SORT 47 (2) July-December 2023, 1-26 DOI:

Simple enough, but not simpler: reconsidering additive logratio coordinates in compositional analysis

V. Nesrstová[1], P. Jašková[1], I. Pavlů[1] K. Hron[1], J. Palarea-Albaladejo[2], A. Gába[3], J. Pelclová[4] and K. Fačevicová[5]

Abstract

Compositional data, multivariate observations carrying relative information, are popularly expressed in additive logratio coordinates which are easily interpretable as they use one of the components as ratioing part to produce pairwise logratios. These coordinates are however oblique and they lead to issues when applying multivariate methods on them, including widely-used techniques such as principal component analysis and linear regression. In this paper we propose a way to redefine alr coordinates with respect to an orthonormal system and we also extend the idea to the case of compositional tables. The new approach is demonstrated in an application to movement behavior data.

MSC: *62H25, 62J05.*

Keywords: *Compositional data, compositional tables, regression, principal component analysis.*

1. Introduction

Compositional data analysis concerns extracting knowledge from data carrying relative information (Aitchison, 1986). Technically this involves the representation of the original components in logratio coordinates which are, fairly naturally, desirable to be

[1] Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University Olomouc, 17. listopadu 12, Olomouc, 771 46, Czech Republic.

[2] Department of Computer Science, Applied Mathematics and Statistics, University of Girona, 17003 Girona, Spain.

[3] Department of Natural Sciences in Kinanthropology, Faculty of Physical Culture, Palacký University Olomouc, Třída Míru 117, Olomouc, 771 00, Czech Republic.

[4] Institute of Active Lifestyle, Faculty of Physical Culture, Palacký University Olomouc, Třída Míru 117, Olomouc, 771 00, Czech Republic.

[5] Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University Olomouc, 17. listopadu 12, Olomouc, 771 46, Czech Republic. Email: kamila.facevicova@upol.cz

Received: November 2022
Accepted: May 2023

somehow interpretable in terms of the original components to facilitate scientific insight. Following on the early additive and centred logratio transformations (alr and clr respectively) proposed by Aitchison (1986), formal developments in the last two decades, crucially the characterisation of the sample space for compositional data (the simplex) as an Euclidean space, have led to the popularisation of so-called isometric logratio (ilr) coordinates (Egozcue et al., 2003). Ilr coordinates have been recently re-coined orthonormal logratio (olr) coordinates to more precisely honour their particular geometrical properties. Any of these logratio representations aim to map the data into real space so that ordinary statistical methods can be applied for analysis. They are all connected to each other by algebraic manipulation and, depending on the method and the purpose of the data analysis, they can be used indistinctly and lead to the same results. However, the olr representation has been in recent times favoured since it is directly deducted from the geometrical structure of the simplex as real-valued coordinates with respect to an orthonormal basis. In first instance this is technically consistent with the own geometry of the simplex but, beyond that, it overcomes some difficulties and inconsistencies with the alr and clr representations (see e.g. Pawlowsky-Glahn et al. (2015)). Even so, sometimes practitioners find the interpretation of olr coordinates challenging in real-world applications. Particular classes of olr coordinates have been proposed aiming to overcome this difficulty. These include so-called balances, which are the flagship approach and roughly interpreted as comparisons between subsets of components (Egozcue and Pawlowsky-Glahn, 2005), and pivot coordinates (Filzmoser, Hron and Templ, 2018), which are aimed at synthesising all the relative information of a component against all the others in a single coordinate.

Even so, some criticism has been recently brought out questioning the role of balances in compositional data analysis and the actual relevance of orthonormality of logratio coordinates in general (Greenacre, 2019b, 2018, 2019a). It is argued in these works that balances result from some form of mathematical funambulism and that orthonormality is not that much necessary and, if required, it can be closely approximated by a much simpler logratio coordinate system. One of these simpler systems is obtained by just using alr coordinates, which result from just dividing all components by one of them (the ratioing part). Specifically, for a D-part composition $\mathbf{x} = (x_1, \ldots, x_D)$, and x_D being the rationing part up to a possible permutation, alr coordinates are given by

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

These logratios are sometimes even identified with the original compositional parts in some fields where the scientific community widely accepts the role of x_D as reference or normalizing part. On the down side, however, they do not preserve distances and angles when moving from the simplex to the real space (they refer to an oblique instead of orthonormal coordinate system) (Pawlowsky-Glahn, Egozcue and Tolosana-Delgado, 2015). This causes problems for both PCA, which is only orthogonal equivariant (Hron et al., 2021), and regression analysis, where the regression coefficients lose their stan-

dard interpretation (Coenders and Pawlowsky-Glahn, 2020). One recent attempt to provide a reasonable alternative is found in Hron et al. (2021), which aims to maintain orthonormality while enabling the use of simpler pairwise logratios. They are called *backwards pivot coordinates* in reference to their relationship with pivot coordinates. Hron et al. (2021) also demonstrate that orthonormality really matters when popular statistical methods such as principal component analysis or regression analysis are applied to compositional data.

In this paper we present backwards pivot coordinates as a valid alternative to additive logratio coordinates, stressing the associated gains in interpretability in the context of principal component analysis (PCA) biplots and regression analysis. Furthermore, we extend the concept to two-factorial compositions, a.k.a. compositional tables (Egozcue et al., 2015; Fačevicová et al., 2018), for which to our best knowledge no appropriate counterpart to alr coordinates is available. In Section 2, backwards pivot coordinates and their extension to compositional tables are introduced. The subsequent sections 3 and 4 are devoted to the development and interpretation of PCA and regression analysis in terms of backwards pivot coordinates. Section 5 demonstrates an application using movement behavior data in the form of both standard vector compositions and compositional tables. Finally, a synthesis of the presented material is outlined.

2. Backwards pivot coordinates and beyond

Proper choice of logratio coordinates is fundamental for reliable analysis of compositional data. Here we briefly review current approaches for ordinary vector compositional data and then proceed to the more general setting of compositional tables.

2.1. Vector compositional data

It is common in compositional data analysis that different logratio representations are used for different purposes. Thus, clr coefficients of the form

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

are commonly used to produce compositional biplots (Aitchison and Greenacre, 2002), where $g(\mathbf{x})$ stands for the geometric mean of the parts of the composition $\mathbf{x}$. Although clr coefficients are isometric with respect to the so-called Aitchison geometry of the sample space of compositions and are easy to interpret (the original components are symmetrically represented with respect to the overall geometric mean at the observation level), they lead to singular covariance matrix in the clr-space and, thus, hinder the use of methods that require regular covariance matrices. Using olr coordinates overcomes this issue and reflects the actual $D-1$ dimensionality of compositions. However, there are infinitely many possibilities to construct olr coordinates (although they are all related by orthogonal rotation) and, as mentioned above, some special cases such as

balances (Egozcue et al., 2005) and pivot coordinates (Filzmoser et al., 2018) facilitate interpretability.

A pivot coordinate system allows to aggregate all the relative information about a given compositional part in the first coordinate (the pivoting coordinate), including the possibility of defining weights for the logratios aggregated therein (Hron et al., 2017). This idea is extended by Hron et al. (2021) to define *backwards pivot coordinates* (bpcs), where a pairwise logratio plays the role of pivoting coordinate. Using bpcs allows to easily capture the information conveyed in alr coordinates while fulfilling orthogonality. Specifically, $D-1$ bpc systems can be written in the form

$$\mathrm{bpc}(\mathbf{x}^{(l)})_i = \sqrt{\frac{i}{i+1}} \ln \frac{x_{i+1}^{(l)}}{\left(\prod_{j=1}^{i} x_j^{(l)}\right)^{1/i}}, \; i = 1, \ldots, D-1, \tag{3}$$

where $\mathbf{x}^{(l)} = (x_1^{(l)}, \ldots, x_D^{(l)})$, $l \in \{1, \ldots, D-1\}$, stands for the permutation of the parts in $\mathbf{x}$ so that the l-th part is placed at the second position and the rationing part (e.g. x_D for the sake of simplicity) is placed at the first position. That is, $\mathbf{x}^{(l)} = (x_D, x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_{D-1})$. Then the pairwise logratio of interest (pivoting coordinate) is given by

$$\mathrm{bpc}(\mathbf{x}^{(l)})_1 = \frac{1}{\sqrt{2}} \ln \frac{x_2^{(l)}}{x_1^{(l)}} = \frac{1}{\sqrt{2}} \ln \frac{x_l}{x_D}. \tag{4}$$

Consequently, by varying the part of interest x_l, which is placed at the second position in the respective permuted vector $\mathbf{x}^{(l)}$, all alr-like coordinates can be obtained.

Following Müller et al. (2018), the requirement of orthonormality can be even relaxed in regression analysis by replacing it by just orthogonality, which in practice means to remove the normalizing constants in (3). This helps to simplify interpretation of regression coefficients. A bpcs representation (3) without normalizing constant $\sqrt{i/(i+1)}$ will be denoted in the following by $\mathbf{z}(\mathbf{x}^{(l)})$. Orthonormality is however important e.g. for principal component analysis because here it guarantees that the total explained variability coincides with that based on the clr (or any other olr) representation.

2.2. Compositional tables

Although for vector compositional data the use of non-orthonormal coordinate systems such as alr can be argued on the basis of practical convenience, this is no longer the case for compositional tables (Fačevicová et al., 2018) nor multi-factorial compositional data in general (Fačevicová, Filzmoser and Hron, 2023; Hron, Machalová and Menafoglio, 2023). For these latter, the concept of orthogonality is crucial for their decomposition into independent and interactive parts and preserving their respective dimensionality.

Moreover, while in ordinary compositional data the most basic information is contained in the pairwise logratios, compositional tables (comprising I rows and J columns) involve two types of elemental objects. On the one hand, the within-factor structure is

given by $\binom{I}{2}$ (or $\binom{J}{2}$) pairwise balances between the respective row (column) levels. On the other hand, the interactions between rows and columns are then naturally preserved in $\binom{I}{2}\binom{J}{2}$ simple (four-part) log odds-ratios, which represent the elemental source of information also by contingency tables (Agresti (2013), Ch. 2.4). This needs to be taken into account for construction of any logratio coordinates aimed to generalize the concept of alr coordinates (in the form of bpcs) to multi-factorial compositions.

With the aim of defining a coordinate system highlighting a specific source of elemental information, we introduce $\mathbf{x}^{(kl)}$ as a permuted version of the original compositional table $\mathbf{x}$. Within this table, normalizing row and column (let say the I-th row and the J-th column for the sake of simplicity) are placed at the first position followed by the k-th row and the l-th column at the second position in the respective dimensions. Consequently, the part x_{IJ}, placed at the position $[1,1]$ in $\mathbf{x}^{(kl)}$, plays the role of normalizing part x_D, while the part x_{kl} (at the position [2,2] in the respective permutation) represents the pivoting part (denoted x_l in the vector case). There are in general $(I-1)(J-1)$ of such permutations, and then there is a coordinate system for each of them, following the idea of vector bpcs and coordinates defined in Fačevicová et al. (2016). It can be defined as follows.

The elemental information on the row factor is contained in the first coordinate from the set of *row backwards pivot balances* (rbpb)

$$\mathrm{rbpb}(\mathbf{x}^{(kl)})_i = \sqrt{\frac{iJ}{i+1}} \ln \frac{g(\mathbf{x}^{(kl)}_{i+1\bullet})}{\left(\prod_{m=1}^{i} g(\mathbf{x}^{(kl)}_{m\bullet})\right)^{1/i}}, \; i = 1,\ldots,I-1. \tag{5}$$

That is, it is in the pairwise row balance

$$\mathrm{rbpb}(\mathbf{x}^{(kl)})_1 = \sqrt{\frac{J}{2}} \ln \frac{g(\mathbf{x}^{(kl)}_{2\bullet})}{g(\mathbf{x}^{(kl)}_{1\bullet})} = \sqrt{\frac{J}{2}} \ln \frac{g(\mathbf{x}_{k\bullet})}{g(\mathbf{x}_{I\bullet})}. \tag{6}$$

Note that here $g(\mathbf{x}_{i\bullet})$ stands for the geometric mean of elements in the i-th row of a table. Using the geometric mean to represent the rows guarantees that rbpbs can be considered an orthogonal projection of the compositional table that accounts for relative information conveyed solely by the rows (Fačevicová et al., 2016). This involves further benefits, including the Pythagorean decomposition of the overall variability while respecting independence and interaction structure of the table. Similarly, $g(\mathbf{x}_{\bullet j})$ is used for the definition of *column backwards pivot balances* (cbpb) that concerns the inner structure of the column factor as

$$\mathrm{cbpb}(\mathbf{x}^{(kl)})_j = \sqrt{\frac{jI}{j+1}} \ln \frac{g(\mathbf{x}^{(kl)}_{\bullet j+1})}{\left(\prod_{n=1}^{j} g(\mathbf{x}^{(kl)}_{\bullet n})\right)^{1/j}}, \; j = 1,\ldots,J-1. \tag{7}$$

Thus, the pairwise column balance computed for each row-column permutation (kl) is given as

$$\mathrm{cbpb}(\mathbf{x}^{(kl)})_1 = \sqrt{\frac{I}{2}} \ln \frac{g(\mathbf{x}_{\bullet 2}^{(kl)})}{g(\mathbf{x}_{\bullet 1}^{(kl)})} = \sqrt{\frac{I}{2}} \ln \frac{g(\mathbf{x}_{\bullet l})}{g(\mathbf{x}_{\bullet J})} \tag{8}$$

and carries information on the ratio between the l-th column of interest and the normalizing column (the J-th column).

Typically, the main interest lies on the relationships between factors. For this purpose, the balances (5) and (7) are accompanied by the system of *odds-ratio* (referred to as *table*) *backwards pivot coordinates*

$$\mathrm{tbpc}(\mathbf{x}^{(kl)})_{ij} = \sqrt{\frac{ij}{(i+1)(j+1)}} \ln \frac{\left(\prod_{m=1}^{i}\prod_{n=1}^{j} x_{mn}^{(kl)}\right)^{1/ij} x_{i+1,j+1}^{(kl)}}{\left(\prod_{m=1}^{i} x_{m,j+1}^{(kl)}\right)^{1/i}\left(\prod_{n=1}^{j} x_{i+1,n}^{(kl)}\right)^{1/j}}, \tag{9}$$

for $i = 1,\ldots,I-1$ and $j = 1,\ldots,J-1$. The analysis is again focused on the first coordinate, the simple log odds-ratio, which here represents the interaction between the first two rows and columns of the respective permutation (kl), i.e.

$$\mathrm{tbpc}(\mathbf{x}^{(kl)})_{11} = \frac{1}{2} \ln \frac{x_{11}^{(kl)} x_{22}^{(kl)}}{x_{12}^{(kl)} x_{21}^{(kl)}} = \frac{1}{2} \ln \frac{x_{IJ} x_{kl}}{x_{Il} x_{kJ}}. \tag{10}$$

The proposed sets of coordinates (5), (7) and (9) can be analyzed also separately when only a specific source of information is of interest. Following Egozcue and Pawlowsky-Glahn (2008) any compositional table can be decomposed orthogonally onto its independent and interactive parts. In our setting, the former is represented by (5) and (7), while the latter is accounted for by (9). Finally, omitting the normalizing constants can also lead here to a more straightforward interpretation in e.g. regression analysis. These orthogonal alternatives to (5), (7) and (9) will be denoted by ${}_r\mathbf{z}(\mathbf{x}^{(kl)})$, ${}_c\mathbf{z}(\mathbf{x}^{(kl)})$ and ${}_t\mathbf{z}(\mathbf{x}^{(kl)})$ following on the notation introduced in the previous subsection.

To facilitate the understanding of the construction of coordinates (for both compositional vectors and tables), Supplementary Material B contains a thorough description of the process together with a graphical illustration.

3. Principal component analysis

Principal component analysis (PCA) is a well-known multivariate technique that aims to reduce the dimension of a data set through the construction of mutually orthogonal linear latent variables (principal components, PCs), which are fundamentally defined by matrices of loadings (coefficients of the linear combinations of the original variables) and scores (the values of the principal components) (Johnson and Wichern, 2007; Varmuza et al., 2009). The ordinary formulation of PCA and the associated biplot display for compositional data is based on clr coefficients (Aitchison and Greenacre, 2002). Following

on the strategy introduced in Kynčlová, Filzmoser and Hron (2016), backwards pivot coordinates can be used for compositional PCA by combining first coordinates from the $D-1$ bpc coordinate systems (3) (Hron et al., 2021). The final matrix of loadings is obtained as a combination of rows from all the $D-1$ loading matrices, picking out from each the row corresponding to the first bpc. This approach is feasible due to the fact that the resulting matrices of scores are the same for all $D-1$ bpc systems. Key features of this representation are the following: (1) the associated biplot display retains (visually) all the properties of ordinary clr-based PCA biplots; (2) the first coordinate of the l-th backwards pivot coordinate system conveys information equivalent to the analogous l-th alr coordinate; and (3) unlike a biplot based on oblique alr coordinates, the results are invariant to orthogonal rotations. In the following subsections we detail the technical formulation of bpc-based PCA biplots for vector compositional data and generalize it to compositional tables.

3.1. Preliminaries

Before we proceed, we briefly summarise the main elements of the theory of PCA and the biplot display that will be required later on. Given a real matrix $\mathbf{X}_{n\times D}$, where n is the number of observations and D stands for the number of variables (assumed to be centered). PCA is commonly performed through singular value decomposition (SVD):

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T, \tag{11}$$

where $\mathbf{U}_{n\times D}$ is an orthogonal matrix, $\mathbf{D}_{D\times k}$ is a matrix of singular values on diagonal, arranged in descending order (i.e. $d_{11} \geq d_{22} \geq \ldots d_{kk} > 0$, $k \leq \min(n,D)$) and $\mathbf{V}_{D\times k}$ is an orthogonal matrix of loadings. As noted above, the scores are the values of new PCs and the loadings are the coefficients describing the importance of each variable on each PC. The singular values in $\mathbf{D}$ are the standard deviations of the scores (Varmuza et al., 2009). The biplot display (Gabriel, 1971) results from an alternative decomposition of $\mathbf{X}$ as

$$\mathbf{X} = \mathbf{G}\mathbf{H}^T, \tag{12}$$

where $\mathbf{G} = \sqrt{n-1}\mathbf{U}$ and $\mathbf{H} = \frac{1}{\sqrt{n-1}}\mathbf{V}\mathbf{D}$. The rows of matrix $\mathbf{G}_{n\times D}$ carry information about the observations, while the information about the variables is contained in the rows of $\mathbf{H}_{D\times k}$. The biplot is commonly based on the first $k = 2$ PCs (those retaining most of the variability of the original data set) and, therefore, only the first two columns of the matrices $\mathbf{G}$ and $\mathbf{H}$ are used. Hence, equation (12) holds only approximately (unless $\mathbf{X}$ is of a rank two or less). This gives rise to the biplot graph where observations are represented by points (with coordinates determined by the PC scores) and variables are represented by arrows from the origin (with angle determined by the loadings vector).

In the compositional case, $\mathbf{X}$ is replaced by $\mathbf{Y} = \mathrm{clr}(\mathbf{X})$, which can be (after column-wise centering) decomposed as in (12). The matrix $\mathbf{G}$ still represents the observations in the PC system, whereas $\mathbf{H}$ now carries information about the importance of the clr coordinates. Following Aitchison and Greenacre (2002) and known properties of the

classical (non-compositional) biplot, general properties of the compositional (clr-)biplot are listed below (note that vectors with indexes $i\bullet$ and $\bullet j$ refer, respectively, to the i-th row and j-th column of a matrix; all vectors are considered as columns):

- The inner product between the rows of $\mathbf{G}$ and $\mathbf{H}$ (from the decomposition of $\mathbf{Y}$) approximates the clr coefficients:

$$\mathbf{g}_{i\bullet}^T \mathbf{h}_{j\bullet} \approx y_{ij} = \ln \frac{x_{ij}}{g(\mathbf{x}_{i\bullet})}$$

- The length of the rays approximate the variability of clr coefficients corresponding to the parts $\mathbf{x}_j$, $j = 1, \ldots, D$:

$$\|\mathbf{h}_{j\bullet}\|^2 \approx \widehat{\mathrm{var}}(\mathbf{y}_j) = \widehat{\mathrm{var}}\left(\ln \frac{\mathbf{x}_j}{g(\mathbf{x})}\right)$$

- The length of the links between the vertices of the rays approximate the variability of a pairwise logratio formed by the respective parts:

$$\|\mathbf{h}_{i\bullet} - \mathbf{h}_{j\bullet}\|^2 \approx \frac{1}{n-1}(\mathbf{y}_{\bullet i} - \mathbf{y}_{\bullet j})^T(\mathbf{y}_{\bullet i} - \mathbf{y}_{\bullet j}) = \widehat{\mathrm{var}}\left(\ln \frac{\mathbf{x}_i}{\mathbf{x}_j}\right)$$

- The projection of the i-th score onto a link approximates the difference between the respective clr coefficients (i.e. logratio between x_{ij} and x_{ik}):

$$\mathbf{g}_{i\bullet}^T(\mathbf{h}_{j\bullet} - \mathbf{h}_{k\bullet}) \approx \ln \frac{x_{ij}}{g(\mathbf{x}_{i\bullet})} - \ln \frac{x_{ik}}{g(\mathbf{x}_{i\bullet})} = \ln \frac{x_{ij}}{x_{ik}}$$

More elaborated derivations of the biplot properties (discussed both here and in the following sections) are available in Supplementary Material A.

3.2. Vector compositional data

The relationship between a pairwise logratio, i.e. the first bpc in the corresponding coordinate system, and clr coefficients (2) contained in $\mathbf{Y}$ can be written as

$$\mathrm{bpc}(\mathbf{x}^{(l)})_1 = \sqrt{\frac{1}{2}} \ln \frac{x_l}{x_D} = \sqrt{\frac{1}{2}} y_l - \sqrt{\frac{1}{2}} y_D = \sqrt{\frac{1}{2}} (y_l - y_D), \tag{13}$$

where x_D is the chosen rationing element. All $D-1$ coordinate systems need to be constructed to obtain all possible pairwise logratios with a chosen ratioing element. Let $\mathbf{Z}^{(l)}$ be a centered matrix containing the bpcs of the l-th system, $l = 1, \ldots, D-1$. The biplot decomposition of $\mathbf{Z}^{(l)}$ can be written as $\mathbf{Z}^{(l)} = \mathbf{G}^{(l)}\mathbf{H}^{(l)T}$. The matrices $\mathbf{G}^{(l)}$ are identical for all systems and in fact equal to the matrix $\mathbf{G}$ resulting from the clr version of PCA as defined above. The suggested biplot representation combines the first rows of

the loading matrices $\mathbf{H}^{(l)}$, corresponding to coordinates $\mathrm{bpc}(\mathbf{x}^{(l)})_1$ and, consequently, to pairwise logratios $\ln(x_l/x_D)$ of interest.

The interpretation of the bpc-based biplot can be derived from the relationship between the first row of a matrix $\mathbf{H}^{(l)}$ and the rows of the clr loading matrix $\mathbf{H}$. From (13) it follows that

$$\mathbf{h}_{1\bullet}^{(l)} = \sqrt{\frac{1}{2}}\mathbf{h}_{l\bullet} - \sqrt{\frac{1}{2}}\mathbf{h}_{D\bullet} = \sqrt{\frac{1}{2}}\left(\mathbf{h}_{l\bullet} - \mathbf{h}_{D\bullet}\right). \tag{14}$$

Given that $\mathbf{G}^{(l)} = \mathbf{G}$ and the relationship in (14), the main properties of the bpc-based biplot are as follows:

- The inner product between a row of the matrix $\mathbf{G}$ and $\mathbf{h}_{1\bullet}^{(l)}$ approximates the pairwise logratio between the l-th component and the ratioing part:

$$\mathbf{g}_{i\bullet}^T\mathbf{h}_{1\bullet}^{(l)} \approx \sqrt{\frac{1}{2}}\ln\frac{x_{il}}{x_{iD}}$$

- The lengths of biplot rays approximate the variability of the logratio between the l-th and the rationing part:

$$\|\mathbf{h}_{1\bullet}^{(l)}\|^2 \approx \frac{1}{2}\widehat{\mathrm{var}}\left(\ln\frac{\mathbf{x}_l}{\mathbf{x}_D}\right)$$

- The length of the links between the vertices of two rays approximate the variability of the logratio between the compositional parts placed in the numerator of the respective pairwise logratios:

$$\|\mathbf{h}_{1\bullet}^{(p)} - \mathbf{h}_{1\bullet}^{(s)}\|^2 \approx \frac{1}{2}\widehat{\mathrm{var}}\left(\ln\frac{\mathbf{x}_p}{\mathbf{x}_s}\right)$$

- The projection of a score onto a link approximates the pairwise logratio of the respective components:

$$\mathbf{g}_{i\bullet}^T(\mathbf{h}_{1\bullet}^{(p)} - \mathbf{h}_{1\bullet}^{(s)}) \approx \sqrt{\frac{1}{2}}\ln\frac{x_{ip}}{x_{is}}$$

Amongst the properties above, it is worth highlighting that lengths of the links between vertices approximate pairwise logratio variances, i.e. the elements of the variation matrix (Aitchison, 1986). Thus, the bpc-based biplot shares this property with the ordinary clr-based biplot, except that, for obvious reasons, now the pairwise logratios containing the ratioing part are not contained in the links. They can however be readily obtained from the lengths of the rays. This implies that the bpc-based biplot can be used as a reasonable alternative to the crude alr-based biplot which suffers from non-orthonormality of the input coordinates.

3.3. Compositional tables

Similarly to the case of vector compositional data, PCA can be modified to summarize the elemental information contained in a set of compositional tables. As noted in Section 2.2, this concerns the interactions between factors, as carried by simple log odds-ratios, and the within-factor structure, as carried by the pairwise balances computed from the entire rows and columns. The modified PCA using the bpc approach is like in the vector case given by the repeated computation of the decompositions (11) and (12), with the subsequent combination of the results from the different bpc systems. In each step, the data matrix $\mathbf{X}$ is replaced by a matrix of coordinates given by (5), (7) and (9). When the I-th row and the J-th column are understood as the normalizing categories, results from $(I-1)(J-1)$ partial computations (based on $(I-1)(J-1)$ permutations $\mathbf{x}^{(kl)}$) have to be combined.

From a methodological point of view, it would be possible to analyze the elemental information from both the independent and interactive parts of a compositional table simultaneously in one biplot. Nevertheless, in the following we focus on the case where these two types of information are displayed and interpreted separately. In our view, this leads to a more straightforward interpretation of the resulting biplots.

3.3.1. Inter-factorial relationships

The ordinary PCA can be modified so that the resulting biplot contains all possible simple log odds-ratios with the rationing element x_{IJ}. Namely, let $\mathbf{Z}^{(kl)}$, $k=1,\ldots,I-1$, $l=1,\ldots,J-1$ be a mean centered $n\times(IJ-1)$ matrix of coordinates from a sample of compositional tables ${}_i\mathbf{x}^{(kl)}$, $i=1,\ldots,n$ (note that in the following a left subscript will stand for the sample index), with the coordinates being ordered as $\mathrm{tbpc}(\mathbf{x}^{(kl)})$, $\mathrm{rbpb}(\mathbf{x}^{(kl)})$ and $\mathrm{cbpb}(\mathbf{x}^{(kl)})$. That is, the log odds-ratio $\ln\left(x_{kl}x_{IJ}/x_{kJ}x_{Il}\right)$ is represented by the first coordinate. When this matrix is decomposed as in (12), $\mathbf{Z}^{(kl)}=\mathbf{G}^{(kl)}\mathbf{H}^{(kl)T}$, the matrix $\mathbf{G}^{(kl)}$ remains the same for all row/column permutations and equals the one obtained for the clr coefficients ($\mathbf{G}$). For the biplot representation, the first row from each of the loading matrices $\mathbf{H}^{(kl)}$ is used, since it corresponds to the first tbpc (10). Each of these coordinates is related to the clr coefficients through

$$\mathrm{tbpc}(\mathbf{x}^{(kl)})_{11}=\frac{1}{2}\ln\frac{x_{IJ}x_{kl}}{x_{Il}x_{kJ}}=\frac{1}{2}\left(y_{[I,J]}+y_{[k,l]}-y_{[I,l]}-y_{[k,J]}\right), \tag{15}$$

where $y_{[i,j]}$ stands for the clr representation of the part x_{ij} of a compositional table $\mathbf{x}$, i.e. $y_{[i,j]}=\ln x_{ij}/g(\mathbf{x})$, $i=1,\ldots,I$ and $j=1,\ldots,J$. The first row of $\mathbf{H}^{(kl)}$ is therefore related to the rows of the clr loading matrix $\mathbf{H}$ as follows:

$$\mathbf{h}_{1\bullet}^{(kl)}=\frac{1}{2}\left(\mathbf{h}_{[I,J]\bullet}+\mathbf{h}_{[k,l]\bullet}-\mathbf{h}_{[I,l]\bullet}-\mathbf{h}_{[k,J]\bullet}\right), \tag{16}$$

with $\mathbf{h}_{[i,j]\bullet}$ denoting the row of $\mathbf{H}$ corresponding to the clr coefficient $y_{[i,j]}$.

Based on the aforementioned decomposition, the biplot properties in this case can be summarized as follows:

- The inner product between the rows of the score matrix $\mathbf{G}$ and the first row of $\mathbf{H}^{(kl)}$ approximates the respective simple log odds-ratio:

$$\mathbf{g}_{i\bullet}^T \mathbf{h}_{1\bullet}^{(kl)} \approx \frac{1}{2} \ln \frac{{}_i x_{IJ}\, {}_i x_{kl}}{{}_i x_{Il}\, {}_i x_{kJ}}.$$

- Due to the centering of the clr coordinate matrix $\mathbf{Y}$ the lengths of the biplot rays approximate the variability of the simple log odds-ratios:

$$\|\mathbf{h}_{1\bullet}^{(kl)}\|^2 \approx \frac{1}{4} \widehat{\mathrm{var}} \left(\ln \frac{x_{IJ} x_{kl}}{x_{Il} x_{kJ}} \right).$$

- The length of the links between two vertices can in general be understood in terms of difference between variation of the respective simple log odds-ratios. Moreover, a more convenient interpretation is available for some combinations. Thus, the links between rays related to odds-ratios sharing two common elements approximate the variation of a new simple log odds-ratio. Considering e.g. the log odds-ratios sharing the elements at positions $[k,J]$ and $[I,J]$, and differing in the column permutation index l (represented by $\mathbf{h}_{1\bullet}^{(kl_1)}$ and $\mathbf{h}_{1\bullet}^{(kl_2)}$, $l_1 \neq l_2$), the distance between the corresponding rays verifies that

$$\|\mathbf{h}_{1\bullet}^{(kl_1)} - \mathbf{h}_{1\bullet}^{(kl_2)}\|^2 \approx \frac{1}{4} \widehat{\mathrm{var}} \left(\ln \frac{x_{kl_1} x_{Il_2}}{x_{Il_1} x_{kl_2}} \right).$$

  A similar derivation can be given for odds-ratios sharing the same column indices, i.e. $x_{k_1 l} x_{IJ} / x_{k_1 J} x_{Il}$ and $x_{k_2 l} x_{IJ} / x_{k_2 J} x_{Il}$ for $k_1 \neq k_2$, where the corresponding link estimates the variance of $\ln(x_{k_1 l} x_{k_2 J} / x_{k_1 J} x_{k_2 l})$. Consequently, when a biplot collects results from all coordinate systems defined for a fixed rationing part x_{IJ}, it preserves also the information on the variability of the odds-ratios containing parts either from the I-th row or the J-th column. On the other hand, the characteristics of the other odds-ratios are not represented in this setting.

- In the case of the projection of a score to a link, only pairs of odds-ratios sharing elements from the same row or column are worth investigating. E.g. for the same pair as considered in the previous point, it holds that

$$\mathbf{g}_{i\bullet}^T (\mathbf{h}_{1\bullet}^{(kl_1)} - \mathbf{h}_{1\bullet}^{(kl_2)}) \approx \frac{1}{2} \ln \frac{{}_i x_{kl_1}\, {}_i x_{Il_2}}{{}_i x_{Il_1}\, {}_i x_{kl_2}}.$$

Interestingly, the links (partially) approximate the elemental covariance structure, represented here by variances of simple log odds-ratios. This property, which was also observed for the bpc biplot in the vector case, is thus transferred to a more general setting by a proper choice of coordinate representation of the interaction part of the compositional table.

3.3.2. Intra-factorial relationships

The other main source of information carried by a compositional table lies in its independent part, whose elemental representation is given by pairwise row or column balances.

When the I-th row of a compositional table $\mathbf{x}$ is understood as the normalizing element (category of the row factor), all pairwise row balances with this element can be given by the first rbpb, computed from different permutations $\mathbf{x}^{(kl)}$, $k = 1, \ldots, I-1$. The following relationship holds between these coordinates and the clr representation:

$$\mathrm{rbpb}(\mathbf{x}^{(kl)})_1 = \sqrt{\frac{J}{2}} \ln \frac{g(\mathbf{x}_{k\bullet})}{g(\mathbf{x}_{I\bullet})} = \sqrt{\frac{1}{2J}} \left(y_{[k,1]} + \ldots + y_{[k,J]} - y_{[I,1]} - \ldots - y_{[I,J]} \right). \tag{17}$$

The construction of a biplot with pairwise row balances is based again on the decomposition of the mean centered coordinate matrix $\mathbf{Z}^{(kl)}$ into matrices $\mathbf{G}^{(kl)}$ and $\mathbf{H}^{(kl)}$, so it can be computed along with the one for the interaction part. Here the biplot representation collects the rows of the loading matrices $\mathbf{H}^{(kl)}$, $k = 1, \ldots, I-1$, standing at the position $[(I-1)(J-1)+1]$ (the column permutation l does not play a relevant role in this case). Let the position be denoted by $r1$ in the following (in reference to the fact that it is the position corresponding to the first rbpb). According to (17) the following relation holds between a row of $\mathbf{H}^{(kl)}$ and rows of the clr loading matrix $\mathbf{H}$:

$$\mathbf{h}^{(kl)}_{r1\bullet} = \sqrt{\frac{1}{2J}} \left(\mathbf{h}_{[k,1]\bullet} + \ldots + \mathbf{h}_{[k,J]\bullet} - \mathbf{h}_{[I,1]\bullet} - \ldots - \mathbf{h}_{[I,J]\bullet} \right). \tag{18}$$

Accordingly, the biplot properties are now as follows:

- The inner product between a row of matrix $\mathbf{G}$ and $\mathbf{h}^{(kl)}_{r1\bullet}$ approximates the pairwise row balance:

$$\mathbf{g}^T_{i\bullet} \mathbf{h}^{(kl)}_{r1\bullet} \approx \sqrt{\frac{J}{2}} \ln \frac{g\left({}_i\mathbf{x}_{k\bullet}\right)}{g\left({}_i\mathbf{x}_{I\bullet}\right)}.$$

- The lengths of the biplot rays give an approximation of the variability of the pairwise row balances:

$$\|\mathbf{h}^{(kl)}_{r1\bullet}\|^2 \approx \frac{J}{2} \widehat{\mathrm{var}} \left(\ln \frac{g(\mathbf{x}_{k\bullet})}{g(\mathbf{x}_{I\bullet})} \right).$$

- The length of the links between the vertices of the rays approximate the variability of the balance between row categories standing in the numerator of the respective coordinates:

$$\|\mathbf{h}^{(k_1 l)}_{r1\bullet} - \mathbf{h}^{(k_2 l)}_{r1\bullet}\|^2 \approx \frac{J}{2} \widehat{\mathrm{var}} \left(\ln \frac{g(\mathbf{x}_{k_1\bullet})}{g(\mathbf{x}_{k_2\bullet})} \right).$$

- A specific pairwise row balance can be for a given observation approximated by the projection to a link:

$$\mathbf{g}_{l\bullet}^T \left(\mathbf{h}_{r1\bullet}^{(k_1 l)} - \mathbf{h}_{r1\bullet}^{(k_2 l)} \right) \approx \sqrt{\frac{J}{2}} \ln \frac{g\left({}_i\mathbf{x}_{k_1\bullet}\right)}{g\left({}_i\mathbf{x}_{k_2\bullet}\right)}.$$

Analogous properties can be derived for the biplot constructed from the rows at the positions $[(I-1)J+1]$ of matrices $\mathbf{H}^{(kl)}$, computed for different column permutations $l = 1, \ldots, J-1$. These rows refer to the first cbpbs of $\mathbf{x}^{(kl)}$, and therefore the interpretation relates to the pairwise column balances with the J-th (rationing) column level or, alternatively, between the other levels of the column factor.

4. Regression analysis

We now demonstrate the bpc approach focusing on the elemental information in vector- or table-type compositional data in the context of regression analysis. More specifically, we focus on linear regression models with real-valued response variable and explanatory composition. The coordinate representation of the composition challenges the interpretation of regression parameters. As Coenders et al. (2020) pointed out, the standard interpretation of regression coefficients in terms of "increasing one regressor while keeping the others constant" is violated when non-orthonormal coordinates are used. We add that, even when an olr coordinate system is used, the idea of keeping regressors constant needs to be understood correctly. In this section, following Hron et al. (2021), we elaborate on the interpretation of regression parameters in the vector composition case using a bpc representation to, subsequently, extend the concept to compositional tables.

4.1. Vector compositional data

We focus on the regression analysis problem where a real-valued variable Y is modelled in terms of a D-part compositional vector $\mathbf{x} = (x_1, \ldots, x_D)$. Thus, a system of (orthogonal) bpcs $\mathbf{z}(\mathbf{x}^{(l)}) = (z_1^{(l)}, \ldots, z_{D-1}^{(l)})$ can be used to represent the composition in a regression model as

$$E\left[Y|\mathbf{z}(\mathbf{x}^{(l)})\right] = \beta_0 + \beta_1^{(l)} z_1^{(l)} + \ldots + \beta_{D-1}^{(l)} z_{D-1}^{(l)}. \tag{19}$$

Even though the definition of bpcs in (3) refers to the usual natural logarithm, i.e. logarithm with the base of e, following Müller et al. (2018) this can be replaced by logarithms with any other base κ. An adequate choice of the base of the logarithm can facilitate the interpretation of the regression parameters.

While the intercept β_0 and global characteristics of the regression model such as residuals, overall F-statistic, and coefficient of determination remain the same in any coordinate system, all the other regression coefficients vary with the choice of logratio basis. When bpcs with x_D as the normalizing part are of interest, the effects of the

simple logratios $\ln(x_l/x_D)$ can be investigated by changing the index l ($l = 1,\ldots,D-1$), each represented by the corresponding regression parameter $\beta_1^{(l)}$ associated to the first bpc coordinate $z_1^{(l)}$ of the system. Therefore, the main result from such a modelling is formed by estimates $\hat{\beta}_0$ and $\hat{\beta}_1^{(l)}$, $l = 1,\ldots,D-1$.

Regardless the specific permutation (l), β_0 gives an expected value of the response when all compositional parts x_i, $i = 1,\ldots,D$, are equal. However, the interpretation of the $\beta_1^{(l)}$, $l = 1,\ldots,D-1$, is related to the underlying model and coordinate system it is coming from. For a given l the parameter $\beta_1^{(l)}$ estimates the effect of the κ-times growth of the ratio x_l/x_D on the response, while keeping the remaining coordinates $z_2^{(l)},\ldots,z_{D-1}^{(l)}$ constant. For example, using $\kappa = 2$ leads to the interpretation in terms of doubling the ratio of interest. In order to keep the remaining coordinates unaffected, the κ-times increase in x_l/x_D has to be equally distributed between both parts in the ratio. More specifically, the only scenario leading to the required change is the $\sqrt{\kappa}$-times increase in x_l accompanied by the same decrease in x_D. Each of the regression parameters $\beta_1^{(l)}$, $l = 1,\ldots,D-1$, therefore models the effect of the increase in the part of interest x_l at the expense of the rationing part x_D, while keeping the rest of the composition constant.

Obviously, there are several other ways to achieve the κ-times increase in the ratio involving x_l and x_D. Particularly interesting is the case when the increase is caused by a κ-times increase in x_l only. Even though this does not affect any other pairwise logratio with the normalizing part x_D, it leads to change in the remaining coordinates $z_2^{(l)},\ldots,z_{D-1}^{(l)}$ from the l-th system, which decrease by $1/2, 1/3, \ldots, 1/(D-1)$ respectively. The effect of the κ-times increase of x_l would therefore need to be accounted for from all the regression coefficients $\beta_i^{(l)}$, $i = 1,\ldots,D-1$.

4.2. Compositional tables

For the purpose of regression modeling, a table $\mathbf{x}$ of dimensions $I \times J$ can be represented by a vector of olr coordinates (Fačevicová et al., 2021). When, additionally, the elemental information related to the I-th row and the J-th column is of interest, such coordinates need to be constructed as in Section 2.2. The explanatory variables can be the orthogonal versions of the row and column backwards pivot balances (5) and (7) and the odds-ratio bpcs (9), i.e. ${}_r\mathbf{z}(\mathbf{x}^{(kl)})$, ${}_c\mathbf{z}(\mathbf{x}^{(kl)})$ and ${}_t\mathbf{z}(\mathbf{x}^{(kl)})$ respectively. As noted in the previous section, the natural logarithm can be replaced by any other with base κ. A linear model of dependence between a real variable Y and a $I \times J$ compositional table $\mathbf{x}$ can therefore be formulated as

$$\begin{aligned} E\left[Y|\mathbf{z}(\mathbf{x}^{(kl)})\right] = \beta_0 &+ {}_r\beta_1^{(kl)}\,{}_rz_1^{(kl)} + \ldots + {}_r\beta_{I-1}^{(kl)}\,{}_rz_{I-1}^{(kl)} + \\ &+ {}_c\beta_1^{(kl)}\,{}_cz_1^{(kl)} + \ldots + {}_c\beta_{J-1}^{(kl)}\,{}_cz_{J-1}^{(kl)} + \\ &+ {}_t\beta_{11}^{(kl)}\,{}_tz_{11}^{(kl)} + \ldots + {}_t\beta_{I-1,J-1}^{(kl)}\,{}_tz_{I-1,J-1}^{(kl)}. \end{aligned} \tag{20}$$

The parameter β_0 does not depend on the olr basis chosen and it has the usual interpretation, i.e. it is expected value of the response when all the regressors are set to

zero. This corresponds to the case of no relationship between row and column factors (i.e. ${}_t\mathbf{z}(\mathbf{x}^{(kl)})$ being zero) and no informative categorization of the individual factors (i.e. ${}_r\mathbf{z}(\mathbf{x}^{(kl)})$ and ${}_c\mathbf{z}(\mathbf{x}^{(kl)})$ being zero). The value and interpretation of the remaining regression parameters depend on the coordinate system from which they are estimated. As we focus on the elemental sources of information, we suggest to combine the results of the regression modeling (20) for each of the row-column permutations of the table $\mathbf{x}$, i.e. $\mathbf{x}^{(kl)}$, with $k = 1,\ldots,I-1$ and $l = 1,\ldots,J-1$. As in the vector case in Section 4.1, the global characteristics of the model do not depend on the specific permutation applied. The main part of the outcome is then formed by the estimates of coefficients corresponding to the orthogonal versions of the first rbpbs ${}_r\hat{\beta}_1^{(kl)}$, $k = 1,\ldots,I-1$, the first cbpbs ${}_c\hat{\beta}_1^{(kl)}$, $l = 1,\ldots,J-1$ and the first odds-ratio backwards pivot coordinates ${}_t\hat{\beta}_{11}^{(kl)}$, $\forall k,l$.

The final summary of the model collects estimates of $I-1$ regression coefficients ${}_r\beta_1^{(kl)}$, which quantify the effect of a unit change of the ratios

$$ {}_rz_1^{(kl)} = \ln \frac{g(\mathbf{x}_{k\bullet})}{g(\mathbf{x}_{I\bullet})}, \quad k = 1,\ldots,I-1, \tag{21} $$

on the response variable Y while keeping the remaining ones constant. But, as we are actually combining results from several regression models, that refers to the other coordinates included in the respective model and not to those listed in the final summary. The logratios (21) can be understood as a special case of the pairwise logratios studied in Section 4.1. Thus, the interpretation of ${}_r\beta_1^{(kl)}$ is analogous. The unit increase in ${}_rz_1^{(kl)}$ means a κ-times increase in the ratio $g(\mathbf{x}_{k\bullet})/g(\mathbf{x}_{I\bullet})$. This is achieved through a proportional $\sqrt{\kappa}$-times increase in parts from the k-th row accompanied by a proportional decrease in parts from the rationing row I by the same constant.

The second set of regression coefficients obtained from the regression analysis are the ${}_c\beta_1^{(kl)}$, which analogously to the above quantify the effect of a unit change of

$$ {}_cz_1^{(kl)} = \ln \frac{g(\mathbf{x}_{\bullet l})}{g(\mathbf{x}_{\bullet J})}, \quad l = 1,\ldots,J-1, \tag{22} $$

on the response variable Y. The κ-times increase in the respective ratio can be (under the condition of unaltered remaining coordinates) achieved here through a proportional $\sqrt{\kappa}$-times increase in parts from the l-th column accompanied by a proportional decrease in parts from the J-th column by the same constant.

The last group of regression coefficients are the ${}_t\beta_{11}^{(kl)}$. These $(I-1)(J-1)$ coefficients quantify the effect of changes in the interactive structure of the compositional table, as each of them is related to one log odds-ratio of the form

$$ {}_tz_{11}^{(kl)} = \ln \frac{x_{IJ}x_{kl}}{x_{Il}x_{kJ}}. \tag{23} $$

A κ-times increase in this case means a $\sqrt[4]{\kappa}$-times increase in the parts x_{IJ} and x_{kl} accompanied by a simultaneous $\sqrt[4]{\kappa}$-times decrease in x_{Il} and x_{kJ}. Note that even though

this ensures that the remaining coordinates from the system remain unchanged, it still affects some of the elemental odds-ratios with rationing part x_{IJ}. More specifically, the following happens:

- The odds-ratios sharing the pair of elements x_{IJ} and x_{kJ} or x_{IJ} and x_{Il}, i.e.

$$\frac{x_{IJ}x_{kj}}{x_{Ij}x_{kJ}} \quad \text{or} \quad \frac{x_{IJ}x_{il}}{x_{iJ}x_{Il}}, \quad i=1,\ldots,I-1, \quad i\neq k \quad \text{and} \quad j=1,\ldots,J-1, \quad j\neq l,$$

  turn out to increase $\sqrt{\kappa}$ - times.

- The odds-ratios sharing with the one of interest only the part x_{IJ},

$$\frac{x_{IJ}x_{ij}}{x_{Ij}x_{iJ}}, \quad i=1,\ldots,I-1, \quad i\neq k \quad \text{and} \quad j=1,\ldots,J-1, \quad j\neq l,$$

  increase $\sqrt[4]{\kappa}$-times.

Similarly to the vector composition case, the unit increase in the coordinate of interest admits an alternative interpretation. If the condition that the other coordinates must remain unchanged is relaxed, we can consider the case in which only the part of the interest x_{kl} observes a κ-times increase. This implies that the other coordinates from the system are affected and the overall effect on the response variable is therefore a combination of all regression parameters for the (kl) model. In particular, the row backwards pivot balances ${}_rz_i^{(kl)}$, $i=2,\ldots,I-1$, decrease by $1/iJ$, while ${}_rz_1^{(kl)}$ increases by $1/J$. Similarly, the column backwards pivot balances ${}_cz_j^{(kl)}$, $j=2,\ldots,J-1$, decrease by $1/jI$ and ${}_cz_1^{(kl)}$ is increased by $1/I$. Finally, the odds-ratio backwards pivot coordinates ${}_tz_{1j}^{(kl)}$, $j=2,\ldots,J-1$, or ${}_tz_{i1}^{(kl)}$, $i=2,\ldots,I-1$, decrease by $1/j$ and $1/i$ respectively. The remaining ${}_tz_{ij}^{(kl)}$, $i=2,\ldots,I-1$ and $j=2,\ldots,J-1$, are increased by $1/ij$.

Another interesting case of a unit change in ${}_tz_{11}^{(kl)}$ is that derived from change in one of the odds constituting the corresponding odds-ratio (while keeping the other unchanged). Thus, when for example the odd x_{kJ}/x_{IJ} decreases κ-times in a way such that x_{IJ} becomes $\sqrt{\kappa}x_{IJ}$ and x_{kJ} decreases proportionally to $x_{kJ}/\sqrt{\kappa}$, this change propagates to coordinates ${}_tz_{1j}^{(kl)}$, $j=1,\ldots,J-1$, which increase by $1/j$, and to the row balance ${}_rz_1^{(kl)}$, which increases by $1/J$. The effect of such a change on the response variable is therefore given by a combination of the parameters ${}_t\beta_{1j}^{(kl)}$, $j=1,\ldots,J-1$, and ${}_r\beta_1^{(kl)}$.

5. Illustrative applications

The presented methodology is illustrated in the following subsections using two real-world data sets from the field of time-use epidemiology, where compositional methods have been notably introduced in recent years (see e.g. (Dumuid et al., 2020)). The

data analyses were conducted using R (R Core Team, 2020). All methods introduced in this manuscript are implemented in the R package `robCompositions` (Templ, Hron and Filzmoser, 2011), with associated functions called `bpc`, `bpcPca`, `bpcReg`, `bpcTabWrapper`, `bpcTabPca` and `bpcTabReg`.

5.1. Movement behavior patterns in children and adolescents — a vector approach

The first data set describes the distribution of 24-hour movement behaviors of 336 children and adolescents aged between 8 and 18 (Gába et al., 2021). For each of the participants the times spent in *sedentary behavior* (SB), *light physical activity* (LPA), *moderate physical activity* (MPA) and *vigorous physical activity* (VPA) were collected together with sleep time using wrist-worn accelerometers. Note that sleep is a natural ratioing part, so we will focus on the four pairwise logratios representing time spent in SB, LPA, MPA and VPA behaviors relative to sleep.

5.1.1. Principal component analysis

Compositional PCA and associated biplot based on backwards pivot coordinates as described in Section 3.2 are applied here to the movement composition, focusing on the representation of pairwise logratios including sleep as reference behavior. The resulting biplot display is shown in Figure 1, where it can be observed that the data variability is mainly driven by the VPA-to-sleep logratio. The individuals (biplot points) are represented by their ID number in the database. Thus, while participant no. 1287 spent substantially more time sleeping than in VPA (the raw values are VPA $= 0.12$ min/day and sleep $= 478.35$ min/day), for example participant no. 1656 (at the opposite side of the biplot) reported a higher absolute (and relative) amount of VPA (VPA = 29.14 min/day, sleep = 497.66 min/day). Another important source of variability is the logratio between MPA and sleep, while the relative time spent in SB or LPA is rather consistent in comparison to the former two logratios. Looking at the links, the logratio between VPA and MPA also stands out as relevant source of variation between participants, followed by logratios between VPA and SB, VPA and LPA or MPA and SB.

To illustrate possible caveats of the ordinary alr approach (considering sleep as reference element as in the case of backwards pivot coordinates), Figure 2 shows the resulting alr-based biplot. Two main issues are noticeable. Firstly, the scores of this biplot are distorted with respect to the biplot based on backwards pivot coordinates. Although the latter could be closely approximated using alr coordinates through an adequate choice of reference element (Greenacre, 2018), this option is not feasible when the reference element is chosen to have a concrete interpretation in the context of the problem at hand (as it is the case here). Secondly, and related to the above, the loadings are also dramatically different to those obtained using backwards pivot coordinates, particularly showing an exaggerated variability of the logratio between VPA and sleep. This variability is better represented relative to the other elements of the multivariate structure when orthonormal coordinates as seen in Figure 1.

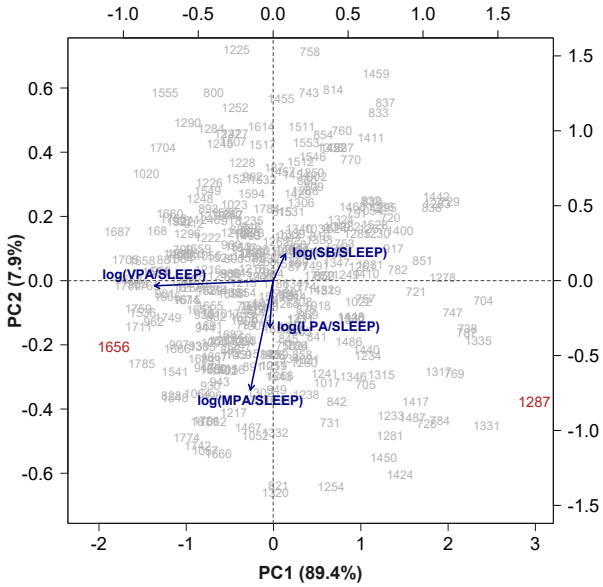

Figure 1. *PCA biplot based on backwards pivot coordinates for the movement behavior vector composition with sleep used as reference behavior. The numbers correspond to participant ID. Individuals no. 1287 and no. 1656 (highlighted in red color) differ substantially in time spent in VPA relative to sleep.*

5.1.2. Regression

The above movement distribution was accompanied by information on body fat percentage as response variable. Regression analysis as described in Section 4 was conducted to investigate their relationships. Namely, four regression models were required, each concerning the association with the response variable of each of the pairwise logratios with sleep as rationing part. Moreover, the logarithmic base κ was set to 2 for interpretation in terms of doubling the respective ratios. Note that body fat percentage was also represented in $\log_2$-scale as commonly done in practice to symmetrize its distribution. The results from all four models are summarized in Table 1.

The value of parameter $\hat{\beta}_0$ indicates that the average body fat percentage of an individual, who distributes time fairly equally over all five behaviors, is approximately 18 % ($2^{\hat{\beta}_0} = 2^{4.169}$). The pairwise logratios to sleep do not seem to play an important role overall, with the exception of time in MPA relative to sleep that is statistically significant at the usual 5% significance level. This implies that doubling this ratio (a $\sqrt{2}$-times increase in MPA at the expense of a similar decrease in sleep) results in an increase in body fat percentage of more than one third ($2^{\hat{\beta}_1^{(3)}} = 2^{0.446} = 1.362$). Note that this trend represents an average behavior over the entire data set. As the participants in this study span a fairly wide age range, it is expectable that a more structured analysis, which is

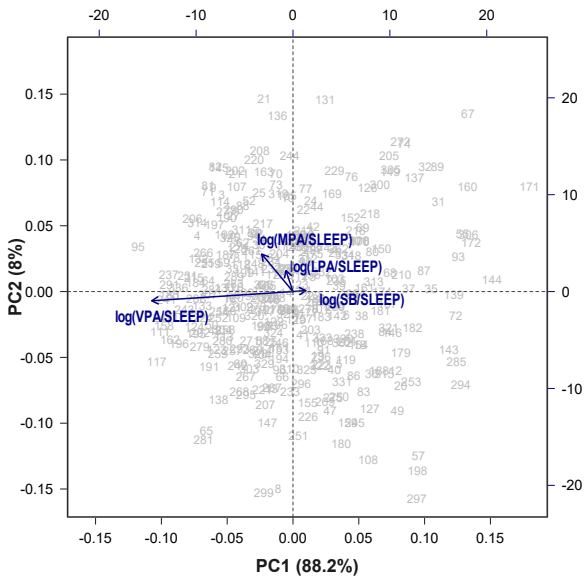

Figure 2. *PCA biplot based on alr coordinates for the movement behavior vector composition with sleep used as reference behavior. The numbers correspond to participant IDs.*

beyond the scope of this more methodologically-orientated manuscript, would lead to more specific results.

Following on Section 4.1, when only time dedicated to MPA doubles while all the remaining behaviors are unchanged (before closure), the overall effect on body fat percentage is derived from all the β coefficients in the respective model. In this case, they are the coefficients from $\beta_1^{(3)}$ to $\beta_4^{(3)}$ estimated for the third ($l = 3$) model, which are $(0.446, -0.036, -0.398, -0.193)$. Therefore, doubling the time spent in MPA is associated to an increase in body fat percentage by approximately one fifth ($2^{0.446-0.036/2-0.398/3-0.193/4} = 1.187$).

Finally, it can be compared to a simpler model that considers only the MPA-to-sleep logratio as explanatory variate, ignoring time devoted to other behaviors. This latter gives a markedly lower performance (adjusted $R^2 = 0.013$) and doubling the MPA-to-sleep ratio here implies a slight increase in body fat percentage ($\hat{\beta}_1 = 0.172$ with associated p-value equal to 0.020). The predicted body fat percentage when time spent in MPA is equal to sleep time is estimated to be 26 % ($\hat{\beta}_0 = 4.688$ with p-value < 0.001).

5.2. Movement behavior for older adults — a compositional table approach

The second data set focuses on the older adult population from a study conducted in 2016-2019 (Cuberek et al., 2019). The structure of movement behaviors during weekend waking time was assessed by hip-worn accelerometers and its association with visceral fat area (VFA) was studied based on 161 individuals aged between 60 and 84. For each

Table 1. *Regression analysis of body fat percentage on movement behavior composition. Summary of the four regression models needed to extract all orthogonal logratios with sleep as reference behavior. Common overall significance and R^2 measures: F-statistic $= 12.304$ (p-value < 0.001), $R^2 = 0.129$ and adjusted $R^2 = 0.119$.*

(l)	Variable	Estimate	Std. error	t-value	p-value
	Intercept	4.169	0.219	19.077	< 0.001
(1)	$\log_2(\text{SB}/\text{SLEEP})$	0.196	0.153	1.283	0.200
(2)	$\log_2(\text{LPA}/\text{SLEEP})$	-0.051	0.154	-0.335	0.738
(3)	$\log_2(\text{MPA}/\text{SLEEP})$	0.446	0.114	3.918	< 0.001
(4)	$\log_2(\text{VPA}/\text{SLEEP})$	0.027	0.097	0.278	0.781

participant, a two-factorial composition was available since activity (with categories SB, LPA and MVPA (moderate to vigorous physical activity aggregated)) was also split by part of the day (LM - *late morning* (9-12 am), N - *noon* (12 am - 3 pm), and A - *afternoon* (3-6 pm)). Moreover, each of the 3×3 compositional tables was accompanied by information about the visceral fat area (in cm^2) for the individual (see Table 2 for an example). Considering SB and LM as normalizing categories, the analysis focuses on (1) pairwise row balances between LPA or MVPA and SB, and (2) pairwise column balances between N or A and LM. Additionally, the interaction structure is studied focusing on the simple four-part log odds-ratios with [SB, LM] serving as the reference.

Table 2. *Example compositional table showing the distribution of movement behaviors during weekend (min./part of the day) for a senior individual presenting 71.36 cm^2 of visceral fat area.*

	LM	N	A
SB	110.5	73.0	38.5
LPA	65.5	88.5	105.5
MVPA	4.0	18.5	36.0

5.2.1. Principal component analysis

Figure 3 shows biplots for the row and column pairwise balances (left) and simple log odds-ratios (right) resulting from the four PCAs required here, each based on a permutation of the 3×3 compositional table $\mathbf{x}^{(kl)}$.

The left biplot suggests that the data variability is mostly driven by the MVPA-to-SB logratio (averaged over parts of the day). Even though SB time overall dominates MVPA time amongst participants, the more active ones are represented on the left-hand side of the biplot (e.g. participant no. 97159 has a paiwise row balance equal to -0.279).

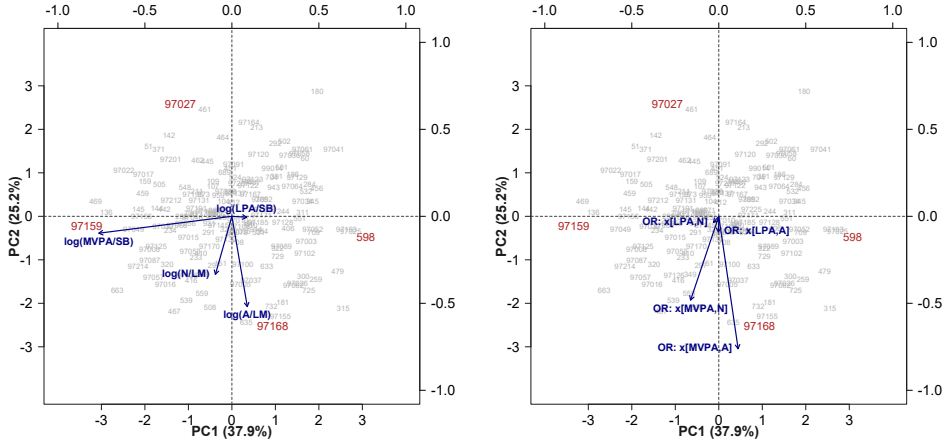

Figure 3. *PCA biplots based on backwards pivot coordinates for the movement behavior compositional table with sedentary behavior (SB) and late morning (LM) used as normalizing categories. The numbers correspond to participant ID. Participants further discussed in main text are highlighted in red.*

Contrarily, participants whose amount of MVPA time is relatively small (e.g. no. 598 has a pairwise row balance of -7.275) appear on the right-hand side. An important contribution to the overall variability is given by the simple log odds-ratio comparing the MVPA-to-SB ratio between afternoon and late morning. Thus, participants at the top of the right-hand side biplot tend to spend more time in MVPA (relatively to SB) in the late morning than in the afternoon (e.g. participant no. 97027, with log odds-ratio equal to -5.843). On the contrary, those at the bottom typically exhibit the opposite behavior (e.g. participant no. 97168 has log odds-ratio equal to 5.010). Moreover, the lengths of the rays indicate that a non-negligible variability involves the averaged A-to-LM and N-to-LM logratios (i.e. there is a good deal of variability in movement behavior during the day) and the log odds-ratio comparing the MVPA-to-SB ratios between noon and late morning. It can be observed that all logratios including LPA are fairly stable across participants. Thus, the LPA-to-SB logratio is fairly stable during the day and, considering the results from the regression analysis, change in LPA could be related with reduction in fat. Finally, looking at the link between the vertices of the odds-ratios including MVPA, we can conclude that the odds-ratio comparing the MVPA-to-SB ratio between afternoon and noon is markedly variable.

5.2.2. Regression

The relationship between amount of visceral fat and structure of weekend activities can be studied using a regression model of the form (20) as introduced in Section 4.2. Four permutations $\mathbf{x}^{(kl)}$ were considered for each individual table, where $k = 1,2$ distinguishes whether either LPA ($k = 1$) or MVPA ($k = 2$) is set as second row, and $l = 1,2$ refers to permutations where N ($l = 1$) or A ($l = 2$) is placed in the second column. For each

permutation, a system of orthogonal coordinates $\mathbf{z}\left(\mathbf{x}^{(kl)}\right)$ was computed following (5), (7) and (9) . They were used as covariates in the regression model. As in the previous example, a base $\kappa = 2$ was set for the logarithm and the response variable VFA was log-transformed to meet distributional assumptions. The results from all four models are summarized in Table 3. Note that, e.g., notation $\log_2(\mathrm{g(MVPA)/g(SB)})$ refers to the pairwise row balance between MVPA and SB, while $\log_2$OR: x[MVPA, N] denotes simple log odds-ratio with the pivoting element at the position [MVPA, N].

Table 3. *Regression analysis of visceral fat area on weekend movement behavior composition in adults according to intensity and part of a day. Summary of the four regression models needed with sedentary behavior (SB) and late morning (LM) as normalizing categories. Common overall significance and R^2 measures: F-statistic $= 2.943$ (p-value $= 0.004$), $R^2 = 0.134$, adjusted $R^2 = 0.089$ (see text for details).*

(kl)	Covariate	Estimate	Std. error	t-value	p-value
	Intercept	6.246	0.126	49.566	< 0.001
(1•)	$\log_2(\mathrm{g(LPA)/g(SB)})$	-0.029	0.048	-0.607	0.545
(2•)	$\log_2(\mathrm{g(MVPA)/g(SB)})$	-0.105	0.035	-2.987	0.003
(•1)	$\log_2(\mathrm{g(N)/g(LM)})$	0.068	0.157	0.436	0.664
(•2)	$\log_2(\mathrm{g(A)/g(LM)})$	-0.069	0.131	-0.530	0.597
(11)	$\log_2$OR: x[LPA, N]	-0.088	0.047	-1.878	0.062
(12)	$\log_2$OR: x[LPA, A]	-0.063	0.039	-1.626	0.106
(21)	$\log_2$OR: x[MVPA, N]	-0.066	0.042	-1.562	0.120
(22)	$\log_2$OR: x[MVPA, A]	-0.024	0.035	-0.667	0.506

The results suggest that VFA is mostly related to the MVPA-SB ratio, with the corresponding regression coefficient being ${}_r\hat{\beta}_1^{(2\bullet)} = -0.105$. That is, doubling the average MVPA-SB ratio is related to a decrease in VFA by approximately 7% $(1 - 2^{-0.105})$. Change in the respective row balance is considered under the condition of constant remaining coordinates. Therefore, it has to happen across the whole day by simultaneous $\sqrt{2}$-increase of time spent in MVPA at the expense of SB. Alternatively, we can consider a 2-time increase in MVPA over the day, without simultaneous decrease in SB nor LPA and before closure. Similarly to the vector case, such a change affects the second rbpb and, thus, the overall effect on VFA is $2^{\left({}_r\hat{\beta}_1^{(2\bullet)} - \frac{1}{2}\,{}_r\hat{\beta}_2^{(2\bullet)}\right)} = 2^{(-0.105+0.031/2)} = 0.940$ (approx. 6% decrease). If LPA and the interactions were ignored, the simple model of VFA on the MVPA-to-SB balance would estimate an effect of -0.120, leading to a similar conclusion: doubling the mean MVPA time (with respect to SB without any other condition) is associated with a decrease in VFA of about 8 %. Note that the overall performance of this simpler model is also fairly poor (adjusted $\mathrm{R}^2 = 0.087$).

Another interesting regression coefficient is ${}_t\hat{\beta}_{11}^{(11)}$, which quantifies the effect of change in the odds-ratio comparing the LPA-to-SB ratio at noon and late morning. Even though the estimate is marginally non-significant at the usual 5% significance level (p-value = 0.062), its value suggests that doubling the odds-ratio decreases VFA by more than five percent ($2^{-0.088} = 0.941$). This change has to be proportionally distributed over all parts in the odds-ratio, meaning that while time spent in LPA at noon and in SB at late morning increases $\sqrt[4]{2}$-times, this is at the expense of time devoted to the same behaviors at the complementary part of a day. It can be therefore understood as a transfer of LPA time from late morning to noon, compensated by a transfer of SB time in the opposite direction. Alternative scenarios as discussed in Section 4.2 can be considered. For instance, when the change affects only the late morning LPA-to-SB ratio (decreasing in a half by reducing LPA time at the favor of SB time), the effect on VFA is equal to $2^{\left({}_t\hat{\beta}_{11}^{(11)}+\frac{1}{2}\,{}_t\hat{\beta}_{12}^{(11)}+\frac{1}{3}\,{}_r\hat{\beta}_{1}^{(11)}\right)}$, i.e. $2^{(-0.088-0.026/2-0.029/3)} = 0.926$ (a 7% decrease approx.). Finally, the simpler model of VFA on the discussed log odds-ratio gives an estimated β coefficient equal to -0.101 (p-value $= 0.024$). Therefore, when the other covariates are neglected, doubling the odds-ratio comparing the LPA-to-SB ratio at noon and late morning relates to a decrease in VFA by approximately 7%.

6. Final remarks

Some recent developments in compositional data analysis suggest that there is a demand amongst practitioners for simple, interpretable logratio representations of compositional data. The classic alr coordinates, although having some issues related to the fact that they define an oblique system of coordinates, are indisputably a key representative of this kind. Moreover, orthonormality of logratio coordinates is a desirable property which is very much linked to the Aitchison geometry of compositional data, contributing to guarantee consistent and reliable results. In this paper we present backwards pivot coordinates as an orthonormal alternative to alr coordinates. It is demonstrated how they can be used with widely-used techniques such as principal component analysis and linear regression analysis. Just taking into account that the results are originated from multiple coordinate representations simultaneously, the interpretation results to be simple and natural while orthonormality is satisfied. Additionally, the approach is extended in this contribution to the case of compositional tables, where orthonormality of coordinates is required to enable (orthogonal) decomposition into independent and interactive parts.

We then consider that the approach in the present work opens up new possibilities in compositional data analysis, offering simplicity of interpretation while respecting the well-established geometrical framework for both vector compositional data and multifactorial compositions. Computer implementations of the methods are made freely available to facilitate use by practitioners on the R software for statistical computing.

Acknowledgements

This work was supported by the project Nr. DSGS-2021-0141 *Analysis of multifactorial relative data* established within the framework of the project OP RDE: *Improving schematics of Doctoral student grant competition and their pilot implementation*, CZ.02.2.69/0.0/0.0/19_073/0016713 [to VN, PJ and IP], the Czech Science Foundation (Project 22-15684L) [to KF and KH], (Projects 18-09188S and 22-02392S) [to AG], (Project 22-02392S) [to JP] and the Spanish Ministry of Science and Innovation (MCIN/AEI/10.13039/501100011033) and ERDF A way of making Europe (Grant PID2021-123833OB-I00) [to KH and JP-A].

References

Agresti, A. (2013). *Categorical Data Analysis.* John Wiley & Sons, Hoboken.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Chapman and Hall, London.

Aitchison, J. and Greenacre, M. (2002). Biplots of compositional data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 51(4), 375–392.

Coenders, G. and Pawlowsky-Glahn, V. (2020). On interpretations of tests and effect sizes in regression models with a compositional predictor. *SORT-Statistics and Operations Research Transactions*, 44(1), 201–220.

Cuberek, R., Pelclová, J., Gába, A., Pechová, J., Svozilová, Z., Přidalová, M., Štefelová, N. and Hron, K. (2019). Adiposity and changes in movement-related behaviors in older adult women in the context of the built environment: a protocol for a prospective cohort study. *BMC Public Health*, 19(1), 1–7.

Dumuid, D., Pedišič, Ž., Palarea-Albaladejo, J., Martín-Fernández, J. A., Hron, K. and Olds, T. (2020). Compositional Data Analysis in Time-Use Epidemiology: What, Why, How. *International Journal of Environmental Research and Public Health*, 17(7), 2220. doi:10.3390/ijerph17072220

Egozcue, J.J. and Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37(7), 795–828.

Egozcue, J.J. and Pawlowsky-Glahn, V. (2008). Compositional analysis of bivariate discrete probabilities. In: Daunis-i Estadella, J. and Martín-Fernández, J. A. (editors). *Proceedings of CODAWORK'08, The 3rd Compositional Data Analysis Workshop*. University of Girona, Spain.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3), 279–300.

Egozcue, J.J., Pawlowsky-Glahn, V., Templ, M. and Hron, K. (2015). Independence in contingency tables using simplicial geometry. *Communications in Statistics - Theory and Methods*, 44(18), 3978-3996.

Fačevicová, K., Filzmoser, P. and Hron, K. (2023). Compositional cubes: a new concept for multi-factorial compositions, *Statistical Papers*, 64(3), 955-985.

Fačevicová, K., Hron, K., Todorov, V. and Templ, M. (2016). Compositional tables analysis in coordinates. *Scandinavian Journal of Statistics*, 43(4), 962–977.

Fačevicová, K., Hron, K., Todorov, V. and Templ, M. (2018). General approach to coordinate representation of compositional tables. *Scandinavian Journal of Statistics*, 45(4), 879–899.

Fačevicová, K., Kynčlová, P. and Macků, K. (2021). Geographically Weighted Regression Analysis for Two-Factorial Compositional Data. In: *Advances in Compositional Data Analysis*, pages 103–124. Springer, Cham.

Filzmoser, P., Hron, K. and Templ, M. (2018). *Applied Compositional Data Analysis.* Springer, Cham.

Gába, A., Dygrỳn, J., Štefelová, N., Rubín, L., Hron, K. and 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), 1–9. https://doi.org/10.1186/s12199-021-00932-6

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

Greenacre, M. (2018). *Compositional Data in Practice.* CRC Press, Boca Raton.

Greenacre, M. (2019a). Comments on: Compositional data: the sample space and its structure. *TEST*, 28(3), 644–652.

Greenacre, M. (2019b). Variable selection in compositional data analysis using pairwise logratios. *Mathematical Geosciences*, 51, 649–682.

Hron, K., Coenders, G., Filzmoser, P., Palarea-Albaladejo, J., Faměra, M. and Matys Grygar, T. (2021). Analysing pairwise logratios revisited. *Mathematical Geosciences*, 53(7), 1643–1666.

Hron, K., Filzmoser, P., Caritat, P. d., Fišerová, E. and Gardlo, A. (2017). Weighted pivot coordinates for compositional data and their application to geochemical mapping. *Mathematical Geosciences*, 49, 797–814.

Hron, K., Machalová, J. and Menafoglio, A. (2023). Bivariate densities in Bayes spaces: orthogonal decomposition and spline representation. *Statistical Papers*, 64, 1629-1669.

Johnson, R.A. and Wichern, D.W. (2007). *Applied Multivariate Statistical Analysis. 6th Edition*. Pearson Prentice Hall, Upper Saddle River.

Kynčlová, P., Filzmoser, P. and Hron, K. (2016). Compositional biplots including external non-compositional variables. *Statistics*, 50(5), 1132–1148.

Müller, I., Hron, K., Fišerová, E., Šmahaj, J., Cakirpaloglu, P. and Vančáková, J. (2018). Interpretation of compositional regression with application to time budget analysis. *Austrian Journal of Statistics*, 47(2), 3–19.

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

R Core Team (2020). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing.* Vienna, Austria. https://www.R-project.org/

Templ, M., Hron K. and Filzmoser, P. (2011). robCompositions: An R-package for Robust Statistical Analysis of Compositional Data. *Compositional Data Analysis, Theory and Applications,* 341–355, John Wiley & Sons, Ltd.

Varmuza, K. and Filzmoser, P. (2009). *Introduction to multivariate statistical analysis in chemometrics.* CRC Press, Boca Raton.