PALACKÝ UNIVERSITY IN OLOMOUC
FACULTY OF SCIENCE

DISSERTATION THESIS

Compositional approach to the analysis of data in biostatistics

Supervisor: **prof. RNDr. Karel Hron, Ph.D.**
Co-supervisor: **Dr. Javier Palarea-Albaladejo**
Author: **Mgr. Nikola Štefelová**
Study program: P1104 Applied Mathematics
Field of study: Applied Mathematics
Form of study: Full-time
The year of submission: 2021

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Mgr. Nikola Štefelová

Název práce: Kompoziční přístup v analýze biostatistických dat

Typ práce: disertační práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

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

Školitel specialista: Dr. Javier Palarea-Albaladejo

Rok obhajoby práce: 2021

Abstrakt:

Mnoho typů dat v biostatistice má podobu kompozičních dat, tj. jde o mnohorozměrná pozorování obsahující kladné složky, které reprezentují části nějakého celku a nesou relativní informaci. Logpodílová metodika, zohledňující specifické vlastnosti kompozic, slouží jako vhodný prostředek k jejich analýze. Tato práce představuje metodologické inovace a aplikace kompozičního přístupu v oborech biostatistiky, konkrétně v oblasti regresní analýzy a vizualizace dat, a to v metabolomice a při zkoumání vlivu pohybového chování na zdraví. Je zde prezentována nová robustní metoda pro regresi s kompozičními vysvětlujícími proměnnými, jež je schopna efektivně pracovat s pozorováními, která jsou odlehlá jako celek, i s těmi, kde se odlehlost projevuje pouze na prvkové úrovni. Také je tu představena vylepšená procedura pro identifikaci statisticky významných proměnných ve vysoce-dimenzionálních kompozičních datech. Je založena na regresi metodou částečných nejmenších čtverců (PLS regresi), při níž se pro reprezentaci kompozic využívají vážené pivotové souřadnice s novou strategií pro vážení danou povahou problému. V kontextu výzkumu pohybového chování je kladen zvláštní důraz na vhodnou souřadnicovou reprezentaci dat o pohybovém chování. Navržený souřadnicový systém bere v potaz to, že mezi kategoriemi pohybového chování, kompozičními proměnnými, existuje přirozené uspořádání.

Klíčová slova: kompoziční data, logpodílová metodika, bilance, pivotové souřadnice, vážené pivotové souřadnice, regresní analýza, PLS biplot, robustní statistika, odlehlá pozorování na úrovni buněk, data o pohybovém chování, metabolomická data

Počet stran: 94

Počet příloh: 0

Jazyk: anglický

BIBLIOGRAPHICAL IDENTIFICATION

Author: Mgr. Nikola Štefelová

Title: Compositional approach to the analysis of data in biostatistics

Type of thesis: dissertation thesis

Department: Department of Mathematical Analysis and Application of Mathematics

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

Co-supervisor: Dr. Javier Palarea-Albaladejo

The year of presentation: 2021

Abstract: Many types of data in biostatistics meet properties of compositional data, i.e. multivariate observations comprising positive parts of a whole carrying relative information. Given their specific properties, the logratio methodology serves as a proper tool for the analysis of compositions. This thesis presents methodological developments and applications of the compositional approach in fields of biostatistics, specifically in relation to regression analysis and data visualization as applied to metabolomics and time-use epidemiology. A novel method for regression with compositional explanatory variables is introduced, which is robust against rowwise as well as against cellwise outliers. Further, a procedure for improved biomarker discovery in high-dimensional compositional data is presented. It is based on partial least squares (PLS) regression using a weighted pivot coordinate representation for compositions with a new, task-driven, strategy for weighting. In the context of time-use research, special relevance is given to a suitable coordinate representation of time-use data. The proposed coordinate system aims to reflect the fact that there is a natural ordering in time-use categories, the compositional variables.

Key words: compositional data, logratio methodology, balances, pivot coordinates, weighted pivot coordinates, regression analysis, PLS biplot, robust statistics, cellwise outliers, time-use data, metabolomic data

Number of pages: 94

Number of appendices: 0

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. and Dr. Javier Palarea-Albaladejo. 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 or the same academic title.

In Olomouc

Acknowledgment

I would like to thank to all who helped and supported me during my Ph.D. study and research, especially to my supervisors Karel Hron and Javier Palarea-Albaladejo. I am very grateful to all co-authors of my scientific papers. Also, I greatly appreciate my working visits to Edinburgh kindly allowed by BioSS.

Contents

Introduction

Compositional data (compositions) are essentially characterized by their relative nature. They represent vectors of stricly positive values describing parts of some whole. Accordingly, the relevant information is contained in the ratios between the compositional parts. Due to specific sample space of compositional data and their geometry, compositions require different statistical processing than standard multivariate observations conveying absolute information (in terms of interval scale). A suitable approach for their analysis is the logratio methodology (Aitchison, 1986; Pawlowsky-Glahn et al., 2015). Its cornerstone lies in the construction of logratio coordinates that enable to express compositions as real-valued vectors, to which standard statistical methods can be applied. The choice of interpretable coordinates leading to meaningful results is of particular importance.

The aim of this thesis is to present the compositional approach and innovative methods within the logratio methodology suited to the analysis of biostatistical data, i.e. data involving living systems. There are many types of biostatistical data of compositional character. Here, two cases are considered: 1) molecular biology data concerning metabolites, i.e. small molecules involved in metabolism and 2) time-use movement behaviour data which reflect how people spend their time in terms of sleep, sedentary behaviour and physical activity of various intensities.

The first chapter of the thesis provides an overview of compositional data analysis. It introduces compositional data, their properties, sample space and geometry. Next, different coordinate representations for compositions are discussed with emphasis on particular isometric logratio coordinates - balances, pivot coordinates and weighted pivot coordinates. Finally, fundamental ideas behind several types of linear regression with explanatory variables including a composition (here termed as *compositional linear regression*) are presented.

In the second chapter, a novel method for robust compositional regression is introduced that is able to deal not only with outlying observations comprising whole observations (rowwise outliers) but also with outliers in individual cells (cellwise outliers) (Štefelová et al., 2021a). The proposed algorithm is described in detail, its use is demonstrated in application to metabolomic data and its performance is further assessed by a simulation study.

The third chapter presents a new weighting strategy for the construction of weighted pivot coordinates that is particularly suitable for PLS-based marker discovery in high-dimensional compositional biomolecular data (Štefelová et al., 2021b). The benefits of the proposal are illustrated using real metabolomic data as well as using simulated datasets.

The fourth chapter demonstrates the use of the compositional approach in the context of time-use epidemiology. Robust techniques for compositional descriptive statistics, visualization and linear regression are applied to analyse wake-time movement behaviour data (Štefelová et al., 2018). Strong emphasis is placed on a proper coordinate representation of time-use data considering a natural ordering of the given compositional parts. A new concept of pivoting balances is developed that, in combination with an adapted formulation of compositional PLS biplot, enables meaningful visualization of more complex time-use patterns and their relationships with an outcome variable (Štefelová et al., 2021c).

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

- Štefelová N, Dygrýn J, Hron K, Gába A, Rubín L, Palarea-Albaladejo J (2018) Robust compositional analysis of physical activity and sedentary behavior data. International Journal of Environmetal Research and Public Health 15(10):2248, DOI 10.3390/ijerph15102248

- Štefelová N, Alfons A, Palarea-Albaladejo J, Filzmoser P, Hron K (2021) Robust regression with compositional covariates including cellwise outliers. Advances in Data Analysis and Classification, DOI 10.1007/s11634-021-00436-9

- Štefelová N, Palarea-Albaladejo J, and Hron K (2021) Weighted pivot coordinates for PLS-based marker discovery in high-throughput compositional data. *Under review*

- Štefelová N, Palarea-Albaladejo J, Hron K, Gába A, Dygrýn J (2021) Compositional PLS biplot based on pivoting balances: a graphical tool to examine the association between 24-hour movement behaviours and health outcomes. *Under review*

In addition to these methodological papers, below are listed further papers from an interdisciplinary work in time-use epidemiology, where the focus was primarily to investigate a concrete problem in time-use movement behavior research using the logratio methodology:

- Pelclová J, Štefelová N, Pechová J, Dygrýn J, Gába A, Zając-Gawlak I (2018) Reallocating Time from Sedentary Behavior to Light and Moderate-to-Vigorous Physical Activity: What Has a Stronger Association with Adiposity in Older Adult Women? International Journal of Environmetal Research and Public Health 15(7):1444, DOI 10.3390/ijerph15071444

- Cuberek R., Pelclová J, Gába A, Pechová J, Svozilová Z, Přidalová M, Štefelová N Hron K (2019) Adiposity and changes in movement-related behaviors in older adult women in the context of the built environment: a protocol for prospective cohort study. BMC Public Health 19:1522, DOI 10.1186/s12889-019-7905-8

- Pelclová J, Štefelová N, Dumuid D, Pedišić Ž, Hron K, Gába A, Olds T, Pechová J, Zając-Gawlak I, Tlučáková L (2020) Are longitudinal reallocations of time between movement behaviors associated with adiposity among elderly women? A compositional isotemporal substitution analysis. International Journal of Obesity 44(4):857–864, DOI 10.1038/s41366-019-0514-x

- Gába A, Pedišić Ž, Štefelová N, Dygrýn J, Hron K, Dumuid D, Tremblay M (2020) Sedentary behavior patterns and adiposity in children: A study based on compositional data analysis. BMC Pediatric 20:147, DOI 10.1186/s12887-020-02036-6

- Gába A, Dygrýn J, Štefelová N, Rubín L, Hron K, Jakubec J, Pedišić Ž (2020) How do short sleepers use extra waking hours? A compositional analysis of 24-hour time-use patterns among children and adolescents. International Journal of Behavioral Nutrition and Physical Activity 17:104, DOI 10.1186/s12966-020-01004-8

- Gába A, Pelclová J, Štefelová N, Přidalová M, Zając-Gawlak I, Tlučáková L, Pechová J, Svozilová Z (2020) Prospective study on sedentary behavior patterns and changes in body composition parameters in older women:

A compositional and isotemporal substitution analysis. Clinical Nutrition, DOI 10.1016/j.clnu.2020.10.020

- Gába A, Dygrýn J, Štefelová N, Rubín L, Hron, K, Jakubec L (2021) Replacing school and out-of-school sedentary behaviors with physical activity, and its associations with adiposity in children and adolescents: A compositional isotemporal substitution analysis. Environmental Health and Preventive Medicine, 26(1):16, DOI 10.1186/s12199-021-00932-6

- Germano-Soares AH, Tassitano R, Farah B, Andrade-Lima A, Correia M, Gába A, Štefelová N, Puech-Leão P, Wolosker N, Cucato G, Ritti-Dias R (2021) Reallocating time from sedentary behavior to physical activity in patients with peripheral artery disease: analyzing the effects on walking capacity using compositional data analysis. Journal of Physical Activity & Health, DOI 10.1123/jpah.2020-0487

- Pelclová J, Štefelová N, Olds T, Dumuid D, Hron K, Chastin S, Pedišić Ž (2021) A study on prospective associations between adiposity and 7-year changes in movement behaviors among older women based on compositional data analysis. BMC Geriatrics, 21(1):203, DOI 10.1186/s12877-021-02148-3

- Gallo J, Lošťák J, Gába A, Dygrýn J, Baláž L, Štefelová N (2021) Accelerometer-based measures of 24-hour movement behaviors in patients before total knee arthroplasty: A study based on compositional data analysis. *Under review*

1 Compositional data analysis

A vector $\boldsymbol{x} = (x_1, \ldots, x_D)^\top$ is called a D-part composition when all its elements are strictly positive real numbers that carry relative information (Aitchison, 1986; Pawlowsky-Glahn et al., 2015). Accordingly, the absolute values of the parts are not important for the analysis and the relevant information is captured in the ratios between them. The compositional parts, representing quantitatively contributions to some whole, are co-dependent as within a given representation the change in one part necessarily affects the relative values of the remaining ones.

Compositional data are *scale invariant* which means that if the composition is multiplied by a positive number, the ratios between its parts are not altered. Consequently, the sample space of compositions is formed by equivalence classes of proportional vectors (Pawlowsky-Glahn et al., 2015). Therefore, compositions can be represented without loss of information as vectors with an arbitrary sum of components (typically 1 or 100 in case of proportions or percentages, respectively). The operation of rescaling the initial vector so that the components add up to a constant κ is called a *closure* with the formula

$$\mathcal{C}(\boldsymbol{x}) = \left(\frac{\kappa \cdot x_1}{\sum_{d=1}^{D} x_d}, \cdots, \frac{\kappa \cdot x_D}{\sum_{k=1}^{D} x_d} \right)^\top .$$

The resulting sample space of such constrained representation is a *simplex* defined as

$$\mathcal{S}^D = \{\boldsymbol{x} = (x_1, \ldots, x_D)^\top : x_1 > 0, \ldots, x_D > 0;\ \sum_{d=1}^{D} x_d = \kappa\},$$

which is a $(D-1)$-dimensional subset of the ordinary real space. Another property of compositional data is *permutation invariance* meaning that reordering of the compositional parts does not affect the information they contain. Lastly, *subcompositional coherence* refers to the fact that if only a subset of compositional parts (i.e. a subcomposition) is available, the information conveyed by this subcomposition should not be in contradiction with that coming from the full (original) composition, and more specifically, the distance between two subcompositions is not greater than the distance between the two original compositions

(we refer to *subcompositional dominance*) (Pawlowsky-Glahn et al., 2015; Filzmoser et al., 2018).

Compositions obey the so-called *Aitchison geometry* on the simplex (Pawlowsky-Glahn et al., 2015). Two basic operations within this geometry are called *perturbation* and *powering*. These are analogous to addition and scalar multiplication in the real space, respectively. For $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}^D$ and $\alpha \in \mathbb{R}$ they are defined as

$$\boldsymbol{x} \oplus \boldsymbol{y} = \mathcal{C}(x_1 \cdot y_1, \ldots, x_D \cdot y_D)^\top \quad \text{and} \quad \alpha \odot \boldsymbol{x} = \mathcal{C}(x_1^\alpha, \ldots, x_D^\alpha)^\top.$$

The triplet $\left(\mathcal{S}^D, \oplus, \odot\right)$ forms a vector space. The Euclidean vector space structure of the simplex is completed by *the Aitchison inner product, norm and distance* defined as

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_\mathrm{A} = \frac{1}{2D} \sum_{c=1}^{D} \sum_{d=1}^{D} \ln \frac{x_c}{x_d} \ln \frac{y_c}{y_d}, \quad \|\boldsymbol{x}\|_\mathrm{A} = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_\mathrm{A}} \quad \text{and} \quad \mathrm{d}_\mathrm{A}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} \ominus \boldsymbol{y}\|_\mathrm{A},$$

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

When analysing compositional data, their specific nature should be taken into account. The direct use of standard statistical methods relying on the Euclidean geometry in real space would lead to misleading results and conclusions (Filzmoser et al., 2018). One approach is to develop the counterparts to the standard methods within the Aitchison geometry on the simplex. For example, having a (N, D)-matrix $\mathbf{X} = (x_{nd})$ representing a compositional dataset of N observed D-part compositions, the compositional mean, called *center*, is computed as a (closed) column-wise geometric mean

$$\overline{\boldsymbol{x}} = \mathcal{C}\left(\left(\prod_{n=1}^{N} x_{n1} \right)^{\frac{1}{N}}, \ldots, \left(\prod_{n=1}^{N} x_{nD} \right)^{\frac{1}{N}} \right)^\top.$$

That is, it replaces the standard column-wise arithmetic mean by the one in line with the Aitchison geometry: the sum is swapped with the product and the scalar multiplicator with the power. Furthermore, dividing each row of the dataset by the center (or in other words, perturbing each row by the center

powered by -1), yields *centered compositional data* with the new center shifted to $\mathcal{C}(1,\ldots,1)^\top$ called *barycenter* (Pawlowsky-Glahn et al., 2015; Filzmoser et al., 2018) and corresponding to the neutral element on the simplex.

1.1 Logratio coordinates

Another approach to the analysis of compositional data is the use of real-valued logratio coordinates. Since the characterisation of the simplex as Euclidean vector space allows to build an (orthonormal) basis of $\mathcal{S}^D$, $\boldsymbol{x} \in \mathcal{S}^D$ can be represented by coordinates with respect to such a basis. The key idea of the logratio methodology is to map compositions from the simplex into real space via logratio coordinates and then proceed with the statistical processing there (Filzmoser et al., 2018). If necessary, results can be mapped back to the simplex. Using logratios, instead of simply ratios as bearers of the elemental information, is advantageous as they map the range of a ratio from the positive real space onto the entire real space, symmetrise their values around zero and, moreover, inverse logratios provide the same information up to the sign, i.e. $\ln(x_c/x_d) = -\ln(x_d/x_c)$.

To obtain a generating system for building a basis of $\mathcal{S}^D$, we can take exponentials of the canonical basis of $\mathbb{R}^D$, i.e. $\{\tilde{\boldsymbol{e}}_1,\ldots,\tilde{\boldsymbol{e}}_D\}$, where

$$\tilde{\boldsymbol{e}}_j = \mathcal{C}\left(\underbrace{1,\ldots,1}_{j-1}, e, \underbrace{1,\ldots,1}_{D-j}\right)^\top, \quad j = 1,\ldots,D.$$

Then, D different bases can be built from this generating system that are given by its $D-1$ compositions, e.g. $\{\tilde{\boldsymbol{e}}_1,\ldots,\tilde{\boldsymbol{e}}_{D-1}\}$. Further, the Gram-Schmidt procedure (Egozcue et al., 2003) can be applied resulting into an orthonormal basis $\{\boldsymbol{e}_1,\ldots,\boldsymbol{e}_{D-1}\}$. Note that this will be just one of infinitely many possible orthonormal bases (Pawlowsky-Glahn et al., 2015).

Three basic coordinate systems are commonly used within the logratio methodology - *additive logratio coordinates (alr)*, *centered logratio coefficients (clr)* (Aitchison, 1986) and *isometric logratio coordinates (ilr)* (Egozcue et al., 2003)

that are defined as follows

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

$$\operatorname{clr}(\boldsymbol{x}) = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{d=1}^{D} x_d}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{d=1}^{D} x_d}} \right)^\top, \tag{2}$$

$$\operatorname{ilr}(\boldsymbol{x}) = (\langle \boldsymbol{x}, \boldsymbol{e}_1 \rangle_{\mathrm{A}}, \ldots, \langle \boldsymbol{x}, \boldsymbol{e}_{D-1} \rangle_{\mathrm{A}},)^\top = \boldsymbol{\Psi} \cdot \ln(\boldsymbol{x}), \tag{3}$$

where the rows of the $(D-1, D)$-matrix $\boldsymbol{\Psi}$, called *logcontrast coefficients*, are given by $\operatorname{clr}(\boldsymbol{e}_j),\ j = 1, \ldots, D-1$.

While alr are coordinates with respect to the basis of the simplex (but not to an orthonormal basis) and clr coefficients represent just coefficients with respect to the generating system (which leads to singular covariance matrix), ilr are coordinates with respect to an orthonormal basis of the simplex (Pawlowsky-Glahn et al., 2015). Note that in the alr coordinates defined in (1), x_D plays the role of the reference (ratioing) part, which corresponds to the given basis $\{\tilde{\boldsymbol{e}}_1, \ldots, \tilde{\boldsymbol{e}}_{D-1}\}$. But actually D different alr coordinate systems can be constructed taking different $x_d,\ d = 1, \ldots, D$ as the reference part (thus corresponding to the basis omitting $\tilde{\boldsymbol{e}}_d$ from the generating system). On the other hand, in case of ilr coordinates, there are infinitely many options for their construction (depending on the orthonormal basis chosen), thus $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{D-1}$ and $\boldsymbol{\Psi}$ in (3) refer to an arbitrarily chosen basis of the simplex and the associated matrix of logcontrast coefficients, respectively.

The mapping $\operatorname{alr} : \mathcal{S}^D \to \mathbb{R}^{D-1}$ is an isomorphism (but not an isometry) between $\mathcal{S}^D$ and $\mathbb{R}^{D-1}$; $\operatorname{clr} : \mathcal{S}^D \to \mathcal{A} \subset \mathbb{R}^D,\ \dim(\mathcal{A}) = D-1$, is an isomorphism as well as isometry between $\mathcal{S}^D$ and the $(D-1)$-dimensional subspace of $\mathbb{R}^D$; and $\operatorname{ilr} : \mathcal{S}^D \to \mathbb{R}^{D-1}$ is an isomorphism as well as isometry between $\mathcal{S}^D$ and $\mathbb{R}^{D-1}$ (Pawlowsky-Glahn et al., 2015). Accordingly, whereas operations in the simplex are translated into operations in the real vector space using any of the three mappings, inner product (and consequently norm and distance) are preserved

only with clr and ilr mappings. That is, for $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}^D$ and $\alpha, \beta \in \mathbb{R}$ it holds that

$$\begin{aligned}
\mathrm{alr}(\alpha \odot \boldsymbol{x} \oplus \beta \odot \boldsymbol{y}) &= \alpha \cdot \mathrm{alr}(\boldsymbol{x}) + \beta \cdot \mathrm{alr}(\boldsymbol{y}), \\
\mathrm{clr}(\alpha \odot \boldsymbol{x} \oplus \beta \odot \boldsymbol{y}) &= \alpha \cdot \mathrm{clr}(\boldsymbol{x}) + \beta \cdot \mathrm{clr}(\boldsymbol{y}), \\
\mathrm{ilr}(\alpha \odot \boldsymbol{x} \oplus \beta \odot \boldsymbol{y}) &= \alpha \cdot \mathrm{ilr}(\boldsymbol{x}) + \beta \cdot \mathrm{ilr}(\boldsymbol{y}),
\end{aligned}$$

but

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathrm{A}} = \langle \mathrm{clr}(\boldsymbol{x}), \mathrm{clr}(\boldsymbol{y}) \rangle = \langle \mathrm{ilr}(\boldsymbol{x}), \mathrm{ilr}(\boldsymbol{y}) \rangle \neq \langle \mathrm{alr}(\boldsymbol{x}), \mathrm{alr}(\boldsymbol{y}) \rangle.$$

Furthermore, an inverse mapping can be applied to transfer the compositions back to the simplex as

$$\mathrm{alr}^{-1}\left(\mathrm{alr}(\boldsymbol{x})\right) = \mathcal{C}\left(\exp\left(\mathrm{alr}(\boldsymbol{x})\right)^{\top}, 1\right)^{\top} = \boldsymbol{x}, \tag{4}$$

$$\mathrm{clr}^{-1}\left(\mathrm{clr}(\boldsymbol{x})\right) = \mathcal{C}\left(\exp\left(\mathrm{clr}(\boldsymbol{x})\right)\right) = \boldsymbol{x}, \tag{5}$$

$$\mathrm{ilr}^{-1}\left(\mathrm{ilr}(\boldsymbol{x})\right) = \mathcal{C}\left(\exp\left(\mathrm{ilr}(\boldsymbol{x})^{\top} \cdot \boldsymbol{\Psi}\right)\right)^{\top} = \boldsymbol{x}. \tag{6}$$

Although alr coordinates and clr coefficients are quite easily interpretable and are used in specific contexts, they are not compatible with certain multivariate statistical methods. The former are not eligible for techniques based on a metric assumption and the latter for methods where a singular covariance matrix represents an issue. The ilr coordinates avoid drawbacks of the former two representations and importantly, they are orthonormal coordinates. The fact, that different ilr coordinate systems are just orthogonal rotations of each other, is a useful property in statistical analysis. For example, in regression analysis, it enables the use of an arbitrary choice of ilr coordinates to obtain the required (unique) output. Moreover, affine equivariant robust (regression) estimators provide results invariant to the choice of ilr coordinates (Filzmoser et al., 2018).

Going back to the compositional mean, it can be equivalently computed within the logratio methodology by calculating column-wise arithmetic mean of the compositional dataset expressed in any logratio coordinates followed by the corresponding inverse mapping (and closure). If we want to reduce the influence of possible outliers in the dataset, we can compute *robust center* by applying robust Minimum Covariance Determinant (MCD) estimator of location,

which is computed from the subset of observations of a chosen size whose sample covariance matrix has the smallest determinant (Maronna et al., 2002; Filzmoser et al., 2018). Because of the affine equivariance, the MCD estimator of location applied on $\mathbf{X}$ in any ilr coordinates followed by the respective inverse mapping gives the same robust center regardless the choice of coordinate system. Accordingly, *robustly centered compositional data* can be obtained by dividing each row of $\mathbf{X}$ by the robust center.

For all the reasons mentioned above, ilr coordinates are preferable in most cases. Then, the crucial challenge is to construct interpretable coordinates tailored to the scientific question at hand.

1.1.1 Balances

The procedure known as sequential binary partition (SBP) can be applied to construct customized ilr coordinates called (compositional) balances (Egozcue and Pawlosky-Glahn, 2005). In the first step of the SBP process, the entire collection of compositional parts is divided into two disjoint subsets, with each subset summarised by the geometric mean of its components and going into the numerator and denominator, respectively, of a normalized logratio constituting the first balance. In the next steps, these subsets are further split into two mutually exclusive subgroups going into the numerator and denominator, respectively, of the subsequent balances. This process continues until only one-part subsets remain and $D-1$ balances are constructed.

The balance coordinates are represented by a real vector $\boldsymbol{b} = (b_1, \ldots, b_{D-1})^\top$ with

$$b_j = \sqrt{\frac{r_j s_j}{r_j + s_j}} \ln \frac{\sqrt[r_j]{\prod_{i=1}^{r_j} x_{j_i}^+}}{\sqrt[s_j]{\prod_{i=1}^{s_j} x_{j_i}^-}}, \quad j = 1, \ldots, D-1, \tag{7}$$

where $x_{j_i}^+$ and $x_{j_i}^-$ refers to the parts selected for the numerator and denominator, respectively, in the jth balance while r_j and s_j stands for the respective number of parts (Egozcue and Pawlosky-Glahn, 2005; Pawlowsky-Glahn et al., 2015).

The associated matrix of logcontrast coefficients $\boldsymbol{\Psi}$ has elements

$$\psi_{jd} = \begin{cases} +\frac{1}{r_j}\sqrt{\frac{r_j s_j}{r_j+s_j}}, & \text{if } x_d \in \{x_{j_i}^+,\ i = 1,\dots,r_j\}, \\ -\frac{1}{s_j}\sqrt{\frac{r_j s_j}{r_j+s_j}}, & \text{if } x_d \in \{x_{j_i}^-,\ i = 1,\dots,s_j\}, \\ 0 & \text{otherwise}, \end{cases}$$

$$j = 1,\dots,D-1,\ d = 1,\dots,D.$$

Two exemplary SBP for a 5-part composition are illustrated in Table 1.

Table 1: Example of two possible SBP for 5-part composition which results in (a) general balances and (b) special balances called pivot coordinates. Parts chosen for the numerator and denominator of the jth balance are coded $+$ and $-$, respectively; 0 indicates that the part is not included in the respective balance.

j	x_1	x_2	x_3	x_4	x_5	r_j	s_j
1	+	−	−	−	+	2	3
2	+	0	0	0	−	1	1
3	0	+	−	−	0	1	2
4	0	0	+	−	0	1	1

(a)

j	x_1	x_2	x_3	x_4	x_5	r_j	s_j
1	+	−	−	−	−	1	4
2	0	+	−	−	−	1	3
3	0	0	+	−	−	1	2
4	0	0	0	+	−	1	1

(b)

Balance coordinates are interpreted, as their name indicates, in terms of a balance (contrast) between two subsets of parts represented by their respective geometric means (Egozcue and Pawlosky-Glahn, 2005; Pawlowsky-Glahn et al., 2015). They can be constructed according to the scientific questions of interest and based on domain-specific knowledge, e.g. to represent meaningful trade-offs.

Sometimes we are interested in various balances which summarize information about the whole composition in different ways. In other words, we want to analyse within one statistical model the first balances from L different coordinate systems. Then, we can use effectively the rotation between orthonormal coordinate systems and construct desirable balances $\boldsymbol{b}^{(l)} = \left(b_1^{(l)},\dots,b_{D-1}^{(l)}\right)^\top$, $l = 1,\dots,L$ (the superscript here refers to the balance coordinate system) while our focus lies only on the first balance $b_1^{(l)}$ in each of the systems (Štefelová et al., 2021c). This leads to the idea of pivot coordinates.

1.1.2 Pivot coordinates

The procedure of extracting unique information from different orthonormal coordinate system is particularly applied with special balances called pivot coordinates (Fišerová and Hron, 2011). These are intended to highlight the role of a single compositional part relative to all the others in one (the first) coordinate. In SBP, one part is always set against the remaining ones as illustrated in Table 1b.

Given a composition $\boldsymbol{x}$, we can rearrange it so that the lth part is put at the first place and denote that composition as

$$\boldsymbol{x}^{(l)} = \left(x_1^{(l)}, \ldots, x_D^{(l)}\right)^\top = (x_l, x_2, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)^\top, \quad l = 1, \ldots, D.$$

Then, the corresponding pivot coordinates define a real vector $\boldsymbol{z}^{(l)} = \left(z_1^{(l)}, \ldots, z_{D-1}^{(l)}\right)^\top$, where

$$\begin{aligned} z_j^{(l)} &= \sqrt{\frac{D-j}{D-j+1}} \ln \frac{x_j^{(l)}}{\sqrt[D-j]{\prod_{d=j+1}^{D} x_d^{(l)}}} \\ &= \frac{1}{\sqrt{(D-j+1)(D-j)}} \left[\ln\left(\frac{x_j^{(l)}}{x_{j+1}^{(l)}}\right) + \cdots + \ln\left(\frac{x_j^{(l)}}{x_D^{(l)}}\right)\right] \\ &= \boldsymbol{u}_j^\top \ln\left(\boldsymbol{x}^{(l)}\right), \quad j = 1, \ldots, D-1, \; l = 1, \ldots, D, \end{aligned} \tag{8}$$

with

$$\boldsymbol{u}_j = \sqrt{\frac{D-j}{D-j+1}} \left(\underbrace{0, \ldots, 0}_{j-1}, 1, -\frac{1}{D-j}, \ldots, -\frac{1}{D-j}\right)^\top$$

representing the vectors of logcontrast coefficients, i.e. the rows of matrix $\boldsymbol{\Psi}$ associated with pivot coordinates (Filzmoser et al., 2018; Hron et al., 2017) .

Each first coordinate $z_1^{(l)}$ in the pivot coordinate system contains all the relative information about the lth compositional part. Specifically, this first coordinate is a scaled logratio of the part x_l of interest to the geometric mean of all the other $D-1$ parts, which is equivalent to the scaled sum of the $D-1$ pairwise logratios including the part of interest in the numerator as shown in (8). Thus, it

can be interpreted in terms of dominance of the l-th part with respect to an average (geometric mean) of the other parts (Fišerová and Hron, 2011; Filzmoser et al., 2018). Furthermore, there is a notable relation to clr coordinates as the lth clr coefficient equals to $\sqrt{(D-1)/D} \cdot z_1^{(l)},\ l = 1, \dots D$. Eventually, pivot coordinates can be expressed as standard logcontrasts $\boldsymbol{u}_j^\top \ln\left(\boldsymbol{x}^{(l)}\right)$, with $\boldsymbol{u}_j^\top \mathbf{1} = 0$, where vectors $\boldsymbol{u}_j$ are orthonormal, i.e. $\boldsymbol{u}_i^\top \boldsymbol{u}_j = \delta_{ij},\ i, j = 1, \dots, D-1$, with δ_{ij} being 1 if $i = j$ and 0 otherwise.

1.1.3 Weighted pivot coordinates

In the representation of the first pivot coordinate as a scaled sum of the $D-1$ pairwise logratios of $x_1^{(l)}$ over the other parts, the logratios are treated equally. However, the collection of logratios aggregated into that coordinate can include information from completely different processes. Therefore, a weighted counterpart to the ordinary pivot coordinates was introduced, namely the weighted pivot coordinates (Hron et al., 2017). These enable to weight the logratios aggregated into the first coordinate according to their relevance for the purpose of the analysis.

Accordingly, by using $\gamma_2^{(l)}, \dots, \gamma_D^{(l)}$ to denote the weights, the first weighted pivot coordinate $w_1^{(l)}$ is constructed by taking the weighted sum of pairwise logratios with $x_1^{(l)}$,

$$\gamma_2^{(l)} \ln\left(\frac{x_1^{(l)}}{x_2^{(l)}}\right) + \ldots + \gamma_D^{(l)} \ln\left(\frac{x_1^{(l)}}{x_D^{(l)}}\right),\ \gamma_2^{(l)}, \dots, \gamma_D^{(l)} > 0,\ \sum_{d=2}^{D} \gamma_d^{(l)} = 1,$$

which, after rescaling to a standard logcontrast, leads to the coordinate

$$w_1^{(l)} = \frac{1}{\sqrt{1 + \sum_{d=2}^{D} \left(\gamma_d^{(l)}\right)^2}} \ln \frac{x_1^{(l)}}{\prod_{d=2}^{D} \left(x_d^{(l)}\right)^{\gamma_d^{(l)}}} = \left(\boldsymbol{v}_1^{(l)}\right)^\top \ln\left(\boldsymbol{x}^{(l)}\right), \quad l = 1, \dots, D \tag{9}$$

with

$$\boldsymbol{v}_1^{(l)} = \frac{1}{\sqrt{1 + \sum_{d=2}^{D} \left(\gamma_d^{(l)}\right)^2}} \left(1, -\gamma_2^{(l)}, \dots, -\gamma_D^{(l)}\right)^\top$$

representing the first vector of logcontrast coefficients, i.e the first row of matrix $\boldsymbol{\Psi}^{(l)}$ associated with the lth system of weighted pivot coordinates (Hron et al., 2017).

The remaining elements to form a real vector of weighted pivot coordinates $\boldsymbol{w}^{(l)} = \left(w_1^{(l)}, \ldots, w_{D-1}^{(l)}\right)^\top$ are obtained sequentially by considering the orthonormal property of the logcontrast coefficients and the requirement for standard logcontrasts. That is, $w_j^{(l)} = \left(\boldsymbol{v}_j^{(l)}\right)^\top \ln\left(\boldsymbol{x}^{(l)}\right)$, $\left(\boldsymbol{v}_1^{(l)}\right)^\top \mathbf{1} = 0$, $\left(\boldsymbol{v}_i^{(l)}\right)^\top \boldsymbol{v}_j^{(l)} = \delta_{ij}$, $i, j = 1, \ldots, D-1$, $l = 1, \ldots, D$. Note that unlike in the case of ordinary pivot coordinates, weighted pivot coordinates, using the construction from Hron et al. (2017), contain two coordinates which capture information about the part of interest: $w_1^{(l)}$ and $w_{D-1}^{(l)}$. However, the former coordinate contains the relevant information, whereas the latter corresponds to just a redundant remainder (Hron et al., 2017).

1.2 Compositional linear regression

Regression analysis is one of the most widely used techniques in practical data analysis and statistical modelling. The object of linear regression is to model linear relationship between response (dependent) variable and explanatory (independent) variables, also called covariates or predictors (Härdle and Simar, 2012). The compositional data framework has three basic regression problems. These concern the relation between the real-valued response and compositional covariates, compositional response and real covariates, or between compositional parts themselves. In all instances, the logratio methodology serves as useful tool as, with compositions expressed in proper logratio coordinates, standard regression methods can be applied and interpretable results obtained (Filzmoser et al., 2018). Because of their properties, the ilr coordinates are preferable, especially balances or the (weighted) pivot coordinates.

Throughout this thesis, we mainly deal with the cases where explanatory variables are formed by, or at least include, a composition. In that case, we consider two data structures: column vector $\mathbf{y}$ of size N and $(N, D+P)$-matrix $\mathbf{A} = (\mathbf{1}, \mathbf{i}_1, \ldots, \mathbf{i}_{D-1}, \mathbf{c}_1, \ldots, \mathbf{c}_P)$. The vector $\mathbf{y}$ describes values of the response va-

riable on N objects. The first column of the so-called design matrix $\mathbf{A}$ is formed by ones (for the intercept term parameter) and the remaining columns combine values on the $D-1$ ilr coordinates and the P non-compositional covariates corresponding to the same N observations. The resulting linear regression model has the form

$$\mathbf{y} = \mathbf{A}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{10}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots \beta_{D-1+P})^\top$ is a vector of unknown $K = D + P$ regression coefficients and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_N)^\top$ is an error vector (Härdle and Simar, 2012; Filzmoser et al., 2018).

Often, the focus lies on L different first coordinates conveying information about compositional parts in a desirable way. Then, L different regression models are examined and information associated with the first coordinate from each system is extracted. That is, we have L models

$$\mathbf{y} = \mathbf{A}^{(l)}\boldsymbol{\beta}^{(l)} + \boldsymbol{\varepsilon}, \quad l = 1, \dots L, \tag{11}$$

where the design matrix $\mathbf{A}^{(l)} = \left(\mathbf{1}, \mathbf{i}_1^{(l)}, \dots, \mathbf{i}_{D-1}^{(l)}, \mathbf{c}_1, \dots, \mathbf{c}_P\right)$ contains values on the lth set of ilr coordinates and the regression coefficient vector $\boldsymbol{\beta}^{(l)} = \left(\beta_0, \beta_1^{(l)}, \dots, \beta_{D-1}^{(l)}, \beta_D, \dots \beta_{D-1+P}\right)^\top$ has, due to the orthogonality of different ilr coordinate systems, the same intercept term β_0 and the same coefficients corresponding to the non-compositional covariates in each model. Consequently, the vector of estimates $\left(\hat{\beta}_0, \hat{\beta}_1^{(1)}, \dots, \hat{\beta}_1^{(D)}, \hat{\beta}_D, \dots, \hat{\beta}_{D-1+P}\right)^\top$ is used for interpretation purposes. Note that also model fit measures like the coefficient of determination are invariant to the choice of orthonormal coordinate system. Commonly, the L different ilr systems represent D sets of pivot coordinates so that each time the emphasis is put on the coordinate isolating the relative information about one compositional part (Hron et al., 2012; Filzmoser et al., 2018).

Considering the case when the response is of compositional nature, it is converted in practice to the case of real response by using any specific coordinate from an ilr representation of the composition, e.g., the first pivot coordinate. Then, $\mathbf{y}$ in (10) is given by $\mathbf{i}_1$, $\mathbf{A} = (\mathbf{1}, \mathbf{c}_1, \dots, \mathbf{c}_P)$ and $\boldsymbol{\beta}$ is of size $K = P + 1$. Again, we can consider L different models – now with different first ilr coordinate

as the response. Of course, then the regression coefficients and model fit measures differ in each model (Müller et al., 2018; Filzmoser et al., 2018).

1.2.1 OLS compositional regression

The basic method for estimating coefficients in a linear regression model is the ordinary least squares (OLS) technique. It produces estimates that minimize the sum of squared residuals, i.e. the sum of squared differences between observed and predicted values of the response variable (Härdle and Simar, 2012).

Denoting the vector of residuals as $\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_N)$, with $\hat{\varepsilon}_n(\boldsymbol{\beta})$ implying that the n-th residual, $n = 1, \ldots, N$, depends on the parameter $\boldsymbol{\beta}$, the least square solution for (10) is given by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{n=1}^{N} \left(\hat{\varepsilon}_n(\boldsymbol{\beta})\right)^2 = \left(\mathbf{A}^\top \mathbf{A}\right)^{-1} \mathbf{A}^\top \mathbf{y}.$$

Accordingly, the sum of squared residuals is computed as

$$SSR = \sum_{n=1}^{N} \left(\hat{\varepsilon}_n\left(\hat{\boldsymbol{\beta}}\right)\right)^2 = \left(\mathbf{y} - \mathbf{A}\hat{\boldsymbol{\beta}}\right)^\top \left(\mathbf{y} - \mathbf{A}\hat{\boldsymbol{\beta}}\right).$$

Assuming independent identically distributed errors with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, we can test whether the explanatory variables are significant in explaining the response variable. Estimating σ^2 by

$$\hat{\sigma}^2 = \frac{SSR}{N - K - 1},$$

we can compute the variance estimate of $\beta_k,\ k = 0, 1, \ldots, K-1$ as

$$\widehat{\mathrm{var}}(\beta_k) = \hat{\sigma}^2 \left\{\left(\mathbf{A}^\top \mathbf{A}\right)^{-1}\right\}_{k+1,k+1},$$

with $\{\}_{k+1,k+1}$ denoting the $(k+1)$th element at the diagonal of the respective matrix. Then, the explanatory variable corresponding to β_k is considered significant in relation to the response if

$$\left|\frac{\hat{\beta}_k}{\sqrt{\widehat{\mathrm{var}}(\beta_k)}}\right| > t_{1-\alpha/2;N-K-1},$$

where $t_{1-\alpha/2;N-K-1}$ denotes $(1-\alpha/2)$-quantile of Student's t-distribution with $N-K-1$ degrees of freedom (Härdle and Simar, 2012) . The usual choice of the significance level is $\alpha = 0.05$.

1.2.2 MM compositional regression

In practice, a common issue is that the observed dataset contains outliers, i.e. individual values or entire multivariate observations that deviate considerably from the main cloud of data points. Unfortunately, outliers can greatly influence ordinary estimates of model parameters and may lead to unreliable results. A number of regression methods robust against outlying observations have been developed (Maronna et al., 2002). Among those, MM-regression (Yohai, 1987) is a popular choice as it produces highly efficient estimates (i.e. with small variance and thus high precision) with a high breakdown-point, concretely up to 0.5 (meaning that reliable results can be obtained even with 50% observations being contaminated).

For (10), the MM-estimator is determined as the M-estimator

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{n=1}^{N} \rho\left(\frac{\hat{\varepsilon}_n(\boldsymbol{\beta})}{\hat{\sigma}}\right), \quad l = 1, \ldots, D,$$

with $\hat{\sigma}$ being the scale M-estimator (thus the double M in the title) defined as solution of

$$\frac{1}{N}\sum_{n=1}^{N} \rho^*\left(\frac{\hat{\varepsilon}_n}{\hat{\sigma}}\right) = \delta,$$

where $\rho(\cdot)$ and $\rho^*(\cdot)$ are appropriate bounded loss functions and δ is a given constant (Maronna et al., 2002). The optimal estimator is found via the IRWLS (iteratively reweighted least squares) algorithm. Robust estimate with high breakdown point but possibly inefficient, e.g. S-estimator (Yohai, 1987), is taken for the initial value $\hat{\boldsymbol{\beta}}^{[0]}$. Until convergence is reached, the estimator is updated in each iteration t with the weighted least square equation

$$\hat{\boldsymbol{\beta}}^{[t]} = \left(\mathbf{A}^\top \boldsymbol{\Omega}^{[t-1]} \mathbf{A}\right)^{-1} \mathbf{A}^\top \boldsymbol{\Omega}^{[t-1]} \mathbf{y},$$

where $\mathbf{\Omega}^{[t-1]}$ is a diagonal (N, N)-matrix with weights $\omega_1^{[t-1]}, \ldots, \omega_N^{[t-1]}$ as entries given by

$$\omega_n^{[t-1]} = \frac{\rho'\left(\frac{\hat{\varepsilon}_n^{[t-1]}}{\hat{\sigma}^{[t-1]}}\right)}{\frac{\hat{\varepsilon}_n^{[t-1]}}{\hat{\sigma}^{[t-1]}}}, \quad n = 1, \ldots, N,$$

with ρ' denoting the derivative of ρ. Throughout this work we use Tukey's biweight loss function, with the initial estimator tuned for maximum breakdown point and the final estimator tuned for 95% efficiency.

1.2.3 PLS compositional regression and biplot

Partial least squares (PLS) regression enjoys wide popularity in areas such as chemometrics (Höskuldson, 1988), especially in the case where the number of explanatory variables is significantly larger than the number of observations. It aims to fit the relationship between response variable(s) and potentially many and/or highly correlated explanatory variables by finding a small number of latent factors that synthesize the relationship in lower dimension. The underlying assumption is that the observed data are generated by a process driven by this small number of latent factors, also known as PLS components. The values on the PLS components (scores) are linear combinations of the explanatory variables with parameters (loadings) determined in such a way that they maximize the covariance between the response and the explanatory variables. Once the model is fitted in the latent space, the regression coefficients associated with the original explanatory variables can be subsequently worked out and their significance investigated. Even if PLS regression is particularly useful for the analysis of high-dimensional data, it offers other features that make the method also appealing for datasets with a relatively small to moderate number of explanatory variables. This includes the capacity to handle multicollinearity and highly correlated explanatory variables, the ability to separate main information from noise, the no requirement of distributional assumptions for error terms and, last but not least, the possibility of visualizing data in low dimensions via a PLS biplot.

Before fitting the PLS model, the data are usually mean-centered so that the intercept is excluded from further considerations. Accordingly, the design matrix omits the column of ones. That is, for (10) we have centered $\mathbf{y}$, column-

centered $\mathbf{A} = (\mathbf{i}_1, \ldots, \mathbf{i}_{D-1}, \mathbf{c}_1, \ldots, \mathbf{c}_P)$ and $\boldsymbol{\beta} = (\beta_1, \ldots \beta_{D-1+P})^\top$ of size $K = D - 1 + P$. PLS decomposes the design matrix as

$$\mathbf{A} = \mathbf{F}\mathbf{G}^\top + \mathbf{E}_\mathbf{A},$$

where $\mathbf{F}$ (the score matrix) is of size $N \times Q$, $\mathbf{G}$ (the loading matrix) is of size $K \times Q$, with Q being the number of PLS components (Varmuza and Filzmoser, 2009). Q is usually selected based on cross-validated (CV) prediction performance assessed by root mean squared error of prediction (RMSEP) and coefficient of determination R^2. One concrete option to determine the optimal number of PLS components is the randomization test approach (van der Voet, 1994). In brief, given a reference model chosen according to the absolute minimum in the CV curve, the procedure tests for the significance of increments in the squared prediction errors in models with fewer components. The selected model is the one with the smallest number of components that is not significantly worse than the reference model.

A number of procedures have been proposed to estimate the PLS model coefficients so that the covariance between the scores and the response is maximized. One of the most popular methods producing uncorrelated scores is the NIPALS algorithm (Varmuza and Filzmoser, 2009), that can be summarized in the following steps. For $q = 1, \ldots, Q$:

1. $\mathbf{o}_q^* = \mathbf{A}_q^\top \mathbf{y} / \left(\mathbf{y}^\top \mathbf{y}\right)$, with $\mathbf{A}_1 = \mathbf{A}$
2. $\mathbf{o}_q = \mathbf{o}_q^* / \|\mathbf{o}_q^*\|$, with $\|\cdot\|$ denoting the Euclidean norm
3. $\mathbf{f}_q = \mathbf{A}_q \mathbf{o}_q$
4. $\mathbf{g}_q = \mathbf{A}_q^\top \mathbf{f}_q / \left(\mathbf{f}_q^\top \mathbf{f}_q\right)$
5. $u_q = \mathbf{y}^\top \mathbf{f}_q / \left(\mathbf{f}_q^\top \mathbf{f}_q\right)$
6. $\mathbf{A}_{q+1} = \mathbf{A}_q - \mathbf{f}_q \mathbf{g}_q$

Then, the regression coefficients are estimated by

$$\hat{\boldsymbol{\beta}} = \mathbf{O}\left(\mathbf{F}^\top \mathbf{O}\right)\mathbf{u},$$

where the matrix $\mathbf{O}$ is formed by columns $\mathbf{o}_q$, matrix $\mathbf{F}$ by columns $\mathbf{f}_q$ and column vector $\mathbf{u}$ by elements $u_q,\ q = 1, \ldots, Q$.

To determine the individual statistical significance of the explanatory variables, bootstrap-based significance testing of the standardised PLS regression coefficients can be applied (Kalivodová et al., 2015). That is, denoting by μ_k and υ_k respectively the mean and the standard deviation of $\hat{\beta}_k,\ k = 1, \ldots, K$ over B bootstrap resamples, the estimated bootstrap standardised coefficients μ_k/υ_k are compared with the $(\alpha/2)$- and $(1-\alpha/2)$-quantile of a standard normal distribution. This means that with the usual choice $\alpha = 0.05$ as statistical significance level, the kth explanatory variable is considered significant in relation to the response variable if $|\mu_k/\upsilon_k| > 1.96,\ k = 1, \ldots, K$.

Using PLS regression allows to project the data onto a 2-dimensional PLS biplot corresponding to the first two PLS components (Oyedele and Gardner-Lubbe, 2015) . That is, the representation of the N observations (using points) is given by the rows of the matrix $\mathbf{F}_{(2)} = (\mathbf{f}_1, \mathbf{f}_2)$ and the representation of the K explanatory variables (using arrows from the origin) is given by the rows of the matrix $\mathbf{G}_{(2)} = (\mathbf{g}_1, \mathbf{g}_2)$. The scores represent the projection of the observations onto the space defined by the PLS components, while the loadings represent the effect of the explanatory variables on the directions of the projections. Therefore, a PLS biplot provides a single graphical representation of the observations alongside the explanatory variables which, unlike ordinary biplots based on PCA, accounts for the relationship with the response variable. The observations in the direction of an arrow are characterised by higher values on the corresponding explanatory variable (hence in case of a balance by dominance of the parts in the numerator over those in the denominator of the logratio). The sign of the relationship with the outcome variable determines the direction of the arrow.

Often, instead of one model (10), L models (11) are examined with the focus on the first coordinate and the respective estimated bootstrap standardised coefficient $\mu_1^{(l)}/\upsilon_1^{(l)}, l = 1, \ldots, L$ together with the estimates $\mu_D/\upsilon_D, \ldots, \mu_{D-1+P}/\upsilon_{D-1+P}$ corresponding to the P non-compositional variables (Kalivodová et al., 2015; Štefelová et al., 2021b,c). Note that these are invariant to the specific choice of balances due to the orthogonality of the coordinate re-

presentation and linearity of PLS regression (Helland, 2010). This property also leads to the fact that the decomposition of the matrix $\mathbf{A}^{(l)}$ yields the same score matrix $\mathbf{F}$ in each of the L models. Therefore $\mathbf{F}_{(2)}$ from any given model can be used for the visualization of the observations in PLS biplot. Of course, different matrix of loadings $\mathbf{G}^{(l)}$ is obtained each time. Then in a compositional PLS biplot, we display only loadings corresponding to the first coordinate from each system, together with the loadings associated with the non-compositional variables. That is, denoting by $\mathbf{G}^{(l)}_{(2)} = (\mathbf{g}^{(l)}_1, \mathbf{g}^{(l)}_2)$ the matrix of loadings corresponding to the first two PLS components in the lth coordinate system, the first row of $\mathbf{G}^{(l)}_{(2)}$ is used for the representation of the first coordinate (e.g. $b^{(l)}_1, z^{(l)}_1$or $w^{(l)}_1$, $l = 1, \ldots, L$) and the last P rows from any given $\mathbf{G}^{(l)}_{(2)}$ are used to visualize the non-compositional covariates. When interpreting the biplot, similarly to the case for PCA biplots (Kynčlová et al., 2016), we need to take into account that the loadings are generated from different PLS models. Thus, interpretation of relationships between arrows (loadings) corresponding to the different first ilr coordinates could lead to misleading conclusions.

Especially in case of high-dimensional data, it is convenient to automatize the choice of ilr coordinates by considering $L = D$ (weighted) pivot coordinate systems so that the relative importance of each of the D compositional parts is assessed (Kalivodová et al., 2015; Štefelová et al., 2021b). Further, with a large number of explanatory variables it is reasonable to include adjustment for multiple testing, e.g. using Bonferroni's adjustment (Kalivodová et al., 2015), the estimated bootstrap standardised coefficients are compared with the $(\alpha_{adj}/2)$- and $(1-\alpha_{adj}/2)$-quantile of a standard normal distribution, where $\alpha_{adj} = \alpha/(D+P)$.

2 Robust regression on compositional covariates including cellwise outliers

Traditionally, robust statistical methods have been designed to deal with *rowwise outliers*, i.e. entire observations being contaminated, assuming that there is a majority of non-contaminated observations in the dataset. This includes robust regression techniques, such as MM-estimation described in Section 1.2.2. However, atypical observations often exhibit outlying values only in a single variable or a small subset of variables (Rousseeuw and Van den Bossche, 2018). When contamination occurs at the cell level of a data matrix, it is actually possible that the majority of rows contain some outlying cells. Thus, treating entire observations as outliers might lead to an unacceptable loss of useful information (Alqallaf et al., 2009). Recent literature has focused on this latter type of outliers, referred to as *cellwise outliers*. Figure 1 illustrates the two types of outliers that can be found in a data matrix.

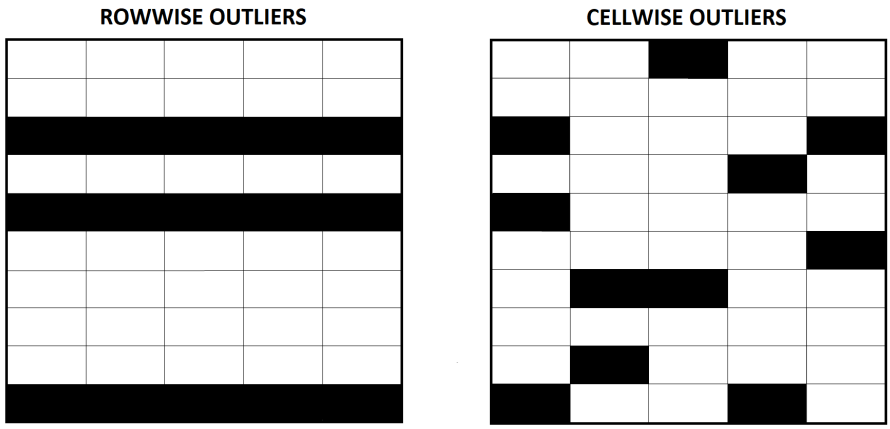


Figure 1: Illustration of rowwise outliers (left) and cellwise outliers (right).

In the context of linear regression, a few methods have been introduced recently that are robust against cellwise outliers such as shooting S-estimator (Öllerer et al., 2016) and 3-step regression estimator (Leung et al., 2016). The former, which combines a coordinate descent algorithm with simple robust regression, deals with deviating cells by weighing the components of an observation differently. The latter, robust also against rowwise outliers, starts by filtering outlying cells and then apply rowwise robust estimator for incomplete data. However, both

methods have some limitations when it comes to working with compositional data. Neither of them is suitable for regression with ilr coordinate representation of compositions as detailed in Section 1.2. The reason is that one outlying compositional part can affect several logratio coordinates so cellwise contamination easily propagates throughout.

In this chapter, we present a robust estimation procedure for a linear regression model with a real-valued response and compositional explanatory variables, possibly accompanied by additional real-valued covariates, that is designed to handle both cellwise and rowwise outliers (Štefelová et al., 2021a). The method is developed for the regular case with more observations than explanatory variables. It is similar in spirit to the 3-step regression estimator as it filters cellwise outliers and apply rowwise robust regression technique. But since a construction of an appropriate coordinate system for compositions is not feasible for incomplete data, our procedure makes use of an imputation step after the filtering. Imputation uncertainty is then reflected on regression coefficients estimates via multiple imputation scheme.

Section 2.1 gives a detailed description of the proposed algorithm, Section 2.2 illustrates its use in a bio-environmental science application and its relative performance in comparison to other regression methods is assessed by simulation in Section 2.3. The results indicate that our procedure, which maximizes the use of the information contained in the dataset, can cope with moderate levels of cellwise and rowwise contamination, and yields better or comparable estimates than its competitors: the aforementioned shooting S-estimator and 3-step regression estimator, as well as the rowwise robust MM-estimator and the OLS estimator. Moreover, our procedure allows to perform regression analysis in any ilr coordinate system that provides suitable interpretability of the results, whereas the predicted values do not depend on the particular coordinate representation.

2.1 Proposed algorithm

Here we address three challenges for regression analysis: (i) the inclusion of compositional explanatory variables, possibly complemented by real-valued explanatory variables; (ii) the presence of cellwise outliers; and (iii) the presence of rowwise outliers. Each one creates its own set of particular issues for statistical

modelling, and regardless of their occurrence in isolation or in combination, ignoring these issues can lead to unreliable and biased results. Therefore, the proposed method consists of three stages:

1. Detect outlying cells in the dataset (that are not part of entire outlying observations).
2. Replace them by sensible values via rowwise robust imputation.
3. Conduct rowwise robust compositional regression with multiple imputation estimates.

These stages are discussed in more detail in the following subsections.

2.1.1 Detection of cellwise outliers

The detection of deviating cells is based on the bivariate filter of Rousseeuw and Van den Bossche (2018). The foremost assumption of this method is that the data matrix is generated from a multivariate normal population, but some cell values are contaminated at random and become outliers. The procedure is briefly sketched in the following:

1. First, all variables (columns) are robustly standardized, e.g., by subtracting the median and dividing by the median absolute deviation (MAD).
2. Then deviating cells in single variables are marked, i.e., those containing absolute values higher than the cut-off value $\sqrt{\chi^2_{1,\tau}}$, where $\chi^2_{1,\tau}$ is the τ-quantile of the χ^2 distribution with one degree of freedom.
3. For each variable, the correlated variables are determined, i.e., those with absolute robust correlation higher than 0.5. Predictions for every cell are made based on each correlated variable that has a nonmarked cell in the same observation (row). If multiple nonmarked cells are available, the weighted mean of the corresponding predictions can be taken as the predicted value. A deshrinkage step is subsequently applied to obtain the final prediction. If all other cells of the row are marked as well, the prediction is set to 0 (which is the location estimate of the variable since all variables are

standardized). A cell for which the observed value differs too much from its prediction is marked.

4. The cells marked in step 2 or 3 are considered to be cellwise outliers.

5. Finally, rowwise outliers are identified. The n-th row of the data matrix is marked as an outlier if the absolute value of a robustly standardized statistic T_n exceeds the cut-off value $\sqrt{\chi^2_{1,\tau}}$. The statistic T_n is defined as the average (over m) of $F(\Delta^2_{nm})$, where F stands for the cumulative distribution function of the χ^2 distribution with one degree of freedom, and Δ_{nm} denotes the robustly standardized difference between the value in the cell with indices (n, m) and its prediction (from step 3).

Denote the dataset at hand as $(N, D + P + 1)$-matrix $\boldsymbol{\mathcal{X}} = (\mathbf{x}_1, \ldots, \mathbf{x}_D, \mathbf{c}_1, \ldots, \mathbf{c}_P, \mathbf{y}) = (\mathbf{x}_1, \ldots, \mathbf{x}_D, \mathbf{r}_1, \ldots, \mathbf{r}_{P+1})$ whose columns combine D-part compositional dataset $\mathbf{X} = (x_{nd})$ and values on $P + 1$ real-valued variables (P non-compositional covariates and 1 response variable) corresponding to N observations. For compositional data, we search for deviating cells through pairwise logratios where the elemental information is contained. Since the inverse logratios differs just on the sign, only $D(D-1)/2$ logratios have to be considered. Clearly, if a form of contamination generates an outlying value in a compositional part x_{nd}, this will affect all pairwise logratios where x_{nd} is contained. On the other hand, data contamination that generates aberrant pairwise logratio $\ln(x_{nc}/x_{nd})$ might have been originated from two outlying compositional parts, namely x_{nc} and x_{nd}. These considerations need to be taken into account when determining cellwise outliers in compositional dataset.

Accordingly, we apply the bivariate filter to the $(N, D(D-1)/2 + P + 1)$-matrix $\boldsymbol{\mathcal{L}}$, which contains the relevant pairwise logratios of the compositions along with potential real-valued covariates and the response variable, i.e., $\boldsymbol{\mathcal{L}} = (\ln(\mathbf{x}_1/\mathbf{x}_2), \ldots, \ln(\mathbf{x}_{D-1}/\mathbf{x}_D), \mathbf{r}_1, \ldots, \mathbf{r}_{P+1})$. The next task is to transfer the information about the cellwise outliers in $\boldsymbol{\mathcal{L}}$ to $\boldsymbol{\mathcal{X}}$. While this is identical for the real-valued variables, we propose to mark a compositional part x_{nd} in $\boldsymbol{\mathcal{X}}$ as a cellwise outlier (and subsequently set its value to missing to be imputed) if at least half of the logratios containing x_{nd} are identified as outliers by the bivariate filter. After extensive simulation experiments, we found this condition strict enough

to detect outlying compositional parts but not overly strict. As a matter of fact, many outlying cells would not be detected if we required that all logratios including a particular part had to be marked as outliers. We set $\tau = 0.99$ in the cutoff value $\sqrt{\chi^2_{1,\tau}}$ of the outlier filter as recommended in Rousseeuw and Van den Bossche (2018) since it gave favorable results in our simulations.

Note that the purpose of the initial filter is to avoid that the subsequent regression modelling is influenced by cellwise outliers. However, while cellwise outlier filters perform well in detecting individual outlying cells, they are not as effective in detecting rowwise outliers (Leung et al., 2016; Rousseeuw and Van den Bossche, 2018). Hence it is still crucial to protect against rowwise outliers in the subsequent stages of the procedure. Moreover, observations that have a large number of outlying cells are likely to be rowwise outliers. In our view, it is thus better not to impute those data cells and instead have the entire observation downweighted by a robust regression estimator in the following stages. Hence, at this point we treat an observation as a rowwise outlier if step 5 of the bivariate filter identifies the corresponding row in $\boldsymbol{\mathcal{L}}$ as a rowwise outlier, or if at least 75% cells of the corresponding row in $\boldsymbol{\mathcal{X}}$ are marked as cellwise outliers. The final index set $\mathcal{O}$ contains the indices (n, m) of all cellwise outliers that are not part of rowwise outliers. Cells of $\boldsymbol{\mathcal{X}}$ indicated by $\mathcal{O}$ are treated as missing values to be imputed in the next stage.

2.1.2 Imputation of cellwise outliers

Since compositional data are projected onto $\mathbb{R}^{D-1}$ through logratios involving several parts, missing parts as derived from the cellwise outlier filter can easily result in an unmanageable amount of missing logratios. We therefore impute the affected cells beforehand, so that subsequent compositional regression based on logratio coordinates can be conducted as usual on the imputed data matrix. For this purpose, we modify the iterative model-based imputation procedure of Hron et al. (2010) for compositional data to allow for a mixture of compositional and real-valued variables. This method uses a representation of the compositional data in pivot coordinates, and imputes the missing cells by estimates of expected values conditional on the observed part of the data. Such conditional expected values are modeled by linear regression models (with the assumption

that the error terms have expected value equal to zero), which are fitted using the rowwise robust MM-estimator (Yohai, 1987). As MM-regression allows to reduce the influence of rowwise outliers on the estimation of the imputation model, the imputed values should reflect the structure of the majority of the available data.

The imputation of outlying cells starts by separately sorting compositional parts and real-valued variables in decreasing order according to the amount of missing values. To simplify notation, we assume that this sorting does not change the original position of any compositional part or real-valued variable. Following Hron et al. (2010), the imputation algorithm is initialized with the simultaneous k-nearest-neighbor ($k$nn) method, which is based on the Aitchison distance between neighbors for the compositional parts and on the Euclidean distance between neighbors for the real-valued variables.

Each iteration of the imputation algorithm consists of at most $D + P + 1$ steps. The first steps involve the imputation of the compositional parts (up to D), whereas the remaining steps involve the imputation of the real-valued variables (up to $P + 1$). The procedure is summarized as follows:

1. For each compositional part $\mathbf{x}_l$ that contains outlying cells, $l = 1, \ldots, D$, pivot coordinates $\mathbf{Z}^{(l)} = \left(z_{nj}^{(l)}\right)$ are obtained (Section 1.1.2) to sequentially fit regression models of the first pivot coordinate on the remaining $D - 2$ coordinates plus the $P + 1$ non-compositional variables as covariates, while observations with no outlying cell in $\mathbf{x}_l$ are used for model fitting. The estimated regression coefficients are obtained using MM estimation such that they are robust against rowwise outliers. Furthermore, MM-regression also protects against poorly initialized missing value imputation (Hron et al., 2010). The coefficient estimates are then used to compute predicted values $\hat{z}_{n1}^{(l)},\ (n, l) \in \mathcal{O}$.

2. For $(n, l) \in \mathcal{O}$, imputed compositional parts $\hat{x}_{n1}, \ldots, \hat{x}_{nD}$ are obtained via the inverse mapping $\text{ilr}^{-1}\left(\hat{z}_{n1}^{(l)}, z_{i2}^{(l)}, \ldots, z_{n,D-1}^{(l)}\right)$. Note that the ratios between the non-outlying parts are not affected by this procedure.

3. Next, each real-valued variable that contains outlying cells is imputed

in an analogous way by sequentially serving as response in MM-regression on the remaining variables as predictors, including the compositional parts through pivot coordinates. Note that it does not matter which particular pivot coordinate system is used here. They all yield the same predictions due to the fact that they are orthogonal rotations of each other.

This is repeated iteratively until the sum of the squared relative changes in the imputed values are smaller than a threshold η. Following Hron et al. (2010), η is set at 0.5. Only a few iterations were typically needed to reach convergence in our simulations. The iterative procedure results in an imputed dataset $\tilde{\boldsymbol{\mathcal{X}}} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_D, \tilde{\mathbf{r}}_1, \ldots, \tilde{\mathbf{r}}_{P+1}) = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_D, \tilde{\mathbf{c}}_1, \ldots, \tilde{\mathbf{c}}_P, \tilde{\mathbf{y}})$, whose columns combine D-part compositional imputed dataset $\tilde{\mathbf{X}} = (\tilde{x}_{nd})$ and imputed values on P non-compositional covariates and the response variable corresponding to N observations. $\tilde{\boldsymbol{\mathcal{X}}}$ serves as input for the subsequent stage.

The performance of the imputations in the stages 1 and 3 above can often be improved by applying some form of variable selection to fit the corresponding regression models. To keep the computational burden low, we use a simple initial variable screening technique: before starting the iterative imputation procedure, we identify the most correlated variables for each variable to be imputed. We thereby compute robust correlations via bivariate winsorization (Khan et al., 2007) based on pairwise complete observations. However, initial simulations suggest that variable screening may not be necessary if the number of variables and the amount of filtered cells are both relatively small (e.g., $D + P + 1 \leq 10$ and less than 10% filtered cells). Moreover, when the number of variables is small, a smaller correlation threshold should be used to ensure that enough variables survive the screening process. Our procedure therefore implements the following default behavior as a compromise: if $D+P+1 \leq 10$, only variables with absolute correlations higher than 0.2 are used, otherwise the threshold is set to 0.5.

2.1.3 Robust compositional regression with multiple imputation estimates

After imputing cellwise outliers, and possibly other missing values in the dataset, the actual regression modelling is conducted. However, it is well-known that measures of variability like standard errors can be underestimated when

the usual formulas are applied to imputed data (Little and Rubin, 2002). Consequently, statistical significance tests in relation to the regression coefficients tend to be anticonservative. The reason is that the uncertainty derived from imputing the filtered cells is not taken into account. A well-established solution to this problem is using multiple imputation (MI) (Rubin and Schenker, 1986). The basic idea is that instead of a single imputed dataset, H different imputed datasets are actually analysed. It has been shown that by aggregating estimates from all these datasets, better estimates of the standard errors are obtained, as they reflect the additional uncertainty from the imputation process (Little and Rubin, 2002; Van Buuren, 2012; Cevallos Valdiviezo and Van Aelst, 2015). We adopt this approach and, following Bodner (2009) and White et al. (2011), we consider the number of imputed datasets H to be the rounded percentage of rows in the data matrix affected by cellwise outliers.

Each of the H datasets is obtained from $\tilde{\boldsymbol{\mathcal{X}}}$ by adding random noise to the estimated values resulting from the imputation procedure. That is, rather than imputing the filtered cells with the conditional expected value, we impute them by a random draw from the estimated conditional distribution. For compositional data, the noise is not added directly to the compositional part $\tilde{x}_{nl}, (n,l) \in \mathcal{O}$, as this would be incoherent with the geometry of the simplex, but to the first pivot coordinate $\tilde{z}_{n1}^{(l)}$ obtained from the composition $\left(\tilde{x}_{n1}^{(l)}, \ldots, \tilde{x}_{nD}^{(l)}\right)$. The corresponding values of the compositional parts are then obtained by the inverse mapping. More specifically, consider the m-th step of the last iteration of the imputation procedure (Section 2.1.2), with $m = 1, \ldots, D + P + 1$. Missing values in the m-th variable are imputed by robust regression using all the other variables as predictors. Following Templ et al. (2011), random noise is added to the imputed value by drawing H random values from $\mathcal{N}(0, \hat{\sigma}_m^2(1 + \iota_m/N))$, where $\hat{\sigma}_m$ is a robust residual scale estimate from the corresponding regression fit and ι_m denotes the number of values to be imputed in the m-th variable.

Afterwards, regression analysis is performed for each of the H imputed datasets with compositions expressed in proper ilr coordinates (Section 1.2). Since we still need to protect against rowwise outliers after dealing with cellwise outliers, we apply the robust and highly efficient MM-estimator (Section 1.2.2). Note that this estimator is designed to handle rowwise outliers only, and it could easily fail

if applied directly to data containing cellwise outliers by skipping the previous cellwise outlier detection and imputation stages. We denote the k-th regression coefficient estimate, $k = 0, 1, \ldots, D-1+P$, from the h-th imputed dataset, $h = 1, \ldots, H$, as $\hat{\beta}_k^{\{h\}}$ and the corresponding estimated variance as $\hat{\phi}_k^{\{h\}}$. Following Rubin (1987) and Barnard and Rubin (1999), a final point estimate and variance for each regression coefficient is then obtained as

$$\hat{\beta}_k = \frac{1}{H}\sum_{h=1}^{H}\hat{\beta}_k^{\{h\}} \qquad \text{and} \qquad \hat{\phi}_k = \hat{\zeta}_k + \frac{H+1}{H}\hat{\xi}_k,$$

respectively, where $\hat{\zeta}_k = \frac{1}{H}\sum_{h=1}^{H}\hat{\phi}_k^{\{h\}}$ is the average within-imputation variance and $\hat{\xi}_k = \frac{1}{H-1}\sum_{h=1}^{H}\left(\hat{\beta}_k^{\{h\}} - \hat{\beta}_k\right)^2$ is the between-imputation variance.

The entire procedure is summarized in the following pseudocode.

Algorithm 1 Detection of cellwise outliers

Input: Data matrix $\boldsymbol{\mathcal{X}} = (\mathbf{x}_1, \ldots, \mathbf{x}_D, \mathbf{r}_1, \ldots, \mathbf{r}_{P+1})$ of compositional parts and real-valued variables

Output: Index set $\mathcal{O}$ of outlying cells and index set $\mathcal{R}$ of outlying rows

▷ Cellwise outlier detection on pairwise logratios and real-valued variables

$\boldsymbol{\mathcal{L}} \leftarrow (\ln(\mathbf{x}_1/\mathbf{x}_2), \ldots, \ln(\mathbf{x}_{D-1}/\mathbf{x}_D), \mathbf{r}_1, \ldots, \mathbf{r}_{P+1})$

Apply bivariate filter of Rousseeuw and Van den Bossche (2018) to $\boldsymbol{\mathcal{L}}$

Store index set $\mathcal{O}_{\boldsymbol{\mathcal{L}}} \leftarrow \{(n, j) :$ cell in row n and column j of $\boldsymbol{\mathcal{L}}$ is marked as cellwise outlier$\}$

Store index set $\mathcal{R}_{\boldsymbol{\mathcal{L}}} \leftarrow \{n :$ row n of $\boldsymbol{\mathcal{L}}$ is marked as rowwise outlier$\}$

▷ Mark outlying cells in compositional parts

Initialize empty set $\mathcal{O}$ ▷ set of indices (n, m) of cells in $\boldsymbol{\mathcal{X}}$ to be marked as cellwise outliers

Initialize empty set $\mathcal{R}$ ▷ set of indices n of rows in $\boldsymbol{\mathcal{X}}$ to be marked as rowwise outliers

for $d \in \{1, \ldots, D\}$ **do**

    Obtain index set $J_d \leftarrow \{j :$ column j of $\boldsymbol{\mathcal{L}}$ contains a logratio involving $x_d\}$

    for $n \in \{1, \ldots, N\}$ **do**

        if $\frac{1}{(D-1)}\sum_{j \in J_d} I_{\mathcal{O}_{\boldsymbol{\mathcal{L}}}}((n, j)) \geq 0.5$ **then**

$\mathcal{O} \leftarrow \mathcal{O} \cup \{(n, d)\}$
end if
end for
end for
▷ Adopt outlying cells in real-valued variables from bivariate filter
for $p \in \{1, \ldots, P+1\}$ **do**
for $n \in \{1, \ldots, N\}$ **do**
if $(n, D(D-1)/2 + p) \in \mathcal{O}_{\boldsymbol{\mathcal{L}}}$ **then**
$\mathcal{O} \leftarrow \mathcal{O} \cup \{(n, D+p)\}$
end if
end for
end for
▷ Mark outlying rows and only mark outlying cells that are not part of outlying rows
for $n \in \{1, \ldots, N\}$ **do**
if $n \in \mathcal{R}_{\boldsymbol{\mathcal{L}}}$ or $\frac{1}{D+P+1} \sum_{m=1}^{D+P+1} I_{\mathcal{O}}((n, m)) >= 0.75$ **then**
▷ Marked as rowwise outlier in $\boldsymbol{\mathcal{L}}$ or at least 75% of cells marked as cellwise outliers in $\boldsymbol{\mathcal{X}}$
$\mathcal{R} \leftarrow \mathcal{R} \cup \{n\}$
$\mathcal{O} \leftarrow \mathcal{O} \setminus \{(n, m) : m = 1, \ldots, D+P+1\}$
end if
end for
return Index sets $\mathcal{O}$ and $\mathcal{R}$

Algorithm 2 Initial $k$nn imputation for compositional data and real-valued variables

Input: Data matrix $\boldsymbol{\mathcal{X}} = (\mathbf{x}_1, \ldots, \mathbf{x}_D, \mathbf{r}_1, \ldots, \mathbf{r}_{P+1})$ of compositional parts and real-valued variables with missing values (outlying cells)

Output: Imputed data matrix $\tilde{\boldsymbol{\mathcal{X}}}$

Apply simultaneous $k$nn imputation with Aitchison distance to $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_D)$
Store imputed data matrix as $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_D)$
Compute pivot coordinates $\tilde{\mathbf{z}}_1^{(1)}, \ldots, \tilde{\mathbf{z}}_{D-1}^{(1)}$ from $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_D$
Apply simultaneous $k$nn imputation with Euclidean distance to $\left(\mathbf{r}_1, \ldots, \mathbf{r}_{P+1}, \tilde{\mathbf{z}}_1^{(1)}, \ldots, \tilde{\mathbf{z}}_{D-1}^{(1)}\right)$

5: Store imputed real-valued variables as $\tilde{\mathbf{R}} = (\tilde{\mathbf{r}}_1, \ldots, \tilde{\mathbf{r}}_{P+1})$

6: **return** Imputed data matrix $\tilde{\boldsymbol{\mathcal{X}}} = (\tilde{\mathbf{X}}, \tilde{\mathbf{R}})$

Algorithm 3 Model-based imputation for compositional data and real-valued variables

---

Input: Data matrix $\boldsymbol{\mathcal{X}} = (\mathbf{x}_1, \ldots, \mathbf{x}_D, \mathbf{r}_1, \ldots, \mathbf{r}_{P+1})$ of compositional parts and real-valued variables with missing values (outlying cells)

Output: Imputed data matrix $\tilde{\boldsymbol{\mathcal{X}}}$, residual scale estimates $\hat{\sigma}_1, \ldots, \hat{\sigma}_{D+P+1}$ from imputation models

1: ▷ Initializations

2: Rearrange first D columns of $\boldsymbol{\mathcal{X}}$ by sorting compositional parts by decreasing amount of missing values

3: Rearrange last $P+1$ columns of $\boldsymbol{\mathcal{X}}$ by sorting real-valued variables by decreasing amount of missing values

4: Obtain index sets $\kappa_m \leftarrow \{n : \text{cell in row } n \text{ and column } m \text{ of } \boldsymbol{\mathcal{X}} \text{ is missing}\}$, $m = 1, \ldots, D+P+1$

5: Obtain index sets $\tau_m \leftarrow \{n : \text{cell in row } n \text{ and column } m \text{ of } \boldsymbol{\mathcal{X}} \text{ is observed}\}$, $m = 1, \ldots, D+P+1$

6: Initialize counter $t \leftarrow 0$ and convergence criterion $\eta \leftarrow \infty$

7: Initialize $\boldsymbol{\mathcal{X}}^{[0]} = \left(\mathbf{x}_1^{[0]}, \ldots, \mathbf{x}_D^{[0]}, \mathbf{r}_1^{[0]}, \ldots, \mathbf{r}_{P+1}^{[0]}\right)$ by applying $k$nn imputation from Algorithm 2 to $\boldsymbol{\mathcal{X}}$

8: ▷ Iterative model-based imputations

9: **while** $\eta \geq 0.5$ **do**

10: $t \leftarrow t + 1$

11: $\boldsymbol{\mathcal{X}}^{[t]} = \left(\mathbf{x}_1^{[t]}, \ldots, \mathbf{x}_D^{[t]}, \mathbf{r}_1^{[t]}, \ldots, \mathbf{r}_{P+1}^{[t]}\right) \leftarrow \boldsymbol{\mathcal{X}}^{[t-1]} = \left(\mathbf{x}_1^{[t-1]}, \ldots, \mathbf{x}_D^{[t-1]}, \mathbf{r}_1^{[t-1]}, \ldots, \mathbf{r}_{P+1}^{[t-1]}\right)$

12: ▷ Imputations in compositional data

13: **for** $d \in \{1, \ldots, D\}$ **do**

14: Compute pivot coordinates $z_{n1}^{(d)}, \ldots, z_{n,D-1}^{(d)}$ from $x_{n1}^{[t]}, \ldots, x_{nD}^{[t]}$, $n = 1, \ldots, N$

15: Perform MM-regression of $z_{n1}^{(d)}$ on $z_{n2}^{(d)}, \ldots, z_{n,D-1}^{(d)}, r_{n1}^{[t]}, \ldots, r_{n,P+1}^{[t]}$, $n \in \tau_d$

16: Compute prediction $\hat{z}_{n1}^{(d)}$ from $z_{n2}^{(d)}, \ldots, z_{n,D-1}^{(d)}, r_{n1}^{[t]}, \ldots, r_{n,P+1}^{[t]}$, $n \in \kappa_d$

17: Replace $x_{n1}^{[t]}, \ldots, x_{nD}^{[t]}$ with the inverse mapping of $\hat{z}_{n1}^{(d)}, z_{n2}^{(d)}, \ldots, z_{n,D-1}^{(d)}$,

$n \in \kappa_d$

Compute robust residual scale estimate $\hat{\sigma}_d$ from MM-regression fit

end for

▷ Imputations in real-valued variables

Compute pivot coordinates $z_{n1}^{(1)}, \ldots, z_{n,D-1}^{(1)}$ from $x_{n1}^{[t]}, \ldots, x_{nD}^{[t]}$, $n = 1, \ldots, N$

for $p \in \{1, \ldots, P+1\}$ **do**

Perform MM-regression of $r_{np}^{[t]}$ on $z_{n1}^{(1)}, \ldots, z_{n,D-1}^{(1)}, r_{n1}^{[t]}, \ldots, r_{n,p-1}^{[t]}, r_{n,p+1}^{[t]}, r_{n,P+1}^{[t]}$, $n \in \tau_p$

Replace $r_{np}^{[t]}$ with prediction $\hat{r}_{np}^{[t]}$ from $z_{n1}^{(1)}, \ldots, z_{n,D-1}^{(1)}, r_{n1}^{[t]}, \ldots, r_{n,p-1}^{[t]}, r_{n,p+1}^{[t]}, r_{n,P+1}^{[t]}$, $n \in \kappa_p$

Compute robust residual scale estimate $\hat{\sigma}_{D+p}$ from MM-regression fit

end for

▷ Update convergence criterion

$\eta \leftarrow \sum_{n=1}^{N} \left[\sum_{d=1}^{D} \left(\frac{x_{nd}^{[t-1]} - x_{nd}^{[t]}}{x_{nd}^{[t]}} \right)^2 + \sum_{p=1}^{P+1} \left(\frac{r_{np}^{[t-1]} - r_{np}^{[t]}}{r_{np}^{[t]}} \right)^2 \right]$

end while

Obtain $\tilde{\boldsymbol{\mathcal{X}}}$ by rearranging columns of $\boldsymbol{\mathcal{X}}^{[t]}$ from last iteration according to original order of columns in $\boldsymbol{\mathcal{X}}$

Rearrange residual scale estimates $\hat{\sigma}_1, \ldots, \hat{\sigma}_{D+P+1}$ accordingly

return Imputed data matrix $\tilde{\boldsymbol{\mathcal{X}}}$ and residual scale estimates $\hat{\sigma}_1, \ldots, \hat{\sigma}_{D+P+1}$

Algorithm 4 Cellwise and rowwise robust compositional regression with bivariate filter and multiple imputation

---

Input: Compositional data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_D)$, real-valued covariates $\mathbf{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_P)$, real-valued response $\mathbf{y}$

Output: Regression coefficient estimates and corresponding variance estimates

▷ Detect cellwise outliers

Obtain index set $\mathcal{O}$ of cellwise outliers by applying Algorithm 1 to $\boldsymbol{\mathcal{X}} = (\mathbf{x}_1, \ldots, \mathbf{x}_D, \mathbf{c}_1, \ldots, \mathbf{c}_P, \mathbf{y})$

▷ Special case of no cellwise outliers

if $\mathcal{O} = \emptyset$ **then**

Compute ilr coordinates $\mathbf{i}_1, \ldots, \mathbf{i}_{D-1}$ from $\mathbf{x}_1, \ldots, \mathbf{x}_D$

Perform MM-regression of $\mathbf{y}$ on $\mathbf{i}_1, \ldots, \mathbf{i}_{D-1}, \mathbf{c}_1, \ldots, \mathbf{c}_P$

return Coefficient estimates and corresponding variance estimates

end if

▷ Filter and impute cellwise outliers

Replace cells of $\boldsymbol{\mathcal{X}}$ with indices in $\mathcal{O}$ by missing values

Apply model-based imputation with Algorithm 3 to $\boldsymbol{\mathcal{X}} = (\mathbf{x}_1, \ldots, \mathbf{x}_D, \mathbf{c}_1, \ldots, \mathbf{c}_P, \mathbf{y})$

Store imputed data matrix as $\tilde{\boldsymbol{\mathcal{X}}} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_D, \tilde{\mathbf{c}}_1, \ldots, \tilde{\mathbf{c}}_P, \tilde{\mathbf{y}})$

Store residual scale estimates from imputation models as $\hat{\sigma}_1, \ldots, \hat{\sigma}_{D+P+1}$, respectively

▷ Rowwise robust compositional regression with multiple imputation

$N_{\text{out}} \leftarrow N - \sum_{n=1}^{N} \prod_{m=1}^{D+P+1} (1 - I_{\mathcal{O}}((n,m)))$ ▷ Number of observations with outlying cells

$H \leftarrow \max(2, \text{round}(100 \cdot N_{\text{out}}/N))$ ▷ Number of imputations

Obtain $\iota_m \leftarrow \sum_{n=1}^{N} I_{\mathcal{O}}((n,m))$, $m = 1, \ldots, D+P+1$ ▷ Number of outlying cells per variable

for $h \in \{1, \ldots, H\}$ **do**

▷ Add random noise to imputations

Initialize $\tilde{\boldsymbol{\mathcal{X}}}^{\{h\}} = \left(\tilde{\mathbf{x}}_1^{\{h\}}, \ldots, \tilde{\mathbf{x}}_D^{\{h\}}, \tilde{\mathbf{c}}_1^{\{h\}}, \ldots, \tilde{\mathbf{c}}_P^{\{h\}}, \tilde{\mathbf{y}}^{\{h\}}\right)$ by $\tilde{\mathcal{X}} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_D, \tilde{\mathbf{c}}_1, \ldots, \tilde{\mathbf{c}}_P, \tilde{\mathbf{y}})$

for $(n,m) \in \mathcal{O}$ **do**

Draw random noise term $e \sim N(0, \hat{\sigma}_m^2(1 + \iota_m/N))$

if $m \in \{1, \ldots, D\}$ **then** ▷ Compositional parts

Compute pivot coordinates $\tilde{z}_{n1}^{(m)}, \ldots, \tilde{z}_{n,D-1}^{(m)}$ from $\tilde{x}_{n1}, \ldots, \tilde{x}_{nD}$

$\tilde{z}_{n1}^{(m)} \leftarrow \tilde{z}_{n1}^{(m)} + e$

Replace $\tilde{x}_{n1}^{\{h\}}, \ldots, \tilde{x}_{nD}^{\{h\}}$ with the inverse mapping of $\tilde{z}_{n1}^{(m)}, \ldots, \tilde{z}_{n,D-1}^{(m)}$

else if $m \in \{D+1, \ldots, D+P\}$ **then** ▷ Real-valued variables

$\tilde{c}_{n,m-D}^{\{h\}} \leftarrow \tilde{c}_{n,m-D} + e$

else ▷ Response variable

$\tilde{y}_n^{\{h\}} \leftarrow \tilde{y}_n + e$

end if

end for

▷ Rowwise robust compositional regression

Compute ilr coordinates $\mathbf{i}_1, \ldots, \mathbf{i}_{D-1}$ from $\tilde{\mathbf{x}}_1^{\{h\}}, \ldots, \tilde{\mathbf{x}}_D^{\{h\}}$

Perform MM-regression of $\tilde{\mathbf{y}}^{\{h\}}$ on $\mathbf{i}_1, \ldots, \mathbf{i}_{D-1}, \tilde{\mathbf{c}}_1^{\{h\}}, \ldots, \tilde{\mathbf{c}}_P^{\{h\}}$

Store coefficient estimates as $\left(\hat{\beta}_0^{\{h\}}, \ldots, \hat{\beta}_{D-1+P}^{\{h\}}\right)^\top$

Store variance estimates as $\left(\hat{\phi}_0^{\{h\}}, \ldots, \hat{\phi}_{D-1+P}^{\{h\}}\right)^\top$

end for

▷ Aggregate results from multiple imputation

Compute final coefficient estimates $\hat{\beta}_k \leftarrow \frac{1}{H}\sum_{h=1}^{H} \hat{\beta}_k^{\{h\}}$, $k = 0, \ldots, D-1+P$

Compute average within-imputation variances $\hat{\zeta}_k \leftarrow \frac{1}{H}\sum_{h=1}^{H} \hat{\phi}_k^{\{h\}}$, $k = 0, \ldots, D-1+P$

Compute between-imputation variances $\hat{\xi}_k \leftarrow \frac{1}{H-1}\sum_{h=1}^{H} \left(\hat{\beta}_k^{\{h\}} - \hat{\beta}_k\right)^2$, $k = 0, \ldots, D-1+P$

Compute variance estimates $\hat{\phi}_k \leftarrow \hat{\zeta}_k + \frac{H+1}{H}\hat{\xi}_k$, $k = 0, \ldots, D-1+P$

return Coefficient estimates $(\hat{\beta}_0, \ldots, \hat{\beta}_{D-1+P})^\top$ and corresponding variance estimates $(\hat{\phi}_0, \ldots, \hat{\phi}_{D-1+P})^\top$

2.2 Application to low-dimensional metabolomic data

We apply the proposed compositional MM-regression with a bivariate cellwise outlier filter and multiple imputation (BF-MI algorithm) to investigate the association between livestock methane emissions from individual animals and their ruminal volatile fatty acid (VFA) composition, while accounting for the potential effects of other animal and diet-related covariates. The dataset contains $N = 239$ observations originating from the study carried out in Palarea-Albaladejo et al. (2017). The concentrations of 6-part VFA composition consisting of acetate, propionate, butyrate, isobutyrate, isovalerate and valerate were determined by high-performance liquid chromatography from rumen fluid samples taken using a stomach tube. The quality of the chromatography determines the precision of the measurements, and outlying measurements may be related to unstable baselines, noisy detectors, poor resolution of the components, or errors

on the part of the operator in preparing the solution or performing the measurement. Animal methane yield (CH_4 in grams per kilogram of dry matter intake) was measured using indirect respiration chambers. Further, animal diet metabolizable energy (ME), dry matter intake (DMI), weight and type of diet (either concentrate or mixed) were recorded .

All four positive-valued variables in the dataset (CH_4, ME, DMI and weight) are log-transformed and thus mapped into real space to better accommodate model assumptions. Moreover, the dataset is split by diet type before the bivariate outlier filter (Section 2.1.1) is applied separately to each resulting subset of data. Overall, 1.26% of rows are marked as rowwise outliers, while 1.96% of cells in the remaining observations are marked as cellwise outliers. Figure 2 highlights these in each numerical variable, as well as the marked rows, in red color. Note that both the imputation step (Section 2.1.2) and the regression step

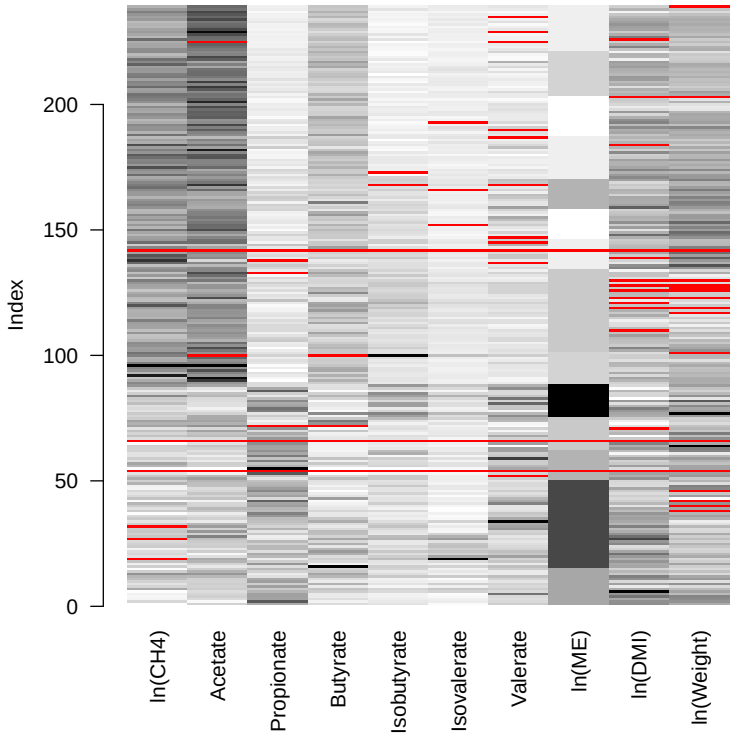


Figure 2: Cellwise and rowwise outliers detected by the bivariate filter in the VFA dataset. Outlying cells/rows are colored in red. The grey color scheme reflects the values of compositional parts and real-valued variables (the higher the value, the darker the color).

(Section 2.1.3) of our procedure work with categorical variables in the usual way

by including dummy variables. Here we consider dummy variable MD, which takes the value 1 for mixed diet and 0 for a concentrate diet. We skip the variable screening in the imputation step, as the number of variables is rather small and fewer than 2% of cells are filtered. The final estimates are obtained for $L = D = 6$ models (11) with CH_4 (in log-scale) set as a response and VFA composition expressed in pivot coordinates $z_1^{(l)}, \ldots, z_5^{(l)}$, $l = 1, ..., 6$, as explanatory variables so we can examine the relative role (dominance) of each of the six parts through the regression coefficient at the first pivot coordinate in each system. ME, DMI, weight (all three in log-scale) and MD are put as additional covariates. For comparison, we also fit the regression model using OLS estimation (Section 1.2.1) and MM estimation (Section 1.2.2). Note that in this application we are interested in an interpretation of the results in terms of pivot coordinates, therefore it is not meaningful to apply other methods such as the shooting S-estimator (Öllerer et al., 2016) or 3-step regression (Leung et al., 2016).

Table 2 displays the relevant results extracted from the six models using the three estimation procedures considered. Focusing on the VFA composition, OLS estimation does not result in a statistically significant association between the dominance of ruminal acetate and methane yield (p-value $= 0.127$). The MM-estimator (without the cellwise outlier filter) provides only a weakly significant positive association between animal methane emission and the relative production of ruminal acetate (p-value $= 0.053$). Moreover, a statistically significant negative association is concluded in both cases between methane yield and the dominance of propionate (p-value < 0.001). The results from using our proposed BF-MI method are comparable in terms of overall directions of the associations, but the statistical significance of the acetate related term is notably higher (p-value < 0.001), which further stresses the role of the contrast between acetate and propionate as a driver of the association between the ruminal VFA composition and methane emission, which is in agreement with biological knowledge (Wolin, 1960; Palarea-Albaladejo et al., 2017).

Table 2: Regression coefficient estimates, standard errors and p-values for the VFA dataset: compositional OLS estimation, compositional MM estimation without a cellwise outlier filter, and proposed compositional MM estimation with a bivariate cellwise outlier filter and multiple imputation (BF-MI).

	OLS			MM			BF-MI		
Covariate	Coeff.	Std. error	p-value	Coeff.	Std. error	p-value	Coeff.	Std. error	p-value
$z_1^{(\text{Acetate})}$	0.125	0.082	0.127	0.203	0.104	0.053	0.301	0.084	< 0.001
$z_1^{(\text{Propionate})}$	-0.247	0.048	< 0.001	-0.304	0.067	< 0.001	-0.385	0.054	< 0.001
$z_1^{(\text{Butyrate})}$	0.093	0.051	0.072	0.070	0.054	0.193	0.025	0.050	0.617
$z_1^{(\text{Isobutyrate})}$	-0.015	0.047	0.744	-0.023	0.052	0.664	-0.014	0.055	0.794
$z_1^{(\text{Isovalerate})}$	0.006	0.032	0.848	0.005	0.034	0.890	0.015	0.034	0.662
$z_1^{(\text{Valerate})}$	0.038	0.039	0.322	0.049	0.037	0.195	0.059	0.064	0.350
ln(ME)	0.725	0.484	0.136	0.999	0.512	0.052	0.755	0.481	0.118
ln(DMI)	-0.413	0.064	< 0.001	-0.408	0.064	< 0.001	-0.397	0.072	< 0.001
ln(Weight)	0.627	0.147	< 0.001	0.651	0.165	< 0.001	0.689	0.186	< 0.001
DM	0.328	0.040	< 0.001	0.308	0.048	< 0.001	0.245	0.048	< 0.001

2.3 Simulation study

In order to assess the performance of our procedure in comparison to other (robust) methods for compositional regression, we perform a simulation study. The parameters for the simulation design are partly inspired by the VFA dataset from Section 2.2. As the main novelty of our procedure is the inclusion of compositional covariates in the context of robust regression with cellwise and rowwise outliers, we assume for simplicity that there are only compositional covariates involved. We set $N \in \{50, 100, 200\}$ as the number of observations and $D \in \{5, 10, 20\}$ as the number of compositional parts. The simulated compositions are generated through pivot coordinates. In order to obtain a realistic covariance structure in the pivot coordinate system, we chose an initial covariance matrix $\boldsymbol{\Sigma}_0 = \left(0.5^{|i-j|}/10\right)_{1 \leq i,j \leq D-1}$, with entries being similar in magnitude to the ones observed in the VFA case study. To investigate the effects of adding more variability to the data matrix, we consider the covariance matrix in pivot coordinates $\boldsymbol{\Sigma}$ as a multiple of the initial covariance matrix, i.e., $\boldsymbol{\Sigma} = c\boldsymbol{\Sigma}_0$ with $c \in \{1, 2, 3\}$.

We examine a scenario with both rowwise and cellwise outliers. Specifically, we consider the case where outlying rows (entire observations) and outlying cells (in the compositional parts and the response variable) both occur with probability $\theta \in \{0, 0.02, 0.05, 0.1, 0.2\}$. We first generate entire outlying observations (rows) and, subsequently, outlying cells only in non-outlying rows. We perform 1000 simulation runs for each configuration. In each simulation run, the data are generated as follows:

1. Pivot coordinates are sampled as $\boldsymbol{z}_n = (z_{n1}, \ldots, z_{n,D-1}) \sim \mathcal{N}_{D-1}(\mathbf{0}, \boldsymbol{\Sigma})$, $n = 1, \ldots, N$.

2. The values of the response variable are obtained in the pivot coordinate system as

   $$y_n = \beta_0 + \beta_1 z_{n1} + \ldots + \beta_{D-1} z_{n,D-1} + \varepsilon_n, \quad \varepsilon_n \sim \mathcal{N}(0, 0.25^2), \quad n = 1, \ldots, N,$$

   with regression parameters $\beta_0 = 0$ and $(\beta_1, \ldots, \beta_{D-1}) = (1, 0, 1, 0, \ldots)$. The variance of the error terms ε_n is chosen to roughly mimic the signal-to-noise ratio observed in the VFA data.

3. The pivot coordinates $\boldsymbol{z}_n$ are transformed according to (1.1) to obtain the corresponding compositions $\boldsymbol{x}_n = (x_{n1}, \ldots, x_{nD}) = \text{ilr}^{-1}(\boldsymbol{z}_n)$, $n = 1, \ldots, N$.

4. Observations are randomly selected with probability θ to be turned into rowwise outliers. We first generate outliers in the pivot coordinates along the smallest principal component. Let $\mathcal{R} \subseteq \{1, \ldots, N\}$ denote the set of indices of the rowwise outliers, and let $\boldsymbol{a}_n = (a_{n1}, \ldots, a_{n,D-1})$ denote the principal component scores corresponding to $\boldsymbol{z}_n$. For $n \in \mathcal{R}$, we change the value of the last component $a^*_{n,D-1} = a_{n,D-1} + 5\sqrt{c}$. Note that the factor $\sqrt{c}$ ensures that the outlier shift is of the same magnitude for the different scalings of the covariance matrix $\boldsymbol{\Sigma} = c\boldsymbol{\Sigma}_0$. After transforming the scores $\boldsymbol{a}^*_n = (a_{n1}, \ldots, a_{n,D-2}, a^*_{n,D-1})$ back to pivot coordinates to obtain outlying $\boldsymbol{z}^*_n = (z^*_{n1}, \ldots, z^*_{n,D-1})$, we change the respective values of the response variable to

$$y^*_n = \beta^*_0 + \beta^*_1 z^*_{n1} + \ldots + \beta^*_{D-1} z^*_{n,D-1} + \varepsilon_n, \quad n \in \mathcal{R},$$

with regression parameters $\beta^*_0 = 0$ and $\beta^*_k = -1$, $k = 1, \ldots, D-1$. Using regression coefficients that are very different to those from clean observations ensures that the rowwise outliers are bad leverage points. Finally, the outlying pivot coordinates $\boldsymbol{z}^*_n = (z^*_{n1}, \ldots, z^*_{n,D-1})$ are transformed according to (1.1) to obtain the corresponding outlying compositions $\boldsymbol{x}^*_n = (x^*_{n1}, \ldots, x^*_{nD}) = \text{ilr}^{-1}(\boldsymbol{z}^*_n)$, $n \in \mathcal{R}$.

5. Cells corresponding to non-outlying observations $(x_{n1}, \ldots, x_{nD}, y_n)$, $n \notin \mathcal{R}$, are randomly selected with probability θ to be turned into cellwise outliers. Let $\mathcal{O}$ denote the set of indices (n, m) of the outlying cells. For any pair $(n, m) \in \mathcal{O}$, we change the cell value to $x^{**}_{nm} = 10 \cdot x_{nm}$ if $m \in \{1, \ldots, D\}$ or to $y^{**}_n = 10 \cdot y_n$ if $m = D + 1$. The multiplicative factor was chosen to minimize the chance that outlying cells overlap with noise that occurs naturally in the composition or the real-valued response.

The resulting observations with rowwise and cellwise outliers are denoted by $\boldsymbol{x}^\star_n =$

$(x^\star_{n1}, \ldots, x^\star_{nD})'$ and $y^\star_n$, where

$$x^\star_{nd} = \begin{cases} x^*_{nd}, & \text{if } n \in \mathcal{R}, \\ x^{**}_{nd}, & \text{if } (n,d) \in \mathcal{O}, \\ x_{nd}, & \text{otherwise,} \end{cases} \qquad n = 1, \ldots, N, \;\; d = 1, \ldots, D,$$

and

$$y^\star_n = \begin{cases} y^*_n, & \text{if } n \in \mathcal{R}, \\ y^{**}_n, & \text{if } (n, D+1) \in \mathcal{O}, \\ y_n, & \text{otherwise,} \end{cases} \qquad n = 1, \ldots, N.$$

Below we give a brief description of the methods that participate in the evaluation, together with the abbreviations we use to refer to them:

OLS: ordinary compositional least squares regression described in Section 1.2.1 (with no treatment for outliers).

MM: robust compositional MM-regression described in Section 1.2.2 (with no treatment for cellwise outliers).

ShS: shooting S-estimator (Öllerer et al., 2016) obtained from the $D(D-1)/2$ unique pairwise logratios. The shooting S-estimator is designed to cope with cellwise contamination by weighing the components of an observation differently. Note that the results can only be compared in terms of prediction and not in terms of parameter estimation. We used both Tukey's biweight loss function and the skipped Huber loss function: the former yields continuous weights in $[0, 1]$ while the latter leads to binary weights in $\{0, 1\}$. We only report the results for Tukey's biweight loss function, as it generally gave better and more stable results than the skipped Huber loss function.

3S: 3-step regression (Leung et al., 2016) fitted to alr coordinates defined in (1), while in each simulation run, the reference part is selected randomly. Note that the use of $D(D-1)/2$ pairwise logratios as covariates is not possible here since the algorithm requires a full-rank data matrix. 3-step regression first uses a consistent univariate filter to eliminate outlying cells; second, it applies a robust estimator of multivariate location and scatter to the filtered data to downplay outlying rows; and third, it computes robust regression coefficients from the previous step. As with the shooting S-estimator, the results are compared only in terms of prediction.

It is important to note that the predicted values depend on the choice of the reference part. For example, an outlying value in a cell x_{n1} results in a rowwise outlier in the observation $(\ln(x_{n2}/x_{n1}), \ldots, \ln(x_{nD}/x_{n1}))$, but only in a cellwise outlier in $(\ln(x_{n1}/x_{nD}), \ldots, \ln(x_{n,D-1}/x_{nD}))$. These cases will be handled differently by 3-step regression, yielding different predictions of the response variable. Although this leads to somewhat limited practical applicability, it is still informative to include this approach here in order to compare its general performance.

BF-MI: this is our proposed method which applies the bivariate filter (BF) followed by multiple imputation (MI). In the imputations, we use the default behavior for variable screening (see Section 2.1.2).

IF-MI: this represents a hypothetical situation where an ideal filter (IF) is able to perfectly identify all outlying cells (and only those). The remaining steps of our method are afterwards applied using multiple imputation (MI). We use the same settings for variable screening as used for BF-MI. This case is included for benchmarking purposes only, as it is generally unattainable in practice.

Note that all methods except the shooting S-estimator and 3-step regression consider pivot coordinates to represent the compositional covariates. By construction, the shooting S-estimator and the 3-step regression method require the use of pairwise logratios and alr coordinates, respectively.

The performance of the methods is assessed in terms of the mean squared error (MSE) of the coefficient estimates, computed as

$$MSE = \frac{1}{D}\sum_{k=0}^{D-1}(\hat{\beta}_k - \beta_k)^2.$$

Further evaluation is made in terms of prediction error. For this purpose, N additional clean test observations $\boldsymbol{x}_n^{test}$ and y_n^{test}, $n = 1, \ldots, N$, are generated in each simulation run according to steps 1–3 of our data generating process. On the test data, the mean squared error of prediction (MSEP) is calculated as

$$MSEP = \frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n^{test} - y_n^{test})^2,$$

where $\hat{y}_n^{test}$ denote the predicted values of y_n^{test}.

For different numbers of compositional parts D, Figures 3–5 contain plots of the average MSE against the contamination level θ for various sample sizes N and scaling factors c of the covariance matrix in pivot coordinates. Similarly, the average MSEP is displayed in Figures 6–8.

Regarding coefficient estimates, all methods are accurate when there is no contamination ($\theta = 0$). As contamination increases, OLS is quickly influenced by the outliers, yielding the highest MSE of all methods. The MSE of MM also increases continuously for increasing contamination level, which is expected since MM is only robust to rowwise outliers but not to cellwise outliers. Our proposed method BF-MI is however very accurate for up to 5% contamination and close to the hypothetical IF-MI case using an ideal outlier filter. While the MSE of BF-MI increases for larger contamination levels, it is generally still lower than that of MM, although the difference between the two becomes small as variability in the data increases (increasing c). The MSE of IF-MI remains fairly low for 10% contamination, which indicates that the outlier filtering step is crucial for the performance of our proposed method, but under 20% contamination the MSE of IF-MI increases as well. All in all, the assessment based on MSE suggests that BF-MI offers improved performance over existing techniques for regression analysis with compositional covariates.

As to prediction performance, the results are comparable to the above. OLS in general has the highest MSEP, and BF-MI outperforms MM. In many settings, the MSEP of ShS is comparable to that of BF-MI or somewhat higher, but ShS is unstable if the ratio of N/D is small. Furthermore, ShS cannot be applied for $D = 20$ and $N = 50$ or $N = 100$, since the number of pairwise logratios is larger than the number of observations in those cases. 3S is also similar to BF-MI in terms of MSEP while the contamination level is 5% or lower, but each method is performing slightly better than the other in some settings with higher amounts of contamination. While 3S predicts better for lower values of D when the data are more scattered (higher values of c), BF-MI has lower MSEP for $D = 20$.

Note that we also considered counterparts to IF-MI and BF-MI that use single imputation instead of multiple imputation. The results were very similar. This is actually expected, as the main purpose of multiple imputation is to improve

standard errors (Little and Rubin, 2002; Van Buuren, 2012; Cevallos Valdiviezo and Van Aelst, 2015), but there should not be large differences in the point estimates of the coefficients (compared to single imputation). Consequently, the bias component of the MSEP should be similar, and the MSEP can only be improved by reducing the variance of the predictions. In multiple imputation, such reduction in variance would in turn require to decrease the correlation between predictions based on different imputed datasets. However, when the number of imputed cells is rather small, the predictions based on different imputed datasets are still highly correlated. An improvement in prediction performance via multiple imputation can only be expected for larger fractions of imputed cells (cf. results and recommendations of Cevallos Valdiviezo and Van Aelst, 2015), where the correlation between imputed data sets is sufficiently reduced.

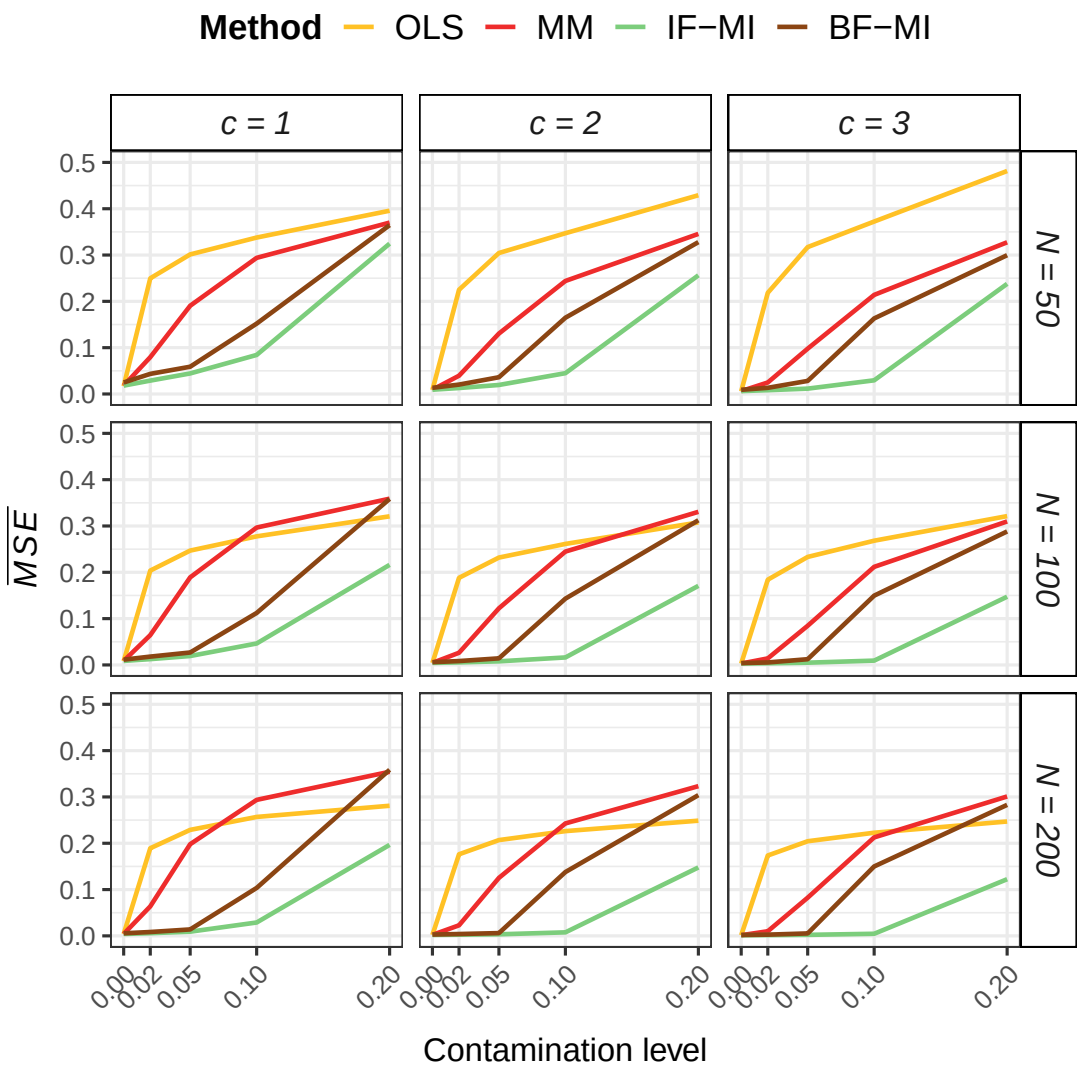


Figure 3: Results from 1000 simulation runs for the scenario with $D = 5$ compositional parts: the average MSE of coefficient estimates from regression methods in pivot coordinates is plotted against the contamination level θ for various sample sizes N and scaling factors c of the covariance matrix in pivot coordinates.

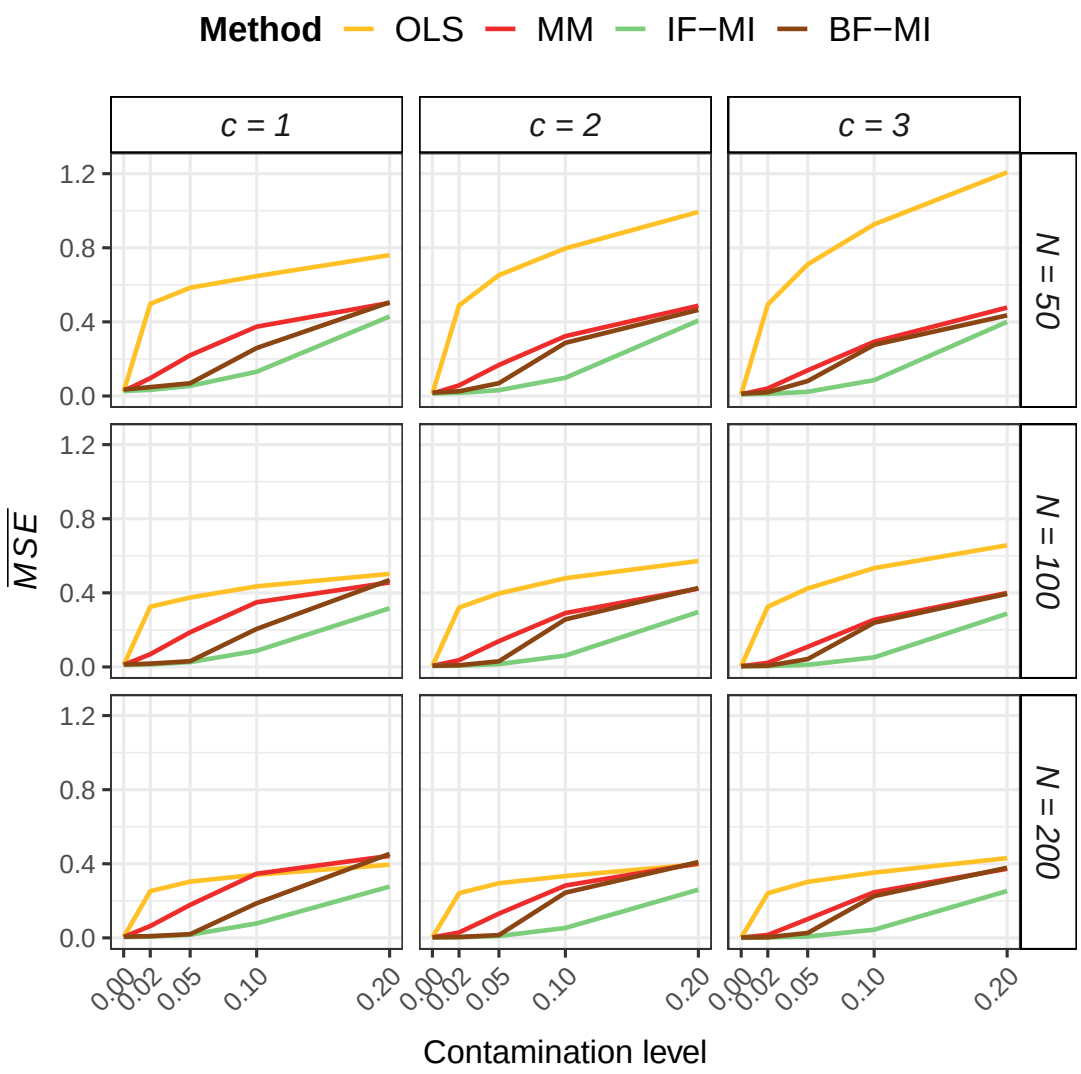


Figure 4: Results from 1000 simulation runs for the scenario with $D = 10$ compositional parts: the average MSE of coefficient estimates from regression methods in pivot coordinates is plotted against the contamination level θ for various sample sizes N and scaling factors c of the covariance matrix in pivot coordinates.

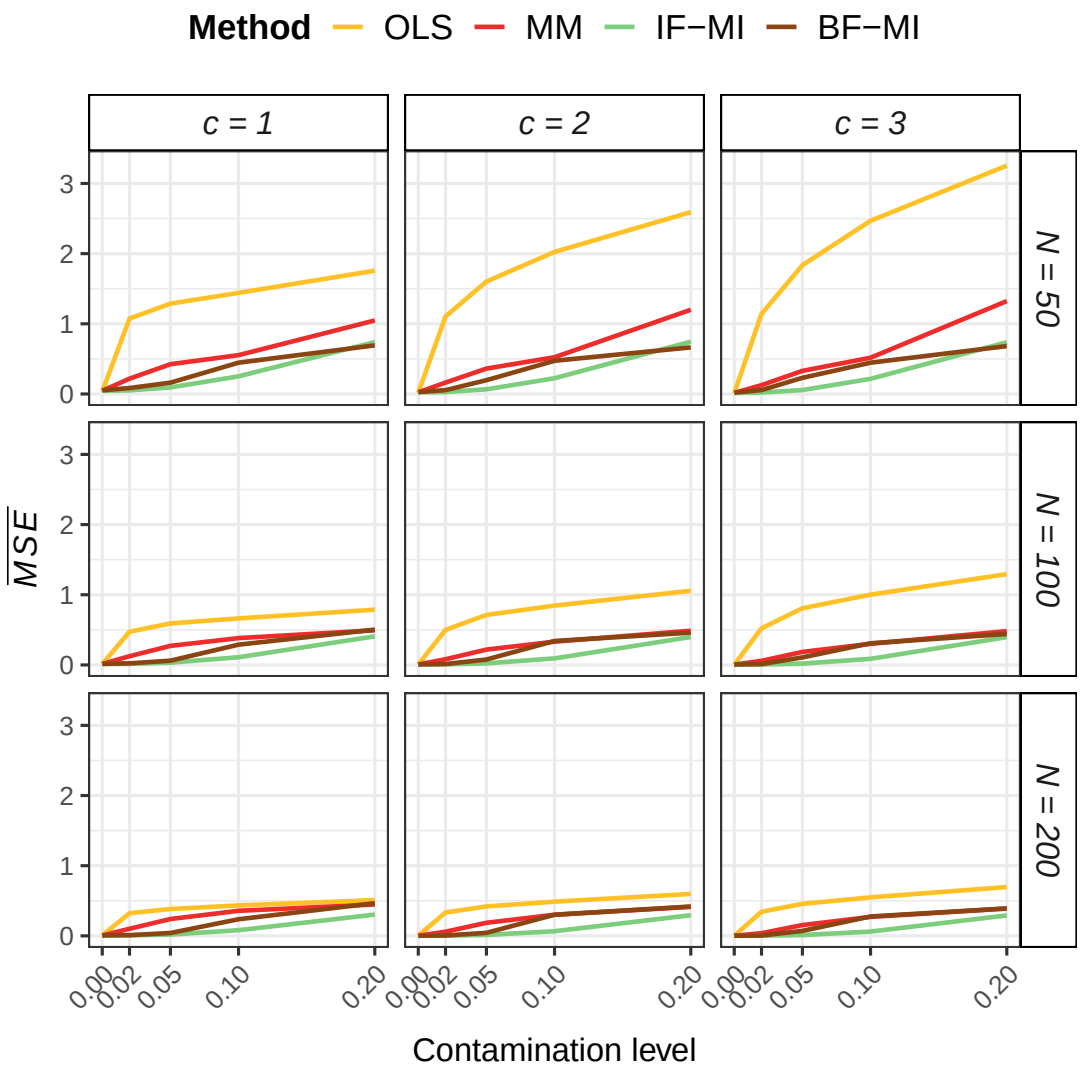


Figure 5: Results from 1000 simulation runs for the scenario with $D = 20$ compositional parts: the average MSE of coefficient estimates from regression methods in pivot coordinates is plotted against the contamination level θ for various sample sizes N and scaling factors c of the covariance matrix in pivot coordinates.

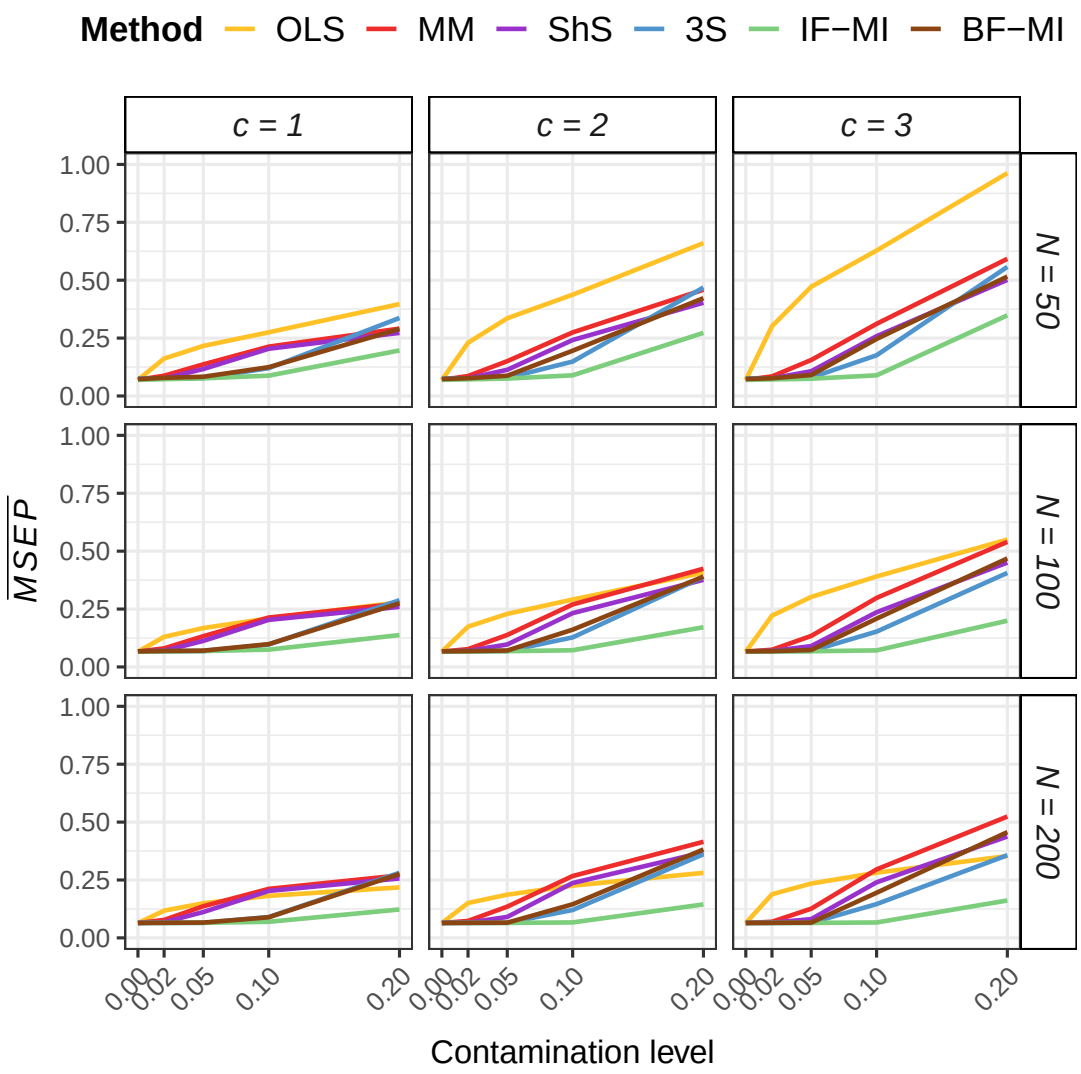


Figure 6: Results from 1000 simulation runs for the scenario with $D = 5$ compositional parts: the average MSEP for different regression methods is plotted against the contamination level θ for various sample sizes N and scaling factors c of the covariance matrix in pivot coordinates.

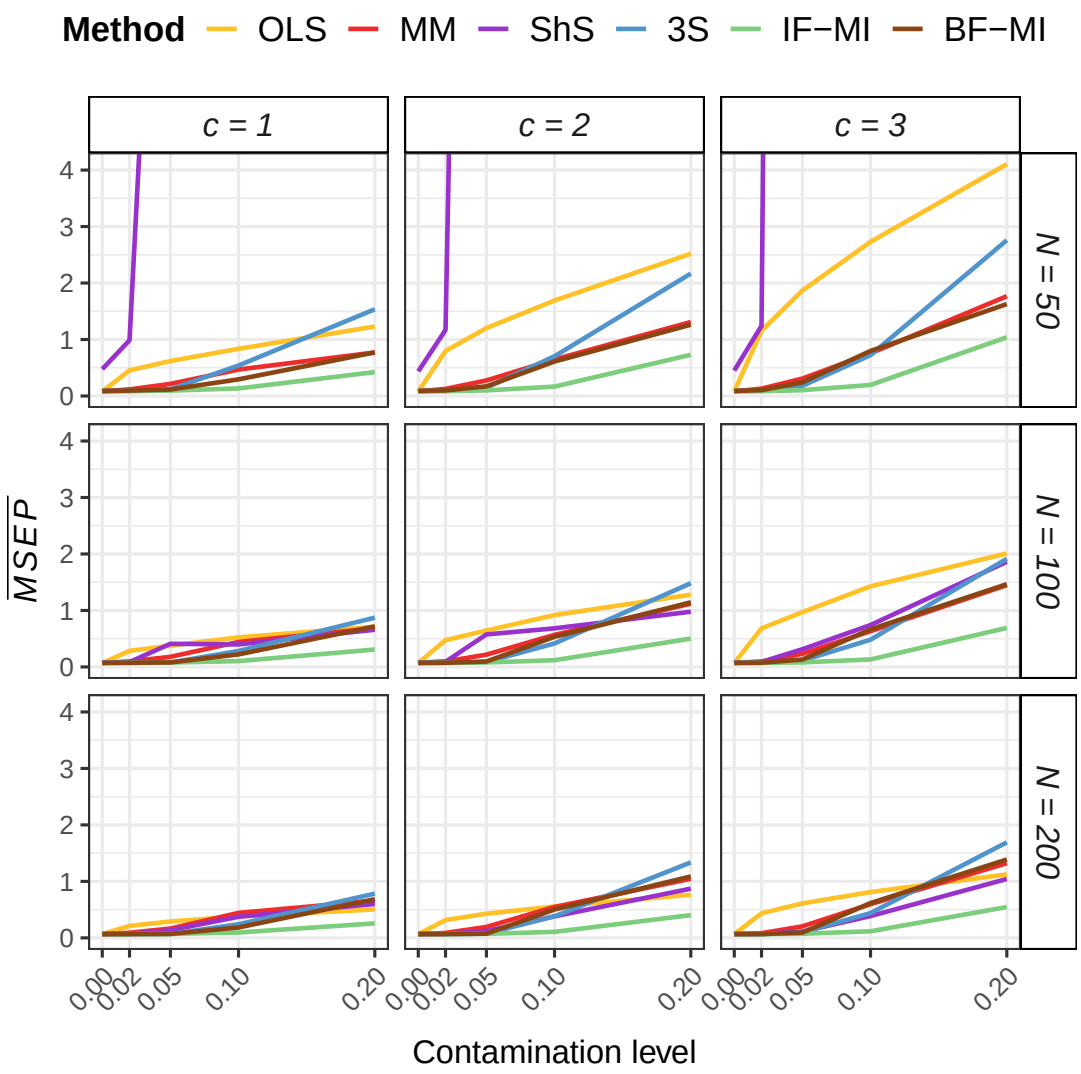


Figure 7: Results from 1000 simulation runs for the scenario with $D = 10$ compositional parts: the average MSEP for different regression methods is plotted against the contamination level θ for various sample sizes N and scaling factors c of the covariance matrix in pivot coordinates.

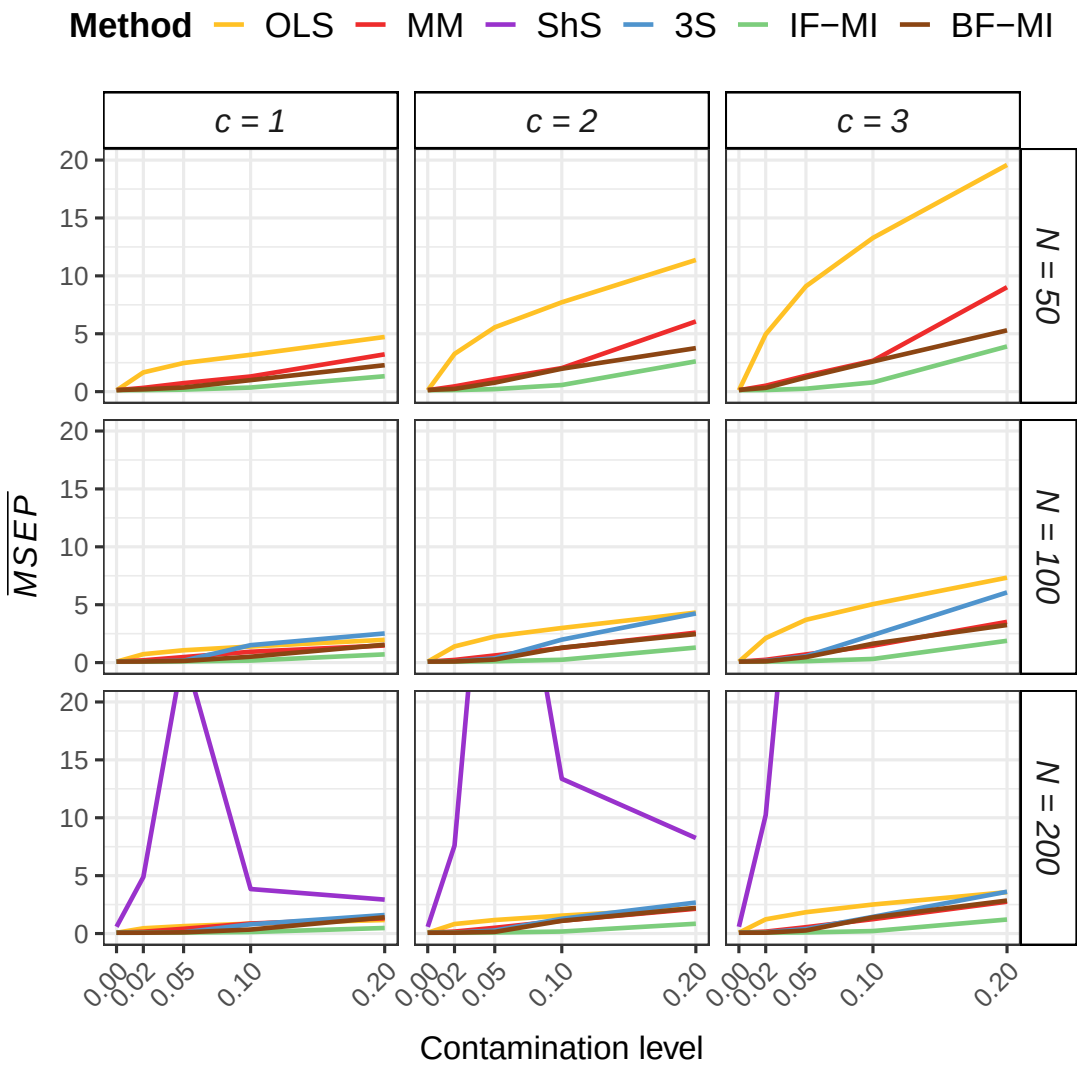


Figure 8: Results from 1000 simulation runs for the scenario with $D = 20$ compositional parts: the average MSEP for different regression methods is plotted against the contamination level θ for various sample sizes N and scaling factors c of the covariance matrix in pivot coordinates. Note that the shooting S-estimator (ShS) cannot be applied for $N = 50$ and $N = 100$, as the number of pairwise logratios is larger than the number of observations. In addition, the 3-step regression estimator (3S) is unstable for $N = 50$, yielding an average MSEP that is outside the depicted range on the y-axis.

3 Weighted pivot coordinates for PLS-based marker discovery in high-dimensional compositional data

As discussed in Section 1.2.3, PLS regression is a well-established method to identify which (in a large set of) explanatory variables are significant (markers) in relation to a response variable of interest, including cases where covariates are of compositional nature. Using the pivot coordinate representation (Section 1.1.2) for compositional explanatory variables allows to investigate each compositional part in terms of its relative importance, as used e.g. in Kalivodová et al. (2015) for PLS discriminant analysis (PLS-DA).

The method presented in this chapter extends previous work in PLS modelling with compositional data by using weighted pivot coordinates (Section 1.1.3) instead of the ordinary ones with a newly introduced weighting strategy aiming to enhance the identification of markers (Štefelová et al., 2021b). This is achieved by defining weights which focus on the correlation structure between a real-valued response variable and pairwise logratios aggregated into the first pivot coordinate in order to downplay the effect of irrelevant logratios and enhance the most relevant ones in relation to the outcome variable (Section 3.1). The practical relevance of the proposed model is demonstrated by its application to the identification of metabolite signals associated with the emission of greenhouse gases from cattle (Section 3.2). Its performance is further investigated through a simulation study (Section 3.3). The results provide evidence of the overall improved ability of the proposed weighted pivot coordinates approach to distinguish between markers and non-markers, increasing sensitivity, although resulting in slightly worse specificity.

3.1 Proposed weighting scheme

Although weighted pivot coordinates were introduced in Hron et al. (2017), the weighting schemes suggested there are not appropriate for regression analysis because they were only meant to downplay parts which were not proportional enough (have poor relative relationship) to the pivot part. In order to make a sensible choice of weights for a regression purpose, we must first determine what

we understand as a marker in our context. We aim for a compositional part to be identified as a marker if a relatively significant number of pairwise logratios including that part are strongly associated with the response variable Y. Moreover, considering the pairwise logratios where the part of interest is in the numerator, that strong association should be (possibly with a few exceptions) in one direction, either positive or negative.

Accordingly, we propose to construct weighted pivot coordinates (9) using weights $\gamma_d^{(l)},\ d = 2, \ldots, D,\ l = 1, \ldots, D$, defined as follows:

$$\gamma_d^{(l)} = \frac{\tilde{\gamma}_d^{(l)}}{\sum_{d=2}^{D} \tilde{\gamma}_d^{(l)}}, \tag{12}$$

with

$$\tilde{\gamma}_d^{(l)} = \left| \int_0^{r_d^{(l)}} \hat{f}^{(l)}(\lambda)\, d\lambda \right|, \quad r_d^{(l)} = \operatorname{cor}\left(Y, \ln \frac{x_1^{(l)}}{x_d^{(l)}}\right),$$

$$\hat{f}^{(l)}(\lambda) = \frac{1}{\nu(D-1)} \sum_{d=2}^{D} \mathcal{K}\left(\frac{\lambda - \tilde{r}_d^{(l)}}{\nu}\right), \quad \tilde{r}_d^{(l)} = \begin{cases} 0, & \text{if } \left|r_d^{(l)}\right| < o^{(l)}, \\ r_d^{(l)}, & \text{otherwise,} \end{cases}$$

$$o^{(l)} = 2 \times \min\left(\frac{\sum_{d=2}^{D} \mathcal{I}(r_d^{(l)} \geq 0)}{D-1}, \frac{\sum_{d=2}^{D} \mathcal{I}(r_d^{(l)} < 0)}{D-1} \right),$$

where $\hat{f}$ is a kernel density estimator, $\mathcal{K}$ is a Gaussian kernel function (defined as $\mathcal{K}(\lambda) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\lambda^2}$), ν (set to $\nu = 0.05$) is the bandwidth used and $\mathcal{I}$ is an indicator function.

Thus, for the lth part, rearranged into the first position as $x_1^{(l)}$, the set of correlations $r_2^{(l)}, \ldots r_D^{(l)}$ is smoothed using kernel density estimation (Everitt and Hothorn, 2011), with the correlations under the given threshold being set to zero in order to estimate the density, and the weight $\tilde{\gamma}_d^{(l)}$ is computed as the area under the curve from zero to the value of the correlation $r_d^{(l)}$. The rationale behind this weighting scheme is to minimise the influence of logratios that are not related to the response Y, so that higher weights are given to logratios strongly correlated to Y. Among these, the procedure emphasises those logratios representing the main trend in the distribution of $r_2^{(l)}, \ldots r_D^{(l)}$ by using a kernel density. In order

to prevent from false positives, correlations with absolute value smaller than a cut-off value $o^{(l)}$ are set to zero when conducting kernel density estimation. The value of $o^{(l)}$ modulates the effect of the weighting, which is downplayed with increasing values of $o^{(l)}$. Therefore, the value given to $o^{(l)}$ is higher when there is no clear trend in the distribution of the correlations and vice versa. For instance, when all correlations are positive, then $o^{(l)} = 0$, the density is estimated from the unaltered set of correlations and, as a consequence, the logratios strongly correlated with Y are highlighted. On the other hand, when half of the correlations are positive and half are negative, then $o^{(l)} = 1$, and all correlations are taken to be zero for the density estimation. Thus, the value of the area under the curve from 0 to $r_d^{(l)}$ is practically the same for any d (apart from the cases where $r_d^{(l)}$ are the closest to 0), so only logratios very weakly correlated with Y are suppressed, while the rest are treated equally. The final normalised weight results from dividing each $\tilde{\gamma}_d^{(l)}$ by the sum of all of them.

We use some results from the exemplary case study in Section 3.2 to illustrate the functioning of the proposed weighting scheme. The compositional parts consist of a collection of integral values (normalised areas under the peaks of spectra generated by nuclear magnetic resonance). The interest is in identifying integrals which are potentially relevant markers associated with methane yield from cattle as response variable. The upper graphs in the subfigures of Figure 9 display the histograms of the correlations between the response variable and the pairwise logratios including the part (integral) in the numerator. Figures 9a and 9b correspond to two parts that should be identified as meaningful markers, while Figures 9c and 9d shows two parts that should not be considered as meaningful markers. The graphs at the bottom compare the respective weights assigned to the logratios when using ordinary pivot coordinates (PC) and weighted counterparts with cut-off value $o^{(l)}$ as defined in (12) (WPC), or using the extreme cases $o^{(l)} = 0$ (WPC 0) and $o^{(l)} = 1$ (WPC 1). Note that for PC, each logratio is applied the uniform weight $1/(D-1)$. In the case of an obvious trend (Fig. 9a), the irrelevant logratios are downplayed, while the meaningful ones are enhanced. In Fig. 9b, the histogram suggests that the relative importance of the part in the composition is positively associated with the response variable. However, there are two deviating logratios with strong negative correlation that are attenua-

ted using the proposed weighting scheme (12). When no or only a few relevant logratios are present (Fig. 9c), then giving higher weight to those with higher correlation should not affect the overall lack of significance of the weighted pivot coordinate. Finally, when strong correlations are found in both directions with no clear trend (Fig. 9d), the scheme implies a neutralising effect by weighting both sides similarly.

3.2 Application to high-dimensional metabolomic data

To illustrate the functioning of the proposed PLS regression model based on weighted pivot coordinates we use a real dataset kindly provided by the Scotland's Rural College (UK). It consists of high-throughput spectral profiles, representative of metabolite signals, acquired by nuclear magnetic resonance (NMR) spectrometry on rumen fluid samples from cattle. The raw samples went through a number of ordinary pre-processing stages, including phase and baseline correction, binning to integrate the area under the signal peaks, and normalisation by referencing all the integrals to a same integral (corresponding to methyl of propionate), which resulted in $D = 127$ integrals per animal sample ($N = 211$ samples in total). A few cases of zero integrals were assumed to correspond to values below the limit of detection and were imputed based on the information from the other signals using the logratio expectation-maximisation (EM) algorithm (Palarea-Albaladejo and Martín-Fernández, 2008). Methane yield (CH_4 in grams per kilogram of dry matter intake) was also measured for each individual animal using respiration chambers. Information about the diet type used to feed the animals was also recorded (either concentrate, mixed or forage based diet).

Ruminants are known to be important contributors to the world production of greenhouse gases, particularly methane which is strongly implied in global warming. Methane production is mainly associated with fermentation of feed in the rumen. The purpose of this case study is to identify the most relevant metabolite signals (markers) associated with cattle methane emissions. PLS regression modelling is an adequate approach given the large number of signals and the multicollinearity between them. $L = D = 127$ models (11) are considered with signals represented through weighted pivot coordinates (Section 1.1.3) with weights constructed as proposed in Section 3.1. Then, each rotated coordi-

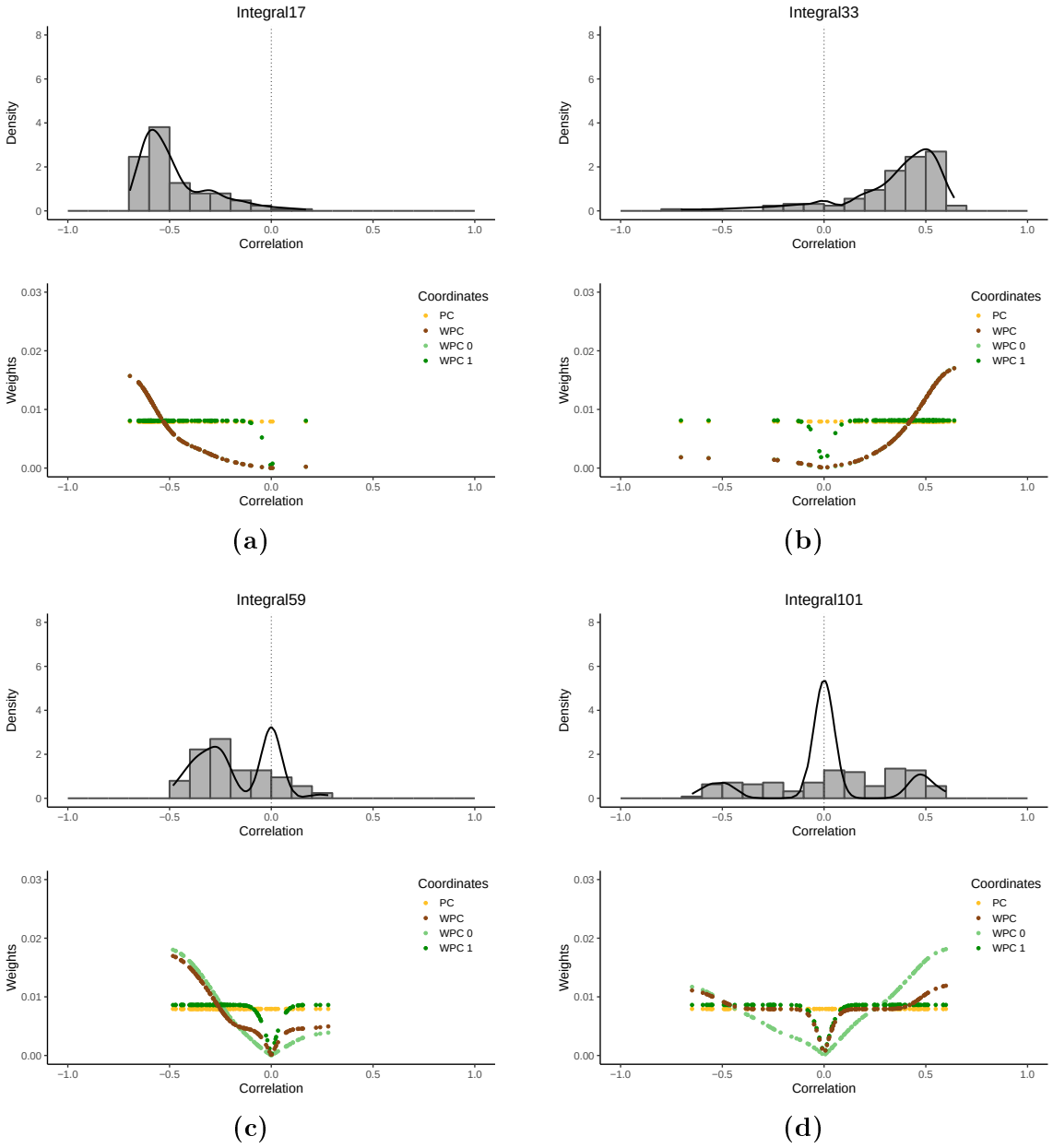


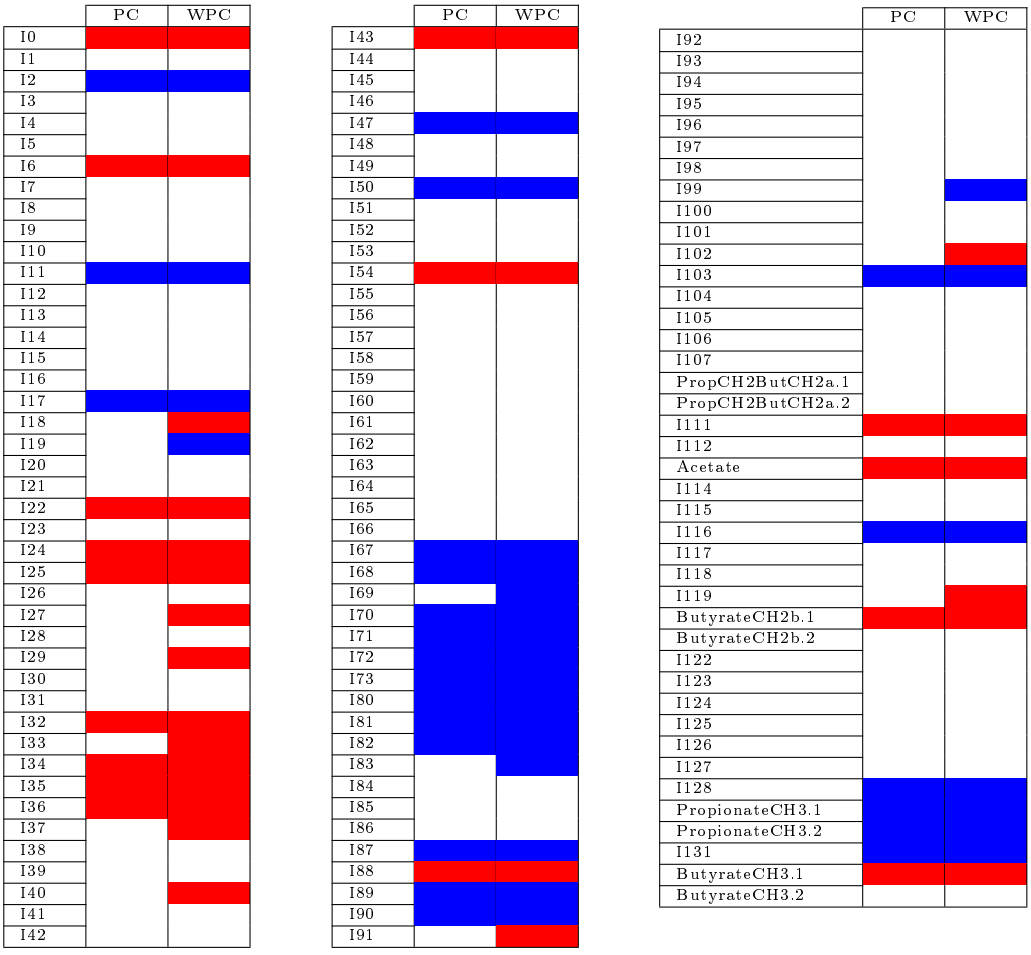
Figure 9: Illustration of the proposed weighting scheme to identify potential markers–(a) and (b)–and non-markers–(c) and (d)–in a collection of NMR integral values. The upper section of each subfigure displays the histogram of the correlations between the response variable and the pairwise logratios containing the integral in the numerator. The curve represents the respective estimated kernel density. The lower section of each subfigure compares the weights assigned to the logratios when using ordinary (PC) and weighted pivot coordinates with cut-off value $o^{(l)}$ as defined in (12) (WPC) or equal to either 0 (WPC 0) or 1 (WPC 1) as extreme cases.

nate system highlights the relative role of one integral in the NMR composition. For comparison, signals are then represented through ordinary pivot coordinates (Section 1.1.2). The response variable CH_4 is also mapped into real space using a simple log-transformation to better accommodate its scale and model assumptions. PLS modelling is conducted as described in Section 1.2.3. The optimal models (determined by the randomization test approach) consist of two PLS components (CV RMSEP = 0.17 and CV R^2 = 0.51). The estimated bootstrap standardised regression coefficients are computed from $B = 1000$ bootstrap resamples. Bonferroni's correction is applied for statistical significance testing. Table 3 compares integrals identified as markers using ordinary pivot coordinates (PC) and weighted pivot coordinates (WPC). The integral signals are distinguished using the letter I followed by a numeric ID, although note that some which represent known metabolites are named after these. PC and WPC based results differ in 13 integrals (39 vs. 52 respectively). Namely, these are all integrals only identified when the proposed weighting scheme is applied, i.e. they are missed when using ordinary pivot coordinates, and their corresponding histograms of correlations (Figure 10) in fact suggest that they are associated with the methane yield.

In agreement with previous modelling work based on only the ruminal volatile fatty acids (VFA) composition (Palarea-Albaladejo et al., 2017), both procedures identify as markers the signals of acetate and species of butyrate and propionate (with the direction of the associations being also coincident with previous results). Moreover, integrals I22 and I125 (forming the VFA called valerate) are both non-significant, which is also consistent with the previous results based on VFA only. On top of this, the current analysis using high-throughput data provides further insight by identifying, amongst others, integrals I67-I73, I80-I83 and I87, which are known to belong to glucose protons. However, note that some signals in these regions are only identified when weighted pivot coordinates are used (I69 and I83), hence suggesting that this approach provides a higher level of sensitivity.

Figure 11 shows heatmaps of the correlations between pairwise logratios and response variable. We can observe that, in general, the pairwise logratios of a marker deviating from the main trend are defined over another marker (in the denominator) associated with the response variable in the same direction. For example, the logratios of I33 are almost all positively correlated with methane production,

Table 3: Peak integrals in NMR spectral data identified as markers using ordinary pivot coordinates (PC) and weighting pivot coordinates (WPC). Red (resp. blue) colour refers to significant markers in positive (resp. negative) direction.



many of them even strongly correlated. But two of them are strongly correlated in negative direction (see Fig. 9b). These two are the logratios of I33 over other integrals positively associated with the response (I34 and I35). Note that I33 is identified as marker only when weighted pivot coordinates are used. Moreover, the deviating pairwise logratio of I35 is the one over I34.

Compositional PLS biplot is constructed using the loadings of the D first weighted pivot coordinates (Figure 12). The biplot shows that markers negatively associated with methane yield (in blue) are mostly linked to the concentrate diet, whereas markers showing a positive association (in red) are more related with mixed and forage diets.

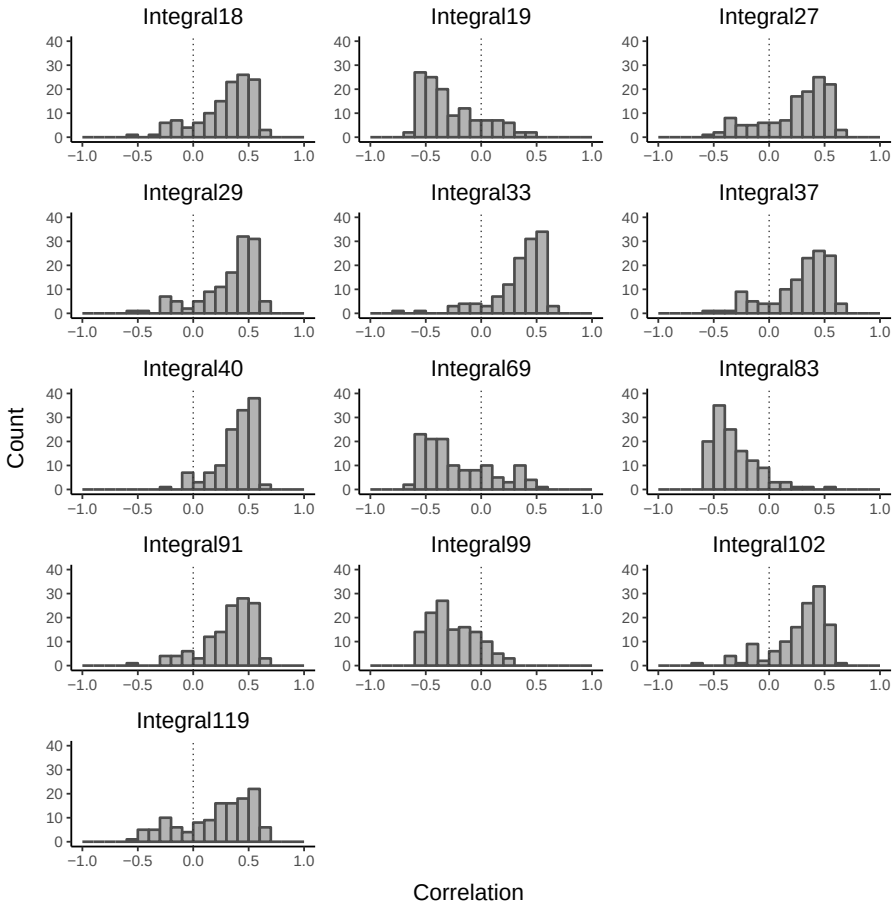


Figure 10: Histograms of correlations between pairwise logratios of NMR integrals and methane yield for integrals only identified as markers by using weighted pivot coordinates.

3.3 Simulation study

A simulation study is conducted to assess and compare the performance of ordinary and weighted pivot coordinates for marker identification through PLS regression across a range of parameter settings. In particular, we set $D = \{100, 200, 300\}$ as number of compositional parts, $n = \{D/2, D, 2D\}$ as number of observations, $M = \{D/25, D/10, D/5\}$ as number of markers associated in positive and negative direction. Results from each combination of parameter settings are assessed over 500 simulation runs.

For each simulation run, the data are simulated so that the first M odd compositional parts represent markers associated with the response in positive direction, while the first M even compositional parts represent markers associated

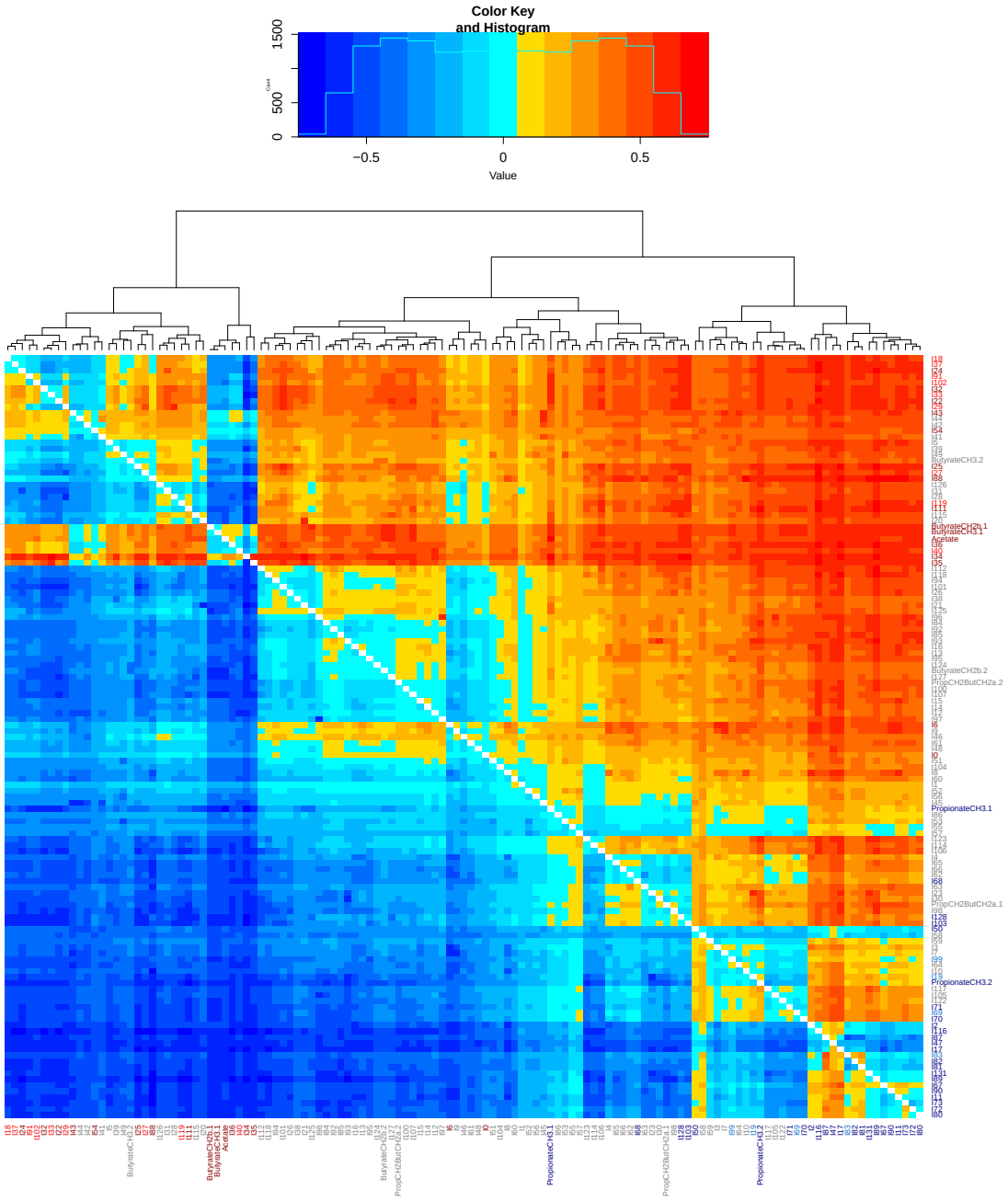


Figure 11: Heatmap of the correlations between pairwise logratios of NMR integrals and methane yield. The y- (resp. x-) axis show the integral used in the numerator (resp. denominator). Identified markers using either ordinary pivot coordinates (PC) or weighted pivot coordinates (WPC) are coloured in red or blue according to the direction of the relationship with the response variable (positive or negative respectively). For each colour, dark shade indicates markers identified by both methods, whereas light shade refers to those identified only by WPC. Labels in grey refer to signals not identified by any method.

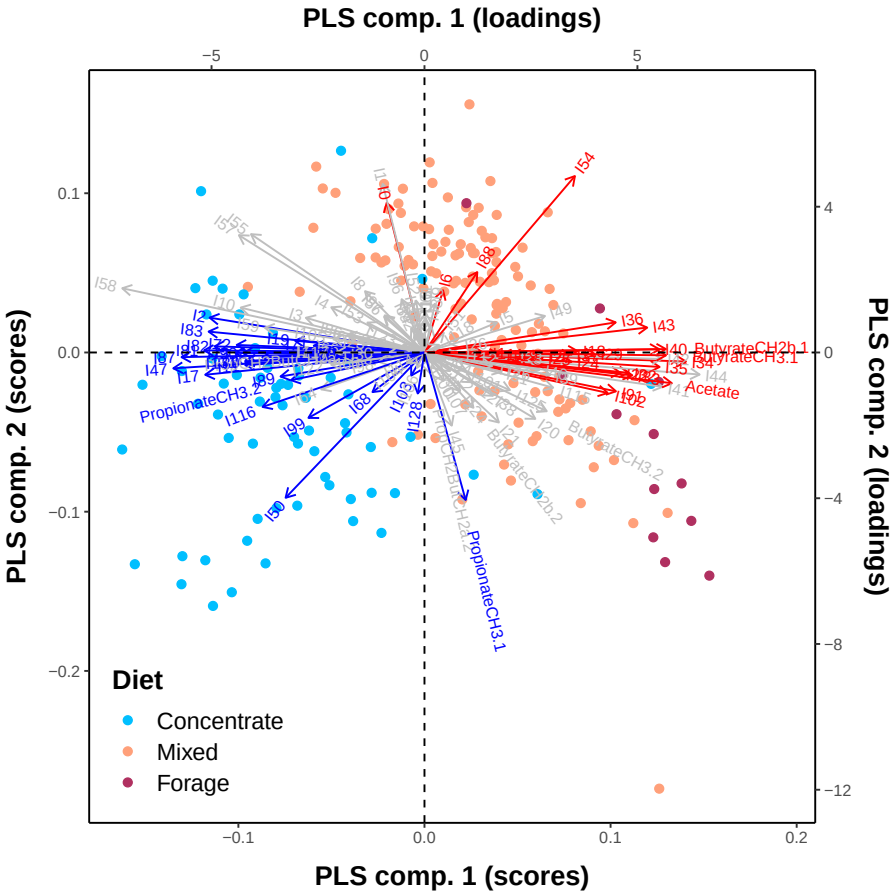


Figure 12: PLS biplot using weighted pivot coordinates. Identified markers in positive (resp. negative) direction are coloured in red (resp. blue). The points are distinguished according to diet type. Rays in grey refer to signals not identified as markers. The dashed lines indicate the origin for the first and second PLS components (PLS comp. 1 and PLS comp. 2). A 39.96% of explanatory data variance (resp. 56.64% of response data variance) is explained by the first two PLS components: 31.74% by PLS comp. 1 and 8.22% by PLS comp. 2 (resp. 41.96% by PLS comp.1 and 14.68% by PLS comp. 2).

with the response in negative direction. Compositions are generated through pivot coordinates having a multivariate normal distribution. The covariance matrix of the pivot coordinates is chosen so that there is: 1) positive covariances between every pair from the first M odd pivot coordinates, 2) positive covariances between every pair from the first M even pivot coordinates and 3) negative covariances between each of the first M odd pivot coordinates and each of the first M even pivot coordinates. Moreover, variances for the first $2M$ pivot coordinates are set higher than for the remaining ones. The values of the response variable are

generated through a regression equation, with the first $2M$ pivot coordinates used as explanatory variables, where positive (resp. negative) regression coefficients are set for the first M odd (resp. even) pivot coordinates. The optimal number of PLS components is determined using the randomization test approach as described in Section 1.2.3 (the same number of PLS components is used for both, ordinary pivot and weighted pivot coordinates).

The simulated data generation process is outlined in the following points:

1. Pivot coordinates are generated as $\boldsymbol{z}_n = (z_{n1}, \ldots, z_{n,D-1})$ from a $\mathcal{N}_{D-1}(\mathbf{0}, \boldsymbol{\Sigma})$ distribution, $n = 1, \ldots, N$, with covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})$ containing elements

$$\sigma_{ij} = \begin{cases} 2, & \text{if } i = j \leq 2M, \\ 1, & \text{if } i = j > 2M, \\ 0.5 \times (-1)^{i+j} & \text{if } i \neq j,\ i, j \leq 2M, \\ 0, & \text{otherwise,} \end{cases} \qquad i, j = 1, \ldots, D-1.$$

2. The pivot coordinates $\boldsymbol{z}_n$ are transformed according to (1.1) to obtain the associated compositions $\boldsymbol{x}_n = (x_{n1}, \ldots, x_{nD}) = \text{ilr}^{-1}(\boldsymbol{z}_n),\ n = 1, \ldots, N$.

3. Values for the response variable are obtained as

$$y_n = \beta_1 z_{n1} - \beta_2 z_{n2} + \ldots + \beta_{2M-1} z_{n,2M-1} - \beta_{2M} z_{n,2M} + \varepsilon_n,$$

with the error terms $\varepsilon_n \sim \mathcal{N}(0, 1),\ n = 1, \ldots, N$, and regression coefficients $\beta_k \sim \mathcal{U}(0.1, 1),\ k = 1, \ldots, 2M$.

This scheme ensures that markers and non-markers in each simulation scenario can be clearly distinguished (see Fig. 13), while marker identification using the PLS regression model is still not perfect. For the simulated markers, it can be observed that a marker's pairwise logratios most correlated with the response variable are those defined over other markers associated with the response variable in opposite direction. Moreover, the least correlated (or deviating from the main trend) logratios for a marker are those defined over other markers associated with the response variable in the same direction. In general, this resembles the relationships observed in the NMR dataset used in Section 3.2 (see Fig. 11).

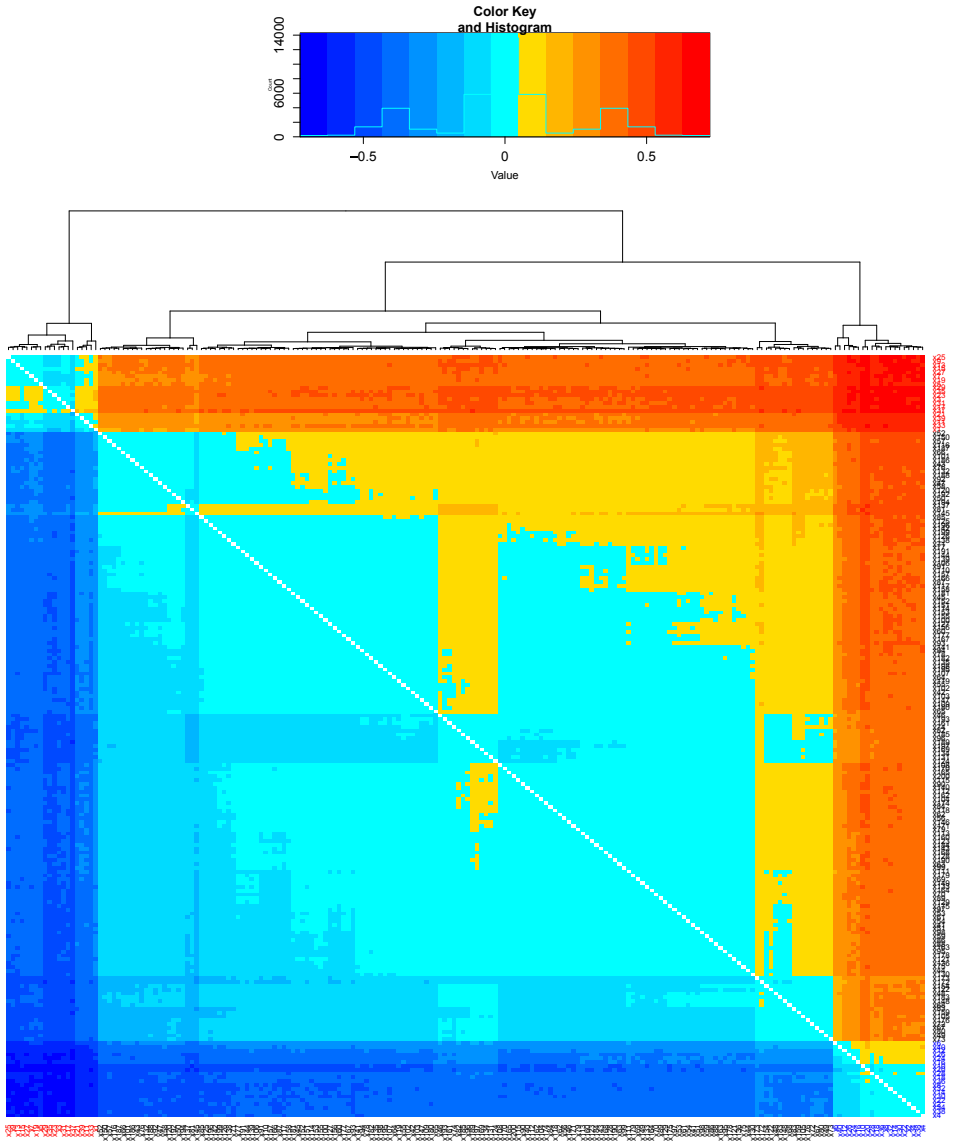


Figure 13: Examplary heatmap of correlations between pairwise log-ratios and response variable in simulation scenario. The y- (resp. x-) axis corresponds to the part in the numerator (resp. denominator). Markers are coloured red or blue according to the relationship with the response variable (positive or negative respectively).

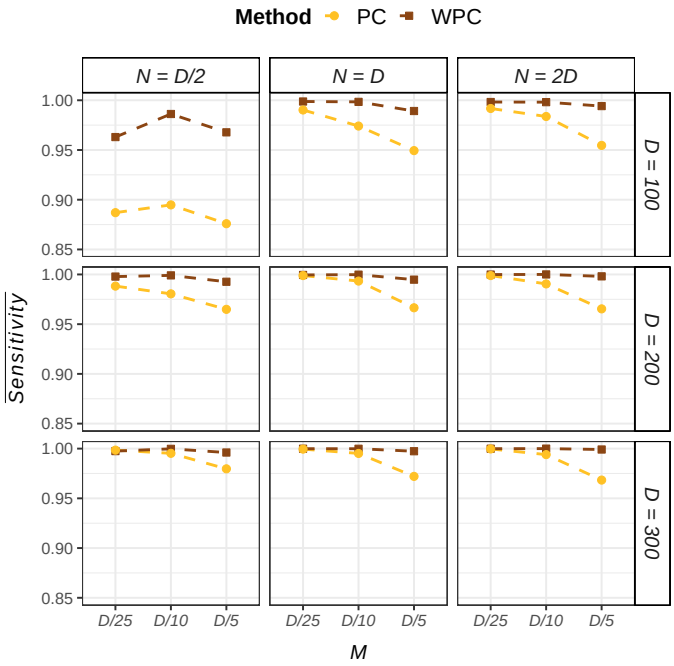
For example, parts x_1, x_3 and x_5 in the simulation scenario would be equivalent to integrals I34, I35 and I33 in the case study.

The ordinary and weighted pivot coordinates approaches are compared according to their ability to distinguish between genuine markers and non-markers. Note that we can consider a binary classifier (i.e marker or non-marker) since

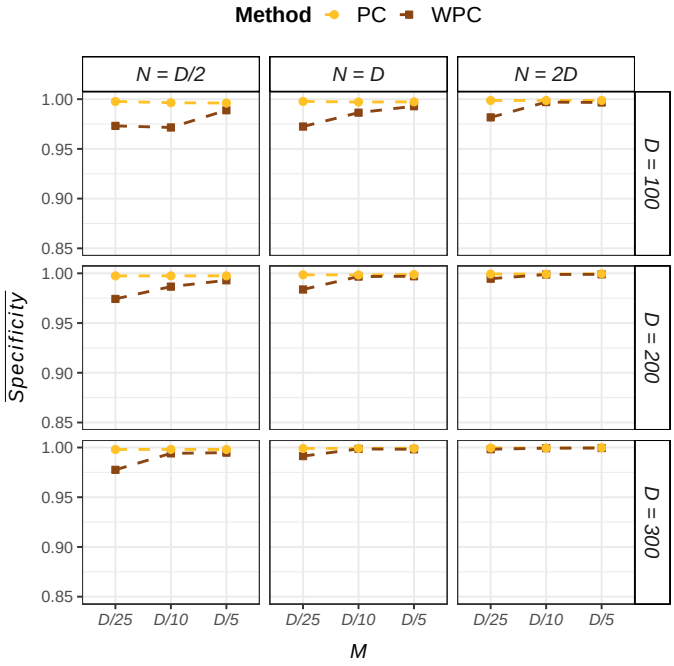
in the simulations a positive marker is never identified as negative or vice versa. The performance is compared in terms of

1. *Sensitivity*, i.e true positive rate (ability to detect actual markers),
2. *Specificity*, i.e true negative rate (ability to not detect non-markers as markers).

For reference, the nearer the values of these measures are to 1 the better the performance is. The results for each scenario are summarised by the means of these two measures across simulation runs and are displayed in Figure 14 (PC and WPC denote ordinary and weighted pivot coordinates respectively). In all the scenarios it can be observed that using WPC provides higher sensitivity, but some lower specificity. The benefit of WPC can particularly be seen in cases with higher percentage of markers, when PC are outperformed in terms of sensitivity, while specificity is close to 1.



(a)



(b)

Figure 14: Results from 500 simulation runs: the average (a) sensitivity and (b) specificity for the two different approaches is plotted against the number of markers associated in positive and negative direction M for various numbers of compositional parts D and sample sizes N.

4 Compositional approach in time-use epidemiology

Time-use epidemiology is a subfield of biostatistics which focuses on the relationship between health and movement behavior patterns in populations. Daily movement behavior (time-use) data are usually reported as amounts of time spent in various activities during a certain time period. Nowadays, the data are generally collected from people wearing some sort of accelerometers, commonly within one week. The raw signals of the accelerations measured by the accelerometer are processed and evaluated by various methods to obtain the required data (Burchartz et al., 2020), e.g. using the `GGIR` package (Migueles et al., 2019) of the R software for statistical computing (R Core Team, 2021). The basic partition of behaviors is made in terms of sleep, sedentary behavior (SB) and physical activity (PA) of different intensities: light (LPA), moderate (MPA) and vigorous (VPA). However, with some devices, the distinction between sleep and SB is not possible, therefore in such cases they are not worn overnight and the corresponding analysis, limited to wake-time day, does not include sleep. Note that the accelerometers are usually not worn during water-based activities, or they are taken off during a day for another reasons, so to compute sleep duration as the time adding up to 24 hours per day would be inadequate.

In either way, whether the observations available represent 24-hours movement behavior vectors (MB) or wake-time movement behavior vectors (WMB), they meet properties of compositions. More time spent in one activity necessarily causes less time spent in another one(s). As the parts of (W)MB are interrelated, they should not be analysed independently of each other. Accordingly, the logratio methodology is appropriate tool for the analysis of time-use data. Hence, the particular scale in which the parts of (W)MB are measured (e.g. using hours/week, minutes/day, or their expression in percentage) is irrelevant. The first study, where compositional approach to the analysis of movement behavior data was introduced and discussed in a comprehensive and statistically-principled means, was the work by Chastin et al. (2015). Over the last few years, there has been an increasing awareness of the suitability of compositional approach in time-use epidemiology and a large number of studies examining movement behavior patterns using compositional data analysis have been published.

During my Ph.D. study, I have collaborated in several articles from time-use epidemiology that are listed in the Introduction. These were mostly related to the projects "Application of a novel compositional data analysis approach for the evaluation of combined effects of 24-hour lifestyle behaviors on childhood obesity" and "Influence of obesity on changes in long-term physical activity of older adults women in context of built environment: a prospective study" funded by the Czech Science Foundation under reg. No. GA18-09188S, respectively No. GA18-16423S. To a great extend, the studies followed practice depicted in Chastin et al. (2015), Dumuid et al. (2017b) and Dumuid et al. (2017a) applying compositional descriptive statistics, basic visualization, linear regression models based on ilr coordinates and compositional version of isotemporal substitution analysis that allows to estimate a theoretical change in a health outcome resulting from a change in the duration of one type of behavior in favour of another one. My main methodological contributions to the field have been 1) demonstrating robust compositional analysis of time-use data (Štefelová et al., 2018), which is a novel aspect in the context of movement behavior research and 2) presenting an advanced visualization technique suitable for time-use epidemiology in the form of compositional PLS biplot based on newly introduced pivoting balances (Štefelová et al., 2021c). The former study is described in more detail in Section 4.1, with an extensive discussion about proper coordinate representation of WMB composition, while the latter analysis including the reasoning for coordinate representation of MB composition is summarized in Section 4.2.

4.1 Robust compositional analysis of wake-time movement behavior data

In the first application, we investigate wake-time movement behaviors of Czech adolescents, concretely its association with adiposity and the role of age in the behavior patterns, similarly as in Štefelová et al. (2018). Visualization techniques for compositional data and linear regression within logratio methodology are used for this purpose. These were firstly demonstrated in the context of time-use data in Chastin et al. (2015). Here, robust statistics is used instead in order to lessen the influence of possible outliers. The real dataset comes from the Faculty of Physical Culture in Olomouc. The participants were $N = 420$ healthy

adolescents (169 boys and 251 girls). The amount of time spent in WMB parts, i.e. in SB, LPA, MPA and VPA, were measured by hip-worn ActiGraph GT3X accelerometer (ActiGraph LLC, Pensacola, FL, USA). The distinction between the particular activities were made based on the Evenson's cut-points (Evenson et al., 2008). Body Mass Index (BMI) was calculated from the self-reported height and weight. Age and sex-adjusted BMI (zBMI) was used as an adiposity indicator. The calculation of zBMI and the respective weigh categories was done according to the WHO guideline (de Onis et al., 2007). Due to the lack of individuals in each category, only two groups were defined: underweight/normal and overweight/obese (with 344 vs. 76 adolescents).

First, we examine the differences in movement behaviors between adolescents from different weight groups. Table 4 shows the robust compositional center (computed as described in Section 1.1) of WMB computed for all adolescents, compared with that in the underweight/normal and overweight/obese groups.

Table 4: Robust center (expressed in %) of adolescents wake-time movement behavior data for the whole sample and for underweight/normal and overweight/obese subgroups.

Group	SB	LPA	MPA	VPA
All	60.54	33.66	3.83	1.97
Underweight/normal	60.57	33.52	3.85	2.05
Overweight/obese	60.81	33.89	3.74	1.57

The relative difference between the groups is visualized by the robust compositional mean barplot (Figure 15). The graph displays the ratio between each group's robust center and the overall robust center after the data are robustly centered (so that the comparison is made towards the barycenter). Thus, in the overweight/obese group, the proportion of time spent in VPA is reduced by 15.6% relatively to the overall mean composition. Accordingly, VPA stands out as a key driver of the difference between the underweight/normal and overweight/obese groups, stressing the lack of VPA time in adolescents with adiposity issues.

The continuous character of zBMI is used to obtain more precise information about the association between movement behaviors and adiposity using adequate

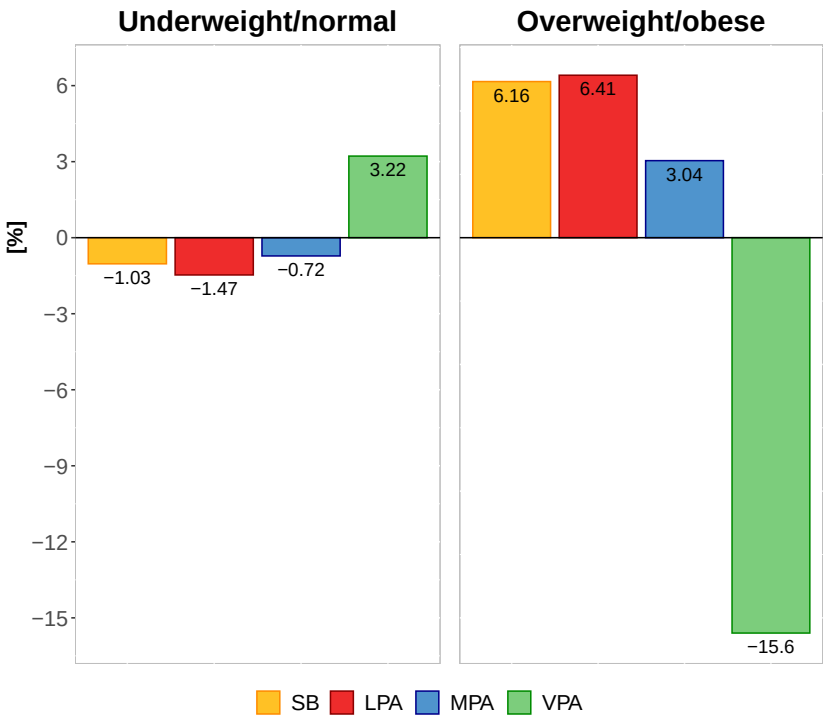


Figure 15: Robust compositional mean barplots for the underweight/normal and overweight/obese adolescents group.

regression models with zBMI set as the response variable and the 4-part WMB composition expressed by three ilr coordinates as covariates. In Štefelová et al. (2018), $L = D = 4$ models (11) are used with pivot coordinates representation $\left(z_1^{(l)}, z_2^{(l)}, z_3^{(l)}\right)^\top$, $l = 1, \ldots, 4$ for WMB composition resulting from sequential placement of each part originally at position l at the first position (Section 1.1.2). In each model, the focus lies on the coefficient estimate corresponding to the first coordinate which refers to the dominance of one particular part within the given composition. We give those coordinates of interest the symbolic notation SB_LPA.MPA.VPA, LPA_SB.MPA.VPA, MVPA_SB.LPA.VPA and VPA_SB.LPA.MPA (i.e. using an underscore to separate the parts in the numerator and the denominator of the logratio and a point symbol to split out the parts into the respective subgroup). MM-estimation of regression coefficients is performed (Section 1.2.2). The results extracted from the four models are summarized in Table 5. Statistically significant regression coefficients (at 5% significance level) are obtained for SB_LPA.MPA.VPA (with a positive sign)

and VPA_SB.LPA.MPA (with a negative sign). In other words, the findings reveal that the relative dominance of SB (with respect to the average contribution of the other parts) is in positive relationship with zBMI, while the relative dominance of VPA is in inverse relationship with zBMI.

Table 5: MM-regression coefficient estimates, standard errors and p-values associated with the first pivot coordinates extracted from the 4 models assessing the relationship between adolescents' zBMI and wake-time behaviors.

Covariate	Coeff.	Std. error	p-value
SB_LPA.MPA.VPA	0.376	0.190	0.048
LPA_SB.MPA.VPA	−0.308	0.196	0.118
MPA_SB.LPA.VPA	0.139	0.144	0.334
VPA_SB.LPA.MPA	−0.207	0.081	0.011

Although the process of extracting information about the first pivot coordinates from D models is common practice that enables to obtain relative information about each compositional part (Filzmoser et al., 2018), it has some drawbacks regarding time-use epidemiology. Here, it is reasonable to take into account the ordinal character of wake-time behaviors to construct the appropriate ilr coordinates as it is generally accepted that a higher health benefit is obtained from more physically demanding activities. Then, e.g. coordinate LPA_SB.MPA.VPA aggregates logratios with potentially contrary association with health outcome, namely ln(LPA/SB) vs. ln(LPA/MPA) and ln(LPA/VPA). Therefore, we further consider two models (10) – one with composition (SB, LPA, MPA, VPA) and the second one with composition (VPA, MPA, LPA, SB) expressed in pivot coordinates $\left(z_1^{(1)}, z_2^{(1)}, z_3^{(1)}\right)^\top$. Thus, we have coordinates with a symbolic notation SB_LPA.MPA.VPA, LPA_MPA.VPA and MPA_VPA, respectively VPA_MPA.LPA.SB, MPA_LPA.SB and LPA_SB (so, the two first coordinates are the same as in the previous case), where each one corresponds to the dominance of one activity with respect to the average of the more intense, respectively the less intense, activities. MM-estimation of regression coefficients is performed for the two models. The results are summarized in Table 6. In addition to the significance of SB_LPA.MPA.VPA and VPA_MPA.LPA.SB, only a weak significance (at 10% significance level) is observed for MPA_VPA (with a positive

sign) and LPA_SB (with a negative sign).

Table 6: MM-regression coefficient estimates, standard errors and p-values associated with the "ordinal" pivot coordinates from the two models assessing the relationship between adolescents' zBMI and wake-time behaviors.

Covariate	Coeff.	Std. error	p-value
SB_LPA.MPA.VPA	0.376	0.190	0.048
LPA_MPA.VPA	−0.194	0.166	0.243
MPA_VPA	0.212	0.114	0.063
VPA_MPA.LPA.SB	−0.207	0.081	0.011
MPA_LPA.SB	0.074	0.146	0.612
LPA_SB	−0.418	0.220	0.058

If we want coordinates that reflect the ordinal character of WMB but capture the information about the whole wake-time day structure, we can use so-called *pivoting balances*. These combine ideas behind balances (Section 1.1.1) and pivot coordinates (Section 1.1.2). They result from L balance coordinate systems in which a balance of interest is isolated in the first coordinate. These were introduced in Štefelová et al. (2021c) in the context of MB data. In this case, $L = 3$ models (11) are considered with WMB expressed in balances $\left(b_1^{(l)}, b_2^{(l)}, b_3^{(l)}\right)^\top$, $l = 1, 2, 3$, while the focus lies always in the first balance. In the first system, SB is set against the remaining (active) parts in the initial partition. In the following systems, the initial subgroup consisting of SB is subsequently accompanied by the other activities from the least to the most intense. Thus, the three first balances of interest have a symbolic notation SB_LPA.MPA.VPA, SB.LPA_MPA.VPA and SB.LPA.MPA_VPA. Table 7 illustrates an exemplary SBP to obtain the required set of balances. Note that the reciprocal balances, i.e. swapping the subsets of behaviors in the logratio, differ only by the sign. So in fact, the three first balance coordinates provide information also about the balances LPA.MPA.VPA_SB, MPA.VPA_SB.LPA and VPA_SB.LPA.MPA. As for the interpretation, e.g. SB.LPA_MPA.VPA is a contrast of time spent in the two least physically demanding activities against the two most intense activities. MM-estimation of regression coefficients is performed. The results extracted from the three models are summarized in Table 8. Apart of the already revealed significance of SB_LPA.MPA.VPA and SB.LPA.MPA_VPA (respectively

LPA.MPA.VPA_SB and VPA_SB.LPA.MPA) no additional significant result is found here. We can conclude that the problem with the adolescents' obesity is related with spending relatively too little time in VPA and on the other hand relatively too much time in SB.

Table 7: Examplary SBP for WMB composition which results in the required pivoting balance systems with the (first) balance of interest as noted in the captions. Parts chosen for the numerator and denominator of the jth balance are coded $+$ and $-$, respectively; 0 indicates that the part is not included in the respective balance.

j	x_1	x_2	x_3	x_4	r_j	s_j
1	$+$	$-$	$-$	$-$	1	3
2	0	$+$	$-$	$-$	1	2
3	0	0	$+$	$-$	1	1

(a) SB_LPA.MPA.VPA

j	x_1	x_2	x_3	x_4	r_j	s_j
1	$+$	$+$	$-$	$-$	2	2
2	$+$	$-$	0	0	1	1
3	0	0	$+$	$-$	1	1

(b) SB.LPA_MPA.VPA

j	x_1	x_2	x_3	x_4	r_j	s_j
1	$+$	$+$	$+$	$-$	3	1
2	$+$	$-$	$-$	0	2	1
3	0	$+$	$-$	0	1	1

(c) SB.LPA.MPA_VPA

Table 8: MM-regression coefficient estimates, standard errors and p-values associated with the pivoting balances extracted from the 3 models assessing the relationship between adolescents' zBMI and wake-time behaviors.

Covariate	Coeff.	Std. Error	p-value
SB_LPA.MPA.VPA	0.376	0.190	0.048
SB.LPA_MPA.VPA	0.059	0.123	0.632
SB.LPA.MPA_VPA	0.207	0.081	0.011

Next, we investigate how age is associated with the structure of adolescents' wake-time movement behaviors. To get an initial insight into the problem, we display the data in a ternary diagram, which is a standard tool for visualization of the simplex sample space for a three-part (sub)composition (Pawlowsky-Glahn et al., 2015). Here, four three-part subcompositions are available. A color gradient is used to distinguish the points by age (Figure 16). For further insight, it

is useful to plot centered data, particularly when the data are concentrated near the borders of the ternary diagram. The reason for this is the relative scale of compositional data; near the borders, the ratios between the components change substantially more than near the barycenter and this is reflected by larger distances between points in terms of the Aitchison geometry (von Eynatten et al., 2002). This means that near the borders, outlying observations might easily be overlooked due to the small relative values of some compositional parts. Moreover, in order to prevent from possible masking of outliers, robustly centered data are displayed in Figure 17. The diagrams indicate that the proportion of time spent in SB and VPA is associated with higher age, whereas the effect is the opposite for LPA and MPA.

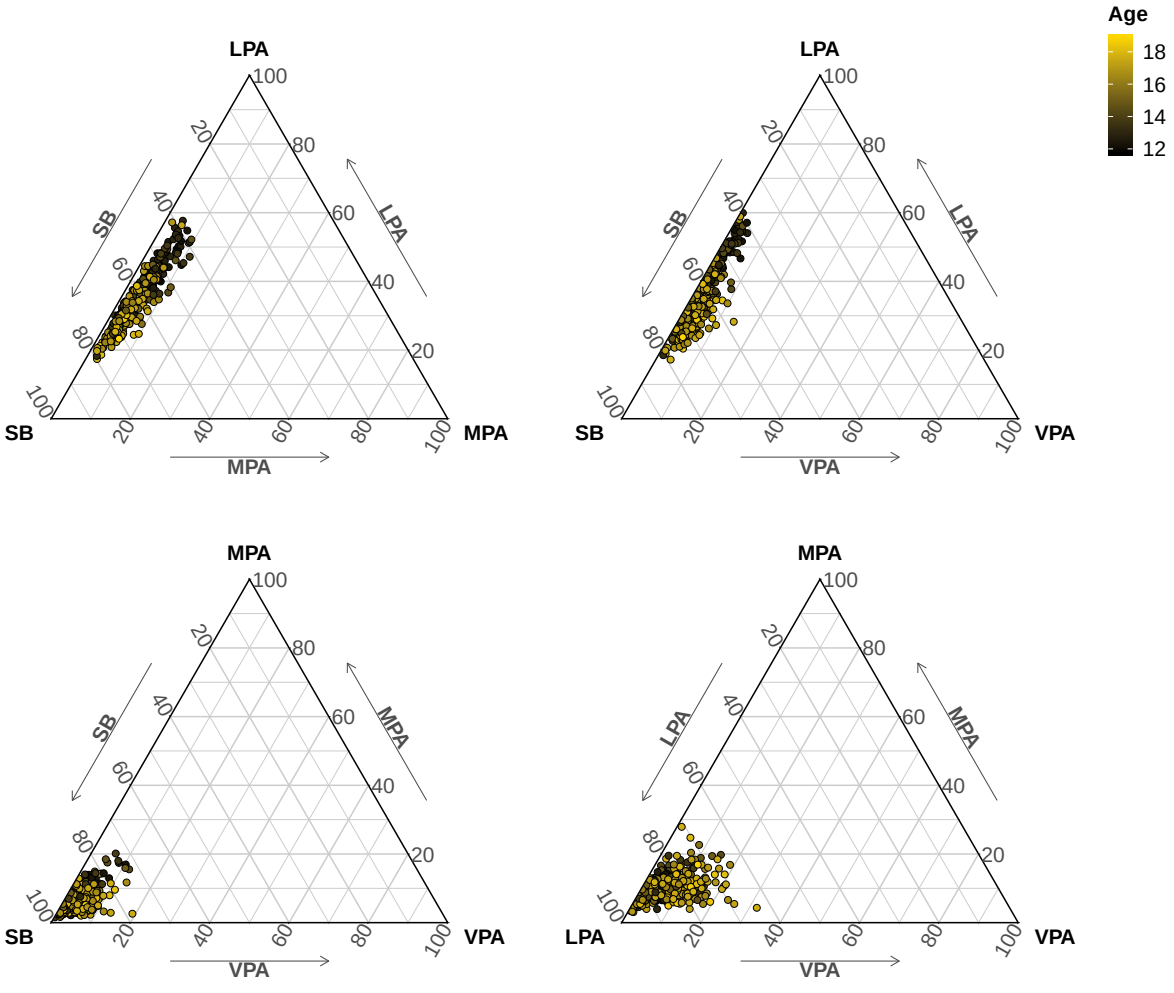


Figure 16: Ternary diagrams visualizing how adolescents' wake-time movement behaviors change with increasing age (the lighter the point, the higher the age).

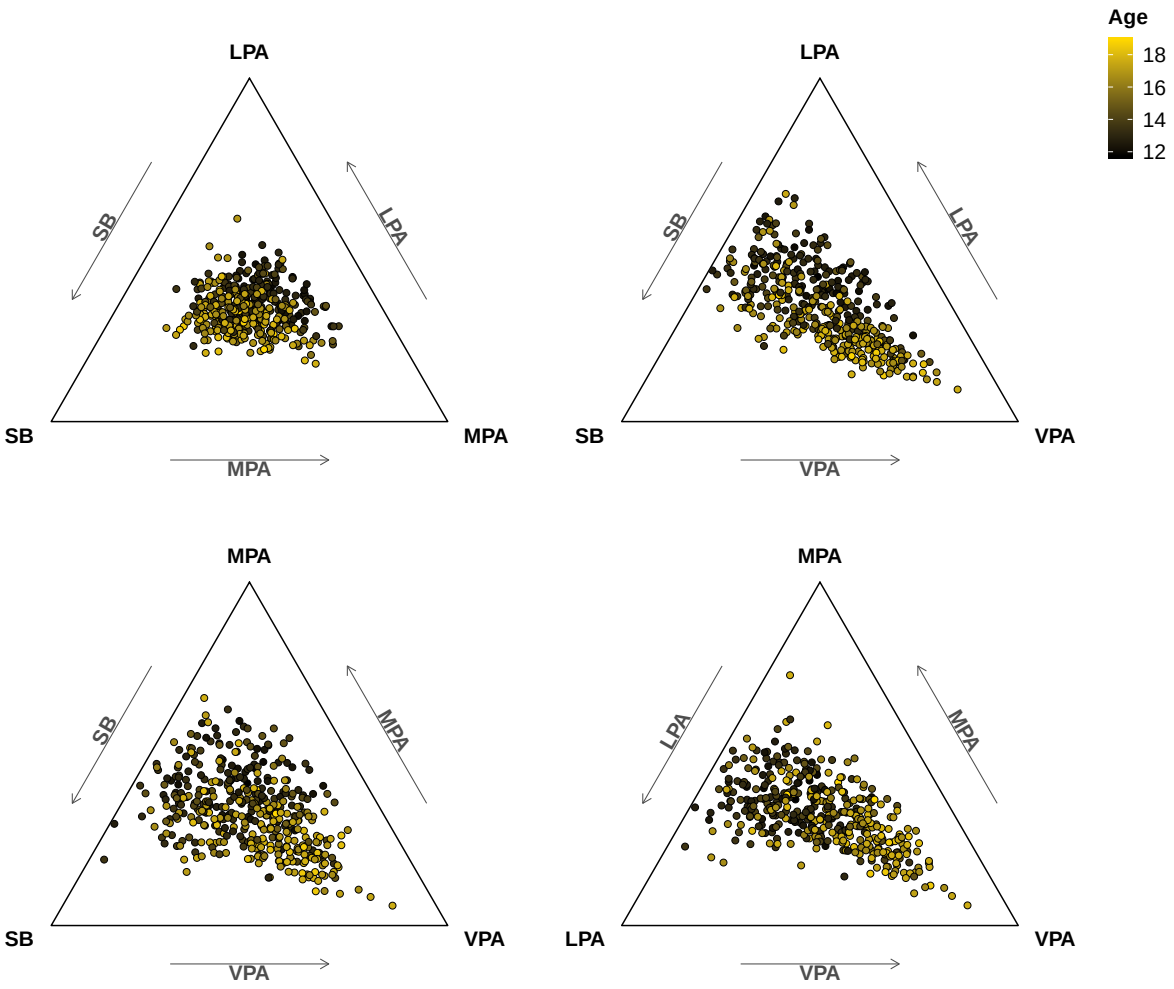


Figure 17: Ternary diagrams with robustly centered data visualizing how adolescents' wake-time movement behaviors change with increasing age (the lighter the point, the higher the age).

We further examine the relationship between WMB and age via regression analysis. We conduct four regression models, each time with different pivot coordinate (isolating different behavior) as a response variable and age (mapped into real space using log-transformation) as a covariate. Thus, we obtain information about increasing/decreasing dominance of the individual parts (with respect to the remaining parts). Alternatively, we can exploit the findings from the ternary diagrams and set balance SB.VPA_LPA.MPA as a response, dividing the behaviors into two groups that are apparently affected by age in an opposite way. MM-regression is conducted for the five models and the results are

displayed in Table 9. As anticipated, positive (negative, respectively) association is observed between age and SB_LPA.MPA.VPA, VPA_SB.LPA.MPA and SB.VPA_LPA.MPA (LPA_SB.MPA.VPA and MPA_SB.LPA.VPA, respectively). So the older they get, Czech adolescents tend to spend more time in SB and VPA at the expense of LPA and MPA. It is interesting to see that the two behaviors relatively increasing with age are the two behaviors associated with adiposity in opposite direction.

Table 9: MM-regression coefficient estimates, standard errors and p-values associated with the age from the 5 models assessing the relationship between adolescents' wake-time behaviors and age.

Response	Coeff.	Std. error	p-value
SB_LPA.MPA.VPA	0.320	0.133	0.016
LPA_SB.MPA.VPA	−0.157	0.113	< 0.001
MPA_SB.LPA.VPA	−1.072	0.131	< 0.001
VPA_SB.LPA.MPA	2.199	0.024	< 0.001
SB.VPA_LPA.MPA	2.272	0.156	< 0.001

4.2 Examining the association between 24-hour behaviors and health outcome via compositional PLS biplot based on pivoting balances

In the second application, we demonstrate the use of compositional PLS regression and biplot based on pivoting balances in the context of evaluating the association between 24-hour movement behavior patterns and health indicator (Štefelová et al., 2021c). Concretely, the methods are applied to examine the combined association of 24-hour behaviors on fat mass (FM), i.e. adiposity-related parameter, from a sample of Czech school-aged girls. The real dataset comes from the Faculty of Physical Culture in Olomouc. The participants were $N = 414$ healthy girls. The amount of time spent in parts of MB, i.e. in sleep, SB, LPA, MPA and VPA were measured by wrist-worn tri-axial ActiGraph accelerometers (ActiGraph LLC, Pensacola, FL, USA) wGT3X-BT and GT9X Link for children and adolescents, respectively. Raw data were processed with the `GGIR` package (Migueles et al., 2019) of the R software for statistical computing (R Core

Team, 2021). Time spent in the particular wake-time activities was classified using cut-points for non-dominant wrist (Hildebrand et al., 2017). The default algorithm guided by participants sleep log was used to detect sleep time (van Hees et al., 2015). FM was measured by means of a multifrequency bioelectrical impedance analysis using the InBody 720 device (InBody Co., Seoul, Korea). Additionally, height and age were recorded.

When exploring the relationship between a response variable and the time-use composition, ternary diagrams can serve well in many situations. However, similarly to the ordinary scatterplot, its practical usefulness is limited as the number of compositional parts increases, as only 3-part (sub)compositions can be displayed at once in a ternary plot. Thus, if we consider a 5-part composition MB, 10 ternary diagrams would be needed to visualize all possible combinations of 3-part subcompositions. Therefore, data dimension reduction techniques can provide more useful insight into the problem at hand. Specifically, using compositional PLS regression based on pivoting balances allows to project the data onto a 2-dimensional biplot display which represents the observations alongside the relevant time-use balances while accounting for the relationship with the outcome variable.

In our study, we consider $L = 7$ regression models (11) with FM (in log-scale) set as the response and 5-part composition MB expressed in balances $\left(b_1^{(l)}, b_2^{(l)}, b_3^{(l)}, b_4^{(l)}\right)^\top$, $l = 1, \ldots, 7$, as explanatory variables. Naturally, fat mass depends on height and age respectively (with these two being highly correlated for individuals in school age). Accordingly, age and height (both mapped into real space using log-transformation) are put in as additional covariates. For the construction of pivoting balances we take into account that the ordination of MB parts is not so straightforward as when dealing with only wake-time behaviors. Therefore, we consider seven different coordinate systems with different initial partitions into two subgroups, starting with sleep against the rest of behaviors and subsequently enlarging the subset of parts in the numerator from the least to the most intense activities; and vice versa, sleep in the numerator subsequently accompanied by the other activities from the most to the least intense. With the aim of examining the whole 24-hour behavior pattern, in each system we focus only on the first balance

involving all five parts, i.e. on the pivoting balances with symbolic notation Sleep_SB.LPA.MPA.VPA, Sleep.SB_LPA.MPA.VPA, Sleep.SB.LPA_MPA.VPA, Sleep.SB.LPA.MPA_VPA, Sleep.VPA_SB.LPA.MPA, Sleep.MPA.VPA_SB.LPA, Sleep.LPA.MPA.VPA_SB (and the corresponding reciprocals). Table 10 illustrates an exemplary SBP to obtain the required set of balances. As for the interpretation, e.g. the balance Sleep_SB.LPA.MPA.VPA compares time spent in sleep relative to waking-time behaviors, Sleep.SB_LPA.MPA.VPA is a contrast of time spent in non-active behaviors against physical activities, and so on.

PLS modelling is performed as described in Section 1.2.3. The optimal model, selected based on the randomized test approach, consists of two PLS components (CV RMSEP = 0.49 and CV R^2 = 0.32). Table 11 shows the bootstrap standardized regression coefficients estimated from $B = 1000$ bootstrap resamples for each of the first balances and their reciprocals, as well as for the two additional covariates. Statistically significant variables in positive relationship to FM (at 5% significance level) are obtained for the following balances (listed in decreasing order according to the estimated standardised regression coefficient): SB.LPA.MPA_Sleep.VPA, SB.LPA.MPA.VPA_Sleep, SB.LPA_Sleep.MPA.VPA, Sleep.SB.LPA.MPA_VPA, SB_Sleep.LPA.MPA.VPA. Hence, their reciprocals are significant in negative direction. Thus, according to the results, for obesity prevention (in school-aged girls) it would be beneficial to spend relatively less time sitting. Moreover, the results suggest that the two non-active behaviors, sleep and SB, have contrary association with fat. That is, sleep (unlike SB) would play a positive role in fat reduction. As expected, both non-compositional covariates (age and height) are positively associated with FM.

Furthermore, a PLS biplot is constructed based on the first compositional balances and their reciprocals (together with the non-compositional variables) from the seven coordinate systems (Figure 18). It provides further insight into what would be more recommendable movement behaviour patterns within the 24-hour period associated with lower adiposity. The arrows representing the covariates are coloured according to the sign of their respective associations with the FM outcome (positive in red and negative in blue for statistically significant associations, grey for insignificant association). A color gradient is used to distinguish the points according to the individuals' FM (in log-scale).

Table 10: Examplary SBP for MB composition which results in the required pivoting balance systems with the (first) balance of interest as noted in the captions. Parts chosen for the numerator and denominator of the jth balance are coded $+$ and $-$, respectively; 0 indicates that the part is not included in the respective balance.

j	x_1	x_2	x_3	x_4	x_5	r_j	s_j
1	$+$	$-$	$-$	$-$	$-$	1	4
2	0	$+$	$-$	$-$	$-$	1	3
3	0	0	$+$	$-$	$-$	1	2
4	0	0	0	$+$	$-$	1	1

(a) Sleep_SB.LPA.MPA.VPA

j	x_1	x_2	x_3	x_4	x_5	r_j	s_j
1	$+$	$+$	$-$	$-$	$-$	2	3
2	$+$	$-$	0	0	0	1	1
3	0	0	$+$	$-$	$-$	1	2
4	0	0	0	$+$	$-$	1	1

(b) Sleep.SB_LPA.MPA.VPA

j	x_1	x_2	x_3	x_4	x_5	r_j	s_j
1	$+$	$+$	$+$	$-$	$-$	3	2
2	$+$	$-$	$-$	0	0	1	2
3	0	$+$	$-$	0	0	1	1
4	0	0	0	$+$	$-$	1	1

(c) Sleep.SB.LPA_MPA.VPA

j	x_1	x_2	x_3	x_4	x_5	r_j	s_j
1	$+$	$+$	$+$	$+$	$-$	4	1
2	$+$	$-$	$-$	$-$	0	1	3
3	0	$+$	$-$	$-$	0	1	2
4	0	0	$+$	$-$	0	1	1

(d) Sleep.SB.LPA.MPA_VPA

j	x_1	x_2	x_3	x_4	x_5	r_j	s_j
1	$+$	$-$	$-$	$-$	$+$	2	3
2	$+$	0	0	0	$-$	1	1
3	0	$+$	$-$	$-$	0	1	2
4	0	0	$+$	$-$	0	1	1

(e) Sleep.VPA_SB.LPA.MPA

j	x_1	x_2	x_3	x_4	x_5	r_j	s_j
1	$+$	$-$	$-$	$+$	$+$	3	2
2	$+$	0	0	$-$	$-$	1	2
3	0	0	0	$+$	$-$	1	1
4	0	$+$	$-$	0	0	1	1

(f) Sleep.MPA.VPA_SB.LPA

j	x_1	x_2	x_3	x_4	x_5	r_j	s_j
1	$+$	$-$	$+$	$+$	$+$	4	1
2	$+$	0	$-$	$-$	$-$	1	2
3	0	0	$+$	$-$	$-$	1	2
4	0	0	0	$+$	$-$	1	1

(g) Sleep.LPA.MPA.VPA_SB

The points are fairly well distinguished according to the relationships of the significant variables with the FM outcome along both PLS component axes. The position of a point along the horizontal axis approximately reflects the contrast be-

Table 11: Estimated bootstrap standardized coefficients (and 95% confidence intervals) from PLS regression fit to school-age girls' fat mass on movement behaviour pivoting balances and their reciprocals (plus age and height).

Variable	Boot. std. coeff. (CI)	Variable	Boot. std. coeff. (CI)
SB.LPA.MPA_Sleep.VPA	6.36 (4.17, 8.16)	Sleep.VPA_SB.LPA.MPA	−6.36 (−8.16, −4.17)
SB.LPA.MPA.VPA_Sleep	5.60 (3.64, 7.61)	Sleep_SB.LPA.MPA.VPA	−5.60 (−7.61, −3.64)
SB.LPA_Sleep.MPA.VPA	3.34 (1.57, 5.39)	Sleep.MPA.VPA_SB.LPA	−3.34 (−5.39, −1.57)
Sleep.SB.LPA.MPA_VPA	3.16 (1.13, 4.95)	VPA_Sleep.SB.LPA.MPA	−3.16 (−4.95, −1.13)
SB_Sleep.LPA.MPA.VPA	3.02 (1.18, 5.04)	Sleep.LPA.MPA.VPA_SB	−3.02 (−5.04, −1.18)
MPA.VPA_Sleep.SB.LPA	1.35 (−0.83, 3.00)	Sleep.SB.LPA_MPA.VPA	−1.35 (−3.00, 0.83)
Sleep.SB_LPA.MPA.VPA	0.67 (−1.62, 2.35)	LPA.MPA.VPA_Sleep.SB	−0.67 (−2.35, 1.62)
ln(Age)	6.36 (4.46, 8.42)		
ln(Height)	5.72 (3.90, 7.70)		

tween active lifestyle and SB. The vertical axis is largely related to age, height and the lack (or the deficiency) of sleep, with the lack of sleep being associated

with higher age and height. The arrows of the variables with significantly positive regression coefficients point roughly towards the top-right quadrant, which includes mainly individuals having higher fat mass. The variables with significantly negative regression coefficients point to opposite direction (i.e. roughly bottom-left quadrant), where the individuals with lower fat mass mostly concentrate. A few outlying cases are observed in the bottom-right quadrant, particularly two cases of lower fat mass corresponding to the very young individuals with low level of physical activity. The two balances that most clearly indicate the division between lower and higher fat mass are SB.LPA.MPA_Sleep.VPA and its reciprocal. We can conclude from these results that a beneficial strategy for obesity prevention (for the school-aged girls) is to spend more time in VPA and sleep, considered in combination, with respect to the other behaviours. This result also illustrates the advantage of pivoting balances over ordinary pivot coordinates in the context of movement behavior research. Although the corresponding pivot coordinates SB.LPA.MPA.VPA_Sleep and Sleep.SB.LPA.MPA_VPA (and their reciprocals) would also indicate a significant association of the relative contributions of the (single) components Sleep and VPA to adiposity, the combined association of both relative to the other behaviours could not be assessed. This is only possible with the balance SB.LPA.MPA_Sleep.VPA, indicating that including the two movement behaviours into one group leads to even stronger evidence of their association with adiposity.

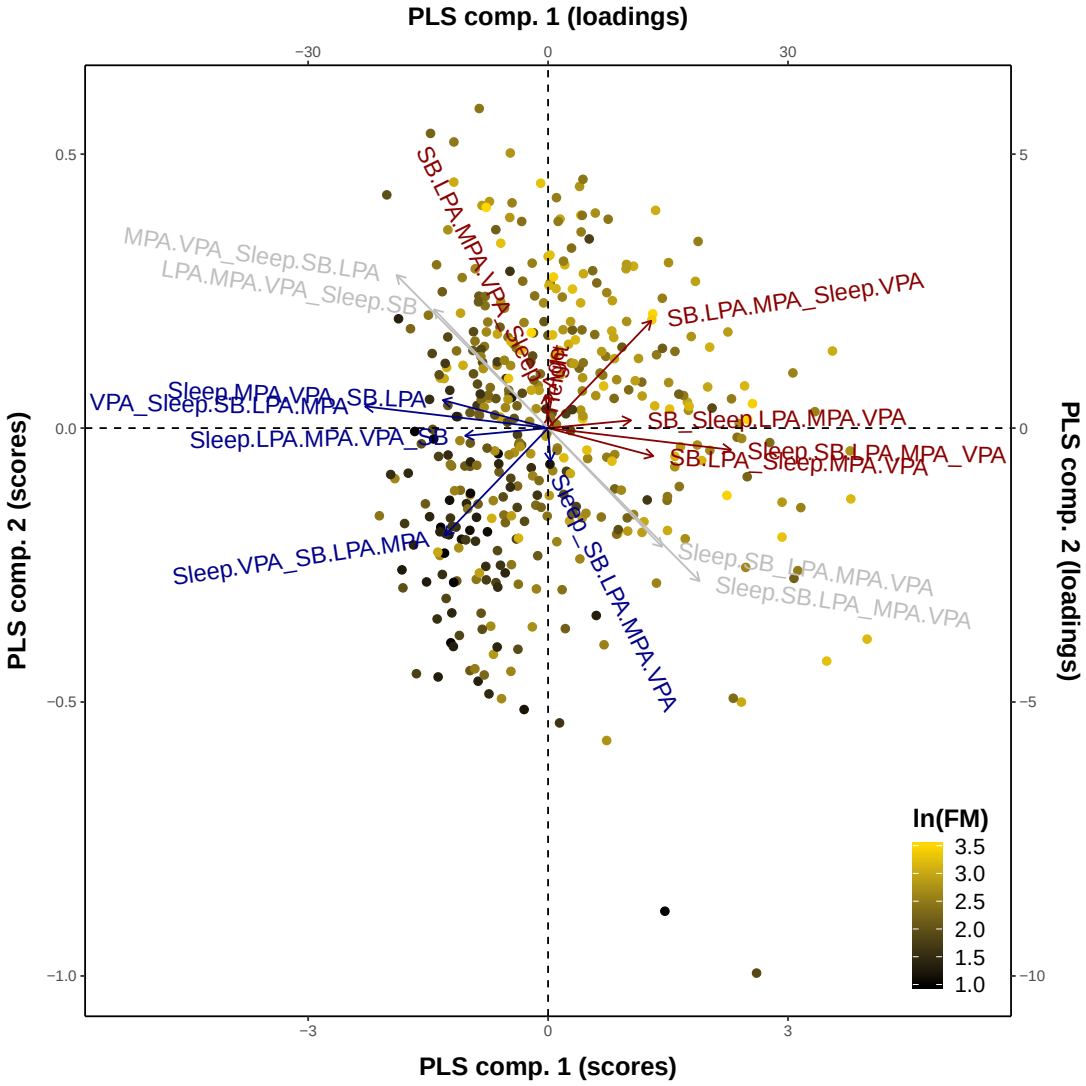


Figure 18: PLS biplot for school-aged girls data based on the first compositional balances and their reciprocals from seven coordinate systems. Significant variables in positive (resp. negative) direction are coloured in red (resp. blue), grey refers to insignificant covariates. The lighter the point is, the higher the fat mass of the respective individual is. The dashed lines indicate the origin for the first and second PLS components (PLS comp. 1 and PLS comp. 2). A 92.25% of explanatory data variance (resp. 33.31% of response data variance) is explained by the first two PLS components: 88.52% by PLS comp. 1 and 3.73% by PLS comp. 2 (resp. 20.48% by PLS comp.1 and 12.83% by PLS comp. 2).

Concluding remarks

This thesis contributes novel methods for the processing of data in biostatistics that are of compositional nature, i.e. data conveying relative information. The specific features of compositions call for an adequate approach to their statistical analysis. The logratio methodology is used for their proper statistical treatment. The application of the developments is demonstrated in time-use (physical) and livestock greenhouse gas emission research. Additionally, the work conducted during this Ph.D. project contributed new scientific insights through a number of interdisciplinary collaborations.

Section 1 revised the basics and fundamental properties of compositional data and principles for their analysis. The main emphasis was put on different logratio coordinates which allow to express compositions as real-valued vectors. It turns out that so-called isometric logratio coordinate systems, preferable from the theoretical perspective, lead also to great flexibility with respect to the choice of interpretable coordinates. Specifically, balances and (weighted) pivot coordinates can be interpreted in terms of a contrast between two subsets of compositional parts. In the case of pivot coordinates, a particular part is examined in contrast to the remaining ones. Most of the methods introduced in the thesis are related to regression modelling, what is of primary importance in a biostatistical context. Regression analysis with compositional data, particularly the case with real response and compositional explanatory variables, was thoroughly discussed.

A new method for compositional regression that is robust against cellwise and rowwise outliers was introduced in Section 2. Cellwise outliers are first filtered and then imputed by robust estimates. Afterwards, rowwise robust compositional regression using a multiple imputation scheme is performed to obtain model coefficient estimates. An application to bio-environmental data relating biological processes in livestock rumen with methane emissions revealed that the proposed procedure (compared to other regression methods) leads to conclusions that are best aligned with established scientific knowledge. An extensive simulation study shows that the procedure generally outperforms a traditional rowwise-only robust regression method (MM-estimator). Moreover, our procedure yields better or comparable results to recently proposed cellwise robust regression methods (shooting S-estimator, 3-step regression) while it is preferable for interpretation

through the use of appropriate coordinate systems for compositional data.

A new weighting strategy for the construction of weighted pivot coordinates was proposed in Section 3. Designed to improve PLS-based marker discovery in high-dimensional compositional data, it draws on the correlation between response variable and pairwise logratios aggregated into the first coordinate. The illustrative application to investigate the association between ruminal high-throughput metabolite signals and methane emission in cattle extended the study in Section 2 to the high-dimensional case. It demonstrated the practical relevance and potential of the proposed approach, providing results compatible with previous knowledge along with a higher sensitivity to identify meaningful markers. A simulation study provided additional evidence that this proposed logratio coordinate representation enhances the discovery of markers, although it results in slightly worse specificity.

In Section 4, the compositional approach within time-use epidemiology was studied. Proper coordinate representation for movement behavior data was discussed given the ordinal character of daily activities. In the first application, wake-time movement behavior data were examined via robust linear regression and visualization tools such as compositional mean barplots and ternary diagrams. The second application demonstrated how an adapted version of compositional PLS regression and biplot based on the newly introduced concept of pivoting balances could be employed to evaluate the association between 24-hours behavior patterns and a health marker.

All computation in this work were performed within the R environment for statistical computing (R Core Team, 2021). The related codes are available at `https://github.com/aalfons/lmcrCoda` (Section 2), `https://github.com/StefelovaN/Weighted-pivot-coordinates` (Section 3), `https://github.com/StefelovaN/Robust-CoDA-WMB` (Section 4.1) and `https://github.com/StefelovaN/Balance-based-PLS-biplot` (Section 4.2).

I truly hope that the presented thesis helps in further development of the logratio methodology, not solely in the biostatistical context, but others where similar statistical modelling challenges are presented.

Bibliography

Aitchison J (1986) The Statistical Analysis of Compositional Data. Chapman & Hall, London, DOI 10.1007/978-94-009-4109-0

Alqallaf F, Van Aelst S, Yohai V, Zamar R (2009) Propagation of outliers in multivariate data. The Annals of Statistics 37(1):311–331, DOI 10.1214/07-AOS588

Barnard J, Rubin D (1999) Small-sample degrees of freedom with multiple imputation. Biometrika 86(4):948–955, DOI 10.1093/biomet/asm028

Bodner T (2009) What improves with increased missing data imputations? Structural Equation Modeling: A Multidisciplinary Journal 15(4):651–675, DOI 10.1080/10705510802339072

Burchartz A, Anedda B, Auerswald T, Mall C, Giurgiu M, Hill H, Ketelhut S, Kolb S, Manz K, Nigg C, Reichert M, Sprengeler O, Wunsch K, Matthews C (2020) Assessing physical behavior through accelerometry âFIXME“ state of the science, best practices and future directions. Psychology of Sport & Exercise 49:101703, DOI 10.1016/j.psychsport.2020.101703

Cevallos Valdiviezo H, Van Aelst S (2015) Tree-based prediction on incomplete data using imputation or surrogate decisions. Information Sciences 311:163–181, DOI 10.1016/j.ins.2015.03.018

Chastin S, Palarea-Albaladejo J, Dontje M, Skelton D (2015) Combined effects of time spent in physical activity, sedentary behaviors and sleep on obesity and cardio-metabolic health markers: A novel compositional data analysis approach. PLoS ONE 10:1–37, DOI 10.1371/journal.pone.0139984

de Onis M, Onyango A, Borghi E, Siyam A, Nishida C, Siekmann J (2007) Development of a who growth reference for school-aged children and adolescents. Bulletin of the World Health Organization 85(9):660–667, DOI 10.1590/S0042-96862007000900010

Dumuid D, Pedišić v, Stanford T, Martín-Fernández J, Hron K, Maher C, Lewis L, Olds T (2017a) The compositional isotemporal substitution model: A method for estimating changes in a health outcome for reallocation of time be-

tween sleep, physical activity and sedentary behaviour. Statistical Methods in Medical Research 28:846–857, DOI 10.1177/0962280217737805

Dumuid D, Stanford T, Olds T, Lewis L, Martín-Fernández J, Pedišić v, Hron K, Katzmarzyk P, Barreira T, Broyles S, Chaput J, Fogelholm M, Hu G, Lambert E, Maia J, Sarmiento O, Standage M, Tremblay M, Tudor-Locke C, Maher C (2017b) Compositional data analysis for physical activity, sedentary time and sleep research. Statistical Methods in Medical Research 27(12):3726–3738, DOI 10.1186/s12889-018-5207-1

Egozcue J, Pawlosky-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 J, 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

Evenson K, Catellier D, Ondrak K, McMurray R (2008) Calibration of two objective measures of physical activity for children. Journal of Sports Sciences 26(14):1557–1565, DOI 10.1080/02640410802334196

Everitt B, Hothorn T (2011) An Introduction to Applied Multivariate Analyses with R. Springer, New York, DOI 10.1007/978-1-4419-9650-3

Filzmoser P, Hron K, Templ M (2018) Applied Compositional Data Analysis. 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(4):455–468, DOI 10.1007/s11004-011-9333-x

Härdle W, Simar L (2012) Applied Multivariate Statistical Analysis. Springer, Heidelberg, DOI 10.1007/978-3-662-45171-7

Helland I (2010) Steps Towards a Unified Basis for Scientific Models and Methods. World Scientific, Singapore, DOI 10.1142/7404

Hildebrand M, Hansen B, van Hees V, Ekelund U (2017) Evaluation of raw acceleration sedentary thresholds in children and adults. Scandinavian Journal of Medicine & Science in Sports 27(12):1814–1823, DOI 10.1111/sms.12795

Höskuldson A (1988) PLS regression methods. Journal of Chemometrics 2:211–228, DOI 10.1002/cem.1180020306

Hron K, Templ M, Filzmoser P (2010) Imputation of missing values for compositional data using classical and robust methods. Computational Statistics & Data Analysis 54(12):3095–3107, DOI 10.1016/j.csda.2009.11.023

Hron K, Filzmoser P, Thompson K (2012) Linear regression with compositional explanatory variables. Journal of Applied Statistics 39(5):1115–1128, DOI 10.1080/02664763.2011.644268

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

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

Khan J, Van Aelst S, Zamar R (2007) Robust linear model selection based on least angle regression. Journal of the American Statistical Association 102(480):1289–1299, DOI 10.1198/016214507000000950

Kynčlová P, Filzmoser P, Hron K (2016) Compositional biplots including external non-compositional variables. Statistics 50:1–17, DOI 10.1080/02331888.2015.1135155

Leung A, Zhang H, Zamar R (2016) Robust regression estimation and inference in the presence of cellwise and casewise contamination. Computational Statistics & Data Analysis 99:1–11, DOI 10.1016/j.csda.2016.01.004

Little R, Rubin D (2002) Statistical Analysis with Missing Data, 2nd edn. John Wiley & Sons, Chichester, DOI 10.2307/2531606

Maronna R, Martin R, Yohai V (2002) Robust Statistics: Theory and Methods. John Wiley & Sons, Chichester, DOI 10.1002/0470010940

Migueles J, Rowlands A, Huber F, S S, van Hees V (2019) GGIR: A research community-driven open source R package for generating physical activity and sleep outcomes from multi-day raw accelerometer data. Journal for the Measurement of Physical Behaviour 2(3):188–196, DOI 10.1123/jmpb.2018-0063

Müller I, Hron K, Fišerová E, Šmahaj J, Cakirpaloglu P, Vančáková J (2018) Interpretation of compositional regression with application to time budget analysis. Austrian Journal of Statistics 47(2):3–19, DOI 10.17713/ajs.v47i2.652

Öllerer V, Alfons A, Croux C (2016) The shooting S-estimator for robust regression. Computational Statistics 31(3):829–844, DOI 10.1007/s00180-015-0593-7

Oyedele O, Gardner-Lubbe S (2015) The construction of a partial least-squares biplot. Journal of Applied Statistics 42:1–12, DOI 10.1080/02664763.2015.1043858

Palarea-Albaladejo J, Martín-Fernández J (2008) A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. Computers & Geosciences 34(8):902–917, DOI 10.1016/j.cageo.2007.09.015

Palarea-Albaladejo J, Rooke JA, Nevison IM, Dewhurst RJ (2017) Compositional mixed modeling of methane emissions and ruminal volatile fatty acids from individual cattle and multiple experiments. Journal of Animal Science 95(6):2467–2480, DOI 10.2527/jas2016.1339

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

R Core Team (2021) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.r-project.org

Rousseeuw P, Van den Bossche W (2018) Detecting deviating data cells. Technometrics 60(2):135–145, DOI 10.1080/00401706.2017.1340909

Rubin D (1987) Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, Chichester, DOI 10.1002/9780470316696

Rubin D, Schenker M (1986) Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. Journal of the American Statistical Association 81(394):366–374, DOI 10.1080/01621459.1986.10478280

Templ M, Kowarik A, Filzmoser P (2011) Iterative stepwise regression imputation using standard and robust methods. Computational Statistics & Data Analysis 55(10):2793–2806, DOI 10.1016/j.csda.2011.04.012

Van Buuren S (2012) Flexible Imputation of Missing Data. Chapman & Hall/CRC, Boca Raton, DOI 10.1201/9780429492259

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

van Hees V, S S, Anderson K, Denton S, Oliver J, Catt M, Abell J, Kivimaki M, Trenell M, Singh-Manoux A (2015) A novel, open access method to assess sleep duration using a wrist-worn accelerometer. PLoS One 10(11), DOI 10.1371/journal.pone.0142533

Varmuza K, Filzmoser P (2009) Introduction to Multivariate Statistical Analysis in Chemometrics. Taylor & Francis, New York, DOI 10.1201/9781420059496

von Eynatten H, Pawlowsky-Glahn V, Egozcue J (2002) Understanding perturbation on the simplex: A simple method to better visualize and interpret compositional data in ternary diagrams. Mathematical Geology 34(2):249–257, DOI 10.1023/A:1014826205533

Štefelová N, Dygrýn J, Hron K, Gába A, Rubín L, Palarea-Albaladejo J (2018) Robust compositional analysis of physical activity and sedentary behaviour data. International Journal of Environmental Research and Public Health 15(10):2248, DOI 10.3390/ijerph15102248

Štefelová N, Alfons A, Palarea-Albaladejo J, Filzmoser P, Hron K (2021a) Robust regression with compositional covariates including cellwise outliers. Advances in Data Analysis and Classification DOI 10.1007/s11634-021-00436-9

Štefelová N, Palarea-Albaladejo J, Hron K (2021b) Weighted pivot coordinates for PLS-based marker discovery in high-throughput compositional data. *Under review*

Štefelová N, Palarea-Albaladejo J, Hron K, Gába A, Dygrýn J (2021c) Compositional PLS biplot based on pivoting balances: a graphical tool to examine the association between 24-hour movement behaviours and health outcomes. *Under review*

White I, Royston P, Wood A (2011) Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine 30(4):377–399, DOI 10.1186/2196-0739-1-4

Wolin M (1960) A theoretical rumen fermentation balance. Journal of Dairy Science 43:1452–1459, DOI 10.3168/jds.S0022-0302(60)90348-9

Yohai V (1987) High breakdown point and high efficiency robust estimates for regression. The Annals of Statistics 15(2):642–656, DOI 10.1214/aos/1176350366

PALACKÝ UNIVERSITY IN OLOMOUC
FACULTY OF SCIENCE

DISSERTATION THESIS SUMMARY

Compositional approach to the analysis of data in biostatistics

Supervisor: **prof. RNDr. Karel Hron, Ph.D.**
Co-supervisor: **Dr. Javier Palarea-Albaladejo**
Author: **Mgr. Nikola Štefelová**
Study program: P1104 Applied Mathematics
Field of study: Applied Mathematics
Form of study: Full-time
The year of submission: 2021

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. Nikola Štefelová**
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

Co-supervisor: **Dr. Javier Palarea-Albaladejo**
Biomathematics and Statistics Scotland
United Kingdom

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

doc. PaedDr. RNDr. Stanislav Katina, Ph.D.
Institute of Mathematics and Statistics
Faculty of Science
Masaryk University Brno

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

Many types of data in biostatistics meet properties of compositional data, i.e. multivariate observations comprising positive parts of a whole carrying relative information. Given their specific properties, the logratio methodology serves as a proper tool for the analysis of compositions. This thesis presents methodological developments and applications of the compositional approach in fields of biostatistics, specifically in relation to regression analysis and data visualization as applied to metabolomics and time-use epidemiology. A novel method for regression with compositional explanatory variables is introduced, which is robust against rowwise as well as against cellwise outliers. Further, a procedure for improved biomarker discovery in high-dimensional compositional data is presented. It is based on partial least squares (PLS) regression using a weighted pivot coordinate representation for compositions with a new, task-driven, strategy for weighting. In the context of time-use research, special relevance is given to a suitable coordinate representation of time-use data. The proposed coordinate system aims to reflect the fact that there is a natural ordering in time-use categories, the compositional variables.

Key words: compositional data, logratio methodology, balances, pivot coordinates, weighted pivot coordinates, regression analysis, PLS biplot, robust statistics, cellwise outliers, time-use data, metabolomic data

Abstrakt v českém jazyce

Mnoho typů dat v biostatistice má podobu kompozičních dat, tj. jde o mnohorozměrná pozorování obsahující kladné složky, které reprezentují části nějakého celku a nesou relativní informaci. Logpodílová metodika, zohledňující specifické vlastnosti kompozic, slouží jako vhodný prostředek k jejich analýze. Tato práce představuje metodologické inovace a aplikace kompozičního přístupu v oborech biostatistiky, konkrétně v oblasti regresní analýzy a vizualizace dat, a to v metabolomice a při zkoumání vlivu pohybového chování na zdraví. Je zde prezentována nová robustní metoda pro regresi s kompozičními vysvětlujícími proměnnými, jež je schopna efektivně pracovat s pozorováními, která jsou odlehlá jako celek, i s těmi, kde se odlehlost projevuje pouze na prvkové úrovni. Také je tu představena vylepšená procedura pro identifikaci statisticky významných proměnných ve vysoce-dimenzionálních kompozičních datech. Je založena na regresi metodou částečných nejmenších čtverců (PLS regresi), při níž se pro reprezentaci kompozic využívají vážené pivotové souřadnice s novou strategií pro vážení danou povahou problému. V kontextu výzkumu pohybového chování je kladen zvláštní důraz na vhodnou souřadnicovou reprezentaci dat o pohybovém chování. Navržený souřadnicový systém bere v potaz to, že mezi kategoriemi pohybového chování, kompozičními proměnnými, existuje přirozené uspořádání.

Klíčová slova: kompoziční data, logpodílová metodika, bilance, pivotové souřadnice, vážené pivotové souřadnice, regresní analýza, PLS biplot, robustní statistika, odlehlá pozorování na úrovni buněk, data o pohybovém chování, metabolomická data

1 Introduction

Compositional data (compositions) are essentially characterized by their relative nature. They represent vectors of stricly positive values describing parts of some whole. Accordingly, the relevant information is contained in the ratios between the compositional parts. Due to specific sample space of compositional data and their geometry, compositions require different statistical processing than standard multivariate observations conveying absolute information (in terms of interval scale). A suitable approach for their analysis is the logratio methodology (Aitchison, 1986; Pawlowsky-Glahn et al., 2015). Its cornerstone lies in the construction of logratio coordinates that enable to express compositions as real-valued vectors, to which standard statistical methods can be applied. The choice of interpretable coordinates leading to meaningful results is of particular importance.

This thesis focus on the compositional approach and innovative methods within the logratio methodology suited to the analysis of biostatistical data, i.e. data involving living systems. There are many types of biostatistical data of compositional character. Here, two cases are considered: 1) molecular biology data concerning metabolites, i.e. small molecules involved in metabolism and 2) time-use movement behaviour data which reflect how people spend their time in terms of sleep, sedentary behaviour and physical activity of various intensities.

First, a novel method for robust compositional regression is introduced that is able to deal not only with outlying observations comprising whole observations (rowwise outliers) but also with outliers in individual cells (cellwise outliers) (Štefelová et al., 2021a). Next, a new weighting strategy for the construction of weighted pivot coordinates is presented that is particularly suitable for PLS-based marker discovery in high-dimensional compositional biomolecular data (Štefelová et al., 2021b). Finally, the use of the compositional approach in the context of time-use epidemiology is demonstrated (Štefelová et al., 2018, 2021c). Strong emphasis is placed on a proper coordinate representation of time-use data considering a natural ordering of the given compositional parts. A new concept of pivoting balances is developed that, in combination with an adapted formulation of compositional PLS biplot, enables meaningful visualization of more complex time-use patterns and their relationships with an outcome variable.

2 Summary of the state of the art

2.1 Compositional data

A vector $\boldsymbol{x} = (x_1, \ldots, x_D)^\top$ is called a D-part composition when all its elements are strictly positive real numbers that carry relative information (Aitchison, 1986; Pawlowsky-Glahn et al., 2015). Accordingly, the absolute values of the parts are not important for the analysis and the relevant information is captured in the ratios between them. The compositional parts, representing quantitatively contributions to some whole, are co-dependent as within a given representation the change in one part necessarily affects the relative values of the remaining ones.

Compositional data are *scale invariant* which means that if the composition is multiplied by a positive number, the ratios between its parts are not altered. Consequently, the sample space of compositions is formed by equivalence classes of proportional vectors (Pawlowsky-Glahn et al., 2015). Therefore, compositions can be represented without loss of information as vectors with an arbitrary sum of components (typically 1 or 100 in case of proportions or percentages, respectively) on a simplex.

Compositions obey the so-called *Aitchison geometry* on the simplex (Pawlowsky-Glahn et al., 2015). When analysing compositional data, their specific nature should be taken into account. The direct use of standard statistical methods relying on the Euclidean geometry in real space would lead to misleading results and conclusions (Filzmoser et al., 2018).

The key idea of the logratio methodology is to map compositions from the simplex into real space via logratio coordinates and then proceed with the statistical processing there (Filzmoser et al., 2018). Using logratios, instead of simply ratios as bearers of the elemental information, is advantageous as they map the range of a ratio from the positive real space onto the entire real space, symmetrise their values around zero and, moreover, inverse logratios provide the same information up to the sign, i.e. $\ln(x_c/x_d) = -\ln(x_d/x_c)$.

Among the different types of logratio coordinates proposed, the ilr (isometric logratio) coordinates are preferred as they allow to express compositions

in an orthonormal coordinate system (Egozcue et al., 2003). There are infinitely many options for their construction. The fact that different ilr coordinate systems are just orthogonal rotations of each other is a useful property in statistical analysis. For example, in regression analysis, it enables the use of an arbitrary choice of ilr coordinates to obtain the required (unique) output. Moreover, affine equivariant robust (regression) estimators provide results invariant to the choice of ilr coordinates (Filzmoser et al., 2018). The crucial challenge is to construct interpretable coordinates tailored to the scientific question at hand.

2.2 Balances

The procedure known as sequential binary partition (SBP) can be applied to construct customized ilr coordinates called (compositional) balances (Egozcue and Pawlosky-Glahn, 2005). In the first step of the SBP process, the entire collection of compositional parts is divided into two disjoint subsets, with each subset summarised by the geometric mean of its components and going into the numerator and denominator, respectively, of a normalized logratio constituting the first balance. In the next steps, these subsets are further split into two mutually exclusive subgroups going into the numerator and denominator, respectively, of the subsequent balances. This process continues until only one-part subsets remain and $D-1$ balances are constructed.

The balance coordinates are represented by a real vector $\boldsymbol{b} = (b_1, \ldots, b_{D-1})^\top$ with

$$b_j = \sqrt{\frac{r_j s_j}{r_j + s_j}} \ln \frac{\sqrt[r_j]{\prod_{i=1}^{r_j} x_{j_i}^+}}{\sqrt[s_j]{\prod_{i=1}^{s_j} x_{j_i}^-}}, \quad j = 1, \ldots, D-1, \tag{1}$$

where $x_{j_i}^+$ and $x_{j_i}^-$ refers to the parts selected for the numerator and denominator, respectively, in the jth balance while r_j and s_j stands for the respective number of parts (Egozcue and Pawlosky-Glahn, 2005; Pawlowsky-Glahn et al., 2015).

Balance coordinates are interpreted, as their name indicates, in terms of a balance (contrast) between two subsets of parts represented by their respective geometric means (Egozcue and Pawlosky-Glahn, 2005; Pawlowsky-Glahn et al., 2015). They can be constructed according to the scientific questions of interest

and based on domain-specific knowledge, e.g. to represent meaningful trade-offs.

2.3 Pivot coordinates

The procedure of extracting unique information from different orthonormal coordinate system is particularly applied with special balances called pivot coordinates (Fišerová and Hron, 2011). These are intended to highlight the role of a single compositional part relative to all the others in one (the first) coordinate. In SBP, one part is always set against the remaining ones.

Given a composition $\boldsymbol{x}$, we can rearrange it so that the lth part is put at the first place and denote that composition as

$$\boldsymbol{x}^{(l)} = \left(x_1^{(l)}, \ldots, x_D^{(l)}\right)^\top = (x_l, x_2, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)^\top, \quad l = 1, \ldots, D.$$

Then, the corresponding pivot coordinates define a real vector $\boldsymbol{z}^{(l)} = \left(z_1^{(l)}, \ldots, z_{D-1}^{(l)}\right)^\top$, where

$$\begin{aligned} z_j^{(l)} &= \sqrt{\frac{D-j}{D-j+1}} \ln \frac{x_j^{(l)}}{\sqrt[D-j]{\prod_{d=j+1}^{D} x_d^{(l)}}} \\ &= \frac{1}{\sqrt{(D-j+1)(D-j)}} \left[\ln\left(\frac{x_j^{(l)}}{x_{j+1}^{(l)}}\right) + \cdots + \ln\left(\frac{x_j^{(l)}}{x_D^{(l)}}\right)\right] \qquad (2) \\ &= \boldsymbol{u}_j^\top \ln\left(\boldsymbol{x}^{(l)}\right), \quad j = 1, \ldots, D-1, \; l = 1, \ldots, D, \end{aligned}$$

with

$$\boldsymbol{u}_j = \sqrt{\frac{D-j}{D-j+1}} \left(\underbrace{0, \ldots, 0}_{j-1}, 1, -\frac{1}{D-j}, \ldots, -\frac{1}{D-j}\right)^\top$$

representing the vectors of logcontrast coefficients (Filzmoser et al., 2018; Hron et al., 2017) .

Each first coordinate $z_1^{(l)}$ in the pivot coordinate system contains all the relative information about the lth compositional part. It can be interpreted in terms of dominance of the l-th part with respect to an average (geometric mean) of the other parts (Fišerová and Hron, 2011; Filzmoser et al., 2018).

2.4 Weighted pivot coordinates

In the representation of the first pivot coordinate as a scaled sum of the $D-1$ pairwise logratios of $x_1^{(l)}$ over the other parts, the logratios are treated equally. However, the collection of logratios aggregated into that coordinate can include information from completely different processes. Therefore, a weighted counterpart to the ordinary pivot coordinates was introduced, namely the weighted pivot coordinates (Hron et al., 2017). These enable to weight the logratios aggregated into the first coordinate according to their relevance for the purpose of the analysis.

Accordingly, by using $\gamma_2^{(l)}, \ldots, \gamma_D^{(l)}$ to denote the weights, the first weighted pivot coordinate $w_1^{(l)}$ is constructed by taking the weighted sum of pairwise logratios with $x_1^{(l)}$,

$$\gamma_2^{(l)} \ln\left(\frac{x_1^{(l)}}{x_2^{(l)}}\right) + \ldots + \gamma_D^{(l)} \ln\left(\frac{x_1^{(l)}}{x_D^{(l)}}\right), \ \gamma_2^{(l)}, \ldots, \gamma_D^{(l)} > 0, \ \sum_{d=2}^{D} \gamma_d^{(l)} = 1,$$

which, after rescaling to a standard logcontrast, leads to the coordinate

$$w_1^{(l)} = \frac{1}{\sqrt{1 + \sum_{d=2}^{D} \left(\gamma_d^{(l)}\right)^2}} \ln \frac{x_1^{(l)}}{\prod_{d=2}^{D} \left(x_d^{(l)}\right)^{\gamma_d^{(l)}}} = \left(\boldsymbol{v}_1^{(l)}\right)^{\top} \ln\left(\boldsymbol{x}^{(l)}\right), \quad l = 1, \ldots, D \tag{3}$$

with

$$\boldsymbol{v}_1^{(l)} = \frac{1}{\sqrt{1 + \sum_{d=2}^{D} \left(\gamma_d^{(l)}\right)^2}} \left(1, -\gamma_2^{(l)}, \ldots, -\gamma_D^{(l)}\right)^{\top}$$

representing the first vector of logcontrast coefficients (Hron et al., 2017).

The remaining elements to form a real vector of weighted pivot coordinates $\boldsymbol{w}^{(l)} = \left(w_1^{(l)}, \ldots, w_{D-1}^{(l)}\right)^{\top}$ are obtained sequentially by considering the orthonormal property of the logcontrast coefficients and the requirement for standard logcontrasts. Note that unlike in the case of ordinary pivot coordinates, weighted pivot coordinates, using the construction from Hron et al. (2017), contain

two coordinates which capture information about the part of interest: $w_1^{(l)}$ and $w_{D-1}^{(l)}$. However, the former coordinate contains the relevant information, whereas the latter corresponds to just a redundant remainder (Hron et al., 2017).

2.5 Compositional linear regression

Regression analysis is one of the most widely used techniques in practical data analysis and statistical modelling. The object of linear regression is to model linear relationship between response (dependent) variable and explanatory (independent) variables, also called covariates or predictors (Härdle and Simar, 2012). The compositional data framework has three basic regression problems. These concern the relation between the real-valued response and compositional covariates, compositional response and real covariates, or between compositional parts themselves. In all instances, the logratio methodology serves as useful tool as, with compositions expressed in proper logratio coordinates, standard regression methods can be applied and interpretable results obtained (Filzmoser et al., 2018). Because of their properties, the ilr coordinates are preferable, especially balances or the (weighted) pivot coordinates.

Throughout this thesis, we mainly deal with the cases where explanatory variables are formed by, or at least include, a composition. In that case, we consider two data structures: column vector $\mathbf{y}$ of size N and $(N, D+P)$-matrix $\mathbf{A} = (\mathbf{1}, \mathbf{i}_1, \ldots, \mathbf{i}_{D-1}, \mathbf{c}_1, \ldots, \mathbf{c}_P)$. The vector $\mathbf{y}$ describes values of the response variable on N objects. The first column of the so-called design matrix $\mathbf{A}$ is formed by ones (for the intercept term parameter) and the remaining columns combine values on the $D-1$ ilr coordinates and the P non-compositional covariates corresponding to the same N observations. The resulting linear regression model has the form

$$\mathbf{y} = \mathbf{A}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{4}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots \beta_{D-1+P})^\top$ is a vector of unknown $K = D + P$ regression coefficients and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N)^\top$ is an error vector (Härdle and Simar, 2012; Filzmoser et al., 2018).

Often, the focus lies on L different first coordinates conveying information about compositional parts in a desirable way. Then, L different regression models

are examined and information associated with the first coordinate from each system is extracted. That is, we have L models

$$\mathbf{y} = \mathbf{A}^{(l)}\boldsymbol{\beta}^{(l)} + \boldsymbol{\varepsilon}, \quad l = 1, \dots L, \tag{5}$$

where the design matrix $\mathbf{A}^{(l)} = \left(\mathbf{1}, \mathbf{i}_1^{(l)}, \dots, \mathbf{i}_{D-1}^{(l)}, \mathbf{c}_1, \dots, \mathbf{c}_P\right)$ contains values on the lth set of ilr coordinates and the regression coefficient vector $\boldsymbol{\beta}^{(l)} = \left(\beta_0, \beta_1^{(l)}, \dots, \beta_{D-1}^{(l)}, \beta_D, \dots \beta_{D-1+P}\right)^\top$ has, due to the orthogonality of different ilr coordinate systems, the same intercept term β_0 and the same coefficients corresponding to the non-compositional covariates in each model. Consequently, the vector of estimates $\left(\hat{\beta}_0, \hat{\beta}_1^{(1)}, \dots, \hat{\beta}_1^{(D)}, \hat{\beta}_D, \dots, \hat{\beta}_{D-1+P}\right)^\top$ is used for interpretation purposes. Note that also model fit measures like the coefficient of determination are invariant to the choice of orthonormal coordinate system. Commonly, the L different ilr systems represent D sets of pivot coordinates so that each time the emphasis is put on the coordinate isolating the relative information about one compositional part (Hron et al., 2012; Filzmoser et al., 2018).

The basic method for estimating coefficients in a linear regression model is the ordinary least squares (OLS) technique. It produces estimates that minimize the sum of squared residuals, i.e. the sum of squared differences between observed and predicted values of the response variable (Härdle and Simar, 2012). In practice, a common issue is that the observed dataset contains outliers, i.e. individual values or entire multivariate observations that deviate considerably from the main cloud of data points. Unfortunately, outliers can greatly influence ordinary estimates of model parameters and may lead to unreliable results. A number of regression methods robust against outlying observations (i.e. *rowwwise outliers*) have been developed (Maronna et al., 2002). Among those, MM-regression (Yohai, 1987) is a popular choice as it produces highly efficient estimates (i.e. with small variance and thus high precision) with a high breakdown-point, concretely up to 0.5 (meaning that reliable results can be obtained even with 50% observations being contaminated). A few methods have been introduced recently that are robust against *cellwise outliers* (i.e. designed to deal with contamination occurring at the cell level of a data matrix) such as shooting S-estimator (Öllerer et al., 2016) and 3-step regression estimator (Leung et al., 2016). However, both

methods have some limitations when it comes to working with compositional data. Neither of them is suitable for regression with ilr coordinate representation of compositions. The reason is that one outlying compositional part can affect several logratio coordinates so cellwise contamination easily propagates throughout.

Partial least squares (PLS) regression enjoys wide popularity in areas such as chemometrics (Höskuldson, 1988), especially in the case where the number of explanatory variables is significantly larger than the number of observations. It aims to fit the relationship between response variable(s) and potentially many and/or highly correlated explanatory variables by finding a small number of latent factors that synthesize the relationship in lower dimension. The underlying assumption is that the observed data are generated by a process driven by this small number of latent factors, also known as PLS components. The values on the PLS components (scores) are linear combinations of the explanatory variables with parameters (loadings) determined in such a way that they maximize the covariance between the response and the explanatory variables. Once the model is fitted in the latent space, the regression coefficients associated with the original explanatory variables can be subsequently worked out and their significance investigated. Even if PLS regression is particularly useful for the analysis of high-dimensional data, it offers other features that make the method also appealing for datasets with a relatively small to moderate number of explanatory variables. This includes the capacity to handle multicollinearity and highly correlated explanatory variables, the ability to separate main information from noise, the no requirement of distributional assumptions for error terms and, last but not least, the possibility of visualizing data in low dimensions via a PLS biplot. PLS regression is a well-established method to identify which (in a large set of) explanatory variables are significant (markers) in relation to a response variable of interest, including cases where covariates are of compositional nature. Using the pivot coordinate representation for compositional explanatory variables allows to investigate each compositional part in terms of its relative importance, as used e.g. in Kalivodová et al. (2015) for PLS discriminant analysis (PLS-DA).

3 Thesis objectives

This thesis aims to introduce methodological inovations of the compositional approach in fields of biostatistics as applied to metabolomics and time-use epidemiology. Most of the presented methods are related to regression modelling, what is of primary importance in a biostatistical context. The developments touch upon subjects such as presence of cellwise outliers in dataset and proper coordinate representation of compositions, including the case of high-dimensional data and data of compositional variables with an intrinsic ordering.

4 Theoretical framework and applied methods

4.1 Cellwise and rowwise robust compositional regression

We present a robust estimation procedure for a linear regression model with a real-valued response and compositional explanatory variables, possibly accompanied by additional real-valued covariates, that is designed to handle both cellwise and rowwise outliers (Štefelová et al., 2021a). The method is developed for the regular case with more observations than explanatory variables. It is similar in spirit to the 3-step regression estimator (Leung et al., 2016) as it filters cellwise outliers and apply rowwise robust regression technique. But since a construction of an appropriate coordinate system for compositions is not feasible for incomplete data, our procedure makes use of an imputation step after the filtering. Imputation uncertainty is then reflected on regression coefficients estimates via multiple imputation scheme. The entire procedure is summarized in the following pseudocode involving a number of algorithms for its different stages.

Algorithm 1 Detection of cellwise outliers

Input: Data matrix $\boldsymbol{\mathcal{X}} = (\mathbf{x}_1, \ldots, \mathbf{x}_D, \mathbf{r}_1, \ldots, \mathbf{r}_{P+1})$ of compositional parts and real-valued variables

Output: Index set $\mathcal{O}$ of outlying cells and index set $\mathcal{R}$ of outlying rows

1: ▷ Cellwise outlier detection on pairwise logratios and real-valued variables

2: $\boldsymbol{\mathcal{L}} \leftarrow (\ln(\mathbf{x}_1/\mathbf{x}_2), \ldots, \ln(\mathbf{x}_{D-1}/\mathbf{x}_D), \mathbf{r}_1, \ldots, \mathbf{r}_{P+1})$

3: Apply bivariate filter of Rousseeuw and Van den Bossche (2018) to $\boldsymbol{\mathcal{L}}$

Store index set $\mathcal{O}_{\boldsymbol{\mathcal{L}}} \leftarrow \{(n, j)$: cell in row n and column j of $\boldsymbol{\mathcal{L}}$ is marked as cellwise outlier$\}$
Store index set $\mathcal{R}_{\boldsymbol{\mathcal{L}}} \leftarrow \{n$: row n of $\boldsymbol{\mathcal{L}}$ is marked as rowwise outlier$\}$
▷ Mark outlying cells in compositional parts
Initialize empty set $\mathcal{O}$ ▷ set of indices (n, m) of cells in $\boldsymbol{\mathcal{X}}$ to be marked as cellwise outliers
Initialize empty set $\mathcal{R}$ ▷ set of indices n of rows in $\boldsymbol{\mathcal{X}}$ to be marked as rowwise outliers
for $d \in \{1, \ldots, D\}$ **do**
    Obtain index set $J_d \leftarrow \{j$: column j of $\boldsymbol{\mathcal{L}}$ contains a logratio involving $x_d\}$
    for $n \in \{1, \ldots, N\}$ **do**
        if $\frac{1}{(D-1)} \sum_{j \in J_d} I_{\mathcal{O}_{\boldsymbol{\mathcal{L}}}}((n, j)) \geq 0.5$ **then**
            $\mathcal{O} \leftarrow \mathcal{O} \cup \{(n, d)\}$
        end if
    end for
end for
▷ Adopt outlying cells in real-valued variables from bivariate filter
for $p \in \{1, \ldots, P + 1\}$ **do**
    for $n \in \{1, \ldots, N\}$ **do**
        if $(n, D(D-1)/2 + p) \in \mathcal{O}_{\boldsymbol{\mathcal{L}}}$ **then**
            $\mathcal{O} \leftarrow \mathcal{O} \cup \{(n, D + p)\}$
        end if
    end for
end for
▷ Mark outlying rows and only mark outlying cells that are not part of outlying rows
for $n \in \{1, \ldots, N\}$ **do**
    if $n \in \mathcal{R}_{\boldsymbol{\mathcal{L}}}$ or $\frac{1}{D+P+1} \sum_{m=1}^{D+P+1} I_{\mathcal{O}}((n, m)) >= 0.75$ **then**
        ▷ Marked as rowwise outlier in $\boldsymbol{\mathcal{L}}$ or at least 75% of cells marked as cellwise outliers in $\boldsymbol{\mathcal{X}}$
        $\mathcal{R} \leftarrow \mathcal{R} \cup \{n\}$
        $\mathcal{O} \leftarrow \mathcal{O} \setminus \{(n, m) : m = 1, \ldots, D + P + 1\}$
    end if
end for
return Index sets $\mathcal{O}$ and $\mathcal{R}$

Algorithm 2 Initial $k$nn imputation for compositional data and real-valued variables

Input: Data matrix $\boldsymbol{\mathcal{X}} = (\mathbf{x}_1, \ldots, \mathbf{x}_D, \mathbf{r}_1, \ldots, \mathbf{r}_{P+1})$ of compositional parts and real-valued variables with missing values (outlying cells)

Output: Imputed data matrix $\tilde{\boldsymbol{\mathcal{X}}}$

1: Apply simultaneous $k$nn imputation with Aitchison distance to $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_D)$

2: Store imputed data matrix as $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_D)$

3: Compute pivot coordinates $\tilde{\mathbf{z}}_1^{(1)}, \ldots, \tilde{\mathbf{z}}_{D-1}^{(1)}$ from $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_D$

4: Apply simultaneous $k$nn imputation with Euclidean distance to $\left(\mathbf{r}_1, \ldots, \mathbf{r}_{P+1}, \tilde{\mathbf{z}}_1^{(1)}, \ldots, \tilde{\mathbf{z}}_{D-1}^{(1)}\right)$

5: Store imputed real-valued variables as $\tilde{\mathbf{R}} = (\tilde{\mathbf{r}}_1, \ldots, \tilde{\mathbf{r}}_{P+1})$

6: **return** Imputed data matrix $\tilde{\boldsymbol{\mathcal{X}}} = (\tilde{\mathbf{X}}, \tilde{\mathbf{R}})$

Algorithm 3 Model-based imputation for compositional data and real-valued variables

Input: Data matrix $\boldsymbol{\mathcal{X}} = (\mathbf{x}_1, \ldots, \mathbf{x}_D, \mathbf{r}_1, \ldots, \mathbf{r}_{P+1})$ of compositional parts and real-valued variables with missing values (outlying cells)

Output: Imputed data matrix $\tilde{\boldsymbol{\mathcal{X}}}$, residual scale estimates $\hat{\sigma}_1, \ldots, \hat{\sigma}_{D+P+1}$ from imputation models

1: ▷ Initializations

2: Rearrange first D columns of $\boldsymbol{\mathcal{X}}$ by sorting compositional parts by decreasing amount of missing values

3: Rearrange last $P+1$ columns of $\boldsymbol{\mathcal{X}}$ by sorting real-valued variables by decreasing amount of missing values

4: Obtain index sets $\kappa_m \leftarrow \{n : \text{cell in row } n \text{ and column } m \text{ of } \boldsymbol{\mathcal{X}} \text{ is missing}\}$, $m = 1, \ldots, D+P+1$

5: Obtain index sets $\tau_m \leftarrow \{n : \text{cell in row } n \text{ and column } m \text{ of } \boldsymbol{\mathcal{X}} \text{ is observed}\}$, $m = 1, \ldots, D+P+1$

6: Initialize counter $t \leftarrow 0$ and convergence criterion $\eta \leftarrow \infty$

7: Initialize $\boldsymbol{\mathcal{X}}^{[0]} = \left(\mathbf{x}_1^{[0]}, \ldots, \mathbf{x}_D^{[0]}, \mathbf{r}_1^{[0]}, \ldots, \mathbf{r}_{P+1}^{[0]}\right)$ by applying $k$nn imputation from Algorithm 2 to $\boldsymbol{\mathcal{X}}$

▷ Iterative model-based imputations

while $\eta \geq 0.5$ **do**

$t \leftarrow t + 1$

$\boldsymbol{\mathcal{X}}^{[t]} = \left(\mathbf{x}_1^{[t]}, \ldots, \mathbf{x}_D^{[t]}, \mathbf{r}_1^{[t]}, \ldots, \mathbf{r}_{P+1}^{[t]}\right) \leftarrow \boldsymbol{\mathcal{X}}^{[t-1]} = \left(\mathbf{x}_1^{[t-1]}, \ldots, \mathbf{x}_D^{[t-1]}, \mathbf{r}_1^{[t-1]}, \ldots, \mathbf{r}_{P+1}^{[t-1]}\right)$

▷ Imputations in compositional data

for $d \in \{1, \ldots, D\}$ **do**

Compute pivot coordinates $z_{n1}^{(d)}, \ldots, z_{n,D-1}^{(d)}$ from $x_{n1}^{[t]}, \ldots, x_{nD}^{[t]}$, $n = 1, \ldots, N$

Perform MM-regression of $z_{n1}^{(d)}$ on $z_{n2}^{(d)}, \ldots, z_{n,D-1}^{(d)}, r_{n1}^{[t]}, \ldots, r_{n,P+1}^{[t]}$, $n \in \tau_d$

Compute prediction $\hat{z}_{n1}^{(d)}$ from $z_{n2}^{(d)}, \ldots, z_{n,D-1}^{(d)}, r_{n1}^{[t]}, \ldots, r_{n,P+1}^{[t]}$, $n \in \kappa_d$

Replace $x_{n1}^{[t]}, \ldots, x_{nD}^{[t]}$ with the inverse mapping of $\hat{z}_{n1}^{(d)}, z_{n2}^{(d)}, \ldots, z_{n,D-1}^{(d)}$, $n \in \kappa_d$

Compute robust residual scale estimate $\hat{\sigma}_d$ from MM-regression fit

end for

▷ Imputations in real-valued variables

Compute pivot coordinates $z_{n1}^{(1)}, \ldots, z_{n,D-1}^{(1)}$ from $x_{n1}^{[t]}, \ldots, x_{nD}^{[t]}$, $n = 1, \ldots, N$

for $p \in \{1, \ldots, P+1\}$ **do**

Perform MM-regression of $r_{np}^{[t]}$ on $z_{n1}^{(1)}, \ldots, z_{n,D-1}^{(1)}, r_{n1}^{[t]}, \ldots, r_{n,p-1}^{[t]}, r_{n,p+1}^{[t]}, r_{n,P+1}^{[t]}$, $n \in \tau_p$

Replace $r_{np}^{[t]}$ with prediction $\hat{r}_{np}^{[t]}$ from $z_{n1}^{(1)}, \ldots, z_{n,D-1}^{(1)}, r_{n1}^{[t]}, \ldots, r_{n,p-1}^{[t]}, r_{n,p+1}^{[t]}, r_{n,P+1}^{[t]}$, $n \in \kappa_p$

Compute robust residual scale estimate $\hat{\sigma}_{D+p}$ from MM-regression fit

end for

▷ Update convergence criterion

$\eta \leftarrow \sum_{n=1}^{N} \left[\sum_{d=1}^{D} \left(\frac{x_{nd}^{[t-1]} - x_{nd}^{[t]}}{x_{nd}^{[t]}} \right)^2 + \sum_{p=1}^{P+1} \left(\frac{r_{np}^{[t-1]} - r_{np}^{[t]}}{r_{np}^{[t]}} \right)^2 \right]$

end while

Obtain $\tilde{\boldsymbol{\mathcal{X}}}$ by rearranging columns of $\boldsymbol{\mathcal{X}}^{[t]}$ from last iteration according to original order of columns in $\boldsymbol{\mathcal{X}}$

31: Rearrange residual scale estimates $\hat{\sigma}_1, \ldots, \hat{\sigma}_{D+P+1}$ accordingly

32: **return** Imputed data matrix $\tilde{\boldsymbol{\mathcal{X}}}$ and residual scale estimates $\hat{\sigma}_1, \ldots, \hat{\sigma}_{D+P+1}$

Algorithm 4 Cellwise and rowwise robust compositional regression with bivariate filter and multiple imputation

Input: Compositional data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_D)$, real-valued covariates $\mathbf{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_P)$, real-valued response $\mathbf{y}$

Output: Regression coefficient estimates and corresponding variance estimates

▷ Detect cellwise outliers

Obtain index set $\mathcal{O}$ of cellwise outliers by applying Algorithm 1 to $\boldsymbol{\mathcal{X}} = (\mathbf{x}_1, \ldots, \mathbf{x}_D, \mathbf{c}_1, \ldots, \mathbf{c}_P, \mathbf{y})$

▷ Special case of no cellwise outliers

if $\mathcal{O} = \emptyset$ **then**

    Compute ilr coordinates $\mathbf{i}_1, \ldots, \mathbf{i}_{D-1}$ from $\mathbf{x}_1, \ldots, \mathbf{x}_D$

    Perform MM-regression of $\mathbf{y}$ on $\mathbf{i}_1, \ldots, \mathbf{i}_{D-1}, \mathbf{c}_1, \ldots, \mathbf{c}_P$

    return Coefficient estimates and corresponding variance estimates

end if

▷ Filter and impute cellwise outliers

Replace cells of $\boldsymbol{\mathcal{X}}$ with indices in $\mathcal{O}$ by missing values

Apply model-based imputation with Algorithm 3 to $\boldsymbol{\mathcal{X}} = (\mathbf{x}_1, \ldots, \mathbf{x}_D, \mathbf{c}_1, \ldots, \mathbf{c}_P, \mathbf{y})$

Store imputed data matrix as $\tilde{\boldsymbol{\mathcal{X}}} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_D, \tilde{\mathbf{c}}_1, \ldots, \tilde{\mathbf{c}}_P, \tilde{\mathbf{y}})$

Store residual scale estimates from imputation models as $\hat{\sigma}_1, \ldots, \hat{\sigma}_{D+P+1}$, respectively

▷ Rowwise robust compositional regression with multiple imputation

$N_{\text{out}} \leftarrow N - \sum_{n=1}^{N} \prod_{m=1}^{D+P+1} (1 - I_{\mathcal{O}}((n, m)))$ ▷ Number of observations with outlying cells

$H \leftarrow \max(2, \text{round}(100 \cdot N_{\text{out}}/N))$ ▷ Number of imputations

Obtain $\iota_m \leftarrow \sum_{n=1}^{N} I_{\mathcal{O}}((n, m))$, $m = 1, \ldots, D + P + 1$ ▷ Number of outlying cells per variable

for $h \in \{1, \ldots, H\}$ **do**

    ▷ Add random noise to imputations

    Initialize $\tilde{\boldsymbol{\mathcal{X}}}^{\{h\}} = \left(\tilde{\mathbf{x}}_1^{\{h\}}, \ldots, \tilde{\mathbf{x}}_D^{\{h\}}, \tilde{\mathbf{c}}_1^{\{h\}}, \ldots, \tilde{\mathbf{c}}_P^{\{h\}}, \tilde{\mathbf{y}}^{\{h\}}\right)$ by $\tilde{\mathcal{X}} =$

$(\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_D, \tilde{\mathbf{c}}_1, \ldots, \tilde{\mathbf{c}}_P, \tilde{\mathbf{y}})$

for $(n, m) \in \mathcal{O}$ **do**

  Draw random noise term $e \sim N(0, \hat{\sigma}_m^2(1 + \iota_m/N))$

  if $m \in \{1, \ldots, D\}$ **then** ▷ Compositional parts

    Compute pivot coordinates $\tilde{z}_{n1}^{(m)}, \ldots, \tilde{z}_{n,D-1}^{(m)}$ from $\tilde{x}_{n1}, \ldots, \tilde{x}_{nD}$

    $\tilde{z}_{n1}^{(m)} \leftarrow \tilde{z}_{n1}^{(m)} + e$

    Replace $\tilde{x}_{n1}^{\{h\}}, \ldots, \tilde{x}_{nD}^{\{h\}}$ with the inverse mapping of $\tilde{z}_{n1}^{(m)}, \ldots, \tilde{z}_{n,D-1}^{(m)}$

  else if $m \in \{D+1, \ldots, D+P\}$ **then** ▷ Real-valued variables

    $\tilde{c}_{n,m-D}^{\{h\}} \leftarrow \tilde{c}_{n,m-D} + e$

  else ▷ Response variable

    $\tilde{y}_n^{\{h\}} \leftarrow \tilde{y}_n + e$

  end if

end for

▷ Rowwise robust compositional regression

Compute ilr coordinates $\mathbf{i}_1, \ldots, \mathbf{i}_{D-1}$ from $\tilde{\mathbf{x}}_1^{\{h\}}, \ldots, \tilde{\mathbf{x}}_D^{\{h\}}$

Perform MM-regression of $\tilde{\mathbf{y}}^{\{h\}}$ on $\mathbf{i}_1, \ldots, \mathbf{i}_{D-1}, \tilde{\mathbf{c}}_1^{\{h\}}, \ldots, \tilde{\mathbf{c}}_P^{\{h\}}$

Store coefficient estimates as $\left(\hat{\beta}_0^{\{h\}}, \ldots, \hat{\beta}_{D-1+P}^{\{h\}}\right)^\top$

Store variance estimates as $\left(\hat{\phi}_0^{\{h\}}, \ldots, \hat{\phi}_{D-1+P}^{\{h\}}\right)^\top$

end for

▷ Aggregate results from multiple imputation

Compute final coefficient estimates $\hat{\beta}_k \leftarrow \frac{1}{H} \sum_{h=1}^{H} \hat{\beta}_k^{\{h\}}$, $k = 0, \ldots, D-1+P$

Compute average within-imputation variances $\hat{\zeta}_k \leftarrow \frac{1}{H} \sum_{h=1}^{H} \hat{\phi}_k^{\{h\}}$, $k = 0, \ldots, D-1+P$

Compute between-imputation variances $\hat{\xi}_k \leftarrow \frac{1}{H-1} \sum_{h=1}^{H} \left(\hat{\beta}_k^{\{h\}} - \hat{\beta}_k\right)^2$, $k = 0, \ldots, D-1+P$

Compute variance estimates $\hat{\phi}_k \leftarrow \hat{\zeta}_k + \frac{H+1}{H} \hat{\xi}_k$, $k = 0, \ldots, D-1+P$

return Coefficient estimates $(\hat{\beta}_0, \ldots, \hat{\beta}_{D-1+P})^\top$ and corresponding variance estimates $(\hat{\phi}_0, \ldots, \hat{\phi}_{D-1+P})^\top$

4.2 Weighted pivot coordinates for a regression task with high-dimensional compositional data

The method presented here extends previous work in PLS modelling with compositional data by using weighted pivot coordinates (Section 2.4) instead of the ordinary ones with a newly introduced weighting strategy aiming to enhance the identification of markers (Štefelová et al., 2021b). This is achieved by defining weights which focus on the correlation structure between a real-valued response variable and pairwise logratios aggregated into the first pivot coordinate in order to downplay the effect of irrelevant logratios and enhance the most relevant ones in relation to the outcome variable.

In order to make a sensible choice of weights for a regression purpose, we must first determine what we understand as a marker in our context. We aim for a compositional part to be identified as a marker if a relatively significant number of pairwise logratios including that part are strongly associated with the response variable Y. Moreover, considering the pairwise logratios where the part of interest is in the numerator, that strong association should be (possibly with a few exceptions) in one direction, either positive or negative.

Accordingly, we propose to construct weighted pivot coordinates (3) using weights $\gamma_d^{(l)},\ d = 2, \ldots, D,\ l = 1, \ldots, D$, defined as follows:

$$\gamma_d^{(l)} = \frac{\tilde{\gamma}_d^{(l)}}{\sum_{d=2}^{D} \tilde{\gamma}_d^{(l)}}, \tag{6}$$

with

$$\tilde{\gamma}_d^{(l)} = \left| \int_0^{r_d^{(l)}} \hat{f}^{(l)}(\lambda)\, d\lambda \right|, \quad r_d^{(l)} = \operatorname{cor}\left(Y, \ln \frac{x_1^{(l)}}{x_d^{(l)}} \right),$$

$$\hat{f}^{(l)}(\lambda) = \frac{1}{\nu(D-1)} \sum_{d=2}^{D} \mathcal{K}\left(\frac{\lambda - \tilde{r}_d^{(l)}}{\nu} \right), \quad \tilde{r}_d^{(l)} = \begin{cases} 0, & \text{if } \left| r_d^{(l)} \right| < o^{(l)}, \\ r_d^{(l)}, & \text{otherwise,} \end{cases}$$

$$o^{(l)} = 2 \times \min\left(\frac{\sum_{d=2}^{D} \mathcal{I}(r_d^{(l)} \geq 0)}{D-1}, \frac{\sum_{d=2}^{D} \mathcal{I}(r_d^{(l)} < 0)}{D-1} \right),$$

where $\hat{f}$ is a kernel density estimator, $\mathcal{K}$ is a Gaussian kernel function (defined

as $\mathcal{K}(\lambda) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\lambda^2}$), ν (set to $\nu = 0.05$) is the bandwidth used and $\mathcal{I}$ is an indicator function.

Thus, for the lth part, rearranged into the first position as $x_1^{(l)}$, the set of correlations $r_2^{(l)}, \ldots r_D^{(l)}$ is smoothed using kernel density estimation (Everitt and Hothorn, 2011), with the correlations under the given threshold being set to zero in order to estimate the density, and the weight $\tilde{\gamma}_d^{(l)}$ is computed as the area under the curve from zero to the value of the correlation $r_d^{(l)}$. The rationale behind this weighting scheme is to minimise the influence of logratios that are not related to the response Y, so that higher weights are given to logratios strongly correlated to Y. Among these, the procedure emphasises those logratios representing the main trend in the distribution of $r_2^{(l)}, \ldots r_D^{(l)}$ by using a kernel density. In order to prevent from false positives, correlations with absolute value smaller than a cut-off value $o^{(l)}$ are set to zero when conducting kernel density estimation. The value of $o^{(l)}$ modulates the effect of the weighting, which is downplayed with increasing values of $o^{(l)}$. Therefore, the value given to $o^{(l)}$ is higher when there is no clear trend in the distribution of the correlations and vice versa. For instance, when all correlations are positive, then $o^{(l)} = 0$, the density is estimated from the unaltered set of correlations and, as a consequence, the logratios strongly correlated with Y are highlighted. On the other hand, when half of the correlations are positive and half are negative, then $o^{(l)} = 1$, and all correlations are taken to be zero for the density estimation. Thus, the value of the area under the curve from 0 to $r_d^{(l)}$ is practically the same for any d (apart from the cases where $r_d^{(l)}$ are the closest to 0), so only logratios very weakly correlated with Y are suppressed, while the rest are treated equally. The final normalised weight results from dividing each $\tilde{\gamma}_d^{(l)}$ by the sum of all of them.

4.3 Pivoting balances and their use with movement behaviour compositions

Specific feature of movement behavior (time-use) compositions is their ordinal character. In case of wake-time compositions WMB = (SB, LPA, MPA, VPA), the parts (standing for sedentary behavior, light, moderate and vigorous physical

activity) are thus placed in ascending order according to their intensity. With 24-hour compositions MB = (Sleep, SB, LPA, MPA, VPA), the situation is a little more complicated – particularly when assessing the relationship between the MB composition and health outcomes. For it is generally accepted that a higher health benefit is obtained from more physically demanding activities, but the role of sleep is less clear.

Therefore for the representation of time-use composition we propose so-called *pivoting balances* (Štefelová et al., 2021c). These combine ideas behind balances (Section 2.2) and pivot coordinates (Section 2.3). They result from L balance coordinate systems in which a balance of interest is isolated in the first coordinate.

In case of WMB, $L = 3$ sets of coordinates $\left(b_1^{(l)}, b_2^{(l)}, b_3^{(l)}\right)^\top$, $l = 1, 2, 3$ are considered. In the first system, SB is set against the remaining (active) parts in the initial partition. In the following systems, the initial subgroup consisting of SB is subsequently accompanied by the other activities from the least to the most intense. Thus, the three first balances of interest have a symbolic notation SB_LPA.MPA.VPA, SB.LPA_MPA.VPA and SB.LPA.MPA_VPA. Table 1 illustrates an exemplary SBP to obtain the required set of balances. Note that the reciprocal balances, i.e. swapping the subsets of behaviors in the logratio, differ only by the sign. So in fact, the three first balance coordinates provide information also about the balances LPA.MPA.VPA_SB, MPA.VPA_SB.LPA and VPA_SB.LPA.MPA. As for the interpretation, e.g. SB.LPA_MPA.VPA is a contrast of time spent in the two least physically demanding activities against the two most intense activities.

In case of MB, $L = 7$ sets of coordinates $\left(b_1^{(l)}, b_2^{(l)}, b_3^{(l)}, b_4^{(l)}\right)^\top$, $l = 1, \ldots, 7$ are considered. In the first system, sleep is set against the remaining parts in the initial partition. In the following systems, the initial subgroup consisting of sleep is subsequently accompanied by the other activities from the least to the most intense; and vice versa, sleep is subsequently accompanied by the other activities from the most to the least intense. With the aim of examining the whole 24-hour behavior pattern, in each system we focus only on the first balance involving all five parts, i.e. on the pivoting balances with symbolic notation Sleep_SB.LPA.MPA.VPA, Sleep.SB_LPA.MPA.VPA, Sleep.SB.LPA_MPA.VPA,

Table 1: Examplary SBP for WMB composition which results in the required pivoting balance systems with the (first) balance of interest as noted in the captions. Parts chosen for the numerator and denominator of the jth balance are coded $+$ and $-$, respectively; 0 indicates that the part is not included in the respective balance.

j	x_1	x_2	x_3	x_4	r_j	s_j
1	$+$	$-$	$-$	$-$	1	3
2	0	$+$	$-$	$-$	1	2
3	0	0	$+$	$-$	1	1

(a) SB_LPA.MPA.VPA

j	x_1	x_2	x_3	x_4	r_j	s_j
1	$+$	$+$	$-$	$-$	2	2
2	$+$	$-$	0	0	1	1
3	0	0	$+$	$-$	1	1

(b) SB.LPA_MPA.VPA

j	x_1	x_2	x_3	x_4	r_j	s_j
1	$+$	$+$	$+$	$-$	3	1
2	$+$	$-$	$-$	0	2	1
3	0	$+$	$-$	0	1	1

(c) SB.LPA.MPA_VPA

Sleep.SB.LPA.MPA_VPA, Sleep.VPA_SB.LPA.MPA, Sleep.MPA.VPA_SB.LPA, Sleep.LPA.MPA.VPA_SB (and the corresponding reciprocals). Table 2 illustrates an exemplary SBP to obtain the required set of balances. As for the interpretation, e.g. the balance Sleep_SB.LPA.MPA.VPA compares time spent in sleep relative to waking-time behaviors, Sleep.SB_LPA.MPA.VPA is a contrast of time spent in non-active behaviors against physical activities, and so on.

Further, as demonstrated in Štefelová et al. (2021c), pivoting balances and their implementation into an adapted formulation of compositional PLS biplot can be used to facilitate a synthetic and meaningful graphical display of compositions and their relationships with outcome variables. The main idea behind the construction of such a PLS biplot is to display only loadings corresponding to the first balance from each coordinate system (possibly together with the loadings from their reciprocals), and the loadings corresponding to non-compositional variables, similarly to the case for PCA biplots (Kynčlová et al., 2016). (Scores are taken from any given coordinate system as these are invariant to the specific choice of balances).

Table 2: Examplary SBP for MB composition which results in the required pivoting balance systems with the (first) balance of interest as noted in the captions. Parts chosen for the numerator and denominator of the jth balance are coded $+$ and $-$, respectively; 0 indicates that the part is not included in the respective balance.

j	x_1	x_2	x_3	x_4	x_5	r_j	s_j
1	$+$	$-$	$-$	$-$	$-$	1	4
2	0	$+$	$-$	$-$	$-$	1	3
3	0	0	$+$	$-$	$-$	1	2
4	0	0	0	$+$	$-$	1	1

(a) Sleep_SB.LPA.MPA.VPA

j	x_1	x_2	x_3	x_4	x_5	r_j	s_j
1	$+$	$+$	$-$	$-$	$-$	2	3
2	$+$	$-$	0	0	0	1	1
3	0	0	$+$	$-$	$-$	1	2
4	0	0	0	$+$	$-$	1	1

(b) Sleep.SB_LPA.MPA.VPA

j	x_1	x_2	x_3	x_4	x_5	r_j	s_j
1	$+$	$+$	$+$	$-$	$-$	3	2
2	$+$	$-$	$-$	0	0	1	2
3	0	$+$	$-$	0	0	1	1
4	0	0	0	$+$	$-$	1	1

(c) Sleep.SB.LPA_MPA.VPA

j	x_1	x_2	x_3	x_4	x_5	r_j	s_j
1	$+$	$+$	$+$	$+$	$-$	4	1
2	$+$	$-$	$-$	$-$	0	1	3
3	0	$+$	$-$	$-$	0	1	2
4	0	0	$+$	$-$	0	1	1

(d) Sleep.SB.LPA.MPA_VPA

j	x_1	x_2	x_3	x_4	x_5	r_j	s_j
1	$+$	$-$	$-$	$-$	$+$	2	3
2	$+$	0	0	0	$-$	1	1
3	0	$+$	$-$	$-$	0	1	2
4	0	0	$+$	$-$	0	1	1

(e) Sleep.VPA_SB.LPA.MPA

j	x_1	x_2	x_3	x_4	x_5	r_j	s_j
1	$+$	$-$	$-$	$+$	$+$	3	2
2	$+$	0	0	$-$	$-$	1	2
3	0	0	0	$+$	$-$	1	1
4	0	$+$	$-$	0	0	1	1

(f) Sleep.MPA.VPA_SB.LPA

j	x_1	x_2	x_3	x_4	x_5	r_j	s_j
1	$+$	$-$	$+$	$+$	$+$	4	1
2	$+$	0	$-$	$-$	$-$	1	2
3	0	0	$+$	$-$	$-$	1	2
4	0	0	0	$+$	$-$	1	1

(g) Sleep.LPA.MPA.VPA_SB

5 Original results and summary

This thesis contributes novel methods for the processing of data in biostatistics that are of compositional nature, i.e. data conveying relative information. The specific features of compositions call for an adequate approach to their statistical analysis. The logratio methodology is used for their proper statistical treatment. The application of the developments is demonstrated in time-use (physical) and livestock greenhouse gas emission research. Additionally, the work conducted during this Ph.D. project contributed new scientific insights through a number of interdisciplinary collaborations.

First, a new method for compositional regression that is robust against cellwise and rowwise outliers was introduced. Cellwise outliers are first filtered and then imputed by robust estimates. Afterwards, rowwise robust compositional regression using a multiple imputation scheme is performed to obtain model coefficient estimates. An application to bio-environmental data relating biological processes in livestock rumen with methane emissions (not included in this summary) revealed that the proposed procedure (compared to other regression methods) leads to conclusions that are best aligned with established scientific knowledge. An extensive simulation study (not included in this summary) shows that the procedure generally outperforms a traditional rowwise-only robust regression method (MM-estimator). Moreover, our procedure yields better or comparable results to recently proposed cellwise robust regression methods (shooting S-estimator, 3-step regression) while it is preferable for interpretation through the use of appropriate coordinate systems for compositional data.

Next, a new weighting strategy for the construction of weighted pivot coordinates was proposed. Designed to improve PLS-based marker discovery in high-dimensional compositional data, it draws on the correlation between response variable and pairwise logratios aggregated into the first coordinate. The illustrative application to investigate the association between ruminal high-throughput metabolite signals and methane emission in cattle (not included in this summary) extended the study in Section 4.1 to the high-dimensional case. It demonstrated the practical relevance and potential of the proposed approach, providing results compatible with previous knowledge along with a higher sensitivity to identify meaningful markers. A simulation study (not included in this summary) pro-

vided additional evidence that this proposed logratio coordinate representation enhances the discovery of markers, although it results in slightly worse specificity.

Finally, the compositional approach within time-use epidemiology was studied. Proper coordinate representation for movement behavior data was discussed given the ordinal character (in terms of physical intensity) of daily activities. In the first application (not included in this summary), wake-time movement behavior data were examined via robust linear regression and visualization tools such as compositional mean barplots and ternary diagrams. The second application (not included in this summary) demonstrated how an adapted version of compositional PLS regression and biplot based on the newly introduced concept of pivoting balances could be employed to evaluate the association between 24-hours behavior patterns and a health marker.

All computation in this work were performed within the R environment for statistical computing (R Core Team, 2021). The related codes are available at `https://github.com/aalfons/lmcrCoda`, `https://github.com/StefelovaN/Weighted-pivot-coordinates`, `https://github.com/StefelovaN/Robust-CoDA-WMB` and `https://github.com/StefelovaN/Balance-based-PLS-biplot`.

List of publications

- Štefelová N, Dygrýn J, Hron K, Gába A, Rubín L, Palarea-Albaladejo J (2018) Robust compositional analysis of physical activity and sedentary behavior data. International Journal of Environmetal Research and Public Health 15(10):2248, DOI 10.3390/ijerph15102248

- Štefelová N, Alfons A, Palarea-Albaladejo J, Filzmoser P, Hron K (2021) Robust regression with compositional covariates including cellwise outliers. Advances in Data Analysis and Classification, DOI 10.1007/s11634-021-00436-9

- Štefelová N, Palarea-Albaladejo J, and Hron K (2021) Weighted pivot coordinates for PLS-based marker discovery in high-throughput compositional data. *Under review*

- Štefelová N, Palarea-Albaladejo J, Hron K, Gába A, Dygrýn J (2021) Compositional PLS biplot based on pivoting balances: a graphical tool to examine the association between 24-hour movement behaviours and health outcomes. *Under review*

- Pelclová J, Štefelová N, Pechová J, Dygrýn J, Gába A, Zając-Gawlak I (2018) Reallocating Time from Sedentary Behavior to Light and Moderate-to-Vigorous Physical Activity: What Has a Stronger Association with Adiposity in Older Adult Women? International Journal of Environmetal Research and Public Health 15(7):1444, DOI 10.3390/ijerph15071444

- Hradilová I, Duchoslav M, Brus J, Pechanec V, Hýbl M, Kopecký P, Smržová L, Štefelová N, Václavek T, Machalová J, Hron K, Bariotakis M, Pirintsos S, Smýkal P (2019) Variation in wild pea (Pisum sativum subsp. elatius) seed dormancy and its relationship to the environment and seed coat traits. PeerJ, 7(e6263), DOI: 10.7717/peerj.6263

- Cuberek R., Pelclová J, Gába A, Pechová J, Svozilová Z, Přidalová M, Štefelová N Hron K (2019) Adiposity and changes in movement-related behaviors in older adult women in the context of the built environment:

a protocol for prospective cohort study. BMC Public Health 19:1522, DOI 10.1186/s12889-019-7905-8

- Pelclová J, Štefelová N, Dumuid D, Pedišić Ž, Hron K, Gába A, Olds T, Pechová J, Zając-Gawlak I, Tlučáková L (2020) Are longitudinal reallocations of time between movement behaviors associated with adiposity among elderly women? A compositional isotemporal substitution analysis. International Journal of Obesity 44(4):857–864, DOI 10.1038/s41366-019-0514-x

- Gába A, Pedišić Ž, Štefelová N, Dygrýn J, Hron K, Dumuid D, Tremblay M (2020) Sedentary behavior patterns and adiposity in children: A study based on compositional data analysis. BMC Pediatric 20:147, DOI 10.1186/s12887-020-02036-6

- Gába A, Dygrýn J, Štefelová N, Rubín L, Hron K, Jakubec J, Pedišić Ž (2020) How do short sleepers use extra waking hours? A compositional analysis of 24-hour time-use patterns among children and adolescents. International Journal of Behavioral Nutrition and Physical Activity 17:104, DOI 10.1186/s12966-020-01004-8

- Gába A, Pelclová J, Štefelová N, Přidalová M, Zając-Gawlak I, Tlučáková L, Pechová J, Svozilová Z (2020) Prospective study on sedentary behavior patterns and changes in body composition parameters in older women: A compositional and isotemporal substitution analysis. Clinical Nutrition, DOI 10.1016/j.clnu.2020.10.020

- Gába A, Dygrýn J, Štefelová N, Rubín L, Hron, K, Jakubec L (2021) Replacing school and out-of-school sedentary behaviors with physical activity, and its associations with adiposity in children and adolescents: A compositional isotemporal substitution analysis. Environmental Health and Preventive Medicine, 26(1):16, DOI 10.1186/s12199-021-00932-6

- Germano-Soares AH, Tassitano R, Farah B, Andrade-Lima A, Correia M, Gába A, Štefelová N, Puech-Leão P, Wolosker N, Cucato G, Ritti-Dias R (2021) Reallocating time from sedentary behavior to physical activity in patients with peripheral artery disease: analyzing the effects on walking ca-

pacity using compositional data analysis. Journal of Physical Activity & Health, DOI 10.1123/jpah.2020-0487

- Pelclová J, Štefelová N, Olds T, Dumuid D, Hron K, Chastin S, Pedišić Ž (2021) A study on prospective associations between adiposity and 7-year changes in movement behaviors among older women based on compositional data analysis. BMC Geriatrics, 21(1):203, DOI 10.1186/s12877-021-02148-3

- Gallo J, Lošťák J, Gába A, Dygrýn J, Baláž L, Štefelová N (2021) Accelerometer-based measures of 24-hour movement behaviors in patients before total knee arthroplasty: A study based on compositional data analysis. *Under review*

List of conferences

- ODAM 2017, 31.5.–2.6.2017, Olomouc (CZ): Regression analysis with compositional covariates in the presence of cellwise contamination (presentation)
- CoDaWork 2017, 5.–9.6.2017, Abbadia San Salvatore (IT): Robust regression with compositional covariates in the presence of cellwise contamination (poster)
- MOVISS 2017, 20.–23.9.2017, Vorau (AT): Regression analysis with compositional covariates in the presence of cellwise outliers (poster)
- ERCIM 2017, 15.–18.12.2017, London (UK): Robust regression on compositional variables including cellwise outliers (presentation)
- ROBUST 2018, 21.–26.1.2018, Rybník (CZ): Robustní regrese s kompozičními vysvětlujícími proměnnými s odlehlostí na úrovni buněk (poster + presentation, in Czech)
- DSSV 2018, 9.–11.7.2018, Vienna (AT): Compositional PLS regression with weighted pivot coordinates and its application to metabolomic data (presentation)
- BioSS annual meeting 2018, 26.–27.11.2018, Edinburgh (UK): Robust regression with compositional explanatory variables including cellwise outliers (presentation)
- ODAM 2019, 29.–31.5.2019, Olomouc (CZ): Weighted pivot coordinates in PLS regression with compositional covariates and its application to metabolomic data (presentation)
- CoDaWork 2019, 3.–8.6.2019, Terrassa (ES): Robust regression with compositional covariates including cellwise outliers (poster, Best poster award)
- INTUE annual meeting and conference 2019, 8.–10.6.2019, Olomouc (CZ): Robust compositional analysis of physical activity and sedentary behavior data (presentation)

Reference

Aitchison J (1986) The Statistical Analysis of Compositional Data. Chapman & Hall, London, DOI 10.1007/978-94-009-4109-0

Egozcue J, Pawlosky-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 J, 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

Everitt B, Hothorn T (2011) An Introduction to Applied Multivariate Analyses with R. Springer, New York, DOI 10.1007/978-1-4419-9650-3

Filzmoser P, Hron K, Templ M (2018) Applied Compositional Data Analysis. 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(4):455–468, DOI 10.1007/s11004-011-9333-x

Härdle W, Simar L (2012) Applied Multivariate Statistical Analysis. Springer, Heidelberg, DOI 10.1007/978-3-662-45171-7

Höskuldson A (1988) PLS regression methods. Journal of Chemometrics 2:211–228, DOI 10.1002/cem.1180020306

Hron K, Filzmoser P, Thompson K (2012) Linear regression with compositional explanatory variables. Journal of Applied Statistics 39(5):1115–1128, DOI 10.1080/02664763.2011.644268

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

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

Kynčlová P, Filzmoser P, Hron K (2016) Compositional biplots including external non-compositional variables. Statistics 50:1–17, DOI 10.1080/02331888.2015.1135155

Leung A, Zhang H, Zamar R (2016) Robust regression estimation and inference in the presence of cellwise and casewise contamination. Computational Statistics & Data Analysis 99:1–11, DOI 10.1016/j.csda.2016.01.004

Maronna R, Martin R, Yohai V (2002) Robust Statistics: Theory and Methods. John Wiley & Sons, Chichester, DOI 10.1002/0470010940

Öllerer V, Alfons A, Croux C (2016) The shooting S-estimator for robust regression. Computational Statistics 31(3):829–844, DOI 10.1007/s00180-015-0593-7

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

R Core Team (2021) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.r-project.org

Rousseeuw P, Van den Bossche W (2018) Detecting deviating data cells. Technometrics 60(2):135–145, DOI 10.1080/00401706.2017.1340909

Štefelová N, Dygrýn J, Hron K, Gába A, Rubín L, Palarea-Albaladejo J (2018) Robust compositional analysis of physical activity and sedentary behaviour data. International Journal of Environmental Research and Public Health 15(10):2248, DOI 10.3390/ijerph15102248

Štefelová N, Alfons A, Palarea-Albaladejo J, Filzmoser P, Hron K (2021a) Robust regression with compositional covariates including cellwise outliers. Advances in Data Analysis and Classification DOI 10.1007/s11634-021-00436-9

Štefelová N, Palarea-Albaladejo J, Hron K (2021b) Weighted pivot coordinates for PLS-based marker discovery in high-throughput compositional data. *Under review*

Štefelová N, Palarea-Albaladejo J, Hron K, Gába A, Dygrýn J (2021c) Compositional PLS biplot based on pivoting balances: a graphical tool to examine the association between 24-hour movement behaviours and health outcomes. *Under review*

Yohai V (1987) High breakdown point and high efficiency robust estimates for regression. The Annals of Statistics 15(2):642–656, DOI 10.1214/aos/1176350366