REGULAR ARTICLE

Compositional cubes: a new concept for multi-factorial compositions

Kamila Fačevicová[1] · Peter Filzmoser[2] · Karel Hron[1]

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

🖄 Springer

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 (Fačevicová 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) $\mathbf{x} = (x_1, \ldots, x_D)'$, 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 $\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, x_2, \ldots, x_D)' \mid x_i > 0, \forall i, \sum_i x_i = \kappa \right\}$. 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,

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, \ldots, x_D y_D)', \ \alpha \odot \mathbf{x} = \mathcal{C}\left(x_1^\alpha, \ldots, x_D^\alpha\right)',$$

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

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 $\mathbf{z} = (z_1, \ldots, z_{D-1})' \in \mathbb{R}^{D-1}$ are obtained as

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

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 $\xi_1 \ln x_1 + \cdots + \xi_D \ln x_D$ 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 $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iD})'$ 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 with 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

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

or

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

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

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

where $g(.)$ stands for the geometric mean, and the values $i_1, \ldots, i_{r_i}$ and $i'_1, \ldots, i'_{s_i}$ 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 Fačevicová 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

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

where s_i and t_i correspond to the respective step of the SBP performed on the levels of the row factor. The indices $(i_1 \bullet, \ldots, i_{s_i} \bullet)$ and $(i'_1 \bullet, \ldots, i'_{t_i} \bullet)$ 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

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

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

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

for $i = 1, 2, \ldots, I-1$ and $j = 1, 2, \ldots, J-1$, where $A_{ij}, \ldots, D_{ij}$ are indices of parts in each group defined by the i-th and j-th step of row and column SBP, respectively, and $|A_{ij}|, \ldots, |D_{ij}|$ 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

$$
\mathbf{x} = \begin{pmatrix} x_{111} & \cdots & x_{1J1} \\ \vdots & \ddots & \vdots \\ x_{I11} & \cdots & x_{IJ1} \end{pmatrix} \cdots \begin{vmatrix} x_{11K} & \cdots & x_{1JK} \\ \vdots & \ddots & \vdots \\ x_{I1K} & \cdots & x_{IJK} \end{vmatrix} , \tag{9}
$$

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

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

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 Pawlowsky-Glahn and Egozcue 2001) modify to

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

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

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

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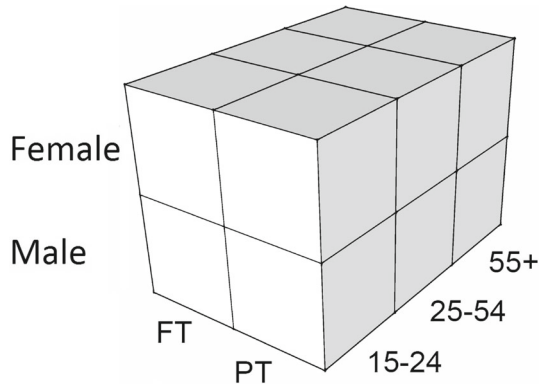
$$x_{ijk}^{\text{ind}} = g(\mathbf{x}_{i\bullet\bullet})g(\mathbf{x}_{\bullet j\bullet})g(\mathbf{x}_{\bullet\bullet k}), \tag{13}$$

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

$$\mathbf{x}^{\text{int}} = \mathbf{x} \ominus \mathbf{x}^{\text{ind}}. \tag{14}$$

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

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

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 $\mathbf{int}^{rc}(\mathbf{x})$, the row-slice interaction cube $\mathbf{int}^{rs}(\mathbf{x})$ has parts

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

Similar to the case of $\mathbf{int}^{rc}(\mathbf{x})$, 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 $\mathbf{int}^{cs}(\mathbf{x})$ with cells

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

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

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

with parts

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

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 $\mathbf{int}^{rc}(\mathbf{x})$ 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 $\mathbf{int}^{rcs}(\mathbf{x})$, which involves also the effect of age, which was suppressed in $\mathbf{int}^{rc}(\mathbf{x})$. 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 $IJK - 1$ 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 $IJK - 1$ 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

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

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 $\boldsymbol{\xi}_i^{\text{r}}$ (of length IJK) 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 $\xi_{i-}^{\mathrm{r}} = -\sqrt{\frac{s_i}{t_i(s_i+t_i)JK}}$ 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,

$$z_i^{\mathrm{r}} = \sqrt{\frac{s_i t_i JK}{s_i + t_i}} \ln \frac{\left[g(\mathbf{x}_{i_1\bullet\bullet}) \cdots g(\mathbf{x}_{i_{s_i}\bullet\bullet})\right]^{1/s_i}}{\left[g(\mathbf{x}_{i_1'\bullet\bullet}) \cdots g(x_{i_{t_i}'\bullet\bullet})\right]^{1/t_i}}, \quad \text{for} \quad i = 1, 2, \ldots, I - 1. \quad (21)$$

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 $\boldsymbol{\xi}_j^{\mathrm{c}}$ with entries $\xi_{j+}^{\mathrm{c}} = \sqrt{\frac{v_j}{u_j(u_j+v_j)IK}}$, $\xi_{j-}^{\mathrm{c}} = -\sqrt{\frac{u_j}{v_j(u_j+v_j)IK}}$ 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

$$z_j^{\mathrm{c}} = \sqrt{\frac{u_j v_j IK}{u_j + v_j}} \ln \frac{\left[g(\mathbf{x}_{\bullet j_1\bullet}) \cdots g(\mathbf{x}_{\bullet j_{u_j}\bullet})\right]^{1/u_j}}{\left[g(\mathbf{x}_{\bullet j_1'\bullet}) \cdots g(\mathbf{x}_{\bullet j_{v_j}'\bullet})\right]^{1/v_j}}, \quad \text{for} \quad j = 1, 2, \ldots, J - 1, (22)$$

which form the second group in the coordinate system representing the whole compositional cube $\mathbf{x}$ as well as its independent part $\mathbf{x}^{\mathrm{ind}}$. 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 $\boldsymbol{\xi}_k^{\mathrm{s}}$ with entries $\xi_{k+}^{\mathrm{s}} = \sqrt{\frac{n_k}{m_k(m_k+n_k)IJ}}$, $\xi_{k-}^{\mathrm{s}} = -\sqrt{\frac{m_k}{n_k(m_k+n_k)IJ}}$ 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

$$z_k^{\mathrm{s}} = \sqrt{\frac{m_k n_k IJ}{m_k + n_k}} \ln \frac{\left[g(\mathbf{x}_{\bullet\bullet k_1}) \cdots g(\mathbf{x}_{\bullet\bullet k_{m_k}})\right]^{1/m_k}}{\left[g(\mathbf{x}_{\bullet\bullet k_1'}) \cdots g(\mathbf{x}_{\bullet\bullet k_{n_k}'})\right]^{1/n_k}}, \quad \text{for} \quad k = 1, 2, \ldots, K - 1. \quad (23)$$

Note here that the row, column and slice balances form a complete coordinate representation of $\mathbf{x}^{\mathrm{ind}}$, 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 $\mathbf{x}^{\mathrm{ind}}$ (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 $\mathbf{x}^{\mathrm{int}}$. For this purpose the normalized Hadamard (entry wise) product ($\circ$) of the vectors $\boldsymbol{\xi}_i^{\mathrm{r}}$, $\boldsymbol{\xi}_j^{\mathrm{c}}$ and $\boldsymbol{\xi}_k^{\mathrm{s}}$ and Eq. (3) is used. The vectors $\boldsymbol{\xi}_{ij}^{\mathrm{rc}} = \boldsymbol{\xi}_i^{\mathrm{r}} \circ \boldsymbol{\xi}_j^{\mathrm{c}}$, $i = 1, \ldots, I - 1, j = 1, \ldots, J - 1$ determine $(I - 1)(J - 1)$ coordinates of type

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

for $i = 1, 2, \ldots, I-1$ and $j = 1, 2, \ldots, J-1$, 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, $i = 1, \ldots, I-1, k = 1, \ldots, K-1$, leads to coordinates

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

for $i = 1, 2, \ldots, I-1$ and $k = 1, 2, \ldots, K-1$, 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 $\boldsymbol{\xi}_j^{\text{c}}$ and $\boldsymbol{\xi}_k^{\text{s}}$, $j = 1, \ldots, J-1, k = 1, \ldots, K-1$ lead to vectors $\boldsymbol{\xi}_{jk}^{\text{cs}}$ and coordinates

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

for $j = 1, 2, \ldots, J-1$ and $k = 1, 2, \ldots, K-1$. 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, $\boldsymbol{\xi}_i^{\text{r}}, \boldsymbol{\xi}_j^{\text{c}}$ and $\boldsymbol{\xi}_k^{\text{s}}$, $i = 1, \ldots, I-1, j = 1, \ldots, J-1, k = 1, \ldots, K-1$, and have the general form

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

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

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

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 **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 $\mathbf{z}^{rc}$ of the interaction cube $\mathbf{int}^{rc}(\mathbf{x})$. More specifically, e.g. the cube $\mathbf{int}^{rc}(\mathbf{x})$ 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 **z** of the original compositional cube the following relation holds,

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

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 **T**, whose rows are formed by coefficients of the respective logarithmized parts of vec(**x**) in the desirable log-ratios. According to (4), the vectorized form of a compositional cube vec(**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
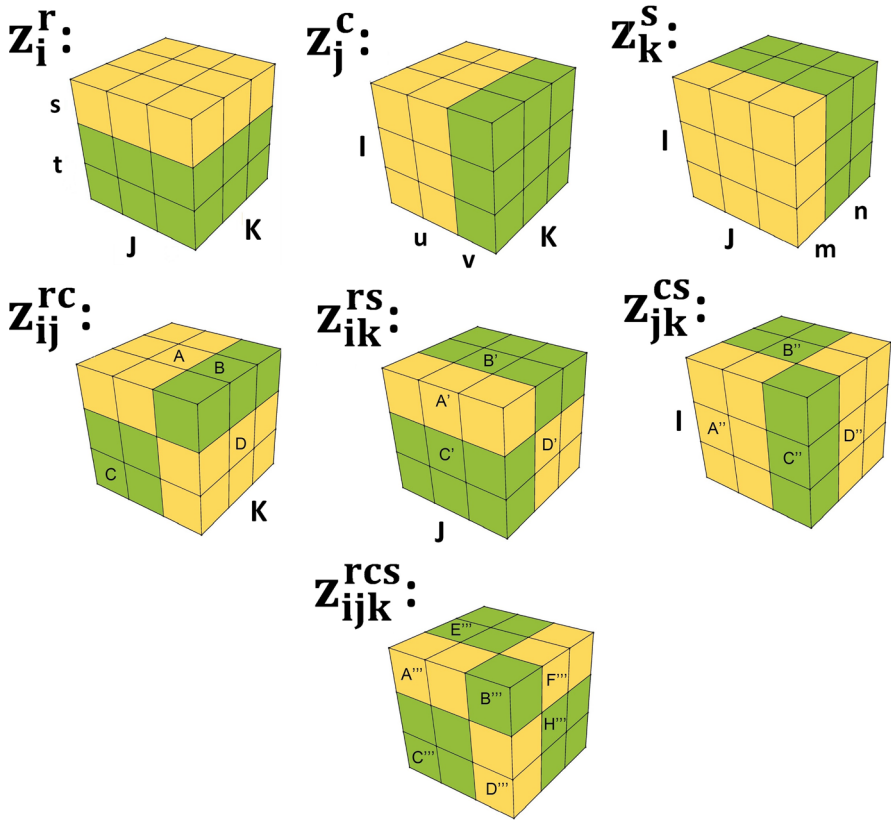
Fig. 2 Graphical representation of groups of parts involved in each type of coordinates forming the whole coordinate system designed for compositional cubes

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,

$$\xi_{11}^{rc} = \sqrt{\frac{1}{12}}(1, 1, 1, -1, -1, -1, -1, -1, -1, 1, 1, 1)' \propto \xi_1^r \circ \xi_1^c,$$

$$\xi_{11}^{rs} = \sqrt{\frac{1}{6}}(1, -0.5, -0.5, 1, -0.5, -0.5, -1, 0.5, 0.5, -1, 0.5, 0.5)' \propto \xi_1^r \circ \xi_1^s,$$

$$\xi_{12}^{rs} = \sqrt{\frac{1}{8}}(0, 1, -1, 0, 1, -1, 0, -1, 1, 0, -1, 1)' \propto \xi_1^r \circ \xi_2^s,$$

$$\xi_{11}^{cs} = \sqrt{\frac{1}{6}}(1, -0.5, -0.5, -1, 0.5, 0.5, 1, -0.5, -0.5, -1, 0.5, 0.5)' \propto \xi_1^c \circ \xi_1^s,$$

$$\xi_{12}^{cs} = \sqrt{\frac{1}{8}}(0, 1, -1, 0, -1, 1, 0, 1, -1, 0, -1, 1)' \propto \xi_1^c \circ \xi_2^s,$$

$$\xi_{111}^{rcs} = \sqrt{\frac{1}{6}}(1, -0.5, -0.5, -1, 0.5, 0.5, -1, 0.5, 0.5, 1, -0.5, -0.5)' \propto \xi_1^r \circ \xi_1^c \circ \xi_1^s,$$

$$\xi_{112}^{rcs} = \sqrt{\frac{1}{8}}(0, 1, -1, 0, -1, 1, 0, -1, 1, 0, 1, -1)' \propto \xi_1^r \circ \xi_1^c \circ \xi_2^s. \tag{31}$$

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

$$z_1^r = \sqrt{3} \ln \frac{g(\mathbf{x}_{1\bullet\bullet})}{g(\mathbf{x}_{2\bullet\bullet})} \qquad z_{11}^{rs} = \sqrt{\frac{2}{3}} \ln \frac{g(\mathbf{x}_{1\bullet 1})\sqrt{g(\mathbf{x}_{2\bullet 2})g(\mathbf{x}_{2\bullet 3})}}{g(\mathbf{x}_{2\bullet 1})\sqrt{g(\mathbf{x}_{1\bullet 2})g(\mathbf{x}_{1\bullet 3})}}$$

$$z_1^c = \sqrt{3} \ln \frac{g(\mathbf{x}_{\bullet 1\bullet})}{g(\mathbf{x}_{\bullet 2\bullet})} \qquad z_{12}^{rs} = \sqrt{\frac{1}{2}} \ln \frac{g(\mathbf{x}_{1\bullet 2})g(\mathbf{x}_{2\bullet 3})}{g(\mathbf{x}_{2\bullet 2})g(\mathbf{x}_{1\bullet 3})}$$

$$z_1^s = \sqrt{\frac{8}{3}} \ln \frac{g(\mathbf{x}_{\bullet\bullet 1})}{\sqrt{g(\mathbf{x}_{\bullet\bullet 2})g(\mathbf{x}_{\bullet\bullet 3})}} \qquad z_{11}^{cs} = \sqrt{\frac{2}{3}} \ln \frac{g(\mathbf{x}_{\bullet 11})\sqrt{g(\mathbf{x}_{\bullet 22})g(\mathbf{x}_{\bullet 23})}}{g(\mathbf{x}_{\bullet 21})\sqrt{g(\mathbf{x}_{\bullet 12})g(\mathbf{x}_{\bullet 13})}}$$

$$z_2^s = \sqrt{2} \ln \frac{g(\mathbf{x}_{\bullet\bullet 2})}{g(\mathbf{x}_{\bullet\bullet 3})} \qquad z_{12}^{cs} = \sqrt{\frac{1}{2}} \ln \frac{g(\mathbf{x}_{\bullet 12})g(\mathbf{x}_{\bullet 23})}{g(\mathbf{x}_{\bullet 13})g(\mathbf{x}_{\bullet 22})}$$

$$z_{11}^{rc} = \sqrt{\frac{3}{4}} \ln \frac{g(\mathbf{x}_{11\bullet})g(\mathbf{x}_{22\bullet})}{g(\mathbf{x}_{12\bullet})g(\mathbf{x}_{21\bullet})} \qquad z_{111}^{rcs} = \sqrt{\frac{1}{6}} \ln \frac{x_{111}x_{221}\sqrt{x_{122}x_{123}}\sqrt{x_{212}x_{213}}}{x_{211}x_{121}\sqrt{x_{112}x_{113}}\sqrt{x_{222}x_{223}}}$$

$$z_{112}^{rcs} = \sqrt{\frac{1}{8}} \ln \frac{x_{112}x_{222}x_{123}x_{213}}{x_{122}x_{212}x_{113}x_{223}}$$

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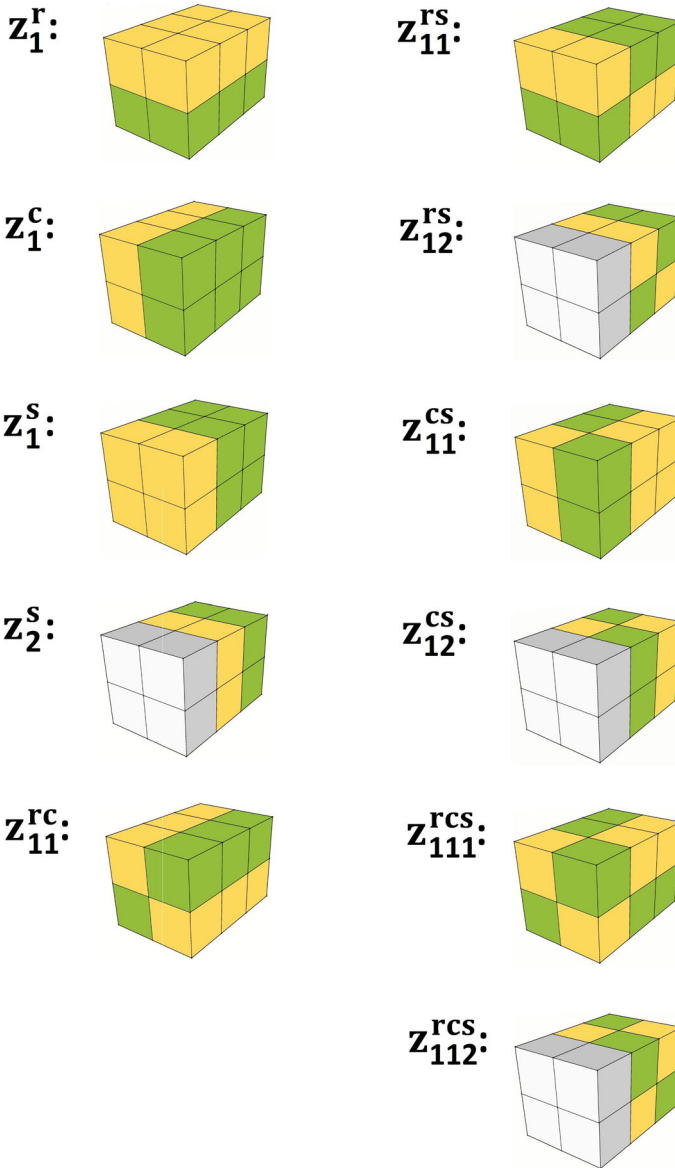
Fig. 3 Graphical representation of the groups of parts involved in particular coordinates. Yellow parts constitute the numerator, green the denominator and white parts are not included in the respective log-ratio. (Color figure online)

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

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

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

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_{111}^{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_{111}^{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`
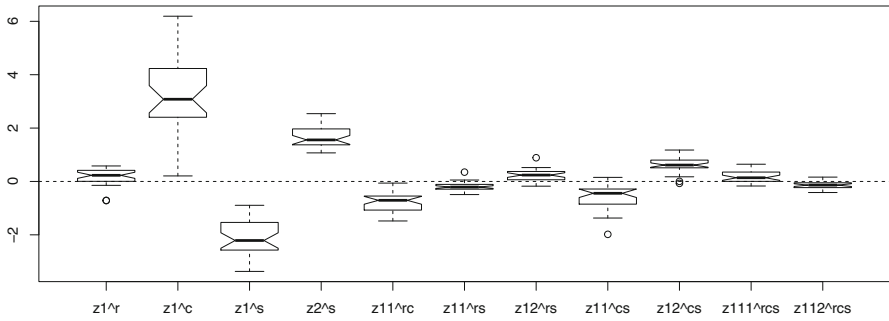
Fig. 4 Boxplots of the coordinates describing the employment structure

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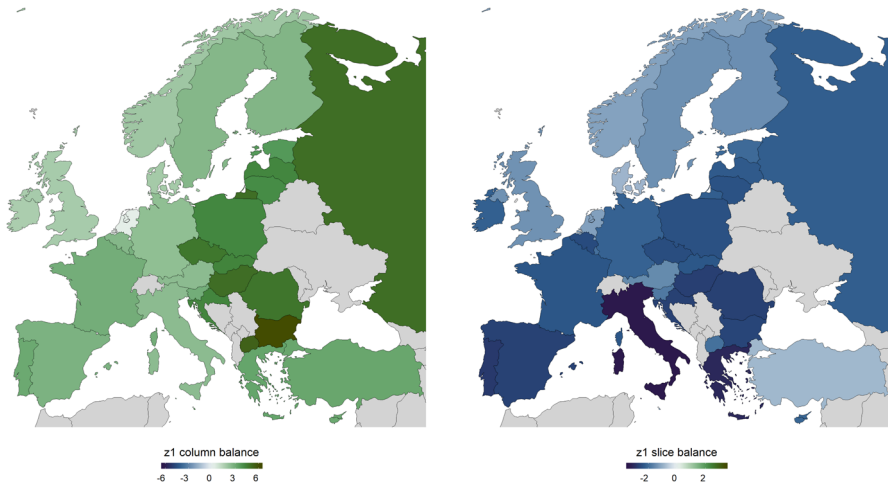
Fig. 5 Values of the coordinate z_1^c representing the log-ratio between full-time and part-time contracts (left) and the coordinate z_1^s, which represents the log-ratio between the youngest group of employees and the rest (right)

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
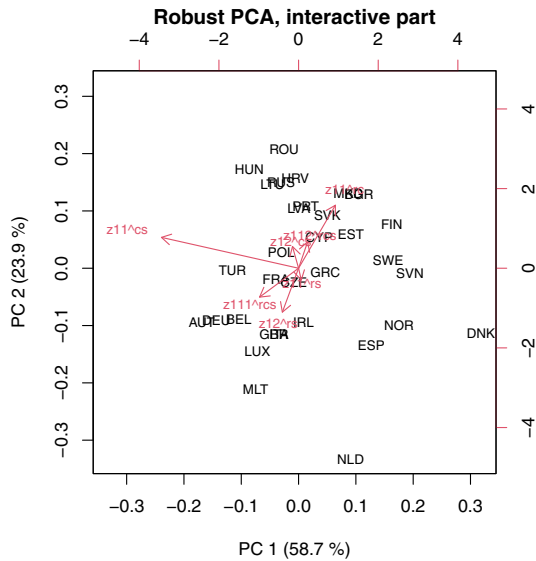
deviations from independence, are the coordinates z_{11}^{rc} and z_{11}^{cs}. The left map in Fig. 7 visualizes the values of the coordinate z_{11}^{rc}. This coordinate is negative for every country in the sample, and therefore the FT/PT ratio is always higher within male employees compared to females. 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}.
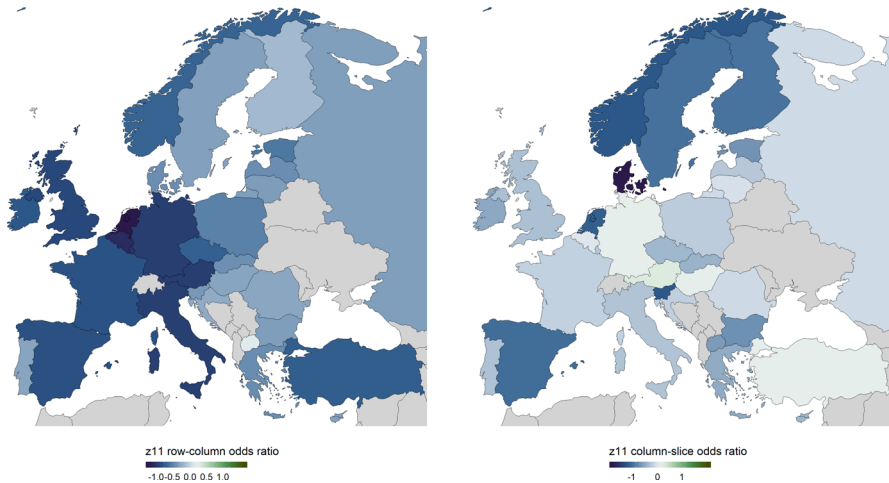
Fig. 7 Values of the coordinate z_{11}^{rc}, comparing the FT/PT ratio between female and male employees (left), and the coordinate z_{11}^{cs}, which compares the same ratio but between the youngest group of employees and the rest (right)

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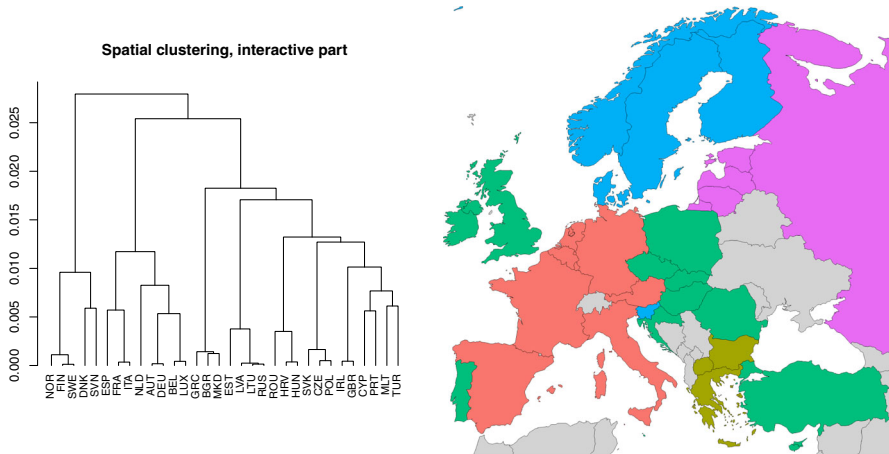
Fig. 8 Result of spatial clustering based on coordinates z_{11}^{rc}–z_{112}^{rcs} representing the interactive part of compositional cubes

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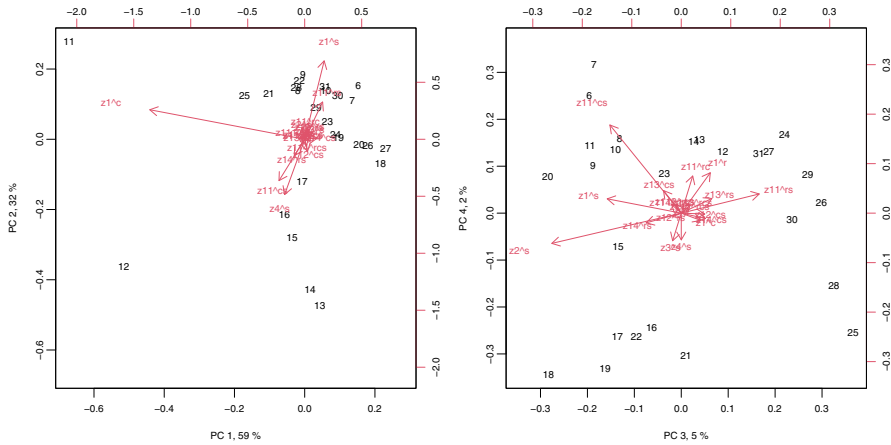
Fig. 9 Biplots based on the first four principal components computed for the mobility data. The numbers represent the week numbers in 2020

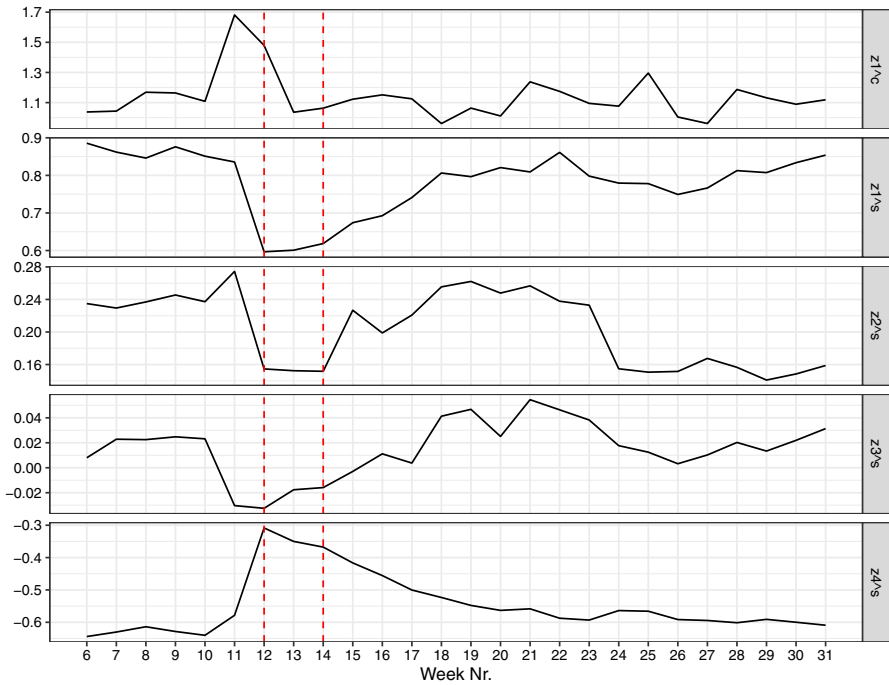**Fig. 10** Loadings of the principal components computed for the mobility data

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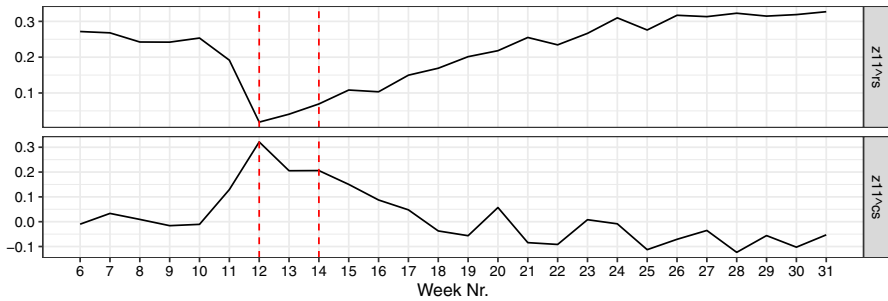
Fig. 12 Selected odds-ratio coordinates (without the normalizing constant) computed for the mobility data and their development over time. The red dashed lines define the first lockdown period in Austria

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 phenomenons. 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.

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{\tfrac{1}{12}} & \sqrt{\tfrac{1}{12}} & \sqrt{\tfrac{1}{12}} & \sqrt{\tfrac{1}{12}} & \sqrt{\tfrac{1}{12}} & \sqrt{\tfrac{1}{12}} & -\sqrt{\tfrac{1}{12}} & -\sqrt{\tfrac{1}{12}} & -\sqrt{\tfrac{1}{12}} & -\sqrt{\tfrac{1}{12}} & -\sqrt{\tfrac{1}{12}} & -\sqrt{\tfrac{1}{12}} \\
\sqrt{\tfrac{1}{12}} & \sqrt{\tfrac{1}{12}} & \sqrt{\tfrac{1}{12}} & -\sqrt{\tfrac{1}{12}} & -\sqrt{\tfrac{1}{12}} & -\sqrt{\tfrac{1}{12}} & \sqrt{\tfrac{1}{12}} & \sqrt{\tfrac{1}{12}} & \sqrt{\tfrac{1}{12}} & -\sqrt{\tfrac{1}{12}} & -\sqrt{\tfrac{1}{12}} & -\sqrt{\tfrac{1}{12}} \\
\sqrt{\tfrac{1}{6}} & -\sqrt{\tfrac{1}{24}} & -\sqrt{\tfrac{1}{24}} & \sqrt{\tfrac{1}{6}} & -\sqrt{\tfrac{1}{24}} & -\sqrt{\tfrac{1}{24}} & \sqrt{\tfrac{1}{6}} & -\sqrt{\tfrac{1}{24}} & -\sqrt{\tfrac{1}{24}} & \sqrt{\tfrac{1}{6}} & -\sqrt{\tfrac{1}{24}} & -\sqrt{\tfrac{1}{24}} \\
0 & \sqrt{\tfrac{1}{8}} & -\sqrt{\tfrac{1}{8}} & 0 & \sqrt{\tfrac{1}{8}} & -\sqrt{\tfrac{1}{8}} & 0 & \sqrt{\tfrac{1}{8}} & -\sqrt{\tfrac{1}{8}} & 0 & \sqrt{\tfrac{1}{8}} & -\sqrt{\tfrac{1}{8}} \\
\sqrt{\tfrac{1}{12}} & \sqrt{\tfrac{1}{12}} & \sqrt{\tfrac{1}{12}} & -\sqrt{\tfrac{1}{12}} & -\sqrt{\tfrac{1}{12}} & -\sqrt{\tfrac{1}{12}} & -\sqrt{\tfrac{1}{12}} & -\sqrt{\tfrac{1}{12}} & -\sqrt{\tfrac{1}{12}} & \sqrt{\tfrac{1}{12}} & \sqrt{\tfrac{1}{12}} & \sqrt{\tfrac{1}{12}} \\
\sqrt{\tfrac{1}{6}} & -\sqrt{\tfrac{1}{24}} & -\sqrt{\tfrac{1}{24}} & \sqrt{\tfrac{1}{6}} & -\sqrt{\tfrac{1}{24}} & -\sqrt{\tfrac{1}{24}} & -\sqrt{\tfrac{1}{6}} & \sqrt{\tfrac{1}{24}} & \sqrt{\tfrac{1}{24}} & -\sqrt{\tfrac{1}{6}} & \sqrt{\tfrac{1}{24}} & \sqrt{\tfrac{1}{24}} \\
0 & \sqrt{\tfrac{1}{8}} & -\sqrt{\tfrac{1}{8}} & 0 & \sqrt{\tfrac{1}{8}} & -\sqrt{\tfrac{1}{8}} & 0 & -\sqrt{\tfrac{1}{8}} & \sqrt{\tfrac{1}{8}} & 0 & -\sqrt{\tfrac{1}{8}} & \sqrt{\tfrac{1}{8}} \\
\sqrt{\tfrac{1}{6}} & -\sqrt{\tfrac{1}{24}} & -\sqrt{\tfrac{1}{24}} & -\sqrt{\tfrac{1}{6}} & \sqrt{\tfrac{1}{24}} & \sqrt{\tfrac{1}{24}} & \sqrt{\tfrac{1}{6}} & -\sqrt{\tfrac{1}{24}} & -\sqrt{\tfrac{1}{24}} & -\sqrt{\tfrac{1}{6}} & \sqrt{\tfrac{1}{24}} & \sqrt{\tfrac{1}{24}} \\
0 & \sqrt{\tfrac{1}{8}} & -\sqrt{\tfrac{1}{8}} & 0 & -\sqrt{\tfrac{1}{8}} & \sqrt{\tfrac{1}{8}} & 0 & \sqrt{\tfrac{1}{8}} & -\sqrt{\tfrac{1}{8}} & 0 & -\sqrt{\tfrac{1}{8}} & \sqrt{\tfrac{1}{8}} \\
\sqrt{\tfrac{1}{6}} & -\sqrt{\tfrac{1}{24}} & -\sqrt{\tfrac{1}{24}} & -\sqrt{\tfrac{1}{6}} & \sqrt{\tfrac{1}{24}} & \sqrt{\tfrac{1}{24}} & -\sqrt{\tfrac{1}{6}} & \sqrt{\tfrac{1}{24}} & \sqrt{\tfrac{1}{24}} & \sqrt{\tfrac{1}{6}} & -\sqrt{\tfrac{1}{24}} & -\sqrt{\tfrac{1}{24}} \\
0 & \sqrt{\tfrac{1}{8}} & -\sqrt{\tfrac{1}{8}} & 0 & -\sqrt{\tfrac{1}{8}} & \sqrt{\tfrac{1}{8}} & 0 & -\sqrt{\tfrac{1}{8}} & \sqrt{\tfrac{1}{8}} & 0 & \sqrt{\tfrac{1}{8}} & -\sqrt{\tfrac{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,

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

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

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

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

$$\boldsymbol{\xi}_1^{\mathrm{s*}} = \sqrt{\frac{1}{6}}(-0.5, -0.5, 1, -0.5, -0.5, 1, -0.5, -0.5, 1, -0.5, -0.5, 1)'$$

and

$$\boldsymbol{\xi}_2^{\mathrm{s*}} = \sqrt{\frac{1}{8}}(-1, 1, 0, -1, 1, 0, -1, 1, 0, -1, 1, 0)'.$$

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

$$\boldsymbol{\xi}_1^{\mathrm{r*}} = \boldsymbol{\xi}_1^{\mathrm{r}}, \quad \boldsymbol{\xi}_1^{\mathrm{c*}} = \boldsymbol{\xi}_1^{\mathrm{c}}, \quad \boldsymbol{\xi}_1^{\mathrm{rc*}} = \boldsymbol{\xi}_1^{\mathrm{rc}}.$$

All the remaining vectors change as follows:

$$\boldsymbol{\xi}_{11}^{\text{rs}*} = \sqrt{\frac{1}{6}}(-0.5, -0.5, 1, -0.5, -0.5, -1, 0.5, 0.5, -1, 0.5, 0.5, -1)' \propto \boldsymbol{\xi}_1^{\text{r}*} \circ \boldsymbol{\xi}_1^{\text{s}*},$$

$$\boldsymbol{\xi}_{12}^{\text{rs}*} = \sqrt{\frac{1}{8}}(-1, 1, 0, -1, 1, 0, 1, -1, 0, 1, -1, 0)' \propto \boldsymbol{\xi}_1^{\text{r}*} \circ \boldsymbol{\xi}_2^{\text{s}*},$$

$$\boldsymbol{\xi}_{11}^{\text{cs}*} = \sqrt{\frac{1}{6}}(-0.5, -0.5, 1, 0.5, 0.5, -1, -0.5, -0.5, 1, 0.5, 0.5, -1)' \propto \boldsymbol{\xi}_1^{\text{c}*} \circ \boldsymbol{\xi}_1^{\text{s}*},$$

$$\boldsymbol{\xi}_{12}^{\text{cs}*} = \sqrt{\frac{1}{8}}(-1, 1, 0, 1, -1, 0, -1, 1, 0, 1, -1, 0)' \propto \boldsymbol{\xi}_1^{\text{c}*} \circ \boldsymbol{\xi}_2^{\text{s}*},$$

$$\boldsymbol{\xi}_{111}^{\text{rcs}*} = \sqrt{\frac{1}{6}}(-0.5, -0.5, 1, 0.5, 0.5, -1, 0.5, 0.5, -1, -0.5, -0.5, 1)' \propto \boldsymbol{\xi}_1^{\text{r}*} \circ \boldsymbol{\xi}_1^{\text{c}*} \circ \boldsymbol{\xi}_1^{\text{s}*},$$

$$\boldsymbol{\xi}_{112}^{\text{rcs}*} = \sqrt{\frac{1}{8}}(-1, 1, 0, 1, -1, 0, 1, -1, 0, -1, 1, 0)' \propto \boldsymbol{\xi}_1^{\text{r}*} \circ \boldsymbol{\xi}_1^{\text{c}*} \circ \boldsymbol{\xi}_2^{\text{s}*}.$$

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

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

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:

$$z_1^{\mathrm{r}*} = \sqrt{3} \ln \frac{g(\mathbf{x}_{1\bullet\bullet})}{g(\mathbf{x}_{2\bullet\bullet})}$$

$$z_{11}^{\mathrm{rs}*} = \sqrt{\frac{2}{3}} \ln \frac{g(\mathbf{x}_{1\bullet3})\sqrt{g(\mathbf{x}_{2\bullet1})g(\mathbf{x}_{2\bullet2})}}{g(\mathbf{x}_{2\bullet3})\sqrt{g(\mathbf{x}_{1\bullet1})g(\mathbf{x}_{1\bullet2})}}$$

$$z_1^{\mathrm{c}*} = \sqrt{3} \ln \frac{g(\mathbf{x}_{\bullet1\bullet})}{g(\mathbf{x}_{\bullet2\bullet})}$$

$$z_{12}^{\mathrm{rs}*} = \sqrt{\frac{1}{2}} \ln \frac{g(\mathbf{x}_{1\bullet2})g(\mathbf{x}_{2\bullet1})}{g(\mathbf{x}_{2\bullet2})g(\mathbf{x}_{1\bullet1})}$$

$$z_1^{\mathrm{s}*} = \sqrt{\frac{8}{3}} \ln \frac{g(\mathbf{x}_{\bullet\bullet3})}{\sqrt{g(\mathbf{x}_{\bullet\bullet1})g(\mathbf{x}_{\bullet\bullet2})}}$$

$$z_{11}^{\mathrm{cs}*} = \sqrt{\frac{2}{3}} \ln \frac{g(\mathbf{x}_{\bullet13})\sqrt{g(\mathbf{x}_{\bullet21})g(\mathbf{x}_{\bullet22})}}{g(\mathbf{x}_{\bullet23})\sqrt{g(\mathbf{x}_{\bullet11})g(\mathbf{x}_{\bullet12})}}$$

$$z_2^{\mathrm{s}*} = \sqrt{2} \ln \frac{g(\mathbf{x}_{\bullet\bullet2})}{g(\mathbf{x}_{\bullet\bullet1})}$$

$$z_{12}^{\mathrm{cs}*} = \sqrt{\frac{1}{2}} \ln \frac{g(\mathbf{x}_{\bullet12})g(\mathbf{x}_{\bullet21})}{g(\mathbf{x}_{\bullet11})g(\mathbf{x}_{\bullet22})}$$

$$z_{11}^{\mathrm{rc}*} = \sqrt{\frac{3}{4}} \ln \frac{g(\mathbf{x}_{11\bullet})g(\mathbf{x}_{22\bullet})}{g(\mathbf{x}_{12\bullet})g(\mathbf{x}_{21\bullet})}$$

$$z_{111}^{\mathrm{rcs}*} = \sqrt{\frac{1}{6}} \ln \frac{x_{113}x_{223}\sqrt{x_{121}x_{122}}\sqrt{x_{211}x_{212}}}{x_{123}x_{213}\sqrt{x_{111}x_{112}}\sqrt{x_{221}x_{222}}}$$

$$z_{112}^{\mathrm{rcs}*} = \sqrt{\frac{1}{8}} \ln \frac{x_{112}x_{222}x_{121}x_{211}}{x_{122}x_{212}x_{111}x_{221}}$$

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, Pawlowsky-Glahn V (2005) Groups of parts and their balances in compositional data analysis. Math Geol 37:795–828

Egozcue JJ, Pawlowsky-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

Fačevicová 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

Fačevicová K, Hron K, Todorov V et al (2016) Compositional tables analysis in coordinates. Scand J Stat 43:962–977

Fačevicová K, Hron K, Todorov V et al (2018) General approach to coordinate representation of compositional tables. Scand J Stat 45(4):879–899

Fačevicová 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

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