

PALACKÝ UNIVERSITY OLOMOUC
FACULTY OF SCIENCE

HABILITATION THESIS

Analysis of Multi-factorial Compositional
Data and Related Methodological
Developments



Mgr. Kamila Fačevicová, Ph.D.
Dept. of Mathematical Analysis and Applications of Mathematics
Applied Mathematics, 2026

I declare that I have prepared the submitted habilitation thesis independently using properly cited literary sources.

In Olomouc on 4. 9. 2026

Mgr. Kamila Fačevicová, Ph.D.

Contents

1	Introduction	5
2	Compositional data	10
2.1	Why should we care about compositional data?	10
2.2	Basics of compositional data analysis	16
2.2.1	Descriptive statistics	16
2.2.2	Working in coordinates	19
3	Multi-factorial compositional data	23
4	Outlier detection and robust analysis of compositional data	27
5	Results	31
5.1	Compositional cubes (Fačevicová et al., 2023)	32
5.2	Correspondence analysis (Fačevicová et al., 2025)	34
5.3	Backwards pivot coordinates (Nesrstová et al., 2023)	36
5.4	Geographically weighted regression (Fačevicová et al., 2021)	40
5.5	Independent component analysis (Muehlmann et al., 2021)	42
5.6	Robust blind source separation (Bábek et al., 2022)	44
5.7	Cellwise robust covariance estimator (Rieser et al., 2023)	46
5.8	Robust principal component analysis (de Sousa et al., 2021)	48
5.9	Zero-handling strategies (Tang et al., 2026)	50
5.10	Application to geochemistry (Fačevicová et al., 2016)	52
5.11	Application to eye-tracking (Fačevicová et al., 2026)	54
6	Summary of the main results and future research directions	57
	References	59
A	Full texts of the selected publications	68

I would like to thank my immediate family for their unconditional support, my friends, colleagues, and collaborators, who believed in me when I could not believe in myself, and my students, who give my work its true meaning.

1. Introduction

The habilitation thesis focuses on the analysis of relative-valued data, known as compositions. This type of observations, traditionally defined as D -part positive-valued vectors (Aitchison, 1982), appears in a broad range of applications, ranging from geochemistry and the natural and health sciences to, for example, eye-tracking research (see Buccianti and Gozzi (2021), Gloor et al. (2017), Dumuid et al. (2020) or [Fačevicová et al. \(2026\)](#)). An extensive list of already well established applications, as well as the emerging ones, is provided in Coenders et al. (2023). All of these applications share a common property: the analysis of such positive-valued vectors focuses on their relative structure rather than (or in addition to) their absolute values, which requires specific statistical treatment. The main principles of (vector) compositional data analysis will be introduced in Section 2.

In the traditional definition, each part of a compositional vector can be understood as a category of an underlying factor (e.g. chemical elements, age categories) and its value represents its relative contribution to the whole, while the sum of these values is usually irrelevant and represents only one of infinitely many equivalent representations of the same composition. This definition can be further extended to the case where the relative structure is constructed with respect to two or more underlying factors, leading to the theory of compositional tables and cubes, described in [Fačevicová et al. \(2018\)](#) and [Fačevicová et al. \(2023\)](#). A useful property of multi-factorial compositions is that such objects can be orthogonally decomposed into independent and interaction components. This allows these components to be modelled or analysed separately, leading to a more detailed understanding of the phenomenon under study.

An introduction to analysis of two-factorial compositions will be given

in Section 3. The theory of multi-factorial compositions, including their definition, decomposition, real-valued coordinate representation and analysis, is developed in the manuscripts [Fačevicová et al. \(2023\)](#), [Nesrstová et al. \(2023\)](#), [Fačevicová et al. \(2021\)](#) and [Fačevicová et al. \(2025\)](#), which form the core of this thesis.

Similarly to classical multivariate data, their compositional counterparts are also affected by the presence of outlying observations (Maronna et al., 2019) or deviating cells (Rousseeuw and van den Bossche, 2018). As will be discussed in Section 4, the nature of such anomalous observations differs from that in the classical setting. In accordance with the fundamental definition of compositional data, compositional outliers are typically characterised by a distorted relative structure, that is, by unusual ratios between compositional parts, rather than by atypically low or high observed values. An important part of this thesis, therefore, deals with robust methods for compositional data analysis. More specifically, it presents three approaches to the robust analysis of vector ([Rieser et al., 2023](#)), two-factorial ([de Sousa et al., 2021](#)), and time-dependent compositional data ([Bábek et al., 2022](#)), when the last paper further develops findings from [Muehlmann et al. \(2021\)](#).

The thesis builds on the results achieved by the applicant and her co-authors, which are described in detail in the following manuscripts (listed in chronological order):

Fačevicová et al. (2026)

Fačevicová, K., Vymazal, J., Popelka, S. Advancing eye movement analysis through compositional modeling: A new perspective on Yarbus' classic study. *Behavior Research Methods*, 58(6), 173 (2026). <https://doi.org/10.3758/s13428-026-03054-5>

Tang et al. (2026)

Tang, W., **Fačevicová, K.**, Nordhausen, K., Taskinen, S. A critical comparison of handling zeros in high-dimensional compositional count data. *Chemometrics and Intelligent Laboratory Systems*, 277, 105821 (2026). <https://doi.org/10.1016/j.chemolab.2026.105821>

Fačevicová et al. (2025)

Fačevicová, K., Filzmoser, P., Hron, K. Correspondence analysis from the viewpoint of compositional tables. *Statistical Analysis and Data Mining*, 18(4) (2025). <https://doi.org/10.1002/sam.70023>

Rieser et al. (2023)

Rieser, C., **Fačevicová, K.**, Filzmoser, P. Cell-wise robust covariance estimation for compositions, with application to geochemical data. *Journal of Geochemical Exploration*, 253, 107299 (2023). <https://doi.org/10.1016/j.gexplo.2023.107299>

Nesrstová et al. (2023)

Nesrstová, V., Jašková, P., Pavlů, I., Hron, K., Palarea-Albaladejo, J., Gába, A., Pelclová, J., **Fačevicová, K.** Simple enough, but not simpler: Reconsidering additive logratio coordinates in compositional analysis. *SORT*, 47(2) (2023). <https://doi.org/10.57645/20.8080.02.9>

Fačevicová et al. (2023)

Fačevicová, K., Filzmoser, P., Hron, K. Compositional cubes: A new concept for multi-factorial compositions. *Statistical Papers*, 64(3), 955–985 (2023). <https://doi.org/10.1007/s00362-022-01350-8>

Bábek et al. (2022)

Bábek, O., **Fačevicová, K.**, Žídek, M., Sedláček, J., Muehlmann, C., Nordhausen, K., Hron, K. X-ray fluorescence scanning of soft and wet-sediment cores in terrestrial environments; A robust blind source separation approach. *Journal of Geochemical Exploration*, 243, 107106 (2022). <https://doi.org/10.1016/j.gexplo.2022.107106>

de Sousa et al. (2021)

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

Fačevicová et al. (2021)

Fačevicová, K., Kynčlová, P. and K. Macků. Geographically weighted regression analysis for two-factorial compositional data. In P. Filzmoser, K. Hron, J.A. Martin-Fernández and J. Palarea-Albaladejo, editors, *Advances in Compositional Data Analysis. Festschrift in Honour of Vera Pawlowsky-Glahn*. Springer, Cham, 2021. https://doi.org/10.1007/978-3-030-71175-7_6

Muehlmann et al. (2021)

Muehlmann, C., **Fačevicová, K.**, Gardlo, A., Janečková, H. and K. Nordhausen. Independent component analysis of compositional data. In: A. Daouia, A. Ruiz-Gazen, editors, *Advances in Contemporary Statistics and Econometrics*. Springer, Cham, 2021. https://doi.org/10.1007/978-3-030-73249-3_27

Fačevicová et al. (2016)

Fačevicová, K., Bábek, O., Hron, K., Kumpan, T. Element chemo-

stratigraphy of the Devonian/Carboniferous boundary – A compositional approach. *Applied Geochemistry*, 75 (2016), 211-221. <https://doi.org/10.1016/j.apgeochem.2016.10.002>

The in-text references to these manuscripts are highlighted in blue, and their full texts are provided in Section A.

2. Compositional data

2.1. Why should we care about compositional data?

As already mentioned in the Introduction, modern compositional data analysis (CoDA) originated in geochemistry, which remains one of the principal fields, where this type of data naturally occurs (see, e.g., Buccianti and Gozzi (2021); Tolosana-Delgado (2012), [Fačevicová et al. \(2016\)](#)). Since then, CoDA has found widespread applications in many other research areas, including metabolomics (Kouřil et al., 2023), microbiome research (Gloor et al., 2017), and time-use epidemiology (Fairclough et al., 2017). Recently, eye-tracking research has emerged as another promising application area for the methodology. In this context, the fundamental principles of CoDA are demonstrated in [Fačevicová et al. \(2026\)](#) using a newly published dataset (Popelka et al., 2026). The dataset is based on a replication of the classical experiment by A. L. Yarbus (Yarbus, 1967), which provided the first evidence that eye movements reflect cognitive processes. It also serves as the running example throughout this thesis.

In general, eye-tracking experiments are based on recording eye movements, including the locations and durations of eye fixations and the transitions between them (Duchowski, 2017). One common approach to the analysis of such data is to evaluate the total time spent observing predefined parts of the stimulus, referred to as Areas of Interest (AOIs). This measure is known as the Total Duration of Fixation (TDoF). The running example dataset contains information on the TDoFs recorded for seven AOIs during repeated observations of the same painting by I. E. Repin (Figure 1) under seven different viewing instructions (Tasks). In total, the dataset comprises recordings from 144 observers.

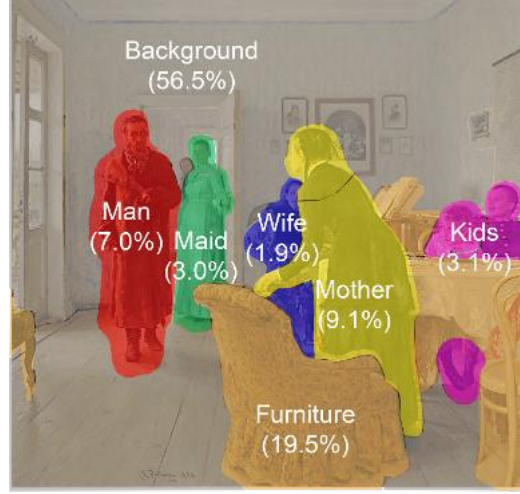


Figure 1: Segmentation of I. E. Repin’s painting ‘They Did Not Expect Him’ into areas of interest. The numbers in brackets indicate the percentage of the painting covered by each respective AOI.

For illustration, consider two observers (P74 and P101) who viewed the painting under the same viewing instruction (Task). For simplicity, suppose that the stimulus is partitioned into only three AOIs—Visitor, Persons, and Background—instead of the original seven. Their eye-tracking recordings are shown in the upper part of Figure 2. The circles represent individual fixations, with their sizes proportional to the corresponding fixation durations, while the connecting lines illustrate the transitions between consecutive fixations.

The same figure also presents scanpaths, i.e., graphical representations of the sequence of fixations, and scarf plots, i.e., graphical representations of the sequence of AOI visits (generated in Vojtechovska and Popelka (2026)), for the same two observers. The lower part of Figure 2 clearly shows that the first observer made substantially more fixations while viewing the painting than the second observer, resulting in higher total fixation durations for each AOI. Specifically, the first observer (P74) spent 8323, 5353, and 4319 ms by

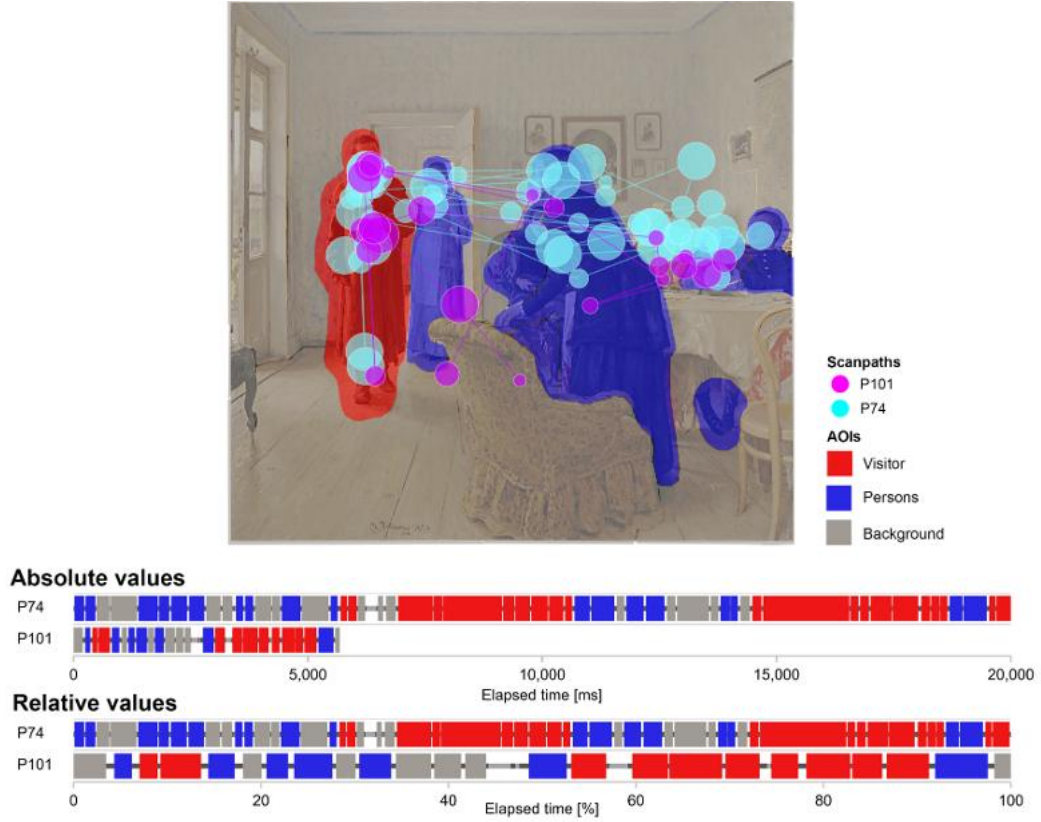


Figure 2: Illustration of fixation distributions for two observers who differ in total fixation duration but allocate their attention equally across AOIs.

observing the Visitor, Persons, and Background AOIs, respectively. For the second observer (P101), the corresponding durations were 2187, 1377.3, and 1036 ms. An analysis based solely on the absolute fixation durations would suggest substantially different viewing behaviour. However, these differences may primarily reflect differences in individual viewing pace or attention span rather than genuinely distinct viewing strategies.

A natural way to mitigate the influence of individual characteristics is to replace absolute fixation counts or dwell times with relative measures. Instead of analyzing the TDoFs per AOI, one can evaluate the proportion of the total viewing time allocated to each AOI. For the observers from

Figure 2, this yields the relative distributions (46.3 %, 29.7 %, 24 %) and (47.5 %, 29.9 %, 22.5 %), which clearly reveal the similarity between the two observers’ fixation patterns. Nevertheless, expressing the time distribution as percentages and the consequent focus on its relative structure brings us to the scope of Compositional Data Analysis (CoDA) (Aitchison, 1982; Coenders et al., 2023; Filzmoser et al., 2018).

When the relative relationships between variables carry more information than their absolute magnitudes, the application of classical statistical methods – originally developed for unconstrained real-valued data – can yield inaccurate or even misleading results. This issue can be in general caused by the presence of spurious correlations between variables, which arise from the constant-sum constraint: an increase in one component (e.g., the proportion of fixations allocated to a given AOI) in the relative data representation necessarily induces a decrease in the remaining components, even if no true association exists between them. A second source of distortion lies, in some cases, in the inappropriate use of Euclidean geometry, which is based on absolute differences between values or observations, rather than on their relative relationships (Pearson, 1896 – 1897; Aitchison, 1982).

To illustrate this issue, consider again the time distributions of observers P74 and P101. When analyzing relative data, the distance between two observations should reflect only differences in their relative structure and therefore remain invariant with respect to the chosen representation. However, the Euclidean distance between the considered vectors depends on the chosen representation: when computed from the raw fixation times, the distance equals 8014.66, whereas for their percentage representation it is 1.97. While this difference is expected due to the change of scale, it reveals that Euclidean geometry does not respect the relative nature of the data. CoDA

offers a principled way to address this limitation. Its methodology is typically based on log-ratio transformations, which extract the pure relative information from each observation. Crucially, this is achieved without requiring prior normalization to a constant sum, since CoDA treats compositions as equivalence classes of positive vectors conveying relative information only. As will be discussed later, compositional methodology relies on the so-called Aitchison geometry, which yields a distance of 0.068 for the discussed pair of observations, regardless of whether the compositions are expressed as raw concentrations, proportions, or percentages. The resulting distance is therefore invariant with respect to scale or normalization, which should not alter the inherent relative structure.

However, the choice of an appropriate analytical approach should always be guided by the research question. In certain situations, a classical analysis based on absolute information may be preferable to a compositional one. Moreover, for certain types of research questions, a parallel analysis of both the relative and absolute information contained in the data may provide a more comprehensive understanding of the phenomenon under study. For a more detailed discussion, see [Fačevicová et al. \(2016\)](#) and [Fačevicová et al. \(2026\)](#).

For the analysis of data carrying relative information, as introduced above, there are several modifications of classical statistical methods that explicitly account for the constrained nature of such data. For instance, the relative structure can be accommodated through an appropriate parametrization of multinomial or mixed-effects models (see, e.g., Vencálek et al. (2020)). Compositional data analysis, however, provides a comprehensive and mathematically grounded framework specifically designed for this type of data. Beyond hypothesis testing, CoDA allows for supervised and unsupervised learning

as well as robust statistical inference. Moreover, the methodology has later been extended to a broader class of relative-valued or distributional data, including multi-factorial (Egozcue et al. (2008) and Fačevićová et al. (2023)) and even functional compositions (van den Boogaart et al., 2014; Pavlů et al., 2023). In this sense, CoDA does not merely adapt existing statistical tools to relative data but provides a coherent geometric and algebraic foundation for their analysis.

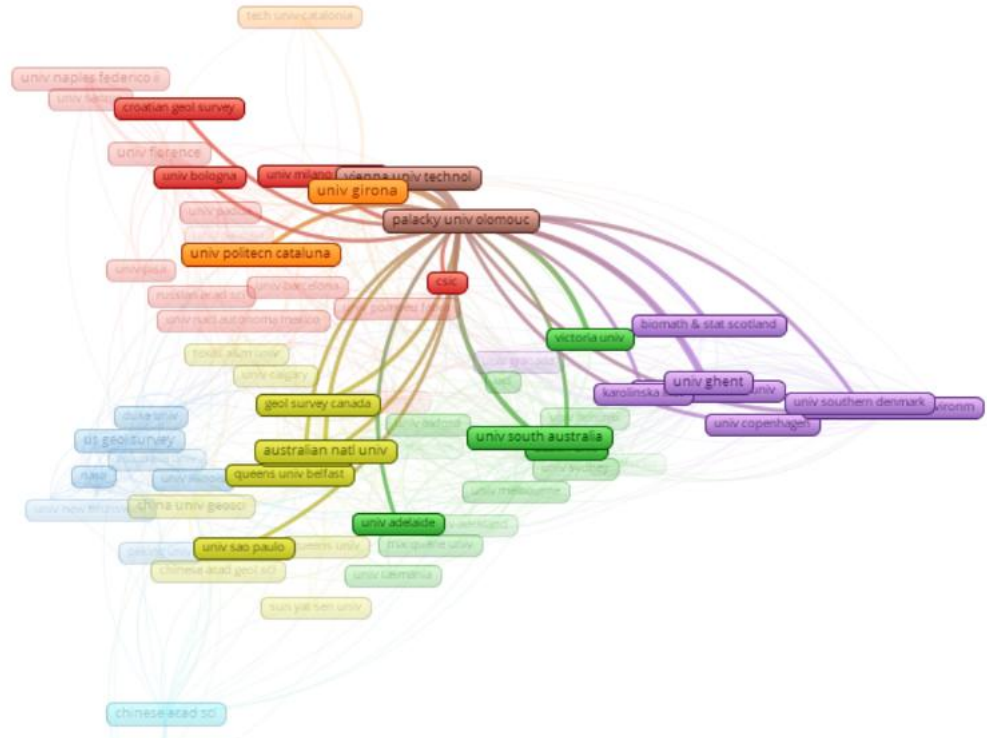


Figure 3: Collaboration network of organizations publishing research on compositional data analysis. The size of each node is proportional to the number of publications, while the thickness of the links represents the strength of collaboration. The collaboration network of Palacký University Olomouc is highlighted. The network was generated using VOSviewer from records retrieved from the Web of Science Core Collection on 29 July 2026.

The field of compositional data analysis represents a vibrant and continuously growing research area. Although a substantial portion of the ideas at the early stages of CoDA development originated from the CoDA group based in Spain, namely at the University of Girona and the Polytechnic University of Catalonia, Figure 3 illustrates that Palacký University Olomouc has become one of the major research centres contributing to the development of CoDA. It is among the five most productive institutions publishing in this field according to the Web of Science database, which reflects the long-term contribution to the development of CoDA and provides the broader context for the methodological contributions presented in this thesis.

2.2. Basics of compositional data analysis

2.2.1. Descriptive statistics

A compositional observation is commonly defined as a D -part vector $\mathbf{x} = (x_1, \dots, x_D)'$ of positive components ($x_i > 0, \forall i$), where the essential information is contained in the ratios between parts, rather than in their absolute magnitudes. This general definition corresponds, for instance, to a situation in which the total duration of fixations spent in D distinct AOIs is recorded with the goal of understanding the relative distribution of attention across these areas. In accordance with this perspective, the total sum of the components is not relevant for the analysis. Normalizing the data to sum to 1 (proportions) or to 100 (percentages) represents only two among many equivalent ways of expressing the same relative information. The underlying relative structure – i.e., the ratios between the parts – remains unchanged regardless of the chosen normalization.

A fundamental principle of compositional analysis is scale invariance

(Egozcue, 2009), which ensures that rescaling should not affect the outcome of the analysis. Nevertheless, using a common total can be practical for numerical comparison between multiple compositions. To facilitate this, we define the closure operation $\odot_{\kappa}(\cdot)$ which rescales a given composition to an equivalent one whose total equals a specified constant κ .

Specific features of compositional data, particularly the scale invariance property, are captured by the Aitchison geometry (Billheimer et al., 2001; Pawlowsky-Glahn and Egozcue, 2001) with Euclidean vector space properties. For compositions $\mathbf{x}, \mathbf{y} \in \mathcal{O}$ and a real constant α , these properties result from defining the operations perturbation, powering, and the Aitchison inner product,

$$\begin{aligned} \mathbf{x} * \mathbf{y} &= \odot(x_i y_i, \dots, x_D y_D)', \quad \alpha \div \mathbf{x} = \odot(x_i^{\alpha}, \dots, x_D^{\alpha})', \\ \mathbf{x}, \mathbf{y} \rhd_{\mathbb{A}} &= \frac{1}{2D} \sum_{i=1}^D \sum_{j=1}^D \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}, \end{aligned} \tag{1}$$

respectively.

As mentioned above, the traditional definition of a compositional vector does not permit any of its components to be zero. However, zero values are frequently encountered in practice and must be properly addressed before applying compositional analysis. This issue can be handled either by merging of categories or by applying suitable zero-imputation techniques, as discussed, for example, in Tang et al. (2026); Hron et al. (2010); Palarea-Albaladejo et al. (2007) and Palarea-Albaladejo and Martín-Fernández (2008).

The basic descriptive statistics of compositional data are compositional (i.e., geometrical) mean, variation matrix and total variance. For a sample of n D -part compositions organized in an $n \times D$ data matrix $\mathbf{X} = (x_{ij})$ these statistics are defined as follows:

The **compositional mean** is given by

$$\mathbf{g}_\star = \odot_\kappa (g_1, \dots, g_D)', \quad \text{where} \quad g_j = \left(\prod_{i=1}^n x_{ij} \right)^{1/n} \quad (2)$$

i.e., (the closure of) the component-wise geometric means across all observations. The $D \times D$ **variation matrix** $\mathbf{T}_\star = (t_{jk})$ has elements given by

$$t_{jk} = S_n^2(\ln(x_j/x_k)) \quad (3)$$

that is, each element represents the sample variance (S_n^2) of the log-ratio between the j -th and k -th components across the sample. Elements of the variation matrix that are close to zero indicate pairs of compositional parts for which $x_{ij} \oplus c \cdot x_{ik}$, with $c \in R^+$ being constant across all observations i . In other words, these parts are nearly proportional, and their mutual relationship remains stable throughout the sample. Conversely, large values of t_{jk} indicate high variability in the corresponding log-ratio, reflecting substantial fluctuations in the mutual relative dominance of the two parts across individual compositions. The classical estimator of the variation matrix is obtained by estimating the variance of each pairwise log-ratio separately. A robust alternative, based on the Qn scale estimator, was proposed in [Rieser et al. \(2023\)](#).

Finally, the overall level of variability captures the **total variance**, defined as

$$\text{totvar}(\mathbf{X}) = \frac{1}{2D} \sum_{j=1}^D \sum_{k=1}^D t_{jk}. \quad (4)$$

The total variance serves as a natural measure of similarity (or homogeneity) within a set of compositions.

A traditional way to visualize 3-part compositions is by means of **ternary diagrams** (one example is given on Figure 5 in Section 4). For higher-dimensional compositions, visualization is typically based on log-ratio transformations or on dimension reduction techniques, such as principal component analysis (PCA), applied to the transformed data. More generally, log-ratio transformations constitute the cornerstone of traditional compositional data analysis, as they provide a real-valued coordinate representation of compositions while preserving their relative information. Consequently, they serve as the starting point for most multivariate statistical methods developed for compositional data.

2.2.2. Working in coordinates

The classical methods of multivariate data analysis, such as hypothesis testing, clustering, or classification (Anderson et al., 1994), are typically not directly applicable to compositional data. These methods generally rely on absolute information carried by data, whereas for compositional analysis the relative structure is of main interest, as outlined above. The standard approach to overcome this discrepancy is to map compositional observations into a system of real-valued coordinates, where the principles of Euclidean geometry apply, like centered (clr) or isometric (ilr) log-ratio coordinates (Pawlowsky-Glahn et al., 2015).

One of the widely used options is the **centered log-ratio (clr) transformation**. This transformation expresses the dominance of each compositional part relative to the overall geometric mean of the respective observation, while logarithmic scaling ensures symmetry. Formally, the clr representation of a D -part compositional vector is defined as

$$\text{clr}(\mathbf{x}) = \left(\ln \frac{x_1}{g(\mathbf{x})}, \dots, \ln \frac{x_D}{g(\mathbf{x})} \right)' = \mathbf{H} \ln(\mathbf{x}) =: \mathbf{y}, \quad (5)$$

with $g(\mathbf{x}) = (x_1 \cdot \dots \cdot x_D)^{1/D}$ and $\mathbf{H}$ being a symmetric centering matrix. If the clr transformation is applied to each row of the compositional data matrix $\mathbf{X}$, i.e., to each individual observation, the resulting $n \times D$ matrix $\mathbf{Y}$ can be directly used in a wide range of standard multivariate techniques, such as clustering or dimension reduction. Moreover, clr coordinates have a convenient interpretation in terms of mean relative dominance of the respective parts. While the positive values represent the more abundant compositional parts, negative values stand for those parts which are in a given composition rather negligible.

Nevertheless, the clr transformation is subject to an inherent zero-sum constraint that follows directly from its definition. Consequently, the covariance matrix of clr-transformed data is singular, which prevents its direct use in statistical procedures requiring a full-rank covariance matrix. For such methods, an alternative coordinate system, namely the isometric log-ratio (ilr) transformation (Egozcue et al., 2003; Egozcue and Pawłowsky-Glahn, 2005), provides an appropriate solution.

Isometric log-ratio coordinates provide a $(D - 1)$ -dimensional real-valued representation of a compositional vector $\mathbf{x}$ and are defined as

$$\text{ilr}(\mathbf{x}) = \mathbf{V} \ln(\mathbf{x}) = \mathbf{V} \mathbf{y} =: \mathbf{z}, \quad (6)$$

where $\mathbf{V}$ is the so-called contrast matrix with orthonormal rows. The ilr transformation represents a flexible class of coordinate systems, whose specific form is determined by the choice of the contrast matrix $\mathbf{V}$. Consequently, the coordinates can be selected according to the objectives of the subsequent

analysis or to facilitate their interpretation.

One of the most widely used systems of ilr coordinates is formed by balances, i.e., log-ratios between two groups of parts defined through a sequential binary partition (SBP). In general, balances are computed as

$$z_i = \sqrt{\frac{r_i s_i}{r_i + s_i}} \ln \frac{g(x_{i_1}, \dots, x_{i_{r_i}})}{g(x_{i'_1}, \dots, x_{i'_{s_i}})}, \quad \text{for } i = 1, \dots, D-1, \quad (7)$$

where the indices $i_1, \dots, i_{r_i}$ and $i'_1, \dots, i'_{s_i}$ denote the parts assigned to the (+) and (−) groups in the respective step of the SBP. Additionally, two subclasses of balances are particularly useful in practice. **Pivot coordinates** (Fišerová and Hron, 2011), defined as

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

are particularly attractive because their first coordinate z_1 quantifies the relative dominance of the first compositional part x_1 . **Backwards pivot coordinates** (Hron et al., 2021) are defined analogously, but emphasize the last coordinate, which is given by $1/\sqrt[D-i]{2} \ln(x_{D-1}/x_D)$. By permuting the parts so that the variable of interest occupies the first or the last position in the composition, pivot and backwards pivot coordinates can be advantageously used, for example, in regression modelling.

The concept of ilr coordinates was extended to two-factorial compositions in Fačevicová et al. (2018) and further generalised to multi-factorial compositions in Fačevicová et al. (2023). In addition, Nesrstová et al. (2023) introduces a modification of backwards pivot coordinates for two-factorial compositions.

An important property shared by both the clr and ilr representations is

that they preserve the algebraic and geometric structure of the simplex. In particular, for two D -part compositional vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ and their clr and ilr representations $\mathbf{y}_1, \mathbf{y}_2$ and $\mathbf{z}_1, \mathbf{z}_2$ it holds that

$$\begin{aligned}\mathbf{x}_1 * \mathbf{x}_2 &= \mathbf{y}_1 + \mathbf{y}_2 = \mathbf{z}_1 + \mathbf{z}_2, \\ \alpha \div \mathbf{x} &= \alpha \cdot \mathbf{y} = \alpha \cdot \mathbf{z}, \\ \mathcal{R}_{\mathbf{x}_1, \mathbf{x}_2} \mathbf{A} &= \mathcal{R}_{\mathbf{y}_1, \mathbf{y}_2} \mathbf{E} = \mathcal{R}_{\mathbf{z}_1, \mathbf{z}_2} \mathbf{E}.\end{aligned}\tag{9}$$

These properties enable the application of classical multivariate statistical methods to compositional data. An overview of the basic methods in the context of eye-tracking research is provided in [Fačevicová et al. \(2026\)](#). The log-ratio methodology is applied in [Fačevicová et al. \(2016\)](#) to obtain a better understanding of the geochemical structure of the Devonian–Carboniferous boundary. Finally, [Muehlmann et al. \(2021\)](#) and [Bábek et al. \(2022\)](#) extend the methodology to blind source separation of iid and time-dependent compositional data.

3. Multi-factorial compositional data

The concept of vector compositional data can be naturally extended to situations where the relative distribution is determined by categories of two factors. In the running eye-tracking example, the same stimulus was repeatedly observed under different viewing instructions. When the total time spent on the experiment is stratified simultaneously by task and by the visited AOIs, the resulting data structure corresponds to the term **compositional table**, as defined in Egozcue et al. (2008).

Formally, an $I \times J$ compositional table is defined as a two-dimensional array of positive entries for which the relative structure is of primary interest. Similarly to the vector case, we can write $\mathbf{x} = (x_{ij})$, $x_{ij} > 0$, $i = 1, \dots, I$, $j = 1, \dots, J$ and extend the basic simplex operations (1) to

$$\mathbf{x} * \mathbf{y} = \odot \begin{pmatrix} x_{11}y_{11} & \dots & x_{1J}y_{1J} \\ \vdots & \ddots & \vdots \\ x_{I1}y_{I1} & \dots & x_{IJ}y_{IJ} \end{pmatrix}, \quad \alpha \div \mathbf{x} = \odot \begin{pmatrix} x_{11}^\alpha & \dots & x_{1J}^\alpha \\ \vdots & \ddots & \vdots \\ x_{I1}^\alpha & \dots & x_{IJ}^\alpha \end{pmatrix} \quad (10)$$

$$\mathbf{x}, \mathbf{y} \equiv_A = \frac{1}{2IJ} \sum_{i,j} \ln \frac{x_{ij}}{x_{kl}} \ln \frac{y_{ij}}{y_{kl}}.$$

A key property of compositional tables is that they can be orthogonally decomposed into an **independent part** ($\text{ind}(\mathbf{x})$) and an **interaction part** ($\text{int}(\mathbf{x})$) (Egozcue et al., 2015),

$$\mathbf{x} = \text{ind}(\mathbf{x}) * \text{int}(\mathbf{x}). \quad (11)$$

In the eye-tracking example, the independent part describes the overall

allocation of attention between AOIs and tasks separately, whereas the interaction part captures task-specific changes in the attention distribution across AOIs. Part $\text{ind}(\mathbf{x})$ mimics the situation of complete independence between row and column factors and its cells are proportional to the product of row and column geometric marginals

$$\text{ind}(\mathbf{x})_{ij} = g(\mathbf{x}_{i\bullet})g(\mathbf{x}_{\bullet j}), \quad i = 1, \dots, I, \quad j = 1, \dots, J. \quad (12)$$

As discussed in Genest et al. (2023), geometric marginals represent orthogonal projections of the original table to the considered row or column subspaces and for purely independent structures these marginals coincide with their arithmetic counterparts. The interaction part $\text{int}(\mathbf{x})$ quantifies the departure from independence and has cells given by

$$\text{int}(\mathbf{x})_{ij} = \frac{x_{ij}}{g(\mathbf{x}_{i\bullet})g(\mathbf{x}_{\bullet j})}, \quad i = 1, \dots, I, \quad j = 1, \dots, J. \quad (13)$$

Consequently, the decomposition enables the independent and interaction structures to be analysed separately. It is therefore advantageous to employ a coordinate system that respects the reduced dimensionality of the decomposed parts. While the sample space dimension of a compositional table is $IJ - 1$, it reduces to $I + J - 2$ for $\text{ind}(\mathbf{x})$ and to $(I - 1)(J - 1)$ for $\text{int}(\mathbf{x})$. The first ilr coordinate system respecting this decomposition was proposed for 2×2 compositional tables in Fačevicová et al. (2014). It was subsequently extended to general $I \times J$ tables in Fačevicová et al. (2016), and a fully general ilr coordinate system was introduced in Fačevicová et al. (2018). The main idea is to define separate sequential binary partitions for the row and column factors. These partitions result in the coordinates of $\text{ind}(\mathbf{x})$, which have the form of balances

$$z_i^r = \sqrt{\frac{stJ}{s+t}} \ln \frac{[g(x_{j_i \bullet}) \cdots g(x_{j_s \bullet})]^{1/s}}{[g(x_{k_i \bullet}) \cdots g(x_{k_t \bullet})]^{1/t}}, \quad (14)$$

$$z_j^c = \sqrt{\frac{uvI}{u+v}} \ln \frac{[g(x_{\bullet l_i}) \cdots g(x_{\bullet l_u})]^{1/u}}{[g(x_{\bullet m_i}) \cdots g(x_{\bullet m_v})]^{1/v}}$$

for $i = 1, \dots, I-1$ and $j = 1, \dots, J-1$. The combination of the row and column partitions further defines sets of four subtables, A_{ij} , B_{ij} , C_{ij} and D_{ij} , which are subsequently used in the construction of the interaction (odds-ratio) coordinates

$$z_{ij}^{OR} = \sqrt{\frac{\subseteq A_{ij} \subseteq D_{ij} \subseteq}{\subseteq A_{ij} \subseteq + \subseteq B_{ij} \subseteq + \subseteq C_{ij} \subseteq + \subseteq D_{ij} \subseteq}} \ln \frac{g(\mathbf{x}_{A_{ij}}) g(\mathbf{x}_{D_{ij}})}{g(\mathbf{x}_{B_{ij}}) g(\mathbf{x}_{C_{ij}})}, \quad (15)$$

$i = 1, \dots, I-1$, $j = 1, \dots, J-1$, which form the ilr representation of $\text{int}(\mathbf{x})$. A more detailed step-by-step description of the coordinate construction is provided in Fačevicová et al. (2018). Figure 4 illustrates how the row and column SBPs jointly determine the groups of cells entering the odds-ratio coordinates.

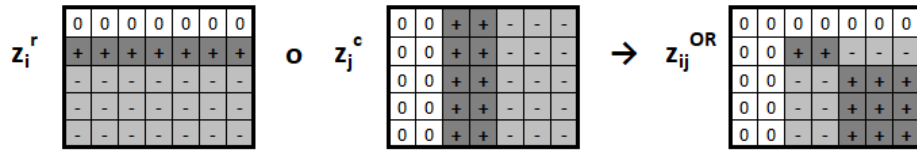


Figure 4: Graphical illustration of the relationship between the row and column SBPs and the groups of cells involved in the construction of the odds-ratio coordinates (15).

Later, a two-factorial analogy of backwards pivot coordinates (Hron et al., 2021) was introduced in Nesrstová et al. (2023). Building on these developments, the same principles were subsequently generalized to multi-factorial

compositions in [Fačevicová et al. \(2023\)](#).

Compositional tables represented in ilr coordinates can then be analysed using a wide range of statistical methods. Examples include geographically weighted regression for spatially dependent compositional tables ([Fačevicová et al., 2021](#)) or robust dimension reduction ([de Sousa et al., 2021](#)). The methodology also exhibits close connections with several classical multivariate methods. [Fačevicová et al. \(2025\)](#) reveals the relationship between the clr representation of $\text{int}(\mathbf{x})$ and correspondence analysis. The same paper also introduces the weighted version of the treatment of compositional tables. Further developments include the analysis of square positive matrices ([Egozcue and Maldonado, 2021](#); [Nakamura et al., 2024, 2025](#)) and the generalization of orthogonal decomposition to the multivariate distributions, including the continuous case, through a proper choice of the reference measure ([Genest et al., 2023](#)).

4. Outlier detection and robust analysis of compositional data

Similarly to real-valued data, compositional data may also be affected by the presence of outliers. However, the nature of compositional outliers differs from that of their real-valued counterparts. While a classical outlier typically exhibits unusually large or small observed values, a compositional outlier is characterised by a distorted relative structure.

Consider again the eye-tracking example introduced in Section 2.1. In addition to observers P74 and P101, whose observation times across the three AOIs (Visitor, Persons and Background) are (8323, 5353, 4319) ms and (2187, 1377.3, 1036) ms, respectively, let us also consider observer P64 with observation times (7330, 4720, 1249.3) ms. From the perspective of classical multivariate analysis, observer P101 would naturally be regarded as the most extreme observation because of the substantially shorter viewing times. In contrast, from the perspective of compositional data analysis, observer P64 appears to be the atypical observation, as the corresponding relative time distribution, (55.1 %, 35.5 %, 9.4 %), differs markedly from those of P74, (46.3 %, 29.7 %, 24 %), and P101, (47.5 %, 29.9 %, 22.5 %). This difference is also apparent from the ternary diagram shown in Figure 5.

Outlier detection and robust analysis of compositional data are typically performed in log-ratio coordinates, where the ilr representation is generally preferred over the clr transformation because the latter yields a singular covariance matrix (Filzmoser and Hron, 2008; Filzmoser and Gregorich, 2020). Most existing robust methodology has therefore been developed as an extension of classical multivariate techniques applied to ilr coordinates, including principal component analysis (Filzmoser et al., 2009), discriminant analysis

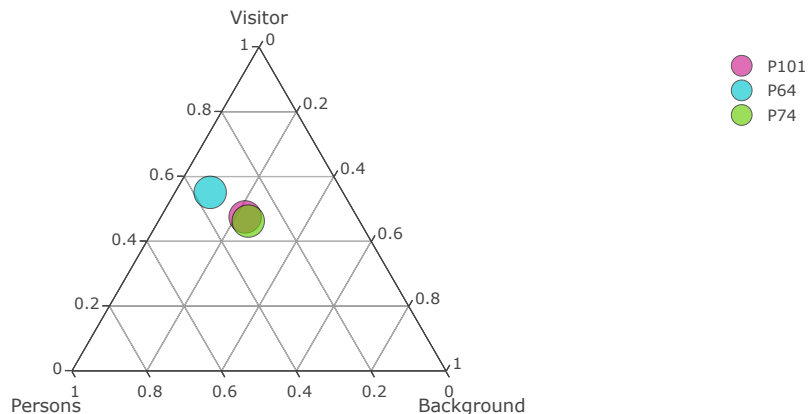


Figure 5: Ternary diagram comparing the relative time distribution of observers P74 and P101 from Figure 2 and an additional observer P64.

(Filzmoser et al., 2011), and regression modelling (van den Boogaart et al., 2021). The same principle is also employed in [de Sousa et al. \(2021\)](#) and [Bábek et al. \(2022\)](#), which aim on robust analysis of compositional tables and time dependent (vector) compositional data, respectively.

Nevertheless, alternative approaches that avoid the standard log-ratio coordinate representation have recently attracted increasing attention. Among them, Sceaaly et al. (2015) proposed a robust principal component analysis based on the power transformation, which provides an alternative representation of compositional data while avoiding the singularity of the clr covariance matrix.

In some situations, particularly in high-dimensional settings, the **rowwise** approach to robust data analysis outlined above, which downweights or removes entire observations, may be overly restrictive. For this reason, Alqalaf et al. (2009) introduced the **cellwise** contamination model, under which only individual cells of the data matrix are assumed to be contaminated. This model has stimulated the rapid development of cellwise robust statistical methods. A comprehensive overview of this field is provided by Ray-

maekers and Rousseeuw (2026). Among the most influential contributions are the work of Rousseeuw and van den Bossche (2018), which introduced a widely used method for detecting cellwise outliers (DDC), and Raymaekers and Rousseeuw (2024), who proposed a cellwise robust alternative to the minimum covariance determinant (MCD) estimator (Rousseeuw, 1985).

Table 1: Observation times of selected participants of the eye-tracking experiment (in ms) and the corresponding pairwise log-ratios.

Participant	Visitor	Persons	Background	$l_{\square}(V./P.)$	$l_{\square}(V./B.)$	$l_{\square}(P./B.)$
P74	8323	5353	4319	0.44	0.66	0.21
P101	2187	1377.3	1036	0.46	0.75	0.28
P64	7330	4720	1249.3	0.44	1.77	1.33

While the methodology for unconstrained data has advanced considerably, extending the concept of cellwise robustness to compositional data is substantially more challenging because a single contaminated cell propagates to all log-ratio coordinates through the geometric mean. This issue is discussed in detail in Templ (2026) and can also be illustrated using the running eye-tracking example. A closer look at the time distribution of observer P64 and Table 1 reveals that it differs from P74 and P101 only in log-ratios involving the Background AOI, whereas the log-ratio between the Visitor and Persons AOIs remains comparable across all three observers. This provides evidence that the main source of outlyingness is the relatively low attention allocated by P64 to the Background AOI. Note that this information becomes considerably less transparent after the clr transformation, because each clr coordinate depends on all compositional parts. The clr coordinates of observers P64, P74 and P101 are respectively $(0.74, 0.30, -1.03)$, $(0.37, -0.08, -0.29)$ and $(0.40, -0.06, -0.34)$. 80 Other coordinate representations, such as ilr, suffer from a similar problem. This example also illustrates why pairwise log-ratios provide a natural basis for cellwise robust procedures: contam-

ination affects only those log-ratios involving the corrupted part, whereas unaffected log-ratios retain their original information.

Current approaches to cellwise robustness for compositional data therefore work primarily with the elemental information contained in pairwise log-ratios. This strategy is adopted, for example, in Walach et al. (2019) and Štefelová et al. (2021), while a pairwise log-ratio-based algorithm for the robust estimation of the variation matrix is introduced in [Rieser et al. \(2023\)](#). Nevertheless, all currently available methods are designed for vector compositional data. The detection and treatment of cellwise contamination in multi-factorial compositional data, introduced in Section 3, remains largely unexplored and therefore represents a natural direction for future methodological research.

5. Results

This chapter presents brief summaries of the eleven articles co-authored by the applicant and listed in Section 1, which constitute the core of this habilitation thesis and are organised according to their topic. Whenever possible, the proposed methodology is illustrated using the running eye-tracking example. The full texts of all manuscripts are provided in the Appendix (Section A).

The manuscript [Fačevicová et al. \(2023\)](#) generalises the concept of compositional tables to multi-factorial compositions, [Fačevicová et al. \(2025\)](#) establishes the connection between compositional tables and correspondence analysis and introduces weighted correspondence analysis in the framework of Bayes spaces, and [Nesrstová et al. \(2023\)](#) generalises backwards pivot coordinates to compositional tables. In addition, [Fačevicová et al. \(2021\)](#) applies the coordinate representation to modelling of spatially dependent data.

The manuscripts [Muehlmann et al. \(2021\)](#) and [Bábek et al. \(2022\)](#) introduce methods for blind source separation of independent and time-dependent vector compositional data, respectively, with the latter providing a robust methodology. Robust methods are also the focus of [Rieser et al. \(2023\)](#) and [de Sousa et al. \(2021\)](#). While the former proposes a cellwise robust estimator of the clr covariance matrix, the latter develops a robust approach to principal component analysis of compositional tables.

Additionally, the manuscript [Tang et al. \(2026\)](#) addresses the problem of handling zeros in high-dimensional compositional count data. Since the methods discussed above generally assume strictly positive components, the choice of an appropriate strategy for handling zeros is crucial for their application.

Finally, the last two papers demonstrate the practical applicability of

compositional data analysis. [Fačevicová et al. \(2016\)](#) applies the methodology to the geochemical analysis of the Devonian–Carboniferous boundary, whereas [Fačevicová et al. \(2026\)](#) introduces compositional data analysis to eye-tracking research. To the best of the applicant’s knowledge, this represents the first application of CoDA in this field and provides a promising starting point for further methodological development.

5.1. Compositional cubes ([Fačevicová et al., 2023](#))

The first paper generalises the results of the applicant’s dissertation thesis, namely the manuscript [Fačevicová et al. \(2018\)](#). As recalled in Section 3, compositional tables describe the relative structure determined by the categories of two underlying factors. In the eye-tracking setting, these factors corresponded to the distribution of visual attention among the AOIs (row factor) while performing different tasks (column factor) on a single stimulus.

The generalisation proposed in this manuscript introduces an additional factor, thereby extending the representation to three dimensions. For example, the same set of tasks may be performed on different devices, such as a smartphone and a desktop computer. Such a three-dimensional object is referred to as a **compositional cube**. Analogously to compositional tables, compositional cubes admit an orthogonal decomposition of the form

$$\mathbf{x} = \text{ind}(\mathbf{x}) * \text{int}(\mathbf{x}) = \text{ind}(\mathbf{x}) * \text{int}^{rc}(\mathbf{x}) * \text{int}^{rs}(\mathbf{x}) * \text{int}^{cs}(\mathbf{x}) * \text{int}^{rcs}(\mathbf{x}), \quad (16)$$

where $\text{ind}(\mathbf{x})$ denotes the independent part, whose cells are determined by the products of the corresponding geometric marginals. The components $\text{int}^{rc}(\mathbf{x})$, $\text{int}^{rs}(\mathbf{x})$, $\text{int}^{cs}(\mathbf{x})$ preserve the pairwise interactions between the corresponding factors, whereas $\text{int}^{rcs}(\mathbf{x})$ captures the interaction shared by all three factors.

The manuscript further shows that the cells of the pairwise interaction components can be obtained analogously to the two-dimensional case, namely as the interaction part of a compositional table obtained by averaging over the remaining dimension. The proposed framework can be readily generalised to relative structures defined by an arbitrary number of factors.

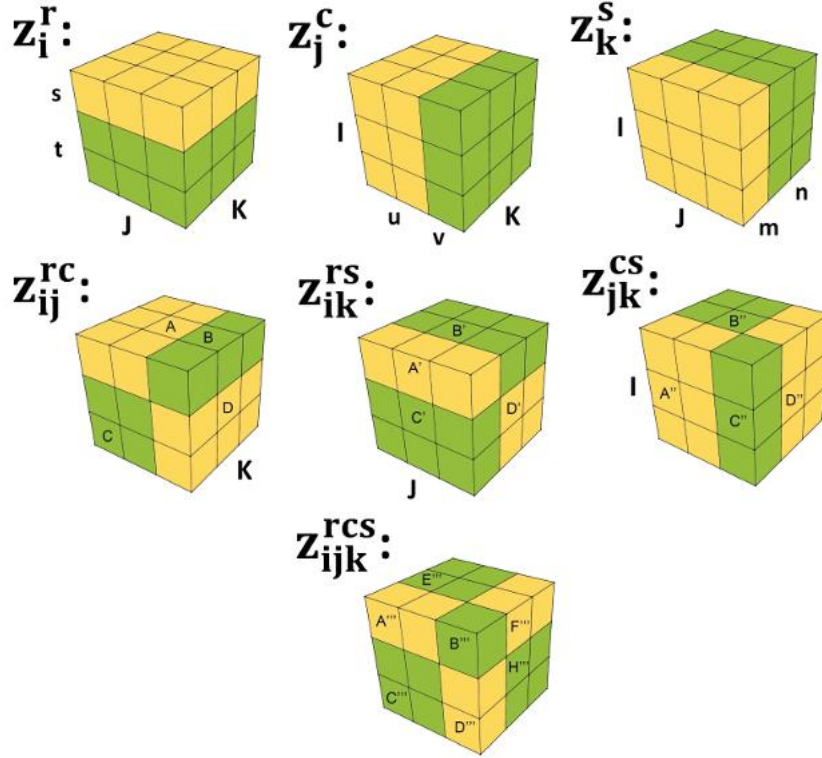


Figure 6: Schematic illustration of the groups of parts involved in the computation of the individual ilr coordinates for compositional cubes.

The analysis of such complex objects can, similarly to classical compositional data, be based on their coordinate representation. The manuscript [Fačevicová et al. \(2023\)](#) introduces a new system of ilr coordinates (see Equation (6)) that respects the decomposition (16) and admits a natural interpretation in terms of balances between the categories of individual factors,

log-odds ratios describing pairwise interactions, and log-ratios of odds-ratios representing the highest-order interactions. This coordinate system not only enables the analysis of the complete three-dimensional structure but also facilitates a separate investigation of its individual components. A schematic illustration of the cells involved in the computation of the respective coordinates is provided in Figure 6.

5.2. Correspondence analysis (Fačevićová et al., 2025)

This manuscript establishes the connection between the analysis of compositional tables, introduced in Section 3, and correspondence analysis (CA), a classical multivariate method for exploring and visualising relationships between two categorical variables represented by a contingency table. It builds on the approach of Greenacre (2010), referred to hereafter as log-ratio analysis (LRA), which formulates CA as a method for analysing a sample of I vector compositions with J parts. The method is based on the singular value decomposition of the doubly centered log-transformed data matrix $\mathbf{X} = (\ln x_{ij})$, yielding the matrix $\mathbf{S}^{\mathbf{R}\mathbf{A}}$ with entries

$$s_{ij}^{\mathbf{R}\mathbf{A}} = \ln x_{ij} - \frac{1}{I} \sum_{i=1}^I \ln x_{ij} - \frac{1}{J} \sum_{j=1}^J \ln x_{ij} + \frac{1}{IJ} \sum_{i=1}^I \sum_{j=1}^J \ln x_{ij}. \quad (17)$$

The manuscript shows that this matrix is mathematically identical to the clr representation of the interaction table (13), i.e., $\mathbf{S}^{\mathbf{R}\mathbf{A}} = \mathbf{V}^{\text{int}}$. Consequently, LRA and singular value decomposition of a clr interaction table yield identical results, despite being based on different statistical models. While LRA interprets $\mathbf{X}$ as a sample of I compositional observations, the compositional-table approach treats the entire matrix as a single observa-

tion.

In the context of eye-tracking, the two paradigms can be interpreted as follows. LRA compares the relative allocation of fixation time across the considered AOIs over a set of experimental tasks, treating each task as one compositional observation. In contrast, the compositional-table approach starts from the relative allocation of the total fixation time of a given observer across both AOIs and tasks. Despite these different interpretations of the data, both approaches lead to exactly the same ordination, illustrated in the left panel of Figure 7.

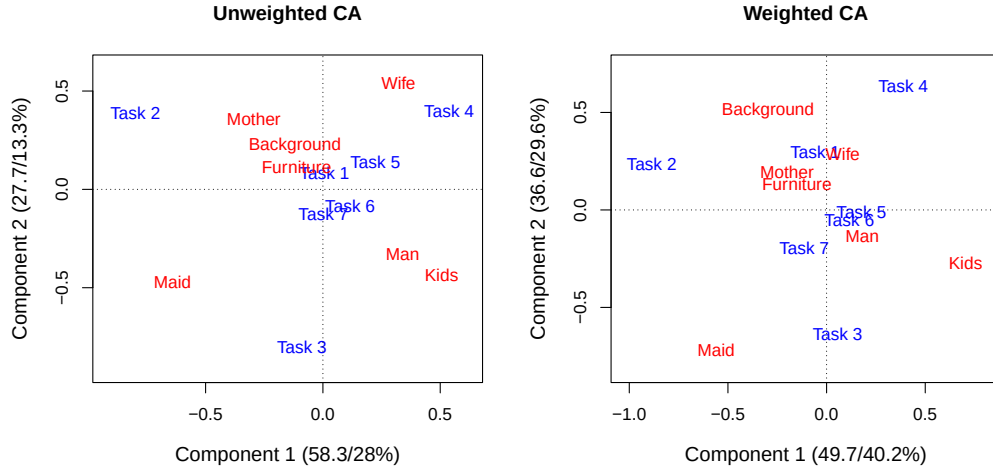


Figure 7: Results of unweighted and weighted correspondence analysis (CA) applied to the distribution of fixation time across all tasks and AOIs for a selected observer. The values in brackets refer to the explained variance within the interaction part and the overall clr table, respectively.

Among other findings, the plot reveals that the attention distribution between AOIs is substantially more imbalanced in Tasks 2, 3, and 4 than in the remaining four tasks. It also shows that the relative attention allocated to the AOIs Maid, Man and Kids varies substantially across tasks, suggesting that these AOIs may be potential task discriminators. The manuscript therefore

demonstrates that two conceptually different statistical models lead to identical analytical results and provides a novel interpretation of correspondence analysis from the perspective of compositional data analysis.

In addition to establishing this equivalence, the manuscript also introduces a weighted framework for the analysis of compositional tables, thereby extending the methodology proposed in Hron et al. (2022). Although this framework is applicable to a broad range of methods for analysing compositional tables, the manuscript explicitly derives the equivalence between weighted LRA, as defined by Greenacre and Lewi (2009), and the SVD of the weighted clr representation of the interaction table, where the weights are given by the products of arithmetic row and column marginals.

An example of the resulting weighted correspondence analysis of the eye-tracking data is shown in the right panel of Figure 7. In terms of explained variance, the weighting reduced the influence of small proportions and resulted in a more representative ordination of the overall clr table, while preserving the discriminating roles of Tasks 2, 3, and 4 and the AOIs Maid and Kids.

5.3. Backwards pivot coordinates (Nesrstová et al., 2023)

This manuscript follows up on the work of Hron et al. (2021), which introduced backwards pivot coordinates for vector compositional data. In general, backwards pivot coordinates emphasise the role of the elemental information, which, in the case of vector compositions, is represented by pairwise log-ratios. The present manuscript extends this idea to compositional tables, where, in accordance with their definition and decomposition (11), the elemental information is represented by pairwise log-ratios of the row and column geometric marginals together with four-part log-odds ratios.

To this end, the manuscript introduces a system of orthonormal (ilr) coordinates consisting of three groups of coordinates: row backwards pivot balances (rbpb($\mathbf{x}^{(kl)}$)), column backwards pivot balances (cbpb($\mathbf{x}^{(kl)}$)), and odds-ratio backwards pivot coordinates (tbpc($\mathbf{x}^{(kl)}$)). In each group, the first coordinate highlights the contribution of the pivotal cell x_{kl} or its corresponding row and column category:

$$\begin{aligned} \text{rbpb}(\mathbf{x}^{(kl)})_i &= \sqrt{\frac{J}{2}} \mathbf{l}_{\mathbf{v}} \frac{g(\mathbf{x}_{k\bullet})}{g(\mathbf{x}_{I\bullet})}, & \text{cbpb}(\mathbf{x}^{(kl)})_i &= \sqrt{\frac{I}{2}} \mathbf{l}_{\mathbf{v}} \frac{g(\mathbf{x}_{\bullet l})}{g(\mathbf{x}_{\bullet J})}, \\ \text{tbpc}(\mathbf{x}^{(kl)})_{i1} &= \frac{1}{2} \mathbf{l}_{\mathbf{v}} \frac{x_{IJ}x_{kl}}{x_{Il}x_{kJ}}. \end{aligned} \quad (18)$$

Similarly to the vector case, multivariate methods such as principal component analysis or regression modelling can be performed in this coordinate system. Repeating the analysis under different permutations of the row and column categories does not affect the numerical results (e.g., principal component scores or fitted regression models), but enables their interpretation with respect to all sources of the elemental information.

The second part of the manuscript develops a detailed interpretation of principal component analysis and regression modelling based on backwards pivot coordinates for both vector compositions and compositional tables. For principal component analysis, it derives the interpretation of the compositional biplot. For regression models with a real-valued response and compositional covariates, it explains the interpretation of the regression coefficients when the covariates are represented in ilr coordinates. Analogous to classical regression, each coefficient describes the effect of changing one ilr coordinate while keeping the remaining coordinates fixed. The manuscript further demonstrates how this interpretation can be translated into changes in the

original compositional parts.

Although the methodology is illustrated in the paper using movement behaviour data, Figure 8 illustrates the methodology using the running eye-tracking example. Among other biplot features, the length of the rays approximates the variability of the respective coordinate. For example, the left panel of Figure 8 suggests that, among all pairwise log-ratios with the AOI Man, the relative attention allocated to Man versus Background varies the most across participants. On the other hand, the variability explained by the first two components is rather small (approximately 20%), suggesting that the biplot provides a substantially simplified representation of this particular sample of 7×7 compositional tables.

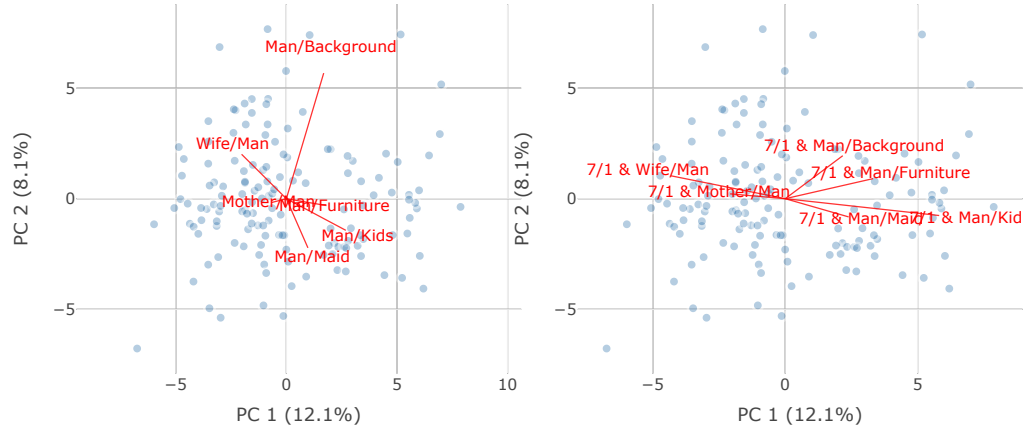


Figure 8: Compositional biplot based on backwards pivot coordinates of the relative time distribution across Tasks and AOIs. To improve readability, the left panel displays only the loadings of column backwards pivot balances for the AOI Man, whereas the right panel shows the loadings of odds-ratio backwards pivot coordinates for the same AOI and Tasks 1 and 7.

5.4. Geographically weighted regression (Fačevicová et al., 2021)

This book chapter further develops the methodology for the analysis of compositional tables by extending regression modelling to spatially dependent data. It thus represents the first step towards the analysis of spatially dependent multi-factorial compositional data. Specifically, it considers a geographically varying relationship between a real-valued response and a two-factor compositional covariate. Spatial heterogeneity is accommodated through geographically weighted regression (GWR), where local models are estimated by weighted least squares according to the spatial proximity of the observations.

Similarly to ordinary regression models with compositional covariates (Müller et al., 2018), GWR cannot be directly applied to the original compositional parts because of their relative nature. The proposed methodology therefore performs the local modelling in the ilr coordinate representation of compositional tables (14) and (15). Formally, the model is given by

$$\begin{aligned} E(Y_i \underline{\mathbf{z}}_i) = & \beta_{i,0} + \beta_{i,1}^r z_{i,1}^r + \cdots + \beta_{i,I-1}^r z_{i,I-1}^r + \beta_{i,1}^c z_{i,1}^c + \cdots + \beta_{i,J-1}^c z_{i,J-1}^c \\ & + \beta_{i,11}^{OR} z_{i,11}^{OR} + \cdots + \beta_{i,I-1,J-1}^{OR} z_{i,I-1,J-1}^{OR}. \end{aligned} \quad (19)$$

The coordinate representation (14) and (15) enables the separate modelling of the effects associated with the row factor, the column factor, and their interaction. Since the regression coefficients are estimated for the ilr coordinates, the chapter derives their interpretation in terms of changes in the original compositional table for a general system of ilr coordinates, extending the results of Nesrstová et al. (2023) beyond backwards pivot coordinates.

The methodology is illustrated using data on the relationship between the risk of falling into poverty and the demographic structure of Germany in 2017, represented by a 2×3 compositional table. Owing to the skewed distribution of the response, the application employs the simplified model

$$E[lr(Y_i)|\underline{z}_i] = \beta_{i,0} + \beta_{i,1}^r z_{i,1}^r + \beta_{i,1}^c z_{i,1}^c + \beta_{i,2}^c z_{i,2}^c + \beta_{i,11}^{OR} z_{i,11}^{OR} + \beta_{i,12}^{OR} z_{i,12}^{OR}, \quad (20)$$

whose estimated local regression coefficients are also interpreted in detail in the chapter. Selected parameter estimates are presented in Figure 9.

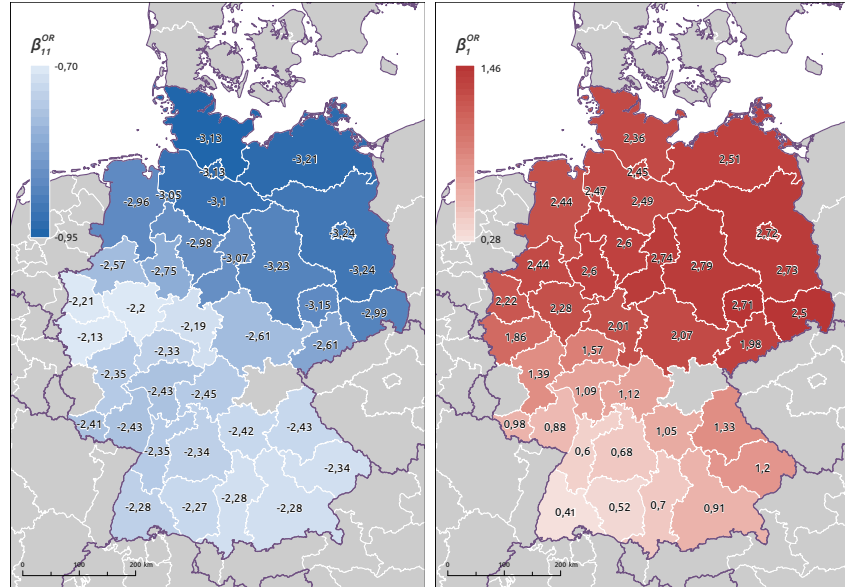


Figure 9: Maps demonstrating estimates of selected regression parameters of the local regression models. The values in the centre of each region correspond to the local t-value.

The proposed methodology combines geographically weighted regression with the orthogonal decomposition of compositional tables, enabling the analysis and interpretation of spatially varying effects associated with both the independent and interaction structures of the compositional covariate.

5.5. Independent component analysis (Muehlmann et al., 2021)

This manuscript extends independent component analysis (ICA) to compositional data, thereby linking compositional data analysis with the field of blind source separation. It considers the case of independent and identically distributed (iid) vector compositional observations. In the classical setting, ICA assumes that the observed vectors $\mathbf{x}$ follow the linear mixing model

$$\mathbf{x} = \mathbf{A}\mathbf{z} + \mathbf{b}, \quad (21)$$

where the latent components $\mathbf{z}$ are mutually independent and at most one of them is Gaussian. Many popular ICA algorithms estimate the unmixing matrix $\mathbf{W}$ by maximizing the non-Gaussianity of the recovered latent components, $\mathbf{W}(\mathbf{x} - \mathbf{b}) = \mathbf{z}$, in contrast to principal component analysis (PCA), which seeks orthogonal directions explaining the maximum variance of the data. For whitened data, the ICA decomposition is identifiable up to a permutation and sign changes of the components. The resulting latent variables can subsequently be used for data visualisation, cluster identification, outlier detection, noise removal, or as a preprocessing step for further statistical analysis.

The adaptation of ICA to iid compositional data naturally follows from their coordinate representation. Since ICA begins by whitening the data, a full-rank representation is required and the ilr coordinates (6) therefore provide a natural choice. Furthermore, different choices of the ilr basis correspond only to orthogonal rotations of the coordinate system and therefore do not affect the ICA solution itself, but only the representation and interpretation of the recovered components. For a more natural interpretation in

terms of the relative dominance of the original compositional parts, the ilr unmixing matrix can subsequently be transformed to the clr representation according to $\mathbb{W}_{clr} = \mathbb{W}_{ilr} \mathbb{V}$. The latent components are thus obtained as

$$\mathbf{z} = \mathbb{W}_{ilr} (\text{ilr}(\mathbf{x}) - \mathbf{b}) = \mathbb{W}_{ilr} (\mathbb{V} \text{clr}(\mathbf{x}) - \mathbf{b}), \quad (22)$$

where the ilr unmixing matrix is estimated using any of the existing ICA algorithms.

The methodology is in [Muehlmann et al. \(2021\)](#) demonstrated using metabolomics data from a newborn screening programme. This application was further extended in [Kouřil et al. \(2023\)](#), where compositional ICA enabled the detection of several rare inherited metabolic disorders. Figure 10 illustrates the output of compositional ICA for the eye-tracking example used throughout this thesis.

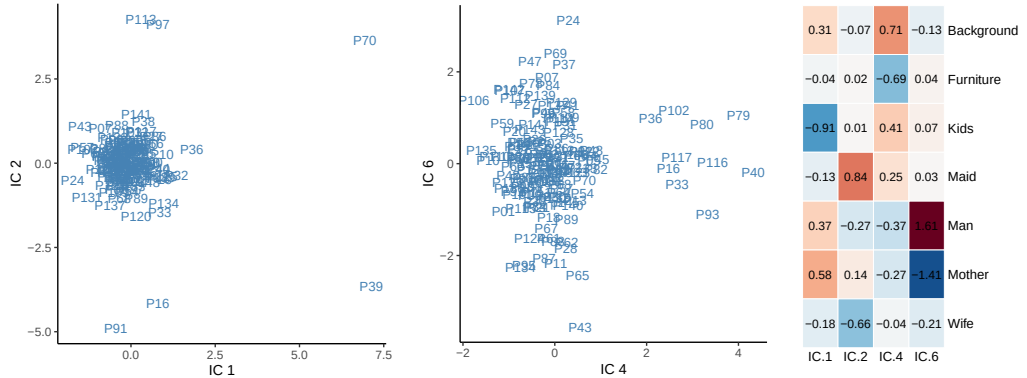


Figure 10: Scatterplots of selected independent components (ICs). The first two components primarily reveal individual outliers, whereas IC4 and IC6 identify a group of observers sharing a similar pattern of attention distribution. The corresponding clr loadings for these four components are shown in the right panel to facilitate their interpretation.

5.6. Robust blind source separation (Bábek et al., 2022)

This manuscript naturally follows [Muehlmann et al. \(2021\)](#). While the previous paper introduced independent component analysis, or more generally blind source separation (BSS), for iid compositional data, this work extends the methodology to serially dependent compositional observations while simultaneously introducing robustness against outlying measurements.

The development of the methodology was motivated by the analysis of X-ray fluorescence (XRF) scans of siliciclastic sediment cores. The study compares measurements obtained using a low-cost hand-held XRF device with traditional laboratory measurements performed on dry powdered samples, which provide higher accuracy but are considerably more expensive. Geochemical concentrations are typically reported in ppm or wt.% and therefore carry relative information. Consequently, as discussed in Section 2, compositional data analysis provides the appropriate methodological framework for their analysis.

In the considered application, two additional characteristics of the data must be taken into account. First, hand-held XRF measurements are affected by measurement noise arising from sample inhomogeneity as well as varying water content and grain size, which may also result in outlying observations. Second, the measurements are collected along the depth profile of a sediment core and can therefore naturally be regarded as a serially dependent process. To accommodate both aspects, the manuscript proposes a robust blind source separation algorithm for compositional time series. Similarly to [Muehlmann et al. \(2021\)](#), the methodology is based on the ilr coordinate representation of the compositional observations, but instead of classical ICA it employs the eSAM-SOBI algorithm (Ilmonen et al., 2015), which exploits a collection of lagged autocovariance matrices to recover the latent components. To increase

robustness, these autocovariance matrices are estimated using robust procedures. As in the iid case, the estimated ilr unmixing matrix is subsequently transformed to its clr counterpart, providing a more natural interpretation of the recovered components in terms of the original compositional parts.

Figure 11 illustrates the practical use of the proposed methodology. The unmixing matrix is estimated solely from the hand-held XRF measurements and subsequently applied to both the hand-held and powdered datasets. The resulting latent components can then be directly compared across the two measurement techniques. While e.g. the first recovered component differs substantially between the datasets and therefore mainly reflects measurement noise, the second component shows a high agreement, indicating that it captures the underlying geochemical signal that can be reliably recovered even from the less accurate but considerably cheaper hand-held measurements.

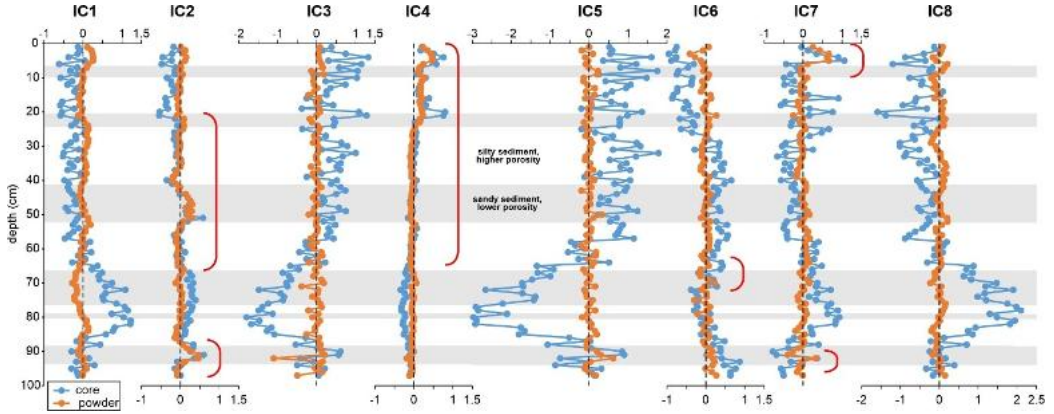


Figure 11: Selected latent components (IC) found by the robust BSS algorithm applied to the hand-held XRF core-scan data (blue dots) and powder XRF data (orange dots).

5.7. Cellwise robust covariance estimator (Rieser et al., 2023)

The issue of robustness is further addressed in manuscript [Rieser et al. \(2023\)](#). As outlined in Section 4, a single corrupted cell in the raw compositional data matrix affects the coordinate representation of the corresponding observation through the geometric mean involved in the log-ratio transformations. Recall that, in the compositional setting, a cellwise outlier corresponds to a distortion of the relative importance of the affected compositional part rather than to an unusually large or small absolute value. The manuscript therefore proposes a method for estimating the covariance structure in the presence of compositional cellwise outliers.

The proposed approach avoids robustifying the clr covariance matrix directly. Instead, it exploits the elemental information carried by pairwise log-ratios, which are affected by a cellwise contamination only locally. As introduced in (3), the variances of pairwise log-ratios form the entries of the variation matrix, one of the fundamental characteristics of a compositional sample. Moreover, the variation matrix $\mathbf{V}$ can be transformed directly to the clr covariance matrix Σ^{clr} , which forms the basis of many multivariate methods for compositional data, including principal component analysis (Filzmoser et al., 2018). More specifically,

$$\Sigma^{\text{clr}} = -\frac{1}{2} \mathbf{H} \mathbf{V} \mathbf{H}, \quad (23)$$

where $\mathbf{H}$ denotes the centering matrix used already in (5). The basic idea of the proposed estimator is therefore to robustly estimate the individual entries of the variation matrix using the Qn estimator (Rousseeuw and Croux, 1993) and subsequently obtain a cellwise robust estimate of the clr covariance

matrix through (23).

Additionally, the manuscript introduces a regularised version of the estimator intended for high-dimensional settings. A comprehensive simulation study demonstrates that, particularly under cellwise contamination, the regularised estimator outperforms both the classical covariance estimator and existing rowwise robust alternatives.

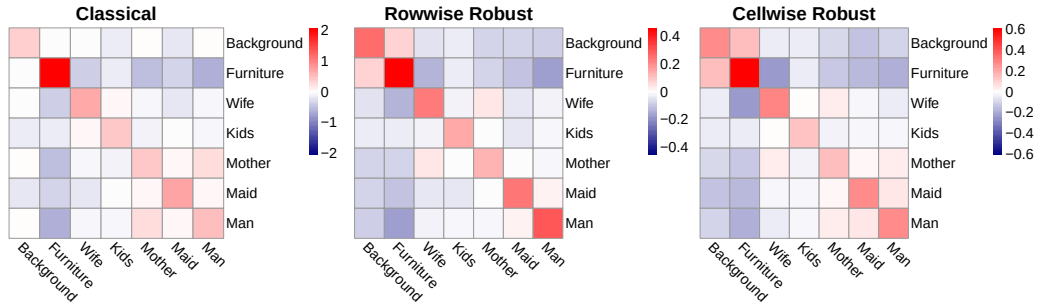


Figure 12: Comparison of different estimates of the clr covariance matrix for the first task of the running eye-tracking example.

The methodology is demonstrated in the manuscript using the well-known GEMAS geochemical dataset (Reimann et al., 2014). For consistency with the examples used throughout this thesis, Figure 12 illustrates the classical, rowwise robust, and proposed cellwise robust estimates of the clr covariance matrix for the first task of the eye-tracking experiment. The heatmaps suggest the presence of anomalous observations in the original data, resulting in an overestimation of most entries of the clr covariance matrix by the classical estimator compared with both robust alternatives. Although the proposed method does not aim to identify specific outlying cells, the most apparent differences are associated with AOI Furniture, whose clr coordinates and, consequently, pairwise log-ratios with other AOIs appear to be substantially affected by anomalous observations. At the same time, the differences between the rowwise and cellwise robust estimates are relatively

small, suggesting that the observed contamination is not limited to a small number of individual cells or log-ratios.

5.8. Robust principal component analysis (de Sousa et al., 2021)

This manuscript provides the first robust methodology for the analysis of two-factorial compositional data by introducing a casewise robust approach to principal component analysis (PCA) of a sample of compositional tables. Classical PCA of compositional data is based on the spectral decomposition of the clr or ilr covariance matrix. Since the clr covariance matrix is singular because of the zero-sum constraint, robust compositional PCA is commonly performed in ilr coordinates using a robust covariance estimator, such as the minimum covariance determinant (MCD), while the results can subsequently be transformed to the clr representation to facilitate their interpretation (Filzmoser et al., 2018).

The methodology proposed in de Sousa et al. (2021) extends this principle to compositional tables. The observations are represented in the ilr coordinate system (14) and (15), robust PCA is based on the MCD estimate of the corresponding covariance matrix, and the resulting loadings are transformed back to the clr representation. The key advantage of this approach is that the coordinate system respects the orthogonal decomposition of compositional tables (11), allowing robust PCA to be performed either for the complete tables or separately for their independent and interaction components.

To facilitate the interpretation of the resulting compositional biplots, the manuscript further derives the relationship between the clr representation of the decomposed table and the corresponding two-factor pivot coordinates

defined in Fačevicová et al. (2016), i.e.

$$\begin{aligned} \text{clr}(\text{ind}(\mathbf{x}))_{kl} &= \sqrt{\frac{I-1}{IJ}} z_i^{r(k)} + \sqrt{\frac{J-1}{IJ}} z_i^{c(l)}, \\ \text{clr}(\text{int}(\mathbf{x}))_{kl} &= \sqrt{\frac{(I-1)(J-1)}{IJ}} z_{11}^{OR(kl)}, \end{aligned} \quad (24)$$

where the superscripts (k) and (l) indicate the permutation of rows and columns, respectively, for which the cell $[k, l]$ is moved to the pivot position $[1, 1]$.

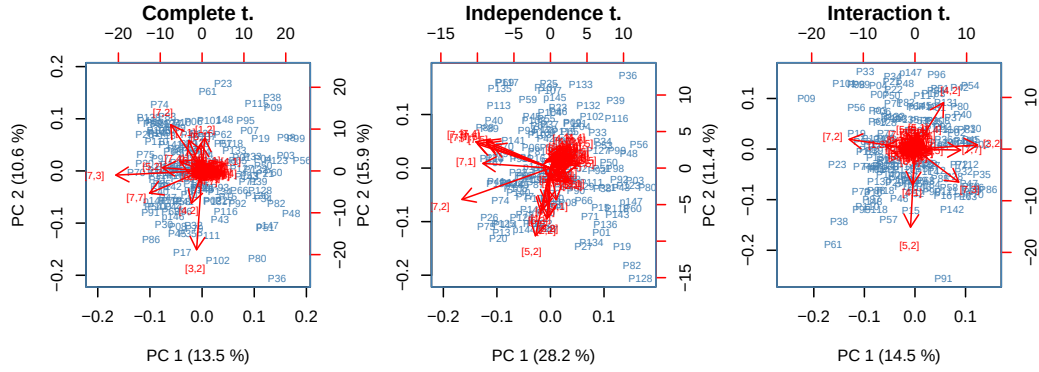


Figure 13: Robust covariance biplots based on the complete compositional tables (left panel) and their decomposed parts (middle and right panels). Observations are labelled by observers' ID, whereas loadings are labelled according to the row and column position of the corresponding cell in the compositional table.

Figure 13 illustrates the results of robust PCA applied to the compositional tables corresponding to the attention distribution in the running eye-tracking example. In addition to the robust PCA of the complete tables, the figure demonstrates the separate analyses of their independent and interaction parts. For illustration, let us briefly focus on the left panel of

Figure 13, which, similarly to Figure 8, provides a two-dimensional representation of the set of 7×7 compositional tables. In this case, however, the rays correspond to the relative importance of individual cells in the clr sense, and the respective loadings are estimated robustly. The resulting biplot suggests that the inter-participant variability is substantially driven by the relative importance of cells $[7, 3]$ and $[3, 2]$, corresponding to AOI Kids in Task 7 and AOI Furniture in Task 3, respectively. Nevertheless, similarly to the example in Section 5.3, the percentage of variability explained by the first two robust principal components is very low. The resulting biplot therefore provides only a limited representation of the overall data structure.

5.9. Zero-handling strategies (Tang et al., 2026)

This manuscript contributes to the methodological treatment of compositional data in the presence of zero values. As already discussed in Section 2, the traditional definition of vector compositional data, as well as its multi-factorial generalisation, assumes strictly positive observations. Besides violating this assumption, zero values also prevent the direct application of log-ratio methodology. Moreover, even small positive values can substantially affect subsequent analyses, as small changes in their magnitude may translate into large changes in log-ratios. The manuscript therefore provides a critical comparison of zero-handling strategies, with a particular focus on high-dimensional compositional count data. The review is motivated by microbiome data, where observations are typically counts and their total is determined by the sequencing and sampling process rather than representing a meaningful biological quantity, which motivates their compositional treatment (Gloor et al., 2017). At the same time, the number of measured features is typically much larger than the number of observations, resulting

in a high-dimensional and often sparse data structure.

More specifically, the manuscript compares several approaches to zero imputation, as well as direct modelling techniques for zero-inflated data and alternative transformations that allow zero values to be retained. The methods are compared both theoretically and empirically, using a high-dimensional microbiome dataset and a simulation study based on real microbial compositions.

The empirical part focuses on a model-agnostic comparison of representative matrix-based imputation methods within the classical log-ratio framework. Their performance is assessed not only in terms of reconstruction accuracy, but also with respect to the preservation of covariance structure and the results of downstream analyses, including PCA and clustering. An important finding is that the reconstruction accuracy alone does not necessarily predict the quality of subsequent statistical analyses. Figure 14 illustrates one outcome of the simulation study, showing how the accuracy of selected methods changes with an increasing proportion of zero values. Among other findings, the results demonstrate the relatively low reconstruction accuracy of the PLS-based imputation method (Templ et al., 2016). This accuracy can be substantially improved by applying a ceiling operation, while the resulting imputed data nevertheless perform well in subsequent downstream analyses.

Overall, the results reveal a trade-off between accurate numerical reconstruction and preservation of the underlying count structure, highlighting the need for methods that simultaneously account for compositionality, zero inflation, and the discrete nature of sequencing data. A better understanding of zero-handling strategies is also an important prerequisite for the further development of methods for multi-factorial compositions, where the occurrence of zero values poses an analogous methodological challenge and may

be particularly relevant for count-valued data.

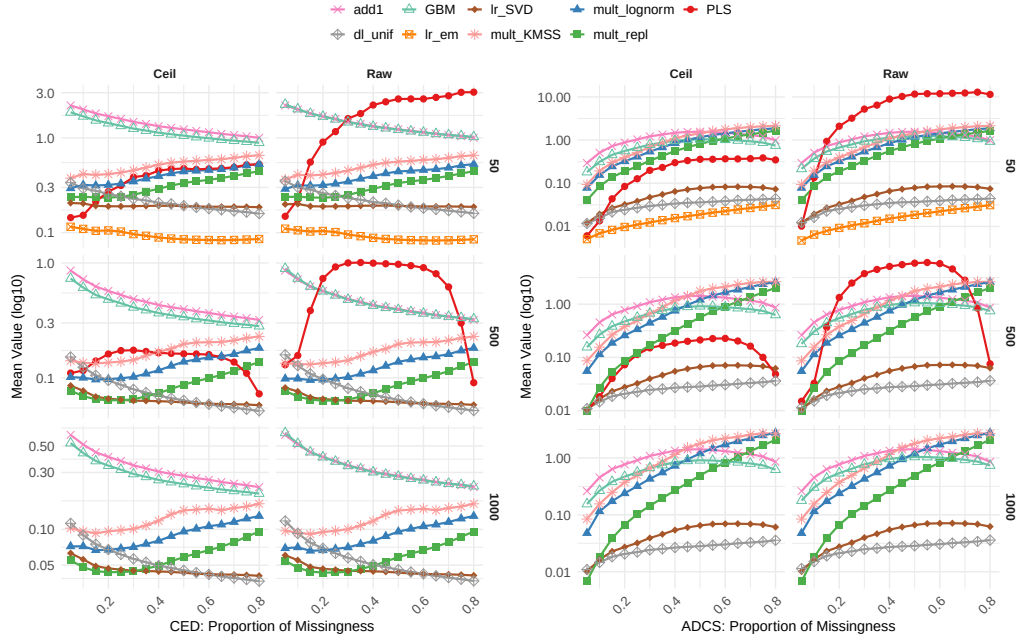


Figure 14: Comparison of imputation methods under varying dimensionality and missingness probabilities.

5.10. Application to geochemistry (Fačevicová et al., 2016)

This manuscript demonstrates the application of compositional data analysis to the geochemical characterisation of the Devonian/Carboniferous boundary based on elemental geochemistry. The analysis is based on 1,884 samples collected from 14 European sections, each described by the concentrations of eighteen chemical elements. Using principal component analysis and Q-mode clustering, i.e. clustering of variables or compositional parts, the manuscript compares the results obtained from raw concentrations, log-transformed data, and the compositional approach. Although compositional methodology provides the theoretically appropriate framework for geochemical data, the comparison shows that, for this particular application, analyses based on raw or

log-transformed concentrations may still provide valuable complementary information. Rather than representing competing approaches, the classical and compositional analyses therefore provide a more comprehensive interpretation when used together.

Principal component analysis and the corresponding compositional bi-plots are primarily employed to explore differences between localities and between geological layers within individual sections. Beyond the exploratory analysis, the manuscript proposes a data-driven strategy for constructing geochemical proxies capable of distinguishing the studied stratigraphic layers. The approach is based on the variation matrix (3), whose entries quantify the variability of pairwise log-ratios between chemical elements. Since the variation matrix measures the similarity of the relative behaviour of individual elements, it naturally serves as the input for Q-mode clustering, which groups together elements exhibiting stable relative relationships throughout the dataset.

Balances (7) constructed between groups of elements belonging to different clusters subsequently serve as candidate geochemical proxies. Their ability to distinguish individual geological layers is illustrated in Figure 15, which shows that the proposed balances clearly separate the four Hangenberg event layers at the Kronhofgraben locality. In contrast, the corresponding plots based on conventionally used geochemical proxies exhibit substantially weaker separation. The proposed workflow therefore provides an objective, data-driven procedure for constructing geochemical proxies while demonstrating that compositional data analysis can yield additional geological insight beyond that obtained from classical geochemical methods.

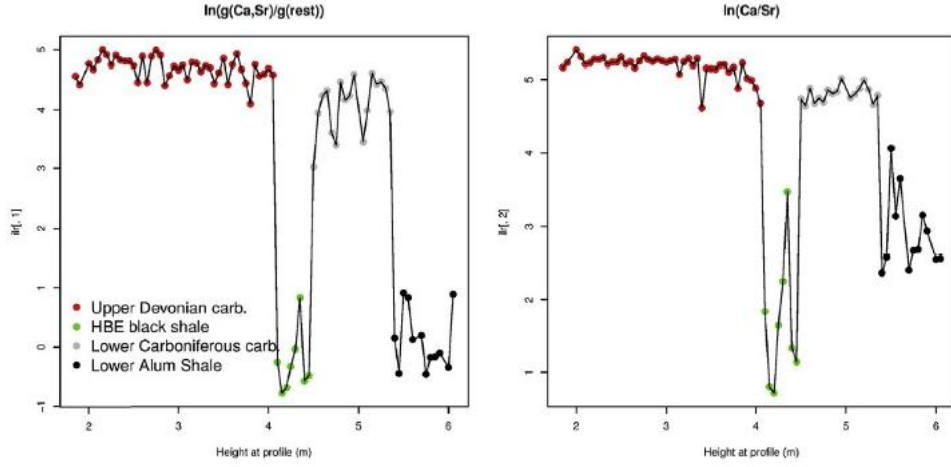


Figure 15: Log-ratios defined based on the Q-mode clustering of measurements from location Kronhofgraben.

5.11. Application to eye-tracking (Fačevicová et al., 2026)

This manuscript represents the first systematic introduction of compositional data analysis to the eye-tracking community. It is conceived as a tutorial, explaining the fundamental principles of vector compositional data analysis and demonstrating them on a newly published dataset (Popelka et al., 2026), which was briefly introduced in Section 2.1 and which served as a running example throughout the thesis. The principal research question inherited from Yarbus’ original experiment is whether the viewing task influences the observers’ visual attention strategy.

Similarly to Fačevicová et al. (2016), the manuscript compares analyses based on absolute and relative information, emphasizing that the appropriate analytical framework should always reflect the underlying research question. While analyses of absolute fixation times are appropriate, for example, for studying the overall complexity or duration of the observation process, the compositional approach focuses on the relative allocation of visual attention across the stimulus and is therefore better suited for investigating differences

in viewing strategies.

The manuscript establishes a complete workflow for the compositional analysis of eye-tracking data, covering descriptive statistics, exploratory methods, hypothesis testing, and regression modelling. It introduces the principal tools of descriptive compositional data analysis, including ternary diagrams, the compositional mean and covariance structure, the variation matrix, and total variance. It further demonstrates the use of multivariate methods such as principal component analysis, cluster analysis, discriminant analysis, MANOVA, and regression modelling within the CoDA framework, providing guidance on the interpretation of their outputs. To facilitate the adoption of the methodology by the eye-tracking community, the complete analysis is accompanied by publicly available R code. Although methodological presentation is the primary aim of the paper, the analyses consistently reveal substantial differences in attention allocation across the experimental tasks, both when absolute fixation times and their compositional representation are analysed. Figure 16 illustrates one example by comparing the effect of the viewing task on the relative and absolute time allocated to the Area of Interest Man.

The manuscript thus establishes compositional data analysis as a comprehensive methodological framework for the statistical analysis of eye-tracking data. At the same time, it naturally opens the way to further methodological developments presented in this habilitation thesis. In particular, repeated observations of the same stimulus under different experimental conditions can be represented as compositional tables, allowing the application of the methods introduced in Fačevicová et al. (2018), [Fačevicová et al. \(2023\)](#), [Fačevicová et al. \(2025\)](#), [Nesrstová et al. \(2023\)](#), and [de Sousa et al. \(2021\)](#). Likewise, the presence of atypical observers or measurement artefacts mo-

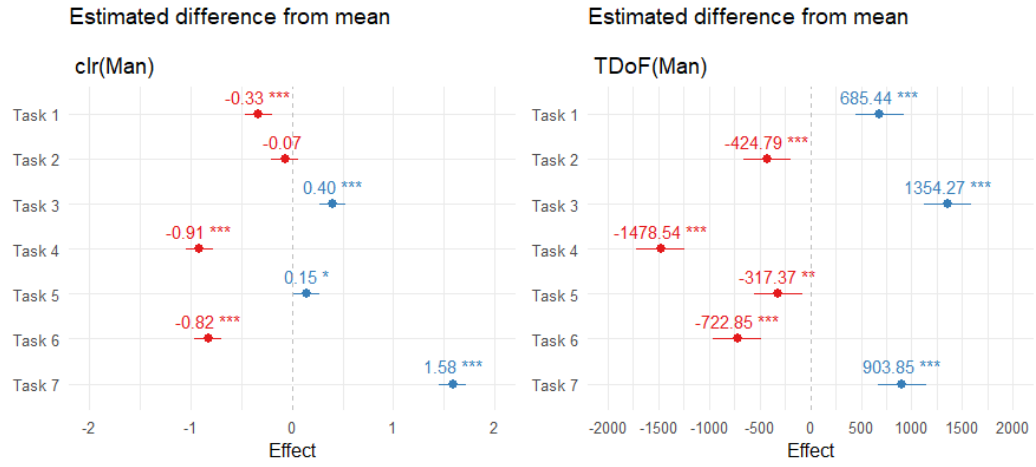


Figure 16: Estimated fixed effects of the linear mixed-effects model for the relative (left) and absolute (right) time spent in the AOI Man across task types.

tivates the use of robust compositional methods, such as those proposed in [Rieser et al. \(2023\)](#). The paper therefore not only introduces CoDA to the eye-tracking community but also establishes a foundation for transferring a broad range of advanced compositional methodology to this rapidly developing research field.

6. Summary of the main results and future research directions

This thesis is based on eleven scientific publications co-authored by the applicant, comprising theoretical, methodological, and applied contributions to compositional data analysis. Their common theme is the analysis of relative-valued data, with particular emphasis on the methodology of multi-factorial compositions.

A substantial part of the presented research contributes to the development of the theory of compositional tables. The manuscripts [Fačevicová et al. \(2021\)](#), [Fačevicová et al. \(2023\)](#), [Nesrstová et al. \(2023\)](#), and [de Sousa et al. \(2021\)](#) extend the original framework introduced in Section 3 towards regression modelling, interpretation, robust dimension reduction, and the analysis of general multi-factorial compositions. In turn, [Fačevicová et al. \(2025\)](#) establishes a theoretical connection between the interaction part of a compositional table and correspondence analysis, thereby linking compositional methodology with a widely used multivariate statistical technique.

Another research direction concerns the development of statistical methodology for vector compositional data with clear potential for future extension to multi-factorial compositions. This includes blind source separation ([Muehlmann et al. \(2021\)](#), [Bábek et al. \(2022\)](#)) and cellwise robust covariance estimation ([Rieser et al. \(2023\)](#)), both of which naturally motivate further methodological developments in the multi-factorial setting. Finally, [Fačevicová et al. \(2026\)](#) introduces compositional data analysis to the field of eye-tracking research. Besides demonstrating the advantages of CoDA for analysing visual attention distributions, it establishes a new application area in which many of the methods presented throughout this thesis can be further

developed and applied.

Despite the substantial methodological progress achieved in recent years, the analysis of multi-factorial compositional data remains an active research area with several open questions. One important topic concerns the treatment of zeros, where the recent work of [Tang et al. \(2026\)](#) provides a promising starting point. Another natural direction is a deeper investigation of the relationship between multi-factorial compositions and log-linear or multinomial regression models. For example, [Fačevicová et al. \(2026\)](#) points out the close connection between log-linear mixed-effects models and the mean clr interaction table, suggesting that these approaches deserve a more systematic comparison. Further opportunities arise from methodological links with tensor data analysis (Kolda and Bader, 2009), while additional extensions may emerge through connections with fields such as ecological inference (Rosen et al., 2001) or dependence measures based on distance correlation (Székely et al., 2007). Together, these directions demonstrate that the methodology of multi-factorial compositional data analysis is still far from complete and continues to offer numerous opportunities for further theoretical development as well as new interdisciplinary applications.

References

- Fačevicová, K., Filzmoser, P., Hron, K.: Compositional cubes: A new concept for multi-factorial compositions. *Statistical Papers* **64**(3), 955–985 (2023) <https://doi.org/10.1007/s00362-022-01350-8>
- Fačevicová, K., Filzmoser, P., Hron, K.: Correspondence analysis from the viewpoint of compositional tables. *Statistical Analysis and Data Mining* **18**(4), 117–130 (2025) <https://doi.org/10.1002/sam.70023>
- Nesrstová, V., Jašková, P., Pavlů, I., Hron, K., Palarea-Albaladejo, J., Gába, A., Pelclová, J., Fačevicová, K.: Simple enough, but not simpler: Reconsidering additive logratio coordinates in compositional analysis. *SORT: Statistics and Operations Research Transactions* **47**(2) (2023) <https://doi.org/10.57645/20.8080.02.9>
- Fačevicová, K., Kynčlová, P., Macků, K.: Geographically weighted regression analysis for two-factorial compositional data. In: Filzmoser, P., Hron, K., Martin-Fernández, J.A., Palarea-Albaladejo, J. (eds.) *Advances in Compositional Data Analysis. Festschrift in Honour of Vera Pawlowsky-Glahn*. Springer, Cham (2021). Chap. 6. https://doi.org/10.1007/978-3-030-71175-7_6
- Muehlmann, C., Fačevicová, K., Gardlo, A., Janečková, H., Nordhausen, K.: Independent component analysis of compositional data. In: Daouia, A., Ruiz-Gazen, A. (eds.) *Advances in Contemporary Statistics and Econometrics*, pp. 525–545. Springer, Cham (2021). Chap. 27. https://doi.org/10.1007/978-3-030-73249-3_27
- Bábek, O., Fačevicová, K., Židek, M., Sedláček, J., Muehlmann, C., Nordhausen, K., Hron, K.: X-ray fluorescence scanning of soft and wet-sediment cores in terrestrial environments; A robust blind source separation approach. *Journal of Geochemical Exploration* **243**, 107106 (2022) <https://doi.org/10.1016/j.gexplo.2022.107106>
- Rieser, C., Fačevicová, K., Filzmoser, P.: Cell-wise robust covariance estimation for compositions, with application to geochemical data. *Journal of Geochemical Exploration* **253**, 107299 (2023) <https://doi.org/10.1016/j.gexplo.2023.107299>
- de Sousa, J., Hron, K., Fačevicová, K., Filzmoser, P.: Robust principal components analysis for compositional tables. *Journal of Applied Statistics* **48**(2), 214–233 (2021) <https://doi.org/10.1080/02664763.2020.1722078>

- Tang, W., Fačevicová, K., Nordhausen, K., Taskinen, S.: A critical comparison of handling zeros in high-dimensional compositional count data. *Chemometrics and Intelligent Laboratory Systems* (277), 105821 (2026) <https://doi.org/10.1016/j.chemolab.2026.105821>
- Fačevicová, K., Bábek, O., Hron, K., Kumpan, T.: Element chemostratigraphy of the Devonian/Carboniferous boundary - A compositional approach. *Applied Geochemistry* **75**, 211–221 (2016) <https://doi.org/10.1016/j.apgeochem.2016.10.002>
- Fačevicová, K., Vymazal, J., Popelka, S.: Advancing eye movement analysis through compositional modeling: A new perspective on Yarbus' classic study. *Behavior Research Methods* **58**(173), 1–28 (2026) <https://doi.org/10.3758/s13428-026-03054-5>
- Aitchison, J.: The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)* **44**(2), 139–160 (1982) <https://doi.org/10.1111/j.2517-6161.1982.tb01195.x>
- Buccianti, A., Gozzi, C.: The whole versus the parts: The challenge of compositional data analysis (coda) methods for geochemistry. In: Filzmoser, P., Hron, K., Martín-Fernández, J.A., Palarea-Albaladejo, J. (eds.) *Advances in Compositional Data Analysis. Festschrift in Honour of Vera Pawlowsky-Glahn*. Springer, Cham (2021). Chap. 13. https://doi.org/10.1007/978-3-030-71175-7_13
- Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., Egozcue, J.J.: Microbiome datasets are compositional: And this is not optional. *Frontiers in Microbiology* **8**(2224) (2017) <https://doi.org/10.3389/fmicb.2017.02224>
- Dumuid, D., Pedišić, , Palarea-Albaladejo, J., Martín-Fernández, J.A., Hron, K., Olds, T.: Compositional data analysis in time-use epidemiology: What, why, how. *International Journal of Environmental Research and Public Health* **17**(7), 2220 (2020) <https://doi.org/10.3390/ijerph17072220>
- Coenders, G., Egozcue, J.J., Fačevicová, K., Navarro-López, C., Palarea, J. Albaladejo, Pawlowsky-Glahn, V., Tolosana-Delgado, R.: 40 years after Aitchison's article "The statistical analysis of compositional data". Where we are and where we are heading. *SORT: Statistics and Operations Research Transactions* **47**(2) (2023) <https://doi.org/10.57645/20.8080.02.6>

- Fačevicová, K., Hron, K., Todorov, V., Templ, M.: General approach to coordinate representation of compositional tables. *Scandinavian Journal of Statistics* **45**(4), 879–899 (2018) <https://doi.org/10.1111/sjos.12326>
- Maronna, R.A., Martin, R.D., Yohai, V.J., Salibián-Barrera, M.: *Robust Statistics: Theory and Methods (with R)*, 2nd edn. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ (2019). <https://doi.org/10.1002/9781119214656>
- Rousseeuw, P., Bossche, W.: Detecting deviating data cells. *Technometrics* **60**(2), 135–145 (2018) <https://doi.org/10.1080/00401706.2017.1340909>
- Tolosana-Delgado, R.: Uses and misuses of compositional data in sedimentology. *Sedimentary Geology* **280**, 60–79 (2012) <https://doi.org/10.1016/j.sedgeo.2012.05.005>
- Kouřil, S., de Sousa, J., Fačevicová, K., Gardlo, A., Muehlmann, C., Nordhausen, K., Friedecký, D., Adam, T.: Multivariate independent component analysis identifies patients in newborn screening equally to adjusted reference ranges. *International Journal of Neonatal Screening* **9**(4) (2023) <https://doi.org/10.3390/ijns9040060>
- Fairclough, S.J., Dumuid, D., Taylor, S., Curry, W., McGrane, B., Stratton, G., Maher, C., Olds, T.: Fitness, fatness and the reallocation of time between children’s daily movement behaviours: An analysis of compositional data. *International Journal of Behavioral Nutrition and Physical Activity* **14**(64), 1–12 (2017) <https://doi.org/10.1186/s12966-017-0521-z>
- Popelka, S., Fačevicová, K., Vymazal, J., Beitlová, M.: Visitor and kangaroo as a large-scale AOI-structured Yabus replication dataset. In: *Proceedings of the 2026 Symposium on Eye Tracking Research and Applications. ETRA ’26*. Association for Computing Machinery, New York, NY, USA (2026). <https://doi.org/10.1145/3797246.3806792>
- Yabus, A.L.: *Eye Movement and Vision*. Springer, New York, NY (USA) (1967). <https://doi.org/10.1007/978-1-4899-5379-7>
- Duchowski, A.T.: *Eye Tracking Methodology: Theory and Practice* (3rd Ed.). Springer, London (2017). <https://doi.org/10.1007/978-3-319-57883-5>
- Vojtechovska, M., Popelka, S.: Gazeplotter: An open-source solution for the automatic generation of scarf plots from eye-tracking data.

- Behavior Research Methods **58**(85) (2026) <https://doi.org/10.3758/s13428-026-02959-5>
- Filzmoser, P., Hron, K., Templ, M.: Applied Compositional Data Analysis. Springer, Cham (2018). <https://doi.org/10.1007/978-3-319-96422-5>
- Pearson, K.: Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London **60** (1896 – 1897)
- Vencálek, O., Hron, K., Filzmoser, P.: A comparison of generalised linear models and compositional models for ordered categorical data. Statistical Modeling **20**(3), 249–273 (2020) <https://doi.org/10.1177/1471082X18816540>
- Egozcue, J.J., Díaz-Barrero, J.L., Pawłowsky-Glahn, V.: Compositional analysis of bivariate discrete probabilities. In: Proceedings of CoDa-Work'08, The 3rd Compositional Data Workshop. University of Girona, Spain (2008)
- van den Boogaart, K.G., Egozcue, J.J., Pawłowsky-Glahn, V.: Bayes Hilbert spaces. Australian and New Zealand Journal of Statistics **56**, 171–194 (2014) <https://doi.org/10.1111/anzs.12074>
- Pavlu, I., Menafooglio, A., Bongiorno, E.G., Hron, K.: Classification of probability density functions in the framework of Bayes spaces: Methods and applications. SORT-Statistics and Operations Research Transactions **47**(2), 295–322 (2023) <https://doi.org/10.57645/20.8080.02.10>
- Egozcue, J.J.: Reply to “On the Harker variation diagrams” by J.A. Cortés. Mathematical Geosciences **41**, 829–834 (2009) <https://doi.org/10.1007/s11004-009-9238-0>
- Billheimer, D., Guttorp, P., Fagan, W.F.: Statistical interpretation of species composition. Journal of the American Statistical Association **96**(456), 1205–1214 (2001)
- Pawłowsky-Glahn, V., Egozcue, J.J.: Geometric approach to statistical analysis on the simplex. Stochastic Environmental Research and Risk Assessment **15**(5), 384–398 (2001) <https://doi.org/10.1007/s004770100077>

- Hron, K., Templ, M., Filzmoser, P.: Imputation of missing values for compositional data using classical and robust methods. *Computational Statistics and Data Analysis* **54**(12), 3095–3107 (2010) <https://doi.org/10.1016/j.csda.2009.11.023>
- Palarea-Albaladejo, J., Martín-Fernández, J.A., Gómez-García, J.A.: Parametric approach for dealing with compositional rounded zeros. *Mathematical Geology* **39**(7), 625–645 (2007) <https://doi.org/10.1007/s11004-007-9100-1>
- Palarea-Albaladejo, J., Martín-Fernández, J.A.: A modified EM algorithm for replacing rounded zeros in compositional data sets. *Computers and Geosciences* **34**(8), 2233–2251 (2008) <https://doi.org/j.cageo.2007.09.015>
- Anderson, T.W., Olkin, I., Fang, K.T. (eds.): *Multivariate Analysis and Its Applications* vol. 24. Institute of Mathematical Statistics, Hayward, CA (USA) (1994)
- Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R.: *Modelling and Analysis of Compositional Data*. John Wiley & Sons, Chichester, UK (2015). <https://doi.org/10.1002/9781119003144>
- Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barcelo-Vidal, C.: Isometric logratio transformations for compositional data analysis. *Mathematical Geology* **35**, 279–300 (2003) <https://doi.org/10.1023/A:1023818214614>
- Egozcue, J.J., Pawlowsky-Glahn, V.: Groups of parts and their balances in compositional data analysis. *Mathematical Geology* **37**(7), 795–828 (2005) <https://doi.org/10.1007/s11004-005-7381-9>
- Fišerová, E., Hron, K.: On interpretation of orthonormal coordinates for compositional data. *Mathematical Geosciences* **43**, 455–468 (2011) <https://doi.org/10.1007/s11004-011-9333-x>
- Hron, K., Coenders, G., Filzmoser, P., Palarea-Albaladejo, J., Faměra, M.: Analysing pairwise logratios revisited. *Mathematical Geosciences* **53**, 1643–1666 (2021) <https://doi.org/10.1007/s11004-021-09938-w>
- Egozcue, J.J., Pawlowsky-Glahn, V., Templ, M., Hron, K.: Independence in contingency tables using simplicial geometry. *Communications in Statistics – Theory and Methods* **44**, 3978–3996 (2015) <https://doi.org/10.1080/03610926.2013.824980>

- Genest, C., Hron, K., Nešlehová, J.: Orthogonal decomposition of multivariate densities in Bayes spaces and relation with their copula-based representation. *Journal of Multivariate Analysis* **198**, 105228 (2023) <https://doi.org/10.1016/j.jmva.2023.105228>
- Fačevicová, K., Hron, K., Todorov, V., Dong, G., Templ, M.: Logratio approach to statistical analysis of 2x2 compositional tables. *Journal of Applied Statistics* **41**(5), 944–958 (2014) <https://doi.org/10.1080/02664763.2013.856871>
- Fačevicová, K., Hron, K., Todorov, M., Templ, M.: Compositional tables analysis in coordinates. *Scandinavian Journal of Statistics* **43**(4), 962–977 (2016) <https://doi.org/10.1111/sjos.12223>
- Egozcue, J.J., Maldonado, W.L.: An interpretable orthogonal decomposition of positive square matrices. In: Filzmoser, P., Hron, K., Martín-Fernández, J.A., Palarea-Albaladejo, J. (eds.) *Advances in Compositional Data Analysis. Festschrift in Honour of Vera Pawlowsky-Glahn*. Springer, Cham (2021). Chap. 1. https://doi.org/10.1007/978-3-030-71175-7_1
- Nakamura, K., Nakagawa, T., Tahata, K.: Symmetry of square contingency tables using simplicial geometry. *Austrian Journal of Statistics* **53**(4), 85–98 (2024) <https://doi.org/10.17713/ajs.v53i4.1845>
- Nakamura, K., Nakagawa, T., Tahata, K.: Quasi-symmetry and geometric marginal homogeneity: A simplicial approach to square contingency tables. *Information Geometry* (2025) <https://doi.org/10.1007/s41884-025-00176-1>
- Filzmoser, P., Hron, K.: Outlier detection for compositional data using robust methods. *Mathematical Geosciences* **40**, 233–248 (2008) <https://doi.org/10.1007/s11004-007-9141-5>
- Filzmoser, P., Gregorich, M.: Multivariate outlier detection in applied data analysis: Global, local, compositional and cellwise outliers. *Mathematical Geosciences* **52**, 1049–1066 (2020) <https://doi.org/10.1007/s11004-020-09861-6>
- Filzmoser, P., Hron, K., Reimann, C.: Principal component analysis for compositional data with outliers. *Environmetrics* **20**, 621–632 (2009) <https://doi.org/10.1002/env.966>

- Filzmoser, P., Hron, K., Templ, M.: Discriminant analysis for compositional data and robust parameter estimation. *Computational Statistics* **27**, 585–604 (2011) <https://doi.org/10.1007/s00180-011-0279-8>
- Boogaart, G., Filzmoser, P., Hron, K., Templ, M., Tolosana-Delgado, R.: Classical and robust regression analysis with compositional data. *Mathematical Geosciences* **53**, 823–858 (2021) <https://doi.org/10.1007/s11004-020-09895-w>
- Sealy, J.L., Caritat, P., Grunsky, E.C., Tsagis, M.T., Welsh, A.H.: Robust principal component analysis for power transformed compositional data. *Journal of the American Statistical Association* **110**(509), 136–148 (2015) <https://doi.org/10.1080/01621459.2014.990563>
- Alqallaf, F., van Aalest, S., Yohai, V.J., Zamar, R.H.: Propagation of outliers in multivariate data. *The Annals of Statistics* **37**(1), 311–331 (2009) <https://doi.org/10.1214/07-A0588>
- Raymaekers, J., Rousseeuw, P.J.: Challenges of cellwise outliers. *Econometrics and Statistics* **38**, 6–25 (2026) <https://doi.org/10.1016/j.ecosta.2024.02.002>
- Raymaekers, J., Rousseeuw, P.J.: The cellwise minimum covariance determinant estimator. *Journal of the American Statistical Association* **548**(119), 2610–2621 (2024) <https://doi.org/10.1080/01621459.2023.2267777>
- Rousseeuw, P.: Multivariate estimation with high breakdown point. In: Grossmann, W., Pflug, G., Vincze, I., Wertz, W. (eds.) *Mathematical Statistics and Applications vol. B*, pp. 283–297. Reidel, Dordrecht (1985)
- Templ, M.: Log-ratio propagation on the simplex: A theory of cellwise contamination for compositional data (2026). <https://arxiv.org/abs/2605.31345>
- Walach, J., Filzmoser, P., Kouřil, Š., Friedecký, D., Adam, T.: Cellwise outlier detection and biomarker identification in metabolomics based on pairwise log ratios. *Journal of Chemometrics* **34**, 3182 (2019) <https://doi.org/10.1002/cem.3182>
- Štefelová, N., Alfons, A., Palarea-Albaladejo, J., Filzmoser, P., Hron, K.: Robust regression with compositional covariates including cellwise outliers. *Advances in Data Analysis and Classification* **15**, 869–909 (2021) <https://doi.org/10.1007/s11634-021-00436-9>

- Greenacre, M.: Log-ratio analysis is a limiting case of correspondence analysis. *Mathematical Geosciences* **42**, 129–134 (2010) <https://doi.org/10.1007/s11004-008-9212-2>
- Hron, K., Menafoglio, A., Palarea-Albaladejo, J., Filzmoser, P., Talská, R., Egozcue, J.: Weighting of parts in compositional data analysis: Advances and applications. *Mathematical Geosciences* **54**, 71–93 (2022) <https://doi.org/10.1007/s11004-021-09952-y>
- Greenacre, M., Lewi, P.: Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio-scale measurements. *Journal of Classification* **26**, 29–54 (2009) <https://doi.org/10.1007/s00357-009-9027-y>
- Müller, I., Hron, K., Fišerová, E., Šmahaj, J., Cakirpaloglu, P., Vančáková, J.: Interpretation of compositional regression with application to time budget analysis. *Austrian Journal of Statistics* **47**(2), 3–19 (2018) <https://doi.org/10.17713/ajs.v47i2.652>
- Ilmonen, P., Nordhausen, K., Oja, H., Theis, F.: An affine equivariant robust second-order BSS method. In: Vincent, E., Yeredor, A., Koldovský, Z., Tichavský, P. (eds.) *Latent Variable Analysis and Signal Separation. LVA/ICA 2015*, pp. 328–335. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22482-4_38
- Rousseeuw, P.J., Croux, C.: Alternatives to the median absolute deviation. *Journal of the American Statistical Association* **88**(424), 1273–1283 (1993) <https://doi.org/10.2307/2291267>
- Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P. (eds.): *Chemistry of Europe's Agricultural Soils - Part A: Methodology and Interpretation of the GEMAS Data Set*. Schweizerbart Science Publishers, Stuttgart, Germany (2014)
- Templ, M., Hron, K., Filzmoser, P., Gardlo, A.: Imputation of rounded zeros for high-dimensional compositional data. *Chemometrics and Intelligent Laboratory Systems* **155**, 183–190 (2016) <https://doi.org/10.1016/j.chemolab.2016.04.011>
- Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. *SIAM Review* **51**(3) (2009) <https://doi.org/10.1137/07070111X>

- Rosen, O., Jiang, W., King, G., Tanner, M.A.: Bayesian and frequentist inference for ecological inference: The $R \times C$ case. *Statistica Neerlandica* **55**, 134–156 (2001) <https://doi.org/10.1111/1467-9574.00162>
- Székely, G.J., Rizzo, M.L., Bakirov, N.K.: Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35**(6), 2769–2794 (2007) <https://doi.org/10.1214/009053607000000505>

A. Full texts of the selected publications

Compositional cubes

Fačevicová, K., Filzmoser, P., Hron, K. Compositional cubes: A new concept for multi-factorial compositions. *Statistical Papers*, 64(3), 955–985 (2023). <https://doi.org/10.1007/s00362-022-01350-8>

[IF 1.5 Q2; AIS 0.586 Q3]

[7 citations on WoS; 3 citations without autocitations]

[Author's contribution: 50%]

Abstract

Compositional data are commonly known as multivariate observations carrying relative information. Even though the case of vector or even two-factorial compositional data (compositional tables) is already well described in the literature, there is still a need for a comprehensive approach to the analysis of multi-factorial relative-valued data. Therefore, this contribution builds around the current knowledge about compositional data a general theoretical framework for k-factorial compositional data. As a main finding it turns out that, similar to the case of compositional tables, also the multi-factorial structures can be orthogonally decomposed into an independent and several interactive parts and, moreover, a coordinate representation allowing for their separate analysis by standard analytical methods can be constructed. For the sake of simplicity, these features are explained in detail for the case of three-factorial compositions (compositional cubes), followed by an outline covering the general case. The three-dimensional structure is analyzed in depth in two practical examples, dealing with systems of spatial and time dependent compositional cubes. The methodology is implemented in the R package `robCompositions`.



Compositional cubes: a new concept for multi-factorial compositions

Kamila Fačevicová¹ · Peter Filzmoser² · Karel Hron¹

Received: 25 January 2022 / Revised: 24 May 2022 / Accepted: 26 July 2022 /

Published online: 11 August 2022

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract

Compositional data are commonly known as multivariate observations carrying relative information. Even though the case of vector or even two-factorial compositional data (compositional tables) is already well described in the literature, there is still a need for a comprehensive approach to the analysis of multi-factorial relative-valued data. Therefore, this contribution builds around the current knowledge about compositional data a general theoretical framework for k -factorial compositional data. As a main finding it turns out that, similar to the case of compositional tables, also the multi-factorial structures can be orthogonally decomposed into an independent and several interactive parts and, moreover, a coordinate representation allowing for their separate analysis by standard analytical methods can be constructed. For the sake of simplicity, these features are explained in detail for the case of three-factorial compositions (compositional cubes), followed by an outline covering the general case. The three-dimensional structure is analyzed in depth in two practical examples, dealing with systems of spatial and time dependent compositional cubes. The methodology is implemented in the R package `robCompositions`.

Keywords Analysis of independence · Compositional data · Coordinate representation · Orthogonal decomposition

Kamila Fačevicová
kamila.facevicova@gmail.com

Peter Filzmoser
Peter.Filzmoser@tuwien.ac.at

Karel Hron
karel.hron@upol.cz

¹ Department of Mathematical Analysis and Applications of Mathematics, Palacký University Olomouc, 17. listopadu 12, 77146 Olomouc, Czech Republic

² Institute of Statistics and Mathematical Methods in Economics, TU Wien, Wiedner Hauptstraße 8-10, 1040 Vienna, Austria

1 Introduction

Consider a data set where the relative structure of parts is of interest. As an example, the age structure of all employees in a given country are to be analyzed. In this case, the ratios between the parts shall be considered for the analysis rather than the absolute values, which are mainly influenced by the size of the country and by other external factors. As it will be shown later, this situation has already been studied extensively in the classical framework of compositional data analysis (Aitchison 1982, 1986). However, when the data structure is determined according to more than one factor, e.g. one can study the employment structure from the perspective of age and gender of employees, the classical theory needs to be properly adjusted in order to cope with these more complex structures. The first attempt towards this goal has been considered by the proposal of compositional tables, two-factorial compositions (Fačevićová et al 2018). Nevertheless, also structures formed by three or even more factors are likely to occur in practice. For instance, in addition to gender and age, one could be interested in analyzing the employment structure according to full-time and part-time employment. Therefore, the manuscript introduces a general framework of dealing with multi-factorial compositional data, hereby extending the concepts and developments for compositional tables.

The goal of compositional data analysis is to process data which carry relative information. This resulted in a concise methodology with a wide range of possible applications, see, e.g., Pawlowsky-Glahn et al (2015), Filzmoser et al (2018) and references therein. A D -part composition is defined as a vector with positive components (parts) $\mathbf{x} = (x_1, \dots, x_D)'$, where the real information content is in the ratios between these parts rather than directly in the measured absolute values. In other words, compositional data describe quantitatively relative contributions of parts on a whole. Consequently, compositional data are scale invariant and can be represented without any loss of information as observations with a prescribed sum of the parts, e.g. in proportions (sum 1) or percentages (sum 100). Accordingly, the sample space of (representations of) compositional data is traditionally considered to be the D -part simplex $\mathcal{S}^D = \{\mathbf{x} = (x_1, x_2, \dots, x_D)' \mid x_i > 0, \forall i, \sum_i x_i = \kappa\}$. Note that the constant $\kappa > 0$, representing the sum of the compositional parts, can be chosen arbitrarily, and it reduces the dimensionality of the sample space to $D - 1$. Specific features of compositional data, particularly the scale invariance property, are captured by the Aitchison geometry (Pawlowsky-Glahn and Egozcue 2001; Billheimer et al 2015) with Euclidean vector space properties. For compositions $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ and a real constant α , these properties result from defining the operations perturbation, powering, and the Aitchison inner product,

$$\begin{aligned} \mathbf{x} \oplus \mathbf{y} &= \mathcal{C}(x_1 y_1, \dots, x_D y_D)', \quad \alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, \dots, x_D^\alpha)', \\ \langle \mathbf{x}, \mathbf{y} \rangle_A &= \frac{1}{2D} \sum_{i=1}^D \sum_{j=1}^D \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}, \end{aligned} \quad (1)$$

respectively, where $\mathcal{C}$ denotes closure and constitutes the equivalence classes of compositional data, which differ only by the sum of their parts κ . As a consequence, a direct application of traditional multivariate statistical methods that rely on the Euclidean geometry in the real space (Eaton 1983) is not appropriate. Even though it would be possible to adapt them to the Aitchison geometry, it is more sensible to find a way how to express compositional data isometrically in the $(D - 1)$ -dimensional real space and proceed there, just by taking into account the specific interpretation of the new variables. In the compositional data analysis context this refers to isometric log-ratio (ilr) coordinates (Egozcue et al 2003), which are orthonormal with respect to the Aitchison geometry. The main idea is to find a system of $D - 1$ orthonormal basis vectors $\mathbf{e}_i$ of $\mathcal{S}^D$, where the new coordinates $\mathbf{z} = (z_1, \dots, z_{D-1})' \in \mathbb{R}^{D-1}$ are obtained as

$$z_i = \langle \mathbf{x}, \mathbf{e}_i \rangle_A, \quad \text{for } i = 1, \dots, D - 1. \quad (2)$$

Since there does not exist a canonical basis on $\mathcal{S}^D$, an option is to use such an ilr coordinate system which has an advantageous interpretation under the given problem setting. From the definition, any ilr coordinate is a log-contrast, i.e. a linear combination $\xi_1 \ln x_1 + \dots + \xi_D \ln x_D$ with $\sum_{i=1}^D \xi_i = 0$. One popular approach for the construction of orthonormal coordinates was defined in Egozcue and Pawlowsky-Glahn (2005). The aim is to construct a sequence of binary partitions of groups of compositional parts in order to obtain coordinates that are interpretable in terms of balances between these groups of parts. Accordingly, sequential binary partitions (SBP) are based on a systematical splitting of the compositional vector into two non-overlapping subcompositions, and the generating process ends after $D - 1$ steps when each subcomposition is formed by only one part. The i -th step of the partition produces one vector of log-contrast coefficients $\boldsymbol{\xi}_i = (\xi_{i1}, \dots, \xi_{iD})'$ with r_i parts $\xi_{i+} = \sqrt{\frac{s_i}{r_i(r_i+s_i)}}$ at the positions corresponding to parts from the first subcomposition formed by this step (denoted with $+$), s_i parts $\xi_{i-} = -\sqrt{\frac{r_i}{s_i(r_i+s_i)}}$ at the positions related to parts from the second subcomposition (denoted with $-$), and 0 elsewhere. The coefficients are closely linked to basis vectors through the relation $\mathbf{e}_i = \exp(\boldsymbol{\xi}_i)$, and the resulting coordinates can be obtained from Eq. (2) or, without the need of enumeration of $\mathbf{e}_i$, directly as the log-contrast

$$z_i = \sum_{j=1}^D \xi_{ij} \ln x_j, \quad \text{for } i = 1, \dots, D - 1, \quad (3)$$

or

$$\mathbf{z} = \mathbf{V} \ln(\mathbf{x}), \quad (4)$$

where the contrast matrix $\mathbf{V}$ of order $(D - 1) \times D$ has rows formed by $\boldsymbol{\xi}_i$. With this notation it is possible to express the ilr coordinates (balances) by

$$z_i = \sqrt{\frac{r_i s_i}{r_i + s_i}} \ln \frac{g(x_{i_1}, \dots, x_{i_{r_i}})}{g(x_{i'_1}, \dots, x_{i'_{s_i}})}, \quad \text{for } i = 1, \dots, D - 1, \quad (5)$$

where $g(\cdot)$ stands for the geometric mean, and the values $i_1, \dots, i_{r_i}$ and $i'_1, \dots, i'_{s_i}$ label the parts from the first (+) and the second (−) subcomposition, respectively. Equation (5) reveals that SBP produces coordinates in form of log-ratios between mean representations of two groups of parts, which has led to the name “balances”. Particularly, in the case of vector compositional data, balances allow for a simple and natural interpretation, and for this reason they are frequently used in applications (Pawlowsky-Glahn et al 2015).

Although balances form a flexible class of orthonormal coordinates for vector compositional data, a further challenge is to develop a coordinate representation for the case when the whole is distributed according to two or more factors. The case of two-factorial compositional data (Egozcue et al 2008, 2015), referred to as compositional tables, has been intensively studied in Fačevićová et al (2014, 2016, 2018), and a general coordinate representation has been derived, which enables to decompose compositional tables into independent and interactive parts. Accordingly, the resulting ilr coordinates have the form of balances between two groups of parts (independent part) and log-odds ratios (interactive part), and they respect the dimensionality of the decomposed parts.

More specifically, compositional tables refer to a setting where the relative structure of the data is determined by two factors. Accordingly, not only the relations within each factor, but also relations between them need to be analyzed. As an example consider an employment structure in a given country, distributed according to the age of the employees and their gender. Three types of questions arise: Is the proportion of females among the employees comparable to the proportion of males? Does any of the age groups outbalance? Does the age structure of the employees depend on their gender? The first two questions focus exclusively on one factor, suppressing the effect of the other one. The last question, on the other hand, links information from both factors together. When the within-factor structure is analyzed, the effect of the other factor can be suppressed by averaging across all its levels. This results in a standard compositional vector, and balances are then a natural way of its coordinate representation. Particularly, consider a table formed by two factors, a row factor with I levels, and a column factor with J levels. The whole information about the relations among the I levels of the row factor is preserved in $I - 1$ coordinates of the form

$$z_i^r = \sqrt{\frac{s_i t_i J}{s_i + t_i}} \ln \frac{[g(\mathbf{x}_{i_1 \bullet}) \cdots g(\mathbf{x}_{i_{s_i} \bullet})]^{1/s_i}}{[g(\mathbf{x}_{i'_1 \bullet}) \cdots g(\mathbf{x}_{i'_{t_i} \bullet})]^{1/t_i}}, \quad \text{for } i = 1, 2, \dots, I - 1, \quad (6)$$

where s_i and t_i correspond to the respective step of the SBP performed on the levels of the row factor. The indices $(i_1 \bullet, \dots, i_{s_i} \bullet)$ and $(i'_1 \bullet, \dots, i'_{t_i} \bullet)$ specify the rows, and $g(\cdot)$ is the geometric mean. Similarly, relations among the J levels of the column factor are preserved in the $J - 1$ balances

$$z_j^c = \sqrt{\frac{u_j v_j I}{u_j + v_j}} \ln \frac{\left[g(\mathbf{x}_{\bullet j_1}) \cdots g(\mathbf{x}_{\bullet j_{u_j}}) \right]^{1/u_j}}{\left[g(\mathbf{x}_{\bullet j'_1}) \cdots g(\mathbf{x}_{\bullet j'_{v_j}}) \right]^{1/v_j}}, \quad \text{for } j = 1, 2, \dots, J-1, \quad (7)$$

constructed with respect to the SBP of the levels of the column factor, where the indices $(\bullet j_1, \dots, \bullet j_{u_j})$ and $(\bullet j'_1, \dots, \bullet j'_{v_j})$ specify the included columns.

The relations between two factors are traditionally described by odds ratios (Agresti 2002). This concept can be adapted to compositional tables, because the last group of coordinates has the form of log-odds ratios between four groups of parts. These groups are uniquely defined by row and column SBPs and represented by geometrical means of their parts. More specifically, these $(I-1)(J-1)$ odds ratio coordinates are given by

$$z_{ij}^{\text{OR}} = \sqrt{\frac{|A_{ij}| |D_{ij}|}{|A_{ij}| + |B_{ij}| + |C_{ij}| + |D_{ij}|}} \ln \frac{g(\mathbf{x}_{A_{ij}}) g(\mathbf{x}_{D_{ij}})}{g(\mathbf{x}_{B_{ij}}) g(\mathbf{x}_{C_{ij}})} \quad (8)$$

for $i = 1, 2, \dots, I-1$ and $j = 1, 2, \dots, J-1$, where $A_{ij}, \dots, D_{ij}$ are indices of parts in each group defined by the i -th and j -th step of row and column SBP, respectively, and $|A_{ij}|, \dots, |D_{ij}|$ are the numbers of parts within these groups. The construction and interpretation of this coordinate system is discussed in detail in Fačevicová et al (2018).

An important feature of compositional tables is the possibility of their orthogonal decomposition. In the special situation when there exists no relationship between row and column factors, all parts of the compositional table would be formed by the product of row and column marginals. This leads to the so called *independence table*. The orthogonal complement to the independence table is called *interaction table*. Since the previously introduced coordinate system respects this decomposition, the independence table is characterized by row and column balances, and the interaction table by odds ratio coordinates, and it is possible to analyze each part separately.

Even though the structure of compositional tables is already well described in the literature, a comprehensive approach for the analysis of multi-factorial relative-valued data is still lacking. Thus, the framework of compositional tables is extended to a general theory to work with k -factorial compositional data. Besides other findings, it turns out that also the multi-factorial structures can be decomposed orthogonally into an independent and several interactive parts and, moreover, a coordinate representation allowing for their separate analysis is provided. Although all considerations in the next section (Sect. 2) are presented just for the case of three-factorial compositional data (called *compositional cubes* in the following), they can be easily generalized to the case of more than three factors.

The construction and interpretation of the proposed coordinate system is explained on an illustrative example in Sects. 2.1 and 3, which also introduces the function implemented in the R package `robCompositions` (Templ et al 2011). The coordinates are used for the analysis of the employment structure of the European OECD countries, and a graphical comparison of the countries as well as a spatial clustering are provided. Moreover, the main sources of differences between the clusters are

investigated by robust principal component analysis. Based on an example of Austrian mobility data, a strategy for the analysis of multi-factorial time series is presented in Sect. 4. The final Sect. 5 concludes.

2 Compositional cubes

In this section we simplify the main findings derived in Fačevicová et al (2018) for the two-factorial situation, and consequently generalize them to the multi-factorial case. Consider a relative structure formed according to three factors with I , J and K levels, respectively. Such a situation can be represented with a compositional cube and written in the form

$$\mathbf{x} = \left(\begin{array}{ccc|ccc} x_{111} & \cdots & x_{1J1} & & & \\ \vdots & & \vdots & & & \\ x_{I11} & \cdots & x_{IJ1} & & & \end{array} \middle| \cdots \middle| \begin{array}{ccc} x_{11K} & \cdots & x_{1JK} \\ \vdots & & \vdots \\ x_{I1K} & \cdots & x_{IJK} \end{array} \right), \quad (9)$$

where $x_{ijk} > 0, \forall i, j, k$. The vertical lines separate the levels of the third factor, called slices in the following. Since compositional cubes form a special case of the concept of $(I \cdot J \cdot K)$ -part vector compositional data, all basic definitions can be accommodated for this case.

The sample space of compositional cubes is a subset of the $(I \cdot J \cdot K)$ -part simplex

$$\mathcal{S}^{IJK} = \left\{ \mathbf{x} = (x_{111}, \dots, x_{IJK})' \mid x_{ijk} > 0, \quad \forall i, j, k; \sum_{i,j,k=1}^{I,J,K} x_{ijk} = \kappa \right\}, \quad (10)$$

which includes only those $(I \cdot J \cdot K)$ -part compositions, which can be recorded into the form of a three-factorial structure with I rows, J columns and K slices. The basic operations of the Aitchison geometry (as given in Pawłowsky-Glahn and Egozcue 2001) modify to

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_{ijk} \cdot y_{ijk})_{i,j,k=1}^{I,J,K} \quad \text{and} \quad \alpha \odot \mathbf{x} = \mathcal{C}(x_{ijk}^\alpha)_{i,j,k=1}^{I,J,K}, \quad (11)$$

and they result again in a compositional cube. The Euclidean vector space structure of the Aitchison geometry is completed by defining the Aitchison inner product for cubes

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \frac{1}{2IJK} \sum_{i,j,k} \sum_{i',j',k'} \ln \frac{x_{ijk}}{x_{i'j'k'}} \ln \frac{y_{ijk}}{y_{i'j'k'}}. \quad (12)$$

2.1 Example: employment structure

The proposed approach will be illustrated with an example where the analysis of the employment structure in several countries is of interest. For this purpose, data from 32 European members of OECD were collected at <http://stats.oecd.org>. For each country in the sample, an estimated number of employees in the year 2015 was available. The data were structured according to gender and age of the employees and the type of their contract. More specifically, we distinguish males (M) and females (F), young (category 15–24), middle-aged (25–54) and older (55+) employees, and full-time (FT) and part-time (PT) contracts. The data at hand thus form a sample of 32 cubes with two rows (gender), two columns (type of contract) and three slices (age), which allow for a deeper analysis of the overall employment structure, not just from the perspective of each factor separately, but also from the perspective of the relations/interactions between them. Besides the global aspects of the employment, the analysis aims also at revealing the national specifics of the countries contained in the sample. An example of one cube from Czech Republic is displayed in Table 1, and a graphical overview of the cubes is depicted in Fig. 1.

Obviously, the counts in the cells of the cubes depend on the population size of the country. When the analysis of structural patterns of the employment in several countries is of interest, the compositional approach appears as appropriate, because the population size is not relevant in this approach.

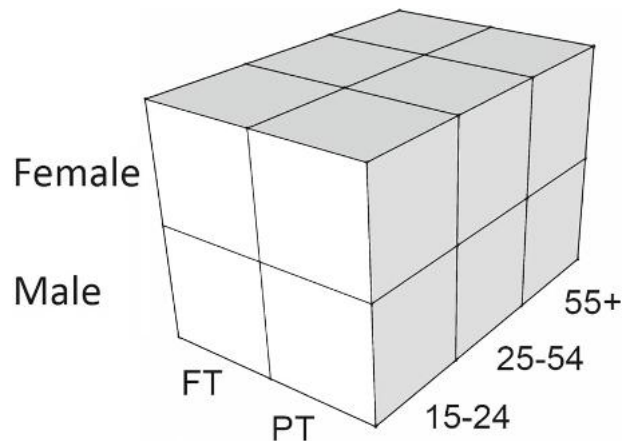
2.2 Decomposition of compositional cubes

Egozcue et al (2008) proposed a decomposition of a compositional table into an independent and an interactive part (still a compositional table), which are mutually orthogonal; their perturbation again leads to the original compositional table. The independent part mimics the independence of the factors. As in the standard case of contingency tables, an assumption of independence means that the whole information about the relative structure of both factors is preserved in the row and column marginals, and each entry of the table can be obtained as their product. In the compositional case, the only difference is that the arithmetic marginals are replaced by the geometric ones. When the factors are not independent, and the compositional table does not equal to the independence one, another table needs to be introduced. The interaction table, simply defined as a residual resulting from the difference between the original and the independence tables in the sense of perturbation, preserves the whole information about the relations between the factors and becomes mainly important when these relations are analyzed. A similar idea can be utilized also in the case of compositional cubes, but due to the presence of pairwise and whole interactions, it is possible to further decompose the interactive part of the cube into additional four cubes, each preserving information about another source of association between the factors.

Similar to the case of compositional tables, also parts of the independence compositional cube are formed by the product of row, column and slice (geometric) marginals

Table 1 Example of one cube from the sample analyzed in Sections 2.1 and 3: employment structure in the Czech Republic in 2015 (in thousands of employees)

Gender	15–24		25–54		55+	
	FT	PT	FT	PT	FT	PT
Female	104.756	17.128	1618.415	90.505	317.031	56.355
Male	169.851	11.165	2127.849	22.759	467.212	38.208

Fig. 1 Graphical representation of the cube structure used in Sects. 2.1 and 3. The rows represent the gender of the employees, the columns the type of contract (FT: full-time, PT: part-time), and the slices separate different age groups

$$x_{ijk}^{\text{ind}} = g(\mathbf{x}_{i\bullet\bullet})g(\mathbf{x}_{\bullet j\bullet})g(\mathbf{x}_{\bullet\bullet k}), \quad (13)$$

where dots in the index indicate an aggregation over the respective factors. For example, the notation $g(\mathbf{x}_{i\bullet\bullet})$ stands for the geometric mean of all parts in the i -th row of the cube, i.e. $(\prod_{j=1}^J \prod_{k=1}^K x_{ijk})^{1/JK}$. In case of perfect independence of all three factors, the original cube would be equal to the independence one. Otherwise, all associations between the factors are preserved in the interactive part

$$\mathbf{x}^{\text{int}} = \mathbf{x} \ominus \mathbf{x}^{\text{ind}}. \quad (14)$$

As mentioned above, the interactive part can be further decomposed. First, the relations between row and column factors are analyzed. Aggregation over values of the slice factor eliminates its impact and reduces the three-dimensional structure to a system of K similar compositional tables, forming K slices of a cube. According to Egozcue et al (2008), the interactive part of a table is extracted by a division of its parts by the respective geometric marginals. These considerations result in the compositional cube $\mathbf{int}^{\text{rc}}(\mathbf{x})$ with cells

$$\text{int}^{\text{rc}}(\mathbf{x})_{ijk} = \frac{g(\mathbf{x}_{ij\bullet})}{g(\mathbf{x}_{i\bullet\bullet})g(\mathbf{x}_{\bullet j\bullet})}. \quad (15)$$

From (15) it follows that the $\mathbf{int}^{\text{rc}}(\mathbf{x})$ is actually formed by K equal slices (compositional tables). Moreover, the row and column geometric marginals (that means compositional tables resulting from an aggregation of cube cells by geometric means

across the respective direction) are uniform, which underlines the favorable structure of the proposed decomposition. The system of marginals is completed in the direction of slices, whose respective marginal table corresponds to (15). In order to extract the pure interaction between row and slice factor, the effect of the column factor needs to be filtered out using the geometric mean, and similarly as $\mathbf{int}^{\text{rc}}(\mathbf{x})$, the row-slice interaction cube $\mathbf{int}^{\text{rs}}(\mathbf{x})$ has parts

$$\mathbf{int}^{\text{rs}}(\mathbf{x})_{ijk} = \frac{g(\mathbf{x}_{i\bullet k})}{g(\mathbf{x}_{i\bullet\bullet})g(\mathbf{x}_{\bullet\bullet k})}. \quad (16)$$

Similar to the case of $\mathbf{int}^{\text{rc}}(\mathbf{x})$, also this cube has uniform marginals. This property holds for the row and slice directions and the marginal table computed across the columns equal to (16). Finally, interactions between column and slice factors are contained in the cube $\mathbf{int}^{\text{cs}}(\mathbf{x})$ with cells

$$\mathbf{int}^{\text{cs}}(\mathbf{x})_{ijk} = \frac{g(\mathbf{x}_{\bullet j k})}{g(\mathbf{x}_{\bullet j \bullet})g(\mathbf{x}_{\bullet\bullet k})}. \quad (17)$$

Since this cube is formed by I identical rows, also row marginals equal to a table with parts (17), however the column and slice marginals are again uniform, i.e., they are composed by the same positive elements. All pairwise interaction cubes are orthogonal, but since there was always one factor omitted from the consideration, the information about the interactive part of the original cube is still not complete. The structure of a compositional cube is completed by considering mutual interactions between all three factors. This corresponds to the cube

$$\mathbf{int}^{\text{rcs}}(\mathbf{x}) = \mathbf{x}^{\text{int}} \ominus \mathbf{int}^{\text{rc}}(\mathbf{x}) \ominus \mathbf{int}^{\text{rs}}(\mathbf{x}) \ominus \mathbf{int}^{\text{cs}}(\mathbf{x}) \quad (18)$$

with parts

$$\mathbf{int}^{\text{rcs}}(\mathbf{x})_{ijk} = \frac{x_{ijk}g(\mathbf{x}_{i\bullet\bullet})g(\mathbf{x}_{\bullet j \bullet})g(\mathbf{x}_{\bullet\bullet k})}{g(\mathbf{x}_{ij\bullet})g(\mathbf{x}_{\bullet j k})g(\mathbf{x}_{i\bullet k})}. \quad (19)$$

Also this cube has an advantageous structure from the perspective of the marginal tables, which are in this case uniform in all three directions. Note here that a similar property holds also for the interactive part of a compositional table, where row and column marginals are from their construction uniform. Such a decomposition of the multi-factorial data can be very useful for an in-depth analysis of the data structure, as demonstrated in Fačevicová et al (2016, 2018); Fačevicová et al (2021) for compositional tables. For instance, in the example from Sect. 2.1, the cube $\mathbf{int}^{\text{rc}}(\mathbf{x})$ preserves interactions between the gender of employees and the type of their contract. The information on relations between these two factors is then completed by $\mathbf{int}^{\text{rcs}}(\mathbf{x})$, which involves also the effect of age, which was suppressed in $\mathbf{int}^{\text{rc}}(\mathbf{x})$. Moreover, from the decomposition it follows that for those interactions, for which the respective cubes in the interaction part are constructed, the respective marginals are uniform, i.e., the

information is fully captured by these cubes and does not propagate further. Consequently, vector (one-factorial) geometric marginals occur in the decomposition indeed only in the independent part, as expected, and more-dimensional nontrivial marginals (here in form of compositional tables) are left for bifactorial interaction cubes.

On the other hand, if each part of the decomposition is considered separately, two challenges need to be taken into account. At first, the dimensionality of the original cube sample space is decomposed as well. More specifically, the overall dimensionality $JK - 1$ turns to $I + J + K - 3$ for the independence cube $\mathbf{x}^{\text{ind}}$, $(I - 1)(J - 1)$ for cube $\mathbf{int}^{\text{rc}}(\mathbf{x})$ (and similarly for the remaining cubes related to paired interactions) and, finally, the dimensionality of the sample space of cube $\mathbf{int}^{\text{rcs}}(\mathbf{x})$ is $(I - 1)(J - 1)(K - 1)$. The altered dimensionality, corresponding to each of the cubes from the decomposition, can cause computational problems when an arbitrary ilr coordinate system (primarily designed for vector compositional data) is used for the representation of independence and interaction cubes. For example, this can be the case for robust statistical analysis (de Sousa et al 2021), but also in general it is desirable to assign to each of the cubes from the decomposition such a number of coordinates (out of the total number $JK - 1$ of them) that reflects their respective dimensionality. The second problem concerns the interpretation of the results. Even though it is usually possible to convert the results back to the simplex, it is convenient to proceed with the analysis in some well-interpretable coordinates. Obviously, balances, defined as a log-ratio between two groups of parts, are not able to capture the multi-factorial nature of the compositional cubes. Although they can help to describe the relative structure of each factor separately, for a description of interactions we need to construct some alternative coordinate system. The construction of such orthonormal coordinates is presented in the following section.

2.3 Coordinate representation of compositional cubes

In this section we will focus on a possible coordinate representation of three-factorial compositional data, compositional cubes, which simplifies substantially the construction of ilr coordinates for compositional tables proposed in Fačevicová et al (2018). A deeper understanding of the structure of this coordinate representation allows its generalization and application to compositional data describing relationships given by more than three factors. In order to keep the construction as simple as possible, we consider a vectorized version of the cube

$$\text{vec}(\mathbf{x}) = (x_{111}, x_{112}, \dots, x_{1JK}, x_{211}, \dots, x_{2JK}, \dots, x_{I11}, \dots, x_{IJK})'. \quad (20)$$

As it was already suggested, balances can help to describe the relative structure within each factor. For this purpose, the whole rows, columns and slices (each represented by the geometric mean across all levels of the remaining factors) should be taken. After $I - 1$ steps of the sequential binary partition applied on the levels of the row factor (SBPr), a system of $I - 1$ vectors ξ_i^{r} (of length JK) is obtained. The i -th generating vector has entries $\xi_{i+}^{\text{r}} = \sqrt{\frac{t_i}{s_i(s_i + t_i)JK}}$ at positions corresponding to parts from the s_i rows of the cube $\mathbf{x}$, which were in the respective step assigned to the +

group, and $\xi_{i-}^r = -\sqrt{\frac{s_i}{t_i(s_i+t_i)JK}}$ at positions corresponding to parts from the t_i rows assigned to the $-$ group, and zero elsewhere. The first group of coordinates is thus simply formed by row balances, which characterize the structure of the row factor when the influence of the other factors is suppressed,

$$z_i^r = \sqrt{\frac{s_i t_i J K}{s_i + t_i}} \ln \frac{[g(\mathbf{x}_{i1\bullet\bullet}) \cdots g(\mathbf{x}_{i s_i \bullet\bullet})]^{1/s_i}}{[g(\mathbf{x}_{i1'\bullet\bullet}) \cdots g(\mathbf{x}_{i t_i' \bullet\bullet})]^{1/t_i}}, \quad \text{for } i = 1, 2, \dots, I-1. \quad (21)$$

An example can be seen in Sect. 3.1, where this first set of coordinates results in balance between male and female employees, regardless of their age or type of the contract. A similar construction can be made for column and slice factors. In the first case, a sequential binary partition of the whole columns (SBPc) results in a system of $J-1$ vectors ξ_j^c with entries $\xi_{j+}^c = \sqrt{\frac{v_j}{u_j(u_j+v_j)IK}}$, $\xi_{j-}^c = -\sqrt{\frac{u_j}{v_j(u_j+v_j)IK}}$ and 0, always corresponding to parts of u_j columns from the $+$ group, v_j columns from the $-$ group, and columns not included to the respective step of SBPc. The inner structure of the column factor is preserved through column balances

$$z_j^c = \sqrt{\frac{u_j v_j I K}{u_j + v_j}} \ln \frac{[g(\mathbf{x}_{\bullet j1\bullet}) \cdots g(\mathbf{x}_{\bullet j u_j \bullet})]^{1/u_j}}{[g(\mathbf{x}_{\bullet j1'\bullet}) \cdots g(\mathbf{x}_{\bullet j v_j' \bullet})]^{1/v_j}}, \quad \text{for } j = 1, 2, \dots, J-1, \quad (22)$$

which form the second group in the coordinate system representing the whole compositional cube $\mathbf{x}$ as well as its independent part $\mathbf{x}^{\text{ind}}$. The third group describes the structure of the slice factor. A sequential binary partition of the whole slices (SBPs) now determines the final system of $K-1$ vectors ξ_k^s with entries $\xi_{k+}^s = \sqrt{\frac{n_k}{m_k(m_k+n_k)IJ}}$, $\xi_{k-}^s = -\sqrt{\frac{m_k}{n_k(m_k+n_k)IJ}}$ and 0, corresponding to m_k slices from group $+$, n_k slices from group $-$, and the remaining slices not included in the k -th step, respectively; the slice balances are

$$z_k^s = \sqrt{\frac{m_k n_k I J}{m_k + n_k}} \ln \frac{[g(\mathbf{x}_{\bullet\bullet k1}) \cdots g(\mathbf{x}_{\bullet\bullet k m_k})]^{1/m_k}}{[g(\mathbf{x}_{\bullet\bullet k1'}) \cdots g(\mathbf{x}_{\bullet\bullet k n_k'})]^{1/n_k}}, \quad \text{for } k = 1, 2, \dots, K-1. \quad (23)$$

Note here that the row, column and slice balances form a complete coordinate representation of $\mathbf{x}^{\text{ind}}$, since their number equals the dimensionality of the independence table. All the remaining $(I-1)(J-1)(K-1)$ coordinates, completing the orthonormal coordinate system of $\mathbf{x}^{\text{ind}}$ (e.g. those given by Eqs. (24)–(27)), are zero.

When row, column and slice SBPs are defined, we can immediately construct the remaining elements of the coordinate system of the cube $\mathbf{x}$, which also correspond to coordinates of $\mathbf{x}^{\text{int}}$. For this purpose the normalized Hadamard (entry wise) product ($\circ$) of the vectors ξ_i^r , ξ_j^c and ξ_k^s and Eq. (3) is used. The vectors $\xi_{ij}^{\text{rc}} = \xi_i^r \circ \xi_j^c$, $i = 1, \dots, I-1$, $j = 1, \dots, J-1$ determine $(I-1)(J-1)$ coordinates of type

$$z_{ij}^{\text{rc}} = \sqrt{\frac{|A_{ij}| |D_{ij}|}{|A_{ij}| + |B_{ij}| + |C_{ij}| + |D_{ij}|}} \ln \frac{g(\mathbf{x}_{A_{ij}})g(\mathbf{x}_{D_{ij}})}{g(\mathbf{x}_{B_{ij}})g(\mathbf{x}_{C_{ij}})}, \quad (24)$$

for $i = 1, 2, \dots, I-1$ and $j = 1, 2, \dots, J-1$, capturing the interactions between row and column factors, which through the geometric mean suppress the influence of the slice factor. Obviously, these coordinates are formed by four groups of parts (denoted as A, B, C, D and represented by their respective geometric means), and they can be interpreted in terms of a log-odds ratio, which is also used in a standard statistical analysis of two-factorial data (Agresti 2002). This system of coordinates, composed into the $(IJK - 1)$ -component vector $\mathbf{z}^{\text{rc}}$ with z_{ij}^{rc} on the positions $(I-1)(J-1)$ corresponding to $\mathbf{int}^{\text{rc}}(\mathbf{x})$ and zeros elsewhere, thus allows for the analysis of the relations exclusively between row and column factors.

The Hadamard product ξ_{ik}^{rs} of ξ_i^r and ξ_k^s , $i = 1, \dots, I-1$, $k = 1, \dots, K-1$, leads to coordinates

$$z_{ik}^{\text{rs}} = \sqrt{\frac{|A'_{ik}| |D'_{ik}|}{|A'_{ik}| + |B'_{ik}| + |C'_{ik}| + |D'_{ik}|}} \ln \frac{g(\mathbf{x}_{A'_{ik}})g(\mathbf{x}_{D'_{ik}})}{g(\mathbf{x}_{B'_{ik}})g(\mathbf{x}_{C'_{ik}})}, \quad (25)$$

for $i = 1, 2, \dots, I-1$ and $k = 1, 2, \dots, K-1$, which capture the information about the relations between row and slice factors, when the influence of the column factor is suppressed. Similar to the case of $\mathbf{z}^{\text{rc}}$, also coordinates contained in the respective vector $\mathbf{z}^{\text{rs}}$ can be interpreted in terms of a log-odds ratio and are utilized, when the relationship between row and slice factors is of primary interest.

Finally, the Hadamard products of ξ_j^c and ξ_k^s , $j = 1, \dots, J-1$, $k = 1, \dots, K-1$ lead to vectors ξ_{jk}^{cs} and coordinates

$$z_{jk}^{\text{cs}} = \sqrt{\frac{|A''_{jk}| |D''_{jk}|}{|A''_{jk}| + |B''_{jk}| + |C''_{jk}| + |D''_{jk}|}} \ln \frac{g(\mathbf{x}_{A''_{jk}})g(\mathbf{x}_{D''_{jk}})}{g(\mathbf{x}_{B''_{jk}})g(\mathbf{x}_{C''_{jk}})}, \quad (26)$$

for $j = 1, 2, \dots, J-1$ and $k = 1, 2, \dots, K-1$. The system of these coordinates $\mathbf{z}^{\text{cs}}$ completes the odds ratio-type coordinates with those concerning relations between column and slice factors.

To complete the original data structure, also full interactions between all three factors need to be contained in the coordinate system. The remaining $(I-1)(J-1)(K-1)$ coordinates are determined by the Hadamard product of all three types of vectors, ξ_i^r , ξ_j^c and ξ_k^s , $i = 1, \dots, I-1$, $j = 1, \dots, J-1$, $k = 1, \dots, K-1$, and have the general form

$$z_{ijk}^{\text{rcs}} = Q_{ijk} \ln \frac{g(\mathbf{x}_{A'''_{ijk}})g(\mathbf{x}_{D'''_{ijk}})g(\mathbf{x}_{F'''_{ijk}})g(\mathbf{x}_{G'''_{ijk}})}{g(\mathbf{x}_{B'''_{ijk}})g(\mathbf{x}_{C'''_{ijk}})g(\mathbf{x}_{E'''_{ijk}})g(\mathbf{x}_{H'''_{ijk}})}, \quad (27)$$

for $i = 1, 2, \dots, I - 1$, $j = 1, 2, \dots, J - 1$, and $k = 1, 2, \dots, K - 1$, where

$$Q_{ijk} = \sqrt{\frac{|A'''_{ijk}| |H'''_{ijk}|}{|A'''_{ijk}| + |B'''_{ijk}| + |C'''_{ijk}| + |D'''_{ijk}| + |E'''_{ijk}| + |F'''_{ijk}| + |G'''_{ijk}| + |H'''_{ijk}|}}, \quad (28)$$

is a constant ensuring orthonormality of the coordinates. Even though the interpretation of this last group of coordinates may be a bit tricky (one possible interpretation is in terms of a log-ratio of two odds ratios), their definition is necessary to complete the system of $IJK - 1$ orthonormal coordinates of the original table $\mathbf{x}$. Moreover, when a sample of compositional cubes is available, these coordinates can be used for instance to test for the presence of full interactions.

For an easier understanding of the coordinate structure, especially the assignment of parts into groups, Fig. 2 provides a graphical representation of each type of proposed coordinates. Moreover, the specific system of generating vectors, respective coordinates, and their interpretation is given in Sect. 3.1.

Besides the benefits of the proposed coordinate system in terms of interpretation, it is important to point out that the coordinates reflect the dimensionality of the sample space of the decomposed parts to which they are assigned, and thus allow to analyze these parts separately. Of course, each decomposed part is still a cube of the same dimension as the original one (with I rows, J columns and K slices) and its coordinate representation must contain $IJK - 1$ components, but the structure of the vector of coordinates follows the one introduced for $\mathbf{z}^{\text{rc}}$ of the interaction cube $\mathbf{int}^{\text{rc}}(\mathbf{x})$. More specifically, e.g. the cube $\mathbf{int}^{\text{rc}}(\mathbf{x})$ is represented in the proposed system with coordinates z_{ij}^{rc} and the remaining coordinates defined in this section equal zero. Accordingly, with respect to the decomposition described in Sect. 2.2, for the coordinate representation $\mathbf{z}$ of the original compositional cube the following relation holds,

$$\mathbf{z} = \mathbf{z}^{\text{r}} + \mathbf{z}^{\text{c}} + \mathbf{z}^{\text{s}} + \mathbf{z}^{\text{rc}} + \mathbf{z}^{\text{rs}} + \mathbf{z}^{\text{cs}} + \mathbf{z}^{\text{rcs}}. \quad (29)$$

Even though the interpretation of the coordinate system is determined by the initial SBPs, any other relationship within the compositional cube is reachable through a transformation matrix $\mathbf{T}$, whose rows are formed by coefficients of the respective log-arithmized parts of $\text{vec}(\mathbf{x})$ in the desirable log-ratios. According to (4), the vectorized form of a compositional cube $\text{vec}(\mathbf{x})$ is equal (after closure) to $\exp(\mathbf{V}'\mathbf{z})$. Therefore, a system of log-contrasts representing a given compositional cube equals

$$\mathbf{z}^* = \mathbf{T}\mathbf{V}'\mathbf{z}. \quad (30)$$

An example of such a transformation for the coordinates constructed in Sect. 3 is provided in Appendix.

2.4 General properties of multi-factorial compositional data

The findings from Sects. 2.2 and 2.3 can be directly extended to a general k -factorial case. k -factorial compositional data are formed by a k -dimensional array of positive

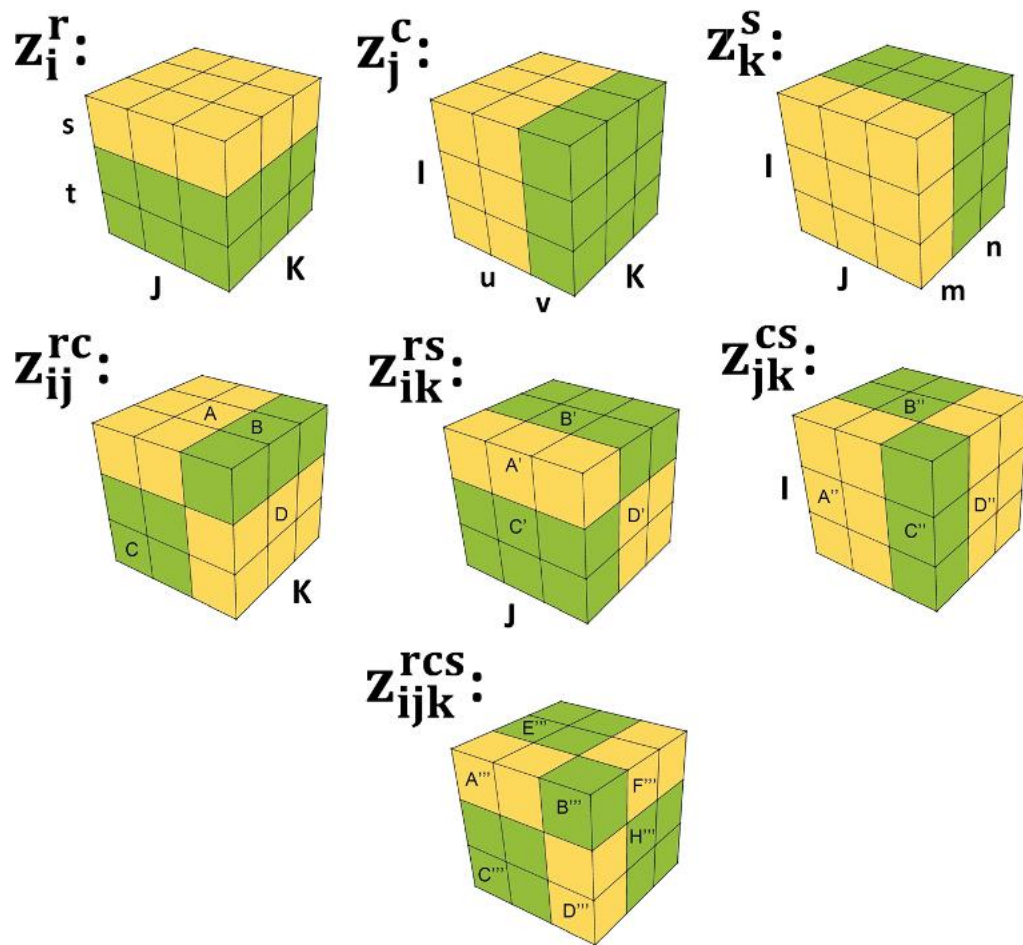


Fig. 2 Graphical representation of groups of parts involved in each type of coordinates forming the whole coordinate system designed for compositional cubes

entries, representing a relative structure given by the levels of k constituting factors. Also such a complex structure contains its independent and interactive parts, where the independent part is equal to the product of (vector) geometrical marginals. The sources of interactions are given by the relations between pairs, triplets, quaternions, etc. of constituting factors, and therefore the interactive part can be further orthogonally decomposed to objects carrying information about each of these sources. The main principle is based on aggregation over the redundant dimensions and expression of interactions within the resultant object. In the case of compositional cubes we have seen that pairwise interactions actually correspond to the interactive part of a compositional table formed by geometric means computed across the levels of the third factor, see e.g. Eq. (15). Similarly, in the case of a four-factorial compositional object, all sources of interactions between a selected triplet of factors can be reached by a decomposition of a cube given by an aggregation over the remaining fourth dimension. When we vary over the fourth dimension, all pairwise and three-way interactions are extracted and, finally, by subtraction of all these parts together with the independent one, the object preserving the full interactions is reached (similar as in the Eq. (18)).

Section 2.3 shows that the whole coordinate representation of a compositional cube is determined by three systems of SBPs, separately given for the levels of row, column and slice factors. Similarly, also k -factorial compositions can be represented in orthonormal coordinates. Balances between levels of the individual factors characterize the independent part of the object. Log-contrasts obtained from the Hadamard product of pairs, triplets, etc. of SBP basis vectors ξ and Eq. (3) then represent the respective sources of interactions.

3 Example: employment structure—continuation

Let us go back to the employment structure data set introduced in Sect. 2.1. According to the proposed methodology, each cube from the sample can be represented by a system of coordinates (Sects. 3.1 and 3.2), which allows for a deeper analysis of the structural patterns (Sect. 3.3).

3.1 Coordinate representation

Following Section 2, the row, column and slice SBPs need to be determined prior to the construction of coordinates. For the factors “type of contract” and “gender”, only their two levels need to be separated. Consequently, the first two generating vectors are

$$\xi_1^r = \sqrt{\frac{1}{12}} (1, 1, 1, 1, 1, 1, -1, -1, -1, -1, -1, -1)'$$

and

$$\xi_1^c = \sqrt{\frac{1}{12}} (1, 1, 1, -1, -1, -1, 1, 1, 1, -1, -1, -1)',$$

where the components of these vectors correspond to the cells of the vectorized form of the cube,

$$\text{vec}(\mathbf{x}) = (x_{111}, x_{112}, x_{113}, x_{121}, x_{122}, x_{123}, x_{211}, x_{212}, x_{213}, x_{221}, x_{222}, x_{223})'.$$

There are more options for the slice SBP, where the analyst can decide which age group has to be separated first. Here, the youngest group was firstly separated from the remaining two groups and, in the next step, the middle-aged group (25–54 years) from the oldest. The other options, starting with the separation of the middle-aged or the oldest group, respectively, would lead to similar results (in terms of presence of interaction between factors), but they would slightly alter the interpretation. In the presented case, the generating vectors are

$$\xi_1^s = \sqrt{\frac{1}{6}} (1, -0.5, -0.5, 1, -0.5, -0.5, 1, -0.5, -0.5, 1, -0.5, -0.5)'$$

and

$$\xi_2^s = \sqrt{\frac{1}{8}}(0, 1, -1, 0, 1, -1, 0, 1, -1, 0, 1, -1)'$$

Following the construction from Sect. 2, the Hadamard product of the above derived generating vectors leads (after their normalization) to the remaining system of vectors. Particularly,

$$\begin{aligned}\xi_{11}^{rc} &= \sqrt{\frac{1}{12}}(1, 1, 1, -1, -1, -1, -1, -1, -1, 1, 1, 1)' \propto \xi_1^r \circ \xi_1^c, \\ \xi_{11}^{rs} &= \sqrt{\frac{1}{6}}(1, -0.5, -0.5, 1, -0.5, -0.5, -1, 0.5, 0.5, -1, 0.5, 0.5)' \propto \xi_1^r \circ \xi_1^s, \\ \xi_{12}^{rs} &= \sqrt{\frac{1}{8}}(0, 1, -1, 0, 1, -1, 0, -1, 1, 0, -1, 1)' \propto \xi_1^r \circ \xi_2^s, \\ \xi_{11}^{cs} &= \sqrt{\frac{1}{6}}(1, -0.5, -0.5, -1, 0.5, 0.5, 1, -0.5, -0.5, -1, 0.5, 0.5)' \propto \xi_1^c \circ \xi_1^s, \\ \xi_{12}^{cs} &= \sqrt{\frac{1}{8}}(0, 1, -1, 0, -1, 1, 0, 1, -1, 0, -1, 1)' \propto \xi_1^c \circ \xi_2^s, \\ \xi_{111}^{rcs} &= \sqrt{\frac{1}{6}}(1, -0.5, -0.5, -1, 0.5, 0.5, -1, 0.5, 0.5, 1, -0.5, -0.5)' \propto \xi_1^r \circ \xi_1^c \circ \xi_1^s, \\ \xi_{112}^{rcs} &= \sqrt{\frac{1}{8}}(0, 1, -1, 0, -1, 1, 0, -1, 1, 0, 1, -1)' \propto \xi_1^r \circ \xi_1^c \circ \xi_2^s.\end{aligned}\quad (31)$$

Finally, according to Eq. (3), these vectors lead to a system of 11 orthonormal coordinates:

$$\begin{aligned}z_1^r &= \sqrt{3} \ln \frac{g(\mathbf{x}_{1\bullet\bullet})}{g(\mathbf{x}_{2\bullet\bullet})} & z_{11}^{rs} &= \sqrt{\frac{2}{3}} \ln \frac{g(\mathbf{x}_{1\bullet 1})\sqrt{g(\mathbf{x}_{2\bullet 2})g(\mathbf{x}_{2\bullet 3})}}{g(\mathbf{x}_{2\bullet 1})\sqrt{g(\mathbf{x}_{1\bullet 2})g(\mathbf{x}_{1\bullet 3})}} \\ z_1^c &= \sqrt{3} \ln \frac{g(\mathbf{x}_{\bullet 1\bullet})}{g(\mathbf{x}_{\bullet 2\bullet})} & z_{12}^{rs} &= \sqrt{\frac{1}{2}} \ln \frac{g(\mathbf{x}_{1\bullet 2})g(\mathbf{x}_{2\bullet 3})}{g(\mathbf{x}_{2\bullet 2})g(\mathbf{x}_{1\bullet 3})} \\ z_1^s &= \sqrt{\frac{8}{3}} \ln \frac{g(\mathbf{x}_{\bullet\bullet 1})}{\sqrt{g(\mathbf{x}_{\bullet\bullet 2})g(\mathbf{x}_{\bullet\bullet 3})}} & z_{11}^{cs} &= \sqrt{\frac{2}{3}} \ln \frac{g(\mathbf{x}_{\bullet 11})\sqrt{g(\mathbf{x}_{\bullet 22})g(\mathbf{x}_{\bullet 23})}}{g(\mathbf{x}_{\bullet 21})\sqrt{g(\mathbf{x}_{\bullet 12})g(\mathbf{x}_{\bullet 13})}} \\ z_2^s &= \sqrt{2} \ln \frac{g(\mathbf{x}_{\bullet\bullet 2})}{g(\mathbf{x}_{\bullet\bullet 3})} & z_{12}^{cs} &= \sqrt{\frac{1}{2}} \ln \frac{g(\mathbf{x}_{\bullet 12})g(\mathbf{x}_{\bullet 23})}{g(\mathbf{x}_{\bullet 13})g(\mathbf{x}_{\bullet 22})} \\ z_{11}^{rc} &= \sqrt{\frac{3}{4}} \ln \frac{g(\mathbf{x}_{11\bullet})g(\mathbf{x}_{22\bullet})}{g(\mathbf{x}_{12\bullet})g(\mathbf{x}_{21\bullet})} & z_{111}^{rcs} &= \sqrt{\frac{1}{6}} \ln \frac{x_{111}x_{221}\sqrt{x_{122}x_{123}\sqrt{x_{212}x_{213}}}}{x_{211}x_{121}\sqrt{x_{112}x_{113}\sqrt{x_{222}x_{223}}}} \\ & & z_{112}^{rcs} &= \sqrt{\frac{1}{8}} \ln \frac{x_{112}x_{222}x_{123}x_{213}}{x_{122}x_{212}x_{113}x_{223}}\end{aligned}$$

A graphical representation of these coordinates is provided in Fig. 3.

3.2 Interpretation

The interpretation of the coordinates can be discussed on the example of the data from the Czech Republic, see Table 1. The set of row, column and slice balances (as a subvector of $\mathbf{z}$) corresponds to

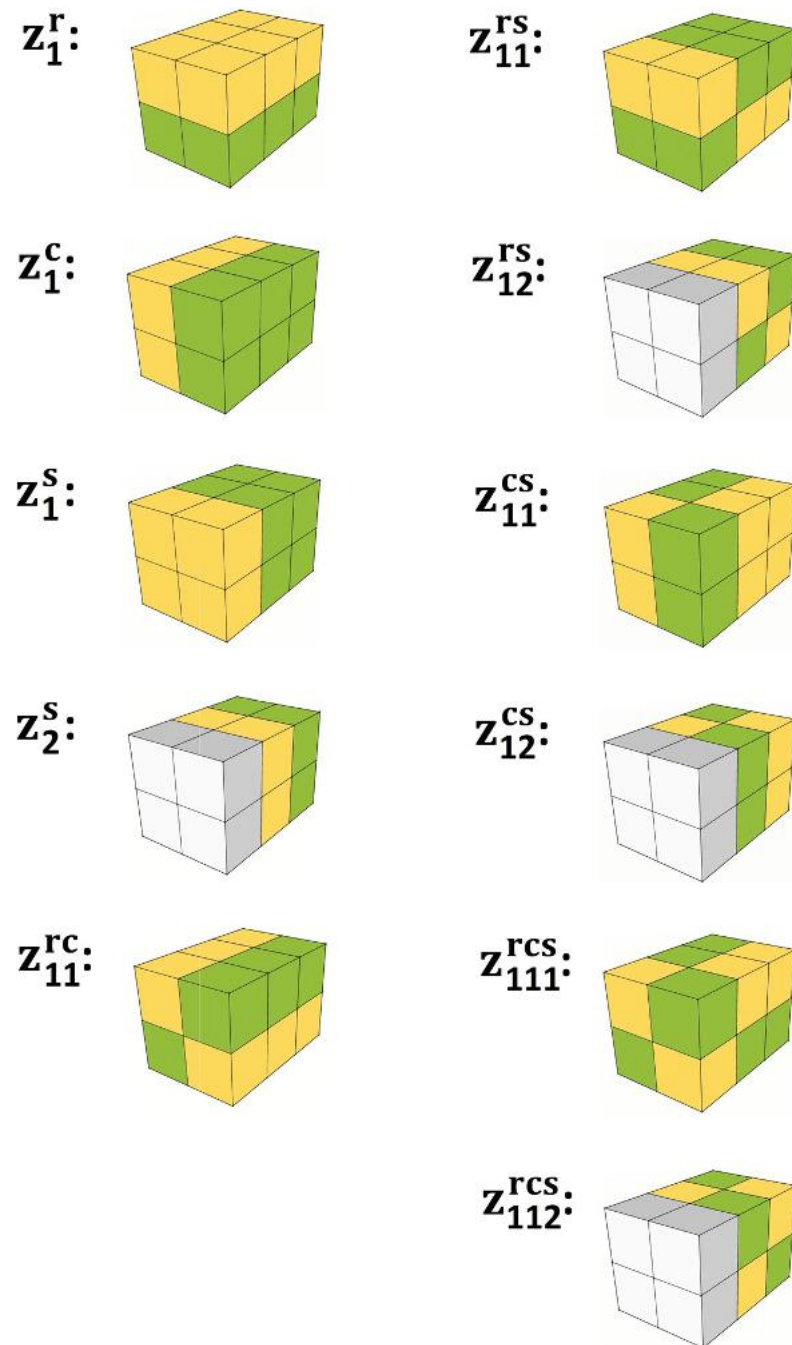


Fig. 3 Graphical representation of the groups of parts involved in particular coordinates. Yellow parts constitute the numerator, green the denominator and white parts are not included in the respective log-ratio. (Color figure online)

$$(z_1^r, z_1^c, z_1^s, z_2^s)' = (0.304, 4.672, -2.487, 1.097)'.$$

These numbers are interpretable, as usual for balances, in terms of a dominance of either the group of cells in the numerator (positive value) or denominator (negative value) of the respective log-ratio. Accordingly, in the Czech Republic the proportion of

female employees slightly dominates over the proportion of males (z_1^r), and full-time contracts clearly dominate over part-time contracts (z_1^c). The slice balances contain information about the age structure of the employees. Due to the high negative value of coordinate z_1^s it can be concluded that the youngest employees are outbalanced by those from the middle age and older groups; within these latter groups, employees aged between 25 and 54 years prevail (coordinate z_2^s). More specifically, the ratio between female and male employees is 1.19 (without the normalizing constant and logarithm), full-time contracts prevail the part-time almost by a factor of fifteen, the group of 25 + employees is about 4.6 times bigger than the group of youngest ones and, finally, there are about twice more employees aged between 25 and 54 than the oldest ones (55 +). Another possible interpretation is in terms of an average log-ratio between the given groups of employees across all combinations of the remaining factors. E.g., if the coordinate z_1^s is divided by its normalizing constant $\sqrt{3}$, it turns out that the average log-ratio between female and male employees across all combinations of age groups and types of contract is 0.176. An important source of information are odds ratio coordinates, which for the Czech Republic result in

$$(z_{11}^{rc}, z_{11}^{rs}, z_{12}^{rs}, z_{11}^{cs}, z_{12}^{cs})' = (-0.965, -0.249, 0.391, -0.528, 1.128)'.$$

Coordinate z_{11}^{rc} compares the type of contract of male and female employees, and the negative value indicates that the proportion of males with full-time contract, compared to those employed on part-time, is higher than the same proportion of females or, alternatively, that the proportion of females is higher within employees with a part-time contract than within those with a full-time contract. The raw odds ratio between these four groups (formed by the geometric mean across all age groups) equals 0.33 ($\exp(-0.965\sqrt{4/3})$) and the mean log-odds ratio across the age groups is -1.12 ($-0.965\sqrt{4/3}$). Coordinates z_{11}^{rs} and z_{12}^{rs} compare the age structure of male and female employees and complete the information carried by the balances z_1^r , z_1^s , z_2^s . The coordinate z_1^s reveals that the youngest group (15–24) is dominated by older employees—the coordinate z_{11}^{rs} adds that this dominance tends to be slightly higher for male employees. On the other hand, the value of the coordinate z_{12}^{rs} indicates that the dominance of the age group 25–54 over 55 + tends to be higher for females. Also the coordinates z_{11}^{cs} and z_{12}^{cs} can be interpreted in the sense of odds ratios, by comparing the proportion between full- and part-time contracts in several age groups. The last group of coordinates is formed by z_{11}^{rcs} and z_{112}^{rcs} , for the Czech Republic with values 0.124 and -0.310 , respectively. These coordinates inform about mutual relations between all three factors and their interpretation becomes a bit tricky. Despite of this complexity (comparable to the complexity of double interaction terms in regression models), the interpretation in the sense of a double odds ratio is still possible. For instance, it was already derived that females are employed more often part-time than males; due to a positive value of z_{11}^{rcs} it can be concluded that this relation differs according to the age of employees, specifically it becomes less visible in the youngest group. The proposed system of orthonormal coordinates is appropriate for the further statistical analysis of the relations within each cube. For a more detailed interpretation, the function `cubeCoordWrapper` of the R package `robCompositions`

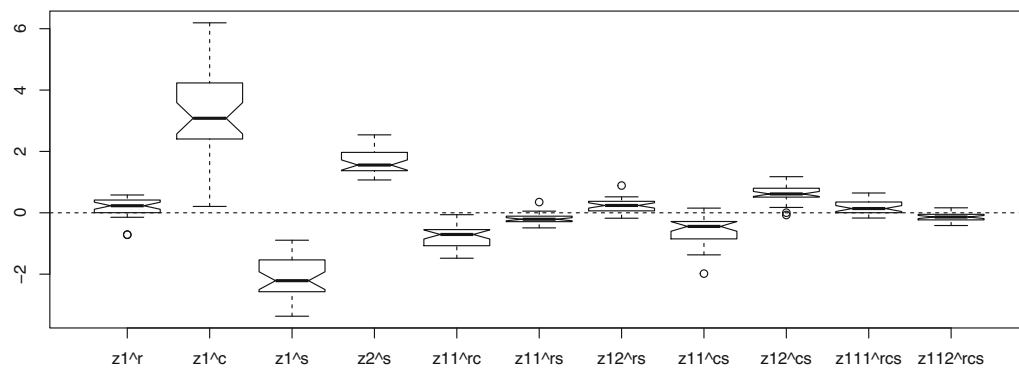


Fig. 4 Boxplots of the coordinates describing the employment structure

Table 2 List of sample means, standard deviations, and 95% bootstrap confidence intervals for the mean of the coordinates describing the employment structure

	Mean	SD	CI		Mean	SD	CI
z_1^r	0.171	0.322	(0.064, 0.271)	z_{11}^{rs}	-0.182	0.164	(-0.235, -0.124)
z_1^c	3.246	1.289	(2.849, 3.697)	z_{12}^{rs}	0.230	0.217	(0.158, 0.307)
z_1^s	-2.102	0.638	(-2.323, -1.903)	z_{11}^{cs}	-0.591	0.490	(-0.752, -0.426)
z_2^s	1.666	0.411	(1.538, 1.809)	z_{12}^{cs}	0.631	0.286	(0.537, 0.724)
z_{11}^{rc}	-0.812	0.333	(-0.919, -0.701)	z_{111}^{rcs}	0.179	0.222	(0.108, 0.255)
				z_{112}^{rcs}	-0.134	0.135	(-0.182, -0.087)

also allows to compute all coordinates without the normalizing constant and therefore to easier quantify the respective relations.

3.3 Statistical analysis

Since a sample of 32 compositional cubes is available, this sample can be investigated in the light of the relative structure. Due to the geographical and economical proximity of some countries, the assumption of independence of the observations seems not to be sufficiently met in this case, and even though the proposed coordinate system is in general designed to allow for any statistical processing, this prevents from using standard inference here. First of all, the behavior of the coordinates in the sample can be described using boxplots, see Fig. 4, and 95% bootstrap confidence intervals for the means (both computed by `cubeCoordWrapper`), which are collected together with the sample mean values and standard deviations in Table 2.

The scale of the different coordinates presented as boxplots in Fig. 4 is comparable since it always refers to the log-ratios of the employment data. It is obvious that the countries in the sample differ mainly in the coordinate z_1^c , comparing the proportionality of full-time and part-time contracts. Larger differences are also visible in coordinate z_1^s , where negative values for all countries are obtained, and thus employees older than 25 years dominate. According to the bootstrap confidence intervals shown in Table 2, the effects of the different factors and factor combinations represented by the coor-

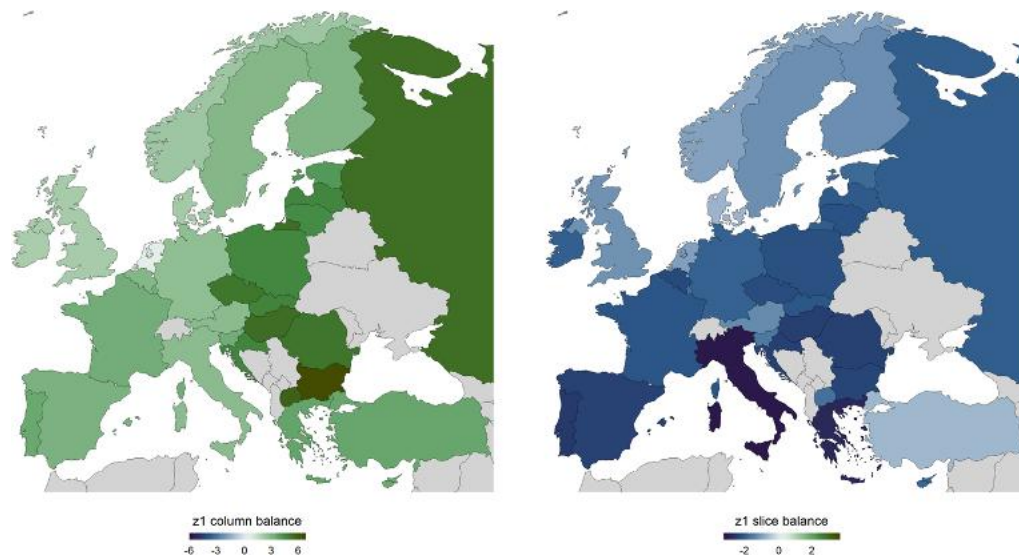


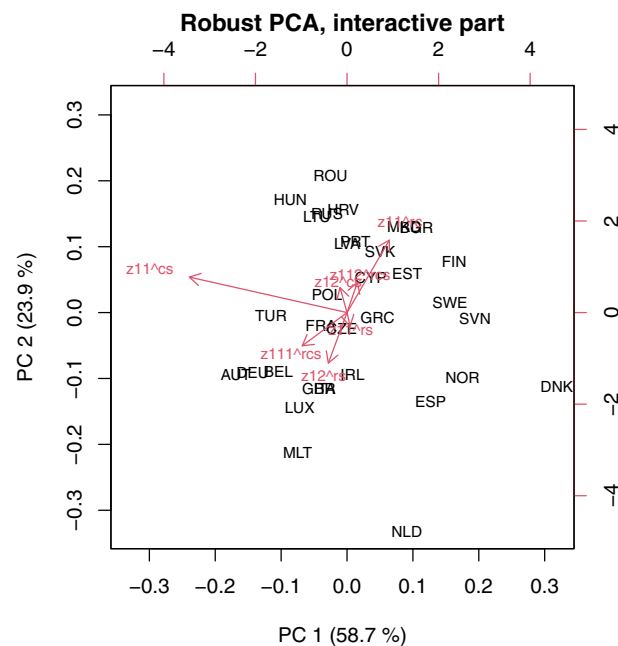
Fig. 5 Values of the coordinate z_1^c representing the log-ratio between full-time and part-time contracts (left) and the coordinate z_1^s , which represents the log-ratio between the youngest group of employees and the rest (right)

dinates are all significant. Thus, not only the previously mentioned simple balances between the factor levels but also interactions between factors strongly influence the overall employment structure.

Besides these general statements, the coordinate representation also allows for a graphical visualization of the regional patterns. For example, the values of the most variable coordinates z_1^c and z_1^s are shown in Fig. 5. From the left map it is clearly visible that high values of coordinate z_1^c , and therefore a big dominance of the full-time contracts, are typical for countries which used to be under the influence of the former Soviet Union. On the other hand, the highest negative values of the coordinate z_1^s , and therefore the biggest prevalence of the older group of employees (25 +), are typical for the southern countries like Italy, Greece or Spain.

Even though the simple balances z_1^c and z_1^s already carry an important piece of information about the employment structure, they suppress the influence of the remaining factors, as described in Sect. 2.3. The possible deviations from the independence between factors are preserved in the coordinates $z_{11}^{rc}—z_{112}^{rcs}$, and the main sources of variability in this regard can be found e.g. by using principal component analysis (PCA) applied on this set of coordinates. Moreover, since the proposed coordinate system respects the dimensionality of the interactive part of the cube, also a robust version of PCA, based on the minimum covariance determinant (MCD) estimates of location and covariance, can be used. This idea was intensively described for the case of compositional tables in de Sousa et al (2021). Note that even though the data can suffer from the presence of a spatial dependence, here classical PCA was used to reveal the main sources of variability. The geographical aspect will be of interest in the later part of the analysis, which is devoted to clustering. The biplot based on the first two robust principal components is shown in Fig. 6. According to this result, we can say that the main sources of differences between the countries in the sample, in terms of

Fig. 6 Biplot of the first two principal components resulting from the robust PCA, performed on the coordinates of the interactive part of the compositional cube



deviations from independence, are the coordinates z_{11}^{rc} and z_{11}^{cs} . The left map in Fig. 7 visualizes the values of the coordinate z_{11}^{rc} . This coordinate is negative for every country in the sample, and therefore the FT/PT ratio is always higher within male employees compared to females. The biggest differences appear in the Netherlands, Belgium, Germany, Austria and Italy; this can be caused e.g. by the popularity of part-time contracts for female employees. On the contrary, in the countries from the eastern part of Europe, where part-time contracts are not very popular in general (see Fig. 5), the difference is less visible. The right map in Fig. 7 shows values of coordinate z_{11}^{cs} . Also this coordinate is mostly negative, the FT/PT ratio is therefore higher for the older group of employees (25 +). Or, conversely, we can say, that the part-time contracts are mostly popular in the age group 15–24. This difference is mostly visible for Denmark and the Scandinavian countries, followed by the Netherlands, Spain and Slovenia. In the remaining countries, the effect of age on the FT/PT ratio is rather negligible.

Robust PCA generates some clusters of countries with similar characteristics in terms of deviations from the independence of the factors. For example, Finland, Sweden, Norway and Denmark have high values on the first component, Romania, Croatia, Bulgaria and Macedonia have high values on the second principal component. A natural question is, whether it is possible to obtain geographically compact clusters of countries with similar employment structure and what are the main features defining these clusters. For this purpose, a clustering method based on two dissimilarity matrices was used Chavent et al (2018). In this method, the first matrix measures the dissimilarity between the numerical variables, in our case the coordinates z_{11}^{rc} – z_{112}^{rcs} , and the second provides information about neighboring countries. The result is shown in Fig. 8. The first cluster is formed by the Scandinavian countries and Slovenia, with high negative values for z_{11}^{cs} and very low, (almost zero) values for coordinate z_{111}^{rcs} .

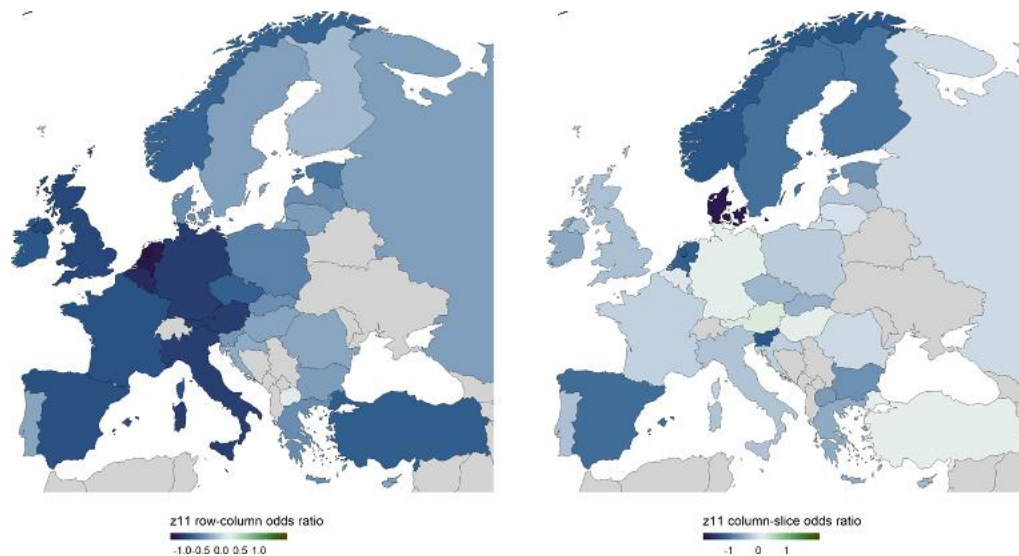


Fig. 7 Values of the coordinate z_{11}^{rc} , comparing the FT/PT ratio between female and male employees (left), and the coordinate z_{11}^{cs} , which compares the same ratio but between the youngest group of employees and the rest (right)

This shows that these countries are characterized by a high popularity of part-time contracts within the youngest group of employees (see the right map on the Fig. 7 for comparison), which moreover holds despite their gender. The second cluster includes countries from Western and Southern Europe, such as Austria, Germany, France, Italy and Spain, with high popularity of part-time contracts for female employees. This property is represented with high negative values of coordinate z_{11}^{rc} and is clearly visible in Fig. 7. Finally, the Baltic countries and Russia form another compact cluster. They are characterized by a low difference in the ratio between female and male employees within the middle age (25–54) and the oldest (55+) group (z_{12}^{rs} close to zero) and also by a small difference in the same ratio within the employees for full-time and part-time contracts (z_{11}^{rc} close to zero).

4 Example: mobility data

In the second example, the interest is in the change of the mobility behavior of people in Austria within the time period February 3rd until August 2nd, 2020, thus during the first period of the COVID'19 pandemic. Mobility is measured through the radius of gyration (ROG), a time-weighted distance of the daily movement locations of mobile phones to the main location of the phone owner, see Heiler et al (2020a, b) for details. The phone owners are classified with respect to gender and five age groups (15–29, 30–44, 45–59, 60–74, 75+). The mobility of each group is represented by the respective median value of ROG. This dataset was already analyzed in Heiler et al (2020a), where the relative differences in the mobility between the age groups were studied through the clr coefficients. This compositional analysis showed an interesting change in the behavior of the youngest (15–24) and oldest (75+) part of the population

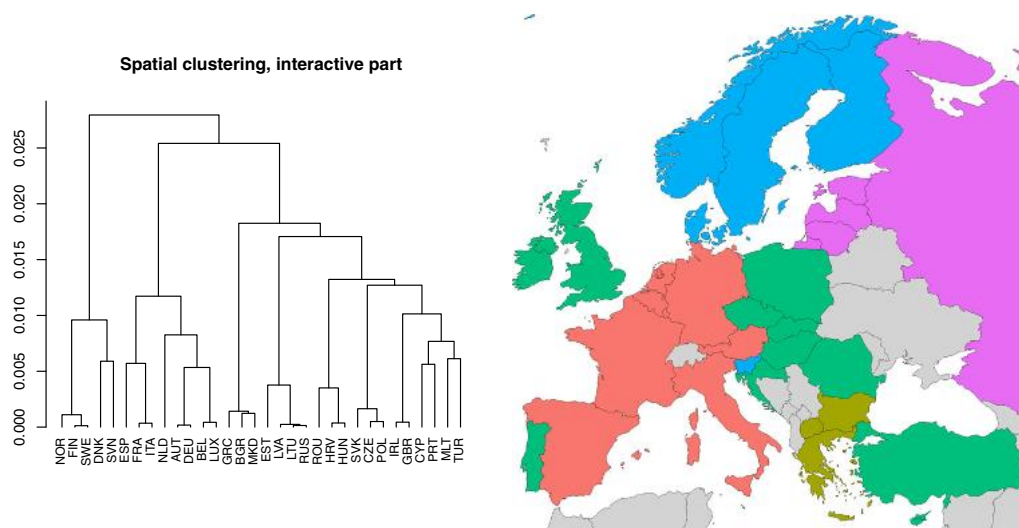


Fig. 8 Result of spatial clustering based on coordinates $z_{11}^{rc} - z_{112}^{rcs}$ representing the interactive part of compositional cubes

during the lockdown period from March 16th to April 6th, 2020 (weeks 12, 13 and 14). The current results are based on the separate analysis of males and females, when a more complex insight can be reached by a simultaneous study of the age and gender structure. Moreover, when the daily records are aggregated according to the part of the week, the relative differences in mobility over weekdays and weekends can be taken into account. The data at hand can therefore be understood as time series of 26 compositional cubes, each representing the relative mobility structure within one week from weeks 6–31 of 2020. The row levels are formed by gender (F—female, M—male), the columns by the parts of the week (WD—weekdays, WE—weekend), and the slices represent the different age groups.

Prior to the main part of the analysis, the SBP of the slice factor needs to be defined. With respect to the findings in Heiler et al (2020a), a separation of the economically active (15–59) and non-active (60+) population seems to be advisable in the first step. The results of the simple clr-based analysis of Heiler et al (2020a) help to define also the remaining steps of the SBP: In the second step, the youngest group is separated from those aged between 25 and 59, and the third step focuses on the relative dominance of group 30–44 over 45–59. Finally, the relative dominance of mobility within the age group 75+ over 60–74 is highlighted by the last step of the slice SBP.

Based on this coordinate representation, some interesting patterns and their sources can be revealed by PCA. As the main aim of the analysis is to illustrate the principles of working with multi-factorial compositional data, the classical approach was applied here. Note, however, that considering the time dependent nature of the data would lead to more accurate results. Figure 9 shows biplots based on the first four principal components, describing 98% of the whole variability.

In order to show the whole structure of the PCA results, the loadings of all PCs are collected in Fig. 10. However, the vast majority of the variability (98%) is explained by the first four principal components, whose values are always mainly driven by a single coordinate:

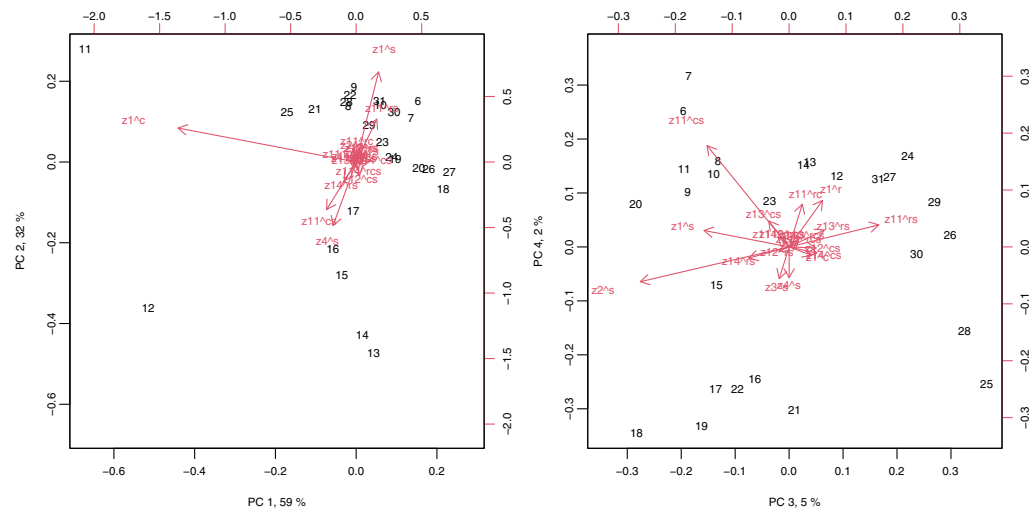


Fig. 9 Biplots based on the first four principal components computed for the mobility data. The numbers represent the week numbers in 2020



Fig. 10 Loadings of the principal components computed for the mobility data

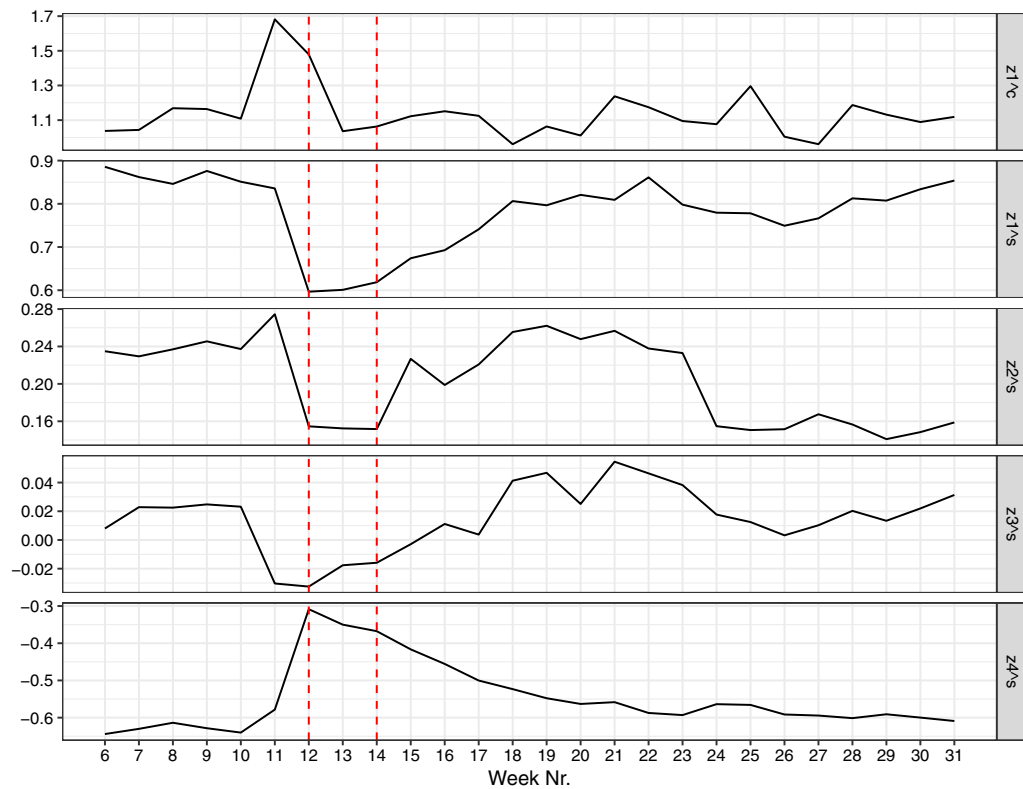


Fig. 11 Selected balances (without the normalizing constant) computed for the mobility data and their development over time. The red dashed lines define the first lockdown period in Austria

- PC1— z_1^c , log-ratio between weekdays and weekend mobility, aggregated over all gender and age groups,
- PC2— z_1^s , log-ratio between mobility of economically active and non-active population, aggregated over all other categories,
- PC3— z_2^s , log-ratio between mobility of age groups 15–29 and 30–59, aggregated over all other categories,
- PC4— z_{11}^{cs} , log-odds-ratio comparing economically active and non-active mobility ratio over weekdays and weekends, aggregated over gender.

The main source of variability is given by the ratio between weekdays and weekend mobility (z_1^c). As it is visible in Fig. 11, this log-ratio was varying over the whole studied period but it was also atypically high during weeks 11 and 12 preceding the lockdown period. This gives an evidence on a decrease of the weekend mobility. The second principal component helps to detect typical characteristics for the weeks during and immediately after lockdown (weeks 12–17). When the relative mobility of the economically active population, quantified by z_1^s , was among the lowest during lockdown, mobility of the oldest group 75+ was nearly comparable to the mobility of the population aged between 60 and 74 (coordinate z_4^s). Moreover, based on coordinate z_{11}^{cs} , Fig. 12 shows a remarkable difference in the economically active and non-active population mobility ratio during weekdays and weekends, when the former tends to be higher during lockdown. The weeks from the end of the observed period (weeks

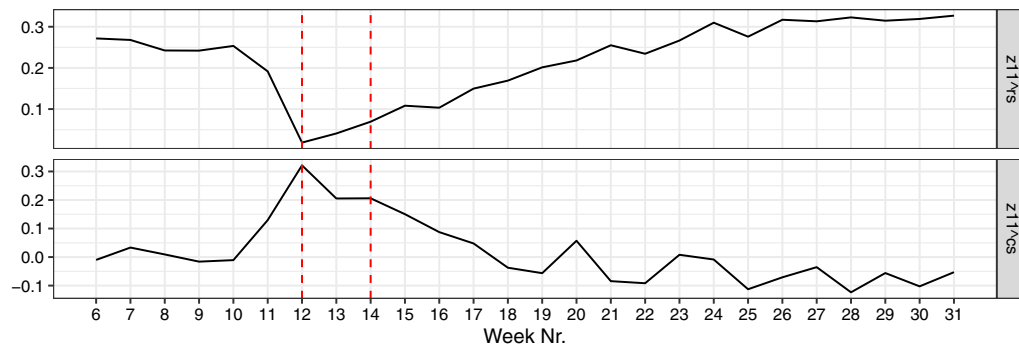


Fig. 12 Selected odds-ratio coordinates (without the normalizing constant) computed for the mobility data and their development over time. The red dashed lines define the first lockdown period in Austria

24–31) are nicely separated by the third principal component. The relative mobility of the youngest group 15–29 is among the lowest in comparison to the mobility of group 30–59 (z_2^S). Moreover, quite stable and high values of coordinate z_{11}^{rS} , comparing the ratio between the economically active and non-active population for females and males, are typical for this period. Finally, PC4 characterizes the behavior during the weeks 16–22 immediately following the lockdown, with the only exception of week 20. According to the respective loadings, the exclusion of weeks 16–22 is not only driven by relatively low values of z_{11}^{CS} , but also by a mixture of other phenomena. For these weeks we observe a high relative dominance of mobility of group 15–29 over the mobility of those aged between 30 and 59 (z_2^S), and a high relative dominance of the mobility of group 30–44 over group 45–59 (z_3^S).

5 Conclusions

It has been demonstrated that the concepts developed for two-factorial compositional data can be extended to compositional cubes, and even to the general k -factorial case. The fundamental idea is to investigate the relative data structure in terms of log-ratios between different factors and factor levels. One advantage of such an approach is scale invariance, which is particularly useful when the reported values of the observations are not comparable (in our example caused by different population sizes of the countries) or if the relative structure is of main interest.

It has been shown that each compositional cube can be decomposed into its independent and interactive parts. Furthermore, the interactive part can be decomposed into cubes representing the pairwise factor interactions and the interaction between all three factors. It turned out that the components of the interactive part have an advantageous property of uniform marginals and, moreover, that the principle of the decomposition can be directly extended to the general case of k -factorial compositions. Since the commonly used systems of orthonormal coordinates are not able to sufficiently describe the multi-factorial nature of cubes and respect the possibility of its decomposition, an alternative system has been proposed. Moreover, this system can be constructed in a flexible manner, basically according to the needs or expert knowledge of the analyst: There might be a certain hypothesis on the relations between factors

and/or factor levels, and based on the principle of sequential binary partitions (SBPs), these combinations can be reflected by the constructed coordinates. Even though the proposed coordinate representation allows to describe the overall relations between factors, similar as in the case of the well developed theory for the analysis of vector compositional data, it is also possible to use them for further analysis with standard statistical methods, or to perform statistical inference with the coordinates, for example, by constructing bootstrap confidence intervals for the mean, in order to determine if the effect conveyed by the coordinate is significant. A proper coordinate representation of the multi-factorial compositional data can therefore be understood as a first step in the analysis, possibly followed by other advanced statistical methods. For example, regression methods with compositional regressors with or without the total (Coenders et al 2017), or any other proper methods of one-factorial (vector) compositional data analysis (Pawlowsky-Glahn et al 2015) can be used (after its possible adaptation to the more complex structure of coordinates).

The idea of modeling interactions between factors using the normalized Hadamard product of vectors, derived from SBPs at the single factor level, works equivalently for a higher number of factors. Therefore, the approach presented here for compositional cubes can be extended in a straightforward manner to higher-order arrays.

Acknowledgements The work was supported by the Czech Science Foundation, Project 22-15684L, and by the Austrian Science Foundation, Project I 5799-N.

Author Contributions All authors contributed to the final version of the manuscript. Kamila Fačevicová developed the main part of the presented theory and prepared the first example. Karel Hron and Peter Filzmoser contributed to the theoretical developments and prepared the second illustrative example. The first draft of the manuscript was written by Kamila Fačevicová and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding The research of KF and KH leading to these results was supported by the Czech Science Foundation, project 22-15684L and work of PF was supported by the Austrian Science Foundation, project I 5799-N.

Data Availability The data analyzed within the first example are freely available at the OECD data repository <http://stats.oecd.org>. The mobility data analyzed in the second example are not publicly available because of data ownership reasons.

Appendix

The SBPs defined in Sect. 3.1, lead to the following contrast matrix,

$$\mathbf{v} = \begin{pmatrix} \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} \\ \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} \\ \sqrt{\frac{1}{6}} & -\sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{24}} & \sqrt{\frac{1}{6}} & -\sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{24}} & \sqrt{\frac{1}{6}} & -\sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{24}} & \sqrt{\frac{1}{6}} & -\sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{24}} \\ 0 & \sqrt{\frac{1}{8}} & -\sqrt{\frac{1}{8}} & 0 & \sqrt{\frac{1}{8}} & -\sqrt{\frac{1}{8}} & 0 & \sqrt{\frac{1}{8}} & -\sqrt{\frac{1}{8}} & 0 & \sqrt{\frac{1}{8}} & -\sqrt{\frac{1}{8}} \\ \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} \\ \sqrt{\frac{1}{6}} & -\sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{24}} & \sqrt{\frac{1}{6}} & -\sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{6}} & \sqrt{\frac{1}{24}} & \sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{6}} & \sqrt{\frac{1}{24}} & \sqrt{\frac{1}{24}} \\ 0 & \sqrt{\frac{1}{8}} & -\sqrt{\frac{1}{8}} & 0 & \sqrt{\frac{1}{8}} & -\sqrt{\frac{1}{8}} & 0 & -\sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & 0 & -\sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} \\ \sqrt{\frac{1}{6}} & -\sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{6}} & \sqrt{\frac{1}{24}} & \sqrt{\frac{1}{24}} & \sqrt{\frac{1}{6}} & -\sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{6}} & \sqrt{\frac{1}{24}} & \sqrt{\frac{1}{24}} \\ 0 & \sqrt{\frac{1}{8}} & -\sqrt{\frac{1}{8}} & 0 & -\sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & 0 & \sqrt{\frac{1}{8}} & -\sqrt{\frac{1}{8}} & 0 & -\sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} \\ \sqrt{\frac{1}{6}} & -\sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{6}} & \sqrt{\frac{1}{24}} & \sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{6}} & \sqrt{\frac{1}{24}} & \sqrt{\frac{1}{24}} & \sqrt{\frac{1}{6}} & -\sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{24}} \\ 0 & \sqrt{\frac{1}{8}} & -\sqrt{\frac{1}{8}} & 0 & -\sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & 0 & -\sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & 0 & \sqrt{\frac{1}{8}} & -\sqrt{\frac{1}{8}} \end{pmatrix}$$

The rows of the matrix correspond to coordinates $\mathbf{z}$ specified in Sect. 3.1, and they are ordered in the following way,

$$\mathbf{z} = (z_1^r, z_1^c, z_1^s, z_2^s, z_{11}^{rc}, z_{11}^{rs}, z_{12}^{rs}, z_{11}^{cs}, z_{12}^{cs}, z_{111}^{rcs}, z_{112}^{rcs})'.$$

The columns correspond to parts of $\ln[\text{vec}(\mathbf{x})]$. Vector $\text{vec}(\mathbf{x})$, the vectorized form of the compositional cube, consists of parts

$$\text{vec}(\mathbf{x}) = (x_{111}, x_{112}, x_{113}, x_{121}, x_{122}, x_{123}, x_{211}, x_{212}, x_{213}, x_{221}, x_{222}, x_{223})'.$$

An alternative partition can be defined for the slice factor—the age of the employees. If the relative dominance of the oldest group over the remaining two is of interest, one can separate this group from the rest first, when defining the slice SBP. The second step can then separate the middle-aged group (25–54 years) from the youngest, which leads to the coordinate quantifying the relative dominance of employees aged between 25 and 54 years over those younger than 25. This strategy would alter the structure of the generating vectors, which now result in

$$\xi_1^{s*} = \sqrt{\frac{1}{6}}(-0.5, -0.5, 1, -0.5, -0.5, 1, -0.5, -0.5, 1, -0.5, -0.5, 1)'$$

and

$$\xi_2^{s*} = \sqrt{\frac{1}{8}}(-1, 1, 0, -1, 1, 0, -1, 1, 0, -1, 1, 0)'$$

The generating vectors which are not related to the values of the slice factor remain the same as in Sect. 3.1. More specifically,

$$\xi_1^{r*} = \xi_1^r, \quad \xi_1^{c*} = \xi_1^c, \quad \xi_1^{rc*} = \xi_1^{rc}.$$

All the remaining vectors change as follows:

$$\begin{aligned}\xi_{11}^{\text{RS}*} &= \sqrt{\frac{1}{6}}(-0.5, -0.5, 1, -0.5, -0.5, -1, 0.5, 0.5, -1, 0.5, 0.5, -1)' \propto \xi_1^{\text{R}*} \circ \xi_1^{\text{S}*}, \\ \xi_{12}^{\text{RS}*} &= \sqrt{\frac{1}{8}}(-1, 1, 0, -1, 1, 0, 1, -1, 0, 1, -1, 0)' \propto \xi_1^{\text{R}*} \circ \xi_2^{\text{S}*}, \\ \xi_{11}^{\text{CS}*} &= \sqrt{\frac{1}{6}}(-0.5, -0.5, 1, 0.5, 0.5, -1, -0.5, -0.5, 1, 0.5, 0.5, -1)' \propto \xi_1^{\text{C}*} \circ \xi_1^{\text{S}*}, \\ \xi_{12}^{\text{CS}*} &= \sqrt{\frac{1}{8}}(-1, 1, 0, 1, -1, 0, -1, 1, 0, 1, -1, 0)' \propto \xi_1^{\text{C}*} \circ \xi_2^{\text{S}*}, \\ \xi_{111}^{\text{RCS}*} &= \sqrt{\frac{1}{6}}(-0.5, -0.5, 1, 0.5, 0.5, -1, 0.5, 0.5, -1, -0.5, -0.5, 1)' \propto \xi_1^{\text{R}*} \circ \xi_1^{\text{C}*} \circ \xi_1^{\text{S}*}, \\ \xi_{112}^{\text{RCS}*} &= \sqrt{\frac{1}{8}}(-1, 1, 0, 1, -1, 0, 1, -1, 0, -1, 1, 0)' \propto \xi_1^{\text{R}*} \circ \xi_1^{\text{C}*} \circ \xi_2^{\text{S}*}.\end{aligned}$$

The new system of coordinates $\mathbf{z}^*$ can be obtained through the Eq. (4), with the contrast matrix

$$\mathbf{T} = \begin{pmatrix} \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} \\ \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} \\ -\sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{24}} & \sqrt{\frac{1}{6}} & -\sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{24}} & \sqrt{\frac{1}{6}} & -\sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{24}} & \sqrt{\frac{1}{6}} & -\sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{24}} & \sqrt{\frac{1}{6}} \\ -\sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & 0 & -\sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & 0 & -\sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & 0 & -\sqrt{\frac{1}{8}} & -\sqrt{\frac{1}{8}} & 0 \\ \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} \\ -\sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{24}} & \sqrt{\frac{1}{6}} & -\sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{24}} & \sqrt{\frac{1}{6}} & \sqrt{\frac{1}{24}} & \sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{6}} & \sqrt{\frac{1}{24}} & \sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{6}} \\ -\sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & 0 & -\sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & 0 & \sqrt{\frac{1}{8}} & -\sqrt{\frac{1}{8}} & 0 & \sqrt{\frac{1}{8}} & -\sqrt{\frac{1}{8}} & 0 \\ -\sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{24}} & \sqrt{\frac{1}{6}} & \sqrt{\frac{1}{24}} & \sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{6}} & -\sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{24}} & \sqrt{\frac{1}{6}} & \sqrt{\frac{1}{24}} & \sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{6}} \\ -\sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & 0 & \sqrt{\frac{1}{8}} & -\sqrt{\frac{1}{8}} & 0 & -\sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & 0 & \sqrt{\frac{1}{8}} & -\sqrt{\frac{1}{8}} & 0 \\ -\sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{24}} & \sqrt{\frac{1}{6}} & \sqrt{\frac{1}{24}} & \sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{6}} & \sqrt{\frac{1}{24}} & \sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{6}} & -\sqrt{\frac{1}{24}} & -\sqrt{\frac{1}{24}} & \sqrt{\frac{1}{6}} \\ -\sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & 0 & \sqrt{\frac{1}{8}} & -\sqrt{\frac{1}{8}} & 0 & \sqrt{\frac{1}{8}} & -\sqrt{\frac{1}{8}} & 0 & -\sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & 0 \end{pmatrix},$$

or, alternatively, as a rotation of vector $\mathbf{z}$. According to (30) the rotation matrix equals

$$\mathbf{TV}' = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{2} & -\sqrt{\frac{3}{4}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\sqrt{\frac{3}{4}} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\frac{1}{2} & -\sqrt{\frac{3}{4}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\sqrt{\frac{3}{4}} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2} & -\sqrt{\frac{3}{4}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\sqrt{\frac{3}{4}} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2} & -\sqrt{\frac{3}{4}} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\sqrt{\frac{3}{4}} & \frac{1}{2} \end{pmatrix}.$$

In both situations, the coordinate system $\mathbf{z}^*$, highlighting the relative dominance of the oldest group of employees, is formed by the following eleven coordinates:

$$\begin{aligned} z_1^{r*} &= \sqrt{3} \ln \frac{g(\mathbf{x}_{1\bullet\bullet})}{g(\mathbf{x}_{2\bullet\bullet})} & z_{11}^{rs*} &= \sqrt{\frac{2}{3}} \ln \frac{g(\mathbf{x}_{1\bullet 3})\sqrt{g(\mathbf{x}_{2\bullet 1})g(\mathbf{x}_{2\bullet 2})}}{g(\mathbf{x}_{2\bullet 3})\sqrt{g(\mathbf{x}_{1\bullet 1})g(\mathbf{x}_{1\bullet 2})}} \\ z_1^{c*} &= \sqrt{3} \ln \frac{g(\mathbf{x}_{\bullet 1\bullet})}{g(\mathbf{x}_{\bullet 2\bullet})} & z_{12}^{rs*} &= \sqrt{\frac{1}{2}} \ln \frac{g(\mathbf{x}_{1\bullet 2})g(\mathbf{x}_{2\bullet 1})}{g(\mathbf{x}_{2\bullet 2})g(\mathbf{x}_{1\bullet 1})} \\ z_1^{s*} &= \sqrt{\frac{8}{3}} \ln \frac{g(\mathbf{x}_{\bullet\bullet 3})}{\sqrt{g(\mathbf{x}_{\bullet\bullet 1})g(\mathbf{x}_{\bullet\bullet 2})}} & z_{11}^{cs*} &= \sqrt{\frac{2}{3}} \ln \frac{g(\mathbf{x}_{\bullet 13})\sqrt{g(\mathbf{x}_{\bullet 21})g(\mathbf{x}_{\bullet 22})}}{g(\mathbf{x}_{\bullet 23})\sqrt{g(\mathbf{x}_{\bullet 11})g(\mathbf{x}_{\bullet 12})}} \\ z_2^{s*} &= \sqrt{2} \ln \frac{g(\mathbf{x}_{\bullet\bullet 2})}{g(\mathbf{x}_{\bullet\bullet 1})} & z_{12}^{cs*} &= \sqrt{\frac{1}{2}} \ln \frac{g(\mathbf{x}_{\bullet 12})g(\mathbf{x}_{\bullet 21})}{g(\mathbf{x}_{\bullet 11})g(\mathbf{x}_{\bullet 22})} \\ z_{11}^{rc*} &= \sqrt{\frac{3}{4}} \ln \frac{g(\mathbf{x}_{11\bullet})g(\mathbf{x}_{22\bullet})}{g(\mathbf{x}_{12\bullet})g(\mathbf{x}_{21\bullet})} & z_{111}^{rcs*} &= \sqrt{\frac{1}{6}} \ln \frac{x_{113}x_{223}\sqrt{x_{121}x_{122}}\sqrt{x_{211}x_{212}}}{x_{123}x_{213}\sqrt{x_{111}x_{112}}\sqrt{x_{221}x_{222}}} \\ & & z_{112}^{rcs*} &= \sqrt{\frac{1}{8}} \ln \frac{x_{112}x_{222}x_{121}x_{211}}{x_{122}x_{212}x_{111}x_{221}} \end{aligned}$$

References

- Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, New York
- Aitchison J (1982) The statistical analysis of compositional data (with discussion). *J R Stat Soc Ser B (Stat Methodol)* 44(2):139–177
- Aitchison J (1986) The statistical analysis of compositional data. Chapman and Hall, London
- Billheimer D, Guttorp P, Fagan WF (2015) Statistical interpretation of species composition. *J Am Stat Assoc* 96(456):1205–1214
- Chavent M, Kuentz-Simonet V, Labenne A et al (2018) Clustgeo: an r package for hierarchical clustering with spatial constraints. *Comput Stat* 33:1799–1822
- Coenders G, Martín-Fernández JA, Ferrer-Rosell B (2017) When relative and absolute information matter: compositional predictor with a total in generalized linear models. *Stat Model* 17(6):494–512
- de Sousa J, Hron K, Fačevicová K et al (2021) Robust principal component analysis for compositional tables. *J Appl Stat* 48(2):214–233
- Eaton ML (1983) Multivariate statistics. A vector space approach. Wiley, New York
- Egozcue JJ, Pawłowsky-Glahn V (2005) Groups of parts and their balances in compositional data analysis. *Math Geol* 37:795–828
- Egozcue JJ, Pawłowsky-Glahn V, Mateu-Figueras G et al (2003) Isometric logratio transformations for compositional data analysis. *Math Geol* 35(3):279–300

- Egozcue JJ, Díaz-Barrero JL, Pawlowsky-Glahn V (2008) Compositional analysis of bivariate discrete probabilities. In: Daunis-i Estadela J, Martín-Fernández JA (eds) Proceedings of CODAWORK'08, The 3rd compositional data analysis workshop, University of Girona, Spain
- Egozcue JJ, Pawlowsky-Glahn V, Templ M et al (2015) Independence in contingency tables using simplicial geometry. *Commun Stat* 44(18):3978–3996
- Fačevićová K, Hron K, Todorov V et al (2014) Logratio approach to statistical analysis of 2×2 compositional tables. *J Appl Stat* 41(5):944–958
- Fačevićová K, Hron K, Todorov V et al (2016) Compositional tables analysis in coordinates. *Scand J Stat* 43:962–977
- Fačevićová K, Hron K, Todorov V et al (2018) General approach to coordinate representation of compositional tables. *Scand J Stat* 45(4):879–899
- Fačevićová K, Kynčlová P, Macků K (2021) Geographically weighted regression analysis for two-factorial compositional data. In: Filzmoser P, Hron K, Martín-Fernández JA, Palarea-Albaladejo J (eds) *Advances in compositional data analysis*. Springer, Cham, pp 103–124
- Filzmoser P, Hron K, Templ M (2018) *Applied compositional data analysis*. Springer series in statistics. Springer, Cham
- Heiler G, Hanbury A, Filzmoser P (2020a) The impact of covid-19 on relative changes in aggregated mobility using mobile-phone data. [arXiv: 2009.03798](https://arxiv.org/abs/2009.03798)
- Heiler G, Reisch T, Hurt J, et al (2020b) Country-wide mobility changes observed using mobile phone data during covid-19 pandemic. In: 2020 IEEE international conference on big data (big data). IEEE Computer Society, Los Alamitos, CA, USA, pp 3123–3132, <https://doi.org/10.1109/BigData50022.2020.9378374>
- Pawlowsky-Glahn V, Egozcue JJ (2001) Geometric approach to statistical analysis on the simplex. *Stochast Environ Res Risk Assess (SERRA)* 15(5):384–398
- Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015) *Modeling and analysis of compositional data*. Wiley, Chichester
- Templ M, Hron K, Filzmoser P (2011) robCompositions: an R-package for robust statistical analysis of compositional data. In: Pawlowsky-Glahn V, Buccianti A (eds) *Compositional data analysis: theory and applications*. Wiley, Chichester, pp 341–355

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Correspondence analysis

Fačevicová, K., Filzmoser, P., Hron, K. Correspondence analysis from the viewpoint of compositional tables. *Statistical Analysis and Data Mining*, 18(4) (2025). <https://doi.org/10.1002/sam.70023>

[IF 1.1 Q3; AIS 0.539 Q3]

[3 citations on WoS; 1 citation without autocitations]

[Author's contribution: 50%]

Abstract

Correspondence analysis (CA), a well-known method for analyzing the relationships between rows and columns of a table, has been reformulated to link to the logratio methodology of compositional data by using the limiting case of the power transformation. The resulting methodology investigates relative rather than absolute information, and it is invariant with respect to rescaling rows or columns. The latter properties also hold for the analysis of compositional tables, where the table is first decomposed into an independent and an interaction part. It is shown that the analysis of the interaction part is equivalent to CA, but in addition, the variance contributions can be determined. Both concepts also allow for an inclusion of weights to suppress undesirable variance, and it is shown that the equivalence between weighted CA and the analysis of weighted compositional tables again holds. This equivalence allows us to make use of the mathematical framework of weighted compositional tables, the so-called Bayes spaces, to get a deeper understanding of CA and to construct extensions to multi-factorial tables (cubes, etc.).

RESEARCH ARTICLE OPEN ACCESS

Correspondence Analysis From the Viewpoint of Compositional Tables

Kamila Fačevicová¹ | Peter Filzmoser²  | Karel Hron¹ ¹Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacky University Olomouc, Olomouc, Czech Republic |²Institute of Statistics and Mathematical Methods in Economics, TU Wien, Vienna, Austria**Correspondence:** Peter Filzmoser (peter.filzmoser@tuwien.ac.at)**Received:** 10 January 2025 | **Revised:** 25 March 2025 | **Accepted:** 29 March 2025**Funding:** This work was supported by the Austrian Science Fund (grant DOI: 10.55776/15799); the European Commission (no. CZ.02.01.01/00/23_025/008686); and the Czech Science Foundation (grant 22-15684L).**Keywords:** centered logratio transformation | compositional data | compositional tables | correspondence analysis | singular value decomposition

ABSTRACT

Correspondence analysis (CA), a well-known method for analyzing the relationships between rows and columns of a table, has been reformulated to link to the logratio methodology of compositional data by using the limiting case of the power transformation. The resulting methodology investigates relative rather than absolute information, and it is invariant with respect to rescaling rows or columns. The latter properties also hold for the analysis of compositional tables, where the table is first decomposed into an independent and an interaction part. It is shown that the analysis of the interaction part is equivalent to CA, but in addition, the variance contributions can be determined. Both concepts also allow for an inclusion of weights to suppress undesirable variance, and it is shown that the equivalence between weighted CA and the analysis of weighted compositional tables again holds. This equivalence allows us to make use of the mathematical framework of weighted compositional tables, the so-called Bayes spaces, to get a deeper understanding of CA and to construct extensions to multi-factorial tables (cubes, etc.).

1 | Introduction

Correspondence analysis (CA) is a prominent method in exploratory data analysis, with the aim to analyze the relationships in a contingency table with either discrete-valued or continuous entries [1–3]. The main idea is to subtract the product of row and column marginals from the proportional representation of the contingency table (referring to the “correspondence matrix”), rescale it to the totals of the marginals, and proceed with a singular value decomposition (SVD). This yields row and column information of the table, which can be visualized in order to study their relationships.

An interesting case is studied in Greenacre [4], where in a first step the elements of a $I \times J$ contingency table $\mathbf{X} = (x_{ij})$ are

transformed with the Box–Cox transformation [5] with power parameter α ,

$$x_{ij}(\alpha) = \begin{cases} (1/\alpha) \left(x_{ij}^\alpha - 1 \right), & \alpha > 0, \\ \ln x_{ij}, & \alpha = 0. \end{cases} \quad (1)$$

If α approaches zero, CA is essentially (in the limiting case) based on log-transformed data, and this provides the link to a compositional data analysis of the table [4]. Generally speaking, compositional data are understood as observations carrying relative information [6], and also CA makes use of a decomposition of a specific kind of relative information. If log-transformed data are used, an SVD of row- and column-wise centered data allows to construct the so-called compositional biplot [7], with the goal

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). Statistical Analysis and Data Mining published by Wiley Periodicals LLC.

to visually study the relationships between rows and columns of the table.

In the context of compositional data, row centering is presented by the so-called centered logratio (clr) coefficients [6], defined for the i th composition $\mathbf{x}_i$, that is, the i th row of a table $\mathbf{X}$ with entries x_{ij} ($i = 1, \dots, I$ and $j = 1, \dots, J$), as

$$\text{clr}(\mathbf{x}_i) = \left(\ln \frac{x_{i1}}{\sqrt{\prod_{j=1}^J x_{ij}}}, \dots, \ln \frac{x_{iJ}}{\sqrt{\prod_{j=1}^J x_{ij}}} \right). \quad (2)$$

With the clr coefficients, the compositional data are moved isometrically from their original sample space, endowed with the Aitchison geometry, to the real space (see, e.g., Pawłowsky-Glahn et al. [8]), where indeed standard SVD after (column-)centering results in a meaningful representation of loadings and scores in a biplot. Since logarithms of ratios are involved, this kind of procedure is generally known under the name logratio approach (see, e.g., Pawłowsky-Glahn et al. [8], Greenacre [9], and Filzmoser et al. [10]).

Recent developments in compositional data analysis, however, enable to proceed further with this limiting case. In particular, the concept of compositional tables opens up new possibilities also for the analysis of relative information in contingency tables [11–13]. As an example, a simple compositional table could be the number of employed people in a region, where the rows are determined by part-time and full-time employment, and the columns by males and females. If one is interested in comparing and analyzing different regions, relative rather than absolute information needs to be considered, as the absolute numbers would essentially be determined by the population size.

To make the link to the mentioned limiting case of correspondence analysis as presented in Greenacre [4, 14], in this paper only one table is considered, and the interest is again in identifying relationships between rows and columns. More specifically, the possibility of a decomposition into two tables, an independent and an interaction part, will be utilized. The former assumes independence of the row and column factors and consists of a product of the respective marginals, and the latter contains the information about their interaction and is numerically equal to the matrix used in correspondence analysis. However, the construction of the interaction table enables to completely filter out in a geometrically meaningful way the independent part from the original table, and thus provides direct pathways to analyzing the remaining interactions. Besides the description of CA from the perspective of compositional tables, the paper also introduces a new mathematical framework for weighting parts of a compositional table and, consequently, for weighted CA, and establishes a solid foundation for an extension to the multi-factorial problem.

The structure of the paper is as follows. In the next section, a key relationship between double-centered log-transformed data and the interaction part of a compositional table (correspondence table) is derived. In Section 3, we generalize the findings to weighted versions of the methods and investigate their equivalence. Section 4 shows that the important property of distributional equivalence also holds for the logratio approach.

Numerical experiments which reveal the advantages of using weights are presented in Section 5, and the final Section 6 concludes and provides an outlook to further extensions.

2 | Unweighted CA

This section recalls classical (unweighted) CA as well as logratio analysis (LRA), performing CA on log-transformed data. Moreover, we present the concept of compositional tables and show the link to LRA.

2.1 | Logratio Analysis (LRA)

Let the $I \times J$ contingency table $\mathbf{X} = (x_{ij})$ be given, either in form of counts or as continuous numbers, and let $\mathbf{P} = \mathbf{X}/n$, $n = \sum_{i,j} x_{ij}$, be the respective matrix of proportions. Classical CA is typically based on an SVD of the residual matrix

$$\mathbf{S}^{\text{CA}} = \mathbf{D}_r^{-1/2} (\mathbf{P} - \mathbf{r}\mathbf{c}') \mathbf{D}_c^{-1/2}, \quad (3)$$

where $\mathbf{r} = \mathbf{P}\mathbf{1}_J$ and $\mathbf{c} = \mathbf{P}'\mathbf{1}_I$ are vectors of row and column marginals of $\mathbf{P}$, with $\mathbf{1}_l$ a vector of l values of 1, and $\mathbf{D}_r = \text{diag}(\mathbf{r})$ and $\mathbf{D}_c = \text{diag}(\mathbf{c})$.

If the rows of the matrix $\mathbf{X}$ represent realizations of a J -part compositional vector, that is, a vector with positive values carrying relative information [6], the unweighted LRA is given by an SVD of the double-centered matrix $\mathbf{L} = (\ln x_{ij})$ [15]. Row-wise centering of $\mathbf{L}$ can be understood as the clr transformation of the rows of $\mathbf{X}$ [8], where the entries of $\text{clr}(\mathbf{x}_i)$, $i = 1, \dots, I$, are given as

$$\tilde{y}_{ij} = \ln x_{ij} - \frac{1}{J} \sum_{j=1}^J \ln x_{ij}, \quad j = 1, \dots, J.$$

After additional column centering we obtain the matrix $\mathbf{S}^{\text{LRA}}$ with entries

$$\begin{aligned} s_{ij}^{\text{LRA}} &= \ln x_{ij} - \frac{1}{J} \sum_{j=1}^J \ln x_{ij} - \frac{1}{I} \sum_{i=1}^I \left[\ln x_{ij} - \frac{1}{J} \sum_{j=1}^J \ln x_{ij} \right] \\ &= \ln x_{ij} - \frac{1}{I} \sum_{i=1}^I \ln x_{ij} - \frac{1}{J} \sum_{j=1}^J \ln x_{ij} + \frac{1}{IJ} \sum_{i=1}^I \sum_{j=1}^J \ln x_{ij}, \end{aligned} \quad (4)$$

which is subsequently decomposed with SVD. The matrix $\mathbf{S}^{\text{LRA}}$ has mathematically the same entries as the clr representation of the so-called interaction table, which is introduced in the following section.

2.2 | Compositional Tables

The two-factorial extension of the concept of compositional data to compositional tables [11, 12] treats $\mathbf{X} = (x_{ij})$ as one data object, which follows the idea of CA. A convenient property of the approach is that the table $\mathbf{X}$ can be orthogonally decomposed into an independent and an interactive part with respect to the Aitchison geometry [11]. This can be written as $\mathbf{X} = \mathbf{X}^{\text{ind}} \oplus \mathbf{X}^{\text{int}}$, with $\oplus$ standing for the entry-wise product, known in the compositional context as the operation of perturbation.

The independent table $\mathbf{X}^{\text{ind}}$ follows the situation of independence between row and column factors, and its entries are given as the product of the geometric marginals,

$$x_{ij}^{\text{ind}} = g(x_{i.}) \cdot g(x_{.j}) = \sqrt[J]{\prod_{j=1}^J x_{ij}} \cdot \sqrt[I]{\prod_{i=1}^I x_{ij}}.$$

Thus, the geometric marginals replace the arithmetic marginals $\mathbf{r}$ and $\mathbf{c}$ used in classical CA, see Equation (3). The geometric marginals have an important property: they are orthogonal projections of the compositional table on the information contained in its rows and columns, see Genest et al. [16] for more details. Note that in case of truly independent row and column factors, both the arithmetic and geometric marginals are equivalent [11], which underpins the reasonability of their definition.

The remainder of $\mathbf{X}$ is contained in the interaction table $\mathbf{X}^{\text{int}}$ with entries

$$x_{ij}^{\text{int}} = \frac{x_{ij}}{g(x_{i.}) \cdot g(x_{.j})}.$$

This table carries information on associations between row and column factors. Therefore, it can serve as a natural compositional alternative to matrices used in classical CA (matrix of standardized residuals or matrix of Pearson contingency ratios).

The clr representation of a compositional table $\mathbf{X}$, $\mathbf{Y} := \text{clr}(\mathbf{X})$, has components

$$y_{ij} = \ln x_{ij} - \frac{1}{IJ} \sum_{i=1}^I \sum_{j=1}^J \ln x_{ij} \quad (5)$$

and the orthogonal decomposition still applies, as $\mathbf{Y}$ can be decomposed into $\mathbf{Y}^{\text{ind}} + \mathbf{Y}^{\text{int}}$. The matrices $\mathbf{Y}^{\text{ind}}$ and $\mathbf{Y}^{\text{int}}$ are the clr representations of $\mathbf{X}^{\text{ind}}$ and $\mathbf{X}^{\text{int}}$, with entries given as

$$y_{ij}^{\text{ind}} = \bar{y}_{i.} + \bar{y}_{.j} \quad \text{and} \quad y_{ij}^{\text{int}} = y_{ij} - y_{ij}^{\text{ind}}.$$

The clr representation turns the geometric marginals of $\mathbf{X}$ into the (scaled) arithmetic margins of $\mathbf{Y}$, which can be written as

$$\bar{y}_{i.} = \frac{1}{J} \sum_{j=1}^J \ln x_{ij} - \frac{1}{J} \frac{1}{I} \sum_{i=1}^I \sum_{j=1}^J \ln x_{ij} = \frac{1}{J} \sum_{j=1}^J \left[\ln x_{ij} - \frac{1}{I} \sum_{i=1}^I \ln x_{ij} \right] \quad (6)$$

and

$$\bar{y}_{.j} = \frac{1}{I} \sum_{i=1}^I \ln x_{ij} - \frac{1}{I} \frac{1}{J} \sum_{i=1}^I \sum_{j=1}^J \ln x_{ij} = \frac{1}{I} \sum_{i=1}^I \left[\ln x_{ij} - \frac{1}{J} \sum_{j=1}^J \ln x_{ij} \right]. \quad (7)$$

The entries of $\mathbf{Y}^{\text{int}}$ are thus equal to

$$\begin{aligned} y_{ij}^{\text{int}} &= y_{ij} - y_{ij}^{\text{ind}} = y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} \\ &= \ln x_{ij} - \frac{1}{J} \sum_{j=1}^J \ln x_{ij} - \frac{1}{I} \sum_{i=1}^I \ln x_{ij} + \frac{1}{IJ} \sum_{i=1}^I \sum_{j=1}^J \ln x_{ij}. \end{aligned} \quad (8)$$

Equations (4) and (8) reveal that

$$\mathbf{S}^{\text{LRA}} = \mathbf{Y}^{\text{int}} \quad (9)$$

and thus LRA is equivalent to an SVD of the clr-transformed interaction table. However, the equivalence is achieved just

numerically, but from a mathematical perspective the models are not the same. While the former approach understands $\mathbf{X}$ as a realization of I observations, the latter treats the entire matrix $\mathbf{X}$ as one observation.

For both the independent and the interaction table, the clr coefficients can also be derived directly from the elements of the original input table $\mathbf{X}$ as

$$y_{ij}^{\text{ind}} = \ln \frac{g(x_{i.})g(x_{.j})}{g(x_{..})^2}, \quad y_{ij}^{\text{int}} = y_{ij} - y_{ij}^{\text{ind}} = \ln \frac{x_{ij}g(x_{..})}{g(x_{i.})g(x_{.j})} \quad (10)$$

de Sousa et al. [17], where $g(x_{..})$ stands for the geometric mean of the entire compositional table. Moreover, $\mathbf{Y}^{\text{int}}$ has uniform (zero) marginals and is thus *margin free* as mentioned in Greenacre [4] (Result 1).

Due to the orthogonality of the independent and interaction tables, Pythagoras' decomposition of squared norms holds [16]. Specifically,

$$\|\mathbf{Y}\|_F^2 = \|\mathbf{Y}^{\text{ind}}\|_F^2 + \|\mathbf{Y}^{\text{int}}\|_F^2, \quad (11)$$

where $\|\cdot\|_F$ is the usual Frobenius norm of a matrix. Accordingly, by computing

$$R_{\Delta}^2(\mathbf{X}) = \frac{\|\mathbf{Y}^{\text{int}}\|_F^2}{\|\mathbf{Y}\|_F^2}, \quad (12)$$

also known under the term “simplicial deviance” [11], one can easily (and in a mathematically justified way) decipher the amount of information stored in the interaction table.

2.3 | CA of a Compositional Table

The equivalence of the double-centered matrix $\mathbf{S}^{\text{LRA}}$ and the clr-transformed interaction table $\mathbf{Y}^{\text{int}}$, see Equation (9), implies that CA is obtained by an SVD of the interaction table,

$$\mathbf{Y}^{\text{int}} = \mathbf{U}\mathbf{D}\mathbf{V}', \quad (13)$$

with orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$, and the (rectangular) diagonal matrix $\mathbf{D} = \text{diag}(d_1, \dots, d_D)$ with $D = \min\{I, J\}$. Double centering of $\mathbf{Y}$ guarantees that the columns $\mathbf{u}_i$ of $\mathbf{U}$, $i = 1, \dots, D$, and the columns $\mathbf{v}_j$ of $\mathbf{V}$, $j = 1, \dots, D$, have zero totals, so they can be perceived as clr transformations of the respective row and column compositions. The SVD can then be written as

$$\mathbf{Y}^{\text{int}} = \sum_{k=1}^D d_k \mathbf{u}_k \mathbf{v}_k',$$

which results in a univariate decomposition of the interaction table (which could possibly be back-transformed to the original space). Clearly,

$$\|\mathbf{Y}^{\text{int}}\|_F^2 = \sum_{k=1}^D d_k^2$$

and therefore $d_k^* = d_k^2 / \sum_{k=1}^D d_k^2$ can be considered as proportion of information (variance) explained by the k th factor—of course,

within the interaction part. By considering Equation (11), one can however relate the factors also to the overall table through $d_k^2/\|\mathbf{Y}\|_F^2$. Finally, plotting the first two columns of $\mathbf{U}$ and $\mathbf{V}$, possibly rescaled by $\sqrt{d_1}$ and $\sqrt{d_2}$, together in a planar graph then informs about interactions between row and column factors, as in classical CA.

3 | Generalization: Weighted CA

In real-world applications, the information on the actual data structure can be blurred due to problems related to sampling of the initial data matrix. The structure can be affected by measurement errors, unbalanced sizes of samples defining the rows, or the presence of cells with low observed value; see, for example, the example in Greenacre and Lewi [15, sec. 2]. Also, in recent work, weighting of rows and columns in CA is proposed and considered useful [18]. In the logratio context, weighting can be used to give less importance in the analysis to components with small proportions that often have high variance on the logratio scale [4, 15, 19]. In the following, we will compare weighted CA/LRA with a weighted CA version for compositional tables.

3.1 | Weighted CA and LRA

Weighting in CA is commonly carried out by introducing row and column weights (typically row and column arithmetic marginals) in the double-centering stage. The vectors $\mathbf{r}$ and $\mathbf{c}$ forming the correspondence matrix $\mathbf{S}^{\text{CA}}$, see Equation (3), are computed with respect to given weights. Consequently, weighting propagates also into the approximation stage, so that fitting is done by weighted least squares [15]. According to Greenacre and Lewi [15], weighted CA can alternatively be motivated by a matrix $\mathbf{Q} = (q_{ij})$ of Pearson contingency ratios $q_{ij} = nx_{ij}/x_i x_j$, with $x_i = \sum_j x_{ij}$, $x_j = \sum_i x_{ij}$, $\forall i, j$. With the weights in form of the row and column marginal vectors $\mathbf{r}$ and $\mathbf{c}$, and the matrices $\mathbf{D}_r = \text{diag}(\mathbf{r})$ and $\mathbf{D}_c = \text{diag}(\mathbf{c})$, see also Equation (3), one can carry out an SVD of $\mathbf{D}_r^{1/2}(\mathbf{I}_I - \mathbf{1}_I \mathbf{r}')\mathbf{Q}(\mathbf{I}_J - \mathbf{c} \mathbf{1}_J')\mathbf{D}_c^{1/2}$ to perform a weighted CA, which then is equivalent to the weighted form of LRA. There are also other important contributions on the equivalence between correspondence analysis and weighted LRA indices: Greenacre [4] provided an empirical description of the respective transformation, Choulakian [20] later presented a mathematical formulation and proof, referring to it as Greenacre's Theorem, and Greenacre [14] subsequently offered an alternative mathematical formulation with a similar proof.

The weighted LRA starts with double centering of $\mathbf{L}$ by weighted means. When $\mathbf{w}_r$ and $\mathbf{w}_c$ are vectors of normalized row and column weights ($\mathbf{1}_I' \mathbf{w}_r = \mathbf{1}_J' \mathbf{w}_c = 1$), weighted LRA is based on an SVD of the matrix

$$\mathbf{S}^{\text{WLRA}} = \mathbf{D}_{w_r}^{1/2}(\mathbf{I}_I - \mathbf{1}_I \mathbf{w}_r')\mathbf{L}(\mathbf{I}_J - \mathbf{w}_c \mathbf{1}_J')\mathbf{D}_{w_c}^{1/2} \quad (14)$$

with $\mathbf{D}_{w_r}$ and $\mathbf{D}_{w_c}$ being diagonal matrices of weights. The result remains unchanged if a constant is added to the rows or columns of $\mathbf{L}$, as it vanishes through the double centering. The matrix $\mathbf{L}$ is therefore equivalent to the matrix of log-transformed Pearson ratios q_{ij} . Moreover, since a logarithm of x can be approximated by $x - 1$ via Taylor approximation, the matrix $\log(\mathbf{Q})$ is similar to

$\mathbf{Q} - \mathbf{1}_I \mathbf{1}_J'$ if all $q_{ij} \rightarrow 1$. Finally, the shift by $\mathbf{1}_I \mathbf{1}_J'$ again vanishes through the double-centering and the weighted LRA and CA approaches are therefore equivalent if the observed and expected values (nearly) coincide, that is, if the analyzed structure is rather independent. The detailed derivation is given in Greenacre and Lewi [15]. Additionally, the weighted LRA is also equivalent to spectral mapping, described, for example, in Lewi [21], and to the weighted version of CA of a compositional table as detailed below.

3.2 | Weighted CA of a Compositional Table

Let $\mathbf{W} = (w_{ij})$ be an $I \times J$ matrix of positive weights satisfying $w_{ij} = w_i^r \cdot w_j^c$, that is, of a structure which is in agreement with weighted LRA [15] and which also corresponds to the product reference measure as used in Genest et al. [16]. Among other options, the vectors of weights $\mathbf{w}^r$ and $\mathbf{w}^c$ can be given by the arithmetic or geometric marginals, and possibly rescaled to unit sum. However, the rescaling would affect merely the scale of the final result, not the (weighted) data structure itself.

The mathematical framework for the weighted analysis of compositional tables, and consequently for the weighted CA of a compositional table, is provided by Bayes spaces [22, 23]. The fundamentals of the weighted analysis of vector compositional data are given in Hron et al. [19] and can be directly generalized to the two-factorial case, as described here. Weighting in the Bayes space framework is in general understood as a shift, in compositional terms a perturbation-subtraction (denoted as $\ominus$) of the object with weights. More specifically, for a compositional table $\mathbf{X}$ and a matrix of weights $\mathbf{W}$, the weighted compositional table $\mathbf{X}^W = \mathbf{X} \ominus \mathbf{W}$ has elements

$$x_{ij}^W = x_{ij}/w_{ij} \quad (15)$$

and it is represented in a weighted Aitchison geometry which honors scale invariance of the table and its weights. The scale of the weights only propagates in metric concepts, like in the weighted Aitchison distance between two tables $\mathbf{X}^W$ and $\mathbf{Z}^W = (z_{ij}^W) = (z_{ij}/w_{ij})$,

$$d(\mathbf{X}^W, \mathbf{Z}^W)^2_w = \frac{1}{2 \sum_{i=1}^I \sum_{j=1}^J w_{ij}} \sum_{i=1}^I \sum_{j=1}^J \sum_{i'=1}^I \sum_{j'=1}^J w_{ij} w_{i'j'} \times \left(\ln \frac{x_{ij}^W}{x_{i'j'}^W} - \ln \frac{z_{ij}^W}{z_{i'j'}^W} \right)^2; \quad (16)$$

depending on rescaling of $\mathbf{W}$ by $c > 0$, the sample space of the tables either shrinks ($c < 1$) or expands ($c > 1$). Note that the weighted Aitchison distance (16) corresponds to quite an extent to the heuristic approach to weighting the distance in compositional data analysis, cf. Greenacre and Lewi [15, sec. 2]. The main difference is the initial shift of the correspondence table by the compositional table of weights, which can be understood as shifting the data to a new representation where the matrix of weights plays the role of the new origin. Accordingly, the weighted observations now indicate how much they differ from the a priori information represented by the weights. This gives an additional theoretical frame to the setting from Greenacre and Lewi [15] which enables to generalize, for example, the

Pythagoras' decomposition or the distributional equivalence from the unweighted case [16, 24]. Despite differences in the weighting pipeline, the final effect is of course the same: cells having higher weights yield more emphasis in the analysis, compare with Talská et al. [24] and fig. 2 therein.

The weights propagate into clr coefficients of the whole table as well as its independent and interaction parts. The entries are given by

$$y_{ij}^W = \sqrt{w_{ij}} \ln \frac{x_{ij}^W}{g_W(x_{..}^W)}, \quad (17)$$

$$\begin{aligned} y_{ij}^{\text{ind},W} &= \sqrt{w_{ij}} \ln \frac{g_W(x_{i.}^W) g_W(x_{.j}^W)}{g_W(x_{..}^W)^2}, \\ y_{ij}^{\text{int},W} &= \sqrt{w_{ij}} \ln \frac{x_{ij}^W g_W(x_{..}^W)}{g_W(x_{i.}^W) g_W(x_{.j}^W)}, \end{aligned} \quad (18)$$

where the weighted geometric mean of $\mathbf{X}$ is defined as

$$g_W(x_{..}^W) = \exp\left(\frac{1}{\sum_{i=1}^I \sum_{j=1}^J w_{ij}} \sum_{i=1}^I \sum_{j=1}^J w_{ij} \ln x_{ij}^W\right),$$

and likewise for row and column geometric means. Note that such definition of marginals corresponds, up to normalizing with the sum of weights, to weighted arithmetic marginals in the log-scale as defined in Greenacre and Lewi [15, sec. 2, Step 1] or Choulakian et al. [25]. The rescaling by $\sqrt{w_{ij}}$ guarantees that the resulting weighted clr coefficients are represented in the (unweighted) Euclidean geometry, and are thus ready for further processing using ordinary tools of multivariate statistics [19]; compare also with Greenacre and Lewi [15, sec. 2, Step 2]. The entries of the interaction table are (up to rescaling by $\sqrt{w_{ij}}$)

$$\begin{aligned} \frac{1}{\sqrt{w_{ij}}} y_{ij}^{\text{int},W} &= \ln x_{ij} - \frac{1}{\sum_{i=1}^I w_i^r} \sum_{i=1}^I w_i^r \ln x_{ij} - \frac{1}{\sum_{j=1}^J w_j^c} \sum_{j=1}^J w_j^c \ln x_{ij} \\ &\quad + \frac{1}{\sum_{i=1}^I w_i^r \sum_{j=1}^J w_j^c} \sum_{i=1}^I \sum_{j=1}^J w_i^r w_j^c \ln x_{ij} \end{aligned} \quad (19)$$

and they play an essential role in CA. Moreover, if we denote sums of the weights $\mathbf{w}^r$ and $\mathbf{w}^c$ by s_r and s_c , respectively, and their normalized versions as $\mathbf{w}_r = \mathbf{w}^r/s_r$ and $\mathbf{w}_c = \mathbf{w}^c/s_c$, the clr representation of the weighted interaction table turns to

$$\begin{aligned} \mathbf{Y}^{\text{int},W} &= \sqrt{s_r s_c} \cdot \mathbf{D}_{w_r}^{1/2} (\mathbf{I} - \mathbf{1} \mathbf{w}_r') \mathbf{L} (\mathbf{I} - \mathbf{w}_c \mathbf{1}') \mathbf{D}_{w_c}^{1/2} \\ &= \sqrt{s_r s_c} \cdot \mathbf{S}^{WLRA}. \end{aligned} \quad (20)$$

Therefore, the weighted version of CA based on a log-transformation (LRA) and the decomposition of a compositional table are equivalent.

It follows directly from the general case [16] that the weighted independent and interaction parts are again orthogonal, and therefore Pythagoras' decomposition

$$\|\mathbf{Y}^W\|_F^2 = \|\mathbf{Y}^{\text{ind},W}\|_F^2 + \|\mathbf{Y}^{\text{int},W}\|_F^2 \quad (21)$$

holds, similar as in Equation (11). Here, a key component is the use of the product reference measure, which is essential also for derivations performed in Genest et al. [16]. Consequently, Equation (21) can be used for diagnostics related to an ordinary SVD of $\mathbf{Y}^W$, analogously as in Section 2.

3.3 | Choice of Weights and Implications

The most appealing case of weighting is definitely the one with the standard arithmetic marginals of the original contingency table or its proportional representation, the correspondence matrix, as w_i^r and w_j^c . Due to scale invariance of weighting in compositional tables, both representations are now equivalent and the rescaling results just in a shrinkage or an expansion of the weighted space [19]. There are good reasons for weighting in the logratio CA: due to the logarithmic scale, row or column factor instances (variables) with small presence engender large variability, which is often a rather undesired effect [15]. The weighting also combines advances of both approaches: simple interpretability of arithmetic marginals as weights is complemented by geometric marginals (i.e., arithmetic margins in the logarithmic scale) which are necessary to develop both important theoretical and practical consequences of the decomposition into independent and interactive parts.

4 | Distributional Equivalence of the Logratio Approach

Distributional equivalence is a natural requirement for CA and, more generally, for analyzing any ratio-scale data, including compositional data and compositional tables. In the former case, this requirement was already emphasized in the seminal work on CA [1], and it was further elaborated by Greenacre and Lewi [15], who used the formulation: *If two columns (resp., two rows) have the same relative values, then merging them does not affect the distances between rows (resp., columns).* An important aspect is what we understand under *merging* in the context of ratio-scale data. As the logarithm naturally moves the data from the ratio-scale to the interval-scale [26], the simple aggregation should be done there, possibly rescaled by the number of components. In the original scale this is just the geometric mean, which is promoted in the literature also for a geometric reasoning [27]. Likewise, in CA, the distributional equivalence is related to an amalgamation of rows/columns in case of their proportionality, which should be done again for similar reasons in the log-scale. Due to scale invariance of compositional tables, from the perspective of the original scale, it is equivalent if the aggregation is done in the log-scale or in the clr space. Then, if any two rows (or columns) of a compositional table carry the same relative information, or in other words, if they are a constant multiple of each other, it is expected that the logratio CA keeps unchanged irrespectively whether these rows (columns) are aggregated.

In case of compositional data, replacing two compositional parts with their respective geometric mean essentially means that the information contained in the ratio between these two parts is removed, and even more, it can be considered as the orthogonal projection of the original composition to the subspace of the

remaining information [28]. Consequently, when considering a sample of compositional data, distances among the original compositions and among their aggregated counterparts remain the same. It is only important to keep the original dimensionality of the data; otherwise, the subcompositional dominance [8] necessarily applies.

The thoughts of Greenacre and Lewi [15] on distributional equivalence can also be reinterpreted to the case of compositional tables. In particular, in case of proportional rows (columns), the interaction part should remain unchanged irrespectively whether these rows (columns) are aggregated or not. This can be easily demonstrated. Without loss of generality, let the first two rows of the compositional table $\mathbf{X} = (x_{ij})$ be proportional, which means that $x_{2j} = c^2 x_{1j}$ for any $c > 0$ and $j = 1, \dots, J$. If each of these rows is replaced by $\sqrt{x_{1j}x_{2j}} = cx_{1j}$, from the scale invariance property it directly follows that the column marginals are the same as in the non-aggregated version of the table. For row marginals we need to consider that

$$g(x_{1.}) = \sqrt[2]{\prod_{j=1}^J x_{1j}}, \quad g(x_{2.}) = \sqrt[2]{\prod_{j=1}^J c^2 x_{1j}} = c^2 g(x_{1.}).$$

The row geometric marginals for a table, where the corresponding elements of the first two rows are replaced by their geometric means, can then be expressed as perturbation of the original column geometric marginal by the I -part composition $(c, 1/c, 1, \dots, 1)$. From the Yule perturbation property (cf. Genet et al. [16, Proposition 7]) it then follows that the interaction table remains unchanged.

Similarly, the same arguments can also be made for the weighted case, with $w_{ij} = w_i^r w_j^c$, $i = 1, \dots, I, j = 1, \dots, J$. By the aggregation of rows we now understand their replacement by the weighted geometric mean, and from the equality

$$\left[(x_{1j}^W)^{w_1^r} (c^2 x_{1j}^W)^{w_2^r} \right]^{\frac{1}{w_1^r + w_2^r}} = c^{\frac{2w_2^r}{w_1^r + w_2^r}} x_{1j}^W = c^r x_{1j}^W$$

it follows that the weighted column marginals are the same irrespectively of the row aggregation. For row marginals, we need to consider that

$$g_W(x_{1.}^W) = \left[\prod_{j=1}^J (x_{1j}^W)^{w_j^c} \right]^{\frac{1}{\sum_j w_j^c}},$$

$$g_W(x_{2.}^W) = \left[\prod_{j=1}^J (c^2 x_{1j}^W)^{w_j^c} \right]^{\frac{1}{\sum_j w_j^c}} = c^2 g_W(x_{1.}^W)$$

and that the geometric mean of an aggregated row equals $c^r g_W(x_{1.}^W)$. The aggregated row marginals can therefore be obtained by the perturbation of the non-aggregated ones by $(c^r, c^r/c^2, 1, \dots, 1)$, which again results in an equivalence of the respective interaction tables. Finally, the proportionality of the rows can be still traced back to the original (unweighted) table, since

$$x_{2j}^W = \frac{x_{2j}}{w_{1j}} = \frac{c^2 x_{1j}}{w_{1j}} = c^2 x_{1j}^W.$$

5 | Numerical Experiments

In this section we aim to complement previous work by Greenacre and Lewi [15] and compare the stability of the results for the unweighted and weighted version of LRA. Accordingly, we perform an SVD of the matrices $\mathbf{Y}^{\text{int}}$ and $\mathbf{Y}^{\text{int},W}$. For the weights we will use the arithmetic marginals of the original table.

5.1 | Bootstrapping Tables

The main idea is to draw bootstrap samples from the original table $\mathbf{X} = (x_{ij})$, where we assume that x_{ij} are integer-valued. Each entry $i = 1, \dots, I$ and $j = 1, \dots, J$ is replicated x_{ij} times, forming the rows of a data matrix in “long format,” with $n = \sum_{i,j} x_{ij}$ rows. Then we draw a bootstrap sample, that is, n observations with replacement, and the resulting long format representation is aggregated to a table format, with the same rows and columns as the original table. Call this bootstrapped table $\mathbf{X}^b$, for $b = 1, \dots, B$, where the number of bootstrap tables B can be large (e.g., 1000). The unweighted version for the original table results in a decomposition of the interaction table with the singular vectors arranged in $\mathbf{U}$ and $\mathbf{V}$, respectively, see Equation (13), and similarly, we obtain the matrices $\mathbf{U}^b$ and $\mathbf{V}^b$ for the interaction table of the bootstrapped table. For the weighted version we obtain the interaction table $\mathbf{Y}^{\text{int},W}$, see Equation (20), and the orthonormal matrices $\mathbf{U}^W$ and $\mathbf{V}^W$ with the left and right singular vectors, respectively. Equivalently, we obtain for each bootstrapped weighted interaction table the corresponding matrices with the singular vectors $\mathbf{U}^{W,b}$ and $\mathbf{V}^{W,b}$.

5.2 | Comparison by the Principal Angle

The results are compared by the *principal angle* between subspaces, as introduced in Björck and Golub [29] and implemented as the function `angle()` in the R package `rospca` [30]. Denote by $\mathcal{U}$ and $\mathcal{V}$ two orthonormal bases, where the number of basis vectors in $\mathcal{U}$ is less or equal the number of basis vectors in $\mathcal{V}$. Then, the principal angle θ between the corresponding subspaces is computed as

$$\theta(\mathcal{U}, \mathcal{V}) = \frac{\sin^{-1}(\sigma_{\max}((\mathbf{I} - \mathbf{U}\mathbf{U}'\mathbf{V}\mathbf{V}')))}{\pi/2}, \quad (22)$$

where $\sigma_{\max}$ corresponds to the largest singular value of the projected matrix [29]. Here, the angle was already scaled to the interval $[0, 1]$, where 0 means that the smaller subspace is embedded in the bigger one (or they span the same space).

In the following experiments, we compare the angle of the results for the original (weighted) table with those for the bootstrapped (weighted) tables, where we use all singular vectors in the first case, but only the first two singular vectors in the second case. Thus, the idea is to see whether the results usually shown in a 2D plot for the bootstrapped version are related or even embedded in the space spanned by the results of the original data version. Note that the angle will be 0 when comparing the singular vectors of the smaller dimension of $\mathbf{X}$, and thus we will report $\max(\theta(\mathbf{U}\mathbf{U}_{1:2}^b), \theta(\mathbf{V}, \mathbf{V}_{1:2}^b))$, and similarly for the weighted versions.

TABLE 1 | Data derived from the Spanish national health survey.

Age group	Very good	Good	Regular	Bad	Very bad
16–24	243	789	167	18	6
25–34	220	809	164	35	6
35–44	147	658	181	41	8
45–54	90	469	236	50	16
55–64	53	414	306	106	30
65–74	44	267	284	98	20
75+	20	136	157	66	17

5.3 | Spanish Health Survey Data

This data set originates from a Spanish health survey and it was analyzed in Greenacre [31]. Table 1 presents the data that have been used for CA. The rows refer to different age groups, and the columns refer to the health status as perceived by the 6371 respondents, with the corresponding frequencies in the cells. In contrast to the analysis presented in Greenacre [31] based on the frequencies, we are here interested in relative information, and thus treat the table as a compositional table. Some categories contain small frequencies, which can introduce a lot of undesirable variability in an unweighted analysis.

Figure 1 presents the results from unweighted and weighted LRA. There are several changes visible, such as an exchange of the positions of “Very bad” and “Regular,” but also of “55–64” and “65–74.” The bootstrapped tables will introduce even more uncertainty in the categories with small frequencies, and the stability of the results will be investigated in the following. It is also interesting to compare the explained variances, shown along the axis legends in the plots of Figure 1: The first numbers refer to the explained variance within the interaction parts, while the second numbers refer to the proportion of explained variance from the overall (weighted) clr table. In the latter case, weighting here leads to a much higher variance proportion because the weights shift the information to

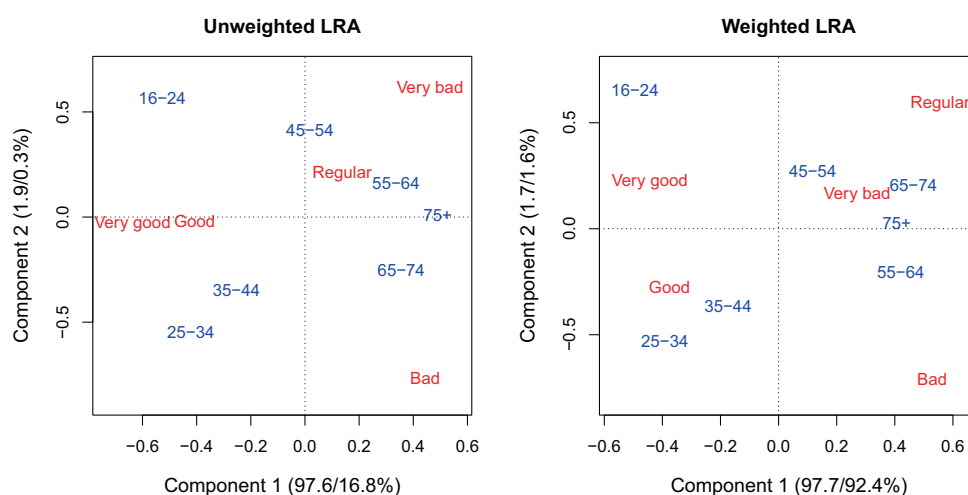
a new origin which removes a lot of the information from the independent part.

Figure 2 presents the results from the bootstrap experiments, as described above. It can happen that zero frequencies occur in a bootstrap table. Such values were replaced simply by 2/3 to add minimum possible variance [32]. The boxplots in the left plot compare the angles of the unweighted LRA with the weighted version, and it can be seen that the angles are clearly smaller for the weighted LRA. The right plot shows for each bootstrap replication the difference of the angle between the unweighted and the angle of the weighted version as boxplots, together with notches for confidence intervals around the median. The boxplots are split up into bootstrap experiments where the smallest value in the bootstrapped table was 0, 1, 2, and so forth, which is shown on the horizontal axis. This reveals that the medians of the differences are positive, and that they tend to be higher if the smallest value is smaller. Thus, weighting stabilizes the results, particularly if there are small frequencies involved.

5.4 | Further Data Sets

We investigate for several other data sets known from the CA literature the stability of the results based on the bootstrap procedure as described above. For the choice of the data sets, we considered different aspects, such as the dimension of the table, values close to zero, low counts in some rows or columns, and so forth. Of course, here we consider the table information as compositional. As before, we will present the results from the bootstrap procedure in terms of the difference between the angle of the unweighted and the weighted version, see Figure 3.

Stores: Age distribution in food stores, used in Chapter 15 of Greenacre [3]. This small data set with counts consists of 5 stores (rows) and 5 age groups (columns). The counts are relatively balanced among the cells, with the smallest value of 8 and the largest of 69. The weighted version shows a slight improvement in stability of the results when compared to the unweighted version (see Figure 3).

**FIGURE 1** | Spanish health data: Results from unweighted (left) and weighted (right) LRA.

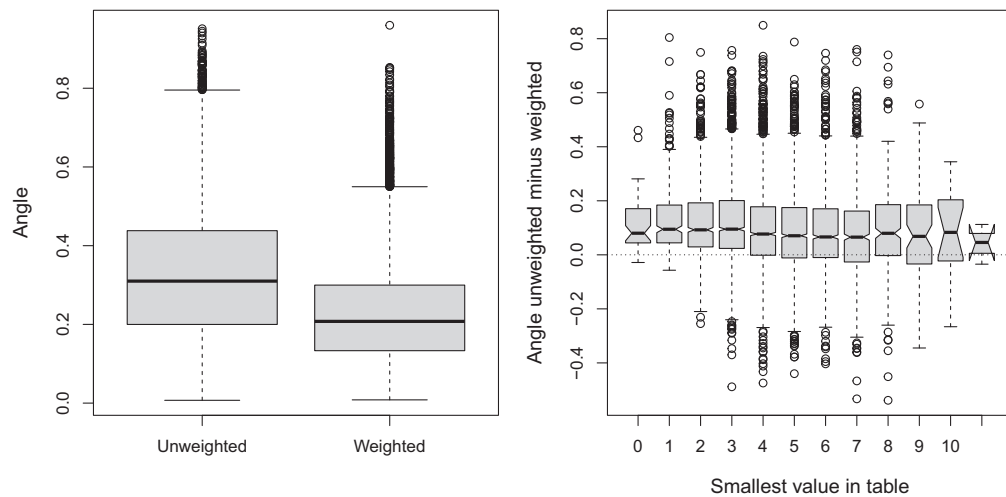


FIGURE 2 | Spanish health data: Angles for unweighted and weighted LRA based on 10,000 bootstrap samples (left), and the difference of the angles of the unweighted and the weighted version, depending on the smallest count in the bootstrapped table (right).

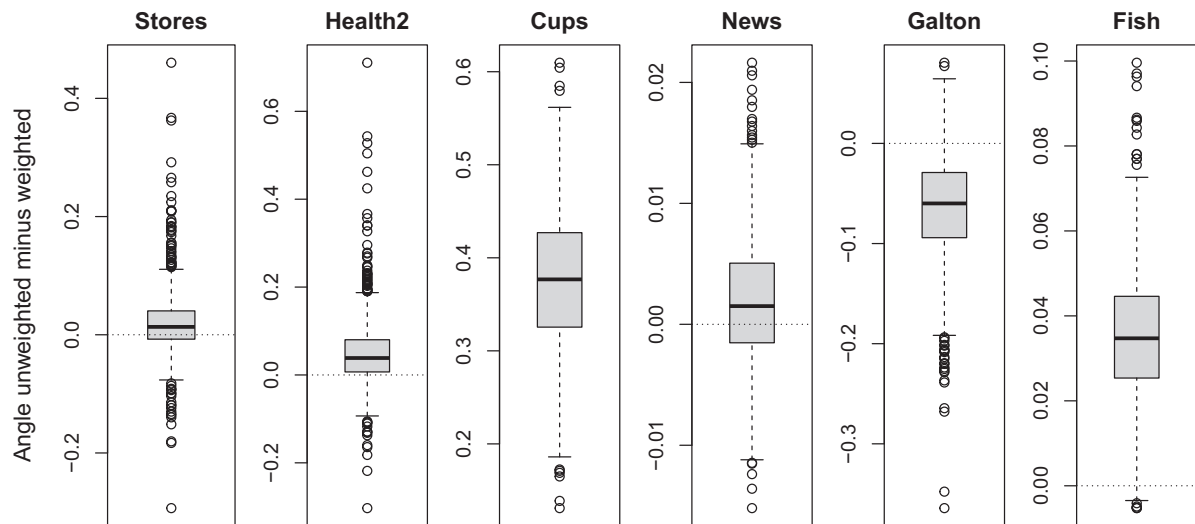


FIGURE 3 | Further data sets: Difference of the angles of the unweighted and the weighted version, based on 1000 bootstrapped tables.

Health2: We consider again the Spanish health data from Table 1, but aggregate the categories “Bad” and “Very bad” in order to avoid small counts. Figure 3 still reveals a clear advantage of weighting, possibly because the categories “Very good” and “75+” have a large variability.

Cups: A data set originating from the analysis of Roman glass cups, see Greenacre and Lewi [15], available as data cups in the R package easyCODA: concentrations of 47 observations for 11 chemical elements. The element “Mn” has very low values due to detection limit problems. This data set was used in Greenacre and Lewi [15] to illustrate the usefulness of weighting. We multiplied the concentrations, reported in % with 2 digits, by 100 to produce integers in order to make the data suitable for our bootstrap procedure. The results in Figure 3 indeed show a huge advantage concerning the stability of the results when using weighting.

News: Data about the news interest in Europe, see Chapter 19 of Greenacre [3]. The table consists of 34 countries (rows) and 18 categories (columns), and the frequencies are in the range from 18 to 652. There are no issues with small frequencies or big variabilities of values for single categories, and thus the results in Figure 3 are not in favor of any of the methods.

Galton: This data set originates from Galton [33], where the body heights of parents and their children are studied. We use the data as aggregated in tab. 1 of Cuadras and Greenacre [34], resulting in a 9×12 table, with 20 cells containing zeros, which are replaced by $2/3$. There are several other cells with small frequencies, distributed among several categories. This is a difficult situation, and here weighting by arithmetic marginals even leads to slightly more instability compared to the unweighted version.

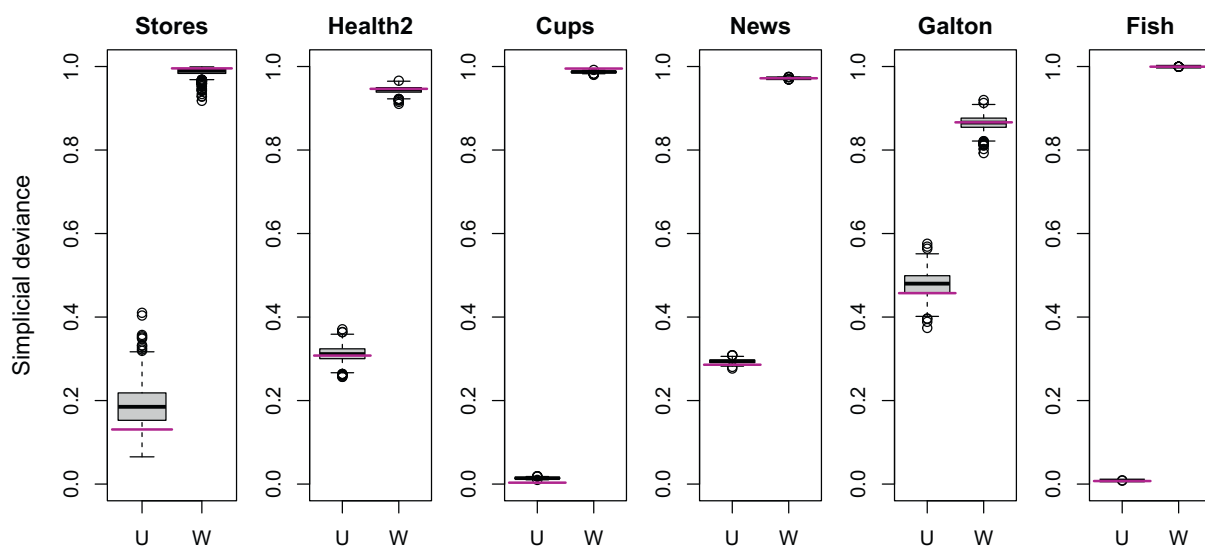


FIGURE 4 | Further data sets: Simplicial deviance of the unweighted (U) and the weighted (W) version, based on 1000 bootstrapped tables. The simplicial deviance for the original data is shown by horizontal lines.

Fish: Morphological data on Arctic charr fish, available as data `fish` in the R package `easyCODA`. We use the 26 morphological measurements for the 75 observations, and multiply the values by 100 to create integers, making them suitable for our bootstrap procedure. There are no particularly small values or categories with large variances in the table, and accordingly, the boxplot shows only a marginal advantage of using weights.

Figure 4 presents the simplicial deviances of the unweighted (U) and weighted (W) versions. The horizontal lines are the values for the original data, and the boxplots present the results of the 1000 bootstrapped tables. In all cases we can see a clear advantage of weighting, which allows to shift much more information to the interaction table.

6 | Summary and Conclusions

CA is generally considered an exploratory data analysis tool. The method is motivated by an algorithm, and there is still a continuous discussion about its mathematical background; see, for example, Breitung [35]. The aim of this paper was to show that the link to the logratio methodology [36] as its limiting case, Greenacre [4] can contribute to build a solid theoretical framework for CA. We have shown that the unweighted LRA, which performs an SVD of the centered logratio represented compositions [4], is equivalent to an analysis of a compositional table. In the latter case, the whole table is considered a composition, and it is not treated as a sample of compositional data. Moreover, the orthogonal decomposition of a compositional table into its independent and interaction parts enables us to assess the explained variability not only within the interaction part (corresponding in the jargon of CA to contingency ratios) but also within the whole (logratio) correspondence table.

Compositional data, and also compositional (correspondence) tables as their two-factorial decomposition, are characterized by

the property of scale invariance [8]. The possibility of approaching unweighted or weighted LRA with power transformation by different representations of the same contingency table (cf. Results 1 and 2 in Greenacre [4]) indicates that scale invariance in a truly compositional sense is not the main strength of CA. Weighting with compositional tables is achieved instead through a change of the reference measure in the respective Bayes space. This leads to the usual choice of weighting in CA with row and column (arithmetic) marginals to new insights, and to a natural increment of the explained variance within the whole table. On top of that, weighting with compositional tables is equivalent to the weighted CA.

The reformulation of CA using compositional tables also guarantees distributional equivalence for both the unweighted and the weighted case, while the unweighted CA so far lacked this feature. Clearly, the aggregation in distributional equivalence needs to be reformulated in terms of the geometric mean, as it is the case for the marginals in a compositional table, and in general as a measure of central tendency in a truly ratio-scale analysis; however, it is merely nothing but the usual aggregation in the log-scale, from the viewpoint of the original scale [16, 26]. Still, for the weighted case also the usual arithmetic marginals play the important role in providing the absolute (interval-scale) information, which is essentially their role in data analysis. Herewith, the benefits of both concepts of marginals can be utilized.

The whole concept can be easily extended to the case of k -factors, $k > 2$, known under the name compositional cubes [37]. They represent the discrete version of the orthogonal decomposition of multivariate densities [16]. Due to the orthogonality of the decomposition, the explained variance of all possible combinations of factors can be assessed which is of particular importance in high-dimensional settings.

Of course, with the logratio approach, the problem of zeros naturally occurs, which needs to be carefully considered. The

zeros in contingency tables can be, however, treated as so-called count zeros where a reasonable imputation by non-zero values is adequate, and approaches for this purpose are available [38]. Moreover, the effect of the imputation (and presence of zeros in general) is naturally downweighted in the weighted logratio CA by lowering the respective marginal values. Still, dealing with zeros in the logratio CA is one of the next challenges to be addressed.

Overall, the logratio approach to CA opens up many new potential avenues for how the field can be further developed. The R source codes for both weighted and unweighted CA using the compositional tables methodology and for the numerical experiments are available at <https://github.com/kfacevicova>.

Author Contributions

K.F., K.H., and P.F. contributed with theory, numerical experiments, and with paper writing.

Acknowledgments

K.H. and K.F. were supported by the Czech Science Foundation (grant 22-15684L) and the project ReDiKid: Resilient Kid in Digital World (reg. no. CZ.02.01.01/00/23_025/008686), co-funded by the European Commission. P.F. was supported by the Austrian Science Fund (grant DOI: 10.55776/15799). Open access funding was provided by Technische Universität Wien/KEMÖ.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

References

1. J.-P. Benzécri, "L'Analyse des Données," in *L'Analyse des Correspondances*, vol. II (Dunod, 1973).
2. L. A. Goodman, "Measures, Models, and Graphical Displays in the Analysis of Cross-Classified Data," *Journal of the American Statistical Association* 86, no. 416 (1991): 1085–1111.
3. M. Greenacre, *Correspondence Analysis in Practice* (Chapman & Hall/CRC, 2017).
4. M. Greenacre, "Log-Ratio Analysis Is a Limiting Case of Correspondence Analysis," *Mathematical Geosciences* 42 (2010): 129–134.
5. G. Box and D. Cox, "An Analysis of Transformations (With Discussion)," *Journal of the Royal Statistical Society, Series B* 35 (1964): 473–479.
6. J. Aitchison, *The Statistical Analysis of Compositional Data* (Chapman & Hall, 1986).
7. J. Aitchison and M. Greenacre, "Biplots of Compositional Data," *Journal of the Royal Statistical Society: Series C: Applied Statistics* 51, no. 4 (2002): 375–392.
8. V. Pawłowsky-Glahn, J. Egozcúe, and R. Tolosana-Delgado, *Modeling and Analysis of Compositional Data* (Wiley, 2015).
9. M. Greenacre, *Compositional Data Analysis in Practice* (CRC Press, 2018).

10. P. Filzmoser, K. Hron, and M. Templ, *Applied Compositional Data Analysis* (Springer, 2018).
11. J. Egozcúe, V. Pawłowsky, M. Templ, and K. Hron, "Independence in Contingency Tables Using Simplicial Geometry," *Communications in Statistics* 44, no. 18 (2015): 3978–3996.
12. J. J. Egozcúe, J. L. Díaz-Barrero, and V. Pawłowsky-Glahn, "Compositional Analysis of Bivariate Discrete Probabilities," in *Proceedings of CODAWORK'08, the 3rd Compositional Data Analysis Workshop*, ed. J. Daunis-i-Estadella and J. Martín-Fernández (University of Girona, 2008).
13. K. Fačevićová, K. Hron, V. Todorov, and M. Templ, "General Approach to Coordinate Representation of Compositional Tables," *Scandinavian Journal of Statistics* 45, no. 4 (2018): 879–899.
14. M. Greenacre, "The Chipower Transformation: A Valid Alternative to Logratio Transformations in Compositional Data Analysis," *Advances in Data Analysis and Classification* 18 (2024): 769–796.
15. M. Greenacre and P. Lewi, "Distributional Equivalence and Subcompositional Coherence in the Analysis of Compositional Data, Contingency Tables and Ratio-Scale Measurements," *Journal of Classification* 26 (2009): 29–54.
16. C. Genest, K. Hron, and J. Nešlehová, "Orthogonal Decomposition of Multivariate Densities in Bayes Spaces and Relation With Their Copula-Based Representation," *Journal of Multivariate Analysis* 198 (2023): 105228.
17. J. de Sousa, K. Hron, K. Fačevićová, and P. Filzmoser, "Robust Principal Component Analysis for Compositional Tables," *Journal of Applied Statistics* 48, no. 2 (2021): 214–233.
18. A. D'Ambra, G. Meccariello, and L. Della Ragione, "Weighted Cumulative Correspondence Analysis Based on a Particular Cumulative Power Divergence Family," *Annals of Operations Research* 342 (2024): 1407–1428.
19. K. Hron, A. Menafoglio, J. Palarea-Albaladejo, P. Filzmoser, R. Talská, and J. Egozcúe, "Weighting of Parts in Compositional Data Analysis: Advances and Applications," *Mathematical Geosciences* 54 (2022): 71–93.
20. V. Choulakian, Scale Invariant Correspondence Analysis (IEEE, 2023).
21. P. Lewi, "Spectral Mapping, a Technique for Classifying Biological Activity Profiles of Chemical Compounds," *Arzneimittel-Forschung/Drug Research* 26, no. 7 (1976): 1295–1300.
22. J. J. Egozcúe, J. L. Díaz-Barrero, and V. Pawłowsky-Glahn, "Hilbert Space of Probability Density Functions Based on Aitchison Geometry," *Acta Mathematica Sinica* 22, no. 4 (2006): 1175–1182.
23. K. van den Boogaart, J. Egozcúe, and V. Pawłowsky-Glahn, "Bayes Hilbert Spaces," *Australian & New Zealand Journal of Statistics* 56, no. 2 (2014): 171–194, <https://doi.org/10.1111/anzs.12074>.
24. R. Talská, A. Menafoglio, K. Hron, J. Egozcúe, and J. Palarea-Albaladejo, "Weighting the Domain of Probability Densities in Functional Data Analysis," *Statistics in Medicine* 9, no. 1 (2020): e283.
25. V. Choulakian, J. de Tibeiro, and P. Sarnacchiaro, "On the Choice of Weights in Aggregate Compositional Data Analysis," *Behaviormetrika* (2024), <https://doi.org/10.1007/s41237-024-00234-5>.
26. G. Mateu-Figueras and V. Pawłowsky-Glahn, "A Critical Approach to Probability Laws in Geochemistry," *Mathematical Geosciences* 40 (2008): 489–502.
27. J. McChesney, "You Should Summarize Data With the Geometric Mean," 2016, <https://jlmc.medium.com/understanding-three-simple-statistics-for-data-visualizations-2619dbb3677a>.

28. J. Egozcue and V. Pawłowsky-Glahn, "Groups of Parts and Their Balances in Compositional Data Analysis," *Mathematical Geology* 37, no. 7 (2005): 795–828.
29. R. Björck and G. H. Golub, "Numerical Methods for Computing Angles Between Linear Subspaces," *Mathematics of Computation* 27, no. 123 (1973): 579–594.
30. T. Reynkens, "rospca: Robust Sparse PCA Using the ROSPCA Algorithm," R Package Version 1.0.4, 2018.
31. M. Greenacre, "Correspondence Analysis of the Spanish National Health Survey," *Gaceta Sanitaria* 16 (2002): 160–170.
32. J. Martín-Fernández, C. Barceló-Vidal, and V. Pawłowsky-Glahn, "Dealing With Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation," *Mathematical Geology* 35, no. 3 (2003): 253–278.
33. F. Galton, "Regression Towards Mediocrity in Hereditary Stature," *Journal of the Anthropological Institute of Great Britain and Ireland* 15 (1886): 246–263.
34. C. Cuadras and M. Greenacre, "A Short History of Statistical Association: From Correlation to Correspondence Analysis to Copulas," *Journal of Multivariate Analysis* 188 (2022): 104901.
35. J. Breitung, "Dr. Strangelove or How I Learned to Stop Worrying and Love Correspondence Analysis," in *Multivariate Scaling Methods and the Reconstruction of Social Spaces*, ed. A. Barth, F. Leßke, R. Atakan, M. Schmidt, and Y. Scheit (Verlag Barbara Budrich, 2023).
36. J. Aitchison, "The Statistical Analysis of Compositional Data (With Discussion)," *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 44, no. 2 (1982): 139–177.
37. K. Fačevicová, P. Filzmoser, and K. Hron, "Compositional Cubes: A New Concept for Multi-Factorial Compositions," *Statistical Papers* 64 (2022): 955–985.
38. J. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser, and J. Palarea-Albaladejo, "Bayesian-Multiplicative Treatment of Count Zeros in Compositional Data Sets," *Statistical Modelling* 15, no. 2 (2015): 134–158.

Backwards pivot coordinates

Nesrstová, V., Jašková, P., Pavlů, I., Hron, K., Palarea-Albaladejo, J., Gába, A., Pelclová, J., **Fačevicová, K.** Simple enough, but not simpler: Reconsidering additive logratio coordinates in compositional analysis. *SORT*, 47(2) (2023). <https://doi.org/10.57645/20.8080.02.9>

[IF 1.6 Q2; AIS 0.561 Q3]

[5 citations on WoS; 5 citations without autocitations]

[Author's contribution: 20%]

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.

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

V. Nesrstová¹, P. Jašková¹, I. Pavlů¹, K. Hron¹, J. Palarea-Albaladejo²,
A. Gába³, J. Pelclová⁴ and K. Fačevicová⁵

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 the 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 deduced 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 rationing part). Specifically, for a D -part composition $\mathbf{x} = (x_1, \dots, x_D)$, and x_D being the rationing part up to a possible permutation, alr coordinates are given by

$$\text{alr}(\mathbf{x}) = \left(\ln \frac{x_1}{x_D}, \dots, \ln \frac{x_{D-1}}{x_D} \right). \quad (1)$$

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

$$\text{clr}(\mathbf{x}) = \left(\ln \frac{x_1}{g(\mathbf{x})}, \ln \frac{x_2}{g(\mathbf{x})}, \dots, \ln \frac{x_D}{g(\mathbf{x})} \right) \quad (2)$$

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

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

where $\mathbf{x}^{(l)} = (x_1^{(l)}, \dots, x_D^{(l)})$, $l \in \{1, \dots, D-1\}$, 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, $\mathbf{x}^{(l)} = (x_D, x_l, x_1, \dots, x_{l-1}, x_{l+1}, \dots, x_{D-1})$. Then the pairwise logratio of interest (pivoting coordinate) is given by

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

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)

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

That is, it is in the pairwise row balance

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

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

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

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

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

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*

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

for $i = 1, \dots, I-1$ and $j = 1, \dots, J-1$. 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.

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

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 Pawłowsky-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 ${}_r{}_c\mathbf{z}(\mathbf{x}^{(kl)})$ 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):

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T, \quad (11)$$

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. $d_{11} \geq d_{22} \geq \dots d_{kk} > 0$, $k \leq \min(n, D)$) 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

$$\mathbf{X} = \mathbf{G}\mathbf{H}^T, \quad (12)$$

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 $\mathbf{Y} = \text{clr}(\mathbf{X})$, 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 $\mathbf{x}_j$, $j = 1, \dots, D$:

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

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

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

- 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

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

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 rationing element. Let $\mathbf{Z}^{(l)}$ be a centered matrix containing the bpcs of the l -th system, $l = 1, \dots, D - 1$. 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 $\text{bpc}(\mathbf{x}^{(l)})_1$ 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

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

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 rationing 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:

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

- 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:

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

- 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 rationing 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 $\mathbf{Z}^{(kl)}$, $k = 1, \dots, I-1$, $l = 1, \dots, J-1$ be a mean centered $n \times (IJ-1)$ matrix of coordinates from a sample of compositional tables $\mathbf{x}^{(kl)}$, $i = 1, \dots, n$ (note that in the following a left subscript will stand for the sample index), with the coordinates being ordered as $\text{tbpc}(\mathbf{x}^{(kl)})$, $\text{rbpb}(\mathbf{x}^{(kl)})$ and $\text{cbpb}(\mathbf{x}^{(kl)})$. That is, the log odds-ratio $\ln(x_{kl}x_{IJ}/x_{kJ}x_{Il})$ 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

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

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

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

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:

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

- 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

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

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 $\mathbf{x}^{(kl)}$, $k = 1, \dots, I - 1$. The following relationship holds between these coordinates and the clr representation:

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

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 $\mathbf{H}^{(kl)}$, $k = 1, \dots, I - 1$, 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 $\mathbf{H}$:

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

Accordingly, the biplot properties are now as follows:

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

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

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

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

- 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:

$$\|\mathbf{h}_{r1\bullet}^{(k_1l)} - \mathbf{h}_{r1\bullet}^{(k_2l)}\|^2 \approx \frac{J}{2} \widehat{\text{var}} \left(\ln \frac{g(\mathbf{x}_{k_1\bullet})}{g(\mathbf{x}_{k_2\bullet})} \right).$$

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

$$\mathbf{g}_{i\bullet}^T (\mathbf{h}_{r1\bullet}^{(k_1l)} - \mathbf{h}_{r1\bullet}^{(k_2l)}) \approx \sqrt{\frac{J}{2}} \ln \frac{g(i\mathbf{x}_{k_1\bullet})}{g(i\mathbf{x}_{k_2\bullet})}.$$

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 $l = 1, \dots, J-1$. 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 $\mathbf{x} = (x_1, \dots, x_D)$. Thus, a system of (orthogonal) bpcs $\mathbf{z}(\mathbf{x}^{(l)}) = (z_1^{(l)}, \dots, z_{D-1}^{(l)})$ can be used to represent the composition in a regression model as

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

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 ($l = 1, \dots, D-1$), 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 $\hat{\beta}_1^{(l)}$, $l = 1, \dots, D-1$.

Regardless the specific permutation (l), β_0 gives an expected value of the response when all compositional parts x_i , $i = 1, \dots, D$, are equal. However, the interpretation of the $\beta_1^{(l)}$, $l = 1, \dots, D-1$, 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 $z_2^{(l)}, \dots, z_{D-1}^{(l)}$ 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 $\beta_1^{(l)}$, $l = 1, \dots, D-1$, 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 $z_2^{(l)}, \dots, z_{D-1}^{(l)}$ from the l -th system, which decrease by $1/2, 1/3, \dots, 1/(D-1)$ respectively. The effect of the κ -times increase of x_l would therefore need to be accounted for from all the regression coefficients $\beta_i^{(l)}$, $i = 1, \dots, D-1$.

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 (Fačevićová 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

$$\begin{aligned} E[Y|\mathbf{z}(\mathbf{x}^{(kl)})] = & \beta_0 + {}_r\beta_1^{(kl)} {}_r z_1^{(kl)} + \dots + {}_r\beta_{I-1}^{(kl)} {}_r z_{I-1}^{(kl)} + \\ & + {}_c\beta_1^{(kl)} {}_c z_1^{(kl)} + \dots + {}_c\beta_{J-1}^{(kl)} {}_c z_{J-1}^{(kl)} + \\ & + {}_t\beta_{11}^{(kl)} {}_t z_{11}^{(kl)} + \dots + {}_t\beta_{I-1,J-1}^{(kl)} {}_t z_{I-1,J-1}^{(kl)}. \end{aligned} \quad (20)$$

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 $k = 1, \dots, I - 1$ and $l = 1, \dots, J - 1$. 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 ${}_r\hat{\beta}_1^{(kl)}$, $k = 1, \dots, I - 1$, the first cbpbs ${}_c\hat{\beta}_1^{(kl)}$, $l = 1, \dots, J - 1$ 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 ${}_r\hat{\beta}_1^{(kl)}$, which quantify the effect of a unit change of the ratios

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

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 ${}_r\hat{\beta}_1^{(kl)}$ 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 ${}_c\hat{\beta}_1^{(kl)}$, which analogously to the above quantify the effect of a unit change of

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

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 ${}_t\hat{\beta}_{11}^{(kl)}$. 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

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

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.

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

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

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

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

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 ${}_r z_i^{(kl)}$, $i = 2, \dots, I-1$, decrease by $1/iJ$, while ${}_r z_1^{(kl)}$ increases by $1/J$. Similarly, the column backwards pivot balances ${}_c z_j^{(kl)}$, $j = 2, \dots, J-1$, decrease by $1/jI$ and ${}_c z_1^{(kl)}$ is increased by $1/I$. Finally, the odds-ratio backwards pivot coordinates ${}_t z_{1j}^{(kl)}$, $j = 2, \dots, J-1$, or ${}_t z_{i1}^{(kl)}$, $i = 2, \dots, I-1$, decrease by $1/j$ and $1/i$ respectively. The remaining ${}_t z_{ij}^{(kl)}$, $i = 2, \dots, I-1$ and $j = 2, \dots, J-1$, are increased by $1/ij$.

Another interesting case of a unit change in ${}_t z_{11}^{(kl)}$ 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 ${}_t z_{1j}^{(kl)}$, $j = 1, \dots, J-1$, which increase by $1/j$, and to the row balance ${}_r z_1^{(kl)}$, which increases by $1/J$. The effect of such a change on the response variable is therefore given by a combination of the parameters ${}_t \beta_{1j}^{(kl)}$, $j = 1, \dots, J-1$, 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.

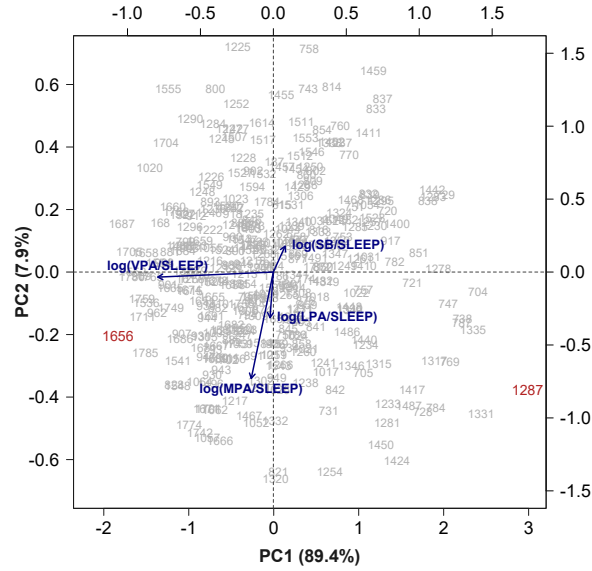


Figure 1. PCA biplot based on backwards pivot coordinates for the movement behavior vector composition with sleep used as reference behavior. The numbers correspond to participant ID. Individuals no. 1287 and no. 1656 (highlighted in red color) differ substantially in time spent in VPA relative to sleep.

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

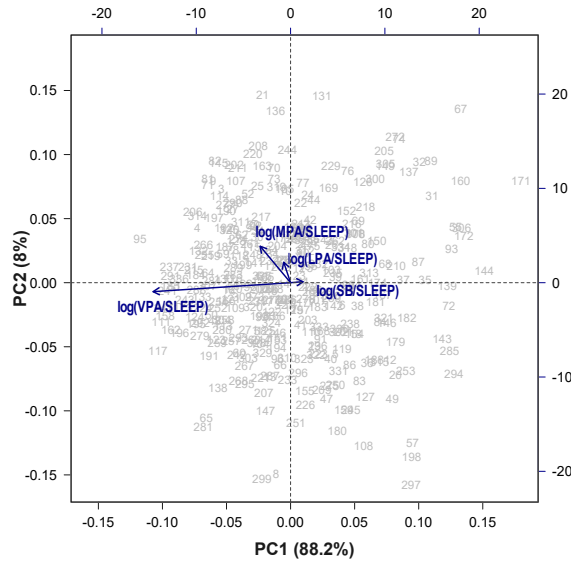


Figure 2. PCA biplot based on *alr* coordinates for the movement behavior vector composition with sleep used as reference behavior. The numbers correspond to participant IDs.

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 $R^2 = 0.119$.

(1)	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 pairwise row balance equal to -0.279).

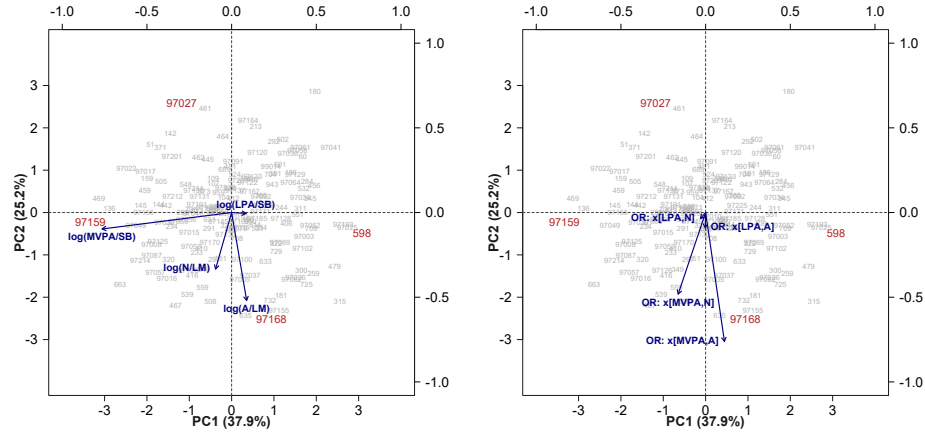


Figure 3. PCA biplots based on backwards pivot coordinates for the movement behavior compositional table with sedentary behavior (SB) and late morning (LM) used as normalizing categories. The numbers correspond to participant ID. Participants further discussed in main text are highlighted in red.

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 $\mathbf{z}(\mathbf{x}^{(kl)})$ 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 $\log_2(g(\text{MVPA})/g(\text{SB}))$ refers to the pairwise row balance between MVPA and SB, while $\log_2\text{OR: } x[\text{MVPA}, N]$ 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●)	$\log_2(g(\text{LPA})/g(\text{SB}))$	-0.029	0.048	-0.607	0.545
(2●)	$\log_2(g(\text{MVPA})/g(\text{SB}))$	-0.105	0.035	-2.987	0.003
(●1)	$\log_2(g(N)/g(\text{LM}))$	0.068	0.157	0.436	0.664
(●2)	$\log_2(g(A)/g(\text{LM}))$	-0.069	0.131	-0.530	0.597
(11)	$\log_2\text{OR: } x[\text{LPA}, N]$	-0.088	0.047	-1.878	0.062
(12)	$\log_2\text{OR: } x[\text{LPA}, A]$	-0.063	0.039	-1.626	0.106
(21)	$\log_2\text{OR: } x[\text{MVPA}, N]$	-0.066	0.042	-1.562	0.120
(22)	$\log_2\text{OR: } x[\text{MVPA}, A]$	-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 $\hat{\beta}_1^{(2\bullet)} = -0.105$. 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 rbp and, thus, the overall effect on VFA is $2^{(\hat{\beta}_1^{(2\bullet)} - \frac{1}{2}, \hat{\beta}_2^{(2\bullet)})} = 2^{(-0.105 + 0.031/2)} = 0.940$ (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 $R^2 = 0.087$).

Another interesting regression coefficient is $\hat{\beta}_{11}^{(11)}$, 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 $2^{\left(\hat{\beta}_{11}^{(11)} + \frac{1}{2}\hat{\beta}_{12}^{(11)} + \frac{1}{3}\hat{\beta}_1^{(11)}\right)}$, 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 multi-factorial 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., Dygrŷn, 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.

Geographically weighted regression

Fačevicová, K., Kynčlová, P., Macků, K. Geographically weighted regression analysis for two-factorial compositional data. In P. Filzmoser, K. Hron, J.A. Martin-Fernández and J. Palarea-Albaladejo, editors, *Advances in Compositional Data Analysis. Festschrift in Honour of Vera Pawlowsky-Glahn*. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-71175-7_6

[3 citations on Scopus; 0 citations without autocitations]

[Author's contribution: 50%]

Abstract

The chapter focuses on the modelling and analysis of spatial dependent two-factorial compositional data. Spatial statistics provides a wide range of methods for the analysis of data with local variations but only a few of them are accommodated for the purposes of modelling relative structures. In this chapter, the geographically weighted regression model is introduced to analyse the relationship between the dependent variable and an explanatory variable reflecting a structure expressed in terms of a compositional table. The methodology is motivated by the problem of modelling local variations of the relationship between at-risk-of-poverty rates and the structure of the highest attained educational level in the German population aged 30–34. The real data study shows how information included in a compositional table and further expressed in real-valued coordinates can be highly valuable in selecting variables and prioritising them with respect to a research interest to facilitate the final interpretation of the model.

Geographically Weighted Regression Analysis for Two-factorial Compositional Data

Kamila Fačevicová and Petra Kynčlová and Karel Macků

Abstract The chapter focuses on modelling and analysis of spatial dependent two-factorial compositional data. Spatial statistics provides a wide range of methods for analysis of data with local variations but only a few of them are accommodated to the purposes of modelling of relative structures. In this chapter, the geographically weighted regression model is introduced to analyse the relationship between the dependent variable and an explanatory variable reflecting a structure expressed in terms of a compositional table. The methodology is motivated by the problem of modelling local variations of the relationship between at-risk-of-poverty rates and the structure of the highest attained educational level in the German population aged 30-34. The real data study shows how information included in a compositional table and further expressed in real-valued coordinates can be highly valuable in selecting variables and prioritising them with respect to a research interest to facilitate the final interpretation of the model.

Kamila Fačevicová

Department of Mathematical Analysis and Applications of Mathematics, Palacký University Olomouc, 17. listopadu 12, 77146 Olomouc, Czech Republic.

e-mail: kamila.facevicova@gmail.com

Petra Kynčlová

Statistics Division, United Nation Industrial Development Organization (UNIDO), Wagramerstr. 5, P.O. Box 300, A-1400 Vienna, Austria.

e-mail: p.kynclova@unido.org

Karel Macků

Department of Geoinformatics, Palacký University Olomouc, 17. listopadu 50, 77146 Olomouc, Czech Republic.

e-mail: karel.macku@upol.cz

1 Introduction

Modelling of spatial dependent data represents one of the well-developed fields of statistical analysis. Methods of spatial statistics are based on the assumption that elements that are spatially close to one another are also more closely related. Thus, it is a direct link to Tobler's first law of geography saying that "Everything is related to everything else, but near things are more related than distant things." [Tobler, 1970]. Spatial data are often mistreated without taking into account their specific characteristics resulting from spatial relationships. These linkages may disrupt the assumptions of common statistical analyses, such as homogeneity, or absence of autocorrelation [Fotheringham, 1997], making standard statistical methods unsuitable [Anselin, 2013].

Many authors have been dealing with the design of spatial statistical methods. There are several well-developed methods and tools for solving spatial autocorrelation problems [Cliff and Ord, 1970], applying regression models on spatial data [Anselin, 2001, Brunsdon et al., 1998], estimating of local spatial and geographically weighted correlations [Kalogirou, 2012] or performing geographically weighted principal component analysis [Harris et al., 2011]. Nevertheless, all these methods are primarily designed for real-valued data and, therefore, if the data at hand are of the relative (compositional) nature, a proper modification of the methods is needed.

Analysis of spatial dependent compositional data is quite frequent in practice. Let us mention here the Kola data set consisting of a geochemical structure of over 600 soil samples from different locations in Norway, Finland and Russia [Reimann et al., 1997] as a well-known example of spatial dependent vector compositional data. Relative-valued data also naturally appear in the econometric studies, when the demographic structure of countries [Gollini et al., 2015] or the election results in different electoral districts are of interest. Finally, not just vector but also more complex compositional structures appear in practice. The compositional approach to analysis of manufacturing output distributed according to technology intensity and resource efficiency was introduced in [Fačevicová et al., 2016]. Furthermore, the educational structure reflecting the field of study and the highest reached degree was analysed in [de Sousa et al., 2020]. In both cases, the set of spatial dependent compositional tables was analysed. In the latter case, a more detailed structure can be considered when inhabitants' gender is taken into account resulting in a set of three factorial compositional data - compositional cubes.

Even though spatial dependency is quite common for compositional data, the methods for its analysis are still not well developed. One of the first steps towards this goal was taken by [Pawlowsky-Glahn et al., 2015, Tolosana-Delgado et al., 2008] who focused on simplicial kriging and cokriging. The authors of [Nordhausen et al., 2015] deal with the task of blind source separation. They propose the use of an independent component analysis method tailored for spatial dependent compositional data. Other authors have recently focused on methods for analysis of spatial dependence [Andreano et al., 2019, Pirzamanbein et al., 2018], spatial regression modelling [Leininger et al., 2013] or multivariate spatial processes [Yoshida and

Tsutsumi, 2018]. However, all the current methods are designed for vector compositional data and do not reflect the complexity of their more-factorial alternatives. This chapter represents the first step towards analysing more-factorial compositional data using spatial statistical methods. The main aim is to introduce a geographically weighted regression model for exploring spatial heterogeneity of the relationship between the dependent variable and an explanatory variable reflecting a structure expressed in terms of a compositional table.

The contribution itself is motivated by the problem of modelling a relationship between at-risk-of-poverty rates and educational attainment structure of the population aged 30-34 in 36 German regions in 2017. As the at-risk-of-poverty rates vary significantly across different German regions, the relationship with the structure of the highest attained education level is explored assuming local variations over the space. The compositional aspect of the educational structure of German population plays an important role in this study. The system of coordinates is thus constructed in such a way to focus primarily on the relative dominance of attained tertiary education as a key factor of at-risk-of-poverty rates.

The chapter is organised as follows. The theory of compositional tables and their coordinate representation is introduced in Sect. 2. Moreover, this section proposes an ordinary linear regression model with a compositional table in the role of the independent variable, and focuses on the interpretation of its coefficients. The conventional geographically weighted regression model is defined in Sect. 3 and further extended to its compositional version in Sect. 4. Sect. 5 demonstrates a practical application of the method followed by concluding remarks in Sect. 6.

2 Two-factorial Compositional Regression Model

Analysis of more-factorial compositional data represents a natural extension of the traditional log-ratio methodology, which is described in detail e.g. in Chapters 1, 2, 3 and 11 of [Pawlowsky-Glahn and Buccianti, 2011]. When the relative structure results from the combination of levels of two factors, it can be expressed in the form of an $I \times J$ compositional table

$$\mathbf{x} = \begin{pmatrix} x_{11} & \cdots & x_{1J} \\ \vdots & \ddots & \vdots \\ x_{I1} & \cdots & x_{IJ} \end{pmatrix}, \quad (1)$$

whose cells $x_{ij} > 0$, for $i = 1, 2, \dots, I$ and $j = 1, 2, \dots, J$ convey relative contributions on the whole. An example of such a structure is shown in Table 1 where each cell represents a percentage of the given educational-gender category on the whole population aged 30-34 in Leipzig in 2017.

Similarly to the case of vector compositional data, the sample space of compositional tables is a subset of an IJ -part simplex S^{IJ} where both basic operations of the

Table 1 The structure of attained education in German population aged 30-34 living in Leipzig in 2017 (in percent). The educational classes are Less than primary, primary and lower secondary education (ED0-2), Upper secondary and post-secondary non-tertiary education (ED3-4) and Tertiary education (ED5-8).

	ED0-2	ED3-4	ED5-8
Women	1.63	25.35	20.94
Men	4.53	27.71	19.84

Aitchison geometry, perturbation of tables $\mathbf{x}$ and $\mathbf{x}'$ and powering by a real constant α , can be extended as follows

$$\mathbf{x} \oplus \mathbf{x}' = C \left(x_{ij} x'_{ij} \right)_{i,j=1}^{I,J} \quad \text{and} \quad \alpha \odot \mathbf{x} = C \left(x_{ij}^\alpha \right)_{i,j=1}^{I,J}, \quad (2)$$

where C denotes closure and constitutes the equivalence classes of compositional tables, which differ only by the sum of their parts κ . The vector space structure is moreover extended with the Aitchison inner product defined for compositional tables as

$$\langle \mathbf{x}, \mathbf{x}' \rangle_A = \frac{1}{2IJ} \sum_{i,j} \ln \frac{x_{ij}}{x_{kl}} \ln \frac{x'_{ij}}{x'_{kl}}. \quad (3)$$

Similarly to compositional vectors, compositional tables are also defined as having strictly positive entries. The situation when some cells equal zero would disable the application of the log-ratio approach to further analysis and therefore its imputation is needed. Various techniques for dealing with zeros or missing values have been introduced, e.g. in Filzmoser et al. [2018], ch. 13. Nevertheless, all of them are primarily designed for vector compositional data and not their more-factorial alternatives.

According to Egozcue et al. [2008, 2015], it is possible to decompose each compositional table into its independent and interactive parts

$$\mathbf{x} = \mathbf{x}_{ind} \oplus \mathbf{x}_{int} \quad (4)$$

and to analyse them separately. More specifically, both $\mathbf{x}_{ind}$ and $\mathbf{x}_{int}$ are also $I \times J$ compositional tables with cells

$$x_{ij}^{ind} = \left(\prod_{k=1}^I \prod_{l=1}^J x_{kl} \right)^{\frac{1}{IJ}} \quad \text{and} \quad x_{ij}^{int} = \left(\prod_{k=1}^I \prod_{l=1}^J \frac{x_{ij}}{x_{kl}} \right)^{\frac{1}{IJ}}, \quad (5)$$

respectively.

Before further analysing the compositional structures, it is convenient to express them in a system of $D - 1$ real-valued coordinates $\mathbf{z} = (z_1, \dots, z_{D-1})'$. More specifically, a vector of isometric log-ratio (ilr) coordinates represents a mapping from

$\mathcal{S}^D$ to $\mathbb{R}^{D-1}$. This real space representation of compositional data is defined as log-contrasts

$$z_i = \sum_{j=1}^D \xi_{ij} \ln x_j, \quad \sum_{j=1}^D \xi_{ij} = 0, \quad i = 1, \dots, D-1, \quad (6)$$

where $\mathbf{e}_i = \exp(\boldsymbol{\xi}_i) = [\exp(\xi_{i1}), \dots, \exp(\xi_{iD})]'$, $i = 1, \dots, D-1$ is a system of orthonormal vectors which forms an orthonormal basis of $\mathcal{S}^D$.

According to Hron et al. [2012], the linear regression model with a D -part compositional vector in the role of the explanatory variable can then be defined as

$$E(Y|\mathbf{z}) = \beta_0 + \beta_1 z_1 + \dots + \beta_{D-1} z_{D-1}. \quad (7)$$

Its alternative form will be mainly used in this chapter as follows

$$E[\ln(Y)|\mathbf{z}] = \beta_0 + \beta_1 z_1 + \dots + \beta_{D-1} z_{D-1}. \quad (8)$$

Different ilr coordinate systems are orthogonal rotations of each other and therefore the general characteristics of the regression models (AIC, R^2 , residuals, β_0 , etc.) are invariant to the specific choice of the coordinate system. Despite this, the parameters $\beta_1, \dots, \beta_{D-1}$ vary according to the structure of coordinates, which is usually set with regards to a convenient interpretation. An ilr system, formed by three groups of coordinates, is defined in Fačevićová et al. [2018] to enable the real space representation of compositional tables. The construction is based on sequential binary partition (SBP) [Egozcue and Pawłowsky-Glahn, 2005] applied separately to levels of each of the constituting factors. The first two groups of coordinates $\mathbf{z}^r$ and $\mathbf{z}^c$ of a compositional table $\mathbf{x}$ are thus simple balances between the whole rows and columns,

$$z_i^r = \sqrt{\frac{stJ}{s+t}} \ln \frac{[g(\mathbf{x}_{j_1 \bullet}) \cdots g(\mathbf{x}_{j_s \bullet})]^{1/s}}{[g(\mathbf{x}_{k_1 \bullet}) \cdots g(\mathbf{x}_{k_t \bullet})]^{1/t}}, \quad \text{for } i = 1, 2, \dots, I-1 \quad (9)$$

and

$$z_j^c = \sqrt{\frac{uvI}{u+v}} \ln \frac{[g(\mathbf{x}_{\bullet l_1}) \cdots g(\mathbf{x}_{\bullet l_u})]^{1/u}}{[g(\mathbf{x}_{\bullet m_1}) \cdots g(\mathbf{x}_{\bullet m_v})]^{1/v}}, \quad \text{for } j = 1, 2, \dots, J-1, \quad (10)$$

where s, t (u, v) are numbers of rows (columns) separated in the i -th (j -th) step of SBP, indices $(j_1 \bullet, \dots, j_s \bullet)$ and $(k_1 \bullet, \dots, k_t \bullet)$ or $(\bullet l_1, \dots, \bullet l_u)$ and $(\bullet m_1, \dots, \bullet m_v)$ denote rows/columns involved in this step and $g(\cdot)$ stands for the geometric mean. These coordinates describe relations between levels of one factor when the influence of the second factor is suppressed. The coordinate system is complemented with a third group of $(I-1)(J-1)$ log-contrasts, which describe relations between factors. The interpretation of these coordinates is closely connected to the odds ratio between four groups of parts, which are uniquely determined by the initial row and column

SBPs and represented by the geometric mean of the respective parts. When a, b, c, d are the numbers of parts within each group and i, j, k, l are indices of parts within each group, these coordinates can be in general defined as

$$z^{OR} = \sqrt{\frac{a \cdot d}{a + b + c + d}} \ln \frac{(x_{i_1} \cdots x_{i_a})^{1/a} (x_{l_1} \cdots x_{l_d})^{1/d}}{(x_{j_1} \cdots x_{j_b})^{1/b} (x_{k_1} \cdots x_{k_c})^{1/c}}. \quad (11)$$

The construction of coordinates $\mathbf{z}^{OR}$ is described in detail in Fačevicová et al. [2018] and the whole coordinate system can be computed with the function `tabCoord` from the R package `robCompositions` [Templ et al., 2011].

As an example, let us consider the 2x3 compositional table from Table 1. In order to represent the structure in orthonormal coordinates, the row and column SBPs need to be defined first. For the purpose of deeper analysis in Sect. 5, it seems to be convenient to focus on the relative dominance of men over women (which is quantified by the coordinate z_1^r). It furthermore seems convenient to concentrate on the overall population aged 30-34 with attained tertiary education (ED5-8) over the geometric mean of the remaining groups (ED0-2 and ED3-4) - z_1^c , complemented by the relative dominance of the upper secondary qualification (ED3-4) over those with the lowest achieved degree of education - z_2^c . The whole relative structure can therefore be represented with the system of coordinates that are listed in Table 2 and whose simple graphical description is presented in Fig. 1.

Table 2 List of orthonormal coordinates and their values used for the practical example from Table 1.

z_1^r	z_1^c	z_2^c	z_{11}^{OR}	z_{12}^{OR}
$\sqrt{\frac{3}{2}} \ln \frac{(x_{21}x_{22}x_{23})^{1/3}}{(x_{11}x_{12}x_{13})^{1/3}}$	$\sqrt{\frac{4}{3}} \ln \frac{(x_{13}x_{23})^{1/2}}{(x_{11}x_{12}x_{21}x_{22})^{1/4}}$	$\ln \frac{(x_{12}x_{22})^{1/2}}{(x_{11}x_{21})^{1/2}}$	$\sqrt{\frac{1}{3}} \ln \frac{(x_{11}x_{12})^{1/2}x_{23}}{(x_{21}x_{22})^{1/2}x_{13}}$	$\frac{1}{2} \ln \frac{x_{11}x_{22}}{x_{12}x_{21}}$
0.4317	1.0120	2.2777	-0.3520	-0.4769

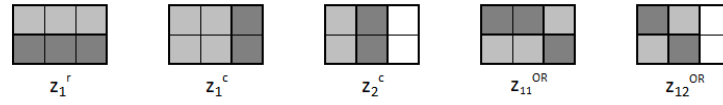


Fig. 1 Graphical representation of orthonormal coordinates defined in Table 2.

According to the values of the coordinate z_1^r , men in Leipzig dominated women across all attained educational classes in 2017. Considering the column balances z_1^c and z_2^c , we can conclude that there is (on average) a higher proportion of people with a university degree than those with primary or secondary educational qualification.

Moreover, the coordinate z_2^c indicates that the population with attained secondary education prevails over those who attained only the lowest level of education.

An even deeper insight into the structure is provided by the odds ratio coordinates z_{11}^{OR} and z_{12}^{OR} . The negative value of z_{11}^{OR} highlights the fact that the proportion of women with a tertiary education (in comparison to the geometric mean of those with the primary or secondary one) is slightly higher than the respective proportion of men with tertiary education. Furthermore, the relative predominance of the secondary education over the primary one is slightly higher for women due to the negative value of z_{11}^{OR} . This difference is presumably caused by a small percentage of women with only a primary qualification.

The linear regression model with a compositional table as the explanatory variable can be written as

$$E(Y|\mathbf{z}) = \beta_0 + \beta_1^r z_1^r + \cdots + \beta_{I-1}^r z_{I-1}^r + \beta_1^c z_1^c + \cdots + \beta_{J-1}^c z_{J-1}^c + \beta_{11}^{OR} z_{11}^{OR} + \cdots + \beta_{I-1,J-1}^{OR} z_{I-1,J-1}^{OR} \quad (12)$$

or by incorporating a log-transformed response with respect to the application introduced in this chapter

$$E[\ln(Y)|\mathbf{z}] = \beta_0 + \beta_1^r z_1^r + \cdots + \beta_{I-1}^r z_{I-1}^r + \beta_1^c z_1^c + \cdots + \beta_{J-1}^c z_{J-1}^c + \beta_{11}^{OR} z_{11}^{OR} + \cdots + \beta_{I-1,J-1}^{OR} z_{I-1,J-1}^{OR}. \quad (13)$$

Similar to the case of classical regression, the intercept β_0 equals the predicted value of Y (or $\ln(Y)$) when all coordinates are zero. Zero values of all odds ratio coordinates correspond to the situation when the row and column factors are completely independent. Moreover, the zero values of all row and column balances reflect a situation when all parts of the compositional table are equal. The remaining parameters quantify the effect of a unit change of the respective coordinate (while fixing the remaining ones) on the response variable. A deeper insight into the model structure is needed to enable interpretation of this effect in terms of the original compositional parts.

Consider a linear regression model (13) with regressors in the form of ilr coordinates of a 2x3 table, defined in Table 2. The model can be written as

$$E[\ln(Y)|\mathbf{z}] = \beta_0 + \beta_1^r z_1^r + \beta_1^c z_1^c + \beta_2^c z_2^c + \beta_{11}^{OR} z_{11}^{OR} + \beta_{12}^{OR} z_{12}^{OR}. \quad (14)$$

The value of the parameter β_1^r quantifies the effect of a k -times growth of the relative dominance of the second level of the row factor. More specifically, this change, with simultaneously fixed values of the remaining coordinates, results in an $\exp\left(\sqrt{3/2}\beta_1^r \ln k\right)$ -times growth of the response variable. A similar reasoning can also be done for the other coefficients. The effects of ratio changes forming the remaining regressors are summarised in Table 3. Note here, that only finite changes of the coordinates are considered within this chapter and its infinitesimal alternatives are for the case of a vector compositional covariate discussed in Morais et al. [2018].

Table 3 Interpretation of regression parameters of model (14). Column z denotes the coordinate corresponding to the given parameter β and z' its changed value. The original prediction y and its new value y' , resulting from the model with z' and other coordinates fixed, are compared in $\ln(y'/y)$. Column $\mathbf{x}'$ introduces possible changes in the compositional table $\mathbf{x}$, which lead to a change in the coordinate z and simultaneously keep the remaining coordinates unchanged.

β	z	z'	$\ln(y'/y)$	$\mathbf{x}'$		
β_1^r	$\sqrt{\frac{3}{2}} \ln \frac{(x_{21}x_{22}x_{23})^{1/3}}{(x_{11}x_{12}x_{13})^{1/3}}$	$\sqrt{\frac{3}{2}} \ln k \frac{(x_{21}x_{22}x_{23})^{1/3}}{(x_{11}x_{12}x_{13})^{1/3}}$	$\sqrt{\frac{3}{2}} \beta_1^r \ln k$	$\frac{x_{11}}{kx_{21}}$	$\frac{x_{12}}{kx_{22}}$	$\frac{x_{13}}{kx_{23}}$
β_1^c	$\sqrt{\frac{4}{3}} \ln \frac{(x_{13}x_{23})^{1/2}}{(x_{11}x_{12}x_{21}x_{22})^{1/4}}$	$\sqrt{\frac{4}{3}} \ln k \frac{(x_{13}x_{23})^{1/2}}{(x_{11}x_{12}x_{21}x_{22})^{1/4}}$	$\sqrt{\frac{4}{3}} \beta_1^c \ln k$	$\frac{x_{11}}{x_{21}}$	$\frac{x_{12}}{x_{22}}$	$\frac{kx_{13}}{kx_{23}}$
β_2^c	$\ln \frac{(x_{12}x_{22})^{1/2}}{(x_{11}x_{21})^{1/2}}$	$\ln k \frac{(x_{12}x_{22})^{1/2}}{(x_{11}x_{21})^{1/2}}$	$\beta_2^c \ln k$	$\frac{1}{\sqrt{k}}x_{11}$ $\frac{1}{\sqrt{k}}x_{21}$	$\sqrt{k}x_{12}$ $\sqrt{k}x_{22}$	x_{13} x_{23}
β_{11}^{OR}	$\sqrt{\frac{1}{3}} \ln \frac{(x_{11}x_{12})^{1/2}x_{23}}{(x_{21}x_{22})^{1/2}x_{13}}$	$\sqrt{\frac{1}{3}} \ln k \frac{(x_{11}x_{12})^{1/2}x_{23}}{(x_{21}x_{22})^{1/2}x_{13}}$	$\sqrt{\frac{1}{3}} \beta_{11}^{OR} \ln k$	$\sqrt[6]{k}x_{11}$ $\frac{1}{\sqrt[6]{k}}x_{21}$	$\sqrt[6]{k}x_{12}$ $\frac{1}{\sqrt[6]{k}}x_{22}$	$\frac{1}{\sqrt[6]{k}}x_{13}$ $\sqrt[6]{k}x_{23}$
β_{12}^{OR}	$\frac{1}{2} \ln \frac{x_{11}x_{22}}{x_{12}x_{21}}$	$\frac{1}{2} \ln k \frac{x_{11}x_{22}}{x_{12}x_{21}}$	$\frac{1}{2} \beta_{12}^{OR} \ln k$	$\sqrt[4]{k}x_{11}$ $\frac{1}{\sqrt[4]{k}}x_{21}$	$\frac{1}{\sqrt[4]{k}}x_{12}$ $\sqrt[4]{k}x_{22}$	x_{13} x_{23}

The condition of fixed regressors implies that the change of the coordinate of interest should be caused by a simultaneous increase of one constituting group and a decrease of the other. For instance, an increase of the coordinate z_2^c can reflect the situation when the proportion of ED3-4 men and women grows together with a decrease of proportions belonging to the group ED0-2 and an unchanged proportion of men and women with tertiary education. Another complex case can be derived when an increase of the coordinate of interest is not caused by simultaneous changes of all involved compositional parts (as described in Table 3) but e.g. by an increase of only those parts, which form the nominator of the respective formula. Such changes of the relative structure would also affect (some of) the other coordinates. The resulting effect on the response will therefore be strengthened or weakened with these supplementary effects. Some of these scenarios are with respect to the model (14) described in Table 4.

3 Geographically Weighted Regression Model

There are situations where the analysed phenomena vary over space. That is when each observation corresponds to some geographic location. An approach to deal with the problem of spatial non-stationarity was given in [Brunsdon et al., 1998] where authors define the geographically weighted regression (GWR) model. The main idea of this approach is that the regression parameters β can vary over space. Therefore a

Table 4 List of possible changes in the relative structure of a compositional table $\mathbf{x}$ (column $\mathbf{x}'$), their impact on ilr coordinates (z'), when the not-listed coordinates remain unchanged, and the final effect on the value of the response ($\ln(y'/y)$).

$\mathbf{x}'$	z'	$\ln(y'/y)$
$\begin{array}{ccc} x_{11} & kx_{12} & x_{13} \\ x_{21} & kx_{22} & x_{23} \end{array}$	$z_2^{c'} = \ln k \frac{(x_{12}x_{22})^{1/2}}{(x_{11}x_{21})^{1/2}}$ $z_1^{c'} = \sqrt{\frac{4}{3}} \ln \frac{1}{\sqrt{k}} \frac{(x_{13}x_{23})^{1/2}}{(x_{11}x_{12}x_{21}x_{22})^{1/4}}$	$\beta_2^c \ln k - \beta_1^c \sqrt{\frac{4}{3}} \ln \sqrt{k}$
$\begin{array}{ccc} \frac{1}{k}x_{11} & x_{12} & x_{13} \\ \frac{1}{k}x_{21} & x_{22} & x_{23} \end{array}$	$z_2^{c'} = \ln k \frac{(x_{12}x_{22})^{1/2}}{(x_{11}x_{21})^{1/2}}$ $z_1^{c'} = \sqrt{\frac{4}{3}} \ln \sqrt{k} \frac{(x_{13}x_{23})^{1/2}}{(x_{11}x_{12}x_{21}x_{22})^{1/4}}$	$\beta_2^c \ln k + \beta_1^c \sqrt{\frac{4}{3}} \ln \sqrt{k}$
$\begin{array}{ccc} x_{11} & x_{12} & x_{13} \\ \frac{1}{\sqrt{k}}x_{21} & \frac{1}{\sqrt{k}}x_{22} & \sqrt{k}x_{23} \end{array}$	$z_{11}^{OR'} = \sqrt{\frac{1}{3}} \ln k \frac{(x_{11}x_{12})^{1/2}x_{23}}{(x_{21}x_{22})^{1/2}x_{13}}$ $z_1^{r'} = \sqrt{\frac{3}{2}} \ln \frac{1}{\sqrt{k}} \frac{(x_{21}x_{22}x_{23})^{1/3}}{(x_{11}x_{12}x_{13})^{1/3}}$ $z_1^{c'} = \sqrt{\frac{4}{3}} \ln \sqrt{k} \frac{(x_{13}x_{23})^{1/2}}{(x_{11}x_{12}x_{21}x_{22})^{1/4}}$	$\beta_{11}^{OR} \sqrt{\frac{1}{3}} \ln k - \beta_1^r \sqrt{\frac{3}{2}} \ln \sqrt[3]{k} + \beta_1^c \sqrt{\frac{4}{3}} \ln \sqrt{k}$
$\begin{array}{ccc} \sqrt{k}x_{11} & \sqrt{k}x_{12} & \frac{1}{\sqrt{k}}x_{13} \\ x_{21} & x_{22} & x_{23} \end{array}$	$z_{11}^{OR'} = \sqrt{\frac{1}{3}} \ln k \frac{(x_{11}x_{12})^{1/2}x_{23}}{(x_{21}x_{22})^{1/2}x_{13}}$ $z_1^{r'} = \sqrt{\frac{3}{2}} \ln \frac{1}{\sqrt{k}} \frac{(x_{21}x_{22}x_{23})^{1/3}}{(x_{11}x_{12}x_{13})^{1/3}}$ $z_1^{c'} = \sqrt{\frac{4}{3}} \ln \frac{1}{\sqrt{k}} \frac{(x_{13}x_{23})^{1/2}}{(x_{11}x_{12}x_{21}x_{22})^{1/4}}$	$\beta_{11}^{OR} \sqrt{\frac{1}{3}} \ln k - \beta_1^r \sqrt{\frac{3}{2}} \ln \sqrt[3]{k} - \beta_1^c \sqrt{\frac{4}{3}} \ln \sqrt{k}$
$\begin{array}{ccc} \sqrt{k}x_{11} & \frac{1}{\sqrt{k}}x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \end{array}$	$z_{12}^{OR'} = \frac{1}{2} \ln k \frac{x_{11}x_{22}}{x_{12}x_{21}}$ $z_2^{c'} = \ln \frac{1}{\sqrt{k}} \frac{(x_{12}x_{22})^{1/2}}{(x_{11}x_{21})^{1/2}}$	$\beta_{12}^{OR} \frac{1}{2} \ln k - \beta_2^c \ln \sqrt{k}$
$\begin{array}{ccc} x_{11} & x_{12} & x_{13} \\ \frac{1}{\sqrt{k}}x_{21} & \sqrt{k}x_{22} & x_{23} \end{array}$	$z_{12}^{OR'} = \frac{1}{2} \ln k \frac{x_{11}x_{22}}{x_{12}x_{21}}$ $z_2^{c'} = \ln \sqrt{k} \frac{(x_{12}x_{22})^{1/2}}{(x_{11}x_{21})^{1/2}}$	$\beta_{12}^{OR} \frac{1}{2} \ln k + \beta_2^c \ln \sqrt{k}$

system of n *local models* is estimated. Each model corresponds to one geographical location i and is based on the observations which are geographically *close enough* to the given one. The local model can be formally written as

$$E(Y_i | \mathbf{v}_i) = \beta_{i0} + \sum_{k=1}^m \beta_{ik} v_{ik}, \quad (15)$$

where Y_i and $\mathbf{v}_i = (v_{i1}, \dots, v_{im})'$ are values of the dependent and independent variables measured at the location i and $\boldsymbol{\beta}_i = (\beta_{i0}, \dots, \beta_{im})'$ are the respective regression coefficients. Within the GWR approach, the observation which is closer to the location i has a greater influence on the model estimation at the location i . The regression parameters are therefore estimated using a weighted least squares formula

$$\hat{\boldsymbol{\beta}}_i = (V'W_iV)^{-1} V'W_i\mathbf{y}, \quad (16)$$

where V is an n times $m + 1$ data matrix, $\mathbf{y} = (Y_1, \dots, Y_n)'$ is vector of values of the response variable and W_i is the diagonal matrix

$$W_i = \begin{pmatrix} w_1^i & 0 & \dots & 0 \\ 0 & w_2^i & \dots & 0 \\ \dots & & \ddots & 0 \\ 0 & 0 & \dots & w_n^i \end{pmatrix}. \quad (17)$$

The diagonal elements w_j^i correspond to the weights of each data point j and are assigned according to its geographical (Euclidean) distance from the actual location i . The geographical weighting is based on the value of a kernel function. The most frequently used ones are summarised in Table 5 according to [Gollini et al., 2015]. When the *global model* incorporates all observations with the same weight, it corresponds to the classical ordinary least squares (OLS) approach. If the Gaussian or Exponential kernel is used, observations which are far from the i -th location contribute to the model estimation with a smaller weight. The bi-square, tri-cube or box-car kernels use only those observations which are within some bandwidth b around the location i . The bandwidth b is one of the key parameters of the local model estimation. It can be defined as the distance from the center point i and is then fixed for all local models. Alternatively, if the adaptive version of the algorithm is used, b can vary according to a given number of local data. The optimal bandwidth choice is usually found based on a cross-validation method or a model performance criterion like Akaike Information Criterion (AIC) or its corrected version (AICc) (see [Brunsdon et al., 1998, Gollini et al., 2015] for details). The bandwidth can also be determined by an expert choice based on the knowledge of real spatial relationships given by behaviour of the observed phenomenon (e.g. the influence of commuting between municipalities).

Local models are characterised with local estimates of the regression coefficients $\boldsymbol{\beta}_i$ supplemented by the respective t-statistics and corresponding local R_i^2 values for each model. These characteristics can be compared with those of the global

model and their proper geographical visualisation can facilitate the exploration of spatial non-stationarity of the modelled phenomenon [Matthews and Yang, 2012]. Moreover, the overall performance of local models can be expressed by the local R^2 values (or alternatively by AIC_{local}) which are based on the residuals in the centre location of each partial model.

Table 5 List of kernel functions frequently used for computation of weights in GWR Gollini et al. [2015]. w_j^i denotes the weight of the j -th observation in the i -th regression model, d_{ij} is the geographical distance between the i -th and j -th location and b is the bandwidth.

Global model	$w_j^i = 1$
Gaussian	$w_j^i = \exp\left(-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right)$
Exponential	$w_j^i = \exp\left(-\frac{ d_{ij} }{b}\right)$
Bi-square	$w_j^i = \begin{cases} \left(1 - \left(\frac{d_{ij}}{b}\right)^2\right)^2 & \text{if } d_{ij} < b, \\ 0 & \text{otherwise} \end{cases}$
Tri-cube	$w_j^i = \begin{cases} \left(1 - \left(\frac{ d_{ij} }{b}\right)^3\right)^3 & \text{if } d_{ij} < b, \\ 0 & \text{otherwise} \end{cases}$
Box-car	$w_j^i = \begin{cases} 1 & \text{if } d_{ij} < b, \\ 0 & \text{otherwise} \end{cases}$

4 Geographically Weighted Regression for Compositional Tables

Similarly to other conventional statistical methods, the GWR approach to modelling of spatial non-stationary relations is primarily designed for real-valued regressors. When a variable of interest is modelled according to a relative structure of some regressor variable, the use of an ilr coordinate representation of the structure is needed. Therefore, the global model for one- or two-dimensional compositional regressors can be described by Equations (8) and (13), respectively. Assuming the case of spatial dependent two-factorial compositional data, the global model can be further extended with a set of local models as follows

$$\begin{aligned}
 E(Y_i|\mathbf{z}_i) = & \beta_{i,0} + \beta_{i,1}^r z_{i,1}^r + \cdots + \beta_{i,I-1}^r z_{i,I-1}^r + \beta_{i,1}^c z_{i,1}^c + \cdots + \beta_{i,J-1}^c z_{i,J-1}^c \\
 & + \beta_{i,11}^{OR} z_{i,11}^{OR} + \cdots + \beta_{i,I-1,J-1}^{OR} z_{i,I-1,J-1}^{OR}
 \end{aligned} \tag{18}$$

or alternatively, for the logarithmic response,

$$\begin{aligned}
E[\ln(Y_i)|\mathbf{z}_i] = & \beta_{i,0} + \beta_{i,1}^r z_{i,1}^r + \cdots + \beta_{i,I-1}^r z_{i,I-1}^r + \beta_{i,1}^c z_{i,1}^c + \cdots + \beta_{i,J-1}^c z_{i,J-1}^c \\
& + \beta_{i,11}^{OR} z_{i,11}^{OR} + \cdots + \beta_{i,I-1,J-1}^{OR} z_{i,I-1,J-1}^{OR}.
\end{aligned} \tag{19}$$

The model coefficients are estimated using the weighted least squares formula (16), where the data matrix V is now an n times IJ matrix of coordinates defined by equations (9), (10) and (11). Similarly to the case of the global model, the performance of its local alternatives does not depend on the specific choice of the coordinate system. The interpretation of the regression coefficients should be carried out with respect to the complexity of the modelled structure and taking into account findings from Sect. 2.

5 Application

The practical part of this chapter focuses on a country example of Germany. The aim of the study is to explore and analyse the relationship between the risk of falling into poverty and the educational attainment structure of the 30-34 aged population and its variation over space.

Poverty and inequality are not only an issue of developing countries. The vulnerability towards falling into poverty is driven by many factors determined by national, sector-specific, community, household or individual characteristics. The level and depth of poverty also increases as a result of any crisis with strong economic impacts such as increase of unemployment inducing a reduction of available income of households [Duiella and Turrini, 2014].

From the global perspective, the number of people living in extreme poverty in Germany is negligible compared to developing countries. However, poverty in Europe is not typically measured in absolute terms. People are considered poor if they have lower income and fewer opportunities than other individuals living in the same society. In fact, the percentage of people living at risk of poverty and social exclusion in Germany in 2017 was 19.0 percent; almost one fifth of the population.

Many research findings suggest that factors that often condition the number of people living at risk of poverty or social exclusion in Germany are: young age, single, lone parent, or regional aspects such as unemployment, low education and the region of East Germany [Brülle, 2018]. People with a lower level of educational background and a limited access to the labour market are highly vulnerable to easily fall into this group. Moreover, high rental costs, especially in urban areas, pose an additional burden to many households [Pieper et al., 2019].

5.1 Data and Methods

The data were processed in the environment of the R Software [R Core Team, 2020] with packages `robCompositions` for computation of coordinates and `GWmodel` and

spdep for estimation of the global and local models [Gollini et al., 2015, Bivand et al., 2013]. The maps were produced using the ArcGIS® software by Esri [ESRI, 2011].

Data used for the analysis are sourced from the regional database of EUROSTAT accessed and processed through the R package eurostat [Lahti et al., 2017]. The data were collected by the Federal Statistical Office Germany through Labour Force Survey (EU-LFS) and Statistics on Income and Living Conditions (EU-SILC). Both types of information are available by NUTS2 (Nomenclature of territorial units for statistics) regions.

The indicator chosen to represent the poverty dimension is the share of people at risk of poverty or social exclusion, i.e. the at-risk-of-poverty rate (RoP). At risk-of-poverty are people with an equivalised disposable income below the risk-of-poverty threshold, which is set at 60 percent of the national median equivalised disposable income (after social transfers).

The compositional tables show the structure of the population aged 30-34 by sex and the highest level of education successfully completed. The level of education attained is classified by International Standard Classification of Education (ISCED). In this chapter, the classes chosen for further analysis are:

- Less than primary, primary and lower secondary education (ED0-2),
- Upper secondary and post-secondary non-tertiary education (ED3-4),
- Tertiary education (ED5-8).

Educational attainment data series are available from 2000 to 2017. The information on the risk of poverty or social exclusion by NUTS2 regions is available only for 2016 and 2017. For the assessment of the regression models, the relationship between the dependent and independent variables is modelled for 36 German regions in the latest available year 2017.

5.2 Poverty Analysis

At-risk-of-poverty rates vary substantially across German regions. Values reported show the minimum of 13.3 percent in Lüneburg and the maximum of 28.5 percent in Bremen in 2017. Fig. 2 demonstrates the distribution of people at risk of poverty or social exclusion in Germany by NUTS2 regions in 2017.

The highest at-risk-of-poverty rates are experienced by urban areas such as Bremen, Berlin and Hamburg, together with regions of the former East Germany, in particular Mecklenburg-Vorpommern, Chemnitz, Thuringia and Saxony-Anhalt. In contrast, Lüneburg followed by all regions of Baden-Württemberg have the lowest shares of people exposed to the risk of poverty or social exclusion.

Education can serve as an elevator for upward economic mobility. Particularly higher levels of education increase chances of participation in the labour market. Level of attained education influences the income distribution and thus has a substantial effect on at-risk-of-poverty rates. As education plays an important role, ap-

plication of the methodology is illustrated on modelling people at risk of poverty or social exclusion by using the information on the structure of educational attainment level by sex within the population aged 30-34.

5.2.1 Comparison of Educational Attainment Structure in Germany Between 2000 and 2017

Germany has one of the highest levels of upper secondary education attainment and tertiary education has expanded in the past decades. In 2019, 86.7 percent of the country's 25-64 year-olds have obtained at least an upper secondary qualification. The expansion of tertiary education is reflected by the number of people with tertiary qualifications which increased from 24.9 percent in 2004 to 29.9 percent in 2019. The differences between generations are noticeable: the share of people with tertiary education in the 30-34 year-old population is 35.5 percent compared to 26.9 percent of 55-64 year-olds. The tertiary education rates are fairly higher for men (32.4 percent in 2019) than women (27.3 percent).

Building on existing research studies [Hanesch, 2019], the focus of this article is placed on the educational attainment structure of population aged 30-34 to explore its relation to at-risk-of-poverty rates. Such constructed compositional tables which focus on the productive age of the population (30-34 year-old) reveal patterns of the attained level of education with respect to the gender characteristics. An example of such a compositional table is illustrated in Table 1 reflecting the educational attainment structure in Leipzig.

The coordinate representation of compositional tables, which is defined in Table 2 and whose schematic structure is given in Fig. 1, also enables us to compare the

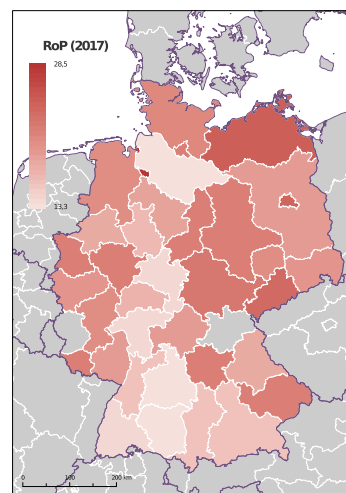


Fig. 2 People at risk of poverty or social exclusion in Germany by NUTS2 regions in 2017 (in percentages).

structure over time. The coordinates calculated for years 2000 and 2017 reveal a clear structural change in the German educational system.

Despite the lack of data for a several regions, particularly interesting results are captured by coordinates z_1^c and z_{11}^{OR} , which quantify the relative prevalence of university educated citizens and further compare their proportion within men and women.

The positive values of z_1^c indicate that the component of population with tertiary education on average dominates the population with secondary or primary education across all women and men aged 30-34 in Germany. Fig. 3 suggests that the importance of tertiary education has grown significantly in Germany in the past decades. While most of German regions showed negative values in 2000 representing a relative dominance of secondary and primary education, this trend reversed in 2017. The majority of former Western German regions show dominance of tertiary education in 2017 with the exception of Lower Saxony and Saarland. Both regions have undergone a structural change moving from mining-oriented economies towards high technology industrial production. Negative values can also be observed in regions of the former East Germany indicating relative predominance of vocational qualifications across the population aged 30-34 in this area in 2017. Moreover, the relative dominance of the upper secondary attainment over the primary one in these regions is supported by positive values of the coordinate z_2^c .

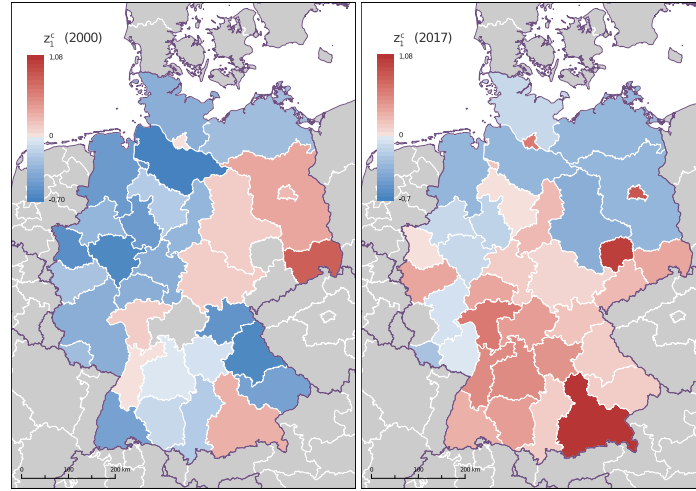


Fig. 3 Comparison of z_1^c in 2000 (left) and 2017 (right)

The coordinate z_{11}^{OR} examines the proportion of men with tertiary education with the proportion of women with tertiary education. The comparison of z_{11}^{OR} in 2000 and 2017 shows a much wider range of the coordinate values in 2000. This indicates greater disparities between men and women in the distribution of attained education in the population aged 30-34 in Germany in 2000. Positive values correspond to the

situation when the proportion of men with tertiary education is higher in the 30-34 aged population than the proportion of women with tertiary education, which is a pattern observed in all regions of former West Germany. In past decades, disparity has been significantly reduced as noticed in Fig. 4. No substantial change has been noticed for the territory of former East Germany where negative values of z_{11}^{OR} still prevail in most of the regions. In contrast, the dominance of tertiary education over the other education levels among 30-34 aged men has become similar to the situation of women in the west of Germany. The most balanced regions are for instance southern Bavaria, Hessen or the eastern part of Lower Saxony, which show z_{11}^{OR} values around 0.

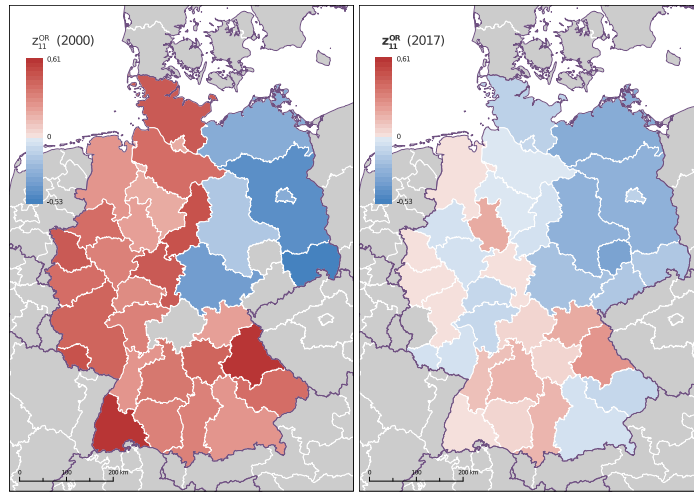


Fig. 4 Comparison of z_{11}^{OR} in 2000 (left) and 2017 (right)

5.2.2 The GWR Model

The introduced methodology is applied to explore whether the relationship of people at risk of poverty or social exclusion and the educational attainment structure by sex varies across German regions in the population of 30-34 years-old. As a first step, the linear regression model (14) with a compositional table as the explanatory variable was fitted to the examined data set. The global model diagnostics is shown in Table 6. According to the estimate of parameter β_0 of the global model and the interpretation of parameters defined in Sect. 2, the basic level of the RoP when all groups are equally frequent is approximately 20 percent ($\exp(3.027)$).

The level of RoP is significantly affected by the OR coordinates z_{11}^{OR} and z_{12}^{OR} and therefore the comparison of the educational structure within the male and female population seems to be important. According to the parameter $\beta_{11}^{OR} = -0.91$ and

Table 6 Results from the global OLS model. The R^2 and R^2_{adj} value are 0.370 and 0.265, residual sum of squares is 0.748 on 30 degrees of freedom and value of the F statistic is 3.527 on 5 and 30 degrees of freedom, AIC equals -23.292 and $AICc -19.292$

	β_0	β_1^r	β_1^c	β_2^c	β_{11}^{OR}	β_{12}^{OR}
Estimates	3.027	1.337	-0.040	-0.119	-0.913	1.250
Significance	***	*			***	**
St. deviations	0.139	0.692	0.070	0.115	0.249	0.457
Coordinates		z_1^r	z_1^c	z_2^c	z_{11}^{OR}	z_{12}^{OR}

findings listed in Table 3, a doubling of the proportion of men with tertiary education (related to the average proportion of those with the secondary or lower degree) in comparison to the same proportion for women is related to the decrease of the RoP by approximately one third ($\exp(-0.91\sqrt{1/3}\ln 2) = 0.69$). In order to keep the remaining coordinates fixed, this change can be caused by a simultaneous increase in groups of Male ED5-8 and Female ED0-2 and ED3-4 together with a proportional decrease in the remaining ones (see Table 3).

The second significant factor, which affects the overall value of at-risk-of-poverty rate, is the proportion of men with secondary education (within the group of all men who attained at least the upper secondary education) in comparison to the same proportion for women. According to the value of the parameter $\beta_{12}^{OR} = 1.25$ and Table 3, a doubling of the respective log-odds ratio is related to an increase of at-risk-of-poverty rate by 50 percent ($\exp(0.5 \cdot 1.25 \ln 2) = 1.54$). This growth can be caused by a simultaneous increase of ED0-2 women and ED3-4 men and a proportional decrease by ED3-4 women and ED0-2 men. In a less complex situation, when the difference is caused by a change in just one category (an increase of the proportion of women with the lowest level of attained education or conversely in an increase of men with at least the secondary qualification), the change affects also the value of the coordinate z_2^c . According to Table 4, the overall change of at-risk-of-poverty rate is given as $0.5\beta_{12}^{OR} \ln k - \beta_2^c \ln \sqrt{k}$. Considering negative values of the parameter β_2^c , an increase of the proportion of ED0-2 women seems to be more critical as it would result in a growth of the at-risk-of-poverty rate by 60 percent ($\exp(0.5 \cdot 1.25 \ln 2 + 0.119 \ln \sqrt{2}) = 1.61$).

The Moran's I test is employed to test the presence of spatial autocorrelation among residuals in our data set. We assume that the disturbances are normally distributed and therefore the Moran's I test under normality is conducted [Leung et al., 2000]. The resulting Moran's I statistic is 0.021 with the standard deviation of 0.448. The p-value of 0.327 suggests that the statistic does not significantly differ from zero and there is thus no spatial autocorrelation among the residuals. Nevertheless, we assume that the relationship between at-risk-of-poverty rates and the educational attainment structure might vary geographically. The GWR model for compositional tables is thus used to explore variations of this relationship across German regions.

It should be highlighted that the choice of the coordinate system does not affect the resulting outcome from either the global and local regression models. The coordinate system is chosen with respect to the convenience of results' interpretation. Our application builds on the coordinates constructed with the intention of exploring the relative dominance of attained tertiary education in the overall population aged 30-34 as an apparent determinant of at-risk-of-poverty rates.

The local regression model has been selected with the following settings of input parameters. The bi-square kernel type function was set to ADAPTIVE as it fits our data the best using the corrected AIC values (AICc). All 36 regions were considered in the estimation of each set of coefficients. The summary diagnostics of our local GWR model is provided in Table 7.

Table 7 Results from the local GWR model with the adaptive bandwidth of 36 (all regions), the given p-values are related to the Monte Carlo test of spatial variability of the coefficients. The R^2 and R^2_{adj} value are 0.514 and 0.254, residual sum of squares 0.577 with effective degrees of freedom 23.812. The effective number of parameters is 12.19, AIC equals -36.756 and AICc -14.210

	β_0	β_1^r	β_1^c	β_2^c	β_{11}^{OR}	β_{12}^{OR}
Min	2.778	-0.230	-0.233	-0.189	-0.946	0.283
1st Qu.	2.813	0.285	-0.146	-0.112	-0.869	0.658
Median	2.923	0.973	-0.060	-0.015	-0.764	1.248
3rd Qu.	3.025	1.419	0.012	0.083	-0.728	1.330
Max	3.122	1.479	0.019	0.148	-0.692	1.462
MCtest p-value	0.071	0.217	0.010	0.059	0.750	0.287
Coordinates		z_1^r	z_1^c	z_2^c	z_{11}^{OR}	z_{12}^{OR}

The comparison of both the global and the local regression model can be carried out through the AIC or AICc values ([Fotheringham et al., 2002], p. 96). The AIC value of the local model $AIC_{local} = -36.756$ is compared to the global one $AIC_{global} = -23.292$, $AICc_{local} = -14.210$ to the global one $AICc_{global} = -19.292$ respectively. The resulting statistics indicate lower value of AIC for the local model, however the AICc shows slightly better performance for the global one. Alternatively, the coefficient of determination R^2 and its adjusted version R^2_{adj} are examined. The global R^2 is 0.370 and the local R^2 0.514, the global R^2_{adj} is 0.265 and the local R^2_{adj} 0.254 respectively. Although the coefficient R^2 has increased for the GWR model, the gain is probably driven by the difference in the number of parameters and degrees of freedom as the adjusted R^2_{adj} values remain almost unchanged.

Table 7 summarises the main statistics of estimated parameters generated using the local model. The median values of parameters preserve the same signs as the estimates resulting from the global model (Table 6). The local measure of goodness-of-fit R^2 , whose geographical distribution is visualised in Fig. 5, varies between 0.48

and 0.58, which is satisfactory but with limitations. The R^2 values show a better fit for the model in the part of former East Germany and in the south of the country. In addition, each regression coefficient is accompanied by the p-value of the Monte Carlo test of its spatial variability Brunsdon et al. [1998]. The test results complement the ambiguous comparison of the local and the global model provided by AIC s and R^2 s. The majority of coefficients does not significantly vary across locations. However, the hypothesis that the parameter is globally fixed was rejected for β_1^c and values for β_0 and β_2^c are quite marginal. Hence, local differences are apparently driven by the effect of changes in the structure of population with the highest level of attained education. Values of β_1^c indicate that the growth of the mean proportion of the population with tertiary education has a positive effect on at-risk-of-poverty rates in the north of Germany, while it leads to an at-risk-of-poverty rate decrease in the southern regions. Moreover, the parameter β_2^c suggests that the growth of the (mean) relative proportion of people with only primary qualification (in comparison to those with the secondary education) contributes to higher at-risk-of-poverty rates in Northern Germany but it results in lower rates in the south.

According to the local regression model, apparent interactions between coordinates and at-risk-of-poverty rates are observed. Higher at-risk-of-poverty rates are associated with negative values of the coordinate z_{11}^{OR} and positive values of z_{12}^{OR} . Regions exhibiting this particular behaviour are Hamburg, Arnsberg and Saarland. All these regions reveal the at-risk-of-poverty rate higher than the overall national rate of 19.0 percent reported by Germany in 2017. However, the coordinates z_{11}^{OR} and z_{12}^{OR} attain values very close to zero and they all lie in the area with lower goodness-of-fit measures R^2 . The strongest association is observed in Hamburg with at-risk-of-poverty rate of 21.6 percent. The region experienced one of the lowest values of the coordinate $z_{11}^{OR} = -0.085$ and the second highest value of $z_{12}^{OR} = 0.180$. Furthermore, Hamburg lies in the area showing high significant t-values of parameters (-2.96 and 2.44 respectively) given by the local regression models, see Fig. 5.

In contrast, positive values of the coordinate z_{11}^{OR} and negative values of z_{12}^{OR} lead to lower at-risk-of-poverty rates. This trend has been experienced by regions of Weser-Ems, Detmold, Kassel, Unterfranken, Karlsruhe and Freiburg. Except for Weser-Ems, all mentioned regions experienced at-risk-of-poverty rates below the German national rate of 19.0 percent in 2017. Examining a specific example of Detmold, the trend is further illustrated. The share of people at risk of poverty or social exclusion in Detmold was 16.9 percent in 2017. The coordinate $z_{11}^{OR} = 0.186$ is the second highest one and the value of $z_{12}^{OR} = -0.045$ belongs to the lowest scores among all regions. Moreover, Detmold is situated in the area where both parameters are significant with t-values -2.75 and 2.44 respectively.

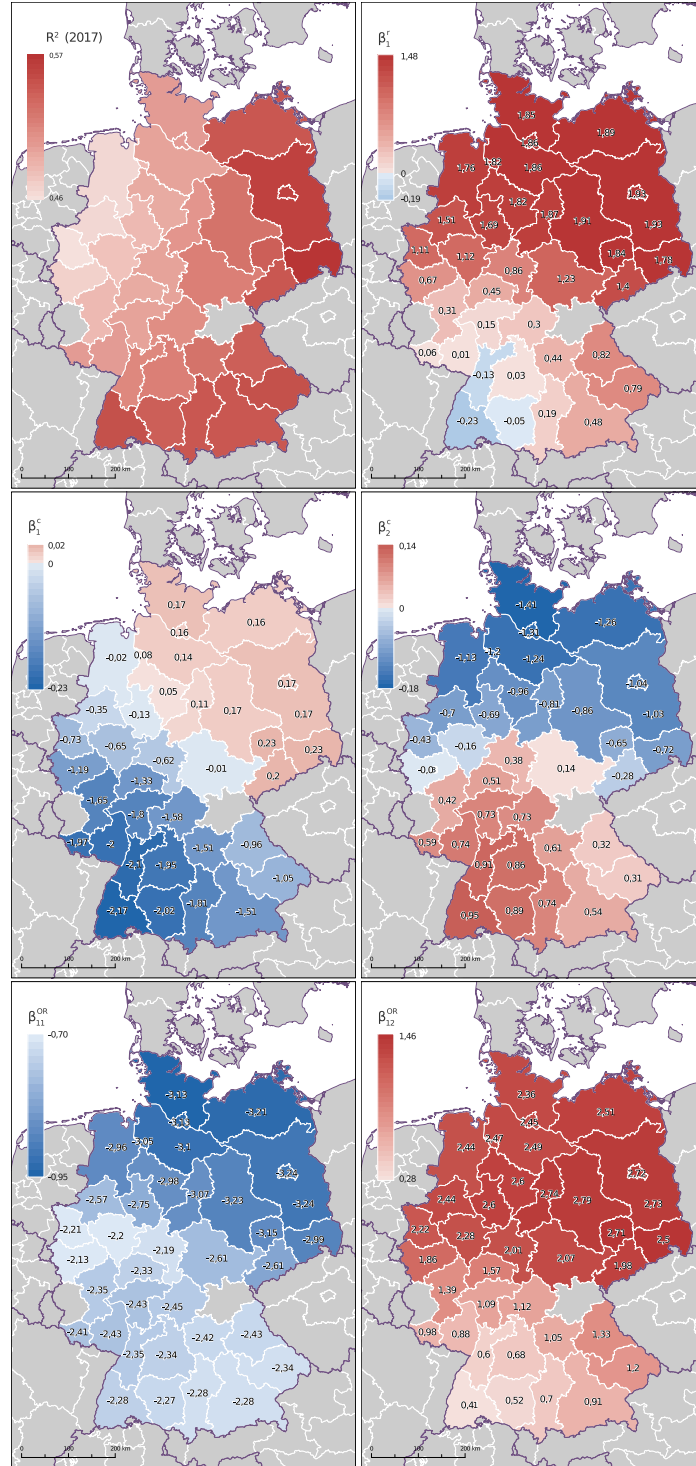


Fig. 5 Maps demonstrating estimates of regression parameters of the local regression models and values of the respective R^2 characteristic. The local t-value of the respective estimate is displayed in the centre of each region.

6 Conclusions

In this contribution, we introduced a novel approach to regression with spatial dependent compositional tables. Even though some methods for analysis of spatial dependent compositional data have already been developed, all of them are primarily designed for its one-factorial (vector) version. Hence, this chapter serves as the first step towards modelling more complex relative structures. The main contribution of this chapter is that it serves as an introduction of a geographically weighted regression model designed in order to inspect the relationship between the dependent variable and the explanatory variable expressed in the form of a compositional table.

It turned out that the local regression model was much more suitable for the analysis of the relationship between the risk of falling into poverty and the educational attainment structure of the 30-34 aged population in Germany. The comparison of both the global and local regression model revealed important aspects of local variations in our real data example. The selection of the coordinate system based on the desired research interest, in our case the dominance of tertiary education over the entire population, enabled a convenient interpretation of the regression outcomes with respect to the variable of interest, the at-risk-of-poverty rates.

The future research initiatives might be devoted to accommodation of spatial or spatiotemporal autoregressive models for the case of more-factorial compositional data as well as more complex or robust alternatives of the GWR model.

Acknowledgements This work was supported by The Czech Science Foundation under grant No.: 18-05432S.

References

- M. Simona Andreano, Roberto Benedetti, and Federica Piersimoni. A distance correlation index of spatial dependence for compositional data. *Pap. Reg. Sci.*, 98 (6):2371–2389, 2019.
- Luc Anselin. Spatial econometrics. In Badi H. Baltagi, editor, *A companion to theoretical econometrics*. Blackwell Publishing Ltd., 2001.
- Luc Anselin. *Spatial econometrics: methods and models (Vol. 4)*. Springer Science Business Media, 2013.
- Roger S. Bivand, Edzer Pebesma, and Virgilio Gomez-Rubio. *Applied spatial data analysis with R, Second edition*. Springer, NY, 2013. URL <http://www.asdar-book.org/>.
- Jan Brülle. *Poverty Trends in Germany and Great Britain*. Springer VS, Wiesbaden, 2018.
- Chris Brunsdon, Stewart Fotheringham, and Martin Charlton. Geographically weighted regression-modelling spatial non-stationarity. *J. Royal Stat. Soc., Ser. D Stat.*, 47(3):431–443, 1998.

- Andrew D. Cliff and Keith Ord. Spatial autocorrelation: A review of existing and new measures with applications. *Econ. Geogr.*, 46(sup1):269–292, 1970.
- Julie de Sousa, Karel Hron, Kamila Fačevicová, and Peter Filzmoser. Robust principal component analysis for compositional tables. *J. Appl. Stat.*, pages 1–20, 2020. doi: 10.1080/02664763.2020.1722078.
- Matteo Duiella and Alessandro Turrini. Poverty developments in the EU after the crisis: a look at main drivers. In *ECFIN Economic Briefs*. Directorate-General for Economic and Financial Affairs (European Commission), 2014.
- Juan José Egozcue and Vera Pawlowsky-Glahn. Groups of Parts and Their Balances in Compositional Data Analysis. *Math. Geol.*, 37:795–828, 2005.
- Juan José Egozcue, José Luis Díaz-Barrero, and Vera Pawlowsky-Glahn. Compositional analysis of bivariate discrete probabilities. In Josef Daunis i Estadella and Jose Antonio Martín-Fernández, editors, *Proceedings of CODAWORK’08, The 3rd Compositional Data Analysis Workshop*. University of Girona, Spain, 2008.
- Juan José Egozcue, Vera Pawlowsky-Glahn, Matthias Templ, and Karel Hron. Independence in contingency tables using the simplicial geometry. *Commun. in Stat. Theory Methods*, 44(18):3978–3996, 2015.
- ESRI. *ArcGIS Desktop: Release 10*. Environmental Systems Research Institute, Redlands, CA, 2011.
- Kamila Fačevicová, Karel Hron, Valentin Todorov, and Matthias Templ. Compositional tables analysis in coordinates. *Scand. J. Stat.*, 43(4):962–977, 2016.
- Kamila Fačevicová, Karel Hron, Valentin Todorov, and Matthias Templ. General approach to coordinate representation of compositional tables. *Scand. J. Stat.*, 45(4):879–899, 2018.
- Peter Filzmoser, Karel Hron, and Matthias Templ. *Applied Compositional Data Analysis*. Springer, Cham, 2018.
- A. Stewart Fotheringham. Trends in quantitative methods I: stressing the local. *Prog. Hum. Geogr.*, 21(1):88–96, 1997.
- A. Stewart Fotheringham, Chris Brunsdon, and Martin Charlton. *Geographically weighted regression: the analysis of spatially varying relationships*. Wiley, Chichester, NJ, 2002.
- Isabella Gollini, Binbin Lu, Martin Charlton, Christopher Brunsdon, and Paul Harris. GWmodel: An R Package for Exploring Spatial Heterogeneity Using Geographically Weighted Models. *J. Stat. Softw.*, 63(17):1–50, 2015.
- Walter Hanesch. ESPN Thematic Report on In-work poverty – Germany, 2019.
- Paul Harris, Chris Brunsdon, and Martin Charlton. Geographically weighted principal components analysis. *Int. J. Geogr. Inf. Sci.*, 25(10):1717–1736, 2011.
- Karel Hron, Peter Filzmoser, and Katherin Thompson. Linear regression with compositional explanatory variables. *J. Appl. Stat.*, 39(5):1115–1128, 2012.
- Stamatis Kalogirou. Testing local versions of correlation coefficients. *Review of Regional Research: Jahrbuch für Regionalwissenschaft*, 32(1):45–61, 2012.
- Leo Lahti, Janne Huovari, Markus Kainu, and Przemyslaw Biecek. eurostat R package, 2017. URL <https://journal.r-project.org/archive/2017/RJ-2017-019/index.html>. Version 3.6.1.

- Thomas J. Leininger, Alan E. Gelfand, Jenica M. Allen, and John A. Silander. Spatial regression modeling for compositional data with many zeros. *J. Agric. Biol. Environ. Stat.*, 18(3):314–334, 2013.
- Yee Leung, Chang-Lin Mei, and Wen-Xiu Zhang. Testing for Spatial Autocorrelation among the Residuals of the Geographically Weighted Regression. *Environ. Plan. A: Econ. Space*, 32(5):871–890, 2000. doi: 10.1068/a32117.
- Stephen Matthews and Tse-Chuan Yang. Mapping the results of local statistics: Using geographically weighted regression. *Demogr. Res.*, S13(6):151–166, 2012.
- Joanna Morais, Christine Thomas-Agnan, and Michel Simioni. Interpretation of Explanatory Variables Impacts in Compositional Regression Models. *Austrian J. Stat.*, 47(5):1–25, 2018.
- Klaus Nordhausen, Hannu Oja, Peter Filzmoser, and Clemens Reimann. Blind source separation for spatial compositional data. *Math. Geosci.*, 47(7):753–770, 2015.
- Vera Pawlowsky-Glahn and Antonella Buccianti. *Compositional Data Analysis, Theory and Applications*. John Wiley Sons, Ltd, 2011. ISBN 9781119976462.
- Vera Pawlowsky-Glahn, Juan José Egozcue, Ricardo A. Olea, and Eulogio Pardo-Igúzquiza. Cokriging of compositional balances including a dimension reduction and retrieval of original units. *J. South. Afr. Inst. Min. Metall.*, 115:59 – 72, 2015. ISSN 2225-6253.
- Jonas Pieper, Ulrich Schneider, Wiebke Schröder, and Gwendolyn Stilling. 30 Jahre Mauerfall – Ein viergeteiltes Deutschland. Der Paritätische Armutsbericht 2019, 2019.
- Behnaz Pirzamanbein, Johan Lindström, Anneli Poska, and Marie-José Gaillard. Modelling spatial compositional data: Reconstructions of past land cover and uncertainties. *Spat. Stat.*, 24:14–31, 2018.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL <https://www.R-project.org/>.
- Clemens Reimann, M. Ayas, V. A. Chekushin, I. Bogatyrev, R. Boyd, P. de Caritat, R. Dutter, T.E. Finne, J.H. Halleraker, and Ø Jæger. *Environmental geochemical atlas of the central Barents region*. Norges geologiske undersøkelser, 1997.
- Matthias Templ, Karel Hron, and Peter Filzmoser. robCompositions: An R-package for Robust Statistical Analysis of Compositional Data. In Vera Pawlowsky-Glahn and Antonella Buccianti, editors, *Compositional Data Analysis, Theory and Applications*, chapter 25, pages 341–355. John Wiley Sons, Ltd, 2011. ISBN 9781119976462.
- Waldo R. Tobler. A computer movie simulating urban growth in the detroit region. *Econ. Geogr.*, 46(sup1):234–240, 1970.
- Raimon Tolosana-Delgado, Vera Pawlowsky-Glahn, and Juan José Egozcue. Indicator kriging without order relation violations. *Math. Geosci.*, 40:327–347, 2008.
- Takahiro Yoshida and Morito Tsutsumi. On the effects of spatial relationships in spatial compositional multivariate models. *Letters in Spatial and Resource Sciences*, 11(1):57–70, 2018.

Independent component analysis

Muehlmann, C., **Fačevicová, K.**, Gardlo, A., Janečková, H., Nordhausen, K. Independent component analysis of compositional data. In: A. Daouia, A. Ruiz-Gazen, editors, *Advances in Contemporary Statistics and Econometrics*. Springer, Cham (2021).

https://doi.org/10.1007/978-3-030-73249-3_27

[7 citations on Scopus; 5 citations without autocitations]

[Author's contribution: 30%]

Abstract

Compositional data represent a specific family of multivariate data, where the information of interest is contained in the ratios between parts rather than in absolute values of single parts. The analysis of such specific data is challenging as the application of standard multivariate analysis tools on the raw observations can lead to spurious results. Hence, it is appropriate to apply certain transformations prior to further analysis. One popular multivariate data analysis tool is independent component analysis. Independent component analysis aims to find statistically independent components in the data and as such might be seen as an extension to principal component analysis. In this paper, we examine an approach of how to apply independent component analysis on compositional data by respecting the nature of the latter and demonstrate the usefulness of this procedure on a metabolomics dataset.

Independent Component Analysis for Compositional Data



Christoph Muehlmann, Kamila Fačevicová, Alžběta Gardlo, Hana Janečková,
and Klaus Nordhausen

Abstract Compositional data represent a specific family of multivariate data, where the information of interest is contained in the ratios between parts rather than in absolute values of single parts. The analysis of such specific data is challenging as the application of standard multivariate analysis tools on the raw observations can lead to spurious results. Hence, it is appropriate to apply certain transformations prior to further analysis. One popular multivariate data analysis tool is independent component analysis. Independent component analysis aims to find statistically independent components in the data and as such might be seen as an extension to principal component analysis. In this paper, we examine an approach of how to apply independent component analysis on compositional data by respecting the nature of the latter and demonstrate the usefulness of this procedure on a metabolomics dataset.

C. Muehlmann · K. Nordhausen (✉)
Institute of Statistics & Mathematical Methods in Economics,
Vienna University of Technology, Wiedner Hauptstr. 7, 1040 Vienna, Austria
e-mail: klaus.nordhausen@tuwien.ac.at

C. Muehlmann
e-mail: christoph.muehlmann@tuwien.ac.at

K. Fačevicová
Department of Mathematical Analysis and Applications of Mathematics,
Palacký University Olomouc, 17. listopadu 12, 771 46 Olomouc, Czech Republic
e-mail: kamila.facevicova@gmail.com

A. Gardlo
Department of Clinical Biochemistry, University Hospital Olomouc and Palacký University
Olomouc, I.P. Pavlova 185/6, 779 00 Olomouc, Czech Republic
e-mail: alzbetagardlo@gmail.com

H. Janečková
Laboratory for Inherited Metabolic Disorders, Department of Clinical Biochemistry,
University Hospital Olomouc and Palacký University Olomouc, I.P. Pavlova 185/6,
779 00 Olomouc, Czech Republic
e-mail: janeckovah@gmail.com

1 Introduction

Independent component analysis (ICA) is a well-established data analysis method in signal processing with the goal of recovering hidden signals that are usually meant to have a physical meaning. In recent years, ICA methods have attracted increasing interest in the statistics community as an extension of normality-based multivariate methods that only use second-order moments. In principle, ICA can be seen as a refinement of principal component analysis where, after removing second-order information, higher order moments are used to search for hidden structures which are not visible in the principal components. Classical ICA methods are mainly developed for independent and identically distributed observations in a Euclidean space. Nevertheless, these methods are also applied, for example, on time series, spatial data, etc. but to the best of our knowledge not on iid compositional data.

Compositional data is special in the way that the entries (parts) of a d -variate vector are positive and carry relative rather than absolute information about the respective observation of interest. Moreover, the parts of the compositional vector are by nature not independent and in some specific situations, e.g. when all parts are bounded by a constant sum constraint, a spurious correlation between them is present. Therefore, compositional data lies on a simplex and does not follow the real Euclidean geometry. Examples of compositional data are geochemical data where the chemical composition of soil samples is of interest, the composition of nutrients of food intake or the distribution of market shares. For further details and examples of compositional data, see, for example, Aitchison (1986), Egozcue and Pawlowsky-Glahn (2019), Fačevićová et al. (2016), Filzmoser et al. (2018), Morais et al. (2018), Pawlowsky-Glahn and Buccianti (2011), Trinh et al. (2019).

It is well established that standard multivariate methods should not be applied directly to compositional data. Either methods which take the geometry of compositional data into account or methods that transform compositional data in such a way that standard multivariate analysis tools can be applied are appropriate. In this paper, we take the latter approach.

We review some basic ICA methods in Sect. 2. In Sect. 3, we describe compositional data and methods to transform such data into the real space. Based on the former two sections, we present how ICA can be performed on compositional data in Sect. 4 and conclude the paper with the analysis of a metabolomics dataset from healthy newborns in Sect. 5 and a discussion in Sect. 6.

2 Independent Component Analysis

From a statistical perspective, independent component analysis is usually formulated as a latent variable model as follows.

Definition 1 An observable p -vector $\mathbf{x}$ follows the independent component (IC) model if

$$\mathbf{x} = \mathbf{A}\mathbf{z} + \mathbf{b},$$

where $\mathbf{A}$ is a $p \times p$ non-singular matrix, $\mathbf{b}$ a p -vector, and the latent p -variate random vector $\mathbf{z}$ satisfies

- (A1) $\mathbf{E}(\mathbf{z}) = \mathbf{0}$ and $\mathbf{COV}(\mathbf{z}) = \mathbf{I}_p$,
- (A2) the components of $\mathbf{z}$ are independent, and
- (A3) at most one component of $\mathbf{z}$ is Gaussian.

Thus $\mathbf{E}(\mathbf{x}) = \mathbf{b}$ and $\mathbf{COV}(\mathbf{x}) = \mathbf{A}\mathbf{A}^\top$. The goal of ICA is to find a $p \times p$ matrix $\mathbf{W}$ such that $\mathbf{W}\mathbf{x}$ has independent components. Note however that in general it will not hold that $\mathbf{W}(\mathbf{x} - \mathbf{b}) = \mathbf{z}$ as the IC model assumptions only fix the location and scale of $\mathbf{z}$ but not the signs or the order of the components. Therefore, for every solution $\mathbf{W}$, also $\mathbf{P}\mathbf{J}\mathbf{W}$ is a solution, where $\mathbf{P}$ is a $p \times p$ permutation matrix (1 per row and column, 0 elsewhere) and $\mathbf{J}$ is a $p \times p$ sign-change matrix (a diagonal matrix with ± 1 on its diagonal).

There are many suggestions in the literature on how to estimate $\mathbf{W}$ based on a sample $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, and for recent reviews see, for example, Comon and Jutten (2010), Nordhausen and Oja (2018). Almost all ICA methods make, however, use of the following result:

Key result Let $\mathbf{x}$ follow the IC model and denote $\mathbf{x}^{st} = \mathbf{COV}(\mathbf{x})^{-1/2}(\mathbf{x} - \mathbf{E}(\mathbf{x}))$, then there exists an orthogonal $p \times p$ matrix $\mathbf{U}$ such that

$$\mathbf{U}^\top \mathbf{x}^{st} = \mathbf{z}.$$

This result implies that after estimating $\mathbf{COV}(\mathbf{x})$ and $\mathbf{E}(\mathbf{x})$, the problem is reduced from finding a general $p \times p$ matrix to a $p \times p$ orthogonal matrix. Also note that this means that the performance of ICA methods does not depend on the values of $\mathbf{A}$ and $\mathbf{b}$, as these are accounted for when standardizing the data. An unmixing matrix estimate is therefore obtained as $\mathbf{W} = \mathbf{U}^\top \mathbf{COV}(\mathbf{x})^{-1/2}$ and different ICA approaches differ in the way they estimate $\mathbf{U}$. In the following, we will show how some popular ICA methods estimate this rotation.

2.1 FOBI

Fourth-order blind identification (FOBI), presented in Cardoso (1989), was one of the first ICA methods but is still popular as it has a closed-form solution. For FOBI, we need to define the scatter matrix of fourth-order moments

$$\mathbf{COV}_4(\mathbf{x}) = \frac{1}{p+2} \mathbf{E} \left((\mathbf{x} - \mathbf{E}(\mathbf{x}))^\top \mathbf{COV}(\mathbf{x})^{-1} (\mathbf{x} - \mathbf{E}(\mathbf{x})) (\mathbf{x} - \mathbf{E}(\mathbf{x})) (\mathbf{x} - \mathbf{E}(\mathbf{x}))^\top \right).$$

Then we can define the following:

Definition 2 The FOBI unmixing matrix is $\mathbf{W}_{\text{FOBI}} = \mathbf{U}_{\text{FOBI}}^\top \mathbf{COV}(\mathbf{x})^{-1/2}$ where the columns of $\mathbf{U}_{\text{FOBI}}$ are given by the eigenvectors of $\mathbf{COV}_4(\mathbf{x}^{st})$.

From denoting $\mathbf{U}_{\text{FOBI}} \mathbf{D} \mathbf{U}_{\text{FOBI}}^\top$ to be the eigendecomposition of $\mathbf{COV}_4(\mathbf{x}^{st})$ which is needed to compute $\mathbf{W}_{\text{FOBI}}$, it is obvious that FOBI is only unique when the eigenvalues contained in the diagonal matrix $\mathbf{D}$ are distinct. One can actually show that these eigenvalues are linked to the kurtosis values of the independent components. For FOBI to be well-defined, Assumption (A3) from the IC model needs to be replaced by the stronger assumption:

(A4) The kurtosis values of the independent components must be distinct.

FOBI is often the first ICA method applied as it is quick to compute, gives a fast first impression, and its statistical properties are well known; see, for example, Miettinen et al. (2015), Nordhausen and Virta (2019) for more details. FOBI can also be of interest outside the IC model and can be seen as an invariant coordinate selection method (Tyler et al. 2009).

2.2 JADE

Assumption (A4) is considered highly restrictive. Joint approximate diagonalization of eigenmatrices (JADE) can be seen as an extension of FOBI which relaxes this strict assumption, Cardoso and Souloumiac (1993).

For JADE, we have to define the fourth-order cumulant matrices

$$\mathbf{C}_{ij}(\mathbf{x}) = \mathbf{E} \left((\mathbf{x}^{st\top} \mathbf{E}_{ij} \mathbf{x}^{st}) \mathbf{x}^{st} \mathbf{x}^{st\top} \right) - \mathbf{E}_{ij} - \mathbf{E}_{ij}^\top - \text{tr}(\mathbf{E}_{ij}) \mathbf{I}_p,$$

where $\mathbf{E}_{ij} = \mathbf{e}_i \mathbf{e}_j^\top$ with $\mathbf{e}_i$ being a vector of dimension p with the i th element equals 1 and 0 otherwise. As i and j range from 1 to p , there are in total p^2 such cumulant matrices. In the IC model, $\mathbf{C}_{ij}(\mathbf{z}) = 0$ if $i \neq j$ and for the case where $i = j$ $\mathbf{C}_{ii}(\mathbf{z})$ corresponds to the kurtosis of the i th component. The matrix of fourth moments can actually be expressed as

$$\mathbf{COV}_4(\mathbf{x}) = \frac{1}{p+2} \sum_{i=1}^p \mathbf{C}_{ii}(\mathbf{x}) + (p+2) \mathbf{I}_p,$$

meaning that it uses not all possible cumulant information. The idea of JADE is to exploit the information contained in all cumulant matrices.

Definition 3 The JADE unmixing matrix is $\mathbf{W}_{\text{JADE}} = \mathbf{U}_{\text{JADE}}^\top \mathbf{COV}(\mathbf{x})^{-1/2}$ where $\mathbf{U}_{\text{JADE}}$ is the maximizer of

$$\sum_{i=1}^p \sum_{j=1}^p \|\text{diag}(\mathbf{U}^\top \mathbf{C}_{ij}(\mathbf{x}^{st}) \mathbf{U})\|_F^2.$$

Thus, JADE tries to maximize the diagonal elements of $\mathbf{U}^\top \mathbf{C}_{ij}(\mathbf{x}^{st}) \mathbf{U}$ which is equivalent to minimize the off-diagonal elements by the orthogonal invariance of the Frobenius norm $\|\cdot\|_F$. As in the IC model, only $\mathbf{C}_{ii}(\mathbf{z})$ is non-zero and corresponds to the kurtosis of z_i . This means that JADE relaxes the FOBI assumption (A4) to the following:

(A5) At most one independent component can have zero kurtosis.

For a finite sample, the joint diagonalization of more than two matrices needs to be carried out approximately; many algorithms that jointly diagonalize two or more matrices are available; see, for example, Illner et al. (2015). For the purpose of this paper, we will use an algorithm based on Givens rotations, Clarkson (1988).

The statistical properties of JADE are, for example, given in Miettinen et al. (2015); from an asymptotic point of view, FOBI is never superior compared to JADE. JADE is however computationally more expensive, especially when the number of independent components grows, as p^2 matrices need to be computed and jointly diagonalized.

As a compromise, k-JADE was suggested in Miettinen et al. (2013). The idea is to use not all matrices $\mathbf{C}_{ij}$, but only those whose indices are not too far apart, i.e. $|i - j| < k$. This requires however that the first step, the whitening step, is not done using just the covariance matrix but using $\mathbf{W}_{\text{FOBI}}$.

Definition 4 Denote $\mathbf{x}^{st'} = \mathbf{W}_{\text{FOBI}}(\mathbf{x} - \mathbf{E}(\mathbf{x}))$ and choose an integer $1 \leq k \leq p$, then the k-JADE unmixing matrix is $\mathbf{W}_{\text{kJADE}} = \mathbf{U}_{\text{kJADE}}^\top \mathbf{W}_{\text{FOBI}}$ where $\mathbf{U}_{\text{kJADE}}$ is the maximizer of

$$\sum_{|i-j|<k}^p \|\text{diag}(\mathbf{U}^\top \mathbf{C}_{ij}(\mathbf{x}^{st'}) \mathbf{U})\|_F^2.$$

The value k is basically a tuning parameter. The intuition is that the multiplicities of the distinct non-zero kurtosis values of the independent components are at most k , and that there is at most one component having kurtosis zero. Usually, k is simply chosen by the user based on expert knowledge. In Virta et al. (2020), some guidelines for the selection are offered, which are however not very practical. The statistical properties of k-JADE are given in Miettinen et al. (2013), Virta et al. (2020). It can be shown that for a value of k which fulfills the multiplicity condition, k-JADE is asymptotically as efficient as JADE but has, if k is small, a much smaller computational complexity.

2.3 FastICA

FOBI, JADE, and k-JADE are often called algebraic ICA methods. Another large group of ICA methods is based on projection pursuit ideas, where the most prominent one is FastICA. It was originally suggested in Hyvärinen (1999a). Some of the many FastICA variants are discussed below.

The general idea of FastICA is to find the column vectors $\mathbf{u}_1, \dots, \mathbf{u}_p$ of $\mathbf{U}$ which maximize the non-Gaussianity of the components of $\mathbf{U}^\top \mathbf{x}^{st}$. Non-Gaussianity of a univariate random variable x is measured by $|E(G(x))|$ with some twice continuously differentiable and non-quadratic function G that satisfies $E(G(y)) = 0$ for $y \sim N(0, 1)$. The most popular choices for G are

pow3: $G(x) = (x^4 - 3)/4$,
 tanh: $G(x) = \log(\cosh(x)) - c_t$, and
 gauss: $G(x) = -\exp(-x^2/2) - c_g$.

The constants $c_t = E(\log(\cosh(y))) \approx 0.375$ and $c_g = E(-\exp(-y^2/2)) \approx -0.707$ are normalizing constants. The derivatives of G , denoted as g , are called non-linearities and are the name givers as pow3 : $g(x) = x^3$, tanh : $g(x) = \tanh(x)$ and gauss : $g(x) = x \exp(-x^2/2)$.

2.3.1 Deflation-Based FastICA

FastICA was first suggested in Hyvärinen and Oja (1997) using the non-linearity pow3 and finding the column vectors of $\mathbf{U}_{DF}$ one after another which is now known as deflation-based FastICA.

Definition 5 The deflation-based FastICA unmixing matrix is defined as $\mathbf{W}_{DF} = \mathbf{U}_{DF}^\top \mathbf{COV}(\mathbf{x})^{-1/2}$, where the k th column of $\mathbf{U}$, $\mathbf{u}_k$, maximizes

$$|E[G(\mathbf{u}_k^\top \mathbf{x}_{st})]|$$

under the constraints $\mathbf{u}_k^\top \mathbf{u}_k = 1$ and $\mathbf{u}_j^\top \mathbf{u}_k = 0$, $j = 1, \dots, k-1$.

To obtain estimates, a modified Newton-Raphson algorithm is used which iterates the following steps until convergence:

$$\begin{aligned} \mathbf{u}_k &\leftarrow \mathbf{E}[g(\mathbf{u}_k^\top \mathbf{x}_{st}) \mathbf{x}_{st}] - \mathbf{E}[g'(\mathbf{u}_k^\top \mathbf{x}_{st})] \mathbf{u}_k \\ \mathbf{u}_k &\leftarrow \left(\mathbf{I}_p - \sum_{l=1}^{k-1} \mathbf{u}_l \mathbf{u}_l^\top \right) \mathbf{u}_k \\ \mathbf{u}_k &\leftarrow \|\mathbf{u}_k\|^{-1} \mathbf{u}_k. \end{aligned}$$

The last two steps perform the Gram-Schmidt orthonormalization.

The properties of deflation-based FastICA have been studied in detail in Ollila (2010), Nordhausen et al. (2011). One issue with deflation-based FastICA is that besides the global maximum it has many local maxima and the order in which the vectors $\mathbf{u}_k$ are found depends heavily on the initial value of the algorithm, where in turn the estimation performance depends on the order in which the vectors $\mathbf{u}_k$ are found. Using asymptotic arguments, Nordhausen et al. (2011) suggested reloaded

Table 1 Table of default candidate set of non-linearities of adaptive deflation-based FastICA, where $(x)_+ = x$ if $x > 0$ and 0 otherwise, and $(x)_- = x$ if $x < 0$ and 0 otherwise

$g_1(x) = x^3$	$g_6(x) = (x)_+^2 + (x)_-^2$	$g_{11}(x) = (x - 1.0)_+^2 + (x + 1.0)_-^2$
$g_2(x) = \tanh(x)$	$g_7(x) = (x - 0.2)_+^2 + (x + 0.2)_-^2$	$g_{12}(x) = (x - 1.2)_+^2 + (x + 1.2)_-^2$
$g_3(x) = x \exp(-x^2/2)$	$g_8(x) = (x - 0.4)_+^2 + (x + 0.4)_-^2$	$g_{13}(x) = (x - 1.4)_+^2 + (x + 1.4)_-^2$
$g_4(x) = (x + 0.6)_-^2$	$g_9(x) = (x - 0.6)_+^2 + (x + 0.6)_-^2$	$g_{14}(x) = (x - 1.6)_+^2 + (x + 1.6)_-^2$
$g_5(x) = (x - 0.6)_+^2$	$g_{10}(x) = (x - 0.8)_+^2 + (x + 0.8)_-^2$	

FastICA, which estimates first the independent components using FOBI or k-JADE and then derives an optimal order based on the estimated independent components.

The idea of reloaded FastICA to fix the extraction order based on asymptotic arguments was extended in Miettinen et al. (2014) to also select an optimal non-linearity for each component out of a candidate set of possible non-linearities. This is known as adaptive deflation-based FastICA. We will denote the adaptive deflation-based FastICA unmixing matrix as $\mathbf{W}_{\text{ADF}}$. The candidate set of non-linearities suggested in Miettinen et al. (2014) contains, for example, the non-linearities presented in Table 1.

2.3.2 Symmetric FastICA

A FastICA variant estimating all directions in parallel was suggested in Hyvärinen (1999b).

Definition 6 The symmetric FastICA estimator $\mathbf{W}_{\text{SF}} = \mathbf{U}_{\text{SF}}^\top \mathbf{COV}(\mathbf{x})^{-1/2}$ uses as a criterion for $\mathbf{U}_{\text{SF}}$

$$\sum_{j=1}^p |\mathbf{E}[G(\mathbf{u}_j^\top \mathbf{x}_{st})]|$$

which should be maximized under the orthogonality constraint $\mathbf{U}_{\text{SF}}^\top \mathbf{U}_{\text{SF}} = \mathbf{I}_p$.

The steps of the iterative algorithm to compute $\mathbf{U}_{\text{SF}}$ are

$$\begin{aligned} \mathbf{u}_k &\leftarrow \mathbf{E}[g(\mathbf{u}_k^\top \mathbf{x}_{st}) \mathbf{x}_{st}] - \mathbf{E}[g'(\mathbf{u}_k^\top \mathbf{x}_{st})] \mathbf{u}_k, \quad k = 1, \dots, p \\ \mathbf{U}_{\text{SF}}^\top &\leftarrow (\mathbf{U}_{\text{SF}}^\top \mathbf{U}_{\text{SF}})^{-1/2} \mathbf{U}_{\text{SF}}^\top. \end{aligned}$$

The first update step of the algorithm is similar to that of the deflation-based FastICA estimator. The orthogonalization step can be interpreted as taking an average

over the vectors of the first step. This differs from the deflation-based approach where errors made in the k th direction carry on to the following directions and therefore the errors accumulate. This is often the reason why symmetric FastICA is usually considered superior to the deflation-based FastICA. However, there are also cases where the accumulation is preferable to the averaging. This occurs when some independent components are easier to find than the others. Statistical properties of symmetric FastICA are given in Miettinen et al. (2015), Wei (2015), Miettinen et al. (2017).

2.3.3 Squared Symmetric FastICA

One of the most recent variants of FastICA is the squared symmetric FastICA estimator (Miettinen et al. 2017). The idea of this estimator is to replace the absolute values in the objective function of the symmetric FastICA with squared values.

Definition 7 The squared symmetric FastICA estimator $\mathbf{W}_{\text{S2F}} = \mathbf{U}_{\text{S2F}}^\top \mathbf{COV}(\mathbf{x})^{-1/2}$ obtains $\mathbf{U}_{\text{S2F}}$ as the maximizer of

$$\sum_{j=1}^p (\mathbf{E}[G(\mathbf{u}_j^\top \mathbf{x}_{st})])^2$$

under the orthogonality constraint $\mathbf{U}_{\text{S2F}}^\top \mathbf{U}_{\text{S2F}} = \mathbf{I}_p$.

The steps of the resulting algorithm are

$$\begin{aligned} \mathbf{u}_k &\leftarrow \mathbf{E}[G(\mathbf{u}_k^\top \mathbf{x}_{st})](\mathbf{E}[g(\mathbf{u}_k^\top \mathbf{x}_{st})\mathbf{x}_{st}] - \mathbf{E}[g'(\mathbf{u}_k^\top \mathbf{x}_{st})]\mathbf{u}_k), \quad k = 1, \dots, p, \\ \mathbf{U}_{\text{S2F}}^\top &\leftarrow (\mathbf{U}_{\text{S2F}}^\top \mathbf{U}_{\text{S2F}})^{-1/2} \mathbf{U}_{\text{S2F}}^\top. \end{aligned}$$

Thus, the first step of the algorithm equals the first step in the symmetric algorithm with an additional multiplication by $\mathbf{E}[G(\mathbf{u}_k^\top \mathbf{x}_{st})]$. Hence, the squared symmetric variant puts more weight on components that are “more” non-Gaussian, which most often, but not always, is advantageous. The properties of the squared symmetric FastICA estimator as well as comparisons to the deflation-based and symmetric FastICA methods are given in Miettinen et al. (2017). In Miettinen et al. (2017), it is also shown that if the non-linearity `pow3` is used, symmetric squared FastICA is asymptotically equivalent to JADE.

Besides assumptions (A1)–(A3), deflation-based, symmetric, and squared symmetric FastICA need further assumptions based on G to ensure consistency. Assuming the order of the components is fixed as $|\mathbf{E}[G(z_1)]| \geq \dots \geq |\mathbf{E}[G(z_p)]|$, then it is required that for any $\mathbf{z} = (z_1, \dots, z_p)^\top$ with independent and standardized components and for any orthogonal matrix $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_p)$, the following holds.

For deflation-based FastICA:

(A6) For all $k = 1, \dots, p$, $|\mathbf{E}[G(\mathbf{u}_k^\top \mathbf{z})]| \leq |\mathbf{E}[G(z_k)]|$, when $\mathbf{u}_k^\top \mathbf{e}_j = 0$ for all $j = 1, \dots, k-1$, where $\mathbf{e}_i$ is a p -vector with i th element one and others zero,

for symmetric FastICA

(A7) $|\mathbf{E}[G(\mathbf{u}_1^\top \mathbf{z})]| + \dots + |\mathbf{E}[G(\mathbf{u}_p^\top \mathbf{z})]| \leq |\mathbf{E}[G(z_1)]| + \dots + |\mathbf{E}[G(z_p)]|$,

and for squared symmetric FastICA

(A8) $(\mathbf{E}[G(\mathbf{u}_1^\top \mathbf{z})])^2 + \dots + (\mathbf{E}[G(\mathbf{u}_p^\top \mathbf{z})])^2 \leq (\mathbf{E}[G(z_1)])^2 + \dots + (\mathbf{E}[G(z_p)])^2$.

It was proven, for example, in Miettinen et al. (2015), that all three conditions are fulfilled with pow3. On the other hand, in the case of non-linearities like tanh and gauss some of these conditions might be violated for certain source distributions.

From a computational point of view, the advantage of both symmetric versions is that the initial value of $\mathbf{U}$ is not important when the sample size is large, as the algorithms converge usually to the global maxima.

To conclude this section we can point out that FOBI, JADE, k-JADE, symmetric FastICA, and squared symmetric FastICA are affine equivariant ICA methods which means that their performance does not depend on the mixing matrix. So, from this point of view, only deflation-based FastICA differs, which can be overcome when the reloaded version or adaptive version is used. Affine equivariance will be of relevance later when applying the ICA methods to compositional data.

3 Compositional Data and Its Real Space Representation

A specific family of d -dimensional vectors is present when each entry (part) of a vector is positive and carries information about its contribution to the whole. In the following, such multivariate observations are called (vector) compositional data, whose specifics were already described, utilized, and analyzed in a wide range of applications (Pawlowsky-Glahn and Buccianti 2011). The main property of compositional data is its relative nature, when the relevant information is contained in the ratios between parts rather than in the absolute values of the parts. Consider, e.g. a vector describing a geochemical structure of soil, where each part represents the quantity of the given element in the sample. The quantity can be given either in absolute scale, like in mg of the component contained in the sample, or some of its relative alternatives, typically ppm. While the mg representation depends on the overall size of the sample, the ppm one does not, despite the ratios between parts remaining unchanged. Both representations are therefore from the compositional point of view equivalent.

Due to the relative nature of compositional data, the sample space of representations of a d -part compositional vector $\mathbf{x}$ forms a d -part simplex

$$\mathcal{S}^d = \left\{ \mathbf{x} = (x_1, \dots, x_d)^\top, \sum_{i=1}^d x_i = \kappa, \kappa > 0 \right\},$$

where the Aitchison geometry holds. The whole sample space is formed by equivalence classes of proportional vectors (Pawlowsky-Glahn et al. 2015, Chaps. 2, 3). Since most of the standard statistical methods are designed for real-valued data following the usual Euclidean geometrical structure, it is favorable to express compositional data in real coordinates prior to their analysis. One of the possible representations is the centered log-ratio (clr) transformation from $\mathcal{S}^d$ to $\mathbb{R}^d$ given by

$$\text{clr}(\mathbf{x})_i = \ln \frac{x_i}{g_m(\mathbf{x})} = \frac{1}{d} \sum_{j=1}^d \ln \frac{x_i}{x_j}, \quad \text{for } i = 1, \dots, d,$$

where $g_m(\mathbf{x})$ denotes the geometrical mean of all parts. The parts of the resulting clr vector can be interpreted in terms of the dominance of the compositional part in the numerator within the whole composition or equivalently as its mean dominance over each part of the whole composition. The use of logarithm symmetrizes this relationship. Let us stress here that the clr values depend on the set of compositional parts used for its computation and therefore the above interpretation holds true only when the whole composition is considered. Within the whole manuscript, the clr transformation based on all compositional parts will be of interest. On the other hand, from its construction, the clr coefficients/variables are not linearly independent, as they sum up to zero and, therefore, the whole clr vector falls in a $(d - 1)$ -dimensional subspace of $\mathbb{R}^d$. This feature prevents direct use of the clr representation within methods that require full rank data, like robust PCA (Filzmoser et al. 2009) or the above stated ICA methods.

One possible workaround is the isometric log-ratio (ilr) transformation, which represents the compositional vector $\mathbf{x}$ in a system of $d - 1$ orthonormal real coordinates. This system can be obtained directly from the clr vector as

$$\text{ilr}(\mathbf{x}) = \mathbf{V}^\top \text{clr}(\mathbf{x}),$$

where the columns of the $d \times d - 1$ log-contrast matrix $\mathbf{V}$ are given as $\mathbf{v}_i = \text{clr}(\xi_i)$ and the vectors ξ_i , $i = 1, \dots, d - 1$ constitute an orthonormal basis in $\mathcal{S}^d$. See Pawlowsky-Glahn and Buccianti (2011), Ch. 11 for details.

The system of basis vectors $\{\xi_1, \dots, \xi_{d-1}\}$ is not uniquely given and can be chosen according to the purpose of further analysis. Since each system of ilr coordinates can be obtained as an orthogonal rotation of the others, its specific choice does not affect the results of their analysis, like predictions of the regression model with a compositional regressor or scores of the robust PCA model (Filzmoser et al. 2009; Hron et al. 2012). When it is required, a specific coordinate system can be selected by some data-driven method, like hierarchical clustering of the compositional parts, or using expert knowledge. In both cases, the main aim is to obtain such an interpretation of the coordinates at hand, which is favorable according to the given problem (Egozcue

and Pawlowsky-Glahn 2005). Since a specific interpretation of the ilr coordinates is not the main purpose here, the same system as in Nordhausen et al. (2015) is used. The basis vectors ξ_i have the value $\exp(\sqrt{1/i(i+1)})$ at the first i positions, $\exp(-\sqrt{1/i(i+1)})$ at the position $i+1$, and 1 at the remaining ones. Consequently, the columns of the log-contrast matrix are

$$\mathbf{v}_i = \sqrt{\frac{i}{i+1}} \left(\frac{1}{i}, \dots, \frac{1}{i}, -1, 0, \dots, 0 \right)^\top, \quad i = 1, \dots, d-1.$$

The ilr coordinates have the form of balances between the i th part of the composition and all parts with lower indices

$$\text{ilr}(\mathbf{x})_i = \sqrt{\frac{i}{i+1}} \ln \left(\frac{(x_1 \cdots x_i)^{1/i}}{x_{i+1}} \right), \quad \text{for } i = 1, \dots, d-1.$$

Finally, the clr and ilr representations are mutually transferable through the contrast matrix $\mathbf{V}$

$$\text{clr}(\mathbf{x}) = \mathbf{V} \text{ilr}(\mathbf{x})$$

and also the back-transformation to the simplex is possible by using

$$\mathbf{x} = \exp(\text{clr}(\mathbf{x})) = \exp(\mathbf{V} \text{ilr}(\mathbf{x})).$$

4 ICA for Compositional Data

As described above, ICA is not reasonable for data following the Aitchison geometry in its raw form. Therefore, it is natural to transform the data first into the Euclidean space. As ICA methods start with whitening and therefore require full rank data, the ilr space is the most natural representation. Due to the affine equivariance property of the discussed ICA methods, the particular used basis for the ilr transformation at most affects the order and signs of the estimated independent components. Hence, for compositional ICA we have the following model assumption:

$$\text{ilr}(\mathbf{x}) = \mathbf{A}_{\text{ilr}} \mathbf{z} + \mathbf{b},$$

where $\mathbf{A}_{\text{ilr}}$ is a $(d-1) \times (d-1)$ full rank mixing matrix specific for a chosen ilr basis, $\mathbf{b}$ a $d-1$ -dimensional location vector, and $\mathbf{z} = (z_1, \dots, z_{d-1})^\top$ a random vector with independent components, which are standardized so that $\mathbf{E}(\mathbf{z}) = \mathbf{0}$ and $\mathbf{COV}(\mathbf{z}) = \mathbf{I}_{d-1}$. When the unmixing matrix $\mathbf{W}_{\text{ilr}}$ is estimated using one of the ICA methods described in Sect. 2, the system of independent components is given by

$$\mathbf{z} = \mathbf{W}_{\text{ilr}}(\text{ilr}(\mathbf{x}) - \mathbf{b}) = \mathbf{W}_{\text{ilr}}(\mathbf{V}^\top \text{clr}(\mathbf{x}) - \mathbf{b}).$$

As ilr coordinates are not directly related to the dominance of the original parts within the considered composition, the relationship between ilr and clr spaces can be exploited yielding a $(d - 1) \times d$ “clr” loading matrix $\mathbf{W}_{\text{clr}} = \mathbf{W}_{\text{ilr}} \mathbf{V}^\top$, allowing interpretation of the independent components in the clr space. In the context of principal component analysis performed in the clr space, principal components lead to a new system of ilr coordinates (Pawlowsky-Glahn et al. 2011). This is not the case for ICA, as the unmixing matrix $\mathbf{W}_{\text{ilr}}$ (and consequently also $\mathbf{W}_{\text{clr}}$) is generally not restricted to be orthogonal. Even if the independent component model does not hold, ICA transformations remain affine equivariant which means that $\mathbf{z}$ can be seen as an intrinsic data representation with a coordinate system, whose components are as independent as possible.

After performing ICA, one is usually interested in either using $\mathbf{z}$ itself for further analysis, such as classification and outlier identification, with possible interpretation in ilr or clr space using the former defined loading matrices $\mathbf{W}_{\text{ilr}}$ or $\mathbf{W}_{\text{clr}}$, or, using ICA for noise or artifact removal. For that purpose, the components of $\mathbf{z}$ are divided into a signal part $\mathbf{z}_s$ and a noise/artifact part $\mathbf{z}_n$. This defines also the partition of the unmixing matrix $\mathbf{W}_{\text{ilr}}$ into $\mathbf{W}_{\text{ilr}}^s$ and $\mathbf{W}_{\text{ilr}}^n$ and the mixing matrix $\mathbf{A}_{\text{ilr}} = (\mathbf{W}_{\text{ilr}})^{-1}$ into $\mathbf{A}_{\text{ilr}}^s$ and $\mathbf{A}_{\text{ilr}}^n$. $\mathbf{A}_{\text{ilr}}^s$ is formed only by those columns of $\mathbf{A}_{\text{ilr}}$ that correspond to the signal components $\mathbf{z}_s$. The pure signal can then be restored in the ilr, clr, and original space by using

$$\text{ilr}(\mathbf{x})_s = \mathbf{A}_{\text{ilr}}^s \mathbf{z}_s + \mathbf{b}, \quad \text{clr}(\mathbf{x})_s = \mathbf{V} (\mathbf{A}_{\text{ilr}}^s \mathbf{z}_s + \mathbf{b}), \quad \text{and} \quad \mathbf{x}_s = \exp [\mathbf{V} (\mathbf{A}_{\text{ilr}}^s \mathbf{z}_s + \mathbf{b})],$$

respectively.

5 A Case Study in Metabolomics

In order to demonstrate the above-described methods, the data from a neonatal screening program in the Czech Republic was analyzed. Anonymous data were obtained from a retrospective study approved by the Ethics Committee of the University Hospital Olomouc which was part of a larger international study described in Fleischman et al. (2013). Newborn screening is a preventive program that allows for early detection of a selected spectrum of inborn metabolic diseases. At an age of 48–72 hours after birth, several drops of blood from the heel of the child were sampled on a special paper and sent for analysis to the screening laboratory. The data at hand were constituted by the metabolite profile of over 10 000 healthy newborns. For each neonate, the values of 48 metabolites were measured. Moreover, information about sex and birth weight was available. More specifically, the birth weight ranged from 300 to 5 570 grams and for newborns with very low birth weight (less than 1500 grams) a different metabolite structure can be expected, due to their prematurity and the artificial nutrition they receive. One of the main goals of metabolomics is to investigate interactions between metabolites, their dynamic changes, and responses to stimuli. Biofluids, e.g. blood or urine, and also tissues are used for the analysis. On the one

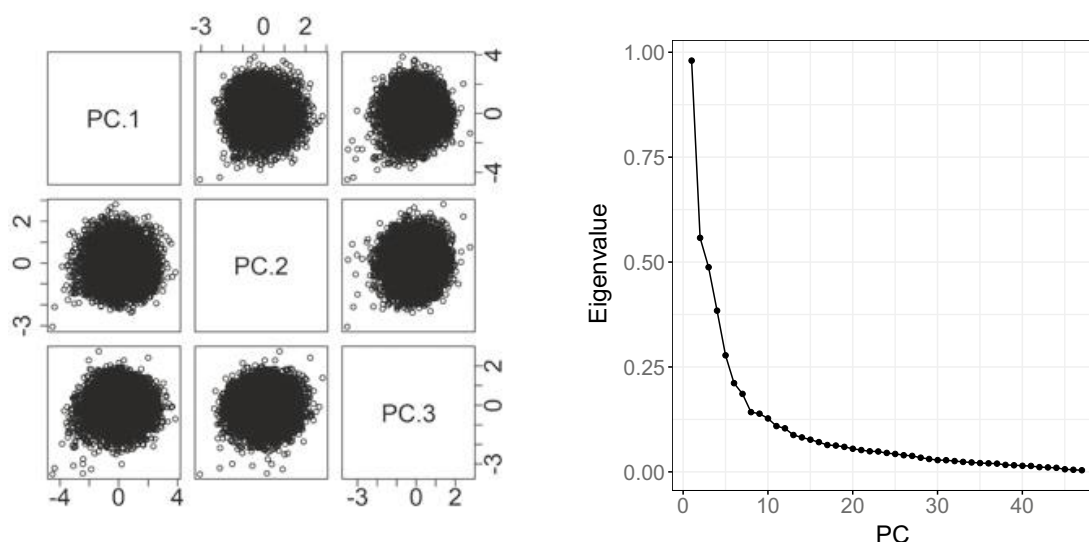


Fig. 1 Scatterplots of the first three principal components resulting from the compositional PCA (left) and scree plot of the respectively explained variability (right)

hand, the most frequently used approach for the data analysis is done through comparison of absolute values of biomarkers and reference ranges (data from the healthy population). On the other hand, the new trend of data evaluation is based on the use of ratios of metabolite data. Relative changes are more relevant/informative than absolute values in diagnostics based on profiling. Therefore, metabolomic data can be considered as observations carrying relative information, i.e. as compositional data (Kalivodová et al. 2018), and as such the above-discussed methods can be applied.

The following analysis was carried out in R 3.6.1 (R Core Team 2019) with the help of the packages JADE (Miettinen et al. 2017), fICA (Miettinen et al. 2018), compositions (van den Boogaart et al. 2019), and robCompositions (Templ et al. 2011). As the first step, standard principal component analysis (PCA) was performed on the clr transformed data. There were no significant patterns visible within the first three principal components; see Fig. 1, left. The whole dataset forms one quite compact cluster with no outliers. Moreover, the variance explained by the first components is low (around 20 % for the first PC) (Fig. 1, right), and therefore PCA does not seem to deal well with the issue of outlier detection, grouping, as well as dimension reduction in that case.

As PCA seems not to reveal any clear structure, we applied FOBI, k-JADE, with $k = 5$, and adaptive deflation-based FastICA to the ilr representation of the data (the dimension $p = 47$ was already too large for JADE). For easier comparison, the components from all three ICA methods were ordered according to their kurtosis values. As all three ICA methods showed similar results, we focus on our presentation and discussion of the components on those from adaptive deflation-based FastICA.

Due to the kurtosis ordering, the first components show heavy-tailed distributions, and they are expected to find outliers or small groupings, while the last components show light-tailed distributions and hence might find more balanced groupings. Scores of the first and last three independent components are plotted in Fig. 2, and the chosen

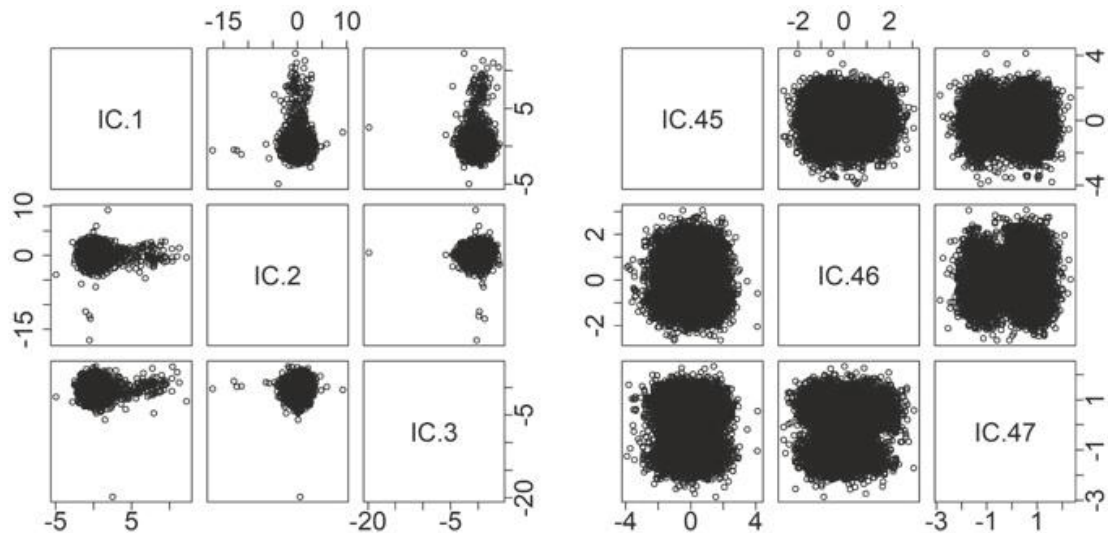


Fig. 2 Scatterplots of the first (left) and last (right) three independent components resulting from the compositional FastICA, using adaptive deflation-based FastICA

non-linearities are given in Table 2 for all independent components. According to the left plot of Fig. 2, one outlier is clearly detected due to its high negative value in the third component (IC.3). According to its loadings, which are collected in Table 3, IC.3 mostly reflects the relative dominance (with respect to concentrations of all 48 measured metabolites) of phenylalanine (Phe), hexadecanoylcarnitine (C16), octadecenoylcarnitine (C18:1), valine (Val), and hexadecenoyl- and octadecanoylcarnitines in the form of C16:1 and C18, respectively, when the higher dominance of the first three metabolites results in a decrease of IC.3 and vice versa for the last three stated metabolites. The high loadings of the clr coefficients of these six metabolites imply that IC.3 reflects mostly (but not solely) the balance between subcompositions formed by Phe, C16, C18:1, and Val, C16:1, C18. The value of this balance was for the outlier significantly lower than that within the rest of the sample. After a deeper investigation of the outlying sample, it turned out that it belongs to a newborn suffering from Phenylketonuria, a metabolic disease which is typically followed by distinctly high absolute blood concentrations of phenylalanine. The measured value was $1014.7 \mu\text{mol/l}$, which significantly exceeds the upper norm value set on $120 \mu\text{mol/l}$ (van Wegberg et al. 2017) and which is represented with the respective high clr value 6.76. The levels of the remaining metabolites were comparable with the other samples, but particularly the atypical high dominance of Phe over all measured metabolites, which for the rest of samples ranged from 5.72 to 3.58 for their clr values, resulted in the high negative value of the third component, and therefore clear identification of this non-standard observation.

The next interesting feature is presented by IC.1. According to Fig. 2, the values of this component are not very homogeneous across the whole dataset and therefore some specific groups of neonates might be identified. A deeper graphical analysis of the first component (presented in Fig. 3) shows that for newborns with a birth weight smaller than 1500 grams, higher values of IC.1 are typical. The independent

Table 2 Chosen non-linearities g_i for each independent component computed with the adaptive deflation-based FastICA algorithm. Non-linearities are ordered according to kurtosis values of the corresponding ICs. In the original ordering, IC.44 was the last component, thus no non-linearity is given. See Table 1 for the definitions of the functions g_i

IC	g_i	IC	g_i	IC	g_i	IC	g_i	IC	g_i
IC.1	g_2	IC.11	g_2	IC.21	g_5	IC.31	g_1	IC.41	g_5
IC.2	g_2	IC.12	g_9	IC.22	g_6	IC.32	g_5	IC.42	g_4
IC.3	g_6	IC.13	g_9	IC.23	g_9	IC.33	g_{14}	IC.43	g_4
IC.4	g_2	IC.14	g_8	IC.24	g_1	IC.34	g_5	IC.44	–
IC.5	g_6	IC.15	g_6	IC.25	g_1	IC.35	g_5	IC.45	g_5
IC.6	g_8	IC.16	g_{10}	IC.26	g_4	IC.36	g_5	IC.46	g_3
IC.7	g_8	IC.17	g_5	IC.27	g_{14}	IC.37	g_{12}	IC.47	g_6
IC.8	g_8	IC.18	g_6	IC.28	g_5	IC.38	g_{14}		
IC.9	g_9	IC.19	g_{10}	IC.29	g_{10}	IC.39	g_4		
IC.10	g_8	IC.20	g_6	IC.30	g_8	IC.40	g_{11}		

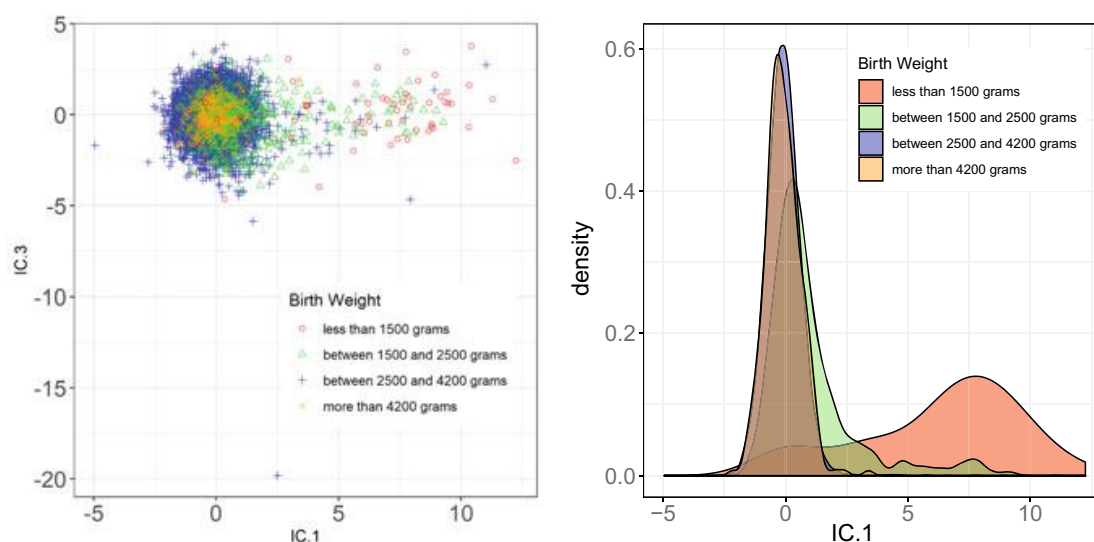


Fig. 3 Scatterplots of IC.1 and IC.3 (left) and the kernel density plot of IC.3 (right) with the groups defined according to the birth weight

component IC.1 is mostly formed by clr values of acylcarnitines dodecanoylcarnitine (C12), C16, and C18:1, whose high relative dominance over all measured metabolites results in low values of the component and, e.g. clr values of acylcarnitines isovalerylcarnitine/methylbutyrylcarnitine C5, and linoleoylcarnitine (C18:2) increase the IC.1 values. Even though there are also other metabolites contributing with a high weight to the values of IC.1 (all clr loadings are collected in Table 3), the clr values of the selected ones systematically differ for the group of the newborns with low birth weight, and therefore these acylcarnitines seem to be responsible for their separation from the remaining neonates. The differences in the selected metabolites are clearly visible in Fig. 4. Let us stress here that the immature neonates tend

to have different diet supplementation, therefore the metabolic profile can substantially differ within this group, but despite the proposed ICA method being able to find some similar patterns, detect the important metabolites, and separate the low birth weight newborns from the remaining ones. More specifically, artificial nutrition consists of amino acids, lipids, sugars, vitamins, etc. Essential unsaturated fatty acids including linoleic acid may be responsible for increased C18:2. The increased blood concentration of the long-chain acylcarnitines (C12, 16, C18:1) as well as of the short-chain C5 carnitine, which then results in high respective clr values, corresponds with previous studies. In Gucciardi et al. (2015), the significantly lower amounts of acylcarnitines except the branched-chain acylcarnitines (e.g. C5), which were significantly higher in preterm infants, were described. The latter mentioned are direct products of branched-chain amino acid (BCAA) catabolism, therefore its elevated levels may be related to BCAA overfeeding (Gucciardi et al. 2015; Wilson et al. 2014). The difference of several amino acids measured for the premature newborns compared to the others agrees with findings in Wilson et al. (2014), where increased levels of several amino acids (arginine, leucine, Orn, Phe, and Val) in the blood spots of premature infants were described. This observation may be related to the catabolic state of organisms in these children, amino acid supplementation, and immaturity of preterm infants (hepatic maturation, renal insufficiency, etc.) (Wilson et al. 2014; te Braake et al. 2005). The raw concentrations of valine (Val) and leucine/isoleucine (Xle) are known to be highly positively correlated, therefore the opposite signs of the respective loadings of IC.1 seem to be counter-intuitive at the first glance. However, the values of the loadings suggest that the resulting value of IC.1 is affected by the difference of clr values of the respective metabolites, or equivalently by the log-ratio of their measured concentrations, when the higher relative dominance of Val over Xle results in a higher value of IC.1. These findings agree with the data, since slightly higher values of the Val-Xle log-ratio are typical for newborns with a low birth weight (see Fig. 4). Finally, an even more complex interpretation can be based on the ilr loading matrix $\mathbf{W}_{\text{ilr}}$. According to the values of this matrix, IC.1 is mainly influenced by the balance between C18 and subcompositions C18:1, C18:OH, C18:2, and C18:2OH. This balance corresponds to the highest positive loading, and its values are systematically higher for the group of newborns with low birth weight than for the rest of the samples.

An even better visible pattern is formed by the last independent component IC.47, which clearly divides the whole dataset into two groups as seen in Fig. 5. According to the loadings (collected in Table 3), the most contributing are clr values of metabolites Xle, ornithine (Orn), and lysine (Lys) with a negative effect and methionine (Met), proline (Pro), and valine (Val) with a positive one. This suggests that the value of IC.47 is highly affected by the balance between subcompositions Met, Pro, Val and Xle, Orn, Lys. The dataset is roughly separated into two groups of observations with values of IC.47 higher and lower than -0.34 ; this value was chosen as the corresponding value of IC.47 at the local minimum in the middle of the density presented in Fig. 5 (this density was computed with Gaussian kernels and a bandwidth selection with Silverman's rule of thumb). The relative dominance of the six above-mentioned metabolites itself over all measured concentrations does not significantly differ in its

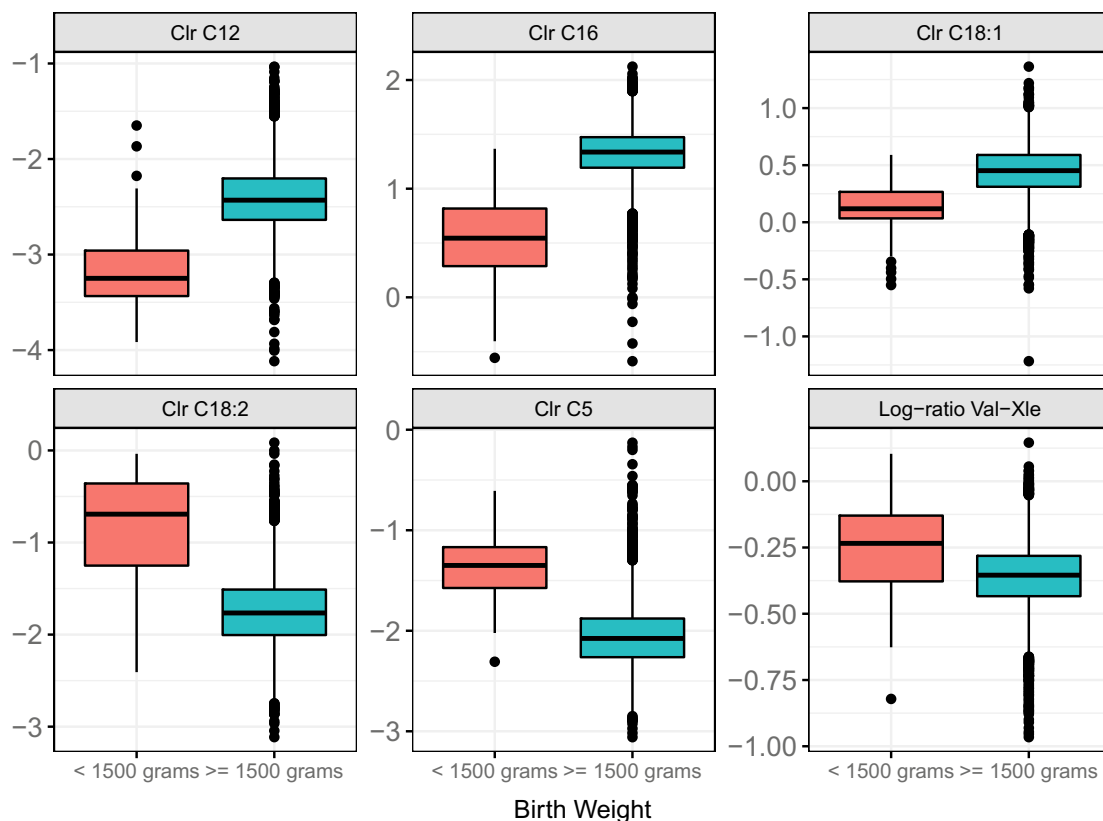
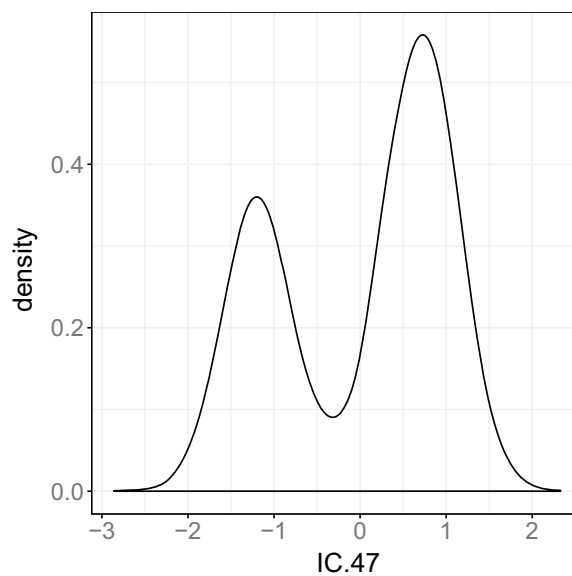


Fig. 4 Boxplots of clr as well as log-ratio values of the selected metabolites, which significantly differ for newborns with very low (< 1500 g) and normal (≥ 1500 g) weight at birth

Fig. 5 Density plot of IC.47, the bimodal shape shows a clear grouping



values between the two groups. Therefore, the grouping effect of IC.47 is hidden in some of their more complex combinations, e.g. the suggested balance between Met, Pro, Val and Xle, Orn, Lys, which is distinctly higher by cases with IC.47 higher than -0.34 .

Table 3 The list of loadings for IC.1, IC.3, and IC.47 computed with the adaptive deflation-based FastICA algorithm regarding clr transformed data

	IC.1	IC.3	IC.47		IC.1	IC.3	IC.47
Ala	0.02	0.24	0.41	C5DC/C6OH	-0.08	0.14	0.04
Arg	0.13	-0.33	0.04	C5:1	-0.05	0.11	0.13
ArgSucc	0.00	-0.09	-0.01	C6	-0.22	0.22	0.07
Cit	-0.05	0.38	0.01	C8	0.84	-0.05	0.21
Glu	-0.16	0.70	0.20	C8:1	-0.05	-0.16	0.09
Gly	0.08	1.33	-0.16	C10	0.25	-0.21	-0.09
His	-0.56	0.44	0.55	C10:1	-1.05	-0.01	-0.21
Lys	0.20	0.38	-0.46	C10:2	0.28	0.20	0.06
Met	0.53	1.03	1.01	C12	-1.27	-0.01	-0.13
Orn	0.48	0.40	-0.78	C12:1	-0.09	0.10	-0.09
Phe	1.02	-7.30	-0.40	C14	-0.39	0.45	-0.38
Pro	-1.15	-0.66	0.65	C14:1	0.88	-0.46	0.04
Thr	-0.17	-1.71	-0.01	C14:2	0.15	-0.01	0.17
Trp	-1.33	-0.77	0.14	C14OH	0.03	0.27	0.08
Tyr	-0.28	1.25	-0.02	C16	-2.00	-1.49	-0.02
Val	3.09	3.86	0.59	C16:1	0.97	1.60	0.09
Xle	-1.73	-0.15	-1.03	C16OH	0.24	0.09	0.01
C0	0.51	0.60	0.17	C16:1OH	0.05	0.01	0.24
C2	0.08	-0.30	-0.25	C18	2.38	1.72	0.12
C3	-0.57	-0.34	-0.01	C18:1	-3.46	-2.64	0.00
C3DC/C4OH	0.29	-0.23	0.34	C18:2	1.80	0.41	0.16
C4	0.05	0.14	0.01	C18:1OH	-0.00	0.27	-0.15
C4DC/C5OH	0.07	0.49	-1.42	C18:2OH	0.19	0.23	-0.05
C5	0.35	-0.32	-0.08	C18OH	-0.33	0.20	0.09

6 Discussion

In this paper, we reviewed some classical independent component analysis methods and showed how these can be applied to compositional data. The key finding here is that when the ICA methods are affine equivariant it is most natural to use an ilr transformation, as the choice of the basis constituting the ilr coordinate system does not matter. For interpretability, the link between ilr coordinates and clr coefficients/variables can be easily exploited, which allows interpreting the results either in terms of the dominance of single compositional parts with respect to the whole composition, or, e.g. based on values of balances between subcompositions formed according to values of clr loadings. Finally, since the clr loadings are derived from the ilr ones, it is also possible to provide the interpretation directly in terms of the ilr coordinates. The proposed technique is demonstrated on a metabolomics dataset

where PCA, which is probably the most used multivariate transformation, reveals no specific feature on the first few components while ICA reveals several interesting features visible when exploiting the higher order moments information. Independent component analysis belongs to the larger class of blind source separation methods where for the separation of the latent components often also temporal or spatial information is used. In the context of compositional data such blind source separation methods are, for example, discussed in Nordhausen et al. (2015), Nordhausen et al. (2020). But these methods would not be applicable to the metabolomics dataset from Sect. 5 as there is no temporal or spatial information present. The current results, which were discussed mostly in terms of the relative dominance of a single compositional part respective to the highest loading of an IC, open new challenges for further research. An alternative interpretation can be reached, e.g. by adaptation the approach based on principal balances (Pawlowsky-Glahn et al. 2011). However, the loadings of ICs are in general not orthonormal and therefore the principal balances approach is not as straightforward as in the case of PCA. Finally, an extension of the dataset with a group of blood samples collected from neonates with a diagnosed disease can further prove the usefulness of the method.

Acknowledgements The work of CM and KN was supported by the Austrian Science Fund (FWF) Grant no. P31881-N32. The work of KF was supported by The Czech Science Foundation Grant no. 19-07155S. The work of HJ and AG was supported by The Czech Science Foundation Grant no. 18-12204S, the grant of Internal Grant Agency, Palacký University Olomouc no. IGA_LF_2019_006 and grant of the Ministry of Health no. NU20-08-00367. We also thank V. Pawlowsky-Glahn and an anonymous referee for their comments which helped to improve this article.

References

- Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman & Hall.
- Cardoso, J.-F. (1989). Source separation using higher order moments. *International Conference on Acoustics, Speech, and Signal Processing, 1989*, 2109–2112.
- Cardoso, J.-F., & Soudoumiac, A. (1993). Blind beamforming for non-gaussian signals. *IEEE Proceedings F (Radar and Signal Processing)*, 140, 362–370.
- Clarkson, D. B. (1988). A least squares version of algorithm AS 211: The FG diagonalization algorithm. *Journal of the Royal Statistical Society C*, 37, 317–321.
- Comon, P., & Jutten, C. (2010). *Handbook of blind source separation: Independent component analysis and applications*. Amsterdam: Academic.
- Egozcue, J. J., & Pawlowsky-Glahn, V. (2005). Groups of parts and their balances. *Mathematical Geology*, 37, 795–828.
- Egozcue, J. J., & Pawlowsky-Glahn, V. (2019). Compositional data: The sample space and its structure. *Test*, 28, 599–638.
- Fačevicová, K., Bábek, O., Hron, K., & Kumpan, T. (2016). Element chemostratigraphy of the devonian/carboniferous boundary - a compositional approach. *Applied Geochemistry*, 75, 211–221.
- Filzmoser, P., Hron, K., & Reimann, C. (2009). Principal component analysis for compositional data with outliers. *Environmetrics*, 20, 621–632.
- Filzmoser, P., Hron, K., & Templ, M. (2018). *Applied compositional data analysis*. Cham: Springer.

- Fleischman, A., Thompson, J. D., & Glass, M. (2013). Systematic data collection to inform policy decisions: Integration of the region 4 stork (r4s) collaborative newborn screening database to improve ms/ms newborn screening in Washington State. In J. Zschocke, K. M. Gibson, G. Brown, E. Morava, & V. Peters (Eds.), *JIMD reports - case and research reports* (Vol. 13, pp. 15–21). Berlin: Springer.
- Gucciardi, A., Zaramella, P., Costa, I., Pirillo, P., Nardo, D., Naturale, M., et al. (2015). Analysis and interpretation of acylcarnitine profiles in dried blood spot and plasma of preterm and full-term newborns. *Pediatric Research*, 77(1–1), 36–47.
- Hron, K., Filzmoser, P., & Thompson, K. (2012). Linear regression with compositional explanatory variables. *Journal of Applied Statistics*, 39(5), 1115–1128.
- Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10, 626–634.
- Hyvärinen, A. (1999). Gaussian moments for noisy independent component analysis. *IEEE Signal Processing Letters*, 6, 145–147.
- Hyvärinen, A., & Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9, 1483–1492.
- Illner, K., Miettinen, J., Fuchs, C., Taskinen, S., Nordhausen, K., Oja, H., et al. (2015). Model selection using limiting distributions of second-order blind source separation algorithms. *Signal Processing*, 113, 95–103.
- Kalivodová, A., Hron, K., Filzmoser, P., Najdekr, L., Janečková, H., & Adam, T. (2018). PLS-DA for compositional data with application to metabolomics. *Journal of Chemometrics*, 29, 21–28.
- Miettinen, J., Nordhausen, K., Oja, H., & Taskinen, S. (2013). Fast equivariant JADE. In *IEEE international conference on acoustics, speech and signal processing (ICASSP) 2013* (pp. 6153–6157).
- Miettinen, J., Nordhausen, K., Oja, H., & Taskinen, S. (2014). Deflation-based FastICA with adaptive choices of nonlinearities. *IEEE Transactions on Signal Processing*, 62, 5716–5724.
- Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S., & Virta, J. (2017). The squared symmetric FastICA estimator. *Signal Processing*, 131, 402–411.
- Miettinen, J., Nordhausen, K., & Taskinen, S. (2017). Blind source separation based on joint diagonalization in R: The packages JADE and BSSasymp. *Journal of Statistical Software*, 76(2), 1–31.
- Miettinen, J., Nordhausen, K., & Taskinen, S. (2018). fICA: FastICA algorithms and their improved variants. *The R Journal*, 10, 148–158.
- Miettinen, J., Taskinen, S., Nordhausen, K., & Oja, H. (2015). Fourth moments and independent component analysis. *Statistical Science*, 30, 372–390.
- Morais, J., Thomas-Agnan, C., & Simioni, M. (2018). Interpretation of explanatory variables impacts in compositional regression models. *Austrian Journal of Statistics*, 47, 1–25.
- Nordhausen, K., Fischer, G., & Filzmoser, P. (2020). Blind source separation for compositional time series. *To appear in Mathematical Geosciences* (pp. 1–21).
- Nordhausen, K., Ilmonen, P., Mandal, A., Oja, H., Ollila, E.: Deflation-based FastICA reloaded. In *Proceedings of 19th European signal processing conference* (pp. 1854–1858).
- Nordhausen, K., & Oja, H. (2018) Independent component analysis: A statistical perspective. *WIREs: Computational Statistics*, 10, e1440.
- Nordhausen, K., Oja, H., Filzmoser, P., & Reimann, C. (2015). Blind source separation for spatial compositional data. *Mathematical Geosciences*, 47, 753–770.
- Nordhausen, K., & Virta, J. (2019). An overview of properties and extensions of FOBI. *Knowledge-Based Systems*, 173, 113–116.
- Ollila, E. (2010). The deflation-based FastICA estimator: Statistical analysis revisited. *IEEE Transactions on Signal Processing*, 58, 1527–1541.
- Pawlowsky-Glahn, V., & Buccianti, A. (Eds.). (2011). *Compositional Data Analysis, Theory and Applications*. Chichester: Wiley.

- Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R. (2011). Principal balances. In J.J. Egozcue, R. Tolosana-Delgado, & M.I. Ortego (Eds.), *Proceedings of the 4th International Workshop on Compositional Data Analysis*.
- Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (Eds.). (2015). *Modelling and analysis of compositional data*. Chichester: Wiley.
- R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Te Braake, F. W., van den Akker, C. H., Wattimena, D. J., Huijmans, J. G., van Goudoever, J. B. (2005). Amino acid administration to premature infants directly after birth. *The Journal of Pediatrics*, 147, 457–461.
- Templ, M., Hron, K., & Filzmoser, P. (2011). robCompositions: an R-package for robust statistical analysis of compositional data. In *Compositional data analysis: Theory and applications* (pp. 341–355). New York: Wiley.
- Trinh, H. T., Morais, J., Thomas-Agnan, C., & Simioni, M. (2019). Relations between socio-economic factors and nutritional diet in Vietnam from 2004 to 2014: New insights using compositional data analysis. *Statistical Methods in Medical Research*, 28, 2305–2325.
- Tyler, D., Critchley, F., Dümbgen, L., & Oja, H. (2009). Invariant coordinate selection. *Journal of Royal Statistical Society B*, 71, 549–592.
- van den Boogaart, K. G., Tolosana-Delgado, R., & Bren, M. (2019). *Compositions: Compositional data analysis*. R package version 1.40-3.
- van Wegberg, A., MacDonald, A., Ahring, K., Bélanger-Quintana, A., Blau, N., Bosch, A. M., et al. (2017). The complete European guidelines on phenylketonuria: Diagnosis and treatment. *Orphanet Journal of Rare Diseases*, 12, 162.
- Virta, J., Lietzen, N., Ilmonen, P., Nordhausen, K. (2020). Fast tensorial JADE. *To appear in Scandinavian Journal of Statistics*.
- Wei, T. (2015). A convergence and asymptotic analysis of the generalized symmetric FastICA algorithm. *IEEE Transactions on Signal Processing*, 63, 6445–6458.
- Wilson, K., Hawken, S., Ducharme, R., Potter, B. K., Little, J., Thébaud, B., et al. (2014). Metabolomics of prematurity: Analysis of patterns of amino acids, enzymes, and endocrine markers by categories of gestational age. *Pediatric Research*, 75, 367–373.

Robust blind source separation

Bábek, O., **Fačevicová, K.**, Žídek, M., Sedláček, J., Muehlmann, C., Nordhausen, K., Hron, K. X-ray fluorescence scanning of soft and wet-sediment cores in terrestrial environments; A robust blind source separation approach. *Journal of Geochemical Exploration*, 243, 107106 (2022).

<https://doi.org/10.1016/j.gexplo.2022.107106>

[IF 4.6 Q1; AIS 0.795 Q2]

[6 citations on WoS; 6 citations without autocitations]

[Author's contribution: 20%]

Abstract

X-ray fluorescence (XRF) scanning of split cores using automated cores scanners is rapid but expensive and semi-quantitative method in sedimentary geochemistry. While searching for a low-cost and quantitative solution in environmental geochemistry, we investigated the application of hand-held XRF device in scanning on two siliciclastic-sediment cores from dam-reservoirs in Slovakia. The core-scan data were compared with XRF data from dry, powdered sample aliquots, grain size, water content, and magnetic susceptibility data, which were used as independent variables. The aim of the manuscript is to investigate the applicability of time-series statistical methods to reveal the physical and chemical signals from hand-held XRF scanning of wet-sediment cores. The geochemical data were processed using the centred log-ratio (clr) technique, and then by the Second-Order Blind Identification (SOBI) method, which is a subtype of robust Blind Source Separation (BSS) techniques commonly used to enhance the signal-to-noise ratio in time series analysis of data in medicine or acoustics. So far, BSS has been rarely applied in geochemistry. Raw element concentrations (ppm/wt%) of core

scans and sediment powders are essentially incomparable, but their clr coefficients show statistically significant correlation for Al, P, K, Ca, Mn, Fe, Cu, As, Sr, Zr, and Pb, which reduces the difference between the two measurement methods and justifies previous attempts to calibrate the core-scanning data using log-ratio techniques. Scores of three of eight latent components (IC2, IC4, and IC7) found by SOBI in the core scan- and powder data show significant statistical correlation demonstrating the applicability of the core-scan technique. Their geochemical interpretation is based on the IC loadings by element clr, and the stratigraphic correlation with the sediment grain size and water content. These ICs explain geochemical variability due to grain-size contrast between sandy and silty layers, input of biogenic productivity-sensitive elements (Ca and P), eutrophication, and accumulation of anthropogenic elements. The robust BSS approach represents a promising method of processing of large datasets in environmental geochemistry.



Contents lists available at ScienceDirect

Journal of Geochemical Exploration

journal homepage: www.elsevier.com/locate/jgexplo

X-ray fluorescence scanning of soft and wet-sediment cores in terrestrial environments; A robust blind source separation approach

Ondřej Bábek^{a,*}, Kamila Fačevicová^b, Martin Židek^a, Jan Sedláček^a, Christoph Muehlmann^c, Klaus Nordhausen^d, Karel Hron^b^a Department of Geology, Palacký University of Olomouc, 17. listopadu 12, 77146 Olomouc, Czechia^b Department of Mathematical Analysis and Applications of Mathematics, Palacký University of Olomouc, 17. listopadu 12, 77146 Olomouc, Czechia^c Institute of Statistics & Mathematical Methods in Economics, Vienna University of Technology, Austria^d Department of Mathematics and Statistics, University of Jyväskylä, Finland

ARTICLE INFO

Keywords:

X-ray fluorescence
Core scanning
Dam-reservoir sediments
Geochemistry
Log-ratio analysis

ABSTRACT

X-ray fluorescence (XRF) scanning of split cores using automated cores scanners is rapid but expensive and semi-quantitative method in sedimentary geochemistry. While searching for a low-cost and quantitative solution in environmental geochemistry, we investigated the application of hand-held XRF device in scanning on two siliciclastic-sediment cores from dam-reservoirs in Slovakia. The core-scan data were compared with XRF data from dry, powdered sample aliquots, grain size, water content, and magnetic susceptibility data, which were used as independent variables. The aim of the manuscript is to investigate the applicability of time-series statistical methods to reveal the physical and chemical signals from hand-held XRF scanning of wet-sediment cores. The geochemical data were processed using the centred log-ratio (clr) technique, and then by the Second-Order Blind Identification (SOBI) method, which is a subtype of robust Blind Source Separation (BSS) techniques commonly used to enhance the signal-to-noise ratio in time series analysis of data in medicine or acoustics. So far, BSS has been rarely applied in geochemistry. Raw element concentrations (ppm/wt%) of core scans and sediment powders are essentially incomparable, but their clr coefficients show statistically significant correlation for Al, P, K, Ca, Mn, Fe, Cu, As, Sr, Zr, and Pb, which reduces the difference between the two measurement methods and justifies previous attempts to calibrate the core-scanning data using log-ratio techniques. Scores of three of eight latent components (IC2, IC4, and IC7) found by SOBI in the core scan- and powder data show significant statistical correlation demonstrating the applicability of the core-scan technique. Their geochemical interpretation is based on the IC loadings by element clr, and the stratigraphic correlation with the sediment grain size and water content. These ICs explain geochemical variability due to grain-size contrast between sandy and silty layers, input of biogenic productivity-sensitive elements (Ca and P), eutrophication, and accumulation of anthropogenic elements. The robust BSS approach represents a promising method of processing of large datasets in environmental geochemistry.

1. Introduction

Rapid geochemical analysis and statistical processing of large data sets are one of the current trends in geochemistry (Reimann et al., 2012; de Caritat and Grunsky, 2013; McKinley et al., 2016). The use of X-ray fluorescence core-scanning (XRF-CS) devices shorten analytical time to seconds and allow analysing of cores with sub-mm stratigraphic resolution (Jenkins, 1999; Rothwell and Croudace, 2015). XRF core scanners are now widely used for analysis of sediment provenance,

palaeoclimatology, tephrostratigraphy, environmental risk assessment, geochemistry of organic-rich sediments, and even biological tissues (Kelloway et al., 2014; Miller et al., 2015; Gadd et al., 2018; Peti et al., 2019; Hsiung et al., 2021). Perhaps of similar importance are hand-held XRF analysers, which can be used for direct scanning of elemental composition from surfaces of hard rock and soft sediment and soils in outcrops and split cores for anthropogenic impact assessment (Matys Grygar et al., 2014).

Although the detection limits of XRF-CS have significantly improved

* Corresponding author.

E-mail address: ondrej.babek@upol.cz (O. Bábek).<https://doi.org/10.1016/j.jgexplo.2022.107106>

Received 1 April 2022; Received in revised form 22 September 2022; Accepted 14 October 2022

Available online 19 October 2022

0375-6742/© 2022 Elsevier B.V. All rights reserved.

over time and may detect a few ppm for selected trace elements such as Rb and Sr. However, the XRF-CS and hand-held XRF scanning are strongly affected by surface geometry of the scanned rocks, their grain size, porosity, water content, and limited time for a single analysis (Croudace et al., 2006; MacLachlan et al., 2015; Weltje and Tjallingii, 2008). Given in counts per second the XRF-CS results are incomparable to conventional geochemical analysis such as ICP-MS and wet-chemical methods, which are represented by concentration units (wt%, ppm). Consequently, the XRF-CS is generally considered as a semi-quantitative method.

Attempts have been made to calibrate the XRF-CS results with independent analytical methods based on linear regression of raw XRF-CS data, element normalization to Si, Ca and coherent/incoherent X-ray scatter ratios (CIR), correction of XRF-CS data to water content, pairwise log-ratios, and multivariate log-ratio calibration (MLC) (Weltje and Tjallingii, 2008; Weltje et al., 2015; Gregory et al., 2019; Löwemark et al., 2019). In particular, the MLC technique using Partial Least Square (PLS) regression performed very well in predicting absolute concentrations of major and trace elements in scanned cores as well as comparing scan results from different labs (Weltje et al., 2015; Dunlea et al., 2019; Hennekam et al., 2019). The relationship between element concentration and grain size was investigated by Bloemsa et al. (2012) and Henares et al. (2019) who used the PLS regression to decompose the grain size, porosity and permeability effects from other factors contributing to the total geochemical variability. However, calibrations based on ICP-MS analysis of sample aliquots are hampered by different scaling between the high-resolution scanning (typically mm scale across stratigraphy) and a minimum volume of sample required for conventional chemical analysis although XRF scanning of discrete samples may provide a solution to the scaling difference (Gregory et al., 2019; Profe and Ohlendorf, 2019). Data from the hand-held XRF devices are typically given in concentration units (ppm), but the negative effects of grain size, porosity, and water/air content during direct scanning are not eliminated.

Use of XRF-CS and hand-held XRF devices in the analysis of trace elements Co, Ni, Cu, Zn, As, Cd, Hg and Pb, which are important for environmental risk assessment, is thus limited to semi-quantitative monitoring of contamination (Rodríguez-Germade et al., 2015). Although external factors such as porosity or water content can affect the measurements obtained by these devices, it allows for a rapid and cost-effective measurement of large datasets. Combined with a proper data processing, this makes these tools a promising alternative to the traditional geochemical methods. In general, given the large volume of data and the importance of relative changes in element intensities, the XRF scanning provides a valuable source of data for unconventional statistical methods in geochemistry, e.g. the compositional data analysis (CoDA) (Pawlowsky-Glahn and Buccianti, 2011). In most of the CoDA methods, the geochemical information (single element concentrations) is first extracted using element (log-)ratios and then processed by multivariate statistical analysis (Weltje et al., 2015; Fačevićová et al., 2016; Henares et al., 2019). With the hand-held XRF measurements in cores, two additional features have to be respected. Due to the sample inhomogeneity and sensitivity to grain size, non-typical (outlying) observations are likely to appear in the dataset, which can be suppressed by robust statistical approach (Matys Grygar et al., 2020). Second, the depth-distribution of samples across a sediment core generally follows the principles of time-series analysis and we may assume that samples located close to each other tend to be more similar than those occurring at greater distance are. Therefore, our aim is to find latent components reflecting the common physical processes controlling sediment geochemistry such as lithology, provenance, grain size, diagenesis, and anthropogenic contamination (Sageman and Lyons, 2003; Vijver et al., 2008; Grosbois et al., 2012; von Eynatten et al., 2016; Lacey et al., 2017) as well as the traits inherent to the measurement methods using a family of blind source separation (BSS) techniques. The BSS is a well-established multivariate statistical tool that builds on a linear latent

variable model. We assume that the observed data are controlled by linear combinations of unknown (latent) components, which are uncorrelated or even independent. Based on the principles of signal processing the aims of the BSS are to find physical signals of the measured (time series) data, for example brain signals based on fMRI measurements (Comon and Jutten, 2010; Pan et al., 2021). The robust BSS is also widely used in acoustic noise reduction and speech quality enhancement (Mukai et al., 2003; Zhang and Zhao, 2013). However, applications of this method in Earth Sciences are still rare and often not taking into account the relationships between adjacent measurements, which, in our perspective, is one of the key properties of the data (Nordhausen et al., 2015, 2021).

In this study, we focus on XRF scanning of discrete samples from split cores of soft, wet sediments retrieved from dam reservoirs, Váh River, Slovakia, potentially impacted by anthropogenic contamination by toxic metals, and statistical processing of the data by CoDA and robust BSS. Our aim is to investigate whether this rapid and cost-effective but alternative approach to geochemical analysis is capable of providing reasonable interpretations of near-surface Earth processes that can be imprinted in the sediments (research question, RQ 1). The scanned cores were subjected to subsequent EDXRF analysis, grain-size analysis, and porosity measurements in order to evaluate the effects of primary signals inherent to the sediment composition, texture, and anthropogenic contamination, and secondary signals inherent to the XRF-CS method. Our aim is to separate primary geogenic signals from measurement noise using the robust BSS method and investigate whether and under what conditions this novel method can be applied to sedimentary geochemistry (RQ 2).

2. Geographic and geological settings

Váh River, a left-hand tributary of the Danube River, is the longest river in Slovakia, which drains a predominantly mountainous catchment with very diverse geology spreading from Slovakia to Poland and Czechia. The Váh river basin has an area of 19,660,977 km². The altitude difference between the river origin (664 m above sea level) and its outflow to the River Danube (102 m a.s.l.) is 562 m. The river flows through a very rugged relief of the Western Carpathians reaching elevations of almost 2500 m above sea level, and through lowlands of the West Pannonian Basin. The river valleys are relatively narrow. Geology of the catchment comprises Cenozoic siliciclastic sedimentary rocks of the Flysch Zone of Outer Western Carpathians, metamorphic and igneous rocks of the crystalline basement of Central Western Carpathians with its upper Palaeozoic and Mesozoic sedimentary cover (mainly carbonates), and largely unconsolidated, Neogene siliciclastic sediments of the Pannonian basin and its northern promontories.

There are 22 dams constructed along the 406 km long course of the Váh River (one dam per 18.5 km of river course on average) known as the Váh Cascade (Fig. 1). Owing to the frequency of dam reservoirs and the very diverse geology, the Váh River provides a good opportunity to study the effects of damming on the transport, accumulation, architecture, and composition of the reservoir sediments. Sediments found in the reservoirs are commonly a mix of soil particles and rock fragments from the catchment transported by the river and deposited from atmospheric sources. This paper addresses two sediment cores from two reservoirs, the Hričov reservoir completed in 1962 with dam height 19 m, upstream length 6 km, water storage capacity 8.47 million m³, maximum depth 8 m, and dam height 19 m, and the Nosice reservoir, completed in 1957 with dam height 21 m, water storage capacity 35.9 million m³, and maximum depth of 18 m.

3. Material and methods

3.1. Physical properties and geochemistry of analysed sediments

Two underwater sediment cores with 6.3 cm inner diameter were

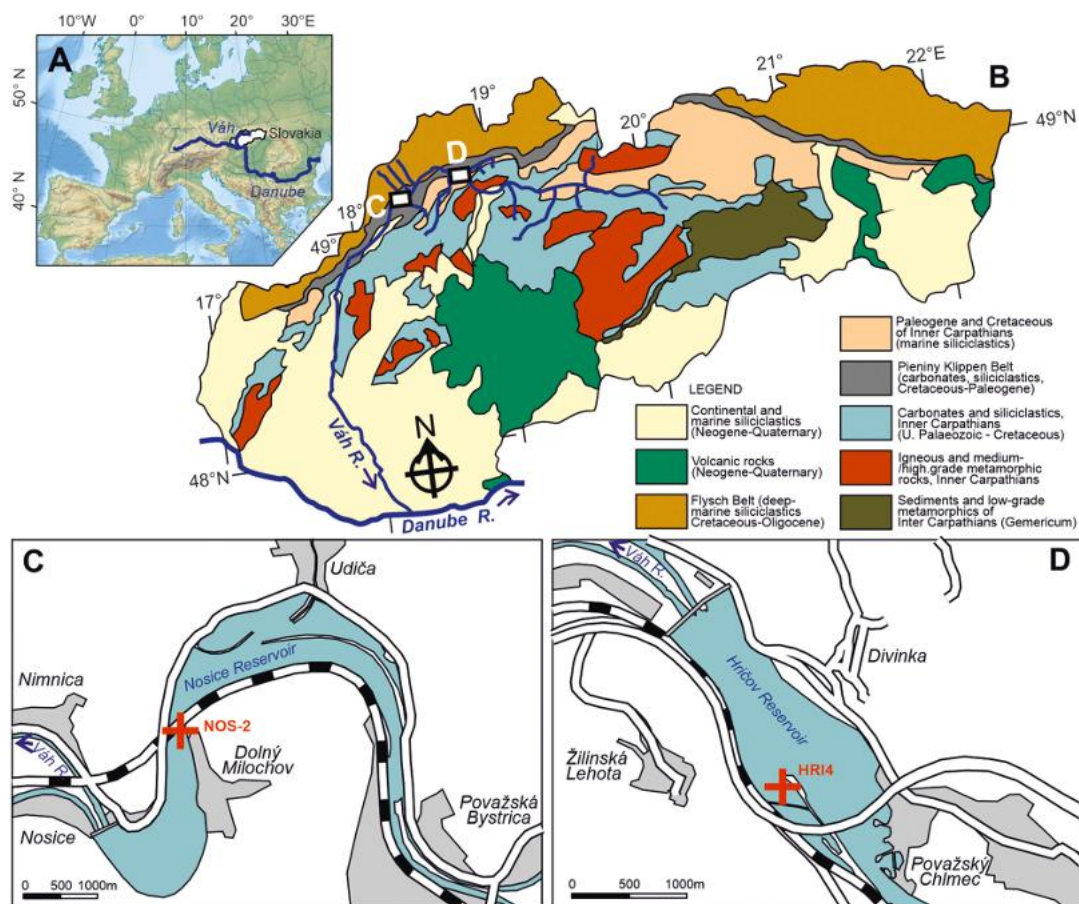


Fig. 1. Map of the study area: (A) Geographic context of Europe; (B) Geology of Slovakia with Váh river course and location of the Nosice and Hričov dam reservoirs; (C,D) location of the drill cores in the Hričov and Nosice reservoirs.

retrieved from Nosice (NOS 2, water depth 0.6 m) and Hričov (HRI 4, water depth 1.8 m) dam reservoirs located on the Váh River, Slovakia, using UWITEC gravity corer sampler (Uwitec sampling equipment, Austria) with hydraulic core catcher. The cores were split lengthwise, photographed, covered by a Mylar foil, and scanned while still wet by a Delta (Innov-X, USA) portable spectrometer with 1 cm stratigraphic interval (XRF core scan data). The spectrometer with 25 mm² circular analysis spot size was operating in Geochem mode (analytical beam operating at 1 to 40 kV, and 0.1 to 4 mA for V, Cr, Fe, Co, Ni, Cu, Zn, W, Hg, As, Se, Au, Pb, Bi, Rb, U, Sr, Y, Zr, Th, Mo, Ag, Cd, Sn, and Sb; beam operating at 2 to 10 kV, and 0.4 to 2 mA for Mg, Al, Si, P, S, Cl, K, Ca, Ti, and Mn). The assay time was set to 240 s. The assay results from the Delta portable spectrometer are given in concentration units (ppm). A simple device with fixed scaled core holder and manually operated sliding holder for the XRF spectrometer attached to a computer was specially developed for the core scanning (Fig. 2).

After the scanning, the scanned half-core was sliced to produce sample aliquots corresponding to the scanning intervals. A smaller part of each aliquot, while still wet, was subjected to grain-size analysis by wet-suspension laser grain-size analysis using a Fritsch Analysette 22 MicroTec Plus laser particle sizer (0.08–2000 µm) with an attached ultrasound dispersion unit. The remaining part was weighted, dried in a lab oven at 40 °C temperature for 48 h, and then weighted again. Water content was determined as the difference between the wet and dry mass in the NOS-2 core. Water content was not measured in the HRI-4 core,



Fig. 2. Manual X-ray fluorescence scanning of a wet core covered by Mylar foil using Delta (Innov-X, USA) portable EDXRF spectrometer attached to a specially designed core holder.

which was drilled earlier, and wet samples are no more available. However, we use the XRF, MS, and grain-size results from HRI-4 core as reference material for the NOS-2 core.

The dry samples were powdered to analytical fineness (< 63 µm), put

into plastic containers sealed by Mylar foil, and their elemental composition analysed by a Delta Premium (Innov-X, U.S.A.) EDXRF spectrometer with a large-area silicon-drift detector, with 2×120 s acquisition time (beam) at accelerating voltages of 15 kV and 40 kV (XRF powder data). Fifteen elements (Al, Si, P, K, Ca, Ti, Mn, Fe, Cu, Zn, As, Rb, Sr, Zr, and Pb) and 188 samples were analysed in total. Thirty-one samples (each pooled from three neighbouring EDXRF samples) from the HRI-4 core were analysed by inductively coupled optical emission- (ICP-OES) and mass spectrometry (ICP-MS) after total digestion in Aqua Regia at Bureau Veritas Minerals, Acme Labs, Vancouver, Canada. Eleven oxides of major elements (SiO_2 , Al_2O_3 , Fe_2O_3 , MgO , CaO , Na_2O , K_2O , TiO_2 , P_2O_5 , MnO , and Cr_2O_3) and 47 elements were analysed. Analytical precision and accuracy of powdered EDXRF analyses and ICP-OES/MS analyses from dam reservoir sediments, using the same instrumentation, are described in our previous papers (Bábek et al., 2021). Magnetic susceptibility (MS) of the powdered samples was measured using KLY-2 Kappabridge (Agico, Czech Republic) at a magnetic field intensity of $300 \text{ A}\cdot\text{m}^{-1}$, 920 Hz operating frequency, and a sensitivity of $4 \times 10^{-8} \text{ SI}$. The bulk MS (SI) was converted to mass-specific MS ($\text{m}^3\cdot\text{kg}^{-1}$).

3.2. Statistical analysis

The measured concentrations represent a set of positive D -dimensional multivariate observations. Since the nature of data is relative, element concentrations from adjacent depths are dependent, and the presence of outliers is expectable due to the measurement method, principles of robust statistics, time series, and compositional data analysis will be combined.

In order to respect the relative nature of measured element concentrations, and to follow the common practice in geochemistry, the raw element concentrations will be replaced with two coordinate systems. The core of the analysis runs on $D-1$ pivot coordinates (where D is the number of measured elements), a specific type of orthonormal log-ratio (olr, also abbreviated as ilr – isometric log-ratio) coordinates whose advantageous geometrical structure allows us to perform any statistical method (including the robust ones) without the need of its further adjustment (Fišerová and Hron, 2011; Matys Grygar et al., 2020). A more favourable interpretation of relative dominance of the given element within the whole composition is provided by centred log-ratio coefficients (clr), which will be therefore used for the final presentation of results. Note that the vectors of olr coordinates $\mathbf{x}^{\text{olr}}$ and clr coordinates $\mathbf{x}^{\text{clr}}$ are interconnected though a log-contrast matrix $\mathbf{V}$ by $\mathbf{x}^{\text{olr}} = \mathbf{V}' \mathbf{x}^{\text{clr}}$, see Pawłowsky-Glahn and Buccianti (2011), Ch. 11 for details.

As a usual data pre-processing step, Principal Component Analysis (PCA) is used in order to reduce the dimension of the problem and filter out the noise. More specifically, robust PCA, utilizing the deterministic Minimum Covariance Determinant (MCD) estimates of location and scatter, will be applied (Hubert et al., 2012) on pivot coordinates and first k components explaining 99 % of the total variability will be kept for the further analysis.

The core of the statistical analysis is based on the time series BSS methodology, when the measurements from different equidistant depths are treated as equidistant points in time. Following the methodology described by Ilmonen et al. (2015) the measurements from a given depth s ($\mathbf{x}_s$) are assumed to be linked to latent components $\mathbf{z}_s$ through the mixing matrix $\mathbf{A}$ and the theoretical model $\mathbf{x}_s = \mathbf{A} \mathbf{z}_s$. The aim of BSS is to recover the latent components by $\mathbf{W} \mathbf{x}_s$ given only $\mathbf{x}_s$, where $\mathbf{W}$ is an unknown unmixing matrix (e.g., $\mathbf{A}^{-1}$).

Note that the latent components representing the real geochemical signal or a noise inherent in the measurement method are found as a linear combination of $\mathbf{x}_s$ and the interpretation of the result is therefore similar to interpretation of scores and loadings resulting from PCA. The latent components are independent of prior linear transformations, which makes the methodology especially appealing in conjunction with the olr coordinates present in compositional data analysis, this is

detailed for example by Muehlmann et al. (2021) and Nordhausen et al. (2021). The basic BSS method is the Second-Order Blind Identification (SOBI) method (Belouchrani et al., 1997; Miettinen et al., 2016), which finds latent components maximizing the second-order serial dependence. Its robust version eSAM-SOBI, working with robust estimates of location and auto-covariance matrix (Ilmonen et al., 2015), will be used in our application. Finally, the unmixing matrix will be back-transformed to the clr space. The analysis is carried out in R 4.1.0 (R Core Team, 2021) with the help of the packages JADE (Miettinen et al., 2017), robCompositions (Templ et al., 2011) and robustbase (Maechler et al., 2021).

4. Results

4.1. Grain size, water content and magnetic susceptibility

Majority of the sediments from the NO2 core (Fig. 3) are composed of unimodally distributed silt (~ 83 to ~ 90 %) with minor admixture of clay (~ 5 to ~ 9 %) and very fine- to fine sand (~ 2 to ~ 12 %). Median grain size varies from ~ 17 to $\sim 30 \mu\text{m}$. These background sediments alternate with ~ 4 to ~ 9 cm thick sandy beds composed of very fine- to coarse sand (~ 21 to ~ 66 %), silt (~ 29 to ~ 74 %) and clay (~ 4 %) with median grain size varying from ~ 32 to $\sim 85 \mu\text{m}$, and unimodal or polymodal distribution. The HRI-4 core is composed of very uniform silty sediment with ~ 9 to ~ 20 wt% of clay fraction and very little admixture of very fine sand (up to 0.5 wt%), and no sandy beds.

The sample weights in the NOS-2 core varied from ~ 13 to ~ 35 g of wet sediment, and from ~ 5 to ~ 22 g of dry sediment. The calculated water content varies from ~ 25 to ~ 61 wt% (Fig. 3). Generally, the water content decreases from the top to the bottom of the core. The silty background sediments have higher water content from ~ 40 to ~ 55 wt% while the sandy beds have lower water content, typically from ~ 25 to ~ 45 wt% (Supplementary material, Table S1).

Mass-specific MS values vary from ~ 2.0 to $\sim 5.6 \times 10^{-6} \text{ m}^3\cdot\text{kg}^{-1}$. The MS distribution in the HRI-4 core is relatively uniform with a change from ~ 4 to $\sim 3 \times 10^{-6} \text{ m}^3\cdot\text{kg}^{-1}$ at about 70 cm depth. The MS distribution in the NOS-2 core is also relatively uniform with the exception of a prominent peak at 21 cm depth (Fig. 3).

4.2. Element geochemistry of raw data

The ICP-MS element concentrations and XRF data from cores scans and powdered samples are shown in Tables S2, and S3, respectively (Supplementary material). The XRF core scan data show systematically lower concentrations than the XRF powder data. The Al concentrations vary between ~ 0.26 and ~ 1.6 wt% in the core scans, and between ~ 1.62 and ~ 4.82 wt% in the powder data. The other elements show similar difference (Figs. 4, 5). The average concentrations of XRF scan data from the HRI-4 core reach about 16 to 41 % of their powder-sample counterparts.

The XRF powder concentrations of Al and Rb, which are frequently considered lithogenic, grain-size sensitive elements, show statistically significant correlation (logarithmic regression) with the grain-size data (concentration of clay fraction). However, for XRF core scan data, no correlation or much less significant correlation was found (Fig. 4). Several elements (K, Rb, and Ti) show good covariance with the lithogenic Al in both the XRF core scan and XRF powder analyses. Their correlation coefficients (R^2) vary from ~ 0.3 to ~ 1.0 but usually they are higher than 0.8 (Fig. 5). On the other hand, Ca and P show negative correlation with Al in the powder data, but generally, no correlation in the core scans (Fig. 5). Arsenic, zinc, copper and lead show weaker correlation with Al; their correlation coefficients vary between ~ 0.2 and ~ 0.7 . The lower concentrations and higher correlations with Al in the XRF scan data are likely controlled by the presence and dilution effect, respectively, of pore water and organic matter in the sediments. Organic matter was not analysed in the studied cores, but it is considered low.

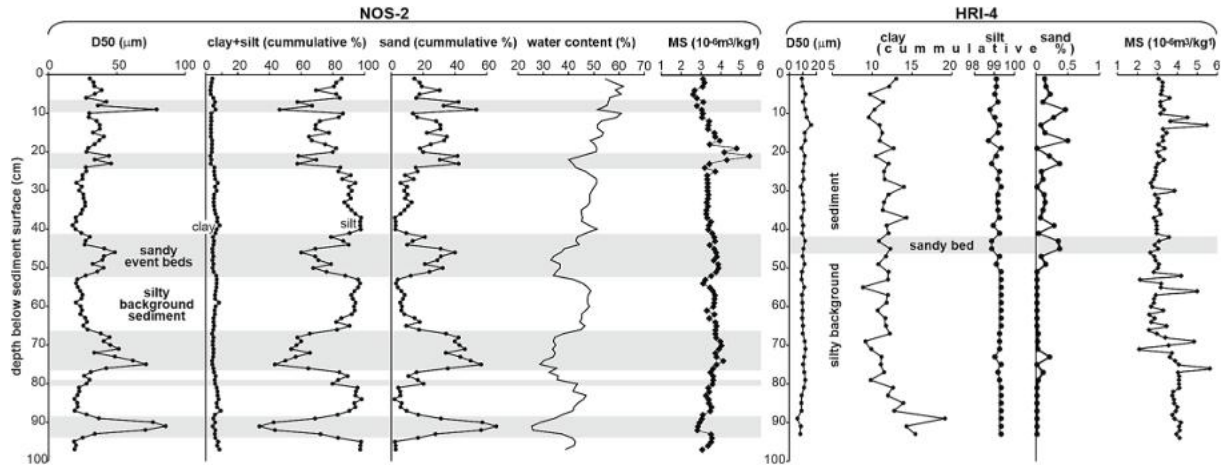


Fig. 3. Grain-size and mass-specific magnetic susceptibility (MS) data for NOS-2 and HRI-4 core; note distribution of less porous, sandy event layers (event beds) on the background of silty sediments with higher water content (porosity) in the NOS-2 core.

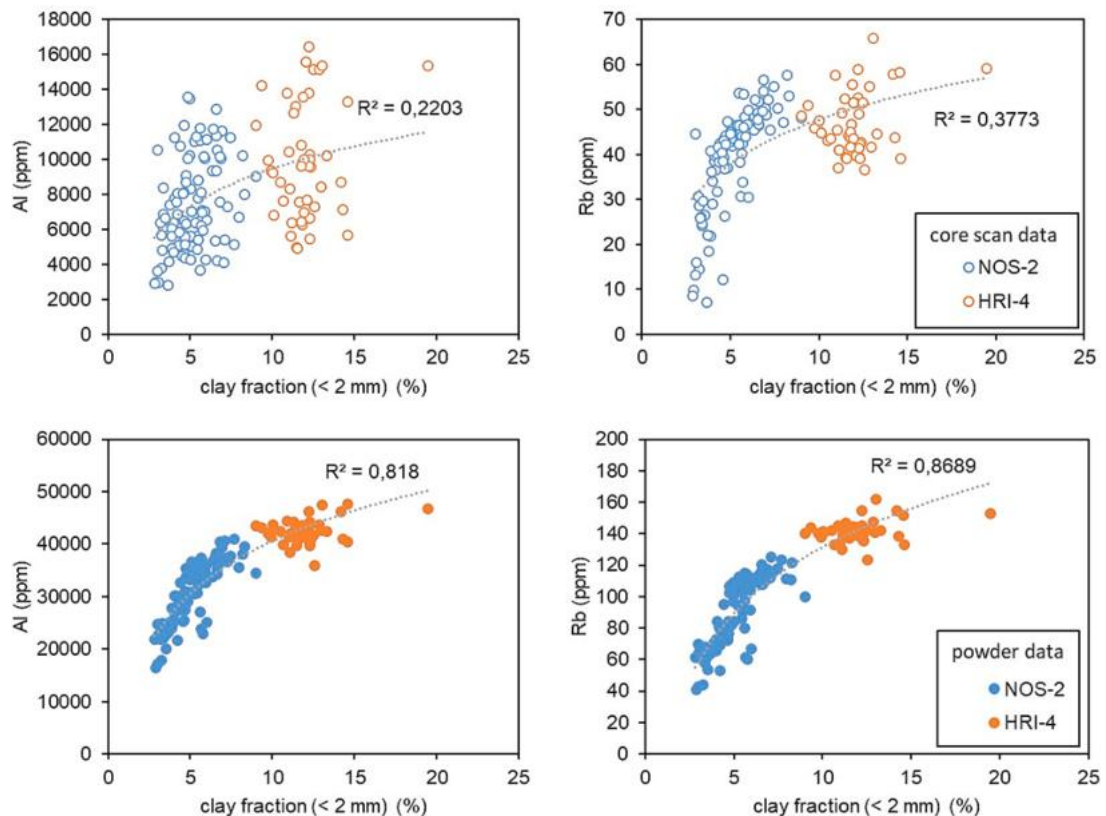


Fig. 4. Bivariate plots of XRF concentrations of Al and Rb (raw data) vs. grain size (concentration of clay fraction). Note very good (logarithmic) correlation of powder XRF data, and lack of correlation of core scan data.

Another core from the Nosice reservoir located about 1000 m S of the NOS-2 core contained 1.9 wt% on average (from 1.2 to 2.5 wt%, $n = 27$) of total organic carbon (TOC) (Ondřej Bábek, unpublished data).

Concentrations of the same elements from XRF core scan and powder data show a poor match; linear regression coefficients (R^2) of four

elements, Ca, As, Zr, and Rb, vary from 0.30 (Ca) to 0.81 (Zr) (Fig. 6). Lower correlation coefficients were found for Al, Si, P, K, Ti, Fe, Mn, Cu, Zn, Sr, and Pb. Conformable with the above-mentioned observations, vertical distribution of elements generally shows a little match between XRF core scan- and powder data (Fig. 7). In general, the powder data (Al,

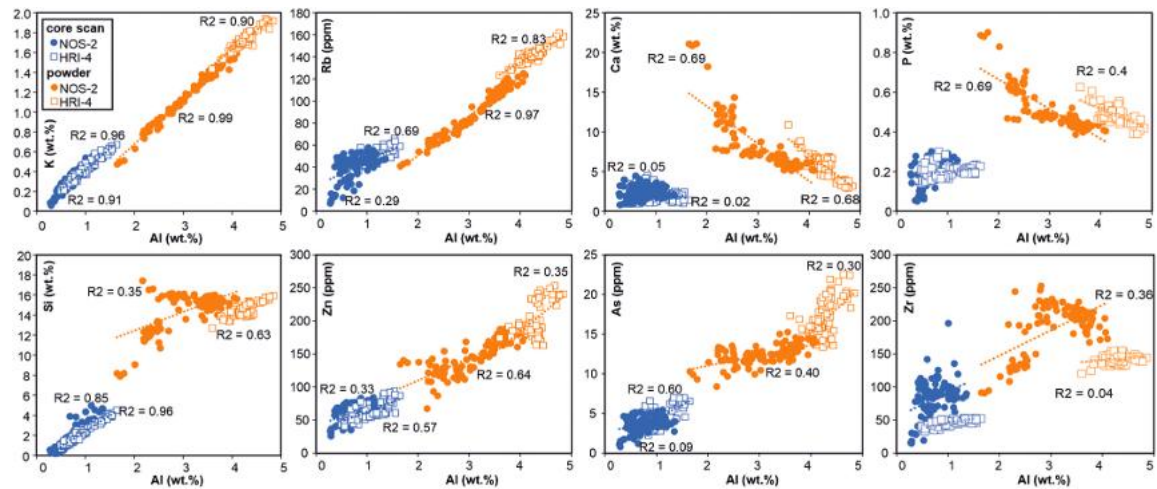


Fig. 5. Patterns of covariance of selected elements with Al in XRF core scan data (blue dots) and XRF powder data (orange dots) in the HRI-4 and NOS-2 cores. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

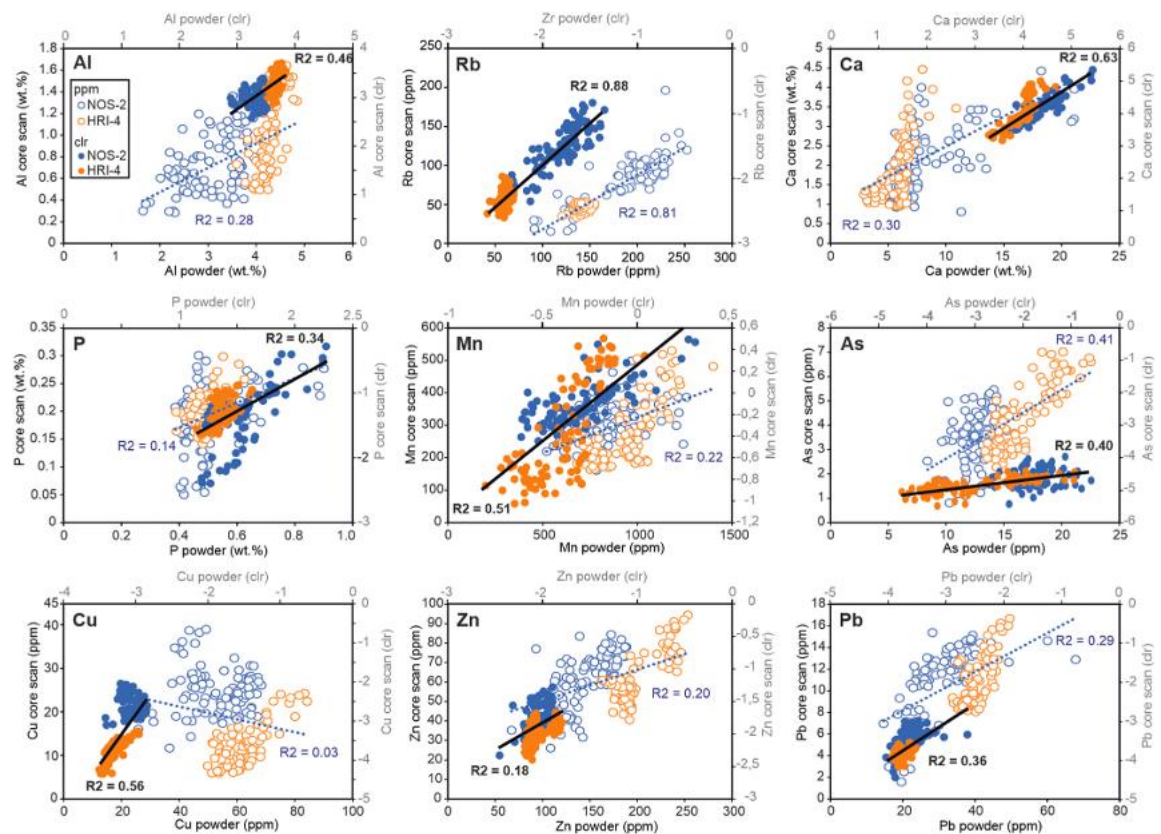


Fig. 6. Bivariate plots and linear-regression correlation coefficients of EDXRF of powdered samples (horizontal axes) with EDXRF core scans (vertical axes) in selected elements from the NOS-2 and HRI-4 cores. Data are represented as concentration units (ppm, blank circles) and their clr-transformed counterparts (full circles).

Ca, Zn, Fig. 7, but also K, Ti, Fe, Cu, Rb, and Pb) show higher variability due to changing grain size and water content than the core scan data in both cores. While for example, Al, K, Ti, Fe, and Rb have higher

concentrations in the silty layers, Ca and P have higher concentrations in the sandy layers. The powder data in the uniform silty-clayey sediments of the HRI-4 core are markedly uniform, with relatively high Al and K

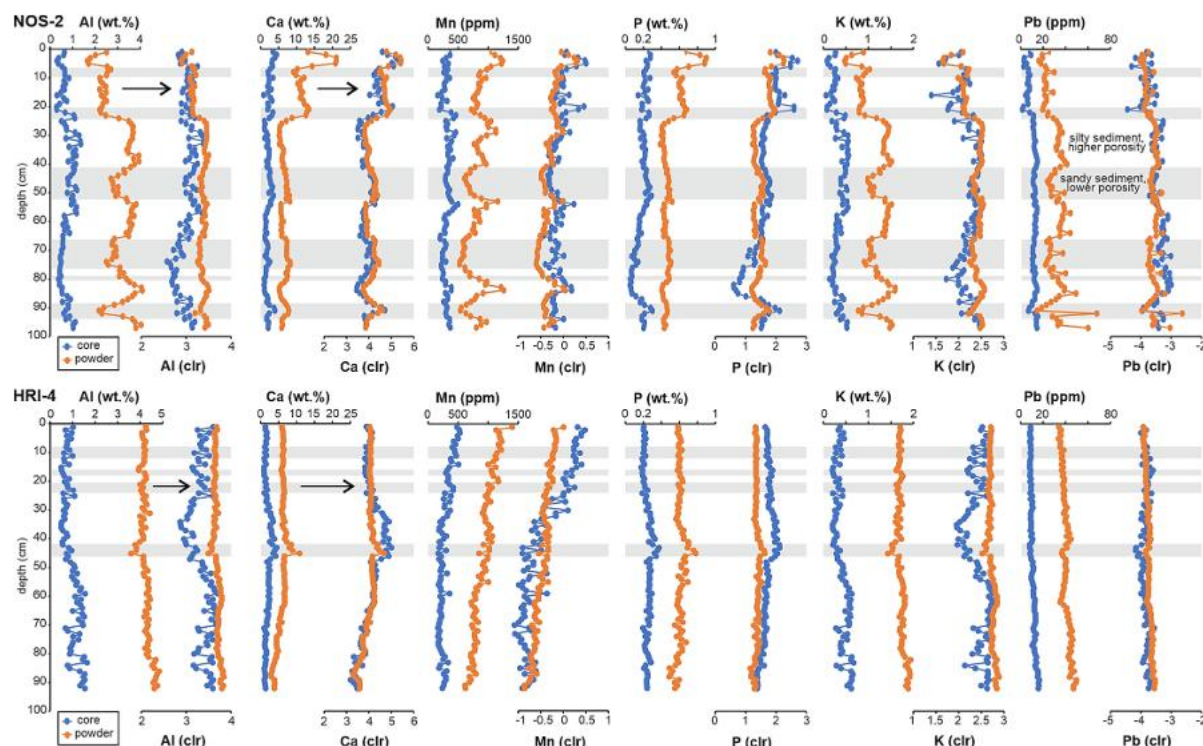


Fig. 7. Effects of clr coordinate representation on vertical distribution of selected elements in the NOS-2 core for XRF core scan data (blue dots) and XRF powder data (orange dots). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

concentrations and relatively low Ca concentrations. The core scan data are also variable but their variability does not coincide with the changes in grain size in both cores and the water content in the NOS-2 core (Fig. 7).

4.3. Log-ratio transformation of geochemical data

Centred log-ratio (clr) transformation of the same elements from the core scan and powder data resulted in higher correlations (R^2 from 0.33 to 0.88) in eleven elements: Al(clr), P(clr), K(clr), Ca(clr), Mn(clr), Fe(clr), Cu(clr), As(clr), Sr(clr), Zr(clr), and Pb(clr). Correlation of Ca(clr), Mn(clr), Cu(clr), Sr(clr), and Zr(clr) exhibit R^2 values higher than 0.5.

The clr coordinate representation partly reduces the variability due to lithology of the powder data, and reduces the difference between the powder and core-scan data (Figs. 4, 6). The median clr concentrations of core-scan and powder data are 3.04 and 3.63 for Al, 2.23 and 2.38 for K, 3.97 and 4.2 for Ca, and -4.59 and -4.48 for As, respectively. Although the clr values are much more similar for core-scan and powder data, the resulting shapes of the depth distribution curves are poorly comparable in none of the elements except for calcium (Fig. 7) and partly for K, Ti, Zr, and Mn. In case of Ca, the clr coefficients are a suitable data representation for calibration of XRF core-scan data.

4.4. Robust BSS results

Following the methodology described in Section 3.2, the XRF powder data were pre-processed with robust PCA based on the MCD estimate of location and covariance. The proportion of observations used for the covariance matrix estimation was set to 0.8, which represents a reasonable compromise between stability and usability of the whole algorithm. We run the eSAM-SOBI algorithm on the first eight robust principal components representing 99 % of total variability, which

resulted in finding of eight latent components (ICs). The lag parameter was set to five and therefore the measurements in the range of 5 cm were considered as mutually influential. The latent components are shown in Fig. 8. The orange dots represent the latent components (IC) for powder XRF data found by the algorithm, while the blue dots show the XRF core scan measurements expressed in the same latent variables.

The IC1, and IC3 – IC5 latent components of XRF powder data are uniform and most of them are independent on the grain size and porosity (Fig. 8). The IC2 shows variation, which coincides with the grain-size changes and the associated pore water variations. The IC4 reveals a trend towards higher values in the top 25 cm of the NOS-2 core. The latent components of XRF core-scan data (IC1, IC3, and IC5 to IC8) are highly variable but their variation is not related to the lithological boundaries. The IC3 with IC5 and IC1 with IC8 show similar depth distributions. The correlation between IC scores for powder and core-scan data is high for IC4 ($r = 0.786$), lower but significant for IC2 (0.465) and IC7 (0.383) and low or even negative for IC1 (-0.286), IC3 (0.075), IC5 (-0.01), IC6 (0.282) and IC8 (-0.045). The former three ICs may therefore carry information on real physical or chemical signals while the later ones are likely affected by the method of measurement. The vertical distribution of IC2 and IC4 shows a high degree of similarity between the XRF powder and core scan data. In contrast, the IC1, IC3, IC5, and IC8 show large-scale differences, while the IC6 and IC7 contain parts that are similar and parts that are highly dissimilar. The later three components clearly show that particularly the concentrations between 60 and 90 cm depth are influenced by the method of measurement. The structure of the latent components is explained by their respective clr loadings, which help to identify the main elemental sources for each IC (Table 1).

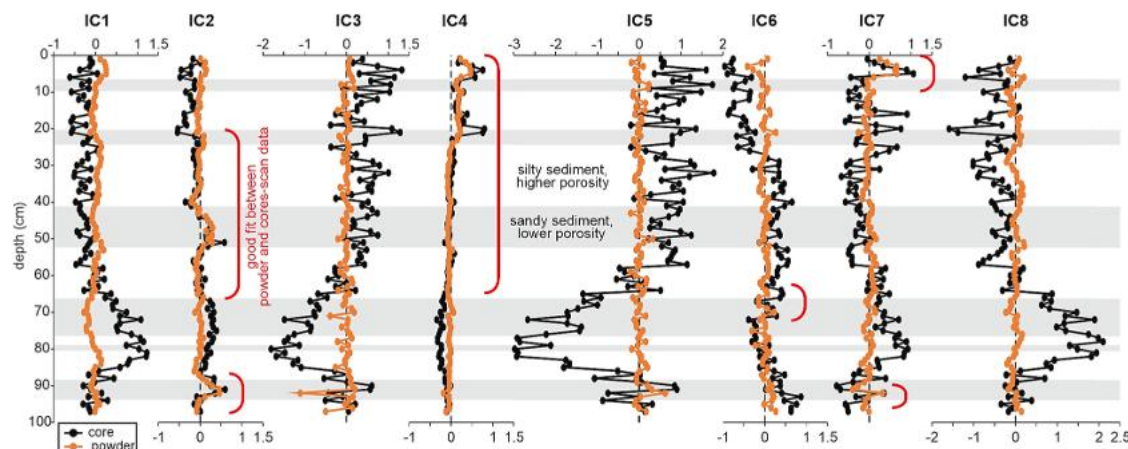


Fig. 8. Selected latent components (IC) found by the robust BSS algorithm applied to the NOS-2 XRF core-scan data (blue dots) and powder XRF data (orange dots). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5. Discussion

5.1. Interpretation of raw data: lithogenic sources of geochemical variation

Element compositions of the HRI-4 and NOS-2 cores differ. The main cause of the geochemical difference is the contrasting grain size, which is demonstrated by the significant correlation between the Al and Rb concentrations in XRF powder data and the proportion of the clay fractions (Fig. 4).

The Al-normalized XRF powder data (Fig. 5), which reduce the effects of different grain-size in the both cores (cf. Sageman and Lyons, 2003; Bábek et al., 2015), indicate that sediments in the NOS-2 core are enriched in Si, and Zr while the Hričov reservoir is enriched in Ca, P, Rb and As. Silicon and zirconium are typically lithogenic elements bound to coarse-grained (silt- and sand-sized) grains, whereas Ca and P are largely bound to authigenic carbonate rocks (Sageman and Lyons, 2003; Ross and Bustin, 2009; Bábek et al., 2016).

This compositional variability is presumably due to different lithogenic sources in their upstream catchments. Catchments of tributaries located upstream of the Hričov reservoir drain crystalline rocks and Mesozoic carbonates of the Central Carpathians (Varínka R.), Palaeogene deep-marine siliciclastics of the Flysch Zone of Outer Western Carpathians (Kysuca River), and Neogene unconsolidated siliciclastics of intramontane basins (Turiec R.). In contrast, tributaries located downstream of the Hričov and upstream of the Nosice reservoir, drain predominantly the Palaeogene deep-water siliciclastics of the Outer and Central Western Carpathians (Štiavnik River, Papradnianska R., Marikovský potok R., Domanizanka R.). The element geochemistry suggests that reservoirs provide barriers to downstream sediment dispersal as suggested by previous authors (Vörösmarty et al., 2003; Arnaud et al., 2019; Vauclin et al., 2019).

Magnetic susceptibility has been demonstrated to correlate with grain size, industrial contamination by fly ashes, and enrichment in selected heavy metals in fluvial settings by several authors (Jordanova et al., 2006; Bábek et al., 2011; Faměra et al., 2013). However, the magnetic susceptibility in our dataset shows little correlation with the grain size changes, raw XRF data, clr of XRF data, and the latent components. Consequently, it cannot be used as independent variable to assess the applicability of the clr and the robust BSS algorithms in our dataset.

5.2. Interpretation of clr coefficients

Centred log-ratio coefficients significantly reduce the difference between the core-scan and powdered EDXRF readings (Fig. 7) and thus justify previous attempts to calibrate the core-scanning data using log-ratio techniques (Weltje and Tjallingii, 2008). A good correlation ($R^2 > 0.5$) between clr of the core-scans and powdered samples was observed for Ca, Mn, Cu, Sr, and Zr (two cores, $n = 189$) and partly also for P, K and Pb in the NOS-2 core. All elements in the powdered samples, with exception of Si, As, Sr and Zr, show high variability in concentrations consistent with the lithology changes across the NOS-2 core (Figs. 3, 6). This means that the layers with different grain size and porosity also have different mineral composition. Surprisingly, although these parameters are known to strongly affect the results of the XRF core-scans (Croudace et al., 2006; MacLachlan et al., 2015) they are preferentially picked up by XRF of the powdered samples and not by the core scans in our data set.

5.3. Geochemical interpretation of the latent components

The IC2 and IC4 latent components show statistical correlation between the core scan and powdered XRF data. We assume that this similarity reflects a common control on both the cores scan and powdered data and thus they are not largely affected by the method of measurement. Loadings of latent components by element clr, found by eSAM-SOBI are shown in Table 1. The loadings explain the elemental contributions in the latent components and thus they can be used in geochemical interpretation.

The vertical distribution of the IC2 scores is consistent with the grain size and porosity changes thus indicating that the IC2 primarily reflects the sediment lithology. Maximum positive loadings of IC2 correspond to Zr(clr) and Mn(clr) while negative loadings correspond to Cu(clr) and Zn(clr) (Table 1), therefore the log-ratio between these two pairs of elements seems to be the main building block of this component. Moreover, the vertical distributions of Si/Al and Zr/Al element ratios (Fig. 9), which are used as grain-size proxies (Vijver et al., 2008; Bábek et al., 2015), show high degree of similarity with the IC2 on the powder XRF data. Zirconium is a refractory element, which is typically enriched in silt- and sand-sized sediments and thus is grain-size sensitive (Morton and Hallsworth, 1999; Novák et al., 2017). On the other hand, Mn as a redox-sensitive element is vertically mobilized in recurrent cycles of reduction / dissolution and oxidation / precipitation in layers representing redox boundaries (Davison, 1993; Staubwasser et al., 2006). Redox boundaries may develop at lithological boundaries between

Table 1
Loadings of selected latent components found by eSAM-SOBI. Each entry is related to the clr value of the given element and therefore quantify the effect of the relative dominance of the given element to the actual value of IC.

	Al (clr)	Si (clr)	P (clr)	K (clr)	Ca (clr)	Ti (clr)	Mn (clr)	Fe (clr)	Cu (clr)	Zn (clr)	As (clr)	Rb (clr)	Sr (clr)	Zr (clr)	Pb (clr)
IC1	-0.168	-0.486	-0.147	-0.18	0.076	0.295	0.487	0.226	-0.004	-0.323	-0.08	0.089	-0.065	0.289	-0.007
IC2	-0.152	-0.051	-0.017	-0.202	0.031	0.222	0.375	0.058	-0.214	-0.375	-0.076	-0.101	-0.065	0.581	-0.015
IC3	0.459	0.24	0.235	0.676	0.079	-0.184	-0.113	-0.403	-0.251	-0.939	-0.921	-0.273	-0.008	-0.125	-0.896
IC4	0.019	0.02	0.127	0.081	0.063	-0.179	0.088	-0.177	-0.049	0.313	-0.035	-0.013	0.002	-0.167	-0.093
IC5	0.434	0.997	0.359	0.313	-0.132	-0.422	0.16	-0.342	-0.121	0.1	-1.168	-0.181	0.024	-0.532	0.508
IC6	0.33	-0.04	0.12	0.65	0.12	-0.12	-0.51	-0.34	-1.48	0.96	0.18	0.31	0.01	-0.27	0.06
IC7	-0.32	-0.44	0.17	-0.05	0.01	-0.16	0.19	-0.66	0	1.07	-0.42	-0.13	-0.22	0.69	0.27
IC8	-0.04	-0.97	-0.25	-0.06	0.65	0.67	-0.59	0.58	0.19	-0.64	-0.34	0.4	0.23	0.04	0.15

Highest values (bold) indicate the highest contribution of the elements.

more- and less-porous sediments being affected by groundwater flow and oxygen saturation. The IC2 reflects geochemical changes associated with grain size and hereto related development of redox barriers facilitating precipitation of redox-sensitive elements.

The IC4 component markedly increases in the topmost, about 25 cm thick sediment layer of the NOS-2 core, which is consistent with the rise in Ca/Al and P/Al ratios and enrichment factors of As (Fig. 8). In contrast to the element ratios, enrichment factors (EF) and centred log-ratios the IC4 completely filters out the effects of grain size. Maximum positive loadings of IC4 correspond to P(clr) and Zn(clr) (Table 1). In the topmost five centimetres, the IC4 peak overlaps with a peak in IC7, and elevated values of Zn and Cu (Fig. 8). These two elements can be derived from anthropogenic sources as indicated by their EF values (XRF powder data) ranging between 5 and 10. Elements correlating with the IC2 and IC4, Ca, P, Cu, and Zn are usually associated with increased organic productivity in lacustrine and marine environments and partly (Cu and Zn) with reducing conditions in water column of near the sediment–water interface (Sageman and Lyons, 2003; Dean, 2007; Tribouillard et al., 2006; Ettler et al., 2017). We assume the IC4 to reflect increased organic input into the sediments, which was presumably associated with eutrophication and accumulation of anthropogenic Cu and Zn.

The vector of element concentrations is analysed as a whole throughout the BSS algorithm. Consequently, complex relations between elements, typical for data of a compositional nature, are therefore preserved and taken into account. Although the presented methodology focuses on eSAM-SOBI, an algorithm for time-dependent data with possible presence of outliers, the family of BSS methods also contains algorithms for iid or space-dependent data. These methods therefore represent a flexible framework for restoring hidden structures from geochemical measurements. The time-series approach of the eSAM-SOBI seems to be suitable for identification of stratigraphic trends in sediment geochemistry although it can have certain disadvantages. For example, identification of thin event-beds with contrasting composition (e.g., ash layers) may be limited, or strongly affected by the sampling interval, using the eSAM-SOBI. However, the time-series statistics of core-scan data may represent a complementary method to the multivariate log-ratio calibration, which is independent on the stratigraphic position in a core.

6. Conclusions

The studied sediment cores penetrated layers of clayey silt and silty sands with highly variable grain size and variable water content (40 to 61 wt% in silts; ~25 to ~45 wt% in sands), both of them supposedly affecting the results of XRF core scanning. Sediments in the Nosice reservoir are enriched in Si and Zr and depleted in Ca, P, Rb, and As, as compared to their Hričov reservoir counterparts. Their compositional contrast is presumably due to different geology in catchments of upstream tributaries of the Váh River, which suggests that reservoirs act as important barriers to downstream sediment dispersal.

The XRF core scan data show systematically lower element concentrations, on average only 16 to 41 % of their XRF powder counterparts. Several elements (K, Rb, and Ti) show relatively strong relationship to Al both in the XRF core scan- and powder analyses. This demonstrates the effect of changing contents of aluminosilicates (clay minerals) as a function of changing grain size. This grain-size effect is stronger (higher R^2) in the powder data than in the core scans, which is surprising because grain size is particularly known to affect the results of the XRF core-scans strongly.

As expected, the element concentrations in XRF cores scans and powder data are poorly correlatable. Centred log-ratio (clr) coefficients improved for reinterpreted data (in terms of relative dominance of elements) the correlations (R^2 from 0.33 to 0.88, $n = 179$) in Al, P, K, Ca, Mn, Fe, Cu, As, Sr, Zr, and Pb. In addition, the clr coefficients significantly reduce the absolute difference between the XRF core-scan and powder data, which justifies previous attempts to calibrate the core-

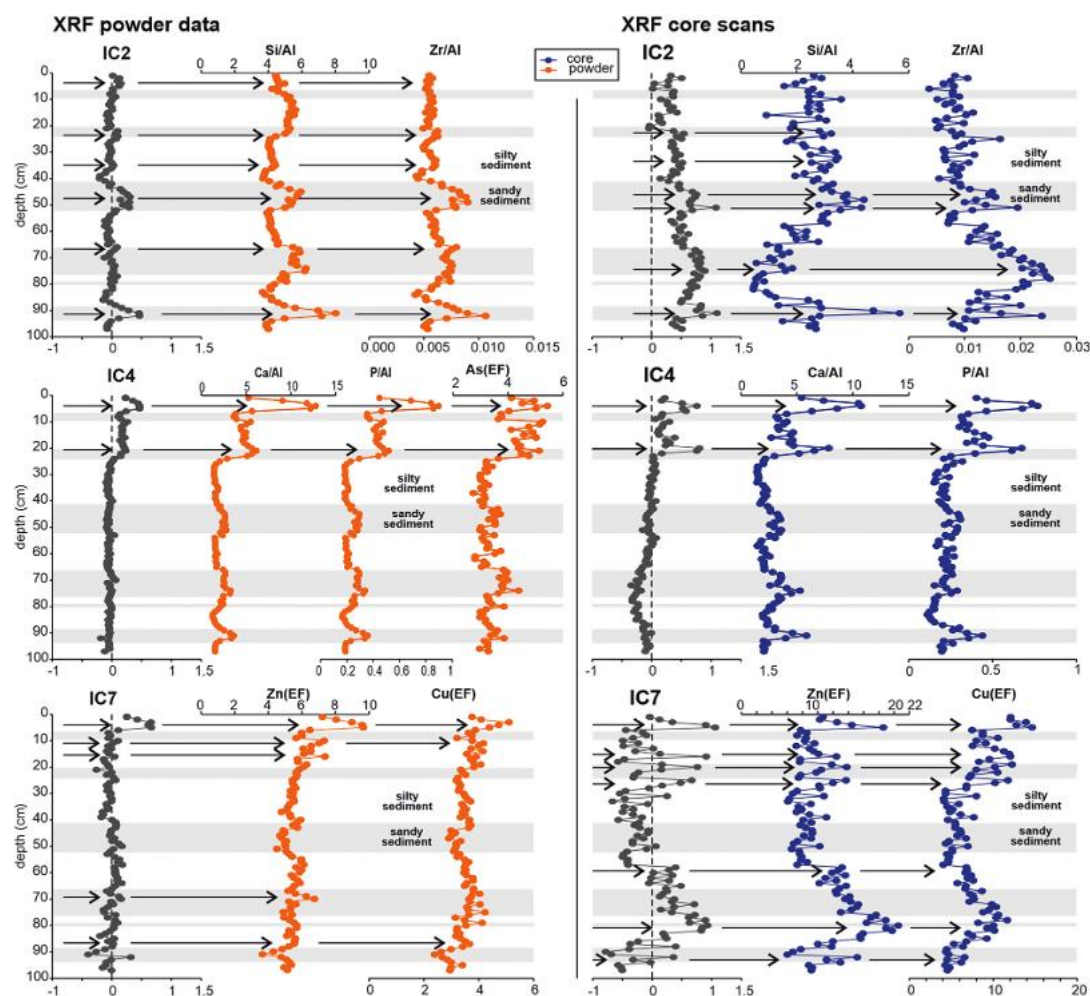


Fig. 9. Vertical distribution of latent components IC2, IC4 and IC7 and their correlation with selected element ratios (Si/Al, Zr/Al, Ca/Al, and P/Al, raw data) and enrichment factors (EF) of typical anthropogenic pollutants (As, Cu, and Zn).

scanning data using log-ratio techniques. Geochemical core scanning using hand-held XRF device is a method, which allows reasonable interpretation of primary geogenic signals in soft sediments in terrestrial, e.g., fluvial environments (RQ 1).

Application of the eSAM-SOBI algorithm, a time-series statistical technique, yielded eight latent components (IC1 to IC8) representing 99 % of total variability. The scores of IC4, IC2, and IC7 of the XRF powder- and core scan data show significant statistical correlation and similar depth distribution, which reflects a common control on the data and the fact that these latent components are not largely affected by the method of measurement. Their geochemical interpretation is based on the IC loadings by element clr, and stratigraphic correlation with the sediment grain size and water content. The IC2 corresponds to high Zr(clr) loadings and it primarily reflects the compositional contrast between the sandy and silty layers. The IC4 and IC7 components correspond to high P(clr), Ca(clr), and Zn(clr) loadings while they are not affected by grain size. These latent components presumably reflect increased organic input, eutrophication, and accumulation of anthropogenic elements. Being capable of picking up the effects of grain size and anthropogenic enrichment, the robust BSS approach represents a promising time-series statistical method to separate signal from noise. The BSS method allows to process large datasets in environmental geochemistry, which can be

obtained by XRF-CS without the need of prior calibration to the measurements resulting from the conventional methods (RQ 2).

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.gexplo.2022.107106>.

CRedit authorship contribution statement

Ondřej Bábek: Conceptualization, Methodology, Writing – Original Draft, Review and Editing, Visualisation, Supervision Kamila Fačevicová: Conceptualization, Methodology, Formal Analysis, Writing – Original Draft, Review and Editing Martin Židek: Investigation, Writing – Review and Editing Jan Sedláček: Investigation, Resources, Writing – Review and Editing Christoph Muehlmann: Methodology, Formal Analysis, Writing – Review and Editing Klaus Nordhausen: Methodology, Formal Analysis, Writing – Review and Editing Karel Hron: Methodology, Writing – Review and Editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This research was supported by the Czech Science Foundation (GACR) research project 19-01768S. The Austrian Science Fund P31881-N32 supported the work of CM and KN. We thank Vodohospodářská výstavba, s.p. company, Bratislava, Slovakia for permission to access the Hričov Reservoir. Jan Sedláček participated in sediment sampling and grain-size measurements. Martin Židek measured the XRF, water content data, and participated in grain-size measurements. Kamila Fačevicová, Christoph Muehlmann, and Klaus Nordhausen provided the statistical analyses and interpretations. Karel Hron contributed to the statistical interpretations and co-ordinated the research project. Ondřej Bábek designed the study, processed and interpreted the geochemical and grain-size data, and wrote the manuscript. All authors contributed to the final interpretations.

References

- Arnaud, F., Schmitt, L., Johnstone, K., Rollet, A.J., Piégay, H., 2019. Engineering impacts on the 572 Upper Rhine channel and floodplain over two centuries. *Geomorphology* 330, 13–27.
- Bábek, O., Faměra, M., Hilscherová, K., Kalvoda, J., Dobrovolný, P., Sedláček, J., Machát, J., Holoubek, I., 2011. Geochemical traces of flood layers in the fluvial sedimentary archive; implications for contamination history analyses. *Catena* 87, 281–290.
- Bábek, O., Matys Grygar, T., Faměra, M., Hron, K., Nováková, T., Sedláček, J., 2015. Geochemical background in polluted river sediments: how to separate the effects of sediment provenance and grain size with statistical rigour? *Catena* 135, 240–253.
- Bábek, O., Kumpan, T., Kalvoda, J., Matys Grygar, T., 2016. Devonian/Carboniferous boundary glacioeustatic fluctuations in a platform-to-basin direction: a geochemical approach of sequence stratigraphy in pelagic settings. *Sedim. Geol.* 337, 81–99.
- Bábek, O., Sedláček, J., Lenáková, Z., Elzníková, Z., Tolaszová, J., Pacina, J., 2021. Historical pond systems as long-term composite archives of anthropogenic contamination in the Vrchlice River. *Czechia. Anthropocene*. 33, 100283 <https://doi.org/10.1016/j.ancene.2021.100283>.
- Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F., Moulines, E., 1997. A blind source separation technique using second-order statistics. *IEEE Trans. Signal Process.* 45 (2), 434–444. <https://doi.org/10.1109/78.554307>.
- Bloemsma, M.R., Zabel, M., Stuut, J.B.W., Tjallingii, R., Collins, J.A., Weltje, G.J., 2012. Modelling the joint variability of grain size and chemical composition in sediments. *Sedim. Geol.* 280, 135–148.
- de Caritat, P., Grunsky, E.C., 2013. Defining element associations and inferring geological processes from total element concentrations in Australian catchment outlet sediments: Multivariate analysis of continental-scale geochemical data. *Appl. Geochem.* 33, 104–126.
- Comon, P., Jutten, C., 2010. *Handbook of blind Source Separation: independent Component Analysis and applications*. Academic Press, Amsterdam.
- Croudace, I.W., Rindby, A., Rothwell, R.G., 2006. ITRAX: description and evaluation of a new multi-function X-ray core scanner. In: Rothwell, R.G. (Ed.), *New Techniques in Sediment Core Analysis*, 267. Geol. Soc., London, Spec. Publ., pp. 51–63.
- Davison, W., 1993. Iron and manganese in lakes. *Earth Sci. Rev.* 34, 119–163.
- Dean, W.E., 2007. Sediment geochemical records of productivity and oxygen depletion along the margin of western North America during the past 60,000 years: teleconnections with Greenland Ice and the Cariaco Basin. *Quat. Sci. Rev.* 26, 98–114.
- Dunlea, A.G., Murray, R.W., Tada, R., Alvarez-Zarikian, C.A., Anderson, C.H., Gilli, A., Giosan, L., Gorgas, T., Hennekam, R., Irino, T., Murayama, M., Peterson, L.C., Reichart, G.-J., Seki, A., Zheng, H., Ziegler, M., 2019. Intercomparison of XRF Core Scanning results from seven Labs and Approaches to Practical Calibration. *Geochem. Geophys. 21* e2020GC009248.
- Ettler, V., Chren, M., Mihaljevic, M., Drahota, P., Kribek, B., Veselovsky, F., Sracek, O., Vanek, A., Penizek, V., Komarek, M., Mapani, B., Kamona, F., 2017. Characterization of Fe-Mn concentric nodules from Luvisol irrigated by mine water in a semi-arid agricultural area. *Geoderma* 299, 32–42.
- von Eynatten, H., Tolosana-Delgado, R., Karius, V., Bachmann, K., Caracciolo, L., 2016. Sediment generation in humid Mediterranean setting: grain-size and source-rock control on sediment geochemistry and mineralogy (Sila Massif, Calabria). *Sedim. Geol.* 336, 68–80.
- Fačevicová, K., Bábek, O., Hron, K., Kumpan, T., 2016. Element chemostratigraphy of the Devonian/Carboniferous boundary – a compositional approach. *Appl. Geochem.* 75, 211–221.
- Faměra, M., Bábek, O., Matys Grygar, T., Nováková, T., 2013. Distribution of heavy-metal contamination in regulated river-channel deposits: a magnetic susceptibility and grain-size approach; River Morava. *Czech Republic. Water Air Soil Poll.* 141, 35–54.
- Fišerová, E., Hron, K., 2011. On the Interpretation of Orthonormal Coordinates for Compositional Data. *Math. Geosci.* 43, 455–468.
- Gadd, P., Gobi, K., Sammut, J., Saintilan, N., Crawford, J., Mazumder, D., 2018. Itrax micro X-ray fluorescence (μXRF) for soft biological tissues. *MethodsX* 5, 1267–1271. <https://doi.org/10.1016/j.mex.2018.10.001>. ISSN 2215–0161.
- Gregory, B.R.B., Patterson, R.T., Reinhardt, E.G., Galloway, J.M., Roe, H.M., 2019. An evaluation of methodologies for calibrating Itrax X-ray fluorescence counts with ICP-MS concentration data for discrete sediment samples. *Chem. Geol.* 521, 12–27.
- Grosbois, C., Meybeck, M., Lestel, L., Lefevre, I., Moatier, F., 2012. Severe and contrasted polymetallic contamination patterns (1900–2009) in the Loire River sediments (France). *Sci. Tot. Env.* 435, 290–305.
- Henares, S., Donselaar, M.E., Bloemsma, M.R., Tjallingii, R., De Wijn, B., Weltje, G.J., 2019. Quantitative integration of sedimentological core descriptions and petrophysical data using high-resolution XRF scans. *Mar. Pet. Geol.* 110, 450–462.
- Hennekam, R., Sweere, T., Tjallingii, R., de Lange, G.J., Reichart, G.J., 2019. Trace metal analysis of sediment cores using a novel X-ray fluorescence core scanning method. *Quatern. Int.* 514, 55–67.
- Hsiung, K.H., Kanamatsu, T., Ikehara, K., Usami, K., Horng, C.S., Ohkouchi, N., Ogawa, N.O., Saito, S., Murayama, M., 2021. X-ray fluorescence core scanning, magnetic signatures, and organic geochemistry analyses of Ryukyu Trench sediments: turbidites and hemipelagites. *Prog. Earth Planet. Sci.* 8, 2. <https://doi.org/10.1186/s40645-020-00396-2>.
- Hubert, M., Rousseeuw, P.J., Verdonck, T., 2012. A deterministic algorithm for robust location and scatter. *J. Comput. Graph. Stat.* 21, 618–637.
- Ilmonen, P., Nordhausen, K., Oja, H., Theis, F., 2015. An affine equivariant robust second-order BSS method. In: Vincent, E., Yeredor, A., Koldovský, Z., Tichavský, P. (Eds.), *Latent variable analysis and signal separation: 12th international conference LVA/ICA 2015*. Springer, pp. 328–335.
- Jenkins, R., 1999. *X-Ray Fluorescence Spectroscopy*, Second edition. John Wiley & Sons, New Jersey.
- Jordanova, D., Jordanova, N., Hoffmann, V., 2006. Magnetic mineralogy and grain-size dependence of hysteresis parameters of single spherulites from industrial waste products. *Phys. Earth Planet. Interiors* 154, 255–265.
- Kelloway, S., Ward, C., Marjo, C., Wainwright, I., Cohen, D., 2014. Calibration for ED-XRF profiling of coal cores for the Itrax Core Scanner. *Powder Diffract.* 29, 28–34. <https://doi.org/10.1017/S088571561400089X>.
- Lacey, J.P., Evrard, O., Smith, H.G., Blake, W.H., Olley, J.M., Minella, J.P.G., Owens, P. N., 2017. The challenges and opportunities of addressing particle size effects in sediment source fingerprinting: a review. *Earth-Sci. Rev.* 169, 85–103.
- Löwemark, L., Bloemsma, M., Croudace, I., Daly, J.S., Edwards, R.J., Francus, P., Galloway, J.M., Gregory, B.R.B., Huang, J.-J.S., Jones, A.F., Kylander, M., Luon, Y., MacLachlan, S., Ohlendorf, C., Patterson, R.T., Pearce, C., Profe, J., Reinhardt, E.G., Stranne, C., Tjallingii, R., Turner, J.N., 2019. Practical guidelines and recent advances in the Itrax XRF core-scanning procedure. *Quat. Int.* 514, 16–29.
- MacLachlan, S., Hunt, J., Croudace, I., 2015. An Empirical Assessment of Variable Water Content and Grain-size on X-Ray Fluorescence Core-Scanning Measurements of Deep Sea Sediments. In: Croudace, I., Rothwell, R. (Eds.), *Micro-XRF Studies of Sediment Cores. Developments in Palaeoenvironmental Research*, Springer, Berlin Heidelberg, pp. 173–185.
- Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibián-Barrera, M., Verbeke, T., Koller, M., Conceicao, E.L.T., di Palma, M.A., 2021. Robustbase: Basic Robust Statistics. <http://robustbase.r-forge.r-project.org/>.
- Matys Grygar, T., Elzníková, J., Bábek, O., Hošek, M., Engel, Z., Kiss, T., 2014. Obtaining isochrones from pollution signals in a fluvial sediment record: a case study in uranium-polluted floodplain of the Ploučnice River, Czech Republic. *Appl. Geochem.* 48, 1–15.
- Matys Grygar, T., Mach, K., Hron, K., Fačevicová, K., Martinez, M., Zeeden, C., Schnabl, P., 2020. Lithological correction of chemical weathering proxies based on K, Rb, and Mg contents for isolation of orbital signals in clastic sedimentary archives. *Sedim. Geol.* 406, 105717 <https://doi.org/10.1016/j.sedgeo.2020.105717>.
- McKinley, J., Hron, K., Grunsky, E., Reimann, C., de Caritat, P., Filzmoser, P., van den Boogart, K.G., Tolosana-Delgado, R., 2016. The single component geochemical map: fact or fiction. *J. Geochem. Explor.* 162, 16–28.
- Miettinen, J., Illner, K., Nordhausen, K., Oja, H., Taskinen, S., Theis, F., 2016. Separation of uncorrelated stationary time series using autocovariance matrices. *J. Time Ser. Anal.* 37, 337–354.
- Miettinen, J., Nordhausen, K., Taskinen, S., 2017. Blind source separation based on joint diagonalization in R: the packages JADE and BSSasympt. *J. Stat. Software* 76, 1–31.
- Miller, H., Croudace, I.W., Bull, J.M., Cotterill, C.J., Dix, J.K., Taylor, R.N., 2015. Modern Pollution Signals in Sediments from Windermere, NW England, Determined by Micro-XRF and Lead Isotope Analysis. In: Croudace, I.W., Rothwell, R.G. (Eds.), *Micro-XRF Studies of Sediment Cores Applications of a Non-Destructive Tool for the Environmental Sciences, Developments in Palaeoenvironmental Research*. Springer, Berlin Heidelberg, pp. 423–442.
- Morton, A.C., Hallsworth, C.R., 1999. Processes Controlling the Composition of Detrital Heavy Mineral Assemblages in Sandstones. *Sedim. Geol.* 124, 3–29.
- Muehlmann, C., Fačevicová, K., Gardlo, A., Janečková, H., Nordhausen, K., 2021. Independent Component Analysis for Compositional Data. In: Daouia, A., Ruiz-Gazen, A. (Eds.), *Advances in Contemporary Statistics and Econometrics*. Springer, pp. 1–15. https://doi.org/10.1007/978-3-030-73249-3_27.
- Mukai, R., Sawada, H., Shoko, A., Makino, S., 2003. Robust real-time blind source separation for moving speakers in a room. *Proceedings of ICASSP '03 IEEE International Conference on Acoustics, Speech, and Signal Processing* 469–472. <https://doi.org/10.1109/ICASSP.2003.1200008>.
- Nordhausen, K., Oja, H., Filzmoser, P., Reimann, C., 2015. Blind source separation for spatial compositional data. *Math. Geosci.* 47, 753–770.

- Nordhausen, K., Fischer, G., Filzmoser, P., 2021. Blind Source Separation for Compositional Time Series. *Math. Geosci.* 53, 905–924.
- Novák, A., Bábek, O., Kapusta, J., 2017. Late Quaternary tectonic switching of siliciclastic provenance in the strike-slip-dominated foreland of the Western Carpathians; Upper Morava Basin, Bohemian Massif. *Sedim. Geol.* 355, 58–74.
- Pan, Y., Matilainen, M., Taskinen, S., Nordhausen, K., 2021. A review of second-order blind identification methods. *WIREs. Comput. Stat.* e1550 <https://doi.org/10.1002/wics.1550>.
- Pawlowsky-Glahn, V., Buccianti, A. (Eds.), 2011. *Compositional Data Analysis: Theory and Applications*. Wiley, ISBN, ISBN 978-0-470-71135-4.
- Peti, L., Gadd, P.S., Hopkins, J.L., Augustinus, P.C., 2019. Itrax mu-XRF core scanning for rapid tephrostratigraphic analysis: a case study from the Auckland Volcanic Field maar lakes. *J. Quat. Sci.* 35, 54–65.
- Profe, J., Ohlendorf, C., 2019. X-ray fluorescence scanning of discrete samples - an economical perspective. *Quat. Int.* 514, 68–75.
- Reimann, C., Filzmoser, P., Fabian, K., Hron, K., Birke, M., Demetriades, A., Dinelli, E., Ladenberger, A., 2012. The concept of compositional data analysis in practice – Total major element concentrations in agricultural and grazing land soils in Europe. *Sci. Tot. Env.* 426, 196–210.
- Rodríguez-Germade, I., Rubio, B., Rey, D., Borrego, J., 2015. Detection and monitoring of REEs and Related Trace elements with an Itrax™ Core Scanner in the Ría de Huelva (SW Spain). *Water Air Soil Pollut.* 226, 137. <https://doi.org/10.1007/s11270-015-2389-3>.
- Ross, D.J.K., Bustin, R.M., 2009. Investigating the use of sedimentary geochemical proxies for paleoenvironment interpretation of thermally mature organic-rich strata: examples from the Devonian-Mississippian shales, Western Canadian Sedimentary Basin. *Chem. Geol.* 260, 1–19.
- Rothwell, R.G., Croudace, I.W., 2015. Twenty years of XRF Core Scanning Marine Sediments: What do Geochemical Proxies tell Us? In: Croudace, I., Rothwell, R. (Eds.), *Micro-XRF Studies of Sediment Cores. Developments in Palaeoenvironmental Research*, Springer, Berlin Heidelberg, pp. 25–102.
- Sageman, B.B., Lyons, T.W., 2003. Geochemistry of fine-grained sediments and sedimentary rocks. In: Mackenzie, F.T. (Ed.), *Sediments, Diagenesis, and Sedimentary Rocks. Treatise on Geochemistry* vol. 7, 115–158. Elsevier, Amsterdam.
- Staubwasser, M., von Blanckenburg, F., Schoenberg, R., 2006. Iron isotopes in the early marine diagenetic iron cycle. *Geology* 34, 629–632.
- Team, R.Core, 2021. *A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>. (Accessed 20 December 2021).
- Templ, M., Hron, K., Filzmoser, P., 2011. robCompositions: An R-package for robust statistical analysis of compositional data. In: Pawlowsky-Glahn, V. (Ed.), *Compositional Data Analysis: Theory and Applications*. Wiley, New York, pp. 341–355.
- Tribouillard, N., Algeo, T.J., Lyons, T., Riboulleau, A., 2006. Trace metals as paleoredox and paleoproductivity proxies: an update. *Chem. Geol.* 232, 12–32.
- Vauclin, S., Mourier, B., Tena, A., Piégay, H., Winiarski, T., 2019. Effects of river infrastructures on 875 the floodplain sedimentary environment in the Rhône River. *J. Soils Sedim.* 2019 <https://doi.org/10.1007/s11368-019-02449-6>.
- Vijver, M.G., Spijker, J., Vink, J.P.M., Postuma, L., 2008. Determining metal origins and availability in fluvial deposits by analysis of geochemical baselines and solid–solution partitioning measurements and modelling. *Env. Pollution* 156, 832–839.
- Vörösmarty, C.J., Meybeck, M., Fekete, B., Sharma, K., Green, P., Syvitski, J.P., 2003. Anthropogenic sediment retention: major global impact from registered river impoundments. *Glob. Planet. Chang.* 39, 169–190.
- Weltje, G.J., Tjallingii, R., 2008. Calibration of XRF core scanners for quantitative geochemical logging of sediment cores: theory and application, *Earth Planet. Sci. Lett.* 274, 423–438.
- Weltje, G.J., Bloemsa, M.R., Tjallingii, R., Heslop, D., Röhl, U., Croudace, I.W., 2015. Prediction of geochemical composition from XRF core scanner data: a new multivariate approach including automatic selection of calibration samples and quantification of uncertainties. In: Croudace, I.W., Rothwell, R.G. (Eds.), *Micro-XRF Studies of Sediment Cores, Developments in Palaeoenvironmental Research*. Springer, Berlin Heidelberg, pp. 507–534.
- Zhang, Y., Zhao, Y., 2013. Modulation domain blind speech separation in noisy environments. *Speech Comm.* 55, 1081–1099.

Cellwise robust covariance estimator

Rieser, C., **Fačevicová, K.**, Filzmoser, P. Cell-wise robust covariance estimation for compositions, with application to geochemical data. *Journal of Geochemical Exploration*, 253, 107299 (2023).

<https://doi.org/10.1016/j.gexplo.2023.107299>

[IF 4.6 Q1; AIS 0.795 Q2]

[3 citations on WoS; 3 citations without autocitations]

[Author's contribution: 25%]

Abstract

Cell-wise outliers are outliers in single entries of a compositional data matrix, and they can lead to a certain bias in the statistical analysis. Traditional row-wise robust methods downweight outlying observations for the estimation, independent of how many or which cells of an observation are contaminated. Cell-wise robustness still makes use of the information contained in non-contaminated cells. Here, cell-wise robustness is used for the estimation of the variation and the covariance matrix. For higher dimensional data also a regularized estimator is introduced. The advantages of the cell-wise robust estimators are demonstrated in simulation experiments and in a geochemistry application in the context of clustering and principal component analysis.



Contents lists available at ScienceDirect

Journal of Geochemical Exploration

journal homepage: www.elsevier.com/locate/gexplo

Cell-wise robust covariance estimation for compositions, with application to geochemical data

Christopher Rieser^a, Kamila Fačevicová^b, Peter Filzmoser^{a,*}^a TU Wien, Institute of Statistics and Mathematical Methods in Economics, Wiedner Hauptstrasse 8-10, Vienna, Austria^b Palacký University Olomouc, Department of Mathematical Analysis and Applications of Mathematics, 17. listopadu 12, Olomouc, Czech Republic

ARTICLE INFO

Keywords:

Cell-wise outliers
Covariance matrix
Geochemistry
Log-ratio analysis

ABSTRACT

Cell-wise outliers are outliers in single entries of a compositional data matrix, and they can lead to a certain bias in the statistical analysis. Traditional row-wise robust methods downweight outlying observations for the estimation, independent of how many or which cells of an observation are contaminated. Cell-wise robustness still makes use of the information contained in non-contaminated cells. Here, cell-wise robustness is used for the estimation of the variation and the covariance matrix. For higher dimensional data also a regularized estimator is introduced. The advantages of the cell-wise robust estimators are demonstrated in simulation experiments and in a geochemistry application in the context of clustering and principal component analysis.

1. Introduction

The field of compositional data analysis (CoDA) has received a lot of attention after the publication of the book (Aitchison, 1986) of John Aitchison in 1986. The approach followed there is often called log-ratio methodology, as the building blocks are pairwise log-ratios between the variables (compositional parts). If D denotes the number of parts, then there are $D(D-1)/2$ (relevant) pairwise log-ratios for the analysis, thus each part is involved in $D-1$ different log-ratios. These pairwise log-ratios are usually not used separately as new variables within a multivariate statistical analysis, but they are appropriately aggregated, e.g. to form a coordinate system with $D-1$ coordinates (Filzmoser et al., 2018). This, however, creates difficulties if an observation has outliers only in single parts, since their effect in the coordinate representation will very much depend on the type of aggregation, as well as on the parts which are contaminated.

Outliers in single variables of traditional non-compositional data are called cell-wise outliers. If an observation contains cell-wise outliers, the remaining non-contaminated cells usually still contain valuable information for the analysis. This can be very different for compositional data. Contamination in single parts of a composition can be caused by atypical values in this part, or it can result from an improper imputation of zeros or missing values. When forming pairwise log-ratios with contaminated cells, also the resulting log-ratio pairs can be outlying, and the same is true for other types of aggregation with contaminated parts.

Robust multivariate statistical methods commonly downweight complete observations if they are deviating, independent of how many cells (variables) of such observations are contaminated. Since observations are usually arranged in the rows of a data matrix, this procedure is often called row-wise robustness (Maronna et al., 2019). Especially for a growing number of variables, this would lead to a severe loss of valuable information if only few cells of an observation are deviating. Therefore, methods have been developed to identify and downweight deviating cells rather than only outlying rows, and this is called cell-wise robustness (Rousseeuw and Van Den Bossche, 2018).

Row-wise robustness is well established in the context of CoDA (Filzmoser et al., 2018). Since many multivariate methods are based on an estimated covariance matrix, a robust version can be obtained by expressing the composition in coordinates, and performing robust covariance estimation with any of the proposed methods (Maronna et al., 2019, p 195–224), for example with the widely used MCD (minimum covariance determinant) estimator (Rousseeuw, 1985; Rousseeuw and Van Driessen, 1999). Since a complete row is downweighted (or not), it does not matter which parts are contaminated, and which form of coordinate representation is used.

It is not straightforward to implement cell-wise robustness for the CoDA case, especially after the compositions are expressed in coordinates. Depending on the parts which are contaminated, as well as on the coordinate representation used, it might be hardly possible to trace back which compositional parts caused cell-wise outliers, detected in the

* Corresponding author.

E-mail address: peter.filzmoser@tuwien.ac.at (P. Filzmoser).<https://doi.org/10.1016/j.gexplo.2023.107299>

Received 15 May 2023; Received in revised form 19 July 2023; Accepted 11 August 2023

Available online 24 August 2023

0375-6742/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

space of the coordinates. It is much more straightforward to investigate pairwise log-ratios in order to identify those parts which caused the outlyingness. Several attempts in this direction have already been done, e.g. in Štefelová et al. (2021) and Walach et al. (2020) in the context of regression and outlier detection.

In this paper we focus on cell-wise robustness, but limited to the problem of robust covariance estimation. Nevertheless, a robustly estimated covariance matrix opens the door to many multivariate methods, and in the application part we will focus on cell-wise robust principal components. The main idea is to estimate the elements of the variation matrix robustly, and these elements contain the estimated variances of pairwise log-ratios (Aitchison, 1986). Afterwards, we make use of the well-known relationship between the variation matrix and the compositional covariance matrix (Aitchison, 1986).

The paper is organized as follows. Section 2 recalls some of the CoDA concepts which are needed to introduce the cell-wise robust covariance estimators in Section 3. Simulation studies in Section 4 reveal robustness properties of the new proposals, and a comparison of the methods based on geochemical data is provided in Section 5. The final Section 6 concludes.

2. Some CoDA concepts

Denote by $\mathbb{R}_+^D$ the space of D dimensional strictly positive real values. In the classical approach of CoDA the D -part simplex is of central importance: it is defined as the space $\mathcal{S}^D = \{\mathbf{x} = (x_1, \dots, x_D)^T \in \mathbb{R}_+^D \mid \sum_{i=1}^D x_i = 1\}$. Considering also a composition $\mathbf{y} = (y_1, \dots, y_D)^T$, the D -part simplex is equipped with the Aitchison inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{A}} := \frac{1}{2D} \sum_{i,j=1}^D \ln \left(\frac{x_i}{x_j} \right) \ln \left(\frac{y_i}{y_j} \right)$$

and the so-called perturbation $\mathbf{x} \oplus \mathbf{y} := \frac{1}{\sum_{j=1}^D x_j y_j} (x_1 y_1, \dots, x_D y_D)^T$ and powering $\alpha \odot \mathbf{x} := \frac{1}{\sum_{j=1}^D x_j^\alpha} (x_1^\alpha, \dots, x_D^\alpha)^T$ operation, for any $\alpha \in \mathbb{R}$, see Pawłowsky-Glahn and Egozcue (2001) and Billheimer et al. (2001). With these definitions it can be shown that $(\mathcal{S}^D, \langle \cdot, \cdot \rangle_{\mathcal{A}}, \oplus, \odot)$ is a Hilbert space with neutral element $\frac{1}{D}(1, \dots, 1)^T \in \mathbb{R}_+^D$ and norm $\|\mathbf{x}\|_{\mathcal{A}} := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{A}}}$, see Pawłowsky-Glahn et al. (2015), p.23–30. Of central importance in CoDA is the clr (centered log-ratio)-map

$$\text{clr} : \mathcal{S}^D \rightarrow \mathbb{R}^D, \quad \text{clr}(\mathbf{x}) := \left(\ln \left(\frac{x_1}{\sqrt[D]{\prod_{j=1}^D x_j}} \right), \dots, \ln \left(\frac{x_D}{\sqrt[D]{\prod_{j=1}^D x_j}} \right) \right)^T. \quad (1)$$

It is well known that the clr-map is distance preserving on $\mathcal{S}^D$, see Pawłowsky-Glahn et al. (2015), p 35. Additionally, the clr-map possesses the following properties,

$$\text{clr}(\mathbf{x} \oplus \mathbf{y}) = \text{clr}(\mathbf{x}) + \text{clr}(\mathbf{y}) \quad (2)$$

$$\text{clr}(\alpha \odot \mathbf{x}) = \alpha \text{clr}(\mathbf{x}) \quad (3)$$

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{A}} = \langle \text{clr}(\mathbf{x}), \text{clr}(\mathbf{y}) \rangle_2 \quad (4)$$

with $\langle \cdot, \cdot \rangle_2$ denoting the standard inner product in $\mathbb{R}^D$, and $\alpha \in \mathbb{R}$ (Pawłowsky-Glahn et al., 2015, p 35). A drawback of the clr-map is that it is not bijective onto $\mathbb{R}^D$ and its components are constrained to sum to zero. The clr-map can also be written as a linear map of the coordinate-wise log-transformed variables of $\mathbf{x}$, i.e. $\text{clr}(\mathbf{x}) = \mathbf{L}_{\mathcal{A}} \ln(\mathbf{x})$, where $\mathbf{L}_{\mathcal{A}}$ is the so-called centering matrix

$$\mathbf{L}_{\mathcal{A}} := \frac{1}{D} \begin{pmatrix} D-1 & -1 & & -1 \\ -1 & \ddots & & \vdots \\ \vdots & & \ddots & -1 \\ -1 & & & D-1 \end{pmatrix}$$

which will play a role in the next section. Alternatively, the so-called ilr (isometric log-ratio)-map (Egozcue et al., 2003) is often considered as it preserves properties (2)–(4) and is bijective. It is defined as

$$\text{ilr}_{\mathbf{V}} : \mathcal{S}^D \rightarrow \mathbb{R}^{D-1}, \quad \text{ilr}_{\mathbf{V}}(\mathbf{x}) := \mathbf{V}^T \text{clr}(\mathbf{x}), \quad (5)$$

where $\mathbf{V} \in \mathbb{R}^{D \times (D-1)}$ is any matrix with orthogonal columns that span the $D-1$ dimensional subspace $\{\mathbf{z} \in \mathbb{R}^D \mid \sum_{j=1}^D z_j = 0\} \subset \mathbb{R}^D$. The centering matrix has the following important properties that also connect it to any admissible $\mathbf{V}$ as just described,

$$\mathbf{L}_{\mathcal{A}} = \mathbf{L}_{\mathcal{A}}^T, \quad \mathbf{L}_{\mathcal{A}} \mathbf{1} = \mathbf{0}, \quad \mathbf{L}_{\mathcal{A}} = \mathbf{V} \mathbf{V}^T,$$

where $\mathbf{1}$ is a vector of ones.

There are multiple ways to characterize the dependence structure between different compositional parts of $\mathbf{x}$ in CoDA depending on the transformation used. Given two different random vectors $\mathbf{u}, \mathbf{v}$, then the covariance matrix is given as $\text{Cov}(\mathbf{u}, \mathbf{v}) := \mathbb{E}((\mathbf{u} - \mathbb{E}(\mathbf{u}))(\mathbf{v} - \mathbb{E}(\mathbf{v}))^T)$, where $\mathbb{E}$ denotes the expectation. Then we write in the following $\Sigma^{\text{clr}} := \text{Cov}(\text{clr}(\mathbf{x}), \text{clr}(\mathbf{x}))$ for the covariance matrix of $\text{clr}(\mathbf{x})$, $\Sigma^{\text{ln}} := \text{Cov}(\ln(\mathbf{x}), \ln(\mathbf{x}))$ for the covariance matrix of $\ln(\mathbf{x})$ and $\Sigma^{\text{ilrv}} := \text{Cov}(\text{ilr}_{\mathbf{V}}(\mathbf{x}), \text{ilr}_{\mathbf{V}}(\mathbf{x}))$ for the covariance matrix of $\text{ilr}_{\mathbf{V}}(\mathbf{x})$, for any fixed $\mathbf{V}$. Additionally, we write $\mathbf{T} \in \mathbb{R}^{D \times D}$ for the variation matrix which is defined entry-wise for any $i, j = 1, \dots, D$ as $T_{ij} = \text{Cov}\left(\ln\left(\frac{x_i}{x_j}\right), \ln\left(\frac{x_i}{x_j}\right)\right) = \text{Var}\left(\ln\left(\frac{x_i}{x_j}\right)\right)$. Thus, the entries of the variation matrix are the variances of all pairwise log-ratios. Values near zero represent proportionality of the respective parts and consequently also their close relationship.

As noted in Aitchison (1986), there are the following relations between the covariance matrix Σ^{clr} and the variation matrix $\mathbf{T}$,

$$\mathbf{T} = \mathbf{1}\sigma^{2T} + \sigma^2 \mathbf{1}^T - 2\Sigma^{\text{clr}}, \quad (6)$$

$$\Sigma^{\text{clr}} = -\frac{1}{2} \mathbf{L}_{\mathcal{A}} \mathbf{T} \mathbf{L}_{\mathcal{A}}, \quad (7)$$

where $\sigma^2 = \text{diag}(\Sigma^{\text{clr}})$ denotes the vector of the diagonal elements of Σ^{clr} . These relationships will be very helpful for cell-wise robustness, as they link associations between pairs of compositional parts with covariances.

3. Cell-wise robust covariance estimation

Denote by $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathcal{S}^D$ a sample of N independent observations which are arranged as rows in the data matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$. The classical sample covariance matrix estimate of Σ^{clr} is given as $\hat{\Sigma}^{\text{clr}} = \frac{1}{N-1} \sum_{n=1}^N (\text{clr}(\mathbf{x}_n) - \hat{\boldsymbol{\mu}}^{\text{clr}})(\text{clr}(\mathbf{x}_n) - \hat{\boldsymbol{\mu}}^{\text{clr}})^T$, where $\hat{\boldsymbol{\mu}}^{\text{clr}} := \frac{1}{N} \sum_{n=1}^N \text{clr}(\mathbf{x}_n)$. It is a well-known fact that any contaminated data point $\mathbf{x}_n$ can have a huge effect on the classical sample covariance estimate of Σ^{clr} (Maronna et al., 2019, p 196–198). The same holds true if only a single entry of $\mathbf{x}_n$, say the k -th entry, is contaminated. A part of a composition is understood as contaminated, when it has an unusually high (or low) absolute value and which therefore distorts the whole relative structure of the respective composition. As we can write any l -th entry of $\text{clr}(\mathbf{x}_n)$ also in terms of pairwise log-ratios

$$\text{clr}(\mathbf{x}_n)_l = \ln \left(\frac{x_{nl}}{\sqrt[D]{\prod_{j=1}^D x_{nj}}} \right) = \frac{1}{D} \left(\ln \frac{x_{nl}}{x_{n1}} + \dots + \ln \frac{x_{nl}}{x_{nk}} + \dots + \ln \frac{x_{nl}}{x_{nD}} \right),$$

we can see that a contamination in x_{nk} propagates to $\text{clr}(\mathbf{x}_n)_l$, for any

$l \in \{1, \dots, D\}$, which in consequence also can have an impact on $\hat{\mu}^{\text{clr}}$, see also Mert et al. (2016). Something similar also holds for the ilr-map. Therefore, if a row of $\mathbf{X}$ contains only a single cell that displays outlying behavior, taking the clr- or ilr-map of this row can lead to a whole outlying vector. The goal is to find a covariance estimator which protects against outlying cells in compositional data.

The main idea to achieve cell-wise robustness is thus to restrict the influence of cell-wise outliers at the log-ratio level. The relationship (7) is of particular interest as it permits to estimate Σ^{clr} in terms of the variation matrix, which consists solely of the variances at the log-ratio level. To construct a robust estimator of Σ^{clr} , or equivalently Σ^{lr} , we suggest to replace each entry of the variation matrix $\mathbf{T}$, $t_{ij} = \text{Var}\left(\ln\left(\frac{x_i}{x_j}\right)\right)$, by a robust counterpart. This simplifies things not only because we focus on pairwise log-ratios, but also because robust variance estimation is a simpler task compared to robust covariance estimation.

Several robust estimators of the variance have been proposed in the literature. A highly robust and efficient estimator of scale (standard deviation) is the Qn estimator (Rousseeuw and Croux, 1993), which is essentially defined as the first quartile of the absolute pairwise differences. To be more precise, for a vector $\mathbf{z} = (z_1, \dots, z_N)^T$, the Qn estimator of scale is defined as

$$\rho(\mathbf{z}) = \rho_{n=1, \dots, N}(z_n) = 2.219 \cdot \left\{ |z_i - z_j|; i < j \right\}_{(k)}, \quad (8)$$

where $k = \binom{h}{2} \approx \binom{N}{2} / 4$ and $h = \lceil N / 2 \rceil + 1$, and (k) is the k -th value of the sorted absolute pairwise differences (in ascending order) (Rousseeuw and Croux, 1993). A further advantage of this estimator is that it does not require a (robust) location estimation. These properties make the Qn estimator preferable over other robust scale estimators, such as the more well-known median absolute deviation (MAD).

3.1. The cell-wise robust Qn covariance estimator

We will use the Qn estimator for the elements of the variation matrix to obtain $t_{ij}^Q = \rho_{n=1, \dots, N}\left(\ln\left(\frac{x_{ni}}{x_{nj}}\right)\right)^2$. We denote the resulting robust variation matrix by $\hat{\mathbf{T}}_Q := t_{ij}^Q$. To construct a valid estimate of Σ^{clr} we then transform $\hat{\mathbf{T}}_Q$ according to Eq. (7). This estimate should however fulfill the following three properties that Σ^{clr} has.

- Σ^{clr} is a positive semi-definite matrix.
- Σ^{clr} has an eigenvector $\mathbf{1}$ to the zero eigenvalue, i.e. $\Sigma^{\text{clr}}\mathbf{1} = \mathbf{0}$.
- Σ^{clr} is invariant under multiplication with $\mathbf{L}_{\mathcal{A}}$, thus $\Sigma^{\text{clr}} = \mathbf{L}_{\mathcal{A}}\Sigma^{\text{clr}}\mathbf{L}_{\mathcal{A}}$.

Note that the second point is a direct consequence of the third point, and the third point directly follows from the idempotency of $\mathbf{L}_{\mathcal{A}}$ and the representation of the clr-map through the latter. Therefore, to construct a positive semi-definite estimator we propose to take the nearest positive semidefinite matrix to $-\frac{1}{2}\mathbf{L}_{\mathcal{A}}\hat{\mathbf{T}}_Q\mathbf{L}_{\mathcal{A}}$ and adjust the median quantile. All together we perform the following steps:

1. Estimate the elements t_{ij}^Q of $\hat{\mathbf{T}}_Q$ by $t_{ij}^Q = \rho_{n=1, \dots, N}\left(\ln\left(\frac{x_{ni}}{x_{nj}}\right)\right)^2$, where $i \neq j$ (otherwise, the entries are zero).
2. Project $\hat{\mathbf{T}}_Q$ on the appropriate subspace by $\hat{\Sigma}_0^{\text{clr}} := -\frac{1}{2}\mathbf{L}_{\mathcal{A}}\hat{\mathbf{T}}_Q\mathbf{L}_{\mathcal{A}}$.
3. Find the nearest positive semi-definite matrix (p.s.d) $\hat{\Sigma}_1^{\text{clr}}$ by solving $\min_{\Sigma^{\text{clr}} \text{ p.s.d}} \|\hat{\Sigma}_0^{\text{clr}} - \Sigma^{\text{clr}}\|_F$, where $\|\cdot\|_F$ denotes the Frobenius norm. Note that also other measures could be considered that take into account the symmetry of the matrix.

4. The final estimator is $\hat{\Sigma}^{\text{clr}} = c\hat{\Sigma}_1^{\text{clr}}$ where c is chosen such that the median of the squared Mahalanobis distances is equal to the quantile 0.5 of the χ^2 distribution with degrees of freedom equal to the rank of $\hat{\Sigma}_1^{\text{clr}}$. This alignment allows to use the quantile 0.975 of this χ^2 distribution as outlier cutoff value for the squared Mahalanobis distances (Maronna et al., 2019, p 206).

For these steps to lead to an admissible estimate of $\hat{\Sigma}^{\text{clr}}$ we need to check that the three properties as described above hold. Clearly $\hat{\Sigma}_1^{\text{clr}}$ is positive semi-definite by definition and so we only need to check the latter two properties. As $\mathbf{L}_{\mathcal{A}}\mathbf{1} = \mathbf{0}$ holds, we can follow directly $\hat{\Sigma}_0^{\text{clr}}\mathbf{1} = \mathbf{0}$. Step 3. does not affect the eigenvectors, as the solution to the nearest positive semi-definite matrix problem is equal to a zero thresholding of the eigenvalues, see Higham (1988). Therefore, $\hat{\Sigma}_1^{\text{clr}}\mathbf{1} = \mathbf{0}$ holds, and consequently the same is true for $\hat{\Sigma}^{\text{clr}}$. The third property directly follows from the second one by $\mathbf{L}_{\mathcal{A}}\hat{\Sigma}^{\text{clr}}\mathbf{L}_{\mathcal{A}} = (\mathbf{I} - \frac{1}{D}\mathbf{1}\mathbf{1}^T)\hat{\Sigma}^{\text{clr}}(\mathbf{I} - \frac{1}{D}\mathbf{1}\mathbf{1}^T) = \hat{\Sigma}^{\text{clr}} - \frac{1}{D}\hat{\Sigma}^{\text{clr}}\mathbf{1}\mathbf{1}^T - \frac{1}{D}\mathbf{1}\mathbf{1}^T\hat{\Sigma}^{\text{clr}} + \frac{1}{D^2}\mathbf{1}\mathbf{1}^T\hat{\Sigma}^{\text{clr}}\mathbf{1}\mathbf{1}^T = \hat{\Sigma}^{\text{clr}}$.

In the following, we will refer to $\hat{\Sigma}^{\text{clr}}$ as the cell-wise robust Qn estimator, abbreviated simply as “Qn”.

3.2. The cell-wise robust regularized Qn covariance estimator

In the case where $N \leq D$, it is known in the non-compositional data setting, that non-robust as well as robust covariance estimates display an increased estimation error, and that estimates can be approved upon by adding a shrinkage term, often in form of a unit matrix, see Ledoit and Wolf (2004); Chen et al. (2011) and Hubert et al. (2018). Note, however, that also other shrinkage matrices can be used. Inspired by these ideas we also propose a shrinkage version. We first perform steps 1. to 3. as described above obtaining $\hat{\Sigma}_1^{\text{clr}}$. Instead of continuing to step 4., we now define $\hat{\Sigma}_2^{\text{clr}} := (1 - \alpha)\hat{\Sigma}_1^{\text{clr}} + \alpha\mathbf{L}_{\mathcal{A}}$, where α is a fixed chosen constant, see below. Lastly, we obtain the final estimator as $\hat{\Sigma}^{\text{clr}} := c\hat{\Sigma}_2^{\text{clr}}$, where c is chosen such that the median of the squared Mahalanobis distances is equal to the quantile 0.5 of the χ^2 distribution with degrees of freedom equal to the rank of $\hat{\Sigma}_2^{\text{clr}}$. Note that equivalently we could also compute instead step 1. to 2., replace $\hat{\Sigma}_0^{\text{clr}}$ by a shrunk version $(1 - \alpha)\hat{\Sigma}_1^{\text{clr}} + \alpha\mathbf{I}$ and then continue with step 3. to 4., with $(1 - \alpha)\hat{\Sigma}_1^{\text{clr}} + \alpha\mathbf{I}$ instead of $\hat{\Sigma}_1^{\text{clr}}$. The motivation for adding $\mathbf{L}_{\mathcal{A}}$ instead of the unit matrix is that the latter does not fulfill the properties of an admissible covariance estimator. $\mathbf{L}_{\mathcal{A}}$ however trivially does. One can also argue that $\mathbf{L}_{\mathcal{A}}$ can be considered the compositional pendant to the unit matrix. This can easily be seen from the fact that if a compositional variable $\mathbf{x}$ is standard Gaussian distributed, $\text{ilr}_{\mathbf{V}}(\mathbf{x}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then the covariance of $\ln(\mathbf{x})$ is given by $\mathbf{V}^T\mathbf{V} = \mathbf{L}_{\mathcal{A}}$ as $\text{ilr}_{\mathbf{V}}(\mathbf{x})^T \text{ilr}_{\mathbf{V}}(\mathbf{x}) = \ln(\mathbf{x})\mathbf{V}\mathbf{V}^T\ln(\mathbf{x})$ and $\mathbf{V}\mathbf{V}^T = (\mathbf{V}\mathbf{V}^T)^+$, where $+$ is the Moore-Penrose generalized inverse. This shrinkage step introduces an additional tuning parameter α . In the non-compositional literature, optimal choices of α are well-known. For an unbiased estimate $\hat{\Sigma}$ of Σ , the optimal choice of α such that the expected Frobenius distance between Σ and $(1 - \alpha)\hat{\Sigma} + \alpha\mathbf{I}$, i.e. $\mathbb{E}\left(\|\Sigma - ((1 - \alpha)\hat{\Sigma} + \alpha\mathbf{I})\|_F^2\right)$, is minimized, with $\mathbf{I}$ being a shrinkage target matrix, is given by

$$\alpha = \frac{\sum_{i,j=1}^D (\text{Var}(\hat{\Sigma}_{ij}) - \text{Cov}(\hat{\Gamma}_{ij}, \hat{\Sigma}_{ij}))}{\sum_{i,j=1}^D \mathbb{E}\left((\hat{\Gamma}_{ij} - \hat{\Sigma}_{ij})^2\right)},$$

see Schäfer and Strimmer (2005), where subscripts ij denote the corresponding matrix entries. Taking $\mathbf{I} = \mathbf{L}_{\mathcal{A}}$ we can drop $\text{Cov}(\hat{\Gamma}_{ij}, \hat{\Sigma}_{ij})$ due to

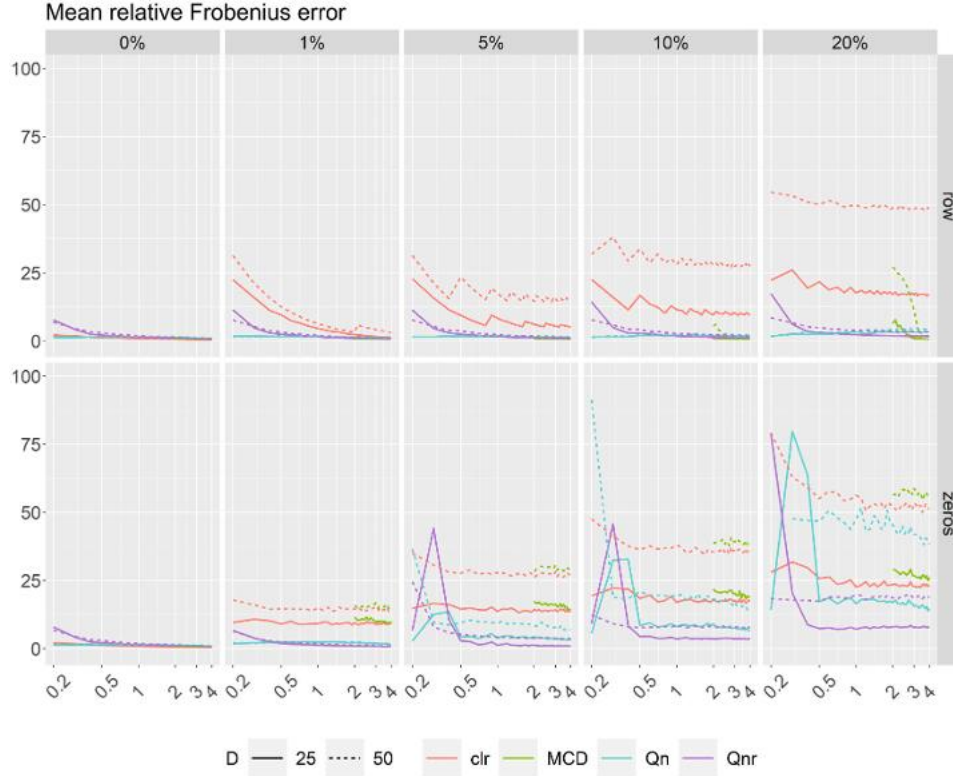


Fig. 1. Results of the simulation study in terms of the mean relative Frobenius error. The classical estimate based on the clr representation (clr) and row-wise robust estimate (MCD) are compared with two newly proposed cell-wise robust estimates (Qn and Qnr). The column blocks correspond to different levels of corruption and the horizontal axis is given by the value of $k = N/D$, with D being number of compositional parts.

Γ having constant expectation. With this choice of Γ and taking $\hat{\Sigma} = \hat{\Sigma}_1^{\text{clr}}$ (here, for simplicity, $\hat{\Sigma}_1^{\text{clr}}$ denotes the estimator in terms of random variables), the expectation in the denominator becomes

$$\Sigma_{i,j=1}^D \mathbb{E} \left(\left(\Gamma_{ij} - (\hat{\Sigma}_1^{\text{clr}})_{ij} \right)^2 \right) = \Sigma_{i,j=1}^D \mathbb{E} \left(\left(-\frac{1}{D} - (\hat{\Sigma}_1^{\text{clr}})_{ij} \right)^2 \right) +$$

$$\Sigma_{i=1}^D \mathbb{E} \left(\left(\left(1 - \frac{1}{D} \right) - (\hat{\Sigma}_1^{\text{clr}})_{ii} \right)^2 \right), \text{ leading to}$$

$$\alpha = \frac{\Sigma_{i,j=1}^D \text{Var} \left((\hat{\Sigma}_1^{\text{clr}})_{ij} \right)}{\Sigma_{i,j=1}^D \mathbb{E} \left(\left((\hat{\Sigma}_1^{\text{clr}})_{ij} + \frac{1}{D} \right)^2 \right) + \Sigma_{i=1}^D \mathbb{E} \left(\left((\hat{\Sigma}_1^{\text{clr}})_{ii} + \frac{1}{D} - 1 \right)^2 \right)}.$$

Recently, also in the non-robust compositional setting, such shrinkage estimators have been proposed (Jin et al., 2022). As an estimate of α requires knowledge of the variance of terms involving $\hat{\Sigma}_1^{\text{clr}}$, and as these might be error-prone due to cell-wise outliers in the data, we propose to replace the variance and expectation by robust counterparts, and to estimate α by

$$\alpha = \frac{\Sigma_{i,j=1}^D \left(\rho_{k=1, \dots, K} \left((\hat{\Sigma}_1^{\text{clr},k})_{ij} \right) \right)^2}{\Sigma_{i,j=1}^D \text{med}_{k=1, \dots, K} \left(\left((\hat{\Sigma}_1^{\text{clr},k})_{ij} + \frac{1}{D} \right)^2 \right) + \Sigma_{i=1}^D \text{med}_{k=1, \dots, K} \left(\left((\hat{\Sigma}_1^{\text{clr},k})_{ii} + \frac{1}{D} - 1 \right)^2 \right)},$$

where med denotes the median, and $\hat{\Sigma}_1^{\text{clr},k}$ is the k -th estimate, for $k \in \{1, \dots, K\}$, and K is the number of bootstrap samples of size N (with replacement), see Wilcox (2010) and Efron and Tibshirani (1994), p 87–108. More precisely, we sample with replacement N times full observations, not cell-wise, from the observed samples $x_1, \dots, x_N$ and

calculate $\hat{\Sigma}_1^{\text{clr}}$ by the proposed algorithm, which we denote instead by $\hat{\Sigma}_1^{\text{clr},k}$, where k denotes the number of bootstrap samples we draw; in our computations we set $K = 500$.

We will refer to the resulting estimator as the cell-wise robust regularized Qn estimator, abbreviated simply as “Qnr”.

4. Simulation study

In this section we conduct a simulation study to compare the performance of the proposed estimators in comparison with other existing non-robust and row-wise robust versions. For different dimensions $D \in \{25, 50\}$ and different numbers of samples $N = Dk$ with $k \in \{0.2, 0.4, 0.6, 0.8, 1, 1.5, 2, 3, 4\}$, we simulate a random covariance matrix Σ^{clr} and a mean vector $\mu^{\text{clr}} \in \mathbb{R}^D$ from a standard Gaussian distribution. More precisely, we generate a $D - 1$ dimensional correlation matrix Σ as in Agostinelli et al. (2015), with code adapted from the R cellWise package (Raymaekers, 2022), and turn it into an admissible covariance matrix in clr coordinates by $\Sigma^{\text{clr}} = V \Sigma V^T$, where V is the same matrix as for an ilr coordinate function. Note that our setup differs in that we include an additional step to turn these into admissible covariance matrices in clr coordinates. Following this, we generate an $N \times D$ data matrix X such that the log-transformation of each row $x_n \in \mathcal{S}^D$ is singular normally distributed $\ln(x_n) \sim \mathcal{N}(\mu^{\text{clr}}, \Sigma^{\text{clr}})$. To corrupt the data, two different setups are used, which we denote in the following by “row-wise” and “zero” corruption.

- Row-wise corruption: A given percentage of observations is randomly selected and replaced by observations that are sampled as

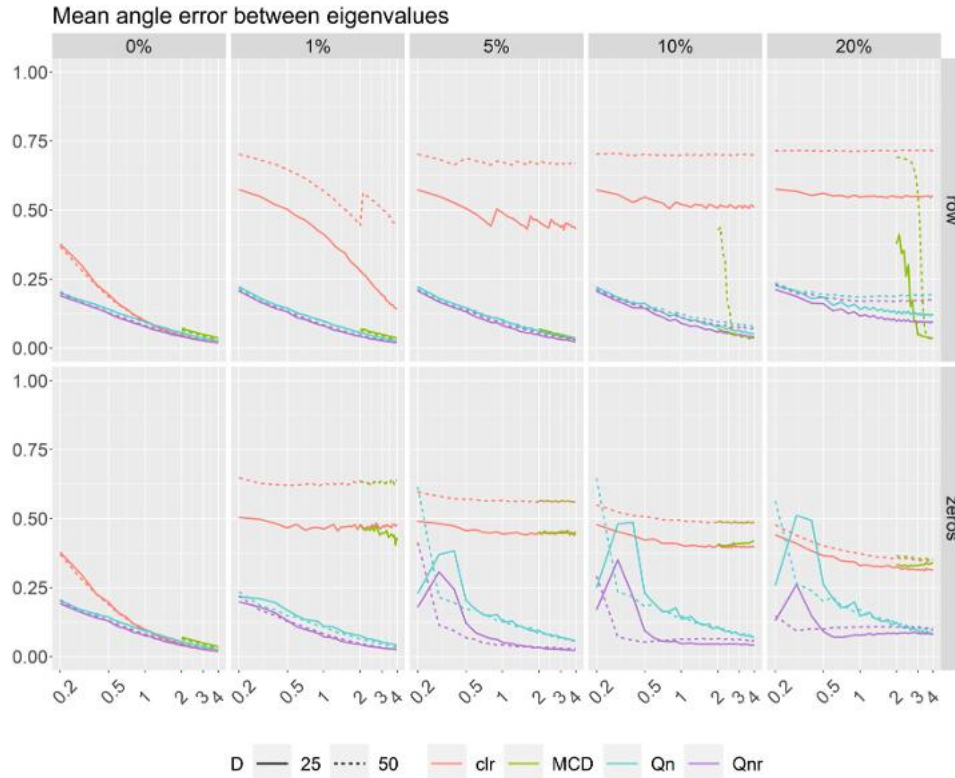


Fig. 2. Results of the simulation study in terms of the mean angle error between eigenvalues. The classical estimate based on the clr representation (clr) and row-wise robust estimate (MCD) are compared with two newly proposed cell-wise robust estimates (Qn and Qnr). The column blocks correspond to different levels of corruption and the horizontal axis is given by the value of $k = N/D$, with D being number of compositional parts.

above, but from another randomly generated covariance matrix, and a different mean sampled from a standard Gaussian distribution.

- Zero corruption: We randomly select a certain percentage of the columns of the data matrix X and set for each selected column all entries below the quantile 0.25 to zero. Finally, in accordance with the common CoDA practice, we replace the zeros by 0.5 times the smallest non-zero entry of the corresponding values of the column.

Thus, with row-wise corruption we would expect that the new estimators show similar performance as a row-wise robust estimator. The setting with zero corruption mimics a typical scenario observed with geochemical data, where a proportion of samples falls below a detection limit for specific variables. Here we should see the benefits of a cell-wise robust method.

Each simulation is repeated 500 times, and we will report averages over these replications. To evaluate the performance of the methods we use mean values of two different error measures between an estimated $\hat{\Sigma}^{\text{clr}}$ and the true $\hat{\Sigma}^{\text{clr}}$. These measures are

- Relative Frobenius error: $\frac{\|\hat{\Sigma}^{\text{clr}} - \Sigma^{\text{clr}}\|_F}{\|\Sigma^{\text{clr}}\|_F}$
- Angle error between eigenvalues: $1 - \frac{\hat{\mathbf{a}}^T \mathbf{a}}{\sqrt{\hat{\mathbf{a}}^T \hat{\mathbf{a}}} \sqrt{\mathbf{a}^T \mathbf{a}}}$

where $\hat{\mathbf{a}}$ resp. $\mathbf{a}$ are the vectors of eigenvalues of $\hat{\Sigma}^{\text{clr}}$ resp. Σ^{clr} . Big values of the relative Frobenius error indicate difficulties in the estimation of the elements of the covariance matrix. The angle error between the eigenvalues is in the interval $[0, 1]$, and a big value means that the shape of

the covariance matrix is not appropriately estimated. From this point of view, the two measures complement each other to a certain extent.

Figs. 1 and 2 compare the considered estimators in terms of mean relative Frobenius error and mean angle error between eigenvalues, respectively. The top panels refer to row-wise corruption, and the bottom panels to zero corruption. In the panels from left to right we increase the amount of contamination. The horizontal axes show the ratio $k = N/D$, and thus in the left part we have a higher-dimensional low-sample size situation, in the right part “classical” tall data matrices with $N > D$. Note that some repetitions (less than 1 %) of the simulation resulted in extremely high values of the relative Frobenius error for the Qn estimator, which can particularly happen in the setting with zeros, leading to degenerate solutions for robust scale estimation. These failures were removed from the following considerations. Based on the respective plots, the performance of the estimators differs between scenarios with and without contamination. When no contamination is present (left column) and $D \leq N$, the classical estimate, denoted in the plots as clr, performs well. Moreover, the mean values of both errors considered are similar to errors obtained for the Qn and Qnr estimators. Slightly worse but still comparable values of the mean angle error are obtained for the MCD estimate. Even in the case of no contamination, but for $D > N$, we can already see an advantage of Qn and Qnr over clr; in this case, the MCD does not give a solution since it is ill-defined for such a situation.

When contamination increases, the classical estimate is naturally distorted by the presence of outliers. For both corruption types, this effect gets more serious with increasing number of variables D and, on the contrary, is quite stable with respect to the N to D ratio. The performance of the other considered estimators depends on the corruption type. The use of the MCD estimator is appropriate for the case of row-wise contamination. The upper panel of Fig. 2 clearly shows that

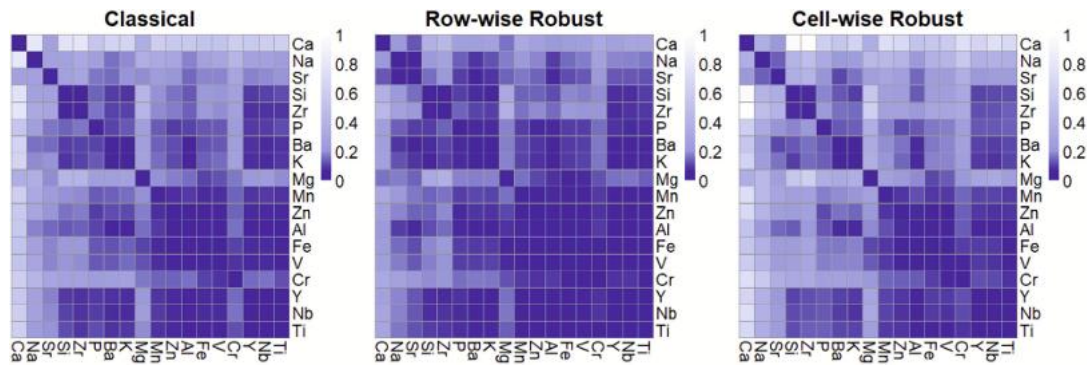


Fig. 3. Variation matrices scaled by the maximum of the entries over all three methods.

particularly with higher percentage of outlying rows, the MCD estimator results in the lowest mean angle error between eigenvalues (if N is sufficiently large). On the other hand, for a lower percentage of the outlying rows, both the Qn estimator and its regularized version Qnr give comparable results as MCD, but these two estimators are also able to cope with settings where D is close or even bigger than N .

The lower panels of Figs. 1 and 2 mimic the scenario of cell-wise outliers, where downweighting of the whole observations may yield too much information loss. The MCD estimator therefore results in a very high error, which is in term of the mean relative Frobenius distance even worse than the one of the classical estimate. The Qn estimator copes with the cell-wise outliers significantly better. This estimator brings valuable results for both high- and low-dimensional settings and its performance enhances with increasing number of observations. The Qn estimate can be refined with its regularized version Qnr, which, particularly under the high-dimensional setting, results in a very low error rate. For all levels of corruption considered and the N to D ratio above 0.5, the performance of the Qnr estimator is comparable to the situation where there is no corruption in the data.

5. Application to a dataset from geochemistry

The performance of the proposed covariance estimators will be demonstrated on a dataset which originates from the large-scale geochemical mapping project GEMAS. This dataset, which is freely available in the R package robCompositions (Templ et al., 2011), consists of the concentration of chemical elements, measured at 2108 locations in most countries of Europe, see Reimann et al. (2014) for details. Here we consider elements measured by XRF (as total element concentrations), and they should provide a close link to the lithology. Elements with more than 3 % of values below the detection limit were excluded, and thus we considered the 18 elements Al, Ba, Ca, Cr, Fe, K,

Mg, Mn, Na, Nb, P, Si, Sr, Ti, V, Y, Zn, and Zr. Our goal here is to compare different versions of estimating the variation matrix and the corresponding covariance matrix, and their effect on Q-mode clustering and on principal component analysis (PCA). All computations were done in the software environment R (R Core Team, 2023).

It is unknown if this dataset contains outliers (row-wise or cell-wise) or not. However, according to the results from the simulations we might not have to expect a huge difference between the row-wise and cell-wise robust covariance estimations.

5.1. Estimates of covariance and variation matrices

As a first step, we compare the estimated covariance and variation matrices:

- *Classical:* Σ^{clr} is estimated by the sample covariance matrix of the clr-transformed data. The estimated variation matrix is obtained by the relationship given in Eq. (6).
- *Row-wise robust:* The data are first expressed in (any) ilr coordinates. Then the MCD estimator is used to obtain a row-wise robust covariance estimate, which is then transformed to the clr representation. The row-wise robust variation matrix is again obtained through Eq. (6).
- *Cell-wise robust:* The cell-wise robust variation and covariance matrices are computed as explained in Section 3.

Note that here we do not compare with a robust covariance based on shrinkage, as introduced in Section 3, because the estimated value of α is essentially zero, implying that no shrinkage would be done, and the number of observations is big enough in comparison to the number of measured elements.

The resulting variation and covariance matrices are visualized as

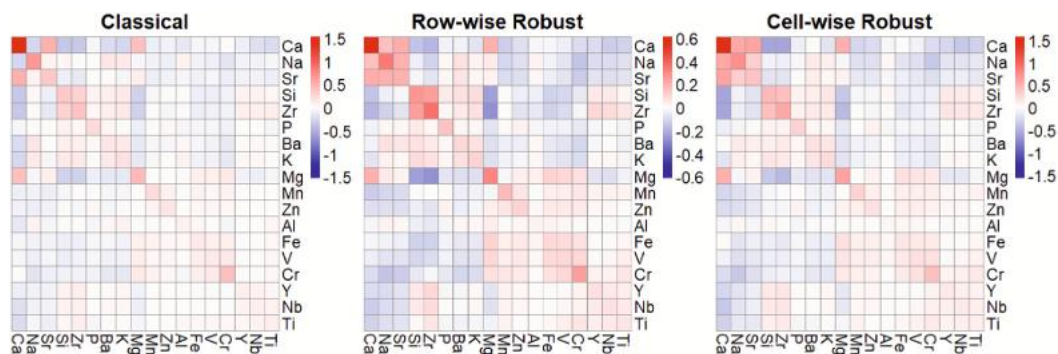


Fig. 4. Clr covariance matrices estimated by the different methods.

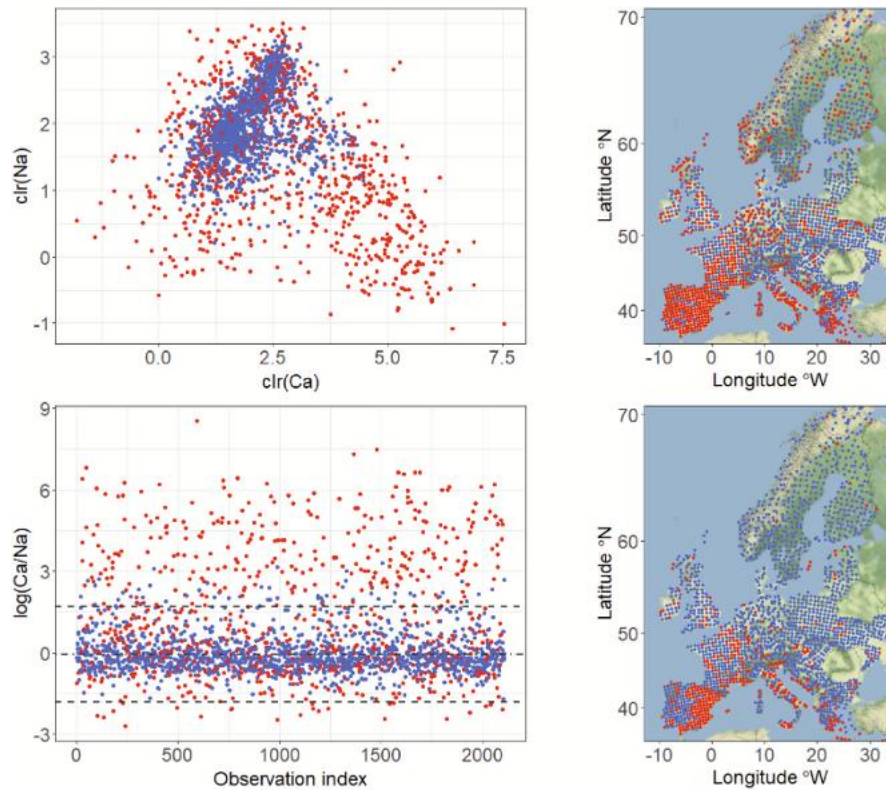


Fig. 5. Top left: clr-transformed Ca versus Na, with row-wise outliers in red; classical covariance is negative, robust one is positive. Top right: row-wise outliers in red in the map. Bottom left: log-ratio Ca/Na versus index, with extremes indicated by values outside median plus/minus 2 times Q_n , with color as above. Bottom right: extremes detected according to value of the Ca to Na log-ratio in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

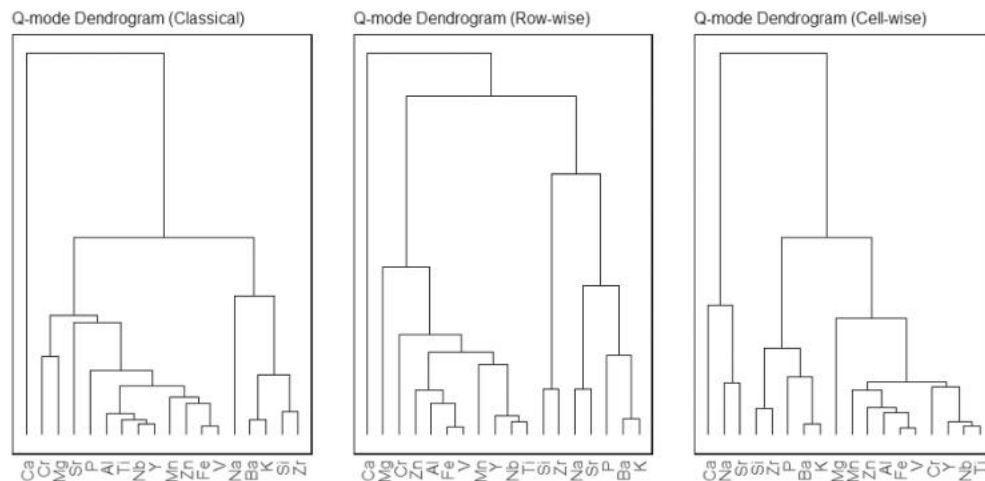


Fig. 6. Dendrograms resulting from the Q-mode clustering, based on the classical (left), row-wise robust (middle) and cell-wise robust (right) estimate of the variation matrix.

heatmaps in Figs. 3 and 4. Note that the elements have been reordered by the order given from the cell-wise robust dendrogram for the Q-mode clustering, see Section 5.2, which might simplify the comparison. The estimated variation matrices, visualized in Fig. 3 with a common normalized scale, reveal that Ca is to some extent different from most of the other elements, and to a lesser extent also Mg and Cr.

The most obvious difference in the covariances in Fig. 4 is the negative relationship between clr representations of Ca and Na in the

classical version, while the robust estimates indicate a positive covariance. This phenomenon is explained visually in Fig. 5. The upper left panel shows the clr-transformed variables Ca and Na, and the red color refers to observations which are identified as outliers with the row-wise robust MCD estimator. Indeed, the two variables show a negative trend if the outliers are included, and a positive if they are downweighted. The upper right panel shows the color information at the sample locations in the map. As an example, a row-wise robust analysis would downweight

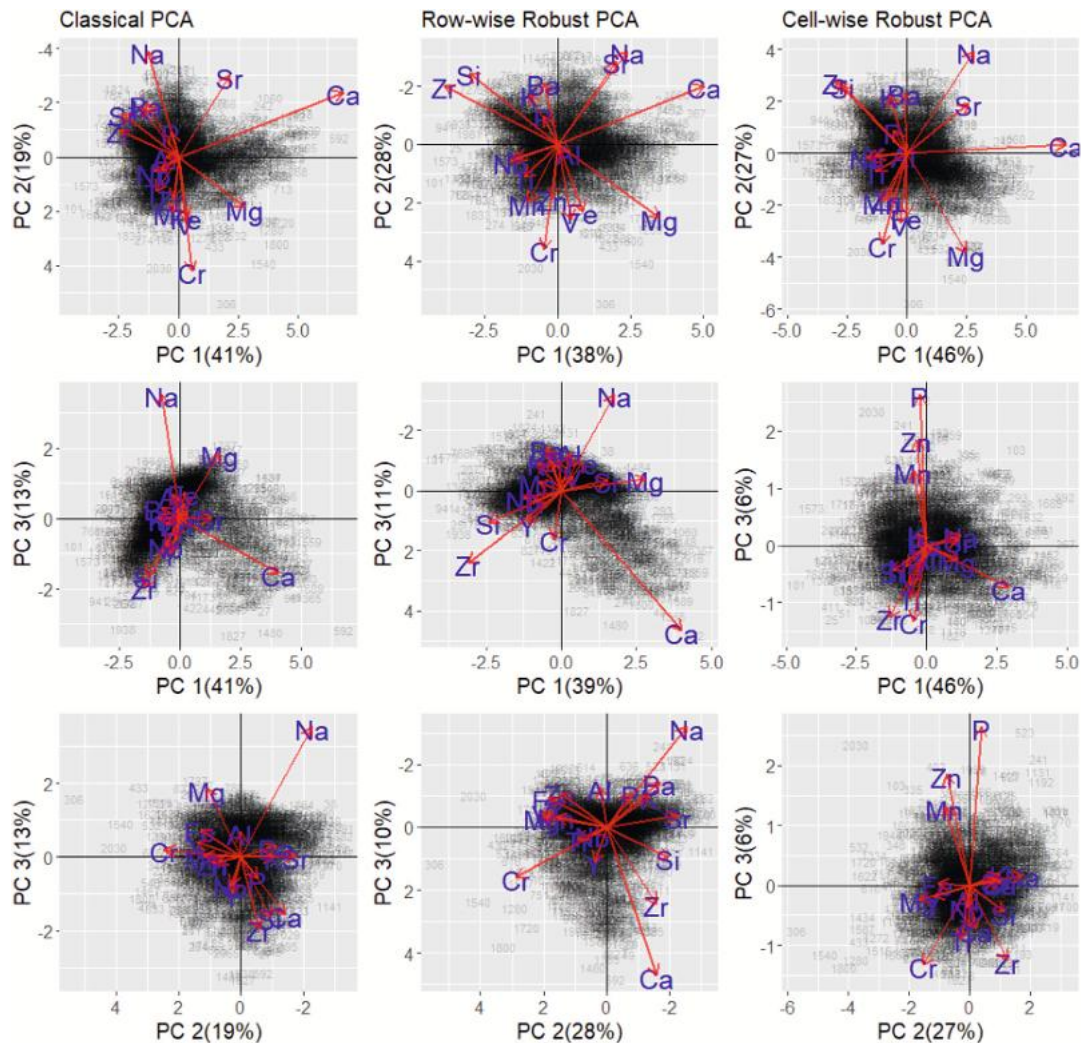


Fig. 7. Biplot of the PCA results based on the classical (left), row-wise robust (middle) and cell-wise robust (right) estimates of the clr covariance matrix.

almost all samples from Spain. The bottom plots try to show the differences to a cell-wise treatment. The lower left panel shows the log-ratio Ca versus Na, for which the Qn estimator is used. The horizontal lines indicate median (dashed), and median ± 2 times the Qn estimator of this log-ratio (dot-dashed). Thus, values outside the solid lines are “extremes” of this log-ratio. The color information is the same as before, and thus we can see that row-wise outliers are not only found in the extremes but in the entire range. The map at the bottom right has as color information the extremes in red and “normal” values of the log-ratio in blue, and it clearly differs from the coloring in the map above. Here, Ca-rich sediments, e.g. in the eastern and southern part of Spain are highlighted and thus downweighted in a cell-wise robust analysis. The cell-wise robust method thus tries to incorporate information of the trend visible in the data majority, but acts pairwise, and thus incorporates also other relevant information of a particular log-ratio.

5.2. Q-mode clustering

Q-mode clustering generally performs a grouping of the variables. In the CoDA context, the basis for Q-mode clustering is the estimated variation matrix, see Fig. 3, and usually Ward’s hierarchical clustering is

used (Filzmoser et al., 2018, p 113). The results are presented as dendrograms in Fig. 6. As already seen in Fig. 3, Ca makes the most important difference in the three versions: while Ca joins as a single element at the very last stage for the classical and the row-wise version, it is in a cluster with Na and Sr in the cell-wise version. From a geochemical point of view, this cluster is meaningful, and we will discuss these and other relationships in more detail for the PCA results.

5.3. Principal component analysis

The estimated covariance matrix is the basis for computing the principal components. Accordingly, differences identified in Fig. 4 are also supposed to be visible in the different versions of PCA (classical, row-wise and cell-wise robust). The corresponding results are shown as biplots in Fig. 7, and the PCA scores are plotted in the maps in Figs. 8–10. The results of the row-wise robust version can also be compared to the results in Reimann et al. (2014), where almost the same elements have been used for row-wise robust PCA.

When comparing the different versions of PCA, we can immediately see a bigger difference from classical PCA to the robust counterparts, which obviously is based on the effect of outliers on the covariance

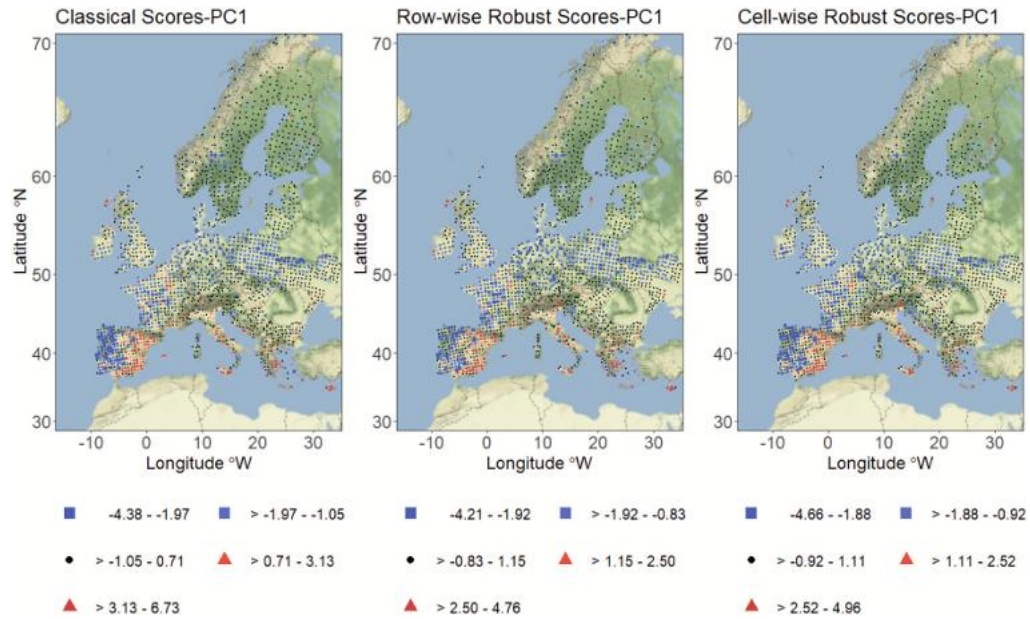


Fig. 8. Scores of PC1 based on the classical (left), row-wise robust (middle) and cell-wise robust (right) estimates of the clr covariance matrix. Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

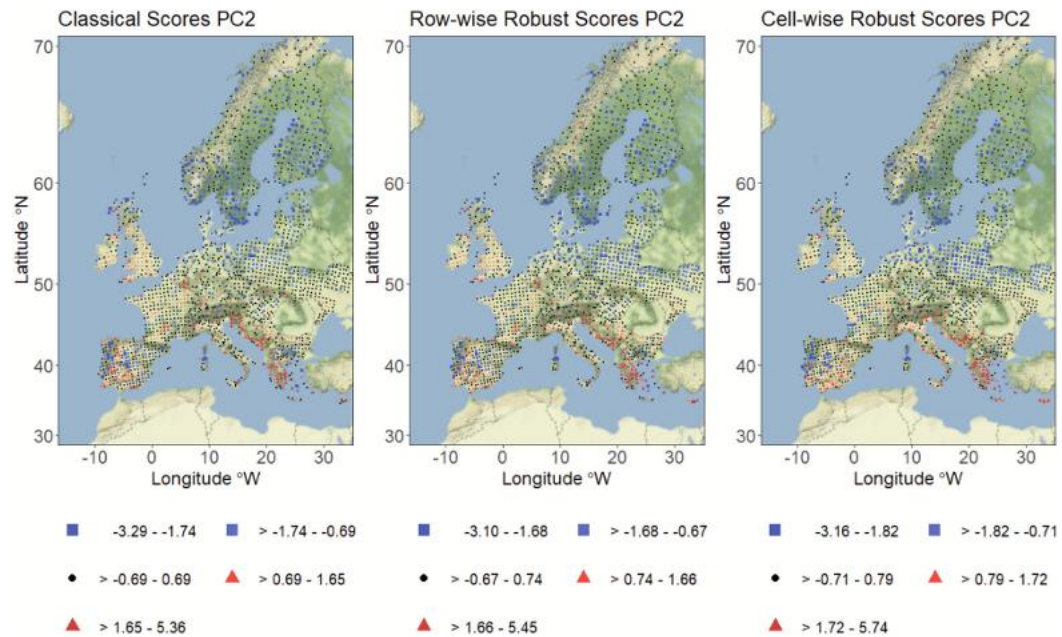


Fig. 9. Scores of PC2 based on the classical (left), row-wise robust (middle) and cell-wise robust (right) estimates of the clr covariance matrix. Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

estimation. Differences between row-wise and cell-wise robust PCA are not so clearly pronounced, but the element Ca plays a key role: in contrast to the row-wise robust version, the element Ca is practically not affecting PC2 and PC3 in the cell-wise robust version. Ca is related to calcareous rocks, which are very diverse as a parent material in the form of chalk, limestone, dolomite, or marble. Thus, having a component which is not involving Ca might be of advantage to better distinguish the parent material.

Focusing now on the results of the cell-wise robust PCA version, we can conclude:

- PC1 has positive loadings for Ca, Sr, Na, Mg, which characterize calcareous rocks (e.g. in eastern and southern Spain, south France, the Apennines in Italy), ophiolites (e.g. in Greece and Cyprus), greenstones and the crystalline less weathered rocks in Scandinavia. Negative contributions from Zr and Si relate to parent materials as

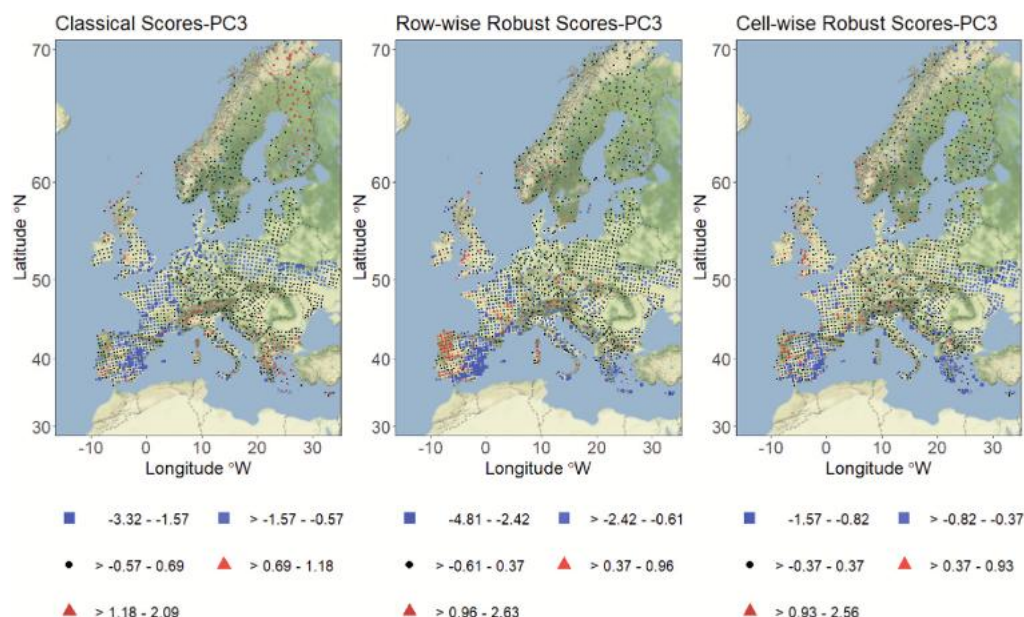


Fig. 10. Scores of PC3 based on the classical (left), row-wise robust (middle) and cell-wise robust (right) estimates of the clr covariance matrix. Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

coarse grained, sandy soil, and soil developed on granite, typically in central and southern Europe.

- PC2 has positive loadings for Na, Sr, Zr, Si, Ba, K, Sr, which are indicative of felsic lithologies. Negative loadings are for Mg, Cr, V, Fe, Zn, Mn, which indicate mafic bedrocks. Overall, this component shows a separation into northern and southern Europe, following the boundary of the last glaciation.
- PC3 is unique, with high loadings for P, Zn and Mn. This might be related to climatic conditions (north-western parts of Spain, western part of UK, south-west of Scandinavia), where wet and cold climate favor the build-up of organic material. Negative loadings are for Ca, Cr, Zr, Ti, indicative of ophiolite (e.g. Greece, Balkan States, Cyprus), but also limestone (southern/eastern Spain).

6. Summary and conclusions

In the field of robust statistics, research in cell-wise robustness, i.e., robustness with respect to outliers in single cells of a data matrix, received a lot of attention in the last years. This started with an investigation on error propagation of outliers in single cells (Alqallaf et al., 2009), where it has been demonstrated that the effect becomes worse especially if the number of variables increases. Even a small proportion of contaminated cells could then result in a high proportion of rows containing outliers (i.e. outlying cells), which could cause breakdown of traditional row-wise robust methods (Rousseeuw and Van Den Bossche, 2018; Agostinelli et al., 2015).

Cell-wise outliers play also an important role in CoDA, and depending on the specific log-ratio transformation they may also propagate to the entire row (observation) (Mert et al., 2016). The goal is to avoid this form of propagation, and to still make use of the non-contaminated cells (log-ratios) in a statistical analysis.

The focus here was on cell-wise robust covariance estimation by robustly estimating the elements of the variation matrix and by making use of the well-known relationship between both matrices. Robust estimation of the variation matrix is straightforward, by simply applying robust variance estimation to all elements of this matrix. What is not so straightforward is to obtain a covariance estimator which (a) fulfills the properties of a clr covariance matrix and (b) has reasonable robustness

properties. We proposed two estimators which are appropriate in the sense of (a), both based on the highly robust and efficient Qn scale estimator. One proposed estimator has an additional regularization, which has advantages especially if the ratio of the number of observations to the number of variables gets small (smaller than 1). Simulations (Section 4) have shown that even without contamination, the estimators are comparable to the classical clr sample covariance estimator, or even preferable if the dimension gets larger than the sample size. In case of contamination, they are clearly preferable over the classical estimator, but in many cases also preferable over a row-wise robust estimator. Particularly if the dimension is close to or even bigger than the sample size, the regularized cell-wise version shows excellent properties, and it achieves good results in cases where the row-wise estimator even cannot be computed. The simulation settings with zero corruption might be of special interest to applications, where detection limit problems occur, and it is surprising how well the regularized Qn estimator performs in this situation.

A covariance estimate is the basis for many multivariate methods, such as PCA, and in CoDA also the variation matrix estimate is very useful, e.g. for Q-mode clustering. Both methods have been demonstrated at a geochemistry dataset, based on classical and robust versions of the estimates. Generally speaking, the principle of robustness is to fit the model to the data majority, and thus to downweight observations which deviate from this majority trend (Maronna et al., 2019, p 195–224). While in row-wise robustness the downweighting is done for complete observations, cell-wise methods here in the variation matrix context only downweight deviating values of particular pairwise log-ratios, and thus they incorporate all “useful” information of the observations from pairwise log-ratios.

One of the proposed estimators of the variation matrix was a shrinkage estimator. Note that the same type of estimator could also be defined in a non-robust manner. The resulting estimator could then serve as a classical counterpart in situations where the dimension exceeds the number of observations. This use-case is of increasing importance in areas where more and more variables are measured, e.g. in bioinformatics (Gloor et al., 2017). In our future research we will investigate this case in more detail, also in combination with sparsity constraints.

CRedit authorship contribution statement

The authors declare that this paper is their own work, and that nobody else is involved.

Declaration of competing interest

The authors declare that they do not have any conflict of interests.

Data availability

The data are already available.

Acknowledgements

KF was supported by the Czech Science Foundation (grant number 22-15684L), and CR and PF by the Austrian Science Fund (project number I 5799-N).

References

- Agostinelli, C., Leung, A., Yohai, V.J., Zamar, R.H., 2015. Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test* 24 (3), 441–461.
- Aitchison, J., 1986. *The Statistical Analysis of Compositional Data*. Chapman & Hall, London.
- Alqallaf, F., Van Aelst, S., Yohai, V.J., Zamar, R.H., 2009. Propagation of outliers in multivariate data. *Ann. Stat.* 37 (1), 311–331.
- Billheimer, D., Guttorm, P., Fagan, W.F., 2001. Statistical interpretation of species composition. *J. Am. Stat. Assoc.* 96 (456), 1205–1214.
- Chen, Y., Wiesel, A., Hero, A.O., 2011. Robust shrinkage estimation of high-dimensional covariance matrices. *IEEE Trans. Signal Process.* 59 (9), 4097–4107.
- Efron, B., Tibshirani, R.J., 1994. *An Introduction to the Bootstrap*. CRC press.
- Egozcue, J.J., Pawłowsky-Glahn, V., Mateu-Figueras, G., Barcelo-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. *Math. Geol.* 35 (3), 279–300.
- Filzmoser, P., Hron, K., Templ, M., 2018. *Applied Compositional Data Analysis: With Worked Examples in R*. Springer.
- Gloor, G.B., Macklaim, J.M., Pawłowsky-Glahn, V., Egozcue, J.J., 2017. Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* 8, 2224.
- Higham, N.J., 1988. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra Appl.* 103, 103–118. [https://doi.org/10.1016/0024-3795\(88\)90223-6](https://doi.org/10.1016/0024-3795(88)90223-6). URL: <https://www.sciencedirect.com/science/article/pii/0024379588902236>.
- Hubert, M., Debruyne, M., Rousseeuw, P., 2018. Minimum covariance determinant and extensions. *Wiley Interdiscip. Rev.: Comput. Stat.* 10 (3), e1421.
- Jin, S., Notredame, C., Erb, I., 2022. Compositional Covariance Shrinkage and Regularised Partial Correlations. <https://doi.org/10.48550/ARXIV.2212.00496>. URL: <https://arxiv.org/abs/2212.00496>.
- Ledoit, O., Wolf, M., 2004. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.* 88 (2), 365–411.
- Maronna, R.A., Martin, R.D., Yohai, V.J., Salibián-Barrera, M., 2019. *Robust Statistics: Theory and Methods (With R)*. John Wiley & Sons.
- Mert, C., Filzmoser, P., Hron, K., 2016. Error propagation in compositional data analysis: theoretical and practical considerations. *Math. Geosci.* 48 (8), 941–961.
- Pawłowsky-Glahn, V., Egozcue, J.J., 2001. Geometric approach to statistical analysis on the simplex. *Stoch. Env. Res. Risk A.* 15, 384–398.
- Pawłowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2015. *Modeling and Analysis of Compositional Data*. John Wiley & Sons.
- R Core Team, 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>.
- Raymaekers, J., 2022. *cellWise: Analyzing Data With Cellwise Outliers*, r Package Version 2.5.0. URL: <https://CRAN.R-project.org/package=cellWise>.
- Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P., 2014. *Chemistry of Europe's Agricultural Soils - Part A: Methodology and Interpretation of the GEMAS Data Set*. Schweizerbart.
- Rousseeuw, P., 1985. Multivariate estimation with high breakdown point. In: Grossmann, W., Pflug, G., Vincze, I., Wertz, W. (Eds.), *Mathematical Statistics and Applications*, pp. 283–297.
- Rousseeuw, P., Croux, C., 1993. Alternatives to the median absolute deviation. *J. Am. Stat. Assoc.* 88 (424), 1273–1283. <https://www.tandfonline.com/doi/pdf/10.1080/01621459.1993.10476408>. URL: <https://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476408>.
- Rousseeuw, P., Van Den Bossche, W., 2018. Detecting deviating data cells. *Technometrics* 60 (2), 135–145.
- Rousseeuw, P., Van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 41 (3), 212–223.
- Schäfer, J., Strimmer, K., 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* 4 (1).
- Štefelová, N., Alfons, A., Palarea-Albaladejo, J., Filzmoser, P., Hron, K., 2021. Robust regression with compositional covariates including cellwise outliers. *ADAC* 15 (4), 869–909.
- Templ, M., Hron, K., Filzmoser, P., 2011. *robCompositions: an R-package for robust statistical analysis of compositional data*. In: Pawłowsky-Glahn, V., Buccianti, A. (Eds.), *Compositional Data Analysis: Theory and Applications*, pp. 341–355.
- Walach, J., Filzmoser, P., Koutil, Š., Friedecký, D., Adam, T., 2020. Cellwise outlier detection and biomarker identification in metabolomics based on pairwise log ratios. *J. Chemom.* 34 (1).
- Wilcox, R.R., 2010. *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy*. Springer.

Robust principal component analysis

de Sousa, J., Hron, K., **Fačevicová, K.**, Filzmoser, P. Robust principal component analysis for compositional tables. *Journal of Applied Statistics*, 48(2), 214-233 (2021). <https://doi.org/10.1080/02664763.2020.1722078>

[IF 1.1 Q3; AIS 0.474 Q3]

[5 citations on WoS; 3 citations without autocitations]

[Author's contribution: 20%]

Abstract

A data table arranged according to two factors can often be considered a compositional table. An example is the number of unemployed people, split according to gender and age classes. Analyzed as compositions, the relevant information consists of ratios between different cells of such a table. This is particularly useful when analyzing several compositional tables jointly, where the absolute numbers are in very different ranges, e.g. if unemployment data are considered from different countries. Within the framework of the logratio methodology, compositional tables can be decomposed into independent and interactive parts, and orthonormal coordinates can be assigned to these parts. However, these coordinates usually require some prior knowledge about the data, and they are not easy to handle for exploring the relationships between the given factors. Here we propose a special choice of coordinates with direct relation to centered logratio coefficients, which are particularly useful for an interpretation in terms of the original cells of the tables. With these coordinates, robust principal component analysis is performed for dimension reduction, allowing to investigate relationships between the factors. The link between orthonormal coordinates and clr coefficients enables to apply rPCA, which would otherwise suffer from the singularity of clr coefficients.



Robust principal component analysis for compositional tables

J. de Sousa ^a, K. Hron ^a, K. Fačevicová ^a and P. Filzmoser ^b

^aDepartment of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, Olomouc, Czech Republic; ^bInstitute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria

ABSTRACT

A data table arranged according to two factors can often be considered a compositional table. An example is the number of unemployed people, split according to gender and age classes. Analyzed as compositions, the relevant information consists of ratios between different cells of such a table. This is particularly useful when analyzing several compositional tables jointly, where the absolute numbers are in very different ranges, e.g. if unemployment data are considered from different countries. Within the framework of the logratio methodology, compositional tables can be decomposed into independent and interactive parts, and orthonormal coordinates can be assigned to these parts. However, these coordinates usually require some prior knowledge about the data, and they are not easy to handle for exploring the relationships between the given factors. Here we propose a special choice of coordinates with direct relation to centered logratio (clr) coefficients, which are particularly useful for an interpretation in terms of the original cells of the tables. With these coordinates, robust principal component analysis (rPCA) is performed for dimension reduction, allowing to investigate relationships between the factors. The link between orthonormal coordinates and clr coefficients enables to apply rPCA, which would otherwise suffer from the singularity of clr coefficients.

ARTICLE HISTORY

Received 13 November 2018
Accepted 21 January 2020

KEYWORDS

Compositional data; compositional table; robust principal component analysis; independence table; interaction table; pivot coordinates

1. Introduction

Many practical data sets (e.g. in economics [12,13], biology [5,19], or sociology [8,32]) consist of observations which contain inherently relative information about the distribution according to two factors (i.e. two random variables in case of distributional data). From a mathematical perspective, this leads to a two-factorial extension of vector compositional data [1,35] carrying information about a relationship between and within these (row and column) factors. In [8], it was shown that such a structure, called a compositional table $\mathbf{x}$,

$$\mathbf{x} = \begin{pmatrix} x_{11} & \cdots & x_{1J} \\ \vdots & \ddots & \vdots \\ x_{I1} & \cdots & x_{IJ} \end{pmatrix}, \quad x_{ij} > 0, i = 1, \dots, I, j = 1, \dots, J, \quad (1)$$

can be decomposed into its independent and interactive parts where their separate analysis can be advantageous for further interpretation concerning both factors and their relationships. Similar as compositional data, compositional tables are driven by the Aitchison geometry [33] which inhibits a direct application of multivariate statistical tools. Instead, an appropriate real coordinate representation, which enables a decomposition into independent and interactive parts, needs to be established with respect to the Aitchison geometry [13].

A frequent primary task in multivariate statistics is to reduce the dimensionality of the data at hand, done using principal component analysis (PCA). As stated above, in case of compositional data, and consequently also compositional tables, this needs to be done in a proper coordinate representation that maps the Aitchison geometry of compositions to a standard Euclidean geometry [33]. To eliminate the influence of outlying observations in PCA, in [16] it was proposed to estimate the covariance matrix for robust PCA by the Minimum Covariance Determinant (MCD) estimator [25]. Since centered logratio (clr) coefficients [1], that aggregate relative information on single compositional parts, lead to singularity and are not appropriate for most robust methods including the MCD estimator, loadings and scores of PCA need to be computed from isometric logratio (ilr) coordinates [10] of the compositional data and then transformed back to clr coefficients for a better interpretation of the resulting compositional biplot.

Accordingly, the aim of this paper is to generalize the previous considerations on dimension reduction of vector compositional data and to propose a robust approach to principal component analysis of compositional tables. While it is obvious that the work can be started from [13,16], it is on the other hand not immediate to see if there is a possibility to identify relationship between clr coefficients and ilr coordinates as it is done in the case of vector compositions. The first issue here is posed by the dimensionality of compositional tables which is much lower than the number of clr coefficients, and the second obstacle is formed by the decomposition where not all choices of coordinates allow for a satisfactory interpretation of both the original and the decomposed tables at the same time (i.e. providing also a way to capture the relationships between them which is necessarily contributing to a better insight into the structure of the tables). It will be shown that the so-called pivot coordinates present a favorable choice of ilr coordinates in line with the previous thoughts and that they can keep a good interpretability when properly linked to clr coefficients.

In Sections 2 and 3, the logratio methodology of vector compositional data and their dimension reduction using principal component analysis, respectively, are briefly reviewed. Compositional tables and their interpretable processing using robust PCA are introduced in Section 4. The new methodology is illustrated in Sections 5 and 6 on real data sets from OECD Statistics using the statistical software R, namely the *robCompositions* package. Data from several different countries containing unemployment information with gender distribution and age structure are processed as a set of 2×4 compositional tables. Therefore, a robust compositional biplot is a possible tool to analyze the distribution of unemployment rates in these countries as well as gender and age differences. Data from the area of education, carrying relative information about fields of study and the resulting degree in given countries, are approached as larger 3×8 compositional tables, and results for men and women are compared.

2. Logratio methodology of compositional data

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

A positive (row) vector $\mathbf{x} = (x_1, x_2, \dots, x_D)$ is defined to be a D -part composition if it carries relative information, i.e. the ratios between the components are informative [1,35]. Any compositional vectors with equal number of parts are considered to be representatives of the same *equivalence class* if one vector is obtained from another by a positive scalar multiplication [35]. Accordingly, equivalence classes of compositional data are represented without loss of information in a D -part simplex,

$$\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, \dots, x_D) \mid x_i > 0, i = 1, \dots, D, \sum_{i=1}^D x_i = \kappa \right\} \quad (2)$$

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

$$\mathcal{C}(\mathbf{x}) = \left(\frac{\kappa \cdot x_1}{\sum_{i=1}^D x_i}, \frac{\kappa \cdot x_2}{\sum_{i=1}^D x_i}, \dots, \frac{\kappa \cdot x_n}{\sum_{i=1}^D x_i} \right) \quad (3)$$

can be applied to rescale the data to a given constant sum (κ) representation. Accordingly, the D -part simplex is a sample space of (representatives of equivalence classes of) compositions. As an interesting consequence of the constant sum representation, possibly outlying observations from the main data cloud are characterized by aberrant ratios rather than by significantly deviating absolute values of components [17].

As mentioned above, relevant information in data with compositional nature is contained only in ratios. Consequently, results of statistical processing should not depend on the sum κ of compositional parts and instead of Euclidean distances, relative differences are used to express distances between observations. This principle called *scale invariance* is the first of three basic compositional principles [35]. Often the original data contain

some non-informative part(s) in the compositional vector that are not of interest. Hence, we do not expect any change of results concerning the respective subcomposition when these parts are removed from the data. *Subcompositional coherence* is a principle stating that results obtained from a d -part subcomposition, $d < D$, are not in contradiction with results obtained by an analysis of the original D -part composition. Finally, *permutation invariance* states that the results are independent from a chosen order of parts within the composition, an expectable assumption for any reasonable statistical processing.

The above principles and the relative scale of compositional data should be captured by a meaningful geometric structure, preferably following the properties of the Euclidean vector space. This is provided by the Aitchison geometry [10]. Operations of *perturbation* and *power transformation*, being analogous to a sum of two vectors and a multiplication of a vector by a scalar in the real Euclidean geometry, are defined as

$$\mathbf{x} \oplus \mathbf{y} = (x_1 y_1, \dots, x_D y_D), \quad \alpha \odot \mathbf{x} = (x_1^\alpha, \dots, x_D^\alpha), \quad (4)$$

where $\mathbf{x}$ and $\mathbf{y}$ are D -part compositions, and α is a real constant. Accordingly, operations of perturbation and power transformation form a $(D - 1)$ -dimensional vector space $(\mathcal{S}^D, \oplus, \odot)$ [35].

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

$$\begin{aligned} \langle \mathbf{x}, \mathbf{y} \rangle_A &= \frac{1}{2D} \sum_{i=1}^D \sum_{j=1}^D \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}, \quad \|\mathbf{x}\|_A = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_A}, \\ d_A(\mathbf{x}, \mathbf{y}) &= \|\mathbf{x} \ominus \mathbf{y}\|_A, \end{aligned} \quad (5)$$

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

Given the introduced specifics of compositional data endowed with the Aitchison geometry, standard multivariate statistical methods cannot be applied directly on raw data. It is advisable to employ the *working on coordinates* principle; that is to represent the compositional data in real coordinates before starting with a statistical analysis [26,34]. There are two types of isometric coordinate representations with respect to the Aitchison geometry. *Centered logratio coefficients* [1] are defined as

$$\text{clr}(\mathbf{x}) = \left(\ln \frac{x_1}{g(\mathbf{x})}, \ln \frac{x_2}{g(\mathbf{x})}, \dots, \ln \frac{x_D}{g(\mathbf{x})} \right), \quad (6)$$

where $g(\mathbf{x})$ stands for the geometrical mean of the whole composition. Each clr coefficient aggregates all pairwise logratios with a given compositional part, thus enabling for a simple and meaningful interpretation in terms of dominance of that part with respect to the other parts *on average*. Consequently, clr coefficients are useful for a graphical interpretation of compositional data including compositional biplots as a result of a dimension reduction using PCA [2]. On the other hand, clr coefficients sum up to zero which leads to a singular covariance matrix, being inappropriate for processing using common robust statistical methods [16,17].

Fortunately, *isometric logratio coordinates* $\mathbf{z} \in \mathbf{R}^{D-1}$ representing orthonormal coordinates with respect to the Aitchison geometry can be derived as

$$\mathbf{z} = \text{ilr}(\mathbf{x}) = (\langle \mathbf{x}, \mathbf{e}^1 \rangle_A, \langle \mathbf{x}, \mathbf{e}^2 \rangle_A, \dots, \langle \mathbf{x}, \mathbf{e}^{D-1} \rangle_A), \quad (7)$$

where D -part compositions $\mathbf{e}^i = \mathcal{C}(e_1^i, e_2^i, \dots, e_D^i)$, $i = 1, \dots, D-1$, form an orthonormal basis on the simplex.

Obviously, the interpretation of ilr coordinates might be more tricky than in the case of clr coefficients as there are infinitely many possibilities of constructing ilr coordinates depending on the choice of basis vectors $\mathbf{e}^i$, $i = 1, \dots, D-1$. *Sequential binary partitioning* (SBP) of compositional parts is one possibility for providing a meaningful choice of $\mathbf{e}^i$ for the practitioner which is corresponding to the prior knowledge about compositions and resulting in coordinates called balances [9]. Moreover, there is a linear transformation between ilr coordinates and clr coefficients, done through a $D \times (D-1)$ matrix $\mathbf{V}$ of clr representations of the ilr basis vectors,

$$\text{clr}(\mathbf{x}) = \mathbf{V}\mathbf{z} = [\text{clr}(\mathbf{e}_1)^T, \text{clr}(\mathbf{e}_2)^T, \dots, \text{clr}(\mathbf{e}_{D-1})^T] \cdot \text{ilr}(\mathbf{x}). \quad (8)$$

Recently, *pivot coordinates* [18,20]

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^D x_j^{(l)}}} \quad (9)$$

were introduced as a special case of ilr coordinates. They are appropriate especially in situations where no prior knowledge about how to perform SBP is available, e.g. in [3,4,7,23]. Here, $x_i^{(l)}$ refers to the i -th part of the re-ordered composition $(x_l, x_1, \dots, x_{l-1}, x_{l+1}, \dots, x_D)$ which can be rewritten as $(x_1^{(l)}, x_2^{(l)}, \dots, x_l^{(l)}, x_{l+1}^{(l)}, \dots, x_D^{(l)})$. In each of the D coordinate systems, a permutation of compositional parts needs to be performed, so that the l -th part of $\mathbf{x}$ ($l = 1, \dots, D$) stands at the first position. Accordingly, the first pivot coordinate in each system, $z_1^{(l)}$, then clearly explains all relative information about part x_l and, additionally, it is proportional to the respective clr coefficient

$$\text{clr}(x_l^{(l)}) = \sqrt{\frac{D}{D-1}} z_1^{(l)}. \quad (10)$$

Expression (8) holds also for pivot coordinates which is particularly useful in the context of this paper.

3. Robust principal component analysis for compositional data

When dealing with large-scale data sets, dimension reduction is often of primary interest. PCA is one of the widely used methods for this purpose also in a compositional approach, converting possibly correlated original variables from the data at hand into a smaller set of linearly uncorrelated variables called principal components (PCs). Additionally, the first component accounts for the largest variance of the given data, the second one for a maximum of the remaining variance, etc., under the constraint of being orthogonal to all the previous principal components [22].

The covariance matrix C estimated from a real data matrix X can be spectrally decomposed into

$$C = GLG^T, \quad (11)$$

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

$$X^* = (X - \mathbf{1}t^T)G, \quad (12)$$

where t is the location estimator and $\mathbf{1}$ is a vector of ones with length n . The columns of the matrix X^* , the coordinates of the principal components, are called *scores* and the columns of G , containing the respective basis vectors, are called *loadings*. Typically, only the first few principal components are considered for further analysis. Taking into account only two PCs, a graphical outcome called *biplot* can depict both loadings as arrows and scores as points in one plot, where associations can be revealed.

It is common to take t as the arithmetic mean and C as the sample covariance matrix, however, both are very sensitive to outlying observations. Robust alternatives can be obtained by using the *Minimum Covariance Determinant (MCD)* estimators of location and covariance [25]. However, this approach inquires working in ilr coordinates to obtain full rank data in order to get the MCD estimate of the covariance matrix and the respective matrix of eigenvectors G . In addition, ilr coordinates ensure subcompositional coherence and enable to keep affine equivariance of the results to the change of basis.

Accordingly, robust principal component analysis of compositional data based on the MCD estimator requires ilr coordinates z_i as an input, and the scores z_i^* are given by

$$z_i^* = G^T(z_i - t). \quad (13)$$

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

$$G_{\text{clr}} = VG,$$

accounting for compositional biplot construction with meaningful interpretation, whereas the scores remain identical and only a column of zeros is added to the end. Clr coefficients are also worth as such for their simple construction as an amalgamation of pairwise logratios of a given part. Due to the zero-sum constraint of clr coefficients, their covariance structure is distorted, thus the interpretation of the biplot in terms of the correlation between coefficients (through angles between arrows) might be misleading. Instead, the focus is on links between vertices of arrows as they stand for a proportionality between the original compositional parts [2]. On the other hand, due to the link with pivot coordinates, the single clr variables (or the respective loadings) can be used to identify observations with a high dominance of the respective parts in a compositional vector [24].

4. Compositional tables

Compositional tables describe quantitatively relative contributions to a given whole that is distributed according to two (row and column) factors. Mathematically, this leads to a two-factorial extension of vector compositional data, carrying information about a relationship between these factors. A compositional table (1) thus can be represented, e.g. either as a

contingency table (with sufficiently high numbers of counts in the cells) or as a table of the same order with ML estimates of the respective probabilities – due to scale invariance, the relative information (contained in the ratios between the cells) is the same in both cases [8,11]. Hence, the concept of compositional tables (1) covers both the discrete case of contingency tables and its continuous counterpart (e.g. input–output tables, see [12]). Nevertheless, in the compositional context, a concrete table represents usually just one realization in a sample from a multivariate continuous distribution. Due to the decision to treat the data from such a distribution compositionally, the possible order of the factor categories (e.g. age levels) is ignored, making this a relevant subject for future research.

Since compositional tables form a direct extension of vector compositional data, all the principles and operations introduced in Section 2 apply, up to some minor modifications due to the two-factorial structure of the tables.

Accordingly, the closure operation

$$\mathcal{C}(\mathbf{x}) = \begin{pmatrix} \frac{\kappa x_{11}}{\sum_{i,j} x_{ij}} & \cdots & \frac{\kappa x_{1J}}{\sum_{i,j} x_{ij}} \\ \vdots & \ddots & \vdots \\ \frac{\kappa x_{I1}}{\sum_{i,j} x_{ij}} & \cdots & \frac{\kappa x_{IJ}}{\sum_{i,j} x_{ij}} \end{pmatrix} \quad (14)$$

is used to represent a compositional table $\mathbf{x}$ in an IJ -part simplex $\mathcal{S}^{IJ}$ of vectorized tables $\text{vec}(\mathbf{x}) = (x_{11}, \dots, x_{I1}, \dots, x_{IJ})$. Perturbation $\mathbf{x} \oplus \mathbf{y}$, powering $\alpha \odot \mathbf{x}$, and the Aitchison inner product $\langle \mathbf{x}, \mathbf{y} \rangle_A$ of two tables $\mathbf{x}, \mathbf{y}$ and a real number α can be defined analogously [8,11]:

$$\begin{aligned} \mathbf{x} \oplus \mathbf{y} &= \begin{pmatrix} x_{11}y_{11} & \cdots & x_{1J}y_{1J} \\ \vdots & \ddots & \vdots \\ x_{I1}y_{I1} & \cdots & x_{IJ}y_{IJ} \end{pmatrix}, \quad \alpha \odot \mathbf{x} = \begin{pmatrix} x_{11}^\alpha & \cdots & x_{1J}^\alpha \\ \vdots & \ddots & \vdots \\ x_{I1}^\alpha & \cdots & x_{IJ}^\alpha \end{pmatrix}, \\ \langle \mathbf{x}, \mathbf{y} \rangle_A &= \frac{1}{2IJ} \sum_{i,j} \sum_{k,l} \ln \frac{x_{ij}}{x_{kl}} \ln \frac{y_{ij}}{y_{kl}}. \end{aligned} \quad (15)$$

It is straightforward to derive that the dimension of the simplex $\mathcal{S}^{IJ}$ is $IJ-1$, corresponding to the dimensionality of $(I \times J)$ -compositional tables.

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

To analyze compositional tables, it is beneficial to work also with the so-called *independence* and *interaction* tables. These can be obtained through an orthogonal decomposition [8]

$$\mathbf{x} = \mathbf{x}_{\text{ind}} \oplus \mathbf{x}_{\text{int}}. \quad (16)$$

Here, the independence table is constructed to extract all the relative information about row and column factors under the assumption that the original compositional table is a product of its row and column geometric marginals, and the interaction table contains information about the relationships between the row and column factors. Therefore, in

case of actual independence in the data at hand (in the above sense), all the entries of the interaction table are the same, since there is no remaining information left in the data after the decomposition; the interaction table thus forms a neutral element with respect to the Aitchison geometry of compositional tables. Otherwise, the interactive part describes the nature of the deviation from an independent situation. Needless to say, analyzing each of these new tables separately allows for a deeper insight into the original data.

It turns out that the introduced decomposition can be easily derived from row and column projections of the compositional table onto marginal subspaces (for further details, see [8]),

$$\text{row}^\perp(\mathbf{x}) = \begin{pmatrix} g(x_{11}, \dots, x_{1I}) & \cdots & g(x_{11}, \dots, x_{1J}) \\ \cdots & \cdots & \cdots \\ g(x_{I1}, \dots, x_{IJ}) & \cdots & g(x_{I1}, \dots, x_{IJ}) \end{pmatrix}, \quad (17)$$

$$\text{col}^\perp(\mathbf{x}) = \begin{pmatrix} g(x_{11}, \dots, x_{1I}) & \vdots & g(x_{1J}, \dots, x_{IJ}) \\ \vdots & \vdots & \vdots \\ g(x_{I1}, \dots, x_{I1}) & \vdots & g(x_{IJ}, \dots, x_{IJ}) \end{pmatrix}, \quad (18)$$

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

Recalling the case of independence in probability tables, it is instant to get the independence table simply by perturbing both these projections, $\mathbf{x}_{\text{ind}} = \text{row}^\perp(\mathbf{x}) \oplus \text{col}^\perp(\mathbf{x})$. From (16), it follows that the interaction table is just a decomposition remainder in $\mathbf{x}_{\text{int}} = \mathbf{x} \ominus \mathbf{x}_{\text{ind}}$. For practical calculations, the following formulas are used to obtain the single entries of these tables

$$\begin{aligned} x_{ij}^{\text{ind}} &= \left(\prod_{k=1}^I \prod_{l=1}^J x_{kl} x_{il} \right)^{1/IJ} \propto \left(\prod_{k=1}^I x_{kj} \right)^{1/I} \left(\prod_{l=1}^J x_{il} \right)^{1/J}, \\ x_{ij}^{\text{int}} &= \left(\prod_{k=1}^I \prod_{l=1}^J \frac{x_{ij}}{x_{kl} x_{il}} \right)^{1/IJ}. \end{aligned} \quad (19)$$

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

As stated in the previous sections, a coordinate representation which respects the sample space dimensionality is needed to perform robust PCA of compositional data. In case of compositional tables (and particularly their decomposed parts), a generalization of balance coordinates needs to consider two SBPs according to each factor [14]. Even for moderate numbers of rows and columns, the interpretation of this coordinate representation gets rather complex without a deeper expert knowledge. Therefore, only a two-factorial alternative to pivot coordinates is usually used in practice [13,14]. Interestingly, the coordinates of the entire compositional table can be divided into two groups according to the

dimensionality of the independence and interaction tables, respectively. This becomes the main advantage also when using pivot coordinates for robust PCA since it allows for a comparison of the results from the whole table and its decomposed parts.

Generally, there are three types of pivot coordinates corresponding to the row, column and ‘odds ratio’ partitioning of the compositional table [13]. The first two types together form a coordinate representation of the independence table, the third one is used for the interaction table. Altogether, they form a coordinate representation of the original compositional table. In case of row and column types of coordinates, the entire first row or column, respectively, is taken as the pivot element and separated from the rest. In the next step, this pivot is not considered anymore and the following row or column is taken as the new pivot element, and so on, until the following $I + J - 2$ coordinates are obtained,

$$z_i^r = \sqrt{\frac{(I-i)J}{1+I-i}} \ln \frac{g(\mathbf{x}_{i\bullet})}{[g(\mathbf{x}_{i+1\bullet}), \dots, g(\mathbf{x}_{I\bullet})]^{1/(I-i)}}, \quad i = 1, \dots, I-1, \quad (20)$$

$$z_j^c = \sqrt{\frac{I(J-j)}{1+J-j}} \ln \frac{g(\mathbf{x}_{\bullet j})}{[g(\mathbf{x}_{\bullet j+1}), \dots, g(\mathbf{x}_{\bullet J})]^{1/(J-j)}}, \quad j = 1, \dots, J-1, \quad (21)$$

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

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

$$z_{rs}^{\text{OR}} = \sqrt{\frac{1}{(I-r)(J-s)(I-r+1)(J-s+1)}} \ln \prod_{i=r+1}^I \prod_{j=s+1}^J \frac{x_{ij}x_{rs}}{x_{is}x_{rj}}. \quad (22)$$

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

For the sake of completeness, permutations of the entire rows or columns following the same principle as stated in Section 2 could be performed. Hereby for all combinations of rows and columns, different coordinate systems consisting of $z_i^{r(k)}$, $z_j^{c(l)}$ and $z_{rs}^{\text{OR}(kl)}$, where (kl) , $k = 1, \dots, I$, $l = 1, \dots, J$, defines row and column permuted to the pivot position within the whole table, would be gained [13]. Following (10), also the first coordinates of the three types from each system can then be expressed as proportional (up to a constant) to respective clr coefficients,

$$\text{clr}(\mathbf{x}_{\text{ind}})_{kl} = \sqrt{\frac{I-1}{IJ}} z_1^{r(k)} + \sqrt{\frac{J-1}{IJ}} z_1^{c(l)}, \quad (23)$$

$$\text{clr}(\mathbf{x}_{\text{int}})_{kl} = \sqrt{\frac{(I-1)(J-1)}{IJ}} z_{11}^{\text{OR}(kl)}, \quad (24)$$

which is an important fact for the interpretation of the analysis. The resulting clr coefficients, computed originally from the elements of the independence and interaction tables,

$$\text{clr}(\mathbf{x}_{\text{ind}})_{ij} = \ln \frac{x_{ij}^{\text{ind}}}{g(\mathbf{x}_{\bullet\bullet}^{\text{ind}})}, \quad \text{clr}(\mathbf{x}_{\text{int}})_{ij} = \ln \frac{x_{ij}^{\text{int}}}{g(\mathbf{x}_{\bullet\bullet}^{\text{int}})}, \quad (25)$$

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

$$\text{clr}(\mathbf{x}_{\text{ind}})_{ij} = \ln \frac{g(\mathbf{x}_{i\bullet})g(\mathbf{x}_{\bullet j})}{g(\mathbf{x}_{\bullet\bullet})^2}, \quad \text{clr}(\mathbf{x}_{\text{int}})_{ij} = \ln \frac{x_{ij}g(\mathbf{x}_{\bullet\bullet})}{g(\mathbf{x}_{i\bullet})g(\mathbf{x}_{\bullet j})}, \quad (26)$$

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

Furthermore, note that each coordinate $\text{clr}(\mathbf{x}_{\text{ind}})_{ij}$ is formed by the sum of clr coefficients of the respective row and column marginals, $1/J \sum_j \text{clr}(\mathbf{x}_{\text{ind}})_{ij} = \ln g_{i\bullet}/g_{\bullet\bullet}$ and $1/I \sum_i \text{clr}(\mathbf{x}_{\text{ind}})_{ij} = \ln g_{\bullet j}/g_{\bullet\bullet}$, which amount to zero. Thus, there are only $I + J - 2$ linearly independent clr coefficients, reflecting the dimensionality of the sample space of independence tables again. A similar feature holds also for clr coefficients of interaction tables that sum up to zero across each row or column. Consequently, in the case of an interaction table, the number of linearly independent clr coefficients reduces to $(I-1)(J-1)$. Since this dependency makes it impossible to use the clr coefficients for the robust PCA of independence and interaction tables, the strategy to perform robust PCA for compositional tables is the same as in case of vector compositional data: PCA loadings and scores are computed in ilr coordinates and then back-transformed using relation (8) to the clr space, where the loadings can be interpreted in terms of dominance of single cells. Here, clr coefficients of basis vectors for rows $\mathbf{e}^r$, columns $\mathbf{e}^c$ and interactions $\mathbf{e}^{\text{OR}}$, forming the columns of the matrix $\mathbf{V}$, are defined as follows,

$$\text{clr}(\mathbf{e}^r) = \begin{cases} \sqrt{\frac{I-i}{(I-i+1)J}} & \text{for the elements in pivot row } i, \\ -\sqrt{\frac{1}{(I-i+1)J(I-i)}} & \text{for the elements in rows } i+1, \dots, I, \\ 0 & \text{otherwise,} \end{cases} \quad (27)$$

$$\text{clr}(\mathbf{e}^c) = \begin{cases} \sqrt{\frac{J-j}{(J-j+1)I}} & \text{for the elements in pivot column } j, \\ -\sqrt{\frac{1}{(I-i+1)J(I-i)}} & \text{for the elements in columns } j+1, \dots, J, \\ 0 & \text{otherwise,} \end{cases} \quad (28)$$

and

$$\text{clr}(\mathbf{e}^{\text{OR}}) = \begin{cases} \sqrt{\frac{1}{rs(r-1)(s-1)}} & \text{for the elements on positions } i = r + 1, \dots, I, \\ & j = s + 1, \dots, J \\ \sqrt{\frac{(r-1)(s-1)}{rs}} & \text{for the pivot elements } rs \\ -\sqrt{\frac{r-1}{rs(s-1)}} & \text{for the elements in pivot row } r, \\ & j = s + 1, \dots, J, \\ -\sqrt{\frac{s-1}{rs(r-1)}} & \text{for the elements in pivot column } s, \\ & i = r + 1, \dots, I, \\ 0 & \text{otherwise,} \end{cases} \quad (29)$$

reinterpreting the expressions from [13]. As a result of (8), row-wise clr coefficients of the whole table are obtained for the $IJ-1$ columns of the matrix $\mathbf{V}$. Alternatively, if the matrix $\mathbf{V}$ has just $I + J - 2$ columns formed by clr coefficients of basis vectors corresponding to the ilr representation of the independence table (20), (21), its respective clr coefficients are derived (and similarly for the interaction table with its coordinates (22)). Finally, the transformed loadings and scores can be used to construct a biplot in order to reveal the multivariate structure of the sample of compositional tables and relations between both factors.

5. Unemployment data analysis

In the following, the methodological results are applied to two real-world data sets from the field of economy in order to illustrate the main features and possible limitations of the approach. However, it is important to acknowledge here the potential of the proposed methodology also across other research areas since compositional tables (eventually in form of their count counterparts, i.e. contingency tables) occur in many applications and sciences. For example, in environmental management, the presence of two-factor compositional data was recognized already by Aitchison [1] where areal compositions given by simplified 2×2 tables of vegetation (thick and thin) and animals (dense, sparse) abundance in different regions were provided as an example of compositional data with more complex structures. From rather up-to-date environmental management problems, analyses of e.g. 5×4 tables of material resources [30] given by the extraction type (domestic extraction, import, export, direct material input, and domestic material consumption) and group of the resources (biomass, fossil energy carriers, non-metallic minerals, and metals); 2×7 tables of protected areas [31] characterized by the domain of biodiversity (terrestrial or marine) and designation of the protected area (e.g. nature reserve, wilderness, protected landscape, etc.); or 34×4 tables of carbon emissions embodied in trade [29] diversified by sector (e.g. agriculture, mining, food products, etc.) and measure (imported, exported, consumption-based and production-based emissions), available always for the vast majority of OECD member states, could be mentioned as more relevant representatives.

The first analyzed data set, the Unemployed data set, is coming again from OECD Statistics and contains aggregated data from more than 150 million unemployed people from 42

Table 1. Unemployed people in thousands partitioned according to their gender and age groups [27].

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

different countries in 2010 [27]. It is analyzed using the statistical software environment R [36].

The data contain the numbers of unemployed people together with their gender and age category for the following countries: Australia, Austria, Belgium, Canada, Chile, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Mexico, Netherlands, New Zealand, Norway, Poland, Portugal, Slovakia, Slovenia, Spain, Sweden, Switzerland, Turkey, United Kingdom, United States, Colombia, Costa Rica, Latvia, Lithuania, China, India, Indonesia, Russian Federation and South Africa. An example of (transposed) raw data from the first four countries is shown in Table 1. The numbers in the tables are basically counts of unemployed people according to two factors. As the population size varies among the countries, the interest here is not in the absolute values of the counts in the single countries, but rather the relative structure of unemployment. Particularly, ratios of men and women and ratios among age groups 15–24, 25–39, 40–54 and 55+, as well as proportionality among countries will be analyzed. Since outliers can be anticipated due to completely different economics, education levels, gender balance and also traditions of the listed countries, the analysis will be carried out in a robust manner.

All compositional tables in this example have two rows and four columns, i.e. gender is the row factor and age structure is the column factor. The sample space of tables thus has dimension 7 out of which independence tables account for a dimension of 4 with pivot coordinates z_1^r , z_1^c , z_2^c and z_3^c , while the remaining coordinates z_{11}^{OR} , z_{12}^{OR} and z_{13}^{OR} correspond to the interaction tables with a dimension of 3.

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

In order to perform PCA in ilr coordinates, the standard function `princomp` in R can be used, where the parameter `covmat` is set to `covMcd` (MCD estimator of covariance) in case of robust PCA. Thereafter, loadings need to be transformed to clr coefficients as described in Section 2 using the matrix V with columns defined by (27) – (29) for the entire compositional table, and by (27) and (28), or by (29) for its independent and interactive part, respectively. The resulting classical biplots are depicted on the right-hand side of Figure 1, while the robust PCA output is on the left.

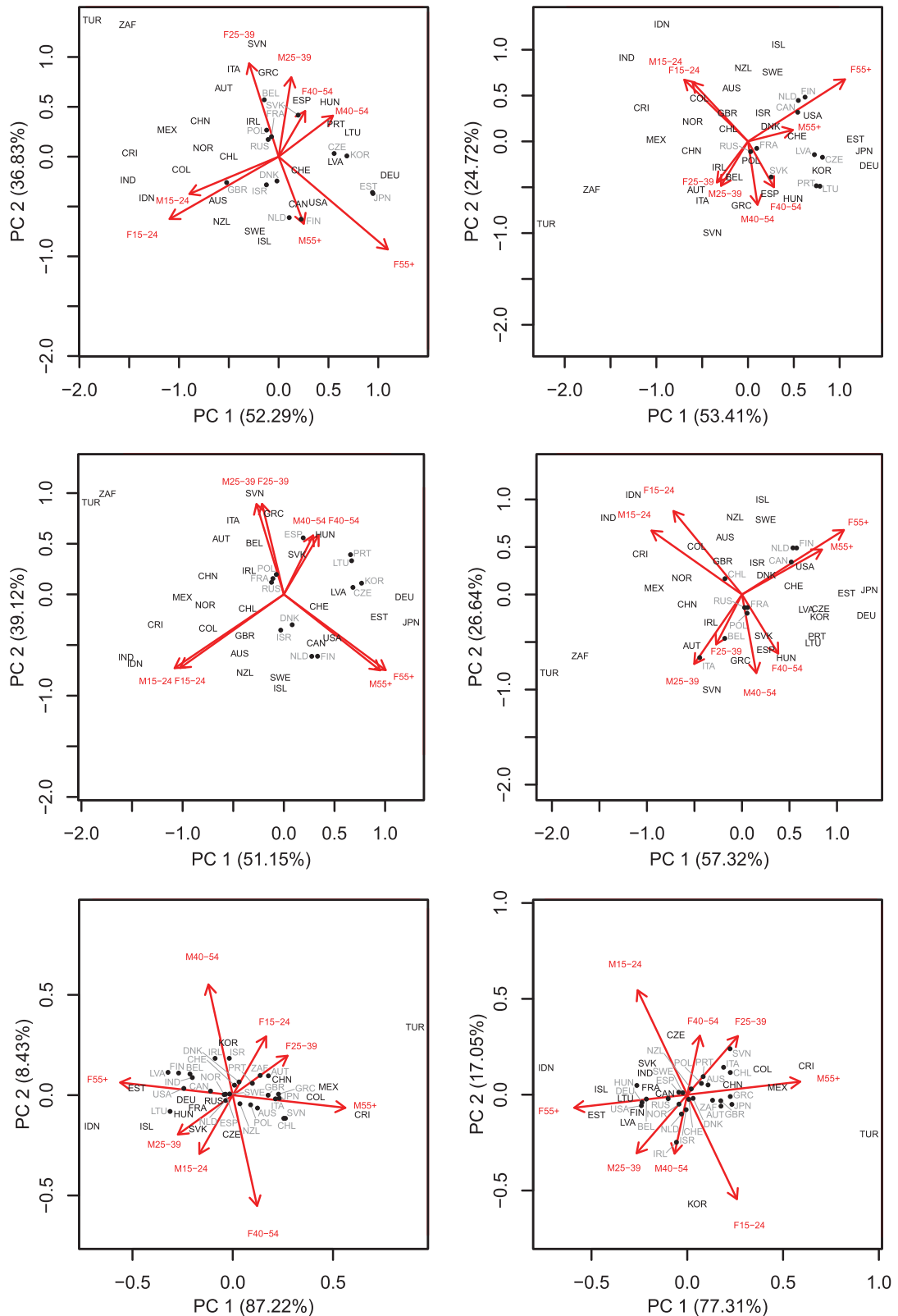


Figure 1. Robust (left column) and classical (right column) covariance biplots of the Unemployment compositional (upper row), independence (middle row) and interaction tables (lower row).

Assessing Figure 1, it can be noticed that in all three cases, robust PCA performs better in terms of explained variability by the first two principal components. As mentioned above, some outliers might be present in the data, and even from the classical biplot of the whole compositional tables (upper right corner of Figure 1), at least two outlying tables (Turkey-TUR and South Africa-ZAF) could be expected, and thus the robust approach should provide more meaningful results. However, an outlier detection is performed additionally in order to confirm these expectations. In the R package *robCompositions* [37], there is a function `outCoDa` defined for this purpose, based on robust Mahalanobis distances computed from *ilr* transformed data [15]. Using the pivot coordinates and applying the 0.975 quantile of the chi-squared distribution as the common cut-off value, 15 out of all 42 countries are identified as outlying observations, clearly supporting the choice of robust analysis. Note that similarly, 10 observations from the set of independence tables and six from the interaction tables were detected as potential outliers.

Additionally, from the same part of the figure, it is easy to identify from the direction of the arrows which countries tend to have relatively higher unemployment among younger people and which ones have a rather higher rate in the opposite situation. Although no clear compact clusters are visible, it seems that most European countries together with the USA and Canada tend to have more likely problems with employing older people, say 40+, while for Central and South America together with China, India, and Indonesia the unemployment depending on age structure has rather opposite tendencies. In the latter case, this generalized finding corresponds to the values of the youth unemployment rate (i.e. $YUR = (\text{unemployed between 15 and 24 years}) / (\text{all unemployed})$) in different countries reported by the United Nations in 2011 [38]. Nevertheless, there are still big differences between European countries, even more apparent in the robust analysis. Moreover, also some gender differences can be observed, except for the youngest generation. The structure in the classical biplot (upper right plot) is similar but driven by the identified outlying observations.

The left plot in the middle part of Figure 1 shows the ‘ideal’ situation in case the relationships between gender and age factors would be filtered out. While the positions of the countries are not apparently changed compared to the previously discussed covariance biplot (upper right corner), the general relationships between the factors are remarkably illustrative. In case of independence, nearly gender equity would be achieved, while on the contrary, relationships among the age levels would be disproportionately weaker. Also, a bigger difference between results provided by robust and classical PCA is present here. One can easily understand how the classical approach does not handle outliers and how those can affect the output; the biplot on the right side is quite far away from picturing the same ideal situation.

As demonstrated, the independence table captures the hypothetical balanced state with each *clr* interpreted in terms of dominance of a given combination of factors in case of independence. However, this dominance is then either amplified, or weakened according to the interaction table; in terms of *clr* coefficients, it depends on whether the logratio dominance is shifted in a positive, or in a negative direction. Note that information obtained from the interaction table only does not provide a complete picture about the dominance of single cells in the table. For example, in the lower left biplot (robust biplot of interaction tables), Costa Rica (CRI) is placed towards the loading ‘male 55+’, but this does not necessarily lead to a conclusion that unemployment in this group is higher in general in

this country; it simply marks the cell whose dominance causes imbalance for Costa Rica, although the actual proportion of unemployment for this age group can be lower than its average dominance. Therefore, the conclusion about the higher dominance of unemployed men in the oldest group than expected in the hypothetical case of independence can be stated only after looking at the biplot of the independence tables. For the compositional tables with dimension $2 \times J$ (or alternatively $I \times 2$), this feature is nicely illustrated by the depicted loadings themselves, placed along a line corresponding to increasing dominance of one factor value at the expense of the latter value of the same factor. The opposite relation between the respective clr coefficients is clearly visible from both the biplots and the form of $\text{clr}(\mathbf{x}_{\text{int}})_{ij}$ in (26): as it was already discussed in Section 4, clr coefficients of the interaction table sum up to zero across each row and column which results in the identity $\text{clr}(\mathbf{x}_{\text{int}})_{1j} = -\text{clr}(\mathbf{x}_{\text{int}})_{2j}$, holding for each j when $I = 2$ (and similarly for $J = 2$). While in case of higher data dimension the property is no longer visible in the graphs, in this example it can be seen that the two possible values of the gender factor lead to precisely contradictory loadings for any chosen value of the age factor. Thus they might only carry the information about the origin of the dominance shift, but no longer about the direction of the shift for which the difference from the independence table has to be consulted.

6. Education data analysis

It was illustrated in the previous example how outliers can affect results of classical PCA. Especially the gender equity achieved in the robust biplot of the independence tables would not be present in the classical one. Hence, in this second example, only robust analysis outputs are discussed. The data set contains information about more than seven million female and nearly six million male students, divided according to eight different fields of study, being Education, Humanities and arts, Social sciences, business and law, Science, mathematics and computing, Engineering, manufacturing and construction, Agriculture and veterinary, Health and welfare, and Services [28]. The information about the achieved degree (bachelor, master, and doctoral, respectively) is recorded as well for about 30 different countries.

Compositional tables are analyzed for both genders separately in order to allow for a comparison of possible differences between them later on. Biplots as graphical PCA outcomes of the whole compositional table as well as independence and interaction tables are collected in Figure 2. Due to a larger dimensionality of compositional tables than in previous case (3×8), the biplots contain three times more variables and an objective interpretation becomes more difficult. An additional aspect is that since it is necessary to go many dimensions down to achieve the PCA projection using the first two principal components, it is expected to obtain more approximative picture of the multivariate data structure in the biplot. However, for data of similar or even bigger size, the proposed methods still offer an extremely useful rank-two approximation capturing the relationships between both factors. The performance of robust PCA is still good enough (at least 54.08%) for this concrete case in terms of cumulative variability explained by the first two principal components.

Despite the previous interpretational doubts, it can be seen that the effect of the chosen final degree is possibly stronger than the effect of the study field since the loadings tend to create a quite clear division of bachelors, masters, and doctors for most of the biplots.

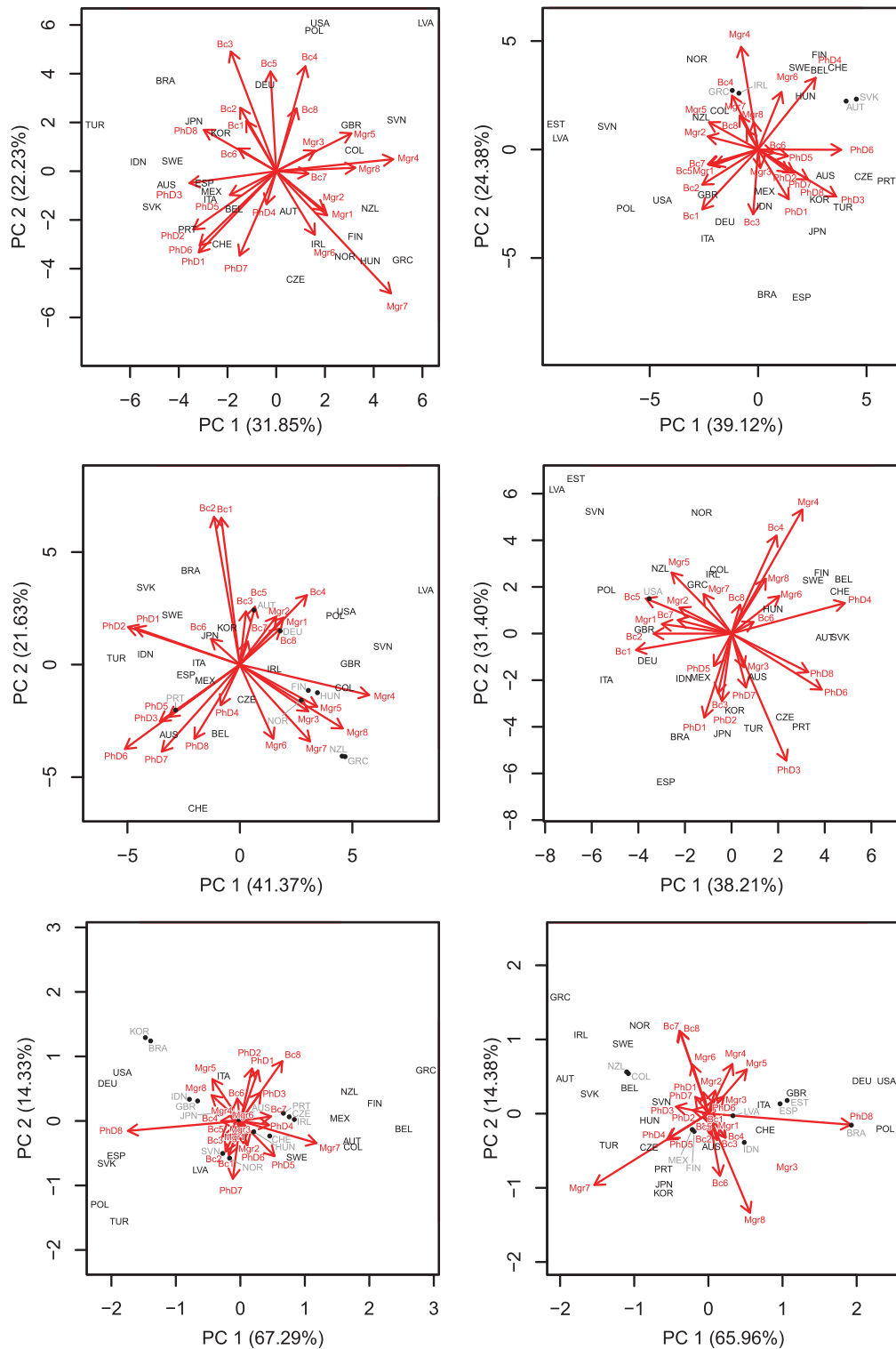


Figure 2. Robust covariance biplots of the Education compositional (upper row), independence (middle row) and interaction tables (lower row) for men (left column) and women (right column), respectively. Study fields are marked as follows: 1 = Education, 2 = Humanities and arts, 3 = Social sciences, business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

This property is more obvious for men while for women the difference between bachelors and masters is partially wiped out, maybe also due to a less compact data structure. Finally, employing the `outCoDa` function [15,37] again, some outliers for the data set of the whole compositional tables are detected: Austria, Norway and Spain for men, and United Kingdom, Turkey and United States for women.

From the second row of the figures, some overall idea about the hypothetical state of independence between degree and study field factors might be obtained. A stronger effect of the chosen degree and a weaker effect of the study field would still be apparent for men, and one new feature could be observed: there would be a strong relationship between Education and Humanities and arts study fields for each degree. For women, e.g. a similarity of educational systems in Sweden, Finland, Belgium and Switzerland is reflected. Also, in the case of independence, higher occurrence of outliers would be present for both men and women.

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

7. Conclusions

Vector compositional data, or their generalization to compositional tables, are characterized by their relative scale and scale invariance properties captured by the Aitchison geometry. Accordingly, a representation of the compositional tables isometrically in clr coefficients or ilr coordinates is essential for the further statistical analysis using popular multivariate methods. Compositional tables can be decomposed onto their independence and interaction parts; a statistical analysis of both is recommended to get insight into the ideal situation when relationships between factors are filtered away, as well as into interactions between factors forming the compositional table.

As most real-world data sets contain outlying observations, robust methods requiring an orthonormal coordinate representation have been considered. To reduce the dimension of data at hand, a robust PCA using the MCD estimator can be applied to pivot coordinates of compositional, independence and interaction tables (but also any other ilr coordinates which respect the decomposition of compositional tables could be used). The necessity of respecting dimensionality of independence and interaction tables forms the main difference to (vector) compositional data where such feature does not occur. In case of pivot coordinates, it is also possible to find a simple relationship between them and the respective clr coefficients as well as to enhance interpretation of the latter variables. It is precisely this specific choice of orthonormal coordinates where coordinates of independence and interaction tables form together coordinates of the entire compositional tables

(with the additional benefit of linkage to clr coefficients constructed in the same manner) which prepare a solid ground for their further generalization in the sense of various weighting techniques. An analogous approach is already being used for vector compositional data [20] where the paper [24] provides a way for their respective PCA analysis by constructing biplots from the first pivot coordinates resulting from different (mutually orthogonal) sets of (weighted) coordinates.

Here, loadings obtained in (ordinary) pivot coordinates for the robust PCA are transformed back to clr coefficients where they are used for the construction of biplots and their meaningful analysis. In case of $(2 \times J)$ table dimensions, an additional feature can be observed in the graphical output of interaction tables, which can be traced back to the interpretation of the clr coefficients. As a next step, three factors of a compositional cube $(I \times J \times K)$ could be analyzed by robust PCA and appropriate 3D graphical tools employed. Consequently, the educational data sets for men and women could be merged into a single data set and approached accordingly. This work could be perhaps started from a recent idea of the robust Parafac model for three-way compositional data as a generalization of the PCA, proposed in [6], combined with a robust approach to PCA as presented in [21]. The case of compositional cubes and also its further extension to generally N factors belong to primary research interests of the authors.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by COST Action CRONoS [grant number IC1408] and Mathematical Models of the Internal Grant Agency of the Palacký University in Olomouc [grant number IGA_PrF_2019_006].

ORCID

J. de Sousa  <http://orcid.org/0000-0001-8067-5742>

K. Hron  <http://orcid.org/0000-0002-1847-6598>

K. Fačevicová  <https://orcid.org/0000-0002-5294-1115>

P. Filzmoser  <https://orcid.org/0000-0002-8014-4682>

References

- [1] J. Aitchison, *The Statistical Analysis of Compositional Data*, Chapman and Hall, London, 1986.
- [2] J. Aitchison and M. Greenacre, *Biplots of compositional data*, J. R. Stat. Soc. C-Appl. 51 (2002), pp. 375–392.
- [3] F. Bruno, F. Greco and M. Ventrucci, *Spatio-temporal regression on compositional covariates: modeling vegetation in a gypsum outcrop*, Environ. Ecol. Stat. 22 (2015), pp. 445–463.
- [4] A. Buccianti, J.J. Egozcue and V. Pawłowsky-Glahn, *Variation diagrams to statistically model the behavior of geochemical variables: Theory and applications*, J. Hydrol. (Amst) 519 (2014), pp. 988–998.
- [5] T. Dickhaus, K. Straßburger, D. Schunk, C. Morcillo-Suarez, T. Illig and A. Navarro, *How to analyze many contingency tables simultaneously in genetic association studies*, Stat. Appl. Genet. Mol. Biol. 11 (2012), pp. 3026–3034.

- [6] M.A. Di Palma, P. Filzmoser, M. Gallo and K. Hron, *A robust Parafac model for compositional data*, J. Appl. Stat. 45 (2018), pp. 1347–1369.
- [7] D. Dumuid, T.E. Stanford, J.A. Martín-Fernández, Ž. Pedišić, C.A. Maher, L.K. Lewis, K. Hron, P.T. Katzmarzyk, J.P. Chaput, M. Fogelholm, G. Hu, E.V. Lambert, J. Maia, O.L. Sarmiento, M. Standage, T.V. Barreira, S.T. Broyles, C. Tudor-Locke, M.S. Tremblay and T. Olds, *Compositional data analysis for physical activity, sedentary time and sleep research*, Stat. Methods. Med. Res. (2018). doi:10.1177/0962280217710835.
- [8] J.J. Egozcue, J.L. Díaz-Barrero and V. Pawlowsky-Glahn, *Compositional analysis of bivariate discrete probabilities*, in *Proceedings of CODAWORK'08, The 3rd Compositional Data Analysis Workshop*, J. Daunis-i-Estadella, J.A. Martín-Fernández, eds., University of Girona, Spain, 2008.
- [9] J.J. Egozcue and V. Pawlowsky-Glahn, *Groups of parts and their balances in compositional data analysis*, Math. Geol. 37 (2005), pp. 795–828.
- [10] J.J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras and C. Barceló-Vidal, *Isometric logratio transformations for compositional data analysis*, Math. Geol. 35 (2003), pp. 279–300.
- [11] J.J. Egozcue, V. Pawlowsky-Glahn, M. Templ and K. Hron, *Independence in contingency tables using simplicial geometry*, Commun. Stat. Theory 44 (2015), pp. 3978–3996.
- [12] K. Fačevicová, K. Hron, V. Todorov, D. Guo and M. Templ, *Logratio approach to statistical analysis of 2 x 2 compositional tables*, J. Appl. Stat. 41 (2014), pp. 944–958.
- [13] K. Fačevicová, K. Hron, V. Todorov and M. Templ, *Compositional tables analysis in coordinates*, Scand. J. Stat. 43 (2016), pp. 962–977.
- [14] K. Fačevicová, K. Hron, V. Todorov and M. Templ, *General approach to coordinate representation of compositional tables*, Scand. J. Stat. 45 (2018), pp. 879–899.
- [15] P. Filzmoser and K. Hron, *Outlier detection for compositional data using robust methods*, Math. Geosci. 40 (2008), pp. 233–248.
- [16] P. Filzmoser, K. Hron and C. Reimann, *Principal component analysis for compositional data with outliers*, Environmetrics 20 (2009), pp. 621–632.
- [17] P. Filzmoser and K. Hron, *Robustness for compositional data*, in *Robustness and Complex Data Structures*, C. Becker, R. Fried, and S. Kuhnt, eds., Springer, Berlin, 2013, pp. 117–131.
- [18] E. Fišerová and K. Hron, *On interpretation of orthonormal coordinates for compositional data*, Math. Geosci. 43 (2011), pp. 455–468.
- [19] C. Herder, W. Rathmann, K. Strassburger, H. Finner, H. Grallert, C. Huth, C. Meisinger, C. Gieger, S. Martin, G. Giani, W.A. Scherbaum, H.E. Wichmann and T. Illig, *Variants of the PPARG, IGF2BP2, CDKAL1, HHEX, and TCF7L2 genes confer risk of type 2 diabetes independently of BMI in the German KORA studies*, Horm. Metab. Res. 40 (2008), pp. 722–726.
- [20] K. Hron, P. Filzmoser, P. de Caritat, E. Fišerová and A. Gardlo, *Weighted pivot coordinates for compositional data and their application to geochemical mapping*, Math. Geosci. 49 (2017), pp. 797–814.
- [21] M. Hubert, P.J. Rousseeuw and K. Vanden Branden, *ROBPCA: A new approach to robust principal component analysis*, Technometrics 47 (2005), pp. 64–79.
- [22] R. Johnson and D. Wichern, *Applied Multivariate Statistical Analysis*, 6th ed., Prentice-Hall, London, 2007.
- [23] A. Kalivodová, K. Hron, P. Filzmoser, L. Najdekr, H. Janečková and T. Adam, *PLS-DA for compositional data with application to metabolomics*, J. Chemom. 29 (2015), pp. 21–28.
- [24] P. Kynčlová, P. Filzmoser and K. Hron, *Compositional biplots including external non-compositional variables*, Statistics 50 (2016), pp. 1132–1148.
- [25] R. Maronna, R.D. Martin and V.J. Yohai, *Robust Statistics: Theory and Methods*, Wiley, New York, 2006.
- [26] G. Mateu-Figueras, V. Pawlowsky-Glahn and J.J. Egozcue, *The principle of working on coordinates*, in *Compositional Data Analysis: Theory and Applications*, Wiley, Chichester, 2011, pp. 31–42.
- [27] OECD Statistics, *Unemployment by sex and age – 2010*. Available at <http://stats.oecd.org/>.
- [28] OECD Statistics, *Education and training – 2010*. Available at <http://stats.oecd.org/>.

- [29] OECD Statistics, *Policy indicators of trade and environment – Carbon emissions embodied in trade – 2011*. Available at <http://stats.oecd.org/>.
- [30] OECD Statistics, *Environment – Material resources – 2017*. Available at <http://stats.oecd.org/>.
- [31] OECD Statistics, *Environment – Biodiversity of protected areas in km² – 2018*. Available at <http://stats.oecd.org/>.
- [32] M.I. Ortego and J.J. Egozcue, *Bayesian estimation of the orthogonal decomposition of a contingency table*, Aust. J. Stat. 45 (2016), pp. 45–56.
- [33] V. Pawlowsky-Glahn and J.J. Egozcue, *Geometric approach to statistical analysis on the simplex*, Stoch. Env. Res. Risk A. 15 (2001), pp. 384–398.
- [34] V. Pawlowsky-Glahn, *Statistical modelling on coordinates*, Universitat de Girona, 2003. Available at <http://ima.udg.es/Activitats/CoDaWork2003/>.
- [35] V. Pawlowsky-Glahn, J.J. Egozcue and R. Tolosana-Delgado, *Modeling and Analysis of Compositional Data*, Wiley, Chichester, 2015.
- [36] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018. Available at <https://www.R-project.org/>.
- [37] M. Templ, K. Hron and P. Filzmoser, *robCompositions: An R-package for robust statistical analysis of compositional data*, in *Compositional Data Analysis. Theory and Applications*, V. Pawlowsky-Glahn and A. Buccianti, eds., Wiley, Chichester, 2011, pp. 341–355.
- [38] UNdata, *Youth unemployment, both sexes – 2011*. Available at <http://data.un.org/DocumentData.asp/x?id=264#30>.

Zero handling strategies

Tang, W., **Fačevicová, K.**, Nordhausen, K., Taskinen, S. A critical comparison of handling zeros in high-dimensional compositional count data, *Chemo-metrics and Intelligent Laboratory Systems*, 277, 105821 (2026).

<https://doi.org/10.1016/j.chemolab.2026.105821>

[IF 3.8 D1; AIS 0.588 Q3]

[0 citations on WoS; 0 citations without autocitations]

[Author's contribution: 20%]

Abstract

The growing use of high-throughput sequencing (HTS) has enabled the large-scale production of compositional count data, driving progress in microbiome research. However, such count data are often high-dimensional, over-dispersed, and heavily zero-inflated, and they conflict with the continuity assumptions underlying log-ratio-based compositional data analysis (CoDA), creating substantial methodological challenges. This review organizes the scattered literature into a structured overview of established zero-handling strategies, categorizing them into zero-tolerant transformations, imputation approaches for rounded zeros, and statistical models for essential zeros. We further highlight the theoretical and practical consequences of applying continuous log-ratio frameworks to discrete compositional count data. Theoretically, violations of continuity may induce numerical instability and distort the geometric assumptions underlying CoDA. Practically, we examine how existing continuous imputation strategies behave when applied to lattice-valued count data within the classical log-ratio framework, comparing their performance in terms of matrix reconstruction and downstream analyses. Our results show that reconstruction accuracy alone does not always reflect

downstream analytical performance. Overall, this review consolidates current zero-handling strategies and clarifies their appropriate scope of application. It also identifies methodological gaps that motivate future developments for high-dimensional compositional count data.



A critical comparison of handling zeros in high dimensional compositional count data

Wenqiang Tang^{a,*}, Kamil Falcão^b, Klaus Nordhausen^c, Sina Takinen^a

^aDepartment of Mathematics and Statistics, University of Jyväskylä, Finland

^bDepartment of Mathematical Analysis and Applications of Mathematics, Polytechnic University of Coimbra, Portugal

^cDepartment of Mathematics and Statistics, University of Helsinki, Finland

ARTICLE INFO

Direct link: <https://github.com/nitrogenae/CDARtools>

Keywords

Compositional data analysis

Zeros

High-throughput sequencing (HTS)

Microbiome count data

Imputation

ABSTRACT

The growing use of high-throughput sequencing (HTS) has led to the large-scale production of compositional count data, which poses increasing challenges. However, such count data often exhibit high dimensionality, over-dispersion, and heavy zero-inflation, and they conflict with the continuity assumptions underlying log-ratio-based compositional data analysis (CDA), causing substantial methodological challenges. This review organizes the scattered literature into a structured overview of established handling strategies, categorizing them into zero-deletion, transformation, imputation, and probabilistic methods, and statistical models for zero-inflated counts. We further highlight the theoretical and practical consequences of applying continuous log-ratio analysis to discrete compositional count data. Theoretical violations of continuity may induce financial instability and distort the generic assumptions underlying CDA. Practically, we examine how existing continuous imputation strategies have been applied to false-valued count data within the classical log-ratio framework, comparing their performance in terms of matrix reconstruction and downstream analyses. Our results show that continuous imputation may indeed cause a delicate downstream analytical performance. Overall, this review identifies current handling strategies and cautions their application to count data, and it identifies methodological gaps that motivate future developments for high-dimensional compositional count data.

Contents

1	Introduction	2
2	Compositional Data Analysis (CDA)	3
21	Log-ratio analysis and its limitations	3
3	Review of methods dealing with zeros in CDA	4
31	Classical methods for zero handling	4
32	Imputation methods for count data	4
321	Multiplicative replacement	4
322	Geometric regression with log-ratio Minimization (EM) algorithms	4
323	ARD transformation (ARD)	5
324	Log-ratio square-root transformation (LSQ)	5
325	Hybrid statistical model for zero-inflated counts (ANN)	5
33	Deletion of zero-inflated counts	6
34	Transformation of zero-inflated counts	6
341	Over-dispersed count models for compositional data	6
342	Bayesian (directed) transformation for compositional data	7
35	Count models	7
351	Probabilistic models for zero-inflated count compositional data	7
352	Bayesian multiplicative imputation methods	8

* Corresponding author.

E-mail addresses: wentian@jyu.fi (W. Tang), kamil.falcao@polz.pt (K. Falcão), klaus.nordhausen@helsinki.fi (K. Nordhausen), sina.takinen@jyu.fi (S. Takinen).

<https://doi.org/10.1016/j.cml.2025.105821>

Received 6 February 2025; Received in revised form 2 July 2025; Accepted 18 July 2025

Available online 29 July 2026

0169-7439/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

353	Challenges of applying GDA to code court data	18
4	Comparison of imputation and zero handling	18
41	Methodological considerations and evaluation scope	18
42	Terminology	18
43	Discussion and conclusions	18
44	Results of imputation and zero handling	18
45	Assessment of downstream PCA and clustering analysis	18
46	Discussion and recommendations	18
5	Conclusion	18
	Call for further contributions	18
	Declaration of generative AI and AI-assisted technologies in the manuscript preparation process	18
	Declaration of competing interest	18
	Acknowledgment	18
	Appendix A: Simulation design for zero imputation comparison	18
A1	Construction of zero handling simulation dataset	18
A2	Results of imputation and zero handling	18
A3	Results of downstream analysis	18
	Appendix B: Additional diagnostic figures	18
	Appendix C: Additional clustering analysis using Gaussian mixture models	18
	Data availability	18
	References	18

1 Introduction

Recent advances in modern sampling and classification techniques, such as high-throughput sequencing (HTS) [1–3], which allow rapid and large-scale generation of DNA/RNA sequence data [4], have greatly advanced modern research. The data obtained from these techniques typically consist of thousands of millions of sequencing reads of single repeating multiple base pairs within a microbial community. In these studies, the data are often characterized by hundreds of millions of reads across samples [5]. The sequencing data, which quantify the number of sequences assigned to individual taxa or genes, form the basis for downstream analyses linking microbial composition to host phenotypes, environmental factors, and clinical outcomes.

The sequencing approach entails the partitioning of HTS data into biologically meaningful units. In practice, analyses of microbial data are often performed at various taxonomic units (OTUs) or taxonomic units (ASVs). These relative count data exhibit unique statistical characteristics that present analytical challenges. First, sequencing data are compositional because sequencing results are generated based on a fixed sequencing depth. As the reads are distributed among different sequences, only a relative rather than absolute proportion is preserved [4]. This feature necessitates statistical methods inappropriate when applied directly to raw sequencing data without suitable normalization and transformation.

Second, the relative nature of compositional data. As discussed in [6], the relative abundance of taxa in a sample is a key feature of compositional data, but the absolute values (see Section 2). However, when applied to microbial count data, several additional challenges arise. The most fundamental is the presence of zeros or near-zero values resulting from either limited sequencing depth or the true absence of taxa. Extreme sparsity is common in microbial sequencing data and may approach 90% [4] in some datasets. As a result, log-ratio transformations (see Fig. 1) cannot be applied directly, even though they are fundamental for many statistical analyses. Beyond zero inflation, the discrete and skewed nature of count data also poses challenges to the continuity assumption underlying the log-ratio framework, introducing quantization effects that can strongly affect small values and distort analytical results [7,8]. These challenges are further compounded by the high dimensionality of microbial datasets, where the number of taxa features greatly exceeds the number of samples. In addition, technical variation

and differences in sequencing depth introduce additional sources of variability and complicate statistical inference.

This review contributes as a synthesis of zero handling strategies with theoretical and empirical examination of challenges specific to compositional count data. Section 2 provides a brief introduction to the foundations of compositional data analysis (CDA) and the role of log-ratio transformations. The main reviews in Section 3. We first review types of zeros, their underlying mechanisms, and the principles guiding appropriate treatment (Section 3.1). We then summarize zero handling approaches for handling ground zeros (Section 3.2), followed by modeling approaches for zero-inflated data as a prescriptive interpretation of true absences (Section 3.3). Section 3.4 reviews zero-inflated data transformation and related data analysis in the presence of zeros. An overview of the review structure is provided in Fig. 1.

Given the characteristics of microbial count data, Section 3.5 is dedicated to the specific challenges posed by discrete count compositional data. We review current approaches for handling discrete count data, including imputation methods and probabilistic models such as Dirichlet-multinomial and zero-inflated Dirichlet-multinomial, and the theoretical implications of applying classical CDA to discrete count data. Section 4 presents a comparative evaluation of existing imputation methods, single high-dimensional count-based datasets [9] and a simulated dataset to assess the performance of various zero handling techniques. The third part focuses on methodological comparison and the practical application of count data, rather than focusing on the statistical relationship between discrete sequencing counts and continuity assumptions. To assess imputation methods under conditions typical of microbial count sequencing data, we use a self-referential discrete count dataset [10]. Additionally, we highlight the importance of controlling conditions, enabling direct comparison. In addition to evaluating reconstruction accuracy and count preservation, we further compare the complexity of downstream analyses, including PCA and clustering, to examine how reconstruction quality relates to practical analytical performance. We also investigate the temporal behavior of standardized log-ratios on discrete count data and evaluate whether simple integer normalization procedures can improve temporal stability. This critical review offers three primary contributions: first, we establish a structured reference framework that categorizes various zero handling approaches and clarifies their appropriate application based on specific data types. Second, we detail the statistical pitfalls inherent in continuous CDA when applied to discrete counts, outlining why lattice value sequencing data requires distinct statistical considerations. Third, through comparative evaluations on high-dimensional count datasets,

File1

Transformation	Mathematical formula	Analytical properties & limitations
Additive Log Ratio (AR)	$AR_{(x)} = \left(\ln \frac{x_1}{x_j}, \dots, \ln \frac{x_D}{x_j} \right)'$ for $i \neq j$	Simple and interpretable as log fold changes. Not invariant to permutations.
Geometric Log Ratio (GR)	$GR_{(x)} = \left(\ln \frac{x_1}{g(x)}, \dots, \ln \frac{x_D}{g(x)} \right)'$ where $g(x) = \sqrt[p]{\prod_{i=1}^D x_i}$	Symmetric, unique, and strictly isotonic. Variables sum to zero, leading to singular covariance matrix.
Isometric Log Ratio (ILR)	$ILR_{(x)} = H' \cdot GR_{(x)}$ where H is an orthonormal matrix	Strictly preserves distances and preserves linearity with $p-1$ independent variables, coordinates axes stay fixed and one rotates.

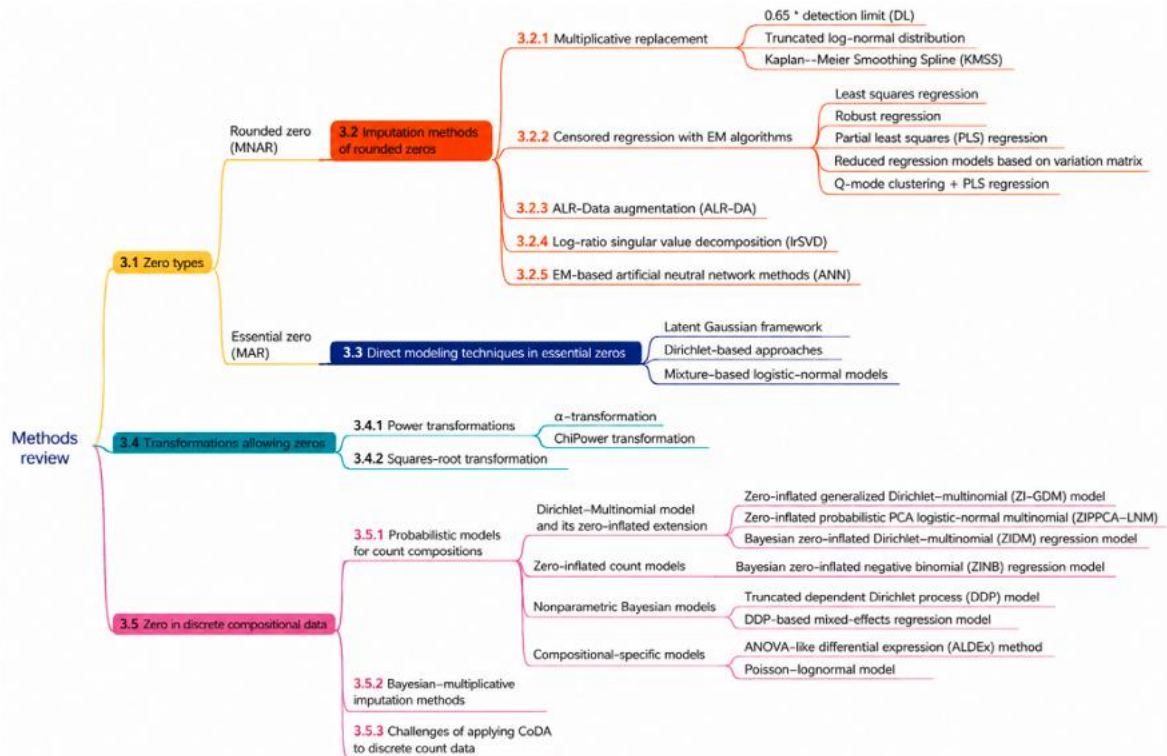


Fig.1 Diagram illustrating the structure of Section 3. This review covers the classification of factors leading to hearing loss arising from different underlying mechanisms and deafness outcomes, and adopted a keyword search to detect empirical data. The diagram illustrates the correspondence between the structure of the review.

[illegible]

2 Compositional Data Analysis (CDA)

Compositional data or *compositions* are vectors of nonnegative values that describe the contribution of D parts contained by a constant sum typically 1 (for proportions) or 100 (for percentages). For a D -part composition $\mathbf{x} = (x_1, \dots, x_D)'$ lies in a $(D - 1)$ -dimensional simplex space S_D^{D-1} :

$$S^D = \left\{ \mathbf{x} = (x_1, \dots, x_D)' \mid x_i > 0, \sum_{i=1}^D x_i = C \right\}.$$

When $n = 3$, a composition is in the orbit of a σ and is finally a simplex s^3 , which is geometrically represented as a tetrahedron containing arbitrarily many interior points. Due to the topology of a 3-sphere, classical statistical methods as, e.g., Haldane's method are inappropriate because they do not take the periodicity of the data into account. In the case of $n = 3$, the simplex s^3 is a tetrahedron. The simplex s^4 is a 4-simplex, i.e., a 4D tetrahedron. The simplex s^5 is a 5-simplex, i.e., a 5D tetrahedron. The simplex s^6 is a 6-simplex, i.e., a 6D tetrahedron. The simplex s^7 is a 7-simplex, i.e., a 7D tetrahedron. The simplex s^8 is a 8-simplex, i.e., a 8D tetrahedron. The simplex s^9 is a 9-simplex, i.e., a 9D tetrahedron. The simplex s^{10} is a 10-simplex, i.e., a 10D tetrahedron. The simplex s^{11} is a 11-simplex, i.e., a 11D tetrahedron. The simplex s^{12} is a 12-simplex, i.e., a 12D tetrahedron. The simplex s^{13} is a 13-simplex, i.e., a 13D tetrahedron. The simplex s^{14} is a 14-simplex, i.e., a 14D tetrahedron. The simplex s^{15} is a 15-simplex, i.e., a 15D tetrahedron. The simplex s^{16} is a 16-simplex, i.e., a 16D tetrahedron. The simplex s^{17} is a 17-simplex, i.e., a 17D tetrahedron. The simplex s^{18} is a 18-simplex, i.e., a 18D tetrahedron. The simplex s^{19} is a 19-simplex, i.e., a 19D tetrahedron. The simplex s^{20} is a 20-simplex, i.e., a 20D tetrahedron. The simplex s^{21} is a 21-simplex, i.e., a 21D tetrahedron. The simplex s^{22} is a 22-simplex, i.e., a 22D tetrahedron. The simplex s^{23} is a 23-simplex, i.e., a 23D tetrahedron. The simplex s^{24} is a 24-simplex, i.e., a 24D tetrahedron. The simplex s^{25} is a 25-simplex, i.e., a 25D tetrahedron. The simplex s^{26} is a 26-simplex, i.e., a 26D tetrahedron. The simplex s^{27} is a 27-simplex, i.e., a 27D tetrahedron. The simplex s^{28} is a 28-simplex, i.e., a 28D tetrahedron. The simplex s^{29} is a 29-simplex, i.e., a 29D tetrahedron. The simplex s^{30} is a 30-simplex, i.e., a 30D tetrahedron. The simplex s^{31} is a 31-simplex, i.e., a 31D tetrahedron. The simplex s^{32} is a 32-simplex, i.e., a 32D tetrahedron. The simplex s^{33} is a 33-simplex, i.e., a 33D tetrahedron. The simplex s^{34} is a 34-simplex, i.e., a 34D tetrahedron. The simplex s^{35} is a 35-simplex, i.e., a 35D tetrahedron. The simplex s^{36} is a 36-simplex, i.e., a 36D tetrahedron. The simplex s^{37} is a 37-simplex, i.e., a 37D tetrahedron. The simplex s^{38} is a 38-simplex, i.e., a 38D tetrahedron. The simplex s^{39} is a 39-simplex, i.e., a 39D tetrahedron. The simplex s^{40} is a 40-simplex, i.e., a 40D tetrahedron. The simplex s^{41} is a 41-simplex, i.e., a 41D tetrahedron. The simplex s^{42} is a 42-simplex, i.e., a 42D tetrahedron. The simplex s^{43} is a 43-simplex, i.e., a 43D tetrahedron. The simplex s^{44} is a 44-simplex, i.e., a 44D tetrahedron. The simplex s^{45} is a 45-simplex, i.e., a 45D tetrahedron. The simplex s^{46} is a 46-simplex, i.e., a 46D tetrahedron. The simplex s^{47} is a 47-simplex, i.e., a 47D tetrahedron. The simplex s^{48} is a 48-simplex, i.e., a 48D tetrahedron. The simplex s^{49} is a 49-simplex, i.e., a 49D tetrahedron. The simplex s^{50} is a 50-simplex, i.e., a 50D tetrahedron. The simplex s^{51} is a 51-simplex, i.e., a 51D tetrahedron. The simplex s^{52} is a 52-simplex, i.e., a 52D tetrahedron. The simplex s^{53} is a 53-simplex, i.e., a 53D tetrahedron. The simplex s^{54} is a 54-simplex, i.e., a 54D tetrahedron. The simplex s^{55} is a 55-simplex, i.e., a 55D tetrahedron. The simplex s^{56} is a 56-simplex, i.e., a 56D tetrahedron. The simplex s^{57} is a 57-simplex, i.e., a 57D tetrahedron. The simplex s^{58} is a 58-simplex, i.e., a 58D tetrahedron. The simplex s^{59} is a 59-simplex, i.e., a 59D tetrahedron. The simplex s^{60} is a 60-simplex, i.e., a 60D tetrahedron. The simplex s^{61} is a 61-simplex, i.e., a 61D tetrahedron. The simplex s^{62} is a 62-simplex, i.e., a 62D tetrahedron. The simplex s^{63} is a 63-simplex, i.e., a 63D tetrahedron. The simplex s^{64} is a 64-simplex, i.e., a 64D tetrahedron. The simplex s^{65} is a 65-simplex, i.e., a 65D tetrahedron. The simplex s^{66} is a 66-simplex, i.e., a 66D tetrahedron. The simplex s^{67} is a 67-simplex, i.e., a 67D tetrahedron. The simplex s^{68} is a 68-simplex, i.e., a 68D tetrahedron. The simplex s^{69} is a 69-simplex, i.e., a 69D tetrahedron. The simplex s^{70} is a 70-simplex, i.e., a 70D tetrahedron. The simplex s^{71} is a 71-simplex, i.e., a 71D tetrahedron. The simplex s^{72} is a 72-simplex, i.e., a 72D tetrahedron. The simplex s^{73} is a 73-simplex, i.e., a 73D tetrahedron. The simplex s^{74} is a 74-simplex, i.e., a 74D tetrahedron. The simplex s^{75} is a 75-simplex, i.e., a 75D tetrahedron. The simplex s^{76} is a 76-simplex, i.e., a 76D tetrahedron. The simplex s^{77} is a 77-simplex, i.e., a 77D tetrahedron. The simplex s^{78} is a 78-simplex, i.e., a 78D tetrahedron. The simplex s^{79} is a 79-simplex, i.e., a 79D tetrahedron. The simplex s^{80} is a 80-simplex, i.e., a 80D tetrahedron. The simplex s^{81} is a 81-simplex, i.e., a 81D tetrahedron. The simplex s^{82} is a 82-simplex, i.e., a 82D tetrahedron. The simplex s^{83} is a 83-simplex, i.e., a 83D tetrahedron. The simplex s^{84} is a 84-simplex, i.e., a 84D tetrahedron. The simplex s^{85} is a 85-simplex, i.e., a 85D tetrahedron. The simplex s^{86} is a 86-simplex, i.e., a 86D tetrahedron. The simplex s^{87} is a 87-simplex, i.e., a 87D tetrahedron. The simplex s^{88} is a 88-simplex, i.e., a 88D tetrahedron. The simplex s^{89} is a 89-simplex, i.e., a 89D tetrahedron. The simplex s^{90} is a 90-simplex, i.e., a 90D tetrahedron. The simplex s^{91} is a 91-simplex, i.e., a 91D tetrahedron. The simplex s^{92} is a 92-simplex, i.e., a 92D tetrahedron. The simplex s^{93} is a 93-simplex, i.e., a 93D tetrahedron. The simplex s^{94} is a 94-simplex, i.e., a 94D tetrahedron. The simplex s^{95} is a 95-simplex, i.e., a 95D tetrahedron. The simplex s^{96} is a 96-simplex, i.e., a 96D tetrahedron. The simplex s^{97} is a 97-simplex, i.e., a 97D tetrahedron. The simplex s^{98} is a 98-simplex, i.e., a 98D tetrahedron. The simplex s^{99} is a 99-simplex, i.e., a 99D tetrahedron. The simplex s^{100} is a 100-simplex, i.e., a 100D tetrahedron. The simplex s^{101} is a 101-simplex, i.e., a 101D tetrahedron. The simplex s^{102} is a 102-simplex, i.e., a 102D tetrahedron. The simplex s^{103} is a 103-simplex, i.e., a 103D tetrahedron. The simplex s^{104} is a 104-simplex, i.e., a 104D tetrahedron. The simplex s^{105} is a 105-simplex, i.e., a 105D tetrahedron. The simplex s^{106} is a 106-simplex, i.e., a 106D tetrahedron. The simplex s^{107} is a 107-simplex, i.e., a 107D tetrahedron. The simplex s^{108} is a 108-simplex, i.e., a 108D tetrahedron. The simplex s^{109} is a 109-simplex, i.e., a 109D tetrahedron. The simplex s^{110} is a 110-simplex, i.e., a 110D tetrahedron. The simplex s^{111} is a 111-simplex, i.e., a 111D tetrahedron. The simplex s^{112} is a 112-simplex, i.e., a 112D tetrahedron. The simplex s^{113} is a 113-simplex, i.e., a 113D tetrahedron. The simplex s^{114} is a 114-simplex, i.e., a 114D tetrahedron. The simplex s^{115} is a 115-simplex, i.e., a 115D tetrahedron. The simplex s^{116} is a 116-simplex, i.e., a 116D tetrahedron. The simplex s^{117} is a 117-simplex, i.e., a 117D tetrahedron. The simplex s^{118} is a 118-simplex, i.e., a 118D tetrahedron. The simplex s^{119} is a 119-simplex, i.e., a 119D tetrahedron. The simplex s^{120} is a 120-simplex, i.e., a 120D tetrahedron. The simplex s^{121} is a 121-simplex, i.e., a 121D tetrahedron. The simplex s^{122} is a 122-simplex, i.e., a 122D tetrahedron. The simplex s^{123} is a 123-simplex, i.e., a 123D tetrahedron. The simplex s^{124} is a 124-simplex, i.e., a 124D tetrahedron. The simplex s^{125} is a 125-simplex, i.e., a 125D tetrahedron. The simplex s

21 Logradiofrankarditslimitations

In practice, the requirement in the traditional definition that compensation be paid to the defendant is not a problem. In the absence of a defendant's list, the presiding judge is not able to make a ruling on a request for compensation. In any case, the defendant's list is not a condition for compensation. In fact, the defendant's list is not a condition for compensation. In fact, the defendant's list is not a condition for compensation.

eros caused deviations from the prescribed total. Second, although many real-world datasets exhibit a compositional character, their total sum often depends on sample-specific technical factors rather than meaningful biological quantities. For example, in microarray data, the total read count of each sample is determined largely by sequencing depth, which reflects technical variation. As emphasized by Gorm et al. [4], sequencing count takes are compositional, but their total read counts have biological interpretation and therefore can be used to compare samples on an absolute scale. Meanwhile, GDA requires scale invariance. Under this principle, multiplying a composition by any positive constant, i.e., x_i/x_j , unchanged the x_i and x_j correlation in the data. However, this property implicitly relies on the continuous nature of the simplex. With the discovery of discrete value counts, scaling generally produces non-integer values that must be rounded back onto the lattice. This process introduces quantization effects and may distort the original data. Verduyn et al. discuss this in detail in Section 3.5.3.

Because individual components are interdependent under the constraint of a total, GDA-based analysis is more challenging than absolute quantities. Huisman et al. [11] emphasize that compositional data should be treated as a network of components and that proportions are not only dependent but also dependent on the information rather than its during characteristics. As summarized in Fig. 1, all standard GDA transformations require strictly positive components. Consequently, classical GDA faces two major challenges in sequencing count data: the presence of zeros and the discrete nature of lattice-valued observations, which are discussed in the following sections.

3 Review of methods dealing with zeros in GDA

Before reviewing methods for handling zeros in compositional data, we list categories of zeros in two parts: with the ideal studied variable and without. Fig. 1 illustrates the categories and groups them into three major lines: (i) ignoring the missing parts to obtain a full data matrix (Section 3.2); (ii) considering statistical models that explicitly account for the missing parts and the zeros (Section 3.3); and (iii) developing zero-deficient transformations (Section 3.4).

3.1 Classification and interpretation of zeros

Distinguishing the nature of zeros is a prerequisite for selecting appropriate handling strategies in compositional data analysis. The first major category is *rounding zeros*, which arise when true positive values fall below the detection limit of an assay and are rounded to zero (e.g., low-level gene expression or sensitivity) or from limited observation windows, during which a part of the count is not detected [13]. For example, in food and budget surveys, some families may report zero spending on alcohol because they do not consume it during the reporting period, even though they do consume it occasionally. These zeros do not reflect the absence of the underlying detection, sampling limitation, or rounding effects [14]. From a GDA perspective, rounding zeros are independent of small positive values and are not typically handled by zero-deficient (imputation) patterns, which estimate small positive values before log-ratio transformations can be applied (Section 3.2).

Although not strictly a different category, *essential zeros* (so-called structural zeros) [6] are essential to the treatment of a component in the structural, biological, or chemical constraints. Unlike rounding zeros, they reflect an irreducible compositional information rather than measurement limitations. In the context of GDA, the presence of essential zeros is a poor indicator of missing information in compositions. Instead, it indicates a component with an essential zero does not contribute to the variation in the composition. And variation containing essential zeros lies on the boundary of the simplex, reducing

its effective dimensionality [16]. Consequently, replacing essential zeros with small positive values is not a rational choice for dimension reduction, which may bias the results. Ignoring missing parts is more appropriate. They therefore require explicit statistical handling, such as missingness-aware methods, or zero-adjusted distributions (Section 3.3), rather than zero replacement.

Next, we meet the classification of zeros in GDA with missing data. In this context, a *homoskedastic* positive missingness is typically described by how it depends on the observed and unobserved parts of the data. Consider a completed data $\mathbf{x} = (x_1, \dots, x_n)$, where x_i is the observed value and x_j is the missing value. Missingness is said to be *Missing Completely At Random* (MCAR) if we think that an entry is missing does not depend on either x_i or x_j . A *value-at-risk* or *informative* missingness is *Missing At Random* (MAR) if we think missingness may depend on the observed values, but not on the unobserved part x_j . *Homoskedastic* missingness, structural zeros, and essential zeros are all *MCAR* in the context of missingness, although their structural interpretation differs substantially [17].

When the possibility that an entry is missing depends on the unobserved values x_j , the data is said to be *Missing Not At Random* (MNAR). MNAR missingness is a challenging problem to address because the required knowledge of why the values are missing is not available. Ignoring missingness in this case is a common but incorrect practice. MAR because the occurrence depends on underlying unobserved values below the detection limit (DL). They can be viewed as left-censored observations with a known threshold and are typically handled using zero-deficient patterns designed for censored compositional data [17].

3.2 Imputation methods for round zeros

Global-purpose imputation methods, such as random forest (missForest) [18], deep generative modeling (mice) [19], and nearest neighbor imputation (VIM) [20], are generally not suitable for round zeros in compositional data. Unlike censored missing values, round zeros represent left-censored observations and must be handled while preserving the relative structure of the composition. Comparative studies have shown that conventional imputation methods often perform poorly under these constraints [21].

The following sections review specialized zero-deficient methods developed for round zeros, which explicitly account for DLs while preserving the compositional structure of the data.

3.2.1 Multiplicative replacement

The multiplicative replacement approach handles round zeros by replacing them with small positive values δ_{ij} and proportionally rescaling the remaining non-zero components to preserve the total. Formally, the replacement is defined by

$$r_{ij} = \begin{cases} \delta_{ij}, & \text{if } x_{ij} = 0, \\ x_{ij} \left(1 - \sum_{k: x_{ik}=0} \frac{\delta_{ik}}{C_i} \right), & \text{if } x_{ij} > 0, \end{cases} \quad (1)$$

where x_{ij} denotes the original composition and r_{ij} the corresponding value after replacement in sample i and component j . The term $\sum_{k: x_{ik}=0} \frac{\delta_{ik}}{C_i}$ is the small positive value imputed in the zero and $C_i = \sum_{j=1}^D x_{ij}$ denotes the original sample total, ensuring that $r_{i \cdot} = C_i$.

Several multiplicative replacement strategies differ in the choice of the imputed value δ_{ij} , as implemented in the R package `zCompositions` [23], summarized in Table 2.

Although multiplicative replacement guarantees positive values below the corresponding DL, the subsequent data transformation may be non-monotonic, as the values may be non-linearly transformed. For example, a point on the boundary of the simplex may be transformed to a point in the interior, which is not desirable. This limitation is noted by Paterlini and Gao et al. [24].

Table 2
Summary of multiplicative replacement strategies in the R package *zCompositions*.

Function	Imputation strategy (δ_{ij})	Advantages	Limitations
multRepl()	Constant fraction of DLs $\delta_{ij} = 0.65 DL_{ij}$	Simple and computationally efficient	May distort covariance when many zeros are present
multLN()	Truncated lognormal distribution $\delta_{ij} \sim \text{LogNormal}$	Accounts for skewed and covariate	Relies on lognormal assumptions
multKM()	Nonparametric kernel-based estimator with smoothing splines (KMS)	No distributional assumptions	Computationally more demanding

Table 3
Regression-based EM methods for imputation of zero composition.

Method	Regression strategy	Software implementation
Modified EMR	AR+Least squares regression	lrEM()
Robust EM	ILR+robust (MM) regression	imputeBDLs(,method="lmrob")
PLSEM	ILR+PLS regression	imputeBDLs(,method="pls")

322 Generalized regression with Expectation-Maximization (EM) algorithms

The second class of robust imputation methods is based on the Expectation-Maximization (EM) algorithm. The modified EM-AR algorithm is proposed by Paerz et al. [25] and Paerz et al. [26] and further extended by Hron et al. [25], as presented in AR-based composition data via multivariate normal distribution and imputation of zero values through generalized regression. Let $\mathbf{x}_{\text{imp}}$ denote the compositional data matrix after replacing zeros with initial or previously imputed values, and let $\mathbf{z}_{\text{ALR}} = \mathbf{Z}_{\text{ALR}}(\mathbf{x}_{\text{imp}})$ denote its AR-based counterpart. For each zero component, a regression model is constructed using the remaining log-ratio coordinates as predictors:

$$\mathbf{z}_{\text{ALR},j} = \mathbf{Z}'_{\text{ALR},-j}\beta + \varepsilon.$$

During each EM iteration, observed values are updated by their conditional expectations under a truncated normal distribution while explicitly incorporating the corresponding DL, after which the imputation of zero values is achieved by taking the softmax. Compared with multiplicative replacement, the EM framework better preserves covariance structure while explicitly accounting for censoring. Building on this framework, Martín-Lamote et al. [27] integrated least squares regression with an alternative M-estimator and applied the RAS algorithm to impute outliers against outliers and extend to null values. Subsequently, Kemp et al. [28] integrated the regression step with partial least squares (PLS) regression to account for high-dimensional settings ($n \gg p$). The same is further proposed using the variation matrix to exploit strong potential relationships before fitting the PLS model. The partial regression coefficients are evaluated in this review as summarized in Table 3. The original EM-AR algorithm is implemented in the *zCompositions* package through the `lrEM()` function, whereas the robust and PLS variants are available in the *robustCompositions* package through the `imputeBDLs()` function.

Subsequent developments further extended this framework. Chen et al. [29] incorporated a more diverse prior to PLS regression to partition components into subpopulations, improving computational efficiency while preserving global covariance structure. Independently, Hron [30] proposed an alternative least squares regression framework as an alternative imputation approach for multivariate imputation. Overall, EM-based methods provide a flexible framework for censored regression while requiring no parametric assumptions.

323 ARDA quantization (ARDA)

The ARDA algorithm is a Bayesian data augmentation (DA) method [31] based on the truncated AN model. Similar to the AR-EM algorithm (Section 322), models are imputed in the AR space under conditional joint constraints. However, instead of deterministic random initialization, ARDA employs Markov Chain Monte Carlo (MCMC) sampling to jointly update the model parameters and model parameters. At the initialization, the missing values are sampled as

from a truncated normal distribution, and the model parameters are sampled as

$$\mathbf{z}_{\text{ALR},ij}^{(l)} \sim \text{TruncNormal}(\mathbf{z}_{\text{ALR},ij}^{(l)} | \beta_j^{(l)}, \sigma_j^{2(l)}, \mathbf{z}_{\text{ALR},ij} < DL_{ij}). \quad (2)$$

While AREM-ARDA provides imputation uncertainty through posterior sampling, making it particularly suitable for small datasets and allowing prior information to be incorporated naturally within the Bayesian framework [23]. The method is implemented in the `lrda()` function of the *zCompositions* package [23]. The same implementation also supports multiple imputation (MI) based on MCMC sampling, while bootstrap sampling has been proposed as a computationally efficient alternative for variance estimation when MCMC is infeasible [24].

324 Low-rank singular value decomposition (LSVD)

Paerz et al. [32] proposed the LSVD algorithm as a method designed for imputing missing values in compositional data. It is considered to be the dominant underlying structure through low-rank matrix approximation in log-ratio space. Specifically, the algorithm minimizes the value of the singular value decomposition (SVD) of the log-ratio matrix $\mathbf{X}$ that best approximates the original data matrix $\mathbf{X}$ (which contains zeros or missing values) in the least squares sense by minimizing the Frobenius norm:

$$\min \|\mathbf{X} - \mathbf{R}\|_F^2 \quad \text{subject to} \quad \text{rank}(\mathbf{R}) \leq r, \quad (3)$$

where r denotes the target rank, i.e., the number of dominant latent components retained in the approximation.

For imputation of zero or great missing values, LSVD first applies multiplicative replacement to obtain a initial complete matrix and then uses the truncated RAS algorithm. If the iterative estimates converge to a low-rank approximation while preserving the observed values. By scaling the singular values over a range, the algorithm can simultaneously accommodate zero values, great missing values, and mixed missingness scenarios. Creating null RAS coordinates preserves imputed compositional properties, including scale invariance and homogeneity constraints.

The LSVD algorithm is implemented in the *zCompositions* package through the `lrSVD()` function. However, as an SVD-based approach, LSVD is sensitive to outliers and reduced scaling values.

325 EM-based quantified random network method (AN)

The method proposed by Kemp [21] integrates a statistical neural network (AN) into an iterative EM framework for imputation. The main idea is to exploit the network's ability to learn complex nonlinear dependencies among compositional parts, while ensuring that the imputed values remain strictly below the DL. The algorithm is initialized using a least squares regression approach based on

Aitchison states and discusses [26]. The shortest EMDs are then carried out in the 1D coordinate system preserving the compositional structure of the data. Unlike principal regression, the ANCOVA model does not require normal distributional assumptions. Instead, it automatically learns complex dependence structures among variables through nonlinear hidden layers. As a drawback, this approach is a substantial computational cost and its sensitivity to outliers, which can limit its robustness in practice.

3.3. Direct modelling techniques in serial zeros

A simple approach to handle serial zeros is amalgamation [6], which merges components containing serial zeros with related non-zero parts. For example, in the soil pH data, “acid” and “toxic” may be combined into a single category. However, this comes at a cost: amalgamation smoothes information about the original parts and is nonlinear in the simplex, limiting interpretability [11]. Since serial zeros carry potentially structural information, they are typically modelled directly, rather than merged. Accordingly, several statistical approaches have been proposed, including mixture-based logit-normal models, latent Gaussian formulations, and Dirichlet-based approaches.

Aitchison and Lee [15] proposed an intuitive extension of the logit-normal distribution. Compositional data are partitioned according to their zero patterns, and each group is modelled separately. This approach is particularly suited to continuous or partially categorical data in which zeros represent true absence. Building on this idea, Sevast'yanov and Lichstein [3] introduced a multiplicative logit-normal mixture model that separately models zero components and their conditional distributions for the non-zero parts. When a component is zero, its location is assumed to follow a (potentially skewed) logit-normal distribution, imposing flexibility on precise mean locations. This framework is often well suited to regression settings and has been directly used by zeros; however, it cannot fully capture covariate effects on compositional components, leaving some dependence structures only partially modelled. Her and Billauer [16] further extended this class of models by pooling compositions into low-dimensional subspaces according to zero patterns and fitting a global logit-normal mixture with shared location and dispersion parameters across groups. While computationally convenient, this approach disregards heterogeneity in where variables are present and assumes identical heteroscedastic structures across patterns, which may not hold in heterogeneous datasets. Overall, mixture-based models for serial zeros follow a common principle: separating “zero” and “non-zero” cases, then modelling the positive components via (extensions of) the logit-normal distribution. While these approaches preserve the structural meaning

of the second class of modelling strategies treats zeros as latent variables within a latent Gaussian framework. These methods assume a Gaussian distribution in the real space and then map it onto the simplex through an appropriate transformation. Biller and Gebrey [34] introduced hidden portions onto the simplex, modelling joint probabilities on the boundaries through a latent multivariate normal distribution to accommodate zeros. However, these methods are limited to low-dimensional cases and do not preserve homogeneity in ratios. Building on this idea, Torges [35] proposed an alternative pooling approach to modelling structural zeros based on the multivariate normal distribution, but the zero components are only considered only as zero or non-zero in the compositional vector, making it difficult to vary the zero status of individual variables. In the context of climate data, Langer et al. [36] introduced a novel scaling transformation that maps compositions into a latent Gaussian space while allowing a local part to have a zero. Spatial dependence is then modelled through a hierarchical power model, making the approach suitable for high-dimensional spatial compositional data. A key limitation of

this approach, however, is that it requires at least one component to remain strictly positive across the spatial domain; otherwise, the transformation is not well defined.

The third class is based on the Dirichlet distribution. Torges and Sevast'yanov [37] proposed a novel model called Zero-Augmented Dirichlet Regression (ZADR). In this approach, observations are grouped according to their zero patterns, and the non-zero components are modelled using truncated Dirichlet distributions. Dirichlet regression, especially ZADR, provides an intuitive framework to model high serial zeros but assumes that all zero patterns share the same parameter set. Its flexibility is further limited by negative correlations and convergence issues. Compared to other methods, it is more practical, although theoretically more flexible. Langer et al. [38] proposed a novel stochastic representation and developed the Dirichlet Composition Distribution (DCD) as a generalization of the Dirichlet distribution. The DCD captures the modelling the probability of zero components from the unobserved non-zero components, effectively addressing serial zeros in compositional data. Although the model is a regression model with high-dimensional covariates remains challenging and needs further study.

All of the above modelling approaches assume that compositional data are continuous and measured on a real scale. However, several practical situations where data are fundamentally discrete; they consist of discrete counts that lie on a discrete simplex. These discrete data types require specialized handling beyond continuous (G)ANOVA. Methods designed specifically for such discrete compositional data are reviewed in detail in Section 3.5.

3.4. Transformations of zeros

In addition to the logit-normal transformations that are incompatible with zeros (see Table 1), several alternative transformation approaches have been designed that can directly accommodate compositional data containing zero values.

3.4.1. Power-based transformations for compositional data

In compositional data analysis, traditional approaches are commonly correlated. One is the log-Ratio Analysis (LRA), such as CLR and ILR , rooted in Aitchison's geometry [6,11]. The other is the use of the components as covariates in linear models, namely Partial Least Squares (PLS). While PLS ignores the compositional constraint, it remains simple and widely used. LRA by contrast, aligns with the (G)ANOVA but cannot be applied directly when zeros are present.

To address this issue, Torges et al. [39] proposed the α -transformation based on a Box-Cox type power transformation. It unifies LRA and PLS within a principled framework by pooling a common unobserved hidden parameter and log-ratio geometry. This class of a practical compromise for zero-related and high-dimensional sparse compositions while addressing the structural zeros of Aitchison's geometry. The transformation consists of two steps. The first applies the α -power transformation to the compositional vector $\mathbf{x}$ to develop the

$$u_{\alpha}(\mathbf{x}) = \left(\frac{x_1^{\alpha}}{\sum_{j=1}^D x_j^{\alpha}}, \dots, \frac{x_D^{\alpha}}{\sum_{j=1}^D x_j^{\alpha}} \right), \quad 0 \leq \alpha \leq 1. \quad (4)$$

The second step linearly maps the transformed composition into a $(D-1)$ -dimensional hidden subspace orthogonal to the unit vector. The α -transformation depends on α and α is a free volume parameter. When $\alpha = 1$, $u_{\alpha}(\mathbf{x})$ reduces to linear scaling of the original components corresponding to RAs. As $\alpha \rightarrow 0$, the transformation converges to $\ln(x_i/g(\mathbf{x}))$, where $g(\mathbf{x}) = \left(\prod_{i=1}^D x_i \right)^{1/D}$ is the geometric mean of the composition corresponding to the CLR/ILR transformation and this to LRA intermediate values ($0 < \alpha < 1$). Power-based family of geometries between these two extremes. Torges et al. [40] applied the α -transformation in discriminant analysis and showed that it is more

However, biotransformation is not as flexible by design as the principle of classical CDA, in which the composition of the biotransformation is determined by the biotransformation parameters, which typically select a dedicated enzyme (e.g., the hydroxylase for covalent modification) according to the downstream task.

Selyadi [38] and Sved [44] proposed a generalization of the Apollonian distribution, d_{A-d-1} , that allows for the presence of non-directional clusters such as the t -distribution [43] or the elliptical symmetric Cauchy-Gauss (ESG) distribution [44]. The directional distributions allow both positive and negative clusters, even components becoming a key limitation of the Hough method, which cannot directly detect positive clusters. However, the generalization of the Apollonian distribution to preserve Anderson's generality because it fails to handle non-directional ratios. Moreover, components below zero determine the probability for directional nodes (e.g. the t -distribution), which may assign a negligible probability to a node outside the positive orthant [44].

By proceeding in this manner, we find that the value of χ naturally rises on a natural lattice studied from a lattice [7]. A d -dimensional natural unit lattice $\mathbb{N}^D$ consists of all pairs of the form $(x_1, \dots, x_D)$ with $x_i \in \mathbb{N}$, for i a degree d shown in the left part of Fig. 2. Later, in this manuscript,

In contrast, classical (C) Alloys assume that compositions are defined in a continuous and singly universal space. All possible compositions with D partition for the $(D-1)$ -dimensional simplex S^D are introduced in Section 2, which is a continuous and infinitely dense domain in length and of Fig. 2, where any two compositions can be connected by infinitely many intermediate points. This idealized representation provides the mathematical foundation for logic operations, data analyses, and the analytical tools in CDA. After doing so, real compositions can be viewed as lattice points embedded within S^D , rather than arbitrary points in the simplex. They therefore occupy discrete sites instead of the simplex whose resolution depends on the sequencing depth. This approach alleviates the discrete lattice and the continuous simplex conflict, the quantization effects and numerical instabilities discussed in the following sections. The relationship between the representation and its statistical analytical in Fig. 2 further discussions provided by Lou et al. [7].

The independent LM distribution provides a principled framework for modeling compositional count data to test the ability to reject the simplex constraint. However, this model suffers from two limitations. (1) its inability to handle zero-inflated data and (2) its restriction to capturing only linear correlations between components. Motivated by the importance of capturing nonlinear dependencies in a high-dimensional and potentially sparse data, this paper extends the LM model that explicitly account for excess zeros and nonlinear correlations. A common feature of these extensions is the option to zero-inflate the data to model zero-inflated distributions. The proposed model is considered as a zero-inflated LM distribution for the zero observations. The model can indicate strong structural zero observations as well as, along the two types of zero, to model sparsity.

Lai and Chen [45] introduced the zero-inflated generalized Dirichlet-multinomial model by modifying the Beta-slicewise representation of the generalized Dirichlet distribution. Zero inflation is incorporated through zero-inflated Beta variables, enabling flexible correlation structures than the classical LMP model. Estimation is performed via an EM algorithm. A limitation of the approach, however, is that it does not naturally extend to multivariate regression settings. Lai et al. [46] proposed the zero-inflated hierarchical Poisson-Gamma-Multinomial (ZIP-GAMM) model, which employs a low-rank log-slice multinomial formulation with an explicit zero-inflation component. This approach captures complex correlation patterns in high-dimensional structured data by leveraging the flexibility of the log-slice multinomial distribution. An iterative procedure integrating empirical Bayes estimation and variational approximation is used for estimation in high-dimensional settings. More recently, Kolosy [47] introduced a Bayesian zero-inflated Dirichlet-multinomial (ZIM) regression model that integrates zero-inflation, compositional constraints, and variable selection in high dimensions. Zero inflation is modeled via a gamma regression distribution, while sparsity and group priors are addressed via a mixture of Dirichlet priors, similar to modern sparse stability and computational efficiency. Taken together, these developments reflect the broad directions in zero-inflated LMM modeling: extending distributional flexibility, embracing complex structures, and integrating zero inflation into high-dimensional regression frameworks. Despite this progress, challenges remain in sparse stability, interpretability of latent mechanisms, and fully modeling zero inflation, overdispersion, and complex correlation patterns in large-scale genomics data.

Byond the Mantel test, a ZINB regression model with site area as predictor differential abundance analysis. However, because the model does not incorporate compositional constraints or link variables to the zero-inflated distribution, estimates may not adequately account for compositional data when applied to compositional count data.

In addition to the practical method mentioned above, Pen et al. [49] proposed an efficient empirical approach in which constitutional disorder is required to exist in pairs of planets. This allows for more complexity to achieve a robust fit on the characteristics of the data, enabling it to capture complex patterns and trends in the data. The method of statistical model selection (L1 and L2) proposed (DD) prior, which more naturally reflects the correlations between components and the sparsity in the data. Later, Pen et al.

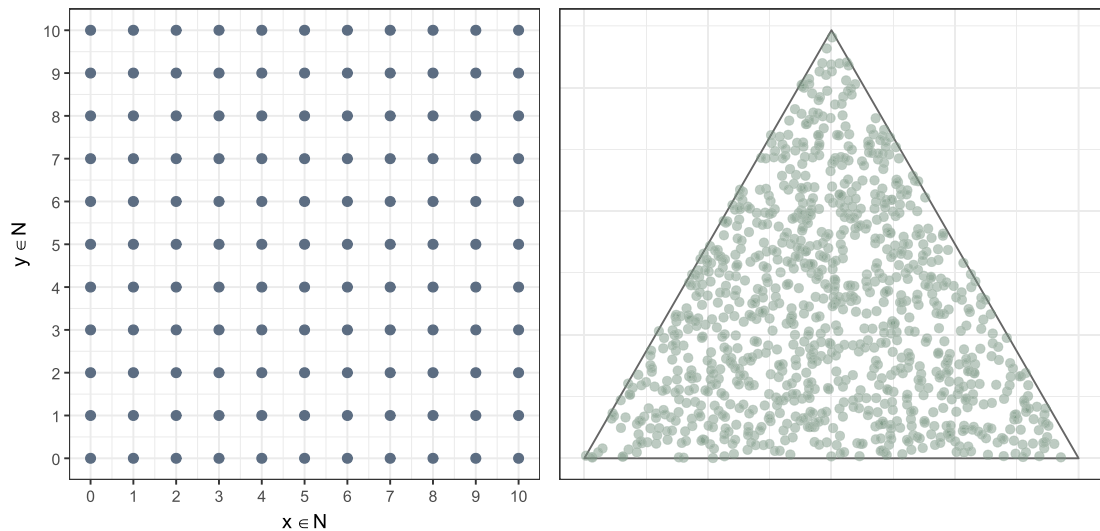


Fig. 2. Interlattice (left) and the continuous 3D simplex (right). Left part: Small simplex $S_N = 1000$ simplex composition generated from asymmetric Dirichlet distribution $D(\theta_1, \theta_2, \theta_3)$ with $\theta_1 = (1, 1, 1)$, plotted on the simplex. Right part: Simplex $S_N = 1000$ simplex composition generated from asymmetric Dirichlet distribution $D(\theta_1, \theta_2, \theta_3)$ with $\theta_1 = (1, 1, 1)$, plotted on the simplex.

extend this approach to mixed-effects regression models to accommodate the need of longitudinal data analysis. However, the method is not explicitly nonlinear and does not consider the correlation between different types of zeros. Additionally, it is designed for small to medium-sized data sets, relying on non-comparative techniques for model selection.

Some additional nonlinear strategies are also worth noting. Ferraz et al. [11] proposed a Bayesian framework for ANOVA-like analysis approach called ANOVA-like Differential Expression (ALDE). It is designed to handle the nonlinear relationship between the response variable and the predictor variables by combining a nonlinear likelihood with a linear prior (with hyperparameter set to 0.5) to obtain the posterior distribution p_i . In this study, the environmental development for a species is zero if its estimated posterior will not be equal to the null hypothesis, which is a binary variable. Subsequently, the posterior p_i is used as a continuous variable in the linear model. This method is designed for statistical analysis and variable selection in high-dimensional data. Finally, Bon-Sure [8] proposed a method based on the Dirichlet distribution applied to the environmental zero as a prior in the composition of the data.

352 Bayesian multiplicative imputation methods

Martínez et al. [5] and Janssens et al. [3] proposed Bayesian multiplicative (BM) imputation methods specifically for compositional data. In the multiplicative replacement approach described in Section 321, BM uses the ratio between the zero component and the non-zero component to impute the zero value. In this study, the ratio is used to impute the zero value. In this study, the ratio is used to impute the zero value. In this study, the ratio is used to impute the zero value.

Let $\mathbf{p} = (p_1, \dots, p_k)$ denote the observed proportions with a cut N . And the prior π is specified as $\pi(\mathbf{p}) = \pi(p_1, \dots, p_k)$ and $\pi(\mathbf{p}) = \pi(p_1, \dots, p_k)$ is the prior density. The Bayesian multiplicative imputation method is defined as follows: $\hat{p}_i = (p_i + s) / (N + s)$, where s is a small positive value. The method is used to impute the zero value. The method is used to impute the zero value. The method is used to impute the zero value.

Table 4
Dirichlet prior and Bayesian posterior estimates.

Prior	S	Posterior mean
Flat	0	$\frac{p_i}{N}$
Rak	1	$\frac{p_i + 1/k}{N + k/2}$
Jeffrey	$\frac{k}{2}$	$\frac{p_i + 1/2}{N + k/2}$
Bayes-Laplace	k	$\frac{p_i + 1}{N + k}$

Accordingly, BM uses the heuristic quantity s_i in Eq. (1) with the corresponding Bayesian posterior estimate, yielding

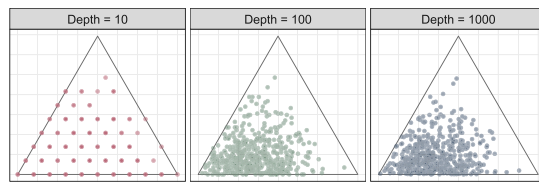
$$p_j^* = \begin{cases} \frac{s_j}{N + s}, & \text{if } p_j = 0, \\ p_j \left(1 - \frac{s}{N + s} \sum_{p_k=0}^k t_k \right), & \text{if } p_j > 0. \end{cases}$$

Because the posterior estimate depends on the sample size N , the method is not strictly scale invariant. Martínez et al. [5] addressed this limitation by introducing a scaled Dirichlet distribution [13].

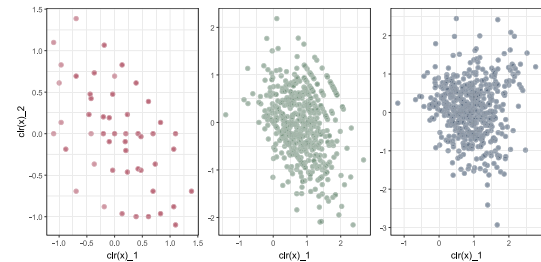
353 Challenges of applying GDA to compositional data

Although many probabilistic and Bayesian methods have been developed specifically for compositional data analysis, system analysis of HSC data—multiplication of ratios and ratios of ratios—and the GDA method is still a challenge. The method is still a challenge. The method is still a challenge. The method is still a challenge.

HSC data are characterized by a combination of biological and technical factors that distinguish them from the continuous composition data used by GDA. First, zero values are common in HSC data, which are not typically present in the continuous composition data. Second, although the data are not strictly scale invariant, the data are not strictly scale invariant. The data are not strictly scale invariant. The data are not strictly scale invariant.



(a) Simplex projection illustrating discretization at different sequencing depths.



(b) CLR coordinates showing sampling noise and geometric distortion under low depth.

Fig. 3 Comparison of compositional geometry across sequencing depths. We simulate a sample size of 500 three-part Dirichlet multinomial compositions with parameters $\alpha = (4, 2, 1)$ under sequencing depths $(10, 100, 1000)$. Left (Fig. 3a) shows the resulting compositional points after applying double centering and distribution to the simplex. Right (Fig. 3b) illustrates the corresponding two components of the CLR standard composition. Overall, increasing sequencing depth systematically reduces the compositional distortion from a highly discrete and sparse lattice approximation to a dense, continuous-like geometric structure, which progressively improves the fidelity and stability of log-ratio analysis.

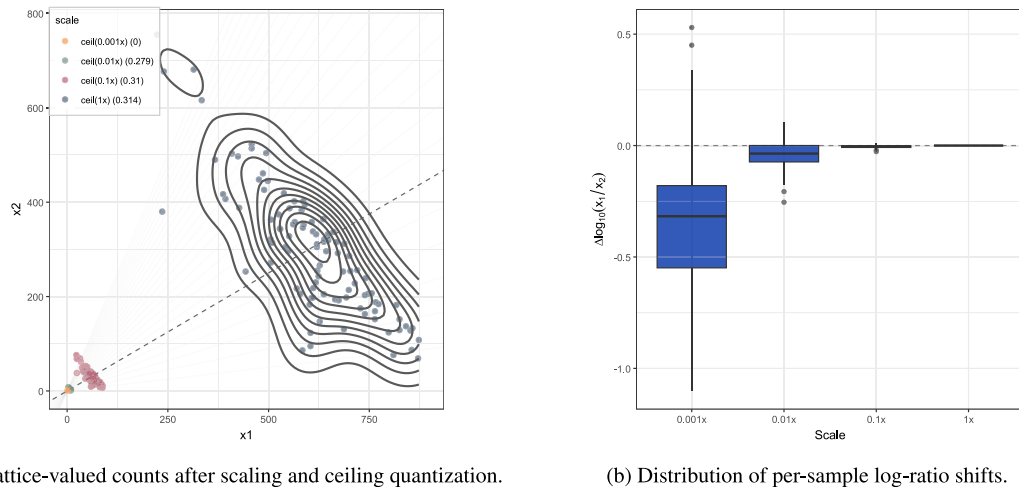
a few outliers or zeros, whereas a small number of highly abundant values may be observed in sparse data sets and latent variable models [5]. Consequently, H-Scout data exhibits a similar scale heterogeneity both within and between samples. Such outliers and imbalances present important challenges for traditional GDA. While GDA emphasizes relative rather than absolute information between components, the considered many-way zero values, the zeros and also very small values cause substantial distortion for log-ratio analysis. The data scaling is typically required for log-ratio analysis for a few samples, which are outliers or the imbalanced data. In sequencing data, the outliers and imbalanced data are common. The scaling procedure is first applied. We now examine the consequences of applying the log-ratio based analytical framework of GDA to compositional data that are inherently continuous for the points.

First, the geometric consequences of applying GDA to lattice-valued data are from the discrete quantization. Low count for different sequencing depths across samples. GDA commonly applies the discrete quantization, which scales each sample to a common total. This does not affect analysis. Although discrete quantization is sample-specific, different sequencing depths and discrete quantization do not alter the relative density of lattice points in the two-part compositions. Consequently, discrete quantization does not affect the relative density of the simplex lattice in the continuous simplex as analyzed by GDA. Moreover, unlike continuous compositions, count values with different numerical scales poses different densities of lattice points. When counts are large, lattice points become much denser, and their positions on the simplex closely approximate the behavior of continuous compositions. In contrast, when counts are small, the corresponding lattice points remain sparse, yielding coarse and unevenly spaced simplex lattice discrete and irregular distributions. The impact of lattice points is illustrated in Fig. 3 (left), where small count compositions form visibly coarse subsets of the simplex whereas large count compositions produce much denser point clouds. The corresponding CLR coordinates in Fig. 3 (right) further illustrate the increased discretization and sampling variability associated with low count compositions. The low density of lattice points can contribute to quantization artifacts and numerical instability in downstream log-ratio analysis [7].

Second, beyond these geometric effects, the discretized count data also has important statistical consequences for log-ratio analysis. The standard log-ratio based GDA procedure relies on the assumption of strictly positive continuous real numbers [6], whereas discrete quantization causes discretization and a lack of homogeneity in

sampling processes such as multinomial or Poisson distributions [7]. Consequently, the uncertainty associated with low count observations is directly propagated into log-ratio calculations. For example, with the same sample, two components with counts of 1 and 2 produce a log-ratio of $\log(2/1) = 0.69$, suggesting a small difference that is largely attributable to sampling variability rather than true biological differences. This illustrates a general point of count data: the influence of quantization or discretization on the quantization of the counts [7]. When counts are small, log-ratio becomes highly sensitive to single unit changes, substantially amplifying sampling noise. As counts increase, this sensitivity rapidly decreases and the log-ratio behavior approaches that of continuous data. Consequently, direct application of GDA to low count compositional data may produce estimates of proportional relationships and associated statistical quantities. The impact of these effects on the evaluation and interpretation of results will be illustrated in the simulation study presented in Section 4.

The third important consequence arises from the discrete assumption of scaling in GDA. A widely used multinomial composition is by a count standard after their relative abundance. Low count data do not naturally satisfy this assumption because their smallest possible increment is 1. When counts are equally scaled down, many distinct compositions collapse onto the same lattice points, creating quantization artifacts and distorting the relative density of the simplex lattice. This discrete quantization leads to quantization artifacts and distorts the relative density of the simplex lattice. The discrete quantization is amplified under low count conditions, making log-ratio analysis increasingly sensitive to sampling variation [7]. To illustrate these effects, we follow an approach analogous to that of Lovell et al. [7]. The part I Dirichlet multinomial composition was generated with a sample size of 100 as a group depth of 100, and $\alpha = (6, 3, 1)$. The simulated counts were then subjected to progressive rounding down by varying the level of quantization, allowing us to visualize the effects of composition reduction and the resulting distortion of log-ratios. The simulation results in Fig. 4a illustrate that the goal of varying the level of quantization causes significant distortion in the Dirichlet multinomial composition to nearest integer levels of lattice points, distorting the standard information loss. The legend reports the non-log-ratio $\log_{10}(\bar{x}_1/\bar{x}_2)$ computed at each scale. As the scaling factor decreases, the non-log-ratios progressively drift away from the true underlying value of $\log_{10}(2) \approx 0.31$, reaching values as low as 0.28, thereby quantifying the increasing distortion induced by aggressive downscaling and discretization. Fig. 4b further shows that quantization induces non-reversible shifts in sample log-ratios. These deviations grow larger



(a) Lattice-valued counts after scaling and ceiling quantization.

(b) Distribution of per-sample log-ratio shifts.

Fig. 4. Effect of global scaling and ceiling quantization on lattice-valued count data. Simulations are conducted using 100 samples from the repeat-lattice multinomial model with sequence depth 100 and $\mathbf{r} = (6, 3, 1)$. (a) Scatter plots show quantized lattice-valued pairs (x_1, x_2) obtained after applying for global scaling factors $s \in \{1, 0.1, 0.01, 0.001\}$. The dashed line indicates the ideal proportionality, $x_2 = 0.5x_1$, which corresponds to the expected ratio $E[x_1/x_2] = 2$ under the given model. Legend lists quantization error $\log_{10}(\bar{x}_1/\bar{x}_2)$ computed at each scale. (b) Box plots show the distribution of per-sample log-ratio shifts $\Delta \log_{10}(x_1/x_2)$ induced by ceiling quantization.

as the scaling factor decreases, confirming that low count regions are especially susceptible to quantization-induced distortion and relative information loss. The impact of these quantization effects on the evaluation of imputation methods will become evident in the simulation study presented in Section 4.

4 Comparison for imputation-based deconvoluting

4.1 Methodological context and data description

The theoretical discussion in Section 3.5.3 suggests that applying the classical log-ratio transformation to skewed count data may introduce non-normal and non-identical means, which are not captured under the continuous assumptions of GCM. These theoretical considerations naturally raise the question of whether such skewness biases are a relevant concern in practical data analysis. The following empirical study is designed to investigate this question through comparative experiments on representative real-world datasets.

In the preceding sections, we introduced several strategies for handling compositional data, with each strategy having its own set of assumptions and potential biases. However, these approaches are typically not designed to handle the complex, high-dimensional data structures encountered in modern sequencing experiments. In this section, we focus on the evaluation of imputation methods, which are designed to handle such data by filling in missing values. The evaluation of imputation methods is a complex task, as it requires a comparison of the methods against a gold standard. In this study, we use a simulated dataset to evaluate the performance of different imputation methods. The simulated dataset is generated using a known true composition, and the imputation methods are evaluated based on their ability to recover the true composition. The results of the evaluation are presented in Section 4.2.

In our comparison study, we designed two sets of experiments to systematically examine how different proportions of zeros and the dimensionality of the data influence the performance of imputation methods. To evaluate the performance of each method, we use two commonly adopted metrics: the compositional error (CE) and the compositional error (CE). The CE is a measure of the difference between the imputed and true compositions, and the CE is a measure of the difference between the imputed and true compositions. The results of the evaluation are presented in Section 4.2.

performance, providing a non-anticipatory comparison of their practical behavior. The simulated datasets were constructed to match the characteristics of the real data, allowing the observed patterns to be independently verified. The corresponding simulation results, presented in Appendix A, corroborate the findings obtained from the real data analysis and strengthen the robustness of our conclusions. All code and analysis used in this work are available at <https://github.com/taoyang-qin/compositional-deconvoluting-framework>.

4.2 Real data context

We evaluate the performance of different imputation strategies using the Rabbit dataset [10]. The dataset consists of $n_s = 3937$ microbial genes measured across $n = 89$ samples collected from two distinct laboratory environments. Crucially, this empirical dataset is fully observed and contains no zeros.

The Rabbit dataset exhibits several characteristics typical of high-throughput sequencing data. It is characterized by high dimensionality, a wide dynamic range, and a non-uniform distribution of values, which challenge traditional compositional imputation methods. As shown in Fig. 5(a), the log-log mean-variance relationship for the genes is non-linear, with the mean increasing sub-linearly as the variance increases. The heatmap in Fig. 5(b), ordered by columns, further illustrates the heterogeneity in the data, with some genes showing high variance and others showing low variance. The results of the evaluation are presented in Section 4.2.

4.3 Data recovery across different frameworks

In our comparison study, we designed two sets of experiments to systematically examine how different proportions of zeros and the dimensionality of the data influence the performance of imputation methods. To evaluate the performance of each method, we use two commonly adopted metrics: the compositional error (CE) and the compositional error (CE). The CE is a measure of the difference between the imputed and true compositions, and the CE is a measure of the difference between the imputed and true compositions. The results of the evaluation are presented in Section 4.2.

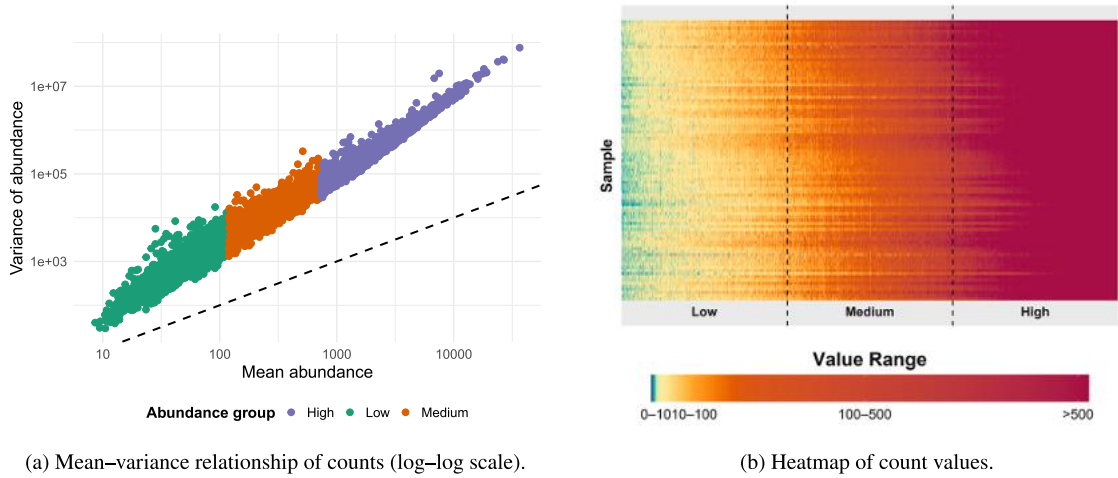


Fig 5 (a) shows the log-log mean-variance relationship across samples illustrating the presence of outliers in compositional data. (b) displays the distribution of count values across samples highlighting the few dynamic range and scale differences between low, medium, and high abundance categories and thereby the effect of the columns.

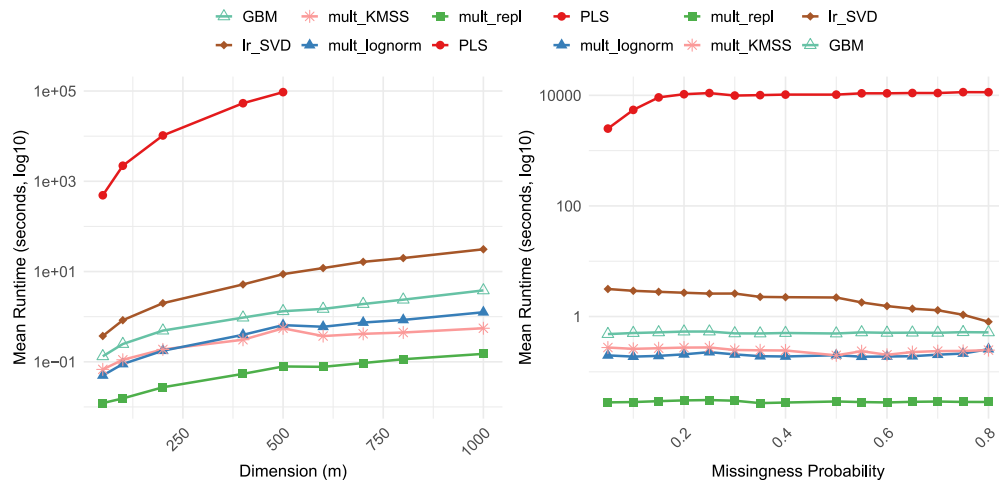


Fig 6 Trends of computational runtime under different settings. The left panel shows the change of mean runtime with increasing dimension (m) at a fixed missingness probability ($p = 0.5$), while the right panel illustrates the runtime variation with different missingness probabilities (p) at a fixed dimension ($m = 200$).

The first experiment examines how imputation accuracy varies with increasing proportions under different dimension settings. For each dimensionality level ($m = 50, 500, 1000$), we randomly sampled columns from the PAB dataset. Random zeros were introduced by replacing all values below the column-specific quantile threshold $Q_p(x_{\cdot j})$ with zero where p denotes the selected quantile level. In our experiments, p varied from 0.05 to 0.8, as we were interested in easy and difficult to predict imputation proportions. Each configuration was replicated 50 times.

The second experiment examines the effect of dimensionality by fixing the zero proportions at $p = 0.2$ and $p = 0.5$ and varying the number of selected columns from $m = 50$ up to $m = 1400$, with 50 replications per configuration. All imputation methods were evaluated using the metrics summarized in Table 5. Because not imputation methods are formulated for continuous data, their outputs are continuous values. Unlike adaptive methods for compositional data, these methods tend to produce values outside the simplex, which is not valid for compositional data. We therefore used the median of the imputed values as the final output.

examined while preserving the structure of sequencing counts influences reconstruction accuracy and numerical stability.

To compute the accuracy of the log-ratio network, let $S = [s_{ij}]$ denote the sample covariance matrix computed from the original data in ILR coordinates, and let $S' = [s'_{ij}]$ denote the corresponding covariance matrix computed after imputation. The AUCS evaluates the preservation of the covariance matrix and is defined as

$$AUCS = \frac{1}{D-1} \|S - S'\|_F, \quad (5)$$

where $D-1$ is the dimensionality of the ILR coordinate space and $\|\cdot\|_F$ denotes the Frobenius norm. A value of $AUCS = 0$ indicates perfect preservation of the covariance structure.

Similarly, to compute the distortion in compositional geometry, let x_M be the M th composition, x_m denote a m th composition, and x_{-m} denote the rest of the fully observed compositions and by normalizing rows in x_M . The GD

Table 5
Overview of imputation methods for left-censored compositional data.

Method (abbr.)	Function	Description
multrep	multRep1()	Multiplicative replacement using fixed fraction (0.05) of the DL, i.e. $Q_p(x_{ij})$.
multlogom	multLN()	Lognormal-based replacement using saturated lognormal model below DL.
multKMSS	multKM()	Multiplicative replacement using a KMSS estimate below DL.
lrda	lrDA()	Log-ratio data augmentation procedure using iterative simulation.
lrsm	lrEM()	EM-based particle replacement using least squares regression.
lrSD	lrSVD()	Low-rank log-ratio imputation using singular value decomposition.
GBM	cmult()	Bayesian multiplicative method based on a Dirichlet multinomial formulation.
lmrob	imputeBDLs(..., method="lmrob")	EM-based particle replacement using robust MM regression.
PLS	imputeBDLs(..., method="PLS")	EM-based particle replacement using PLS regression.
dluf	runif(0.1 * DL, DL)	Random draws from uniform distribution between $0.1 \cdot DL$ and DL .
addl	—	Backvalue replacement where all zeros are set to 1.

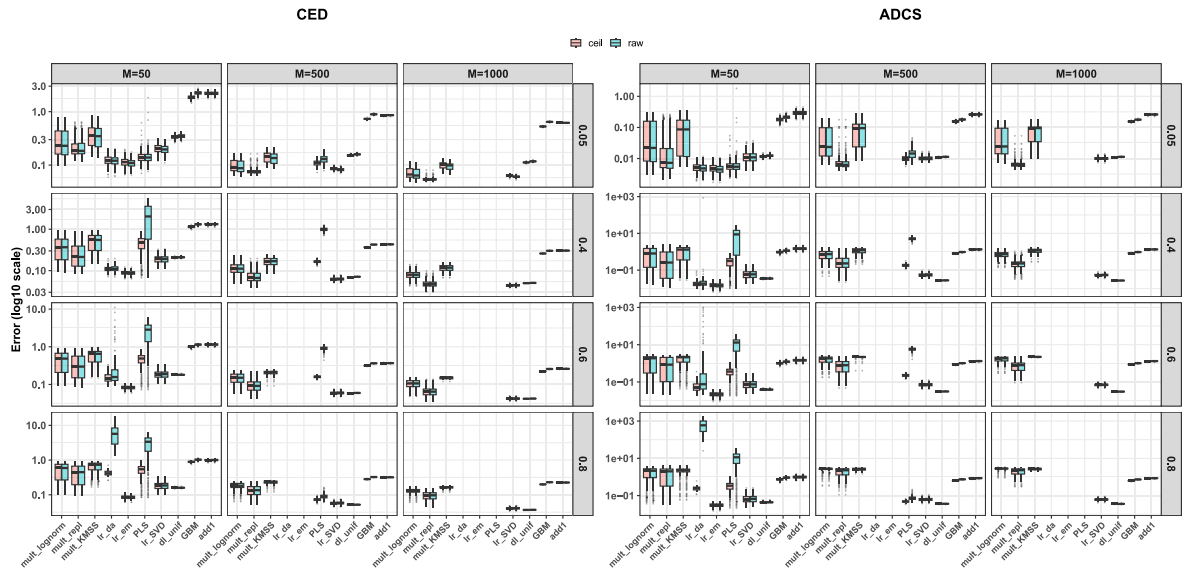


Fig. 7. Comparison of imputation methods across CED (left) and ADCS (right) metrics under varying dimensionality ($m = 50, 500, 1000$) and missingness probabilities.

is defined as

$$CED = \frac{1}{n_M} \sum_{m \in M} \frac{d_A(x_m, \bar{x}_m)}{\max_{x_i, x_j \in X_{-M}} d_A(x_i, x_j)}, \quad (6)$$

where $d_A(\cdot, \cdot)$ denotes the Aitchison distance defined between two compositions x and y as follows

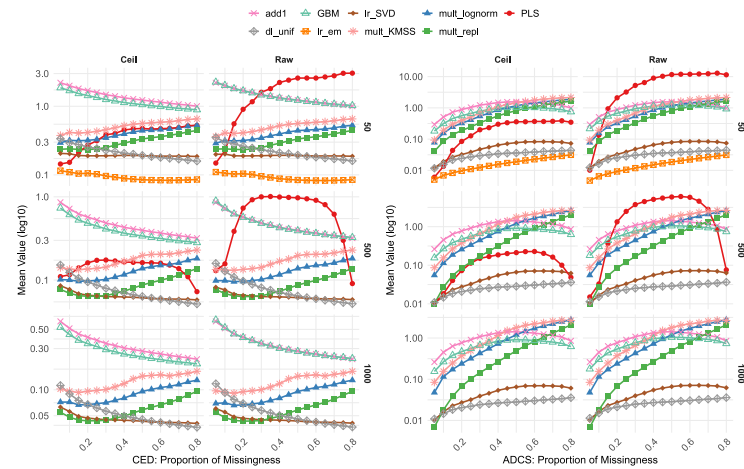
$$d_A(x, y) = \left[\sum_{i=1}^D \left(\ln \frac{x_i}{g(x)} - \ln \frac{y_i}{g(y)} \right)^2 \right]^{1/2}. \quad (7)$$

The denominator in Eq. (6) normalizes the deviation by the maximum observed Aitchison distance between fully observed compositions in X . As a result, CED provides a relative measure of distortion that accounts for the overall compositional variability of the data. The equation generalizes the error proposed by Hironaka et al. [11]. The compositional error proposed in this paper can be computed by CG-11 Gini to Score, llnrad

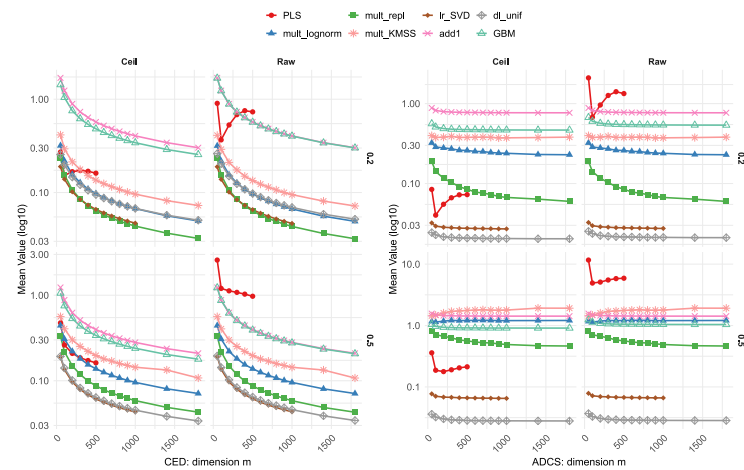
4.4 Results of non-invasive analytical accuracy

From technical perspective, the results revealed considerable limitations in several methods for instance, `lmrob`, `addl`, `plsr` and `dluf` methods were only applicable below dimensional cases (e.g., $m = 50$), and even under this setting, the imputation and simulation methods displayed significant proportion of zeros— from 0.28 to 0.96, $p = 0.05$ up to 0.92, $p = 0.8$ —indicating limited practical applicability. Similarly, the `mult_KMSS` method also showed inconsistent values with high dimensionality while the non-invasive gas analysis at a fixed m was only 0.06 for $m = 50$, if not 0.96 for $m = 1000$. Furthermore, PLS became computationally infeasible at $m = 1000$, raising the cost of time-consuming and another several (Fig. 6).

Figs. 8 and 7 summarize the reconstruction performance of the different imputation methods using the CED and ADCS metrics. Fig. 8 presents the non-metric values across varying dimensions and



(a) Comparison of imputation methods under varying dimensionality and missingness probabilities.



(b) Comparison under fixed missingness probabilities ($p = 0.2$ and $p = 0.5$) across varying dimensions.

Fig 8 Comparison of imputation methods using the CED and ADCS metrics. (a) Mean metric values under varying dimensionality and missingness probabilities. (b) Comparison between Raw and Cell processing under fixed missingness probabilities across different dimensions.

missingness levels, whereas Fig 7 shows the corresponding distributions for simulation metrics, providing a complementary view of the variability and stability of the different methods. In both figures, 'Raw' denotes the continuous imputation metrics, whereas 'Cell' denotes the corresponding results after applying the ceiling operation to the imputed values to enforce focusing on the global performance trends of the continuous (raw) data trends and suppress the structural differences across dimension and sparsity levels. The three multiplicative approaches (mult_repl, mult_lognorm, and mult_KMSS) exhibit moderate and highly similar error growth as the proportion of missingness increases, performing reasonably well in terms of relative correlation. Meanwhile, add1 and GBM consistently maintain higher deviations across missingness levels, though their errors remain relatively stable at high sparsity levels. In contrast, lr_SVD and dl_unif generally maintain the lowest reconstruction errors across different dimensions and missingness levels. As the dimensionality increases and the proportion of missingness increases, the results for some simulation metrics, reflecting the model's computational or numerical limitations under high-dimensional settings.

Importantly, under this continuous raw evaluation framework, PLS performed poorly across all tested dimensions. Specifically, at not even sparsity levels with high dimension, PLS (raw) generated exceptionally high CED and ADCS values.

Evaluating the up-and-down (ceiling) operation levels for different five methods in Fig 8, as illustrated across the panels in Fig 7 and the trend lines in Fig 8, applying the ceiling operation generally provides more stability in model comparison across scenarios for most imputation methods, such as the multiplicative approaches. However, the ceiling operation leads to a substantial performance improvement for the PLS method.

Figs 6 and 7 quantify the instability observed for PLS using two summary metrics: the proportion of replicates with a metric ratio exceeding five ($\text{Metric}_{\text{raw}}/\text{Metric}_{\text{ceiling}} > 5$) and the corresponding median ratio. The median ratio across the first five dimensions of settings was more prone to significant instability. For example, under relatively low sparsity ($p = 0.2$), the proportion of stable ADCS replicates increased from 65% at $M = 50$ to 98% at $M = 500$, while the corresponding median ratio decreased from 1.3 to 1.55.

Table 6
Position of high ratio replicates and median ratios for PLS across varying noise levels (σ).

p	$M = 50$				$M = 500$			
	HD		AOS		HD		AOS	
	Rep	Median	Rep	Median	Rep	Median	Rep	Median
0.05	0000	1007	0000	0998	0000	1161	0000	1337
0.10	0000	1007	0000	1007	0000	1324	0000	1621
0.15	0000	1005	0000	1007	0000	2251	0000	1557
0.20	0000	1005	0000	1007	0000	4668	0000	1858
0.25	0000	1005	0000	1007	0000	5241	0000	1938
0.30	0000	1005	0000	1007	0000	5241	0000	2026
0.35	0000	1005	0000	1007	0000	5241	0000	2026
0.40	0000	1005	0000	1007	0000	5241	0000	2026
0.45	0000	1005	0000	1007	0000	5241	0000	2026
0.50	0000	1005	0000	1007	0000	5241	0000	2026
0.55	0000	1005	0000	1007	0000	5241	0000	2026
0.60	0000	1005	0000	1007	0000	5241	0000	2026
0.65	0000	1005	0000	1007	0000	5241	0000	2026
0.70	0000	1005	0000	1007	0000	5241	0000	2026
0.75	0000	1005	0000	1007	0000	5241	0000	2026
0.80	0000	1005	0000	1007	0000	5241	0000	2026
0.85	0000	1005	0000	1007	0000	5241	0000	2026
0.90	0000	1005	0000	1007	0000	5241	0000	2026
0.95	0000	1005	0000	1007	0000	5241	0000	2026
1.00	0000	1005	0000	1007	0000	5241	0000	2026

Table 7
Position of high ratio replicates and median ratios for PLS across varying dimensions (M).

M	Fixed $p = 0.2$				Fixed $p = 0.5$			
	HD		AOS		HD		AOS	
	Rep	Median	Rep	Median	Rep	Median	Rep	Median
50	0118	100	000	112	000	5407	000	3148
100	0052	100	000	112	000	2336	000	3147
200	0024	200	000	893	000	5380	000	2591
300	0023	365	000	1720	000	5380	100	2591
400	0022	465	000	1957	000	6004	100	2592
500	0041	468	000	1858	000	6027	100	2516

Second, **Table 6** shows that instability was mainly observed under non-Gaussian noise levels roughly between $p = 0.3$ and $p = 0.7$. **Table 7** further demonstrates this pattern by comparing $p = 0.2$ and $p = 0.5$. Under $p = 0.2$, HD remained relatively stable across dimensions while the position of unstable replicates stayed low (43% at $M = 500$). In contrast, under $p = 0.5$, HD instability was already observed in 38% of replicates at $M = 50$ and increased to 98% at $M = 500$. Similar trends were observed for AOS. Initially, AOS was consistently more sensitive than HD, exhibiting higher proportions of unstable replicates and substantially larger median ratios. After applying the collinearity reduction, these inflated ratios were largely diminished, making the performance of PLS comparable with that of most other methods. To examine this specific limitation, we worked with 1000 random outcomes, we created a single representative replicate ($\text{rep} = 1$) under a moderate dimensional setting ($m = 200, p = 0.2$). Without loss of generality, six representative columns were selected from the low, medium, and high burden groups. Rows of all input variables and true values simulated as pos_{PLS} and true_{SVD} are presented in **Fig. 9**. As shown, the scatter plot of PLS imputation lied close to the reference line $y = x$. This indicates that the majority of estimates remained close to their true values despite the poor overall reconstruction metrics. However, PLS greatly overestimated small continuous magnitudes across the entire column while variable density remained relatively unchanged in the pattern of this replicate. After applying the collinearity reduction, these very small continuous values were clustered into points, leading to a substantial reduction in both HD and AOS. This example illustrates one mechanism behind the drastic degradation using a single replicate and six representative columns, similar to other methods may also occur in the remaining unshown columns or other repetitions. Additional examples illustrating how such a small value affects analysis under other dimensionalities and proportions are included in **Appendix B**.

In addition to the wide dynamic range of the ratio data set—spanning several orders of magnitude—the multiplicative nature of mult_repl , mult_lognorm , and mult_KMSS occasionally produced hd (smiles) containing negative imputed values.

4.5 Assessment of downstream PCA and data analysis

While the previous sections focused on reconstruction accuracy and comparison of these methods, we recently evaluated the quality of downstream analyses. We further evaluated the imputed data from the perspective of low-dimensional structure recovery and subsequent analysis. The performance of this was assessed by comparing the hd (smiles) data relative to the estimated hd (smiles) and the true hd (smiles). Specifically, for each simulation replicate, the true component matrix and the imputed matrix were first QR factorized and divided into two PCA scores were obtained from the first two principal components ($k = 2$). The hd (smiles) data was then computed as the minimal squared difference between the estimated and true PCA configurations after optimal rotation and reflection (sign changes), without scaling.

$$m^2 = \sum_{i=1}^n \sum_{r=1}^k (\tilde{f}_{ir, \text{fitted}} - f_{ir, \text{true}})^2, \quad (8)$$

where $f_{ir, \text{true}}$ denotes the PCA score of sample i on principal component r obtained from the complete data and $\tilde{f}_{ir, \text{fitted}}$ denotes the corresponding aligned PCA score after hd (smiles) imputation. This m^2 is the hd (smiles) squared sum of squares with small values indicating better recovery of the low-dimensional sample configuration. Based on the previous analyses, representative combinations of varied dimensions ($M = 50, 100, 200, 300$) and proportions ($p = 0.2, 0.5, 0.8$) were evaluated over 500 independent simulation replicates.

Fig. 10 presents the resulting PCA hd (smiles) (m^2) across these scenarios. The inherent multiplicative properties (mult_lognorm ,

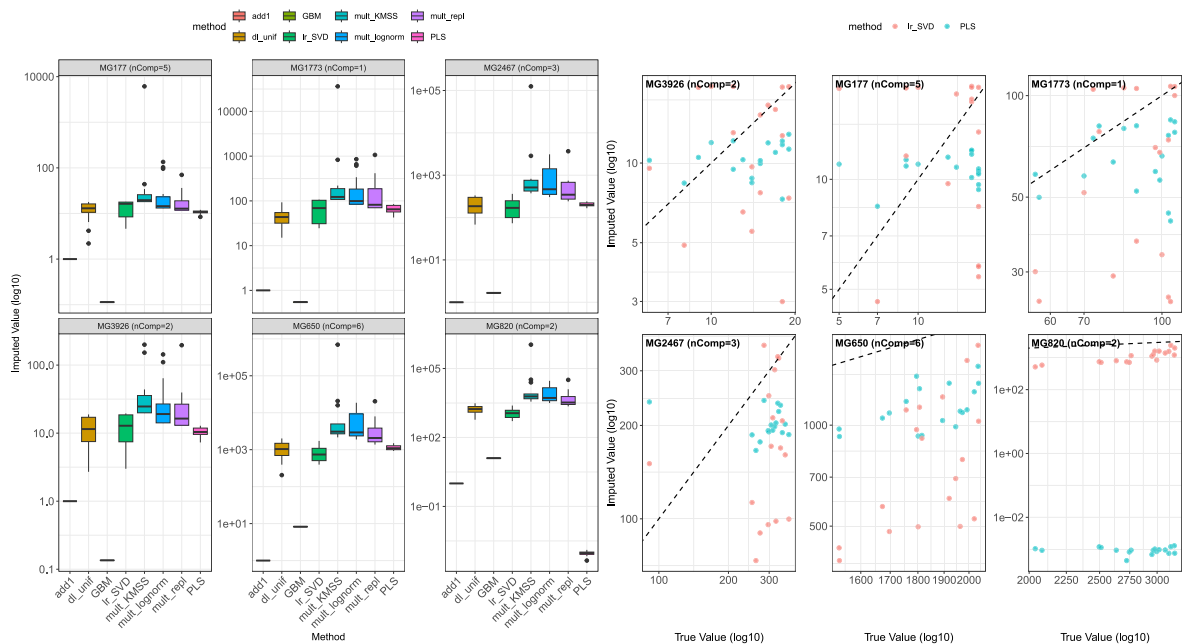


Fig 9 Imputation of missing data under $m = 200$, $p = 0.2$, and $\rho = 1$. The top row of the figure shows the distribution of imputed values across different imputation methods for expression data variables (volcanic data and lowly expressed genes), while the corresponding plots of PLS components (v_{comp}) and clinical data. The bottom row shows the distribution of imputed values across different imputation methods for expression data variables (volcanic data and lowly expressed genes), while the corresponding plots of PLS components (v_{comp}) and clinical data. The bottom row shows the distribution of imputed values across different imputation methods for expression data variables (volcanic data and lowly expressed genes), while the corresponding plots of PLS components (v_{comp}) and clinical data.

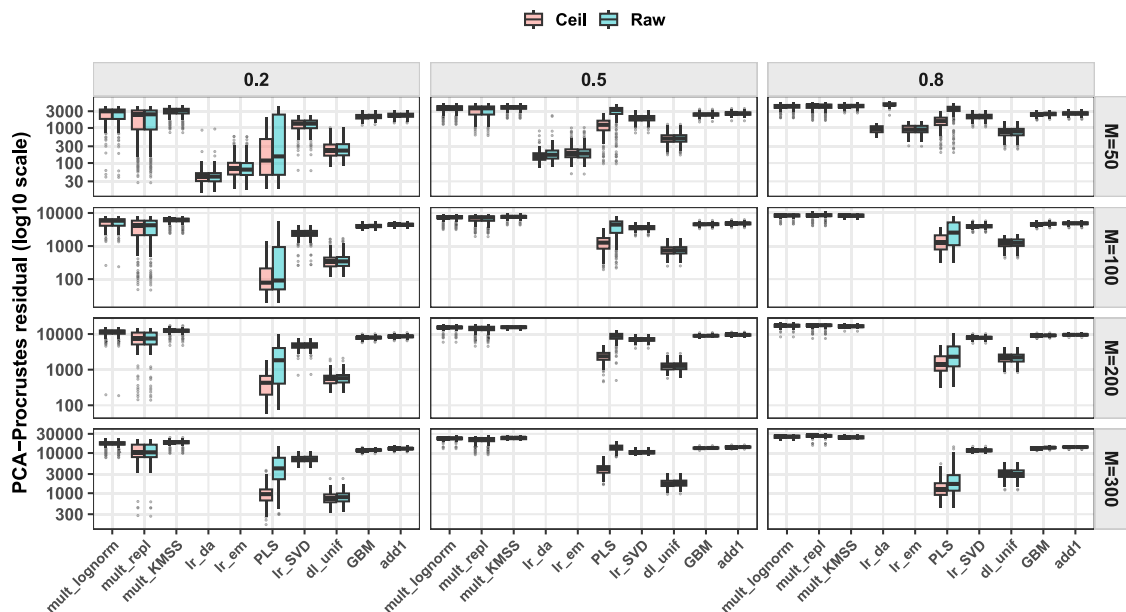
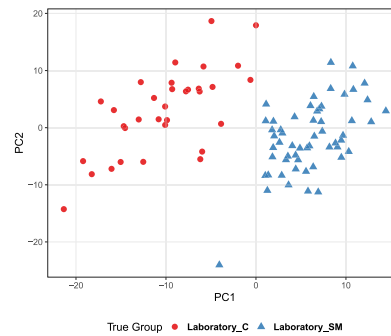
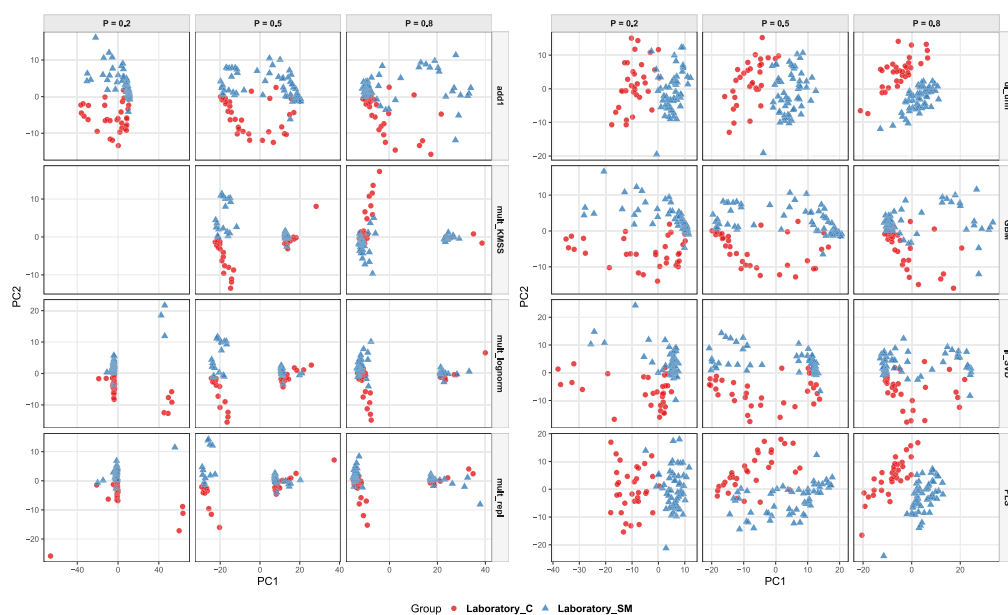


Fig 10 PCA-Procrustes residual (r^2) for different imputation methods across varying datasets (M) and proportions (p). The top row of the figure shows the distribution of imputed values across different imputation methods for expression data variables (volcanic data and lowly expressed genes), while the corresponding plots of PLS components (v_{comp}) and clinical data. The bottom row shows the distribution of imputed values across different imputation methods for expression data variables (volcanic data and lowly expressed genes), while the corresponding plots of PLS components (v_{comp}) and clinical data.



(a) PCA scatter plot of original matrix (ARI = 1.00, NMI = 1.00)



(b) PCA scatter plots of ceil imputed matrix

Fig. 11. Representative PCA scatter plots of the original complete data and the corresponding imputed data for the $M = 400$ features subset under varying proportions ($P = 0.2, 0.5$ and 0.8). Smiles are clustered and ordered according to their laboratory groups (Laboratory_C records, Laboratory_SM, etc.). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

To further assess the downstream impact of imputation, we evaluated the recovery of sample group structure through clustering analysis. The R-part dataset contains samples from two laboratories (33 from Laboratory_C and 67 from Laboratory_SM). Clustering performance was evaluated using two complementary measures: the Adjusted Rand Index (ARI) and the Normalized Mutual Information (NMI). ARI measures the agreement between the predicted clusters and the true laboratory labels while correcting for chance. NMI quantifies the amount of shared information between the predicted and true groupings. For both metrics, values closer to 1 indicate better recovery of the original group structure.

For visual assessment, we first considered a high-dimensional setting with $M = 400$ variables and examined the recovery of the two-dimensional PCA configuration after imputation. A representative subset for which the original complete data showed clear separation between the two laboratory groups (ARI = 1.00 and NMI = 1.00) was selected for visual comparison, as shown in Fig. 11(a). Fig. 11(b)

presents the corresponding PCA scatter plots of the imputed data under different proportions, allowing a qualitative comparison and how well the original sample structures were preserved by each imputation method.

The PCA scatter plots reveal clear differences among the imputation methods in their ability to preserve the original sample structure. The multiplicative approaches (mult_lognorm, mult_repl, and mult_KMSS) exhibit little separation between the two laboratory groups, whereas PLS and al_unif largely preserve the original group structure across different sample sizes. Methods such as lr_SVD, and GBM recover some degree of separation although the group boundaries are less distinct than in the original data.

To further quantify the recovery of downstream structure, we evaluated clustering performance based on the PCA scores for each impute. K-means clustering ($k = 2$) was performed using the first two principal components. To compare the imputed data with the original complete data, we report the difference $\Delta = \text{Base} - \text{Imputed}$, where

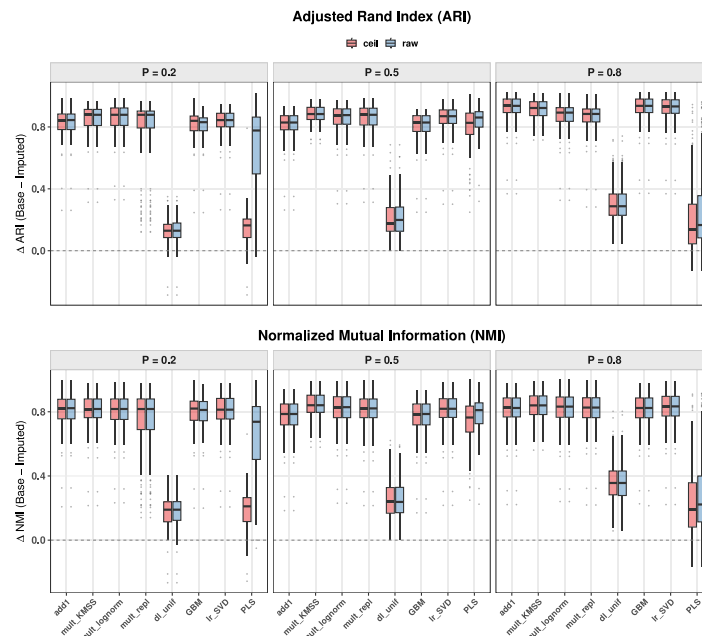


Fig. 12. Difference in performance ($\Delta = \text{Base} - \text{Imputed}$) under Keras using ReLU as shown for the sparsity levels ($p = 0.2, 0.5$ and 0.8). Scale Δ values indicate better recovery of the original structure after imputation.

Table 8. Mean difference in Δ ARI and Δ NMI between the original complete data and the imputed data under Keras using ReLU as shown for the sparsity levels ($p = 0.2, 0.5$ and 0.8). Scale Δ values indicate better recovery of the original structure after imputation. Results are shown as mean Δ for the corresponding results. The reported values are based on 20 independent simulation replicates.

Method	Delta Adjusted Rand Index (Δ ARI)					
	$p = 0.2$		$p = 0.5$		$p = 0.8$	
	Raw	Cell	Raw	Cell	Raw	Cell
add1	0.824	0.824	0.848	0.848	0.920	0.920
mult_KMSS	0.820	0.820	0.825	0.825	0.915	0.915
mult_lognorm	0.801	0.801	0.820	0.820	0.920	0.920
mult_repl	0.748	0.747	0.826	0.826	0.857	0.857
dl_unif	0.738	0.735	0.826	0.824	0.923	0.925
GBM	0.814	0.815	0.843	0.843	0.926	0.921
lr_SVD	0.825	0.825	0.843	0.841	0.923	0.923
HS	0.662	0.455	0.847	0.817	0.928	0.921
Method	Delta Normalized Mutual Information (Δ NMI)					
	$p = 0.2$		$p = 0.5$		$p = 0.8$	
	Raw	Cell	Raw	Cell	Raw	Cell
add1	0.827	0.827	0.849	0.849	0.890	0.890
mult_KMSS	0.827	0.827	0.824	0.824	0.893	0.893
mult_lognorm	0.825	0.825	0.824	0.824	0.896	0.896
mult_repl	0.750	0.75	0.823	0.823	0.837	0.837
dl_unif	0.804	0.805	0.820	0.819	0.845	0.848
GBM	0.799	0.795	0.820	0.819	0.845	0.848
lr_SVD	0.813	0.813	0.821	0.819	0.847	0.845
HS	0.659	0.494	0.855	0.746	0.877	0.822

scale values indicate better recovery of the original structure. Fig. 12 summarizes the distribution of Δ ARI and Δ NMI across the 20 repeated simulation replicates for different imputation methods and sparsity levels. Table 8 summarizes the corresponding mean Δ ARI and Δ NMI values for each method.

Overall, the three multiplicative approaches (mult_lognorm, mult_repl, and mult_KMSS), together with add1, GBM, and lr_SVD, performed relatively large Δ ARI and Δ NMI values across the three sparsity levels. In contrast, dl_unif consistently produced the smallest differences from the original complete data. Adifferent

pattern was observed for PLS. Although it exhibited relatively poor performance according to the reconstruction-based metrics (CD and AOS), this was not reflected in the downstream usage analysis, which was biologically smaller Δ ARI and Δ NMI values were obtained. As shown in Table 8, the effect of the cell types processing varied across the imputation methods. The poor performance of PLS was observed for PLS. At $p = 0.2$, the mean Δ ARI decreased from 0.662 to 0.455, while the mean Δ NMI decreased from 0.659 to 0.494 after cell type correction. Moreover, the results were observed at $p = 0.5$ (Δ ARI: 0.847 to 0.817, Δ NMI: 0.855 to 0.746) and

$p = 0.8$ (AFC 02618 to 0271; NMI 0287 to 0292). A smaller but consistent pattern was also observed for dl_unif at $p = 0.2$ and $p = 0.5$. For the remaining paths, the differences between the two analytical results were negligible.

Be aware that these observations depend on the choice of distribution to model the analysis using a Gaussian mixture model (GMM). The corresponding results are presented in Appendix C and are consistent with those obtained using kernels during

46 Discussion and recommendations

Several findings suggest that caution is required when evaluating imputation methods for compositional count data. Performance rankings based on reconstruction metrics do not always align with downstream analytical performance. Recent machine learning (PLS), which exhibits strong performance under the reconstruction benchmark (GDA and AOS) but is relatively inferior in downstream PCA, Hotelling and clustering analysis. This discrepancy demonstrates that metrics showing large reconstruction errors do not necessarily fail to preserve biologically meaningful gap structure in compositional count data.

A notable characteristic of PLS is the occasional generation of extremely small positive imputed values. Although most imputed values align closely with the reference line $y = x$, indicating strong preservation of potential relationships between zero estimates—such as those observed for variable `mg820` in Fig. 9—enrichment of extremely small values in individual replicates (e.g., GDA and AOS) may deviate from the logarithmic framework. These tiny positive values are disproportionately amplified during their scores. Applying a ceiling operation corrects these values by capping them at a threshold, effectively mitigating logarithmic amplification, consistent with the earlier discussion that quantization may have the strongest impact on extremely small values—particularly those below 1—where values are less than 1 are rounded up to 1 and thus the extreme outliers in the logarithmic transformation are greatly alleviated. This reduction in error in the model is decreased and is less sensitive to outliers than an imputation method in the underlying predictive accuracy. Similar rare value outliers may occur in other rare variables of across different replicates, even though they are not all shown here. As planned, a comparison study based on simulated count data is provided in Appendix A, further confirming that the observed behavior of PLS is not specific to the Rabbit dataset but reflects a more general mechanism. Additional examples illustrating how such rare value outliers are under other distributions and zero positions are included in Appendix B. The PCA, Hotelling and clustering analysis provide further support for this interpretation. One possible explanation is that these downstream analyses primarily assess recovery of the first two principal components, whereas GDA and AOS evaluate the means across the complete structure as a whole. Despite being inferior to performance under GDA and AOS, PLS consistently preserved the low dimensional sample structure and the separation between the two biological groups. The ceiling operation produced the largest downstream impacts for HSS, while the imputation methods diverged only marginally.

Building on these observations, the empirical findings are broadly consistent with the theoretical considerations discussed in Section 3.3. As discussed, the extremely small positive values can be disproportionately amplified by logarithmic transformations, hindering reconstruction and metrics, particularly sensitive to rare zero imputations. The behavior observed for PLS provides an empirical illustration of this mechanism.

These findings also have practical implications for method selection. Reconstruction accuracy alone is insufficient for evaluating imputation methods for compositional count data. Metrics such as $1r_SVD$ provide a better reconstruction performance, whereas PLS may better

preserve downstream structures despite less favorable reconstruction performance. Although dl_unif often achieves competitive results for structural integrity and interpretability, the generally reconstruction metrics should not be interpreted in isolation when analyzing compositional count data. Where possible, conclusions should be verified against downstream analytical objectives. Our results further suggest that merely preserving the discrete nature of sequencing counts, for example through smaller integer mapping procedures, may facilitate the recovery of meaningful data structures and contribute to more stable downstream analysis. Importantly, the ceiling operation applied in this study should be viewed as a simple proof of concept rather than a final solution. The broader implication is that replacement methods may benefit from considering the discrete nature of count compositions by mapping continuous imputations back onto the discrete support. This perspective is consistent with the conceptual framework of discrete replacement proposed by Loel et al. [7], who suggested replacing zeros with positive discrete values below one count derived from the empirical distribution of imputed observations to preserve the discrete nature of count data. Our results suggest that preserving the relative structure of count compositions may be a more important consideration in the context of the specific mapping strategy adopted. Alternative lattice-preserving mapping or discrete imputation strategies warrant further investigation.

5 Conclusion

In this review, we examined the challenges of zero values in compositional data motivated in particular by high dimensional and extremely sparse sequencing data and microbial datasets. As sequencing technologies continue to advance, such data have become increasingly common, yet conventional logarithmic compositional data analysis remains structurally invalid due to the division by zero. The problem becomes even more complex in sequencing count data, where sequences are discrete and lattice-valued rather than continuous, violating key assumptions underlying classical GDA.

We began by testing the generic functions of compositional data analysis, emphasizing the importance of distinguishing between rounded zeros and exact zeros, as they represent different mechanisms and different treatment strategies. Building on this distinction, we summarized the major methodological directions: zero detection, transformation, zero replacement strategies that constructively preserve compositions, and model-based approaches that treat zeros as statistically meaningful quantities. We then reviewed probabilistic methods specifically designed for count compositional data and finally discussed the theoretical consequences of applying continuous logarithmic GDA to discrete lattice-valued count compositions, highlighting the mismatch between classical GDA assumptions and the nature of HSS data.

These theoretical considerations motivated the empirical part of this study. Rather than evaluating specific probabilistic models or objective measures to examine the practical behavior of the classical logarithmic framework in applications to high dimensional sparse HSS count data, and to investigate whether the theoretical limitations discussed above have measurable consequences in practice. To this end, we conducted an end-to-end comparison of several alternative imputation methods operating within the same framework, evaluating their performance in terms of numerical accuracy, preservation of compositional structure, and downstream analytical performance. The comparative experiments were conducted on both real high dimensional datasets and simulated data generated from real microbial compositions. Our results reveal that conventional reconstruction metrics can be misleadingly reflective of downstream analytical performance. Several logarithmic-based methods achieved consistently strong numerical performance. However, methods with similar reconstruction accuracy could still produce markedly different PCA and clustering results. Further qualitative analyses showed that extremely small imputed values can be

The PCA biplots analysis on the simulated data further supports the confounders obtained from the real machine data. As shown in Fig. 16, PLS discriminates feed degradation level under the restriction by dynamics (GD and AOS). Instead, it

19

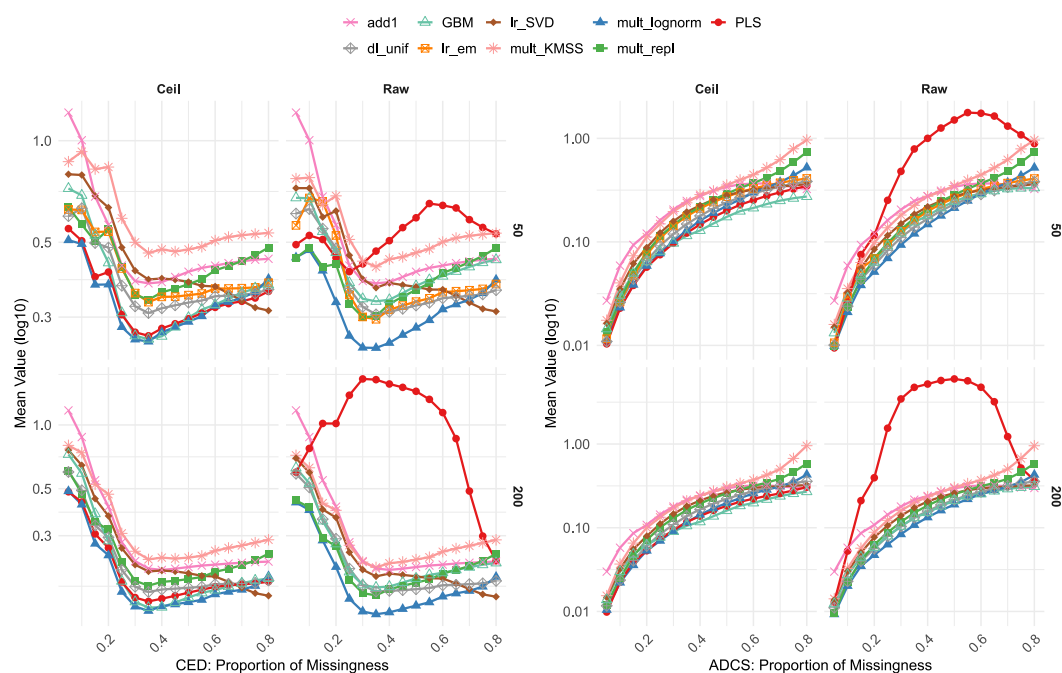


Fig 13 Comparison of imputation methods for CED and ADCS data under varying dimensionality and missingness probabilities with simulation data.

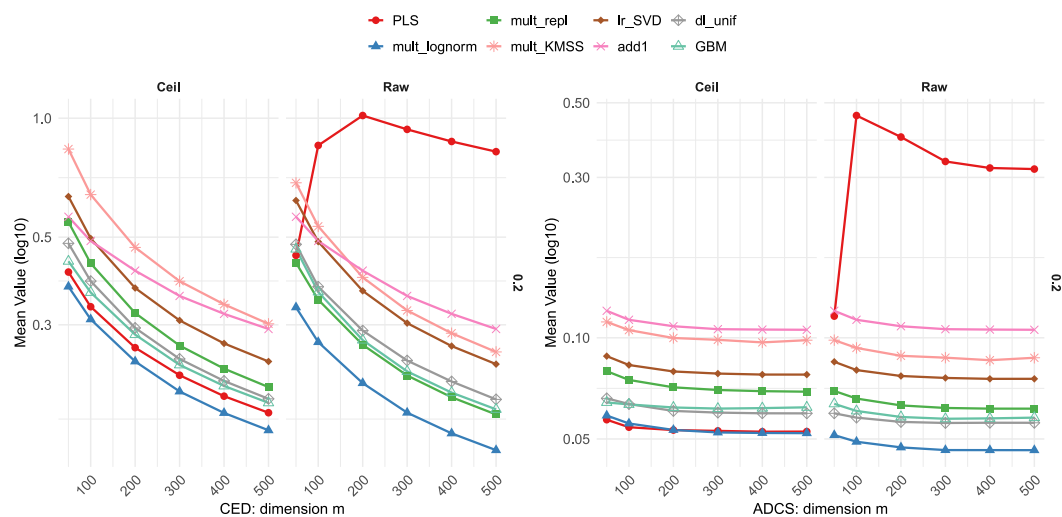
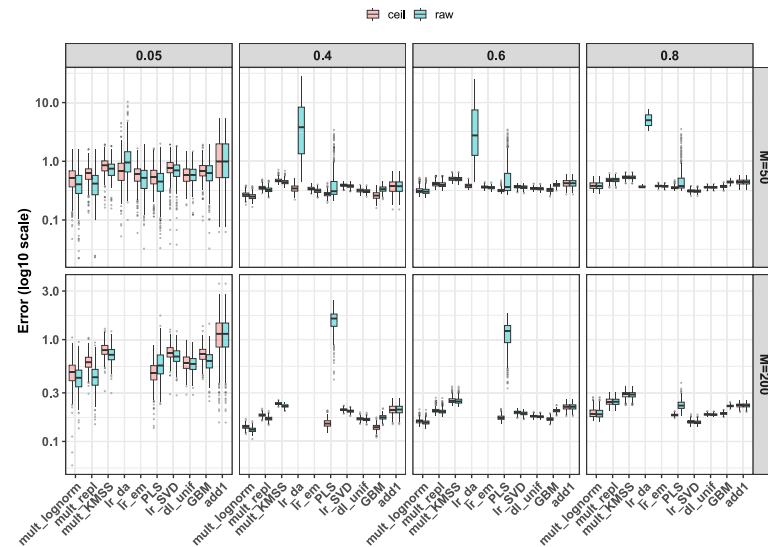
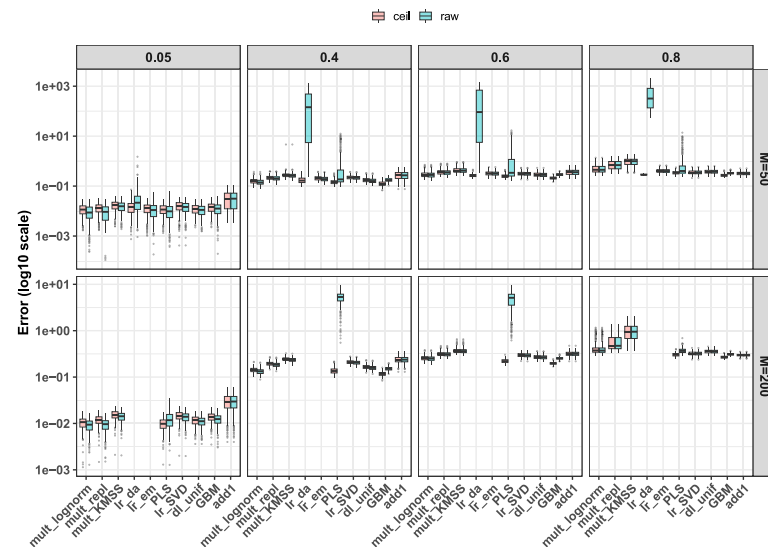


Fig 14 Comparison of CED and ADCS data across different dimensionality and fixed missingness probability of $p = 0.2$ with simulation data.



(a) CED



(b) ADCS

Fig. 15 Simulation comparison of input methods inters of CED (top) and ADCS (bottom) metrics across dimensionality ($m = 50, 200$) and noise levels.

PCA eigenpairs compared to that of seed data points across different noise and sample sizes indicating that large reconstruction errors are easily identified and removed when multivariate prior are used.

The following presents a detailed analysis of the input methods in PCA for seed data points including P_{LS} . Although the proposed method is more robust than the other methods, the results for the seed data points are consistent with the results obtained from the real data, providing additional support to the various input methods in the text.

Appendix B Additional diagnostic figures

To complement the localized examples used in the main text, this appendix provides additional local diagnostic plots that follow the same experimental design as the main body, with the same data generating mechanism and evaluation procedure, but using different combinations of dimensionality (m) and proportion (p). These plots illustrate how robust the proposed input methods are under various settings of dimensionality m and p as detailed

Table 9
Root mean square error (RMSE) and mean absolute error (MAE) for PLS across varying missingness levels (ρ) for the Yarn data.

ρ	$M = 50$				$M = 200$			
	RMSE		MAE		RMSE		MAE	
	Pop	Median	Pop	Median	Pop	Median	Pop	Median
0.05	0.006	0.004	0.005	0.003	0.006	0.004	0.005	0.003
0.1	0.007	0.005	0.006	0.004	0.007	0.005	0.006	0.004
0.15	0.008	0.006	0.007	0.005	0.008	0.006	0.007	0.005
0.2	0.009	0.007	0.008	0.006	0.009	0.007	0.008	0.006
0.25	0.010	0.008	0.009	0.007	0.010	0.008	0.009	0.007
0.3	0.011	0.009	0.010	0.008	0.011	0.009	0.010	0.008
0.35	0.012	0.010	0.011	0.009	0.012	0.010	0.011	0.009
0.4	0.013	0.011	0.012	0.010	0.013	0.011	0.012	0.010
0.45	0.014	0.012	0.013	0.011	0.014	0.012	0.013	0.011
0.5	0.015	0.013	0.014	0.012	0.015	0.013	0.014	0.012
0.55	0.016	0.014	0.015	0.013	0.016	0.014	0.015	0.013
0.6	0.017	0.015	0.016	0.014	0.017	0.015	0.016	0.014
0.65	0.018	0.016	0.017	0.015	0.018	0.016	0.017	0.015
0.7	0.019	0.017	0.018	0.016	0.019	0.017	0.018	0.016
0.75	0.020	0.018	0.019	0.017	0.020	0.018	0.019	0.017
0.8	0.021	0.019	0.020	0.018	0.021	0.019	0.020	0.018
0.85	0.022	0.020	0.021	0.019	0.022	0.020	0.021	0.019
0.9	0.023	0.021	0.022	0.020	0.023	0.021	0.022	0.020
0.95	0.024	0.022	0.023	0.021	0.024	0.022	0.023	0.021
1.0	0.025	0.023	0.024	0.022	0.025	0.023	0.024	0.022

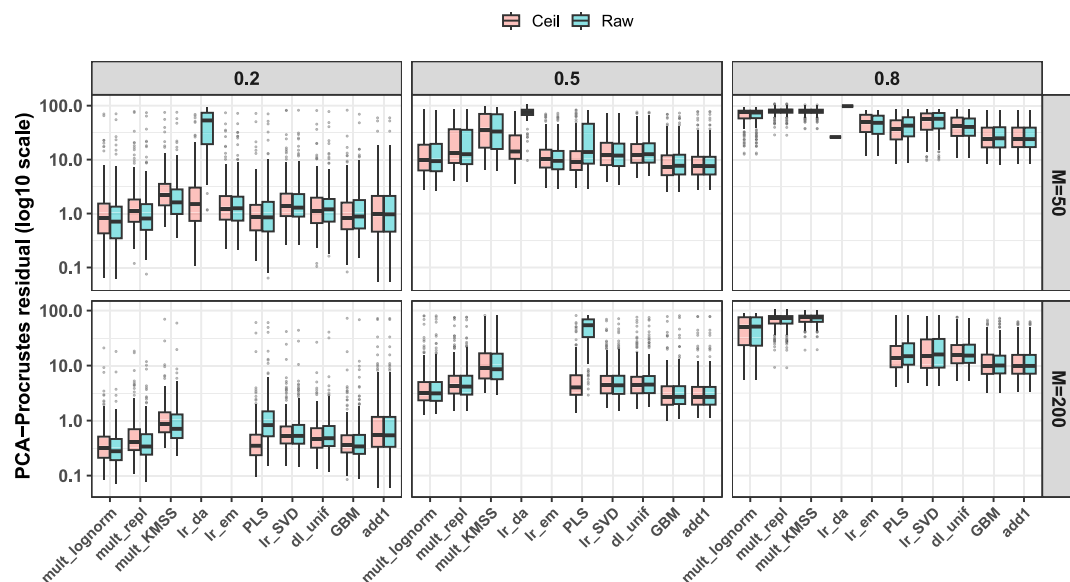


Fig. 16. Simulation-based comparison of PCA Procrustes residuals (Raw and Cell) across different dimensions and missingness probabilities.

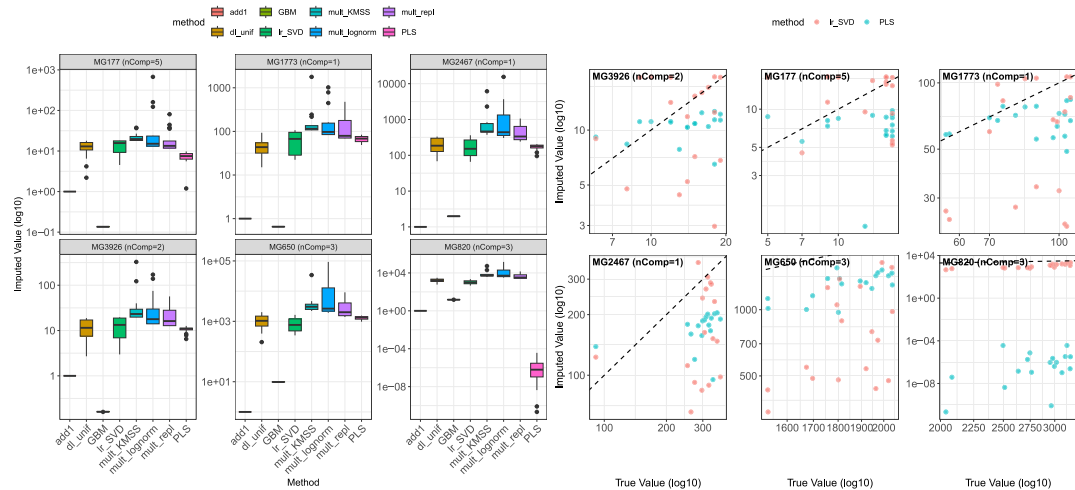
Table 10
Root mean square error (RMSE) and mean absolute error (MAE) for PLS across varying dimensions (M) at fixed missingness ($\rho = 0.2$) for the special pin data.

M	RMSE		MAE	
	Pop	Median	Pop	Median
50	0.04	0.02	0.03	0.02
100	0.04	0.02	0.03	0.02
200	0.04	0.02	0.03	0.02
300	0.04	0.02	0.03	0.02
400	0.04	0.02	0.03	0.02
500	0.04	0.02	0.03	0.02

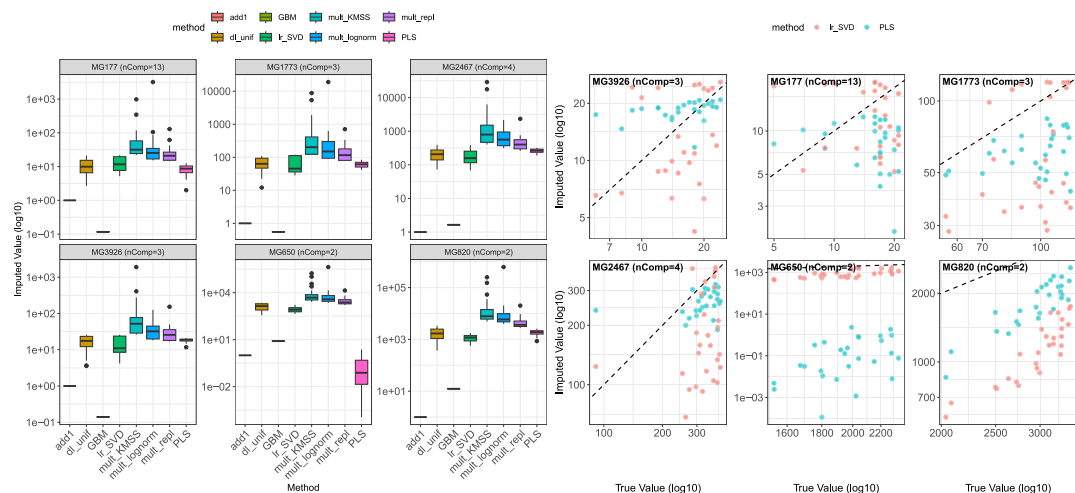
The examples are two purposes. First, they demonstrate that the proposed near-computational strategy greatly reduces the

not limited to the specific case highlighted in the main manuscript (e.g., variable mg820), but also occurs across a broader range of simulation settings. Second, they show that the proposed strategy tends to yield smaller values than other methods, suggesting that this issue deserves further attention when log-ratios are applied to count data. This phenomenon may arise from the underlying computational settings and is not directly related to the particular variable or dimension configuration involved.

Figs. 7(a) and 7(b) present two representative instances showing that, under different dimensions and proportions settings, a single variable can impact the overall simulation results. Under PLS, resulting in the median difference values



(a) $m = 100$, $p = 0.20$, Abnormal variable MG820.



(b) $m = 500$, $p = 0.30$, Abnormal variable MG650.

Fig. 17. (a) corresponds to $m = 100$, $p = 0.20$ with abnormal variable MG820, and (b) corresponds to $m = 500$, $p = 0.30$ with abnormal variable MG650. Each plot summarizes the distribution of imputed values across methods for the variable imputed component for repetitive variables.

Appendix C Additional clustering analysis using Gaussian mixture models

To evaluate whether the conditions described in the first clustering algorithm were replicated by the Gaussian mixture analysis, an alternative comparison of MM features was conducted using a Gaussian mixture model (GMM) to cluster the data and identify the clusters.

Figure 18 shows the distribution of the differences in AIC and BIC between the original complete data and the imputed data over 20 simulation replicates. Overall, the GMM successfully identified the clusters using K-means clustering, but the results were not as consistent as those obtained using the Gaussian mixture model (see Table 1).

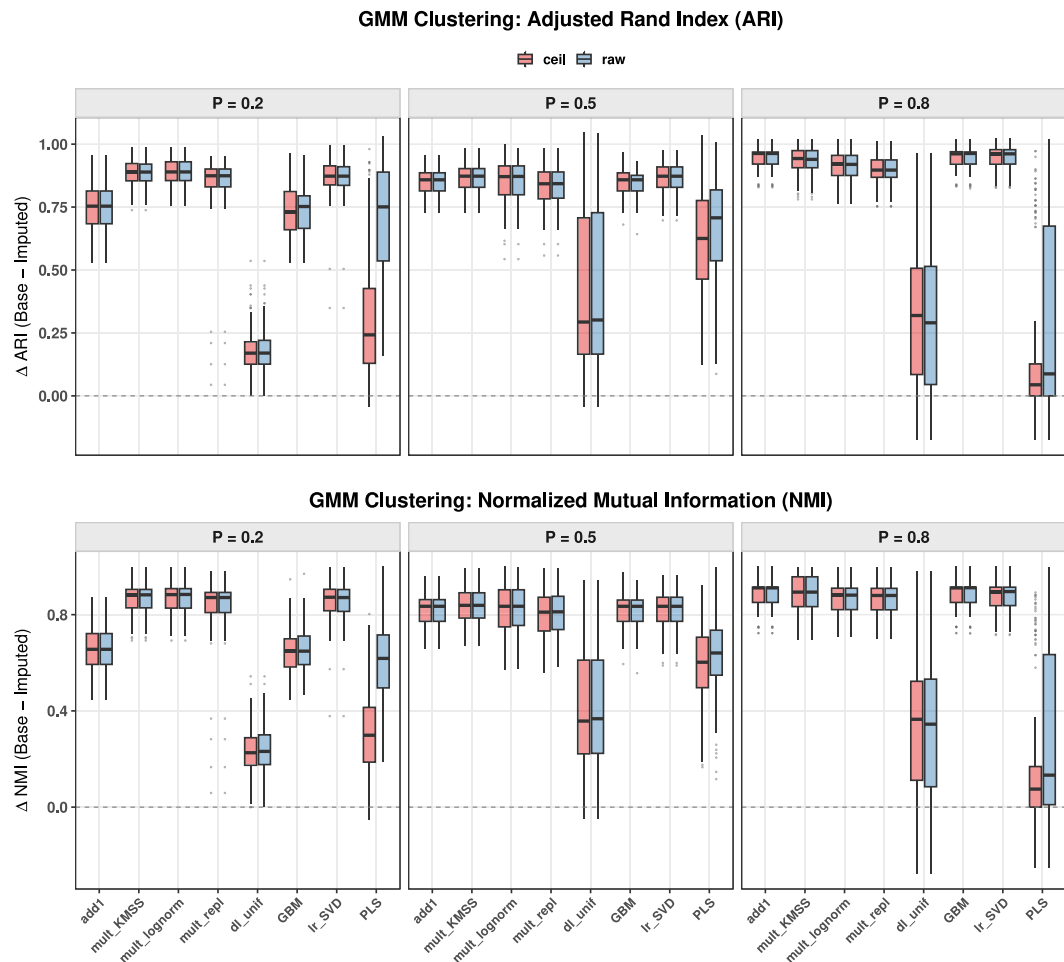


Fig. 18. Differences in clustering performance ($\Delta = \text{Base} - \text{Imputed}$) under Gaussian mixture model (GMM) using Shale. Δ values indicate the recovery of the original data after imputation. The color plots are consistent with the banding known during the Section 4.5. The difference in Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) between the original complete data and the imputed data under Gaussian mixture model (GMM) using the Shale. Δ values indicate the recovery of the original data after imputation. The color plots are consistent with the banding known during the Section 4.5.

Method	Data Adjusted Rand Index (Δ ARI)		Data Normalized Mutual Information (Δ NMI)	
	Raw	Cell	Raw	Cell
p = 0.2				
add1	0.743	0.743	0.672	0.672
mult_KMSS	0.836	0.836	0.876	0.876
mult_lognorm	0.849	0.849	0.874	0.874
mult_rep1	0.852	0.852	0.875	0.875
dl_unit	0.179	0.174	0.243	0.243
GBM	0.736	0.736	0.651	0.651
lr_SVD	0.833	0.833	0.832	0.832
PLS	0.625	0.615	0.632	0.625
p = 0.5				
add1	0.849	0.849	0.832	0.832
mult_KMSS	0.833	0.833	0.832	0.832
mult_lognorm	0.833	0.833	0.832	0.832
mult_rep1	0.833	0.833	0.832	0.832
dl_unit	0.427	0.417	0.474	0.474
GBM	0.833	0.833	0.832	0.832
lr_SVD	0.833	0.833	0.832	0.832
PLS	0.621	0.611	0.626	0.626
p = 0.8				
add1	0.957	0.957	0.895	0.895
mult_KMSS	0.957	0.957	0.895	0.895
mult_lognorm	0.957	0.957	0.895	0.895
mult_rep1	0.957	0.957	0.895	0.895
dl_unit	0.316	0.316	0.316	0.316
GBM	0.957	0.957	0.895	0.895
lr_SVD	0.957	0.957	0.895	0.895
PLS	0.173	0.173	0.173	0.173

- 26

Application to geochemistry

Fačevicová, K., Bábek, O., Hron, K., Kumpan, T. Element chemostratigraphy of the Devonian/Carboniferous boundary – A compositional approach. *Applied Geochemistry*, 75, 211-221 (2016).

<https://doi.org/10.1016/j.apgeochem.2016.10.002>

[IF 3.5 Q1; AIS 0.744 Q2]

[15 citations on WoS; 10 citations without autocitations]

[Author's contribution: 50%]

Abstract

The Devonian/Carboniferous (D/C) boundary is a critical interval in the Phanerozoic history, which is associated with vigorous climatic perturbations, continental glaciation, global sea-level fall and rapidly increased extinction rates in marine realms. In many sections world-wide, these global changes left a marked lithological signature, in particular the Hangenberg black shale (products of deep-shelf anoxia) and the overlying Hangenberg sandstone (sudden siliciclastic influx into predominantly carbonate depositional environments). Both layers bear a distinct geochemical signature. Even though either or both of these two lithologies are absent at many sections, their correlative counterparts can be indicated by subtle geochemical markers. We studied elemental geochemistry of fourteen D/C boundary sections in six key areas across Europe with the aim to select globally correlatable elemental proxy for the D/C boundary. Analysis of raw/log-transformed geochemical data (EDXRF, c.p.s. units), presenting the standard approach here, indicates that concentrations of terrigenous elements (Al, K, Rb, Ti and Zr) are mainly controlled by diluted Ca (carried by marine calcium carbonate) in limestone facies and, accordingly, their variations can be related

to carbonate production in the sea rather than to terrigenous input from continent. Nevertheless, due to the relative nature of geochemical observations, reliance solely on statistical processing of raw data might lead to incomplete picture of multivariate data structure and/or biased interpretations. For this reason, the aim of this contribution is to discuss the logratio alternatives of the standard statistical methods, which may better reflect the relative nature of the data. For this purpose, principal component analysis was employed to reveal main geochemical patterns and while the geochemical signature of the D/C boundary was further analysed using Q-mode clustering that leads to predicative orthonormal logratio coordinates – balances. The comprehensive picture of the multivariate data structure provided by these statistical tools makes them a primary choice for exploratory compositional data analysis. At the same time, it turns out that the standard and compositional approaches have synergic effects. This fact can be extensively used in further geochemical studies.



Contents lists available at ScienceDirect

Applied Geochemistry

journal homepage: www.elsevier.com/locate/apgeochem

Element chemostratigraphy of the Devonian/Carboniferous boundary – A compositional approach

K. Fačevicová^{a,*}, O. Bábek^b, K. Hron^a, T. Kumpan^c^a Department of Mathematical Analysis and Applications of Mathematics – Palacký University Olomouc, Czechia^b Department of Geology – Palacký University Olomouc, Czechia^c Department of Geological Sciences – Masaryk University Brno, Czechia

ARTICLE INFO

Article history:

Received 29 February 2016

Received in revised form

30 September 2016

Accepted 5 October 2016

Available online 6 October 2016

Keywords:

Element geochemistry

Compositional biplot

Dendrogram

Logratio coordinates

Devonian/Carboniferous boundary

Variation matrix

ABSTRACT

The Devonian/Carboniferous (D/C) boundary is a critical interval in the Phanerozoic history, which is associated with vigorous climatic perturbations, continental glaciation, global sea-level fall and rapidly increased extinction rates in marine realms. In many sections world-wide, these global changes left a marked lithological signature, in particular the Hangenberg black shale (products of deep-shelf anoxia) and the overlying Hangenberg sandstone (sudden siliciclastic influx into predominantly carbonate depositional environments). Both layers bear a distinct geochemical signature. Even though either or both of these two lithologies are absent at many sections, their correlative counterparts can be indicated by subtle geochemical markers. We studied elemental geochemistry of fourteen D/C boundary sections in six key areas across Europe with the aim to select globally correlatable elemental proxy for the D/C boundary. Analysis of raw/log-transformed geochemical data (EDXRF, c.p.s. units), presenting the standard approach here, indicates that concentrations of terrigenous elements (Al, K, Rb, Ti and Zr) are mainly controlled by diluted Ca (carried by marine calcium carbonate) in limestone facies and, accordingly, their variations can be related to carbonate production in the sea rather than to terrigenous input from continent. Nevertheless, due to the relative nature of geochemical observations, reliance solely on statistical processing of raw data might lead to incomplete picture of multivariate data structure and/or biased interpretations. For this reason, the aim of this contribution is to discuss the logratio alternatives of the standard statistical methods, which may better reflect the relative nature of the data. For this purpose, principal component analysis was employed to reveal main geochemical patterns and while the geochemical signature of the D/C boundary was further analysed using Q-mode clustering that leads to predicative orthonormal logratio coordinates – balances. The comprehensive picture of the multivariate data structure provided by these statistical tools makes them a primary choice for exploratory compositional data analysis. At the same time, it turns out that the standard and compositional approaches have synergic effects. This fact can be extensively used in further geochemical studies.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

In geochemistry, most of data are compositional in nature (Aitchison, 1986; Pawlowsky-Glahn and Buccianti, 2011; Pawlowsky-Glahn et al., 2015a). It is not just indicated by units, in which observations are measured or expressed, like mg/kg, ppm, or percentages, but inherently also in the fact that ratios between components form the primary source of information. As a consequence, any sum of components (compositional parts) is irrelevant.

In particular, in geochemical practice it rarely happens that the input observations sum up to a constant given by the mode of used representation. For example, to make mg/kg units sum up to unity (one million) it would mean that all elements in the rock were identified and analysed. From this perspective, it is much more convenient to treat the data as compositional, if their parts convey quantitatively expressed relative contributions on a whole, formed by the given composition. The above properties also imply that compositional data carry exclusively relative information (Pawlowsky-Glahn et al., 2015a).

Accordingly, when the relative structure of geochemical observations is of the sole interest, applying standard statistical tools to

* Corresponding author.

E-mail address: kamila.facevicova@gmail.com (K. Fačevicová).

the input concentrations may lead to misleading results, because compositional data obey different geometrical rules. These inherent features can be expressed by principles of compositional data analysis (Egozcue, 2009): scale invariance, permutation invariance and subcompositional coherence. From the practical perspective, the most important one is scale invariance stating that information in a composition does not depend on the particular units in which the composition is represented. Specifically, proportional positive vectors represent the same composition. The latter two principles provide a solid theoretical basis of any reasonable (not exclusively) statistical processing, and they all are needed to build a sound geometrical setting that would reflect inherent properties of geochemical (compositional) data. It is provided by the Aitchison geometry (Billheimer et al., 2001; Pawłowsky-Glahn and Egozcue, 2001; Egozcue et al., 2003), defining an algebraic-geometrical structure of the sample space of compositions, formed by equivalence classes of proportional positive vectors (or by simplex for a given constant-sum representation of compositions). As the majority of statistical methods rely on Euclidean geometry in real space (Eaton, 1983), it is necessary to transform the compositions prior to standard statistical analysis with such as the principal component analysis (PCA) and/or visualization. The definition of compositional data implies that any such transformation (from geometrical reasons, referred to as coordinates) should be formed by ratios between parts that form the elemental information in compositional data. Or, even better, by log-ratios (Aitchison, 1986) that symmetrize positions of components in the ratio and are mathematically easier to handle.

For a standard geochemical analysis, it is preferred to deal with the original compositional parts, not their log-ratios. However, with all consequences that imply from the nature of compositions considered, this is frequently not satisfactory. Nevertheless, if interpretation in terms of single original part is required, centred logratio (clr) coordinates provide a compositional alternative to raw data. Each part of vector of these coordinates.

$$\text{clr}(\mathbf{x}) = [y_1, \dots, y_D] = \left[\ln \frac{x_1}{g(\mathbf{x})}, \ln \frac{x_2}{g(\mathbf{x})}, \dots, \ln \frac{x_D}{g(\mathbf{x})} \right], \quad (1)$$

where $g(\mathbf{x})$ stands for geometric mean.

$$g(\mathbf{x}) = \sqrt[D]{\prod_{i=1}^D x_i} = \exp\left(\frac{1}{D} \sum_{i=1}^D \ln x_i\right), \quad (2)$$

represents a dominance of x_i with respect to the complete composition. Because the resulting vector has D components, a redundancy condition arises, $y_1 + \dots + y_D = 0$, leading to singular covariance matrix of $\text{clr}(\mathbf{x})$. Although this affects applicability of some statistical approaches, like the class of robust methods (Filzmoser et al., 2009a), it is still possible to employ them for most of exploratory tools including PCA. For more discussion about the use of clr coordinates in geochemical context, see McKinley et al. (2016).

Not all geochemical data must necessarily be treated as compositional, especially if total abundances are relevant for the analysis. However, from the methodological perspective, analyzing solely raw data still cannot be recommended. Instead, compositional data should be treated as standard positive observations that induce again specific geometrical features, e.g., their relative scale (Mateu-Figueras and Pawłowsky-Glahn, 2008). Beside the (clr) coordinate representation, the usual log-transformation $\ln(\mathbf{x}) = (\ln x_1, \dots, \ln x_D)$ seems to be meaningful for these situations (Pawłowsky-Glahn et al., 2015b).

The D/C boundary geochemical dataset provides a suitable working material to test the applicability of the compositional

approach. The data comprise marine carbonate rocks, which alternate with several siliciclastic layers each confined to a specific stratigraphic interval. These layers are correlatable across the six studied areas and provide thus a common feature of all the studied sections. They comprise the Hangenberg black shale, sandstone and grey shale, collectively referred to as the Hangenberg event *sensu lato*, HBS s.l. (costatus-kockeli Interregnum, uppermost Famennian) and the Lower Alum Shale, LAS (crenulata Zone, middle Tournaisian) (Kaiser et al., 2011). Their geochemical signature reflects rapid siliciclastic influx into shelf seas (sandstones and grey shales of the HBS s.l.), increased organic productivity in pelagic settings and global development of water bottom anoxia (black shales of the HBS s.l. and LAS) and effects of dilution by biomineralized calcium carbonate (interlayered carbonates) (Kumpan et al., 2014a,b; 2015; Bábek et al., 2016). Developed more-or-less in all of the studied sections, this signature is modulated by such local factors as different source of the siliciclastics and variable depositional settings. All these factors contribute to a complex structure of the geochemical data.

The aim of the manuscript is to demonstrate, that not just statistical analysis based on the compositional approach, being definitively the most relevant from the theoretical perspective, but also analysis of raw data or their log-transformation bring some benefits to the overall analysis. The reason is that the effect of the concrete scale of data cannot be frequently simply removed as it is the case of the compositional approach (scale invariance). Practical experiences indicate that even considering absolute scale of the original (raw) data, being methodologically rather incorrect, can reveal some further interesting features. The aim of the practical part is not to provide an exhaustive comparison, but to demonstrate that just combination of different approaches (though not necessarily theoretically sound) leads to complex understanding of the multivariate data structure. Nevertheless, following the compositional approach that reflects predominant geometrical features of data at hand, one is on the “safe side”. As an adjacent goal, inspired by quasimetric structure of the variation matrix, the paper extends the use of Q-mode clustering, introduced in van den Boogaart and Tolosana-Delgado (2013) for alternative agglomerative clustering procedures. All statistical calculations were conducted with software R (R Core Team, 2016) and its package *compositions* (van den Boogaart et al., 2013; van den Boogaart et al., 2013).

2. Materials and methods

2.1. Data description

The study material includes elemental geochemistry of carbonate and siliciclastic bed successions across the Devonian/Carboniferous (D/C) boundary. The dataset includes 1884 samples from 14 sections studied in six major Variscan massifs of Europe (Rhenish Massif, Germany; Namur-Dinant Basin, Belgium and northern France; Moravo-Silesian Zone of the Bohemian Massif, Czech Republic; Carnic Alps, Austria and Italy; Montagne Noire of the Massif Central, southern France and Pyrenees, southern France). Figure representing position of the studied sections in the major outcrops of Variscan massifs in Europe is attached as a [Supplementary Material](#). The geological settings and stratigraphy of the sections were recently summarized by Kumpan et al. (2014a,b, 2015) and Bábek et al. (2016). Samples for element geochemistry were taken from the studied sections with a vertical step of 5–25 cm (rarely with 1 cm or 50–60 cm) depending on overall section thickness and required detail. The average sampling density was one sample per 11.5 cm.

All the samples were analysed by energy-dispersive X-ray

fluorescence (EDXRF) using a MiniPal 4.0 instrument (PANalytical, Netherlands) with an Rh lamp (30 kV) and Peltier cooled Si PIN detector. The samples were ground to <63 µm particle size and filled into plastic cells 25 mm in diameter and with Mylar foil bottoms. Eighteen elements were analysed (Al, Ca, Cr, Cu, Fe, K, Mn, Ni, P, Pb, Rb, S, Si, Sr, Ti, Y, Zn and Zr). Al and Si signals were acquired for 300 s at 5 kV/400 µA with a Kapton filter under He flush (99.996% purity); K, Ti, Fe, and Mn and Fe for 200 s at 12 kV/200 µA with a thin Al filter in air; and Zr for 500 s at 30 kV/200 µA with an Ag filter in air. Total analysis time was set at 800 s per sample. The EDXRF results are given in counts per second (cps). The EDXRF analytical results for Ca, Si, Al, K, Ti, Mn, Fe, Sr, and Zr (in cps) were calibrated through inductively coupled plasma mass spectrometry (ICP-MS) analysis of 11 samples (Lesní lom, Křtiny, and Grüne Schneid sections) by an accredited analytical laboratory of the Technical University of Ostrava, Czech Republic, using an X Series 2 ICP-MS instrument (Thermo Scientific). The quality of the ICP-MS analytical data was checked by measuring 2709a standard reference material (SRM) (San Joaquin Soil, NIST, USA). EDXRF results for Cr, Rb, Y, Cu, Zn, Ni, and Pb were calibrated by ICP-MS (Element2, Thermo Scientific) at the Geological Institute, Czech Academy of Sciences, Prague, using calibration equations from an external set of 17 samples from lower Devonian carbonates and shales of the Prague Basin, Czech Republic. Data quality from the Geological Institute was checked by analysing 1d SRM (Argillaceous limestone, National Institute of Standards and Technology). Calibration curves of the EDXRF vs. ICP-MS results suggest that all target elements were above EDXRF detection limits while the high correlation coefficients ($R^2 = 0.93$ to 0.99 generally; only slightly lower for Mn: $R^2 = 0.893$, Cr: $R^2 = 0.887$, and Ni: $R^2 = 0.722$) indicate the good reproducibility of the EDXRF signal. P and S were not calibrated. All element ratios and enrichment factors mentioned throughout this paper are based on uncalibrated EDXRF (cps) data.

The elemental composition sensitively reflects basic lithological and paleoenvironmental changes at the D/C boundary. Particularly effective are the following element groups with similar geochemical behavior:

- Ca, which is driven by marine CaCO_3 production in benthic and pelagic settings (Sageman and Lyons, 2005);
- Al, K, Rb and Ti, which are bound to siliciclastic minerals derived from continent, in particular clay minerals and silt-sized heavy minerals and phyllosilicates (Ross and Bustin, 2009; Sageman and Lyons, 2005; Vijver et al., 2008);
- Si and Zr, which tend to concentrate in coarse-grained siliciclastic minerals (Schnetger et al., 2000; Jones et al., 2012) also derived from continent (heavy minerals, quartz, feldspars) but Si is also linked to organic pelagic production in the sea (radiolarians and sponge spicules);
- Zn, Ni, Cu, Pb, S and P - productivity-sensitive elements, which tend to concentrate in organic matter-rich sediments such as the Hangenberg black shale and Lower Alum shale (Bout-Roumazeilles et al., 2013; Fralick and Kronberg, 1997; Sageman and Lyons, 2005; Śliwiński et al., 2010; Tribouillard et al., 2006).
- redox-sensitive elements such as Fe and Mn, which are highly mobile across sub-bottom redox gradients (Haese et al., 1998).

The advantage of the dataset is in the detailed knowledge about distinct sedimentary layers with expected geochemical behavior, which include: layer 1) Upper Devonian carbonates; layer 2) Uppermost Devonian black shales of the Hangenberg (HBE) event interval; layer 3) Uppermost Devonian sandstones and non-black shales of the Hangenberg (HBE) interval; layer 4) Lower

Carboniferous carbonates and layer 5) Lower Carboniferous black shales and cherts of the Lower Alum Shale interval. Figure representing chronostratigraphy, biostratigraphy, lithostratigraphy and thickness of the studied sections is attached as a [Supplementary Material](#).

2.2. Exploration of compositional variation structure

Covariance structure of compositional data reflects the fact that the source information is contained in pairwise log-ratios. Accordingly, in the compositional context the multivariate variability is captured by the variation matrix (Aitchison, 1986), defined as

$$\mathbf{T} = \begin{bmatrix} 0 & \text{var}\left(\ln \frac{x_1}{x_2}\right) & \dots & \text{var}\left(\ln \frac{x_1}{x_D}\right) \\ \text{var}\left(\ln \frac{x_2}{x_1}\right) & 0 & \dots & \text{var}\left(\ln \frac{x_2}{x_D}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \text{var}\left(\ln \frac{x_D}{x_1}\right) & \text{var}\left(\ln \frac{x_D}{x_2}\right) & \dots & 0 \end{bmatrix} = \left[\text{var}\left(\ln \frac{x_i}{x_j}\right) \right] = [t_{ij}]. \quad (3)$$

Its interpretation is intuitive. A non-diagonal element of the variation matrix is zero, or nearly so, if and only if the respective compositional parts are proportional, or nearly so. In other words, proportionality here replaces covariance (correlation) between variables from standard multivariate statistics. Consequently, elements of variation matrix can also be used as a measure of dissimilarity between compositional parts, for example for the purpose of clustering of compositional parts (Montero-Serrano et al., 2010; Pawłowsky-Glahn et al., 2011; van den Boogaart and Tolosana-Delgado, 2013; McKinley et al., 2016). From the basic (metric) properties of distances the following are obviously fulfilled,

$$t_{ij} \geq 0, \quad t_{ij} = 0 \Leftrightarrow i = j; \quad t_{ij} = t_{ji}, \quad (4)$$

i.e., non-negativity, identity of indiscernibles and symmetry. On the other hand, triangular inequality is not fulfilled in general, only its generalized form

$$t_{ij} \leq \alpha(t_{ik} + t_{kj}) \quad (5)$$

for a constant $\alpha \geq 1$ that corresponds to quasimetrics (Xia, 2009) can be derived. According to Fiserová and Hron (2011) it holds true that

$$t_{ij} = 2(t_{ik} + t_{kj}) - \text{var}\left(2 \ln \frac{\sqrt{x_i x_j}}{x_k}\right), \quad (6)$$

resulting in $\alpha=2$. It is shown (Xia, 2009) that many well-known results for the usual metrics still hold true in quasimetric space that makes them a natural generalization of the basic metric settings. These findings are inspirational for Q-mode clustering, introduced in the next section. Finally, a matrix relationship exists between the variation matrix and covariance matrix of clr coordinates (Aitchison, 1986), which is useful for practical computations.

Variation structure of compositional data can be visualized using compositional biplots (Aitchison and Greenacre, 2002). Similarly as for a standard PCA biplot (Gabriel, 1971), which is used to

reduce dimensionality of input data, compositional biplot displays as well scores and loadings of the first two principal components in one planar graph. The scores are usually marked as points in order to capture multivariate data structure; loadings are represented by arrows and stand for the input variables. While in the standard biplot original (log-transformed) variables are considered, in the compositional case clr coordinates are usually represented. This also affects the interpretation of loading vectors. While in the standard case the length of the arrow and the cosine of the angle between two arrows approximate standard deviation of the corresponding variable and correlation between variables, respectively, interpretation of clr variables needs to be taken into account in the compositional case. Consequently, the length of the arrow cannot be interpreted as a single original part but as a representation of its dominance to an “average part” in the composition (Filzmoser et al., 2012). Instead of interpreting correlation between two clr coordinates, which is affected by the zero sum constraint of variables, it is preferred to consider links between vertices, approximating pairwise logratio variances (elements of the variation matrix). In particular, link between vertices of y_i and y_j approximates $\text{var}\left(\ln \frac{x_i}{x_j}\right)$; if the vertices coincide, or nearly so, then x_i and x_j are proportional, or nearly so. An enhanced interpretation of compositional biplot in terms of covariance structure of pairwise log-ratios can be found, e.g., in Pawłowsky-Glahn et al. (2015a).

2.3. Pattern identification

In order to assess the log-ratio patterns in the geochemical signature of the D/C boundary rocks, it is necessary to search for such variables which are responsible for the general geochemical patterns. In particular they include the detrital-input sensitive elements (Al, Ti, Rb, Zr), grain-size sensitive ones (Si, Zr), organic productivity sensitive ones (P, S, Ca, Ni, Zn, Cu) and redox-sensitive ones (Fe, Ni, Mn) (see Section 2.1). For the case, when total abundances are informative as well, the log-transformed original variables or their combinations would be appropriate for this purpose. Nevertheless, it seems that for carbonate rocks rather relative contributions of the EDXRF signal (in cps) are of interest. This is reflected also by recent practice in the field, where ratios between elementary components such as Zr/Rb, K/Al and Rb/K are considered to indicate the lithological changes across the D/C boundary (Kumpan et al., 2015; Bábek et al., 2016). The previous section implies that, referring to relative structure of compositional data, single compositional parts are not appropriate to serve as feature variables, because they necessarily (directly or indirectly) rely on the other components in the actual composition. All relative information about single compositional parts is extracted using clr coordinates (1); on the other hand, this information might be too complex due to different and possibly antagonistic patterns of log-ratios aggregated there (Reimann et al., 2012; McKinley et al., 2016). Taking simply all possible pairwise log-ratios into account would need an exhaustive search, therefore not very practical for data with moderate or even larger number of components. Although expert knowledge can be used to select such log-ratios, a data-based approach can facilitate finding possible further interesting geochemical markers by mutual considering specific geochemical behavior of elements. Consequently, such an unsupervised approach can help to extract log-ratios (not necessarily belonging to any of previously mentioned extreme cases, pairwise log-ratios versus clr coordinates) able to recognize the geochemical signature.

For this purpose, in van den Boogaart and Tolosana-Delgado (2013) a compositional Q-mode clustering was proposed. This clustering method is based on idea to obtain easily interpretable

orthonormal coordinates with respect to the Aitchison geometry, referred to as principal balances (Pawłowsky-Glahn et al., 2011). It is essentially hierarchical clustering, where the variation matrix plays the role of a measure of association between compositional parts. The resulting graphical output, dendrogram, can be applied to define a sequential binary partition of compositional parts into groups of parts (Egozcue and Pawłowsky-Glahn, 2005); each horizontal link is used to set a new variable that expresses balance between the corresponding groups of compositional parts. By denoting those r on the left side of the link by plus sign and s parts on the right side by minus sign, the balance (orthonormal coordinate) is defined as

$$z = \sqrt{\frac{rs}{r+s}} \ln \frac{g(x^+)}{g(x^-)}, \quad (7)$$

where $g(x^+)$ and $g(x^-)$ stand for the geometric mean of parts from the first and second group, respectively; together $D-1$ coordinates are obtained by such a procedure. Their interpretation can be enhanced by suppressing the normalization constant and changing the base of logarithm. It is also worth to note that each balance aggregates all pairwise log-ratios between both groups of parts (Fišerová and Hron, 2011), the fact that can be used to search for simpler log-ratios responsible for the geochemical signature of the D/C boundary. According to McKinley et al. (2016), the resulting clusters of compositional parts will contain elements behaving proportionally throughout the dataset. Log-ratios between parts of two different clusters should thus be similar to other log-ratios of the elements of the same clusters. Therefore, as indicated above, one of these log-ratios or a balance of one cluster against the other might be representative for many log-ratios, and consequently may represent a process influencing many elements in the same way. Balances of elements within the cluster will filter out these large-variability effects and focus on differences between elements behaving similarly with respect to major processes.

As a default agglomerative clustering procedure the Ward method (Ward, 1963) is used in Pawłowsky-Glahn et al. (2011) and van den Boogaart and Tolosana-Delgado (2013), where those two clusters are fused which result in the least increase in the sum of the (squared) distances from each observation to the centroid of cluster contained it. Consequently, this method leads to spherical, tightly bound clusters that might in general be intuitively interpretable. Moreover, because the Ward criterion corresponds to minimizing the total within-cluster variance, it seems to be recommendable for clustering of variables (compositional parts). On the other hand, as the variation matrix has all properties of quasi-metrics, it is meaningful to consider also other clustering methods that might lead even to better interpretable clusters. For example, smaller clusters might be geologically easily interpreted. Large numbers of small, tightly bound clusters are obtained using complete linkage, therefore being a candidate for such an alternative agglomerative clustering algorithm. By considering all these aspects, the dendrogram output might help to reveal such coordinates (balances) that contain log-ratios responsible for the geochemical signature of the D/C boundary, represented by presence of HBE black shale and sandstone layer.

3. Results and discussion

Despite of strong limitations of univariate analysis of raw compositional data (Filzmoser et al., 2009b; McKinley et al., 2016), summaries of elemental concentrations are popular starting point of any geochemical study. According to Pawłowsky-Glahn and Egozcue (2002) and Mateu-Figueras and Pawłowsky-Glahn

(2008), geometric mean is used to compute mean concentrations; for a raw impression about variability of elements, interquartile ranges (IQR) are applied.

The mean concentrations of the elements in the composition (in the order of its geometric mean) are as follow: Ca (geometric mean = 15.55%; IQR = 21.57%), Si (5.05%; 16.33%), Al (0.75%; 2.69%), Fe (0.70%; 1.80%), K (0.29%; 0.77%), Ti (734.04 ppm; 2110.46 ppm), Mn (557.11 ppm; 897.58 ppm), Sr (208.40 ppm; 181.30 ppm), Zr (39.19 ppm; 84.14 ppm), Cr (28.39 ppm; 65.26 ppm), Rb (25.76 ppm; 81.29 ppm), Ni (22.19 ppm; 33.53 ppm), Y (21.02 ppm; 18.99 ppm), Zn (15.50 ppm; 27.16 ppm), Cu (11.65 ppm; 16.21 ppm), and Pb (8.58 ppm; 10.21 ppm). Taking the relative scale into account, the element concentrations show a strong variance across the studied areas and sections. This largely reflects the principal lithology where pure carbonate, shale, sandstone, and siliceous sediments represent the ideal end members.

The Upper Devonian carbonate-dominated successions (layer 1) are generally characterized by high concentrations of Ca (geometric mean = 22.60%) and low concentrations of Al, Si, K, Ti, Fe, Rb, and other elements. The highest concentrations of Ca were detected in the pelagic successions of the Rhenish Massif, Carnic Alps and Montagne Noire. In contrast, coeval strata of the Namur–Dinant Basin (Gendron–Celles, Les Ardennes sections) have relatively lower Ca concentrations.

In the following HBE black shale and sandstone interval (layers 2 and 3), the element composition at the majority of sections changes rapidly, consistently with the lithology change. Compared to the underlying strata, the mean concentrations of Ca in the HBE shales and sandstones are extremely low (geometric mean 3.58%) while the mean concentrations of typically terrigenous elements are much higher (Al = 4.62%, K = 1.31%, Fe = 3.16%, Ti = 4500 ppm, Rb = 190 ppm, S = 15 cps). In addition, the HBE black shales (layer 2) have very high mean concentrations of Zr (221 ppm), S (59 cps), Zn (51 ppm), Ni (114 ppm) and Pb (95 ppm).

The lower Tournaisian carbonate successions (layer 4) have relatively high concentrations of Ca (25.13%) and relatively low concentrations of elements such as Al (0.39%), K (0.16%), Fe (0.43%), Ti (470 ppm) and Rb (15 ppm), which is similar to those of the upper Famennian succession (layer 1).

The youngest strata of the LAS interval (layer 5) were reached only at several sections of the Namur–Dinant Basin (Rivage), Rhenish Massif (Oese, Drewer), Montagne Noire (Puech de la Suque), Pyrenees (Saubette) and Carnic Alps (Kronhofgraben). These sediments are shaly, often siliceous and phosphatic. They have high mean concentrations of Al (2.69%), Si (25.05%), K (0.65%), Fe (2.46%), Ti (0.22%), Rb (104 ppm) and low concentrations of Ca (3.69%). Compared to the all underlying layers, they are markedly enriched in elements such as P (11 cps), Cu (47 ppm), Zn (44 ppm) and Ni (85 ppm).

In order to reveal both overall and specific geochemical patterns, standard biplots of raw or log-transformed data and compositional biplots were used at each stage of analysis and this section presents those options which led to the most interesting results. Fig. 1 shows a standard biplot of raw data (left) and its compositional counterpart (right). In both plots, a cluster of terrigenous elements (Al, Ti, Rb, K, Fe) stands in opposition to Ca along the first principal component, which is interpreted as the effect of dilution of terrigenous input by marine CaCO₃ production. The first principal component (PC) in the biplot depicting raw data explains 56.9% of total variability indicating that the effect of CaCO₃ dilution is the dominant geochemical pattern in the whole dataset. The second PC in the standard biplot shows the contents of such productivity- and redox-sensitive elements as S, P, and Mn, while other productivity-sensitive elements (Zn, Cu) largely overlap with the terrigenous ones. The effect of Ca dilution seems to override the subtle

variations in element behaviors in the standard biplot. On the other hand, the compositional biplot better differentiates between samples from different locations (different provenance) particularly along the second principal component, which represents the effects of provenance and marine productivity. In addition, the typically detrital proxy elements (Al, Ti, Rb, K and Si) are more clearly separated from productivity-sensitive elements (Zn, Pb and Cr) in the compositional biplot than in the standard one. At the same time, the effect of Ca dilution is partly removed in the compositional biplot as indicated by the fact the Ca and terrigenous elements are no more in opposition (Fig. 1).

The biplots showing the whole data set bring important information about the general data structure. However, the clustered distribution of samples from different geographical settings and, hence, different provenance (Fig. 1) suggests that it is appropriate to analyze such data clusters separately.

Log-transformation of the raw geochemical data reflects the relative scale of observations; consequently, it should help to reveal further patterns, possibly hidden behind the raw data. The Fig. 2 shows the application of standard biplot on log-transformed data with the distinct stratigraphic layers, from Upper Devonian carbonates to Lower Alum Shale, indicated. As it will be shown later, such geochemical markers as element ratios are instrumental in distinguishing these layers. The Fig. 2, left, shows a standard biplot of log-transformed data from the Oberrödinghausen section, Rhenish Massif, Germany with four of the five layers (Uppermost Devonian/Famennian/carbonates, the HBE black shale, HBE sandstone, and Lower Carboniferous/Tournaisian/carbonates) indicated. The layers group together as distinct clusters; there is a general pattern of clustering of samples from the HBE shales/sandstones and carbonates along the PC1 axis, which is interpreted as the dilution effect of calcium carbonate (see above). In addition, samples representing the HBE black shale event occur as distinct outlier observations due to the high concentrations of Pb, Zn, Ni and Cu, typical productivity-sensitive elements. The overlaps between the HBE black shales and HBE shales/sandstones in the area with negative PC1 scores and positive PC2 scores suggests that the black shales and “normal” grey shales can be represented by very similar geochemistry.

Similar patterns are shown for the Oese section, Rhenish Massif (Fig. 2, right). The carbonates are again well separated from the shales and sandstones and there are distinct outliers representing the HBE black shales and Lower Alum Shales (negative scores on PC1 and PC2) suggesting high concentrations of productivity-sensitive elements such as Pb, Zn, Cu, Ni and Y. This suggests that the black shales of the HBE and LAS interval share a similar geochemical composition. Moreover, both the biplots show a distinct clustering of variables (elements), in particular the terrigenous Al, Rb, Ti, K and Si, which is consistent with the summary biplot of all observations (Fig. 1) and their expected geochemical behavior (see Section 2.1).

An even better representation of sample clustering is shown in Fig. 3 providing the D/C boundary data from the Kronhofgraben section, Carnic Alps, Austria. This biplot of log-transformed data clearly separates between the cherts and black cherty shales of the Lower Alum Shale interval, enriched in Cu, Ni, Pb and Si (negative scores on PC1 and positive scores on PC2) and black shales of the HBE black shale interval (negative scores on both PC1 and PC2). This is consistent with the dissimilar lithology patterns of the HBE and LAS layers at the Kronhofgraben section, the former represented by typical black shales and the latter by black cherts and cherty shales (Schönlaub et al., 1992).

The applicability of raw data, which are somewhat inappropriate from purely methodological perspective, and clr coordinates can be demonstrated on samples from the Rhenish Massif (Oese,

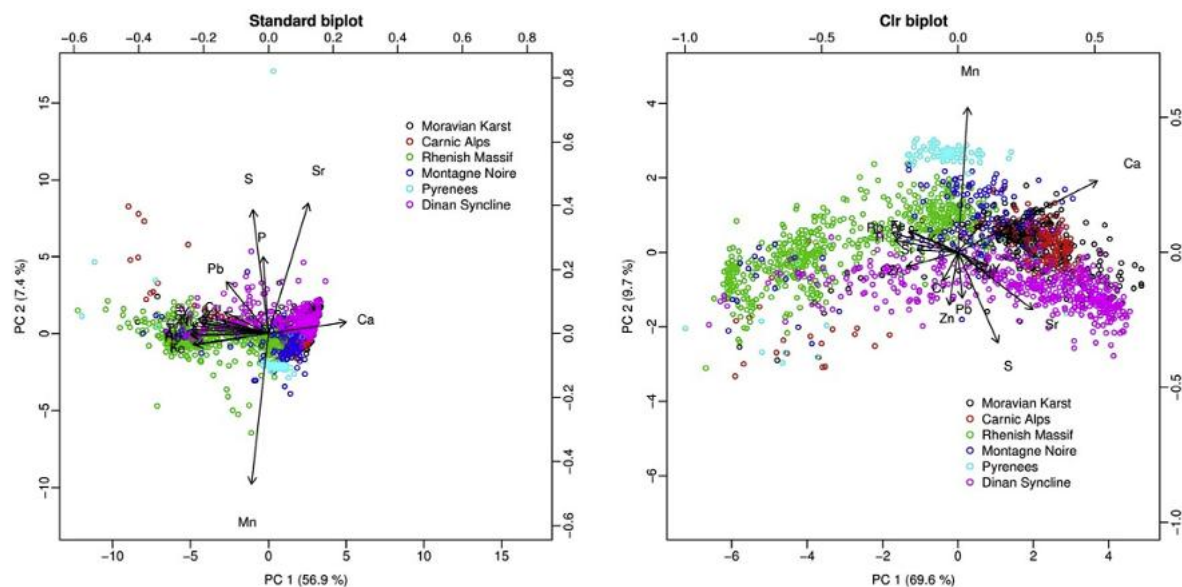


Fig. 1. Standard biplot of raw data (left) and clr biplot (right), with samples distributed according to its location.

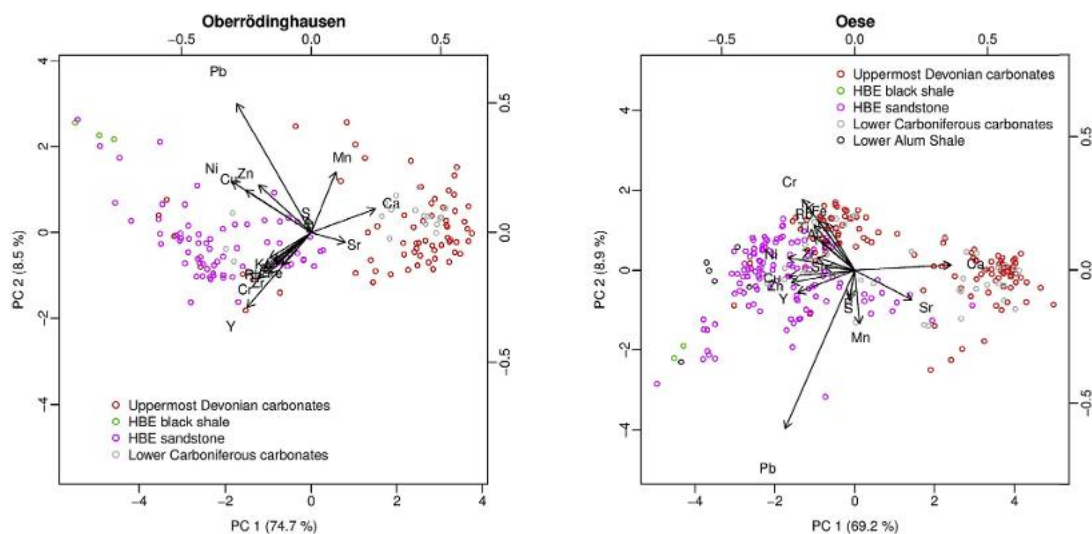


Fig. 2. Standard biplots of log-transformed data from locations Oberrödinghausen (left) and Oese (right), Rhenish Massif, Germany.

Oberrödinghausen and Drewer sections) (Fig. 4). The raw-data biplot (Fig. 4 left) clearly separates the black shales enriched in Ni, Zn, Pb and Cu from the remainder of samples. The latter samples mostly tend to align to a carbonate (Ca) – fine-grained siliciclastic (Al, Rb, K, Ti) component line, which again reflects the degree of carbonate dilution of the siliciclastic detrital input. In the compositional biplot, this pattern is generally lost, while the differences between individual localities are highlighted.

Loadings in the raw-data biplot again show a distinct clustering of elements, depending largely on the expected geochemical behavior. In the Rhenish Massif, this clustering is very well visible in areas where the particular genetic element groups (see above) are

represented by typical lithologies (black shales) with relatively high concentrations of their nominal elements (Zn, Pb, Ni and Cu, Fig. 4, left). There is again the overwhelming negative correlation between Ca on one hand and Al, Rb, K and Ti on the other hand. In the compositional biplot, however, the effects of a strong association with Ca (dilution effect) are strongly suppressed and elements show clustering into several, genetically related groups: Ca and Sr associated with marine calcium carbonate; terrigenous elements associated with detrital rock-forming minerals, mainly phyllosilicates (Al, Si, Rb, Ti, Rb), redox-sensitive elements associated with sulphidic phases (Cu, Ni, S, P) and other redox-sensitive elements (Mn, Fe).

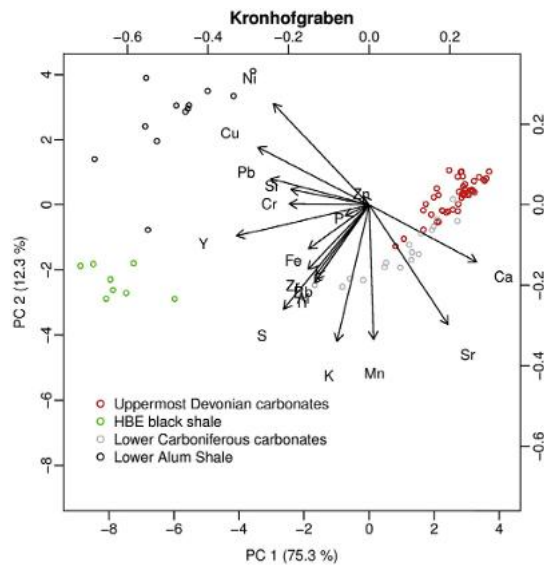


Fig. 3. Biplot of log-transformed data from location Kronhofgraben, Carnic Alps, Austria.

We also tested these patterns in the Ardennes (Gendron Celles section), where the HBE s.l. layer is not represented by the distinct shale/sandstone lithology and the section is carbonate-dominated (Fig. 5). In the raw-data biplot (Fig. 5 left) the effect of carbonate dilution is very strong as indicated by the alignment of samples along the PC1 axis as well as the negative correlation between Ca and the remainder of elements (with the exception of Sr, Mn and Zn). In the compositional biplot, the effect of Ca dilution is again suppressed but the variables (elements) tend to group together according to their expected geochemical behavior (Al + Ti + Rb vs. Si + Zr vs. Pb + Zn + Cr + Mn) despite their very low concentrations. As indicated in the previous sections, this seems to be the effect of relative scale of compositions, captured by their clr coordinates.

Another source of information about the data structure is represented by dendrograms resulting from Q-mode clustering, whose construction was described in Section 2.3. Instead of the originally proposed Ward clustering method, the complete linkage shown here and based on the variation matrix provides more reliable results. Groups of elements identified by this method can facilitate finding log-ratios, which optimally describe the presence of the Hangenberg event layers and thus provide a compositional alternative to the standard proxies as Zr/Al, K/Al and Rb/K ratios (Kumpan et al., 2014b, 2015).

Fig. 6 shows a dendrogram for the Kronhofgraben section as the output of the Q-mode clustering. According to this method, several balance coordinates were defined; three of them, which lead to best separation of groups, are shown in Fig. 7. The last one (Fig. 7, right), which excludes Ca (based on the assumption that Ca is strongly related to the dilution effect), was selected, to see whether it can affect the values of the resulting balance. For comparison, Fig. 8 shows values of standard Hangenberg event proxies Zr/Al, K/Al and Rb/K. The proposed log-ratios are useful to differentiate individual groups of observations; e.g., the low values of $\ln(g(\text{Ca}, \text{Sr})/g(\text{rest}))$, where $g(\text{rest})$ stands for geometric mean of all components except Ca and Sr, and log-ratio between Ca and Sr mark the presence of HBE black shale and the LAS layer. Element ratios such as Zr/Al and K/Al are capable of distinguishing between the HBE black shales and LAS black cherts/cherty shales, the former having much higher Zr/Al and lower K/Al values as compared to the latter (Fig. 8). Nevertheless, when comparing Figs. 7 and 8 we can see that the compositional variables (balances) better distinguish four groups of observations; the changes are really dramatic and provide a clear structure. Note that similar patterns could be also seen for some other balances, indicated by the dendrogram of Q-mode clustering (Fig. 6), but the presented ones seem to provide the best performance.

Q-mode clustering can also help to find suitable proxies only for subsets of observations, for example for similar lithologies (shales and sandstones). Fig. 9 shows the resulting dendrogram for Oese section, in which only the HBE black shale and sandstone and Lower Alum Shales were taken into account.

According to this clustering, combined with geochemical knowledge, we selected first the log-ratio between sub-compositions formed by parts Ti, Al, K and Cr, Rb, respectively

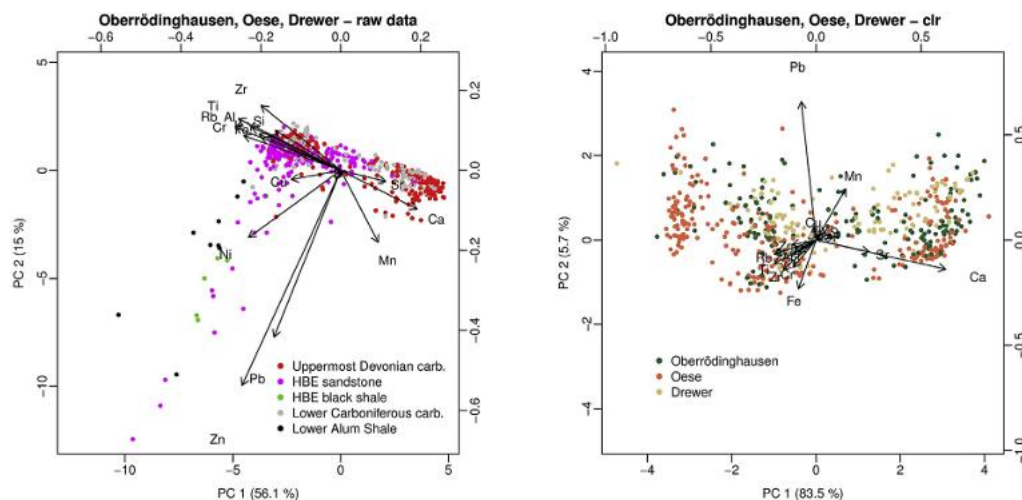


Fig. 4. Biplot of raw data (left) and compositional biplot (right) for locations from Rhenish Massif (Oberrödinghausen, Oese, Drewer sections).

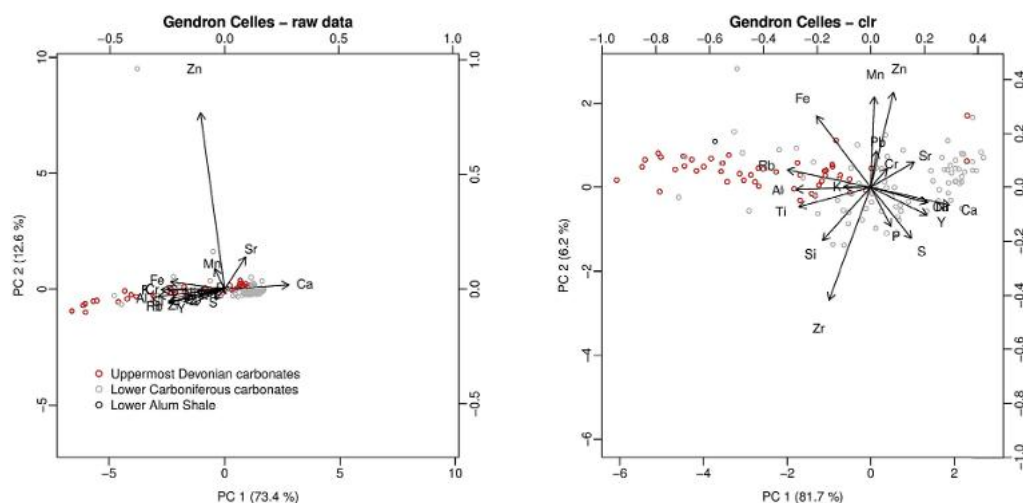


Fig. 5. Biplot of raw data (left) and compositional biplot (right) for location Gendron Celles, Ardennes, Belgium.

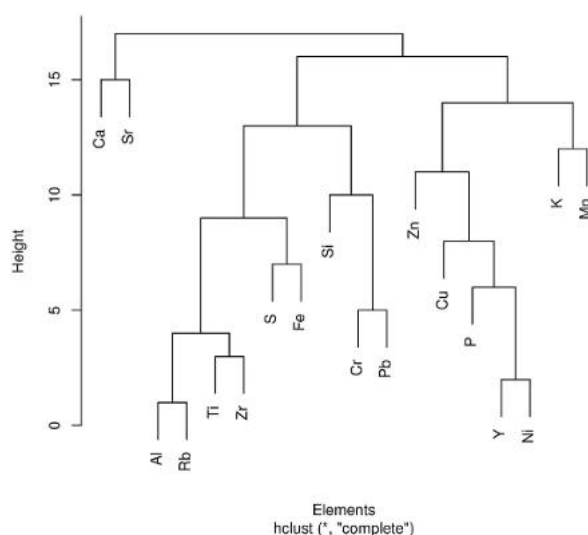


Fig. 6. Dendrogram of Q-mode clustering for location Kronhofgraben.

(Fig. 10, upper left) and one additional balance that links the previous elements with Cu and Zn from the same branch of the dendrogram (Fig. 10, upper right). Both balances clearly separate the Lower Alum Shale layer from the HBE black shales. This represents the advantage of log-ratio approach, compared to standard proxies, whose values are displayed on Fig. 10 (lower row) and which do not distinguish between layers at all. This is a direct consequence of the fact that Q-mode clustering supplements the preliminary geochemical knowledge with further possible candidates (balances) to reveal better the D/C boundary.

In general, statistical analysis of observations including elemental geochemistry of carbonate and siliciclastic bed successions across the D/C boundary using both compositional and non-compositional approaches has revealed several interesting features. Dimension reduction of multivariate geochemical data

through principal component analysis is nowadays a must for any reasonable case study. In stratigraphy, statistical pre-treatment of geochemical data is also a common approach prior to depicting of stratigraphic patterns of element concentrations (Sedláček et al., 2013; Bábek et al., 2015). In the present paper, several approaches how to process the input observations prior to PCA were presented. The work flow proceeds from the raw compositions (which are often inappropriate due to the relative nature of compositional data) to either log-transformation or clr coordinates that consider, or not, total abundances of elements. Interestingly, raw-data biplots were found useful to depict the basic geochemical patterns such as the dilution effect of Ca and enrichment of black shales with productivity-sensitive elements (Cu, Zn, Ni, etc.). The possible reason is that these effects are predominantly driven just by absolute concentrations (or cps signal) of components rather than being inherently contained in their ratios. On the other hand, compositional biplots are capable of filtering out these predominant geochemical trends (such as the Ca dilution effect) and depicting subtle geochemical variations, which are obscured by the major trends.

Following the previous argumentation, depicting of compositional log-ratios in vertical/horizontal logs, which is a common approach in stratigraphy, can be more appropriate than depicting simple element ratios, because subtle geochemical trends can be obscured by the predominant trends such as the Ca dilution. This applies, for example, for nodular limestones (such as the Upper Devonian carbonates of the Rhenish Massif), in which diagenetic Ca redistribution strongly affects the primary geochemical signal. Nevertheless, analysis in the previous section has shown that even the dilution effect of Ca is desirably suppressed, when sufficiently robust log-ratios are considered (Fig. 7). Moreover, with Q-mode clustering it is possible to cover both dimension reduction using PCA, performed in clr coordinates, and D/C layer discrimination under one methodological framework.

4. Conclusions

Our recent experiments clearly indicate that the combination of raw/log-transformed data and compositional analyses are capable of distinguishing D/C boundary sediment layers with

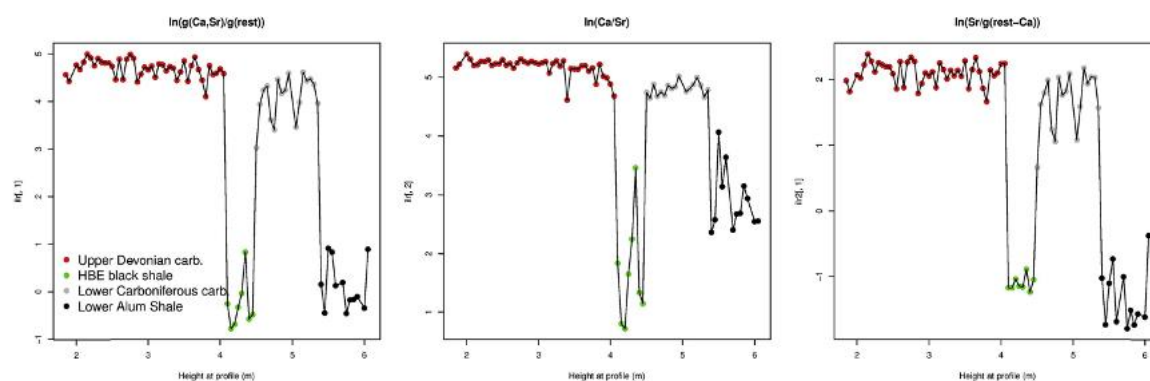


Fig. 7. Proposed log-ratios, based on dendrogram of Q-mode clustering from location Kronhofgraben, which optimally discriminate layers.

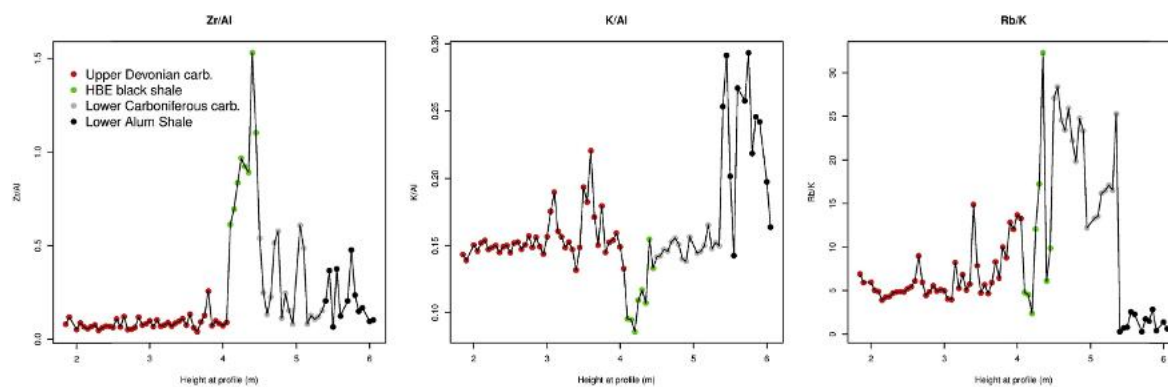


Fig. 8. Standard ratios used for discrimination between layers applied on location Kronhofgraben.

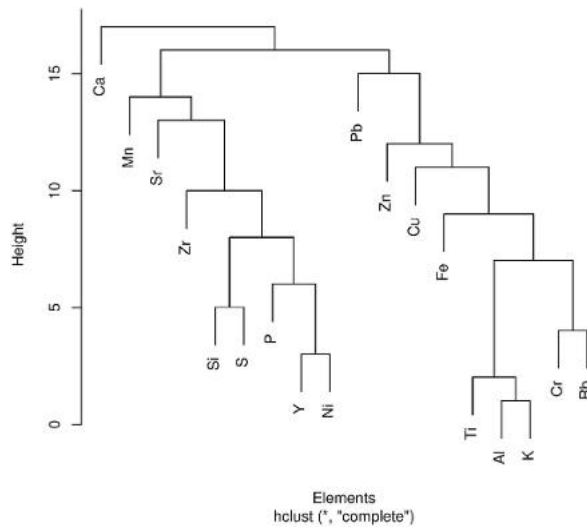


Fig. 9. Dendrogram of Q-mode clustering for location Oese and layers Hangenberg black shale, Hangenberg sandstone and Lower Alum Shale.

specific geochemical signature, in particular the HBE black shales and LAS shales. This is in line with recent contributions from geochemistry (Montero-Serrano et al., 2010; Reimann et al., 2012; Bábek et al., 2015; McKinley et al., 2016) showing that not just the standard processing of geochemical data, represented mostly by log-transformation, but even the compositional data analysis has some limitations that favour the use of the complementary approach. It is unexceptionable that only the logratio methodology would be acceptable, if exclusively relative information out of geochemical data were informative, but it rather seems not to be the case in most practical situations. Consequently, neither the standard nor the compositional approaches have prevalence, but it is advisable to use both of them to discern predominant and subtle geochemical trends in large datasets. Using the fact that the standard and compositional approaches have synergic effects is thus recommendable for future developments, in sedimentology, and also in geochemistry in general. Nevertheless, one must be aware of interpretational dangers by using the standard approaches (log-transformed or even raw data) that do not occur with the logratio methodology; a deep understanding of the underlying geological phenomena (being the case here) is a necessary presumption of their possible use in geochemical practice.

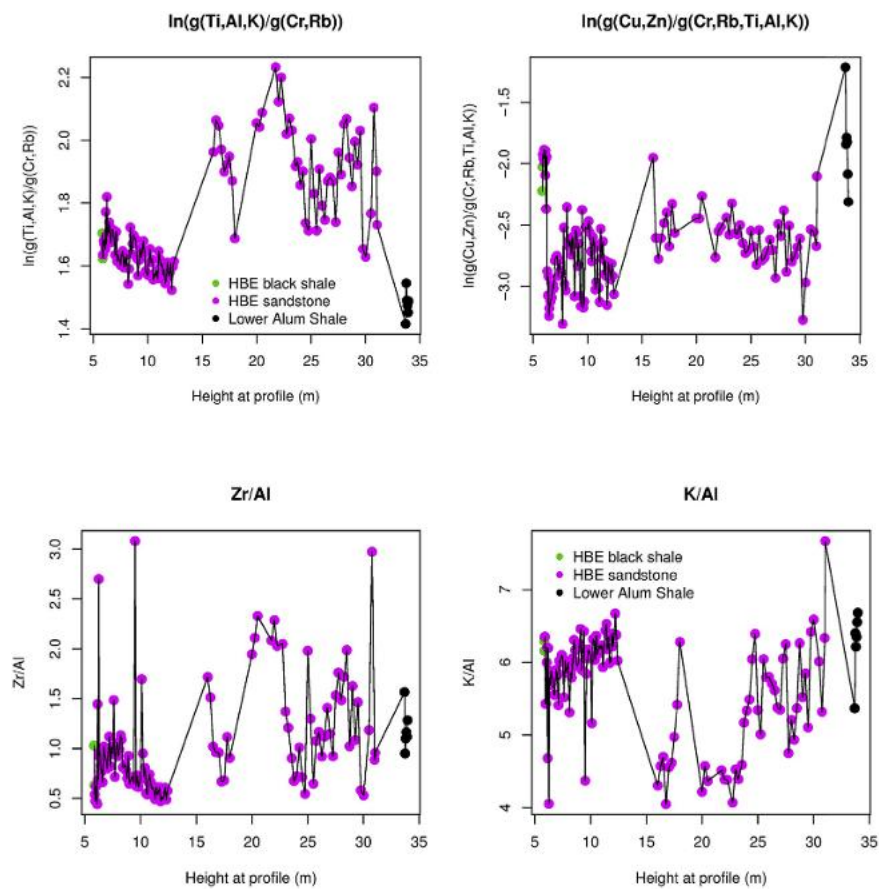


Fig. 10. Proposed log-ratios, based on Q mode clustering from location Oese, which optimally discriminate layers (upper row) and standard ratios used for discrimination between layers (lower row).

Acknowledgements

Karel Hron gratefully acknowledges the support of the grant COST Action CRoNoS IC1408 and the grant IGA_PrF_2016_025 Mathematical Models of the Internal Grant Agency of the Palacký University in Olomouc. This work was partly supported by the Czech Science Foundation research project 14-18183S (O. Bábek). We thank to Tomáš Matys Grygar (Institute of Inorganic Chemistry ASCR, v.v.i.) for providing of EDXRF data.

Appendix A. Supplementary data

Supplementary data related to this article can be found at <http://dx.doi.org/10.1016/j.apgeochem.2016.10.002>.

References

- Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK) (Reprinted in 2003 with additional material by The Blackburn Press).
- Aitchison, J., Greenacre, M., 2002. Biplots of compositional data. *J. R. Stat. Soc. Ser. C Appl. Stat.* 51 (4), 375–392.
- Bábek, O., Grygar, T.M., Faměra, M., Hron, K., Nováková, T., Sedláček, J., 2015. Geochemical background in polluted river sediments: how to separate the effects of sediment provenance and grain size with statistical rigour? *Catena* 135, 240–253.
- Bábek, O., Kumpan, T., Kalvoda, J., Matys, Grygar, T., 15 May 2016. Devonian/Carboniferous boundary glacioeustatic fluctuations in a platform-to-basin direction: a geochemical approach of sequence stratigraphy in pelagic settings. *Sediment. Geol.* 337, 81–99.
- Billheimer, D., Guttorp, P., Fagan, W., 2001. Statistical interpretation of species composition. *J. Am. Stat. Assoc.* 96 (456), 1205–1214.
- Bout-Roumazailles, V., Combourieu-Nebout, N., Desprat, S., Siani, G., Turon, J.-L., Essallami, L., 2013. Tracking atmospheric and riverine terrigenous supplies variability during the last glacial and the Holocene in central Mediterranean. *Clim. Past* 9, 1065–1087.
- Eaton, M.L., 1983. Multivariate Statistics. A Vector Space Approach. John Wiley & Sons, New York (US).
- Egozcue, J.J., Pawłowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. *Math. Geol.* 35, 279–300.
- Egozcue, J.J., Pawłowsky-Glahn, V., 2005. Groups of parts and their balances in compositional data analysis. *Math. Geol.* 37 (7), 795–828.
- Egozcue, J.J., 2009. Reply to “On the Harker Variation Diagrams” by J.A. Cortés. *Math. Geosci.* 41, 829–834.
- Filzmoser, P., Hron, K., Reimann, C., 2009a. Principal component analysis for compositional data with outliers. *Environmetrics* 20, 621–632.
- Filzmoser, P., Hron, K., Reimann, C., 2009b. Univariate statistical analysis of environmental (compositional) data: problems and possibilities. *Sci. Total Environ.* 407, 6100–6108.
- Filzmoser, P., Hron, K., Reimann, C., 2012. Interpretation of multivariate outliers for compositional data. *Comput. Geosci.* 39, 77–85.
- Fišerová, E., Hron, K., 2011. On interpretation of orthonormal coordinates for compositional data. *Math. Geosci.* 43, 455–468.
- Fralick, P.W., Kronberg, B.J., 1997. Geochemical discrimination of clastic sedimentary rock sources. *Sediment. Geol.* 113, 111–124.
- Gabriel, K.R., 1971. The biplot graphical display of matrices with application to principal component analysis. *Biometrika* 58, 453–467.
- Haese, R.R., Petermann, H., Dittert, L., Schulz, H.D., 1998. The early diagenesis of iron in pelagic sediments: a multidisciplinary approach. *Earth Planet. Sci. Lett.* 157, 233–248.
- Jones, A.F., Macklin, M.G., Brewer, P.A., 2012. A geochemical record of flooding on the upper River Severn, UK, during the last 3750 years. *Geomorphology* 179, 89–105.
- Kaiser, S.L., Becker, R.T., Steuber, T., Aboussalam, S.Z., 2011. Climate-controlled mass extinctions, facies, and sea-level changes around the Devonian–Carboniferous boundary in the eastern Anti-Atlas (SE Morocco). *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 310, 340–364.
- Kumpan, T., Bábek, O., Kalvoda, J., Frýda, J., Matys Grygar, T., 2014a. A high-resolution, multiproxy stratigraphic analysis of the Devonian–Carboniferous boundary sections in the Moravian Karst (Czech Republic) and a correlation with the Carnic Alps (Austria). *Geol. Mag.* 151, 201–215.
- Kumpan, T., Bábek, O., Kalvoda, J., Matys Grygar, T., Frýda, J., 2014b. Sea-level and environmental changes around the Devonian–Carboniferous boundary in the Namur–Dinant Basin (S Belgium, NE France): a multi-proxy stratigraphic analysis of carbonate ramp archives and its use in regional and interregional correlations. *Sediment. Geol.* 311, 43–59.
- Kumpan, T., Bábek, O., Kalvoda, J., Matys Grygar, T., Frýda, J., Becker, R.T., Hartenfels, S., 2015. Petrophysical and geochemical signature of the Hangenberg events: an integrated stratigraphy of the Devonian–Carboniferous boundary interval in the Northern Rhenish Massif (Avalonia, Germany). *Bull. Geosci.* 90, 667–694.
- Mateu-Figueras, G., Pawłowsky-Glahn, V., 2008. A critical approach to probability laws in geochemistry. *Math. Geosci.* 40 (5), 489–502.
- McKinley, J.M., Hron, K., Grunsky, E., Reimann, C., de Caritat, P., Filzmoser, P., van den Boogaart, K.G., Tolosana-Delgado, R., 2016. The single component geochemical map: fact or fiction. *J. Geochem. Explor.* 162, 16–28.
- Montero-Serrano, J.C., Palarea-Albaladejo, J., Martín-Fernández, J.A., Martínez-Santana, M., Gutiérrez-Martín, J.V., 2010. Sedimentary chemofacies characterization by means of multivariate analysis. *Sediment. Geol.* 228 (3–4), 218–228.
- Pawłowsky-Glahn, V., Egozcue, J.J., 2001. Geometric approach to statistical analysis on the simplex. *Stoch. Environ. Res. Risk Assess. (SERRA)* 15 (5), 384–398.
- Pawłowsky-Glahn, V., Egozcue, J.J., 2002. BLU estimators and compositional data. *Math. Geol.* 34 (3), 259–274.
- Pawłowsky-Glahn, V., Buccianti, A., 2011. Compositional Data Analysis, Theory and Applications. Wiley, Chichester (UK), 378p.
- Pawłowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2011. Principal balances. In: Egozcue, J.J., Tolosana-Delgado, R., Ortego, M. (Eds.), Proceedings of the 4th International Compositional Analysis Workshop. Sant Feliu de Guixols, Girona, Spain.
- Pawłowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2015a. Modeling and Analysis of Compositional Data. Wiley, Chichester (UK).
- Pawłowsky-Glahn, V., Egozcue, J.J., Lovell, D., 2015b. Tools for compositional data with a total. *Stat. Model.* 15 (2), 175–190.
- R Core Team, 2016. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.
- Reimann, C., Filzmoser, P., Fabian, K., Hron, K., Birke, M., Demetriades, A., Dinelli, E., Ladenberger, A., The GEMAS Project Team, 2012. The concept of compositional data analysis in practice – total major element concentrations in agricultural and grazing land soils in Europe. *Sci. Total Environ.* 426, 196–210.
- Ross, D.J.K., Bustin, R.M., 2009. Investigating the use of sedimentary geochemical proxies for paleoenvironment interpretation of thermally mature organic-rich strata: examples from the Devonian–Mississippian shales, Western Canadian Sedimentary Basin. *Chem. Geol.* 260, 1–19.
- Sageman, B.B., Lyons, T.W., 2005. Geochemistry of fine-grained sediments and sedimentary rocks. In: Mackenzie, F.T. (Ed.), Sediments, Diagenesis, and Sedimentary Rocks. Elsevier, Amsterdam, pp. 115–158.
- Sedláček, J., Bábek, O., Matys Grygar, T., 2013. Trends and evolution of contamination in a well-dated water reservoir sedimentary archive: the Brno Dam, Moravia, Czech Republic. *Environ. Earth Sci.* 69, 2581–2593.
- Schnetger, B., Brumsack, H.-J., Schale, H., Hinrichs, J., Dittert, L., 2000. Geochemical characteristics of deep-sea sediments from the Arabian Sea: a high-resolution study. *Deep-Sea Res. Part II* 47, 2735–2768.
- Schönlaub, H.P., Attrep, M., Boeckelmann, K., Dreesen, R., Feist, R., Fenninger, A., Hahn, G., Klein, P., Korn, D., Kratz, R., Magaritz, M., Orth, C.J., Schramm, J.-M., 1992. The Devonian/Carboniferous boundary in the Carnic Alps (Austria) – a multidisciplinary approach. *Abh. Geol. Bundesanst* 135, 57–98.
- Śliwiński, M.G., Whalen, M.T., Day, J., 2010. Trace element variations in the Middle Frasnian punctata zone (Late Devonian) in the Western Canada sedimentary basin – changes in oceanic bioproductivity and paleoredox spurred by a pulse of terrestrial afforestation? *Geol. Belg.* 13, 459–482.
- Tribouillard, N., Algeo, T.J., Lyons, T., Riboulleau, A., 2006. Trace metals as paleoredox and paleoproductivity proxies: an update. *Chem. Geol.* 232, 12–32.
- van den Boogaart, K.G., Tolosana-Delgado, R., 2013. Analyzing Compositional Data with R. Springer, Heidelberg (Germany), 258p.
- van den Boogaart, K. G., Tolosana-Delgado, R. and Bren, M., 2013. compositions: Compositional Data Analysis. R package version 1.40-1. URL <http://www.stat.boogaart.de/compositions>.
- Vijver, M.G., Spijker, J., Vink, J.P.M., Posthuma, L., 2008. Determining metal origins and availability in fluvial deposits by analysis of geochemical baselines and solid-solution partitioning measurements and modelling. *Environ. Pollut.* 156, 832–839.
- Ward Jr., J.H., 1963. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 58, 236–244.
- Xia, Q., 2009. The geodesic problem in quasimetric spaces. *J. Geom. Anal.* 19 (2), 452–479.

Application to eye-tracking

Fačevicová, K., Vymazal, J., Popelka, S. Advancing eye movement analysis through compositional modeling: A new perspective on Yarbus' classic study. *Behavior Research Methods*, 58(6), 173 (2026).

<https://doi.org/10.3758/s13428-026-03054-5>

[IF 5 D1; AIS 2.371 D1]

[1 citation on WoS; 0 citations without autocitations]

[Author's contribution: 70%]

Abstract

Eye-tracking metrics based on Areas of Interest (AOIs) often represent the relative allocation of visual attention across stimulus regions. Compositional data analysis (CoDA) provides a mathematically principled framework for analyzing such data and enables the application of a wide range of multivariate statistical methods through their representation in log-ratio coordinates. This study demonstrates the utility of CoDA in AOI-based eye-tracking research using a large-scale replication of Yarbus' classic "Unexpected Visitor" experiment. Eye movements of 144 participants were recorded with a high-precision eye-tracker while they viewed Ilja Repin's painting under seven tasks adapted from Yarbus. Total fixation durations within seven AOIs were analyzed either as absolute measures (classical approach) or as compositions representing the relative distribution of viewing time across AOIs. Descriptive CoDA techniques (compositional means, variation matrices, and ternary plots) together with multivariate methods in log-ratio coordinates (principal component analysis, hierarchical clustering, and compositional MANOVA) reproduce the qualitative patterns described by Yarbus and subsequent replications, confirming that task demands strongly shape the relative allocation

of attention. Linear discriminant analysis further shows that the task being performed can be inferred from eye-movement patterns with accuracy above the chance level. The paper is conceived as a tutorial introduction to CoDA in eye-tracking research. The compositional framework is particularly appropriate when AOI metrics represent proportions or when total viewing time is fixed by design, while under unconstrained viewing time, it provides a complementary perspective to classical analyses.



Advancing eye movement analysis through compositional modeling: A new perspective on Yarbus' classic study

Kamila Fačevicová¹ · Jaroslav Vymazal¹ · Stanislav Popelka²

Received: 28 November 2025 / Accepted: 28 April 2026
© The Author(s) 2026

Abstract

Eye-tracking metrics based on Areas of Interest (AOIs) often represent the relative allocation of visual attention across stimulus regions. Compositional data analysis (CoDA) provides a mathematically principled framework for analyzing such data and enables the application of a wide range of multivariate statistical methods through their representation in log-ratio coordinates. This study demonstrates the utility of CoDA in AOI-based eye-tracking research using a large-scale replication of Yarbus' classic "Unexpected Visitor" experiment. Eye movements of 144 participants were recorded with a high-precision eye-tracker while they viewed Ilja Repin's painting under seven tasks adapted from Yarbus. Total fixation durations within seven AOIs were analyzed either as absolute measures (classical approach) or as compositions representing the relative distribution of viewing time across AOIs. Descriptive CoDA techniques (compositional means, variation matrices, and ternary plots) together with multivariate methods in log-ratio coordinates (principal component analysis, hierarchical clustering, and compositional MANOVA) reproduce the qualitative patterns described by Yarbus and subsequent replications, confirming that task demands strongly shape the relative allocation of attention. Linear discriminant analysis further shows that the task being performed can be inferred from eye-movement patterns with accuracy above the chance level. The paper is conceived as a tutorial introduction to CoDA in eye-tracking research. The compositional framework is particularly appropriate when AOI metrics represent proportions or when total viewing time is fixed by design, while under unconstrained viewing time, it provides a complementary perspective to classical analyses.

Keywords Areas of interest · Compositional data · Eye-tracking · Multivariate analysis

Introduction

Eye movements constitute a fundamental mechanism of visual perception and information acquisition. Rather than moving continuously, the eyes alternate between two distinct types of movement. Fixations represent relatively stable periods during which the eyes remain focused on a specific

location, allowing for detailed visual information to be processed. Saccades, by contrast, are rapid ballistic movements that shift the point of gaze from one location to another, thereby enabling the sequential sampling of visual information across a scene (Holmqvist et al., 2011).

A seminal demonstration of how eye movements reflect cognitive processes was provided by Yarbus (1967). In the experiment with Ilja Repin's painting *They Did Not Expect Him*, also known by the name *Unexpected Visitor*, he showed that gaze patterns change markedly depending on the observer's task. When participants viewed the same painting under different instructions, their fixation sequences varied systematically according to the questions they were asked to answer. This finding provided the first empirical evidence that eye movements are not determined solely by the visual properties of a stimulus (bottom-up processes, such as luminance or contrast) but are strongly modulated by the observer's intentions, goals, and expectations (top-down processes). As L. Itti and C. Koch note in Itti and Koch (2001),

Kamila Fačevicová
kamila.facevicova@upol.cz

Jaroslav Vymazal
jarekvmazal@seznam.cz

Stanislav Popelka
stanislav.popelka@upol.cz

¹ Department of Mathematical Analysis and Applications of Mathematics, Palacký University Olomouc, 17. listopadu, 12, Olomouc 779 00, Czech Republic

² Department of Geoinformatics, Palacký University Olomouc, 17. listopadu, 50, Olomouc 779 00, Czech Republic

“subjects selectively direct attention to objects in a scene using both bottom-up, image-based saliency cues and top-down, task-dependent cues” (p. 194). This insight established eye-tracking as a powerful tool for investigating the interaction between perceptual input and cognitive control.

Because fixations mark the loci of overt visual attention, their spatial distribution offers valuable information about attentional allocation, perceptual strategies, and the cognitive mechanisms guiding visual exploration. These distributions can be extensively analyzed and modeled using a range of quantitative approaches, from classical statistical techniques to more advanced analytical frameworks. In many experimental settings, stimuli are subdivided into several disjoint subregions – commonly referred to as areas of interest (AOIs) – and analyses focus on how fixations are distributed across them (Duchowski, 2017). A rigorous statistical treatment of fixation distributions is therefore crucial for extracting meaningful insights about attentional allocation and for distinguishing stimulus-driven effects from individual differences.

When stimuli are structured into AOIs, analyses frequently shift from modeling individual fixations to evaluating aggregated measures, such as the total duration of fixations (TDoF) or total number of fixations within each AOI. Statistical analysis of such data typically relies on classical inferential methods, including *t* tests, ANOVA, or regression models. However, more sophisticated techniques – such as modeling with (hidden) Markov chains (Chuk et al., 2014) or recurrence analysis (RQA) (Anderson et al., 2013; Krejtz et al., 2017) – are also commonly employed in practice. A shared feature of these methods is that they are based on the directly observed data, i.e., the absolute values of dwell time or fixation count recorded for a given participant in a given AOI. In some studies (see, e.g., Altmann and Kamide, 2007), AOI-based metrics are normalized by the size of the respective region, which results, e.g., in fixation duration per unit area, to control for differences in spatial extent. However, in the following, we target the situation when AOIs are defined identically across all stimuli (Holmqvist et al., 2011, p. 226) and the absolute metrics may be influenced more by individual physical or cognitive predispositions of participants than by the saliency, relevance of the stimulus or size of AOI itself. Since the area-normalized metrics are particularly relevant when comparing substantially different AOI sizes across varying stimuli, this situation does not apply to our problem setting.

Consider, for example, an experimental task where AOIs correspond to individual paragraphs in a text, or to problem-solving areas on a map, and participants are allowed unlimited time to complete the task. In such cases, longer dwell time may reflect personal reading speed or information processing habits, rather than a higher level of interest in the AOI or its importance per se. This situation can be illus-

trated on the Yarbush-like experiment, in which the objects on the painting are aggregated in three AOIs – Visitor, Persons and Background, as given in Fig. 1. The same figure provides scanpaths and scarf plots (generated in Vojtechovska et al., 2026) of two observers solving the same task. The first observer made substantially more fixations while observing the painting than the second, resulting in a higher total number of fixations recorded in each AOI. Specifically, the first observer spent 8323, 5353, and 4319 ms in the AOI Visitor, Persons, and Background. For the second observer, the recorded times were 2187, 1377.3, and 1036 ms. Classical analytical methods would likely interpret these results as evidence of substantially different observing behavior. However, these differences may primarily reflect individual observing pace or attention span, rather than truly distinct observing strategies.

A natural way to mitigate the influence of individual characteristics is to replace absolute fixation counts or dwell times with relative measures. Instead of analyzing the TDoFs per AOI, one can evaluate the proportion of an observer’s time that was spent by observing each AOI. For the participants from Fig. 1, this yields the relative distributions (46.3%, 29.7%, 24%) and (47.5%, 29.9%, 22.5%), which clearly reveal the similarity between the two observers’ fixation patterns. The analysis of relative-valued data, typically expressed as percentages or proportions, falls within the scope of Compositional Data Analysis (CoDA) (Coenders et al., 2023). When the relative relationships between variables carry more information than their absolute magnitudes, the application of classical statistical methods – originally developed for unconstrained real-valued data – can yield inaccurate or even misleading results. This issue can be, in general, caused by the presence of spurious correlations between variables, which arise from the constant-sum constraint: an increase in one component (e.g., the proportion of fixations allocated to a given AOI) necessarily induces a decrease in the remaining components, even if no true association exists between them. A second source of distortion lies, in some cases, in the inappropriate use of Euclidean geometry, which is based on absolute differences between values or observations, rather than on their relative relationships (Aitchison, 1982; Pearson, 1897). To illustrate this issue, consider again the time distributions of participants P74 and P101. When analyzing relative data, the distance between two observations should reflect only differences in their relative structure and therefore remain invariant with respect to the chosen representation. However, the Euclidean distance between these two vectors depends on the chosen representation: when computed from the raw fixation times, the distance equals 8014.66, whereas for the percentage representation it is 1.97. While this difference is expected due to the change of scale, it reveals that Euclidean geometry does not respect the relative nature of the data. CoDA offers a principled

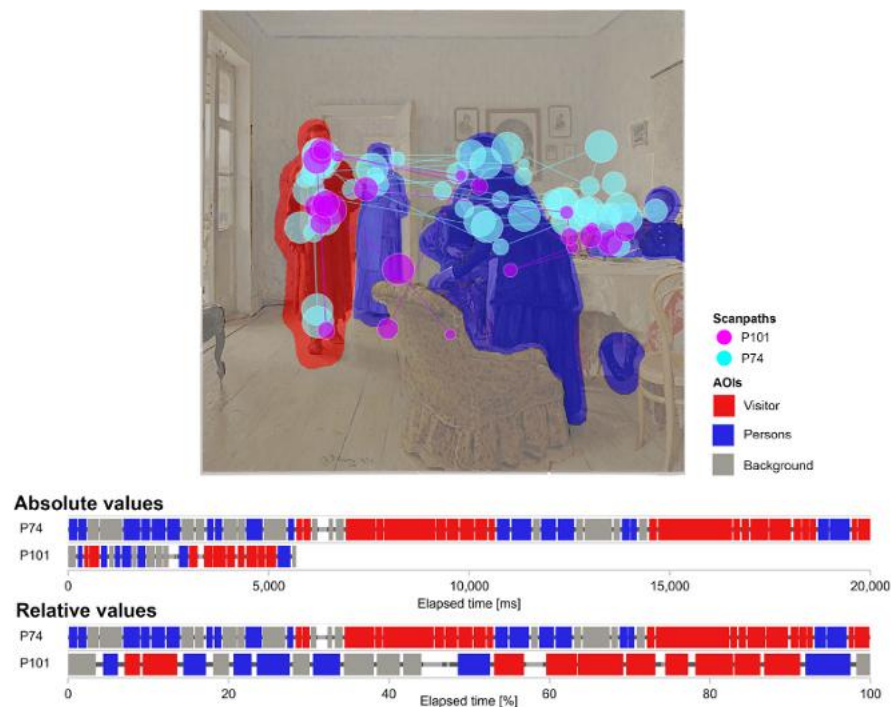


Fig. 1 Illustration of fixation distributions for two observers who differ in total fixation duration but allocate their attention equally across AOIs

way to address this limitation. Its methodology is typically based on log-ratio transformations, which extract the pure relative information from each observation. Crucially, this is achieved without requiring prior normalization to a constant sum, since CoDA treats compositions as equivalence classes of positive vectors conveying relative information only. As will be discussed later, compositional methodology relies on the so-called Aitchison geometry, which yields a distance of 0.068 for the discussed pair of observations, regardless of the specific representation. The resulting distance is therefore invariant with respect to scale or normalization, which should not alter the inherent relative structure.

For an appropriate choice of analytical technique, it is crucial to determine which type of information constitutes the core of the data. Specifically, one must decide whether the analysis should focus on absolute information, based on the observed magnitudes of the measured variable and the differences between them, or on relative information, defined by the ratios among the observed values. As an illustrative example, consider again the TDoFs of observer P74. Recall that the absolute fixation times allocated to the Visitor, Persons, and Background AOIs were 8323, 5353, and 4319 ms, respectively, corresponding to the relative distribution (46.3 %, 29.7 %, 24 %). Now consider another relative time distribution given by (53.6 %, 34.4 %, 12 %). From a classical perspective focused on absolute values, this change would be

interpreted as a halving of the relative time allocated to the third AOI (from 24 % to 12 %), accompanied by increases of 7.3 and 4.7 percentage points in the first and second AOIs, respectively. In contrast, the compositional (relative) perspective interprets the difference between the two percentage distributions in terms of changes in the relative dominance of Visitor and Persons over Background. In the original structure, the ratio of TDoFs between Visitor and Background equals 1.92, meaning that participant P74 observed Visitor approximately twice as long as Background. In the modified structure, this ratio increases to 4.47, indicating a substantial shift in the relative balance of the respective AOIs. Similarly, the ratio between Persons and Background rises from 1.24 to 2.87. Finally, the ratio between Visitor and Persons remains constant at 1.56, demonstrating that their mutual relative importance is preserved.

The aforementioned ratios are invariant with respect to the chosen data representation; the described phenomenon is therefore equivalent whether the data are expressed in TDoFs or, e.g., in percentages. This invariance stands in contrast to the classical interpretation of the data. If the modified relative distribution (53.6 %, 34.4 %, 12.0 %) is considered under the constraint of a fixed total task duration (equal to the original total TDoFs of P74, i.e., 17 995 ms), it would correspond to a redistribution of fixation times to 9637, 6198, and 2159 ms, respectively. In this case, the reduction of time

spent on Background is necessarily compensated by proportional increases in Visitor and Persons, illustrating once again the phenomenon of spurious correlation induced by the constant-sum constraint (data closedness). Alternatively, if total observation time is not constrained, reducing the relative time allocated to Background from 24% to 12%, while keeping the fixation times for Visitor and Persons unchanged, would imply a decrease of the absolute time spent on Background to 1865 ms – less than half of its original value. All three variants of the time distribution can be visually compared on Fig. 2. Such a consequence may appear counter-intuitive and highlights the conceptual differences between absolute and relative interpretations.

For the analysis of data carrying relative information, as introduced above, there exist some modifications of classical statistical methods that explicitly account for the constrained nature of such data. For instance, the relative structure can be accommodated through appropriate parametrization of multinomial or mixed-effects models (see, e.g., Vencálek et al., 2020). Compositional data analysis, however, provides a comprehensive and mathematically grounded framework specifically designed for this type of data. Beyond classical hypothesis testing concerning one or more parameters, CoDA enables the application of a broad spectrum of supervised and unsupervised statistical learning methods, as well as robust inferential procedures. In this sense, it does not merely adjust existing tools to relative data but offers a coherent geometric and algebraic foundation for their analysis.

Even though compositional methodology originally emerged in geochemistry (see, e.g., Tolosana-Delgado, 2012; Fačevićová et al., 2016a), it has since found widespread application across many fields, including metabolomics (Kouřil et al., 2023), microbiome research (Gloor et al., 2017), and time-use epidemiology (Fairclough et al., 2017). This manuscript introduces CoDA in the novel context of eye-tracking data analysis. When combined with complementary approaches such as recurrence quantification analysis (RQA) or spatio-temporal analysis of fixations (Ylitalo et al., 2016), CoDA has the potential to provide a more comprehensive and nuanced characterization of visual attention patterns, as well as their comparison across subjects or experimental conditions.

As stated above, CoDA encompasses a wide range of statistical tools, including methods for hypothesis testing, robust inference, multivariate analysis (both supervised and unsupervised) as well as multi-factorial and even functional (compositional) data analysis (Fačevićová et al., 2022; Filzmoser & Hron, 2011; Pavlů et al., 2023). To the best of our knowledge, this is the first study to apply a compositional approach specifically to the analysis of AOI-based eye-tracking measurements. Accordingly, this manuscript is conceived primarily as a tutorial introduction to compositional data analysis. Its aim is to provide a clear, step-by-step methodological guide to the comprehensive analysis of the relative distribution of visual attention across AOIs. For this reason, we focus on the foundational concepts and practical implementation of CoDA, demonstrating its relevance and applicability in eye-tracking research. The proposed framework is intended to support researchers working with percentage-based time allocations as well as with experiments conducted under fixed total viewing time, and to serve as a practical reference for applying CoDA in eye-tracking studies.

The remainder of the manuscript is structured as follows. Section “**Basics of compositional data analysis**” introduces the main principles of compositional data analysis. Section “**Analysis of problem solving strategies**” presents the replication of Yabus’s experiment together with both the compositional and classical analytical frameworks. Additional results from the practical part are provided in Appendices A and B. The manuscript concludes with Section “**Conclusion**”. All data and source code used in this study are publicly available on GitHub: <https://github.com/kfacevicova/Advancing-Eye-Movement-Analysis-Through-Compositional-Modeling>.

Basics of compositional data analysis

Descriptive statistics

A compositional observation is commonly defined as a D -part vector $\mathbf{x} = (x_1, \dots, x_D)'$ of positive components ($x_i > 0, \forall i$), where the essential information is contained in the

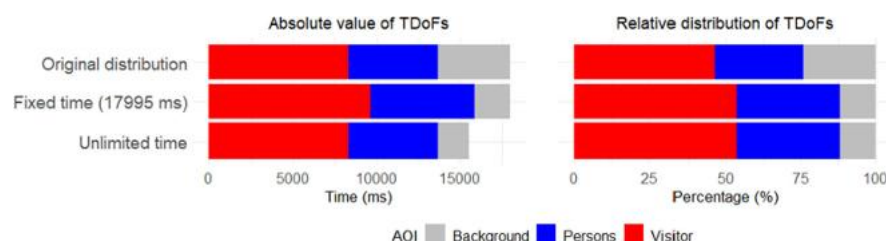


Fig. 2 Comparison of three versions of TDoF distributions and the corresponding compositional structures

ratios between parts, rather than in their absolute magnitudes. This general definition corresponds, for instance, to a situation in which the total duration of fixations spent in D distinct AOIs is recorded with the goal of understanding the relative distribution of attention across these areas. In accordance with this perspective, the total sum of the components is not relevant for the analysis. Normalizing the data to sum to 1 (proportions) or to 100 (percentages) represents only two among many equivalent ways of expressing the same relative information. The underlying relative structure – i.e., the ratios between the parts – remains unchanged regardless of the chosen normalization. A fundamental principle of compositional analysis is scale invariance (Aitchison, 1982), which ensures that rescaling should not affect the outcome of the analysis. Nevertheless, using a common total can be practical for numerical comparison between multiple compositions. To facilitate this, we define the closure operation $C_\kappa(\cdot)$ which rescales a given composition to an equivalent one whose total equals a specified constant κ . It should be noted, however, that in some research contexts, both relative and absolute information may be important – e.g., when analyzing the total duration of a problem-solving task. For the sake of simplicity, we do not address such cases in this manuscript, and instead refer the interested reader to Pawlowsky-Glahn et al. (2015). As mentioned above, the traditional definition of a compositional vector does not permit any of its components to be zero. However, zero values are frequently encountered in practice and must be properly addressed before applying compositional analysis. In the context of eye-tracking data, zeros typically arise when certain AOIs are not visited at all by a given observer. This issue can be handled either by merging sparsely visited areas or by applying suitable zero-imputation techniques, as discussed, for example, in Hron et al. (2010); Palarea-Albaladejo et al. (2007) and Palarea-Albaladejo and Martín-Fernández (2008).

The basic descriptive statistics of compositional data are compositional (i.e., geometrical) mean, variation matrix, and total variance. For a sample on n D -part compositions organized in an $n \times D$ data matrix $\mathbf{X} = (x_{ij})$ these statistics are defined as follows:

The **compositional mean** is given by

$$\mathbf{g}_x = C_\kappa(g_1, \dots, g_D)', \quad \text{where} \quad g_j = \left(\prod_{i=1}^n x_{ij} \right)^{1/n} \quad (1)$$

i.e., (the closure of) the component-wise geometric means across all observations, and the $D \times D$ **variation matrix** $\mathbf{T}_x = (t_{jk})$ has elements given by

$$t_{jk} = S_n^2(\ln(x_j/x_k)) \quad (2)$$

that is, each element represents the sample variance (S_n^2) of the log-ratio between the j -th and k -th components across the

sample. Elements of the variation matrix that are close to zero indicate pairs of compositional parts for which $x_{ij} \approx c \cdot x_{ik}$, with $c \in \mathbb{R}$ being constant across all observations i . In other words, these parts are nearly proportional, and their mutual relationship remains stable throughout the sample. Conversely, large values of t_{jk} indicate high variability in the corresponding log-ratio, reflecting substantial fluctuations in the mutual relative dominance of the two parts across individual compositions. The variation matrix can thus be used to identify AOIs that exhibit inconsistent levels of visual attention across participants, revealing which areas attract interest more variably within the observed population.

Finally, the overall level of variability captures the **total variance**, defined as

$$\text{totvar}(\mathbf{X}) = \frac{1}{2D} \sum_{j=1}^D \sum_{k=1}^D t_{jk}. \quad (3)$$

The total variance serves as a natural measure of similarity (or homogeneity) within a set of compositions and can therefore be used, for example, to compare the stability of visual attention between two groups of observers.

Finally, a traditional way to visualize three-part compositions is through **ternary diagrams**. An example is shown in Fig. 3, where each point represents a single observation, and its position reflects the relative proportions of the three compositional parts. Balanced compositions – in our context, observations where visual attention is distributed equally across all three AOIs – are located near the center of the diagram. As the relative dominance of one part increases, the corresponding points move towards the respective vertex. For instance, observers who focused primarily on the Background are represented by points located near the right vertex of the triangle. Ternary diagrams allow visualization only for $D = 3$. For higher-dimensional data, visualization is typically based on log-ratio transformations, or achieved through dimension reduction techniques such as principal component analysis (PCA) applied to the transformed data.

Working in coordinates

The classical methods of multivariate data analysis, such as hypothesis testing, clustering, or classification (Anderson et al., 1994), are typically not directly applicable to compositional data. These methods generally rely on absolute information carried by data, whereas for compositional analysis, the relative structure is of main interest, as outlined above. The standard approach to overcome this discrepancy is to map compositional observations into a system of real-valued coordinates, where the principles of Euclidean geometry apply, like centered (clr) or isometric (ilr) log-ratio coordinates (Pawlowsky-Glahn & Buccianti, 2011).

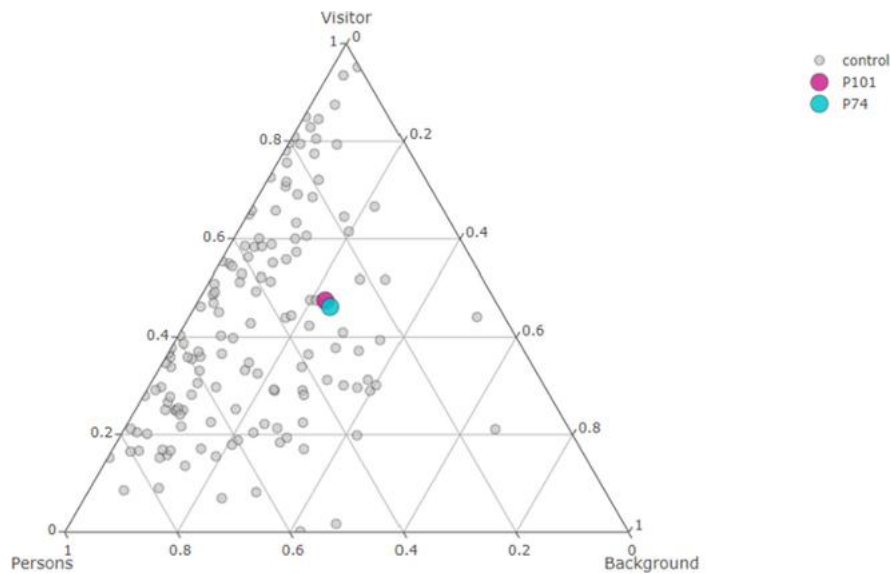


Fig. 3 Ternary diagram of the attention distribution over three AOIs. Colored points correspond to participants P74 and P101, discussed in Section “Introduction”

Centered log-ratio coordinates

One of the widely used options is the **centered log-ratio (clr) transformation**. This transformation expresses the dominance of each compositional part relative to the overall geometric mean of the respective observation, while logarithmic scaling ensures symmetry. Formally, the clr representation of a D -part compositional vector is defined as

$$\text{clr}(\mathbf{x}) = \left(\ln \frac{x_1}{g(\mathbf{x})}, \dots, \ln \frac{x_D}{g(\mathbf{x})} \right)' =: \mathbf{y}, \quad (4)$$

with $g(\mathbf{x}) = (x_1 \cdot \dots \cdot x_D)^{1/D}$. If the clr transformation is applied to each row of the compositional data matrix $\mathbf{X}$, i.e., to each individual observation, the resulting $n \times D$ matrix $\mathbf{Y}$ can be directly used in a wide range of standard multivariate techniques, such as clustering or dimension reduction, as will be discussed later. Moreover, clr coordinates have a convenient interpretation in terms of the mean relative dominance of the respective parts. While positive values represent the more abundant compositional parts, negative values indicate those parts that are rather negligible in a given composition. Thanks to this interpretation, clr coefficients can also be advantageously used as response variables in linear models or their generalizations. The denominator in Eq. 4 serves as a normalizing factor, suppressing between-participant differences in total observation time for a given task. Consequently, the effect of covariates on the (relative) time spent in the AOI of interest can, under the simplest problem setting, be modeled directly without the need for a more complex mixed-effects model.

Nevertheless, a limitation arises from the inherent zero-sum constraint of the clr transformation, which follows directly from its definition. As a consequence, the covariance matrix of clr-transformed data is singular, preventing its direct use in some statistical procedures. For methods relying on full-rank covariance structures, an alternative coordinate system, most notably the isometric log-ratio (ilr) transformation (Egozcue & Pawłowsky-Glahn, 2005; Egozcue et al., 2003), provides a suitable solution.

A widely used technique for data exploration is **principal component analysis (PCA)**, a dimension-reduction technique that represents the data through a small number of linear combinations of the original variables, referred to as principal components (PCs). This approach facilitates low-dimensional visualization of the data and enables the identification of meaningful patterns both among observations and among variables. The key step is to assign coefficients to the variables, thereby forming the principal components. One common way to obtain these coefficients, referred to as loadings in what follows, is through singular value decomposition (SVD) of the observed data matrix.

In order to retain focus on the relative structure, PCA for compositional data is typically performed on its clr representation, i.e., on the SVD of the matrix $\mathbf{Y}$. Analogously to classical PCA, it is common practice to work with a column mean-centered matrix (Filzmoser et al., 2018, p. 133). Let us denote this centered matrix by $\tilde{\mathbf{Y}}$. The SVD of the $n \times D$ matrix $\tilde{\mathbf{Y}}$ can be written as

$$\tilde{\mathbf{Y}} = \mathbf{U}\mathbf{D}\mathbf{V}', \quad (5)$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices of left and right singular vectors, and $\mathbf{D}$ is a diagonal matrix of singular values λ_k , arranged in descending order. Columns of $\mathbf{V}$, $\mathbf{v}_k = (v_{1k}, \dots, v_{Dk})'$, $k = 1, \dots, \min(n, D)$, correspond to loadings, and the k -th principal component (PC_k) is then generally computed as $\mathbf{v}_k' \tilde{\mathbf{y}}$. The matrix of scores, i.e., the values of the principal components for all observations, is given by the matrix product $\mathbf{U}\mathbf{D}$. The proportion of variance explained by the k -th PC is given as $\lambda_k / \sum_j \lambda_j$, and typically the first two components are of primary interest, as they explain the largest proportion of the total variance. Note that the sum of eigenvalues in the denominator equals the total variance of the compositional data matrix $\mathbf{X}$, as defined by Eq. 3.

The result of PCA can be easily interpreted using clr loadings plot and clr (compositional) biplot, as illustrated in Fig. 8. A clr biplot simultaneously represents both scores and loadings, typically in relation to the first two principal components. More specifically, in the context of eye movement analysis, points correspond to participants, while the direction and length of rays represent clr-transformed compositional parts – times spent in individual AOIs, determined by the respective loadings. This visualization enables the representation of a D -part compositional dataset ($D > 2$) in two dimensions, allowing visual inspection for potential outlying observations or clusters – individuals or participant groups with distinct or deviating observation behavior. Furthermore, relationships between compositional parts/AOIs can be explored through the positions of the ray vertices. Unlike in a classical (non-compositional) biplot, where interpretation relies mainly on the length of rays and the angles between them, interpretation of the compositional biplot is based on the links between pairs of vertices (Pawlowsky-Glahn et al., 2007). The length of such a link estimates the variability of the corresponding pairwise log-ratio and can thus be understood as a measure of proportionality between times spent in the respective two AOIs. In extreme cases, coinciding vertices indicate (nearly) proportional times, while very distant vertices suggest AOIs whose mutual relationships vary substantially across the dataset.

The axes of the biplot, and consequently the positions of points, are determined by the first two principal components. The practical interpretation of these components can be supported by clr loadings plot. Such barplots display the loadings of a given component for the individual compositional parts. Due to the zero-sum constraint inherent to clr coefficients, the elements of each loading vector also sum to zero. As a consequence, the principal components can be understood as a special case of log-contrasts (Aitchison & Bacon-Shone, 1984). Finally, in certain cases, they can be interpreted in simplified terms as reflecting the dominance of one group of parts (with high positive loadings) over another (with high negative loadings).

The other typical problem of multivariate statistics is the problem of definition for groups at a level of observations – participants or variables – AOIs. This task can be solved with methods of **cluster analysis**. Within this manuscript, we focus on hierarchical clustering, i.e., an approach which progressively merges smaller clusters into larger ones, with the aim to get in each step a set of disjoint groups of units (observations or variables), which are within each cluster as similar as possible and, simultaneously, have the highest possible level of between-clusters dissimilarity. The result is typically visualized as a dendrogram, which illustrates the nested grouping and the level of similarity at which clusters are joined.

The crucial step in hierarchical clustering is the definition of a dissimilarity measure between pairs of units. In classical (non-compositional) settings, this is typically based on the Euclidean distance. By analogy, hierarchical clustering of compositional observations should employ the Aitchison distance between pairs of compositional vectors (Aitchison, 1986). Importantly, one of the key properties of the clr representation is that it converts the Aitchison distance $d_A(\cdot, \cdot)$ into the Euclidean one $d_E(\cdot, \cdot)$. In particular, for two D -part compositions $\mathbf{x}$ and $\mathbf{z}$, it holds that $d_A(\mathbf{x}, \mathbf{z}) = d_E(\text{clr}(\mathbf{x}), \text{clr}(\mathbf{z}))$. Thanks to this property, the hierarchical clustering of observations that focuses on their relative structure can be simply performed on clr coefficients.

When the research question lies in the clustering of variables, i.e., identifying groups of AOIs that, in terms of TDoFs, exhibit similar behavior across the data sample, the clustering algorithm is typically based on the variation matrix $\mathbf{T}_\mathbf{x}$, see Eq. 2. Recall that the elements of $\mathbf{T}_\mathbf{x}$ correspond to the variances of pairwise log-ratios; large values can therefore serve as a natural sign of dissimilarity of the respective pair of AOIs.

Isometric log-ratio coordinates

The zero-sum constraint prevents the direct use of clr coefficients in a wide range of multivariate methods that rely on a full-rank data matrix and a regular covariance structure. For regression modeling, robust analyses, or hypothesis testing, it is therefore more practical to work with $D - 1$ so-called isometric log-ratio (ilr) coordinates. Although this coordinate system can be defined in multiple ways, the specific choice does not affect the outcomes of subsequent analyses. For simplicity, we focus here on one specific type known as pivot coordinates (Hron et al., 2017). The ilr transformation of a D -part compositional vector $\mathbf{x}$ yields a real vector $\text{ilr}(\mathbf{x}) =: \mathbf{z}$ with components

$$z_j = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{x_j}{\left(\prod_{k=j+1}^D x_k\right)^{\frac{1}{D-j}}}, \quad j = 1, \dots, D-1. \quad (6)$$

The elements of the coordinate system defined by Eq. 6 can be interpreted in terms of the mean relative dominance of a given compositional part over all parts of higher order. More generally, however, the ilr family of coordinate systems is characterized by high flexibility in interpretation, which is typically expressed in terms of dominance between two groups of parts. For this reason, ilr coordinates are often referred to as *balances* (Egozcue & Pawłowsky-Glahn, 2005).

For ilr coordinates, as for clr, the principles of classical Euclidean geometry apply. In contrast to clr, however, this representation does not suffer from the zero-sum constraint and can therefore be readily used in methods that rely on a full-rank covariance structure, such as robust statistical procedures, regression models with compositional covariates, or hypothesis testing. The latter will be outlined in the following paragraph.

The similarity of problem-solving strategies between two or more groups of participants, i.e. similarity of their mean compositions as defined by Eq. 1, can be tested using **compositional MANOVA**. This approach is equivalent to classical multivariate analysis of variance applied to ilr-transformed data (Martín-Fernández et al., 2015). Suppose K groups of a D -part composition are considered, each consisting of n_k observations, $k = 1, \dots, K$, with $\sum n_k = n$. The within-group structure can be expressed as

$$\mathbf{x}_i^k = \mu_k \oplus \mathbf{e}_i^k, \quad i = 1, \dots, n_k, \quad (7)$$

where $\mathbf{x}_i^k$ is the i -th observation in the k -th group, μ_k denotes the group mean, and $\mathbf{e}_i^k$ is a random term with an expected value equal to the unit vector. Symbol $\oplus$ denotes perturbation, one of the basic operations of Aitchison geometry, which corresponds to component-wise multiplication. The null hypothesis of group similarity corresponds to

$$\mu_1 = \dots = \mu_K \quad (8)$$

and is tested under the assumption of log-normality and homoscedasticity of $\mathbf{e}_i^k$, $\forall k$. Moving to ilr transforms the model into a classical MANOVA setting: $\text{ilr}(\mathbf{x}_i^k) = \text{ilr}(\mu_k) + \text{ilr}(\mathbf{e}_i^k)$, $\forall i, k$, with the null hypothesis $\text{ilr}(\mu_1) = \dots = \text{ilr}(\mu_K)$ and the error terms distributed as $\text{ilr}(\mathbf{e}_i^k) \sim N(\mathbf{0}, \Sigma)$, $\forall i, k$. Hypothesis testing is after the ilr transformation performed similarly to the classical approach, which allows for the computation of multivariate test statistics, such as Pillai's trace, for assessing significance, as well as for the calculation of the test effect size, e.g., in the form of partial η^2 (Muller et al., 1992). For further methodological details, see Martín-Fernández et al. (2015).

Similarly to procedure used in MANOVA, the ilr representation of compositional data can be directly employed for classification tasks using, for example **linear** or **quadratic**

discriminant analysis (LDA and QDA). Suppose that the observations can be classified into K groups. If the ilr coordinates $\mathbf{z}$ from the k -th group follow a $D - 1$ -dimensional normal distribution with density $f_k(\mathbf{z})$, $k = 1, \dots, K$, Bayes' theorem can be used to construct the classification rule

$$P(Y = k | \mathbf{Z} = \mathbf{z}) = \frac{\pi_k f_k(\mathbf{z})}{\sum_j \pi_j f_j(\mathbf{z})} \rightarrow \max_k, \quad (9)$$

where $\pi_1, \dots, \pi_K$ denote the a priori class probabilities. In particular, LDA assumes a common covariance structure across all groups, i.e. $N(\mu_k, \Sigma)$, $\forall k$, whereas QDA allows for group-specific covariance matrices Σ_k , $k = 1, \dots, K$, which results in greater model flexibility (James et al., 2021). The unknown parameters μ_k and Σ_k , $k = 1, \dots, K$, or the pooled covariance matrix Σ , are estimated analogously to the classical (non-compositional) case using the sample mean vectors and covariance matrices computed from the training dataset consisting of pairs $(Y_i, \mathbf{z}_i)$, $i = 1, \dots, n$, where Y_i denotes the categorical response and $\mathbf{z}_i$ the corresponding ilr vector. Importantly, the resulting classification performance is invariant to the specific choice of the ilr coordinate system; see Filzmoser et al. (2018) for further details.

Software implementation

In practice, compositional data analysis is well supported in R (R Development Core Team, 2009). The most widely used packages include *compositions* (Boogaart Van Den et al., 2009) and *robCompositions* (Templ et al., 2011). The specific R implementations of the methods described above are listed in Table 1

Analysis of problem solving strategies

The principles of compositional analysis of eye-tracking data will be illustrated using a classic experiment first described by Yarbus (1967). Yarbus was among the first to systematically investigate the relationship between eye movement patterns and high-level cognitive factors (DeAngelus & Pelz, 2009). In his well-known study, participants were asked to view a politically significant painting depicting a Russian revolutionary returning from exile under different task instructions, such as estimating the ages of the people portrayed or remembering their clothing. The resulting scanpaths revealed striking differences in eye movement patterns across task conditions, differences that were more pronounced than those observed among seven participants freely viewing the painting without specific instructions (see (Yarbus, 1967, Fig. 107)). Based on these observations, Yarbus concluded that the eyes “fixate on those elements of an object which carry or may carry essential or useful

Table 1 Functions for compositional data analysis in *compositions* and *robCompositions* packages in R

Method	<i>compositions</i>	<i>robCompositions</i>
Compositional mean	<code>mean(acomp(x))</code>	<code>gmean(x)</code>
Variation matrix	<code>variation(acomp(x))</code>	<code>variation(x, method = "Pivot")</code>
Total variance	<code>mvar(acomp(x))</code>	–
CLR transformation	<code>clr(x)</code>	<code>cenLR(x)</code>
ILR transformation	<code>ilr(x)</code>	<code>balances(x)</code>
Pivot coordinates	–	<code>pivotCoord(x)</code>
PCA	<code>princomp(acomp(x))</code>	<code>pcaCoDa(x, method = "classical")</code>
Hierarchical clustering	<code>CoDaDendrogram(acomp(x))</code>	<code>clustCoDa(x), clustCoDa_qmode(x)</code>
MANOVA	<code>manova(ilr(x) ~ group)</code>	–
Discriminant analysis	<code>lda(ilr(x)), qda(ilr(x))</code>	<code>daCoDa(x)</code>

information” (Yarbus, 1967, p. 211). This experiment is significant because it demonstrates a clear top-down component of visual selection and highlights the active nature of the human visual system. Observers’ cognitive goals and prior experiences interact with the visual stimulus to guide attention and support task-relevant behavior. The visual system is therefore not passive: it does not randomly or uniformly sample the environment, nor does it merely react to external stimuli. Instead, it actively selects and prioritizes information based on the cognitive context. Yarbus’ findings remain one of the most well-known and frequently cited demonstrations in the field of eye-tracking research, providing foundational evidence of the task-dependent nature of saccadic eye movements and continuing to inform contemporary studies of visual attention. The English translation of Yarbus’s book (Yarbus, 1967) became one of the most cited works in visual science, with a surge in citations from the mid-1990s onward (Tatler et al., 2010, p. 23). His findings inspired the development of Active Vision models (Ballard, 1991; Ballard et al., 1992; Findlay & Gilchrist, 2003) and influenced research in artificial intelligence, robotics, and cognitive science beyond psychology (Tatler et al., 2010, p. 23).

Following a description of the data sampling and pre-processing procedures, results obtained by both compositional and classical methods – relying on relative and absolute information, respectively – will be presented and compared.

Data description

Participants

A total of 144 participants (79 female, 65 male) took part in the study. Their ages ranged from 19 to 28 years (mean = 22.20). All participants were fluent Czech speakers, which was essential because all instructions and tasks in the experiment were presented in Czech. Participants were recruited primarily from the student pool of Palacký Univer-

sity Olomouc and volunteered to participate in a battery of eye-tracking experiments.

All participants reported normal or corrected-to-normal vision. At the time of data collection, 44 wore glasses, and seven wore contact lenses. Right-handedness was reported by 127 participants and left-handedness by 16; one participant did not report hand dominance. The dominant eye was right for 104 participants and left for 37 (three participants did not provide this information). Eye dominance was assessed using a self-report measure asking participants which eye they would use to look through a telescope, following (Ehrenstein et al., 2005).

To obtain additional background characteristics, participants completed a short demographic questionnaire. These characteristics were collected in order to enable more detailed analyses beyond the scope of the present paper. Regarding education, 29 participants had completed secondary vocational education with Maturita (the Czech and Slovak school-leaving examination at the end of secondary education), 47 graduated from a grammar school (gymnasium), 56 held a bachelor’s degree, ten a master’s degree (two did not report their education level). Participants also reported their relationship to art, with four describing it as negative, 35 as neutral, and 105 as positive. Since the experiment included distractor stimuli involving logical reasoning, participants were also asked about their relationship to mathematics: 34 described it as negative, 41 as neutral, and 69 as positive.

All participants were screened for suitability to participate in an eye-tracking experiment. The only exclusion criterion was the inability to achieve a successful calibration of the eye-tracker. Four participants were excluded from the initial dataset of 148 participants due to unsuccessful calibration.

The study was conducted in accordance with the ethical standards of the Faculty of Science, Palacký University Olomouc. It was approved by the institutional Ethics Committee (reference no. 22-01), and written informed consent was obtained from all participants prior to the experiment.

Apparatus and material

Eye movements were recorded using a Tobii Pro Spectrum 300 eye-tracker (Tobii AB, Sweden), operating at a sampling frequency of 300 Hz. According to the manufacturer, the system provides a spatial accuracy of approximately 0.3° and a precision of 0.06° . Before the experiment, a nine-point timed calibration was performed for each participant, with calibration points presented in a randomized order as white dots on a grey background. The mean calibration accuracy across participants was 0.197° with a standard deviation of 0.133° , indicating high-quality gaze data.

The main stimulus presented during the experiment was the painting *They Did Not Expect Him (Unexpected Visitor)* by Ilja Repin, obtained from Wikipedia. The image was displayed at a resolution of 1125×1080 pixels, with the remaining area of the screen filled with black.

All stimuli were displayed on a 23-inch LCD monitor with a resolution of 1920×1080 pixels. Participants were seated approximately 65 cm from the screen, and no chin-rest was used. The experiment took place in the eye-tracking laboratory of the Department of Geoinformatics, Palacký University Olomouc. The room had covered windows, and lighting conditions were kept constant for all participants to minimize variability in pupil tracking and visual perception.

Data acquisition and stimulus presentation were controlled using Tobii Pro Lab software. Areas of Interest (AOIs) were defined manually within Tobii Pro Lab, and the resulting gaze data were exported as tab-separated values (TSV) files. The statistical analysis of gaze metrics was subsequently performed in R (R Development Core Team, 2009).

Procedure

This experiment was designed as a replication of the seminal eye-tracking study by Yarbus (1967), using the same stimulus image and task instructions as in the original work. However, some aspects of the procedure differed. In the original study, each viewing lasted three minutes, whereas in our experiment, the observation time was not limited; participants could proceed to the next slide by pressing the space bar once they felt they had completed the task. Additionally, Yarbus presented each task on separate occasions spaced by 1–2 days, while in our experiment, all tasks were presented within a single session, separated by distractor tasks (logical puzzles) to minimize learning and memory effects.

At the beginning of the experimental session, participants were seated in front of the eye-tracker, and the nine-point calibration procedure was performed as described above. Following successful calibration, the experiment commenced with an introductory slide that informed participants that the study focused on logical thinking (related to distractor tasks) and visual attention, referencing the described experiment.

Participants were instructed that they would complete a series of logical tasks interspersed with tasks involving the observation of a painting. They were informed of the specific task for each presentation before the stimulus appeared. After reading the task instructions, participants viewed the stimulus and were instructed to examine it carefully. When ready, they pressed the spacebar to advance to the response screen, where they selected their answer by clicking on the appropriate text option using a mouse. The responses were evaluated based on areas of interest (AOIs) defined around each possible answer.

The main part of the experiment consisted of seven presentations of the same painting, *They Did Not Expect Him (Unexpected Visitor)* by Ilja Repin, presented in a fixed order. The seven tasks were adapted from Yarbus's classic experiment (Yarbus, 1967) and were as follows:

- Task 1 Look carefully at the following image.
- Task 2 Estimate the material circumstances of the family in the picture.
- Task 3 Give the ages of the people.
- Task 4 Surmise what the family has been doing before the arrival of the unexpected visitor.
- Task 5 Remember the clothes worn by the people.
- Task 6 Remember the positions of people and objects in the room.
- Task 7 Estimate how long the unexpected visitor has been away from the family.

Between these presentations, participants completed eight logical thinking tasks that served as distractor stimuli. These tasks were unrelated to the main purpose of the experiment and are therefore not described here in detail.

At the end of the experiment, a debriefing slide was displayed. It thanked participants for their involvement and provided context about the broader aims of the study. The slide linked the experiment to Yarbus's foundational work on eye movements, explained the concepts of top-down and bottom-up visual processing, and highlighted the relevance of such research in various applied fields. This served both as a closure to the experimental session and as a brief educational component that enhanced participants' understanding of the study's significance.

The entire experiment, including calibration and distractor tasks, lasted on average 13.3 minutes.

The areas of interest were marked in the stimuli similarly to those used in DeAngelus and Pelz (2009). For the present analysis, we grouped them into seven AOIs: *Background*, *Furniture*, *Kids*, *Maid*, *Man*, *Mother*, and *Wife*, see Fig. 4. A more detailed segmentation with a larger number of AOIs – and, consequently, a higher data dimensionality – would require a substantially larger sample size to ensure reliable inference. Therefore, the analysis was restricted to those seven areas of interest. All results are based on the

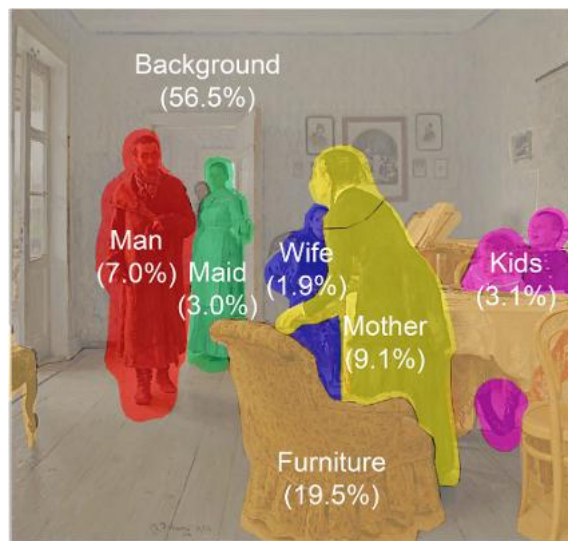


Fig. 4 Segmentation of I. E. Repin's painting 'They Did Not Expect Him' into areas of interest. The numbers in brackets indicate the percentage of the painting covered by each respective AOI

total durations of fixations (TDoFs) recorded within each sub-AOI, which were aggregated prior to analysis according to their assignment to the corresponding AOI.

Since all participants viewed the same stimulus and AOIs were defined identically across all tasks, the spatial extent of each AOI remained constant throughout the experiment. Consequently, comparisons between tasks or participants were not affected by differences in AOI size. For this reason, fixation durations were not normalized by AOI area. Area-normalized metrics are particularly relevant when comparing AOIs of substantially different sizes within varying stimuli; however, this situation does not apply to the present experimental design.

Results

The nature of the experiment, in which participants were given unlimited time to solve each task, allows for a comparison of problem-solving strategies both from the perspective of the relative distribution of attention across AOIs (compositional analysis) and from the perspective of absolute differences in total durations of fixations (classical analysis). The combination of these two approaches can provide a comprehensive understanding of differences between participants (or groups of participants) as well as between tasks. In the following, we primarily focus on the latter case, i.e., differences between tasks.

Analysis of relative time distribution over AOI - compositional analysis

The observation times corresponding to the considered tasks are compared in Fig. 5, revealing substantial differences between them. This finding justifies the use of a relative approach when the primary objective is to compare how visual attention is distributed across AOIs for different tasks.

To gain an initial understanding of the data structure, Fig. 6 (left) presents a comparison of compositional means, see Eq. 1, computed for each task and rescaled to percentages. This comparison already highlights substantial differences in observation strategies, which appear to depend on the specific task type. The most pronounced difference between tasks lies in the attention devoted to Furniture, which was almost entirely neglected in Tasks 3, 5, and 7. In contrast, the hierarchical clustering of means (Fig. 6, right) based on the complete linkage, i.e., distance between the farthest neighbors, which is also used throughout the entire analysis, indicates distinct behavior in Tasks 1, 2, 4, and 6, where participants spent a substantial portion of their time observing Furniture and Background. For Task 1 (free viewing), the distribution is similar to the relative size of the AOIs: approximately one third of the total time was spent on Background, which covers the largest part of the painting (56.5 %). When reasoning about the family activity before the visitor's arrival (Task 4), participants concentrated predominantly on Kids, whereas in estimating the duration of the visitor's absence (Task 7), their attention was directed mainly toward the Man.

The variability of attention devoted to individual AOIs can be examined using the variation matrix, see Eq. 2. Figure 7 presents the variation matrices for Tasks 5 and 7, while the corresponding matrices for the remaining tasks are provided in the Appendix (Fig. 18). The matrices clearly reveal task-specific differences in the distribution of attention. For Task 5, which required focusing on the clothing of the persons, the variation matrix shows a distinct structure. The TDoFs devoted to the persons in the picture – AOIs Kids, Maid, Man, Mother, and Wife – are nearly proportional, pairwise ratios of times spent in these AOIs remain stable across all participants, while high log-ratio variability is observed between Furniture and Background. Consequently, these two AOIs exhibit much higher mutual variability compared to the person-related AOIs. Recall that proportionality between two AOIs does not imply that the percentages of time spent in the respective AOIs are equal; rather, it indicates that their ratio remains stable across the data sample. Conversely, high log-ratio variability reflects an unstable relative allocation of time between the considered pair of AOIs. See Section “Descriptive statistics” for details. In the case of Task 5, this means that the amount of attention directed to Background or

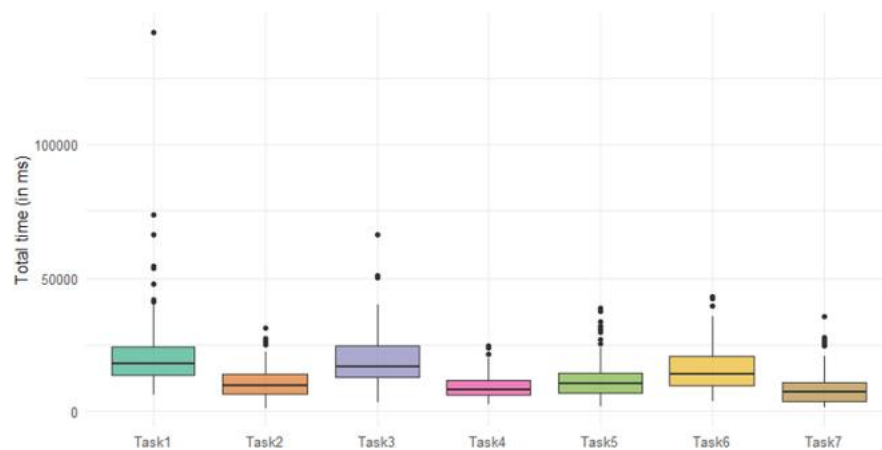


Fig. 5 Comparison of total time distributions

Furniture varies substantially over participants, whereas the relative distribution of the remaining time across the AOIs representing persons remains stable throughout the sample. For Task 7, which focused on estimating the length of the visitor's absence, no clear structure can be observed. All elements of the variation matrix are relatively high, indicating substantial variability in the overall distribution of salience. This conclusion is further supported by the high value of the total variance, see Eq. 3, which for Task 7 equals 17.62. By contrast, the total variance for the remaining tasks, which are given in Table 4, ranges between 5 and 10, with Task 5 specifically reaching 10.03.

Visual inspection of the data structure for potential clusters or outlying observations can be facilitated by PCA, which also provides a more detailed view of the clustering pattern

observed in Fig. 6 (right). PCA reduces the original dataset, comprising seven variables - AOIs, to a two-dimensional representation while retaining the majority of the data variability. The first two principal components of the compositional PCA are illustrated in Fig. 8. Differences between tasks are primarily manifested along the first principal component (PC1). Tasks 1, 2, 4, and 6 generally exhibit negative PC1 scores, whereas Tasks 3, 5, and 7 are associated with higher values. According to the loadings, PC1 is predominantly defined by the log-ratio between Furniture and the mean time spent on AOIs containing persons. Consequently, the shift of Tasks 3, 5, and 7 towards lower attention to Furniture reflects their focus on human figures depicted in the paintings. The second PC clearly separates a group of observations with atypically low values. Based on the loadings, PC2 primarily captures

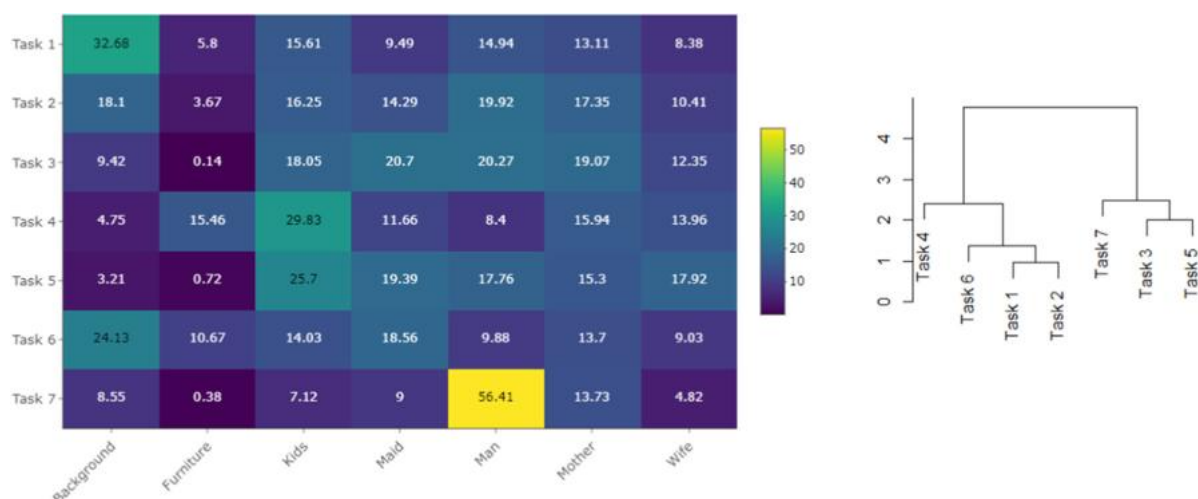


Fig. 6 Compositional means of TDofS spent in each AOI, see Eq. 1, computed separately for each type of task and rescaled to percentages (*left*) and clustering of tasks according to mean time distributions (*right*)

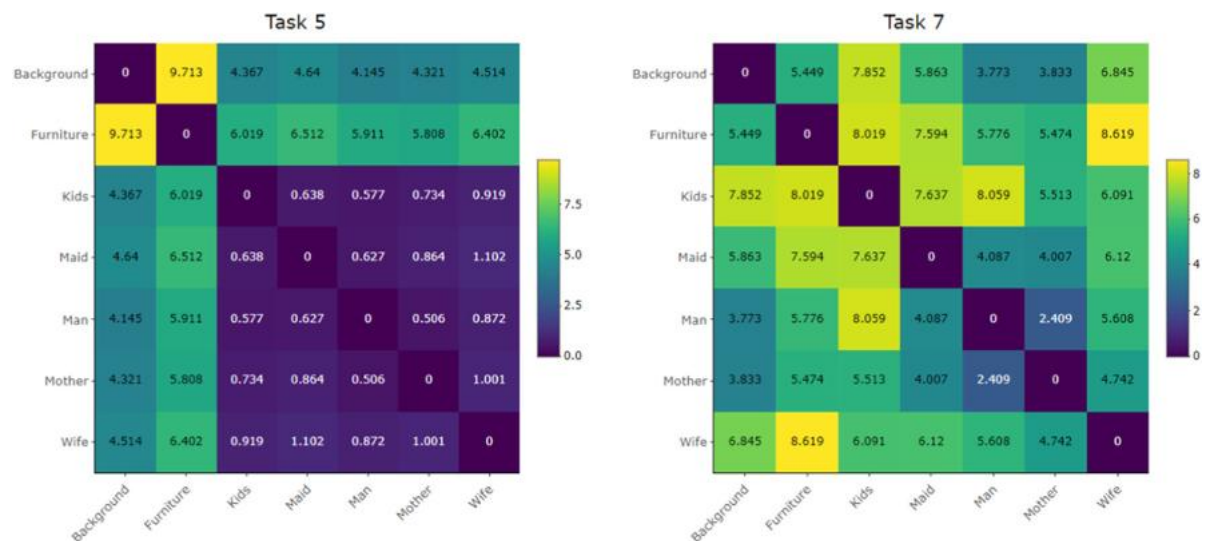


Fig. 7 Variation matrices, see Eq. 2, computed for Tasks 5 (left) and 7 (right). Each cell corresponds to the sample variance of log-ratio between the respective pair of AOIs

the relative dominance of time spent on the Background compared to the Wife and Kids areas of interest. Consequently, the cluster of observations located at the lower end of the plot corresponds to participants who, within the given task, allocated minimal attention to the Background. Finally, the length and orientation of the rays further contribute to the interpretation of the data structure. The longest rays, corresponding to Furniture and Background, indicate substantial variability in the relative time spent on these two AOIs across tasks and participants. The opposite orientation of the Furniture ray, and consequently the considerable distance of its vertex from those representing the remaining AOIs, reflects the pronounced variability in pairwise log-ratios between Furniture and any other AOI. In contrast, the close proximity of vertices associated with person-related AOIs indicates a consistent distribution of attention across these AOIs throughout the experiment, which is in accordance with information provided by variation matrices. It should be noted that the results presented in Fig. 8 are based on the complete dataset and are intended to reveal the main patterns observed across the entire experiment, such as differences between tasks or the most variable AOIs. More detailed insights into participants' behavior could be obtained by conducting a separate PCA on the measurements from a single selected task.

All the analyses presented above indicate that observation strategies differ depending on the specific task. The equivalence of distributions can be statistically assessed using compositional MANOVA, as described in Section “*Isometric log-ratio coordinates*”. More precisely, we test the null hypothesis that the mean distributions of viewing times are

independent of the task and equal to the overall mean reported in Table 2. This hypothesis is rejected (with Pillai's trace = 1, $F = 33.35$, $df = (36, 6006)$, $p \ll 0.001$ and $\eta^2 = 0.167$). The sources of this rejection are clearly illustrated in Fig. 9, which contrasts task-specific geometric means ($g(x)_{AOI}^{task}$) with the overall means under the null hypothesis ($g(x)_{AOI}^{all}$) using log-ratio

$$\ln \left(\frac{g(x)_{AOI}^{task}}{g(x)_{AOI}^{all}} \right). \quad (10)$$

Values of these log-ratios close to zero would indicate agreement with the null hypothesis, whereas larger deviations highlight systematic differences between tasks.

The findings presented in Fig. 9 are consistent with those illustrated in Fig. 6. For example, the portion of time spent on the AOI Furniture shows considerable variability: while Tasks 4 and 6 are characterized by extended viewing times, this duration is substantially reduced in Tasks 5, 7, and particularly in Task 3. Another notable feature is the markedly longer relative time devoted to the AOI Man in the final task, which concerns the duration of the unexpected visitor's absence. In addition to the global test of the main hypothesis, a series of pairwise MANOVA tests was performed. After applying Holm correction, all p values were well below 10^{-5} , indicating significant differences across all task pairs.

The task-specific differences in the relative time allocated to a given AOI can be modelled and tested using a linear model. For example, in order to investigate how the relative time spent in AOI Man varies with respect to the task being solved, we can consider a linear model of

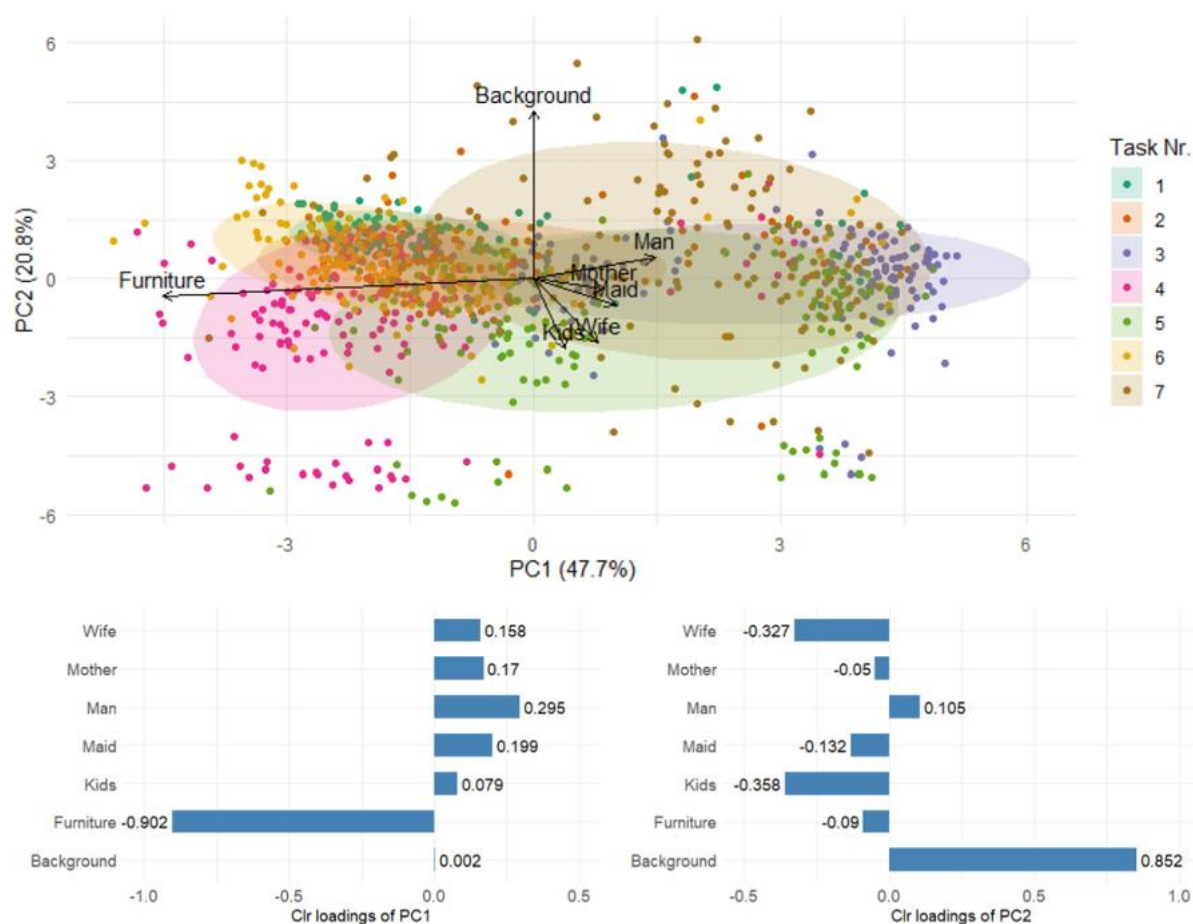


Fig. 8 Clr biplot and loadings corresponding to the first two principal components, with percentages of explained variance provided in the brackets. The ellipses enclose 75 % of the typical observations for each category

the form $\text{clr}(\text{Man}) \sim \text{Task}$. Nevertheless, even though the between-participant differences in total observation time are already suppressed by the clr transformation, the dataset at hand contains grouped observations. Therefore, the use of a mixed-effects model, with participant ID serving as a random effect, is more appropriate. The estimated fixed effects are shown in Fig. 10 (left). The effects are computed as the difference between the respective clr value for a given task and its mean value across all tasks. Negative values therefore correspond to tasks where the attention paid to AOI Man is relatively shorter (Tasks 1, 4, and 6), while positive values, as observed for Tasks 3, 5, and 7, correspond to tasks with relative observation times for AOI Man above

the overall mean. Due to the form of the clr transformation (see Eq. 4), changes in the relative dominance of AOI Man are interpreted multiplicatively. More specifically, the largest effect, estimated as 1.58 for Task 7, means that on average the relative dominance of AOI Man in Task 7 increased by $\exp(1.58) = 4.85$ times compared to the overall mean. Conversely, the relative dominance of this AOI in Task 4 is about $1/\exp(-0.91) = 2.48$ times lower than the mean. Note that the estimated effects slightly differ from the comparison provided in Fig. 9. While Fig. 9 compares the task-specific and overall mean percentage distributions – where the participant effect is taken into account only once, through the total time spent on the given task – the mixed-effects model accounts for

Table 2 Mean time distributions computed from the complete dataset (in percentages)

Background	Furniture	Kids	Maid	Man	Mother	Wife
12.52	2.33	19.21	16.19	20.22	17.71	11.82



Fig. 9 Values of log-ratio defined by Eq. 10 computed for each task and AOI

participant-specific characteristics twice: first when computing the clr representation, where the total solution time for a given task is suppressed, and second when the effects of tasks are estimated relative to the mean clr value observed across all tasks, which may again be participant-specific. With this construction, the mixed-effects model corresponds to the clr representation of the interaction part of a compositional table (Fačevićová et al., 2025, Eq. (8)), which models the joint distribution of TDoFs across the entire experiment, i.e., how they are distributed across combinations of Tasks and AOIs. The methodology of compositional tables is beyond the scope of this manuscript; however, interested readers can find further details in Fačevićová et al. (2016).

To illustrate hierarchical clustering and its outcomes, we focus once again on Task 5, “Remember the clothes worn by the people.” The clustering results of participants, based on the relative distribution of their attention across AOIs, are presented in the upper panel of Fig. 11. The dendrogram

identifies four main participant groups, which correspond to those observed in the clr biplot (Fig. 8); see also Fig. 19 for a biplot restricted to Task 5 with the groups highlighted in color. In Fig. 11, the groups are delineated by colored rectangles. The green group corresponds to participants located in the lower left corner of the biplot, characterized by relatively long viewing times in AOI Furniture and minimal attention to Background. The blue group, situated in the lower right corner, consists of participants focusing primarily on persons rather than on Furniture. The yellow and red groups are no longer marked by atypically low attention to Background but differ in their allocation of attention between Furniture and person-related AOIs: participants in the yellow group focus predominantly on persons, whereas those in the red group devote an atypically high proportion of their attention to Furniture.

Additionally, the variation matrix presented in Fig. 7 can be employed for clustering of AOIs, as shown in Fig. 12,

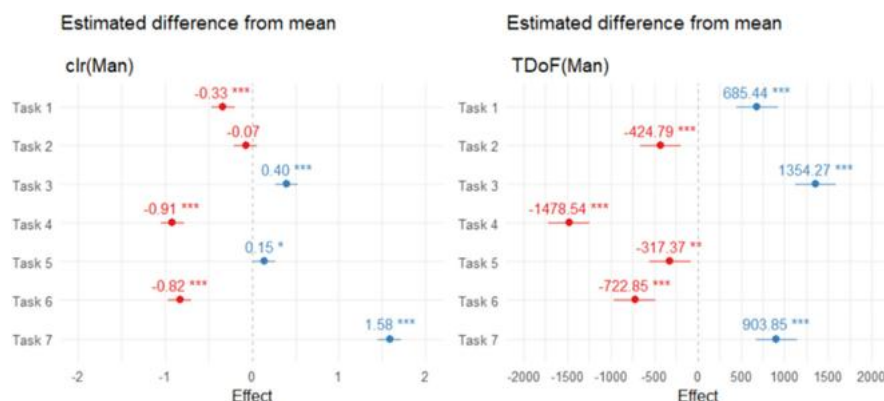


Fig. 10 Estimated fixed effects of the linear mixed-effects model for the relative (left) and absolute (right) time spent in the AOI Man across task types

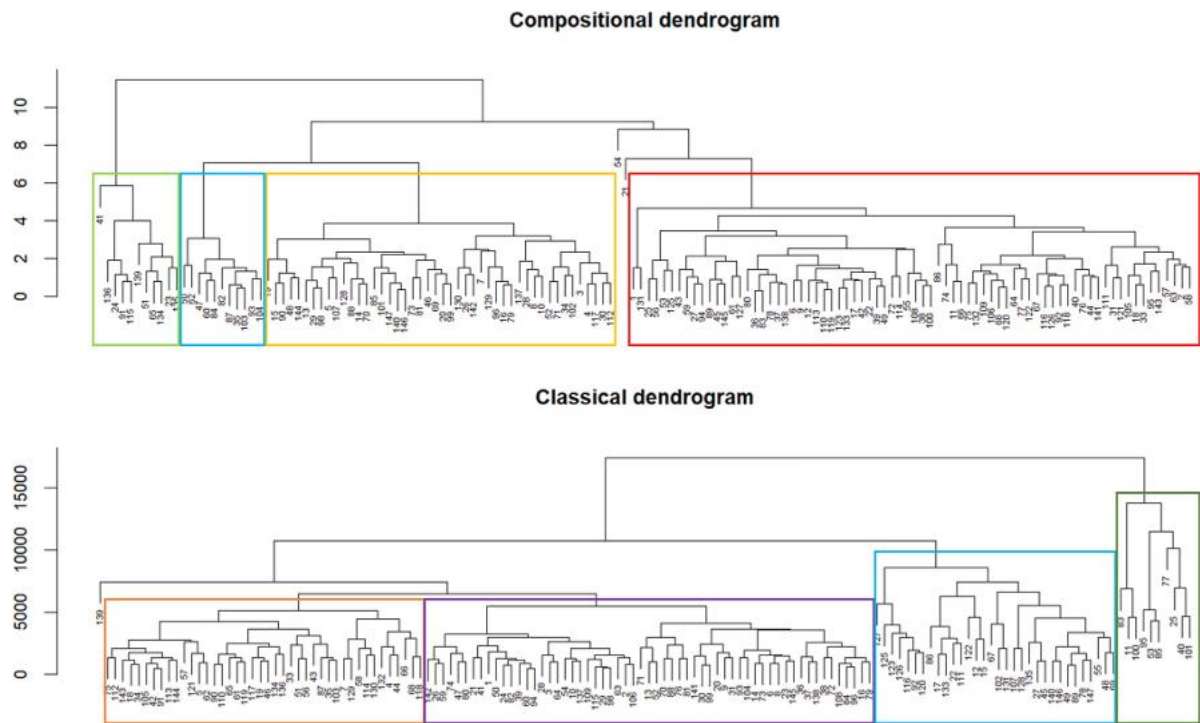


Fig. 11 Dendrograms resulting from hierarchical clustering of participants based on relative (*upper panel*) and absolute (*lower panel*) distributions of attention across AOIs in Task 5

left. The resulting dendrogram reveals a close proximity of person-related AOIs, which remain nearly proportional throughout the dataset. In contrast, the high variances of log-ratios involving Background or Furniture lead to their separation from the person-related AOIs in the clustering structure.

Finally, linear discriminant analysis (LDA) was performed to verify the ability of visual attention distribution to predict the task currently being solved. Although quadratic dis-

criminant analysis (QDA) has a higher potential to capture task-specific differences, its use would likely lead to overfitting due to the large number of estimated parameters relative to the number of available observations. The performance of LDA was evaluated using 10-fold cross-validation and reached an accuracy of 51.7%, which substantially exceeds the chance level. The corresponding confusion matrix is shown in Fig. 13 (left). The results indicate successful classification for Tasks 1, 3, 4, 6, and 7. In contrast, Tasks 2 and

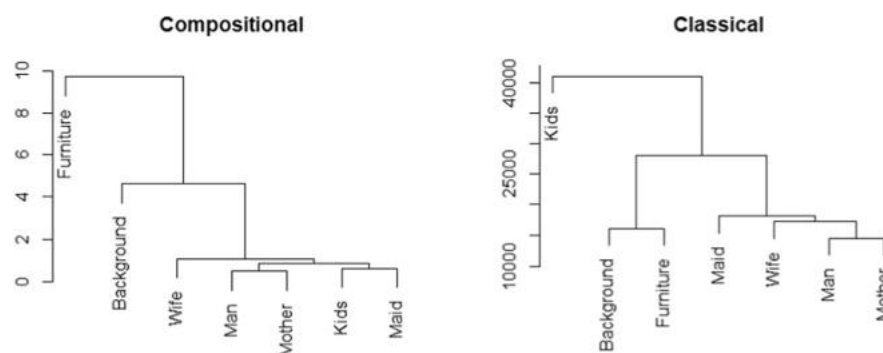


Fig. 12 Dendrograms resulting from hierarchical clustering of AOIs based on relative distributions (*left*) and absolute values (*right*) of TDoF, measured for Task 5

5 appear to be more difficult to distinguish. Attention distributions corresponding to Task 2 are frequently confused with those of Tasks 1 and 6, while observations belonging to Task 5 are often misclassified as Task 3. Nevertheless, these results are to some extent consistent with the patterns previously observed in the PCA and clustering analyses (see Figs. 8 and 6).

Analysis of total duration of fixations in AOIs - classical analysis

In addition to compositional analysis, valuable insights can also be gained from a classical analysis of the total fixation durations. When the primary focus is on the absolute fixation times within AOIs rather than on their relative structure, and when the total time allocated to each task is not restricted, a classical analysis of the directly observed data can be applied. Moreover, in the case of time-unlimited tasks, both absolute and relative approaches can be combined in a complementary manner, providing a more comprehensive description of the phenomenon under study.

The mean fixation times computed for each task and AOI are summarized in Fig. 14. These values are dominated by the mean time spent on the Background during Task 1 (free viewing). In contrast, the mean time devoted to Man in Task 7 is not particularly dominant on the absolute scale: although it represents the longest fixation time within Task 7, it is comparatively moderate when viewed across all tasks, due to the generally shorter observation times in the final task. It should be noted that on the relative scale, Man emerges as the most prominent AOI, accounting for more than 50 % of

observation time in Task 7 (see Fig. 6). Focusing on absolute information therefore leads to a different clustering structure, which is primarily shaped by the total time allocated to individual tasks.

The relationships between AOIs can be further examined using correlation matrices. Figure 15 compares the correlation structures of Tasks 5 and 7 (correlation matrices of the remaining tasks are provided in Fig. 20). As expected, most pairwise correlations are positive, reflecting the overall variability in total times spent across tasks. The correlation matrix of Task 5 closely resembles the corresponding variation matrix in Fig. 7: person-related AOIs are strongly correlated, consistent with their high proportionality, whereas correlations between Furniture and the remaining AOIs are typically low. This aligns with the high variation values observed for Furniture, indicating that its relationship with other AOIs is weak both in absolute and relative terms. By contrast, the correlation matrix of Task 7, similar to its variation counterpart, does not exhibit a clear structure. Nonetheless, relatively high correlations can be observed for Kids. At the same time, Fig. 7 also shows large variances of log-ratios involving this AOI. Thus, although time spent on Kids in Task 7 tends to increase together with times on other AOIs, the relative proportion of attention devoted to this AOI varies substantially across participants.

The results of PCA applied to the absolute values of TDoFs are presented in Fig. 16. In the classical setting, the first two principal components capture the directions of highest variability in the dataset. The first component primarily reflects the overall time spent on the individual tasks, with all loadings being positive and of approximately equal

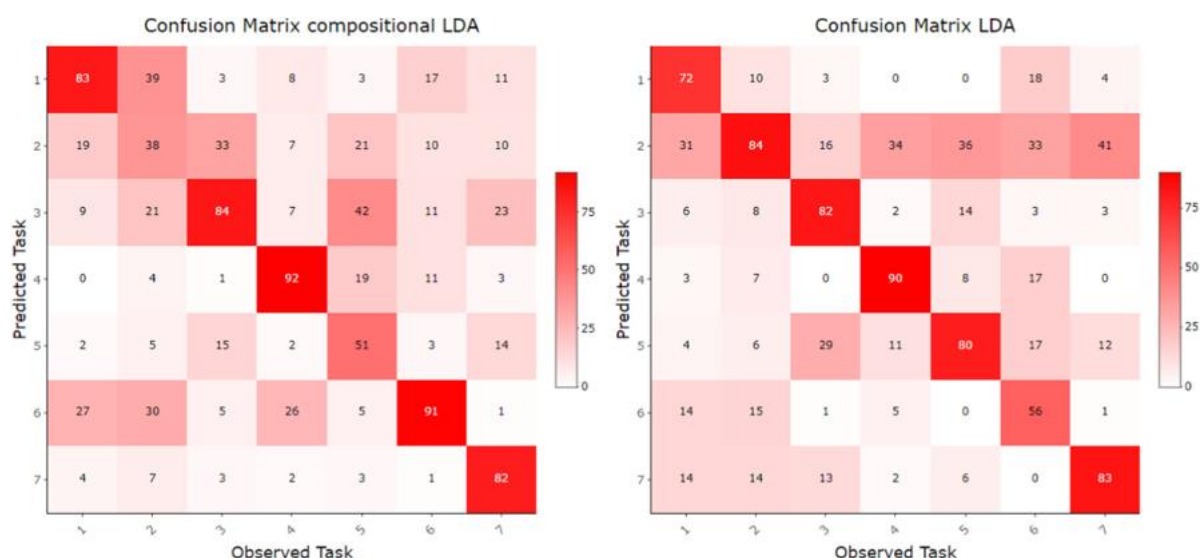


Fig. 13 Confusion matrices resulting from linear discriminant analysis based on relative distributions (*left*) and absolute values (*right*) of TDoF

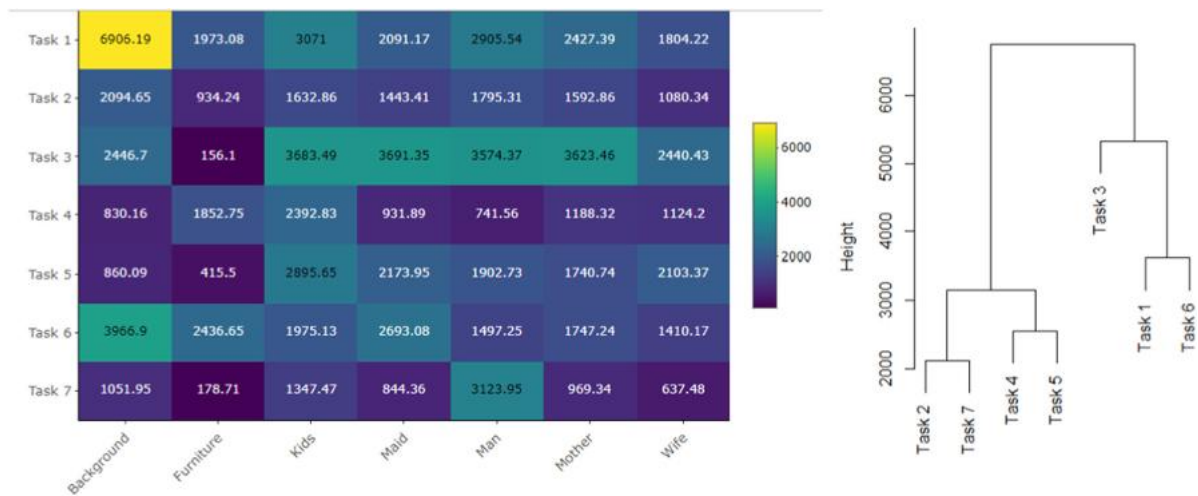


Fig. 14 Arithmetic means of TDoFs spent in each AOI, computed separately for each type of task (*left*) and clustering of tasks according to mean values of TDoFs (*right*)

magnitude; notably long fixation times are observed for Tasks 1 and 3. Task separation becomes apparent along the second component, which contrasts time allocated to Furniture and Background with that devoted to person-related AOIs. This component highlights Tasks 1, 4, and 6, characterized by extended viewing of Background and Furniture, while Task 3 appears on the opposite side with considerably shorter times for these two AOIs. The lengths of the rays indicate that the greatest variability is associated with Furniture and Background. These two, while positively correlated with each

other, are nearly uncorrelated with the person-related AOIs, which once again display strong mutual correlations. These findings are consistent with, and can be further refined by, the results of compositional PCA, which emphasizes the relative distribution of attention across AOIs.

The similarity of tasks with respect to total times spent on AOIs can be evaluated using classical MANOVA, which tests the null hypothesis that the mean AOI-specific TDoFs for each task are identical and equal to the overall means collected in Table 3. The differences between task-specific and

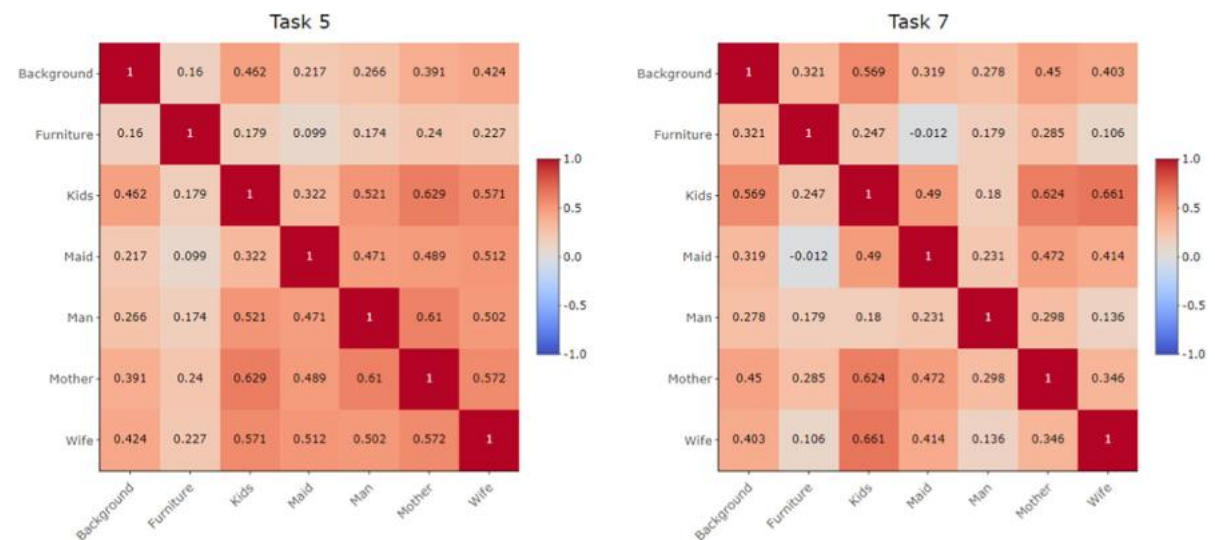


Fig. 15 Correlation matrices computed for TDoFs measured for Tasks 5 (*left*) and 7 (*right*)



Fig. 16 Classical biplot and loadings corresponding to the first two principal components, computed according to the absolute values of TDoFs. The percentages of explained variance are provided in the brackets. The ellipses enclose 75 % of the typical observations for each category

overall means are illustrated in Fig. 17. From the absolute perspective, Tasks 1 and 3 stand out with most mean values substantially exceeding the overall averages. In contrast, Task 7 is characterized by below-average times, with only one exception (AOI Man). The null hypothesis of equality is strongly rejected (with Pillai's trace = 1.24, $F = 37.06$, $df = (42, 6000)$, $p \ll 0.001$ and $\eta^2 = 0.206$), and pairwise post-hoc MANOVA tests confirm significant differences across all task pairs, with all Holm corrected p-values below 10^{-5} . These findings together with the results of compositional analysis indicate that the tasks differ not only in the total time spent on them, but also in the way this time is distributed across AOIs.

When modeling task-specific differences, for example, in the total time spent in AOI Man, the varying total solution time of participants must be taken into account as a random effect in a mixed-effects model. The resulting fixed effects, expressed as mean differences from the overall mean, are visualized in Fig. 10. These effects are similar to the differences shown in Fig. 17; however, the mixed-effects approach additionally provides inferential results demonstrating that all effects are significantly different from zero. While the longest time spent in AOI Man tends to be observed for Task 3 – participants allocate on average 1354.27 ms more to this AOI than their average across all tasks – the shortest time is

Table 3 Mean values of TDoFs computed from the complete dataset (in ms)

Background	Furniture	Kids	Maid	Man	Mother	Wife
2593.81	1135.29	2428.35	1981.32	2220.10	1898.48	1514.32

observed for Task 4, where the mean time spent in this AOI is 1478.54 ms below the overall average.

When participants are clustered within Task 5, the resulting hierarchical structure is shown in the lower plot of Fig. 11. Although the classical biplot in Fig. 16 does not reveal clearly separated groups within Task 5 (a more detailed visualization of PCA results for Task 5 are provided in Fig. 21), the clustering results can still be meaningfully related to the PCA visualization. The clustering is primarily driven by TDoFs in person-related AOIs. The purple group comprises participants with the shortest fixation times in these areas, the orange group represents participants with times around the average, the turquoise group corresponds to participants with above-average times, and the green group consists of participants with substantially longer fixation times.

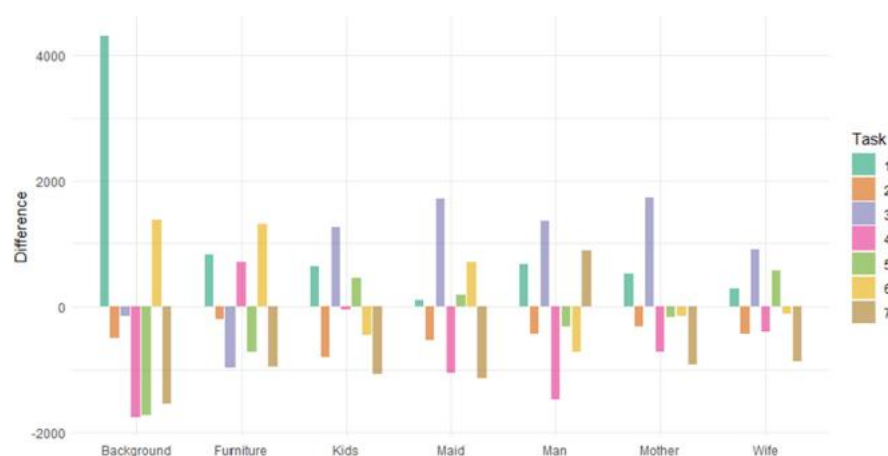
The focus on absolute fixation durations rather than their relative distribution also influences the clustering of AOIs. As shown in Fig. 12 (right), Background and Furniture as well as Man and Mother appear in close proximity, whereas fixation times allocated to Kids differ markedly from those of the remaining AOIs.

Finally, the linear discriminant analysis of TDoFs evaluated using 10-fold cross-validation resulted in a mean classification accuracy of tasks of 54.4%, which is comparable to the accuracy achieved when classifying relative distributions and remains substantially above the chance level. The main difference compared to the compositional approach appears at the level of individual tasks. The confusion matrix

shown in Fig. 13 (right) reveals that the most problematic task for classification was Task 6, which was correctly identified in only 38.9% of cases (56/144). This contrasts with the relative-valued approach, where Task 6 was correctly classified in 63.2% of cases (91/144). On the other hand, Tasks 2 and 5, which were difficult to distinguish based on relative TDoFs distributions, were classified more successfully when absolute fixation durations were used, reaching accuracies of 58.3% and 55.6%, respectively.

Comparison with the original findings of A. Yarbus

The results can be compared with the pioneering study of Yarbus (1967). In the original experiment, a single participant performed all seven tasks in the same order as in our study. Unlike in our case, where participants controlled the duration of stimulus presentation individually using the spacebar, Yarbus set a fixed observation time of three minutes. For scenarios with fixed viewing time, as well as for comparisons of multiple strategies with varying total solution times, the relative distribution of attention provides the most natural basis for comparison. Our comparison with Yarbus' findings is therefore grounded in the results reported in Section "Analysis of relative time distribution over AOI - compositional analysis" and Fig. 6. These are further contrasted with the outcomes of DeAngelus and Pelz (2009), who replicated Yarbus' experiment with 17 observers under unlimited viewing time and presented, on page 803, a visual comparison

**Fig. 17** Differences between group and total arithmetic means of TDoFs, computed for each task and AOI

of attention distributions for each of the seven experimental tasks.

A recent attempt to revisit Yarus' paradigm using contemporary technology was presented by Catrysse in Catrysse (2024). In that study, the classic "An Unexpected Visitor" task was replicated using webcam-based eye-tracking, enabling remote data collection with a larger and more diverse sample (53 participants, of whom usable data were obtained from 26). Despite the lower spatial accuracy of webcam tracking and substantial data loss, the results reproduced the key qualitative pattern originally reported by Yarus – namely, the strong influence of task instructions on gaze distribution.

Although A. Yarus primarily presented eye movement data in the form of graphical representations resembling scanpaths, the clusters of fixations nevertheless clearly identify the areas of highest interest. Most of our findings are consistent with the patterns reported by Yarus. Both approaches reveal strong attention to person-related AOIs in Tasks 3, 5, and 7. In our study, however, Task 7 additionally highlighted Man as a particularly prominent AOI, whereas Yarus' observer distributed attention more evenly across Man, Mother, and Kids. Our results for these tasks are also in agreement with the findings of DeAngelus and Pelz (2009).

The original experiment also indicated more dispersed attention in Tasks 1, 2, 4, and 6, where a substantial portion of fixations was directed at regions outside of person-related AOIs. This pattern is reflected in our data by increased time spent on Background and Furniture. Particularly in Tasks 1 and 6, the scanpaths span almost the entire painting, and the area defined in our study as Background is well covered. In our experiment, participants spent on average 33 % and 24 % of their observation time on this AOI in Tasks 1 and 6, respectively. Interestingly, in DeAngelus and Pelz (2009), participants devoted almost no attention to Background in Task 1.

Task 4 is another notable case, characterized by increased attention to Furniture, especially the tabletop. In DeAngelus and Pelz (2009), this AOI accounted for approximately 50 % of the observation time. At first glance, this contrasts with our findings, which assign only about 15 % to Furniture, while Kids attracted the highest proportion (around 30 %). However, this discrepancy may stem from the spatial proximity of AOI Kids and the tabletop, which in our study was categorized under Furniture.

Another perspective on A. Yarus's study was provided by the works of Greene et al. (2012); Borji and Itti (2014), and Haji-Abolhassani and Clark (2014), which investigated the inverse problem of inferring the observer's task from eye-movement data obtained in Yarus-like experiments. Their conclusions, however, were partly contradictory. While the model proposed in Greene et al. (2012) generally failed to reliably predict the performed task from eye-tracking pat-

terns, Borji and Itti (2014) reanalyzed the same dataset using an alternative set of features and classification methods and demonstrated that task prediction can exceed chance-level performance. Subsequently, Haji-Abolhassani and Clark (2014) repeated the experiment of Greene et al. (2012) and showed that the observer's task can be successfully inferred using a Hidden Markov Model applied to scanpath dynamics.

In this context, our results provide additional evidence that visual attention patterns contain information relevant to the performed task. Using linear discriminant analysis applied to both compositional and absolute representations of TDOFs, we achieved classification accuracies clearly exceeding the chance level. Although the obtained performance is comparable to previously reported studies, it should be emphasized that identifying an optimal classification framework was not the primary objective of this work. Rather, the classification analysis serves mainly as an illustrative demonstration of the information preserved in the proposed compositional representation of eye-tracking data. Despite employing only a relatively simple classifier in the form of LDA, the achieved results can be considered promising and suggest that further improvements could likely be obtained through a more systematic exploration of advanced modeling strategies. Importantly, the newly collected dataset used in this study represents a substantial extension compared to datasets analyzed in Greene et al. (2012) and Haji-Abolhassani and Clark (2014), as it includes observations from a considerably larger number of participants. This richer experimental basis therefore provides favorable conditions for future research focusing specifically on task prediction from eye-movement behavior.

Finally, both Yarus (1967) and DeAngelus and Pelz (2009) reported the occurrence of cyclic eye movement patterns, where observers in the later stages of a task repeated sequences of visited AOIs. Although this phenomenon was not explicitly addressed in our study, the dataset would allow for its verification, for instance, using recurrence analysis (Anderson et al., 2013). On the other hand, the compositional methodology introduced here provides a framework for a more detailed investigation of the phenomenon, making it possible to apply PCA, clustering, and a broad spectrum of multivariate statistical methods as described, for example, in Filzmoser et al. (2018) or Pawlowsky-Glahn and Buccianti (2011).

Conclusion

The presented study demonstrates that compositional data analysis provides a coherent and informative framework for analyzing AOI-based eye-tracking data. Treating total fixation durations as compositions rather than as inde-

pendent absolute measures allows to separate individual differences in overall observing pace from genuine differences in attentional strategy. In the context of a large-scale replication of Yarbus' "Unexpected Visitor" paradigm, this approach recovers the classic task-dependent modulation of gaze behavior and reveals additional structure that remains obscured under classical analysis.

Compositional techniques applied to AOI-wise fixation times highlight several robust regularities. Person-related AOIs form a remarkably stable group, whose mutual relationships are close to proportional across tasks and participants, whereas contextual AOIs, such as background and objects, exhibit much higher variability. Compositional PCA, variation matrices, and clustering collectively show that tasks can be grouped according to their cognitive demands rather than merely by their overall viewing duration. In contrast, a traditional analysis based on absolute fixation times emphasizes differences in total observation time and yields a different task structure, driven primarily by how long participants viewed the stimulus rather than by how they distributed their attention.

The problem can also be viewed conversely: linear discriminant analysis demonstrated the possibility of inferring the task being performed from the eye-movement record. Although both the compositional and classical approaches resulted in classification accuracies above the chance level, their performance varied across tasks. These findings have broader methodological implications for eye-tracking research. Whenever AOI-based metrics reflect relative allocation of attention (e.g., proportions of viewing time or fixation counts across regions), or when the experimental design involves a fixed total viewing time, compositional modeling offers an assumption-consistent alternative to standard multivariate methods that ignore the constrained, relative nature of the data. Additionally, even when viewing time is not limited, as in the present study, the compositional approach can complement classical analyses, enabling a more comprehensive examination of the problem from both relative and absolute perspectives. The techniques introduced here, including descriptive CoDA summaries, log-ratio-based PCA, clustering, compositional MANOVA, regression modeling, and discriminant analysis, can be readily applied to other eye-tracking paradigms in which the structure of gaze distributions is of primary interest. At the same time, several limitations should be acknowledged. We

selected Yarbus study to illustrate the compositional methodology, as it is a historically relevant cornerstone in visual attention research. However, the Yarbus paradigm may not provide a clear normative model specifying where and for how long participants should look, therefore the evaluation against a defined ground truth can be limited. We expect the compositional approach to be even more advantageous in domains with well-established effects such as lexical frequency in reading or saliency in images, where its statistical benefits can be assessed more directly. Additionally, the presented analysis focuses on a single complex scene and on a reduced set of seven AOIs. Future work may extend this approach to more diverse stimuli, dynamic or sequential representations of gaze behavior, and more advanced compositional techniques, including robust (Filzmoser & Hron, 2008, 2011) and functional (Pavlu et al., 2023) CoDA and explicit combinations with methods such as recurrence quantification analysis or hidden Markov models. Such extensions would further refine the understanding of how task demands, stimulus properties, and individual differences jointly shape visual exploration. Compositional modeling emerges as a natural and powerful addition to the methodological toolbox of eye-tracking research. By respecting the relative nature of AOI-based gaze measures, CoDA yields more nuanced and interpretable insights into task-dependent visual attention, providing a principled basis for comparing strategies within and across classic paradigms, such as Yarbus' "Unexpected Visitor" experiment.

Supplementary information

The full dataset and annotated R code for this analysis are available on GitHub: <https://github.com/kfacevicova/Advancing-Eye-Movement-Analysis-Through-Compositional-Modeling>.

Appendix A Supplementary material for the compositional analysis

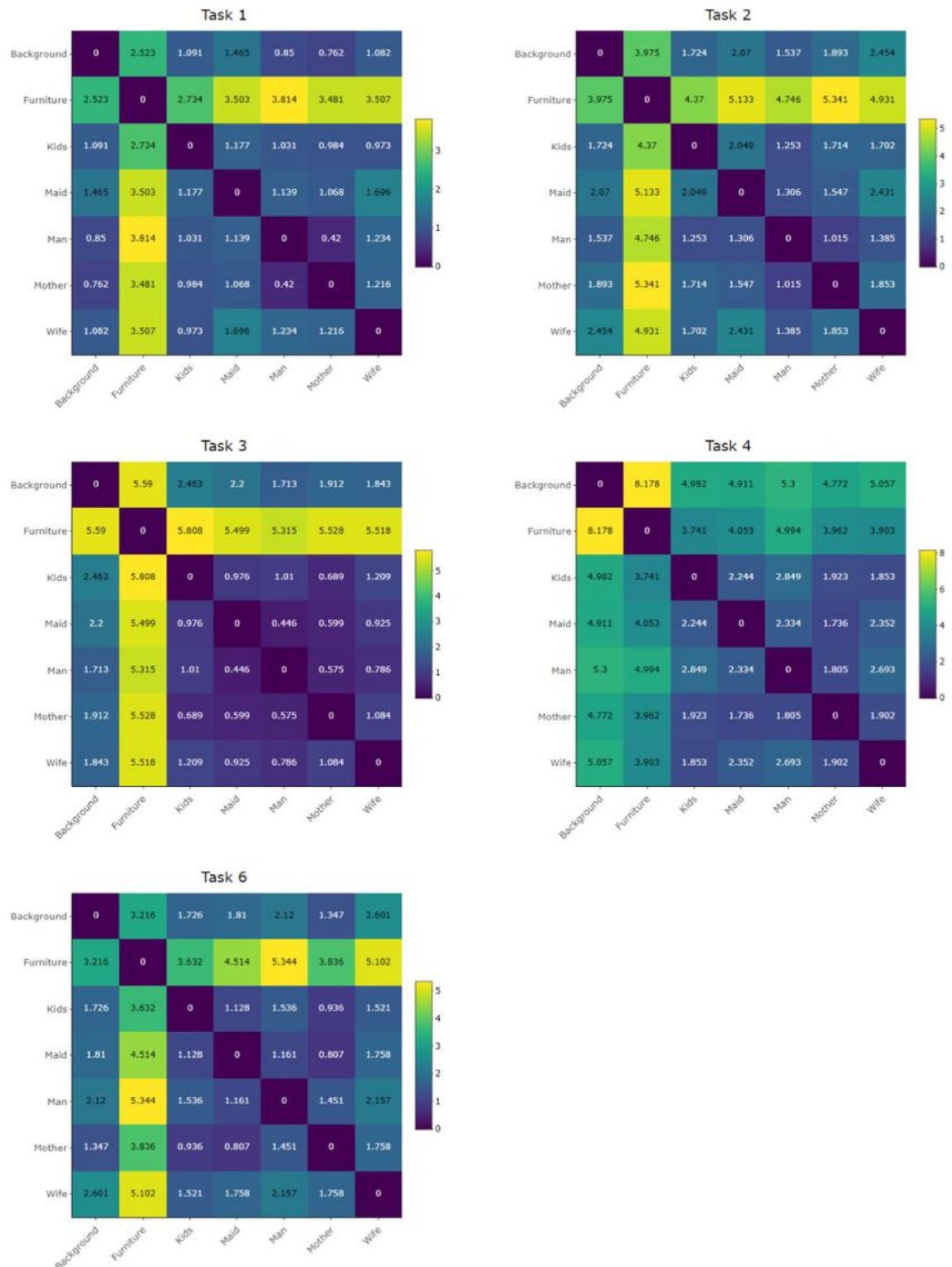


Fig. 18 Variation matrices, defined by Eq. 2, computed for tasks not displayed in Fig. 7. Each cell corresponds to the sample variance of log-ratio between the respective pair of AOIs

Table 4 Total variances, defined by Eq. 3, computed for each task

Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7
5.11	7.76	7.38	10.79	10.03	7.07	17.62

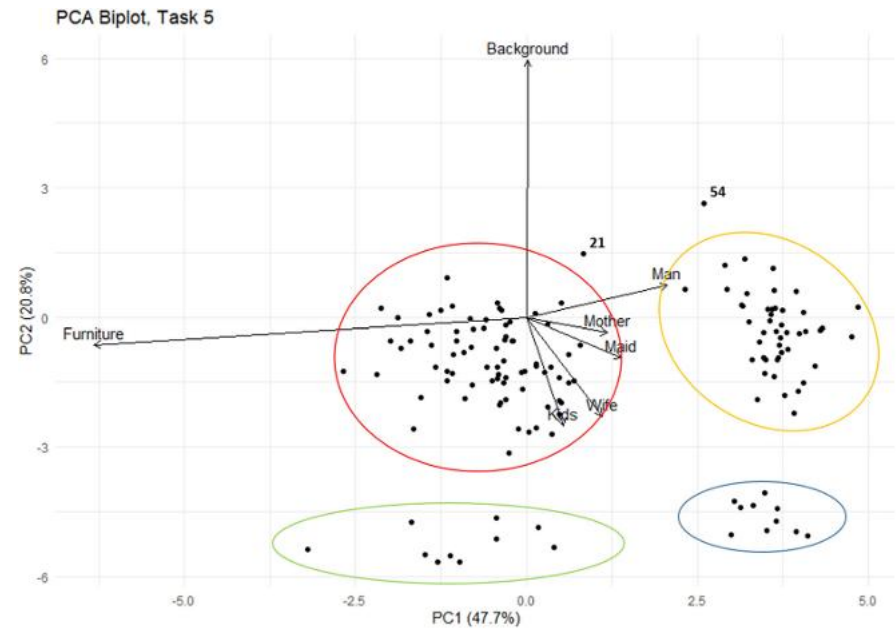


Fig. 19 Compositional biplot (Fig. 8) restricted to observations from Task 5. The *colored ellipses* illustrate the typical locations of the groups defined by hierarchical clustering in the upper part of Fig. 11

Appendix B Supplementary material for the classical analysis

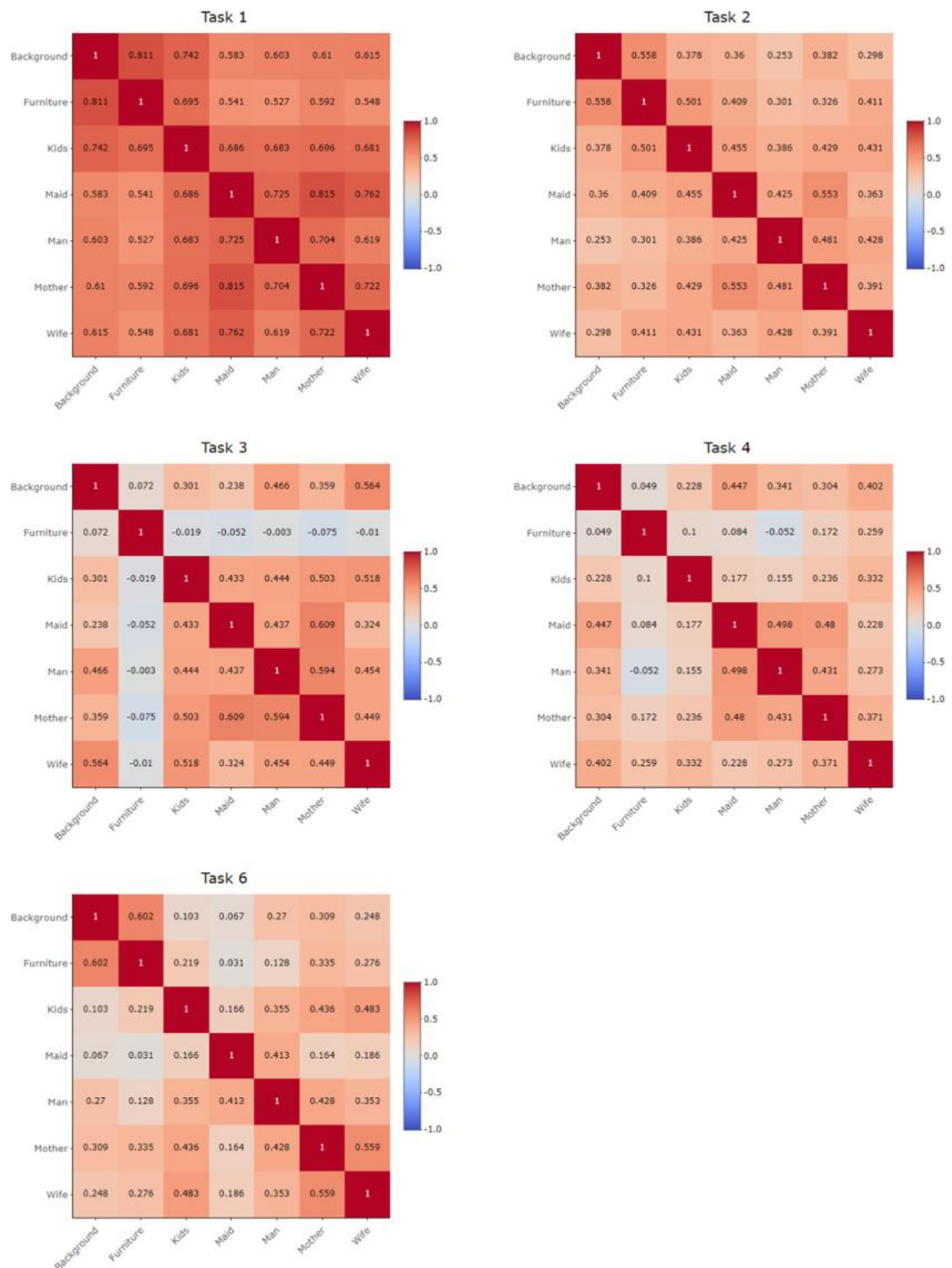


Fig. 20 Correlation matrices computed for tasks not displayed in Fig. 15

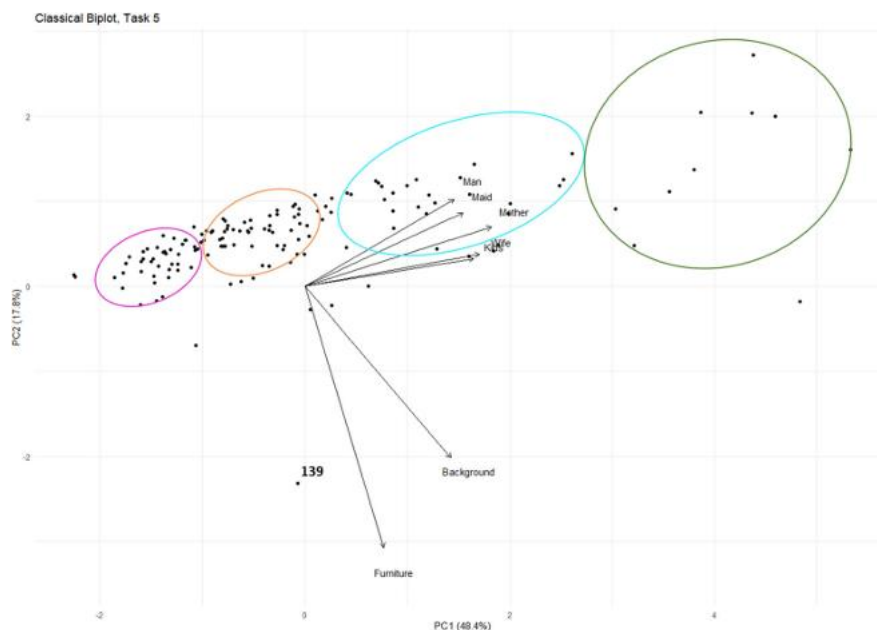


Fig. 21 PCA biplot (Fig. 16) restricted to observations from Task 5. The colored ellipses illustrate the typical locations of the groups defined by hierarchical clustering in the lower part of Fig. 11

Acknowledgements The work was supported by the Czech Science Foundation, Project 23-06187S.

Author Contributions Not applicable.

Funding Open access publishing supported by the institutions participating in the CzechELib Transformative Agreement. The work was supported by the Czech Science Foundation, Project 23-06187S.

Data Availability The complete dataset is available at <https://github.com/kfacevicova/Advancing-Eye-Movement-Analysis-Through-Compositional-Modeling>.

Materials availability Not applicable.

Code availability The R code is available at <https://github.com/kfacevicova/Advancing-Eye-Movement-Analysis-Through-Compositional-Modeling>.

Declarations

Conflicts of Interest/Competing Interests The authors declare that they have no competing interests.

Ethics approval and consent to participate The study was conducted in accordance with the ethical standards of the Faculty of Science, Palacký University Olomouc. It was approved by the institutional Ethics Committee (reference no. 22-01), and written informed consent was obtained from all participants prior to the experiment.

Consent for publication Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of Royal Statistical Society, Series B, Methodological*, 44(2), 139–177.
- Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London: Chapman & Hall.
- Aitchison, J., & Bacon-Shone, J. (1984). Log contrast models for experiments with mixtures. *Biometrika*, 71(2), 323–330. <https://doi.org/10.2307/2336249>
- Altmann, G. T. M., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, 57(4), 502–518. <https://doi.org/10.1016/j.jml.2006.12.004>
- Anderson, T. W., Olkin, I., & Fang, K. T. (Eds.). (1994). *Multivariate Analysis and Its Applications* (Vol. 24). Hayward, CA (USA): Institute of Mathematical Statistics.

- Anderson, N. C., Bischof, W. F., Laidlaw, K. E. W., Risko, E. F., & Kingstone, A. (2013). Recurrence quantification analysis of eye movements. *Behavior Research Methods*, 45(3), 842–856. <https://doi.org/10.3758/s13428-012-0299-5>
- Ballard, D. H. (1991). Animate vision. *Artificial Intelligence*, 48(1), 57–86. [https://doi.org/10.1016/0004-3702\(91\)90080-4](https://doi.org/10.1016/0004-3702(91)90080-4)
- Ballard, D. H., Hayhoe, M. M., Li, F., Whitehead, S. D., Frisby, J. P., Taylor, J. G., & Fisher, R. B. (1992). Hand-eye coordination during sequential tasks. *Philosophical Transactions: Biological Sciences*, 337(1281), 331–339. <https://doi.org/10.1098/rstb.1992.0111>
- Boogaart van den, K. G., Tolosana, R., & Bren, M. (2009). compositions: Compositional Data Analysis. R package version 1.02-1. <http://www.stat.boogaart.de/compositions>
- Borji, A., & Itti, L. (2014). Defending Yarbus: Eye movements reveal observers' task. *Journal of Vision*, 14(3), 1–22. <https://doi.org/10.1167/14.3.29>
- Catrysse, L. (2024). Yarbus revisited: The webcam edition. In *Proceedings of the 22nd European Conference on Eye Movements (ECEM 2024)*. Open Universiteit.
- Chuk, T., Chan, A. B., & Hsiao, J. H. (2014). Understanding eye movements in face recognition using hidden markov models. *Journal of Vision*, 14(11), 1–14. <https://doi.org/10.1167/14.11.8>
- Coenders, G., Egozcue, J. J., Fačevićová, K., Navarro-López, C., Palarea-Albaladejo, J., Pawlowsky-Glahn, V., & Tolosana-Delgado, R. (2023). 40 years after Aitchison's article "The statistical analysis of compositional data". Where we are and where we are heading. *SORT-Statistics and Operations Research Transactions*, 47(2), 207–228. <https://doi.org/10.57645/20.8080.02.6>
- DeAngelus, M., & Pelz, J. B. (2009). Top-down control of eye movements: Yarbus revisited. *Visual Cognition*, 17(6–7), 790–811. <https://doi.org/10.1080/13506280902793843>
- Duchowski, A. T. (2017). *Eye Tracking Methodology: Theory and Practice* (3rd Ed.). Springer, London. <https://doi.org/10.1007/978-3-319-57883-5>
- Egozcue, J. J., & Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37(7), 795–828. <https://doi.org/10.1007/s11004-005-7381-9>
- Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., & Barcelo-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35, 279–300. <https://doi.org/10.1023/A:1023818214614>
- Ehrenstein, W. H., Arnold-Schulz-Gahmen, B. E., & Jaschinski, W. (2005). Eye preference within the context of binocular functions. *Graefes' Archive for Clinical and Experimental Ophthalmology*, 243(9), 926–932. <https://doi.org/10.1007/s00417-005-1128-7>
- Fačevićová, K., Bábek, O., Hron, K., & Kumpan, T. (2016a). Element chemostratigraphy of the Devonian/Carboniferous boundary – a compositional approach. *Applied Geochemistry*, 75, 211–221. <https://doi.org/10.1016/j.apgeochem.2016.10.002>
- Fačevićová, K., Hron, K., Todorov, V., & Templ, M. (2016b). Compositional tables analysis in coordinates. *Scandinavian Journal of Statistics*, 43(4), 962–977. <https://doi.org/10.1111/sjos.12223>
- Fačevićová, K., Filzmoser, P., & Hron, K. (2022). Compositional cubes: A new concept for multi-factorial compositions. *Statistical Papers*, 64(3), 955–985. <https://doi.org/10.1007/s00362-022-01350-8>
- Fačevićová, K., Filzmoser, P., & Hron, K. (2025). Correspondence analysis from the viewpoint of compositional tables. *Statistical Analysis and Data Mining*, 18(4). <https://doi.org/10.1002/sam.70023>
- Fairclough, S. J., Dumuid, D., Taylor, S., Curry, W., McGrane, B., Stratton, G., Maher, C., & Olds, T. (2017). Fitness, fatness and the reallocation of time between children's daily movement behaviours: An analysis of compositional data. *International Journal of Behavioral Nutrition and Physical Activity*, 14(64), 1–12. <https://doi.org/10.1186/s12966-017-0521-z>
- Filzmoser, P., & Hron, K. (2008). Outlier detection for compositional data using robust methods. *Mathematical Geosciences*, 40(3), 233–248. <https://doi.org/10.1007/s11004-007-9141-5>
- Filzmoser, P., & Hron, K. (2011). Robust statistical analysis. In V. Pawlowsky-Glahn & A. Buccianti (Eds.), *Compositional Data Analysis: Theory and Applications* (pp. 59–72). Chichester: John Wiley & Sons.
- Filzmoser, P., Hron, K., & Templ, M. (2018). *Applied Compositional Data Analysis*. Springer, Cham. <https://doi.org/10.1007/978-3-319-96422-5>
- Findlay, J. M., & Gilchrist, I. D. (2003). *Active Vision: The Psychology of Looking and Seeing*. Oxford University Press, Oxford. <https://doi.org/10.1093/acprof:oso/9780198524793.001.0001>
- Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome datasets are compositional: And this is not optional. *Frontiers in Microbiology*, 8(2224). <https://doi.org/10.3389/fmicb.2017.02224>
- Greene, M. R., Liu, T., & Wolfe, J. M. (2012). Reconsidering Yarbus: A failure to predict observers' task from eye movement patterns. *Vision Research*, 62, 1–8. <https://doi.org/10.1016/j.visres.2012.03.019>
- Haji-Abolhassani, A., & Clark, J. J. (2014). An inverse Yarbus process: Predicting observers' task from eye movement patterns. *Vision Research*, 103, 127–142. <https://doi.org/10.1016/j.visres.2014.08.014>
- Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Halszka, J., & Weijer, J. (2011). *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford: Oxford University Press.
- Hron, K., Templ, M., & Filzmoser, P. (2010). Imputation of missing values for compositional data using classical and robust methods. *Computational Statistics and Data Analysis*, 54(12), 3095–3107. <https://doi.org/10.1016/j.csda.2009.11.023>
- Hron, K., Filzmoser, P., Caritat, P., Fišerová, E., & Gardlo, A. (2017). Weighted pivot coordinates for compositional data and their application to geochemical mapping. *Mathematical Geosciences*, 49, 797–814. <https://doi.org/10.1007/s11004-017-9684-z>
- Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3), 194–203. <https://doi.org/10.1038/35058500>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning*. Springer, New York, NY (USA). <https://doi.org/10.1007/978-1-0716-1418-1>
- Kouřil, S., de Sousa, J., Fačevićová, K., Gardlo, A., Muehlmann, C., Nordhausen, K., Friedecký, D., & Adam, T. (2023). Multivariate independent component analysis identifies patients in newborn screening equally to adjusted reference ranges. *International Journal of Neonatal Screening*, 9(4). <https://doi.org/10.3390/ijns9040060>
- Krejtz, K., Çöltekin, A., Duchowski, A. T., & Niedzielska, A. (2017). Using coefficient k to distinguish ambient/focal visual attention during cartographic tasks. *Journal of Eye Movement Research*, 10(2), 1–13. <https://doi.org/10.16910/jemr.10.2.3>
- Martín-Fernández, J.-A., Daunis-i-Estadella, J., & Mateu-Figueras, G. (2015). On the interpretation of differences between groups for compositional data. *SORT-Statistics and Operations Research Transactions*, 39(2), 231–252.
- Muller, K. E., Lavange, L. M., Ramey, S. L., & Ramey, C. T. (1992). Power calculations for general linear multivariate models including repeated measures applications. *Journal of American Statistical Association*, 87(420), 1209–1226. <https://doi.org/10.1080/01621459.1992.10476281>
- Palarea-Albaladejo, J., & Martín-Fernández, J. A. (2008). A modified EM algorithm for replacing rounded zeros in compositional data sets. *Computers and Geosciences*, 34(8), 2233–2251. <https://doi.org/10.1016/j.cageo.2007.09.015>

- Palarea-Albaladejo, J., Martín-Fernández, J. A., & Gómez-García, J. A. (2007). Parametric approach for dealing with compositional rounded zeros. *Mathematical Geology*, 39(7), 625–645. <https://doi.org/10.1007/s11004-007-9100-1>
- Pavlů, I., Menafoglio, A., Bongiorno, E.G., & Hron, K. (2023). Classification of probability density functions in the framework of bayes spaces: methods and applications. *SORT-Statistics and Operations Research Transactions*, 47(2), 295–322. <https://doi.org/10.57645/20.8080.02.10>
- Pawlowsky-Glahn, V., & Buccianti, A.E. (2011). *Compositional Data Analysis*. John Wiley & Sons, Ltd, Chichester, UK. <https://doi.org/10.1002/9781119976462>
- Pavlů, I., Menafoglio, V., Egozcue, J.J., & Tolosana-Delgado, R. (2007). Modeling and Analysis of Compositional Data. Wiley, Chichester. <https://doi.org/10.1002/9781119003144>
- Pawlowsky-Glahn, V., Egozcue, J. J., & Lovell, D. (2015). Tools for compositional data with a total. *Statistical Modelling*, 15(2), 175–190. <https://doi.org/10.1177/1471082X14535526>
- Pearson, K. (1897). Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. In *Proceedings of the Royal Society of London*, 60
- R Development Core Team. (2009). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. R Foundation for Statistical Computing. <http://www.r-project.org>
- Tatler, B.W., Wade, N.J., Kwan, H., Findlay, J.M., & Velichkovsky, B.M. (2010). Yarbus, eye movements, and vision. *i-Perception*1(1), 7–27. <https://doi.org/10.1068/i0382>
- Templ, M., Hron, K., & Filzmoser, P. (2011). robCompositions: An R-package for robust statistical analysis of compositional data. In V. Pawlowsky-Glahn, & A. Buccianti (Eds.) *Compositional Data Analysis: Theory and Applications*, Chichester, pp. 341–355.
- Tolosana-Delgado, R. (2012). Uses and misuses of compositional data in sedimentology. *Sedimentary Geology*, 280, 60–79. <https://doi.org/10.1016/j.sedgeo.2012.05.005>
- Vencálek, O., Hron, K., & Filzmoser, P. (2020). A comparison of generalised linear models and compositional models for ordered categorical data. *Statistical Modelling*, 20(3), 249–273. <https://doi.org/10.1177/1471082X18816540>
- Vojtechovska, M., & Popelka, S. (2026). Gazeplotter: An open-source solution for the automatic generation of scarf plots from eye-tracking data. *Behavior Research Methods*, 58(85). <https://doi.org/10.3758/s13428-026-02959-5>
- Yarbus, A.L. (1967). *Eye Movements and Vision*. Springer, New York. <https://doi.org/10.1007/978-1-4899-5379-7>
- Ylitalo, A. K., Särkkä, A., & Guttorp, P. (2016). What we look at in paintings: A comparison between experienced and inexperienced art viewers. *The Annals of Applied Statistics*, 10(2), 549–574. <https://doi.org/10.1214/16-AOAS921>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.