this kind of observations. One of them is partial least squares regression (PLS regression), concretely its popular special case partial least squares - discriminant analysis (PLS-DA) [30–33]. PLS-DA is devoted to a particular regression problem, where the response is formed by categorical variables, whose values represent single groups that occur in the data set. This approach can also be considered as a compromise between usual discriminant analysis and discriminant analysis on the significant principal components of the predictor variables [34]. Nevertheless, the standard PLS-DA method needs to be adapted to compositional data, because (as mentioned above) using raw observations could lead to useless results. This method will be used in a wide range of applications in this thesis because of the specific characteristics of metabolomic data.

A special modification of the PLS model can also be used. It is called orthogonal - partial least squares (OPLS) method and it works with the orthogonal variation in the data [35, 36]. Results of OPLS are popularly visualized using S-plot which is a scatter plot of correlations and covariance of the data. It is a useful tool for detection of important markers of some disease.

Special techniques of metabolomic (compositional) data processing need to be applied, when they form a three-way structure. This structure arises typically when samples of some biological material are measured at more time points. The resulting data array has a structure of a data cube with samples in rows, variables in columns and time points in slices. This cube may be splitted to individual tables, one table for one time point. For statistical processing of three-way observations well established tools like PARAFAC [37, 38] or Tucker3 [39] exist, they are still rarely used in the compositional context [40–42], with no metabolomics application known.

The crucial parts of this thesis are practical applications of all presented methods. All data from examples were measured in the Laboratory of Metabolomics from the Institute of Molecular and Translational Medicine, Palacký University Olomouc. The last chapter of the thesis consist of practical examples that show a coherent approach of statistical processing of the metabolomic data files. The data preprocessing, including the imputation of rounded zeros, is demonstrated here followed by application of principal component analysis, partial least squares regression and orthogonal partial least squares regression. The interpretation of results is also contained in this chapter.

The whole procedure of complete statistical evaluation of metabolomic data, as presented in this thesis, is used in everyday practice in the Laboratory of Metabolomics. Some parts could be rather elementary for mathematical audience, but they are very useful for people from the outside of the statistical field.

All calculations and graphs were performed by the statistical software R [43] using basic packages. Some special packages were also employed, like *robCompositions* (logratio methodology of compositional data) [44], *xcms* [45–47], *CAMERA* [48], *muma* [49] (previous three for the untargeted analysis of metabolites), *zCompositions* [50], *pls* [51], *PTAk* [52, 53] and *ThreeWay* [54].

This dissertation thesis strongly relies on papers that were published, accepted and submitted during my Ph.D. studies:

- C. Kanagaratham, **A. Kalivodová**, L. Najdekr, D. Friedecký, T. Adam, D. Moreno, J.V. Garmendia, M. Hajduch, J.B. De Sanctis, D. Radzioch, Fenretinide prevents inflammation and airway hyperresponsiveness in a mouse model of allergic asthma, *American Journal of Respiratory Cell and Molecular Biology*, vol. 51, no. 6, pp. 783-792, 2014 [55].

- H. Janečková, **A. Kalivodová**, L. Najdekr, D. Friedecký, K. Hron, P. Bruheim, T. Adam, Untargeted metabolomic analysis of urine samples in the diagnosis of some inherited metabolic disorders, *Biomedical Papers*, vol. 159, no. 4, pp. 582-585, 2015 [56].

- **A. Kalivodová**, K. Hron, P. Filzmoser, L. Najdekr, H. Janečková, T. Adam, PLS-DA for compositional data with application to metabolomics, *Journal of Chemometrics*, vol. 29, pp. 21-28, 2015 [6].

- L. Najdekr, **A. Gardlo**, L. Mádrová, D. Friedecký, H. Janečková, E.S. Correa, R. Goodacre,T. Adam, Oxidized phosphatidylcholines suggest oxidative stress in patients with medium-chain acyl-CoA dehydrogenase deficiency, *Talanta*, vol. 139, pp. 62-66, 2015 [7].

- J. Veleba, J. Kopecký, P. Janovská, O. Kuda, O. Horáková, H. Malinská, L. Kazdova, O. Oliyarnyk, V. Škop, J. Trnovská, M. Hájek, A. Škoch, P. Flachs, K. Bardová, M. Rossmeisl, J. Olza, G. Salim de Castro, P.C. Calder, **A. Gardlo**, E. Fišerová, J. Jensen, M. Bryhn, J. Kopecký, T. Pelikánová, Combined intervention with pioglitazone and n-3 fatty acids in metformin-treated type 2 diabetic patients: improvement of lipid metabolism, *Nutrition & Metabolism*, vol. 12, no. 52, pp. 1–15, 2015 [57].

- M. Templ, K. Hron, P. Filzmoser, **A. Gardlo**, Imputation of rounded zeros for high-dimensional compositional data, accepted to *Chemometrics and Intelligent Laboratory Systems*, 2016 [58].

- **A. Gardlo**, A.K. Smilde, K. Hron, M. Hrdá, R. Karlíková, T. Adam, Normalization techniques for PARAFAC modeling of urine metabolomic data, submitted, 2016 [59].

- O. Horáková, J. Hansíková, K. Bardová, **A. Gardlo**, M. Rombaldová, O. Kuda, M. Rossmeisl, J. Kopecký, Plasma acylcarnitines and amino acid levels as an early complex biomarker of propensity to high-fat diet-induced obesity in mice, submitted, 2016 [60].

- R. Karlíková, J. Široká, P. Jahn, D. Friedecký, **A. Gardlo**, H. Janečková, F. Hrdinová, Z. Drábková, T. Adam, Atypical myopathy of grazing horses: a metabolic study, submitted, 2016 [61].

- R. Karlíková, J. Široká, D. Friedecký, E. Faber, M. Hrdá, K. Mičová, I. Fikarová, **A. Gardlo**, H. Janečková, I. Vrobel, T. Adam, Metabolite profiling of the plasma and leukocytes of chronic myeloid leukemia patients, submitted, 2016 [62].

Chapter 1

Compositional data

1.1. Introduction to compositional data

Data from metabolomics closely follow properties of *compositional data*, and thus, their statistical analysis needs to account for this fact. Such data are characterized by features like scale invariance (the information in a composition does not depend on the particular units in which the composition is expressed) and the relative scale (ratios and not absolute distances are important when dissimilarities of observations are analyzed). The concept of relative scale naturally occurs already for most positive univariate data sets [63]. Although absolute (Euclidean) distances within two pairs of samples taken at two observations, (5; 10 and 100; 105 in ppm) are the same, their interpretation is different. In the first case, most observers would say there is double the total amount in the second observation compared to the first while in the second case they say that the values are high but approximately the same. Another property which is crucial for any meaningful statistical analysis of compositional data is called subcompositional incoherence, i.e. information conveyed by a composition should not be in contradiction with that coming from a subcomposition that involves only a subset of the variables [64].

Representation of compositions through vectors with a constant sum is a very popular attitude. Any composition can be expressed in proportions with an appropriate scaling factor. The operation used for the assigning this constant sum is called the *closure* [1–3] and it is defined for a composition $\mathbf{x} = (x_1, \ldots, x_D)^{'}$ with the formula

$$C(\mathbf{x}) = \left(\frac{\kappa \cdot x_1}{\sum_{i=1}^{D} x_i}, \frac{\kappa \cdot x_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa \cdot x_D}{\sum_{i=1}^{D} x_i} \right)^{'}, \tag{1.1}$$

where $\kappa > 0$ is the sum of components. The choice of κ is often 1 (proportions) or 100 (percentages).

The sample space of compositions is called the *simplex* defined as [1–3]

$$\mathbf{S}^D = \left\{ \mathbf{x} = (x_1, x_2, \ldots, x_D)^{'} \middle| x_i > 0, i = 1, 2, \ldots, D; \sum_{i=1}^{D} x_i = \kappa \right\}, \tag{1.2}$$

Two operations are defined on the simplex. The first operation is called the *perturbation* [1–3] and it represents an analogy to the sum of two real vectors. Let's have two compositions $\mathbf{x} \in \mathbf{S}^D$ and $\mathbf{y} \in \mathbf{S}^D$, the perturbation of $\mathbf{x}$ and $\mathbf{y}$ is

$$\mathbf{x} \oplus \mathbf{y} = C(x_1 y_1, x_2 y_2, \ldots, x_D y_D)^{'} \in \mathbf{S}^D. \tag{1.3}$$

The second operation - *powering* - is formed by powering the vector (composition $\mathbf{x} \in \mathbf{S}^D$) by a constant $\alpha \in R$,

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

The natural geometry of compositions, called the Aitchison geometry, accounts for all the features mentioned above (see, e.g. [2,3,65] for details). The Aitchison geometry has all the usual properties that are known from the Euclidean geometry, for which standard statistical methods are designed [66]. However, operations of the Aitchison geometry, like the above mentioned perturbation and powering, are different from the Euclidean geometry case. For this reason, usual multivariate statistical methods like principal component analysis, factor analysis or correlation analysis cannot be directly applied to compositional data, since otherwise interpretations of the results and conclusions can be misleading [3, 7, 65, 67–69].

The Euclidean vector space structure of for the Aitchison geometry is completed by the inner product, norm and distance. The *Aitchison inner product* of two compositions $\mathbf{x}, \mathbf{y} \in \mathbf{S}^D$ is defined as [1–3]

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}. \tag{1.5}$$

The *Aitchison norm* for a composition $\mathbf{x} \in \mathbf{S}^D$ is characterized by the formula

$$\|\mathbf{x}\|_a = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_a} = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} \right)^2}. \tag{1.6}$$

The *Aitchison distance* between $\mathbf{x} \in \mathbf{S}^D$ and $\mathbf{y} \in \mathbf{S}^D$ is given as

$$d_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2}. \tag{1.7}$$

1.2. Logratio methodology

Statistical data analysis is usually carried out in the Euclidean geometry and not in the Aitchison geometry. Thus, the central idea is to express compositions from the simplex in real coordinates and then to apply the standard multivariate methods. From a mathematical point of view, we search for a basis (or generating system) with respect to the Aitchison geometry in order to express compositional data in coefficients of such a basis (coordinate system). As these coefficients are build up using logarithms of ratios of compositional parts, we refer to logratio coordinates. Currently, three basic logratio coordinate systems occur in the literature: additive, centered and isometric logratio coordinates. Nevertheless, only the latter two can be recommended in general, because they map the Aitchison geometry to the Euclidean one isometrically. The use of logratio coordinates preserves the relative scale property of compositions, which is of primary importance in chemometrics, and follow all requirements for a meaningful analysis of compositions as mentioned above. For more detailed discussions on these issues, see [1, 70].

The *centered logratio (clr) coordinates* [1, 3] are defined for a composition $\mathbf{x} = (x_1, \ldots, x_D)'$ as

$$clr(\mathbf{x}) = \mathbf{r} = (r_1, \ldots, r_D)' = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right)'. \qquad (1.8)$$

Although the resulting variables are quite easily interpretable (each of them corresponds to one of the original compositional parts), clr coordinates are coefficients with respect to a generating system on the simplex. For this reason, the resulting covariance matrix of a random composition in clr coordinates is singular [1, 65]. This is a serious limitation for many standard multivariate statistical methods [67]. The singularity restriction of the clr coordinates is overcome by the *isometric logratio (ilr) coordinates*, resulting in $D-1$ coordinates with respect to an orthonormal basis. Unfortunately, it is thus not possible to assign a coordinate to each of the original compositional parts simultaneously, as it was the case of clr coordinates. Nevertheless, as there are infinitely many ways to construct an orthonormal basis, its proper choice [71, 72] allows to construct coordinates with an intuitive interpretation. Thus we get a $(D-1)$-dimensional real vector $ilr(\mathbf{x}) = \mathbf{z} = (z_1, \ldots, z_{D-1})'$ [3, 65, 71], where

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}}, \quad i = 1, \ldots, D-1. \qquad (1.9)$$

The inverse mapping of $\mathbf{z}$ back to the original composition $\mathbf{x}$ is then given by

$$\begin{aligned}
x_1 &= \exp\left(\frac{\sqrt{D-1}}{\sqrt{D}} z_1\right), \\
x_i &= \exp\left(-\sum_{j=1}^{i-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} z_j + \frac{\sqrt{D-i}}{\sqrt{D-i+1}} z_i\right), \qquad (1.10) \\
x_D &= \exp\left(-\sum_{j=1}^{D-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} z_j\right),
\end{aligned}$$

Afterward, a possible closure operation in order to get a prescribed constant sum constraint of the components can be applied.

With the above orthonormal (ilr) coordinates (1.9), the variable z_1 carries all the relevant information about the compositional part x_1, because it explains all the ratios between x_1 and the other parts of $\mathbf{x}$ [15,71]. In z_1, this is expressed by the logratio between x_1 and the remaining parts in the composition, represented by their geometric mean. Obviously, if we permute the parts $x_2, \ldots, x_D$ in (1.9), the interpretation of z_1 remains unchanged. The interpretation of z_1 holds also when the remaining coordinates are constructed with respect to another orthonormal basis on the simplex [70,71].

Now we can proceed to construct such an orthonormal basis, where the first ilr coordinate explains the relative information about a compositional part of interest. For this purpose, the indices in formula (1.9) are just permuted such that the part of interest plays the role of x_1. Accordingly, in order to assign such coordinates to each compositional part x_l, $l = 1, \ldots, D$, we need to construct D different ilr coordinate systems, where the D-tuple $(x_1, \ldots, x_D)'$ in (1.9) is replaced by $(x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)' =: (x_1^{(l)}, x_2^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})'$ [71]. The corresponding ilr coordinates are thus

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}}, \quad i = 1, \ldots, D-1. \qquad (1.11)$$

As a special case we get $z_i^{(1)} = z_i$, for $i = 1, \ldots, D-1$. Obviously, the vector

$\mathbf{z}^{(l)} = (z_1^{(l)}, \ldots, z_{D-1}^{(l)})'$ is again a vector of orthonormal coordinates. Finally, later on we will see the advantage to relate the clr coefficients and the ilr coordinates linearly as $\mathbf{r} = \mathbf{V}\mathbf{z}$. The matrix $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_{D-1})$ has dimension $D \times (D-1)$ and its columns are formed by the orthonormal basis vectors in clr coordinates,

$$\mathbf{v}_i = \sqrt{\frac{D-i}{D-i+1}}\left(0, \ldots, 0, 1, -\frac{1}{D-i}, \ldots, -\frac{1}{D-i}\right)', \quad i = 1, \ldots, D-1. \tag{1.12}$$

Interestingly, $r_l = \sqrt{\frac{D-1}{D}} z_1^{(l)}, l = 1, \ldots, D$, i.e. r_l is proportional to $z_1^{(l)}$, and thus each clr variable (separately) captures all the relative information about the compositional part x_l as well.

The third basic logratio coordinate system is called the *additive logratio (alr) coordinates*. This system is not very often used because results of statistical processing in alr coordinates might depend on the denominator used in the formula and they represents coordinates with respect to a basis that is not orthonormal. As a consequence, alr coordinates do not form an isometric mapping [2, 3]. Though, as we can see in the following text, these coordinates can naturally occur as a result of combining transformations and normalizations used in metabolomics.

The definition of the alr coordinates for a composition $\mathbf{x} = (x_1, \ldots, x_D)'$ is as follows [1, 3]:

$$alr(\mathbf{x}) = \mathbf{w} = (w_1, \ldots, w_{D-1})' = \left(\ln\frac{x_1}{x_D}, \ln\frac{x_2}{x_D}, \ldots, \ln\frac{x_{D-1}}{x_D}\right)'. \tag{1.13}$$

The resulting coordinates are not symmetric [2,3], because the denominator used in (1.13), x_D, can be replaced by any other compositional part. As a consequence, alr coordinates are not invariant under permutation of components, that forms the final principle of compositional data analysis [64]. The way out is to use clr or ilr coordinates instead of alr [3]; this strategy is followed throughout the thesis.

1.3. Regression for compositional data

Very important part of the logratio methodology is the regression analysis. Its aim is to explain the response (real) variable Y by using explanatory variables $x_1, \ldots, x_D$. A linear regression model can be written using a conditional expected value as

$$\mathrm{E}(Y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_D x_D, \tag{1.14}$$

where $\beta_0, \ldots, \beta_D$ are the unknown regression coefficients that are to be estimated [71]. Direct estimation of these parameters in (1.14) using the standard least squares method could be misleading in case compositional explanatory variables due to their specific geometrical properties. Therefore, several approaches for regression models with compositional explanatory variables were suggested. As a special case, when compositional data are considered as observations with a unit constant sum constraint, the problem is known as experiments with mixtures [73–75]. However, apart from numerical problems, this approach does not follow the basic principles of a meaningful compositional data analysis [1, 71].

Regression with compositional explanatory variables can be carried out by first applying ilr coordinates to covariate composition. For a regression model between Y and $\mathbf{x}$ (composition) we use coordinates $\mathbf{z}$ by applying formula (1.9). The standard multiple linear regression of Y on the explanatory variables $\mathbf{z} = (z_1, \ldots, z_{D-1})'$ is thus obtained,

$$\mathrm{E}(Y|\mathbf{z}) = \gamma_0 + \gamma_1 z_1 + \ldots + \gamma_{D-1} z_{D-1}. \tag{1.15}$$

As in formula (1.11), we can consider the lth ilr basis, for $l = 1, \ldots, D$, resulting in a regression model

$$\mathrm{E}(Y|\mathbf{z}) = \gamma_0 + \gamma_1^{(l)} z_1^{(l)} + \ldots + \gamma_{D-1}^{(l)} z_{D-1}^{(l)}. \tag{1.16}$$

Since $z_1^{(l)}$ explains all the relative information about part $x_1^{(l)}$, also the interpretation of the coefficient $\gamma_1^{(l)}$ can be associated to this part. The interpretation of the other regression coefficients (except γ_0) is not straightforward, because the corresponding explanatory variables (coordinates) do not fully represent one particular part of the composition. Consequently, a possible way to evaluate the contribution of each compositional part for explaining the response Y separately is to consider D regression models according to (1.16) by taking $l \in \{1, \ldots, D\}$, and to interpret the coefficients $\gamma_1^{(l)}$, representing the relative information on parts $x_1^{(l)}$ [71].

Taking a sample with n observations of the response and the explanatory variables, $(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n)$, with $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD})'$ for $i = 1, \ldots, n$, we get the sample version of the regression model (1.15) as

$$\mathrm{Y}_i = \gamma_0 + \gamma_1 z_{i1} + \ldots + \gamma_{D-1} z_{i,D-1} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.17}$$

where the explanatory variables $\mathbf{z}_i = (1, z_{i1}, \ldots, z_{i,D-1})'$ result from the ilr coordinates of $\mathbf{x}_i$ (also 1 for the intercept term is added), and ε_i represents the error term. Without loss of generality, we develop just the sample version of the model (1.15); its generalization for model (1.16) is straightforward. Using the notation $\mathbf{Y} = (\mathrm{Y}_1, \ldots, \mathrm{Y}_n)'$ for the observation vector, $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)'$ for the $n \times D$ design matrix, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ for the error term, the model (1.17) can be rewritten in matrix form,

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \tag{1.18}$$

Accepting the standard assumptions on the random variables ε_i (uncorrelated, with the same variance σ^2), the regression coefficients $\boldsymbol{\gamma}$ can be estimated with the least squares method as

$$\widehat{\boldsymbol{\gamma}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}. \tag{1.19}$$

Note that a regression model with the ilr variables $z_1^{(1)}, \ldots, z_1^{(D)}$ as covariates, which seems to be advantageous for interpretation purposes, would not be appropriate because it results in singularity (remember that $z_1^{(1)}, \ldots, z_1^{(D)}$ are proportional to the clr variables). Namely, the corresponding design matrix of the regression model would not have full rank in columns and (1.19) could not be used for parameter estimation. In this case, the theory of singular regression models would have to be considered [71, 76], which is rarely done in practice due to its complexity. Note also that the regression model (1.17) can be extended to the multivariate case, where more than one response, say $q > 1$ response variables, are considered. Then, $\mathbf{Y}$ just stands for an $n \times q$ data matrix and $\boldsymbol{\gamma}$ denotes a $D \times q$ matrix of regression parameters; their estimation formula (1.19) remains (formally) unaltered.

Another case leading to the singularity of the regression model (1.17), that frequently occurs in chemometrics, comes with more explanatory variables to be involved in the analysis than the number of observations (samples). Here, partial least squares regression seems to be advantageous for estimation of the regression parameters (it will be introduced in Section 3.3).

Chapter 2

Metabolomics

2.1. Introduction to metabolomics

Metabolomics is a quite new field of biochemistry and it aims at studying metabolites, their dynamic changes, interactions and responses to stimuli. It is a science which studies the complex profile of low-molecular weight metabolites present in biological samples at a specific time [56]. Metabolomics is connected with the study of metabolites and metaboloms. The metabolome is the set of small molecular mass organic compounds found in a given biological material. The metabolite is a conception which includes all organic substances naturally occurring from the metabolism of all living organisms. Metabolites are the end products of all cellular processes and are a direct outcome of enzymatic and protein activity [77]. Metabolomics is the analysis of metabolome in a given condition [78]. The term ”of given condition”is very important here because metabolites may change in different environments (for example metabolome is different in healthy and sick organisms).

The metabolomics is focused on two main groups of organisms - human (and animals) and plants. This thesis will target on the metabolomics of human and animals.

The metabolomics is one part of the family of ”omics”disciplines. It is closely connected with lipidomics, glycomics and proteomics at the cellular base. This connection is better visible from Figure 2.1 [79].

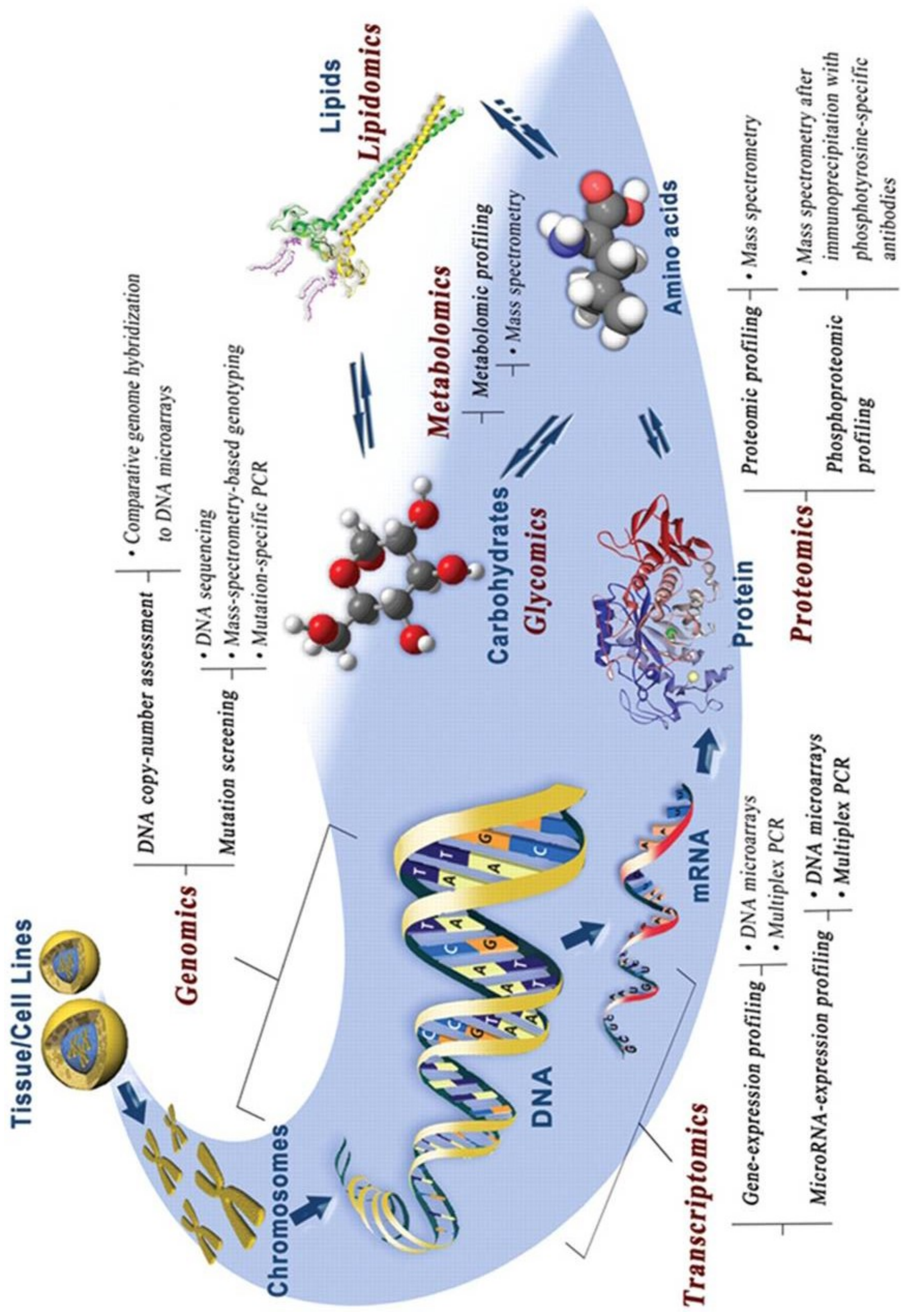


Figure 2.1: The connection between metabolomics and the other ”omics”disciplines [79].

Thanks to a new technology it is possible to measure thousands of metabolites simultaneously from only minimal amounts of sample in presence [80]. This possibility allows defining different attitudes to the analysis of metabolomic samples. The classical division is done by targeted and untargeted approaches [80]. Both of these methods have a specific pros and cons and their choice depends on the experimental objective, measuring instrument, available time and the expert [81].

In the *targeted* analysis the list of metabolites, which are measured, is done before the analysis. The amount of metabolites is not so large (one or two hundreds), because this analysis is very often related only to one specific biochemical pathway. All metabolites are identified before the analysis, therefore, the interpretation of the targeted analysis is easier [78, 80, 81]. This approach is used for pharmacokinetic studies of drug metabolism and for measuring the influence of therapeutics or genetic modifications on a specific enzyme. The targeted analyzes were used as first methods applied in metabolomics and they provide a highly sensitive and robust method for measuring a significant number of biologically important metabolites with relatively high efficiency [80]. The targeted screening of blood samples of newborn babies is also used as a preventive program in hospitals all around the world [78]. The disadvantage of the targeted method may be the price because metabolites used in the analysis must be once adjusted to the machine before the first analysis. This process is done with commercially available chemical standards, but a majority of metabolites are not available commercially or their standards are very expensive. The use of these standards may be a source of the measuring bias [81].

The *untargeted* metabolomic methods are global in scope and have the aim of simultaneously measuring as many metabolites as possible from biological samples without bias [78, 80]. These metabolites are not known before the experiment [81]. This method enables to detect almost all known metabolites. The first disadvantage of the untargeted method is that the output of the analysis is the list of thousands potential metabolites (so called chromatographic peaks) [80, 81].

These peaks must be identified. This procedure is not so easy, despite the fact that several softwares for the identification exist (for example the sequential use of R packages *xcms* [45–47], *CAMERA* [48] and *muma* [49]). Some peaks are misleading, some are only the fractions of real metabolites and some metabolites may produce more than only one peak [80]. On the other hand, there is a higher chance to find some new biomarker of some disease by this attitude, because of the better complexity of the data table. The second disadvantage of this method is a higher presence of zero values that need to be imputed prior to further statistical analysis. Nevertheless, the untargeted metabolomics has great potential to provide insights into fundamental biological processes [80]. Moreover, the sensitivity and specificity of this method are not so high like in the targeted analysis [81].

Some authors also work with the third term - the *semi-targeted* analysis [81]. This is a type of targeted analysis, accordingly it defines metabolites to be tested before the experiment and set the method to detect these specific metabolites with high accuracy, precision, sensitivity and specificity [81]. The difference between targeted and semi-targeted methods is in a number of defined metabolites. In this terminology, the targeted analysis is done only for small amount of metabolites (often less than 20), the semi-targeted analysis works with a larger group of metabolites [81]. We won't use this term in this thesis.

Data sets from metabolomics have a specific structure. The data matrix $\mathbf{X}$ consist of n observations in rows and D variables in columns. Usually more variables (in hundreds) than samples (only in tens) is presented in data ($n < D$) [77]. Data with this structure are called high-dimensional data. Statistical analysis of this type of data requires special methods like partial least squares regression that will be presented in Section 3.3. Data from metabolomics also closely follow the properties of compositional data, so ratios between metabolites instead of their absolute values form the source of relevant information [6]. As a consequence, it is recommended to perform their analysis using the logratio methodology.

Almost all examples used in this thesis are related to the diagnosis of inherited metabolic disorders. They represent a large group of diseases caused by gene mutations resulting in dysfunctional enzymes. The defects are presented biochemically by elevated levels of substrates of the dysfunctional enzymes (in the normal healthy situation these enzymes convert the substrates into products keeping the homeostasis). These diagnostic metabolites are discovered from a random sample of observations and can also be deduced from the knowledge of biochemical pathways. The majority of genetic enzyme defects has been discovered by diagnostic biochemists, who noticed an unusually high concentration of particular metabolites. This observation led to the theory of causative defects which was subsequently confirmed by enzyme assays. Changes which might accompany the defects, which are not easily deductible from biochemistry, are hard to be discovered by traditional ways. Our preliminary results [7] suggest that methods of supervised metabolomics provide potentially effective ways for the recognition of such phenomena.

2.2. Data normalization in metabolomics

Data analysis in metabolomics is a very specific process. It is closely related to the material, which is measured and processed (cells, blood, urine, plasma, . . .). Accordingly, a specific approach is used also for the analysis of urine samples as urine volume can vary widely based on upon water consumption and other physiological factors. Consequently, the concentrations of metabolites in urine vary substantially and proper normalizing for these effects is necessary [21]. Two methods of data normalization are used in practice - creatinine normalization and normalization by the area under the curve [21–26].

The first method of normalization is related to a very specific metabolite, called creatinine, which is presented in all urine samples. Creatinine is a chemical waste product in the blood (a by-product of normal muscle contractions) which passes through kidneys to be filtered and eliminated in urine. It comes through creatine, a supplier of energy to the muscle. Under normal conditions, urinary creatinine output is relatively constant and measurable. As a result, it has become common practice to normalize urinary analyte levels to this metabolite. However, creatinine production does vary and excretion can be impacted by an external stressor such as kidney impairment. In these cases, normalization to creatinine is obviously not warranted [21–23] and can even lead to strongly biased results, when any such kidney disease is not a priori known. Moreover, in practice the level of creatinine is different in various samples, thus, each sample is divided by a different scaling constant. Although this type of creatinine variation is exactly the reason, for which creatinine was employed for normalization purposes, it leads to oblige (biased) coordinates with respect to the logratio methodology (alr coordinates), where scale invariance is automatically an inherent feature of any coordinate system [1, 3].

The second normalization is performed through the area under the curve (AUC) of all peaks, identified with metabolite concentrations in the analysis. By this popular approach, coefficients of metabolites are rescaled by the average AUC. Each mass spectrum (metabolite) is thus divided by average variable area across observations [24–26]. Therefore, up to a constant (resulting from taking the average), it is the well-known total sum normalization, where all elements of a given fingerprint are divided by the total sum of this fingerprint. The average AUC can be computed several ways; while the standard option is formed by the arithmetic mean, in the case of positive data the geometric mean of AUC seems to be more preferable as it also would correspond to centering of log-transformed data across metabolites.

After the normalization step, data are popularly transformed using the log-transformations (popular in metabolomics), or alternatively taking the proposed logratio coordinates. It is important to note that although the popular log-transformation of the input data removes the relative (measurement) scale effects, the scale invariance of compositions is destroyed. The reason is that any normalization applied to the original data prior to log-transforming would alter ratios between components that form the source of relevant information in urine metabolomic data.

Chapter 3

Multidimensional statistical analysis

3.1. Data preprocessing

3.1.1. The use of quality control samples

A very important part of the statistical evaluation is formed by preprocessing of metabolomic data. The first reason for doing this procedure is the fact that measuring instruments have some limitations and measurement errors can be present in the data, for example pressure in the machine can diverge or temperature in the room can change. Special statistical methods must be used to correct these errors. The quality control (QC) samples are used for this purpose. QC samples are mixtures of all samples from the specific analysis. They are measured continuously in the whole analysis on the first ten positions and then as every fourth sample in a way. It is known that signal of these QC samples must be stable in time; if there is some trend, it must be revised. The signal correction by LOESS (LOcal regrESSion) method is used for this purpose [11,12]. The LOESS curve is fitted to the QC samples with respect to the order of injection. A correction curve for the whole analytical run is then interpolated, to which the total data set for that feature is normalized [12].

The LOESS curve fitting is a combination of linear least squares regression and nonlinear regression. It fits simple models to localized subsets of the data to build up a function that describes the deterministic part of the variation in the data, point by point. In this way, there is no requirement of specifying a global function of any form to fit a model to the data, but only to fit segments of the data. In this implementation, the local polynomials that are fitted to each subset of the data are constrained to be either first or second degree (the second degree is used in this thesis). The resulting polynomial is fitted using weighted least squares [12, 82]. In this implementation, the standard tri-cubic weight function is used [11, 12]. The last parameter which is used in LOESS method is called the smoothing parameter. It determines how much of the data is used to fit each local polynomial. This parameter is a number from the interval $\langle(\lambda + 1)/n, 1\rangle$, where λ is the degree of the local polynomial and n denotes the total number of QC samples. The value of this parameter is the proportion of data used in each fit [12]. This parameter is often set up to 0.75 in our analyzes.

The second correction of LOESS curve is done through the comparison of the maximum and minimum interpolated value of QC samples. If the ratio of the maximum and the minimum of new QC values is higher than 10, the particular metabolite is deleted from the data set, because its values are much differentiated and the measurement error in the data is still high. Metabolites with negative values of this ratio (probably caused by the instrument error) are also skipped.

The procedure of using LOESS is visible from Figures 3.1 - 3.3. In Figure 3.1 the time flow of one particular metabolite - the N-acetylaspartate - is shown. The axis x is formed by the indexes of individual samples in the machine, the axis y is the peak area of individual samples, the black points (●) are samples of healthy controls, the blue squares (■) denote individual samples of patients with a specific disease and the red rectangles (▲) are QC samples. Groups of patients and controls are separated only in Figure 3.1, in the other plots they are denoted together as black points. The structure of the data is the following: first eight QC samples are not used in the analysis (they are measured only for the stabilization of the system), so the first two rectangles in the plot are QC samples with numbers 9 and 10, then sequentions of three samples (patients and controls mixed together)

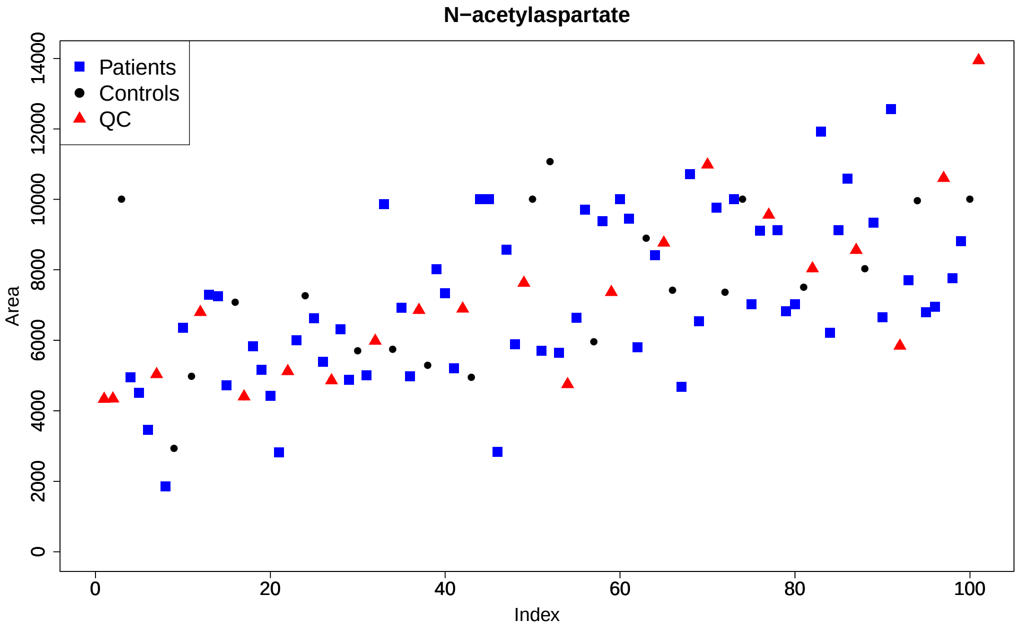


Figure 3.1: Time flow of N-acetylaspartate - the raw data.

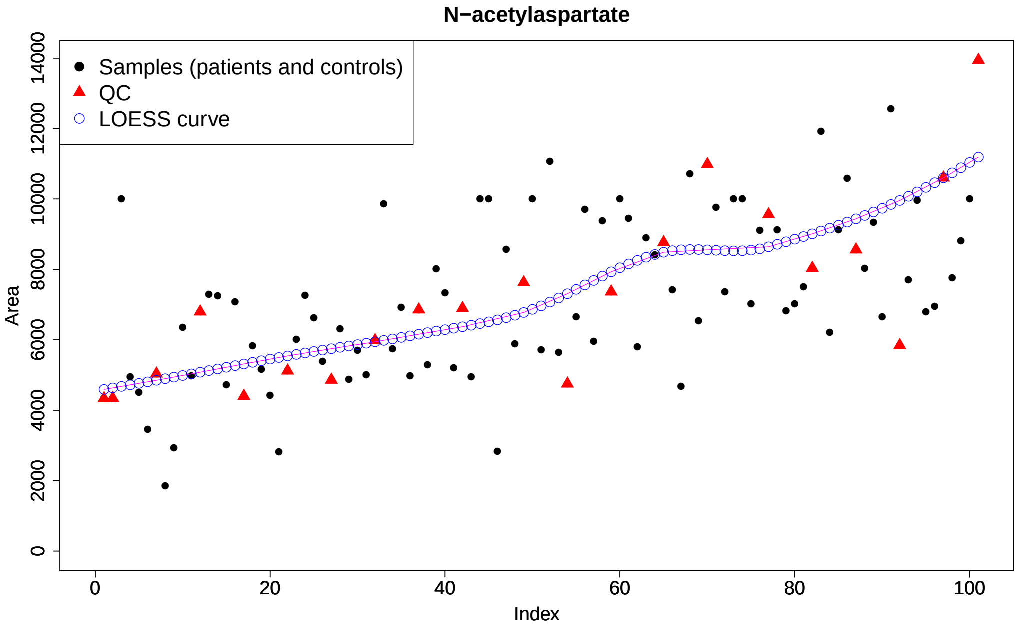


Figure 3.2: Time flow of N-acetylaspartate - the LOESS curve.

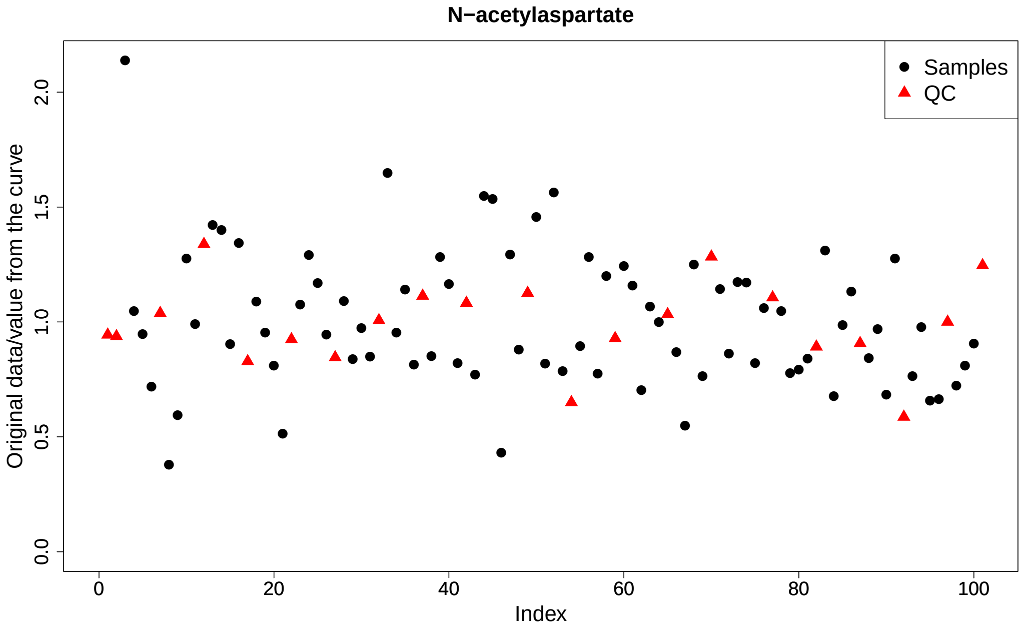


Figure 3.3: Time flow of N-acetylaspartate - the final data.

and one QC follow. The increasing trend in the data is visible. QC samples should be stable in time so this trend must be reduced. In Figure 3.2 the same data are shown with the fitted LOESS curve to QC samples - the blue circled line. Then the corresponding interpolated values from the curve are evaluated for each sample. These interpolated values are represented by blue circles on the fitted line. The result is shown in Figure 3.3. The scale of the y axis has changed, but this doesn't lead to any problem because we are only interested in ratios of metabolites, not in absolute values. QC samples are spaced around the value 1 now; moreover, the position of the outlier (the first sample in the analysis - the black point in the left upper corner of the graph) is preserved. This process is done with all metabolites to be analyzed, so values from y axis of Figure 3.3 are taken to the resulting data table, used for further statistical analysis.

The next step of the preprocessing is the computation of the coefficients of variation (CV) of QCs. All features with CV higher than 30 % are rejected from further processing.

The last step of the preprocessing of the data is an imputation of zeros. This issue is discussed separately in the following section.

3.1.2. Imputation of missing values and rounded zeros

The second very common problem of all chemometric methods is the presence of missing entries in data tables. Standard statistical methods are not able to work with missing values, therefore, they must be imputed by reasonable items in advance. A number of types of missing entries occur in data, we will focus only on two of them - standard missing values and so called rounded zeros. The origin of *missing values* may be caused by several reasons, for instance, the values are not reasonable for some particular variables or variables can not be measured in some samples because of some technical problems. It is not reasonable to simply discard such observations or remove the corresponding variables. The better option is to replace these data by reasonable values [83]. Standard missing values are not frequently present in chemometric data.

The second type of missing entries - *rounded zeros* - are more common in data sets from metabolomics. These zeros are connected with limitations of measurement devices. Every measuring device has a threshold of adjustment which is called the detection limit. Values below this threshold are not recorded and the instrument evaluates them as zeros. These zeros must be imputed with respect to the detection limit. The ascribed value must not exceed this threshold [18]. Some methods of imputation were published [17, 18, 50, 84] but none of them deal with high-dimensional compositional data.

The simplest way to impute missing values of a part is to replace them by the geometric mean of all available data in this part, since the geometric mean reflects the best linear unbiased estimator, see [85] for details. A similar approach is to fill a missing entries with a small predefined value or with the median of the rest values across all samples (evidently with exclusion of all missing values) or with a two thirds (or a half) of the minimum found in the appropriate column of non-missing data [86–88]. The specific approach is an imputation of zero values by two thirds of minimal value in particular feature (metabolite), where the imputation is done within a group of samples. This means that zeros are replaced individually for each group (controls/patients) per each feature. Similarly as in the case of standard multivariate data, these approaches would completely ignore the multivariate data structure and would underestimate the covariance structure of the data set. For these reasons, they should not be used in practice.

Several algorithms are available for the imputation of missing values in compositional data. For example, a modified version of the k-nearest neighbour ($k$nn) imputation can be used, where the missing entries in a composition are replaced by using the available variable information of the k neighbouring observations with respect to an appropriate distance measure, the Aitchison distance, and using an adjustment of the imputations [15, 86]. However, $k$nn imputation still does not fully account for the multivariate relations between the compositional parts as this is only considered indirectly when searching for the k-nearest neighbors. For this reason, among other alternatives, as a second approach which fully accounts for the multivariate relations between the compositional parts, a regression-based imputation procedure was introduced in [15]. It consists of an iterative regression-based algorithm, where the ilr coordinates (1.9) are repeatedly and sequentially applied for $i = 1, \dots, D$ in order to improve the starting solution, represented by results of the $k$nn imputation.

However, these algorithms do not account for the problem of rounded zeros. The replacement of rounded zeros represents a constrained version of missing values imputation. Namely, when x_{ij} represents a rounded zero for a particular observation i and a variable j, it holds that $x_{ij} < e_{ij}$, where e_{ij} is a threshold. The above mentioned regression-based algorithm thus can be used, where the particular (constrained) case should be taken into account. The initialization of the iterative procedure is provided by taking 2/3 of the detection limit for the affected data entries [16, 58]. Note that for more than 10% of rounded zeros this might result in a serious distortion of the multivariate data structure, even new outlying observations might arise. Thus, a substantial improvement of the initial imputation is necessary; hereat, the crucial point is to express the threshold values in coordinates [18]. This guarantees that the estimated values are placed below the detection limit throughout the estimation process.

Although in the regression step of the algorithm the usual least squares regression can be also replaced by a proper robust counterpart, suppressing possible outlying observations, neither of them is able to cope with high-dimensional data. The way out, based on the use of PLS regression, is introduced in Section 3.5.

3.2. Principal component analysis

One of the basic methods used in multivariate data analysis (especially for visualization of the data structure) is definitely *principal component analysis* (PCA). The aim of this method is to reduce the dimensionality of data by preserving the most information identified with variability contained in the data set. Its main principle is to construct an orthogonal coordinate system, which is formed by latent variables, so that only the first few variables explain most of variability in data. The goal of PCA is also the reduction of the effect of measurement error and elimination of components associated with the noise [28].

Principal component analysis belongs to the family of *unsupervised* methods. This means that the algorithm does not know anything about specific groups in samples.

As PCA is not scale invariant, proper scaling of samples must be performed prior to the analysis [27].

Let's have a real data matrix $\mathbf{X}$ of dimension $n \times D$, i.e., with n observations and D variables, which is centered and possibly also scaled. The direction of the highest variability in data is called the first principal component (PC1) and it is defined by a loading vector $\mathbf{p}_1 = (p_1, \ldots, p_D)^{'}$ [27]. The length of loading vectors is normalized to 1; that means $\mathbf{p}_1^{'}\mathbf{p}_1 = 1$. The corresponding scores are linear combinations of loadings and sample vectors. Let's have i-th observation $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD})^{'}$, the score t_{i1} of PC1 is given as

$$t_{i1} = x_{i1}p_1 + \ldots + x_{iD}p_D = \mathbf{x}_i^{'}\mathbf{p}_1, \quad i = 1, \ldots, n, \tag{3.1}$$

and for the whole data matrix $\mathbf{X}$ the score vector $\mathbf{t}_1$ is obtained by

$$\mathbf{t}_1 = \mathbf{X}\mathbf{p}_1. \tag{3.2}$$

The second principal component (PC2) is formed as an orthonormal direction to PC1 and again acquires the maximum possible variability of scores. The following principal components are orthogonal to all previous components and their direction has to cover the maximum possible variance of the data projected on this direction [27]. In the standard analysis, usually only first two (or maximum three) principal components are considered for practical reasons with the hope that they contain most of the total variance in the data set. Nevertheless, in general, the number of principal components is limited only by the number of variables.

Loading vectors of all principal components are orthogonal to each other, which means that the data transformation by PCA is a rotation of the coordinate system. For the orthogonal vectors holds that the scalar product is zero, so we have $\mathbf{p}_j^{'}\mathbf{p}_k = 0, j, k = 1, \ldots, D, j \neq k$. PCA scores are also orthogonal to each other, resulting in $\mathbf{t}_j^{'}\mathbf{t}_k = 0$, for $j, k = 1, \ldots, D, j \neq k$ and $\mathbf{t}_j = (t_{1j}, \ldots, t_{nj})^{'}$. The loading matrix, denoted as $\mathbf{P}$, is formed by all loading vectors and all score vectors result in the score matrix, $\mathbf{T}$. Now it is possible to build a model

$$\mathbf{T} = \mathbf{X}\mathbf{P}. \tag{3.3}$$

The data matrix $\mathbf{X}$ can be reconstructed from the score matrix $\mathbf{T}$ using only a first few principal components corresponding to the main structure of the data (the matrix $\mathbf{T}$ has now less columns than by considering all D principal components). The result is an approximation of the input matrix, denoted as $\mathbf{X}_{app}$, with reduced noise

$$\mathbf{X}_{app} = \mathbf{T}\mathbf{P}^{'}, \qquad \mathbf{X} = \mathbf{T}\mathbf{P}^{'} + \mathbf{E}, \qquad \mathbf{E} = \mathbf{X} - \mathbf{X}_{app}, \tag{3.4}$$

where $\mathbf{E}$ is the error (residual) matrix.

PCA is connected with *singular value decomposition* (SVD). Let's have a data matrix $\mathbf{X}$ of dimension $n \times D$ and rank $k \leq \min(n, D)$. This matrix can be decomposed into a product of three different matrices [89, 90]

$$\mathbf{X} = \mathbf{UDV}^{'}, \tag{3.5}$$

where $\mathbf{U}$ is a $n \times k$ matrix with orthonormal columns containing the left singular vectors, $\mathbf{D}$ is a diagonal matrix of dimension $k \times k$ containing singular values and $\mathbf{V}$ is a $D \times k$ matrix with orthonormal columns containing right singular values. Columns of the last matrix are loadings. Scores are formed by the product of the first two matrices

$$\mathbf{X} = (\mathbf{UD})\mathbf{V}^{'} = \mathbf{TP}^{'}. \tag{3.6}$$

Loadings give weights of the original variables in the principal components. Scores (columns of the matrix $\mathbf{T}$) contribute the coordinates in the space of latent variables. Columns of the matrix $\mathbf{U}$ give the same coordinates in a normalized form (their variances are unit), whereas columns of $\mathbf{T}$ have variances corresponding to the variance of each particular principal component [90]. These variances are denoted as λ_i

$$\lambda_i = \frac{d_i^2}{n-1}, \quad i = 1, \ldots, k, \tag{3.7}$$

where $d_i, i = 1, \ldots, k$ are diagonal elements of the matrix $\mathbf{D}$ sorted in descending order. The variance explained by the i-th principal component is expressed as fraction $\lambda_i / \sum_{j=1}^{k} \lambda_j$.

The compositional approach to PCA is very similar to the standard one. The only difference is in the data matrix, which is used for the analysis. In the logratio approach, the original (compositional) data matrix $\mathbf{X}$ is expressed in centered clr coordinates (1.8), which is denoted as $\mathbf{R}$. Then the PCA machinery can be performed [3]. Ilr coordinates (1.9) may also be used, but their specific interpretation needs to be taken into account for interpretation of loadings (scores corresponding to nonzero singular values are the same). For this reason, ilr (orthonormal) coordinates are still not much popular in the context of compositional PCA.

A graphical representation of PCA is called *biplot*. It is a planar graph used for the projection of scores and loadings of the first two principal components into one plot. Scores, which represent observations, are displayed as points. Loadings, which represent variables, are displayed by arrows (rays) in the same plot [3,29].

The interpretation of the compositional biplot is quite different from the interpretation of the standard one. The loading matrix $\mathbf{P}$ (corresponding to the matrix $\mathbf{R}$) represents clr coordinates of the original variables in compositional biplot. Let‘s denote elements of the matrix $\mathbf{R}$ (respectively $\mathbf{X}$) as r_{ij} (x_{ij}), its rows $\mathbf{r}_{i.}(\mathbf{x}_{i.}), i = 1, \dots, n$, and columns $\mathbf{r}_j(\mathbf{x}_j), j = 1, \dots, D$. The same notation is used also for matrices $\mathbf{T}$ and $\mathbf{P}$. The inner product of rows of matrices $\mathbf{T}$ and $\mathbf{P}$ approximates the matrix of clr coordinates $\mathbf{R}$

$$\mathbf{t}_{i.}^{'}\mathbf{p}_{j.} \approx r_{ij} = \ln \frac{x_{ij}}{g(\mathbf{x}_i)}, \quad i = 1, \dots, n, \; j = 1, \dots, D, \tag{3.8}$$

where $g(\mathbf{x}_i) = \sqrt[D]{\prod_{j=1}^{D} x_{ij}}$ denotes the geometric mean of the given D-part composition $\mathbf{x}_{i.}$ (row of the matrix $\mathbf{X}$).

The single clr variables can be interpreted as those capturing all the relative information (in term of ratios) about the corresponding compositional parts. However, the geometric mean in the denominator of clr coordinates can be driven by possible distortion of involved parts, therefore, the interpretation of clr variables in the sense of original compositional parts requires a careful selection of parts [91].

The first important element of the biplot is the origin, which represents the center (the geometric mean of parts used in clr coordinates) of the data.

The position of the origin is zero for the first two principal components using the centered data set [3].

The other properties are connected with rays and links between vertices of the biplot. They provide an information about the relative variability of a logratio in a data set. The length of rays estimates the standard deviation of clr coordinates

$$\|\mathbf{p}_{j.}\|^2 = \mathbf{p}_{j.}^{'}\mathbf{p}_{j.} \approx \frac{1}{n-1}\mathbf{r}_j^{'}\mathbf{r}_j = \frac{1}{n-1}\sum_{l=1}^{n}\left(\ln\frac{x_{lj}}{g(\mathbf{x}_l)}\right)^2 = \mathrm{var}\left(\ln\frac{\mathbf{x}_{.j}}{g(\mathbf{x}_i)}\right). \tag{3.9}$$

The link between two vertices estimates the standard deviation of the logratio between the corresponding compositional parts [3, 91].

$$\begin{aligned}\|\mathbf{p}_{i.}-\mathbf{p}_{j.}\|^2 &\approx \frac{1}{n-1}(\mathbf{r}_i-\mathbf{r}_j)^{'}(\mathbf{r}_i-\mathbf{r}_j) = \frac{1}{n-1}\sum_{l=1}^{n}(\mathbf{r}_{li}-\mathbf{r}_{lj})^2\\ &= \frac{1}{n-1}\sum_{l=1}^{n}\left(\ln\frac{x_{li}}{g(\mathbf{x}_l)}-\ln\frac{x_{lj}}{g(\mathbf{x}_l)}\right)^2 = \frac{1}{n-1}\sum_{l=1}^{n}\left(\ln\frac{x_{li}}{x_{lj}}\right)^2\\ &= \mathrm{var}\left(\ln\frac{\mathbf{x}_i}{\mathbf{x}_j}\right).\end{aligned} \tag{3.10}$$

Consequently $\mathrm{var}\left(\ln\frac{\mathbf{x}_i}{\mathbf{x}_j}\right)$ can be approximated as the (squared) length of a link.

The connection between observations and variables in the plot is performed by the projection of a score onto a link, which represents an approximate difference between two clr coordinates r_{ij} and r_{ik}. It is the logratio between original values x_{ij} and x_{ik}.

$$\mathbf{t}_{i.}^{'}(\mathbf{p}_{j.}-\mathbf{p}_{k.}) \approx r_{ij}-r_{ik} = \ln\frac{x_{ij}}{g(\mathbf{x}_i)} - \ln\frac{x_{ik}}{g(\mathbf{x}_i)} = \ln\frac{x_{ij}}{x_{ik}}. \tag{3.11}$$

The last property is connected with the Euclidean distance of two score vectors (rows of the matrix $\mathbf{T}$), which approximates the Mahalanobis distance between clr coefficients in the full space with the estimated covariance matrix $\mathbf{S_R}$ of clr coordinates,

$$\|\mathbf{t}_{i.} - \mathbf{t}_{j.}\|^2 \approx (\mathbf{r}_{i.} - \mathbf{r}_{j.})^{'} \mathbf{S}_{\mathbf{R}}^{-1} (\mathbf{r}_{i.} - \mathbf{r}_{j.}), \tag{3.12}$$

see [91] for further details.

3.3. Partial least squares regression

3.3.1. Theoretical background

Very common problem of metabolomic data sets is their high-dimensionality (= the presence of more variables than observations). Therefore, suitable methods must be used for their analysis. One of them is *partial least squares* (PLS) regression which is a class of methods for modeling relations between sets of explanatory and response variables by means of latent variables [92]. It can be viewed as a combination of principal component analysis and multiple regression [93, 94], nevertheless, instead of finding hyperplanes of minimum variance between the response and independent variables using directly principal component regression, it finds a linear regression model by projecting the predicted variables and the observable variables to a new space. PLS can be used for both regression and classification purposes and it can be employed also for reducing the dimensionality of the data. The intrinsic assumption of all PLS methods is that the observed data are generated by a system or process which is guided by a small number of latent (not directly observed or measured) variables [92]. PLS is a widely used method in chemometrics for multivariate calibration and finds increasing interest also in other areas, like when dealing with highly collinear predictor variables [27, 94–96]. Partial least squares - discriminant analysis (PLS-DA) is a special type of a regression analysis where the response variables represent group labels. PLS-DA is one of *supervised* methods. The information about grouping in samples is used in the analysis.

Consider explanatory and response variables, whose sample values are recorded in the matrix $\mathbf{X}$ of dimension $n \times D$ and in the matrix $\mathbf{Y}$ of size $n \times q$. Data in rows of the matrix $\mathbf{X}$ represent n objects with D features (explanatory variables), $\mathbf{Y}$ describes for the same n objects q properties (response variables). In PLS-DA, the matrix $\mathbf{Y}$ consists of binary variables describing the different categories (e.g. zeroes and ones in the case of two categories). The number of dependent variables is equal to the number of categories [34].

Partial least squares applied to the multivariate case ($q > 1$) is also known under the term PLS2, whereas the case $q = 1$ is denoted by PLS1. The aim of PLS2 regression is to find a linear relationship between the response and explanatory variables, using an $D \times q$ matrix $\mathbf{B}$ of regression coefficients, and an error matrix $\mathbf{E}$ [27, 96],

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}. \tag{3.13}$$

In PLS1 regression, formula (3.13) has the form $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$, where $\mathbf{b}$ are the regression coefficients and $\mathbf{e}$ is a vector of errors. The columns of $\mathbf{X}$ and $\mathbf{Y}$ are assumed to be mean-centered before parameter estimation is performed. Consequently, the absolute term parameters are omitted from further considerations.

Instead of directly estimating the regression coefficients in the relation (3.13), $\mathbf{X}$ and $\mathbf{Y}$ can be modeled by linear latent variables according to the regression models

$$\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E}_X \tag{3.14}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}' + \mathbf{E}_Y, \tag{3.15}$$

where $\mathbf{E}_X$ and $\mathbf{E}_Y$ are matrices of residuals. The matrices $\mathbf{T}$ and $\mathbf{U}$ represent score matrices and the matrices $\mathbf{P}$ and $\mathbf{Q}$ are loading matrices. All of these matrices have a columns, where $a \leq \min(D, q, n)$ is the number of PLS components [27, 35, 92] (to be chosen by the user). The scores in $\mathbf{T}$ are linear combinations of the explanatory variables and can be considered as good summaries. The same relationship holds for the response variables and the matrix $\mathbf{U}$ [27].

Then the relationship between the scores becomes

$$\mathbf{U} = \mathbf{T}\mathbf{D} + \mathbf{H}, \tag{3.16}$$

where $\mathbf{D}$ is a diagonal matrix with elements $\mathrm{d}_1, \ldots, \mathrm{d}_a$, and $\mathbf{H}$ is the residual matrix [27]. Since all quantities in Equation (3.16) are unknown (latent variable problem), the parameter estimation needs to be based on an additional criterion. In case of PLS2, this criterion is the maximization of the covariance between scores, corresponding to explanatory and response variables. The requirements of high (total) explained variance of $\mathbf{X}$ and high correlation between $\mathbf{X}$ and $\mathbf{Y}$ are both included in this criterion. Consider a weight vector $\mathbf{d}$ for the explanatory variables ($\mathbf{t} = \mathbf{X}\mathbf{d}$), and a weight vector $\mathbf{c}$ vector for the response variables ($\mathbf{u} = \mathbf{Y}\mathbf{c}$) [27, 92]. Then the maximization problem can be written as

$$cov(\mathbf{t}, \mathbf{u}) = cov(\mathbf{X}\mathbf{d}, \mathbf{Y}\mathbf{c}) \rightarrow \max_{\|\mathbf{t}\|=\|\mathbf{u}\|=1} . \tag{3.17}$$

The solution of the maximization problem is formed by the first score vectors $\mathbf{t}_1$ and $\mathbf{u}_1$, columns of the corresponding score matrices (their unit length is required for uniqueness of the solution). For the next score vectors we impose orthogonality constraints to the previous score vectors, i.e., $\mathbf{t}_j^{'}\mathbf{t}_l = 0$ and $\mathbf{u}_j^{'}\mathbf{u}_l = 0$ for $1 \leq j < l \leq a$ [27]. Finally, the score matrices $\mathbf{T}$ and $\mathbf{U}$ (together with matrices formed by weight vectors $\mathbf{d}$ and $\mathbf{c}$) are used for the estimation of the regression parameters $\mathbf{B}$.

There are several algorithms for solving the PLS problem. One proposal is based on nonlinear iterative partial least squares (NIPALS) algorithm. Another one - the Kernel algorithm - is named from using eigen-decompositions of so-called kernel matrices, being products of $\mathbf{X}$ and $\mathbf{Y}$. The SIMPLS algorithm avoids deflation steps at each iteration of PLS procedure. Orthogonal projection to latent structures (OPLS) aims at removing variation from $\mathbf{X}$ that is orthogonal to the response variables [35, 92, 97, 98]. OPLS will be discussed in Section 3.4.

Some variables from the matrix $\mathbf{X}$ can be important for the modeling of $\mathbf{Y}$. These variables have typically large absolute values of regression coefficients.

A summary of this importance of variables from $\mathbf{X}$ for both $\mathbf{Y}$ and $\mathbf{X}$ are provided by VIP scores (variable importance in the projection) [94, 99]. VIP scores for the case of PLS1 will be defined, therefore we have only vector $\mathbf{y}$ and the formula (3.15) can be rewritten as $\mathbf{y} = \mathbf{Tv} + \mathbf{f}$ [99], where $\mathbf{v}$ is the vector of regression coefficients of the matrix $\mathbf{T}$. Projections are done for a latent variables. Let's have $k = 1, \ldots, a$, $\mathbf{t}_k$ stands for the k-th column of the matrix $\mathbf{T}$. VIP for the j-th explanatory variable overall latent variables measures the contribution of each predictor variable to the model by taking into account the covariance between $\mathbf{X}$ and $\mathbf{y}$, expressed as weight w_{jk}, which is obtained by the use of NIPALS algorithm, [99, 100]

$$\mathrm{VIP}_j = \sqrt{D \sum_{k=1}^{a} (v_k^2 \mathbf{t}_k' \mathbf{t}_k) w_{jk}^2 / \sum_{k=1}^{a} v_k^2 \mathbf{t}_k' \mathbf{t}_k}, \quad j = 1, \ldots, D, \tag{3.18}$$

where $v_k = \mathbf{t}_k^{'} \mathbf{y}_{(k)} / \mathbf{t}_k^{'} \mathbf{t}_k$ is obtained for each column of the score matrix $\mathbf{T}$, $\mathbf{y}_{(k)}$ is vector $\mathbf{y}$ for the k-th latent variable from the NIPALS algorithm (for further details see [99]). The average of squared VIP scores is equal to one, therefore VIP scores greater than one are often chosen as important variables [99]. This is not a statistically justified limit and can be shown to be very sensitive to the presence of non-relevant information referring to $\mathbf{X}$ [100].

In the following text is described, how PLS-DA can be used in the case of compositional explanatory variables [6]. The algorithm is optimized for the balanced case. This means that there is the same amount of members in each category. We will employ the approach of [101] which uses the clr coordinates for the representation of compositional covariates but tries to avoid the resulting additional zero constant sum constraint by using the orthonormal (ilr) coordinates. Let $\mathbf{X}$ be an $n \times D$ matrix of compositional data (sampled compositional parts $x_1, \ldots, x_D$) and $\mathbf{Y}$ be an $n \times q$ matrix of responses representing the groups. As in the standard case, the columns of $\mathbf{Y}$ are mean-centered. However, mean-centering of the compositions $\mathbf{X}$ is done with respect to the Aitchison geometry, i.e. the centering is performed in ilr coordinates. Concretely, compositions are expressed in ilr coordinates, e.g. (1.9), using, the resulting variables are mean-centered and then transformed back to the original space with (1.10).

Following the case of linear regression with compositional explanatory variables (Section 1.3, [71]), where applying the clr coordinates leads to a biased estimation of the regression coefficients due to the singular covariance matrix of the clr variables, the ilr coordinates may be used for the purpose of PLS modeling. Subsequently, the matrix $\mathbf{X}$ is fisrtly expressed in ilr coordinates $\mathbf{Z}$, e.g., using Equation (1.9). The PLS regression problem has now the form

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{E}, \tag{3.19}$$

where $\boldsymbol{\Gamma}$ stands for a $(D-1) \times q$ matrix of regression coefficients.

Nevertheless, the ilr coordinates (1.9) allows only for a meaningful interpretation of the elements in the first row of $\boldsymbol{\Gamma}$, because just the first column of $\mathbf{Z}$ can be associated with one particular compositional part (here x_1). The interpretation of the other regression coefficients is not straightforward because the corresponding explanatory variables (coordinates) do not fully represent one particular part of the composition. For associations also to the other parts, we thus need to use a permutation of the parts, leading to the general setting (1.11) and to data matrices $\mathbf{Z}^{(l)}$. Each of the resulting first ilr coordinates, the observations of $z_1^{(l)},\ l = 1, \ldots, D$, describe all the relative information about the compositional part x_l.

Consequently, a possible way to evaluate the contribution of each compositional part for explaining the response variables $\mathbf{Y}$ separately is to consider D PLS regression models

$$\mathbf{Y} = \mathbf{Z}^{(l)}\boldsymbol{\Gamma}^{(l)} + \mathbf{E}^{(l)}, \tag{3.20}$$

according to (1.11), by taking $l \in \{1, \ldots, D\}$, and to interpret the coefficients of the first row of the parameter matrix $\boldsymbol{\Gamma}^{(l)}$, representing the part $x_1^{(l)}$ [71]. The outlined procedure thus suggests employing PLS regression D times, such that each compositional part is once at the first position in the permuted composition. Since such a procedure would lead to a high computational complexity, the orthogonal relation between the different ilr coordinates can be employed [65]. As an advantage, the regression coefficients need to be estimated just for one regression model and then derived for the other models by using orthogonal transformations of the regression parameters. Note, however, that coefficients of $z_1^{(l)}$ should be always treated individually as they come from individual PLS models.

The final procedure is as follows. We use the matrix $\mathbf{V}$ from (1.12), with rows $\mathbf{v}_{i\cdot}, i = 1, \ldots, D$, that relates the clr coordinates (1.8) and the ilr coordinates (1.9). Consequently, we form $D \times (D-1)$ matrices $\mathbf{V}^{(l)}$, for $l \in \{1, \ldots, D\}$,

$$\begin{aligned}\mathbf{V}^{(1)} &= (\mathbf{v}_{1\cdot}, \mathbf{v}_{2\cdot}, \ldots, \mathbf{v}_{D-1,\cdot}, \mathbf{v}_{D\cdot})' = \mathbf{V}\\ \mathbf{V}^{(l)} &= (\mathbf{v}_{l\cdot}, \mathbf{v}_{l-1,\cdot}, \ldots, \mathbf{v}_{1\cdot}, \mathbf{v}_{l+1,\cdot}, \ldots, \mathbf{v}_{D\cdot})',\ l = 2, \ldots, D-1;\\ \mathbf{V}^{(D)} &= (\mathbf{v}_{D\cdot}, \mathbf{v}_{D-1\cdot}, \ldots, \mathbf{v}_{2\cdot}, \mathbf{v}_{1\cdot})',\end{aligned}$$

and define a new orthogonal matrix $\mathbf{Q}^{(l)}$,

$$\mathbf{Q}^{(l)} = \mathbf{V}'\mathbf{V}^{(l)}. \tag{3.21}$$

The matrices $\mathbf{Z}^{(l)}$ corresponding to ilr coordinates (1.11) are related to $\mathbf{Z}$ by

$$\mathbf{Z}^{(l)} = \mathbf{Z}\mathbf{Q}^{(l)}, \tag{3.22}$$

see [65]. Substituting (3.22) into the model (3.19) gives

$$\mathbf{Y} = \mathbf{Z}^{(l)}(\mathbf{Q}^{(l)})'\boldsymbol{\Gamma} + \mathbf{E}^{(l)} = \mathbf{Z}^{(l)}\boldsymbol{\Gamma}^{(l)} + \mathbf{E}^{(l)}. \tag{3.23}$$

Thus, the estimated regression coefficients $\boldsymbol{\Gamma}$ from the model (3.19), $\widehat{\boldsymbol{\Gamma}}$, can be used to estimate coefficients in regression models that correspond to coordinates $\mathbf{Z}^{(l)}$,

$$\widehat{\boldsymbol{\Gamma}}^{(l)} = (\mathbf{Q}^{(l)})'\widehat{\boldsymbol{\Gamma}}, \quad l = 1, \ldots, D. \tag{3.24}$$

Finally, to complete the estimation process with respect to the above interpretation, we collect the first rows of the matrices $\widehat{\boldsymbol{\Gamma}}^{(l)}$ as rows of a new $D \times q$ matrix of regression coefficients. Specially, for $q = 1$ (the response variable is univariate) we thus, get a vector $\mathbf{g} = (\widehat{\gamma}_1^{(1)}, \widehat{\gamma}_1^{(2)}, \ldots, \widehat{\gamma}_1^{(D)})'$.

3.3.2. Evaluation

A further evaluation of the resulting regression model can be done by testing for significance of the regression parameters. For PLS-DA, it is common to use resampling techniques for this purpose, like the jack-knife procedure [102,103] or bootstrap [104]. In the balanced case, jack-knife works as follows. One observation from each group is taken out from the data matrix $\mathbf{X}$ and from the matrix $\mathbf{Y}$. Then the regression parameters are estimated from PLS-DA. This process is carried out in turn by omitting another observation from each group in both data sets. Finally, the variability of the regression parameters is evaluated by their standard deviation. The process for the unbalanced case is basically the same [104, 105]. The idea of the bootstrap procedure is to draw random samples with replacement from each group of the original data, where the bootstrap group samples have the same size as the original groups. This results in a bootstrap data set for the explanatory variables and the response, where PLS-DA is applied to estimate the parameters. Repeating this procedure many times allows estimating the variability of the regression parameters [104,106]. The standardized regression estimates are then obtained by dividing the regression parameters of the original data by the estimated standard deviations (obtained from jack-knife or bootstrap), and they can be compared with quantiles of the standard normal distribution. For the case $q = 1$ and the estimated parameters $\mathbf{g} = (\widehat{\gamma}_1^{(1)}, \widehat{\gamma}_1^{(2)}, \ldots, \widehat{\gamma}_1^{(D)})'$, the estimates are recomputed using jack-knife or bootstrap, and from the results the standard deviations $s_1, \ldots, s_D$ are computed. The significance of the standardized regression estimates, $\widehat{\gamma}_1^{(1)}/s_1, \widehat{\gamma}_1^{(2)}/s_2, \ldots, \widehat{\gamma}_1^{(D)}/s_D$, is evaluated by comparing them with $\alpha/2$ and $1 - \alpha/2$ quantiles of the standard normal distribution (typically, $\alpha = 0.05$ is chosen). In order to reduce the risk of false positives, a Bonferroni correction is applied, resulting in an adjusted α-level of significance, $\alpha_{adj} = \frac{\alpha}{D}$, that is used further in Section 3.3.4. If the standardized regression coefficient is outside the mentioned interval, the regression coefficient is significantly different from zero, and thus, the corresponding variable contributes to the discrimination task.

3.3.3. Estimation of the number of components

An appropriate estimation of the number of PLS components that avoids underfitting as well as overfit is essential, for example for the performance of an imputation algorithm (Section 3.5).

Two measures of evaluation of optimal number of components are shown in this thesis. The first one, simpler, is based on the mean squared error of prediction (MSEP) criterion. The principle of the method follows. The data set for $q = 1$ is denoted as $L = \{(y_i, x_{i1}, x_{i2}, \dots, x_{iD}),\ i = 1, \dots, n\}$. L is divided randomly in K segments $L_k, k = 1, \dots, K$, of roughly equal size. Let f_K be the predictor from L/L_k (all observation are not from L_k). The K-fold cross-validation estimate is [107]

$$\mathrm{MSEP} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i \in L_k}(f_k(\mathbf{x}_i) - y_i)^2.$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{iD})$. The bias of MSEP is of order $(K-1)^{-1}n^{-1}$.

In practice, partial least squares regression is applied on data, using $1, \dots, l$ components, and MSEP is computed for each number of components, using $K = 10$. The minimum of these MSEPs is chosen, and the optimal number of components is the number corresponding with this minimum. The estimation of the number of components by MSEP is used in the practical example in Section 3.3.4.

The second algorithm for the estimation of the number of components is similar to ideas of [108], where a bootstrap procedure was taken. This procedure is used in the imputation algorithm is Section 3.5.1.

1. Based on a sample of $i = 1, \dots, n$ observations $(y_i, x_{i1}, x_{i2}, \dots, x_{iD}),\ R$ bootstrap data sets, each consisting of n samples with replacement, are taken jointly from the pairs of response and predictors, resulting in the paired data sets $(\mathbf{y}_r^*, \mathbf{X}_r^*)$, for $r = 1, \dots, R$.

2. Partial least squares regression is applied to each pair, using $1, \dots, k$ components. The predicted error sum of squares (PRESS) criterion is computed, using a 10-fold cross validation procedure. PRESS is the sum of squares of the prediction errors, where each fitted value, $\hat{y}_{i,-i}$, is obtained from the remaining n–1 observations, then using the fitted regression function to obtain the predicted value for the ith observation [107]:

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{i,-i})^2.$$

3. For each number of components, the arithmetic mean of the PRESS values overall bootstrap samples is calculated. The minimum of these arithmetic means is chosen, and the standard deviation of the PRESS values is calculated for that number of components determining this minimum. A threshold for the imputation of rounded zeros is fixed given by this minimum plus one standard deviation.

4. The final PLS model is determined with the smallest number of components, for which the mean PRESS value is still below the threshold. This ensures the selection of a parsimonious model that is not significantly worse than the possibly larger model with the smallest cross-validation prediction error.

The above mentioned procedure to find the optimal number of PLS components is illustrated in Figure 3.4. The plot represents prediction errors for 100 bootstrap samples based on a simulated data set (100 variables). Shown are the prediction errors (PRESS) for the bootstrap samples for different numbers of PLS components (the results are connected by the gray lines). The parallel boxplots show the variability of these errors for each number of components. The lower horizontal line is placed at the minimum of the mean PRESS values (marked by white-filled quadrangles) for each number of components. The upper horizontal line determines the threshold given by this minimum plus one standard deviation of the bootstrap results corresponding to the number of components at

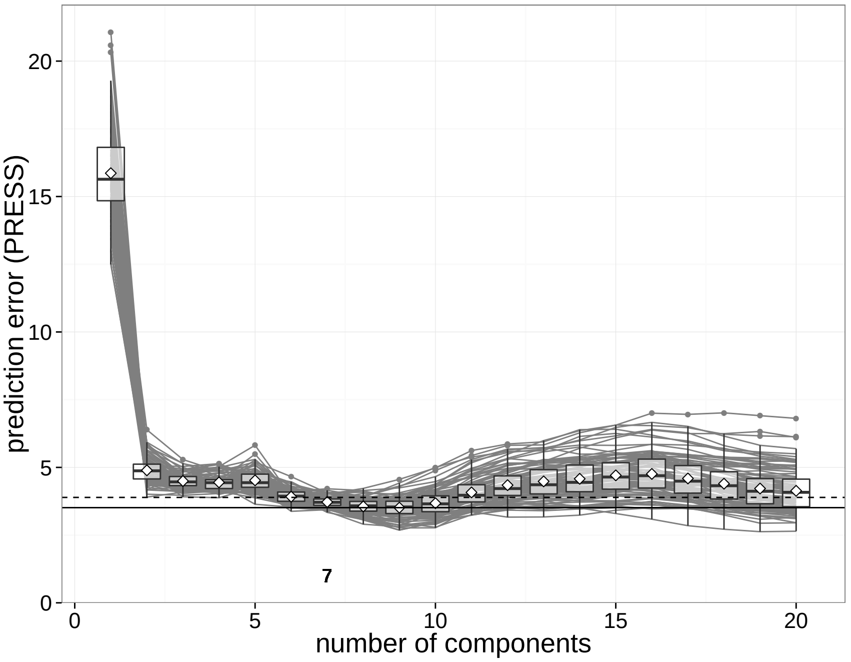


Figure 3.4: Simulation of a bootstrap procedure used for the estimation of the optimal number PLS components.

the minimum. The most parsimonious model, here seven PLS components, for which the mean is below this threshold is selected [58].

3.3.4. Example with data from metabolomics

In the following text, PLS-DA algorithm (accommodated for compositional data) is applied to a real data set from metabolomics connected with inherited metabolic disorders [6]. The problem is related to untargeted analysis of a disease named medium chain acyl-CoA dehydrogenase deficiency (MCADD) [7]. Based on newborn screenings from reports from Australia, Germany and the USA [109], fatty acid oxidation disorders are one of the most prevalent groups of metabolic diseases. At least 15 different disorders of fatty acid metabolism are recently known [110]. MCADD is one of the most common fatty oxidation defects and it is inherited in the autosomal recessive trait. The main marker is octanoylcarnitine (C8) and secondary markers are hexanoylcarnitine (C6), decanoylcarnitine (C10), and decenoylcarnitine (C10:1) [111].

In this application, we examine dry blood spots which were obtained from a screening program of newborns. The data set contains a group of healthy controls ($n = 23$) and a group of patients suffering from an MCAD deficiency ($n = 23$). Quality control samples are used.

The result of the outlined procedure is a table of peak areas (areas of possible metabolites); more than 500 peaks are detected by the untargeted method. These peaks represent possible metabolites, which might be important markers of the disease. Raw data from Orbitrap Elite are processed by Bioconductor packages *xcms* [45–47] and *CAMERA* [48]. *xcms* is used for peak detection and alignment across the samples. The *CAMERA* package is used to exclude isotopic patterns. In order to remove possible systematic errors, LOESS signal correction is applied on quality control samples [11, 12]; no scaling of data is performed. Rounded zeros occur in the data set and they are replaced by 2/3 of the minimum value from a particular group (patients/controls) in single metabolites (without any adjustment of the non-zero metabolite intensities). A small number of zeros occurred in the concrete example (4.5%). Now we proceed to analyze the peaks for statistical significance. Because not the absolute values of peak areas, but rather their relative contributions are of interest, they represent compositional data and should be handled accordingly. Nevertheless, for the sake of comparison, both the standard and the compositional PLS-DA method are performed.

The MSEP for the standard approach is displayed in Table 3.1. From this table, it is visible that the optimal number of PLS components equals to 3.

MSEP	1 comp	2 comp	3 comp	4 comp	5 comp	6 comp
Standard approach	0.0461	0.0310	0.0284	0.0288	0.0285	0.0321
Logratio approach	0.0474	0.0182	0.0158	0.0139	0.0122	0.0128

Table 3.1: Mean squared error of prediction (MSEP) for the MCADD data, for different numbers of PLS components for the standard and the logratio approach.

The significance of the standardized regression coefficients is analyzed using

PLS-DA with three components and bootstrap with 100 replications [6]. The results are displayed in Figure 3.5. From this plot, it is visible that almost one fifth of the corresponding regression coefficients are marked as significant (they are located above/below the cut-off line, represented by the Bonferroni-corrected quantile of the standard normal distribution). The most significant peaks for patients (situated above the line - in the upper part of the graph) are denoted with codes UL1, U1 and UL2. The first and the third are unknown lipids. Within this study, the second is the other unknown metabolite. These three possible metabolites need further research. Known markers of the MCADD are C6 located on the 13th position, C8 on the 16th position, C10 on the 19th position and C10:1 on the 21th position. These results could be considered as satisfactory, because the known markers are placed on the upper positions in the resulting significance plot.

As a second step, the compositional approach to PLS-DA is applied, including bootstrap with 100 replications in order to analyze the significance of the regression parameters. The resulting MSEP is displayed in Table 3.1. From the second line of this table, it is visible that MSEP has smaller values in the case of the logratio approach. Although the appropriate number of components is 5, the same number 3 of components as for the standard case is used for the comaprison of both approaches. Figure 3.6 shows the results of the significance analysis of the regression coefficients. The structure of the outcome is similar to the standard case; approximately one fifth of the corresponding regression coefficients are marked as significant (with the same interpretation of this feature as in the first plot). Nevertheless, the compositional approach is more sensitive and returns better results with respect to the known markers of the disease. In particular, the above mentioned acylcarnitines C8 (the main marker of the disease) and C6 are on the second and the third position in the graph. The first position is occupied by an unknown marker denoted as U2. This feature also needs further research. Also C10:1 (10th) and C10 (25th) are higher compared to the standard approach to PLS-DA. According to the MSEP, the proper number of PLS components is

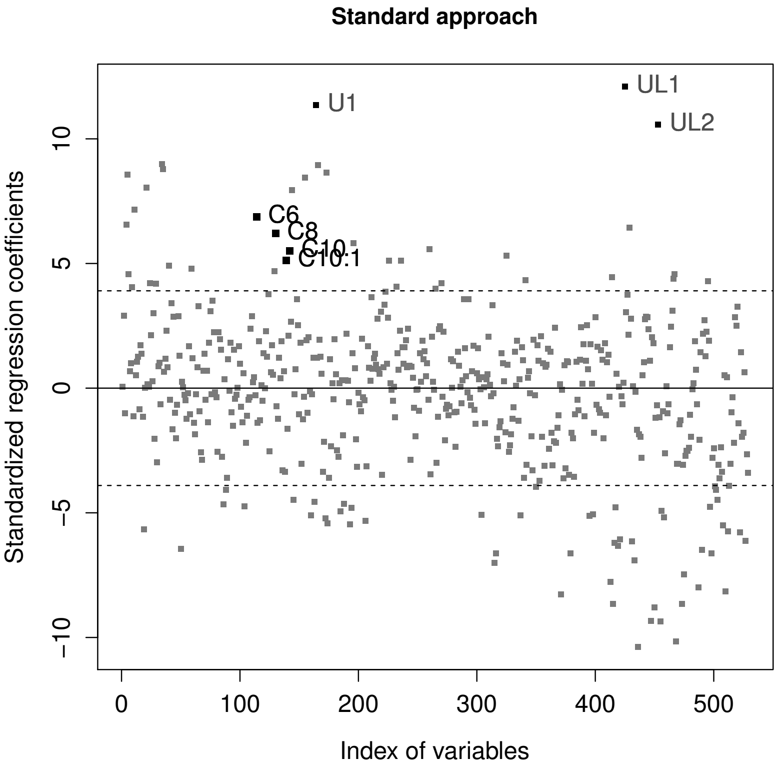


Figure 3.5: Standardized regression coefficients in PLS-DA for the original MCADD data set (after centering).

five in the logratio approach. Looking at the resulting outputs for 5 components (the figure is skipped in this thesis) we could conclude that the results would be even better than in Figure 3.6. Marker C8 is moved to the first position and all markers are placed higher above the cut-off line than in Figure 3.6. This confirms once again our preliminary finding that the logratio approach to compositional (metabolomics) data analysis better detects important peaks, related to the particular disease.

As a result, the compositional PLS-DA procedure turned out to be more accurate in identifying significant metabolites. Based on these experiments, PLS-DA seems to be more reliable and accurate with respect to expert knowledge for the compositional approach [6].

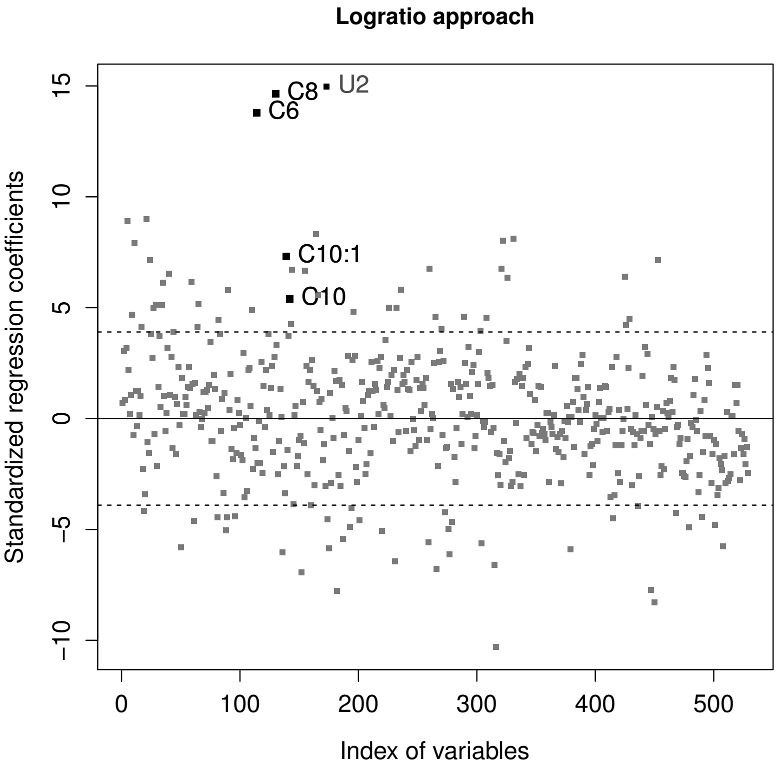


Figure 3.6: Standardized regression coefficients in PLS-DA for the original MCADD data set with application of compositional approach.

3.4. Orthogonal partial least squares regression

Orthogonal partial least squares regression (OPLS) is a modification of the PLS regression [35]. Simultaneously as PLS, it belongs to the group of supervised methods. The known information about data (e.g. the separation of observations to groups) is contained in the matrix **Y**.

The idea of OPLS is to separate the systematic variation in data matrix **X** into two parts. The first one is linearly related to **Y** and represents between class variation, the second one is unrelated (orthogonal) to **Y** and refers to as the uncorrelated variation, which forms the within class variation [35, 36]. The matrix **Y** is connected to the additional information provided by the matrix **X**. For classification purposes, OPLS is often called orthogonal partial least squares -

discriminant analysis (OPLS-DA).

The model includes two modeled variations - the Y-predictive ($\mathbf{T}_P\mathbf{P}'_P$) and the Y-orthogonal ($\mathbf{T}_O\mathbf{P}'_O$) components. Only the first one is used for the modeling of $\mathbf{Y}$ [35]

$$\mathbf{X} = \mathbf{T}_P\mathbf{P}'_P + \mathbf{T}_O\mathbf{P}'_O + \mathbf{E}_X, \tag{3.25}$$

$$\mathbf{Y} = \mathbf{T}_P\mathbf{C}'_P + \mathbf{E}_Y, \tag{3.26}$$

where $\mathbf{E}_X$ and $\mathbf{E}_Y$ are residual matrices of $\mathbf{X}$ and $\mathbf{Y}$, respectively, $\mathbf{T}$ is the score matrix and $\mathbf{P}$ is the loading matrix.

OPLS-DA can be used in the case of compositional explanatory variables, where ilr coordinates (1.11) are taken to obtain the matrix of covariate values $\mathbf{Z}^{(l)}, l = 1, \ldots, D$. Accordingly, formulas (3.25) and (3.26) change to

$$\mathbf{Z}^{(l)} = \mathbf{T}_P^{(l)}\mathbf{P}_P^{(l)'} + \mathbf{T}_O^{(l)}\mathbf{P}_O^{(l)'} + \mathbf{E}_{Z^{(l)}}^{(l)}, \quad l = 1, \ldots, D, \tag{3.27}$$

$$\mathbf{Y} = \mathbf{T}_P^{(l)}\mathbf{C}_P^{(l)'} + \mathbf{E}_Y^{(l)}, \quad l = 1, \ldots, D. \tag{3.28}$$

Both methods, PLS-DA and OPLS-DA, have its own properties. The between class variation and the within-class variation are separated by OPLS-DA but not by PLS-DA. Both have been used for modeling two classes of data to increase the class separation, simplify interpretation and find potential biomarkers. For the two-class problem, OPLS-DA is recommended to obtain a clearer and more straightforward interpretation. It can also provide an understanding of the interclass variation [35]. The advantage of OPLS-DA is also that the model is rotated so that class separation is found in the first predictive component, also referred to as the correlated variation, and variation not related to class separation is seen in orthogonal components, also referred to as the uncorrelated variation. This separation of predictive and orthogonal components facilitates model interpretation [36].

The difference between PLS-DA and OPLS-DA methods is also visible from Figure 3.7 (the picture comes from [36]). Squares and circles denote two different groups of samples (for example patients and controls). The OPLS-DA model is rotated; thanks to this fact, the between class variation (the difference between patients and controls) is found in the predictive component $\mathbf{t}_p$ and within class variation is visible from the first y-orthogonal component $\mathbf{t}_o$ [36].

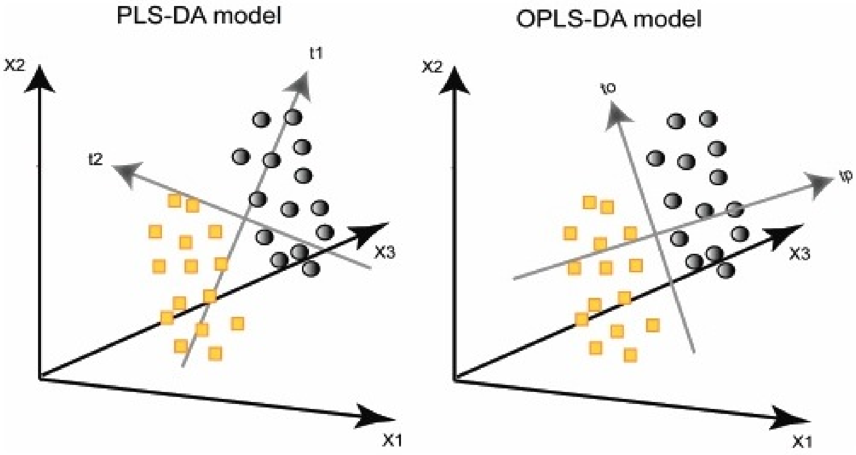


Figure 3.7: The difference between PLS-DA and OPLS-DA methods [36].

The possibility, how to visualize results of OPLS-DA, is a special graph called S-plot. It visualizes the covariance and correlation between variables and the modeled class designation. Accordingly, the influence in the model is captured. The S-plot helps by identifying statistically significant and potentially biochemically significant metabolites, based both on contributions to the model and their reliability [36]. The S-plot is a scatter plot that combines the covariance (the contribution of the magnitude of model component scores) and correlation (the reliability of model component scores) loading profiles resulting from a projection-based model. The S-plot uses these two vectors in the scatter structure [36]

$$Cov(\mathbf{t}, \mathbf{x}_i) = \frac{\mathbf{t}'\mathbf{x}_i}{N-1}, \tag{3.29}$$

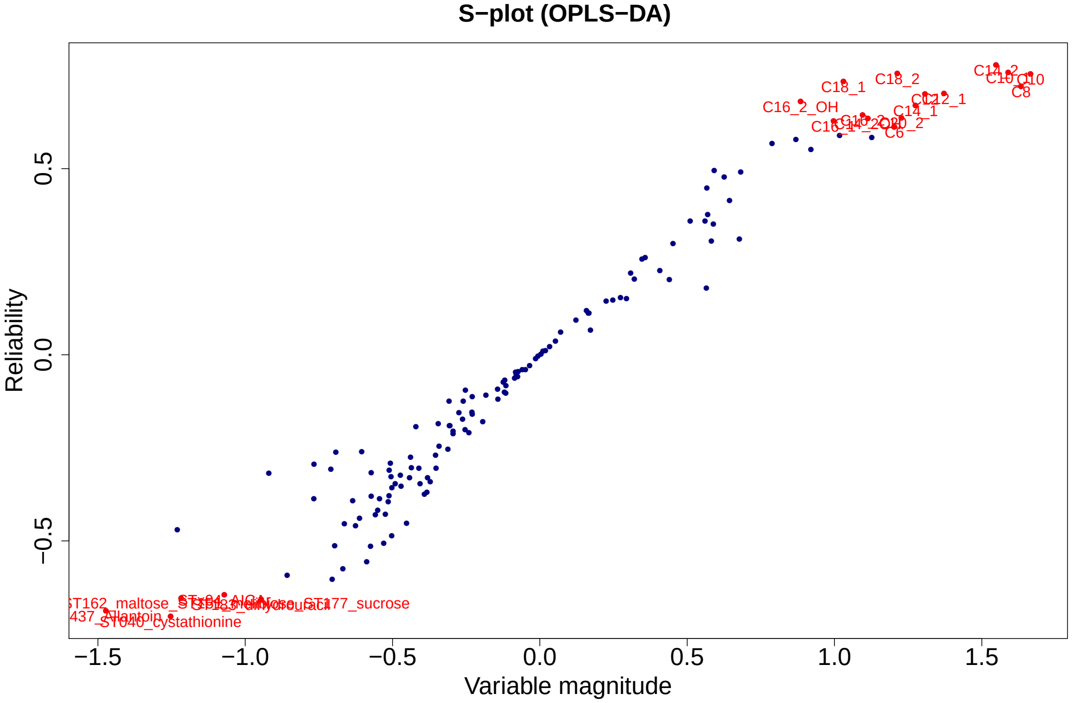


Figure 3.8: The S-plot.

$$Corr(\mathbf{t}, \mathbf{x}_i) = \frac{Cov(\mathbf{t}, \mathbf{x}_i)}{s_t s_{\mathbf{x}_i}}, \tag{3.30}$$

where $\mathbf{t}$ is the score vector in the model, $\mathbf{x}_i$ denotes the centered variable from the data matrix $\mathbf{X}$ (column of the matrix $\mathbf{X}$) and s stands for an estimate of the standard deviation. The name S-plot comes from the shape of the graph. In the optimal situation, the scatter plot should have the shape of the letter S. Metabolites which are in the left lower corner and right upper corner are denoted as the most significant. The S-plot is used for the comparison of only two groups of samples, where the significant metabolites distinguish between these two groups. A complimentary tool for identification of interesting compounds is to plot the loading vector, $Cov(\mathbf{t}, \mathbf{X})$, with its corresponding jack-knifed confidence intervals as these provide additional information about metabolite variability [36].

The example of the S-plot is shown in Figure 3.8. This plot was used for the OPLS-DA analysis of two groups of samples (patients and controls). From the point of view of reliability, the importance of metabolites is evaluated mainly from y-axis. First twenty significant metabolites are highlighted by red color with their names. Blue points represent the rest of metabolites. We can compare these results with the other graphs and try to find important markers of the disease.

3.5. Parametric models for imputation of rounded zeros

3.5.1. Theoretical background for imputation model

As mentioned in Section 3.1.2, only a few algorithms exist for the imputation of rounded zeros in high-dimensional compositional data sets. The crucial point for building up a reasonable imputation procedure is to find interpretable orthonormal coordinates in order to enable further processing in the standard Euclidean geometry. Since there is no canonical basis on the simplex, a set of orthonormal coordinate systems (1.11) needs to be employed sequentially in order to perform the imputation for each of the original compositional parts. The procedure needs to be able to capture both the relative information, conveyed by the compositional data themselves, and the absolute nature of the corresponding detection limits, for a meaningful imputation of rounded zeros.

Based on previous considerations and following the structure of the imputation procedure in [18], an iterative regression-based algorithm for the replacement of rounded zeros is introduced in Algorithm 3.5.1. It is based on PLS regression, introduced in Section 3.3 with the evaluation of the optimal number of components (based on PRESS values) announced in Section 3.3.3, thus able to cope also with high-dimensional compositional data sets [58]. In addition, some notation is given beforehand.

To avoid complicated notation in Algorithm 3.5.1, we assume that $\mathcal{M}(\mathbf{x}_1) \geq \mathcal{M}(\mathbf{x}_2) \geq \ldots \geq \mathcal{M}(\mathbf{x}_D)$, with $\mathcal{M}(\mathbf{x}_j)$ denoting the number of rounded zero cells in variable $\mathbf{x}_j$. Denote $m_l \subset \{1, \ldots, n\}$ the indices of the rounded zeros in variable

Algorithm 3.5.1 PLS

```
for $j \in \{1, ..., D\}$ do                                ▷ INITIALIZATION OF ROUNDED ZEROS
    Initialize all $\mathbf{x}_{ij}$, $i \in m_j$ with 2/3 of the corresponding detection limit.
end for
Sort variables based on $\mathcal{M}(\mathbf{x}_1) \geq \mathcal{M}(\mathbf{x}_2) \geq \ldots \geq \mathcal{M}(\mathbf{x}_D)$. For easier notation, we
assume that the variables are already sorted.                               ▷ SORTING
Let $c$ be large, e.g. $c = 9999999$, and $\epsilon$ small, e.g. $\epsilon = 0.1$, set $r = 1$.
function ESTIMATE THE OPTIMAL NUMBER OF COMPONENTS
    Run the function REGRESSION from below to determine the optimal number
    of components (see Section 3.3.3) for each variable including rounded zeros
                                          ▷ INITIALIZATION OF NUMBER OF COMPONENTS
end function
while $c > \epsilon$ do
    $r \leftarrow r + 1$
    for $l \in \{1, ..., D\}$ do
        function COORDINATE
            Take $\mathbf{X}^{(l)}$ ($l$-th variable at first position) with elements $x_{ij}^{(l)}$;
            compute coordinate representation $\mathbf{z}_1^{(l)}$ and $\mathbf{Z}_{-1}^{(l)}$.
            Let $e_l$ be the detection limit of the $l$-th part; compute coordinates
                $\psi_i^{(l)} = \sqrt{\frac{D-1}{D}} \ln \frac{e_l}{\sqrt[D-1]{\prod_{j=2}^{D} x_{ij}^{(l)}}}$    for    $i \in m_l$.    (3.31)
                                               ▷ REPRESENTATION IN COORDINATES
        end function
        function REGRESSION
            With previously estimated optimal number of components, estimate the
            regression coefficients $\boldsymbol{\beta}$ with PLS regression:
            $\mathbf{z}_1^{(l)} = \mathbf{Z}_{-1}^{(l)}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$   with   $\mathbf{Z}_{-1}^{(l)} = \mathbf{T}\mathbf{P}^T$          ▷ PLS REGRESSION
        end function
        function Replacement
            Use the estimated regression coefficients $\hat{\boldsymbol{\beta}}$ to impute the rounded zeros:
                $\hat{z}_{i1}^{(l)} = \hat{\boldsymbol{\beta}}^T \mathbf{z}^{(l)}{}_{i,-1} - \hat{\sigma}\frac{\phi\left(\frac{\psi_i^{(l)}-\hat{\boldsymbol{\beta}}^T\mathbf{z}^{(l)}{}_{i,-1}}{\hat{\sigma}}\right)}{\Phi\left(\frac{\psi_i^{(l)}-\hat{\boldsymbol{\beta}}^T\mathbf{z}^{(l)}{}_{i,-1}}{\hat{\sigma}}\right)} \quad \text{for } i \in m_l,$    (3.32)
            corresponds to the rounded zeros in $\mathbf{z}_1^{(l)}$, and $\phi$ and $\Phi$ are density and distribution function of the standard normal distribution, respectively; $\hat{\sigma}$ is the estimated conditional standard deviation of variable $\mathbf{z}_1^{(l)}$.    ▷ REPLACEMENT
        end function
        function Inverse mapping
            Use Equation (1.10) to express back in the original sample space; reorder the variables.
            The values that were originally rounded zeros in the cells $m_l$ in variable $\mathbf{x}_l$ are updated.    ▷ INVERSE MAPPING
        end function
        function Re-Scaling
            Due to the nature of this inverse mapping, the scale of variables is changed. Call $M_i$ the set with the cells of the $i$-th observation that were rounded zeros, and $O_i = \{1, \ldots, D\} \backslash M_i$. A cell $x_{ij}$, for any $j \in M_i$, is adjusted (multiplied) by the factor $f_{ij} = \frac{\sum_{o \in O_i} x_{io}}{\sum_{o \in O_i} \hat{x}_{io}}$, where $\hat{x}_{io}$ denote the inverse mapped values from the previous step.    ▷ ADJUSTMENT
        end function
    end for
    function Update criteria
        Update $c$ as the sum of squared differences of the elements of $\mathbf{X}$ in the $r$-th and the $(r-1)$-th iteration.
    end function
end while
Bring the variables to the original order    ▷ UNDO SORTING
```

$\mathbf{x}_l$, and $o_l = \{1, \ldots, n\} \backslash m_l$ the indices corresponding to the remaining cells of $\mathbf{x}_l$. Denote $\mathbf{z}_1^{(l)}$ as the first coordinate according to (1.11), and $\mathbf{Z}_{-1}^{(l)}$ containing the remaining $D-2$ coordinates. The first column of $\mathbf{Z}_{-1}^{(l)}$ consists of ones, taking care of an intercept term in PLS regression, and the observations (rows) are denoted by $\mathbf{z}_{i,-1}^{(l)}$, for $i = 1, \ldots, n$.

Note that due to the complexity of the above algorithm, a rigorous proof of convergence is not available. Nevertheless, our practical experience shows that usually just a few iterations are necessary to reach the convergence criterion.

3.5.2. Modification with variation matrix

The second possible algorithm for the imputation of rounded zeros in high-dimensional data makes use of the variation matrix [1] for selecting variables to reduce the dimension of the data. A slightly modified algorithm of [18] is then used to replace rounded zeros.

The covariance structure of compositions is described by the variation matrix $\mathbf{T}$ [1, 3]. The entries of this matrix are variances of log-ratios of two-part sub-compositions, $t_{jk} = \text{var}\big(\ln \frac{x_{ij}}{x_{ik}}\big)$, for $i = 1, \ldots, n$, and $j, k = 1, \ldots, D$. Here, “var” denotes the empirical variance. Low values in the variation matrix indicate strong association between the parts in terms of their proportionality. When replacing rounded zeros in a particular compositional part, an optimal prediction model with a subcomposition of the remaining variables is identified, using a ranking from the variation matrix elements. The number of predictor variables in the model is kept low, and thus, the rounded zeros imputation is based on ordinary least-squares (OLS) regression. For more details see [58].

3.5.3. Alternative approaches

The available methods for rounded zeros imputation are collected in the R-package *zCompositions* [50]. These methods will be briefly introduced and employed in the simulation study in Section 3.5.5. Some of these methods are not able to work with high-dimensional data, but they are useful in the simulation part.

One method used for the imputation of rounded zeros is called multiplicative replacement (mult repl), which imputes left-censored compositional values by a given fraction of the corresponding detection limit. The default fraction is 0.65 times the detection limit of a variable. The multiplicative adjustment is applied in such a manner that the row-wise sums are made equal to the original values including rounded zeros whenever the data are in closed form, i.e. if they have to sum up to a constant. In this case, the absolute values are not preserved. The multiplicative replacement does not modify the original values above the detection limit if the data are not presented in a closed form [58].

Also multiplicative log-normal replacement (mult lognorm) is used for the imputation, where [112] consider the univariate log-odds for the i-th variable (for values above detection limit). They model the compositions using a multiplicative logistic normal mixture for this purpose.

Multiplicative Kaplan-Meier smoothing spline replacement (mult KMSS) is another way how to impute rounded zeros. This method replaces left-censored rounded zeros by averaging (geometric mean) random draws from a cubic smoothing spline fit. This spline is fit to the inverse Kaplan-Meier empirical cumulative distribution function to values below the corresponding limit of detection or censoring threshold, and the values below detection limit are replaced by the fitted values. Note that this method works in a univariate manner, applied independently to each compositional part containing values below detection limit. However, afterward multiplicative adjustment is applied to preserve the multivariate compositional properties of the samples. Unfortunately, the implementation in the R-package *zCompositions* [50] frequently leads to errors when the smoothing spline is fit, and is thus, problematic in use with high-dimensional data.

Log-ratio data augmentation algorithm (lr da) is simulation-based data augmentation algorithm and it uses $D-1$ additive log-ratio coordinates [1] for the representation of a D-part compositional data set and an MCMC (Markov chain Monte Carlo) approach. Consequently, left-censored compositional parts are imputed by simulated values from their posterior predictive distributions. A common conjugate normal inverted-Wishart distribution with non-informative prior is assumed for the model parameters in the coordinate space [113].

The last option for the imputation of rounded zeros, presented in this thesis, is called additive log-ratio EM algorithm (lr em). As for the data augmentation algorithm, this method expresses compositional data in additive log-ratio coordinates, where an EM algorithm is applied sequentially, i.e. $D-1$ regressions are performed for one iteration. This method thus does not work in situations with more variables than observations. The main difference to approaches like [15] and [18] is that just one coordinate system is taken for the whole iteration process and when convergence is reached, the coordinates are expressed back in the original space. For this reason, the approach needs, at least, one compositional part without rounded zeros. Finally, a correction factor based on the residual covariance obtained by censored regression is applied.

3.5.4. Validation criteria

The numerical properties of the proposed algorithm are investigated in the following for two kinds of simulated data and for a data set from metabolomics. In order to compare with other approaches, some evaluation criteria are introduced [15, 18].

The validation criteria defined in this section [18] assume that the complete data information is available. After introducing rounded zeros, a comparison can be made with the imputed data sets. The imputed data set is denoted by the symbol *.

Average difference in covariance structure (ADCS)

Let $\mathbf{S} = [s_{ij}]$ be the sample covariance matrix of the original observations in ilr coordinates z_{ij} and $\mathbf{S}^* = [s^*_{ij}]$ denote the sample covariance matrix computed with the same ilr observations where all the rounded zeros have been imputed. The measure of the average difference between both covariance matrices [15], based on the Frobenius matrix norm $\| \cdot \|_F$, is

$$ADCS = \sqrt{\frac{1}{(D-1)^2}\sum_{i=1}^{D-1}\sum_{j=1}^{D-1}\left(s_{ij}-s_{ij}^*\right)^2} = \frac{1}{D-1}\|\mathbf{S}-\mathbf{S}^*\|_F \quad . \tag{3.33}$$

Compositional error deviation (CED)

The criterion

$$\frac{\frac{1}{n_M}\sum\limits_{k\in M} d_a(\mathbf{x}_k, \mathbf{x}_k^*)}{\max\limits_{\{\mathbf{x}_i,\mathbf{x}_j\in\mathbf{X}\}}\{d_a(\mathbf{x}_i,\mathbf{x}_j)\}} \tag{3.34}$$

is a generalization of the measure applied in [15] and used in [18]. Here, n_M is the number of samples $\mathbf{x}_k$ containing at least one rounded zero, M is the index set referring to such samples, and $\mathbf{x}_k^*$ is the completed observation. The denominator is the maximum distance in the original data set. Here, d_a stands for the Aitchison distance (1.7).

3.5.5. The practical application of imputation model

Simulation study

For simulating compositional data, the so-called normal distribution on the simplex is used in combination with a latent model. A random composition $\mathbf{X}$ follows a multivariate normal distribution on the simplex if, and only if, the vector of ilr coordinates $\mathbf{Z} = ilr(\mathbf{X})$ (its elements are constructing by formula (1.9)) follows a multivariate normal distribution on $\mathbf{R}^{D-1}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ [3,63]. Thus, $\mathbf{X} \sim \mathcal{N}_{\mathcal{S}}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a D-part composition $\mathbf{X}$ that is multivariate normally distributed on the simplex.

The choice of the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in the simulation study determine the shape of the raw (compositional) data on the simplex. The further away $\boldsymbol{\mu}$ is from the null vector (indicates equilibrium on the simplex), the closer the compositions are to the border of the simplex. The choice of the covariance matrix

$\mathbf{\Sigma}$ determines how elongated the data points appear in the sample space. Finally, the choice of a latent model expresses further relationships between variables [58].

For the purpose of the study, a data set $\mathbf{Z}$ with n observation (compositions in ilr coordinates) and D original parts is simulated by the latent model

$$\mathbf{Z} = \mathbf{T}\mathbf{B}^T + \mathbf{E}, \tag{3.35}$$

where the columns of $\mathbf{E}$ are independently normally distributed with $\mathcal{N}(0, 0.01)$. The columns of the $n \times k$ matrix $\mathbf{T}$ are drawn from a standard normal distribution, and the elements of $\mathbf{B}$ are drawn from a uniform distribution in $[-1, 1]$. The obtained matrix $\mathbf{Z}$ is then expressed in the simplex as matrix $\mathbf{X}$ using the inverse isometric log-ratio mapping given in equation (1.10).

Rounded zeros are placed in every second column of $\mathbf{X}$ for all values below a certain quantile, $\mathbf{x}_j < Q_d(\mathbf{x}_j)$, where d is varied in the simulation between 0 and 0.3 (see for example Figure 3.9). In each step of the simulation, 100 data sets are produced and the average results are reported.

In the following we present three scenarios [58]:

(a) Low-dimensional scenario: $k = 3$ components (latent variables) are used to generate the data $\mathbf{X}$ with $n = 50$ observations and $D = 16$ variables. With this scenario, the methods are compared in the low-dimensional case. The fraction (quantile) of values below detection limit is varied between 0 and 0.3 and in every second variable the respective values are replaced by zeros.

(b) High-dimensional scenario: $k = 6$ components are used to generate the data with $n = 50$ observations and $D = 128$ variables. Again, rounded zeros are introduced in every second variable with varying fractions.

(c) Fixed amount (10%) of rounded zeros, changing dimension: $k = 6$ components are used to generate data with $n = 50$ observations and varying amounts of compositional parts (2, 4, 8, 16, 32, 64, 128, 256).

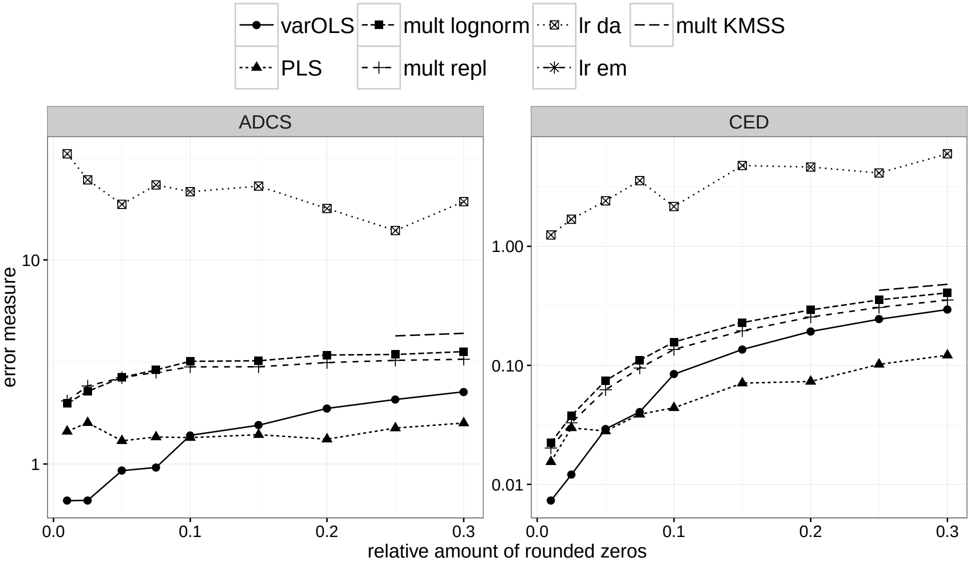


Figure 3.9: Low-dimensional scenario (a). Simulation results for a three-component model with 16 variables and with varying fractions of rounded zeros. Validation criterias are ADCS (left plot) and CED (right plot).

All methods discussed above are applied; the abbreviations in figures below represent the following approaches: *varOLS* refers to the regression method from Section 3.5.2, *PLS* to the PLS method from Section 3.5.1, *mult lognorm* is the multiplicative log-normal replacement method (Section 3.5.3), *mult repl* is the multiplicative replacement method, *lr da* and *lr em* the log-ratio data augmentation algorithm and the additive log-ratio EM algorithm, and finally the abbreviation *mult KSS* belongs to the multiplicative Kaplan-Meier smoothing spline replacement from Section 3.5.3.

Figure 3.9 shows the simulation results in the low-dimensional case. For both measures, the alternative variant with the variation matrix (the *varOLS* method) gives best results for small amounts of rounded zeros; for higher amounts, the PLS approach (Algorithm 3.5.1) is preferable. Multiplicative replacement and multiplicative log-normal replacement show similar behavior and are still reasonable (however, not that the vertical axis is log-transformed). *lr da* leads to much poorer results, and *lr em* and *mult KMSS* had numerical difficulties, mostly without any outcome.

In the higher-dimensional setting, the method *PLS* clearly outperforms the other methods, see Figure 3.10. Not all methods provide results. For example, the additive log-ratio EM algorithm cannot deal with high-dimensional data since a least-squares regression is used with $D-1$ predictors but n is smaller than D. The replacement methods and *varOLS* are relatively comparable in performance.

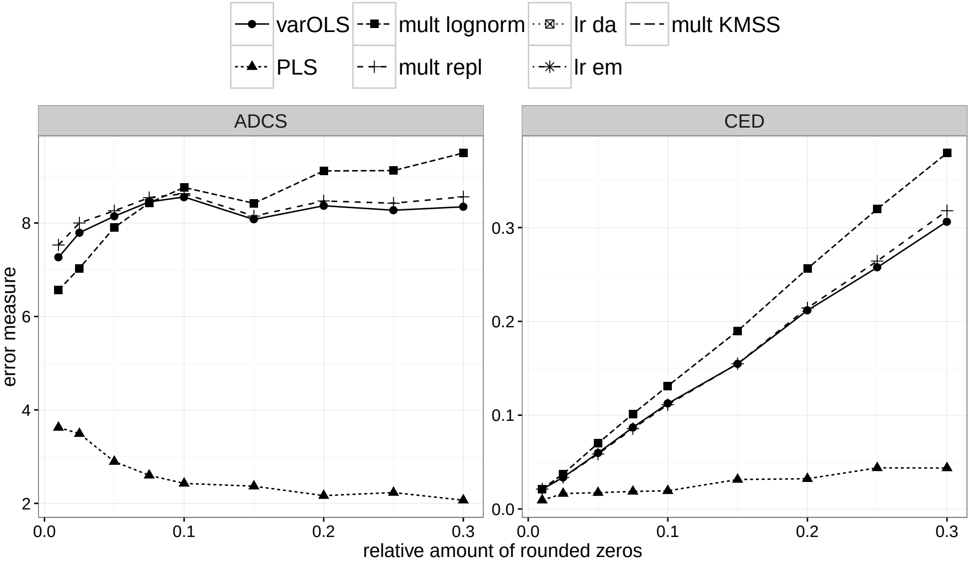


Figure 3.10: High-dimensional scenario (b). Simulation results for a six-component model with varying fractions of rounded zeros. Validation criterias are ADCS (left plot) and CED (right plot).

Figure 3.11 presents results from the last scenario with varying numbers of variables and a fixed fraction of rounded zeros (0.1). As it was already visible in the previous results, *varOLS* performs well in low-dimensional situations, whereas *PLS* clearly outperforms other methods when the number of variables is higher.

One of the reasons for the good performance of the PLS method in the simulations is the way how the data were simulated: Since we already used a latent variable model for simulating the data, it can be assumed that a method like PLS which makes use of this latent structure will be more successful than me-

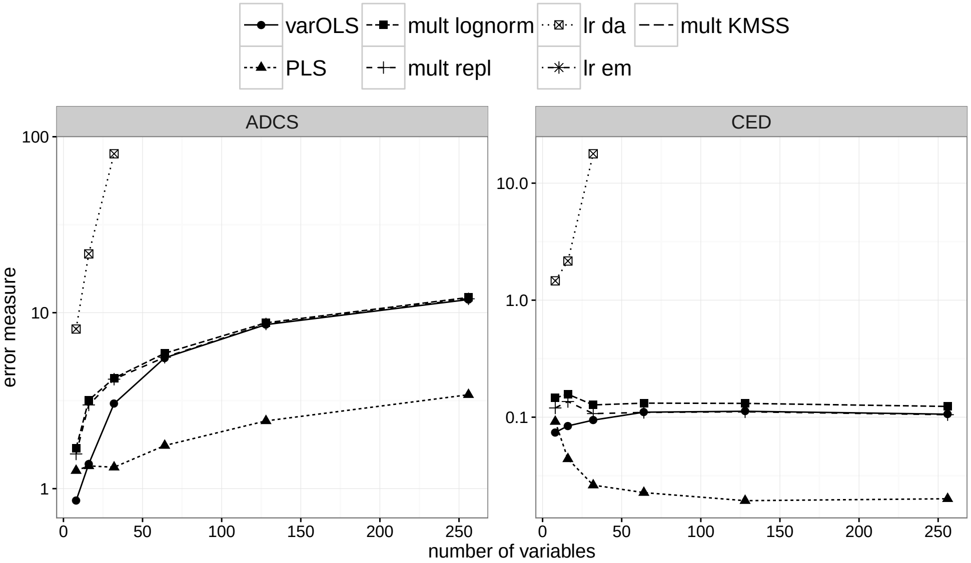


Figure 3.11: Varying dimension of data scenario (c). Simulation results for a three-component model and a fixed fraction of rounded zeros (0.1). Rounded zeros are placed again in every second variable. Validation criterias are ADCS (left plot) and CED (right plot).

thods that do not make use of this fact. In real high-dimensional data situations, however, one can often assume such an underlying model, because PLS regression turned out to be very successful for prediction in high-dimensional applications in chemometrics, metabolomics, etc. In the following section, a data set from metabolomics is used as a basis for simulations, the different methods are again compared.

Although the methods discussed in Section 3.5.3 are quick to compute, they have either limitations concerning the dimensionality, concerning the numerical stability, or concerning the quality of the results. In fact, for the simulation in the metabolomic data set, it turned out that for a relative amount of rounded zeros of more than 0.1, only the methods *varOLS* and *PLS* gave results. The drawback of these methods is higher computational effort: on a standard PC, *varOLS* takes around 10 minutes for one run, and *PLS* about 30 minutes. However, this is because rounded zeros were included in every second variable [58].

Data set from metabolomics

The presented imputation procedure is applied to a data set from metabolomics. Here LC-MS spectra from dried blood spots of samples from patients suffered from MCADD and healthy controls (data used in the example in Section 3.3.4) are considered [7]. Here only 278 metabolites were used, because of the computational severity of the imputation algorithm. These chosen metabolites also does not contain rounded zeros in original data table.

Figure 3.12 shows results from the data where the detection limit is artificially increased in every second variable (with a fraction of rounded zeros from 0.05 to 0.25 in steps of 0.05). Not all methods can be applied on data with higher number of variables than observations, e.g. the data augmentation method (*lr da*) as well as the expectation-maximization method (*lr em*)). These methods are thus, excluded from Figure 3.12. Only few methods give results for higher fractions (≥ 0.1) of rounded zeros, e.g. *mult KMSS* and *mult lognormal* provide only results for a fraction of rounded zeros equal to 0.05.

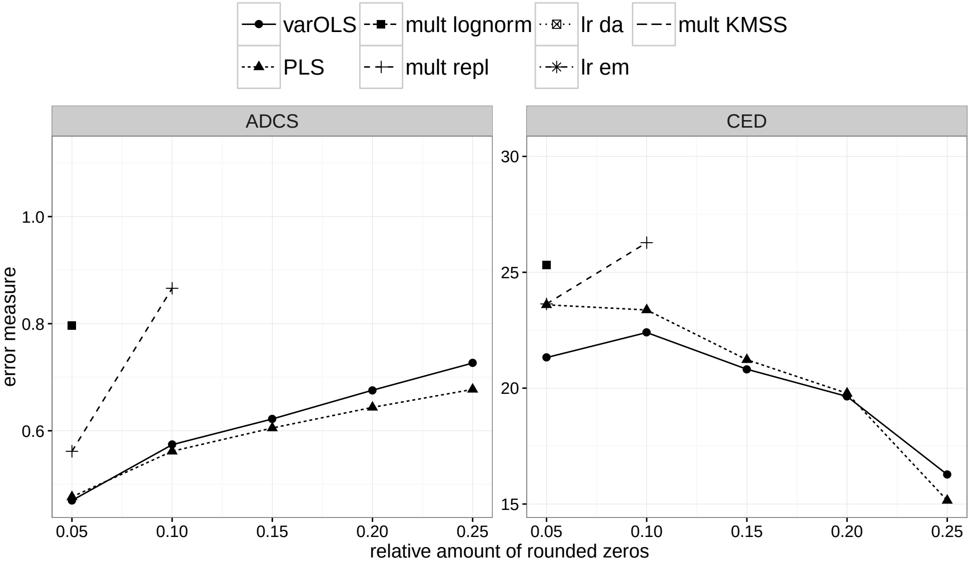


Figure 3.12: Results from the replacement of rounded zeros for the MCADD data. Validation criterias are ADCS (left plot) and CED (right plot).

For the precision measure *CED*, the *varOLS* method shows slightly better performance than the *PLS* method; for the covariance comparisons using *ADCS*, the *PLS* method gives slightly better results. *PLS* uses the full multivariate information for imputation, while *varOLS* uses at most 25 predictors, i.e. not even 10% of the available predictors. A reason for this could be that the predictors also included imputed rounded zeros; using them all covers the full data structure but leads to a possible cumulation of errors; using only a subset will lead to a loss of information but avoids errors.

Our procedures are available in the R package *robCompositions* [44] at github.

3.6. PARAFAC

3.6.1. Building up the model

As mentioned in the introductory section, metabolite (compositional) data may form a three-way structure. The typical example are repeated measurements of samples in time. Similarly as for PCA, also in this context clr coordinates are preferable. In general, multi-way data are characterized by several sets of variables that are measured in a crossed fashion [114]; as an example, the same set of variables is measured in different times. In practice, three-way data are of primary interest, especially also in the form of three-way compositions. Let's have $I \times J \times K$ data array (cube): we have I samples and J variables (compositional parts), every sample is measured K times [41]. Consequently, each of K tables of dimension $I \times J$ (slices of the cube) can be considered as a compositional data matrix, ready to be processed using the logratio methodology. In the following text of this section, the whole data cube is denoted as $\underline{\mathbf{X}}$, for slices the notation $\mathbf{X}_k, k = 1, \ldots, K$ is used. A graphical representation of three-way data is displayed in Figure 3.13.

By following the previous considerations, the first mode is represented by samples, placed in rows of the data cube. The second mode is formed by variables (columns) and the third mode, frequently represented by time, form slices or tubes [114–116]. For the possibility to deal with three-way data in a statistical

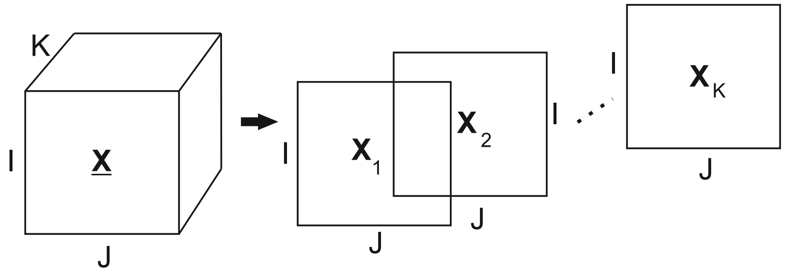


Figure 3.13: Graphical representation of the data array (cube) $\underline{\mathbf{X}}$ and its slices $\mathbf{X}_k$.

software and also to ease the notation, the data cube is matricized into the form of two-way matrix [115, 117]. Matricizing is done by concatenating matrices for different levels of the third mode next to each other. The column-dimension of the resulting matrix thus becomes quite large in the mode consisting of two prior modes, i.e., the final matrix has dimension $I \times JK$. This structure is visible in Figure 3.14.

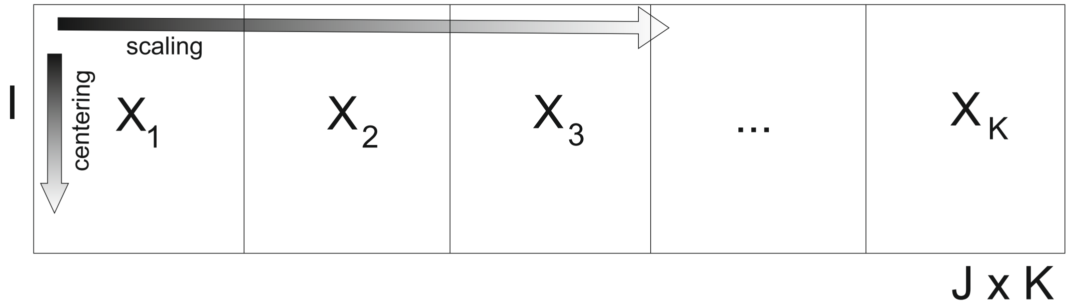


Figure 3.14: The principle of unfolding the data cube $\underline{\mathbf{X}}$ into data matrix $\mathbf{X}$. The principle of centering and scaling of the unfolded data matrix $\mathbf{X}$.

Also preprocessing of three-way data, that is of particular importance in the chemometric context [114,118], must take account specific structure of the observations. The centering is done by the procedure called the single-centering when the unfolded data matrix of dimension $I \times JK$ is centered across the first mode, i.e. single columns are centered. It is possible to center also in more modes simultaneously (e.g., centering the first mode and then the second mode), but it is rather avoided as such centering scheme can destroy the multilinear behavior of the data. Note that result of centering of compositional data in log-coordinates across rows in single slices is nothing else than clr coordinates of the respective observations; it is easy to see, if we rewrite clr coordinates of a composition $\mathbf{x} = (x_1, \ldots, x_D)'$ as $y_i = \ln(x_i) - \frac{1}{D}\sum_{i=1}^{D} \ln(x_i)$, $i = 1, \ldots, D$. The scaling is usually done through rows of the unfolded data matrix, so we refer to scaling within the first mode. If some variable of the second mode is scaled, it is necessary to scale all columns where this variable occurs. Scaling in more modes is also possible, but not recommended for practical applications. The principle of centering and scaling three-way data is shown in Figure 3.14. For particular metabolomic applications, scaling within the first mode is replaced by specific approaches in each slice, like the AUC normalization or normalization to creatinine, mentioned in Section 2.2.

PARAllel FACtor analysis (PARAFAC) as a special version of the three-way PCA is one of popular decomposition methods for three-way data in chemometrics [59,119–121]. The PARAFAC model was invented by R. Harshman [38] and by J. Carroll [37], who named the model CANDECOMP (CANonical DECOMPosition), being an alternative to previously introduced Tucker3 model [114,115,122]. The difference between bilinear PCA and PARAFAC is that PARAFAC result of one score matrix and two loading matrices; moreover, in the case of PARAFAC, the requirement of orthogonality of loadings is not needed to identify the model [115].

The PARAFAC model is structural model with score matrix $\mathbf{A}_{I\times F}$ and two loadings matrices $\mathbf{B}_{J\times F}$, and $\mathbf{C}_{K\times F}$ with elements a_{if}, b_{jf}, and c_{kf}, for $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, and $f = 1, \ldots, F$, where F denotes the number of factors that are extracted. The PARAFAC model in terms of single elements of the data cube $\underline{\mathbf{X}} = (x_{ijk})$ (i.e. for ith observation of jth variable in kth time), can be written as [115,122–125]:

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \; i = 1, \dots, I, \; j = 1, \dots, J, \; k = 1, \dots, K; \tag{3.36}$$

here e_{ijk} stand for residuals. The structure of the model is also visible in Figure 3.15.

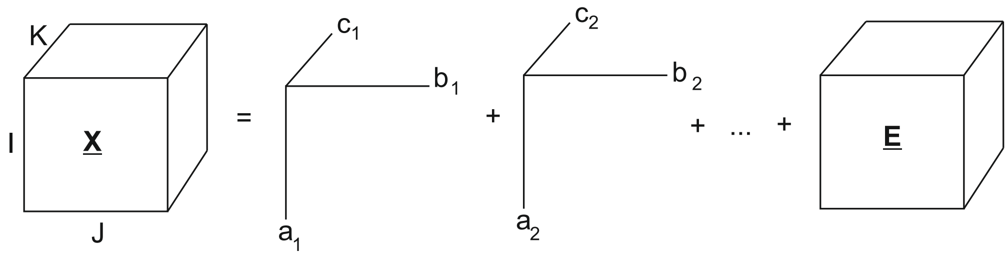


Figure 3.15: Graphical representation of the formula (3.36).

By considering F factors, the PARAFAC model consists of $F(I + J + K)$ parameters. The advantage of the PARAFAC model is the uniqueness of the solution; consequently, there is no problem with rotational freedom like for PCA. Then, if the data is indeed trilinear (if two modes are fixed, then the third mode is linear), the underlying spectra (or whatever constitute the variables) will be found according to the number of components employed and appropriate signal-to-noise ratio [114, 126]. Unique solutions can be expected if the loading vectors are linear independent in two of the modes, and furthermore, in the third mode, the less restrictive condition that no two loading vectors are linearly dependent must be fulfilled. The mathematical meaning of uniqueness is that the estimated PARAFAC model cannot be rotated without a loss of fit, as opposed to two-way analysis (PCA), where one may rotate scores and loadings without changing the fit of the model. A unique solution, therefore, means that no restrictions are necessary to identify the estimate the model apart from trivial variations of scale and column order [114].

The solution of the model (3.36) is obtained using the *alternating least squares (ALS)* algorithm. The principle of ALS is through breaking up iteratively the model into three sets of parameters, such that it is linear in each set given fixed values for the other two sets [123]. Furthermore, we assume that the loadings in two modes are known and then the unknown set of parameters of the last mode are estimated [114, 122]. Explicitly, we define $\mathbf{M} = \left[\text{vec}(\mathbf{b}_1\mathbf{c}_1^{'}), \ldots, \text{vec}(\mathbf{b}_F\mathbf{c}_F^{'})\right]$ and proceed to minimization problem [114, 115]

$$\min_{\mathbf{AM}} \|\mathbf{X} - \mathbf{AM}'\|_F^2, \tag{3.37}$$

where $\|\mathbf{X}\|_F^2 = tr(\mathbf{X}'\mathbf{X})$ denotes the Frobenius norm of $\mathbf{X}$ [114,115,125]. The model for estimation of scores $\mathbf{A}$ is

$$\mathbf{X} = \mathbf{AM} + \mathbf{E}_A, \tag{3.38}$$

where $\mathbf{X}$ represents unfolded matrix $\underline{\mathbf{X}}$ and $\mathbf{E}_A$ errors of the model, both being of dimension $I \times JK$. The conditional least squares estimate of $\mathbf{A}$ is then

$$\mathbf{A} = \mathbf{XM}(\mathbf{M}'\mathbf{M})^+ \tag{3.39}$$

with the Moore-Penrose inverse $(\mathbf{M}'\mathbf{M})^+$ of $\mathbf{M}'\mathbf{M}$. The loading matrices $\mathbf{B}$ and $\mathbf{C}$ are estimated analogously [114,115,122]. The algorithm is repeated until convergence (i.e., when the changes of scores and loadings from two consecutive steps are small enough) that can be achieved much faster by setting proper initialization values [115]. Note that stability of the ALS algorithm (faster convergence, avoiding local minima) can be strengthened also by setting further constraints to the model (mostly connected to orthonormality of loadings), but they may sometimes lead to problems with interpretation of loadings of the model [114, 115] for any kind of data preprocessing.

In the context of PARAFAC modeling of three-way metabolomic data, the normalization is done for each slice that replaces scaling within the first mode (i.e. just centering across the first mode is performed).

3.6.2. Practical application of PARAFAC

Multi-way data occur also in metabolomics in the form of repeated measurements at several time points. A specific case is represented by urine samples of newborn babies, suffered from asphyxia, which are analyzed in this section. Asphyxia is caused by lack of oxygen during the birth that may cause health problems (brain damage, etc.) and some changes in metabolites of the baby [127]. Urine samples were collected from 9 patients at five time points (7, 28, 52, 76 and 100 hours after the birth). Samples were processed using targeted metabolomic analysis by HPLC-MS/MS (High-Performance Liquid Chromatography - Mass Spectrometry), resulting in 179 analyzed metabolites. As a consequence, a structure of the three-way data cube has dimension 9 × 179 × 5 (samples × metabolites × time points).

Data were processed in the R software [43]; in addition to standard packages also those on PARAFAC modeling [52,54] and analysis of compositional data [44] were employed. At first, quality control-based robust LOESS signal correction method was applied [12], and consequently also zero replacement according to [17] in order to enable further processing based on logarithmic calculations.

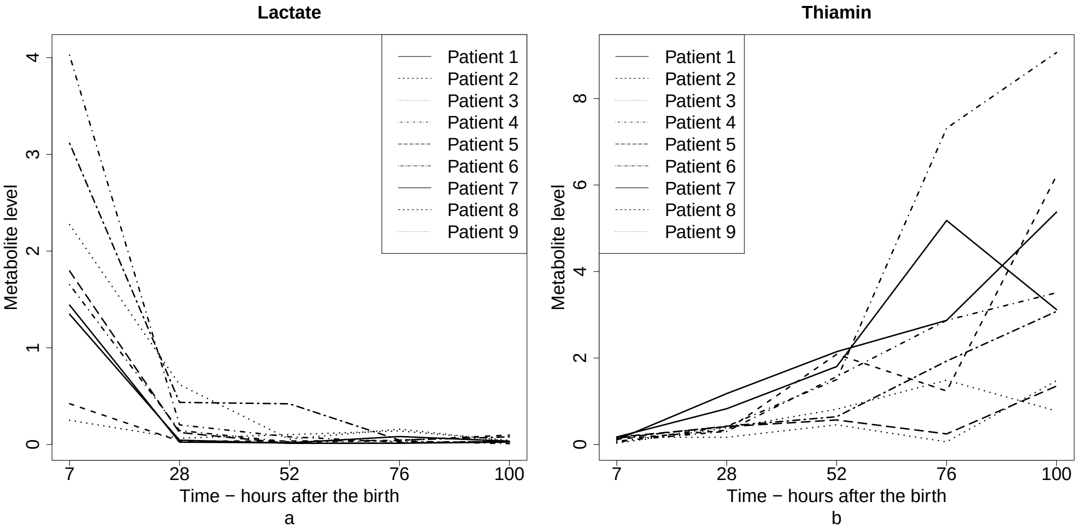


Figure 3.16: Trends of lactate (a) and thiamine (b) levels in time points for each patient.

It is expected that levels of some metabolites change substantially shortly after the birth. It is demonstrated for the cases of lactate and thiamine, see Figure 3.16, for AUC normalized samples. Accordingly, the level of lactate is very high just after the birth (concretely after seven hours), then it sharply decreases, as is clearly visible in the case of 28 hours after the birth and beyond. This trend is clearly caused by lactic acidoses due to anaerobic conditions (well known by chemical consequences of asphyxia). An opposite pattern can be observed for thiamine, which level is very small at the beginning, but then it mostly increases and substantially differentiates among patients. The trend is caused by treatment of asphictic newborns with a parental infusion of vitamin cocktail (that includes thiamine); this treatment was the same for all babies. Note that results after taking the normalization to creatinine would be very similar, thus, they are not displayed here.

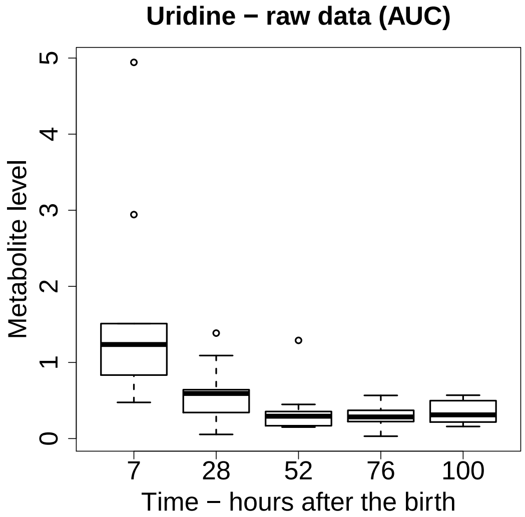
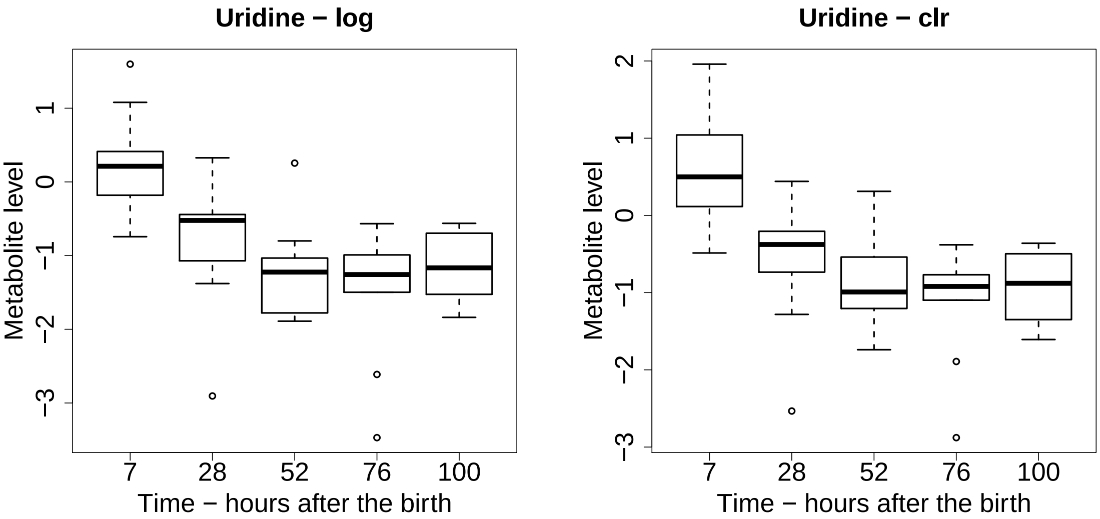


Figure 3.17: Boxplots of uridine - AUC normalized raw data.

The trend of metabolite levels in time can be also visualized using boxplots. Such boxplot series for the case of uridine is displayed in Figure 3.17. We can observe similar patterns like for lactate with highest values in the first 28 hours of babies' life for the majority of the other metabolites. The differen-



a

b

Figure 3.18: Boxplots of uridine after log-transformation (a) or clr coordinates (b).

ces in further time points (with rather small absolute values) will be highlighted by taking any logarithm-based data transformation that captures the relative (ratio based) changes of metabolite levels. For this purpose, both the standard log-transformation and the clr coordinates (1.8) are applied, the latter honoring also the principle of scale invariance of the original observations (i.e. that not factual metabolite levels, but rather their relative contributions are of primary interest). The results, displayed in Figure 3.18, are very similar, small differences in both location and scale are caused by aggregation of metabolites by geometric mean (clr coordinates) instead of arithmetic one (resulting from the AUC normalization) in the denominator of logratios.

The absolute scale of observations, obtained by using either log-transformation or taking the clr coordinates, is also a necessary assumption to proceed with PARAFAC modeling. The reason is that most of the standard statistical methods rely on the Euclidean geometry in real space [66] that is not coherent with the relative character of the original urine metabolomic data. For this purpose, three preprocessing options are compared:

A) clr coordinates are applied, irrespective to previous scaling (AUC, normalization by creatinine);

B) data normalized by AUC, then log-transformation applied;

C) data normalized by creatinine, then log-transformation applied.

The normalization of the level of creatinine and the AUC in combination with the clr coordinates yield the same results, due to scale invariance [59]. For this reason, they are included in one option denoted as A. Option B differs just very slightly (being also a general experience); the reason is the AUC normalization, where the original metabolite levels are divided by their average (arithmetic mean) and then taking the log-transformation. If the arithmetic mean would be replaced by the similar geometric mean as discussed above, then the resulting observations were exactly the same (clr coordinates). On the other hand, option C differs substantially due to log-transformation of ratios of all metabolites with one common denominator; this corresponds to additive logratio (alr) coordinates (1.13). Note that, in addition to geometric incovencies, results of PARAFAC modeling in alr coordinates might depend on the chosen common denominator, and, therefore, this must be chosen carefully.

The PARAFAC model was built for two components and the resulting loadings were plotted in order to ease recognition of patterns. No orthogonality constraints were applied, because the results were already nicely interpretable using default setting; possible nonnegativity constraints became irrelevant after taking clr coordinates/log-transformations. The variation explained by the PARAFAC model (27.4%) was almost the same for the options A and B. The explained variation for the option C was slightly higher - 32.2%.

The first two scores (of the first mode) and loadings (of the second and the third mode) are displayed in this section. The first mode represents patients (Figure 3.19), the second mode shows metabolites (Figure 3.20) and the third mode displays changes in time points (Figure 3.21). As the results for option B are almost the same as for A, just the latter are shown here and option B is

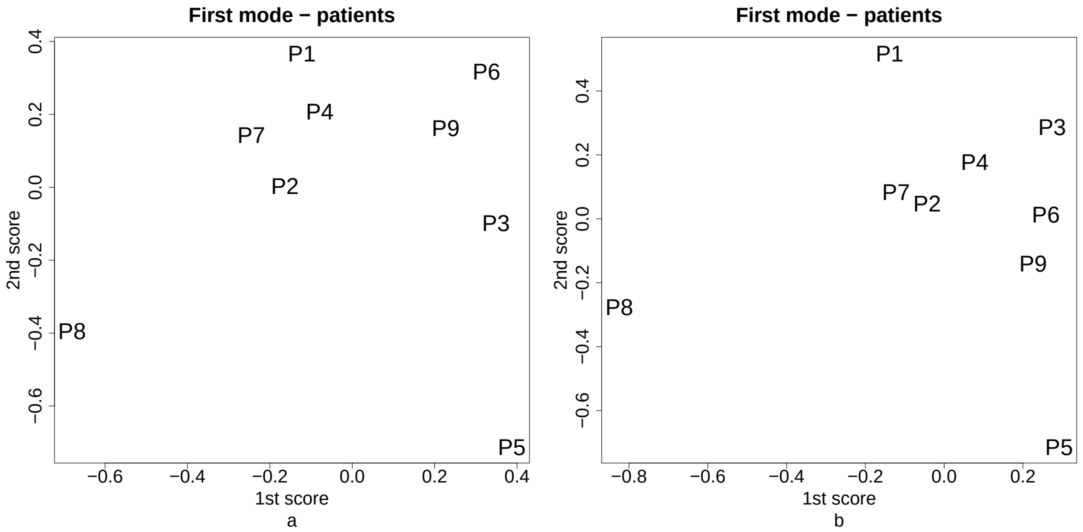


Figure 3.19: Scores of the first mode of PARAFAC for preprocessing options A and B (the results are the same) (a) and C (b).

assigned to the A; on the other hand, PARAFAC outputs for option C are substantially different. The smallest difference is in the first mode. Patients numbers P5 and P8 are placed as outliers in both variants. The reason for the remoteness of the patient P8 is caused by the fact that P8 had the smallest gestational age and it had very serious brain damage. On the other hand, the patient P5 suffered from asthma. Patients P3, P6 and P9 are characterized by similar patterns. Patients P1, P2, P4 and P7 form another cluster in the first mode, just P1 tend to exhibit outlyingness for the latter option. The reason might be that structure of metabolites in a time of this patient differs the most by taking the creatinine normalization (which is also visible directly from the raw data).

Metabolites in the second mode are represented only by abbreviations, most of them hidden in the main cluster. The biggest difference is in ST049, ST368 and ST220; while the first one is a clear outlier for the option C, the remaining form outliers for A and B. ST049 is N-acetylserotonin, which is most probably irregular data, because its values vary from chemical noise (10^4 arbitrary units) to very high numbers (10^7); consequently, they are not associated with specific time points but with particular patients. Moreover, Human Metabolome Data-

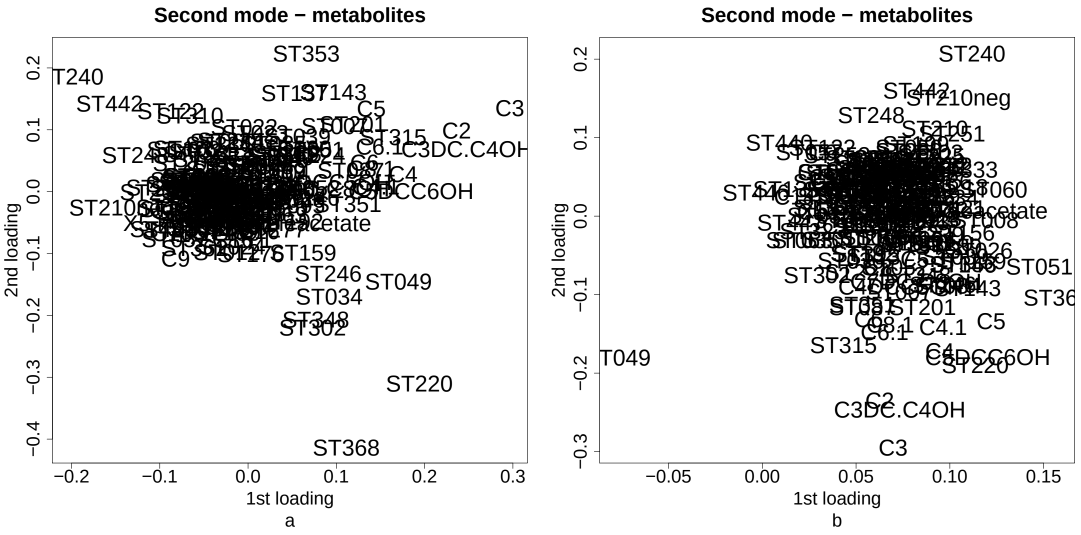


Figure 3.20: Loadings of the second mode of PARAFAC for preprocessing options A and B (the results are the same) (a) and C (b).

base considers ST049 as undetectable in human urine. Note that options A and B are not influenced by this metabolite. ST220 (homovanillate) and ST368 (3-methoxytyramine) have very small chromatographic peaks and they are prone to the variability of instrument measurements. Their solitude is true because their structure is very different from the other metabolites. Accordingly, ST220 and ST368 have very small values at the first time point and these values are growing in time. Also metabolites discussed above show an interesting behavior with respect to preprocessing done. Lactate (ST137) moved from the border to the middle of the main cluster for the option C, while thiamine (ST230) and uridine (ST218) are contained in the cluster of the majority of metabolites in both cases. Metabolites with high values shortly after the birth (ST137, ST143, ST201, ST007) are clustered near the main cluster in Figure 3.20a and tend to be more widespread in Figure 3.20b.

The structure of the time mode is the most interesting. The trend of loadings for options A and B corresponds to trends from Figures 3.16-3.18 and general preliminary expectations with the somehow exceptional position of the first time point. This differs dramatically from C, where no such patterns can be observed,

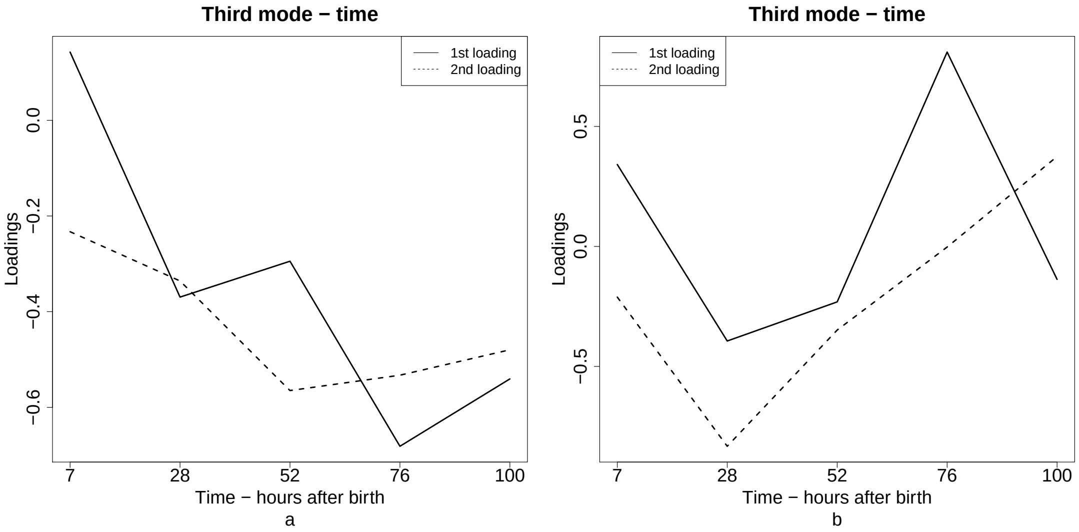


Figure 3.21: Loadings of the third mode of PARAFAC for preprocessing options A and B (the results are the same) (a) and C (b).

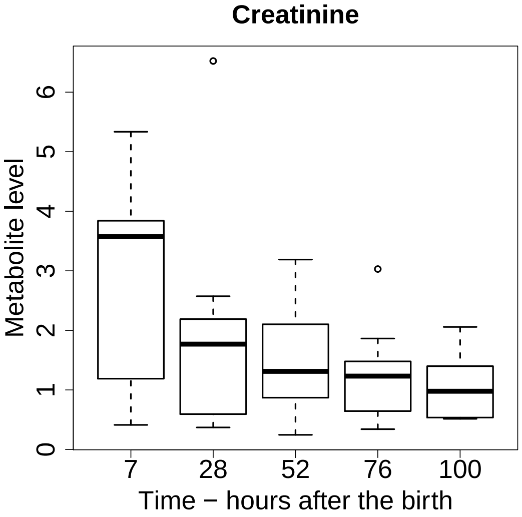


Figure 3.22: Boxplots of creatinine levels.

the loadings thus, seem not reflect the true reality of the data. The reason might be a specific trend of creatinine in time that affects the other metabolites through pairwise logratios in the alr coordinates. The level of creatinine is very high in the time of 7 hours after the birth comparing to the other metabolites, which is visible from boxplots in Figure 3.22 and corresponds also to the position of creatinine by newborns in general [128]. This fact probably causes the strange behavior of time loadings in Figure 3.21b, where the specific role of pairwise logratios with creatinine is pronounced.

Chapter 4

Practical application

This chapter is a summarization of all methods presented in the previous chapters in term of real metabolomic data. The whole methodology is demonstrated on data sets analyzed by targeted and untargeted analyzes. All results are discussed and compared.

4.1. Data introduction

Data sets chosen for this chapter were not measured on human. They represent serum samples of grazing horses evaluated by targeted and untargeted analyzes. The samples are divided into two groups - healthy controls and patients influenced by atypical myopathy (acquired equine multiple acyl-CoA dehydrogenase deficiency). Atypical myopathy is a fatal muscle disease and it is probably caused by ingestion of maple seeds containing toxic hypoglycin A [61, 129]. 12 controls and 10 patients are involved in the analysis, but repeated investigation of some patients has led to more samples. Accordingly, the final number of patient samples is 19.

The additional data set for a deeper comparison of methods consist of urine samples, evaluated by targeted analysis. These samples were not able for all horses, but results of the statistical evaluation are still very informative. Only 5 controls and 6 patients (every measured only once) were used for this analysis.

4.2. Targeted analysis of serum samples

The first data set is formed by serum samples, evaluated by targeted analysis. As already mentioned, 12 controls and 19 patients are involved in the analysis. The number of metabolites is 176, so high-dimensional data are obtained; therefore, statistical processing must be performed with special methods like partial least squares regression.

The QC samples are very stable (generally in all targeted analyzes), so the influence of signal correction is very small. This trend is visible for example in Figure 4.1 which represents the time flow of acylcarnitine C10:2. This data set grows also a different tendency - peak areas are much smaller in control samples (denoted as black points in the figure) in comparison with patients (blue squares) in a lot of metabolites. The QC samples are often placed in the middle of the graph (see Figure 4.1) and this fact may indicate a potential marker of the disease.

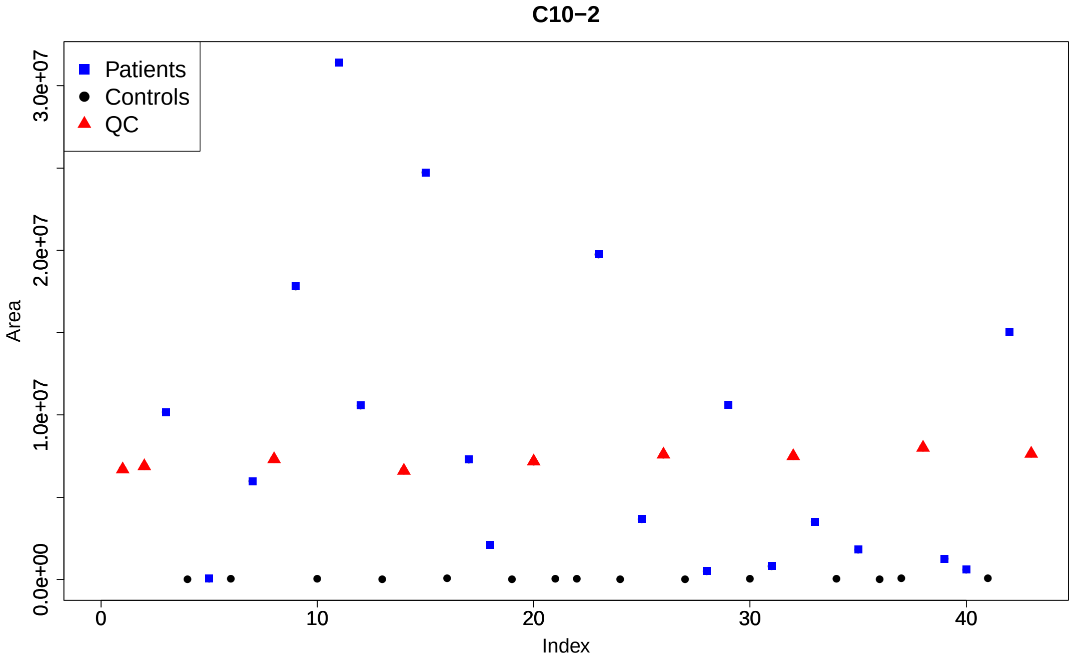


Figure 4.1: Time flow of acylcarnitine C10:2 - raw data.

Interpolated values (after LOESS procedure) of QC samples are very small (from 1 to 2). Almost all coefficients of variation (CV) are lower than 30 %. Only one metabolite (succinyladenosine) had $CV = 36.3$ % and thus, it was removed from the analysis. These facts also support prescription on the stability of the data set.

In this step of the analysis, the final size of the data table was obtained - 31 observations × 175 metabolites. Now we are not interested in time flow of the data and thus, we can reorder the data table in a more digested way: the samples are reordered by groups of controls and patients.

The next step of the analysis is the imputation of zeros. In the targeted analysis, zeros are not often present; it is caused by the fact that all measured metabolites are known and the measurement device is calibrated on their values. Consequently, levels of all metabolites are detected. Also for this data table, all values are higher than zero.

The last step of preprocessing is the application of clr coordinates (1.8) and centering the resulting data column-wise.

Now we can evaluate graphical and numerical results of the analysis. Both types of methods - supervised and unsupervised - were used. Green points (●) are used for labelling controls, blue triangles (▲) denote patients.

The first graph used for the analysis of the metabolomic (compositional) data set is the score plot of the principal component analysis (PCA) in Figure 4.2. Data ellipses are also visualized in this plot. These ellipses are made for first two scores separately for patient and control samples. They are based on Cholesky decomposition of the covariance matrix of these scores. These ellipses correspond to 75% quantile of F distribution with 2 and $n - 2$ degrees of freedom. When variables have a multivariate normal distribution, data ellipses represent estimated probability contours [130]. Ellipses are made only for a better interpretation and visualization of separation of groups (controls/patients) in samples. The same approach is used also in the following examples. In Figure 4.2 two clusters can

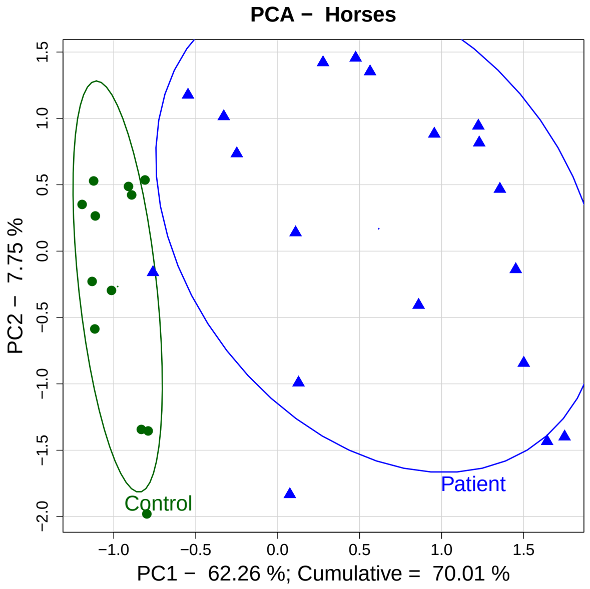


Figure 4.2: Score plot of principal component analysis of serum samples analyzed by the targeted approach.

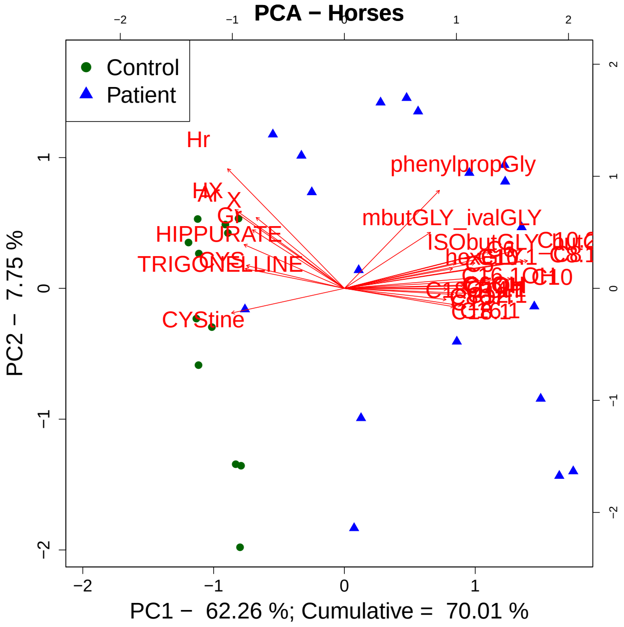
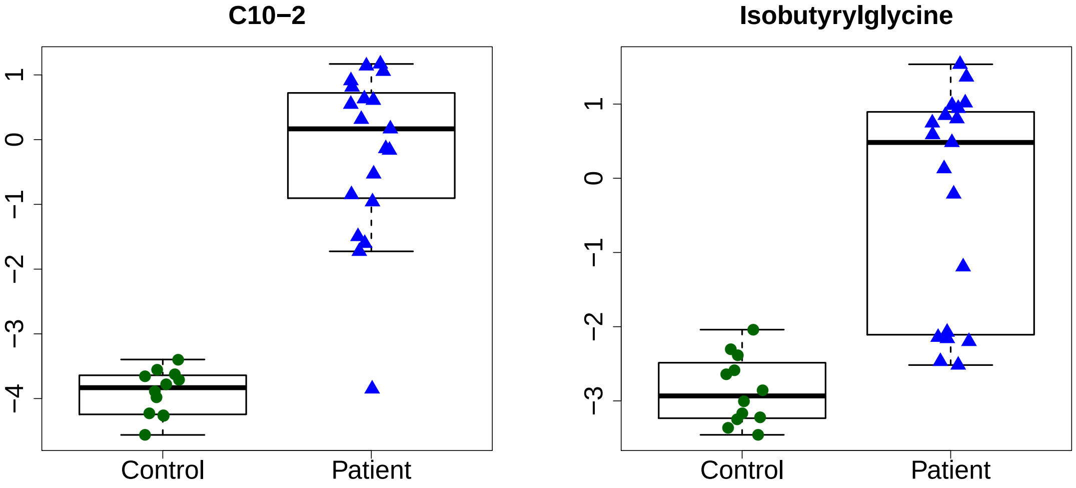


Figure 4.3: PCA biplot of serum samples analyzed by the targeted approach.



a

b

Figure 4.4: Boxplots of acylcarnitine C10:2 (a) and isobutylglycine (b) in terms of clr coordinates.

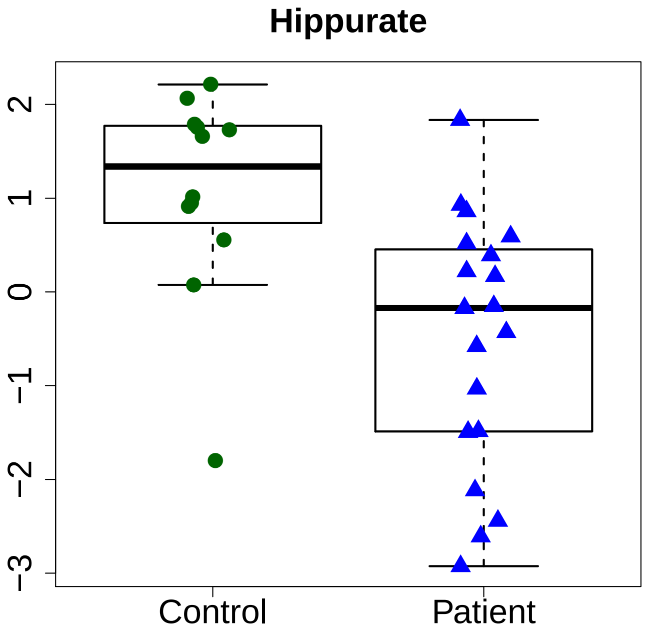


Figure 4.5: Boxplots of hippurate in terms of clr coordinates.

be clearly observed - patients are separated from controls. The variability explained by the first two principal components is 70.1% which is quite high and results seem reasonable. Only one patient is in the cluster of controls - it is sample denoted as "Patient 19", which is a horse performed at the early stage of the disease. The group of patients is more differentiated, this might be due to different severity of the disease. The other differences are visible from PCA biplot (Figure 4.3). Here also rays representing metabolites are contained - for better visualization only the first 30 most important loading vectors are used. A heuristic rule for the selection of the most important metabolites is used. Rays, representing metabolites, are sorted by their lengths and first 30 with longest rays are chosen. A large difference between controls and patients seems to be due to a big group of acylcarnitines (C6, C10, C10:2, C12:1 etc.) - all patients have higher relative values in these acylcarnitines than controls. Also other metabolites can be assigned to this group of potential markers, for example, isobutylglycine (denoted as ISObutGLY_butGLY in biplot) or phenylpropionylglycine (denoted as phenylpropGly). The second group of metabolites (hippurate, cystine, ...) is not so expressive. This trend is also visible from boxplots of the original data in clr coordinates (alternatively $z^{(l)}$ from (1.11)); in Figure 4.4 boxplots of acylcarnitine C10:2 and isobutylglycine are displayed. Both metabolites are characterized by higher values in patient samples. On the other hand, in Figure 4.5 boxplot of hippurate is showed; here the level of this metabolite is increased in patients, but the difference between two groups of samples is not so significant.

The difference between groups of patients and controls is more visible from the score plot of partial least squares - discriminant analysis (PLS-DA) which is displayed in Figure 4.6. This method works a priory with the information about classification of both groups (patients, controls). The outlyingness of Patient 19 is preserved also here. Differences between metabolites are also visible from biplot of PLS-DA in Figure 4.7, the above described structure is preserved. VIP scores are also evaluated, but as was written in Section 3.3.1, their values are slightly sensitive. With the use of "greater than one"rule 60 metabolites would

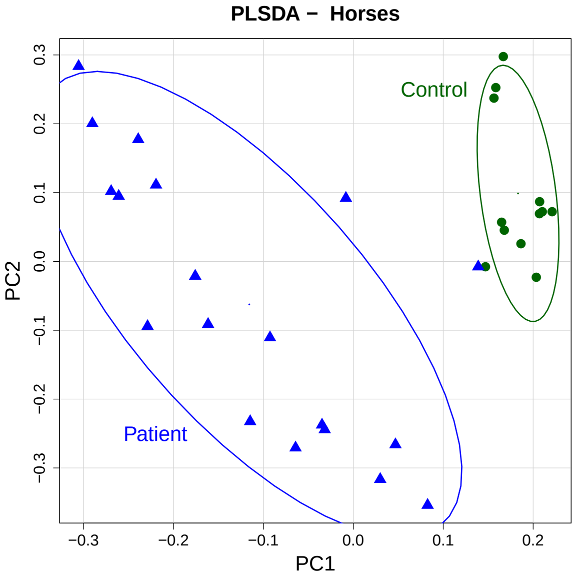


Figure 4.6: Score plot of partial least squares - discriminant analysis of serum samples analyzed by the targeted approach.

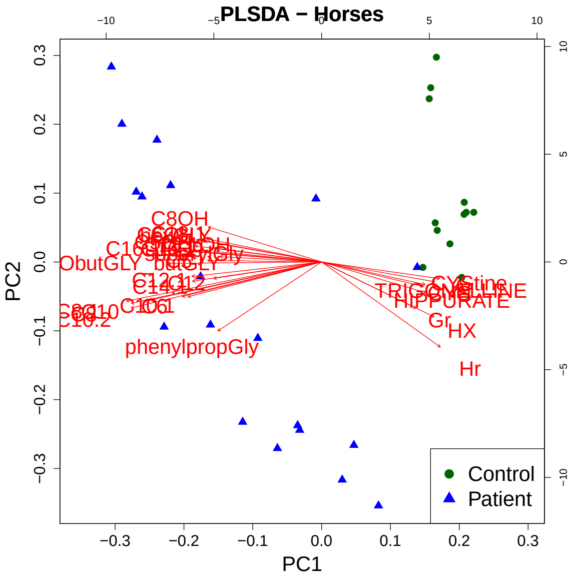


Figure 4.7: Biplot of partial least squares - discriminant analysis of serum samples analyzed by the targeted approach.

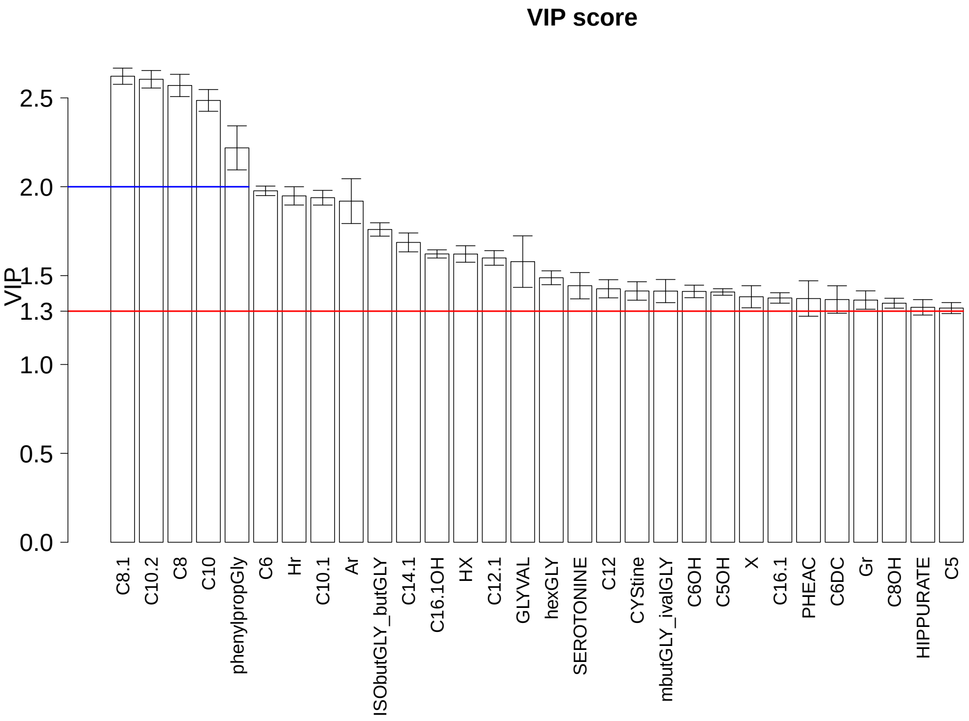


Figure 4.8: VIP scores of PLS-DA of serum samples analyzed by the targeted approach.

be chosen (32 are acylcarnitines), therefore, a heuristic rule for the selection of important metabolites is used. A list of first 30 metabolites with the highest VIP scores from PLS-DA is presented in Figure 4.8. All metabolites in this list have VIP scores greater than 1.3, which stands for the important difference between groups. The level of 1.3 is denoted by a red line. Mean ± one standard deviation is marked in these barplots. The first five metabolites have scores even greater than 2, which is marked by a blue line in the graph. On the second position the above mentioned acylcarnitine C10:2 (written as C10.2) is placed, isobutylglycine appears in the 10th position. Hippurate is also in this list, but its position is less important (29th). High values of the first 30 metabolites can be caused by the fact that matrix of clr coefficients is used instead of the standard real data matrix.

Data were also evaluated by orthogonal partial least squares - discriminant

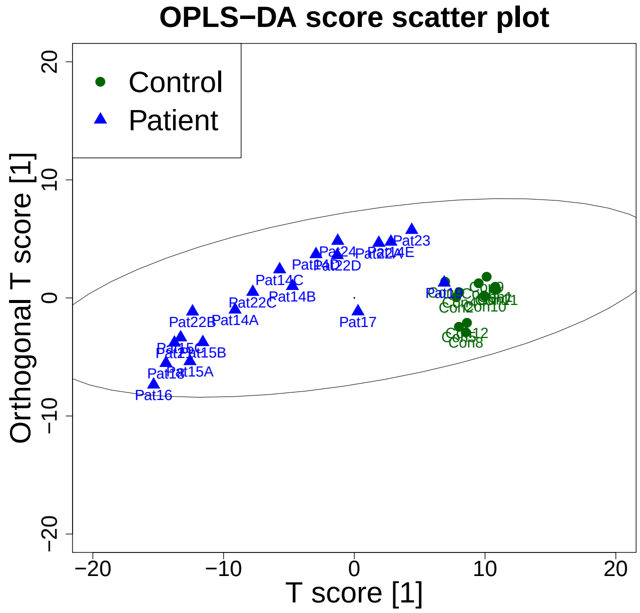


Figure 4.9: Score plot of orthogonal partial leas squares - discriminant analysis of serum samples analyzed by the targeted approach.

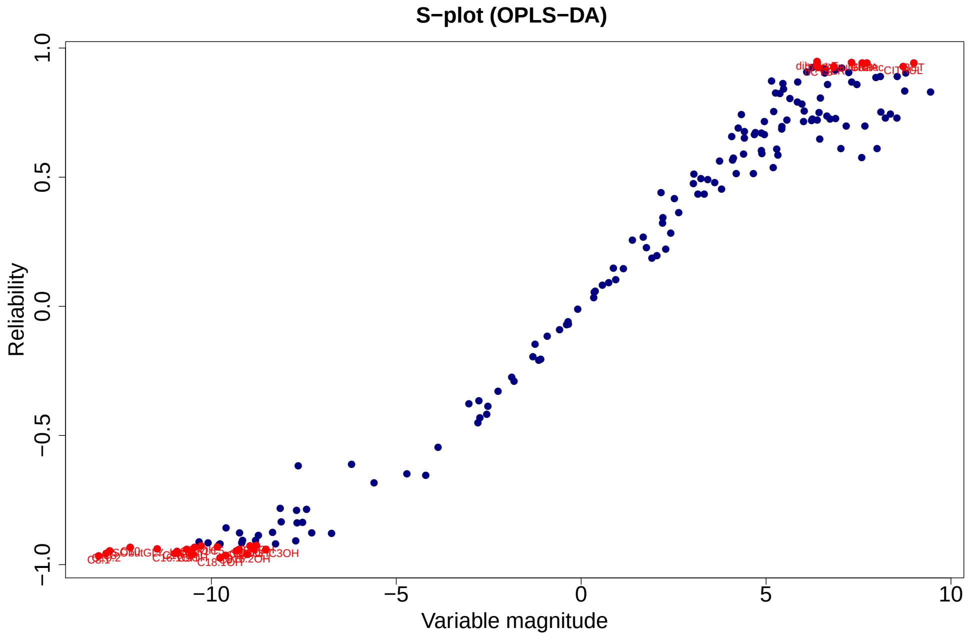


Figure 4.10: S-plot of serum samples analyzed by the targeted approach.

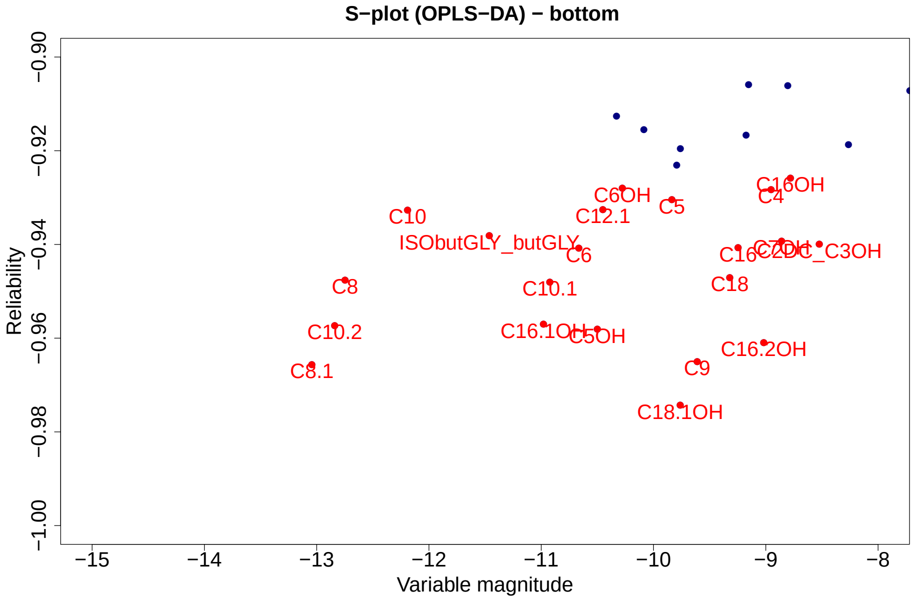


Figure 4.11: S-plot of serum samples analyzed by the targeted approach - left bottom part of Figure 4.10.

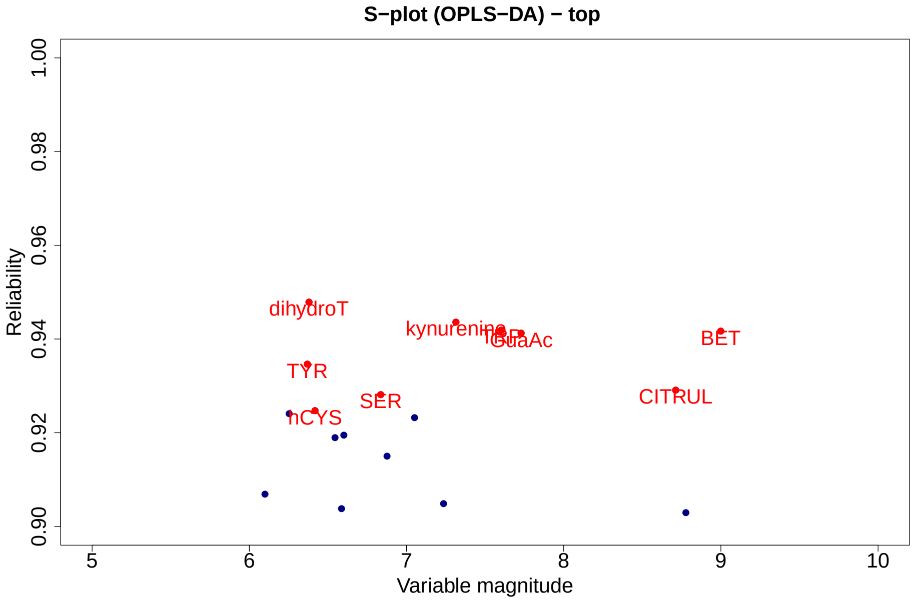


Figure 4.12: S-plot of serum samples analyzed by the targeted approach - right top part of Figure 4.10.

analysis (OPLS-DA) which also confirms a clear differentiation between our groups. The resulting score plot is displayed in Figure 4.9 and the exceptional position of Patient 19 can be observed again. The interpretation of metabolites is obtained through S-plot (Figure 4.10). Here we can see 30 most important metabolites that lead to separation of patients and controls. These are ordered by absolute values of reliability and metabolites with highest values are chosen as important markers. The importance of magnitude is evaluated as minor marker of significance. For better visualization also the zoom of S-plot is provided: left bottom part in Figure 4.11 and right top part in Figure 4.12. All red metabolites in the left bottom corner are typical for the group of patients. Again the same group of acylcarnitines with isobutylglycine can be seen in Figure 4.11 - all these metabolites are potential markers of the disease and they are elevated in patient samples. Metabolites in Figure 4.12 are typical for the group of controls. This group is much smaller than the previous one and metabolites are slightly different than in other methods.

The position of acylcarnitines in samples of horse patients with the atypical myopathy are validated in the literature [129]. Our results seem to be reasonable in comparison with their conclusions. A quite big group of acylcarnitines thus can be considered as markers of this disease.

4.3. Untargeted analysis of serum samples

In this section, we try to point out some differences between metabolomic data coming from targeted and untargeted approaches because the untargeted approach returns slightly different results than the targeted approach. The difference is in number of metabolites included in the analysis which can cause also differences in results. We can find, for example, some new markers of the disease by the untargeted approach in metabolites which are not included in the targeted one. Accordingly, the second data set connected with atypical myopathy of horses comes from analysis of serum samples by the untargeted approach. As a consequence, at the beginning 777 peaks (potential metabolites marked by numerical codes) are considered for 12 control and 19 patient samples in this approach. The QC samples are very different, so the influence of using signal correction is very different in particular peaks. For example, peak numbered 152.061966 has an increasing trend as visible from Figure 4.13. The final shape without any trend after the signal correction is displayed in Figure 4.14. We can see that internal relationship between samples is preserved; only the increasing trend is corrected.

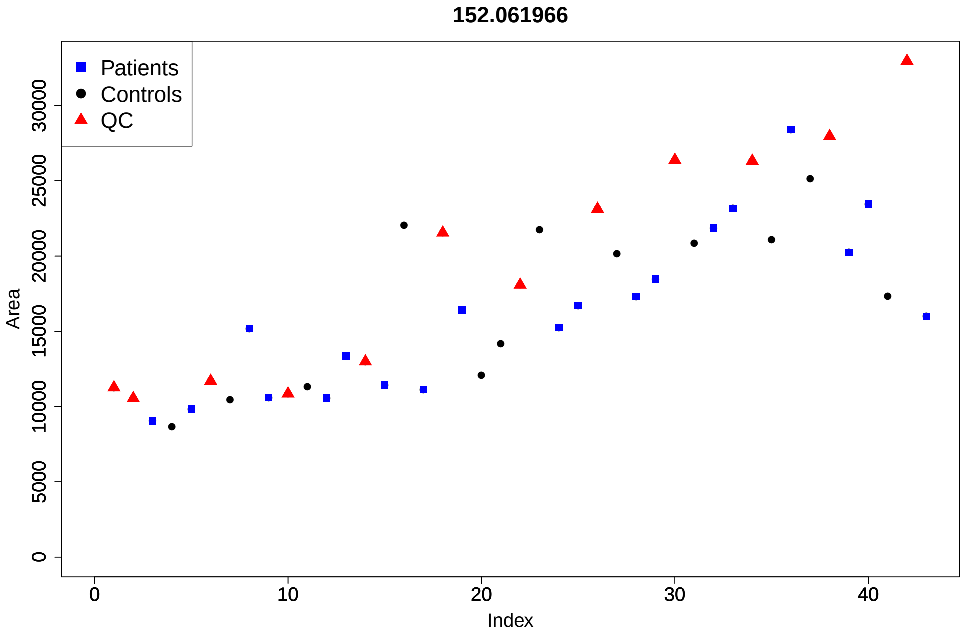


Figure 4.13: Time flow of the peak 152.061966 - raw data.

The range of interpolated values of QC samples is very broad, some values are even negative. This is caused by the fact that the first QC has zero values for some peaks and its interpolated values processed by LOESS are calculated as negative. Seven peaks with this property are excluded from the analysis. The next step consists of a comparison between the maximum and minimum interpolated values of QC samples. Ratios of these values vary between 1 to 184, all peaks with this ratio higher than 10 are also excluded (finally 15 of them). Furthermore, peaks with CV higher than 30 % are omitted as well, together 207 of them in this data set. The final size of the data matrix is thus, 548 peaks and 31 samples,

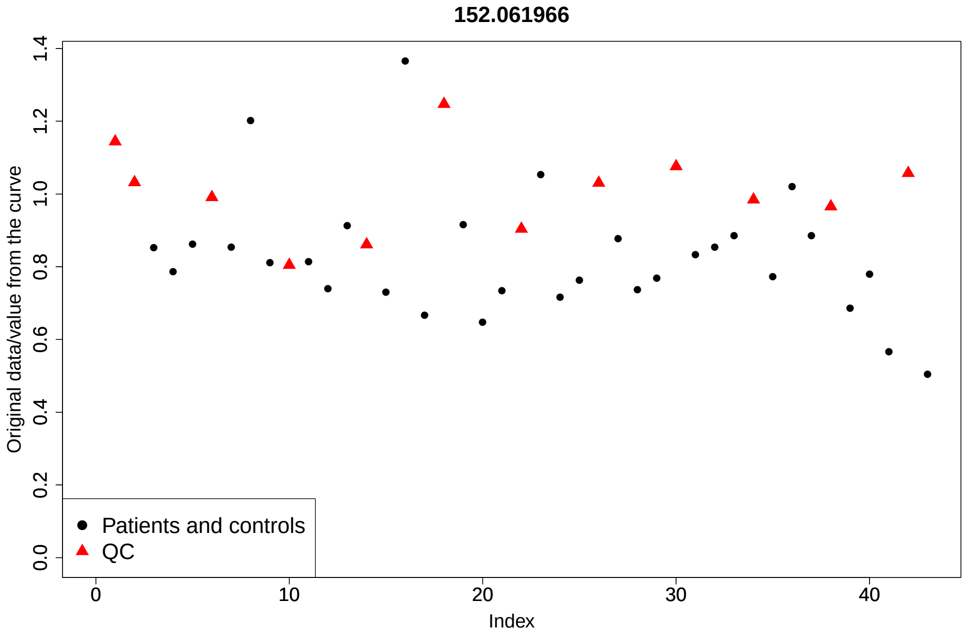


Figure 4.14: Time flow of the peak 152.061966 - corrected data.

which are grouped as controls and patients.

The last, very important part of the preprocessing consists of the imputation of zero values. A number of zeros was not high - here only 0.3% of all values. Zeros were imputed by the Algorithm 3.5.1. At the end, clr coordinates are applied and data table is mean centered.

The statistical analysis of this data set is more difficult than is the case of targeted approach. It is caused by a big amount of peaks, furthermore, some of them are not even metabolites but only fractions or noise. We are not able to distinguish between them and correct metabolites; it is just possible to find some known metabolites in a public database on the Internet, but this must be done manually for every peak. Peaks are denoted by numerical codes in our plots. These codes are called m/z (mass-to-change ratio) and represent the distribution of ions by mass in a sample. The characteristics m/z is very precise and it is presented in seven and more decimal numbers.

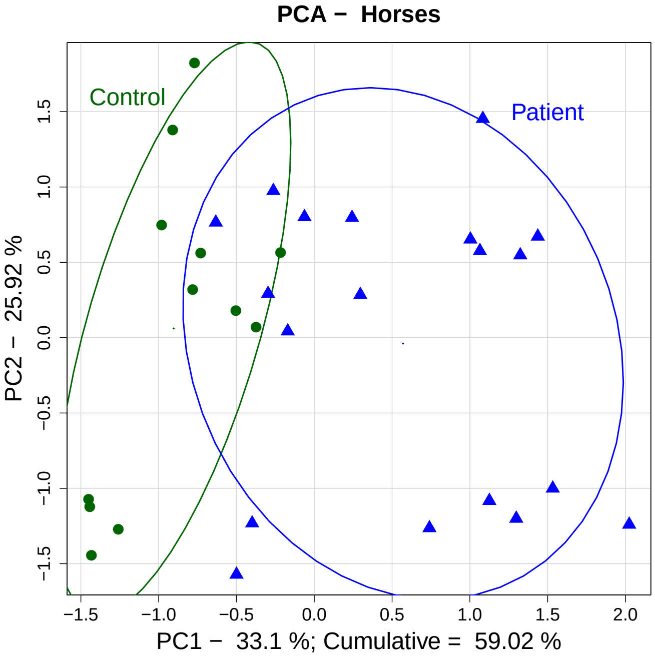


Figure 4.15: Score plot of principal component analysis of serum samples analyzed by the untargeted approach.

Principal component analysis is not as clear as in the case of targeted approach, but we can still distinguish between groups of controls and patients - see score plot in Figure 4.15. The percentage of explained variability is 59.02% which is sufficient for data with this dimension. A horizontal partition of samples also can be seen here. This trend is connected with peaks used in the analysis and becomes more visible from biplot in Figure 4.16, where again thirty loadings with longest rays that correspond to clr coefficients are observed. A group of peaks going to the lower part of the plot can be observed. These peaks are suspect to be the reason of the horizontal separation of samples, but we are not able to define them now. The only metabolites which can be identified in this plot are listed in Table 4.1. All metabolites are the same carnitines as in the targeted approach, whose values are increased in patient samples.

The difference between groups of controls and patients in term of acylcarnitines is also confirmed by boxplots. For example, a boxplot of the peak 316.24799

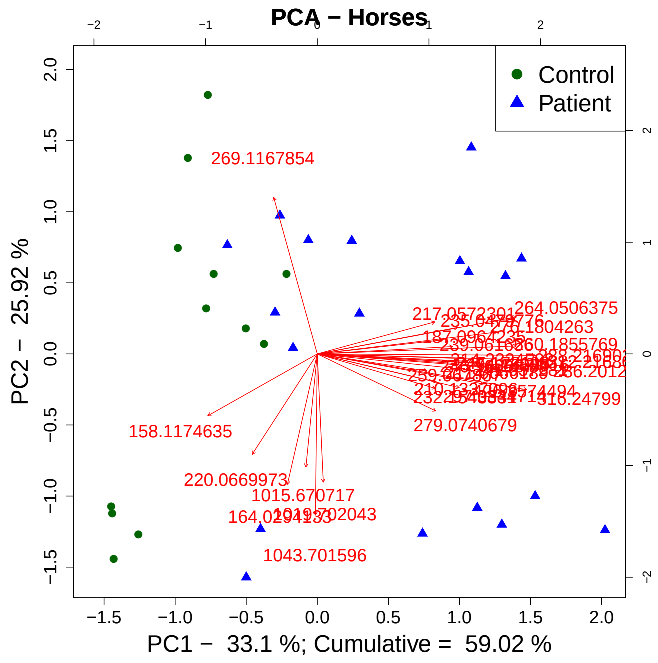


Figure 4.16: PCA biplot of serum samples analyzed by the untargeted approach.

m/z	Potential metabolite
260.1855769	C6
288.2169034	C8
314.2324508	C10:1
316.24799	C10

Table 4.1: Potential metabolites and their mass-to-change ratios.

(C10) is presented in Figure 4.17.

Groups of patients and controls are better splitted in partial least squares - discriminant analysis (score plot in Figure 4.18 and biplot in Figure 4.19), where the information on grouping of samples is taken into account as well. The outlyingness of Patient 19 pronounces now similarly as in the targeted approach; elevated acylcarnitines are also preserved. The problem with high VIP scores is also present in the untargeted approach. VIP greater than one is performed in 80 metabolites, only first 30 of them are displayed in Figure 4.20 and their values are greater than 2.2. The first VIP score is even higher than 6. In the VIP scores

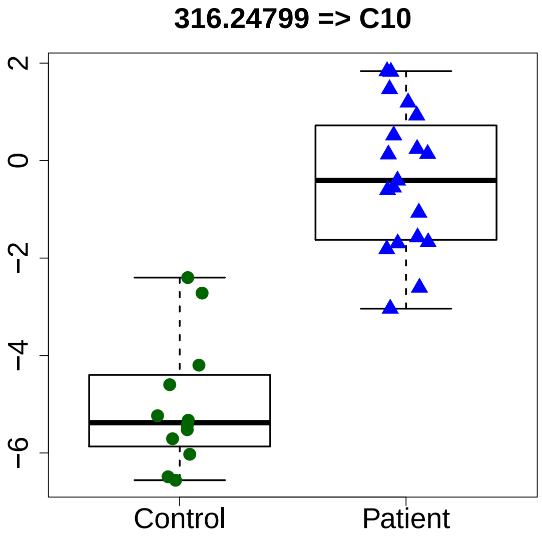


Figure 4.17: Boxplots of the peak 316.24799.

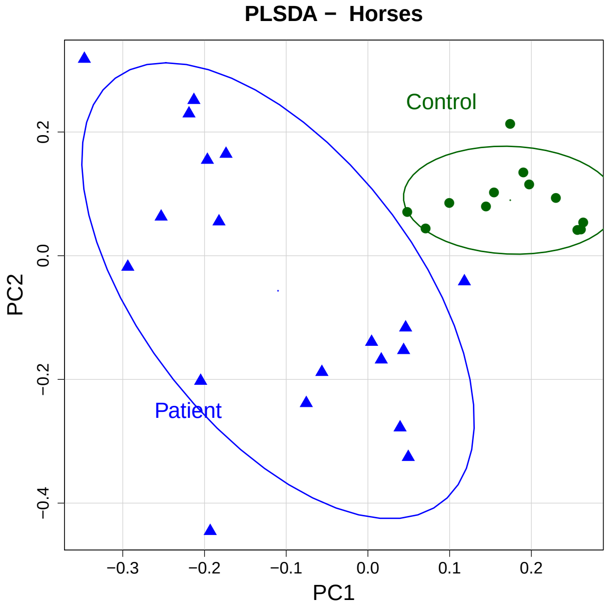


Figure 4.18: Score plot of partial leas squares - discriminant analysis of serum samples analyzed by the untargeted approach.

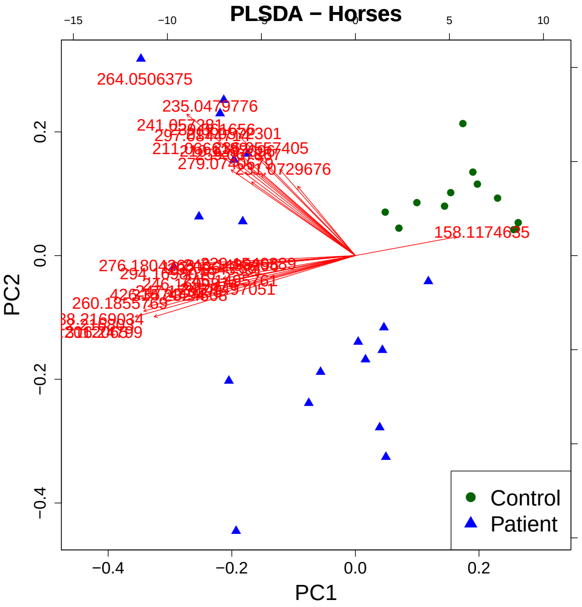


Figure 4.19: Biplot of partial leas squares - discriminant analysis of serum samples analyzed by the untargeted approach.

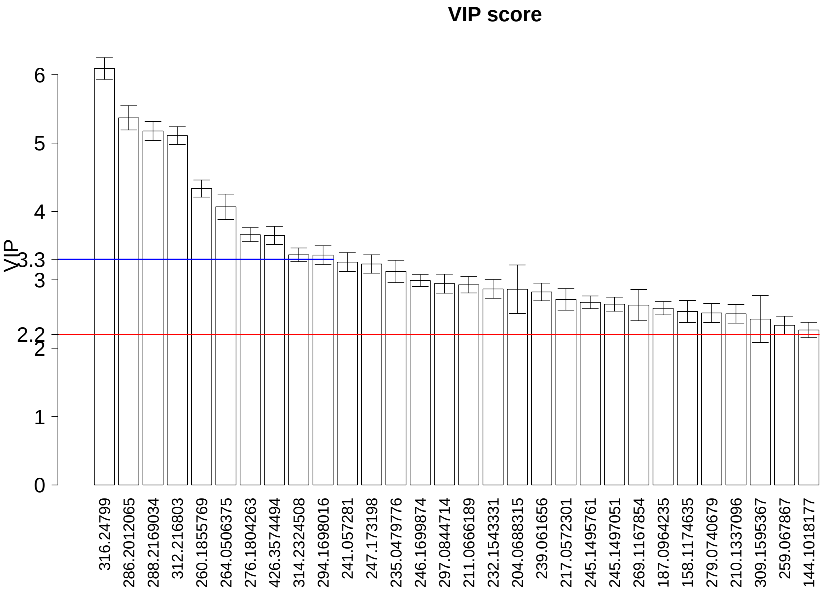


Figure 4.20: VIP scores of PLS-DA of serum samples analyzed by the untargeted approach.

plot elevated acylcarnitines are contained in the group of the first ten highest VIP scores (VIP > 3.3).

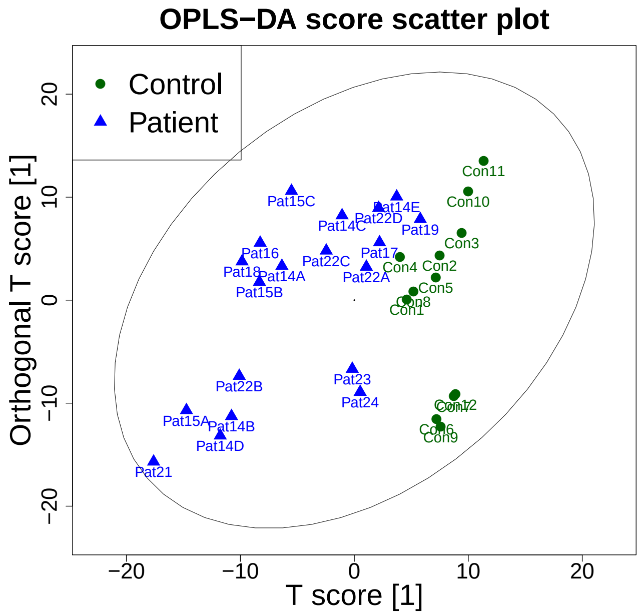


Figure 4.21: Score plot of orthogonal partial leas squares - discriminant analysis of serum samples analyzed by the untargeted approach.

The score plot of OPLS-DA (Figure 4.21) shows a similar structure of samples as PCA - one layer cluster divided into patient and control samples and three smaller clusters - through separation of controls and patients is clearly better now. Two of these smaller clusters are formed by patient samples and one remaining is formed by control samples. Patients 23 and 24 from the smallest cluster had a specific progression of the disease, separation of the other two smaller clusters is not so clear. Because this pattern didn't occur for the targeted approach, we can assume that levels of some specific peaks are very different for these smaller groups. Unfortunately, we are not able to identify them now.

If we analyze S-plots (Figures 4.22 - 4.24), previously identified carnitines can be found in positions in the left bottom corner of the plot. These peaks are very important for the group of patients; particularly, C8 is in the first position from the bottom, C10 is 4th, C6 is 6th and C10:1 occupies the 7th position.

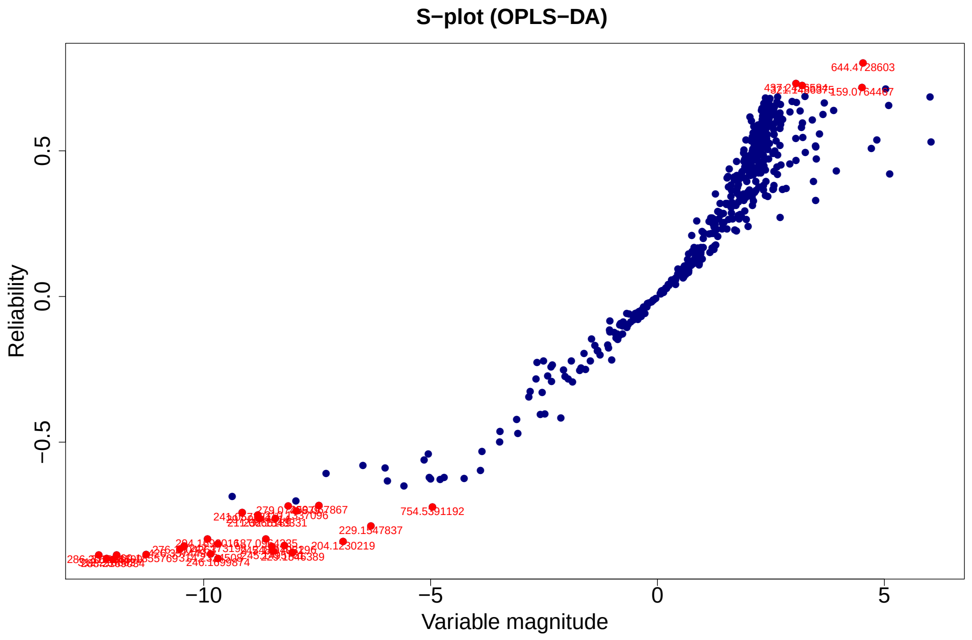


Figure 4.22: S-plot of serum samples analyzed by the untargeted approach.

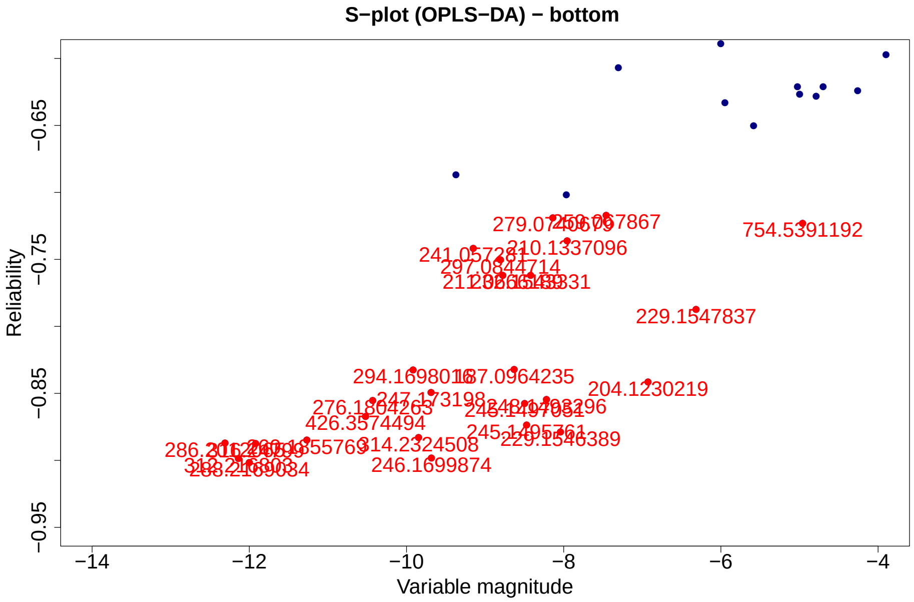


Figure 4.23: S-plot of serum samples analyzed by the untargeted approach - left bottom part of Figure 4.22.

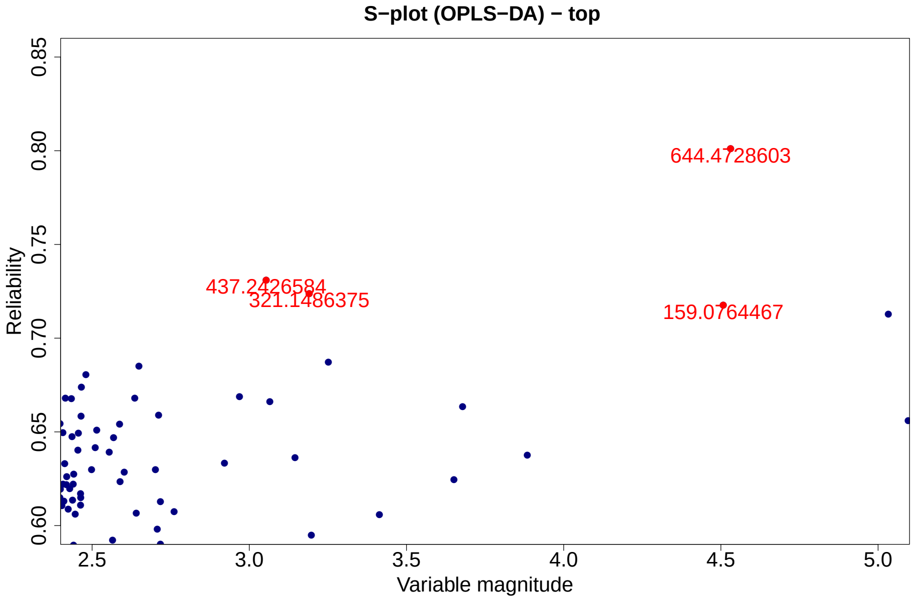


Figure 4.24: S-plot of serum samples analyzed by the untargeted approach - right top part of Figure 4.22.

The conclusion is the same as for the targeted approach. Chosen acylcarnitines are possibly very important markers of the atypical myopathy because they are placed in the first positions in lists of important metabolites in all methods used for the statistical analysis. We are able to identify them from the group of 548 peaks that further supports relevancy of the conclusion. On the other hand, some new peaks (not identified yet) occur in this data set, which cause the separation of patients with specific progression of the disease. These peaks are not visible in the targeted analysis, which illustrates the main differences between targeted and untargeted approaches.

4.4. Targeted analysis of urine samples

As a third approach, the targeted analysis of urine samples of horses suffered from atypical myopathy is performed. This approach is presented for the complex view on the data. The urine was not able for all horses, thus, only 5 controls and 6 patients are included in the analysis. The structure of metabolites presented in urine is different than is serum; accordingly, the slightly different list of metabolites is considered. The resulting data matrix has 11 rows and 165 columns. Data were normalized by AUC. Though, we are interested, how different origins of metabolites (serum, urine) affect results of their statistical processing using the logratio methodology.

Because of the targeted analysis, results of the preprocessing are very similar to those from Section 4.2. Accordingly, QC samples are very stable, also, the range of interpolated values is the same as in the case of serum samples. All coefficients of variation (CV) are lower than 30 % and zero values are not present in data.

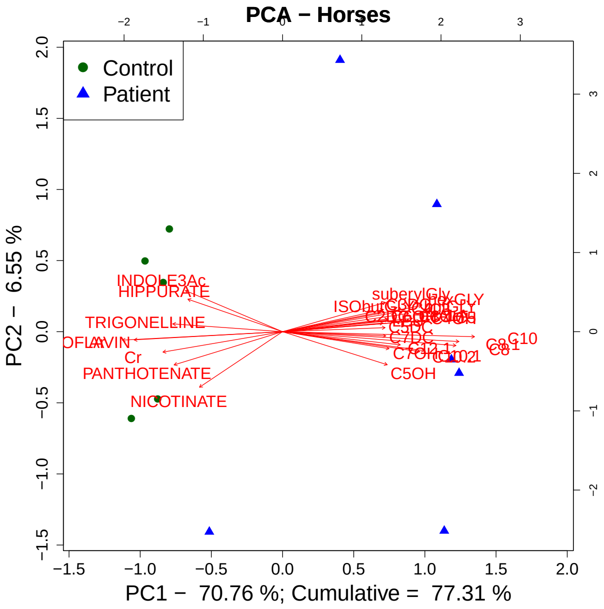


Figure 4.25: PCA biplot of urine samples analyzed by the targeted approach.

Only selected graphical results are discussed in this section. Firstly, PCA biplot is displayed in Figure 4.25. The difference between groups of samples and controls is nicely visible; the outlier close to the control group is the known Patient 19. The specific Patient 19 is preserved in all data sets on purpose. It is here for the demonstration of stability of all our methods. This sample enables us to compare all our methods and types of analyzes. The list of important metabolites (in terms of their expressive clr coefficients) for the patient group is very similar in the case of serum samples. Particularly, the group of acylcarnitines and isobutylglycine are preserved here. On the other hand, arrows in the direction of controls are slightly different, as a result of different materials (serum and urine). The reason is that metabolomic composition of urine of healthy controls naturally differs from the serum of the same controls. The percentage of explained variability is high (77.31%) which enables for very realistic conclusions.

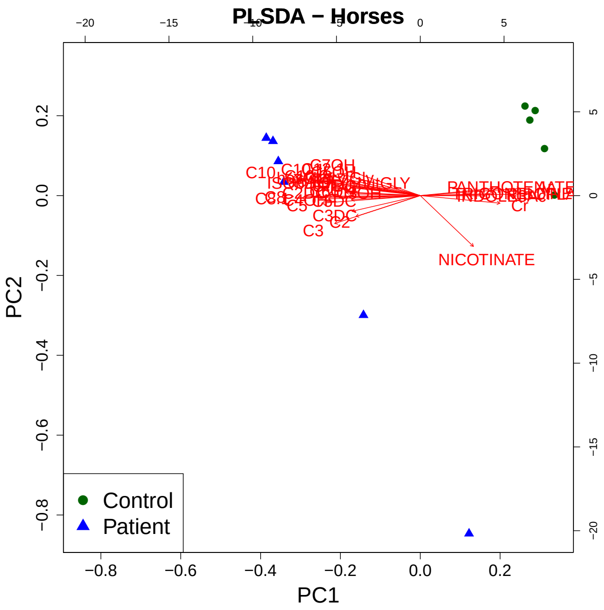


Figure 4.26: Biplot of partial least squares - discriminant analysis of urine samples analyzed by the targeted approach.

Similar results are provided also by PLS-DA biplot, see Figure 4.26. The only difference concerns Patient 19, being now an outlier of all samples in the right bottom corner of the plot. The reason for this behavior may be metabolite called nicotinate. The concentration of this metabolite is 33 times higher in the urine of Patient 19 than in the other samples. The only exception is Control 6, which has a concentration of this metabolite 12 times higher than average. This control is the last point in the cluster of controls close to the outlying patient. The loading

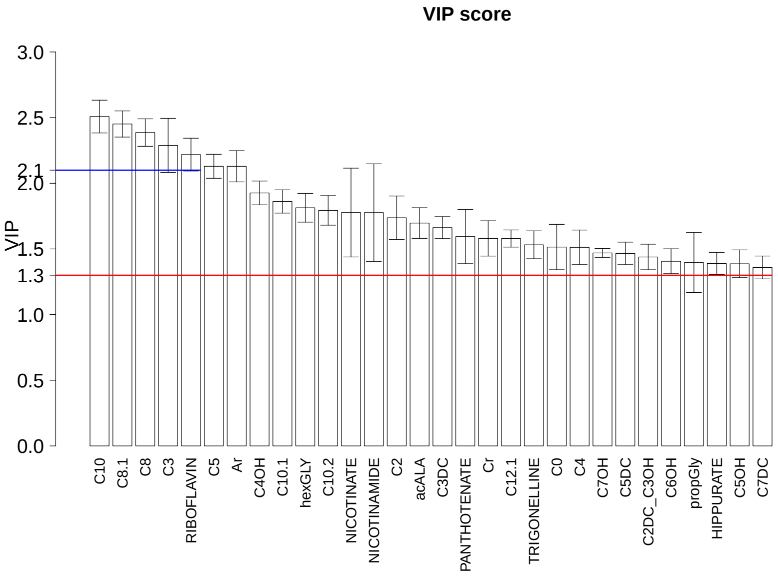


Figure 4.27: VIP scores of PLS-DA of urine samples analyzed by the targeted approach.

vector of nicotinate is visible through the respective arrow in both previous plots. Note that the position of acylcarnitines is still preserved in Figure 4.26.

The position of leading metabolites, acylcarnitines, is also confirmed by the VIP scores plot in Figure 4.27. First four metabolites with the highest VIP scores (VIP > 2.1) are acylcarnitines, followed by riboflavin. From data matrix in clr coordinates, it is easy to see that riboflavin has high values for all controls and Patient 19. On the contrary, concentrations of riboflavin in patient samples are very low. As a consequence, riboflavin may be a further marker of atypical myopathy in the case of urine samples. The arrows of riboflavin are also visible in Figures 4.25 and 4.26.

Last three plots in this section are connected with the S-plot of OPLS-DA, see Figures 4.28 - 4.30. In Figure 4.29 the left bottom part of this plot is zoomed. It includes metabolites important for the group of patients and here acylcarnitines are placed. By zooming the left upper part of the S-plot in Figure 4.30 riboflavin occurs in the first position of important metabolites for controls.

Riboflavin was also measured in the case of serum samples. The difference

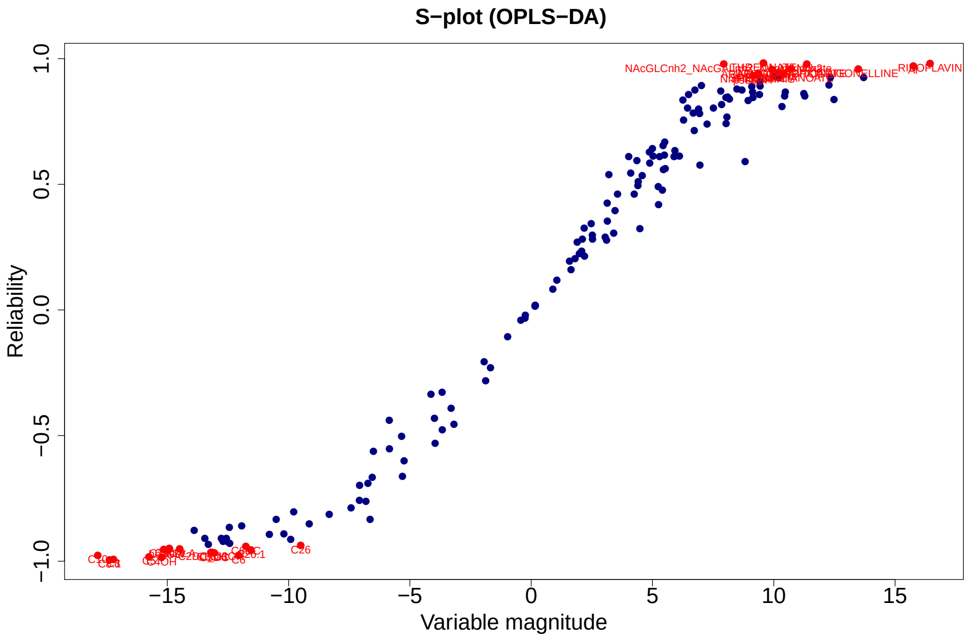


Figure 4.28: S-plot of urine samples analyzed by the targeted approach.

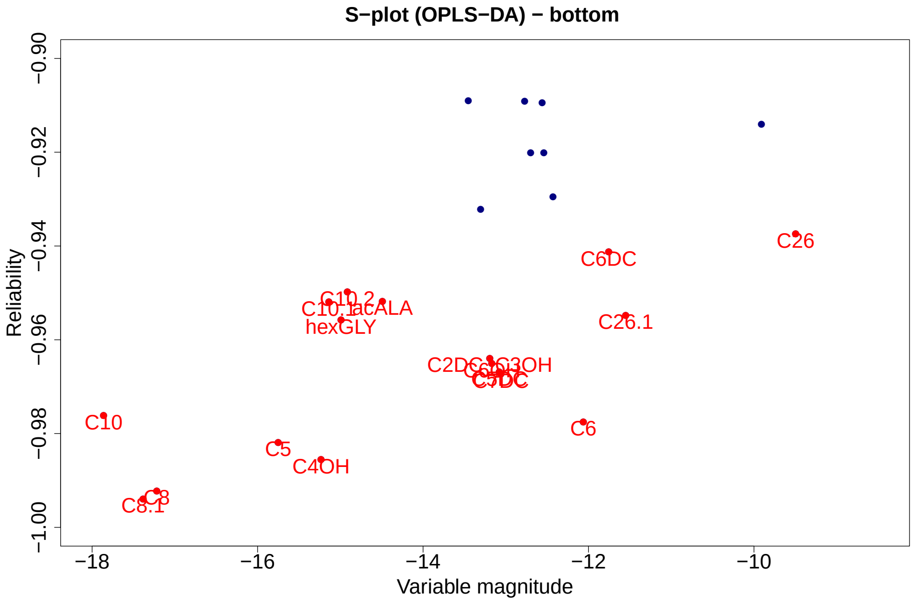


Figure 4.29: S-plot of urine samples analyzed by the targeted approach - left bottom part of Figure 4.28.

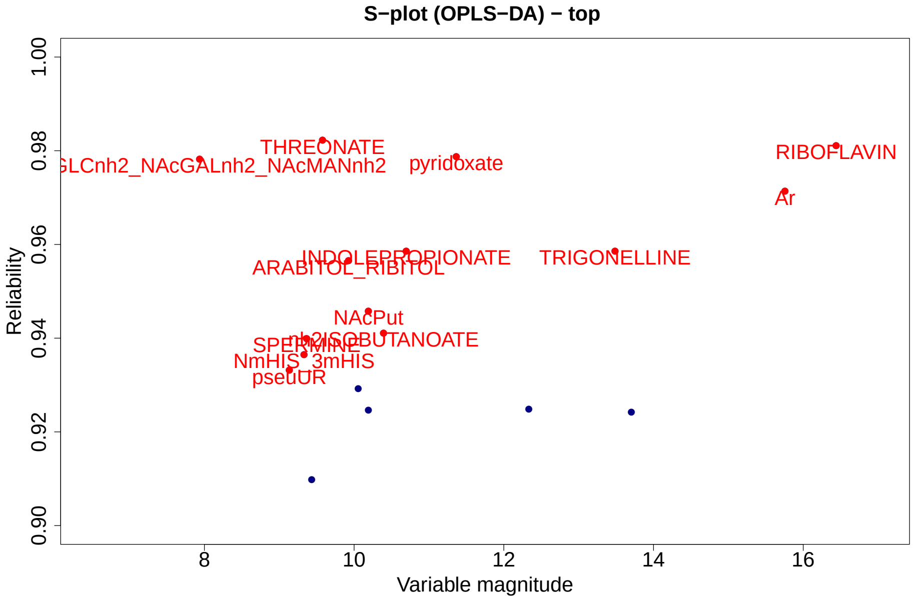


Figure 4.30: S-plot of urine samples analyzed by the targeted approach - right top part of Figure 4.28.

between controls and samples was not present - concentrations are rather different inside groups of patients and controls. Nicotinate was not measured in serum samples.

To conclude, acylcarnitines are confirmed as important markers of atypical myopathy using the logratio approach to statistical analysis of metabolomic (compositional) data. We also found supplementary marker - riboflavin. It is detected also in serum samples, but in urine samples, this metabolite appears to be much more expressive.

4.5. Comparison with the other transformations

Two comparisons with the other transformations are shown in this section. Firstly, log-transformation is applied to metabolomic data. The second transformation; that appears recently to be popular in chemometrics, is called probabilistic quotient normalization (PQN) [131, 132]. Data used in this section are serum samples evaluated by targeted analysis in Section 4.2, resulting in stable and complex outputs.

4.5.1. Logarithmic transformation

The application of the log-transformation is a very common method how to transform data in chemometrics. Without using any transformations (such as clr coordinates) metabolomic data are often skewly distributed and the logarithm is used to symmetrize the data distribution before centering and possible further processing. Nevertheless, it is not appropriate transformation for metabolomic data that have compositional nature; particularly, the logarithm does not take into account scale invariance of compositions.

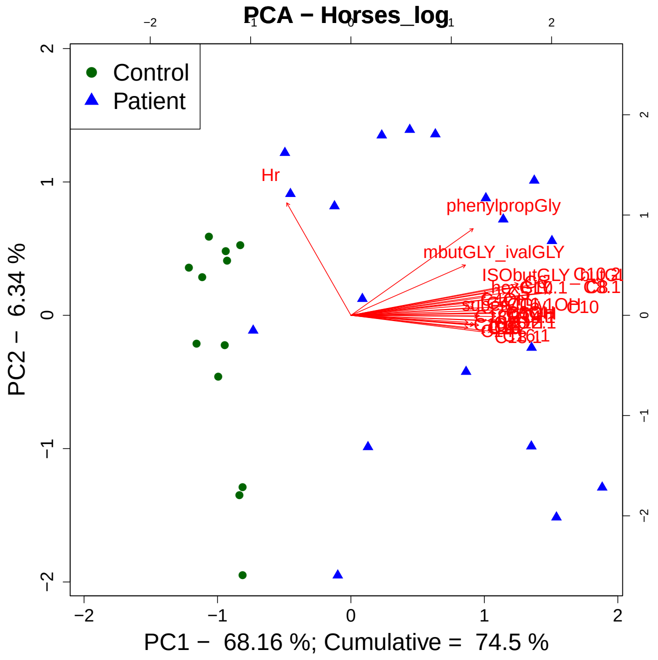


Figure 4.31: PCA biplot of serum samples analyzed by the targeted approach with log-transformation.

In this section, only basic outputs are presented that enable to recognize major differences between the logratio approach and clr coordinates. In Figure 4.31 the PCA biplot is displayed and it is easy to see that some differences, e.g. concerning acylcarnitines, occur. In Figure 4.3 fewer acylcarnitines are in the group

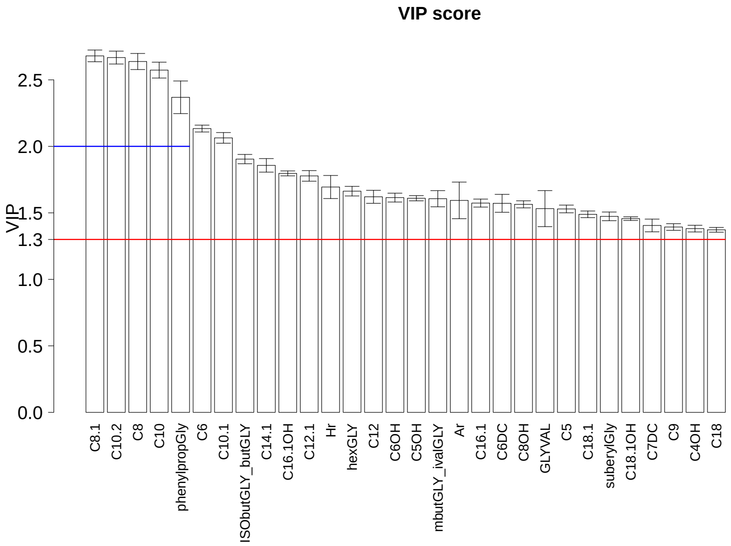


Figure 4.32: VIP scores of PLS-DA of serum samples analyzed by the targeted approach with log-transformation.

of most important metabolites and some metabolites are located also in the left part of the plot that enables to recognize metabolite characteristic for the group of control samples. In Figure 4.31 these metabolites from the left disappeared and more acylcarnitines are in the right part of the plot. Particularly, the PCA biplot of log-transformed metabolomic data is characterized by a distorted structure, where almost all arrows follow the same direction, which indicates damage of the covariance structure. This effect is known also from geochemistry [67] and indicates that using log-transformation might be misleading here.

The same trend as in PCA is also visible from the VIP scores plot of OPLS-DA which is displayed in Figure 4.32. In the list of 30 most important metabolites (with the highest VIP scores), more acylcarnitines than for the logratio approach are contained. This trend is also present in S-plot, which is not shown here.

4.5.2. The probabilistic quotient normalization

The second popular way of transformation of chemometric data, discussed here, is the probabilistic quotient normalization (PQN) [131, 132]. PQN is based on the calculation of a most probable dilution factor by looking at the distribution of the quotients of the amplitudes of a test spectrum by those of a reference spectrum [131]. According to recent studies [132], PQN appears to be the best transformation for metabolomic data sets.

The principle of the PQN is as follows. The size effect s_i is calculated for every sample by the median of the ratios of parts in the i-th sample, $i = 1, \ldots, n$, with the corresponding elements of a reference sample $\mathbf{x}^{ref}$: $s_i = median(x_{i1}/x_1^{ref}, \ldots, x_{in}/x_n^{ref})$, where x_j^{ref} is the median of the jth metabolite. The final transformed data are given as [132]

$$\mathbf{x}_i^{PQN} = (x_{i1}/s_i, \ldots, x_{in}/s_i).$$

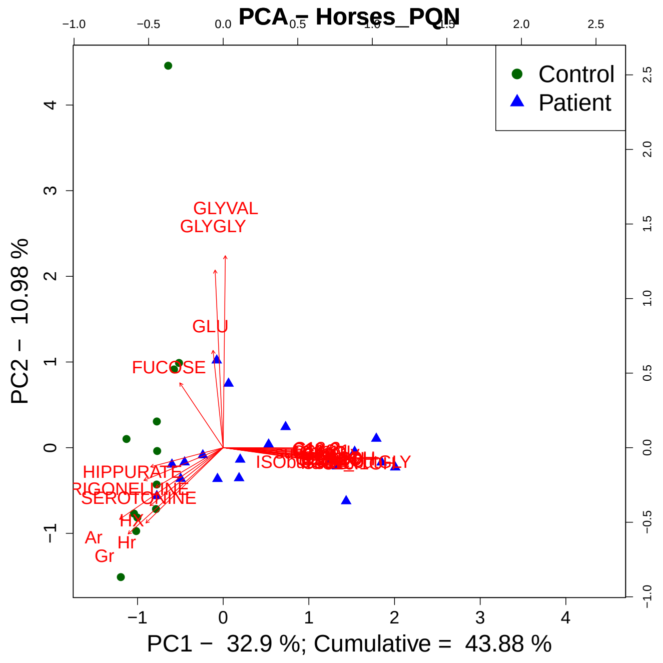


Figure 4.33: PCA biplot of serum samples analyzed by the targeted approach with PQN.

The PQN is applied to serum samples evaluated by the targeted approach, similarly as in the previous section. PCA biplot for centered data is shown in Figure 4.33. The difference is very clear in comparison with Figure 4.3. Differen-

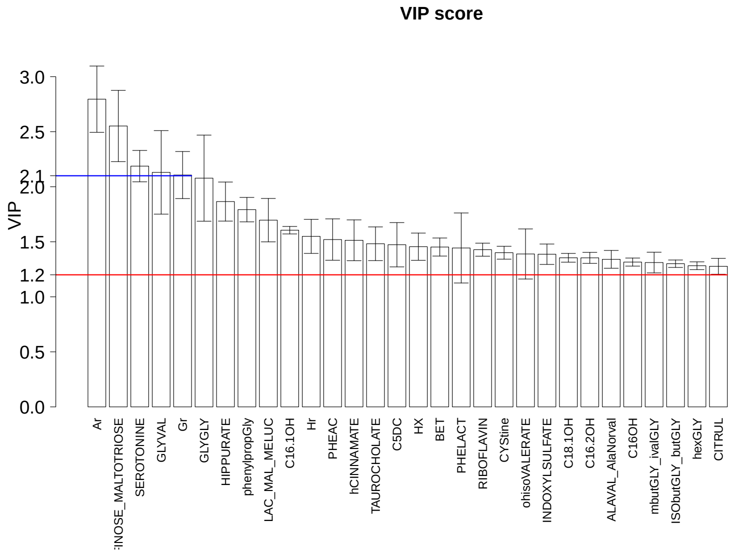


Figure 4.34: VIP scores of PLS-DA of serum samples analyzed by the targeted approach with PQN.

ces between clusters of patients and controls are still present, but the structure of metabolites (loadings) has damaged. Rays of acylcarnitines in the direction of patient samples are the same, but rays of the other metabolites have no clear meaning. For example, oligopeptides glycine (GLY), glycyl-valine (GLYVAL), glycyl-glycine (GLYGLY) and even fucose are highlighted as important, but these metabolites seem to be driven by only one control sample (the green point in the left upper corner of the plot). This specific position of this control is not visible in all previous PCA graphs, which means that the use of logratio approach and log-transformation are somehow robust in the case of the presence of individual outliers. Also, metabolites in the left part of the plot (near to the group of controls) are rather unusual. Finally, the percentage of variability explained by the model is much lower.

The difference in the first thirty important metabolites is also visible from the VIP scores plot in Figure 4.34. Only five acylcarnitines are in the list, the rest of metabolites mentioned there are less meaningful than metabolites from Figure 4.8. Also, the above mentioned glycyl-valine and glycyl-glycine are contained in the plot. The same trend is visible also from the S-plot, which is not shown in this thesis.

Note that logarithmic transformation applied to data transformed by PQN would lead to results comparable with those from Section 4.5.1. Moreover, PQN leads just to a particular representation of the input metabolomic (compositional) data, so their clr coordinates would be exactly the same as for Section 4.2. From this point of view, PQN can not be recommended as transformation for metabolomic data.

Conclusions

This thesis contains a comprehensive guide to the statistical analysis of metabolomic data using the logratio methodology as developed during my study at the Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University Olomouc and by cooperation with the Laboratory of Metabolomics, Institute of Molecular and Translational Medicine, Palacký University Olomouc. Procedures and algorithms introduced in this thesis are in everyday use in the Laboratory of Metabolomics [6, 7, 56].

Compositional data and specific coordinates that enable the statistical analysis in Euclidean space were introduced in Section 1. The specific field of biochemistry, called metabolomics, was proposed in Section 2. Section 3 is the largest in this thesis and presents specific tools used for the statistical analysis of compositional data. The basic preprocessing of such data including elemental algorithms for imputation of missing and zero values were introduced in Section 3.1. Multivariate methods for the statistical analysis were suggested starting with principal component analysis in Section 3.2. Partial least squares regression as a tool that is appropriate for regression analysis of high-dimensional data was introduced in Section 3.3 and the extension of this method, orthogonal partial least squares regression, was presented in Section 3.4. Special parametric models for imputation of zero values (rounded zeros) using partial least squares regression were presented in Section 3.5. For statistical analysis of three-way compositional data, the PARAFAC model was adapted in Section 3.6. The last Section 4 consist of the practical applications of all presented methods to real data sets from metabolomics.

While some methods, discussed in this thesis, are popular even in the context of compositional data, other need to be adapted with any support of previous developments. This was definitely the case of partial least squares regression in ilr coordinates, appropriate tool for statistical analysis of high-dimensional compositional data [6]. Similarly, although PARAFAC modeling of three-way compositional data was previously discussed in literature [40–42], its adaptation to metabolomics for the specific case of urine metabolomic data (including comparisons to standard approaches) wasn't published yet [59]. Finally, another novel part of the thesis is a parametric model for imputation of rounded zeros based on partial least squares regression and logratio methodology [58]. The presence of rounded zeros in metabolomic (and also chemometric) data is quite common and this algorithm may help to solve problems with data in a wide range of practical applications.

The most difficult part of this thesis was the necessity of complex view on data. Metabolomic data have a lot of specific features (they are high-dimensional, with specific covariance structure and mostly of compositional nature) and any reasonable method must take care of all of them. Proper statistical analysis of metabolomic samples is crucial for the reliability of the results for further interpretation and processing. In the chemometric and also metabolomic communities, compositional data are still considered as observations with a fixed constant sum constraint, although this is just a possible representation of the relative information, carried by the compositional parts, not an inherent property of the data. Note that the popular logarithmic transformation would solve the problem of moving the relative scale to the absolute one (necessary for a further reasonable statistical analysis), but just for single metabolites, without considering their relative multivariate relations to the other metabolites in the data set. Consequently, application of standard statistical techniques to raw or rescaled metabolomic data often leads to biased results due to ignoring the mathematical implications. On the other hand, the logratio approach to statistical analysis of compositional data is a well mathematically justified methodology that could provide a concise approach to the statistical treatment of biomarkers in metabolomics. Although absolute values of biomarkers compared to reference ranges (data from the healthy population) is the most frequently used approach, the nature of the resulting multivariate observations is a relative one, i.e. relative contributions of metabolites are of primary interest.

More possibilities for future extension of this thesis exist. The first one concerns robustness aspects of the above mentioned statistical analysis that enable to suppress the influence of outlier observations. They might destroy a picture of the multivariate structure of the observations as often results from classical statistical methods. Especially the robust version of logratio partial least squares regression may be very helpful. Also, a focus on sparse counterparts to methods like principal component analysis and partial least squares, that provide a substantial simplification by interpretation of the results, can be useful, but are particularly challenging in the case of high-dimensional compositional data. The last extension can be focused on the development of new interpretable logratio coordinates. Although recent experiences show clear advantages of logratio coordinates where the first coordinate aggregates information from log-ratios for a particular compositional part of interest, their usefulness is limited if there are distortions like rounding errors or other data "problems"in the involved parts. A possible way out is to use a "robust"(weighted) version of these coordinates, called weighted balances, where the remaining parts (with respect to the part of interest) in the first coordinate are weighted in a way that is relevant to the aims of the statistical analysis. Such weights can be, e.g., derived according to quality assessment analysis and elements of classical/robust variation matrix of compositions. Finally, a new library in the software R [43] summarizing the logratio approach to the statistical analysis of metabolomic data would be very useful.

I hope that the presented thesis helps to expansion of the logratio methodology also to the important field of metabolomic data, and to chemometrics in general.

References

[1] J. Aitchison, *The statistical analysis of compositional data.* Chapman & Hall, London, 1986.

[2] V. Pawlowsky-Glahn and A. Buccianti, *Compositional data analysis: Theory and applications.* Wiley, Chichester, 2011.

[3] V. Pawlowsky-Glahn, J. Egozcue, and R. Tolosana-Delgado, *Modeling and analysis of compositional data.* Wiley, Chichester, 2015.

[4] A. Beddek, P. Rawson, L. Peng, R. Snell, K. Lehnert, H. Ward, and T. Jordan, "Profiling the metabolic proteome of bovine mammary tissue," *Proteomics*, vol. 8, pp. 1502–1515, 2008.

[5] H. Janečková, K. Hron, P. Wojtowicz, E. Hlídková, A. Barešová, D. Friedecký, L. Žídková, P. Hornik, D. Behúlová, D. Procházková, H. Vinohradská, K. Pešková, P. Bruheim, V. Smolka, S. Šťastná, and T. Adam, "Targeted metabolomic analysis of plasma samples for the diagnosis of inherited metabolic disorders," *Journal of Chromatography A*, vol. 1226, pp. 11–17, 2012.

[6] A. Kalivodová, K. Hron, P. Filzmoser, L. Najdekr, H. Janečková, and T. Adam, "PLS-DA for compositional data with application to metabolomics," *Journal of Chemometrics*, vol. 29, pp. 21–28, 2015.

[7] L. Najdekr, A. Gardlo, L. Mádrová, D. Friedecký, H. Janečková, E. Correa, R. Goodacre, and T. Adam, "Oxidized phosphatidylcholines suggest oxidative stress in patients with medium-chain acyl-CoA dehydrogenase deficiency," *Talanta*, vol. 139, pp. 62–66, 2015.

[8] J. Denes, E. Szabo, S. Robinette, I. Szatmari, L. Szonyi, J. Kreuder, E. Rauterberg, and Z. Takats, "Metabonomics of newborn screening dried blood spot samples: A novel approach in the screening and diagnostics of inborn errors of metabolism," *Analytical Chemistry*, vol. 84, pp. 10113–10120, 2012.

[9] A. Fernie, R. Trethewey, A. Krotzky, and L. Willmitzer, “Metabolite profiling: from diagnostics to systems biology,” *Nature Reviews Molecular Cell Biology*, vol. 5, pp. 763–769, 2004.

[10] M. Sysi-Aho, M. Katajamaa, L. Yetukuri, and M. Orešič, “Normalization method for metabolomics data using optimal selection of multiple internal standards,” *Bioinformatics*, vol. 8, no. 93, 2007.

[11] W. Cleveland, “Robust locally weighted regression and smoothing scatterplots,” *Journal of the American Statistical Association*, vol. 74, no. 368, pp. 829–836, 1979.

[12] W. B. Dunn, D. Broadhurst, P. Begley, E. Zelena, S. Francis-McIntyre, N. Anderson, M. Brown, J. Knowles, A. Halsall, J. Haselden, A. W. Nicholls, I. Wilson, D. Kell, and R. Goodacre, “Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry,” *Nature Protocols*, vol. 6, no. 7, pp. 1060–1083, 2011.

[13] R. Little and D. Rubin, *Statistical analysis with missing data.* Wiley, Hoboken, 2002.

[14] B. Walczak and D. Massart, “Dealing with missing data. Part I,” *Chemometrics and Intelligent Laboratory Systems*, vol. 58, pp. 15–27, 2001.

[15] K. Hron, M. Templ, and P. Filzmoser, “Imputation of missing values for compositional data using classical and robust methods,” *Computational Statistics & Data Analysis*, vol. 54, no. 12, pp. 3095–3107, 2010.

[16] J. Martín-Fernández, C. Barceló-Vidal, and V. Pawlowsky-Glahn, “Dealing with zeros and missing values in compositional data sets using nonparametric imputation,” *Mathematical Geology*, vol. 35, no. 3, pp. 253–278, 2003.

[17] J. A. Martín-Fernández, J. Palarea-Albaladejo, and R. A. Olea, “Dealing with zeros,” in *Compositional data analysis: Theory and applications* (V. Pawlowsky-Glahn and A. Buccianti, eds.), pp. 43–58, Wiley, Chichester, 2011.

[18] J. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser, and J. Palarea-Albaladejo, “Model-based replacement of rounded zeros in compositional data: Classical and robust approaches,” *Computational Statistics and Data Analysis*, vol. 56, no. 9, pp. 2688–2704, 2012.

[19] R. Brereton, *Chemometrics for pattern recognition.* Wiley, Chichester, 2009.

[20] R. Goodacre, D. Broadhurst, A. Smilde, B. Kristal, J. Baker, R. Beger, C. Bessant, S. Connor, G. Capuani, C. Craig, T. Ebbels, D. Kell, C. Manetti, G. Newton, J. Paternostro, G. Somorjai, M. Sjöström, J. Trygg, and F. Wulfert, "Proposed minimum reporting standards for data analysis in metabolomics," *Metabolomics*, vol. 3, pp. 231–241, 2007.

[21] B. Warracka, S. Hnatyshyna, K. Otta, M. Reilya, M. Sandersa, H. Zhanga, and D. M. Drexler, "Normalization strategies for metabonomic analysis of urine samples," *Journal of Chromatography B*, vol. 877, pp. 547–552, 2009.

[22] S. Waikar, V. S. Sabbisetti, and J. Bonventre, "Normalization of urinary biomarkers to creatinine during changes in glomerular filtration rate," *Kidney International*, vol. 78, no. 5, pp. 486–494, 2010.

[23] Y. Chen, G. Shen, R. Zhang, J. He, Y. Zhang, J. Xu, W. Yang, X. Chen, Y. Song, and Z. Abliz, "Combination of injection volume calibration by creatinine and MS signals' normalization to overcome urine variability in LC-MS-based metabolomics studies," *Analytical Chemistry*, vol. 85, pp. 7659–7665, 2013.

[24] A. Sauve and T. Speed, "Normalization, baseline correction and alignment of high-throughput mass spectrometry data.," *Proceedings of the Genomic signal processing and statistics workshop, Baltimore, MO, USA, May 26-27*, pp. http://stat–www.berkeley.edu/users/terry/Group/publications/Final2Gensips 2004Sauve.pdf, 2004.

[25] E. T. Fung and C. Enderwick, "Proteinchip clinical proteomics: Computational challenges and solutions," *Computational Proteomics Supplement*, vol. 32, pp. S34–S41, 2002.

[26] O. Haglund, "Qualitative comparison of normalization approaches in maldims," *Master of science thesis, Royal Institute of Technology, Stockholm, Sweden*, 2008.

[27] K. Varmuza and P. Filzmoser, *Introduction to multivariate statistical analysis in chemometrics.* Taylor & Francis, New York, 2009.

[28] P. Gemperline, *Practical guide to chemometrics, 2nd edition.* Taylor & Francis, Boca Raton, 2006.

[29] J. Aitchison and M. Greenacre, "Biplots of compositional data," *Journal of the Royal Statistical Society*, vol. 51, no. 4, pp. 375–392, 2002.

[30] M. Bogdanov, W. Matson, L. Wang, T. Matson, R. Saunders-Pullman, S. Bressman, and M. Beal, "Metabolomic profiling to develop blood biomarkers for Parkinson's disease," *Brain*, vol. 131, no. 2, pp. 389–396, 2008.

[31] J. Jonsson, J. Gullberg, A. Nordström, M. Kusano, M. Kowalczyk, M. Sjöström, and M. Moritz, “A strategy for identifying differences in large series of metabolomic samples analyzed by GC/MS,” *Analytical Chemistry*, vol. 76, pp. 1738–1745, 2004.

[32] S. Bijlsma, I. Bobeldijk, E. Verheij, R. Ramaker, S. Kochhar, I. Macdonald, B. van Ommen, and A. Smilde, “Large-scale human metabolomics studies: A strategy for data (pre-) processing and validation,” *Analytical Chemistry*, vol. 78, pp. 567–574, 2006.

[33] E. Szymańska, E. Saccenti, A. Smilde, and J. Westerhuis, “Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies,” *Metabolomics*, vol. 8, pp. S3–S16, 2012.

[34] M. Pérez-Enciso and M. Tenenhaus, “Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach,” *Human Genetics*, vol. 112, pp. 581–592, 2003.

[35] J. Trygg, E. Holmes, and E. Lundstedt, “Chemometrics in metabonomics,” *Journal of Proteome Research*, pp. 469–479, 2007.

[36] S. Wiklund, E. Johansson, L. Sjöström, E. Mellerowicz, U. Edlund, J. Shockcor, J. Gottfries, T. Moritz, and J. Trygg, “Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models,” *Analytical Chemistry*, vol. 80, pp. 115–122, 2008.

[37] J. Carroll and J. Chang, “Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition,” *Psychometrika*, vol. 35, pp. 283–319, 1970.

[38] R. Harshman, “Foundations of the Parafac procedure: Models and conditions for an ”explanatory”multimodal factor analysis,” *UCLA Working Papers in Phonetics*, vol. 16, pp. 1–84, 1970.

[39] L. Tucker, “Some mathematical notes on three-mode factor analysis,” *Psychometrika*, vol. 31, pp. 279–311, 1966.

[40] M. Gallo, “Log-ratio and parallel factor analysis: An approach to analyze three-way compositional data,” in *Advanced dynamic modeling of economic and social systems* (A. N. Proto, M. Squillante, and J. Kacprzyk, eds.), vol. 448 of *Studies in Computational Intelligence*, pp. 209–221, Springer, Heidelberg, 2013.

[41] M. A. Engle, M. Gallo, K. T. Schroeder, N. J. Geboy, and J. W. Zupancic, “Three-way compositional analysis of water quality monitoring data,” *Environmental and Ecological Statistics*, vol. 21, no. 3, pp. 565–581, 2014.

[42] A. Di Palma, M. Gallo, P. Filzmoser, and K. Hron, "A robust Candecomp/Parafac model for compositional data," *Submitted*, 2015.

[43] R Development Core Team, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

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

[45] C. Smith, E. Want, G. O'Maille, R. Abagyan, and G. Siuzdak, "XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification," *Analytical Chemistry*, vol. 78 (3), pp. 779–787, 2006.

[46] R. Tautenhahn, C. Bottcher, and S. Neumann, "Highly sensitive feature detection for high resolution LC/MS," *BMC Bioinformatics*, vol. 9, p. 504, 2008.

[47] H. Benton, E. Want, and T. Ebbels, "Correction of mass calibration gaps in liquid chromatography - mass spectrometry metabolomics data," *Bioinformatics*, vol. 26, no. 19, pp. 2488–2489, 2010.

[48] C. Kuhl, R. Tautenhahn, C. Bottcher, T. Larson, and S. Neumann, "CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets," *Analytical Chemistry*, vol. 84, no. 1, pp. 283–289, 2012.

[49] E. Gaude, F. Chignola, D. Spiliotopoulos, S. Mari, A. Spitaleri, and M. Ghitti, *muma: Metabolomics univariate and multivariate analysis*, 2012. R package version 1.4.

[50] J. Palarea-Albaladejo and J. Martín-Fernández, "zCompositions — R package for multivariate imputation of left-censored data under a compositional approach," *Chemometrics and Intelligent Laboratory Systems*, vol. 143, pp. 85–96, 2015.

[51] B. Mevik, R. Wehrens, and K. Liland, *pls: Partial least squares and principal component regression*, 2013. R package version 2.4-3.

[52] D. Leibovici and R. Sabatier, "A singular value decomposition of k-way array for a principal component analysis of multiway data, PTA-k," *Linear Algebra and its Applications*, vol. 269, pp. 307–329, 1998.

[53] D. Leibovici, "Spatio-temporal multiway decompositions using principal tensor analysis on k-modes: the R package PTAk," *Journal of Statistical Software*, vol. 34, no. 10, pp. 1–34, 2010.

[54] P. Giordani, H. Kiers, and M. Del Ferraro, "Three-way component analysis using the R package ThreeWay," *Journal of Statistical Software*, vol. 57, no. 7, pp. 1–23, 2014.

[55] C. Kanagaratham, A. Kalivodová, L. Najdekr, D. Friedecký, T. Adam, D. Moreno, J. V. Garmendia, M. Hajduch, J. B. De Sanctis, and D. Radzioch, "Fenretinide prevents inflammation and airway hyperresponsiveness in a mouse model of allergic asthma," *American Journal of Respiratory Cell and Molecular Biology*, vol. 51, no. 6, pp. 783–792, 2014.

[56] H. Janečková, A. Kalivodová, L. Najdekr, D. Friedecký, K. Hron, P. Bruheim, and T. Adam, "Untargeted metabolomic analysis of urine samples in the diagnosis of some inherited metabolic disorders," *Biomedical Papers*, vol. 159, no. 4, pp. 582–585, 2015.

[57] J. Veleba, J. Kopecký, P. Janovská, O. Kuda, O. Horáková, H. Malinská, L. Kazdova, O. Oliyarnyk, V. Škop, J. Trnovská, M. Hájek, A. Škoch, P. Flachs, K. Bardová, M. Rossmeisl, J. Olza, G. Salim de Castro, P. Calder, A. Gardlo, E. Fišerová, J. Jensen, M. Bryhn, J. Kopecký, and T. Pelikánová, "Combined intervention with pioglitazone and n-3 fatty acids in metformin-treated type 2 diabetic patients: improvement of lipid metabolism," *Nutrition & Metabolism*, vol. 12, no. 52, pp. 1–15, 2015.

[58] M. Templ, K. Hron, P. Filzmoser, and A. Gardlo, "Imputation of rounded zeros for high-dimensional compositional data," *accepted to Chemometrics and Intelligent Laboratory Systems*, 2016.

[59] A. Gardlo, A. Smilde, K. Hron, M. Hrdá, R. Karlíková, and T. Adam, "Normalization techniques for PARAFAC modeling of urine metabolomics data," *submitted*, 2016.

[60] O. Horáková, J. Hansíková, K. Bardová, M. Gardlo, A. Rombaldová, O. Kuda, M. Rossmeisl, and J. Kopecký, "Plasma acylcarnitines and amino acid levels as an early complex biomarker of propensity to high-fat diet-induced obesity in mice," *submitted*, 2016.

[61] R. Karlíková, J. Široká, P. Jahn, D. Friedecký, A. Gardlo, H. Janečková, F. Hrdinová, Z. Drábková, and T. Adam, "Atypical myopathy of grazing horses: a metabolic study," *submitted*, 2016.

[62] R. Karlíková, J. Široká, D. Friedecký, E. Faber, M. Hrdá, K. Mičová, I. Fikarová, A. Gardlo, H. Janečková, I. Vrobel, and T. Adam, "Metabolite profiling of the plasma and leukocytes of chronic myeloid leukemia patients," *submitted*, 2016.

[63] G. Mateu-Figueras and V. Pawlowsky-Glahn, “A critical approach to probability laws in geochemistry,” *Mathematical Geosciences*, vol. 40, no. 5, pp. 489–502, 2008.

[64] J. Egozcue, “Reply to “On the Harker variation diagrams; …” by J.A. Cortés,” *Mathematical Geosciences*, vol. 41, no. 7, pp. 829–834, 2009.

[65] J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal, “Isometric logratio transformations for compositional data analysis,” *Mathematical Geology*, vol. 35, no. 3, pp. 279–300, 2003.

[66] M. Eaton, *Multivariate statistics. A vector space approach.* John Wiley & Sons, New York, 1983.

[67] P. Filzmoser, K. Hron, and C. Reimann, “Principal component analysis for compositional data with outliers,” *Environmetrics*, vol. 20, pp. 621–632, 2009.

[68] K. Hron, M. Jelínková, P. Filzmoser, R. Kreuziger, P. Bednář, and P. Barták, “Statistical analysis of wines using a robust compositional biplot,” *Talanta*, vol. 90, pp. 46–50, 2012.

[69] M. Korhoňová, K. Hron, D. Klimčíková, L. Müller, P. Bednář, and P. Barták, “Coffee aroma - statistical analysis of compositional data,” *Talanta*, vol. 80, pp. 710–715, 2009.

[70] J. Egozcue and V. Pawlowsky-Glahn, “Groups of parts and their balances in compositional data analysis,” *Mathematical Geology*, vol. 37, pp. 795–828, 2005.

[71] K. Hron, P. Filzmoser, and K. Thompson, “Linear regression with compositional explanatory variables,” *Journal of Applied Statistics*, vol. 39, pp. 1115–1128, 2012.

[72] E. Fišerová and K. Hron, “On interpretation of orthonormal coordinates for compositional data,” *Mathematical Geosciences*, vol. 43, no. 4, pp. 455–468, 2011.

[73] J. A. Cornell, *Experiments with mixtures: designs, models, and the analysis of the mixture data.* John Wiley & Sons, New York, 2002.

[74] H. Scheffé, “Experiments with mixtures,” *Journal of the Royal Statistical Society*, vol. Series B 20, pp. 344–360, 1958.

[75] H. Scheffé, “The simplex-centroid design for experiments with mixtures,” *Journal of the Royal Statistical Society*, vol. Series B 25, pp. 235–263, 1963.

[76] E. Fišerová, K. Kubáček, and P. Kunderová, *Linear statistical models - regularity and singularities.* Academia, Praha, 2007.

[77] B. Worley and R. Powers, “Multivariate analysis in metabolomics,” *Current Metabolomics*, vol. 1, pp. 92–107, 2013.

[78] A. Roux, D. Lison, C. Junot, and J. Heilier, “Applications of liquid chromatography coupled to mass spectrometry - based metabolomics in clinical chemistry and toxicology: A review,” *Clinical Biochemistry*, vol. 44, pp. 119–135, 2011.

[79] R. Wu, X. Zhao, Z. Wang, M. Zhou, and Q. Chen, “Novel molecular events in oral carcinogenesis via integrative approaches,” *Journal of Dental Research*, vol. 90, pp. 561–572, 2011.

[80] G. Patti, O. Yanes, and G. Siuzdak, “Metabolomics: the apogee of the omics trilogy,” *Nature Reviews Molecular Cell Biology*, vol. 13, pp. 263–269, 2012.

[81] M. Lämmerhofer and W. Wolfram, *Metabolomics in practice.* Wiley-VCH, Weinheim, 2013.

[82] P. Huber, *Robust statistics.* John Wiley, New York, 1981.

[83] T. Aittokallio, “Dealing with missing values in large-scale studies: microarray data imputation and beyond,” *Briefings in Biostatistics*, vol. 2, no. 2, pp. 253–264, 2009.

[84] J. Martín-Fernández, C. Barceló-Vidal, and V. Pawlowsky-Glahn, “Dealing with zeros and missing values in compositional data sets using nonparametric imputation,” *Mathematical Geology*, vol. 35, pp. 253–278, 2003.

[85] V. Pawlowsky-Glahn and J. Egozcue, “BLU estimators and compositional data,” *Mathematical Geology*, vol. 34, no. 3, pp. 259–274, 2002.

[86] O. Hrydziuszko and M. Viant, “Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline,” *Metabolomics*, vol. 8, pp. 161–174, 2012.

[87] R. Steuer, K. Morgenthal, W. Weckwerth, and J. Selbig, *Metabolomics: Methods and protocols*, ch. A gentle guide to the analysis of metabolomic data, pp. 105–129. New Jersey: Humana Press., 2007.

[88] J. Xia, N. Psychogios, N. Young, and D. S. Wishart, “Metaboanalyst: A web server for metabolomics data analysis and interpretation,” *Nucleic Acids Research*, vol. 37, pp. 652–660, 2009.

[89] G. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," *Numerische Mathematik*, vol. 14, no. 5, pp. 403–420, 1970.

[90] R. Wehrens, *Chemometrics with R.* Springer, Heidelberg, 2011.

[91] P. Kynčlová, P. Filzmoser, and K. Hron, "Compositional biplots including external non-compositional variables," *Statistics*, vol. to appear, p. xx, 2016.

[92] R. Rosipal and N. Krämer, *Overview and recent advances in partial least squares.* SLSFS, Springer, 2006.

[93] H. Abdi, *Partial least square regression.* Encyclopedia of measurement and statistics, cognition and neurosciences, 2007.

[94] S. Wold, M. Sjöströma, and E. L., "PLS-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, 2001.

[95] S. Van Huffel, *Partial least square regression. Recent advances in total least squares techniques and errors-in-variables modeling.* SIAM, Philadelphia, 1997.

[96] B. Mevik and R. Wehrens, "The pls package: principal component and partial least squares regression in R," *Journal of Statistical Software*, vol. 18, no. 2, 2007.

[97] D. M. Haaland and E. V. Thomas, "Partial least squares methods for spectral analyses 1," *Anal. Chem.*, vol. 60, pp. 1193–1202, 1988.

[98] S. de Jong and A. Phatak, *Recent advances in total least squares techniques and errors-in-variables modeling.* Siam, Leuven, 1996.

[99] I. G. Chong and C. H. Jun, "Performance of some variable selection methods when multicollinearity is present," *Chemometrics and Intelligent Laboratory Systems*, vol. 78, pp. 103–112, 2005.

[100] T. N. Tran, N. L. Afanador, L. M. C. Buydens, and L. Blanchet, "Interpretation of variable importance in partial least squares with significance multivariate correlation (smc)," *Chemometrics and Intelligent Laboratory Systems*, vol. 138, pp. 153 – 160, 2014.

[101] M. Gallo, "Discriminant partial least squares analysis on compositional data," *Statistical Modelling*, vol. 10, pp. 41–56, 2010.

[102] H. Martens and M. Martens, "Modified jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR)," *Food Quality and Preference*, vol. 11, pp. 5–16, 2000.

[103] M. Rubingh, S. Bijlsma, E. Derks, I. Bobeldijk, E. Verheij, S. Kochhar, and A. Smilde, “Assessing the performance of statistical validation tools for megavariate metabolomics data,” *Metabolomics*, vol. 2, no. 2, pp. 53–61, 2006.

[104] P. Bastien, V. Vinzi, and M. Tenenhausc, “PLS generalised linear regression,” *Computational Statistics & Data Analysis*, vol. 48, pp. 17–46, 2005.

[105] C. Wu, “Jackknife, bootstrap and other resampling methods in regression analysis,” *Annals of Statistics*, vol. 14, no. 4, pp. 1261–1295, 1986.

[106] A. Krishnan, J. Williams, A. McIntoshc, and H. Abdi, “Partial least squares (PLS) methods for neuroimaging: A tutorial and review,” *NeuroImage*, vol. 56, pp. 455–475, 2011.

[107] B.-H. Mevik and H. Cederkvist, “Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR),” *Journal of Chemometrics*, vol. 18, no. 9, pp. 422–429, 2004.

[108] P. Filzmoser, B. Liebmann, and K. Varmuza, “Repeated double cross validation,” *Journal of Chemometrics*, vol. 23, no. 4, pp. 160–171, 2009.

[109] M. Lindner, G. Hoffmann, and D. Matern, “Newborn screening for disorders of fatty-acid oxidation: experience and recommendations from an expert meeting,” *Journal of Inherited Metabolic Disease*, vol. 33, pp. 521–526, 2010.

[110] N. Gregersen, B. Andresen, C. Pedersen, R. Olsen, T. Corydon, and P. Bross, “Mitochondrial fatty acid oxidation defects-remaining challenges,” *Journal of Inherited Metabolic Disease*, vol. 31, pp. 643–657, 2008.

[111] E. Maier, J. Pongratz, A. Muntau, B. Liebl, U. Nennstiel-Ratzel, U. Busch, R. Fingerhut, B. Olgemöller, A. Roscher, and W. Röschinger, “Validation of MCADD newborn screening,” *Clinical Genetics*, vol. 76, pp. 179–187, 2009.

[112] C. Stewart and C. Field, “Managing the essential zeros in quantitative fatty acid signature analysis,” *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 16, no. 1, pp. 45–69, 2011.

[113] J. Palarea-Albaladejo, J. Martín-Fernández, and R. Olea, “A bootstrap estimation scheme for chemical compositional data with nondetects,” *Journal of Chemometrics*, vol. 28, no. 7, pp. 585–599, 2014.

[114] R. Bro, “PARAFAC. Tutorial and applications,” *Chemometrics and Intelligent Laboratory Systems*, vol. 38, pp. 149–171, 1997.

[115] R. Bro, *Multi-way analysis in the food industry - models, algorithms and applications*. PhD thesis, Universiteit van Amsterdam, The Netherlands, 1998.

[116] T. Kolda and B. W. Bader, “Tensor decompositions and applications,” *Siam Review*, vol. 51, no. 3, pp. 455–500, 2009.

[117] A. L. Kiers, “Towards a standardized notation and terminology in multiway analysis,” *J. Chemometr.*, vol. 14, pp. 105–122, 2000.

[118] R. Bro and A. Smilde, “Centering and scaling in component analysis,” *Journal of Chemometrics*, vol. 17, no. 1, pp. 16–33, 2003.

[119] C. Andersson, L. Munck, R. Henrion, and G. Henrion, “Analysis of N-dimensional data arrays from fluorescence spectroscopy of an intermediary sugar product,” *Fresenius Journal of Analytical Chemistry*, vol. 359, pp. 138–142, 1997.

[120] P. Paatero and S. Juntto, “Determination of underlying components of a cyclical time series by means of two-way and three-way factor analytic techniques,” *Journal of Chemometrics*, vol. 14, pp. 241–259, 2000.

[121] V. Pravdova, C. Boucon, S. de Jong, B. Walczak, and D. Massart, “Three-way principal component analysis applied to food analysis: an example,” *Analytica Chimica Acta*, vol. 462, pp. 133–148, 2002.

[122] A. Smilde, R. Bro, and P. Geladi, *Multi-way analysis with applications in the chemical sciences*. John Wiley & Sons, Chichester, UK, 2004.

[123] R. Harshman and M. Lundy, “PARAFAC: Parallel factor analysis,” *Computational Statistics & Data Analysis*, vol. 18, pp. 39–72, 1994.

[124] M. Bosco, M. Garrido, and M. Larrechi, “Determination of phenol in the presence of its principal degradation products in water during a TiO2-photocatalytic degradation process by three-dimensional excitation emission matrix fluorescence and parallel factor analysis,” *Analytica Chimica Acta*, vol. 559, pp. 240–247, 2006.

[125] M. Hubert, J. Van Kerckhoven, and T. Verdonck, “Robust PARAFAC for incomplete data,” *Journal of Chemometrics*, vol. 26, no. 6, pp. 290–298, 2012.

[126] J. Kruskal, “Three-way arrays: Rank and uniqueness of trilinear decomposition, with application to arithmetic complexity and statistics,” *Linear Algebra and its Applications*, vol. 18, pp. 95–138, 1977.

[127] B. Carter, A. Haverkamp, and G. B. Merenstein, “The definition of acurate perinatal asphyxia,” *Clinics in perinatology*, vol. 20, no. 2, pp. 287–304, 1993.

[128] A. Weintraub, A. Carey, J. Connors, V. Blanco, and R. Green, “Relationship of maternal creatinine to first neonatal creatinine in infants $<$ 30 weeks gestation,” *Journal of Perinatology*, vol. 15, pp. 401–404, 2015.

[129] C. Westermann, L. Dorland, D. Votion, M. de Sain-van der Velden, I. Wijnberg, R. Wanders, W. Spliet, N. Testerink, R. Berger, J. Ruiter, and J. van der Kolk, “Acquired multiple Acyl-CoA dehydrogenase deficiency in 10 horses with atypical myopathy,” *Neuromuscular Disorders*, vol. 18, pp. 355–364, 2008.

[130] J. Fox and S. Weisberg, *An R companion to applied regression.* SAGE Publications, California, 2011.

[131] F. Dieterle, A. Ross, G. Schlotterbeck, and H. Senn, “Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. application in 1H NMR metabonomics,” *Analytical Chemistry*, vol. 78, pp. 4281–4290, 2006.

[132] P. Filzmoser and B. Walczak, “What can go wrong at the data normalization step for identification of biomarkers?,” *Journal of Chromatography A*, vol. 1362, pp. 194–205, 2014.

PALACKÝ UNIVERSITY OLOMOUC
FACULTY OF SCIENCE

DISSERTATION THESIS SUMMARY

Multidimensional statistical methods for analysis of human metabolome

Department of Mathematical Analysis and Applications of Mathematics
Supervisor: **Doc. RNDr. Karel Hron, Ph.D.**
Author: **Mgr. Alžběta Gardlo (Kalivodová)**
Consultant: **prof. RNDr. Tomáš Adam, Ph.D.**
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 in the Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University Olomouc.

Applicant: **Mgr. Alžběta Gardlo**
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

Consultant: **prof. RNDr. Tomáš Adam, Ph.D.**
Department of Clinical Biochemistry
University Hospital Olomouc
Institute of Molecular and Translational Medicine
Faculty of Medicine and Dentistry
Palacký University Olomouc

Reviewers: **Doc. PaedDr. RNDr. Stanislav Katina, Ph.D.**
Institute of Mathematics and Statistics
Faculty of Science
Masaryk University Brno

Dr. Javier Palarea-Albaladejo
Biomathematics and Statistics Scotland
Edinburgh, United Kingdom

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

List of abbreviations

Throughout the thesis summary, the following standard abbreviations are used. Other nonstandard abbreviations are introduced in the text as it is needed.

ADCS	average difference in covariance structure
alr coordinates	additive logratio coordinates
ALS	alternating least squares
AUC	area under the curve
CED	compositional error deviation
clr coorinates	centered logratio coordinates
CV	coefficient of variation
ilr coordinates	isometric logratio coordinates
LOESS	local regression
MCADD	medium chain acyl-CoA dehydrogenase deficiency
MSEP	mean squared error of prediction
m/z	mass to change ratio
NIPALS	nonliear iterative partial least squares
OPLS-DA	orthogonal partial least squares regression - discriminant analysis
PARAFAC	parallel factor analysis
PCA	principal component analysis
PLS-DA	partial least squares regression - discriminant analysis
PQN	probabilistic quotient normalization
PRESS	predicted error sum of squares
QC	quality control
SVD	singular value decomposition
VIP	variable importance in the projection

1. Abstract

The metabolomics is a quite new field of biochemistry which aims at studying metabolites, their dynamic changes, interactions and responses to stimuli. Because of relative character of metabolomic data, they can be considered as so called compositional data. They are characterized by positive entries, moreover, not their absolute values but ratios between them are of primary interest. In order to analyze statistically compositional data in standard Euclidean space, specific coordinate systems must be used. Furthermore, for the analysis of metabolomic data also the biochemical material must be considered, and finally, also the fact that substantially less observations than variables are available; we refer to so called high-dimensional compositional data. For statistical analysis of such data set, special statistical procedures must be applied. Prior to the statistical analysis itself, preprocessing of compositional data must be carried out, needed for further representation of logratio coordinates (quality control, zero values of compositional parts). So the statistical analysis itself can be performed using a wide range of proper methods. The most popular one is principal component analysis that can be accompanied by partial least squares method and its orthogonal modification. For the analysis of three-way metabolomic data, PARAFAC is recently preferred choice in chemometrics. Methodological outputs are demonstrated on real data from the Laboratory of Metabolomics, Palacký University Olomouc.

Key words: Compositional data, metabolomics, partial least squares regression, multivariate statistical analysis, practical application, imputation of zeros.

2. Abstrakt v českém jazyce

Metabolomika je poměrně novým oborem biochemie zabývající se studiem metabolitů, jejich dynamickými změnami, interakcemi a odpověďmi na podněty. Vzhledem k relativnímu charakteru metabolomických dat na ně může být pohlíženo jako na tzv. kompoziční data. Vektory takovýchto dat mají kladné složky; navíc nás nezajímají jejich absolutní hodnoty, ale podíly mezi nimi. Abychom mohli pracovat s kompozičními daty v klasickém euklidovském prostoru, musíme použít specifické souřadnicové systémy. Dále musíme při analýze metabolomických dat brát v úvahu materiál, který je použit pro měření, a v neposlední řadě i to, že máme k dispozici typicky řádově méně pozorování než proměnných, hovoříme o tzv. vysoce-dimenzionálních datech. Pro analýzu takového souboru musí být použity speciální statistické metody. První částí statistické analýzy je předzpracování dat, související s vyjádřením metabolomických (kompozičních) dat v tzv. logratio souřadnicích. V metabolomice také používáme tzv. kontroly kvality, které nám pomáhají v odstraňování chyb měření. Dalším problémem jsou nulové hodnoty. Většina v současnosti používaných statistických metod pro kompoziční data neumí pracovat s nulovými hodnotami, proto je musíme umět vhodně nahradit. Vlastní statistická analýza může být provedena pomocí celé řady postupů. První, nejpopulárnější, je metoda hlavních komponent. Ta je východiskem pro metodu částečných nejmenších čtverců či její ortogonální podobu. Pokud pracujeme s trojrozměrnými datovými tabulkami, můžeme analýzu provést také pomocí metody PARAFAC. Důležitou součástí disertační práce jsou také praktické příklady na reálných datech z Laboratoře metabolomiky Univerzity Palackého Olomouc.

Klíčová slova Kompoziční data, metabolomika, metoda částečných nejmenších čtverců, mnohorozměrná statistická analýza, praktická aplikace, nahrazování nul.

3. Introduction

Compositional data (or compositions for short) are multivariate observations with positive components, and they can be represented without loss of information as data with a constant sum constraint like proportions or percentages [1, 2]. In such a case, the sum of the compounds (parts) is not important and the only relevant information is contained in ratios between the parts.

Metabolomics aims at studying metabolites, their dynamic changes, interactions and responses to stimuli. It is applied to the metabolism of plants, bacteria, animals and humans. In humans all biological materials from biofluids (blood, urine) till tissues are analyzed. Although absolute values of biomarkers compared with reference ranges (data from the healthy population) is the most frequently used approach, ratios of metabolite data are frequently analyzed in the biochemical diagnostic practice and relative changes are more relevant/informative than absolute values. It suggests that metabolomic data can indeed be considered as observations carrying relative information, i.e. as compositional data [3].

Very important part of the statistical evaluation is the preprocessing of metabolomic data. The measuring instruments have some limitations and measuring errors can be present in data. To correct these errors, special statistical methods must be used. The signal correction by the LOESS method based on the quality control samples must be done before statistical processing itself is performed [4, 5].

It is widely common in chemometrics, and particularly in metabolomics, to normalize and scale observations prior to further statistical analysis [6,7]. While most of the normalization techniques are heuristic ones, it is also possible to derive systematic approaches based on natural features of the underlying observations. The relative character of metabolite observations is reflected in practice by many kinds of normalization techniques [6]. Let us mention, e.g., the well-known AUC normalization whose aim is to normalize a group of signals with peaks by standardizing the area under the curve (AUC) to the group median, mean or any other proper representation. Another approach is represented by rationing to landmarks, e.g. to normalization of urine end-product metabolites to creatinine, that is often used also in general in chemometrics. The choice of any such normalization is usually strongly data dependent in practice, which affects the objectivity and makes any further comparisons hardly attainable [8–10]. After the normalization step, data are popularly transformed using the log-transformation (popular in metabolomics), or alternatively (and preferably here) expressed as proper logratio coordinates that capture relative nature of metabolomic (compositional) data.

The statistical analysis of two-way data starts typically with principal component analysis [11, 12]. This method must be adapted to work with compositional data, i.e. a special coordinate system must be used for the analysis [2]. Concerning further statistical analysis, the problem occurs because more metabolites (in hundreds) than biological materials (only tens) are present in these data sets. Therefore, suitable methods must be applied for this kind of observations. One of them is partial least squares regression (PLS regression), concretely its popular special case partial least squares - discriminant analysis (PLS-DA) [13]. The standard PLS-DA method needs to be adapted to compositional data, because using raw observations could lead to useless results. A special modification of the PLS model can also be used. It is called orthogonal - partial least squares (OPLS) method and it works with the orthogonal variation in the data [14]. Results of OPLS regression are popularly visualized using S-plot which is a scatter plot of correlations and covariance of the data. Special techniques of metabolomic (compositional) data processing need to be applied, when they form a three-way structure. This structure arises typically when samples of some biological material are measured at more time points. For statistical processing of three-way observations the PARAFAC model [15] represents one of popular tools in chemometrics.

The final part of the thesis is the practical application of all presented methods. Because of the limitation of this summary, all examples are skipped (for further information see the original dissertation thesis). Examples from the theoretical chapters are also omitted.

The whole procedure of complete statistical evaluation of metabolomic data as presented in the thesis is used in everyday practice in the Laboratory of Metabolomics, Institute of Molecular and Translational Medicine, Palacký University Olomouc. Some parts could be rather elementary for mathematical audience, but they are very useful for people from the outside of the statistical field.

4. Recent state summary

4.1. Compositional data

Compositional data occur in a wide range of applications involving geochemistry, analytical chemistry, and its related fields. Nevertheless, up to now just a few papers following the concept of compositional data were published in the field of metabolomics and proteomics [3, 16–18]. These data are characterized by features like scale invariance (the information in a composition does not depend on the particular units in which the composition is expressed) and the relative scale (ratios and not absolute distances are important when dissimilarities of observations are analyzed). Another property which is crucial for any meaningful statistical analysis of compositional data is called subcompositional incoherence, i.e. information conveyed by a composition should not be in contradiction with that coming from a subcomposition that involves only a subset of variables [19].

The sample space of D-part composition $\mathbf{x} = (x_1, x_2, \ldots, x_D)'$ is called the *simplex* defined as [1, 2]

$$\mathbf{S}^D = \left\{ \mathbf{x} = (x_1, x_2, \ldots, x_D)' \middle| x_i > 0, i = 1, 2, \ldots, D; \sum_{i=1}^{D} x_i = \kappa \right\}. \tag{1}$$

The natural geometry of compositions is called the Aitchison geometry and it has all usual properties that are known from the Euclidean geometry, for which standard statistical methods are designed [20]. However, operations of the Aitchison geometry are different from the Euclidean geometry case. For this reason, usual multivariate statistical methods cannot be directly applied to compositional data, since otherwise interpretations of the results and conclusions can be misleading [2].

Statistical data analysis is usually carried out in the Euclidean geometry and not in the Aitchison geometry. Thus, the central idea is to express compositions from the simplex in real coordinates and then to apply the standard multivariate methods. From a mathematical point of view, we search for a basis (or generating system) with respect to the Aitchison geometry in order to express compositional data in coefficients of such a basis (coordinate system). As these coefficients are build up using logarithms of ratios of compositional parts, we refer to logratio coordinates. Currently, three basic logratio coordinate systems occur in the literature: additive, centered and isometric logratio coordinates. Nevertheless, only the latter two can be recommended in general, because they map the Aitchison geometry to the Euclidean one isometrically. The use of logratio coordinates preserves the relative scale property of compositions, which is of primary importance in chemometrics, and follow all requirements for a meaningful analysis of compositions as mentioned above. For more detailed discussions on these issues, see [1, 21].

The *centered logratio (clr) coordinates* [1,2] are defined for a composition $\mathbf{x} = (x_1, \ldots, x_D)'$ as

$$clr(\mathbf{x}) = \mathbf{r} = (r_1, \ldots, r_D)' = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right)'. \qquad (2)$$

Although the resulting variables are quite easily interpretable (each of them corresponds to one of the original compositional parts), clr coordinates are coefficients with respect to a generating system on the simplex. For this reason, the resulting covariance matrix of a random composition in clr coordinates is singular [1, 22]. The singularity restriction of the clr coordinates is overcome by the *isometric logratio (ilr) coordinates*, resulting in $D - 1$ coordinates with respect to an orthonormal basis. Unfortunately, it is thus not possible to assign a coordinate to each of the original compositional parts simultaneously, as it was the case for clr coordinates. Nevertheless, as there are infinitely many ways to construct an orthonormal basis, its proper choice [23] allows to construct coordinates with an intuitive interpretation. Thus we get a $(D - 1)$-dimensional real vector

$ilr(\mathbf{x}) = \mathbf{z} = (z_1, \ldots, z_{D-1})'$ [2,22], where

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

The inverse mapping of $\mathbf{z}$ back to the original composition $\mathbf{x}$ is then given by

$$\begin{aligned} x_1 &= \exp\left(\frac{\sqrt{D-1}}{\sqrt{D}} z_1\right), \\ x_i &= \exp\left(-\sum_{j=1}^{i-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} z_j + \frac{\sqrt{D-i}}{\sqrt{D-i+1}} z_i\right), \\ x_D &= \exp\left(-\sum_{j=1}^{D-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} z_j\right). \end{aligned} \tag{4}$$

With the orthonormal (ilr) coordinates (3), the variable z_1 carries all the relevant information about the compositional part x_1, because it explains all the ratios between x_1 and the other parts of $\mathbf{x}$ [24].

Now we can proceed to construct such an orthonormal basis, where the first ilr coordinate explains the relative information about a compositional part of interest. For this purpose, the indices in formula (3) are just permuted such that the part of interest plays the role of x_1. Accordingly, in order to assign such coordinates to each compositional part x_l, $l = 1, \ldots, D$, we need to construct D different ilr coordinate systems, where the D-tuple $(x_1, \ldots, x_D)'$ in (3) is replaced by $(x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)' =: (x_1^{(l)}, x_2^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})'$ [24]. The corresponding ilr coordinates are thus

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}}, \quad i = 1, \ldots, D-1. \tag{5}$$

Obviously, the vector $\mathbf{z}^{(l)} = (z_1^{(l)}, \ldots, z_{D-1}^{(l)})'$ is again a vector of orthonormal coordinates. The relation between the clr coefficients and the ilr coordinates is linear $\mathbf{r} = \mathbf{V}\mathbf{z}$. The matrix $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_{D-1})$ has dimension $D \times (D-1)$ and its columns are formed by the orthonormal basis vectors in clr coordinates,

$$\mathbf{v}_i = \sqrt{\frac{D-i}{D-i+1}}\left(0,\ldots,0,1,-\frac{1}{D-i},\ldots,-\frac{1}{D-i}\right)', \quad i=1,\ldots,D-1. \tag{6}$$

The third basic logratio coordinate system is called the *additive logratio (alr) coordinates.* This system is not very often used because results of statistical processing in alr coordinates might depend on the denominator used in the formula and they represents coordinates with respect to a basis that is not orthonormal. As a consequence, alr coordinates do not form an isometric mapping [2]. Though, as we can see in the following text, these coordinates can naturally occur as a result of combining transformations and normalizations used in metabolomics.

The definition of the alr coordinates for a composition $\mathbf{x} = (x_1, \ldots, x_D)'$ is as follows [1,2]:

$$alr(\mathbf{x}) = \mathbf{w} = (w_1, \ldots, w_{D-1})' = \left(\ln\frac{x_1}{x_D}, \ln\frac{x_2}{x_D}, \ldots, \ln\frac{x_{D-1}}{x_D}\right)'. \tag{7}$$

The resulting coordinates are not symmetric [2], because the denominator used in (7), x_D, can be replaced by any other compositional part. As a consequence, alr coordinates are not invariant under permutation of components, that forms the final principle of compositional data analysis [19]. The way out is to use clr or ilr coordinates instead of alr [2].

Very important part of the logratio methodology is the regression analysis. Its aim is to explain the response (real) variable Y by using explanatory variables $x_1, \ldots, x_D$. Regression with compositional explanatory variables can be carried out by first applying ilr coordinates to covariate composition. For a regression model between Y and $\mathbf{x}$ (composition) we use coordinates $\mathbf{z}$ by applying formula (3). The standard multiple linear regression of Y on the explanatory variables $\mathbf{z} = (z_1, \ldots, z_{D-1})'$ is thus obtained,

$$\mathrm{E}(Y|\mathbf{z}) = \gamma_0 + \gamma_1 z_1 + \ldots + \gamma_{D-1} z_{D-1}. \tag{8}$$

As in formula (5), we can consider the lth ilr basis, for $l = 1, \ldots, D$, resulting in a regression model

$$\mathrm{E}(Y|\mathbf{z}) = \gamma_0 + \gamma_1^{(l)} z_1^{(l)} + \ldots + \gamma_{D-1}^{(l)} z_{D-1}^{(l)}. \tag{9}$$

Since $z_1^{(l)}$ explains all the relative information about part $x_1^{(l)}$, also the interpretation of the coefficient $\gamma_1^{(l)}$ can be associated to this part. The interpretation of the other regression coefficients (except γ_0) is not straightforward, because the corresponding explanatory variables (coordinates) do not fully represent one particular part of the composition. Consequently, a possible way to evaluate the contribution of each compositional part for explaining the response Y separately is to consider D regression models according to (9) by taking $l \in \{1, \ldots, D\}$, and to interpret the coefficients $\gamma_1^{(l)}$, representing the relative information on parts $x_1^{(l)}$ [24].

4.2. Metabolomics

Metabolomics is a quite new field of biochemistry and it aims at studying metabolites, their dynamic changes, interactions and responses to stimuli. It is possible to measure thousands of metabolites simultaneously from only minimal amounts of sample in presence [25]. This possibility allows defining different attitudes to the analysis of metabolomic samples. The classical division is done by targeted and untargeted approaches. In the *targeted* analysis the list of metabolites, which are measured, is done before the analysis. The *untargeted* metabolomic methods are global in scope and have the aim of simultaneously measuring as many metabolites as possible from biological samples without bias [25]. These metabolites (here called peaks) are not known before the experiment [26] and must be identified after the analysis.

Data analysis in metabolomics is a very specific process. It is closely related to the material, which is measured and processed (cells, blood, urine, plasma, ...). Accordingly, a specific approach is used, e.g., also for the analysis of urine samples as urine volume can vary widely based on upon water consumption and other physiological factors. Consequently, the concentrations of metabolites in urine vary substantially and proper normalizing for these effects is necessary [8]. Two methods of data normalization are used in practice - creatinine normalization and normalization by the area under the curve [8–10]. The first method of normalization is related to the very specific metabolite, called creatinine, that is presented in all urine samples. Creatinine is a chemical waste product in blood that passes through kidneys to be filtered and eliminated in urine. Under normal conditions, urinary creatinine output is relatively constant and measurable. As a result, it has become common practice to normalize urinary analyte levels to this metabolite. Moreover, in practice, the level of creatinine is different in various samples, thus, each sample is divided by a different scaling constant. The second normalization is performed through the area under the curve (AUC) of all peaks, identified with metabolite concentrations in the analysis. Each mass spectrum (metabolite profile) is thus divided by average variable area across observations [10].

After the normalization step, data are popularly transformed using the log-transformations (popular in metabolomics), or alternatively taking the proposed logratio coordinates. It is important to note that although the popular log-transformation of the input data removes the relative (measurement) scale effects, the scale invariance of compositions is destroyed.

4.3. The use of quality control samples

A very important part of the statistical evaluation is formed by preprocessing of metabolomic data. The first reason for doing this procedure is the fact that measuring instruments have some limitations and measurement errors can be present in the data. Special statistical methods must be used to correct these errors. The quality control (QC) samples are used for this purpose. QC samples are mixtures of all samples from the specific analysis. They are measured continuously in the whole analysis on the first ten positions and then as every fourth sample in a way. It is known that signal of these QC samples must be stable in time; if there is some trend, it must be revised. The signal correction by LOESS (LOcal regrESSion) method is used for this purpose [4, 5]. The LOESS curve is fitted to the QC samples with respect to the order of injection. A correction curve for the whole analytical run is then interpolated, to which the total data set for that feature is normalized [5].

4.4. Imputation of missing values and rounded zeros

The second step of data preprocessing is connected with imputation, because almost none of statistical methods is able to process data that contain measurement artifacts like missing values (pure absence of the measurement in some entries) or values below a detection limit (resulting as the effect of rounding errors or imprecision of the measuring device, we also refer to rounded zeros). Especially, values below detection limit occur frequently in chemometric data. Their proper replacement must precede any further statistical analysis. Although for the case of standard multivariate data a comprehensive methodology exists [27], even applicable to high-dimensional data [28], it fails in case of compositional data.

4.5. Multidimensional statistical analysis

The final step of statistical evaluation of data in metabolomics is the multidimensional analysis itself. One of the basic methods used in multivariate data analysis (especially for visualization of the data structure) is *principal component analysis* (PCA). The aim of this method is to reduce the dimensionality of data by preserving the most information identified with variability contained in the data set. Its main principle is to construct an orthogonal coordinate system, which is formed by latent variables, so that only the first variables explain most of variability in data. The goal of PCA is also the reduction of the effect of measurement error and elimination of components associated with the noise [12].

Very common problem of metabolomic data sets is their high-dimensionality (= the presence of more variables than observations). Therefore, suitable methods must be used for their analysis. One of them is *partial least squares* (PLS) regression, which is a class of methods for modeling relations between sets of explanatory and response variables by means of latent variables [29,30]. PLS can be used for both regression and classification purposes, and it can be employed also for reducing the dimensionality of the data. The intrinsic assumption of all PLS methods is that the observed data are generated by a system or process which is guided by a small number of latent (not directly observed or measured) variables [29]. Partial least squares - discriminant analysis (PLS-DA) is a special type of a regression analysis where the response variables represent group labels.

Finally, the statistical processing of three-way observations is done with *PARAllel FACtor analysis* (PARAFAC), but it is still rarely used in the compositional context [31–33], with no metabolomics application known so far.

5. Thesis objectives

This thesis aims to be a complex guide for the statistical processing of metabolomic (compositional) data sets. The main goal is to explain the possibility of using advanced multivariate statistical methods for the statistical analysis of metabolomic data by using the logratio methodology. Methods like principal component analysis, partial least squared regression or parallel factor analysis (PARAFAC) are adopted for the case of compositional data. The only limitation of the logratio methodology is more complex interpretation of results in logratio coordinates. An important part of the thesis is formed by applications of the logratio methodology to a wide range of data sets from metabolomics (which are skipped in this summary).

6. Theoretical framework

Four multidimensional statistical methods are presented for the analysis of metabolomic data sets - principal component analysis, partial least squares regression - discriminant analysis, its orthogonal extension and PARAFAC model. Finally also the parametric model for imputation of rounded zeros is introduced. Except of principal component analysis, adaptations (developments) of the other methods can be considered as new contributions to statistical processing of metabolomic (compositional) data.

6.1. Principal component analysis

As mentioned in Section 4.5., one of basic methods used in multivariate data analysis is principal component analysis (PCA) which is used for the reduction of dimensionality of data by preserving the most information identified with variability contained in the data set. The standard approach to PCA is commonly known, in the logratio approach, the main difference is the necessity of expressing the input data matrix in centered clr coordinates (2). Ilr coordinates (3) can also be used, but their specific interpretation needs to be taken into account.

The direction of the highest variability in data is captured by the first principal component (PC1), the second principal component (PC2) is formed by an orthonormal direction to PC1 and again acquires the maximum possible variability. Following principal components are orthogonal to all previous components and their directions have to cover the maximum possible variance of the data projected on this direction [11]. In the standard analysis, usually only first two (or maximum three) principal components are considered for practical reasons with the hope that they contain most of the total variance in the data set. Nevertheless, in general, the number of principal components is limited only by the number of variables.

A graphical representation of PCA is called *biplot*. It is a planar graph used for the projection of so called scores (coordinates of principal components) and loadings (the corresponding basis vectors) of the first two principal components into one plot. Scores, which represent the structure of the compositional data set in Euclidean space, are displayed as points and they can be used to visualise grouping in the data. Loadings, which represent the corresponding clr variables, are displayed by arrows (rays) in the same plot [2]. The interpretation of the compositional biplot (in clr coordinates) is slightly different from the interpretation of the standard one and it is presented in detail in the thesis.

6.2. Partial least squares - discriminant analysis

Partial least squares - discriminant analysis (PLS-DA) is a popular classification tool in metabolomics. For the case of compositional data, instead of taking clr coordinates that impose the additional constant sum constraint, ilr coordinates (3) are employed [2]. Let $\mathbf{X}$ be an $n \times D$ matrix of compositional data (sampled compositional parts $x_1, \ldots, x_D$) and $\mathbf{Y}$ be an $n \times q$ matrix of responses representing the groups. As in the standard case, the columns of $\mathbf{Y}$ are mean-centered. However, mean-centering of the compositions $\mathbf{X}$ is done with respect to the Aitchison geometry, i.e. the centering is performed in the ilr coordinates.

Following the case of linear regression with compositional explanatory variables (see Section 4.1, [24]), where applying the clr coordinates leads to a biased estimation of the regression coefficients due to the singular covariance matrix of the clr variables, the ilr coordinates may be used for the purpose of PLS modeling. Subsequently, the matrix $\mathbf{X}$ is firstly expressed in ilr coordinates $\mathbf{Z}$ (3). The PLS regression problem has the form [3]

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{E}, \tag{10}$$

where $\boldsymbol{\Gamma}$ stands for a $(D-1) \times q$ matrix of regression coefficients.

Nevertheless, the ilr coordinates (3) allows only for a meaningful interpretation of the elements in the first row of $\boldsymbol{\Gamma}$, because just the first column of $\mathbf{Z}$ can be associated with one particular compositional part (here x_1). The interpretation of the other regression coefficients is not straightforward because the corresponding explanatory variables (coordinates) do not fully represent one particular part of the composition. For associations also to the other parts, we thus need to use a permutation of the parts, leading to the general setting (5) and to data matrices $\mathbf{Z}^{(l)}$. Each of the resulting first ilr coordinates, the observations of $z_1^{(l)},\ l = 1, \ldots, D$, describes all the relative information about the compositional part x_l.

Consequently, a possible way to evaluate the contribution of each compositional part for explaining the response variables $\mathbf{Y}$ separately is to consider D PLS regression models

$$\mathbf{Y} = \mathbf{Z}^{(l)}\boldsymbol{\Gamma}^{(l)} + \mathbf{E}^{(l)}, \tag{11}$$

according to (5), by taking $l \in \{1, \ldots, D\}$, and to interpret the coefficients of the first row of the parameter matrix $\mathbf{\Gamma}^{(l)}$, representing the part $x_1^{(l)}$ [24]. The outlined procedure thus suggests employing PLS regression D times, such that each compositional part is once at the first position in the permuted composition. Since such a procedure would lead to a high computational complexity, the orthogonal relation between the different ilr coordinates can be employed [22]. As an advantage, the regression coefficients need to be estimated just for one regression model and then derived for the other models by using orthogonal transformations of the regression parameters. Note, however, that coefficients of $z_1^{(l)}$ should be always treated individually as they come from individual PLS models.

The final procedure is as follows. We use the matrix $\mathbf{V}$ from (6), with rows $\mathbf{v}_{i\cdot}, i = 1, \ldots, D$, that relates the clr coordinates (2) and the ilr coordinates (3). Consequently, we form $D \times (D-1)$ matrices $\mathbf{V}^{(l)}$, for $l \in \{1, \ldots, D\}$,

$$\begin{aligned}
\mathbf{V}^{(1)} &= (\mathbf{v}_{1\cdot}, \mathbf{v}_{2\cdot}, \ldots, \mathbf{v}_{D-1,\cdot}, \mathbf{v}_{D\cdot})' = \mathbf{V} \\
\mathbf{V}^{(l)} &= (\mathbf{v}_{l\cdot}, \mathbf{v}_{l-1,\cdot}, \ldots, \mathbf{v}_{1\cdot}, \mathbf{v}_{l+1,\cdot}, \ldots, \mathbf{v}_{D\cdot})', \ l = 2, \ldots, D-1; \\
\mathbf{V}^{(D)} &= (\mathbf{v}_{D\cdot}, \mathbf{v}_{D-1\cdot}, \ldots, \mathbf{v}_{2\cdot}, \mathbf{v}_{1\cdot})',
\end{aligned}$$

and define a new orthogonal matrix $\mathbf{Q}^{(l)}$,

$$\mathbf{Q}^{(l)} = \mathbf{V}'\mathbf{V}^{(l)}. \tag{12}$$

The matrices $\mathbf{Z}^{(l)}$ corresponding to ilr coordinates (5) are related to $\mathbf{Z}$ by

$$\mathbf{Z}^{(l)} = \mathbf{Z}\mathbf{Q}^{(l)}, \tag{13}$$

see [22]. Substituting (13) into the model (10) gives

$$\mathbf{Y} = \mathbf{Z}^{(l)}(\mathbf{Q}^{(l)})'\mathbf{\Gamma} + \mathbf{E}^{(l)} = \mathbf{Z}^{(l)}\mathbf{\Gamma}^{(l)} + \mathbf{E}^{(l)}. \tag{14}$$

Thus, the estimated regression coefficients $\mathbf{\Gamma}$ from the model (10), $\widehat{\mathbf{\Gamma}}$, can be used to estimate coefficients in regression models that correspond to coordinates $\mathbf{Z}^{(l)}$,

$$\widehat{\mathbf{\Gamma}}^{(l)} = (\mathbf{Q}^{(l)})'\widehat{\mathbf{\Gamma}}, \quad l = 1, \ldots, D. \tag{15}$$

Finally, to complete the estimation process with respect to the above interpretation, we collect the first rows of the matrices $\widehat{\mathbf{\Gamma}}^{(l)}$ as rows of a new $D \times q$ matrix of regression coefficients. Specially, for $q = 1$ (the response variable is univariate) we thus, get a vector $\mathbf{g} = (\widehat{\gamma}_1^{(1)}, \widehat{\gamma}_1^{(2)}, \ldots, \widehat{\gamma}_1^{(D)})'$.

A further evaluation of the resulting regression model can be done by testing for significance of the regression parameters. For PLS-DA, it is common to use resampling techniques for this purpose, like bootstrap [34]. The idea of the bootstrap procedure is to draw random samples with replacement from each group of the original data, where the bootstrap group samples have the same size as the original groups. This results in a bootstrap data set for the explanatory variables and the response, where PLS-DA is applied to estimate the parameters. Repeating this procedure many times allows estimating the variability of the regression parameters [34]. The standardized regression estimates are obtained by dividing the regression parameters of the original data by the estimated standard deviations (obtained from bootstrap), and they can be compared with quantiles of the standard normal distribution. For the case $q = 1$ and the estimated parameters $\mathbf{g} = (\widehat{\gamma}_1^{(1)}, \widehat{\gamma}_1^{(2)}, \ldots, \widehat{\gamma}_1^{(D)})'$, the estimates are recomputed using bootstrap, and from the results the standard deviations $s_1, \ldots, s_D$ are computed. The significance of the standardized regression estimates, $\widehat{\gamma}_1^{(1)}/s_1, \widehat{\gamma}_1^{(2)}/s_2, \ldots, \widehat{\gamma}_1^{(D)}/s_D$, is evaluated by comparing them with $\alpha/2$ and $1 - \alpha/2$ quantiles of the standard normal distribution (typically, $\alpha = 0.05$ is chosen). In order to reduce the risk of false positives, a Bonferroni correction is applied, resulting in an adjusted α-level of significance, $\alpha_{adj} = \frac{\alpha}{D}$. If the standardized regression coefficient is outside the mentioned interval, the regression coefficient is significantly different from zero, and thus, the corresponding variable contributes to the discrimination task.

The significance of the standardized regression coefficients is analyzed using PLS-DA with three components and bootstrap with 100 replications in example on real data set in the thesis.

The possible modification of PLS regression is orthogonal partial least squares -

discriminant analysis (OPLS). The idea of OPLS is to separate the systematic variation in data matrix $\mathbf{Z}^{(l)}$ (in ilr coordinates (5)) into two parts. The first one is linearly related to $\mathbf{Y}$ and represents between class variation, the second one is unrelated (orthogonal) to $\mathbf{Y}$ and refers to as the uncorrelated variation, which forms the within class variation [35]. The matrix $\mathbf{Y}$ is thus connected to the additional information provided by the matrix $\mathbf{Z}^{(l)}$. For classification purposes, OPLS is often called orthogonal partial least squares - discriminant analysis (OPLS-DA).

6.3. Parametric model for imputation of rounded zeros

Only few algorithms exist for the imputation of rounded zeros in high-dimensional compositional data sets. The crucial point for building up a reasonable imputation procedure is to find interpretable orthonormal coordinates in order to enable further processing in the standard Euclidean geometry. Since there is no canonical basis on the simplex, a set of orthonormal coordinate systems (5) needs to be employed sequentially in order to perform the imputation for each of the original compositional parts. The procedure needs to be able to capture both the relative information, conveyed by the compositional data themselves and the absolute nature of the corresponding detection limits, for a meaningful imputation of rounded zeros.

Based on previous considerations and following the structure of the imputation procedure in [36], an iterative regression-based algorithm for the replacement of rounded zeros is introduced in Algorithm 6.1. It is based on PLS regression, with the evaluation of the optimal number of components based on PRESS criterion announced in the thesis, thus able to cope also with high-dimensional compositional data sets [37]. In addition, some notation is given beforehand.

To avoid complicated notation in Algorithm 6.1, we assume that $\mathcal{M}(\mathbf{x}_1) \geq \mathcal{M}(\mathbf{x}_2) \geq \ldots \geq \mathcal{M}(\mathbf{x}_D)$, with $\mathcal{M}(\mathbf{x}_j)$ denoting the number of rounded zero cells in variable $\mathbf{x}_j$. Denote $m_l \subset \{1, \ldots, n\}$ the indices of the rounded zeros in variable $\mathbf{x}_l$, and $o_l = \{1, \ldots, n\} \backslash m_l$ the indices corresponding to the remaining cells of $\mathbf{x}_l$. Denote $\mathbf{z}_1^{(l)}$ as the first coordinate according to (5), and $\mathbf{Z}_{-1}^{(l)}$ containing the remaining $D-2$ coordinates. The first co-

Algorithm 6.1 PLS

```
for $j \in \{1, ..., D\}$ do                                    ▷ INITIALIZATION OF ROUNDED ZEROS
    Initialize all $\mathbf{x}_{ij}$, $i \in m_j$ with 2/3 of the corresponding detection limit.
end for
Sort variables based on $\mathcal{M}(\mathbf{x}_1) \geq \mathcal{M}(\mathbf{x}_2) \geq \ldots \geq \mathcal{M}(\mathbf{x}_D)$. For easier notation, we assume
that the variables are already sorted.                                             ▷ SORTING
Let $c$ be large, e.g. $c = 9999999$, and $\epsilon$ small, e.g. $\epsilon = 0.1$, set $r = 1$.
function ESTIMATE THE OPTIMAL NUMBER OF COMPONENTS
    Run the function REGRESSION from below to determine the optimal number
    of components (see Section 3.3.3 in the thesis) for each variable including rounded zeros
                                                ▷ INITIALIZATION OF NUMBER OF COMPONENTS
end function
while $c > \epsilon$ do
    $r \leftarrow r + 1$
    for $l \in \{1, ..., D\}$ do
        function COORDINATE
            Take $\mathbf{X}^{(l)}$ ($l$-th variable at first position) with elements $x_{ij}^{(l)}$;
            compute coordinate representation $\mathbf{z}_1^{(l)}$ and $\mathbf{Z}_{-1}^{(l)}$.
            Let $e_l$ be the detection limit of the $l$-th part; compute coordinates
                $\psi_i^{(l)} = \sqrt{\frac{D-1}{D}} \ln \frac{e_l}{\sqrt[D-1]{\prod_{j=2}^{D} x_{ij}^{(l)}}} \quad \text{for} \quad i \in m_l.$        (16)
                                                ▷ REPRESENTATION IN COORDINATES
        end function
        function REGRESSION
            With previously estimated optimal number of components, estimate the
            regression coefficients $\boldsymbol{\beta}$ with PLS regression:
            $\mathbf{z}_1^{(l)} = \mathbf{Z}_{-1}^{(l)}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{with} \quad \mathbf{Z}_{-1}^{(l)} = \mathbf{T}\mathbf{P}^T$        ▷ PLS REGRESSION
        end function
        function Replacement
            Use the estimated regression coefficients $\hat{\boldsymbol{\beta}}$ to impute the rounded zeros:
            $$\hat{z}_{i1}^{(l)} = \hat{\boldsymbol{\beta}}^T \mathbf{z}^{(l)}{}_{i,-1} - \hat{\sigma} \frac{\phi\left(\frac{\psi_i^{(l)} - \hat{\boldsymbol{\beta}}^T \mathbf{z}^{(l)}{}_{i,-1}}{\hat{\sigma}}\right)}{\Phi\left(\frac{\psi_i^{(l)} - \hat{\boldsymbol{\beta}}^T \mathbf{z}^{(l)}{}_{i,-1}}{\hat{\sigma}}\right)} \quad \text{for } i \in m_l, \qquad (17)$$
            corresponds to the rounded zeros in $\mathbf{z}_1^{(l)}$, and $\phi$ and $\Phi$ are density and
            distribution function of the standard normal distribution, respectively;
            $\hat{\sigma}$ is the estimated conditional standard deviation
            of variable $\mathbf{z}_1^{(l)}$.                                   ▷ REPLACEMENT
        end function
        function Inverse mapping
            Use Equation (4) to express back in the original sample space; reorder
            the variables.
            The values that were originally rounded zeros in the cells $m_l$ in variable
            $\mathbf{x}_l$ are updated.                                  ▷ INVERSE MAPPING
        end function
        function Re-Scaling
            Due to the nature of this inverse mapping, the scale of variables is
            changed. Call $M_i$ the set with the cells of the $i$-th observation that were
            rounded zeros, and $O_i = \{1, \ldots, D\} \backslash M_i$. A cell $x_{ij}$, for any $j \in M_i$,
            is adjusted (multiplied) by the factor $f_{ij} = \frac{\sum_{o \in O_i} x_{io}}{\sum_{o \in O_i} \hat{x}_{io}}$, where $\hat{x}_{io}$ denote
            the inverse mapped values from the previous step.            ▷ ADJUSTMENT
        end function
    end for
    function Update criteria
        Update $c$ as the sum of squared differences of the elements of $\mathbf{X}$ in the $r$-th
        and the $(r-1)$-th iteration.
    end function
end while
Bring the variables to the original order                          ▷ UNDO SORTING
```

lumn of $\mathbf{Z}_{-1}^{(l)}$ consists of ones, taking care of an intercept term in PLS regression, and the observations (rows) are denoted by $\mathbf{z}_{i,-1}^{(l)}$, for $i = 1, \ldots, n$.

Note that due to the complexity of the algorithm, a rigorous proof of convergence is not available. Nevertheless, our practical experience shows that usually just a few iterations are necessary to reach the convergence criterion. The practical application of the algorithm on data from metabolomics and simulation study are presented in the thesis.

6.4. PARAFAC

Metabolite (compositional) data may form a three-way structure. The typical example are repeated measurements of samples in time. In practice, three-way data are of primary interest, especially also in the form of three-way compositions. Let's have $I \times J \times K$ data array (cube): we have I samples and J variables (compositional parts), every sample is measured K times [32]. Consequently, each of K tables of dimension $I \times J$ (slices of the cube) can be considered as a compositional data matrix, ready to be processed using the logratio methodology. In the following text of this section, the whole data cube is denoted as $\underline{\mathbf{X}}$, for slices the notation $\mathbf{X}_k, k = 1, \ldots, K$ is used.

For the possibility to deal with three-way data in a statistical software and also to ease the notation, the data cube is matricized into the form of two-way matrix [38]. Matricizing is done by concatenating matrices for different levels of the third mode next to each other. The column-dimension of the resulting matrix thus becomes quite large in the mode consisting of two prior modes, i.e., the final matrix has dimension $I \times JK$.

Also preprocessing of three-way data, that is of particular importance in the chemometric context [39], must take account specific structure of the observations. The centering is done by the procedure called the single-centering when the unfolded data matrix of dimension $I \times JK$ is centered across the first mode, i.e. single columns are centered. Note that result of centering of compositional data in log-coordinates across rows in single slices is nothing else than clr coordinates of the respective observations; it is easy to see, if we rewrite clr coordinates of a composition $\mathbf{x} = (x_1, \ldots, x_D)'$ as $y_i = \ln(x_i) - \frac{1}{D}\sum_{i=1}^{D} \ln(x_i)$, $i = 1, \ldots, D$. The scaling is usually done through rows of the unfolded data matrix, so we refer to scaling within the first mode. For particular metabolomic applications, scaling within the first mode is replaced by specific approaches in each slice, like the AUC normalization or normalization to creatinine.

PARAllel FACtor analysis (PARAFAC) is one of popular decomposition methods for three-way data in chemometrics [40, 41]. It is a structural model with score matrix $\mathbf{A}_{I\times F}$ and two loadings matrices $\mathbf{B}_{J\times F}$, and $\mathbf{C}_{K\times F}$ with elements a_{if}, b_{jf}, and c_{kf}, for $i = 1, \dots, I$, $j = 1, \dots, J$, $k = 1, \dots, K$, and $f = 1, \dots, F$, where F denotes the number of factors that are extracted. The PARAFAC model in terms of single elements of the data cube $\underline{\mathbf{X}} = (x_{ijk})$ (i.e. for ith observation of jth variable in kth time), can be written as [42]:

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \;\; i = 1, \dots, I, \;\; j = 1, \dots, J, \;\; k = 1, \dots, K; \tag{18}$$

here e_{ijk} stand for residuals. The structure of the model is also visible from Figure 1 [41].

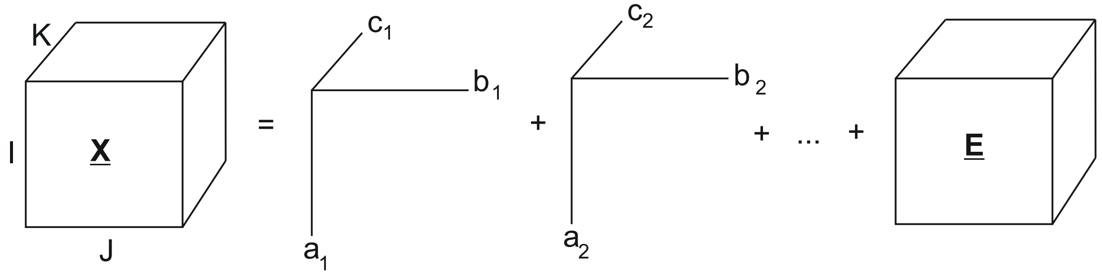


Figure 1: Graphical representation of the formula (18).

By considering F factors, the PARAFAC model consists of $F(I + J + K)$ parameters. The advantage of the PARAFAC model is the uniqueness of the solution; consequently, there is no problem with rotational freedom like for principal component analysis.

The solution of the model (18) is obtained using *alternating least squares (ALS)* algorithm. The principle of ALS is through breaking up iteratively the model into three sets of parameters, such that it is linear in each set given fixed values for the other two sets [42]. Furthermore, we assume that the loadings in two modes are known and then the unknown set of parameters of the last mode are estimated [39, 43]. Explicitly, we define $\mathbf{M} = \left[\mathrm{vec}(\mathbf{b}_1\mathbf{c}_1^{'}), \ldots, \mathrm{vec}(\mathbf{b}_F\mathbf{c}_F^{'})\right]$ and proceed to minimization problem [38, 39]

$$\min_{\mathbf{AM}} \|\mathbf{X} - \mathbf{AM}'\|_F^2, \tag{19}$$

where $\|\mathbf{X}\|_F^2 = tr(\mathbf{X}'\mathbf{X})$ denotes the Frobenius norm of $\mathbf{X}$ [38]. The model for estimation of scores $\mathbf{A}$ is

$$\mathbf{X} = \mathbf{AM} + \mathbf{E}_A, \tag{20}$$

where $\mathbf{X}$ represents unfolded matrix $\underline{\mathbf{X}}$ and $\mathbf{E}_A$ errors of the model, both being of dimension $I \times JK$. The conditional least squares estimate of $\mathbf{A}$ is then

$$\mathbf{A} = \mathbf{XM}(\mathbf{M}'\mathbf{M})^{+} \tag{21}$$

with the Moore-Penrose inverse $(\mathbf{M}'\mathbf{M})^{+}$ of $\mathbf{M}'\mathbf{M}$. The loading matrices $\mathbf{B}$ and $\mathbf{C}$ are estimated analogously [38, 39]. The algorithm is repeated until convergence (i.e., when the changes of scores and loadings from two consecutive steps are small enough) that can be achieved much faster by setting proper initialization values [38].

In the context of PARAFAC modeling of three-way metabolomic data, the normalization is done for each slice that replaces scaling within the first mode (i.e. just centering across the first mode is performed). The practical application of this model is shown in the thesis.

7. Original results and summary

The thesis contains a comprehensive guide to the statistical analysis of metabolomic data using the logratio methodology as developed during my study at the Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University Olomouc and my cooperation with the Laboratory of Metabolomics, Institute of Molecular and Translational Medicine, Palacký University Olomouc. No such broad guide to the analysis of metabolomic (compositional) data does exist yet. Procedures and algorithms introduced in this thesis are in everyday use in the Laboratory of Metabolomics [3,18,44].

Although the logratio methodology was used for the first time with metabolomic data already in [17], this concept was further expanded in publications, based on methodological outputs of the thesis.

In addition to development of a concise procedure to analyze metabolomic data using the logratio methodology, the dissertation thesis newly contributes to three methods for the statistical analysis of multidimensional compositional data. In particular, an adaptation of partial least squares - discriminant analysis to orthonormal logratio coordinates was presented in Section 6.2 of this summary. This concept seems to be very useful for the statistical analysis of high-dimensional compositional data, not just for classification purposes, but also for a range of other purposes, when partial least squares regression represents a proper alternative to the standard least squares model. The second major contribution, presented in the thesis, concerns the parametric model for imputation of rounded zeros based on partial least squares regression and logratio methodology [37] with the Algorithm 6.1. The presence of rounded zeros in metabolomic (and also chemometric) data is quite common and this algorithm overcomes other methods that are currently applied with metabolomic data. The last major contribution concerns PARAFAC modeling of three-way metabolomic (compositional) data. Althought PARAFAC model itself was not updated, a comparison of specific techniques used for the normalization of urine samples in combination with the use of clr coordinates or log-transformation in the context of three-way modeling seems to be of particular importance in metabolomic practice [41].

The most difficult part of this thesis was the necessity of complex view on data. Metabolomic data have a lot of specific features (they are high-dimensional, with specific covariance structure and mostly of compositional nature) and any reasonable method must take care of all of them. Proper statistical analysis of metabolomic samples is crucial for the reliability of the results for further interpretation and processing. In the chemometric and also metabolomic communities, compositional data are still considered as observations with a fixed constant sum constraint, although this is just a possible representation of the relative information, carried by the compositional parts, not an inherent property of the data. Note that the popular logarithmic transformation would solve the problem of moving the relative scale to the absolute one (necessary for a further reasonable statistical analysis), but just for single metabolites, without considering their relative multivariate relations to the other metabolites in the data set. Consequently, application of standard statistical techniques to raw or rescaled metabolomic data often leads to biased results due to ignoring the mathematical implications. On the other hand, the logratio approach to statistical analysis of compositional data is a well mathematically justified methodology that could provide a concise approach to the statistical treatment of biomarkers in metabolomics.

I hope that the presented thesis helps to expansion of the logratio methodology also to the important field of metabolomic data, and to chemometrics in general.

List of publications

Research papers

- C. Kanagaratham, **A. Kalivodová**, L. Najdekr, D. Friedecký, T. Adam, D. Moreno, J.V. Garmendia, M. Hajduch, J.B. De Sanctis, D. Radzioch, Fenretinide prevents inflammation and airway hyperresponsiveness in a mouse model of allergic asthma, *American Journal of Respiratory Cell and Molecular Biology*, vol. 51, no. 6, pp. 783-792, 2014 [45].

- H. Janečková, **A. Kalivodová**, L. Najdekr, D. Friedecký, K. Hron, P. Bruheim, T. Adam, Untargeted metabolomic analysis of urine samples in the diagnosis of some inherited metabolic disorders, *Biomedical Papers*, vol. 159, no. 4, pp. 582-585, 2015 [44].

- **A. Kalivodová**, K. Hron, P. Filzmoser, L. Najdekr, H. Janečková, T. Adam, PLS-DA for compositional data with application to metabolomics, *Journal of Chemometrics*, vol. 29, pp. 21-28, 2015 [3].

- L. Najdekr, **A. Gardlo**, L. Mádrová, D. Friedecký, H. Janečková, E.S. Correa, R. Goodacre,T. Adam, Oxidized phosphatidylcholines suggest oxidative stress in patients with medium-chain acyl-CoA dehydrogenase deficiency, *Talanta*, vol. 139, pp. 62-66, 2015 [18].

- J. Veleba, J. Kopecký, P. Janovská, O. Kuda, O. Horáková, H. Malinská, L. Kazdova, O. Oliyarnyk, V. Škop, J. Trnovská, M. Hájek, A. Škoch, P. Flachs, K. Bardová, M. Rossmeisl, J. Olza, G. Salim de Castro, P.C. Calder, **A. Gardlo**, E. Fišerová, J. Jensen, M. Bryhn, J. Kopecký, T. Pelikánová, Combined intervention with pioglitazone and n-3 fatty acids in metformin-treated type 2 diabetic patients: improvement of lipid metabolism, *Nutrition & Metabolism*, vol. 12, no. 52, pp. 1–15, 2015 [46].

- M. Templ, K. Hron, P. Filzmoser, **A. Gardlo**, Imputation of rounded zeros for high-dimensional compositional data, accepted to *Chemometrics and Intelligent Laboratory Systems*, 2016 [37].

- **A. Gardlo**, A.K. Smilde, K. Hron, M. Hrdá, R. Karlíková, T. Adam, Normalization techniques for PARAFAC modeling of urine metabolomic data, submitted, 2016 [41].

- O. Horáková, J. Hansíková, K. Bardová, **A. Gardlo**, M. Rombaldová, O. Kuda, M. Rossmeisl, J. Kopecký, Plasma acylcarnitines and amino acid levels as an early complex biomarker of propensity to high-fat diet-induced obesity in mice, submitted, 2016 [47].

- R. Karlíková, J. Široká, P. Jahn, D. Friedecký, **A. Gardlo**, H. Janečková, F. Hrdinová, Z. Drábková, T. Adam, Atypical myopathy of grazing horses: a metabolic study, submitted, 2016 [48].

- R. Karlíková, J. Široká, D. Friedecký, E. Faber, M. Hrdá, K. Mičová, I. Fikarová, **A. Gardlo**, H. Janečková, I. Vrobel, T. Adam, Metabolite profiling of the plasma and leukocytes of chronic myeloid leukemia patients, submitted, 2016 [49].

Proceedings papers

- **A. Kalivodová**, K. Hron, M. Župková, H. Janečková, D. Friedecký. Partial least squares for compositional data used in metabolomics. In K. Hron, P. Filzmoser, M. Templ (eds.), *Proceedings of The 5th International Workshop on Compositional Data*, 81–87, 2013.

List of conferences

- ROBUST 2012, 9.-14.9. 2012, Němčičky (CZ): Kompoziční biplot (poster, in Czech, awarded for application theme).

- CoDaWork 2013, 3.-.7.6. 2013, Vorau (AT): Partial least squares for compositional data used in metabolomics (poster).

- ODAM 2013, 12.-14.6. 2013, Olomouc (CZ): Replacement of missing values and rounded zeros in high-dimensional compositional data with application to metabolomics (presentation).

- Škola hmotnostní spektrometrie 2013, 16. - 21.9. 2013, Jeseník (CZ): Statistické metody v MS metabolomice (presentation, in Czech).

- Satelite Workshop on Panomics Data Analysis 2013, 20.11. 2013, Olomouc (CZ): Statistics in metabolomics (presentation).

- ROBUST 2014, 19.-24.1. 2014, Jetřichovice (CZ): Metoda dílčích nejmenších čtverců pro kompoziční data s aplikací v metabolomice (poster+presentation, in Czech, awarded for application theme).

- StatGeo Conference 2014, 17.-20.6. 2014, Olomouc (CZ): Replacement of rounded zeros in high-dimensional compositional data with application to metabolomics (presentation).

- LinStat 2014, 24.-28.8. 2014, Linköping (SW) - Partial least squares discriminant analysis and compositional data applied in metabolomics (presentation, Young Scientists Awards obtained)

- Ercim 2014, 6. - 8.12. 2014, Pisa (IT): PLS-DA for metabolomical (compositional) data using the logratio approach (presentation).

- ODAM 2015, 20.-22.5. 2015, Olomouc (CZ): PARAFAC for compositional data with application to metabolomics (presentation).

- CoDaWork 2015, 1.-5.6. 2015, L´Escala (ES): Imputation of rounded zeros for data from metabolomics (presentation).

Reference

[1] J. Aitchison, *The statistical analysis of compositional data.* Chapman & Hall, London, 1986.

[2] V. Pawlowsky-Glahn, J. Egozcue, and R. Tolosana-Delgado, *Modeling and analysis of compositional data.* Wiley, Chichester, 2015.

[3] A. Kalivodová, K. Hron, P. Filzmoser, L. Najdekr, H. Janečková, and T. Adam, “PLS-DA for compositional data with application to metabolomics,” *Journal of Chemometrics*, vol. 29, pp. 21–28, 2015.

[4] W. Cleveland, “Robust locally weighted regression and smoothing scatterplots,” *Journal of the American Statistical Association*, vol. 74, no. 368, pp. 829–836, 1979.

[5] W. B. Dunn, D. Broadhurst, P. Begley, E. Zelena, S. Francis-McIntyre, N. Anderson, M. Brown, J. Knowles, A. Halsall, J. Haselden, A. W. Nicholls, I. Wilson,

D. Kell, and R. Goodacre, "Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry," *Nature Protocols*, vol. 6, no. 7, pp. 1060–1083, 2011.

[6] R. Brereton, *Chemometrics for pattern recognition.* Wiley, Chichester, 2009.

[7] R. Goodacre, D. Broadhurst, A. Smilde, B. Kristal, J. Baker, R. Beger, C. Bessant, S. Connor, G. Capuani, C. Craig, T. Ebbels, D. Kell, C. Manetti, G. Newton, J. Paternostro, G. Somorjai, M. Sjöström, J. Trygg, and F. Wulfert, "Proposed minimum reporting standards for data analysis in metabolomics," *Metabolomics*, vol. 3, pp. 231–241, 2007.

[8] B. Warracka, S. Hnatyshyna, K. Otta, M. Reilya, M. Sandersa, H. Zhanga, and D. M. Drexler, "Normalization strategies for metabonomic analysis of urine samples," *Journal of Chromatography B*, vol. 877, pp. 547–552, 2009.

[9] S. Waikar, V. S. Sabbisetti, and J. Bonventre, "Normalization of urinary biomarkers to creatinine during changes in glomerular filtration rate," *Kidney International*, vol. 78, no. 5, pp. 486–494, 2010.

[10] O. Haglund, "Qualitative comparison of normalization approaches in maldi-ms," *Master of science thesis, Royal Institute of Technology, Stockholm, Sweden*, 2008.

[11] K. Varmuza and P. Filzmoser, *Introduction to multivariate statistical analysis in chemometrics.* Taylor & Francis, New York, 2009.

[12] P. Gemperline, *Practical guide to chemometrics, 2nd edition.* Taylor & Francis, Boca Raton, 2006.

[13] E. Szymańska, E. Saccenti, A. Smilde, and J. Westerhuis, "Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies," *Metabolomics*, vol. 8, pp. S3–S16, 2012.

[14] J. Trygg, E. Holmes, and E. Lundstedt, "Chemometrics in metabonomics," *Journal of Proteome Research*, pp. 469–479, 2007.

[15] R. Harshman, “Foundations of the Parafac procedure: Models and conditions for an ”explanatory”multimodal factor analysis,” *UCLA Working Papers in Phonetics*, vol. 16, pp. 1–84, 1970.

[16] A. Beddek, P. Rawson, L. Peng, R. Snell, K. Lehnert, H. Ward, and T. Jordan, “Profiling the metabolic proteome of bovine mammary tissue,” *Proteomics*, vol. 8, pp. 1502–1515, 2008.

[17] H. Janečková, K. Hron, P. Wojtowicz, E. Hlídková, A. Barešová, D. Friedecký, L. Žídková, P. Hornik, D. Behúlová, D. Procházková, H. Vinohradská, K. Pešková, P. Bruheim, V. Smolka, S. Šťastná, and T. Adam, “Targeted metabolomic analysis of plasma samples for the diagnosis of inherited metabolic disorders,” *Journal of Chromatography A*, vol. 1226, pp. 11–17, 2012.

[18] L. Najdekr, A. Gardlo, L. Mádrová, D. Friedecký, H. Janečková, E. Correa, R. Goodacre, and T. Adam, “Oxidized phosphatidylcholines suggest oxidative stress in patients with medium-chain acyl-CoA dehydrogenase deficiency,” *Talanta*, vol. 139, pp. 62–66, 2015.

[19] J. Egozcue, “Reply to “On the Harker variation diagrams; . . . ” by J.A. Cortés,” *Mathematical Geosciences*, vol. 41, no. 7, pp. 829–834, 2009.

[20] M. Eaton, *Multivariate statistics. A vector space approach.* John Wiley & Sons, New York, 1983.

[21] J. Egozcue and V. Pawlowsky-Glahn, “Groups of parts and their balances in compositional data analysis,” *Mathematical Geology*, vol. 37, pp. 795–828, 2005.

[22] J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal, “Isometric logratio transformations for compositional data analysis,” *Mathematical Geology*, vol. 35, no. 3, pp. 279–300, 2003.

[23] E. Fišerová and K. Hron, “On interpretation of orthonormal coordinates for compositional data,” *Mathematical Geosciences*, vol. 43, no. 4, pp. 455–468, 2011.

[24] K. Hron, P. Filzmoser, and K. Thompson, "Linear regression with compositional explanatory variables," *Journal of Applied Statistics*, vol. 39, pp. 1115–1128, 2012.

[25] G. Patti, O. Yanes, and G. Siuzdak, "Metabolomics: the apogee of the omics trilogy," *Nature Reviews Molecular Cell Biology*, vol. 13, pp. 263–269, 2012.

[26] M. Lämmerhofer and W. Wolfram, *Metabolomics in practice.* Wiley-VCH, Weinheim, 2013.

[27] R. Little and D. Rubin, *Statistical analysis with missing data.* Wiley, Hoboken, 2002.

[28] B. Walczak and D. Massart, "Dealing with missing data. Part I," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, pp. 15–27, 2001.

[29] R. Rosipal and N. Krämer, *Overview and recent advances in partial least squares.* SLSFS, Springer, 2006.

[30] S. Wold, M. Sjöströma, and E. L., "PLS-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, 2001.

[31] M. Gallo, "Log-ratio and parallel factor analysis: An approach to analyze three-way compositional data," in *Advanced dynamic modeling of economic and social systems* (A. N. Proto, M. Squillante, and J. Kacprzyk, eds.), vol. 448 of *Studies in Computational Intelligence*, pp. 209–221, Springer, Heidelberg, 2013.

[32] M. A. Engle, M. Gallo, K. T. Schroeder, N. J. Geboy, and J. W. Zupancic, "Three-way compositional analysis of water quality monitoring data," *Environmental and Ecological Statistics*, vol. 21, no. 3, pp. 565–581, 2014.

[33] A. Di Palma, M. Gallo, P. Filzmoser, and K. Hron, "A robust Candecomp/Parafac model for compositional data," *Submitted*, 2015.

[34] P. Bastien, V. Vinzi, and M. Tenenhausc, "PLS generalised linear regression," *Computational Statistics & Data Analysis*, vol. 48, pp. 17–46, 2005.

[35] S. Wiklund, E. Johansson, L. Sjöström, E. Mellerowicz, U. Edlund, J. Shockcor, J. Gottfries, T. Moritz, and J. Trygg, "Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models," *Analytical Chemistry*, vol. 80, pp. 115–122, 2008.

[36] J. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser, and J. Palarea-Albaladejo, "Model-based replacement of rounded zeros in compositional data: Classical and robust approaches," *Computational Statistics and Data Analysis*, vol. 56, no. 9, pp. 2688–2704, 2012.

[37] M. Templ, K. Hron, P. Filzmoser, and A. Gardlo, "Imputation of rounded zeros for high-dimensional compositional data," *accepted to Chemometrics and Intelligent Laboratory Systems*, 2016.

[38] R. Bro, *Multi-way analysis in the food industry - models, algorithms and applications.* PhD thesis, Universiteit van Amsterdam, The Netherlands, 1998.

[39] R. Bro, "PARAFAC. Tutorial and applications," *Chemometrics and Intelligent Laboratory Systems*, vol. 38, pp. 149–171, 1997.

[40] V. Pravdova, C. Boucon, S. de Jong, B. Walczak, and D. Massart, "Three-way principal component analysis applied to food analysis: an example," *Analytica Chimica Acta*, vol. 462, pp. 133–148, 2002.

[41] A. Gardlo, A. Smilde, K. Hron, M. Hrdá, R. Karlíková, and T. Adam, "Normalization techniques for PARAFAC modeling of urine metabolomics data," *submitted*, 2016.

[42] R. Harshman and M. Lundy, "PARAFAC: Parallel factor analysis," *Computational Statistics & Data Analysis*, vol. 18, pp. 39–72, 1994.

[43] A. Smilde, R. Bro, and P. Geladi, *Multi-way analysis with applications in the chemical sciences.* John Wiley & Sons, Chichester, UK, 2004.

[44] H. Janečková, A. Kalivodová, L. Najdekr, D. Friedecký, K. Hron, P. Bruheim, and T. Adam, "Untargeted metabolomic analysis of urine samples in the diagnosis of some inherited metabolic disorders," *Biomedical Papers*, vol. 159, no. 4, pp. 582–585, 2015.

[45] C. Kanagaratham, A. Kalivodová, L. Najdekr, D. Friedecký, T. Adam, D. Moreno, J. V. Garmendia, M. Hajduch, J. B. De Sanctis, and D. Radzioch, "Fenretinide prevents inflammation and airway hyperresponsiveness in a mouse model of allergic asthma," *American Journal of Respiratory Cell and Molecular Biology*, vol. 51, no. 6, pp. 783–792, 2014.

[46] J. Veleba, J. Kopecký, P. Janovská, O. Kuda, O. Horáková, H. Malinská, L. Kazdova, O. Oliyarnyk, V. Škop, J. Trnovská, M. Hájek, A. Škoch, P. Flachs, K. Bardová, M. Rossmeisl, J. Olza, G. Salim de Castro, P. Calder, A. Gardlo, E. Fišerová, J. Jensen, M. Bryhn, J. Kopecký, and T. Pelikánová, "Combined intervention with pioglitazone and n-3 fatty acids in metformin-treated type 2 diabetic patients: improvement of lipid metabolism," *Nutrition & Metabolism*, vol. 12, no. 52, pp. 1–15, 2015.

[47] O. Horáková, J. Hansíková, K. Bardová, M. Gardlo, A. Rombaldová, O. Kuda, M. Rossmeisl, and J. Kopecký, "Plasma acylcarnitines and amino acid levels as an early complex biomarker of propensity to high-fat diet-induced obesity in mice," *submitted*, 2016.

[48] R. Karlíková, J. Široká, P. Jahn, D. Friedecký, A. Gardlo, H. Janečková, F. Hrdinová, Z. Drábková, and T. Adam, "Atypical myopathy of grazing horses: a metabolic study," *submitted*, 2016.

[49] R. Karlíková, J. Široká, D. Friedecký, E. Faber, M. Hrdá, K. Mičová, I. Fikarová, A. Gardlo, H. Janečková, I. Vrobel, and T. Adam, "Metabolite profiling of the plasma and leukocytes of chronic myeloid leukemia patients," *submitted*, 2016.