PALACKÝ UNIVERSITY IN OLOMOUC
FACULTY OF SCIENCE

DISSERTATION THESIS

Advanced methods of compositional data analysis

Supervisor: **prof. RNDr. Karel Hron, Ph.D.**
Author: **Mgr. Julie de Sousa**
Study program: P1104 Applied Mathematics
Field of study: Applied Mathematics
Form of study: Full-time
The year of submission: 2023

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Mgr. Julie de Sousa

Název práce: Pokročilé metody analýzy kompozičních dat

Typ práce: Dizertační práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

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

Rok obhajoby práce: 2023

Abstrakt: Celá škála vědeckých oborů produkuje data, u kterých je hlavním zájmem jejich relativní struktura, obsažená ze své podstaty v podílech mezi proměnnými. Vzhledem ke specifickým geometrickým vlastnostem takových (kompozičních) dat je pro jejich relevantní statistickou analýzu nezbytná správná volba reálných souřadnic v rámci logpodílové metodiky. V této dizertační práci jsou představeny nové metody související zejména s využitím tzv. pivotových logpodílových souřadnic v různých oblastech výzkumu generujících datové soubory s vyšší dimenzionalitou nebo komplexností. Jedním z nejzásadnějších úkolů v tzv. -omických vědách je nalezení statisticky významných rozdílů mezi skupinami pacientů a kontrol, které slouží k detekci biomarkerů různých onemocnění s využitím jednorozměrných i mnohorozměrných statistických metod. Je zde představen koncept b-hodnot spolu s bayesovskou verzí populárního nástroje založeného na mnohonásobném testování hypotéz, jež se nazývá vulkánový graf. Díky bayesovské modifikaci lze do grafu zahrnout rovněž zóny vzdálenosti intervalů nejvyšší hustoty (HDI) od nuly. Dále je navržen nový typ souřadnicové reprezentace kompozičních dat, jehož cílem je zlepšit identifikaci biomarkerů. V souladu se svým názvem jsou tyto tzv. selektivní pivotové souřadnice konstruovány tak, že „vodící" souřadnice reprezentující vždy vybranou kompoziční složku agreguje všechny párové logpodíly této složky s ostatními komponentami, s výjimkou aberantních logpodílů. Na souřadnice je následně jako zlatý standard mnohorozměrné analýzy -omických dat aplikována diskriminační analýza metodou částečných nejmenších čtverců. A konečně, složitější strukturu kompozičních dat uspořádaných podle dvou faktorů lze často považovat za kompoziční tabulku. Pro tato data je v práci uvedena speciální volba pivotových souřadnic reflektující možný rozklad tabulky na její nezávislou a interakční část. Za účelem redukce dimenze je pak použita robustní metoda hlavních komponent (PCA), která prostřednictvím přímého vztahu představených souřadnic s centrovanými logpodílovými koeficienty, jenž jsou v kontextu PCA s kompozičními daty tradičně užívány, umožňuje získat lepší vhled do vztahů mezi danými faktory. Teoretické poznatky u prezentovaných metod jsou ilustrovány na analýze reálných datových souborů z metabolomiky a socioekonomie, stejně jako na simulačních studiích demonstrujících přínosy nově navržených nástrojů ve srovnání s těmi v příslušných oborech již etablovanými.

Klíčová slova: kompoziční data, logpodílová metodika, centrované logpodílové koeficienty, pivotové souřadnice, vážené pivotové souřadnice, selektivní pivotové souřadnice, kompoziční tabulky, bayesovská statistika, robustní metoda hlavních komponent, vulkánový graf, metoda částečných nejmenších čtverců – diskriminační analýza, kompoziční biplot, metabolomická data, ekonomická data

Počet stran: 102

Počet příloh: 2

Jazyk: anglický

BIBLIOGRAPHICAL IDENTIFICATION

Author: Mgr. Julie de Sousa

Title: Advanced methods of compositional data analysis

Type of thesis: Dissertation thesis

Department: Dept. of Mathematical Analysis and Applications of Mathematics

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

The year of presentation: 2023

Abstract: An abundance of scientific fields produces data where their relative structure, which is inherently contained in ratios among variables, is of the main interest. Due to the specific geometrical properties of such (compositional) data, a proper choice of real coordinates within the logratio framework is crucial for any sensible statistical analysis. In this thesis, novel methods related particularly to the use of so-called pivot logratio coordinates are presented within different research areas generating data sets of higher dimensionality or complexity. One of the essential tasks in omics sciences is to find statistically significant differences between patient and control groups to detect biomarkers of particular diseases using both univariate and multivariate statistical methods. A concept of b-values is introduced together with a Bayesian version of a widespread tool based on multiple hypotheses testing, the so-called volcano plot, incorporating also distance levels of the posterior highest density intervals from zero. Next, a new type of coordinate representation aiming to enhance the identification of biomarkers is proposed. They are constructed so that the "pivoting" coordinate representing a certain compositional part aggregates all but the deviating pairwise logratios of that part to the remaining ones, in accord with the name selective pivot coordinates. They are further coupled with partial least squares discriminant analysis as a gold standard in the multivariate analysis of omics data. Finally, a data table arranged according to two factors can often be considered a compositional table. Hence, a special choice of pivot coordinates reflecting a decomposition process into independent and interactive parts is presented for compositional data comprising the two-factorial complexity. A robust principal component analysis (PCA) is then performed for dimension reduction, allowing for investigation of the relationships between the given factors through a direct relation of the proposed coordinates to centered logratio coefficients, used traditionally in context of PCA with compositional data. The theoretical background of the presented contributions is illustrated using real data sets from metabolomics and socioeconomy, as well as simulation studies to demonstrate their benefits compared to well-established methods of the respective fields.

Key words: compositional data, logratio methodology, centered logratio coefficients, pivot coordinates, weighted pivot coordinates, selective pivot coordinates, compositional tables, Bayesian statistics, robust principal component analysis, volcano plot, partial least squares discriminant analysis, compositional biplot, metabolomic data, economic data

Number of pages: 102

Number of appendices: 2

Language: English

Statement of originality

I hereby declare that this dissertation thesis has been completed independently, under the supervision of prof. RNDr. Karel Hron, Ph.D. All the materials and resources are cited concerning scientific ethics, copyrights and the laws protecting intellectual property. This thesis or its parts were not submitted to obtain any other nor the same academic title.

In Olomouc

Acknowledgment

I want to thank all who helped and supported me during my Ph.D. study and research, especially my supervisor prof. RNDr. Karel Hron, Ph.D. for his invaluable feedback and motivation. I also could not have undertaken this journey without the members of Laboratory of Metabolomics under the supervision of prof. RNDr. Tomáš Adam, Ph.D. and prof. RNDr. David Friedecký, Ph.D. who generously provided their expertise and interdisciplinary guidance. I am very grateful to all other co-authors of my papers for their knowledge, help, and often moral support, particularly to Mgr. Ondřej Vencálek, Ph.D., the first person who brought me to my love of statistics many years ago. Furthermore, I appreciate all the hospitality and advices of Maria Isabel Ortego, Ph.D. received during my study stay in Barcelona as I am convinced they helped to finish my second article a level better. Lastly, I would be remiss in not mentioning my family and classmates, especially my husband and daughter for their emotional support. Their belief in me has kept my spirits and determination during this process.

Contents

1 Introduction

Compositional data (CoDa) are present in many applications from numerous scientific fields (e.g., economy, sociology, psychology, biology, geochemistry, environmental studies or so-called omics sciences covering metabolomics, genomics, proteomics, transcriptomics, and other branches producing high-throughput data). Logratio methodology based on the Aitchison geometry on simplex (Aitchison, 1986; Pawlowsky-Glahn et al., 2015; Filzmoser et al., 2018) can and should be used as a cornerstone every time the statistician works with strictly positive data carrying relative information. At the same time not only vectors but also more complex structures with the interest lying in several factors can be seen as CoDa. In this dissertation thesis I would like to show the wide potential of the logratio methodology using various different tasks interconnected by a high-dimensionality (or the need for its reduction, respectively) of the relevant data sets using pivot coordinates (Fišerová and Hron, 2011; Filzmoser et al., 2018) or their modifications. Because, to answer any questions about CoDa requires to assign them "appropriate" real coordinates before applying any statistical method (both frequentist and Bayesian) in order to transform the problem from simplex to the real space. Those are, depending on the type of task, especially centered logratio (clr) coefficients allowing for an intuitive interpretation even when dealing with the high-dimensionality and isometric logratio (ilr) coordinates (recently renamed to orthonormal logratio (olr) coordinates to better reflect their geometric properties) – with the special emphasis precisely on pivot coordinates and their weighted counterpart (Hron et al., 2017; Štefelová et al., 2021). The latter are a necessity particularly when some processes in data need to be filtered out in order to obtain better interpretable outputs of the analysis. An overview of logratio methodology for CoDa with the entire development of pivot coordinates, their benefits, and disadvantages is provided in Chapter 2 of this thesis.

The main bottleneck of the statistical analysis and its interpretation in all omics sciences is probably the (ultra) high-dimensionality of their data sets; there are usually hundreds to thousands variables involved with only lower tens (or even less in case of very rare diseases) observations. Another specific of these sciences is the need for a thorough and substantial data pre-processing before any statistical methods can be even applied. This step includes methods used for the

conversion of the original measurements (i.e., chemometric signals) to the data chart, detection of noise variables, and inevitably also data transformation and/or normalization. Naturally, the quality of pre-processing can easily influence the results of the subsequent statistical analysis. For example in the field of untargeted metabolomics or lipidomics where the analysis of clinical patient samples presents a promising way of novel biomarker detection further allowing for a better understanding of pathobiochemical mechanisms and a prediction of various diseases, most of the current publications greatly focus also on the prepocessing steps. For the data transformation, mainly (natural logarithm of) so-called probability quotient normalization (PQN) is used (Dieterle et al., 2006). Here the original data are expressed in terms of ratios to a median of components normalized with respect to some reference sample (usually composed from component-wise medians). The PQN representation is successfully seconded by logratio coordinates where the posed challenge is to find an appropriate counterpart within the logratio methodology to better reflect geometric properties of the relative omics data.

After data pre-processing, tools from both univariate and multivariate statistics are usually used for the analysis in metabolomic experiments aimed at discovering metabolites discriminating the group(s) of patients from healthy controls. In the article

- de Sousa J*, Vencálek O, Hron K, Václavík J, Friedecký D, Adam T (2020) Bayesian Multiple Hypotheses Testing in Compositional Analysis of Untargeted Metabolomic Data. Analytica Chimica Acta 1097: 49–61. DOI: 10.1016/j.aca.2019.11.006.
  * Corresponding author

(Chapter 3 of this thesis) we presented a novel Bayesian approach to a univariate statistical analysis of metabolomic data expressed in first pivot coordinates (or clr coefficients which are up to a scaling constant equal to them) for a multiple hypotheses testing problem. One of the most widespread tools for biomarker identification in omics sciences is the so-called volcano plot (Cui and Churchill, 2003) functioning as a double filter: the size of effect given as a ratio of medians of the patient vs. control data (i.e., a fold-change) is depicted against statistical significance represented by a negative decadic logarithm from p-values obtained

in t-tests of all variables (metabolites). The traditional frequentist way of volcano plot construction, however, suffers from limitations posed by multiple testing of high-dimensional data (Wasserstein and Lazar, 2016) which can result, depending on the choice of the p-value corrections, in a high number of false positives or false negatives leading to a loss of potential biomarkers. On the other hand, the proposed Bayesian approach does not need to rely on any corrections to the number of multiple tests performed from its nature; the decision about a hypothesis is build on highest density intervals (HDI) working with the entire posterior distributions (Kruschke, 2013; Thulin, 2014). Another advantage is the robustness of the method (in Bayesian context) achieved through the prior assumption of the data distribution where a student t-distribution with a possibility of heavier tails is favored over the Gaussian outlier-sensitive option (Kruschke, 2013). For the construction of the volcano plot itself, we suggested to work with the mean values of posterior distributions as a measure of the size of the effect and with newly introduced b-values. The latter provide quite a complex information by taking into account entire posterior distributions and representing them by a singe value substituting the statistical significance. Furthermore, it was shown that a combination of the measures from both axes of the Bayesian volcano plot can be conveniently used in the final assessment of the potential biomarkers. As such, we proposed to construct so-called HDI zones, i.e., distances of the borders of HDI from zero. The entire concept was applied to the analysis of two different inherited rare metabolic diseases, each of them with a bit different specifics and thanks to that also dimensionality, and two simulations. The first one compared the stability of the introduced method and traditional multiple t-testing in case of a loss of samples, while the other scenario considered the influence of the chosen data transformation to a resulting Bayesian volcano plot in case of a systematic error occurring during data measurement.

The results of multivariate statistical methods in metabolomics (or generally also in other omics as well as for example in geochemistry) often suffer from the influence of a handful strong biomarkers on the other variables. This happens regardless to the chosen data transformation due to the nature of mathematical expressions of clr coefficients, pivot coordinates, PQN etc. where such biomarkers impact the coordinates of other components through the (geometric) mean or reference variable, respectively, in the denominator of the coordinate formulas.

An endeavor to eliminate this phenomenon led to a development of selective pivot coordinates (SPCs) presented in the article

- Štefelová N, de Sousa J*, Hron K, Palarea-Albaladejo J, Dobešová D, Kvasnička A, Friedecký D (2023) Selective Pivot Logratio Coordinates for PLS-DA Modelling with Applications in Metabolomics. *Under review.*
  * Corresponding author

(Chapter 4 of this thesis). Pivot coordinates, here termed for better clarity as ordinary pivot coordinates (OPCs), follow a principle where the first ("pivoting") coordinate aggregates all logratios with the compositional part of interest, keeping an easy interpretation just like in the case of clr coefficients (Fišerová and Hron, 2011) (i.e., a dominance of a certain metabolite over the entire metabolome represented by a geometrical mean of all/the rest of metabolites). At the same time, it is possible to create more systems of pivot coordinates (usually the same number as the number of compositional parts) which can be converted to each other by an orthogonal transformation (Filzmoser et al., 2018) and where the part of interest in the first coordinate is permuted. In case of weighted pivot coordinates, the first coordinate from each coordinate system aggregates a relative weighted information about the compositional part of interest (Hron et al., 2017; Štefelová et al., 2021). The atonement for capturing only the relevant and, in other words, more immaculate information in the first coordinate, is generation of a remainder, i.e., another coordinate involving the part of interest where its redundant information is stored. As a weighting technique for classification problems of high-dimensional CoDa, we suggested zero-one weights allowing to fully eliminate aberrant pairwise logratios of the compositional part of interest in its first SPC. The big advantage of such weighting is that there is no specific residual coordinate for the part of interest since the creation of SPCs results in OPC systems with just one difference – the pivoting coordinate of each system is generally no longer the first one. Therefore, SPCs can be seen as a certain orthogonal rotation of the original pivot coordinates which means that the quality of binary classification tasks using multivariate statistical methods on data sets expressed in SPCs does not get deteriorated. As for the particular choice of strategy to assign the weights to the individual compositional parts, we

chose Welch t-statistics with Q_n estimator-based (Rousseeuw and Croux, 1993) confidence intervals to determine individual logratios which should be eliminated from the pivoting coordinate of each set of SPCs (i.e., by assigning zero weights to the respective pairwise logratios). After constructing Welch-based SPCs, partial least squares – discriminant analysis was applied on the data as a well-established method for classification tasks in omics sciences. The estimated regression parameters were further standardized using a bootstrap-based significance test with Benjamini-Hochberg corrections for multiple testing. A comparison of sensitivity and specificity among logarithmized PQN, OPCs and SPCs was provided in a simulation based on the biochemical equation from Filzmoser and Walczak (2014). Especially in case of a higher ratio of potential biomarkers in the total number of metabolites, the newly proposed coordinates outperform the others in both true positive and true negative rate which makes them a very versatile transformation option. The effect of the introduced weighting technique was illustrated on two data sets from targeted lipidomics and untargeted metabolomics, respectively.

More complex CoDa structures where the observations are carrying inherently relative information about data distribution on the basis of two (or even more) factors are not yet common in omics, geochemistry or biology. Nevertheless, to model for example a relative structure of unemployed people depending on their gender and age group, or a relative structure of university students among different study subjects with relation to the obtained university degree, could not be done otherwise. Next to these examples from socioeconomics, other cases of two-factorial compositions might be found e.g., in the field of environmental management, such as mineral resources divided into groups based on their renewability and the type of extraction, or the size of protected areas on land and in the ocean further characterized by the degree of the territorial protection. If we had such data at hand from different countries, the measurements would probably considerably differ depending e.g., on the population size and so the relevant information would be more likely captured by the ratios than absolute values. From the mathematical point of view, we talk about two-factorial extension of vector CoDa, called compositional tables (Egozcue et al., 2008, 2015). Using the logratio methodology, each compositional table can be decomposed into an independent and an interactive part and olr coordinates assigned to each of them (Fačevicová et al., 2016) enabling further statistical processing of compositional tables using

popular multivariate methods. However, the construction of aforesaid coordinates generally requires prior knowledge of the data and it is rather complicated to contemplate the connections among the independence table, interaction table and the entire compositional table where the latter two are influenced by relationships between the two factors. At the same time, it is precisely the connectivity of the individual coordinate systems that should pose as the crucial point in the choice of the coordinates. After all, the comparison of independence and interaction tables is what allows for a better understanding of the original data. For this purpose, in the article

- de Sousa J*, Fačevicová K, Hron K, Filzmoser P (2021) Robust Principal Component Analysis for Compositional Tables. Journal of Applied Statistics 48(2):1–20. DOI: 10.1080/02664763.2020.1722078.
  * Corresponding author

(Chapter 5 of the thesis) we proposed a particular choice of pivot coordinates for all three compositional tables (i.e., the original table and its decomposed parts) with a direct link to clr coefficients including their explicit formulas and interpretation. This is a key step for an application of robust multivariate methods on two-factorial CoDa as well as for a generalization of the entire situation for more than two factors (i.e., multifactorial compositional cubes (Fačevicová et al., 2022)). Regarding the former, we applied on the data expressed in the presented coordinates a robust principal component analysis (rPCA) since one of the most common tasks in statistics is a dimension reduction. To estimate covariation matrix for rPCA, a so-called MCD estimator (Maronna et al., 2006) was used. This approach requires to carry out the computations of loadings and scores using pivot coordinates of vectorized compositional tables, as clr representation leads to singularity, and transform them to clr coefficients only afterward for the purpose of compositional biplots construction. The entire process was illustrated on the two economical data sets mentioned at the beginning of this paragraph.

In addition to the previously mentioned methodological papers, below are listed further papers from an interdisciplinary work in metabolomics, where the focus was primarily

- to improve pre-processing of the measurements by removing data multiplicities using correlation networks:

  Kouřil Š*, de Sousa J*, Václavík J, Friedecký D, Adam T (2020) CROP: Correlation-based Reduction of Feature Multiplicities in Untargeted Metabolomic Data. Bioinformatics 36(9):2941–2942. DOI: 10.1093/bioinformatics/btaa012.

  * Joint first authors

- to identify novel biomarkers in order to help with the description of underlying pathobiochemistry of the studied diseases using the logratio methodology as well as previously mentioned CROP and Bayesian volcano plot:

  Václavík J, Mádrová L, Kouřil Š, de Sousa J, Brumarová R, Janečková H, Jáčová J, Friedecký D, Knapková M, Kluijtmans L A J, Grünert S C, Vaz F M, Janzen N, Wanders R J A, Wevers R A, Adam T (2020) A newborn screening approach to diagnose 3-hydroxy-3-methylglutaryl-CoA lyase deficiency. JIMD Reports 54(1):79–86. DOI: 10.1002/jmd2.12118.

  Mádrová L, Součková O, Brumarová R, Dobešová D, Václavík J, Kouřil Š, de Sousa J, Friedecká J, Friedecký D, Barešová V, Zikánová M, Adam T (2022) Combined Targeted and Untargeted Profiling of HeLa Cells Deficient in Purine De Novo Synthesis. Metabolites 12(3):241. DOI: 10.3390/metabo12030241.

- to enhance the decision process in the newborn screening program of inborn errors of metabolism using a machine learning method coupled with CoDa approach:

  Kouřil Š, de Sousa J, Fačevicová K, Gardlo A, Muehlmann C, Nordhausen K, Friedecký D, Adam T (2023) Multivariate Independent Component Analysis Identifies Patients in Newborn Screening Equally to Adjusted Reference Ranges. *Under review.*

2 Logratio methodology of compositional data

As discussed in the previous chapter, the relative structure of the observations may be more interesting than the absolute values of their components in many real-world applications. For example, when considering numbers of students attending bachelor, master and doctoral studies at different universities, the ratios among these three groups can be more relevant for a statistical analysis than just the empirical values, which might not be comparable because of different total student numbers. In other words, the actual total number of students (sufficiently high so that the impact of a measurement error with small sample sizes can be neglected) might be considered as not informative for the purpose of the analysis. An important point connected to this is the invariance to the change of scale. Suppose that the student numbers are multiplied by a scalar which does not essentially change the information contained in the data. That means, both the original and e.g., percentage representations carry the same information when the relative structure of student degrees is of primary interest. Of course, the challenge is then to process such kind of information in a statistically coherent way. Moreover, if only for example undergraduate student numbers were to enter the analysis, the analysis could be misleading if not conducted carefully to assure results consistent with the findings emerging from the entire composition. Therefore, to work with quantitatively described contributions of a given whole in a concise and meaningful manner, some concepts need to be introduced first.

A positive (row) vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_D)$ is defined to be a D-part composition if it carries relative information, i.e., the ratios between the components are informative (Aitchison, 1986; Pawlowsky-Glahn et al., 2015). Any compositional vectors with equal number of parts are considered to be representatives of the same *equivalence class* if one vector is obtained from another by a positive scalar multiplication (Pawlowsky-Glahn et al., 2015). This is an important point e.g., for some omics sciences where the total often might not be known (i.e., when only peak intensities but not real concentrations are measured). Accordingly, equivalence classes of compositional data are represented without loss of information in a D-part simplex,

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

for any $\kappa > 0$. The choice of κ (being 1 for proportions and 100 for percentages) is irrelevant for the analysis and can also vary throughout the compositional data set. Formally, the closure operation

$$\mathcal{C}(\boldsymbol{x}) = \left(\frac{\kappa \cdot x_1}{\sum_{i=1}^{D} x_i}, \frac{\kappa \cdot x_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa \cdot x_n}{\sum_{i=1}^{D} x_i} \right)$$

can be applied to rescale the data to a given constant sum (κ) representation. Accordingly, the D-part simplex is a sample space of (representatives of equivalence classes of) compositions. As an interesting consequence of the constant sum representation, possibly outlying observations from the main data cloud are characterized by aberrant ratios rather than by significantly deviating absolute values of components (Filzmoser and Hron, 2013).

Therefore, results of statistical processing should not depend on the sum κ of compositional parts and instead of the standard Euclidean distances which rely on absolute (squared) differences between components, relative differences are used to express distances between observations. This principle called *scale invariance* is the first of three basic compositional principles (Pawlowsky-Glahn et al., 2015). Moreover, the original data often contain some non-informative part(s) in the compositional vector that are not of interest. Hence, we do not expect any change of results concerning the respective subcomposition when these parts are removed from the data. *Subcompositional coherence* is a principle declaring that results obtained from a d-part subcomposition, $d < D$, are not in contradiction with results obtained by an analysis of the original D-part composition. Finally, *permutation invariance* states that the results are independent from a chosen order of parts within the composition, an anticipative premise for any reasonable statistical processing and one of the key assumptions for the idea behind the construction of pivot coordinates (Chapter 2.2).

The importance of the compositional principles and possible impacts on the discrepancy of the subsequent analysis are illustrated on a toy example of spurious correlation in Table 1. Let us imagine a metabolomic study where groups of amino acids, organic acids, nucleotides, lipids and other metabolites are measured on samples from 3 healthy controls. Two approaches to the metabolomic analysis are considered and the resulting values are always closed to aliquots; approach A where the mass spectrometry is done with all the metabolites (Table 1a) and ap-

proach B where metabolomics and lipidomics are measured separately, thus lipids are excluded from the metabolomic data set (Table 1b). It is a well expectable assumption for the correlation structure of the groups of metabolites to hold regardless of the chosen analytical approach, however, it is not the case here. It can be seen that the pairwise correlations between amino acids and organic acids, and amino acids and nucleotides, respectively, change when a subcomposition of samples without lipids (after closure operation) is examined (Table 1d). While the former suddenly presents with a positive Pearson correlation coefficient, the latter changes to the negative value; both originally starting with the same zero correlation (Table 1c). Also, correlation coefficient of amino acids with the group of "others" increases from moderate to a very strong between approaches A and B. The reason for this behavior is the choice of the Pearson correlation coefficient for assessing the strength of relationship between compositional parts which is based on Euclidean geometry where the key compositional assumptions generally do not hold.

Table 1: A toy example on spurious correlation of metabolomic data (a whole composition; approach A) and their subcomposition (measurements without lipids; approach B). Closed samples from 3 healthy controls are considered following the distribution of 15 % of amino acids (AA), 18 % of organic acids (OA), 16 % of nucleotides (N), and 13 % of lipids (L) detectable in a human metabolome (Sana et al., 2013) together with the remainder of other metabolites (O).

samples	AA	OA	N	O	L
1	0.15	0.18	0.16	0.39	0.12
2	0.16	0.19	0.15	0.37	0.13
3	0.16	0.17	0.17	0.39	0.11

(a) closed data acquired by approach A

samples	AA	OA	N	O
1	0.17	0.21	0.18	0.44
2	0.18	0.22	0.17	0.43
3	0.18	0.19	0.19	0.44

(b) closed data acquired by approach B

	AA	OA	N	O	L
AA	1	**0**	**0**	**−0.5**	0
OA		1	−1	−0.87	1
N			1	0.87	−1
O				1	−0.87
L					1

(c) correlation matrix for approach A

	AA	OA	N	O
AA	1	**0.31**	**−0.31**	**−0.89**
OA		1	−1	−0.7
N			1	0.7
O				1

(d) correlation matrix for approach B

Therefore, the above principles and the relative scale of CoDa should be captured by a meaningful geometric structure, preferably following the properties of the Euclidean vector space. This is provided by the Aitchison geometry (Pawlowsky-Glahn and Egozcue, 2001; Egozcue et al., 2003). Operations of *perturbation* and *power transformation*, being analogous to a sum of two vectors and a multiplication of a vector by a scalar in the real Euclidean geometry, are defined as

$$\boldsymbol{x} \oplus \boldsymbol{y} = (x_1 y_1, \ldots, x_D y_D) \quad \text{and} \quad \alpha \odot \boldsymbol{x} = (x_1^\alpha, \ldots, x_D^\alpha), \tag{1}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are D-part compositions, and α is a real constant. Accordingly, operations of perturbation and power transformation form a $(D-1)$-dimensional vector space $(\mathcal{S}^D, \oplus, \odot)$ (Pawlowsky-Glahn et al., 2015).

To obtain Euclidean vector space structure, the *Aitchison inner product, norm* and *distance* are defined for D-part compositions $\boldsymbol{x}$ and $\boldsymbol{y}$ as

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_A = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}, \tag{2}$$

$$\|\boldsymbol{x}\|_A = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_A}, \qquad d_A(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} \ominus \boldsymbol{y}\|_A, \tag{3}$$

respectively, where $\boldsymbol{x} \ominus \boldsymbol{y} = \boldsymbol{x} \oplus [(-1) \odot \boldsymbol{y}]$.

Given the introduced specifics of compositional data endowed with the Aitchison geometry, standard statistical methods cannot be applied directly on raw data. Instead of adapting the methods to this specific geometry, it is rather preferred to firstly express compositional data in meaningful real coordinates and then proceed with further statistical processing; i.e., employing the *working on coordinates* principle (Mateu-Figueras et al., 2011).

2.1 Centered logratio coefficients

Generally, there are three types of logratio coordinate representations respecting the Aitchison geometry with interpretation in terms of logratios or their aggregations, *centered logratio coefficients (clr), additive logratio coordinates (alr)* (Aitchison, 1986) and *isometric logratio coordinates (ilr)* (Egozcue et al.,

2003). The former two are defined as

$$\mathrm{alr}(\boldsymbol{x}) = \left(\ln \frac{x_1}{x_D}, \ln \frac{x_2}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D} \right) \tag{4}$$

and

$$\mathrm{clr}(\boldsymbol{x}) = \left(\ln \frac{x_1}{g(\boldsymbol{x})}, \ln \frac{x_2}{g(\boldsymbol{x})}, \ldots, \ln \frac{x_D}{g(\boldsymbol{x})} \right), \tag{5}$$

where $g(\boldsymbol{x})$ stands for the geometrical mean of the whole composition. While alr coordinates could be potentially used in some omics sciences with x_D representing an "anchoring" or "reference" feature, they are only generated with respect to an oblique basis of the simplex (i.e., they form an oblique coordinate system) and do not map the Aitchison inner product (2), norm and distance (3) to the real space resulting also in a violation of the subcompositional coherence (Pawlowsky-Glahn et al., 2015). Since their usage with standard statistical methods thus has some limitations and also represents another streamline in CoDa analysis, they will not be further described for the purpose of this thesis.

On the other hand, clr representation keeps the metric properties of CoDa and enables for a simple and meaningful interpretation in terms of dominance of a given compositional part with respect to the other parts *on average*. Consequently, clr coefficients are useful for a graphical interpretation of compositional data including compositional biplots as a result of a dimension reduction through PCA (Aitchison and Greenacre, 2002) or a multiple hypotheses testing based Bayesian volcano plot (de Sousa et al., 2020). The effect of clr representation is illustrated by another toy metabolomic example (Fig. 1). Imagine the statistical analysis is started with two samples of three metabolites, $\mathrm{M}_1, \mathrm{M}_2, \mathrm{M}_3$ and $\mathrm{M'}_1, \mathrm{M'}_2, \mathrm{M'}_3$, respectively. The measured values vary between these samples (e.g., a situation with differently diluted urine samples), however, the ratios among metabolites are preserved (i.e, 1:16:4). Thus, when the single metabolites are expressed relative to all metabolites (represented by a geometric mean of the entire respective metabolome), the same "absolute" values are achieved for both samples. After applying natural logarithm, the ratios change so the differences in small peaks are exhibited (de Sousa et al., 2020). The logarithmization of the ratios is also a source of some further advantages since it brings symmetry to the data. The values shift from strictly positive onto the entire real space and

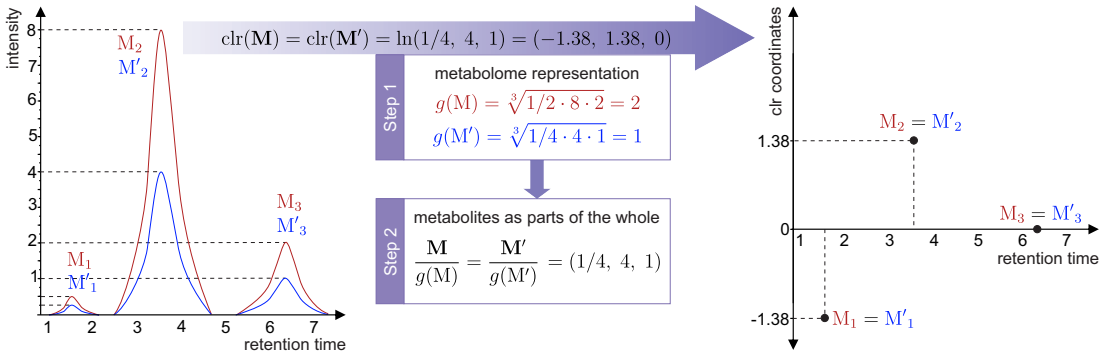

Figure 1: Illustration of the effect of clr coefficients to the (metabolomic) data. Two steps of the workflow can be seen: i) metabolites represented relative to the whole metabolome allow to reveal relative information hidden in the data, ii) application of the natural logarithm provides a way to exhibit differences in smaller peaks.

changing the role of numerator and denominator does not change the information provided by the logratio except for its sign (i.e., $\ln(x_i/x_j) = -\ln(x_j/x_i)$, with $i, j = 1, \ldots, D$).

Another possibility how to understand clr transformed data is through a row-wise centering of logarithmized data,

$$\mathrm{clr}(\boldsymbol{x}) = \left[\left(\ln x_1 - \frac{1}{D}(\ln x_1 + \cdots + \ln x_D)\right), \ldots, \left(\ln x_D - \frac{1}{D}(\ln x_1 + \cdots + \ln x_D)\right)\right].$$

It means that compositional parts after logarithmization are represented in every sample by their arithmetic mean and this mean is subtracted from the logarithmized parts. More importantly, one can see that pairwise logratios to all individual compositional parts are involved in each clr coefficient,

$$\begin{aligned}\mathrm{clr}(\boldsymbol{x}) &= \left(\ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}}\right) \\ &= \frac{1}{D}\left[\left(\ln \frac{x_1}{x_2} + \cdots + \ln \frac{x_1}{x_D}\right), \ldots, \left(\ln \frac{x_D}{x_1} + \cdots + \ln \frac{x_D}{x_{D-1}}\right)\right],\end{aligned}$$

where $1/D$ plays the role of a scaling constant. This is convenient because it guarantees that no information is lost when considering relative contribution of x_i within a given composition. However, it may also lead to biased results caused by presence of either strongly discriminating or "noise" variables (e.g., powerful

biomarkers and values close to detection limit, respectively, in metabolomics) which can heavily influence logratios aggregated into the clr coefficient. Fortunately, while this is quite significant for lower-dimensional data and may even lead to some controversies (Filzmoser and Walczak, 2014), for increasing number of parts the different effects predominantly cancel out for any kind of compositional data (Gardlo et al., 2016; Mert et al., 2016). Moreover, our newly proposed weighting technique used in the construction of SPCs (Štefelová et al., 2023) has the potential to improve the results of the consecutive statistical analysis in terms of both false positive and false negative ratios in such cases.

It is worth noting that clr coefficients sum up to zero which leads to a singular covariance matrix. This reflects dimensionality of compositions, which is just $D-1$ for D-part compositional data. Given the zero-sum condition, it is generally not desirable to analyze any clr part separately without considering the others nor to use clr coefficients with common robust statistical methods (Filzmoser et al., 2009; Filzmoser and Hron, 2013; de Sousa et al., 2021).

2.2 Pivot coordinates

To avoid disadvantages of clr coefficients, ilr coordinates can be used for the mapping of CoDa from simplex to the real space. These orthonormal coordinates (therefore recently proposed to be called rather *orthonormal logratio (olr) coordinates* (Martín-Fernández, 2019)) with respect to the Aitchison geometry, $\boldsymbol{z} \in \mathbf{R}^{D-1}$, can be derived as

$$\boldsymbol{z} = \mathrm{olr}(\boldsymbol{x}) = \left(\left\langle \boldsymbol{x}, \boldsymbol{e}^1 \right\rangle_A, \left\langle \boldsymbol{x}, \boldsymbol{e}^2 \right\rangle_A, \ldots, \left\langle \boldsymbol{x}, \boldsymbol{e}^{D-1} \right\rangle_A\right), \tag{6}$$

where D-part compositions $\boldsymbol{e}^i = \mathcal{C}(e^i_1, e^i_2, \ldots, e^i_D)$, $i = 1, \ldots, D-1$, form an orthonormal basis on the simplex.

Obviously, the interpretation of olr coordinates might be more intricate than in the case of clr coefficients as there are infinitely many possibilities of their construction depending on the choice of basis vectors $\boldsymbol{e}^i$. *Sequential binary partitioning (SBP)* of compositional parts is one possibility for providing a meaningful choice of $\boldsymbol{e}^i$ for the practitioner which is corresponding to the prior knowledge about compositions and resulting in coordinates called *balances* (Egozcue and Pawlowsky-Glahn, 2005). Those can be described as normalized ratios or con-

trasts formed always between two groups of compositional parts which do not overlap until there is nothing left to be further divided anymore.

There is a linear transformation between olr coordinates and clr coefficients, done through a $D \times (D-1)$ matrix $\mathbf{V}$ of clr representations of the olr basis vectors (i.e., *logcontrast coefficients* defined generally as a linear combination of logarithmized parts with zero-sum constraint on the respective coefficients),

$$\mathrm{clr}(\boldsymbol{x}) = \mathbf{V}\boldsymbol{z} = [\mathrm{clr}(\boldsymbol{e}^1)^T, \mathrm{clr}(\boldsymbol{e}^2)^T, \ldots, \mathrm{clr}(\boldsymbol{e}^{D-1})^T] \cdot \mathrm{olr}(\boldsymbol{x})^T. \tag{7}$$

To enable a link to clr coefficients within an olr coordinate system, *(ordinary) pivot coordinates (OPCs)*, $\boldsymbol{z}^{(l)} = (z_1^{(l)}, \ldots, z_{D-1}^{(l)})$, with $z_i^{(l)}$, $i = 1, \ldots, D-1$, given as

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}}, \tag{8}$$

were introduced as a special case of olr coordinates (Fišerová and Hron, 2011; Hron et al., 2017). They are appropriate especially in situations where no prior knowledge about how to perform SBP is available (e.g., in Bruno et al. (2015); Buccianti et al. (2014); Dumuid et al. (2018); Kalivodová et al. (2015)), because they are constructed "semi-automatically". This is certainly an advantage for high-dimensional data and/or multifactorial CoDa structures.

Here, $x_i^{(l)}$ refers to the i-th part of the re-ordered composition $(x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)$ which can be rewritten as $(x_1^{(l)}, x_2^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})$. This indicates that in each of the D coordinate systems, a permutation of compositional parts needs to be performed, so that the l-th part ($l = 1, \ldots, D$) of $\boldsymbol{x}$ stands at the first ("pivoting") position. It ensures that for each part of the original composition, the desired interpretation can be reached in one of the coordinate systems.

Accordingly, the first OPC in each system, $z_1^{(l)}$, then clearly explains all relative information about part x_l and, additionally, it is proportional to the respective clr coefficient from the expression (5) as

$$z_1^{(l)} = \sqrt{\frac{D}{D-1}} \mathrm{clr}(\boldsymbol{x})_l, \tag{9}$$

being an extra asset in case of the univariate statistical analysis. The linear transformation between clr coefficients and olr coordinates (7) naturally holds also for pivot coordinates with logcontrast coefficients

$$\mathrm{clr}(\boldsymbol{e}^i) = \sqrt{\frac{D-i}{D-i+1}}\left(\underbrace{0,\ldots,0}_{i-1},1,\underbrace{-\frac{1}{D-i},\ldots,-\frac{1}{D-i}}_{D-i}\right), \quad i=1,\ldots,D-1.$$

Just as clr coefficients could be rewritten in terms of pairwise logratios, the same can be done with OPCs yielding

$$z_i^{(l)} = \frac{1}{\sqrt{(D-i+1)(D-i)}}\left[\ln\left(\frac{x_i^{(l)}}{x_{i+1}^{(l)}}\right)+\cdots+\ln\left(\frac{x_i^{(l)}}{x_D^{(l)}}\right)\right]. \tag{10}$$

As an alternative to the situation where all pairwise logratios in $z_1^{(l)}$ are treated with the same relevance, *weighted pivot coordinates (WPCs)* were proposed in Hron et al. (2017) with the objective to provide a possibility to enhance or mitigate the effect of some pairwise logratios with the compositional part of interest. If we rewrite the first OPC in the form of the expression (10) with weights $\alpha_j^{(l)}$, $j=2,\ldots,D$ as

$$\alpha_2^{(l)}\ln\frac{x_1^{(l)}}{x_2^{(l)}}+\ldots+\alpha_D^{(l)}\ln\frac{x_1^{(l)}}{x_D^{(l)}}, \quad \alpha_2^{(l)},\ldots\alpha_D^{(l)}\geq 0, \quad \alpha_2^{(l)}+\ldots+\alpha_D^{(l)}=1,$$

the first WPC can be then obtained from here as follows

$$w_1^{(l)} = \frac{1}{\sqrt{1+\sum_{j=2}^{D}\left(\alpha_j^{(l)}\right)^2}}\ln\frac{x_1^{(l)}}{\prod_{j=2}^{D}\left(x_j^{(l)}\right)^{\alpha_j^{(l)}}}. \tag{11}$$

A toll for the non-equal handling of the pairwise logratios with the pivoting compositional part is another coordinate involving $x_1^{(l)}$ where its remaining (relative) information not included in (11) gets stored, i.e., a residual coordinate $w_{D-1}^{(l)}$. While the general formulas for WPC $w_2^{(l)},\ldots,w_{D-1}^{(l)}$ are not provided in the thesis because they are computationally laborious to derive, the way to obtain them is to sequentially apply the orthonormal property of the corresponding logcontrast coefficients, i.e., $\mathrm{clr}(\boldsymbol{e}^{i(l)})\mathrm{clr}(\boldsymbol{e}^{i(l)})^T = 1$ and $\mathrm{clr}(\boldsymbol{e}^{i(l)})\mathrm{clr}(\boldsymbol{e}^{k(l)})^T = 0$,

$i, k = 1, \ldots, D-1, i \neq k$, and the identity $\mathrm{clr}(\boldsymbol{e}^{i(l)})\mathbf{1}^T = 0$, starting with

$$\mathrm{clr}(\boldsymbol{e}^{1(l)}) = \frac{1}{\sqrt{1 + \sum_{j=2}^{D} \left(\alpha_j^{(l)}\right)^2}} \left(1, -\alpha_2^{(l)}, \ldots, -\alpha_D^{(l)}\right), \quad l = 1, \ldots, D.$$

So far, there are two different weighting techniques presented in the literature, both arising from the limitations of OPCs in different practical applications. The first approach published together with the general WPCs formulation in Hron et al. (2017) reflects the need to filter some background noise in geochemical mapping where the calculated concentrations often suffer from measurement errors and imputed rounded zeros. While this could be relatable also for some omics sciences, the chosen weight function

$$\left(\alpha_j^{(l)}\right)^p = \frac{\left(\widetilde{\alpha}_j^{(l)}\right)^p}{\sum_{k=2}^{D} \left(\widetilde{\alpha}_k^{(l)}\right)^p} \qquad \text{with} \quad \left(\widetilde{\alpha}_j^{(l)}\right)^p = \frac{1}{\left(t_{m,n}^{(l)}\right)^p}, \quad j = 2, \ldots, D;\ p > 0$$

based on the variation matrix

$$\mathbf{T}^{(l)} = \left[\mathrm{Var}\left(\ln \frac{x_m^{(l)}}{x_n^{(l)}}\right)\right]_{m,n=1}^{D} = \left[t_{m,n}^{(l)}\right]_{m,n=1}^{D}$$

would generally not work there, as in a majority of situations a certain response variable needs to be considered together with the omics compositional data set.

For regression tasks with high-dimensional compositional explanatory variables, where the response variable is continuous, a weighting approach taking into account the correlation structure of the data was proposed in Štefelová et al. (2021). The weights before normalization, $\widetilde{\alpha}_j^{(l)}$, are defined based on a vector of correlations

$$\boldsymbol{r}^{(l)} = \left(\mathrm{Cor}\left(Y, \ln \frac{x_1^{(l)}}{x_2^{(l)}}\right), \ldots, \mathrm{Cor}\left(Y, \ln \frac{x_1^{(l)}}{x_D^{(l)}}\right)\right)$$

computed between the response variable Y and data expressed in pairwise logratios which are subsequently smoothed by a kernel density estimation and finally integrated from zero to the correlation given by $r_j^{(l)}, j = 2, \ldots, D$.

Both these weighting schemes downplay the parts of the original composition

which have some sort of a poor association with either the pivoting part or the response variable. However, they are not suitable for classification tasks. For the purpose of a categorical response variable coupled with high-dimensional CoDa from metabolomics, another weighting strategy, that can hopefully be seen as the "last piece missing" within the approach where pivot coordinates sophistically aggregate (some) information from all possible pairwise logratios, is presented in Chapter 4.

The geosciences where the usage of pairwise logratios still prevails motivate also the origin of *backwards pivot coordinates* published in Hron et al. (2021a). Employing some kind of "reverse order" in the construction of pivot-like coordinates (i.e., starting with a simple balance of two compositional parts as a scaled pairwise logratio and adding others one by one in an SBP procedure) leads to a possibility of working with the desirable effects of simple logratios without sacrificing the orthonormality of olr coordinates required by many multivariate statistical methods. Starting with a choice of interpretable pairwise logratios (e.g., alr coordinates (4) with x_D as a normalizing geochemical element or any other reference role), an entire set of olr coordinates is built around each of them. This results in systems of $D-1$ backwards pivot coordinates

$$b_i^{(l')} = \sqrt{\frac{i}{i+1}} \ln \frac{\sqrt[i]{\prod_{j=1}^{i} x_j^{(l')}}}{x_{i+1}^{(l')}}, \quad i, l' = 1, \ldots, D-1,$$

which are just orthogonal rotations of each other like in the case of OPCs. The l'-th reordering of the parts of the original composition is chosen in such a way that the pivoting compositional part occupies the first position and the denominator x_D the second one, $\boldsymbol{x}^{(l')} = (x_{l'}, x_D, \ldots, x_{l'-1}, x_{l'+1}, \ldots, x_{D-1})$. With this particular order, a direct link between the first backwards pivot coordinate and the respective l'-th alr coordinate can be expressed similarly to the relationship (9) established between the first OPC and l-th clr coefficient as

$$b_1^{(l')} = \frac{1}{\sqrt{2}} \ln \frac{x_1^{(l')}}{x_2^{(l')}} = \frac{1}{\sqrt{2}} \ln \frac{x_{l'}}{x_D} = \frac{1}{\sqrt{2}} \mathrm{alr}(\boldsymbol{x})_{l'}.$$

For the sake of completeness in the state of the art of the pivot coordinates family, symmetric pivot coordinates (Kynčlová et al., 2017) and their weighted

counterpart (Hron et al., 2021b) should be listed here. As the name suggests, they capture dominance of a compositional part $x_1^{(l)}$ and $x_2^{(l)}$, respectively, over the (weighted) rest of the parts of $\boldsymbol{x}^{(l)}$ in a symmetric way. Their simplified formulas expressed in terms of pairwise logratios are as follows

$$
\begin{aligned}
{}^{s}z_1^{(l)} &= N_1^{(l)} \left(N_2 \ln \frac{x_1^{(l)}}{x_2^{(l)}} + \ln \frac{x_1^{(l)}}{x_3^{(l)}} + \cdots + \ln \frac{x_1^{(l)}}{x_D^{(l)}} \right), \\
{}^{s}z_2^{(l)} &= N_1^{(l)} \left(N_2 \ln \frac{x_2^{(l)}}{x_1^{(l)}} + \ln \frac{x_2^{(l)}}{x_3^{(l)}} + \cdots + \ln \frac{x_2^{(l)}}{x_D^{(l)}} \right), \\
{}^{s}w_1^{(l)} &= N_3^{(l)} \left(\lambda^{(l)} \ln \frac{x_1^{(l)}}{x_2^{(l)}} + \psi_3^{(l)} \ln \frac{x_1^{(l)}}{x_3^{(l)}} + \cdots + \psi_D^{(l)} \ln \frac{x_1^{(l)}}{x_D^{(l)}} \right), \\
{}^{s}w_2^{(l)} &= N_4^{(l)} \left(\lambda^{(l)} \ln \frac{x_2^{(l)}}{x_1^{(l)}} + \delta_3^{(l)} \ln \frac{x_2^{(l)}}{x_3^{(l)}} + \cdots + \delta_D^{(l)} \ln \frac{x_2^{(l)}}{x_D^{(l)}} \right),
\end{aligned}
$$

where $N_1^{(l)}, N_3^{(l)}, N_4^{(l)} > 0$, and $N_2 \xrightarrow{D\to\infty} 1/2$ are normalizing constants and $\lambda^{(l)} > 0, \psi_3^{(l)}, \ldots, \psi_D^{(l)}, \delta_3^{(l)}, \ldots, \delta_D^{(l)}$ are weights, all explicitly defined in Kynčlová et al. (2017) and Hron et al. (2021b) together with the strategies to complete the olr coordinate systems ${}^{s}\boldsymbol{z}^{(l)}$ and ${}^{s}\boldsymbol{w}^{(l)}$, respectively. The need for the symmetrical property of logratio coordinates arises from essentially bivariate statistical methods such as correlation analysis (see e.g., the toy example in Table 1 of Chapter 2) and negative correlation bias of clr coefficients.

2.3 Compositional tables

A considerable amount of practical data sets, such as in econometrics (Fačevicová et al., 2014, 2016), biology (Herder et al., 2008; Dickhaus et al., 2012), or sociology (Egozcue et al., 2008; Ortego and Egozcue, 2016), consist of observations carrying intrinsically relative information about the distribution according to two factors (i.e., two random variables in case of distributional data). From a mathematical perspective, this leads to a two-factorial extension of vector CoDa (Aitchison, 1986; Pawlowsky-Glahn et al., 2015) carrying information about a relationship between and within these (row and column) factors.

Such a structure, called a *compositional table* $\boldsymbol{x}$,

$$\boldsymbol{x} = \begin{pmatrix} x_{11} & \cdots & x_{1J} \\ \vdots & \ddots & \vdots \\ x_{I1} & \cdots & x_{IJ} \end{pmatrix}, \quad x_{ij} > 0, i = 1, \ldots, I, j = 1, \ldots, J, \tag{12}$$

thus can be represented, e.g., either as a contingency table (with sufficiently high numbers of counts in the cells) or as a table of the same order with maximum likelihood estimates of the respective probabilities – due to scale invariance, the relative information (contained in the ratios between the cells) is the same in both cases (Egozcue et al., 2008, 2015). Hence, the concept of compositional tables covers both the discrete case of contingency tables and its continuous counterpart (e.g., input-output tables in Fačevicová et al. (2014)). Nevertheless, in the compositional context, a particular table represents usually just one realization in a sample from a multivariate continuous distribution. Due to the decision to treat such two-factor data compositionally, the possible order of the factor categories (for example age or education levels) is ignored, making this a relevant subject for future research.

Since compositional tables form a direct extension of vector CoDa, all the principles and operations introduced in Chapter 2 apply, up to some minor modifications due to the two-factorial (row and column) structure of the tables.

Accordingly, the closure operation

$$\mathcal{C}(\boldsymbol{x}) = \begin{pmatrix} \frac{\kappa x_{11}}{\sum_{i,j} x_{ij}} & \cdots & \frac{\kappa x_{1J}}{\sum_{i,j} x_{ij}} \\ \vdots & \ddots & \vdots \\ \frac{\kappa x_{I1}}{\sum_{i,j} x_{ij}} & \cdots & \frac{\kappa x_{IJ}}{\sum_{i,j} x_{ij}} \end{pmatrix}$$

is used to represent a compositional table $\boldsymbol{x}$ in an IJ-part simplex $\mathcal{S}^{IJ}$ of vectorized tables $\mathrm{vec}(\boldsymbol{x}) = (x_{11}, \ldots, x_{I1}, \ldots, x_{IJ})$. Perturbation, powering (1), and the Aitchison inner product (2) of two tables $\boldsymbol{x}$, $\boldsymbol{y}$ and a real number α can be defined analogously (Egozcue et al., 2008, 2015),

$$\boldsymbol{x} \oplus \boldsymbol{y} = \begin{pmatrix} x_{11}y_{11} & \cdots & x_{1J}y_{1J} \\ \vdots & \ddots & \vdots \\ x_{I1}y_{I1} & \cdots & x_{IJ}y_{IJ} \end{pmatrix}, \quad \alpha \odot \boldsymbol{x} = \begin{pmatrix} x_{11}^{\alpha} & \cdots & x_{1J}^{\alpha} \\ \vdots & \ddots & \vdots \\ x_{I1}^{\alpha} & \cdots & x_{IJ}^{\alpha} \end{pmatrix},$$

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_A = \frac{1}{2IJ} \sum_{i,j} \sum_{k,l} \ln \frac{x_{ij}}{x_{kl}} \ln \frac{y_{ij}}{y_{kl}},$$

while the Aitchison norm and distance are obtained in the same way as in (3). It is straightforward to derive that the dimension of the simplex $\mathcal{S}^{IJ}$ is $IJ - 1$, corresponding to the dimensionality of $(I \times J)$-compositional tables.

Permutation invariance and subcompositional coherence are valid with respect to the two factors of the compositional tables, allowing to permute and discard entire rows or columns only.

To analyze compositional tables, it is beneficial to work also with the so-called *independence* and *interaction* tables where their separate analysis can be advantageous for further interpretation concerning both factors and their relationships. These independent and interactive parts can be obtained from the original table (12) through an orthogonal decomposition (Egozcue et al., 2008)

$$\boldsymbol{x} = \boldsymbol{x}_{ind} \oplus \boldsymbol{x}_{int}. \tag{13}$$

Here, the independence table is constructed to extract all the relative information about row and column factors under the assumption that the original compositional table is a product of its row and column geometric marginals, and the interaction table contains information about the relationships between the row and column factors. Therefore, in case of actual independence in the data at hand (in the above sense but also in the standard probabilistic sense), all the entries of the interaction table are the same, since there is no remaining information left in the data after the construction of the independence table; the interaction table thus forms a neutral element with respect to the Aitchison geometry of compositional tables. Otherwise, the interactive part describes the nature of the deviation from an independent situation.

It turns out that the introduced decomposition can be easily derived from row and column projections of the compositional table onto marginal subspaces (for further details, see Egozcue et al. (2008)),

$$\mathrm{row}^{\perp}(\boldsymbol{x}) = \begin{pmatrix} g(x_{11}, \ldots, x_{1J}) & \cdots & g(x_{11}, \ldots, x_{1J}) \\ \cdots & \cdots & \cdots \\ g(x_{I1}, \ldots, x_{IJ}) & \cdots & g(x_{I1}, \ldots, x_{IJ}) \end{pmatrix},$$

$$\mathrm{col}^{\perp}(\boldsymbol{x}) = \begin{pmatrix} g(x_{11},\ldots,x_{I1}) & \vdots & g(x_{1J},\ldots,x_{IJ}) \\ \vdots & \vdots & \vdots \\ g(x_{11},\ldots,x_{I1}) & \vdots & g(x_{1J},\ldots,x_{IJ}) \end{pmatrix},$$

where $g(.)$ denotes the geometric mean of the cells in the argument and $\perp$ stands for orthogonality of the projections.

Recalling the case of independence in probability tables, it is instant to get the independence table simply by perturbing both these projections, $\boldsymbol{x}_{ind} = \mathrm{row}^{\perp}(\boldsymbol{x}) \oplus \mathrm{col}^{\perp}(\boldsymbol{x})$. From (13) it follows that the interaction table is just a decomposition remainder in $\boldsymbol{x}_{int} = \boldsymbol{x} \ominus \boldsymbol{x}_{ind}$. For practical calculations, the following formulas are used to obtain the single entries of these tables,

$$x_{ij}^{ind} = \left(\prod_{k=1}^{I}\prod_{l=1}^{J} x_{kj}x_{il}\right)^{\frac{1}{IJ}} \propto \left(\prod_{k=1}^{I} x_{kj}\right)^{\frac{1}{I}} \left(\prod_{l=1}^{J} x_{il}\right)^{\frac{1}{J}},$$

$$x_{ij}^{int} = \left(\prod_{k=1}^{I}\prod_{l=1}^{J} \frac{x_{ij}}{x_{kj}x_{il}}\right)^{\frac{1}{IJ}}. \tag{14}$$

It is crucial to realize that the dimensions of $\boldsymbol{x}_{ind}$ and $\boldsymbol{x}_{int}$ lower to $I+J-2$ for the independence tables, which follows immediately from the dimensions of the row and column projections being, respectively, $I-1$ and $J-1$, and to $(I-1)(J-1)$ for the interaction tables, which is easily obtained from the orthogonality of the decomposition.

Hence, similarly to vector CoDa, an appropriate real coordinate representation of compositional tables, which in addition follows the decomposition into independent and interactive parts, needs to be established with respect to the the sample space dimensionality and the Aitchison geometry (Fačevicová et al., 2016).

In case of compositional tables (and particularly their decomposed parts), a generalization of balance coordinates needs to consider two SBPs according to each factor (Fačevicová et al., 2018). However, even for moderate numbers of rows and columns, the interpretation of such coordinate representation gets rather complex without a deeper expert knowledge. Therefore, only a two-factorial alternative to pivot coordinates is appealing for practice (Fačevicová et al., 2016,

2018).

Generally, there are three types of OPCs corresponding to the row, column and "odds ratio" partitioning of the compositional table (Fačevicová et al., 2016). The first two types jointly form a coordinate representation of the independence table, the third one is used for the interaction table. Altogether, they provide a coordinate representation of the original compositional table. In case of row and column types of coordinates, the entire first row or column, respectively, is taken as the pivoting element and separated from the rest. In the next step, this pivot is not considered anymore and the following row or column is taken as the new (reduced) pivoting element, and so on, until the following $I + J - 2$ coordinates are obtained,

$$z_i^r = \sqrt{\frac{(I-i)J}{1+I-i}} \ln \frac{g(\boldsymbol{x}_{i\bullet})}{[g(\boldsymbol{x}_{i+1\bullet}), \ldots, g(\boldsymbol{x}_{I\bullet})]^{1/(I-i)}}, \quad i = 1, \ldots, I-1,$$

$$z_j^c = \sqrt{\frac{I(J-j)}{1+J-j}} \ln \frac{g(\boldsymbol{x}_{\bullet j})}{[g(\boldsymbol{x}_{\bullet j+1}), \ldots, g(\boldsymbol{x}_{\bullet J})]^{1/(J-j)}}, \qquad j = 1, \ldots, J-1, \tag{15}$$

where $g(\boldsymbol{x}_{i\bullet})$ and $g(\boldsymbol{x}_{\bullet j})$ stand for the geometric mean of the i-th row and j-th column, respectively.

The process of obtaining the remaining $(I-1)(J-1)$ coordinates is based on a division of the original compositional table into four blocks, say upper left A, upper right B, lower left C and lower right D, where A contains always just one (pivot) cell indexed by rs. The odds ratio interpretation should be now easily seen from the following formula, where the elements of blocks A and D are in the numerator, and the elements of blocks B and C in the denominator of the logratio,

$$z_{rs}^{OR} = \sqrt{\frac{1}{(I-r)(J-s)(I-r+1)(J-s+1)}} \ln \prod_{i=r+1}^{I} \prod_{j=s+1}^{J} \frac{x_{ij}x_{rs}}{x_{is}x_{rj}}. \tag{16}$$

To obtain all OPCs of the odds ratio type in a proper order corresponding to the z^r and z^c coordinates (15), the position of the pivoting cell is moving firstly by rows with fixed first column, $r = 1, \ldots, I-1$, then by columns with fixed last row, $s = 1, \ldots, J-1$, and afterward the row position is always leveled back down by one and the column position moves again from 1 to $J-1$ for the given row

until all sizes of the $r \times s$ table are covered.

Finally, permutations of the entire rows or columns following the same principle as stated in Chapter 2.2 could be performed. Hereby for all combinations of rows and columns, different OPC systems consisting of $z_i^{r(k)}$, $z_j^{c(l)}$ and $z_{rs}^{OR(kl)}$, where $(kl), k = 1, \ldots, I, l = 1, \ldots, J$, defines row and column permuted to the pivoting position within the whole table, would be gained (Fačevicová et al., 2016).

3 Bayesian multiple hypotheses testing in compositional analysis of untargeted metabolomic data

Targeted as well as untargeted metabolomic analyses of clinical samples form a promising way to discover new biomarkers allowing better prediction of some diseases. Application of both basic univariate and advanced multivariate statistical methods is a necessary part of all metabolomic experiments aiming to find the most discriminating metabolites between groups of healthy and ill people. After pre-processing the raw untargeted or targeted metabolomic data, respectively, into uniquely characterized features or metabolites, respectively, the differences between patients and controls are often evaluated using t-test or, assuming normality of metabolites is rejected in many cases, its nonparametric alternative – Wilcoxon rank-sum test. The results from multiple testing are compared merely by p-values and fold-changes using a so-called volcano plot. Nevertheless, this approach suffers from the usual frequentist problems, specifically, the high-dimensional character of metabolomic data induces that the multiple simultaneous testing (when used in the correct way, i.e., with p-value corrections) is too strict and tends to produce false negative outputs.

This chapter aims to provide a Bayesian counterpart to the traditional (frequentist) approach. Generally, the methods of Bayesian inference modify prior probabilities of all possible hypotheses or parameter values based on the evidence in the data, until a posterior distribution is obtained (Kruschke, 2014; Gelman et al., 2013). Given fixed parameters, Bayesian t-test assumes t-distributed variables which, since t-distribution is characterized by heavier tails than the normal distribution, results in a robust method in the Bayesian context (Kruschke, 2013). Moreover, it is not needed to consider p-value corrections in Bayesian statistics when running more tests simultaneously since decisions are not based on p-values; Bayesian inference rather relies on the properties of posterior distributions (Gelman et al., 2013; Kruschke and Liddell, 2018). To compare results from multiple hypotheses testing of metabolites and evaluate biomarker candidates, a volcano-like graph using means of posterior distributions together with more sophisticated information provided by the entire posteriors (called b-values) is proposed here. Finally, we suggest incorporating distance levels of the posterior highest density

intervals from zero as an additional feature into the Bayesian version of the volcano plot.

Since a metabolome in an arbitrary biological material can be seen as a complex collection made of ample amount of small molecules (i.e., metabolites), it is rather straightforward to see metabolomic data as compositional in their nature. Also, given that a mass spectrometric response to individual metabolites differs based on their diverse physicochemical properties, signals measured in a metabolomic experiment do not reflect actual concentrations of metabolites. The molar quantification is difficult, laborious, time consuming, and rarely done in current metabolomic experiments since it requires appropriate calibration together with use of suitable internal standards for all metabolites. Hence, the relevant information in metabolomic data is naturally contained not in absolute levels but in ratios/parts of the whole, although the output is anticipated to be interpretable in sense of (groups of) the original metabolites. That is why their analysis needs to be based on the relative structure rather than on absolute values of mass spectrometric measurements even in case of PQN, which have become one of traditional competitors of logratio techniques in omics context (Filzmoser and Walczak, 2014), or other normalizations. The logratio methodology for CoDa introduced in Chapter 2 should be, therefore, an essential step in any statistical treatment of such data including Bayesian analysis. Given the linear transformation (7) between olr coordinates and clr coefficients which gets reduced to the relation (9) for the univariate case, it is sufficient to work with the clr representation of the data in what follows, bearing the respective first OPC in mind instead. Although like in many omics data analyses, also here the interpretability of clr coefficients is satisfactory, the above described mental step leading to pivot coordinates is still needed as univariate analysis with clr coefficients is otherwise inappropriate due to their zero-sum constraint which distorts the covariance structure.

Restrictions of both t-test and its nonparametric version in a multiple hypotheses testing, as well as the limited information they provide, are reminisced in Chapter 3.1. In Chapter 3.2, a Bayesian counterpart to a (non)parametric t-test and its evaluation in the case of multiple testing are provided. Theoretical developments are illustrated in Chapter 3.3 on real data analyses comparing, respectively, plasma samples and dry blood spots of healthy controls and patients suffering from inherited metabolic disorders of 3-hydroxy-3-methylglutaryl-CoA

lyase deficiency (HMGCLD) and medium-chain acyl-CoA dehydrogenase deficiency (MCADD). Lastly, two simulations designed to mimic a loss of samples and a systematical measurement error are used to compare the performance of the traditional methods with the newly proposed approach in Chapter 3.4.

3.1 Limitations of traditional hypothesis testing

Prior to the introduction of the Bayesian approach to hypothesis testing, the procedure of the frequentist approach is recalled, i.e., the case of a parametric two-sample t-test. The null hypothesis suggests that there is no difference between central tendencies of two compared groups, whose observations are described by normal distributions with parameters μ_1, σ_1, and μ_2, σ_2, respectively. A p-value of such a test is a probability to, assuming the null hypothesis is true, obtain the data we have or even more extreme results towards the alternative hypothesis. If this probability is very low, it suggests that the observations are in contradiction with the null hypothesis and therefore the hypothesis should be rejected in favor of the alternative; there is a statistically significant difference between the groups. The threshold of a p-value for rejection is usually set to a significance level $\alpha = 0.05$.

Running more tests simultaneously (under the assumption of independence), say D t-tests for D measured metabolites, results in an unacceptable $(1 - (1 - \alpha)^D) \cdot 100\%$ probability of getting at least one false positive result and, therefore, the significance level needs to be appropriately lowered. There is a number of approaches to do so; a well-known concept is a simple Bonferroni correction, setting $\alpha = 0.05/D$ for each of the multiple comparisons. However, with increasing D this results in a close-to-zero significance level and the Bonferroni correction becomes useless because it tends to produce false negative results; the procedure is not able to detect true biomarkers. Furthermore, in the case of nonparametric testing, i.e., when using the well-known Wilcoxon rank-sum test, the newly set significance level might not be even feasible to overcome. Thus it might be impossible to reject the null hypothesis if only a small number of observations was available; this is anyway a frequent case in metabolomics of rare diseases. A less conservative way, appropriate in particular when dealing with high-dimensional data, is to use some corrections derived from a so-called *false discovery rate (FDR)*, e.g., Benjamini-Hochberg (Benjamini and Hochberg, 1995)

or Benjamini-Yekutieli corrections (Benjamini and Yekutieli, 2001). FDR-based corrections, which are frequently used in the last years, weaken the negative effects of the crude Bonferroni correction but are not fully able to overcome them (i.e., corrected p-values still may not exceed the significance level in case of hundreds of metabolites). Unfortunately, even today numerous publications build the final conclusions only on completely uncorrected results of multiple tests despite the opposite general consensus published in Wasserstein and Lazar (2016).

In addition, (parametric) t-tests are not designed to handle outliers which can easily occur in metabolomic analysis and lead to a distortion of classical statistical procedures including hypothesis testing. Their robust counterparts exist but may become numerically unstable with small sample sizes which are typical for metabolomic data of rare diseases. Hence it is an advantage of Bayesian counterpart to a t-test, introduced in the next chapter, to be a "naturally robust method" due to a proper choice of the prior distribution. In compositional data, moreover, the outlying observations are characterized by deviating logratios, while in standard data sets the same is caused by deviating (absolute) values of the original components. This needs to be taken into account when clr coefficients or OPCs are processed.

Another problem raising from performing multiple t-tests is the absence of a statistically sound decision criterion for an order of the results according to the magnitude of differences between the tested groups (e.g., some criterion ordering the metabolites in consonance with the importance of differences separating patients and healthy controls) and, consequently, identifying the possible biomarkers. Although p-values are still too frequently misinterpreted as a tool for doing so, they do not provide any means of comparability among the rejected hypotheses (Wasserstein and Lazar, 2016). That is why the choice of potential biomarkers should never be done only based on p-values arrangement. On the other hand, *volcano plot* (Cui and Churchill, 2003; Li, 2012), which is a type of a scatter plot used to identify significant changes in large data sets (Fig. 2), already grants a certain way to compare the results of multiple tests thanks to its double-filtering (i.e., by an effect size and a statistical significance). Volcano plot is usually depicting a $\log_2$ fold change of means (medians) of the two groups on x-axis and a $-\log_{10}$ of the t-test (Wilcoxon rank-sum test, ANOVA) p-values on y-axis for every metabolite. Size of the negative logarithm of p-value tends

to increase with an absolute value of the logarithm of the fold change (Cui and Churchill, 2003) and thus all the points in this graph form a "V" shape while potential biomarkers can be found in both upper corners (on one side for healthy controls and on the other one for patients). Yet, the decision based on a volcano graph is to a large extent subjective as there is no general consensus regarding the interpretation of the plot (i.e., importance of the axes and their thresholds) (Li, 2012). We suggest here a Bayesian counterpart that will not suffer from this limitation.

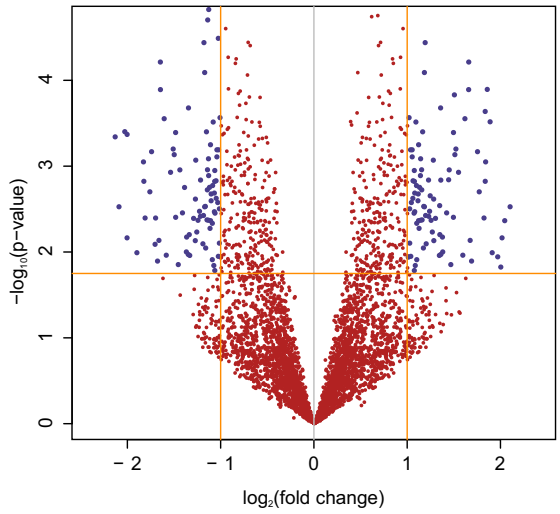

Figure 2: An example of a traditional volcano plot with an effect size expressed by $\log_2$(fold-change) on x-axis and a statistical significance given as $-\log_{10}(p\text{-value})$ on y-axis. Variables depicted in blue are evaluated as significant by employing the most frequent thresholds, $\text{abs}(\log_2(\text{fold-change})) = 1$ and $p\text{-value} = (\text{Bonferroni-corrected } \alpha)$, which are highlighted in orange.

3.2 Bayesian counterpart to a t-test

Methods of Bayesian inference basically reallocate some prior credibility across the space of all possible hypotheses or values of parameters consistently with the data evidence (Kruschke, 2014; Gelman et al., 2013). For the construction of the Bayesian counterpart to the t-test, several steps are needed.

First, as mentioned above, classical t-test assumes a normal distribution of each of the two samples. The normal distribution has light tails and, consequently, it is not appropriate for a description of any data containing outliers.

Here, t-distribution seems to be more convenient because it can be much heavier tailed, depending on degrees of freedom ν. Higher the value of ν is, closer the tails are to those of the normal distribution which is also why ν is called a *normality parameter* in Bayesian statistics (Kruschke, 2013). It turns out that t-distribution is a suitable choice also for the logratio representation (specifically clr coefficients as a workhorse for OPCs) of metabolomic data. Please note that the original measurements (strictly positive data) could hardly be characterized by a t-distribution whose domain is the whole real line. In Bayesian t-test, each of the two groups of samples, i.e., clr represented patients and controls, has its own mean μ_{pat} and μ_{con}, respectively, whose difference is of the main interest, and its own standard deviation σ_{pat} and σ_{con}. The normality parameter is shared by both groups (Kruschke, 2013). To make a qualified decision about the null hypothesis stating no difference in means among the tested samples, all five model parameters need to be inferred.

When choosing prior distributions of the parameters, it is always beneficial to have at least some knowledge about the behavior of the variables (and therefore the parameters), i.e., central tendencies in groups of controls and patients for mass to charge ratios (m/z) of all metabolites, because the initial belief should be ideally reflected in the choice of priors. Conversely, when there are no relevant historical data or expert assumptions, it is generally advised to select as vague priors as possible, named non-informative priors, to allow already a moderate amount of data to deflect the original setting into the direction driven by the evidence (Kruschke, 2014). Additionally, the usage of non-informative priors is supported by utilization of so-called credible sets (Thulin, 2014) which will be defined later. Here, the case of vague priors allowing for their reduced importance during the inference is almost inevitable to follow as in untargeted metabolomics it is prevailing not to have any well-founded prior knowledge for a vast majority of the measured features.

In line with the previous thoughts, priors of the mean value parameters are taken as normally distributed, $\mu_\bullet \sim \mathcal{N}(\overline{x}_\bullet, 1000^2 s_\bullet^2)$, where $\overline{x}_\bullet$ and $s_\bullet^2$ are group sample means and variances from the clr representation of the data at hand with $\bullet$ denoting the group of patients and controls, respectively. In accordance with the non-informative priors philosophy, they are scaled relative to the observations and wide enough not to be confining. The initiatory distri-

bution of the standard deviation is assumed to be uniform on a large enough interval, $\sigma_\bullet \sim \mathcal{U}(1/1000 s_\bullet, 1000 s_\bullet)$, and finally, the prior of the shared parameter ν is exponential with expectation equal to 30 to accommodate the initial credibility evenly between light-tailed data and data with outliers, i.e., $\nu - 1 \sim \text{Exp}(1/29)$ (Kruschke, 2013).

Once the prior assumptions are assigned, the distributions of parameters $\mu_\bullet, \sigma_\bullet$, and ν can be continuously modified with gradually coming observations in terms of conditional "probabilities" of these data X, given the values of the parameters (likelihood). Eventually, this process of credibility reallocation leads towards the posterior distribution. The inference is driven by the Bayes' rule stating the posterior to be proportional (up to an integration constant) to the likelihood times prior,

$$f(\mu_{\text{pat}}, \sigma_{\text{pat}}, \mu_{\text{con}}, \sigma_{\text{con}}, \nu | X) \propto f(X | \mu_{\text{pat}}, \sigma_{\text{pat}}, \mu_{\text{con}}, \sigma_{\text{con}}, \nu) \times f(\mu_{\text{pat}}, \sigma_{\text{pat}}, \mu_{\text{con}}, \sigma_{\text{con}}, \nu), \tag{17}$$

where the joint prior distribution density $f(\mu_{\text{pat}}, \sigma_{\text{pat}}, \mu_{\text{con}}, \sigma_{\text{con}}, \nu)$ can be, assuming independent parameters, rewritten as a product of marginal densities of the single parameters. This assumption permits to take the posterior density simply as a product of prior parameter distribution densities, and t-distributed probability density reflecting the data evidence, making this an important step simplifying the computations.

In practice, posterior density is numerically approximated by a class of *Markov chain Monte Carlo methods (MCMC)* (Gelman et al., 2013) which generates samples from the (non-normalized) posteriors (17),

$$\left\langle \mu^j_{\text{pat}}, \sigma^j_{\text{pat}}, \mu^j_{\text{con}}, \sigma^j_{\text{con}}, \nu^j \right\rangle, \quad j = 1, \ldots, N, \quad N \text{ large},$$

corresponding to both the data and the priors. The only disadvantage of this approach is a possible presence of autocorrelations in the generated sample since for each $j, k = 1, \ldots, N, k \neq j$, the parameter combinations $\left\langle \mu^j_{\text{pat}}, \sigma^j_{\text{pat}}, \mu^j_{\text{con}}, \sigma^j_{\text{con}}, \nu^j \right\rangle$ and $\left\langle \mu^k_{\text{pat}}, \sigma^k_{\text{pat}}, \mu^k_{\text{con}}, \sigma^k_{\text{con}}, \nu^k \right\rangle$ are no longer independent. Thankfully, it can be observed that for a long enough MCMC sample, say $N = 100{,}000$, the autocorrelation is effectively lowered just by the chosen chain length and the estimation of the posterior distribution remains credible (Kruschke, 2014).

The final decision concerning the null hypothesis is very intuitive in Bayesian hypothesis testing with the use of credible sets (Thulin, 2014), for example *highest density interval (HDI)*, which can be formally defined by the following inequality

$$\mathrm{P}(\mu_{\mathrm{pat}} - \mu_{\mathrm{con}} \in \Theta_{\mathrm{HDI}} | X) \geq 1 - \alpha, \tag{18}$$

is constructed to contain 95 % of the most frequented posterior values Θ_{HDI}. Since the resulting MCMC chain of differences between means of clr representation of both original groups of samples can be plotted into a histogram, it may easily be computed where those Θ_{HDI} values are allocated. If this interval does not contain zero, the hypothesis about equality of parameters μ_{pat} and μ_{con} is rejected and the posterior distributions are accepted to be significantly different. Moreover, as can be seen in Fig. 3, the sign of the majority of HDI values further reveals the direction of this difference.

In the same manner as (18), HDI can be constructed also for the difference

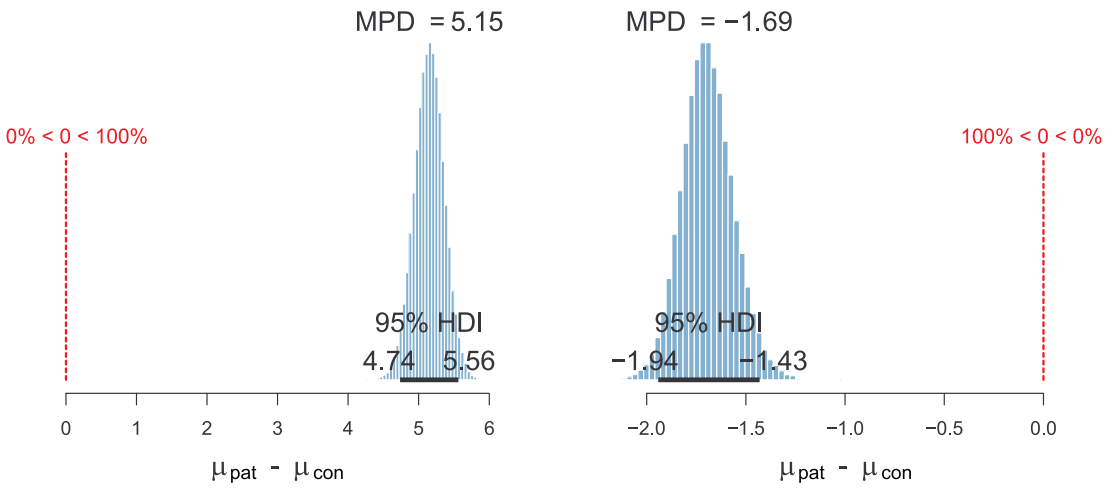

Figure 3: Examples of null hypotheses rejection based on highest density intervals (HDI) where MPDs stand for means of posterior distributions. Differences between Monte Carlo Markov Chain generated posteriors of μ_{pat} and μ_{con} are depicted in blue, where their respective parts in negative and positive values are given by the percentages in red. In both examples, there is strong evidence against the equality of means of the clr represented groups and thus biomarker candidates are detected. On the left, the significant difference is caused by unexpectedly high levels of the metabolite in a group of patients with respect to an average behavior of the whole metabolome; on the right, the opposite tendencies with a relatively high concentration of the depicted metabolite in a metabolome of healthy controls can be observed.

of standard deviations $\sigma_{\text{pat}} - \sigma_{\text{con}}$ instead of central parameters, which produces a very similar output (up to the type of the probability distribution) as in Fig. 3, allowing for a further Bayesian analysis of variance.

An interesting property of the Bayesian t-test is that the null hypothesis can be also accepted. To do so, a concept of a *region of practical equivalence (ROPE)* (Kruschke, 2013) needs to be introduced. Researchers might specify an interval of values being interchangeable with zero for all practical purposes. For example, the difference of 0.5 between groups of patients and controls for a certain metabolite can be determined to be equivalent with no difference at all due to some measurement tolerance. Then it is natural to set the ROPE $= [-0.5, 0.5]$ for such a metabolite. Every time HDI of the posterior distribution happens to be located entirely inside its ROPE and to contain the zero value at the same time, it is a strong enough evidence for accepting the null hypothesis. However, this particularity is impossible to achieve with small ROPE and a small number of observations at the same time, or in other words in untargeted metabolomics of rare diseases.

3.2.1 Multiple Bayesian hypotheses testing

While it was quite immediate, how a decision is made in a single Bayesian t-test, multiple testing complicates the situation a bit. Naturally, except for the hypotheses rejection, we also seek some importance order of metabolites based on the results of the analysis. This can be done simply according to *means of posterior distribution (MPD) criterion* which is a mean of a difference of posteriors of given parameters $\mu_{\text{pat}}, \mu_{\text{con}}$. However, it would lead to a serious loss of information if the complex posterior distribution was reduced just to its MPD value. In addition, empirical probabilities that the differences in μ_{pat} and μ_{con} would have an opposite sign than indicated by posterior distributions can be considered. Even though it is inappropriate to sort the metabolites using just p-values obtained from classical t-tests, some ordering based on the above-mentioned probabilities, which we suggest to call *b-values*, can be performed. Formally, we propose to define

$$b\text{-value} = \min\left\{\mathrm{P}(\mu_{\text{pat}} - \mu_{\text{con}} > 0), \mathrm{P}(\mu_{\text{pat}} - \mu_{\text{con}} < 0)\right\}, \tag{19}$$

where the probabilities are computed from the MCMC posterior distribution. The idea of b-values comes from the concept of discrepancies for posterior predictive assessment of model fitness described in Gelman et al. (1996). Here, instead of judging the fit of some model to the analyzed data, compatibility with the null hypothesis is evaluated by a b-value.

An analogous procedure was proposed to quantify the evidence against the rejected hypothesis when computing the largest credible set which does not contain those values of the tested parameter θ that are valid just under the assumption of the null hypothesis; say credible set Θ_T without values θ_0. Then a probability

$$\mathrm{P}(\theta \notin \Theta_T | X) = \alpha_{\min}, \tag{20}$$

where $\alpha_{\min}$ is the smallest α ensuring that the credible set Θ_T does not contain θ_0, has a very similar meaning to the p-value from a traditional t-test whilst considering the entire posterior distribution, in particular also its tails (De Bragança Pereira and Stern, 1999; Thulin, 2014). The above-suggested b-value (19) could be seen as a certain variation to the idea described by the expression (20), using the smaller part of HDI divided into two intervals by $\theta_0 = \mu_{\mathrm{pat}} - \mu_{\mathrm{con}} = 0$ as an empirical probability of a realization of the posterior on the other side of the zero value.

The b-values are hardly computable when the data from both groups are strongly discriminated. As a consequence, the posterior distributions obtained from the difference of central tendencies of both groups of samples, i.e., $\mu_{\mathrm{pat}} - \mu_{\mathrm{con}}$, are far from zero. They can be even so far from zero that the empirical probabilities of them having opposite signs naturally equal zero (as is also the case of both the examples given in Fig. 3), which may happen for a considerable number of metabolites. Recall that the posterior distributions are acquired by MCMC and as such, the tails are cut at a certain point. Since this can often occur when dealing with real metabolomic data sets (see Chapter 3.3.2), b-values might be alternatively computed using a fitted theoretical distribution to the posterior histograms.

Although a difference of two or more t-distributed probability density functions (pdf) is generally a Behrens-Fisher pdf (i.e., a linear combination of Student's pdf with coefficients formed by sine and cosine of a certain constant which is re-

flecting different population variances), it can be approximated under reasonable conditions by a t-distribution (Patil, 1965; Davis and Scott, 1973). This is also the case of the Bayesian t-test posteriors and so we propose a pooled t-distribution to be fitted on the posterior difference of $\mu_{\text{pat}} - \mu_{\text{con}}$ resulting from each of the multiple tests. Due to the tails of this fit going already to the infinity as opposed to the MCMC result, required b-values are no longer of a zero value albeit they can be fairly small. Subsequently, a version of b-values (empirical or t-distributed pdf–based) which can be used to order the metabolites according to their ability for discrimination is always available.

Both MPD values and b-values are at disposal for the final choice of potential biomarkers from all original metabolites. At this point, some kind of a Bayesian version of the volcano plot may be just the convenient tool for a graphic representation of the results from multiple hypotheses testing, depicting the MPD values on x-axis and $-\log_{10}$ of the b-values on y-axis. Nonetheless, it still remains to a subjective decision which axis should contribute more to the final decision; it is advisable to consider both statistical and metabolomic background. Generally, whenever the variances of posteriors tend to significantly differ among metabolites (e.g., in the analysis of cells), we suggest to rather rely on MPD criterion since the influence of the variance fluctuation affecting the b-values could be even more inflated by the logarithm. In other cases, given some variance stability, the complex information of posterior distributions reflected on the y-axis is favorable.

Another advantage of the Bayesian approach can be seen in the possibility to combine the information from both volcano graph axes, making the interpretation of the plot more straightforward. Whilst there have been recently similar attempts in case of the traditional volcano plot (e.g., in Kumar et al. (2018)) and the field can still be explored more in the future, we suggest a very straightforward idea for the Bayesian counterpart. As was explained in the first part of this chapter, the decision about a single metabolite (in terms of its clr representation) is made through HDI (not) containing the $\mu_{\text{pat}} - \mu_{\text{con}} = 0$ value, under the initial (prior) assumption that the null hypothesis is valid. One could then explore this behavior further and, for those hypotheses that are rejected in the previous step, take the distance of the lower or upper HDI boundary from zero (whichever is in a closer proximity) akin to the measure of evidence against the presumed equality of central behavior of the two groups of samples. The biomarker candi-

dates could then be ordered in accord with these distances of HDI from zero. The information on *HDI distance levels* is easy to incorporate to the final version of the Bayesian volcano plot in a form of colored zones. The shape of the individual HDI levels, which can be seen in Fig. 4 and Fig. 6, indeed confirms that the idea of the distance levels merges both aforementioned criteria for double-filtering of biomarkers.

3.3 Analysis of rare inherited metabolic disorders

Rare diseases are according to the European Medicines Agency (EMA) defined as medical conditions appearing in less than 5 cases per $10,000$. There are over $1,000$ of such conditions varying in incidence where some of them are extremely rare (a few patients only described in the literature). Two rare inherited metabolic disorders were analyzed here; organic aciduria caused by a deficiency of enzyme in leucine metabolism – HMGCLD, and a disorder in beta oxidation of fatty acid metabolism – MCADD. While MCADD is one of the most frequent rare diseases (with the incidence of $1 : 14,600$ (Rhead, 2006)) and it is globally part of the newborn screening, HMGCLD has on the other hand incidence more typical for the rare diseases (less than $1 : 100,000$ (Pié et al., 2007)) and is currently impossible to screen in majority of the countries worldwide. Therefore, HMGCLD represents the type of high-dimensional data that metabolomic experts on rare diseases often have to deal with and so it could conceivably show whether the results of Bayesian multiple hypotheses testing remain valid for a statistically problematic low number of samples. The MCADD data set, on the other hand, provides an opportunity to run at least some small simulations to ensure better comparison of the proposed method with the traditional approach than if evaluated just on the real data analysis itself.

3.3.1 3-Hydroxy-3-methylglutaryl-CoA lyase deficiency

The HMGCLD data set is a result of a recent untargeted metabolomic study (Václavík et al., 2020) performed on plasma samples of 5 patients in a range of 4 days to 8 years of age and 21 age-matched controls. The samples were analyzed by reverse-phase liquid chromatography coupled to orbital ion-trap high-resolution mass spectrometry in positive mode in the range of m/z 90 - 1,000.

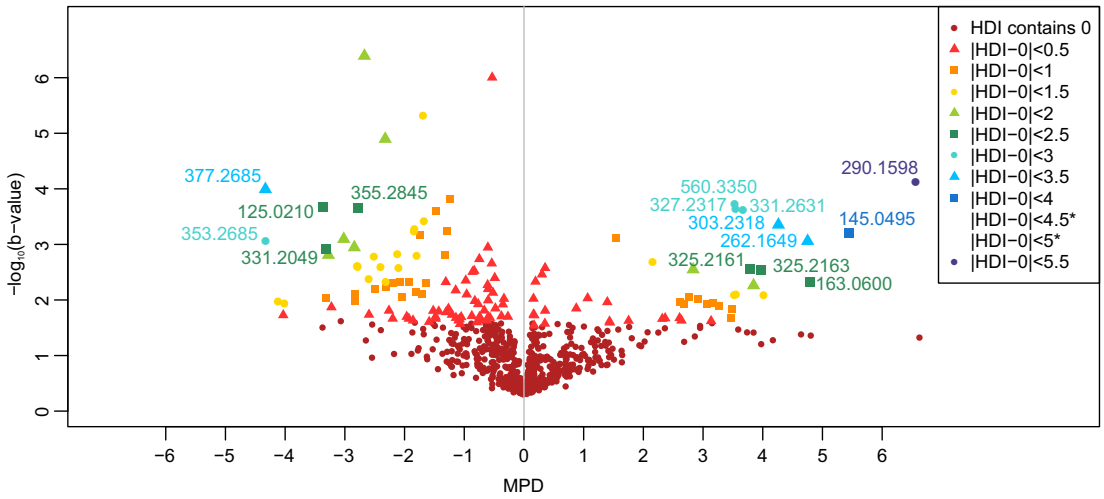

Figure 4: Bayesian volcano plot with colored HDI distance levels for the HMGCLD data set. The dark red points stand for the insignificant compounds where the null hypotheses about equality of the parameters μ_{pat} and μ_{con} were not rejected. The potential biomarkers are depicted in blue-green colors, located in the upper right corner for patients and in the upper left corner for controls, respectively. Metabolites with m/z of 290.1598 and 262.1649 represent previously published biomarkers of the disease.

* – no metabolites detected in these HDI zones

Peak-picking was conducted with Compound Discoverer 3.0 including peak area integration, gap-filling and retention time alignment. Based on the previous steps, a table with metabolomic features (characterized by unique m/z and retention time) and corresponding peak areas of these features in all samples (relative quantitative data) was generated. Peak-picking was followed by removal of adducts, isotopes and ion source fragments applying correlation networks (Kouřil et al., 2020). Afterward, the data were pre-processed in `R` software (R Core Team, 2022) using a package `Metabol` (Gardlo et al., 2019). Employing calculations from quality control samples, locally estimated smoothing signal (LOESS) correction was applied (Sumner et al., 2007) and features with a coefficient of variation higher than 30% were excluded from following data curation. The total amount of unique metabolites after data processing and filtering was 808. The data were then expressed in clr coefficients for further statistical analysis.

There are two previously known diagnostically significant plasma metabolites of HMGCLD, 3-hydroxyisovalerylcarnitine and 3-methylglutarylcarnitine with

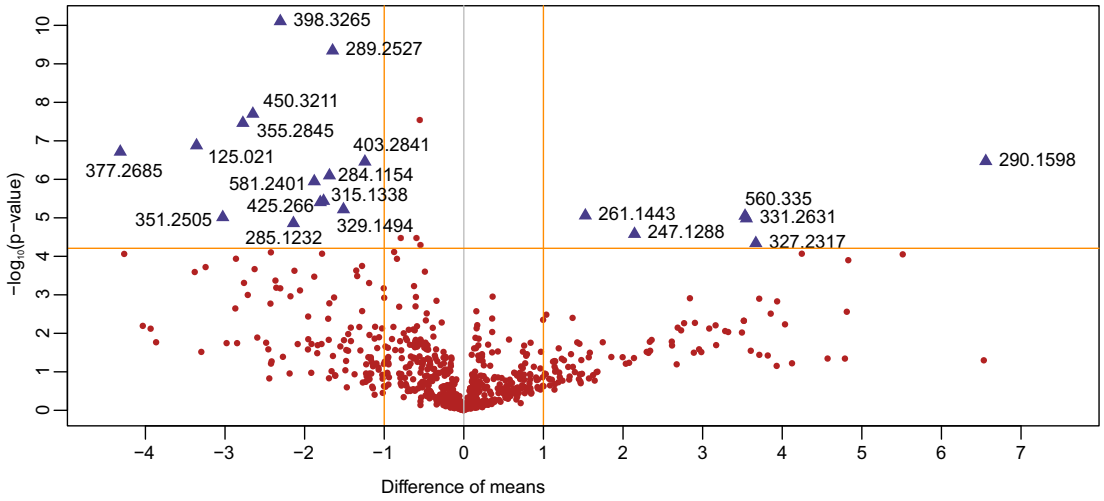

Figure 5: Traditional volcano plot with Bonferroni corrected level of significance resulting from parametric t-test for the clr represented HMGCLD data set. Variables depicted in blue with labels are evaluated as significant by employing the most frequent thresholds, abs(difference of means) = 1 and p-value = (Bonferroni-corrected α), which are highlighted in orange.

m/z of 262.1649 and 290.1598 (Santarelli et al., 2013). However, by applying the traditional parametric univariate analysis (i.e., parametric t-test) and the double-filtering in a volcano plot, only the second biomarker was identified among significant metabolites (Fig. 5). This might be connected to the crucially low number of patients for the analysis involving such a big amount of variables in addition to the fact that the feature 262.1649 was evaluated as important by the logarithm of fold change, however, it did not pass the Bonferroni correction. With this result at hand, the biochemist could easily miss an important biomarker, focusing instead on a group of other (possibly biologically non-significant) metabolites in further steps of the multivariate analysis and feature identification.

When employing the Bayesian approach to the data analysis, both these biomarkers were readily found in the biggest HDI distance levels whilst the compound with m/z = 290.1598 was identified even as the most significant one of the entire graph (Fig. 4). The other metabolites featuring in the five highest HDI levels together with the discussed ones, namely 145.0495 and 303.2318 elevated in patient samples, and 377.2685 elevated in controls, are also meaningful from the biological point of view. Whilst the latter ones are a fatty acid and a monoacylglycerol with a direct connection to pathobiochemistry of the disease, 145.0495

is a novel plasma biomarker published in Václavík et al. (2020). It carries a potential to eventually include HMGCLD in the newborn screening procedure. To conclude, the proposed approach to the volcano plot construction proved to give more accurate and meaningful results for the HMGCLD data in concordance with the published research.

3.3.2 Medium-chain acyl-CoA dehydrogenase deficiency

In a study published by Najdekr et al. (2015), 25 patient dry blood spots and an equal number of control samples were analyzed by untargeted metabolomics approach with reverse-phase liquid chromatography coupled to orbital ion-trap high-resolution mass spectrometry in positive mode in the range of m/z $70 - 1,200$. All experimental details are provided in the original article. Data pre-processing was conducted in `R` software (R Core Team, 2022) with `XCMS` (peak finding, zero imputation; Smith et al. (2006); Tautenhahn et al. (2008); Benton et al. (2010)) and `CAMERA` (isotopes and adducts removal; Kuhl et al. (2011)) packages. Similarly to HMGCLD experiment, LOESS correction was ap-

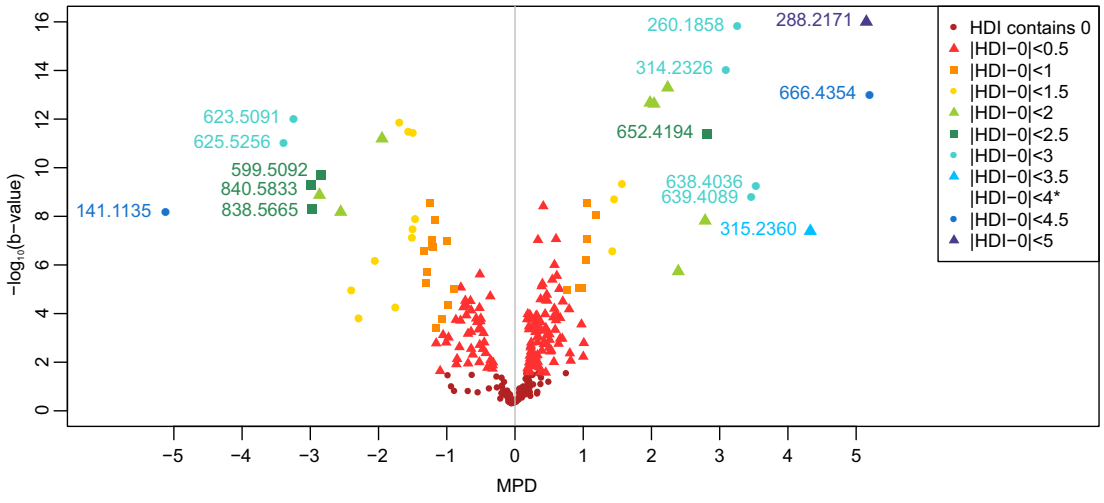

Figure 6: Bayesian volcano plot with HDI distance levels for the MCADD data set. The dark red points stand for the insignificant compounds where the null hypotheses about equality of the parameters μ_{pat} and μ_{con} were not rejected. The potential biomarkers are depicted in blue-green colors, located in the upper right corner for patients and in the upper left corner for controls, respectively.

* – no metabolites detected in this HDI zone

plied (Sumner et al., 2007) and features with a coefficient of variation of quality control samples higher than 30% were excluded from following data curation. The resulting total amount of unique features was 273 where the low number is given by the peak-picking method and by the limited size of dry blood spot samples (analyzed sample corresponds to less than 2 μl of blood).

The analysis of HMGCLD has shown that the traditional and Bayesian approaches must not necessarily lead to similar results and provided some reasoning, why the Bayesian t-test seems to be preferable. Therefore, for next data set, the focus is on the Bayesian approach (Fig. 6; Table 2). Due to the nature of the data, i.e., highly discriminated groups of patients and controls, the t-distributed pdf–based b-values had to be used here. Interestingly, for the MCADD data, solely the b-values provided quite similar results to those based on complex multivariate analysis tools including an S-plot from the orthogonal partial least squares – discriminant analysis (Najdekr et al., 2015). The best biomarker candidates identified by b-values are shown in Table 2 (together with the outcome yielded by MPD). Namely the four known biomarkers (octanoylcarnitine with $m/z = 288.2172$, hexanoylcarnitine with $m/z = 260.1859$, decanoylcarnitine with $m/z = 316.2484$, and decenoylcarnitine with $m/z = 314.2327$) and some oxidative lipids (PAzPC with $m/z = 666.4354$, PC(24:0(COOH)) with $m/z = 652.4194$, and PC(23:0(COOH) with $m/z = 638.4037$) were identified.

Table 2: Ten best biomarkers for patients suffering from MCADD according to MPD and b-value, respectively.

Markers according to	m/z				
MPD	260.1859	288.2172	610.3770	314.2327	652.4194
	638.4037	596.3614	639.4089	666.4354	315.2361
b-value	288.2172	666.4354	791.5634	260.1859	316.2484
	652.4194	772.5488	314.2327	829.6804	638.4037

For the field of rare metabolic diseases, the MCADD data set was processed on a relatively high number of samples in two size-balanced groups. This allowed to carry out simulations (at least up to a certain extent) by considering a loss of samples and a systematic error during measurement. Both issues are discussed in the Chapter 3.4.

3.3.3 Practical aspects of analysis

From practical point of view, researcher should be aware that the proposed method (i.e., compositional approach coupled with the Bayesian model) does not improve the situation with zeros nor the non-linear behavior of analytes above upper limit of quantification and in this way it behaves the same as other statistical tools commonly used for the purpose of metabolomic experiments. Similarly to "standard" metabolomic approach, also here the initial raw data should be pre-processed in order to deal with experimental drift (e.g., by LOESS) and imputation of missing values.

There are generally two types of missing values that might occur in metabolomic data, values under detection limit producing so called rounded zeros and missing values in one statistical class resulting in left-censored data (e.g., for a genetic knock out situation), respectively. In case any of these are present in the measurements, they need to be handled before assigning the coordinate representation to the data. Several approaches to imputation of missing values and rounded zeros in compositional data already exist, for details see e.g., Palarea-Albaladejo and Martin-Fernandez (2015); Templ et al. (2016).

In biological samples differing widely in dilution (physiologically most dominantly observed in urine samples), ratio to a compound representing "concentration" of urine by kidneys (e.g., creatinine) is used to make results clinically comparable. In principle, creatinine could be seen as a substitute for total urinary metabolic content (due to its stable production over time) and thus, the introduced tool seems to be analogous in such situations due to its compositional basis, although it induces also some methodological caveats (the resulting coordinates are oblique). Generally, clr coefficients provide an elegant and due to isometry with the Aitchison geometry also a theoretically reliable way to overcome the problematics of normalization and scaling.

3.4 Simulations

Two simulations were performed to compare results of the proposed method with traditional approaches in case of a loss of samples, where the analysis was carried out on repeatedly randomly chosen half of the samples in both groups, and

for a systematical error occurring during measurements. In the first simulation, the Bayesian t-test was compared with the parametric t-test and the Wilcoxon rank-sum test, whilst the latter simulation evaluated already just the Bayesian method when using different types of data normalization. Due to long computation time of a single run of the multiple hypotheses testing (approx. 4.5 hours of parallel computing on a Windows 10 Home machine equipped with Intel Core i5-7200U, 2.5GHz, 8GB RAM), both simulations were repeated just 20 times. Since the proposed final evaluation by HDI distance intervals is not reasonably comparable with the traditional methods and since for the MCADD data set the b-values produced very good results just by themselves, we decided to use those for the comparisons.

In each step of the simulation, top 20 candidates for biomarkers according to b-value (or p-value for the traditional techniques) were chosen to be represented in the final volcano plots. The fluctuation on both x- and y-axes was then captured using boxplots. The aim of this procedure was to compare the stability of the results for each particular method.

3.4.1 Loss of samples

Results of the first simulation are visualized in Fig. 7 and Fig. 8 with boxplots displaying variation on y-axis and x-axis, respectively. Even though no perspicuous differences are observed, the Bayesian approach fluctuates slightly less with respect to both axes than the traditional parametric approach. For Wilcoxon rank-sum tests, which may seem more stable regarding the y-axis, the results are given by the range of possible p-values; in general, the problem is the same as in the case of t-tests. What is maybe even more important than the fluctuations, the Bayesian approach identifies less different biomarker candidates during distinct simulation steps (chemically unknown compounds with m/z of e.g., 364.2645, 677.5589, 524.3714, 785.6532, and 812.5479 were identified only by the traditional approaches). Overall, it can be concluded that the Bayesian volcano plot is potentially able to preserve more stable results in a situation when some samples are lost, e.g., due to contamination of biological samples.

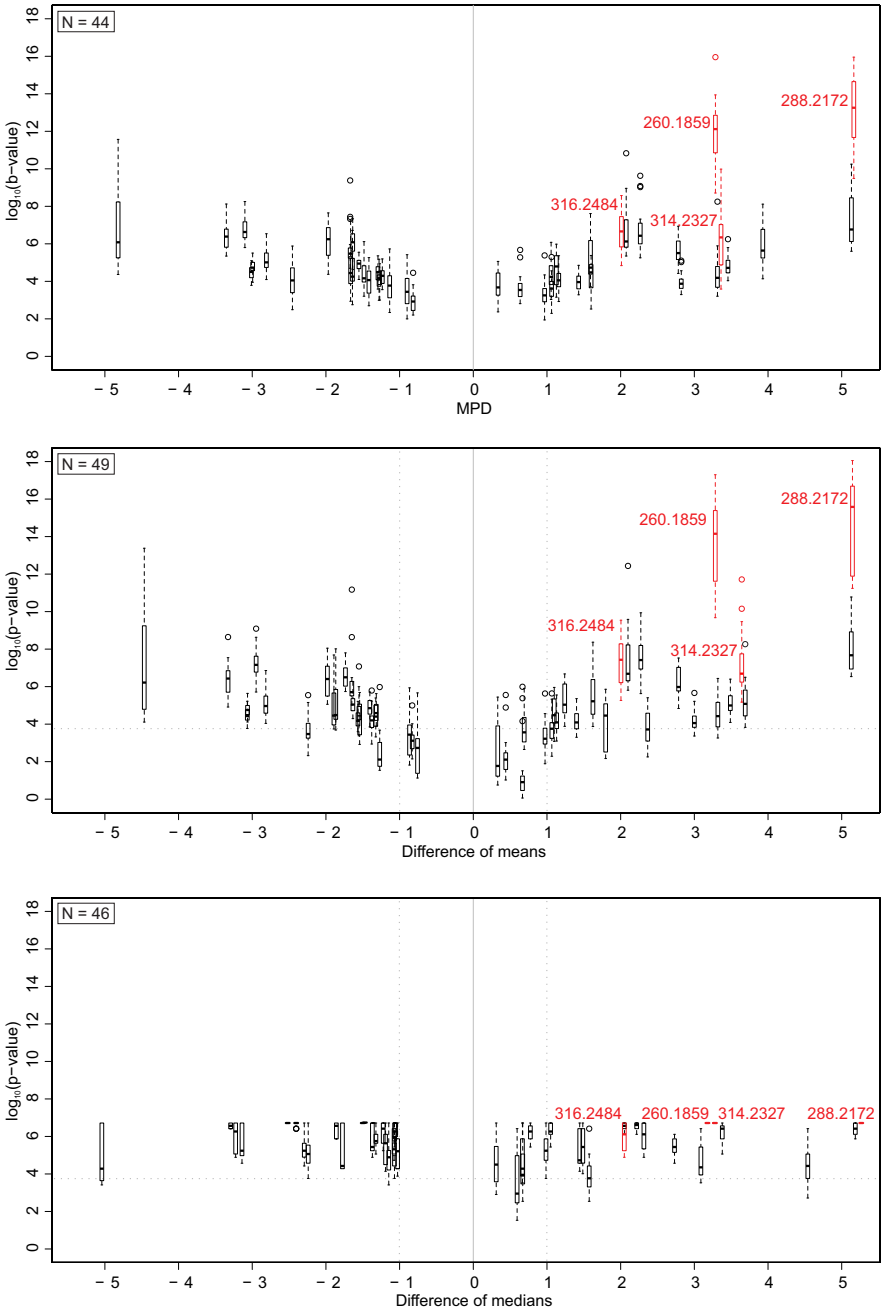

Figure 7: The fluctuation on y-axis during the simulation of loss of samples. Only the candidates that were included among the top 20 candidates at least once (i.e., in one run of the simulation) are depicted in the figure with their total counts in the left upper corners of the graphs. The results from the Bayesian t-test, parametric t-test, and Wilcoxon rank-sum test, respectively, are plotted on the top, in the middle, and at the bottom of the figure, respectively. The four known biomarkers of the disease are shown in red. For the traditional approaches, significance thresholds are depicted by dotted lines.

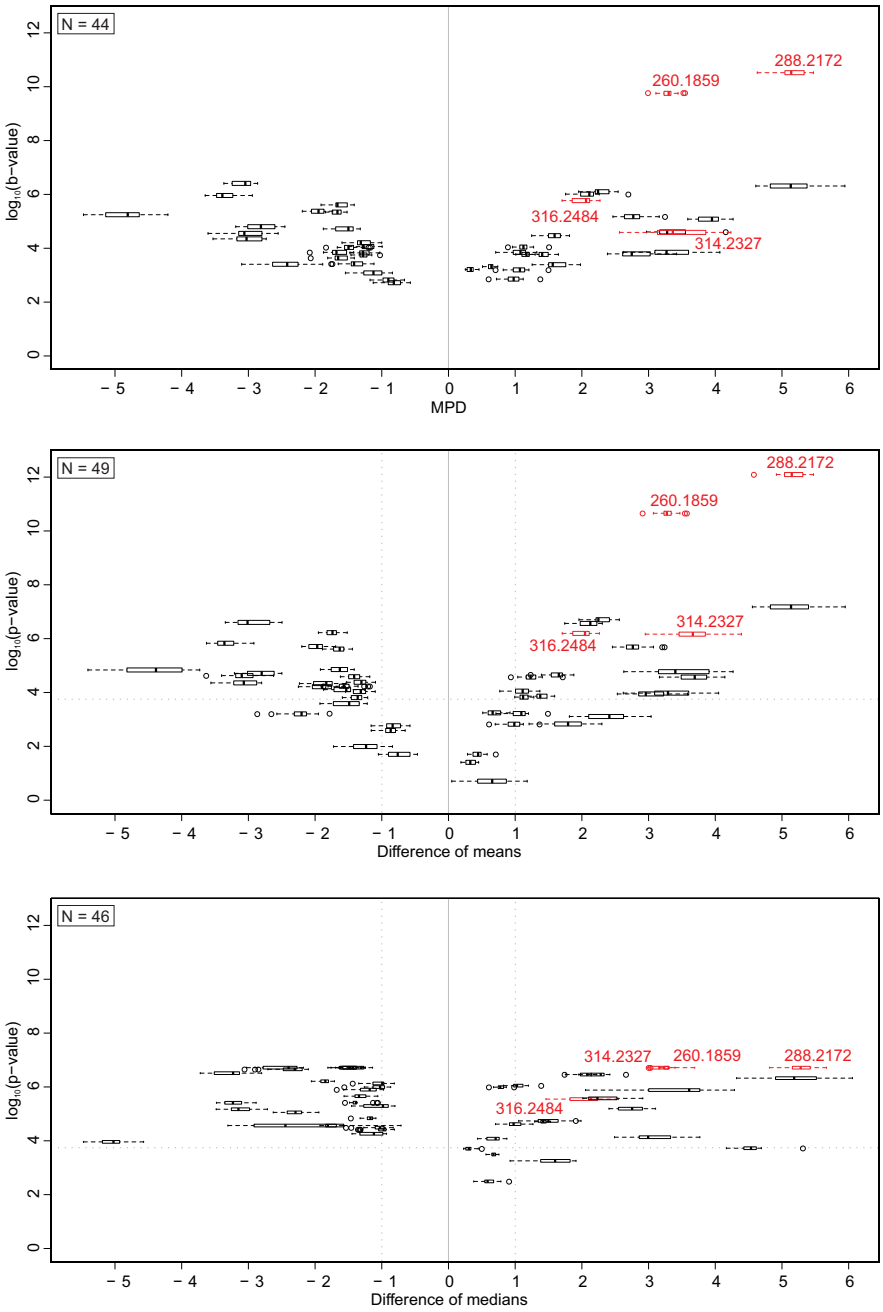

Figure 8: The fluctuation on x-axis during the simulation of loss of samples. Only the candidates that were included among the top 20 candidates at least once (i.e., in one run of the simulation) are depicted in the figure with their total counts in the left upper corners of the graphs. The results from the Bayesian t-test, parametric t-test, and Wilcoxon rank-sum test, respectively, are plotted on the top, in the middle, and at the bottom of the figure, respectively. The four known biomarkers of the disease are shown in red. For the traditional approaches, significance thresholds are depicted by dotted lines.

3.4.2 Systematic error during measurement

The second simulation was constructed to mimic for all sources of technical errors during measurement which are not possible to treat by running quality control samples in order to show the importance of considering a proper data transformation and (or) normalization. In each step of the simulation, the analysis was conducted on samples with a randomly chosen 10 to 30 percent increment. In detail, the original samples were randomly multiplied by 1.1 to 1.3 to reflect technical variability of the measurement, e.g., unstable injection of the samples.

The performance of the Bayesian method accompanied by clr coefficients was compared with the results based on other popular data representations, namely probabilistic quotient normalization (PQN) (Dieterle et al., 2006) and a simple transformation by decadic logarithm. These transformations were applied as suggested in the literature, namely the PQN without any further scaling which is recommended for all univariate methods (Di Guida et al., 2016).

The importance of a particular choice of transformation respecting the relative nature of spectrometric measurements can be seen already in Fig. 9 where

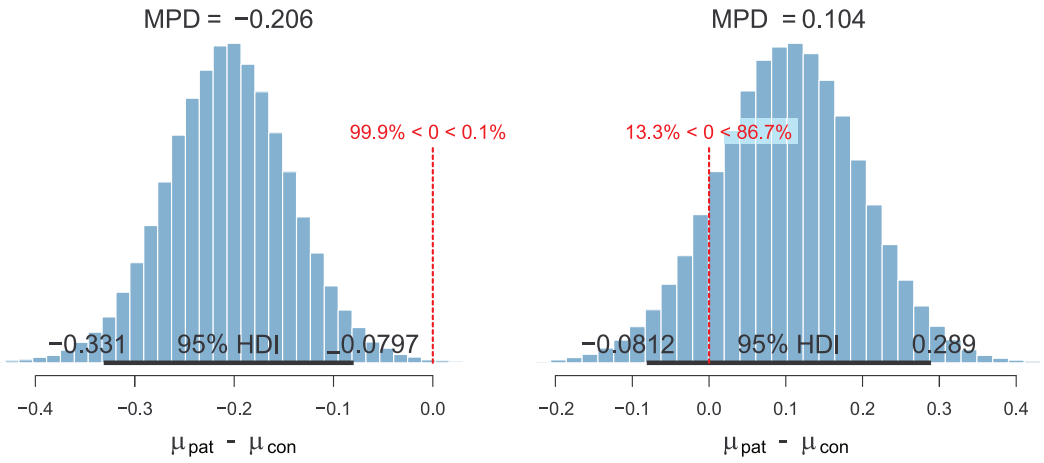

Figure 9: Comparison of a compositional and non-compositional approach. Differences between MCMC generated posteriors of μ_{pat} and μ_{con} are depicted in blue, where their respective parts in negative and positive values are given by the percentages in red. The results of the Bayesian t-test performed on absolute (only pre-processed) values of the metabolite and on the relative values expressed in clr coefficients, respectively, are depicted on the left and right side of the figure, respectively.

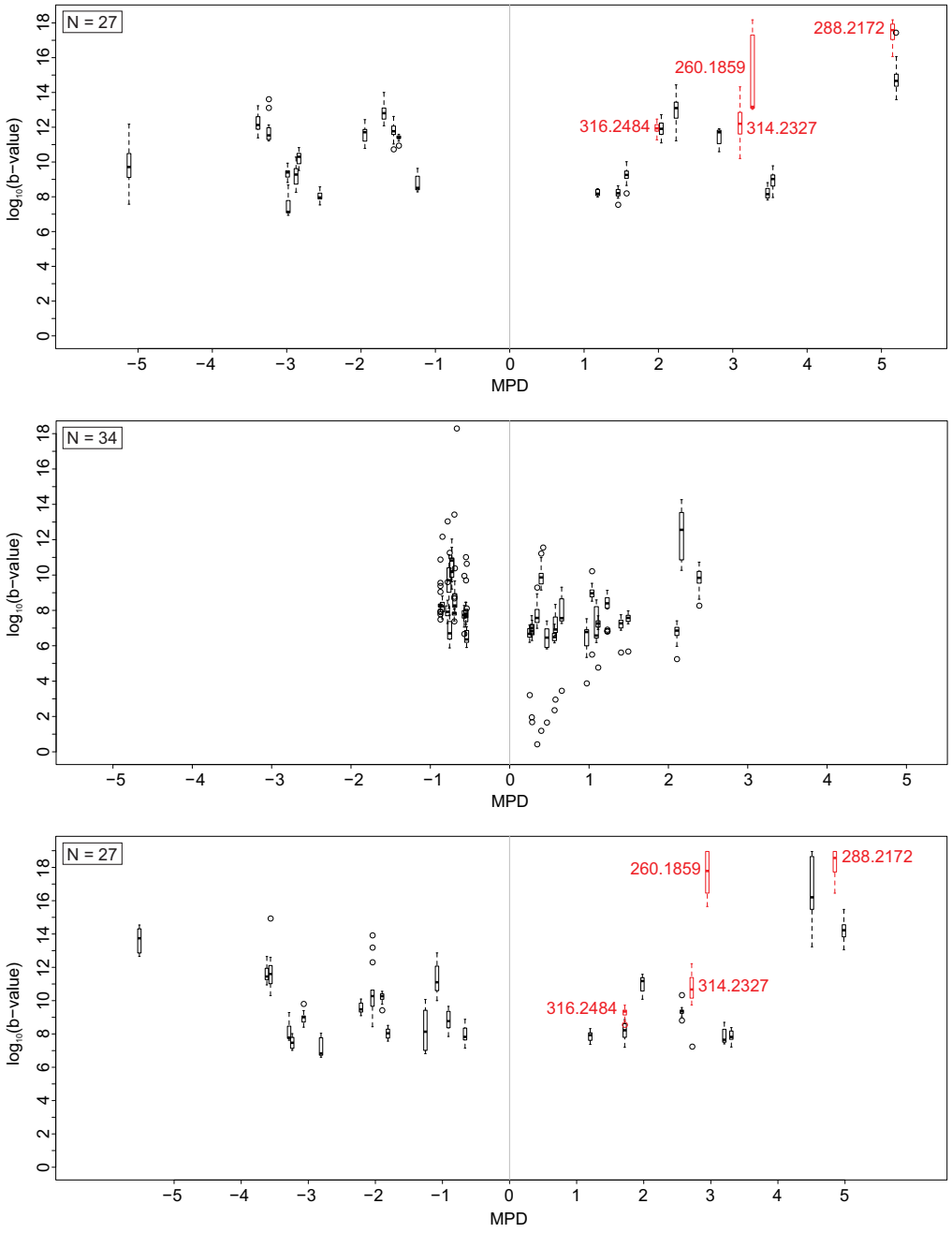

Figure 10: The fluctuation on y-axis for the simulation of systematic error occurrence. Only the candidates that were included among the top 20 candidates at least once (i.e., in one run of the simulation) are depicted in the figure with their total counts in the left upper corners of the graphs. The results from the Bayesian t-test using clr coefficients, PQN transformation and a decadic logarithm of the data, respectively, are plotted on the top, in the middle, and at the bottom of the figure, respectively. The four known biomarkers of the disease are shown in red (where detected among significant).

a single hypothesis test is evaluated based on clr transformed data and raw data, respectively. The conclusion can differ not only in rejecting or not rejecting the hypothesis but even in the direction of the evidence against the initial (null) hypothesis when improper data transformation is used.

The results of the simulation (in Fig. 10 with boxplots illustrating the fluctuations of b-values on y-axis of the Bayesian volcano plot) suggest just subtle differences between clr coefficients and logarithmization in terms of sensitivity to systematic changes of raw data values. Nevertheless, there is quite a notable disagreement when the results are compared with those coming from PQN. The latter normalization is inconsistent, producing contrary to the other transformations substantially more distinct biomarker candidates in the individual steps of the simulation. Moreover, none of these biomarker candidates are true biomarkers (i.e., medium-chain acylcarnitines); the majority of them belongs to lipids. Besides that, this approach generates quite inconsistent b-values. On the other hand, the compositional approach remains persistent and identifies most biomarkers in accordance with published results (Najdekr et al., 2015) even with the incremented data.

4 PLS-DA predictive modeling on selective pivot logratio coordinates and its application in metabolomics

A common issue with metabolomic data in practice, and actually with any omics data in general, is the existence of the so-called size effect. This is related to unavoidable variation in volumes and/or concentrations of the biological material that is processed from one sample to another (Filzmoser and Walczak, 2014) as already described also in the Chapter 3. To downplay such effect in the multivariate tasks of biomarker identification, it is desirable that each original variable (compositional part) is processed in terms of (log)ratios, and preferably so that it gets associated with one specific logratio coordinate. One option would be to apply OPCs, however, previous literature has noted some drawbacks of this strategy that particularly apply to the high-dimensional data case. Namely, by considering the entire collection of pairwise logratios aggregated into a single OPC, it is likely to mix information from completely different processes which leads to confusing insights. In the past, both OPCs and clr coefficients have thus been equally discouraged for classification problems, where usually just some subset(s) of pairwise logratios are responsible for the differences between groups in a given component (Filzmoser and Walczak, 2014; Filzmoser et al., 2018). Consequently, OPCs turned out to be a potential source of false positive (and equally false negative) results. As an alternative to aggregate all pairwise logratios with a component of interest, there have been some attempts to use more robust procedures to extract information from pairwise logratios in order to reveal possible biomarkers. This is the case of Walach et al. (2017), Malyjurek et al. (2019), or alternatively Dieterle et al. (2006), employing the PQN transformation. Nevertheless, using the median of compositional parts in the PQN formulation leads to quite substantial loss of information.

Since in Chapter 2.2 it was shown that the disadvantages of OPCs can be lessened by implementing WPCs (Hron et al., 2017; Štefelová et al., 2021), a novel approach called *selective pivot coordinates* (SPCs) will be introduced as a compromise solution in this chapter. SPCs can be seen as a variant of both the OPC and WPC approaches targeted to binary classification problems. Unlike OPCs,

the new SPCs only aggregate some of the pairwise logratios associated to a given part into the logratio coordinate. In general, the SPCs are designed to facilitate the identification of biomarkers that exhibit different behavior in two groups of samples, typically referring to diseased and control samples. The details of the proposal are thoroughly discussed in Chapter 4.1 and the novel coordinate system is embedded within a partial least squares – discriminant analysis (PLS-DA) model in Chapter 4.2. Chapter 4.3 presents a comparative study of its performance using simulation, while Chapter 4.4 further illustrates the advantages of the new approach using real-world metabolomic data.

4.1 Selective pivot coordinates

In the context of two-group classification involving CoDa, the idea that motivates the development of SPCs is to have logratio coordinates that represent relevant relative information about x_l, but aggregate only informative pairwise logratios including x_l in the first coordinate. That is, given that each pairwise logratio involves two distinct compositional parts, the aim is to include into an SPC, denoted by ${}_{(l)}s$, only those that agree with what the majority of logratios with x_l suggest about its ability to distinguish between the two groups of observations. Namely, in a biomedical setting, having two groups pat (patient) and con (control), a compositional part should be identified as a biomarker candidate if most pairwise logratios involving that part are significantly higher in one group than in the other. Let us discuss some possible scenarios in this setting that contribute to outline the rationale underlying the definition of SPCs:

i) A part x_i is a strongly positive biomarker in group pat and x_j is some other biomarker increased in this group but with a weaker discriminating effect. Then for most logratios $\ln(x_i/x_d),\ d = 1, \ldots, D,\ d \neq i$, it can be expected that their values will be significantly higher in group pat than in group con. However, the behavior of logratios including x_j will partially differ. Thus, values of $\ln(x_j/x_i)$ will be generally lower in group pat, and similar behavior might be observed for some other logratios with x_j in the numerator whenever a stronger biomarker is placed in the denominator. Nevertheless, they should be only a minority that deviates from the prevailing trend. By excluding these from the aggregation in the SPC we expect to increase the

sensitivity of the classifier, since the chance of x_j leading to a false negative in subsequent statistical analysis should be smaller when compared to OPCs that aggregate all logratios. An analogous situation with biomarkers decreased in group pat can be considered.

ii) A part x_i is a strong biomarker increased in group pat and x_j is a strong biomarker decreased in group pat. Thus, it is likely that $\ln(x_i/x_j)$ will have a strong discriminating effect, and similarly for other logratios involving two biomarkers with discriminating effects in opposite directions. These might be flagged as outliers among the logratios involving x_i, respectively x_j, however these are deviating logratios that should be preserved.

iii) A part x_i is not a biomarker. Therefore, it can be anticipated that the logratios $\ln(x_i/x_d),\ d = 1, \ldots, D,\ d \neq i$, will not exhibit a significant difference between groups, except for the case where x_d be a potential biomarker. In this case, excluding deviating logratios should reduce the chances of x_i leading to a false positive in subsequent statistical analysis, and thus increase the specificity of the classifier.

Given a compositional data matrix consisting of N observations from two different groups, we propose to use the ordinary Welch's t-statistic (Welch, 1947) to determine the least relevant logratios. Denoting ${}_{(l)}\boldsymbol{T} = ({}_{(l)}T_1, \ldots, {}_{(l)}T_{l-1}, {}_{(l)}T_{l+1}, \ldots, {}_{(l)}T_D)$ the set of such t-statistics corresponding to logratios $\left(\ln\frac{x_l}{x_1}, \ldots, \ln\frac{x_l}{x_{l-1}}, \ln\frac{x_l}{x_{l+1}}, \ldots, \ln\frac{x_l}{x_D}\right)$, the criterion is to exclude those logratios for which the statistic ${}_{(l)}T_d,\ d = 1, \ldots, D,\ d \neq l$ lays outside the interval $\left[{}_{(l)}\theta_1;\ {}_{(l)}\theta_2\right]$. These boundaries are computed as

$${}_{(l)}\theta_1 = \begin{cases} -\infty, & \text{if } q\left({}_{(l)}\boldsymbol{T}; 1-\xi\right) < t_{N-2}(0.025) \\ \text{med}\left({}_{(l)}\boldsymbol{T}\right) - 2\text{Q}_n\left({}_{(l)}\boldsymbol{T}\right), & \text{otherwise,} \end{cases}$$

and

$${}_{(l)}\theta_2 = \begin{cases} \infty, & \text{if } q\left({}_{(l)}\boldsymbol{T}; \xi\right) > t_{N-2}(0.975) \\ \text{med}\left({}_{(l)}\boldsymbol{T}\right) + 2\text{Q}_n\left({}_{(l)}\boldsymbol{T}\right), & \text{otherwise,} \end{cases}$$

where $q\left({}_{(l)}\boldsymbol{T}; \alpha\right)$ is the α-quantile of ${}_{(l)}\boldsymbol{T}$, $\text{med}\left({}_{(l)}\boldsymbol{T}\right) = q\left({}_{(l)}\boldsymbol{T}; 0.5\right)$ and $t_{N-2}(\alpha)$ is the α-quantile of the Student's t-distribution with $N-2$ degrees of freedom

(the parameter ξ is set to 0.1 by default). Moreover, Q_n stands for the robust scale estimator of Rousseeuw and Croux (1993), i.e., $\mathrm{Q}_n\left({}_{(l)}\boldsymbol{T}\right)$ is given by about the first quartile of the absolute differences $\{|{}_{(l)}T_c - {}_{(l)}T_d|,\ 1 \leq c < d \leq D,\ c, d \neq l\}$. The interval for exclusion then results to be $\left[\mathrm{med}\left({}_{(l)}\boldsymbol{T}\right) \pm 2\mathrm{Q}_n\left({}_{(l)}\boldsymbol{T}\right)\right]$, unless more than 90% of the t-statistic values are roughly either lower than -2 or higher than 2. Where this latter happens, only the upper (resp. lower) cut-off values are used. Note that this additional condition aims to ensure that logratios involving two strong biomarkers with discriminating effects pulling in opposite directions are not excluded from the aggregation (scenario ii) above). A higher value of ξ can be chosen if this undesirable effect is still apparent (as can be seen when visualizing the selected logratios in Chapter 4.4, Fig. 13 and 16).

For the following, let us denote the number of selected logratios including x_l as ${}_{(l)}M$, the parts in the denominator of the selected logratios as ${}_{(l)}x_1^+, \ldots, {}_{(l)}x_{{}_{(l)}M}^+$, and the remaining parts as ${}_{(l)}x_1^-, \ldots, {}_{(l)}x_{D-1-{}_{(l)}M}^-$. To obtain ${}_{(l)}s,\ l = 1, \ldots, D$, the original composition $\boldsymbol{x}$ needs to be rearranged as ${}_{(l)}\boldsymbol{x} = \left({}_{(l)}x_1^-, \ldots, {}_{(l)}x_{D-1-{}_{(l)}M}^-, x_l, {}_{(l)}x_1^+, \ldots, {}_{(l)}x_{{}_{(l)}M}^+\right)$.

Then, an OPC system ${}_{(l)}\boldsymbol{z} = \left({}_{(l)}z_1, \ldots, {}_{(l)}z_{D-1}\right)$ is set up for ${}_{(l)}\boldsymbol{x}$. To define SPCs, the *pivoting* coordinate is no longer the first one but the one at the $(D - {}_{(l)}M)$-th position, denoted by ${}_{(l)}z_{D-{}_{(l)}M}$. Accordingly, the SPC of interest, ${}_{(l)}s$, is obtained as

$$
\begin{aligned}
{}_{(l)}s = {}_{(l)}\, z_{D-{}_{(l)}M} &= \sqrt{\frac{{}_{(l)}M}{{}_{(l)}M+1}} \ln \frac{x_l}{\sqrt[{}_{(l)}M]{\prod_{k=1}^{{}_{(l)}M} {}_{(l)}x_k^+}} \\
&= \frac{1}{\sqrt{({}_{(l)}M+1)\cdot {}_{(l)}M}} \left(\ln \frac{x_l}{{}_{(l)}x_1^+} + \cdots + \ln \frac{x_l}{{}_{(l)}x_{{}_{(l)}M}^+} \right), \quad l = 1, \ldots, D. \qquad (21)
\end{aligned}
$$

The proposed SPCs can also be linked to other alternatives such as WPCs (Hron et al., 2017) given by the expression (11). Note that SPCs are a special case of WPCs where weights of either 1 or 0 are assigned to each logratio involving x_l, depending on whether it is included in the aggregation or not, respectively.

Consequently, ${}_{(l)}s$ can be written as

$$ {}_{(l)}s = \sqrt{\frac{{}_{(l)}M}{{}_{(l)}M+1}} \ln \frac{x_l}{\sqrt[{}_{(l)}M]{\prod_{\substack{d=1 \\ d\neq l}}^{D} (x_d)^{{}_{(l)}\gamma_d}}}, \quad l = 1, \ldots, D, $$

with weights given by

$$ {}_{(l)}\gamma_d = \begin{cases} 1, & \text{if } {}_{(l)}T_d \in \left[{}_{(l)}\theta_1;\ {}_{(l)}\theta_2\right] \\ 0, & \text{otherwise.} \end{cases} $$

Moreover, although WPCs with general non-negative weights as developed for PLS regression (Štefelová et al., 2021) could be extended for the current classification case, the latter is actually more easily handled (no specific weighted coordinate system is required) by reformulating the OPC approach as described above.

4.2 PLS-DA on selective pivot coordinates

Building on the OPC–based approach introduced in Kalivodová et al. (2015), partial least squares discriminant analysis (PLS-DA) through SPCs is used here for the actual identification of biomarker candidates. Thus, D models of the form

$$ Y = \beta_0 + {}_{(l)}\beta_1 \cdot {}_{(l)}z_1 + \ldots + {}_{(l)}\beta_{D-1} \cdot {}_{(l)}z_{D-1} + \varepsilon, \quad l = 1, \ldots, D, \tag{22} $$

are considered, where Y is a binary response representing each of the two groups, the explanatory variables ${}_{(l)}z_1 \ldots, {}_{(l)}z_{D-1}$ are logratio coordinates from the l-th SPC system, $\beta_0, {}_{(l)}\beta_1, \ldots, {}_{(l)}\beta_{D-1}$ are unknown model coefficients and ε is the usual random error term. Note that, unlike with PLS-DA based on OPCs where the procedure can be computationally simplified by fitting one model in clr coefficients and then take advantage of their direct relationship with OPCs (see expression (9) between $z_1^{(l)}$ and $\mathrm{clr}(\boldsymbol{x})_l$), SPCs require the successive models to be fitted individually. Before fitting the PLS model, the data are mean centered so that the intercept β_0 can be excluded from further considerations. The optimal number of PLS components is chosen here based on a randomization test approach (van der Voet, 1994). From each model fit, for $l = 1, \ldots, D$, the estimate

${}_{(l)}\hat{\beta}_{D-{}_{(l)}M}$ associated with the SPC ${}_{(l)}s$ is extracted, and statistical significance is determined by bootstrap-based significance testing on the standardized PLS model coefficients (Kalivodová et al., 2015). The resulting p-values are adjusted using the Benjamini and Hochberg's method (Benjamini and Hochberg, 1995) to control for false discovery rate in multiple testing.

4.3 Comparative simulation study

The performance of PLS-DA applied on SPCs is assessed here by simulation in comparison to previous alternative approaches. Namely, PLS-DA after lnPQN data normalization as a popular reference method (i.e., PQN in combination with log-transformation of the explanatory variables as advocated for multivariate analysis in metabolomics by Di Guida et al. (2016)) and PLS-DA on OPCs as another compositional approach previously proposed. The design of the simulation study follows the setup of Filzmoser and Walczak (2014), aiming to mimic typical high-throughput data sets affected by size effect. Accordingly, data matrices $\mathbf{X} = (x_{nd})$ of size $N \times D$ are generated so that the first $N/2$ rows correspond to samples from group pat while the remaining correspond to group con, with the first R columns representing biomarkers. Each entry is obtained as

$$x_{nd} = (1 - k_n) \cdot \left(\frac{u_d}{v_d} + a_{nd} + f_{nd} \right) \cdot v_d \cdot e^{g_{nd}} + h_{nd}, \quad n = 1, \ldots, N, \; d = 1, \ldots, D,$$

where k_n represents the size effect sampled from a normal distribution $\mathcal{N}(0, 0.3^2)$ and u_d/v_d represents a component concentration with signal abundance u_d sampled from a uniform distribution $\mathcal{U}(1, 100)$ and component absorptivity v_d sampled from a uniform distribution $\mathcal{U}(1, 10)$. Furthermore, a_{nd} determines the higher signal of biomarkers and is defined as

$$a_{nd} = \begin{cases} A, & \text{if } n \leq N/2 \text{ and } d \leq R \\ 0, & \text{otherwise,} \end{cases}$$

and f_{nd}, g_{nd}, and h_{nd} representing different kinds of noise (biological, multiplicative, and background, respectively) are sampled from normal distributions $\mathcal{N}(0, \sigma_f^2)$, $\mathcal{N}(0, \sigma_g^2)$, and $\mathcal{N}(0, 0.05^2)$, respectively. Varying parameters A, σ_f, and σ_g as indicated in Table 3 results in eight different settings. Following the reference design, the number of observations is set to $N = 40$ and the number of variables

(components) to $D = 500$. More options are considered though for the ratio of biomarkers R to the rest of metabolites $(D - R)$, with $R \in \{20, 50, 100\}$. Hence, 24 different simulated scenarios are examined in total, executing 100 simulation runs within each configuration.

As mentioned above, three approaches to PLS-DA–based biomarker identification are compared:

1. **lnPQN**: PLS-DA is applied on data normalized by lnPQN transformation:

   $$x_{nd}^{\text{lnPQN}} = \ln \frac{x_{nd}}{\omega_n}, \quad \omega_n = \text{med}\left(\frac{x_{n1}}{\text{med}\,(x_{11}, \ldots, x_{N1})}, \ldots, \frac{x_{nD}}{\text{med}\,(x_{1D}, \ldots, x_{ND})} \right),$$

   where $n = 1, \ldots, N, \ d = 1, \ldots, D$.

2. **OPC**: PLS-DA is applied on data expressed in OPCs (equivalently in clr coefficients).

3. **SPC**: PLS-DA given by (22) is applied on data expressed in SPCs (21) following the new proposal as detailed in Chapter 4.1.

The performance of these approaches is evaluated and compared in terms of sensitivity (i.e., rate of true biomarkers identified as biomarkers) and specificity (i.e., rate of non-biomarkers correctly identified as non-biomarkers). The results are shown graphically in Fig. 11.

It can be observed that generally PLS-DA based on SPCs outperforms its competitors with regard to both sensitivity and specificity. This superior perfor-

Table 3: Parameter settings for comparative simulation study.

Setting	A	σ_f	σ_g
1	1.8	0.8	0.021
2	1.8	0.8	0.007
3	1.0	0.8	0.021
4	1.8	0.2	0.021
5	1.0	0.2	0.007
6	1.0	0.2	0.021
7	1.8	0.2	0.007
8	1.0	0.8	0.007

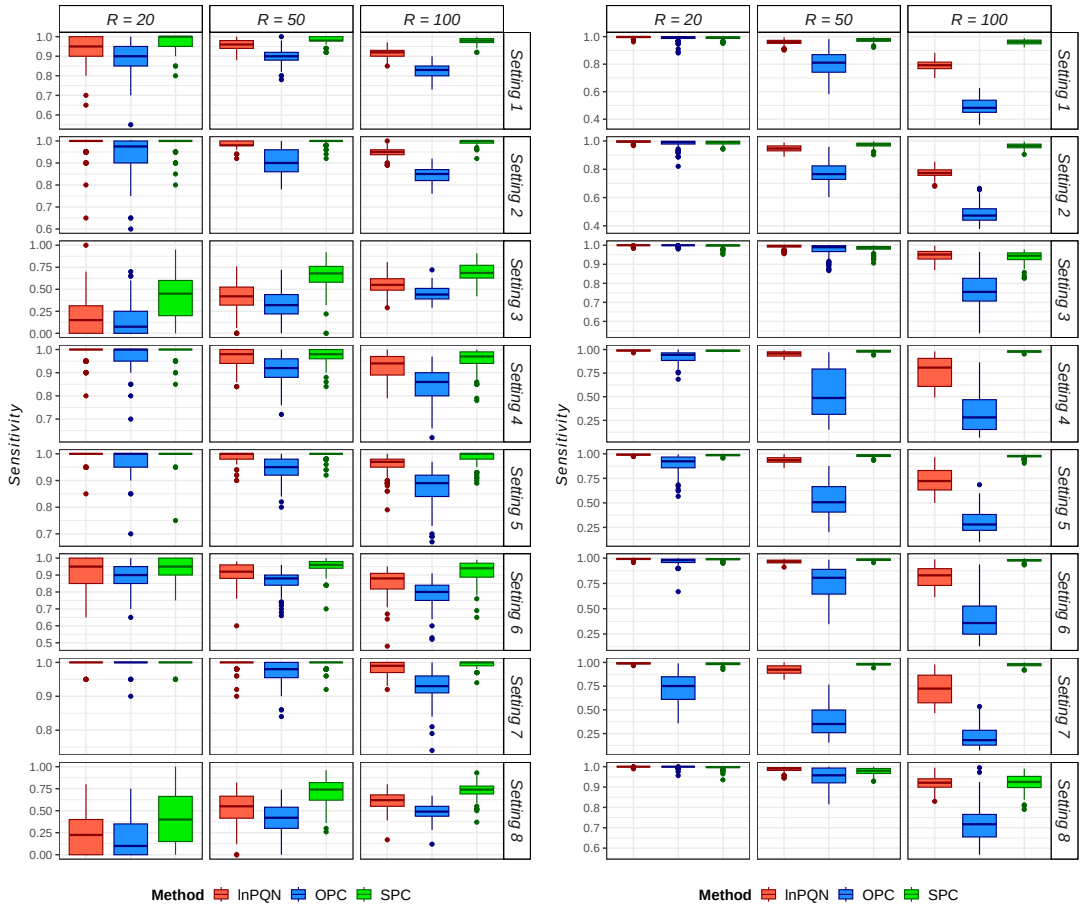

Figure 11: Sensitivity (a) and specificity (b) for different simulation scenarios.

mance is particularly evident in the case with the higher ratio of biomarkers. Thus, when comparing specificity (Fig. 11b), PLS-DA based on OPCs tends to produce too many false positives, which is expected given results from previous studies (Filzmoser and Walczak, 2014). Moreover, when the number of biomarkers is set to $R = 100$, it is apparent that lnPQN also exhibits a poorer performance than SPCs in general, suggesting that PQN tends to oversimplify the data structure here. As to sensitivity (Fig. 11a), the results are influenced by the chosen biomarker ratio only partially; they rather depend on the parameter settings. Notably, in scenarios combining poorer signal and more fluctuating biological noise (i.e., settings 3 and 8, see Table 3), PLS-DA coupled with SPCs clearly surpasses not just the OPC but also the PQN approach.

4.4 Application to metabolomic data

The use of SPCs is illustrated in this section using two different real data sets corresponding to targeted and untargeted metabolomics analyses. Both data sets contain metabolites with higher molecular weight and non-polar properties – lipids, which are studied in the separate field of lipidomics (Gallart-Ayala et al., 2020).

4.4.1 Transgenic rat models with induced tauopathy

The first application concerns data obtained from samples of cerebrospinal fluid (CSF) of transgenic rat models with induced tauopathy. The final data set consisted of $N = 23$ samples: 14 from the TG14 group (14-month-old transgenic rats representing the patient group) and 9 from the TG4 group (4-month-old transgenic rats representing the control group), and $D = 394$ lipids. The abbreviations of the distinguished lipid classes are listed in Table 4 together with their expected (non)biomarker-like behavior based on the previous studies. Further details about data acquisition and pre-processing as well as a short description of the relevant part of the pathobiochemistry of the disease are provided in Appendix A.

Fig. 12 shows Welch's t-statistics for all the individual pairwise logratios, where each row represents the lipid in the numerator. The strongest patient biomarkers should be the ones corresponding to the darkest green color. Biologically, lots of these belong to the classes of CERs, HCERs, LPC(O)s, PC(O)s and SMs as expected. Contrarily, the rows showing mainly light to white color should correspond to non-biomarkers. From the biochemical point of view, this holds for the whole class of DGs, FAs and TGs. Among these, those in a darker brown color are likely to correspond to other patient biomarker candidates that are present in the denominators of the logratios.

Fig. 13 illustrates how our proposal based on SPCs works. Each row indicates which pairwise logratios are and are not included when aggregating into the corresponding SPC. Namely, black stars are used as symbols to indicate those that are not included. It can be seen that the excluded logratios are largely those having a biomarker candidate in the denominator, which is in agreement with the

scenarios discussed in Chapter 4.1 in relation to problematic (deviating) logratios.

Furthermore, as done in the simulation study above, the results obtained by PLS-DA based on SPCs are compared to PLS-DA based on OPCs and lnPQN respectively. In all three cases, a model based on two PLS components was considered adequate based on the randomized test approach. Cross-validated (CV) root mean squared error of prediction (RMSEP) and coefficient of determination (R^2) were comparable: CV RMSEP = 0.11 and CV R^2 = 0.95 for the compositional approaches and CV RMSEP = 0.10 and CV R^2 = 0.96 when using lnPQN. A total of 108 lipids were identified as biomarker candidates using SPCs, whereas they were 156 and 112 using OPCs and lnPQN, respectively. Fig. 14 illustrates the

Table 4: Full names and abbreviations of different classes of lipids present in the transgenic rats data set. Indication of (+) and (0), respectively, is provided in accord with the expectation for them to be detected as biomarkers increased in the TG14 group or as non-biomarkers, respectively, based on the previous studies (Sheikh and Nagai, 2011; Mielke et al., 2014; Ojo et al., 2018; Torretta et al., 2018; Pedersen et al., 2019; Fonteh et al., 2020; Kao et al., 2020) (see Appendix A for more details). Additionaly, (?) marks the cases when the behavior of the entire class is not known.

Full name of the class	Abbreviation	Indication
ceramides	CERs	(+)
diacylglycerols	DGs	(0)
free fatty acids	FAs	(0)
hexosylceramides	HCERs	(+)
lysophosphatidylcholines	LPCs	(+)
plasmanyl/plasmenyl LPCs	LPCOs	(+)
lysophosphatidylethanolamines	LPEs	(0)
plasmanyl/plasmenyl LPEs	LPEOs	(0)
phosphatidylcholines	PCs	(+)
plasmanyl/plasmenyl PCs	PCOs	(+)
phosphatidylethanolamines	PEs	(?)
plasmanyl/plasmenyl PEs	PEOs	(?)
phosphatidylinositols	PIs	(+)
phosphatidylserines	PSs	(?)
sphingomyelins	SMs	(+)
triacylglycerols	TGs	(0)

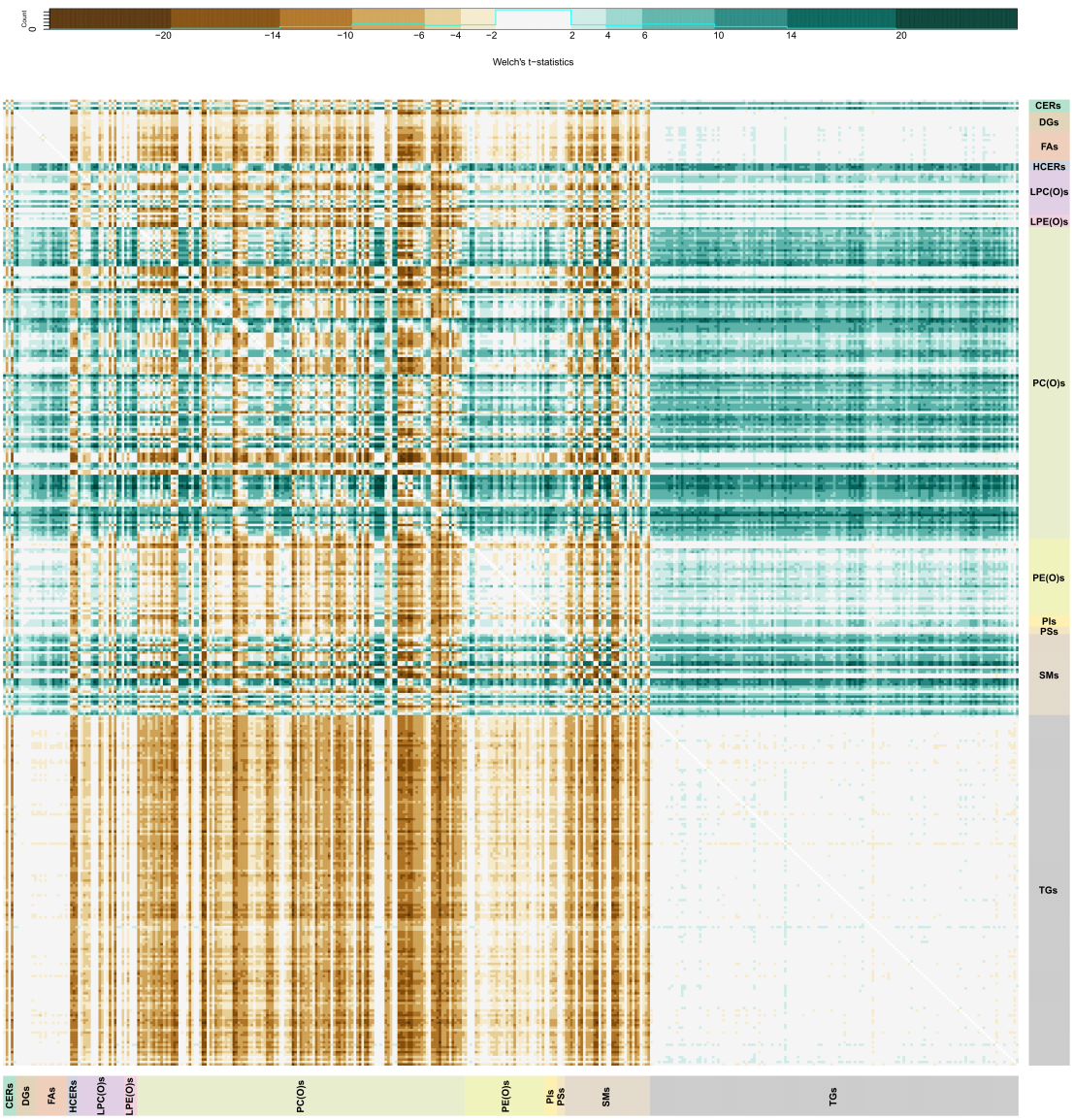

Figure 12: Heatmap of the Welch's t-statistics for pairwise logratios of lipids from the transgenic rats data set. The y- (resp. x-) axis corresponds to the lipid used in the numerator (resp. denominator) of the logratios. Labels on both axes are provided according to clustered classes of lipids given in Table 4.

difference between the two compositional approaches (SPCs and OPCs), while Fig. 15 compares the SPC and lnPQN approaches.

Regarding the comparison of the two compositional approaches, SPCs versus OPCs, all the 17 lipids flagged as biomarker candidates only with SPCs are identified as significantly increased in TG14, whereas those 65 flagged only with

OPCs are identified as significantly decreased in TG14. Fig. 14 confirms that the selective approach to pairwise logratio aggregation indeed helps to i) detect lipids that could be biomarkers with a weaker discriminating effect (as suggested by the

Figure 13: Heatmap of the Welch's t-statistics for pairwise logratios of lipids from the transgenic rats data set illustrating the patterns to choose the logratios for the construction of SPCs. The y- (resp. x-) axis corresponds to the lipid used in the numerator (resp. denominator). Rows and columns are reordered according to their median value of the Welch's t-statistic. Black stars in each row mark logratios not included in the aggregation when constructing the SPC related to the corresponding lipid.

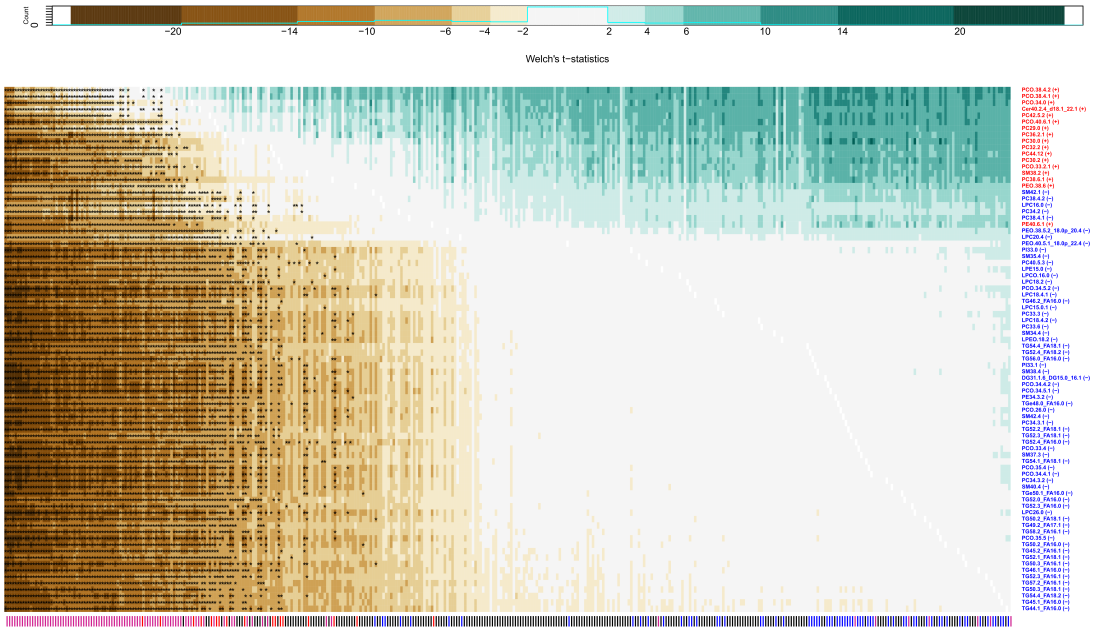

Figure 14: Cut-out from Fig. 13 with the rows limited to the lipids (named at the end of the rows) showing different significance results when using PLS-DA based on either SPCs or OPCs. Biomarkers noted using SPCs and not so using OPCs are colored in red (the opposite case is colored in blue). Lipids identified as biomarkers by both approaches are marked in violet (those identified by none are marked in black). The row labels also denote whether the respective lipid was flagged by the indicated method as a biomarker increased (+) or decreased (−) in the TG14 group.

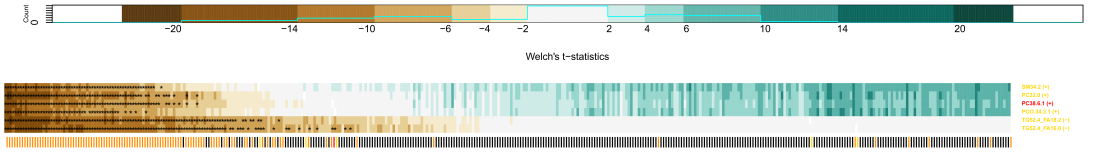

Figure 15: Cut-out from Fig. 13 with rows limited to lipids (named at the end of the rows) showing different significance results when using PLS-DA based on either SPCs or lnPQN. Biomarkers noted using SPCs and not so using lnPQN are colored in red (the opposite case is colored in yellow). Lipids identified as biomarkers by both approaches are marked in orange (those identified by none are marked in black). The row labels also denote whether the respective lipid was flagged by the indicated method as a biomarker increased (+) or decreased (−) in the TG14 group.

profiles in rows marked in red), and ii) to reduce the number of false positives (as suggested by the profiles in rows marked in blue). This finding is consistent with previous studies (see Appendix A), since the majority of the additional potential biomarkers increased in TG14 that are identified using SPCs belong to classes known to be elevated in patients (i.e., CERs, PC(O)s, SMs). However, most of those that were additionally identified when using OPCs (as potential biomarkers *decreased* in TG14) belong to classes of assumed non-biomarkers (i.e., DGs, LPE(O)s, TGs), or even to classes that have been associated with upregulation in patients (i.e. LPC(O)s, PIs, PC(O)s, SMs).

As to SPCs versus lnPQN, both approaches produce similar results from this data set. The one additional potential biomarker increased in TG14 identified using SPCs (belonging to the PCs class), as well as the three additional ones identified using lnPQN (belonging to PC(O)s and SMs classes), are supported by findings in previous studies. Moreover, previous literature suggests that the two additional lipids marked using lnPQN (TGs class) as potential biomarkers decreased in TG14 are most likely false positives (see Appendix A).

4.4.2 SCADD

The second application concerns dry blood spot samples of patients with a hereditary genetic disorder SCADD. The data set consists of $N = 39$ samples: 20 from the patient group and 19 from the control group, and $D = 2011$ features. The abbreviations of some feature classes relevant to the purpose of this thesis

Table 5: Full names and abbreviations of some of the classes of features present in the SCADD data set. Indication of (+), and (0), respectively, is provided in accord with the expectation for them to be detected as potential patient biomarkers and non-biomarkers, respectively, based on the previous studies (Gault et al., 2010; Blom et al., 2011; Nochi et al., 2017) (see Appendix B for more details).

Full name of the class	Abbreviation	Indication
acylated carnitines and other acyl conjugates	aCARs	(+)
polyunsaturated glycerophospholipids	PUFA-PCs	(0)
phosphatidylinositol	PIs	(0)
long-chain sphingomyelin lipids	LC-SMs	(0)

are listed in Table 5 together with their expected (non)biomarker-like behavior. Like in the previous application, further details about the data as well as a short description of the biochemical context are provided in Appendix B.

Figure 16: Heatmap of the Welch's t-statistics for pairwise logratios of features from the SCADD data set illustrating the patterns to choose the logratios for the construction of SPCs. The y- (resp. x-) axis corresponds to the feature used in the numerator (resp. denominator). Rows and columns are reordered according to their median value of Welch's t-statistic. Black stars in each row mark logratios not included in the aggregation when constructing the SPC related to the corresponding feature.

Fig. 16 illustrates which logratios are and are not aggregated into the corresponding SPC. Note that a larger value of the parameter ξ was chosen in this case ($\xi = 1/3$) in order to lessen the undesirable effect depicted by scenario ii) in Chapter 4.1.

Fig. 17 and Fig. 18 show the differences between PLS-DA-based biomarker identification using SPCs instead of OPCs or lnPQN respectively. The randomized test approach used to select the number of PLS components in the PLS-DA model suggested to retain just the first one in all cases (CV RMSEP $= 0.35$ and CV $R^2 = 0.51$). Using SPCs, OPCs and lnPQN, 362, 361 and 351 features were respectively flagged as biomarker candidates. PLS-DA based on SPCs and OPCs differed by 48 features in classifying between groups, whereas such difference was 49 features between SPC and lnPQN. Comparing the results, we can observe that in general the profiles of the features identified solely by using SPCs (rows labeled in red) look more like biomarkers than those identified only by the other methods (rows labeled in blue and yellow respectively), i.e., the discriminating effect is apparent in more logratios with the given part in the numerator.

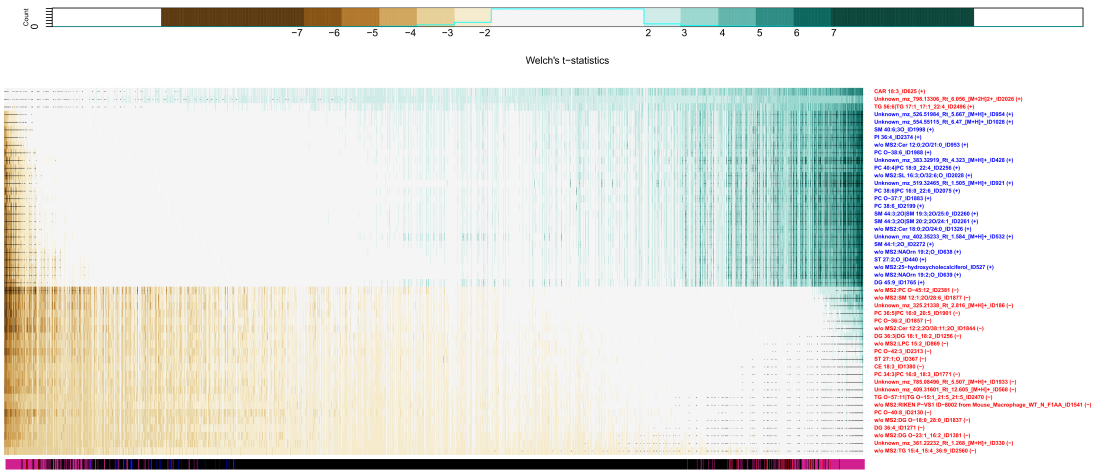

Figure 17: Cut-out from Fig. 16 with the rows limited to the features showing different results in terms of significance when carrying out PLS-DA based on SPCs or OPCs. Features identified as biomarkers by the SPC approach and not by the OPC one are colored in red (the opposite case is colored in blue). Features identified as biomarkers by both approaches are marked in violet (those identified by none are marked in black). The row labels denote whether the respective feature was flagged by the indicated method as a biomarker increased (+) or decreased (−) in the patient group.

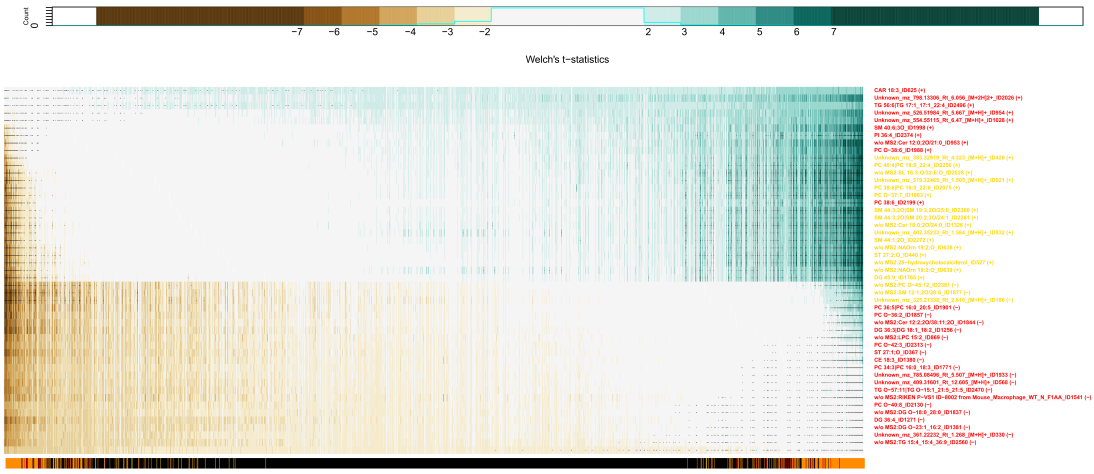

Figure 18: Cut-out from Fig. 16 with the rows limited to the features showing different results in terms of significance when carrying out PLS-DA based on SPCs or lnPQN. Features identified as biomarkers by the SPC approach and not by the lnPQN one are colored in red (the opposite case is colored in yellow). Features identified as biomarkers by both approaches are marked in orange (those identified by none are marked in black). The row labels denote whether the respective feature was flagged by the indicated method as a biomarker increased (+) or decreased (−) in the patient group.

From a biochemical point of view, some interesting comparison can be made in terms of classes that have been identified as biomarkers or non-biomarkers in previous studies (see Appendix B). CAR 18:3 was identified as a significant biomarker solely with SPC-based PLS-DA. This finding is in accord with the previous research on the class of CARs. In contrast, PCs, PIs and SMs were only identified as biomarker candidates when using lnPQN and OPCs in PLS-DA. Based on the literature, these classes of features are probably false positives.

5 Robust principal component analysis for compositional tables

A frequent primary task in multivariate statistics is to reduce the dimensionality of the data at hand, done using principal component analysis (PCA). As stated in Chapter 2, in case of CoDa, and consequently also compositional tables, this needs to be done in a proper coordinate representation that maps the Aitchison geometry of compositions to the standard Euclidean geometry (Pawlowsky-Glahn and Egozcue, 2001). To eliminate the influence of outlying observations in PCA, Filzmoser et al. (2009) proposed to estimate the covariance matrix for robust PCA (rPCA) by the *Minimum Covariance Determinant (MCD)* estimator (Maronna et al., 2006) which has the property of affine equivariance, advantageous in the logratio context. Since clr coefficients (5) lead to singularity and are not appropriate for most robust methods including the MCD estimator, loadings and scores of rPCA need to be computed from olr coordinates (Egozcue et al., 2003) of the compositional data and then transformed back to clr coefficients for a better interpretation of the resulting compositional biplot.

Accordingly, the aim of this chapter is to generalize the previous considerations on dimension reduction of vector CoDa and to propose a robust approach to principal component analysis of compositional tables. While it is obvious that the work can be started from Fačevicová et al. (2016) and Filzmoser et al. (2009), it is on the other hand not immediate to see if there is a possibility to identify a relationship between clr coefficients and olr coordinates as it is done in the case of vector compositions by the equation (7). The first issue here is posed by the dimensionality of compositional tables which is much lower than the number of clr coefficients if computed directly for independence and interaction tables. The second obstacle is formed by the decomposition where not all choices of coordinates allow for a satisfactory interpretation of both the original and the decomposed tables at the same time (i.e., providing also a way to capture the relationships between them which is necessarily contributing to a better insight into the structure of the tables). It will be shown that OPCs present a favorable choice of olr coordinates in line with the previous thoughts and that they can keep a good interpretability when properly linked to clr coefficients.

In Chapter 5.1, the dimension reduction of vector CoDa using rPCA is briefly reviewed. Clr coefficients of compositional tables together with a link to their OPC representation allowing for a well-interpretable processing using rPCA are introduced in Chapter 5.2. The new methodology is illustrated in Chapters 5.3 and 5.4 on real data sets from OECD Statistics using the statistical software `R`, namely the `robCompositions` package. Data from several different countries containing unemployment information with gender distribution and age structure are processed as a set of 2×4 compositional tables. Therefore, a robust compositional biplot is a possible tool to analyze the distribution of unemployment rates in these countries as well as gender and age differences. Data from the area of education, carrying relative information about fields of study and the resulting degree in given countries, are approached as larger 3×8 compositional tables, and results for men and women are compared.

5.1 Robust principal component analysis for compositional data

One of the widely used methods for the purpose of dimension reduction of large-scale data sets in a compositional approach is PCA just like in the case of standard multivariate data analysis. It converts possibly correlated original variables from the data at hand into a smaller set of linearly uncorrelated variables called principal components (PCs). Additionally, the first component accounts for the largest variance of the given data, the second one for a maximum of the remaining variance, etc., under the constraint of being orthogonal to all the previous principal components (Johnson and Wichern, 2007).

The covariance matrix $\mathbf{C}$ estimated from a real data matrix $\mathbf{X}$ can be spectrally decomposed into

$$\mathbf{C} = \mathbf{G}\mathbf{L}\mathbf{G}^T,$$

where $\mathbf{G}$ is a matrix of eigenvectors and $\mathbf{L}$ represents a diagonal matrix of eigenvalues of $\mathbf{C}$. It is then possible to define the PCA transformation as

$$\mathbf{X}^* = (\mathbf{X} - \mathbf{1}^T\boldsymbol{t})\mathbf{G},$$

where $\boldsymbol{t}$ is the (row) location estimator and $\mathbf{1}$ is a vector of ones with length n (number of observations). The columns of the matrix $\mathbf{X}^*$, the coordinates of the principal components, are called *scores* and the columns of $\mathbf{G}$, containing the

respective basis vectors, are called *loadings*. Typically, only the first few principal components are considered for further analysis. Taking into account only two PCs, a graphical outcome called *biplot* can depict both loadings as arrows and scores as points in one plot, where associations can be revealed.

It is common to take $\boldsymbol{t}$ as the arithmetic mean and $\mathbf{C}$ as the sample covariance matrix. However, both are very sensitive to outlying observations. Robust alternatives can be obtained by using the MCD estimators of location and covariance (Maronna et al., 2006). This approach inquires working in olr coordinates to obtain full rank data in order to get the MCD estimate of the covariance matrix and the respective matrix of eigenvectors $\mathbf{G}$. In addition, olr coordinates ensure subcompositional coherence and enable to keep affine equivariance of the results to the change of basis.

Accordingly, rPCA of CoDa based on the MCD estimator requires olr coordinates $\boldsymbol{z}_i$ as an input, and the scores $\boldsymbol{z}_i^*$ are given by

$$\boldsymbol{z}_i^* = (\boldsymbol{z}_i - \boldsymbol{t})\mathbf{G}.$$

Once PCA is performed, the loadings can be transformed back to clr coefficients as

$$\mathbf{G}_{\text{clr}} = \mathbf{V}\mathbf{G},$$

accounting for compositional biplot construction with meaningful interpretation, whereas the scores remain identical and only a column of zeros is added to the end. Clr coefficients are also worth as such for their simple construction as an amalgamation of pairwise logratios of a given part. Due to the zero-sum constraint of clr coefficients, their covariance structure is distorted, thus the interpretation of the biplot in terms of the correlation between coefficients (through angles between arrows) might be misleading. Instead, the focus is on links between vertices of arrows as they stand for a proportionality between the original compositional parts (Aitchison and Greenacre, 2002). On the other hand, due to the relation with OPCs, the single clr variables (or the respective loadings) can be used to identify observations with a high dominance of the respective parts in a compositional vector (Kynčlová et al., 2016).

5.2 Centered logratio representation and its link to pivot coordinates of compositional tables

As stated in the previous chapters, a coordinate representation which respects the sample space dimensionality as well as the decomposition procedure is needed to perform rPCA of compositional tables. Interestingly, the coordinates of the entire compositional table given in (15) and (16) of Chapter 2.3 can be divided into two groups according to the dimensionality of the independence and interaction tables, respectively. This becomes the main advantage also when using OPCs for rPCA since it allows for a comparison of the results from the whole table and its decomposed parts.

Following the link (9) between the first OPC and the respective clr coefficient of vector CoDa, also the first coordinates of the three types from each system can then be expressed as proportional (up to a constant) to respective clr coefficients,

$$\mathrm{clr}(\boldsymbol{x}_{ind})_{kl} = \sqrt{\frac{I-1}{IJ}} z_1^{r(k)} + \sqrt{\frac{J-1}{IJ}} z_1^{c(l)},$$

$$\mathrm{clr}(\boldsymbol{x}_{int})_{kl} = \sqrt{\frac{(I-1)(J-1)}{IJ}} z_{11}^{OR(kl)},$$

which is an important fact for the interpretation of the analysis.

The resulting clr coefficients, computed originally from the elements of the independence and interaction tables (14),

$$\mathrm{clr}(\boldsymbol{x}_{ind})_{ij} = \ln \frac{x_{ij}^{ind}}{g(\boldsymbol{x}_{\bullet\bullet}^{ind})}, \quad \mathrm{clr}(\boldsymbol{x}_{int})_{ij} = \ln \frac{x_{ij}^{int}}{g(\boldsymbol{x}_{\bullet\bullet}^{int})},$$

can thus be expressed also in terms of cells of the input compositional table as

$$\mathrm{clr}(\boldsymbol{x}_{ind})_{ij} = \ln \frac{g(\boldsymbol{x}_{i\bullet})g(\boldsymbol{x}_{\bullet j})}{g(\boldsymbol{x}_{\bullet\bullet})^2}, \quad \mathrm{clr}(\boldsymbol{x}_{int})_{ij} = \ln \frac{x_{ij} g(\boldsymbol{x}_{\bullet\bullet})}{g(\boldsymbol{x}_{i\bullet})g(\boldsymbol{x}_{\bullet j})}, \tag{23}$$

where $g(\boldsymbol{x}_{i\bullet})$, $g(\boldsymbol{x}_{\bullet j})$ and $g(\boldsymbol{x}_{\bullet\bullet})$ stand for the geometric mean of the i-th row ($i = 1, \ldots, I$), the j-th column ($j = 1, \ldots, J$) and the whole compositional table (and its independent and interactive counterparts for $g(\boldsymbol{x}_{\bullet\bullet}^{ind})$, $g(\boldsymbol{x}_{\bullet\bullet}^{int})$), respectively. As a consequence, each $\mathrm{clr}(\boldsymbol{x}_{ind})_{ij}$ expresses a dominance of a given combination of factor values in case of independence. This dominance is then either

amplified or weakened according to the interaction table which depends on whether the interaction is shifted in a positive or a negative direction. The interaction table refers also to sources of departures from independence, nevertheless, the information obtained only from $\mathrm{clr}(\boldsymbol{x}_{int})_{ij}$ does not provide a complete picture about the dominance of the respective cell to all other averaged cells.

Furthermore, note that each coordinate $\mathrm{clr}(\boldsymbol{x}_{ind})_{ij}$ is formed by the sum of clr coefficients of the respective row and column marginals, $1/J \sum_j \mathrm{clr}(\boldsymbol{x}_{ind})_{ij} = \ln g_{i\bullet}/g_{\bullet\bullet}$ and $1/I \sum_i \mathrm{clr}(\boldsymbol{x}_{ind})_{ij} = \ln g_{\bullet j}/g_{\bullet\bullet}$, which amount to zero. Thus, there are only $I+J-2$ linearly independent clr coefficients, reflecting the dimensionality of the sample space of independence tables again. A similar feature holds also for clr coefficients of interaction tables that sum up to zero across each row or column. Consequently, in the case of an interaction table, the number of linearly independent clr coefficients reduces to $(I-1)(J-1)$. Since this dependency makes it impossible to use the clr coefficients for the rPCA of independence and interaction tables, the strategy to perform rPCA for compositional tables is the same as in case of vector CoDa: PCA loadings and scores are computed in olr coordinates (OPCs) and then back-transformed using relation (7) to the clr space, where the loadings can be interpreted in terms of dominance of single cells. Here, clr coefficients of basis vectors for rows $\boldsymbol{e}^r$, columns $\boldsymbol{e}^c$ and interactions $\boldsymbol{e}^{OR}$, forming the columns of the matrix $\mathbf{V}$, are defined as follows,

$$\mathrm{clr}(\boldsymbol{e}^r) = \begin{cases} \sqrt{\frac{I-i}{(I-i+1)J}} & \text{for the elements in pivot row } i, \\ -\sqrt{\frac{1}{(I-i+1)J(I-i)}} & \text{for the elements in rows } i+1,\ldots,I, \\ 0 & \text{otherwise,} \end{cases} \tag{24}$$

$$\mathrm{clr}(\boldsymbol{e}^c) = \begin{cases} \sqrt{\frac{J-j}{(J-j+1)I}} & \text{for the elements in pivot column } j, \\ -\sqrt{\frac{1}{(I-i+1)J(I-i)}} & \text{for the elements in columns } j+1,\ldots,J, \\ 0 & \text{otherwise,} \end{cases} \tag{25}$$

and

$$\operatorname{clr}(\boldsymbol{e}^{OR}) = \begin{cases} \sqrt{\frac{1}{rs(r-1)(s-1)}} & \text{for the elements on positions } i = r+1,\ldots,I, \\ & j = s+1,\ldots,J \\ \sqrt{\frac{(r-1)(s-1)}{rs}} & \text{for the pivot elements } rs \\ -\sqrt{\frac{r-1}{rs(s-1)}} & \text{for the elements in pivot row } r, \\ & j = s+1,\ldots,J, \\ -\sqrt{\frac{s-1}{rs(r-1)}} & \text{for the elements in pivot column } s, \\ & i = r+1,\ldots,I, \\ 0 & \text{otherwise,} \end{cases} \tag{26}$$

reinterpreting the expressions from Fačevicová et al. (2016). As a result of (7), row-wise clr coefficients of the whole table are obtained for the $IJ-1$ columns of the matrix $\mathbf{V}$. Alternatively, if the matrix $\mathbf{V}$ has just $I+J-2$ columns formed by clr coefficients of basis vectors corresponding to the OPC representation of the independence table (15), its respective clr coefficients are derived (and similarly for the interaction table with its coordinates (16)). Finally, the transformed loadings and scores can be used to construct a biplot in order to reveal the multivariate structure of the sample of compositional tables and relations between both factors.

5.3 Unemployment data analysis

In the following, the methodological results are applied to two real-world data sets from the field of economy in order to illustrate the main features and possible limitations of the approach. However, it is important to acknowledge here the potential of the proposed methodology also across other research areas since compositional tables (eventually in form of their count counterparts, i.e, contingency tables) occur in many applications and sciences. For example, in environmental management, the presence of two-factor CoDa was recognized already by Aitchison (1986) where areal compositions given by simplified 2×2 tables of vegetation (thick and thin) and animals (dense, sparse) abundance in different regions were provided as an example of CoDa with more complex structures. From rather up-to-date environmental management problems, analyses of e.g., 5×4 tables of material resources (OECD Statistics, 2017) given by the extraction type

(domestic extraction, import, export, direct material input, and domestic material consumption) and group of the resources (biomass, fossil energy carriers, non-metallic minerals, and metals); 2×7 tables of protected areas (OECD Statistics, 2018) characterized by the domain of biodiversity (terrestrial or marine) and designation of the protected area (e.g., nature reserve, wilderness, protected landscape etc.); or 34×4 tables of carbon emissions embodied in trade (OECD Statistics, 2011) diversified by sector (e.g., agriculture, mining, food products etc.) and measure (imported, exported, consumption-based and production-based emissions), available always for the vast majority of OECD member states, could be mentioned as more relevant representatives.

The first analyzed data set, the Unemployed data set, is coming again from OECD Statistics and contains aggregated data from more than 150 million unemployed people from 42 different countries in 2010 (OECD Statistics, 2010b). It is analyzed using the statistical software environment `R` (R Core Team, 2022).

The data contain the numbers of unemployed people together with their gender and age category for the following countries: Australia, Austria, Belgium, Canada, Chile, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Mexico, Netherlands, New Zealand, Norway, Poland, Portugal, Slovakia, Slovenia, Spain, Sweden, Switzerland, Turkey, United Kingdom, United States, Colombia, Costa Rica, Latvia, Lithuania, China, India, Indonesia, Russian Federation and South Africa. An example of (transposed) raw data from the first four countries is shown in Table 6. The numbers in the tables are basically counts of unemployed people according to two factors. As the population size varies among the countries, the interest here is not in the absolute values of the counts in the single countries, but rather the relative structure of unemployment. Particularly, ratios of men and women and ratios among age groups $15-24$, $25-39$, $40-54$ and $55+$, as well as proportionality among countries will be analyzed. Since outliers can be anticipated due to completely different economics, education levels, gender balance and also traditions of the listed countries, the analysis will be carried out in a robust manner.

All compositional tables in this example have 2 rows and 4 columns, i.e., gender is the row factor and age structure is the column factor. The sample space of tables thus has dimension 7 out of which independence tables account for

Table 6: Unemployed people in thousands partitioned according to their gender and age groups (OECD Statistics, 2010b).

	Australia		Austria		Belgium		Canada	
Age group	Men	Women	Men	Women	Men	Women	Men	Women
15-24	129	111	29	25	53	43	250	178
25-39	90	85	40	35	86	83	241	192
40-54	66	68	36	27	65	52	242	188
55+	37	19	7	3	13	11	121	75

a dimension of 4 with OPCs z_1^r, z_1^c, z_2^c and z_3^c, while the remaining coordinates z_{11}^{OR}, z_{12}^{OR} and z_{13}^{OR} correspond to the interaction tables with a dimension of 3.

To point out the differences between the classical and robust PCA, both are performed and compared through the resulting covariance compositional biplots. Recall that classical PCA can directly be applied on clr coefficients. Nevertheless, since for this data set rPCA may be more relevant because of potential outlying tables, OPCs are used in both cases, and the results are transformed to clr for the biplot construction. This can only yield a different rotation of the classical biplot, however, it obviously does not alter the results.

In order to perform PCA in olr coordinates, the standard function `princomp` in `R` can be used, where the parameter `covmat` is set to `covMcd` (MCD estimator of covariance) in case of rPCA. Thereafter, loadings need to be transformed to clr coefficients as described in Chapter 2.2 using the matrix $\mathbf{V}$ with columns defined by (24) - (26) for the entire compositional table, and by (24) and (25), or by (26) for its independent and interactive part, respectively. The resulting classical biplots are depicted on the right-hand side of Fig. 19, while the rPCA output is on the left.

Assessing Fig. 19, it can be noticed that in all three cases, rPCA performs better in terms of explained variability by the first two PCs. As mentioned above, some outliers might be present in the data, and even from the classical biplot of the whole compositional tables (upper right corner of Fig. 19), at least two outlying tables (Turkey – TUR and South Africa – ZAF) could be expected, and so the robust approach should provide more meaningful results. An outlier detection is performed additionally in order to confirm these expectations. In the `R` package `robCompositions` (Templ et al., 2011), there is a function `outCoDa`

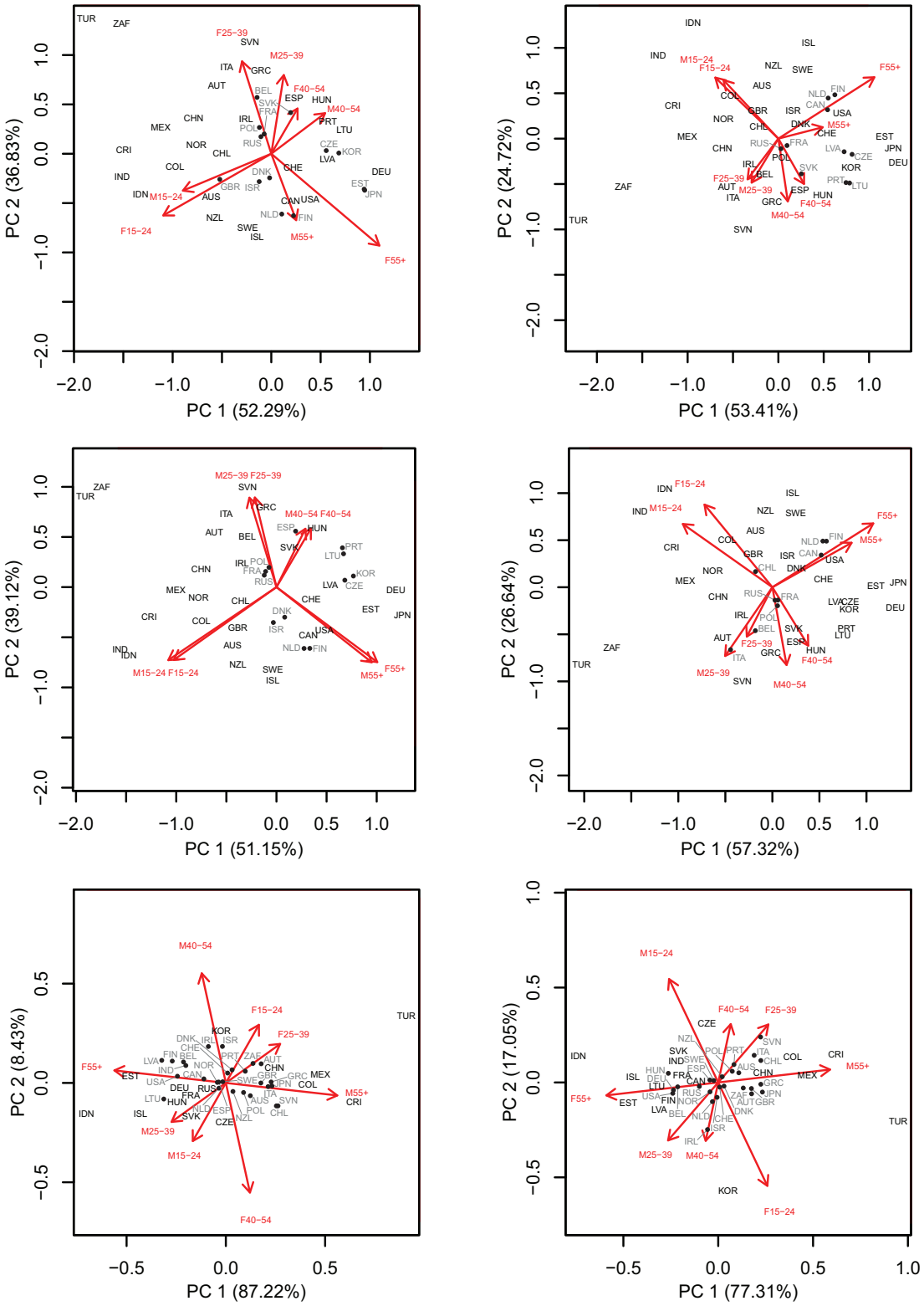

Figure 19: Robust (left column) and classical (right column) covariance biplots of the Unemployment compositional (upper row), independence (middle row) and interaction tables (lower row).

defined for this purpose, based on robust Mahalanobis distances computed from olr transformed data (Filzmoser and Hron, 2008). Using the OPCs and applying the 0.975 quantile of the chi-squared distribution as the common cut-off value, 15 out of all 42 countries are identified as outlying observations, clearly supporting the choice of robust analysis. Note that similarly, 10 observations from the set of independence tables and 6 from the interaction tables were detected as potential outliers.

Additionally, from the same part of Fig. 19, it is easy to identify from the direction of the arrows which countries tend to have relatively higher unemployment among younger people and which ones have a rather higher rate in the opposite situation. Although no clear compact clusters are visible, it seems that most European countries together with the USA and Canada tend to have more likely problems with employing older people, say 40+, while for Central and South America together with China, India, and Indonesia the unemployment depending on age structure has rather opposite tendencies. In the latter case, this generalized finding corresponds to the values of the youth unemployment rate (i.e., YUR $= \frac{\text{unemployed between 15 and 24 years}}{\text{all unemployed}}$) in different countries reported by the United Nations in 2011 (UNdata, 2011). Nevertheless, there are still big differences between European countries, even more apparent in the robust analysis. Some gender differences can be observed as well, except for the youngest generation. The structure in the classical biplot (upper right plot of Fig. 19) is similar but driven by the identified outlying observations.

The left plot in the middle part of Fig. 19 shows the "ideal" situation in case the relationships between gender and age factors would be filtered out. While the positions of the countries are not apparently changed compared to the previously discussed covariance biplot (upper left corner), the general relationships between the factors are remarkably illustrative. In case of independence, nearly gender equity would be achieved, while on the contrary, relationships among the age levels would be disproportionally weaker. Also, a bigger difference between results provided by robust and classical PCA is present here. One can easily understand how the classical approach does not handle outliers and how those can affect the output; the biplot on the right side is quite far away from picturing the same ideal situation.

As demonstrated in Chapter 5.2, the independence table captures the hypo-

thetical balanced state with each clr interpreted in terms of dominance of a given combination of factors in case of independence. However, this dominance is then either amplified, or weakened according to the interaction table; in terms of clr coefficients, it depends on whether the logratio dominance is shifted in a positive, or in a negative direction. Note that information obtained solely from the interaction table does not provide a complete picture about the dominance of single cells in the table. For example, in the lower left graph of Fig. 19 (robust biplot of interaction tables), Costa Rica (CRI) is placed towards the loading "male 55+", but this does not necessarily lead to a conclusion that unemployment in this group is higher in general in this country; it simply marks the cell whose dominance causes imbalance for Costa Rica, although the actual proportion of unemployment for this age group might be lower than its average dominance. Therefore, the conclusion about the higher dominance of unemployed men in the oldest group than expected in the hypothetical case of independence can be stated only after looking at the biplot of the independence tables. For the compositional tables with dimension $2 \times J$ (or alternatively $I \times 2$), this feature is nicely illustrated by the depicted loadings themselves, placed along a line corresponding to increasing dominance of one factor value at the expense of the latter value of the same factor. The opposite relation between the respective clr coefficients is clearly visible from both the biplots and the form of $\mathrm{clr}(\boldsymbol{x}_{int})_{ij}$ in (23): as it was already discussed in Chapter 5.2, clr coefficients of the interaction table sum up to zero across each row and column which results in the identity $\mathrm{clr}(\boldsymbol{x}_{int})_{1j} = -\mathrm{clr}(\boldsymbol{x}_{int})_{2j}$, holding for each j when $I = 2$ (and similarly for $J = 2$). While in case of higher data dimension the property is no longer visible in the graphs, in this example it can be seen that the two possible values of the gender factor lead to precisely contradictory loadings for any chosen value of the age factor. Thus they might only carry the information about the origin of the dominance shift, but no longer about the direction of the shift for which the difference from the independence table has to be consulted.

5.4 Education data analysis

It was illustrated in the previous example how outliers can affect results of classical PCA. Especially the gender equity achieved in the robust biplot of the independence tables would not be present in the classical one. Hence, in this

second example, only robust analysis outputs are discussed. The data set contains information about more than 7 million female and nearly 6 million male students, divided according to 8 different fields of study, being Education, Humanities and arts, Social sciences, business and law, Science, mathematics and computing, Engineering, manufacturing and construction, Agriculture and veterinary, Health and welfare, and Services (OECD Statistics, 2010a). The information about the achieved degree (bachelor, master, or doctoral) is recorded as well for about 30 different countries.

Compositional tables are analyzed for both genders separately in order to allow for a comparison of possible differences between them later on. Biplots as graphical rPCA outcomes of the whole compositional table as well as independence and interaction tables are collected in Fig. 20. Due to a larger dimensionality of compositional tables than in previous case (3×8), the biplots contain three times more variables and an objective interpretation becomes more difficult. An additional aspect is that since it is necessary to go many dimensions down to achieve the PCA projection using the first two PCs, it is expected to obtain more approximative picture of the multivariate data structure in the biplot. However, for data of similar or even bigger size, the proposed methods still offer an extremely useful rank-two approximation capturing the relationships between both factors. The performance of rPCA is still good enough (at least 54.08%) for this particular case in terms of cumulative variability explained by the first two PCs.

Despite the previous interpretational doubts, it can be seen that the effect of the chosen final degree is possibly stronger than the effect of the study field since the loadings tend to create a quite clear division of bachelors, masters, and doctors for most of the biplots. This property is more obvious for men while for women the difference between bachelors and masters is partially wiped out, maybe also due to a less compact data structure. Finally, employing the `outCoDa` function (Templ et al., 2011; Filzmoser and Hron, 2008) again, some outliers for the data set of the whole compositional tables are detected: Austria, Norway and Spain for men, and United Kingdom, Turkey and United States for women.

From the second row of Fig. 20, some overall idea about the hypothetical state of independence between degree and study field factors might be obtained. A stronger effect of the chosen degree and a weaker effect of the study field would still be apparent for men, and one new feature could be observed: there would

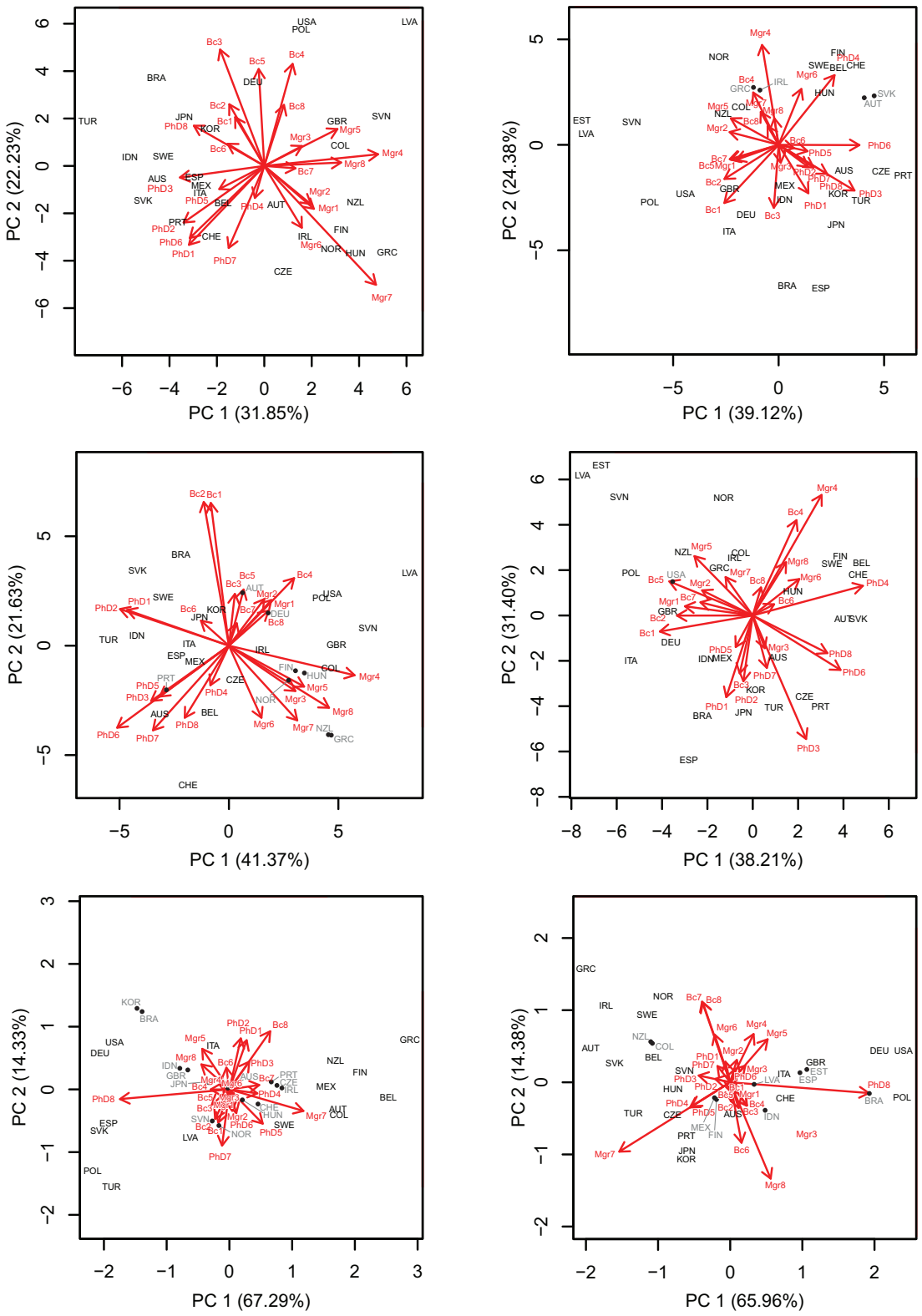

Figure 20: Robust covariance biplots of the Education compositional (upper row), independence (middle row) and interaction tables (lower row) for men (left column) and women (right column), respectively. Study fields are marked as follows: 1 = Education, 2 = Humanities and arts, 3 = Social sciences, business and law, 4 = Science, mathematics and computing, 5 = Engineering, manufacturing and construction, 6 = Agriculture and veterinary, 7 = Health and Welfare, and 8 = Services.

be a strong relation between Education and Humanities and arts study fields for each degree. For women, e.g., a similarity of educational systems in Sweden, Finland, Belgium and Switzerland is reflected. Also, in the case of independence, higher occurrence of outliers would be present for both men and women.

When looking at the biplot for the interaction tables for women (lower right figure), two of the mentioned countries are shifted away from the fields of study that would dominate if independence was achieved, being especially Agriculture and veterinary, Science, mathematics and computing, and Services. These countries are Sweden and Belgium, while Finland would correspond approximately to the independence between the factors, and Switzerland actually accounts for even stronger dominance of those fields (particularly master and doctoral studies in Services). For both men and women, it could be stated that the actual relationships between the factors are quite distant from the relative dominance given in the independence state. Stronger patterns concerning both factors are generally observed for men.

6 Final remarks

As suggested by the name of this thesis, it contributes advanced novel methods in the analysis of so-called compositions, i.e., data carrying relative information. Since the nature of CoDa endowed with Aitchison geometry entails fundamentally different approach to their statistical treatment, logratio methodology is used as a sound and necessary basis for the statistical analysis. The newly introduced tools are applied in the fields of science where high-dimensional data are a daily bread, namely metabolomics and econometrics. Additionally, the work conducted during my Ph.D. study has resulted in a contribution of new scientific insights through several interdisciplinary research collaborations in the field of metabolomics.

Chapter 1 outlined the topics presented in the thesis and their interconnectedness through the complexity and the compositional nature of the related data sets resulting in the need for pivot coordinates. Therefore, in Chapter 2, next to some basic principles of the logratio methodology, the entire genesis of pivot coordinates was provided. Specifically, all the different types of these olr coordinates including the case of compositional tables were summarized together with different real-world data driven examples motivating their origin. Since each of the methods introduced in this thesis at some point relies on the semi–automated choice of coordinates while needing to keep an easy interpretation, a special emphasis was put on the link between clr coefficients and pivot coordinates arising from the form of the respective logcontrasts. The limitations of the clr representation were described together with some fundamental properties of both OPCs and clr coefficients, providing always also a pairwise logratio point of view.

A new approach to univariate statistical analysis of (untargeted) metabolomic data, introducing a Bayesian version of a popular double-filtering graphical tool called volcano plot coupled with logratio data representation was proposed in Chapter 3. Although interpretability of clr coefficients would be fully satisfactory there, the univariate analysis is geometrically only reasonable when first pivot coordinates are used instead. Further, it was explained that the Bayesian counterpart to the multiple hypotheses testing might solve some of the problems occurring in frequentist analysis of high-dimensional data such as the inappropriateness of the routinely used p-value corrections for multiple testing or sensitivity of the traditional methods to outlying observations. Also, even if all limitations

of the frequentist approach were over-passed, the poverty of the information provided as a result of each hypothesis test is notable in the contrast to Bayesian approach producing the whole posterior distribution. Decision made on behalf of Bayesian inference is, therefore, always more competent, because it is based on much richer information compared to a single number from the traditional hypothesis testing. Consequently, the complex information hidden in the posterior distributions was exploited in the construction of the Bayesian version of volcano plot when deriving b-values for the y-axis and simplifying the posteriors to MPD representation for x-axis. Moreover, an additional feature combining information from both axes was provided in the form of HDI distance levels which could generally be used for the final choice of biomarker candidates.

Classification problems with CoDa have led to duly justified criticism of the OPC approach, commonly resulting in poorer sensitivity and specificity than competitors based on data normalization. In the field of metabolomics, this latter group is led by the widely used PQN which represents a more sophisticated alternative to simply using one element for normalization (approach popular e.g., in geochemistry or microbiome data analysis). A new type of pivot coordinates, SPCs, were thus proposed in Chapter 4. They exclude from the aggregation such pairwise logratios that are determined by Welch's t-statistic–based intervals as deviating from the main pattern. Hence, SPCs demonstrate the value in considering more complex logratios involving the compositional part of interest, while still retaining the intuitive idea of aggregating relative information into one (pivoting) logratio coordinate. Moreover, they further stress how the flexibility of the logratio approach built on well-founded geometrical grounds can outperform *ad hoc* solutions. Also, as shown, the method is connected as a particular zero-one weighting case with the broader framework of WPCs, which is able to deal with the drawbacks of OPCs in regression tasks. That is why the SPC approach presented here somehow closes the circle, having now the concept of pivot coordinates covering most common CoDa analysis and modeling situations met in the metabolomics context and beyond. Finally, aiming to enhance the identification of biomarkers in the context of binary classification problems, the novel coordinate system was embedded within a PLS-DA including Benjamini-Hochberg multiple testing adjustment for the bootstrap–based significance testing on the standardized PLS model coefficients.

In Chapter 5, rPCA of compositional tables as a two-factorial generalization of vector CoDa was studied. Given that compositional tables can be decomposed onto their independence and interaction parts, a statistical analysis of both is recommended to get insight into the ideal situation when relationships amid factors are filtered away, as well as into interactions between factors forming the original compositional table. As most practical data sets contain outlying observations, robust methods requiring an orthonormal coordinate representation have been considered. To reduce the dimension of data at hand, rPCA using the MCD estimator can be applied to pivot coordinates of compositional tables according to their decomposition into independence and interaction tables. The necessity of respecting dimensionality of the independent and interactive parts presents the main difference to (vector) CoDa where such feature does not occur. It was precisely this need of specific choice of olr coordinates where coordinates of independence and interaction tables form together coordinates of the entire compositional tables which allowed here for the additional benefit brought by the linkage of OPCs to clr coefficients constructed in the same manner. Thereafter, loadings obtained in OPCs for the rPCA were transformed back to clr coefficients where they were used for the construction of compositional biplots and their meaningful analysis. In case of $(2 \times J)$ table dimensions, an additional feature could be observed in the graphical output of interaction tables, which was traced back to the interpretation of the clr coefficients as well.

The good performance of the novel methods was always shown on the analyses of two different dimension–relatable data sets (from the field of rare metabolic diseases and economy, respectively). For Chapters 3 and 4, simulation studies were also provided to compare the stability and/or performance of the proposed tools with the traditional approaches to the presented tasks. In both cases, the results of the simulations highlighted the potential of the new methods.

All computations in this thesis were performed using the environment of the statistical software `R` (R Core Team, 2022). The related codes are available online at `https://github.com/sousaju/BayesVolcano` for Chapter 3, `https://github.com/sousaju/SPC` for Chapter 4, and `https://github.com/sousaju/rPCA-CoDaTables` for Chapter 5 under GNU GPL.

I truly hope that also thanks to a certain aspect of robustness and high-dimensionality that was present in all three introduced tasks and that is an indispensable part of an ample amount of data sets in practice, will the novel tools have the potential to quickly incorporate alongside the well-established methods from the CoDa analysis. I strongly believe that my dissertation thesis will also render one of the final touches to the research around pivot coordinates by providing the last piece currently missing in the area of weighting techniques.

A Transgenic rat models with induced tauopathy: biological background

The first application concerns transgenic rat models with induced tauopathy. Tauopathies are neurodegenerative disorders, with Alzheimer's disease (AD) being one of the most prevalent tauopathies in humans (Karlíková et al., 2017).

Data obtained from samples CSF collected from transgenic rats at the age of 4 and 14 months were acquired by a targeted lipidomic approach using high-performance liquid chromatography coupled with mass spectrometry (UHPLC-MS). The data set is available at the MassIVE database (Center for Computational Mass Spectrometry, 2023) and it was pre-processed using the `Metabol` package (Gardlo et al., 2019) on the R system for statistical computing (R Core Team, 2022). Based on a mixed sample for quality control (QC; analyzed periodically every 6th sample), locally estimated smoothing signal (LOESS) correction was applied to the data. Lipids whose coefficient of variation calculated from QC aliquots was higher than 30% were excluded from further data processing.

Several previous studies have provided findings about biochemically relevant biomarkers and their role in tauopathic neurodegeneration and AD. In addition to the pathological aggregation of tau protein in tauopathy, amyloid beta plaque formation occurs in AD patients. The whole class of phosphatidylcholines has been detected as upregulated in CSF of cognitively healthy humans with abnormal or pathological tau protein or amyloid beta peptide 42 (Aβ42) levels (Fonteh et al., 2020). Their close metabolic intermediates, the lysophosphatidylcholines, have been positively associated with the formation of Aβ(1-42) fibrils (Sheikh and Nagai, 2011; Pedersen et al., 2019). Next, elevated sphingomyelin concentrations have been linked to membrane breakdown, demyelination, and progressive loss of neuronal cells in brain tissue during the progression of neurofibrillary pathology (Kao et al., 2020). Moreover, no significant alteration in lipids from the class of plasmanyl/plasmenyl lysophosphatidylethanolamines or triacylglycerols has been revealed in patients with AD (Kao et al., 2020) so far.

B SCADD: biological background

The second application concerns dry blood spot samples of patients with a hereditary genetic disorder in β-oxidation of short-chain fatty acids. These patients suffer from enzyme deficiency in short-chain acyl-coenzyme A dehydrogenase (SCAD, EC 1.3.8.1). Owing to the disruption of this pathway, the disease is manifested by increased concentrations of butyric acid residues in the form of butyryl-carnitine, butyryl-coenzyme A, and butyryl-glycine conjugates in the patient's biofluids (Gallant et al., 2012).

As with the previous case study, the data were obtained using UHPLC-MS, and only untargeted lipidomics was applied. Therefore, instead of proper lipids, the variables in the original raw data set refer generically to *features*. These features also represent adducts, source fragments, multimers, and isotopes of as-yet-unidentified molecules (Graça et al., 2022). From the total of 2011 features present in the data set, 761 were fully identified, 693 just partially identified, and the remaining 557 were unknown.

Octadecatrienyl-carnitine (CAR 18:3) was found to be a potential biomarker which is of particular interest because acylated carnitines and other acyl conjugates are related to accumulated intermediates of disrupted beta-oxidation (Nochi et al., 2017). A cascading accumulation of long-chain fatty acids may have occurred due to SCAD deficiency. In contrast, polyunsaturated glycerophospholipids and phosphatidylinositols were also identified as biomarkers by some of the approaches. Because of the failed beta-oxidation in SCAD deficient patients, it can be assumed that many other metabolic pathways taking place in the mitochondria have been affected. However, glycerophospholipids are as such formed in a different compartment of the cell, the endoplasmic reticulum (Blom et al., 2011). Therefore, these lipids are probably false positive findings. Among these potential false positives were long-chain SM lipids, which are also synthesised extra-mitochondrially (Gault et al., 2010).

Bibliography

Aitchison J (1986) The statistical analysis of compositional data. Chapman & Hall, London, DOI 10.1007/978-94-009-4109-0

Aitchison J, Greenacre M (2002) Biplots of compositional data. Journal of the Royal Statistical Society Series C: Applied Statistics 51(4):375–392, DOI 10.1111/1467-9876.00275

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) 57(1):289–300, DOI 10.1111/j.2517-6161.1995.tb02031.x

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics 29(4):1165–1188, DOI 10.1214/aos/1013699998

Benton HP, Want EJ, Ebbels TMD (2010) Correction of mass calibration gaps in liquid chromatography–mass spectrometry metabolomics data. Bioinformatics 26(19):2488–2489, DOI 10.1093/bioinformatics/btq441

Blom T, Somerharju P, Ikonen E (2011) Synthesis and biosynthetic trafficking of membrane lipids. Cold Spring Harbor Perspectives in Biology 3(8), DOI 10.1101/cshperspect.a004713, Article a004713

Bruno F, Greco F, Ventrucci M (2015) Spatio-temporal regression on compositional covariates: modeling vegetation in a gypsum outcrop. Environmental and Ecological Statistics 22:445–463, DOI 10.1007/s10651-014-0305-4

Buccianti A, Egozcue JJ, Pawlowsky-Glahn V (2014) Variation diagrams to statistically model the behavior of geochemical variables: Theory and applications. Journal of Hydrology 519:988–998, DOI 10.1016/j.jhydrol.2014.08.028

Center for Computational Mass Spectrometry (2023) MassIVE dataset MSV000091311. DOI 10.25345/C5D50G70W, URL https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?accession=MSV000091311, accessed on March 17, 2023

Cui X, Churchill GA (2003) Statistical tests for differential expression in cDNA microarray experiments. Genome Biology 4, DOI 10.1186/gb-2003-4-4-210, Article 210

Davis A, Scott A (1973) An approximation to the k-sample Behrens-Fisher distribution. Sankhyā: The Indian Journal of Statistics, Series B 35(1):45–50

De Bragança Pereira C, Stern J (1999) Evidence and credibility: Full Bayesian significance test for precise hypotheses. Entropy 1(4):99–110, DOI 10.3390/e1040099

de Sousa J, Vencálek O, Hron K, Václavík J, Friedecký D, Adam T (2020) Bayesian multiple hypotheses testing in compositional analysis of untargeted metabolomic data. Analytica Chimica Acta 1097:49–61, DOI 10.1016/j.aca.2019.11.006

de Sousa J, Hron K, Fačevicová K, Filzmoser P (2021) Robust principal component analysis for compositional tables. Journal of Applied Statistics 48(2):214–233, DOI 10.1080/02664763.2020.1722078

Di Guida R, Engel J, Allwood JW, Weber RJ, Jones MR, Sommer U, Viant MR, Dunn WB (2016) Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. Metabolomics 12, DOI 10.1007/s11306-016-1030-9, Article 93

Dickhaus T, Straßburger K, Schunk D, Morcillo-Suarez C, Illig T, Navarro A (2012) How to analyze many contingency tables simultaneously in genetic association studies. Statistical Applications in Genetics and Molecular Biology 11(4), DOI 10.1515/1544-6115.1776, Article 12

Dieterle F, Ross A, Schlotterbeck G, Senn H (2006) Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in ^{1}H NMR metabonomics. Analytical Chemistry 78(13):4281–4290, DOI 10.1021/ac051632c

Dumuid D, Stanford TE, Martin-Fernández JA, Pedišić Ž, Maher CA, Lewis LK, Hron K, Katzmarzyk PT, Chaput JP, Fogelholm M, Hu G, Lambert EV,

Maia J, Sarmiento OL, Standage M, Barreira TV, Broyles ST, Tudor-Locke C, Tremblay MS, Olds T (2018) Compositional data analysis for physical activity, sedentary time and sleep research. Statistical Methods in Medical Research 27(12):3726–3738, DOI 10.1177/0962280217710835

Egozcue JJ, Pawlowsky-Glahn V (2005) Groups of parts and their balances in compositional data analysis. Mathematical Geology 37(7):795–828, DOI 10.1007/s11004-005-7381-9

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric logratio transformations for compositional data analysis. Mathematical Geology 35(3):279–300, DOI 10.1023/A:1023818214614

Egozcue JJ, Díaz Barrero JL, Pawlowsky-Glahn V (2008) Compositional analysis of bivariate discrete probabilities. In: Daunis-i Estadella J, Martín-Fernández JA (eds) Proceedings of CODAWORK'08, The 3rd Compositional Data Analysis Workshop, University of Girona, Spain

Egozcue JJ, Pawlowsky-Glahn V, Templ M, Hron K (2015) Independence in contingency tables using simplicial geometry. Communications in Statistics-Theory and Methods 44(18):3978–3996, DOI 10.1080/03610926.2013.824980

Fačevicová K, Hron K, Todorov V, Guo D, Templ M (2014) Logratio approach to statistical analysis of 2× 2 compositional tables. Journal of Applied Statistics 41(5):944–958, DOI 10.1080/02664763.2013.856871

Fačevicová K, Hron K, Todorov V, Templ M (2016) Compositional tables analysis in coordinates. Scandinavian Journal of Statistics 43(4):962–977, DOI 10.1111/sjos.12223

Fačevicová K, Hron K, Todorov V, Templ M (2018) General approach to coordinate representation of compositional tables. Scandinavian Journal of Statistics 45(4):879–899, DOI 10.1111/sjos.12326

Fačevicová K, Filzmoser P, Hron K (2022) Compositional cubes: a new concept for multi-factorial compositions. Statistical Papers DOI 10.1007/s00362-022-01350-8

Filzmoser P, Hron K (2008) Outlier detection for compositional data using robust methods. Mathematical Geosciences 40:233–248, DOI 10.1007/s11004-007-9141-5

Filzmoser P, Hron K (2013) Robustness for compositional data. In: Becker C, Fried R, Kuhnt S (eds) Robustness and Complex Data Structures, Springer, Berlin Heidelberg, DOI 10.1007/978-3-642-35494-6_8

Filzmoser P, Walczak B (2014) What can go wrong at the data normalization step for identification of biomarkers? Journal of Chromatography A 1362:194–205, DOI 10.1016/j.chroma.2014.08.050

Filzmoser P, Hron K, Reimann C (2009) Principal component analysis for compositional data with outliers. Environmetrics: The Official Journal of the International Environmetrics Society 20(6):621–632, DOI 10.1002/env.966

Filzmoser P, Hron K, Templ M (2018) Applied compositional data analysis. Springer Series in Statistics, Springer, Cham, DOI 10.1007/978-3-319-96422-5

Fišerová E, Hron K (2011) On the interpretation of orthonormal coordinates for compositional data. Mathematical Geosciences 43:455–468, DOI 10.1007/s11004-011-9333-x

Fonteh AN, Chiang AJ, Arakaki X, Edminster SP, Harrington MG (2020) Accumulation of cerebrospinal fluid glycerophospholipids and sphingolipids in cognitively healthy participants with Alzheimer's biomarkers precedes lipolysis in the dementia stage. Frontiers in Neuroscience 14, DOI 10.3389/fnins.2020.611393, Article 611393

Gallant NM, Leydiker K, Tang H, Feuchtbaum L, Lorey F, Puckett R, Deignan JL, Neidich J, Dorrani N, Chang E, Barshop BA, Cederbaum SD, Abdenur JE, Wang RY (2012) Biochemical, molecular, and clinical characteristics of children with short chain acyl-CoA dehydrogenase deficiency detected by newborn screening in California. Molecular Genetics and Metabolism 106(1):55–61, DOI 10.1016/j.ymgme.2012.02.007

Gallart-Ayala H, Teav T, Ivanisevic J (2020) Metabolomics meets lipidomics: assessing the small molecule component of metabolism. BioEssays: News and

Reviews in Molecular, Cellular and Developmental Biology 42(12), DOI 10.1002/bies.202000052, Article e2000052

Gardlo A, Smilde AK, Hron K, Hrdá M, Karlíková R, Friedecký D, Adam T (2016) Normalization techniques for PARAFAC modeling of urine metabolomic data. Metabolomics 12, DOI 10.1007/s11306-016-1059-9, Article 117

Gardlo A, Friedecký D, Najdekr L, Karlíková R, Adam T (2019) `Metabol`: The statistical analysis of metabolomic data. DOI 10.5281/zenodo.3235775, URL https://doi.org/10.5281/zenodo.3235775

Gault CR, Obeid LM, Hannun YA (2010) An overview of sphingolipid metabolism: from synthesis to breakdown. In: Chalfant C, Poeta MD (eds) Sphingolipids as Signaling and Regulatory Molecules, Springer Advances in Experimental Medicine and Biology, vol 688, Springer, New York, DOI 10.1007/978-1-4419-6741-1_1

Gelman A, Meng XL, Stern H (1996) Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica 6(4):733–807

Gelman A, Stern HS, Carlin JB, Dunson DB, Vehtari A, Rubin DB (2013) Bayesian data analysis, 3rd edn. Chapman and Hall/CRC press, New York, DOI 10.1201/b16018

Graça G, Cai Y, Lau CHE, Vorkas PA, Lewis MR, Want EJ, Herrington D, Ebbels TMD (2022) Automated annotation of untargeted all-ion fragmentation LC-MS metabolomics data with `MetaboAnnotatoR`. Analytical Chemistry 94(8):3446–3455, DOI 10.1021/acs.analchem.1c03032

Herder C, Rathmann W, Strassburger K, Finner H, Grallert H, Huth C, Meisinger C, Gieger C, Martin S, Giani G, Scherbaum WA, Wichmann HE, Illig T (2008) Variants of the PPARG, IGF2BP2, CDKAL1, HHEX, and TCF7L2 genes confer risk of type 2 diabetes independently of BMI in the German KORA studies. Hormone and Metabolic Research 40(10):722–726, DOI 10.1055/s-2008-1078730

Hron K, Filzmoser P, de Caritat P, Fišerová E, Gardlo A (2017) Weighted pivot coordinates for compositional data and their application to geochemical mapping. Mathematical Geosciences 49(6):797–814, DOI 10.1007/s11004-017-9684-z

Hron K, Coenders G, Filzmoser P, Palarea-Albaladejo J, Faměra M, Matys Grygar T (2021a) Analysing pairwise logratios revisited. Mathematical Geosciences 53(7):1643–1666, DOI 10.1007/s11004-021-09938-w

Hron K, Engle M, Filzmoser P, Fišerová E (2021b) Weighted symmetric pivot coordinates for compositional data with geochemical applications. Mathematical Geosciences 53(4):655–674, DOI 10.1007/s11004-020-09862-5

Johnson RA, Wichern DW (2007) Applied multivariate statistical analysis, 6th edn. Pearson Prentice Hall, Upper Saddle River

Kalivodová A, Hron K, Filzmoser P, Najdekr L, Janečková H, Adam T (2015) PLS-DA for compositional data with application to metabolomics. Journal of Chemometrics 29(1):21–28, DOI 10.1002/cem.2657

Kao YC, Ho PC, Tu YK, Jou IM, Tsai KJ (2020) Lipids and Alzheimer's disease. International Journal of Molecular Sciences 21(4), DOI 10.3390/ijms21041505, Article 1505

Karlíková R, Mičová K, Najdekr L, Gardlo A, Adam T, Majerová P, Friedecký D, Kováč A (2017) Metabolic status of CSF distinguishes rats with tauopathy from controls. Alzheimer's Research & Therapy 9, DOI 10.1186/s13195-017-0303-5, Article 78

Kouřil Š, de Sousa J, Václavík J, Friedecký D, Adam T (2020) CROP: Correlation-based reduction of feature multiplicities in untargeted metabolomic data. Bioinformatics 36(9):2941–2942, DOI 10.1093/bioinformatics/btaa012

Kruschke JK (2013) Bayesian estimation supersedes the t test. Journal of Experimental Psychology: General 142(2):573–603, DOI 10.1037/a0029146

Kruschke JK (2014) Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan, 2nd edn. Academic Press, Boston, DOI 10.1016/B978-0-12-405888-0.09999-2

Kruschke JK, Liddell TM (2018) The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. Psychonomic Bulletin & Review 25(1):178–206, DOI 10.3758/s13423-016-1221-4

Kuhl C, Tautenhahn R, Bottcher C, Larson TR, Neumann S (2011) `CAMERA`: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. Analytical Chemistry 84(1):283–289, DOI 10.1021/ac202450g

Kumar N, Hoque MA, Sugimoto M (2018) Robust volcano plot: identification of differential metabolites in the presence of outliers. BMC Bioinformatics 19(1), DOI 10.1186/s12859-018-2117-2, Article 128

Kynčlová P, Filzmoser P, Hron K (2016) Compositional biplots including external non-compositional variables. Statistics 50(5):1132–1148, DOI 10.1080/02331888.2015.1135155

Kynčlová P, Hron K, Filzmoser P (2017) Correlation between compositional parts based on symmetric balances. Mathematical Geosciences 49:777–796, DOI 10.1007/s11004-016-9669-3

Li W (2012) Volcano plots in analyzing differential expressions with mRNA microarrays. Journal of Bioinformatics and Computational Biology 10(6), DOI 10.1142/S0219720012310038, Article 1231003

Malyjurek Z, de Beer D, Joubert E, Walczak B (2019) Working with log-ratios. Analytica Chimica Acta 1059:16–27, DOI 10.1016/j.aca.2019.01.041

Maronna RA, Martin RD, Yohai VJ (2006) Robust Statistics: Theory and Methods. John Wiley & Sons, New York, DOI 10.1002/0470010940

Martín-Fernández JA (2019) Comments on: Compositional data: the sample space and its structure. Test 28(3):653–657, DOI 10.1007/s11749-019-00672-4

Mateu-Figueras G, Pawlowsky-Glahn V, Egozcue JJ (2011) The principle of working on coordinates. In: Pawlowsky-Glahn V, Buccianti A (eds) Compositional

Data Analysis: Theory and Applications, John Wiley & Sons, Chichester, DOI 10.1002/9781119976462

Mert MC, Filzmoser P, Hron K (2016) Error propagation in isometric logratio coordinates for compositional data: Theoretical and practical considerations. Mathematical Geosciences 48(8):941–961, DOI 10.1007/s11004-016-9646-x

Mielke MM, Haughey NJ, Bandaru VV, Zetterberg H, Blennow K, Andreasson U, Johnson SC, Gleason CE, Blazel HM, Puglielli L, Sager MA, Asthana S, Carlsson CM (2014) Cerebrospinal fluid sphingolipids, β-amyloid, and tau in adults at risk for Alzheimer's disease. Neurobiology of Aging 35(11):2486–2494, DOI 10.1016/j.neurobiolaging.2014.05.019

Najdekr L, Gardlo A, Mádrová L, Friedecký D, Janečková H, Correa ES, Goodacre R, Adam T (2015) Oxidized phosphatidylcholines suggest oxidative stress in patients with medium-chain acyl-CoA dehydrogenase deficiency. Talanta 139:62–66, DOI 10.1016/j.talanta.2015.02.041

Nochi Z, Olsen RKJ, Gregersen N (2017) Short-chain acyl-CoA dehydrogenase deficiency: from gene to cell pathology and possible disease mechanisms. Journal of Inherited Metabolic Disease 40(5):641–655, DOI 10.1007/s10545-017-0047-1

OECD Statistics (2010a) Education and Training. URL http://stats.oecd.org/, accessed on April 24, 2017

OECD Statistics (2010b) Unemployment by sex and age. URL http://stats.oecd.org/, accessed on March 10, 2017

OECD Statistics (2011) Policy indicators of trade and environment – Carbon emissions embodied in trade. URL http://stats.oecd.org/, accessed on October 17, 2019

OECD Statistics (2017) Environment – Material resources. URL http://stats.oecd.org/, accessed on October 12, 2019

OECD Statistics (2018) Environment – Biodiversity of protected areas in km^2. URL http://stats.oecd.org/, accessed on October 16, 2019

Ojo JO, Algamal M, Leary P, Abdullah L, Mouzon B, Evans JE, Mullan M, Crawford F (2018) Disruption in brain phospholipid content in a humanized tau transgenic model following repetitive mild traumatic brain injury. Frontiers in Neuroscience 12, DOI 10.3389/fnins.2018.00893, Article 893

Ortego MI, Egozcue JJ (2016) Bayesian estimation of the orthogonal decomposition of a contingency table. Austrian Journal of Statistics 45(4):45–56, DOI 10.17713/ajs.v45i4.136

Palarea-Albaladejo J, Martin-Fernandez JA (2015) `zCompositions`–R package for multivariate imputation of left-censored data under a compositional approach. Chemometrics and Intelligent Laboratory Systems 143:85–96, DOI 10.1016/j.chemolab.2015.02.019

Patil V (1965) Approximation to the Behrens-Fisher distributions. Biometrika 52(1/2):267–271, DOI 10.2307/2333830

Pawlowsky-Glahn V, Egozcue JJ (2001) Geometric approach to statistical analysis on the simplex. Stochastic Environmental Research and Risk Assessment (SERRA) 15(5):384–398, DOI 10.1007/s004770100077

Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015) Modeling and analysis of compositional data. John Wiley & Sons, Chichester, DOI 10.1002/9781119003144

Pedersen JN, Jiang Z, Christiansen G, Lee JC, Pedersen JS, Otzen DE (2019) Lysophospholipids induce fibrillation of the repeat domain of Pmel17 through intermediate core-shell structures. Proteins and Proteomics 1867(5):519–528, DOI 10.1016/j.bbapap.2018.11.007

Pié J, Lopez-Vinas E, Puisac B, Menao S, Pié A, Casale C, Ramos FJ, Hegardt FG, Gómez-Puertas P, Casals N (2007) Molecular genetics of HMG-CoA lyase deficiency. Molecular Genetics and Metabolism 92(3):198–209, DOI 10.1016/j.ymgme.2007.06.020

R Core Team (2022) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.r-project.org

Rhead WJ (2006) Newborn screening for medium-chain acyl-CoA dehydrogenase deficiency: A global perspective. Journal of Inherited Metabolic Disease 29(2-3):370–377, DOI 10.1007/s10545-006-0292-1

Rousseeuw P, Croux C (1993) Alternatives to the median absolute deviation. Journal of the American Statistical Association 88(424):1273–1283, DOI 10.2307/2291267

Sana TR, Gordon DB, Fischer SM, Tichy SE, Kitagawa N, Lai C, Gosnell WL, Chang SP (2013) Global mass spectrometry based metabolomics profiling of erythrocytes infected with Plasmodium falciparum. PloS one 8(4), DOI 10.1371/journal.pone.0060840, Article e60840

Santarelli F, Cassanello M, Enea A, Poma F, D'Onofrio V, Guala G, Garrone G, Puccinelli P, Caruso U, Porta F, Spada M (2013) A neonatal case of 3-hydroxy-3-methylglutaric-coenzyme A lyase deficiency. Italian Journal of Pediatrics 39, DOI 10.1186/1824-7288-39-33, Article 33

Sheikh AM, Nagai A (2011) Lysophosphatidylcholine modulates fibril formation of amyloid beta peptide. The FEBS Journal 278(4):634–642, DOI 10.1111/j.1742-4658.2010.07984.x

Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G (2006) `XCMS`: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Analytical Chemistry 78(3):779–787, DOI 10.1021/ac051437y

Štefelová N, Palarea-Albaladejo J, Hron K (2021) Weighted pivot coordinates for partial least squares-based marker discovery in high-throughput compositional data. Statistical Analysis and Data Mining: The ASA Data Science Journal 14(4):315–330, DOI 10.1002/sam.11514

Štefelová N, de Sousa J, Hron K, Palarea-Albaladejo J, Dobešová D, Kvasnička A, Friedecký D (2023) Selective pivot logratio coordinates for PLS-DA modelling with applications in metabolomics. *Under review*

Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, Fan TWM, Fiehn O, Goodacre R, Griffin JL, Hankemeier T, Hardy N, Harnly J, Higashi R,

Kopka J, Lane AN, Lindon JC, Marriott P, Nicholls AW, Reily MD, Thaden JJ, Viant MR (2007) Proposed minimum reporting standards for chemical analysis: chemical analysis working group (CAWG) metabolomics standards initiative (MSI). Metabolomics 3:211–221, DOI 10.1007/s11306-007-0082-2

Tautenhahn R, Boettcher C, Neumann S (2008) Highly sensitive feature detection for high resolution LC/MS. BMC Bioinformatics 9(1), DOI 10.1186/1471-2105-9-504, Article 504

Templ M, Hron K, Filzmoser P (2011) `robCompositions`: An R-package for robust statistical analysis of compositional data. In: Pawlowsky-Glahn V, Buccianti A (eds) Compositional Data Analysis: Theory and Applications, John Wiley & Sons, Chichester, DOI 10.1002/9781119976462

Templ M, Hron K, Filzmoser P, Gardlo A (2016) Imputation of rounded zeros for high-dimensional compositional data. Chemometrics and Intelligent Laboratory Systems 155:183–190, DOI 10.1016/j.chemolab.2016.04.011

Thulin M (2014) Decision-theoretic justifications for Bayesian hypothesis testing using credible sets. Journal of Statistical Planning and Inference 146:133–138, DOI 10.1016/j.jspi.2013.09.014

Torretta E, Arosio B, Barbacini P, Casati M, Capitanio D, Mancuso R, Mari D, Cesari M, Clerici M, Gelfi C (2018) Particular CSF sphingolipid patterns identify iNPH and AD patients. Scientific Reports 8, DOI 10.1038/s41598-018-31756-0, Article 13639

UNdata (2011) Youth unemployment, both sexes. URL `http://data.un.org/DocumentData.aspx?id=264#30`, accessed on November 4, 2019

Václavík J, Mádrová L, Kouřil Š, de Sousa J, Brumarová R, Janečková H, Jáčová J, Friedecký D, Knapková M, Kluijtmans LAJ, Grünert SC, Vaz FM, Janzen N, Wanders RJA, Wevers RA, Adam T (2020) A newborn screening approach to diagnose 3-hydroxy-3-methylglutaryl-CoA lyase deficiency. JIMD Reports 54(1):79–86, DOI 10.1002/jmd2.12118

van der Voet H (1994) Comparing the predictive accuracy of models using a simple randomization test. Chemometrics and Intelligent Laboratory Systems 25(2):313–323, DOI 10.1016/0169-7439(94)00084-V

Walach J, Filzmoser P, Hron K, Walczak B, Najdekr L (2017) Robust biomarker identification in a two-class problem based on pairwise log-ratios. Chemometrics and Intelligent Laboratory Systems 171:277–282, DOI 10.1016/j.chemolab.2017.09.003

Wasserstein RL, Lazar NA (2016) The ASA's statement on p-values: context, process, and purpose. The American Statistician 70(2):129–133, DOI 10.1080/00031305.2016.1154108

Welch BL (1947) The generalization of 'Student's' problem when several different population variances are involved. Biometrika 34(1–2):28–35, DOI 10.2307/2332510

PALACKÝ UNIVERSITY IN OLOMOUC
FACULTY OF SCIENCE

DISSERTATION THESIS SUMMARY

Advanced methods of compositional data analysis

Supervisor: **prof. RNDr. Karel Hron, Ph.D.**
Author: **Mgr. Julie de Sousa**
Study program: P1104 Applied Mathematics
Field of study: Applied Mathematics
Form of study: Full-time
The year of submission: 2023

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

Applicant: **Mgr. Julie de Sousa**
Dept. of Mathematical Analysis and Applications of Mathematics
Faculty of Science
Palacký University Olomouc

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

Reviewers: **prof. Josep Antoni Martín-Fernández, Ph.D.**
Dept. of Computer Science, Applied Mathematics and Statistics
University of Girona
Spain

prof. Anne Ruiz-Gazen, Ph.D.
Toulouse School of Economics
University of Toulouse Capitole
France

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 doctoral thesis is available at Study Department of Faculty of Science, Palacký University Olomouc.

Contents

Abstract

An abundance of scientific fields produces data where their relative structure, which is inherently contained in ratios among variables, is of the main interest. Due to the specific geometrical properties of such (compositional) data, a proper choice of real coordinates within the logratio framework is crucial for any sensible statistical analysis. In this thesis, novel methods related particularly to the use of so-called pivot logratio coordinates are presented within different research areas generating data sets of higher dimensionality or complexity. One of the essential tasks in omics sciences is to find statistically significant differences between patient and control groups to detect biomarkers of particular diseases using both univariate and multivariate statistical methods. A concept of b-values is introduced together with a Bayesian version of a widespread tool based on multiple hypotheses testing, the so-called volcano plot, incorporating also distance levels of the posterior highest density intervals from zero. Next, a new type of coordinate representation aiming to enhance the identification of biomarkers is proposed. They are constructed so that the "pivoting" coordinate representing a certain compositional part aggregates all but the deviating pairwise logratios of that part to the remaining ones, in accord with the name selective pivot coordinates. They are further coupled with partial least squares discriminant analysis as a gold standard in the multivariate analysis of omics data. Finally, a data table arranged according to two factors can often be considered a compositional table. Hence, a special choice of pivot coordinates reflecting a decomposition process into independent and interactive parts is presented for compositional data comprising the two-factorial complexity. A robust principal component analysis (PCA) is then performed for dimension reduction, allowing for investigation of the relationships between the given factors through a direct relation of the proposed coordinates to centered logratio coefficients, used traditionally in context of PCA with compositional data.

Key words: compositional data, logratio methodology, centered logratio coefficients, pivot coordinates, weighted pivot coordinates, selective pivot coordinates, compositional tables, Bayesian statistics, robust principal component analysis, volcano plot, partial least squares discriminant analysis, compositional biplot, metabolomic data, economic data

Abstrakt v českém jazyce

Celá škála vědeckých oborů produkuje data, u kterých je hlavním zájmem jejich relativní struktura, obsažená ze své podstaty v podílech mezi proměnnými. Vzhledem ke specifickým geometrickým vlastnostem takových (kompozičních) dat je pro jejich relevantní statistickou analýzu nezbytná správná volba reálných souřadnic v rámci logpodílové metodiky. V této práci jsou představeny nové metody související zejména s využitím tzv. pivotových souřadnic v různých oblastech výzkumu generujících datové soubory s vyšší dimenzionalitou nebo komplexností. Jedním z nejzásadnějších úkolů v tzv. -omických vědách je nalezení statisticky významných rozdílů mezi skupinami pacientů a kontrol, které slouží k detekci biomarkerů různých onemocnění s využitím jednorozměrných i mnohorozměrných statistických metod. Je zde představen koncept b-hodnot spolu s bayesovskou verzí populárního nástroje založeného na mnohonásobném testování hypotéz, nazývaného vulkánový graf. Díky bayesovské modifikaci lze do grafu zahrnout rovněž zóny vzdálenosti intervalů nejvyšší hustoty (HDI) od nuly. Dále je navržen nový typ souřadnicové reprezentace kompozičních dat, jehož cílem je zlepšit identifikaci biomarkerů. V souladu se svým názvem jsou tyto tzv. selektivní pivotové souřadnice konstruovány tak, že „vodící“ souřadnice agreguje všechny párové logpodíly odpovídající kompoziční složky s ostatními komponentami, s výjimkou aberantních logpodílů. Na souřadnice je následně jako zlatý standard mnohorozměrné analýzy -omických dat aplikována diskriminační analýza metodou částečných nejmenších čtverců. A konečně, složitější strukturu kompozičních dat uspořádaných podle dvou faktorů lze často považovat za kompoziční tabulku. Pro tato data je v práci uvedena speciální volba pivotových souřadnic reflektující možný rozklad tabulky na její nezávislou a interakční část. Za účelem redukce dimenze je pak použita robustní metoda hlavních komponent (PCA), která prostřednictvím přímého vztahu představených souřadnic s centrovanými logpodílovými koeficienty, jenž jsou v kontextu PCA s kompozičními daty tradičně užívány, umožňuje získat lepší vhled do vztahů mezi danými faktory.

Klíčová slova: kompoziční data, logpodílová metodika, centrované logpodílové koeficienty, pivotové souřadnice, vážené pivotové souřadnice, selektivní pivotové souřadnice, kompoziční tabulky, bayesovská statistika, robustní metoda hlavních komponent, vulkánový graf, metoda částečných nejmenších čtverců – diskriminační analýza, kompoziční biplot, metabolomická data, ekonomická data

1 Introduction

Compositional data (CoDa) are present in many applications from numerous scientific fields (e.g., economy, sociology, psychology, biology, geochemistry, environmental studies or so-called omics sciences covering metabolomics, genomics, proteomics, transcriptomics, and other branches producing high-throughput data). Logratio methodology based on the Aitchison geometry on simplex (Aitchison, 1986; Pawlowsky-Glahn et al., 2015; Filzmoser et al., 2018) can and should be used as a cornerstone every time the statistician works with strictly positive data carrying relative information. At the same time not only vectors but also more complex structures with the interest lying in several factors can be seen as CoDa.

The main bottleneck of the statistical analysis and its interpretation in all omics sciences is probably the high-dimensionality of their (compositional) data sets. Another specific of these sciences is the need for a thorough and substantial data pre-processing before any statistical methods can be even applied. This step includes also data transformation and/or normalization for which mainly (natural logarithm of) so-called probability quotient normalization (PQN) is used (Dieterle et al., 2006). Here the original data are expressed in terms of ratios to a median of components normalized with respect to some reference sample (usually composed from component-wise medians). The PQN representation is successfully seconded by logratio coordinates where the posed challenge is to find an appropriate counterpart within the logratio methodology to better reflect geometric properties of the relative omics data.

After data pre-processing, tools from both univariate and multivariate statistics are usually used for the analysis in metabolomic experiments aimed at discovering metabolites discriminating the group(s) of patients from healthy controls. In the article de Sousa et al. (2020), we presented a novel Bayesian approach to a univariate statistical analysis of untargeted metabolomic data expressed in first pivot coordinates (or clr coefficients which are up to a scaling constant equal to them) for a multiple hypotheses testing problem. One of the most widespread tools for biomarker identification in omics sciences is the so-called volcano plot (Cui and Churchill, 2003) functioning as a double filter: the size of effect given as a ratio of medians of the patient vs. control data (i.e., a fold-change)

is depicted against statistical significance represented by a negative decadic logarithm from p-values obtained in t-tests of all variables (metabolites). Unlike the traditional frequentist way of volcano plot construction, the proposed Bayesian approach does not need to rely on any p-value corrections to the number of multiple tests performed; the decision about a hypothesis is build on highest density intervals (HDI) working with the entire posterior distributions (Kruschke, 2013; Thulin, 2014). Another advantage is the robustness of the method (in Bayesian context) achieved through the prior assumption of the data distribution (Kruschke, 2013). For the construction of the volcano plot itself, we suggested to work with the mean values of posterior distributions as a measure of the size of the effect and with newly introduced b-values substituting the statistical significance. Furthermore, it was shown that a combination of the measures from both axes of the Bayesian volcano plot can be conveniently used in the final assessment of the potential biomarkers. As such, we proposed to construct so-called HDI zones, i.e., distances of the borders of HDI from zero.

The results of multivariate statistical methods in metabolomics (or generally also in other omics as well as for example in geochemistry) often suffer from the influence of a handful strong biomarkers on the other variables. An endeavor to eliminate this phenomenon led to a development of selective pivot coordinates (SPCs) presented in the article Štefelová et al. (2023). Pivot coordinates, here termed for better clarity as ordinary pivot coordinates (OPCs), follow a principle where the first ("pivoting") coordinate aggregates all logratios with the compositional part of interest, keeping an easy interpretation just like in the case of clr coefficients (Fišerová and Hron, 2011). At the same time, it is possible to create more systems of pivot coordinates (usually the same number as the number of compositional parts) which can be converted to each other by an orthogonal transformation (Filzmoser et al., 2018) and where the part of interest in the first coordinate is permuted. As a weighting technique for classification problems of high-dimensional CoDa, we suggested zero-one weights allowing to fully eliminate aberrant pairwise logratios of the compositional part of interest in its first SPC. The big advantage of such weighting is that SPCs results in OPC systems with just one difference – the pivoting coordinate of each system is generally no longer the first one. Therefore, SPCs can be seen as a certain orthogonal rotation of the original pivot coordinates. As for the particular choice of strategy to assign

the weights to the individual compositional parts, we chose Welch's t-statistics to determine individual logratios which should be eliminated from the pivoting coordinate of each set of SPCs (i.e., by assigning zero weights to the respective pairwise logratios). After constructing Welch-based SPCs, partial least squares – discriminant analysis was applied on the data as a well-established method for classification tasks in omics sciences. A comparison of sensitivity and specificity among logarithmized PQN, OPCs and SPCs was provided in a simulation with the newly proposed coordinates outperforming the others in both true positive and true negative rate, making them a very versatile transformation option.

More complex CoDa structures where the observations are carrying inherently relative information about data distribution on the basis of two (or even more) factors are not yet common in omics, geochemistry or biology. Nevertheless, to model for example a relative structure of unemployed people depending on their gender and age group, or a relative structure of university students among different study subjects with relation to the obtained university degree, could not be done otherwise. From the mathematical point of view, we talk about two-factorial extension of vector CoDa, called compositional tables (Egozcue et al., 2008, 2015). Using the logratio methodology, each compositional table can be decomposed into an independent and an interactive part and olr coordinates assigned to all of them (Fačevicová et al., 2016) enabling further statistical processing of compositional tables using popular multivariate methods. The comparison of independence and interaction tables is what allows for a better understanding of the original data which is why in the article de Sousa et al. (2021) we proposed a particular choice of pivot coordinates for all three compositional tables (i.e., the original table and its decomposed parts) with a direct link to clr coefficients including their explicit formulas and interpretation. This is a key step for an application of robust multivariate methods on two-factorial CoDa and since one of the most common tasks in statistics is a dimension reduction, we applied on the data expressed in the presented coordinates a robust principal component analysis. It requires to carry out the computations of loadings and scores using OPCs of vectorized compositional tables, as clr representation leads to singularity, and transform them to clr coefficients only afterward for the purpose of compositional biplots construction.

2 Summary of the state of the art

2.1 Logratio methodology of compositional data

A positive (row) vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_D)$ is defined to be a D-part composition if it carries relative information, i.e., the ratios between the components are informative (Aitchison, 1986; Pawlowsky-Glahn et al., 2015). Any compositional vectors with equal number of parts are considered to be representatives of the same *equivalence class* if one vector is obtained from another by a positive scalar multiplication (Pawlowsky-Glahn et al., 2015). This is an important point e.g., for some omics sciences where the total often might not be known. Accordingly, equivalence classes of compositional data are represented without loss of information in a D-part simplex,

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

for any $\kappa > 0$. The choice of κ (being 1 for proportions and 100 for percentages) is irrelevant for the analysis and can also vary throughout the compositional data set. The D-part simplex is thus a $(D-1)$-dimensional sample space of (representatives of equivalence classes of) compositions.

A closure operation $\mathcal{C}$ can be applied to rescale the data to a given constant sum representation. Therefore, the results of statistical processing should not depend on the sum κ of compositional parts and instead of the standard Euclidean distances which rely on absolute (squared) differences between components, relative differences are used to express distances between observations. This principle called *scale invariance* is the first of three basic compositional principles (Pawlowsky-Glahn et al., 2015). Moreover, the original data often contain some non-informative part(s) in the compositional vector that are not of interest. Hence, we do not expect any change of results concerning the respective subcomposition when these parts are removed from the data. *Subcompositional coherence* is a principle declaring that results obtained from a d-part subcomposition, $d < D$, are not in contradiction with results obtained by an analysis of the original D-part composition. Finally, *permutation invariance* states that the results are independent from a chosen order of parts within the composi-

tion, an anticipative premise for any reasonable statistical processing and one of the key assumptions for the idea behind the construction of pivot coordinates (Section 2.2).

The above principles and the relative scale of CoDa should be captured by a meaningful geometric structure, preferably following the properties of the Euclidean vector space. This is provided by the Aitchison geometry (Pawlowsky-Glahn and Egozcue, 2001; Egozcue et al., 2003). Instead of adapting the standard statistical methods to this specific geometry, it is rather preferred to firstly express CoDa in meaningful real coordinates and then proceed with further statistical processing; i.e., employing the *working on coordinates* principle (Mateu-Figueras et al., 2011).

Generally, there are three types of logratio coordinate representations respecting the Aitchison geometry with interpretation in terms of log-ratios or their aggregations, *centered logratio coefficients (clr), additive logratio coordinates (alr)* (Aitchison, 1986) and *orthonormal logratio coordinates (olr)* (Egozcue et al., 2003) defined as

$$\operatorname{alr}(\boldsymbol{x}) = \left(\ln \frac{x_1}{x_D}, \ln \frac{x_2}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D} \right), \tag{1}$$

$$\operatorname{clr}(\boldsymbol{x}) = \left(\ln \frac{x_1}{g(\boldsymbol{x})}, \ln \frac{x_2}{g(\boldsymbol{x})}, \ldots, \ln \frac{x_D}{g(\boldsymbol{x})} \right), \tag{2}$$

$$\boldsymbol{z} = \operatorname{olr}(\boldsymbol{x}) = \left(\left\langle \boldsymbol{x}, \boldsymbol{e}^1 \right\rangle_A, \left\langle \boldsymbol{x}, \boldsymbol{e}^2 \right\rangle_A, \ldots, \left\langle \boldsymbol{x}, \boldsymbol{e}^{D-1} \right\rangle_A \right), \tag{3}$$

where $g(\boldsymbol{x})$ stands for the geometrical mean of the whole composition and D-part compositions $\boldsymbol{e}^i = \mathcal{C}(e_1^i, e_2^i, \ldots, e_D^i)$, $i = 1, \ldots, D-1$, form an orthonormal basis on the simplex.

Clr representation keeps the metric properties of CoDa and enables for a simple and meaningful interpretation in terms of dominance of a given compositional part with respect to the other parts *on average*. Consequently, clr coefficients are useful for a graphical interpretation of compositional data including compositional biplots as a result of a dimension reduction through PCA (Aitchison and Greenacre, 2002) or a multiple hypotheses testing based Bayesian volcano plot (de Sousa et al., 2020). However, it is worth noting that clr coefficients sum up to zero which leads to a singular covariance matrix. This reflects dimensionality of com-

positions, which is just $D-1$ for D-part compositional data. Given the zero-sum condition, it is generally not desirable to analyze any clr part separately without considering the others nor to use clr coefficients with common robust statistical methods (Filzmoser et al., 2009; Filzmoser and Hron, 2013; de Sousa et al., 2021).

There is a linear transformation between olr coordinates and clr coefficients, done through a $D \times (D-1)$ matrix $\mathbf{V}$ of clr representations of the olr basis vectors (i.e., *logcontrast coefficients* defined generally as a linear combination of logarithmized parts with zero-sum constraint on the respective coefficients),

$$\mathrm{clr}(\boldsymbol{x}) = \mathbf{V}\boldsymbol{z} = [\mathrm{clr}(\boldsymbol{e}^1)^T, \mathrm{clr}(\boldsymbol{e}^2)^T, \ldots, \mathrm{clr}(\boldsymbol{e}^{D-1})^T] \cdot \mathrm{olr}(\boldsymbol{x})^T. \tag{4}$$

2.2 Pivot coordinates

To enable a link to clr coefficients within an olr coordinate system, *(ordinary) pivot coordinates (OPCs)*, $\boldsymbol{z}^{(l)} = (z_1^{(l)}, \ldots, z_{D-1}^{(l)})$, with $z_i^{(l)}$, $i = 1, \ldots, D-1$, given as

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

were introduced as a special case of olr coordinates (Fišerová and Hron, 2011; Hron et al., 2017). Here, $x_i^{(l)}$ refers to the i-th part of the re-ordered composition $(x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)$ which can be rewritten as $(x_1^{(l)}, x_2^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})$. This indicates that in each of the D coordinate systems, a permutation of compositional parts needs to be performed, so that the l-th part ($l = 1, \ldots, D$) of $\boldsymbol{x}$ stands at the first ("pivoting") position. It ensures that for each part of the original composition, the desired interpretation can be reached in one of the coordinate systems. The first OPC in each system, $z_1^{(l)}$, then clearly explains all relative information about part x_l and, additionally, it is proportional to the respective clr coefficient from the expression (2) as

$$z_1^{(l)} = \sqrt{\frac{D}{D-1}} \mathrm{clr}(\boldsymbol{x})_l, \tag{6}$$

being an extra asset in case of the univariate statistical analysis. Because OPCs are constructed "semi-automatically", they are certainly advantageous for high-

dimensional data and/or multifactorial CoDa structures.

OPCs can be rewritten in terms of pairwise logratios yielding

$$z_i^{(l)} = \frac{1}{\sqrt{(D-i+1)(D-i)}} \left[\ln\left(\frac{x_i^{(l)}}{x_{i+1}^{(l)}}\right) + \cdots + \ln\left(\frac{x_i^{(l)}}{x_D^{(l)}}\right) \right]. \tag{7}$$

As an alternative to this situation where all pairwise logratios in $z_1^{(l)}$ are treated with the same relevance, *weighted pivot coordinates (WPCs)* were proposed in Hron et al. (2017) with the objective to provide a possibility to enhance or mitigate the effect of some pairwise logratios with the compositional part of interest. If we rewrite the first OPC in the form of the expression (7) with weights $\alpha_j^{(l)}$, $j = 2, \ldots, D$ as

$$\alpha_2^{(l)} \ln \frac{x_1^{(l)}}{x_2^{(l)}} + \ldots + \alpha_D^{(l)} \ln \frac{x_1^{(l)}}{x_D^{(l)}}, \quad \alpha_2^{(l)}, \ldots \alpha_D^{(l)} \geq 0, \quad \alpha_2^{(l)} + \ldots + \alpha_D^{(l)} = 1,$$

the first WPC can be then obtained from here as follows

$$w_1^{(l)} = \frac{1}{\sqrt{1 + \sum_{j=2}^{D} \left(\alpha_j^{(l)}\right)^2}} \ln \frac{x_1^{(l)}}{\prod_{j=2}^{D} \left(x_j^{(l)}\right)^{\alpha_j^{(l)}}}. \tag{8}$$

A toll for the non-equal handling of the pairwise logratios with the pivoting compositional part is another coordinate involving $x_1^{(l)}$ where its remaining (relative) information not included in (8) gets stored, i.e., a residual coordinate $w_{D-1}^{(l)}$. While the general formulas for WPC $w_2^{(l)}, \ldots, w_{D-1}^{(l)}$ are computationally laborious to derive, the way to obtain them is to sequentially apply the orthonormal property of the corresponding logcontrast coefficients and the identity $\mathrm{clr}(\boldsymbol{e}^{i(l)})\mathbf{1}^T = 0$.

So far, there are two different weighting techniques presented in the literature, both arising from the limitations of OPCs in different practical applications. The first approach published together with the general WPCs formulation in Hron et al. (2017) reflects the need to filter some background noise in geochemical mapping where the calculated concentrations often suffer from measurement errors and imputed rounded zeros. Although this could be relatable also for some omics sciences, the chosen weight function based on the variation matrix would generally not work there, as in a majority of situations a certain response variable

needs to be considered together with the omics compositional data set. For regression tasks with high-dimensional compositional explanatory variables, where the response variable is continuous, a weighting approach taking into account the correlation structure of the data was proposed in Štefelová et al. (2021). Both these weighting schemes downplay the parts of the original composition which have some sort of a poor association with either the pivoting part or the response variable. However, they are not suitable for classification tasks. For the purpose of a categorical response variable coupled with high-dimensional CoDa from metabolomics, another weighting strategy, that can hopefully be seen as the "last piece missing" within the approach where pivot coordinates sophistically aggregate (some) information from all possible pairwise logratios, is presented in Section 4.2.

The geosciences where the usage of pairwise logratios still prevails motivate also the origin of *backwards pivot coordinates* published in Hron et al. (2021). Employing some kind of "reverse order" in the construction of pivot-like coordinates leads to a possibility of working with the desirable effects of simple logratios without sacrificing the orthonormality of olr coordinates required by many multivariate statistical methods. Starting with a choice of interpretable pairwise logratios (e.g., alr coordinates (1) with x_D as a normalizing geochemical element or any other reference role), an entire set of olr coordinates is built around each of them. This results in systems of $D-1$ backwards pivot coordinates

$$b_i^{(l')} = \sqrt{\frac{i}{i+1}} \ln \frac{\sqrt[i]{\prod_{j=1}^{i} x_j^{(l')}}}{x_{i+1}^{(l')}}, \quad i, l' = 1, \ldots, D-1,$$

which are just orthogonal rotations of each other like in the case of OPCs. The l'-th reordering of the parts of the original composition is chosen in such a way that the pivoting compositional part occupies the first position and the denominator x_D the second one, $\boldsymbol{x}^{(l')} = (x_{l'}, x_D, \ldots, x_{l'-1}, x_{l'+1}, \ldots, x_{D-1})$.

2.3 Compositional tables

Two-factorial extension of vector CoDa (Aitchison, 1986; Pawlowsky-Glahn et al., 2015) carrying information about a relationship between and within row

and column factors is called a *compositional table* $\boldsymbol{x}$,

$$\boldsymbol{x} = \begin{pmatrix} x_{11} & \cdots & x_{1J} \\ \vdots & \ddots & \vdots \\ x_{I1} & \cdots & x_{IJ} \end{pmatrix}, \quad x_{ij} > 0, i = 1, \ldots, I, j = 1, \ldots, J. \tag{9}$$

Since compositional tables form a direct extension of vector CoDa, all the principles introduced in Section 2.1 apply, up to some minor modifications due to the two-factorial structure of the tables. It is straightforward to derive that the dimension of the simplex $\mathcal{S}^{IJ}$ is $IJ - 1$, corresponding to the dimensionality of $(I \times J)$-compositional tables.

To analyze compositional tables, it is beneficial to work also with the so-called *independence* and *interaction* tables where their separate analysis can be advantageous for further interpretation concerning both factors and their relationships. These independent and interactive parts can be obtained from the original table (9) through an orthogonal decomposition $\boldsymbol{x} = \boldsymbol{x}_{ind} \oplus \boldsymbol{x}_{int}$ (Egozcue et al., 2008). Here, the independence table is constructed to extract all the relative information about row and column factors under the assumption that the original compositional table is a product of its row and column geometric marginals, and the interaction table contains information about the relationships between the row and column factors.

It is crucial to realize that the dimensions of $\boldsymbol{x}_{ind}$ and $\boldsymbol{x}_{int}$ lower to $I + J - 2$ and to $(I - 1)(J - 1)$, respectively. Hence, similarly to vector CoDa, an appropriate real coordinate representation of compositional tables, which in addition follows the decomposition into independent and interactive parts, needs to be established with respect to the the sample space dimensionality and the Aitchison geometry (Fačevicová et al., 2016).

Generally, there are three types of OPCs corresponding to the row, column and "odds ratio" partitioning of the compositional table (Fačevicová et al., 2016). The first two types jointly form a coordinate representation of the independence table, the third one is used for the interaction table. Altogether, they provide a coordinate representation of the original compositional table. In case of row and column types of coordinates, the entire first row or column, respectively, is taken as the pivoting element and separated from the rest. In the next step, this pivot is not considered anymore and the following row or column is taken as the new (reduced) pivoting element, and so on, until the following $I + J - 2$

coordinates are obtained,

$$z_i^r = \sqrt{\frac{(I-i)J}{1+I-i}} \ln \frac{g(\boldsymbol{x}_{i\bullet})}{[g(\boldsymbol{x}_{i+1\bullet}),\ldots,g(\boldsymbol{x}_{I\bullet})]^{1/(I-i)}}, \quad i = 1,\ldots,I-1,$$

$$z_j^c = \sqrt{\frac{I(J-j)}{1+J-j}} \ln \frac{g(\boldsymbol{x}_{\bullet j})}{[g(\boldsymbol{x}_{\bullet j+1}),\ldots,g(\boldsymbol{x}_{\bullet J})]^{1/(J-j)}}, \qquad j = 1,\ldots,J-1, \tag{10}$$

where $g(\boldsymbol{x}_{i\bullet})$ and $g(\boldsymbol{x}_{\bullet j})$ stand for the geometric mean of the i-th row and j-th column, respectively.

The process of obtaining the remaining $(I-1)(J-1)$ coordinates is based on a division of the original compositional table into four blocks, say upper left A, upper right B, lower left C and lower right D, where A contains always just one (pivot) cell indexed by rs. The odds ratio interpretation should be now easily seen from the following formula, where the elements of blocks A and D are in the numerator, and the elements of blocks B and C in the denominator of the logratio,

$$z_{rs}^{OR} = \sqrt{\frac{1}{(I-r)(J-s)(I-r+1)(J-s+1)}} \ln \prod_{i=r+1}^{I} \prod_{j=s+1}^{J} \frac{x_{ij}x_{rs}}{x_{is}x_{rj}}. \tag{11}$$

To obtain all OPCs of the odds ratio type in a proper order corresponding to the z^r and z^c coordinates (10), the position of the pivoting cell is moving firstly by rows with fixed first column, $r = 1,\ldots,I-1$, then by columns with fixed last row, $s = 1,\ldots,J-1$, and afterward the row position is always leveled back down by one and the column position moves again from 1 to $J-1$ for the given row until all sizes of the $r \times s$ table are covered.

Finally, permutations of the entire rows or columns following the same principle as stated in Section 2.2 could be performed. Hereby for all combinations of rows and columns, different OPC systems consisting of $z_i^{r(k)}$, $z_j^{c(l)}$ and $z_{rs}^{OR(kl)}$, where $(kl), k = 1,\ldots,I, l = 1,\ldots,J$, defines row and column permuted to the pivoting position within the whole table, would be gained (Fačevicová et al., 2016).

2.4 Robust principal component analysis for compositional data

One of the widely used methods for the purpose of dimension reduction of large-scale data sets in a compositional approach is PCA just like in the case of standard multivariate data analysis. It converts possibly correlated original variables from the data at hand into a smaller set of linearly uncorrelated variables called principal components (PCs). Additionally, the first component accounts for the largest variance of the given data, the second one for a maximum of the remaining variance, etc., under the constraint of being orthogonal to all the previous PCs (Johnson and Wichern, 2007).

The covariance matrix $\mathbf{C}$ estimated from a real data matrix $\mathbf{X}$ can be spectrally decomposed into $\mathbf{C} = \mathbf{GLG}^T$, where $\mathbf{G}$ is a matrix of eigenvectors and $\mathbf{L}$ represents a diagonal matrix of eigenvalues of $\mathbf{C}$. It is then possible to define the PCA transformation as $\mathbf{X}^* = (\mathbf{X} - \mathbf{1}^T\boldsymbol{t})\mathbf{G}$, where $\boldsymbol{t}$ is the (row) location estimator and $\mathbf{1}$ is a vector of ones with length n (number of observations). The columns of the matrix $\mathbf{X}^*$, the coordinates of the PCs, are called *scores* and the columns of $\mathbf{G}$, containing the respective basis vectors, are called *loadings*.

It is common to take $\boldsymbol{t}$ as the arithmetic mean and $\mathbf{C}$ as the sample covariance matrix. However, both are very sensitive to outlying observations. Robust alternatives can be obtained by using the MCD estimators of location and covariance (Maronna et al., 2006). Accordingly, robust principal component analysis (rPCA) of CoDa based on the MCD approach requires olr coordinates $\boldsymbol{z}_i$ as an input to obtain full rank data in order to get the MCD estimate of the covariance matrix and the respective matrix of eigenvectors $\mathbf{G}$. The scores $\boldsymbol{z}_i^*$ are then given by $\boldsymbol{z}_i^* = (\boldsymbol{z}_i - \boldsymbol{t})\mathbf{G}$. Once rPCA is performed, the loadings can be transformed back to clr coefficients as $\mathbf{G}_{\text{clr}} = \mathbf{VG}$, accounting for compositional biplot construction with meaningful interpretation, whereas the scores remain identical and only a column of zeros is added to the end. For the interpretation, the focus is on links between vertices of arrows as they stand for a proportionality between the original compositional parts (Aitchison and Greenacre, 2002). Due to the relation with OPCs, the single clr variables (or the respective loadings) can be used to identify observations with a high dominance of the respective parts in a compositional vector (Kynčlová et al., 2016).

3 Thesis objectives

In this dissertation thesis, the aim is to demonstrate the wide potential of the logratio methodology for statistical analysis of compositional data in various contexts under the common umbrella of complexity, resp. high-dimensionality of the relevant data sets using pivot logratio coordinates (Fišerová and Hron, 2011; Filzmoser et al., 2018) or their modifications. Depending on the type of task, first pivot coordinates can even be pragmatically replaced (especially in high dimensions) by clr coefficients sharing the same interpretation. Conversely, pivot coordinates can also be tuned by weighting to filter out aberrant pairwise logratios in classification tasks, or they can be generalized to the setting of compositional tables (i.e., two-factorial CoDa) and their orthogonal decomposition.

Accordingly, the developments presented in the thesis touch upon subjects such as Bayesian approach to the multiple hypotheses testing of CoDa in metabolomics using first OPCs, construction of a new type of pivot logratio coordinates using zero-one weighting technique for the improvement of biomarker identification, or a particular choice of OPCs for compositional tables complying with the decomposition process of two-factorial CoDa and providing a direct link to the respective clr coefficients in context of dimension reduction using robust principal component analysis. All theoretical developments are accompanied with both simulated data studies and empirical data sets to demonstrate benefits of the new approaches (not included in this summary).

4 Theoretical framework and applied methods

4.1 Bayesian multiple hypotheses testing in compositional analysis of high-dimensional data

We suggest here a Bayesian counterpart to the popular univariate statistical analysis of omics data sets using multiple hypotheses testing. Classical t-test assumes a normal distribution of each of the two groups of samples which is, however, not appropriate for a description of any data containing outliers. Because t-distribution can be much heavier tailed (depending on degrees of freedom ν, called a *normality parameter* in Bayesian statistics (Kruschke, 2013)), it seems

to be more convenient. It turns out that it is a suitable choice also for the logratio representation of metabolomic data (please note that the original measurements, i.e., strictly positive data, could hardly be characterized by a t-distribution whose domain is the whole real line). Given the linear transformation (4) between olr coordinates and clr coefficients, specifically the relation (6) in the univariate context, it is sufficient to work with the clr representation instead of OPCs when the intepretation in terms of dominance of a compositional part with respect to averaged contributions of the others is preferable. However, the mental step leading to considering first pivot coordinates in place of the respective clr coefficients is still recommendable as univariate analysis with clr coefficients is inappropriate due to their zero-sum constraint which distorts the covariance structure.

In Bayesian t-test, each of the two groups of samples, i.e., clr represented patients and controls, has its own mean μ_{pat} and μ_{con}, respectively, whose difference is of the main interest, and its own standard deviation σ_{pat} and σ_{con}. The normality parameter is shared by both groups (Kruschke, 2013). To make a qualified decision about the null hypothesis stating no difference in means among the tested samples, all five model parameters need to be inferred.

Prior distributions of the parameters are taken as non-informative to allow already a moderate amount of data to deflect the original setting into the direction driven by the evidence (Kruschke, 2014). This is in line with the situation of untargeted metabolomics where it is prevailing not to have any well-founded prior knowledge for a vast majority of the measured features.

The inference is driven by the Bayes' rule stating the posterior to be proportional (up to an integration constant) to the likelihood times prior,

$$f(\mu_{\text{pat}}, \sigma_{\text{pat}}, \mu_{\text{con}}, \sigma_{\text{con}}, \nu|X) \propto f(X|\mu_{\text{pat}}, \sigma_{\text{pat}}, \mu_{\text{con}}, \sigma_{\text{con}}, \nu) \times f(\mu_{\text{pat}}, \sigma_{\text{pat}}, \mu_{\text{con}}, \sigma_{\text{con}}, \nu), \qquad (12)$$

where the joint prior distribution density $f(\mu_{\text{pat}}, \sigma_{\text{pat}}, \mu_{\text{con}}, \sigma_{\text{con}}, \nu)$ can be, assuming independent parameters, rewritten as a product of marginal densities of the single parameters. This assumption permits to take the posterior density simply as a product of prior parameter distribution densities, and t-distributed probability density reflecting the data evidence X, making this an important step simplifying the computations.

In practice, posterior density is numerically approximated by a class of *Mar-*

kov chain Monte Carlo methods (MCMC) (Gelman et al., 2013) which generates samples from the (non-normalized) posteriors (12), corresponding to both the data and the priors.

The final decision concerning the null hypothesis is very intuitive in Bayesian hypothesis testing with the use of credible sets (Thulin, 2014); for example *highest density interval (HDI)*, which can be formally defined by inequality $\mathrm{P}(\mu_{\mathrm{pat}} - \mu_{\mathrm{con}} \in \Theta_{\mathrm{HDI}}|X) \geq 1-\alpha$, is constructed to contain 95 % of the most frequented posterior values Θ_{HDI}. Since the resulting MCMC chain of differences between means of clr representation of both original groups of samples can be plotted into a histogram, it may easily be computed where those Θ_{HDI} values are allocated. If this interval does not contain zero, the hypothesis about equality of parameters μ_{pat} and μ_{con} is rejected and the posterior distributions are accepted to be significantly different. Moreover, the sign of the majority of HDI values further reveals the direction of this difference.

Multiple testing complicates the situation since, except for the hypotheses rejection, we also seek some importance order of metabolites based on the results of the analysis. This can be done simply according to *means of posterior distribution (MPD) criterion* which is a mean of a difference of posteriors of given parameters $\mu_{\mathrm{pat}}, \mu_{\mathrm{con}}$. However, it would lead to a serious loss of information if the complex posterior distribution was reduced just to its MPD value. In addition, empirical probabilities that the differences in μ_{pat} and μ_{con} would have an opposite sign than indicated by posterior distributions can be considered. Even though it is inappropriate to sort the metabolites using just p-values obtained from classical t-tests (Wasserstein and Lazar, 2016), some ordering based on the above-mentioned probabilities, which we suggest to call *b-values*, can be performed. Formally, we propose to define

$$b\text{-value} = \min\left\{\mathrm{P}(\mu_{\mathrm{pat}} - \mu_{\mathrm{con}} > 0), \mathrm{P}(\mu_{\mathrm{pat}} - \mu_{\mathrm{con}} < 0)\right\}, \tag{13}$$

where the probabilities are computed from the MCMC posterior distribution. An analogous procedure was proposed to quantify the evidence against the rejected hypothesis when computing the largest credible set which does not contain those values of the tested parameter θ that are valid just under the assumption of the null hypothesis; say credible set Θ_T without values θ_0. Then a probability $\mathrm{P}(\theta \notin \Theta_T|X) = \alpha_{\min}$, where $\alpha_{\min}$ is the smallest α ensuring that the credible set

Θ_T does not contain θ_0, has a very similar meaning to the p-value from a traditional t-test whilst considering the entire posterior distribution, in particular also its tails (De Bragança Pereira and Stern, 1999; Thulin, 2014). The above-suggested b-value (13) could be seen as a certain variation to this idea, using the smaller part of HDI divided into two intervals by $\theta_0 = \mu_{\text{pat}} - \mu_{\text{con}} = 0$ as an empirical probability of a realization of the posterior on the other side of the zero value.

Both MPD values and b-values are at disposal for the final choice of potential biomarkers from all original metabolites.

4.2 Selective pivot logratio coordinates for PLS-DA modeling

In the context of binary classification problems involving CoDa, the idea that motivates the development of *selective pivot coordinates (SPCs)* is to have logratio coordinates that represent relevant relative information about x_l, but aggregate only informative pairwise logratios including x_l in the first coordinate. That is, given that each pairwise logratio involves two distinct compositional parts, the aim is to include into an SPC, denoted by ${}_{(l)}s$, only those that agree with what the majority of logratios with x_l suggest about its ability to distinguish between the two groups of observations. Namely, in a biomedical setting, having two groups (patient and control), a compositional part should be identified as a biomarker candidate if most pairwise logratios involving that part are significantly higher in one group than in the other.

Given a CoDa matrix consisting of N observations from two different groups, we propose to use the ordinary Welch's t-statistic (Welch, 1947) to determine the least relevant logratios. Denoting ${}_{(l)}\boldsymbol{T} = ({}_{(l)}T_1, \ldots, {}_{(l)}T_{l-1}, {}_{(l)}T_{l+1}, \ldots, {}_{(l)}T_D)$ the set of such t-statistics corresponding to logratios $\left(\ln\frac{x_l}{x_1}, \ldots, \ln\frac{x_l}{x_{l-1}}, \ln\frac{x_l}{x_{l+1}}, \ldots, \ln\frac{x_l}{x_D}\right)$, the criterion is to exclude those logratios for which the statistic ${}_{(l)}T_d$, $d = 1, \ldots, D,\ d \neq l$ lays outside the interval $\left[{}_{(l)}\theta_1;\ {}_{(l)}\theta_2\right]$. These boundaries are computed as

$$ {}_{(l)}\theta_1 = \begin{cases} -\infty, & \text{if } q\left({}_{(l)}\boldsymbol{T}; 1-\xi\right) < t_{N-2}(0.025) \\ \text{med}\left({}_{(l)}\boldsymbol{T}\right) - 2\text{Q}_n\left({}_{(l)}\boldsymbol{T}\right), & \text{otherwise,} \end{cases} $$

and

$$ {}_{(l)}\theta_2 = \begin{cases} \infty, & \text{if } q\left({}_{(l)}\boldsymbol{T}; \xi\right) > t_{N-2}(0.975) \\ \text{med}\left({}_{(l)}\boldsymbol{T}\right) + 2\text{Q}_n\left({}_{(l)}\boldsymbol{T}\right), & \text{otherwise,} \end{cases} $$

where $q\left({}_{(l)}\boldsymbol{T}; \alpha\right)$ is the α-quantile of ${}_{(l)}\boldsymbol{T}$, $\text{med}\left({}_{(l)}\boldsymbol{T}\right) = q\left({}_{(l)}\boldsymbol{T}; 0.5\right)$ and $t_{N-2}(\alpha)$ is the α-quantile of the Student's t-distribution with $N-2$ degrees of freedom (the parameter ξ is set to 0.1 by default). Moreover, Q_n stands for the robust scale estimator of Rousseeuw and Croux (1993), i.e. $\text{Q}_n\left({}_{(l)}\boldsymbol{T}\right)$ is given by about the first quartile of the absolute differences $\{|{}_{(l)}T_c - {}_{(l)}T_d|,\ 1 \leq c < d \leq D,\ c, d \neq l\}$. The interval for exclusion then results to be $\left[\text{med}\left({}_{(l)}\boldsymbol{T}\right) \pm 2\text{Q}_n\left({}_{(l)}\boldsymbol{T}\right)\right]$, unless more than 90% of the t-statistic values are roughly either lower than -2 or higher than 2. Where this latter happens, only the upper (resp. lower) cut-off values are used.

Note that this additional condition aims to ensure that logratios involving two strong biomarkers with discriminating effect in opposite directions (i.e., increased and decreased in the group of patients, respectively) are not excluded from the aggregation. Pairwise logratios of such compositional parts would likely have a good discriminating effect in the consequent statistical analysis, nonetheless, they might be flagged as outliers among all the logratios with one of the components in the numerator. Therefore, these are deviating logratios that should be preserved in the respective SPC. A higher value of ξ can be chosen if this undesirable effect is still apparent (for example in data sets with higher ratio of potential biomarkers of the opposite directions).

For the following, let us denote the number of selected logratios including x_l as ${}_{(l)}M$, the parts in the denominator of the selected logratios as ${}_{(l)}x_1^+, \ldots, {}_{(l)}x_{{}_{(l)}M}^+$, and the remaining parts as ${}_{(l)}x_1^-, \ldots, {}_{(l)}x_{D-1-{}_{(l)}M}^-$. To obtain ${}_{(l)}s,\ l = 1, \ldots, D$, the original composition $\boldsymbol{x}$ needs to be rearranged as ${}_{(l)}\boldsymbol{x} = \left({}_{(l)}x_1^-, \ldots, {}_{(l)}x_{D-1-{}_{(l)}M}^-, x_l, {}_{(l)}x_1^+, \ldots, {}_{(l)}x_{{}_{(l)}M}^+\right)$.

Then, an OPC system ${}_{(l)}\boldsymbol{z} = \left({}_{(l)}z_1, \ldots, {}_{(l)}z_{D-1}\right)$ is set up for ${}_{(l)}\boldsymbol{x}$. To de-

fine SPCs, the *pivoting* coordinate is no longer the first one but the one at the $(D - {}_{(l)}M)$-th position, denoted by ${}_{(l)}z_{D-{}_{(l)}M}$. Accordingly, the SPC of interest is obtained as

$$
\begin{aligned}
{}_{(l)}s = {}_{(l)}\, z_{D-{}_{(l)}M} &= \sqrt{\frac{{}_{(l)}M}{{}_{(l)}M+1}}\ln\frac{x_l}{\sqrt[{}_{(l)}M]{\prod_{k=1}^{{}_{(l)}M} {}_{(l)}x_k^+}} \\
&= \frac{1}{\sqrt{({}_{(l)}M+1)\cdot {}_{(l)}M}}\left(\ln\frac{x_l}{{}_{(l)}x_1^+} + \cdots + \ln\frac{x_l}{{}_{(l)}x_{{}_{(l)}M}^+}\right), \quad l = 1,\ldots,D. \qquad (14)
\end{aligned}
$$

The proposed SPCs can also be seen as a special case of WPCs (Hron et al., 2017) introduced in Section 2.2 where weights of either 1 or 0 are assigned to each logratio involving x_l, depending on whether it is included in the aggregation or not, respectively. Consequently, ${}_{(l)}s$ can be written in the form of expression (8) as

$$
{}_{(l)}s = \sqrt{\frac{{}_{(l)}M}{{}_{(l)}M+1}}\ln\frac{x_l}{\sqrt[{}_{(l)}M]{\prod_{\substack{d=1\\ d\neq l}}^{D}(x_d)^{{}_{(l)}\gamma_d}}}, \quad l = 1,\ldots,D,
$$

with weights given by

$$
{}_{(l)}\gamma_d = \begin{cases} 1, & \text{if } {}_{(l)}T_d \in \left[{}_{(l)}\theta_1;\ {}_{(l)}\theta_2\right] \\ 0, & \text{otherwise.} \end{cases}
$$

Building on the OPC-based approach introduced in Kalivodová et al. (2015), partial least squares discriminant analysis (PLS-DA) through SPCs given as (14) is used here for the actual identification of biomarker candidates. Thus, D models of the form

$$
Y = \beta_0 + {}_{(l)}\beta_1 \cdot {}_{(l)}z_1 + \ldots + {}_{(l)}\beta_{D-1} \cdot {}_{(l)}z_{D-1} + \varepsilon, \quad l = 1,\ldots,D,
$$

are considered, where Y is a binary response representing each of the two groups, the explanatory variables ${}_{(l)}z_1 \ldots, {}_{(l)}z_{D-1}$ are logratio coordinates from the l-th SPC system, $\beta_0, {}_{(l)}\beta_1, \ldots, {}_{(l)}\beta_{D-1}$ are unknown model coefficients and ε is the usual random error term. Note that, unlike with PLS-DA based on OPCs where the procedure can be computationally simplified by fitting one model in clr coef-

ficients and then take advantage of their direct relationship with OPCs in form of the equation (6), SPCs require the successive models to be fitted individually. Before fitting the PLS model, the data are mean centred so that the intercept β_0 can be excluded from further considerations. The optimal number of PLS components is chosen here based on a randomization test approach (van der Voet, 1994). From each model fit, for $l = 1, \ldots, D$, the estimate ${}_{(l)}\hat{\beta}_{D-(l)M}$ associated with the SPC ${}_{(l)}s$ is extracted, and statistical significance is determined by bootstrap-based significance testing on the standardized PLS model coefficients (Kalivodová et al., 2015). The resulting p-values are adjusted using the Benjamini and Hochberg's method (Benjamini and Hochberg, 1995) to control for false discovery rate in multiple testing.

4.3 Robust principal component analysis for compositional tables

As stated in the Sections 2.3 and 2.4, such a coordinate representation which respects the sample space dimensionality as well as the decomposition procedure is needed to perform rPCA of compositional tables. Interestingly, the coordinates of the entire compositional table given in (10) and (11) can be divided into two groups according to the dimensionality of the independence and interaction tables, respectively. This becomes the main advantage also when using OPCs for rPCA since it allows for a comparison of the results from the whole table and its decomposed parts.

Following the link (6) between the first OPC and the respective clr coefficient of vector CoDa, also the first coordinates of the three types from each system can then be expressed as proportional (up to a constant) to respective clr coefficients,

$$\mathrm{clr}(\boldsymbol{x}_{ind})_{kl} = \sqrt{\frac{I-1}{IJ}} z_1^{r(k)} + \sqrt{\frac{J-1}{IJ}} z_1^{c(l)},$$

$$\mathrm{clr}(\boldsymbol{x}_{int})_{kl} = \sqrt{\frac{(I-1)(J-1)}{IJ}} z_{11}^{OR(kl)},$$

which is an important fact for the interpretation of the analysis.

The resulting clr coefficients, computed originally from the elements of the

independence and interaction tables,

$$\mathrm{clr}(\boldsymbol{x}_{ind})_{ij} = \ln \frac{x_{ij}^{ind}}{g(\boldsymbol{x}_{\bullet\bullet}^{ind})}, \quad \mathrm{clr}(\boldsymbol{x}_{int})_{ij} = \ln \frac{x_{ij}^{int}}{g(\boldsymbol{x}_{\bullet\bullet}^{int})},$$

can be expressed also in terms of cells of the input compositional table as

$$\mathrm{clr}(\boldsymbol{x}_{ind})_{ij} = \ln \frac{g(\boldsymbol{x}_{i\bullet})g(\boldsymbol{x}_{\bullet j})}{g(\boldsymbol{x}_{\bullet\bullet})^2}, \quad \mathrm{clr}(\boldsymbol{x}_{int})_{ij} = \ln \frac{x_{ij}g(\boldsymbol{x}_{\bullet\bullet})}{g(\boldsymbol{x}_{i\bullet})g(\boldsymbol{x}_{\bullet j})}, \tag{15}$$

where $g(\boldsymbol{x}_{i\bullet})$, $g(\boldsymbol{x}_{\bullet j})$ and $g(\boldsymbol{x}_{\bullet\bullet})$ stand for the geometric mean of the i-th row ($i = 1, \ldots, I$), the j-th column ($j = 1, \ldots, J$) and the whole compositional table (and its independent and interactive counterparts for $g(\boldsymbol{x}_{\bullet\bullet}^{ind})$, $g(\boldsymbol{x}_{\bullet\bullet}^{int})$), respectively. As a consequence, each $\mathrm{clr}(\boldsymbol{x}_{ind})_{ij}$ expresses a dominance of a given combination of factor values in case of independence. This dominance is then either amplified or weakened according to the interaction table which depends on whether the interaction is shifted in a positive or a negative direction. The interaction table refers also to sources of departures from independence, nevertheless, the information obtained only from $\mathrm{clr}(\boldsymbol{x}_{int})_{ij}$ does not provide a complete picture about the dominance of the respective cell to all other averaged cells.

There are only $I + J - 2$ and $(I - 1)(J - 1)$ linearly independent clr coefficients in the case of independence and interaction tables, respectively, reflecting the dimensionality of their sample spaces. Since this dependency makes it impossible to use the clr coefficients for the rPCA of the decomposed tables, the strategy to perform rPCA for compositional tables is the same as in case of vector CoDa: PCA loadings and scores are computed in olr coordinates (OPCs) and then back-transformed using relation (4) to the clr space, where the loadings can be interpreted in terms of dominance of single cells. Here, clr coefficients of basis vectors for rows $\boldsymbol{e}^r$, columns $\boldsymbol{e}^c$ and interactions $\boldsymbol{e}^{OR}$, forming the columns of the matrix $\mathbf{V}$, are defined as follows,

$$\mathrm{clr}(\boldsymbol{e}^r) = \begin{cases} \sqrt{\frac{I-i}{(I-i+1)J}} & \text{for the elements in pivot row } i, \\ -\sqrt{\frac{1}{(I-i+1)J(I-i)}} & \text{for the elements in rows } i+1, \ldots, I, \\ 0 & \text{otherwise,} \end{cases} \tag{16}$$

$$\operatorname{clr}(\boldsymbol{e}^{c}) = \begin{cases} \sqrt{\frac{J-j}{(J-j+1)I}} & \text{for the elements in pivot column } j, \\ -\sqrt{\frac{1}{(I-i+1)J(I-i)}} & \text{for the elements in columns } j+1,\ldots,J, \\ 0 & \text{otherwise,} \end{cases} \tag{17}$$

and

$$\operatorname{clr}(\boldsymbol{e}^{OR}) = \begin{cases} \sqrt{\frac{1}{rs(r-1)(s-1)}} & \text{for the elements on positions } i=r+1,\ldots,I, \\ & j=s+1,\ldots,J \\ \sqrt{\frac{(r-1)(s-1)}{rs}} & \text{for the pivot elements } rs \\ -\sqrt{\frac{r-1}{rs(s-1)}} & \text{for the elements in pivot row } r, \\ & j=s+1,\ldots,J, \\ -\sqrt{\frac{s-1}{rs(r-1)}} & \text{for the elements in pivot column } s, \\ & i=r+1,\ldots,I, \\ 0 & \text{otherwise,} \end{cases} \tag{18}$$

reinterpreting the expressions from Fačevicová et al. (2016). As a result of (4), row-wise clr coefficients of the whole table are obtained for the $IJ-1$ columns of the matrix $\mathbf{V}$. Alternatively, if the matrix $\mathbf{V}$ has just $I+J-2$ columns formed by clr coefficients of basis vectors corresponding to the OPC representation of the independence table (10), its respective clr coefficients are derived (and similarly for the interaction table with its coordinates (11)). Finally, the transformed loadings and scores can be used to construct a biplot in order to reveal the multivariate structure of the sample of compositional tables and relations between both factors.

5 Original results and summary

As suggested by the name of this thesis, it contributes advanced novel methods in the analysis of so-called compositions, i.e., data carrying relative information. Since the nature of CoDa endowed with Aitchison geometry entails fundamentally different approach to their statistical treatment, logratio methodology is used as a sound and necessary basis for the statistical analysis. The newly introduced tools are applied in the fields of science where high-dimensional

data are a daily bread, namely metabolomics and econometrics.

First, a new approach to univariate statistical analysis of (untargeted) metabolomic data, introducing a Bayesian version of a popular double-filtering graphical tool called volcano plot (not included in this summary) coupled with logratio data representation was proposed. Although interpretability of clr coefficients would be fully satisfactory there, the univariate analysis is geometrically only reasonable when first OPCs are used instead. Further, it was explained (not included in this summary) that the Bayesian counterpart to the multiple hypotheses testing might solve some of the problems occurring in frequentist analysis of high-dimensional data such as the inappropriateness of the routinely used p-value corrections for multiple testing or sensitivity of the traditional methods to outlying observations. Also, even if all limitations of the frequentist approach were over-passed, the poverty of the information provided as a result of each hypothesis test is notable in the contrast to Bayesian approach producing the whole posterior distribution. Decision made on behalf of Bayesian inference is, therefore, always more competent, because it is based on much richer information compared to a single number from the traditional hypothesis testing.

Next, since classification problems with CoDa have led to duly justified criticism of the OPC approach, commonly resulting in poorer sensitivity and specificity than competitors based on data normalization (such as PQN in the metabolomic field), a new type of pivot coordinates was proposed in the thesis. These so-called selective pivot coordinates exclude from the aggregation such pairwise logratios that are determined by Welch's t-statistic–based intervals as deviating from the main pattern. Hence, SPCs demonstrate the value in considering more complex logratios involving the compositional part of interest, while still retaining the intuitive idea of aggregating relative information into one (pivoting) logratio coordinate. Moreover, they further stress how the flexibility of the logratio approach built on well-founded geometrical grounds can outperform *ad hoc* solutions. Also, as shown, the method is connected as a particular zero-one weighting case with the broader framework of WPCs, which is able to deal with the drawbacks of OPCs in regression tasks. That is why the SPC approach presented here for classification problems somehow closes the circle, having now covered most common CoDa analysis and modeling situations met in the metabolomics context and beyond.

Finally, rPCA of compositional tables as a two-factorial generalization of vector CoDa was studied. Given that compositional tables can be decomposed onto their independence and interaction parts, a statistical analysis of both is recommended to get insight into the ideal situation when relationships amid factors are filtered away, as well as into interactions between factors forming the original compositional table. As most practical data sets contain outlying observations, robust methods requiring an orthonormal coordinate representation have been considered. To reduce the dimension of data at hand, rPCA using the MCD estimator can be applied to pivot coordinates of compositional tables according to their decomposition into independence and interaction tables. The necessity of respecting dimensionality of the independent and interactive parts presents the main difference to (vector) CoDa where such feature does not occur. It was precisely this need of specific choice of olr coordinates where coordinates of independence and interaction tables form together coordinates of the entire compositional tables which allowed here for the additional benefit brought by the linkage of OPCs to clr coefficients constructed in the same manner. Thereafter, loadings obtained in OPCs for the rPCA were transformed back to clr coefficients where they were used for the construction of compositional biplots and their meaningful analysis. In case of $(2 \times J)$ table dimensions (not included in this summary), an additional feature could be observed in the graphical output of interaction tables, which was traced back to the interpretation of the clr coefficients as well.

The good performance of the novel methods was always shown on the analyses of two different dimension–relatable data sets (from the field of rare metabolic diseases and economy, respectively; not included in this summary). For the Bayesian volcano plot and SPCs, simulation studies (not included in this summary) were also provided to compare the stability and sensitivity and specificity, respectively, of the proposed tools with the traditional approaches to the presented tasks. In both cases, the results of the simulations highlighted the potential of the new methods.

All computations in the thesis were performed using the environment of the statistical software `R` (R Core Team, 2022). The related codes are available online at `https://github.com/sousaju/BayesVolcano`, `https://github.com/sousaju/SPC`, and `https://github.com/sousaju/rPCA-CoDaTables`.

List of publications

- de Sousa J*, Vencálek O, Hron K, Václavík J, Friedecký D, Adam T (2020) Bayesian Multiple Hypotheses Testing in Compositional Analysis of Untargeted Metabolomic Data. Analytica Chimica Acta 1097: 49–61. DOI: 10.1016/j.aca.2019.11.006.
  * Corresponding author

- de Sousa J*, Fačevicová K, Hron K, Filzmoser P (2021) Robust Principal Component Analysis for Compositional Tables. Journal of Applied Statistics 48(2):1–20. DOI: 10.1080/02664763.2020.1722078.
  * Corresponding author

- Štefelová N, de Sousa J*, Hron K, Palarea-Albaladejo J, Dobešová D, Kvasnička A, Friedecký D (2023) Selective Pivot Logratio Coordinates for PLS-DA Modelling with Applications in Metabolomics. *Under review.*
  * Corresponding author

- Kouřil Š*, de Sousa J*, Václavík J, Friedecký D, Adam T (2020) CROP: Correlation-based Reduction of Feature Multiplicities in Untargeted Metabolomic Data. Bioinformatics 36(9):2941–2942. DOI: 10.1093/bioinformatics/btaa012.
  * Joint first authors

- Václavík J, Mádrová L, Kouřil Š, de Sousa J, Brumarová R, Janečková H, Jáčová J, Friedecký D, Knapková M, Kluijtmans L A J, Grünert S C, Vaz F M, Janzen N, Wanders R J A, Wevers R A, Adam T (2020) A newborn screening approach to diagnose 3-hydroxy-3-methylglutaryl-CoA lyase deficiency. JIMD Reports 54(1):79–86. DOI: 10.1002/jmd2.12118.

- Mádrová L, Součková O, Brumarová R, Dobešová D, Václavík J, Kouřil Š, de Sousa J, Friedecká J, Friedecký D, Barešová V, Zikánová M, Adam T (2022) Combined Targeted and Untargeted Profiling of HeLa Cells Deficient in Purine De Novo Synthesis. Metabolites 12(3):241. DOI: 10.3390/metabo12030241.

- Kouřil Š, de Sousa J, Fačevicová K, Gardlo A, Muehlmann C, Nordhausen K, Friedecký D, Adam T (2023) Multivariate Independent Component Analysis Identifies Patients in Newborn Screening Equally to Adjusted Reference Ranges. *Under review.*

List of conferences

- Robust 2016, 11.–16.9.2016, Rejhotice, Loučná nad Desnou (CZ): Analýza kategoriálních dat – problém vícenásobné volby v odpovědi (poster + presentation, in Czech)
- ODAM 2017, 31.5.–1.6.2017, Olomouc (CZ): Dimension reduction for compositional tables using robust principal component analysis (presentation)
- CoDaWork 2017, 5.–9.6.2017, Abbadia San Salvatore (IT): Robust principal component analysis for compositional tables (poster)
- MOVISS 2017, 21.–24.9.2017, Vorau (AT): Bayesian counterpart to t-tests in compositional analysis of metabolomic data (poster)
- ERCIM 2017, 16.–18.12.2017, London (UK): Reducing dimension of compositional tables by robust principal component analysis (presentation)
- MOVISS 2018, 9.–12.9.2018, Vorau (AT): Removing false features in metabolomics data using correlations (poster)
- Robust 2018, 21.–26.1.2018, Rybník, Hostouň (CZ): Bayesovský přístup k t-testům v kompoziční analýze metabolomických dat (poster + presentation, in Czech, Robust award)
- Seminari Aitchison 2018, 17.5.2018, Girona (ES): Bayesian approach to t-tests in compositional analysis of metabolomic data (presentation)
- Seminari de DEIO 2018, 8.6.2018, Barcelona (ES): Compositional analysis of metabolomic data, Bayesian counterpart to t-tests (presentation)
- CzechMS 2019, 27.–29.3.2019, Olomouc (CZ): Bayesian approach to statistical analysis of inherited metabolomic disorders (poster + presentation)
- CRoNoS 2019, 14.–16.4.2019, Limassol (CY): Compositional analysis of untargeted metabolomic data using multiple Bayesian hypotheses testing (presentation)

- ODAM 2019, 29.–31.5.2019, Olomouc (CZ): Bayesian multiple hypotheses testing in compositional analysis of untargeted metabolomic data (presentation)
- CoDaWork 2019, 3.–7.6.2019, Terrassa (ES): Bayesian approach to univariate analysis of inherited metabolomic disorder HMGCLD (poster)

Reference

Aitchison J (1986) The statistical analysis of compositional data. Chapman & Hall, London, DOI 10.1007/978-94-009-4109-0

Aitchison J, Greenacre M (2002) Biplots of compositional data. Journal of the Royal Statistical Society Series C: Applied Statistics 51(4):375–392, DOI 10.1111/1467-9876.00275

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) 57(1):289–300, DOI 10.1111/j.2517-6161.1995.tb02031.x

Cui X, Churchill GA (2003) Statistical tests for differential expression in cDNA microarray experiments. Genome Biology 4, DOI 10.1186/gb-2003-4-4-210, Article 210

De Bragança Pereira C, Stern J (1999) Evidence and credibility: Full Bayesian significance test for precise hypotheses. Entropy 1(4):99–110, DOI 10.3390/e1040099

de Sousa J, Vencálek O, Hron K, Václavík J, Friedecký D, Adam T (2020) Bayesian multiple hypotheses testing in compositional analysis of untargeted metabolomic data. Analytica Chimica Acta 1097:49–61, DOI 10.1016/j.aca.2019.11.006

de Sousa J, Hron K, Fačevicová K, Filzmoser P (2021) Robust principal component analysis for compositional tables. Journal of Applied Statistics 48(2):214–233, DOI 10.1080/02664763.2020.1722078

Dieterle F, Ross A, Schlotterbeck G, Senn H (2006) Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in ^{1}H NMR metabonomics. Analytical Chemistry 78(13):4281–4290, DOI 10.1021/ac051632c

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric logratio transformations for compositional data analysis. Mathematical Geology 35(3):279–300, DOI 10.1023/A:1023818214614

Egozcue JJ, Díaz Barrero JL, Pawlowsky-Glahn V (2008) Compositional analysis of bivariate discrete probabilities. In: Daunis-i Estadella J, Martín-Fernández JA (eds) Proceedings of CODAWORK'08, The 3rd Compositional Data Analysis Workshop, University of Girona, Spain

Egozcue JJ, Pawlowsky-Glahn V, Templ M, Hron K (2015) Independence in contingency tables using simplicial geometry. Communications in Statistics-Theory and Methods 44(18):3978–3996, DOI 10.1080/03610926.2013.824980

Fačevicová K, Hron K, Todorov V, Templ M (2016) Compositional tables analysis in coordinates. Scandinavian Journal of Statistics 43(4):962–977, DOI 10.1111/sjos.12223

Filzmoser P, Hron K (2013) Robustness for compositional data. In: Becker C, Fried R, Kuhnt S (eds) Robustness and Complex Data Structures, Springer, Berlin Heidelberg, DOI 10.1007/978-3-642-35494-6_8

Filzmoser P, Hron K, Reimann C (2009) Principal component analysis for compositional data with outliers. Environmetrics: The Official Journal of the International Environmetrics Society 20(6):621–632, DOI 10.1002/env.966

Filzmoser P, Hron K, Templ M (2018) Applied compositional data analysis. Springer Series in Statistics, Springer, Cham, DOI 10.1007/978-3-319-96422-5

Fišerová E, Hron K (2011) On the interpretation of orthonormal coordinates for compositional data. Mathematical Geosciences 43:455–468, DOI 10.1007/s11004-011-9333-x

Gelman A, Stern HS, Carlin JB, Dunson DB, Vehtari A, Rubin DB (2013) Bayesian data analysis, 3rd edn. Chapman and Hall/CRC press, New York, DOI 10.1201/b16018

Hron K, Filzmoser P, de Caritat P, Fišerová E, Gardlo A (2017) Weighted pivot coordinates for compositional data and their application to geochemical mapping. Mathematical Geosciences 49(6):797–814, DOI 10.1007/s11004-017-9684-z

Hron K, Coenders G, Filzmoser P, Palarea-Albaladejo J, Faměra M, Matys Grygar T (2021) Analysing pairwise logratios revisited. Mathematical Geosciences 53(7):1643–1666, DOI 10.1007/s11004-021-09938-w

Johnson RA, Wichern DW (2007) Applied multivariate statistical analysis, 6th edn. Pearson Prentice Hall, Upper Saddle River

Kalivodová A, Hron K, Filzmoser P, Najdekr L, Janečková H, Adam T (2015) PLS-DA for compositional data with application to metabolomics. Journal of Chemometrics 29(1):21–28, DOI 10.1002/cem.2657

Kruschke JK (2013) Bayesian estimation supersedes the t test. Journal of Experimental Psychology: General 142(2):573–603, DOI 10.1037/a0029146

Kruschke JK (2014) Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan, 2nd edn. Academic Press, Boston, DOI 10.1016/B978-0-12-405888-0.09999-2

Kynčlová P, Filzmoser P, Hron K (2016) Compositional biplots including external non-compositional variables. Statistics 50(5):1132–1148, DOI 10.1080/02331888.2015.1135155

Maronna RA, Martin RD, Yohai VJ (2006) Robust Statistics: Theory and Methods. John Wiley & Sons, New York, DOI 10.1002/0470010940

Mateu-Figueras G, Pawlowsky-Glahn V, Egozcue JJ (2011) The principle of working on coordinates. In: Pawlowsky-Glahn V, Buccianti A (eds) Compositional Data Analysis: Theory and Applications, John Wiley & Sons, Chichester, DOI 10.1002/9781119976462

Pawlowsky-Glahn V, Egozcue JJ (2001) Geometric approach to statistical analysis on the simplex. Stochastic Environmental Research and Risk Assessment (SERRA) 15(5):384–398, DOI 10.1007/s004770100077

Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015) Modeling and analysis of compositional data. John Wiley & Sons, Chichester, DOI 10.1002/9781119003144

R Core Team (2022) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.r-project.org

Rousseeuw P, Croux C (1993) Alternatives to the median absolute deviation. Journal of the American Statistical Association 88(424):1273–1283, DOI 10.2307/2291267

Štefelová N, Palarea-Albaladejo J, Hron K (2021) Weighted pivot coordinates for partial least squares-based marker discovery in high-throughput compositional data. Statistical Analysis and Data Mining: The ASA Data Science Journal 14(4):315–330, DOI 10.1002/sam.11514

Štefelová N, de Sousa J, Hron K, Palarea-Albaladejo J, Dobešová D, Kvasnička A, Friedecký D (2023) Selective pivot logratio coordinates for PLS-DA modelling with applications in metabolomics. *Under review*

Thulin M (2014) Decision-theoretic justifications for Bayesian hypothesis testing using credible sets. Journal of Statistical Planning and Inference 146:133–138, DOI 10.1016/j.jspi.2013.09.014

van der Voet H (1994) Comparing the predictive accuracy of models using a simple randomization test. Chemometrics and Intelligent Laboratory Systems 25(2):313–323, DOI 10.1016/0169-7439(94)00084-V

Wasserstein RL, Lazar NA (2016) The ASA's statement on p-values: context, process, and purpose. The American Statistician 70(2):129–133, DOI 10.1080/00031305.2016.1154108

Welch BL (1947) The generalization of 'Student's' problem when several different population variances are involved. Biometrika 34(1–2):28–35, DOI 10.2307/2332510