General approach to coordinate representation of compositional tables

KAMILA FAČEVICOVÁ

Department of Mathematics, Palacký University Olomouc

KAREL HRON

Department of Mathematical Analysis and Applications of Mathematics, Palacký University Olomouc

VALENTIN TODOROV

United Nations Industrial Development Organisation

MATTHIAS TEMPL

Institute of Data Analysis and Process Design, Zurich University of Applied Sciences

Abstract Compositional tables can be considered a continuous counterpart to the well-known contingency tables. Their cells, which generally contain positive real numbers rather than just counts, carry relative information about relationships between two factors. Hence, compositional tables can be seen as a generalization of (vector) compositional data. Due to their relative character, compositions are commonly expressed in orthonormal coordinates using a sequential binary partition prior to being further processed by standard statistical tools. Unfortunatelly, the resulting coordinates do not respect the two-dimensional nature of compositional tables. Information about relationship between factors is thus not well captured. The aim of this

paper is to present a general system of orthonormal coordinates with respect to the Aitchison geometry, which allows for an analysis of the interactions between factors in a compositional table. This is achieved using logarithms of odds ratios, which are also widely used in the context of contingency tables.

Keywords: Aitchison geometry; compositional tables; odds ratio; orthonormal coordinates.

1 Introduction

In many practical situations, the object of statistical analysis is a table that represents the distribution of a given variable according to two (row and column) factors. If the relative contributions of the cells on the overall distribution are of primary interest rather than the concrete absolute values, we talk about compositional tables (Egozcue *et al.*, 2008, 2015). From this perspective, compositional tables represent a generalization of (vector) compositional data, where only ratios between parts are sufficient to extract all relevant information (Aitchison, 1986; Pawlowsky-Glahn *et al.*, 2015); the specific nature of compositional data is captured by the Aitchison geometry with the structure of finite-dimensional Euclidean vector space. Compositional tables are also closely linked to the well-known contingency tables, which represent the result of a multinomial sampling with cell probabilities $p_{ij} > 0, \sum_i \sum_j p_{ij} = 1$. Namely, the corresponding probability table with entries p_{ij} is only a proportional representation of the compositional table, see Egozcue *et al.* (2015) for details. Even contingency tables can be considered

as compositional tables, if the role of absolute cell values is disregarded in favour of their relative character. Statistical analysis of contingency tables is conducted using Pearson χ^2 statistic, log-linear models for independence testing, or correspondence analysis (Greenacre, 2007). As these methods rely strongly on the assumption of Euclidean geometry (Egozcue *et al.*, 2015) (like most standard statistical methods Eaton, 1983), they are not suitable for compositional tables that are driven by the Aitchison geometry. For correspondence analysis even a link to compositional data exists (Greenacre, 2011), if the absolute values of counts are irrelevant, but it does not utilize all possibilities resulting from considering the Aitchison geometry. Moreover, as is the case of compositional data, it is natural to consider an ensemble of compositional tables that can be analysed with popular multivariate statistical methods (such as principal component analysis, clustering, classification, etc.). This is a distinct difference to the case of contingency tables, where such issues are usually not considered or at most indirectly, through three-way contingency tables and the respective log-linear models.

The key point in statistical analysis of compositional tables is to express them in orthonormal coordinates with respect to the Aitchison geometry, to which the properties of the Euclidean geometry are applicable and which allow to apply statistical methods and calculations that are defined according to the Euclidean geometry. In fact, it follows the idea of odds ratio representation of contingency tables as discussed in Agresti (2002), ch. 2, page 55 and ch. 7, page 276. As there is no standard (natural) basis with respect to the Aitchison geometry, it is of primary importance to derive interpretable coordinates. For the general case of compositional data, it is possible to

construct coordinates in terms of balances between groups of compositional parts (Egozcue & Pawlowsky-Glahn, 2005). However, from the perspective of compositional tables, balances are not satisfactory as they do not follow their two-factor nature and possibility of its decomposition into independent and interactive parts (Egozcue *et al.*, 2015). The first comprehensive system of coordinates suitable for compositional tables was proposed in Fačevicová *et al.* (2016). Its generalization, which allows for the selection of a coordinate system with respect to the nature of the row/column factors and their cell-values and, consequently, achieves better interpretability of the coordinates, is presented here.

The next section summarizes the basics of compositional data, and compositional tables as their two-factor generalization. The third section examines the coordinate representation of compositional data using balances as well as the proposed general coordinates for compositional tables. Since the construction of these new coordinates may seem a bit tricky before understanding the intuitive concept behind it, it is explained step by step using a working example with a 3×5 compositional table, where each step is also illustrated graphically. Another important property of the proposed coordinate system is that it allows to decompose the original compositional table and can thus be used for a detailed analysis of relationships between factors. This feature is discussed in detail at the end of the third section and in the following section concerning macroeconomic analysis.

2 Compositional data and compositional tables

As mentioned above, compositional tables represent a generalization of (vector) compositional data as observations carrying exclusively relative information (Aitchison, 1986). This important aspect drives all considerations presented in the subsequent sections.

2.1 Compositional data

If only ratios between parts are relevant for a statistical analysis of a positive vector, it is common to refer to its compositional nature (Pawlowsky-Glahn *et al.*, 2015). Accordingly, D-part compositional data are defined as vectors with strictly positive components (parts) that quantitatively describe the relative contributions to a whole. Consequently, the sum of parts is not relevant for the analysis, and using the closure operation $\mathcal{C}(.)$, each composition $\mathbf{x} = (x_1, \ldots, x_D)$ can be rescaled to a constant sum ($\kappa > 0$) representation without any loss of information; $\mathcal{C}(\mathbf{x}) = (\kappa x_1 / \sum_i x_i, \ldots, \kappa x_D / \sum_i x_i)$. The sample space of representations of D-part compositional data with an arbitrary, but fixed κ, is a subspace of $\mathbf{R}^D$ called D-part simplex, $\mathcal{S}^D = \{\mathbf{x} = (x_1, x_2, \ldots, x_D) \,|x_i > 0, \forall i, \sum_i x_i = \kappa\}$. The constant sum constraint representation reduces the dimension of $\mathcal{S}^D$ to $D - 1$, i.e. the actual number of parts minus one.

The assumption that only ratios between components carry relevant information about the composition leads to the following principles of compositional data analysis (Pawlowsky-Glahn *et al.*, 2015). The first principle,

scale invariance, states that any rescaling of the original compositional vector $\mathbf{x}$ (using closure $\mathcal{C}(\mathbf{x})$ like the proportional representation) should not alter the results of their analysis. Subcompositional coherence, the second principle of compositional data analysis, requires subcompositions to behave like orthogonal projections in real analyses. For example, the distance between two full compositions must be greater than, or equal to, the distance between them when considering any subcomposition. Similarly, if a non-informative part is removed, the results should not change. Finally, the result of any analysis cannot depend on the order of compositional parts, leading to the permutation invariance principle.

Principles of compositional data analysis led to introducing the Aitchison geometry (Billheimer, 2001; Pawlowsky-Glahn & Egozcue, 2001), which forms the underlying algebraic-geometrical structure of compositions (Euclidean vector space of dimension $D-1$). Its basic operations are perturbation and powering, defined for $\mathbf{x}, \mathbf{y}$ from $\mathcal{S}^D$ and $\alpha \in \mathbf{R}$ as compositions $\mathbf{x} \oplus \mathbf{y} = \mathcal{C}\left(x_1 y_1, x_2 y_2, \ldots, x_D y_D\right)$ and $\alpha \odot \mathbf{x} = \mathcal{C}\left(x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha\right)$, respectively. Consequently, $\mathbf{n} = \mathcal{C}\left(1, 1, \ldots, 1\right)$ represents the neutral element of the perturbation operation. To complete the Euclidean vector space structure, the Aitchison inner product, norm and distance are defined as

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_i \sum_j \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}, \quad \|\mathbf{x}\|_a = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_a}$$

$$\text{and} \quad \mathrm{d}_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_a ,$$

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

2.2 Compositional tables

An $I \times J$ table $\mathbf{x}$, whose cells $x_{ij} > 0$, for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots J$ convey relative contributions to a whole (probability, overall output, etc.) can be considered as a natural extension of vector compositional data and is called compositional table. This type of observations basically conveys relative information on the relationship between two factors with I and J values, respectively. But also the other way around, compositional data can be obtained by vectorization of compositional tables. Therefore, any reasonable analysis of compositional tables should follow the same assumptions as for compositional vectors, but with specific (two-factor) interpretation of their parts; here, a subcomposition of compositional table is realized by omitting entire row(s) and/or column(s) and is called partial table. Basic operations of the Aitchison geometry should also be extended to the case of compositional tables. Perturbation of two compositional tables $\mathbf{x}$ and $\mathbf{y}$ of the same dimension $I \times J$ results in a new compositional table with entries

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C} \begin{pmatrix} x_{11}y_{11} & \cdots & x_{1J}y_{1J} \\ \vdots & \ddots & \vdots \\ x_{I1}y_{I1} & \cdots & x_{IJ}y_{IJ} \end{pmatrix};$$

similarly, by powering of compositional table $\mathbf{x}$ by a constant α, the following table

$$\alpha \odot \mathbf{x} = \mathcal{C} \begin{pmatrix} x_{11}^{\alpha} & \cdots & x_{1J}^{\alpha} \\ \vdots & \ddots & \vdots \\ x_{I1}^{\alpha} & \cdots & x_{IJ}^{\alpha} \end{pmatrix}$$

is obtained. Finally, the Aitchison inner product modifies to

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2IJ} \sum_{i,j} \sum_{k,l} \ln \frac{x_{ij}}{x_{kl}} \ln \frac{y_{ij}}{y_{kl}}. \tag{1}$$

The sample space of (representations of) $I \times J$ compositional tables is a $(IJ-1)$-dimensional simplex $\mathcal{S}^{IJ}$.

3 Coordinate representation of compositional data and compositional tables

Due to the specific nature of compositional data as represented by the above principles, standard statistical methods are not suitable for analysing them. Instead of developing their counterparts within the Aitchison geometry, it seems much more intuitive to express compositions isometrically in real coordinates with respect to this geometry and then to proceed with the usual statistical analysis on these coordinates (Pawlowsky-Glahn *et al.*, 2015). Apparently, the simplest and easiest interpretable case of such coordinates is represented by centred logratio (clr) coefficients, defined for D-part composition $\mathbf{x} = (x_1, \ldots, x_D)$ as $\mathrm{clr}(\mathbf{x}) = (\ln x_1/g(\mathbf{x}), \ln x_2/g(\mathbf{x}), \ldots, \ln x_D/g(\mathbf{x}))$, where $g(\mathbf{x}) = \sqrt[D]{\prod_{i=1}^{D} x_i}$ stands for geometric mean of parts. Even though clr coefficients preserve angles and distances and treat compositional parts symmetrically, they lead to a singular covariance matrix. Apart from purely geometrical disadvantages (like ambiguity of coordinate representation), this fact seriously limits the usability of clr coefficients in many statistical methods. One way out is to apply isometric logratio (ilr) coordinates, i.e. coordinates with respect to an orthonormal basis on the simplex. According

to basic algebraic-geometrical rules and the dimensionality of the Aitchison geometry, the ilr coordinates are defined as

$$\mathbf{z} = \mathrm{ilr}(\mathbf{x}) = \left(\langle \mathbf{x}, \mathbf{e}_1 \rangle_a, \langle \mathbf{x}, \mathbf{e}_2 \rangle_a, \ldots, \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a\right) = (z_1, z_2, \ldots, z_{D-1}), \quad (2)$$

where $\mathbf{e}_i = \mathcal{C}(e_{i1}, e_{i2}, \ldots, e_{iD}), i = 1, 2, \ldots, D-1$ form an orthonormal basis of the simplex. Due to the isometric isomorphism of the ilr coordinates, it immediately follows that

$$\mathrm{ilr}\left((\alpha \odot \mathbf{x}) \oplus (\beta \oplus \mathbf{y})\right) = \alpha \cdot \mathrm{ilr}(\mathbf{x}) + \beta \cdot \mathrm{ilr}(\mathbf{y}), \qquad \langle \mathbf{x}, \mathbf{y} \rangle_a = \langle \mathrm{ilr}(\mathbf{x}), \mathrm{ilr}(\mathbf{y}) \rangle,$$

$$\|\mathbf{x}\|_a = \|\mathrm{ilr}(\mathbf{x})\| \qquad \text{and} \qquad \mathrm{d}_a(\mathbf{x}, \mathbf{y}) = \mathrm{d}(\mathrm{ilr}(\mathbf{x}), \mathrm{ilr}(\mathbf{y})).$$

Clearly, it is not possible to assign an orthonormal coordinate to each of the compositional parts simultaneously, as it was the case with clr coefficients. Therefore, interpretable orthonormal coordinates are of primary interest. Since the coordinates $\mathbf{z}$ correspond to a specific choice of basis vectors (compositions) $\mathbf{e}_i, i = 1, \ldots, D-1$, they can be selected in accordance with the analysis' objectives and possible a priori knowledge about the compositional parts. One popular option for the construction of interpretable orthonormal coordinates is to apply the sequential binary partition (SBP) procedure (Egozcue & Pawlowsky-Glahn, 2005), based on a step-by-step separation of parts into non-overlapping groups. Accordingly, in the first step of SBP, the entire composition is divided into two subcompositions. In the next step, only one of the subcompositions from the previous step is taken and further separated into two groups. This process continues until all groups of parts consist of a single one only. The SBP is done in $D-1$ steps; in each step, one coordinate is obtained,

$$z_i = \sqrt{\frac{uv}{u+v}} \ln \frac{(x_{j_1} x_{j_2} \cdots x_{j_u})^{1/u}}{(x_{k_1} x_{k_2} \cdots x_{k_v})^{1/v}}, \; i = 1, \ldots, D-1. \quad (3)$$

Here, u, v stand for numbers of parts contained in the first and second group, respectively, $\{j_1, \ldots, j_u\}$ and $\{k_1, \ldots, k_v\}$ are their indices. When the parts assigned to the first group are marked with $+$, those in the second group by $-$ and the parts not included in any of the two groups in the i-th step of the partition by 0, SBP can also be illustrated graphically. Table 1 results from one possible SBP for five-part compositional data.

Table 1 about here.

Orthonormal coordinates resulting from SBP (3) can be interpreted in terms of balances between groups of parts, represented by their respective geometrical means. Using a priori expert knowledge, SBP can be chosen with the aim of capturing the most relevant information contained in the ratios between compositional parts and their groups. Particular choice of balances thus depends on the context of data analysis. For example, geochemical data consists of major and minor elements and can be further divided according to a distinct composition of the analysed rock/soil.

As regards vector compositions, balances represent the most popular class of orthonormal coordinates that was recently successfully applied in a number of real-world studies (Pawlowsky-Glahn & Buccianti, 2011). On the other hand, balance interpretation seems unsuitable for compositional tables. Because their cells represent relationships between two factors, only considering two groups of parts into a coordinate would not take account of the two-dimensional nature of these observations. In fact, balances are suitable for extracting information from the two factors individually, thus dealing with the tables's rows/columns. To represent inter-factorial patterns, coordinates in the form of (log) odds ratios between four groups of parts seem

to be preferable, similar to the case of contingency tables (Agresti, 2002). Such coordinates lead to a natural extension of balances for compositional tables. To sum up, balances can be used to capture (log-)ratios within row and column factors, while odds ratios link relative information between the two factors. It turns out (see Section 3.2) that numbers of coordinates in terms of balances of rows/columns and odds ratios reflect the dimensions of tables resulting from a decomposition of a compositional table into its independent and interactive parts (Egozcue *et al.*, 2008).

In line with the above, one specific choice of such odds ratios, representing $(I-1)(J-1)$ orthonormal *pivot coordinates* of an $I \times J$ compositional table, is

$$z_{rc} = \frac{1}{\sqrt{r \cdot c \cdot (r-1) \cdot (c-1)}} \ln \frac{\prod_{i=1}^{r-1} \prod_{j=1}^{c-1} (x_{ij} x_{rc})}{\prod_{i=1}^{r-1} \prod_{j=1}^{c-1} (x_{ic} x_{rj})}. \tag{4}$$

Here, one group of the odds ratio is always formed by a single pivot part $x_{rc}, r = 2, \ldots, I$ and $c = 2, \ldots, J$, which determines the lower right corner of a partial table (see Fačevicová *et al.* (2016), for details). Alternatively, these coordinates can also be seen as a scaled sum of log odds ratios according to some logical scheme, all containing part x_{rc}; this follows directly from (4). It is notable that any possible odds ratio is contained in only one of the coordinates z_{rc}. Given this natural restriction, the coordinates (4) can be further generalized to cover log odds ratios between groups of parts of an arbitrary size. Consequently, the interpretation of such coordinates can be easily adapted in accordance with the specific problem being analysed.

In addition to coordinates in terms of odds ratios, balance-like coordinates must also be determined, so that together $IJ - 1$ orthonormal coordinates are obtained. For the construction of the generalized coordinates of $I \times J$

compositional table, let us first consider SBP of the entire rows (columns) of compositional table $\mathbf{x}$, which is denoted by SBPr (SBPc) in the following. This partition is in line with the nature of the levels of row (column) factors, and follows standard SBP; in each of the $I-1$ ($J-1$) steps, the levels with some common properties are separated from the others. Accordingly, the first $I+J-2$ coordinates $\mathbf{z}^{\mathrm{r}}$ and $\mathbf{z}^{\mathrm{c}}$ of the $I \times J$ compositional table $\mathbf{x}$ are given as

$$z_i^{\mathrm{r}} = \sqrt{\frac{stJ}{s+t}} \ln \frac{[g(\mathbf{x}_{j_1\cdot}) \cdots g(\mathbf{x}_{j_s\cdot})]^{1/s}}{[g(\mathbf{x}_{k_1\cdot}) \cdots g(\mathbf{x}_{k_t\cdot})]^{1/t}}, \quad \text{for} \quad i = 1, 2, \ldots, I-1 \tag{5}$$

and

$$z_j^{\mathrm{c}} = \sqrt{\frac{uvI}{u+v}} \ln \frac{[g(\mathbf{x}_{\cdot l_1}) \cdots g(\mathbf{x}_{\cdot l_u})]^{1/u}}{[g(\mathbf{x}_{\cdot m_1}) \cdots g(\mathbf{x}_{\cdot m_v})]^{1/v}}, \quad \text{for} \quad j = 1, 2, \ldots, J-1, \tag{6}$$

where s, t (u, v) are the numbers of rows (columns) involved in the i-th (j-th) step of SBP, the indices $(j_1\cdot, \ldots, j_s\cdot)$ and $(k_1\cdot, \ldots, k_t\cdot)$, or $(\cdot l_1, \ldots, \cdot l_u)$ and $(\cdot m_1, \ldots, \cdot m_v)$ specify the rows/columns and $g(.)$ stands for the geometric mean.

The remaining coordinates should be orthogonal to these first $I+J-2$ variables, and in order to construct them, some generalization of the basic SBP needs to be introduced. It is based on the partitioning of the parts of the compositional table into four groups (blocks) in a systematic manner that results in coordinates in form of a logarithm of odds ratio between these four groups (marked as A (upper left), B (upper right), C (lower left) and D (lower right))

$$z^{\mathrm{OR}} = \sqrt{\frac{a \cdot d}{a+b+c+d}} \ln \frac{(x_{i_1} \cdots x_{i_a})^{1/a} (x_{l_1} \cdots x_{l_d})^{1/d}}{(x_{j_1} \cdots x_{j_b})^{1/b} (x_{k_1} \cdots x_{k_c})^{1/c}}, \tag{7}$$

where a, b, c, d are the numbers of parts in groups A, B, C and D, respectively and $i_\cdot, j_\cdot, k_\cdot, l_\cdot$ are the indices of those parts. In the following steps, this partition is continued in smaller partial tables in accordance with the starting row and column SBPs.

The separation into subgroups (A–D) and the construction of partial tables should take into account the row and column grouping defined in SBPr and SBPc. Thus, the first four groups are created by the first steps of SBPr and SBPc and determine the first coordinate. If the compositional table consists of more than four parts, a subsequent step should be taken to partition it further. Firstly, the proper partial table should be identified and the only possible partial tables are formed by pairs of groups (A,B), (C,D), (A,C) and (B,D), which are successively analysed. If (A,B) has more than one row, the next coordinate is related to parts of this partial table when the groups are again determined by steps of SBPr and SBPc. The next possible partial table is first sought within the current partial table, but if it only consists of four parts (i.e. the smallest meaningful table), it is necessary to go back an look for another partial table in the bigger superior table from the previous step of the partition. The partial tables with only one row or column or partial tables, which were already analysed in the previous step of partial tables formed by the pairs of groups (A,B), (C,D), (A,C) and (B,D) of each proper partial table are analysed. The process results in $(I-1)(J-1)$ coordinates, each with an interpretation in terms of odds ratios between groups within the respective partial table. Alternatively, each coordinate can also be interpreted as a sum of log odds ratios, each involving four cells only. There are $\binom{I}{2}\binom{J}{2}$ of them in the entire table, each contained exclusively

in one of these new coordinates.

The construction of partial tables and coordinates is done using a combination of row and column SBPs. Although the above description demonstrates how the coordinates are derived, the output can be summarized as follows. For the first step of SBPr applied to the rows of the table, all $J-1$ steps of SBPc are performed. The first $J-1$ coordinates are obtained in accordance with (7). The next $J-1$ coordinates are obtained by applying the second step of SBPr to the rows and all of the steps of the SBPc to the columns, and so on, until all $I-1$ steps of the SBPr have been completed. All $(I-1)(J-1)$ coordinates of z^{OR} thus result from a successive application of all steps of the SBPr combined with repeated use of all steps of the SBPc, or conversely.

For the sake of completeness, the generating vectors from (2) that correspond to the proposed coordinates are

$$\mathbf{e}_i^{\mathrm{r}} \quad \text{with parts} \quad \begin{cases} & \text{for positions corresponding to rows} \\ \exp\left(\sqrt{\frac{t}{Js(s+t)}}\right) & j_1,\ldots,j_s, \\ \exp\left(-\sqrt{\frac{s}{Jt(s+t)}}\right) & k_1,\ldots,k_t, \\ \exp(0) & \text{otherwise,} \end{cases} \tag{8}$$

where $(j_1,\ldots,j_s)$ and $(k_1,\ldots,k_t)$ are indices of rows included in the i-th step of SBPr, for $i=1,\ldots,I-1$,

$$\mathbf{e}_j^{\mathrm{c}} \quad \text{with parts} \quad \begin{cases} & \text{for positions corresponding to rows} \\ \exp\left(\sqrt{\frac{v}{Iu(u+v)}}\right) & l_1, \ldots, l_u, \\ \exp\left(-\sqrt{\frac{u}{Iv(u+v)}}\right) & m_1, \ldots, m_v, \\ \exp(0) & \text{otherwise,} \end{cases} \tag{9}$$

where $(l_1, \ldots, l_u)$ and $(m_1, \ldots, m_v)$ are indices of columns included in the j-th step of SBPc, for $j = 1, \ldots, J-1$ and finally

$$\mathbf{e}_k^{\mathrm{OR}} \quad \text{with parts} \quad \begin{cases} & \text{for positions from group} \\ \exp\left(\sqrt{\frac{d}{a(a+b+c+d)}}\right) & A, \\ \exp\left(-\frac{1}{b}\sqrt{\frac{ad}{a+b+c+d}}\right) & B, \\ \exp\left(-\frac{1}{c}\sqrt{\frac{ad}{a+b+c+d}}\right) & C, \\ \exp\left(\sqrt{\frac{a}{d(a+b+c+d)}}\right) & D, \\ \exp(0) & \text{otherwise,} \end{cases} \tag{10}$$

for $k = 1, \ldots, (I-1)(J-1)$, where A, B, C and D are groups of parts included in the corresponding coordinate and a, b, c, d numbers of these parts, as described above.

3.1 Example - coordinate representation of 3×5 compositional table

To illustrate the above construction of partial tables, let us consider a 3×5 compositional table for which the complete system of orthonormal coordinates is developed. The first six coordinates capture relative information

(logratios) from the row/column factors, built in accordance with (5) and (6). They follow SBPs from Table 2 which are also graphically presented in Figure 1. Accordingly, for the SBPs complete rows and columns are taken and result in coordinates

$$z_1^{\mathrm{r}} = \sqrt{\frac{10}{3}} \ln \frac{g(x_{1.})}{(g(x_{2.})g(x_{3.}))^{1/2}}, \tag{11}$$

$$z_2^{\mathrm{r}} = \sqrt{\frac{5}{2}} \ln \frac{g(x_{2.})}{g(x_{3.})}, \tag{12}$$

$$z_1^{\mathrm{c}} = \sqrt{\frac{18}{5}} \ln \frac{(g(x_{.1})g(x_{.2}))^{1/2}}{(g(x_{.3})g(x_{.4})g(x_{.5}))^{1/3}}, \tag{13}$$

$$z_2^{\mathrm{c}} = \sqrt{\frac{3}{2}} \ln \frac{g(x_{.1})}{g(x_{.2})}, \tag{14}$$

$$z_3^{\mathrm{c}} = \sqrt{\frac{6}{3}} \ln \frac{g(x_{.3})}{(g(x_{.4})g(x_{.5}))^{1/2}}, \tag{15}$$

$$z_4^{\mathrm{c}} = \sqrt{\frac{3}{2}} \ln \frac{g(x_{.4})}{g(x_{.5})}. \tag{16}$$

Table 2 about here

Figure 1 about here

In the next step, relevant partial tables are defined and the remaining eight coordinates are computed according to (7). First, the entries of the entire table are divided into four groups according to steps 1 and I from SBPc and SBPr. This divided table, as well as all the following partial tables, are illustrated in Figure 2 (table **(1)**). According to this separation, the first coordinate is computed as

$$z_1^{\mathrm{OR}} = \frac{2\sqrt{5}}{5} \ln \frac{(x_{11}x_{12})^{1/2} (x_{23}x_{24}x_{25}x_{33}x_{34}x_{35})^{1/6}}{(x_{13}x_{14}x_{15})^{1/3} (x_{21}x_{22}x_{31}x_{32})^{1/4}}.$$

Figure 2 about here

Next, partial tables are created from the parts of the pairs of groups (A_1, B_1)(table **(2a)**), (C_1, D_1)(table **(2b)**), (A_1, C_1)(table **(2c)**) or (B_1, D_1) (table **(2d)**). Since table **(2a)** consists of a single row, it cannot be further separated and we therefore skip it and instead analyse the next possible partial table **(2b)**. This partial table already consists of more than one row and column, thus it represents the first partial table to generate one of the coordinates. The separation of columns within this table still corresponds to step 1 of SBPc. In SBPr, the second and third rows of the compositional tables were separated in step II, thus, the four groups in this partial table are based on steps 1 and II. According to this separation, the next coordinate results in

$$z_2^{\mathrm{OR}} = \sqrt{\frac{3}{5}} \ln \frac{(x_{21}x_{22})^{1/2} (x_{33}x_{34}x_{35})^{1/3}}{(x_{23}x_{24}x_{25})^{1/3} (x_{31}x_{32})^{1/2}}.$$

Figure 3 about here

This table can be further split and the next two coordinates are related to partial tables **(3c)** (formed by groups A_{2b} and C_{2b}) and **(3d)** (groups B_{2b} and D_{2b}), as tables **(3a)** and **(3b)** consist of a single row only, as is evident from Figure 3. In table **(3c)**, the partition of rows has already been achieved through step II of SBPr; furthermore, the columns are separated by applying step 2 of SBPc. Accordingly, the next coordinate

$$z_3^{\mathrm{OR}} = \frac{1}{2} \ln \frac{x_{21}x_{32}}{x_{22}x_{31}}$$

is obtained. Since each group in this table consists by a single part, x_{21}, x_{22}, x_{31} and x_{32}, respectively, this table cannot be further partitioned and we can proceed to the partial table **(3d)**. In this table, row separation is determined by

step II of SBPr while step 3 of SBPc is used for its columns. After assessing the next coordinate

$$z_4^{\mathrm{OR}} = \frac{\sqrt{3}}{3} \ln \frac{x_{23} \left(x_{34} x_{35}\right)^{1/2}}{\left(x_{24} x_{25}\right)^{1/2} x_{33}},$$

the partial table **(3d)** can be further divided. Figure 4 presents all possible partial tables.

Figure 4 about here

Of these, only partial table **(4d)** has more than one row and column, and, consequently, can be used to construct the next coordinate. With respect to steps II and 4 of SBPr and SBPc, this coordinate results in

$$z_5^{\mathrm{OR}} = \frac{1}{2} \ln \frac{x_{24} x_{35}}{x_{25} x_{34}}.$$

The partition of partial table **(2b)** is thus completed and the construction procedure returns to a partial table **(2c)** (Figures 2 and 5). This table, separated using steps I and 2, determines the next coordinate

$$z_6^{\mathrm{OR}} = \frac{\sqrt{3}}{3} \ln \frac{x_{11} \left(x_{22} x_{32}\right)^{1/2}}{x_{12} \left(x_{21} x_{31}\right)^{1/2}}$$

Figure 5 about here

and the only regular partial table contained within it is **(5b)**. Yet since this table is identical with table **(3c)** and has already been analysed, we can skip it and the procedure immediately proceeds to partial table **(2d)**. This table is divided using steps I and 3 of SBPr and SBPc, thus the next coordinate is

$$z_7^{\mathrm{OR}} = \frac{2}{3} \ln \frac{x_{13} \left(x_{24} x_{25} x_{34} x_{35}\right)^{1/4}}{\left(x_{14} x_{15}\right)^{1/2} \left(x_{23} x_{33}\right)^{1/2}}.$$

Figure 6 about here

The consequent possible partial tables are illustrated in Figure 6, which clearly shows that the only regular tables are **(6b)** and **(6d)**. Since **(6b)** corresponds to **(3d)**, the last coordinate is based on table **(6d)** and steps I and 4.

$$z_8^{\mathrm{OR}} = \frac{\sqrt{3}}{3} \ln \frac{x_{14} \left(x_{25} x_{35}\right)^{1/2}}{x_{15} \left(x_{24} x_{34}\right)^{1/2}}$$

For completeness, Figure 7 illustrates the partition of this table, which leads to partial table **(7b)** already obtained above as **(4d)**.

Figure 7 about here

3.2 Decomposition of compositional tables

The construction of coordinates reveals that there are two groups of coordinates. The first $I+J-2$ can be interpreted in terms of balances between rows and columns of the original table $\mathbf{x}$, and the remaining $(I-1)(J-1)$ coordinates are associated with odds ratios between groups of parts. This grouping has a geometric justification, since according to Egozcue *et al.* (2008), Egozcue *et al.* (2015) and Ortego & Egozcue (2016), each compositional table can be decomposed into two parts, the independent and the interactive one, namely independence and interaction tables $\mathbf{x}_{\mathrm{ind}}$ and $\mathbf{x}_{\mathrm{int}}$. Each entry of the independence table is a product of the respective geometric marginals that reminds the usual independence case, known from contingency tables. Consequently, the interaction table accounts for the relations between row and column factors. Both $\mathbf{x}_{\mathrm{ind}}$ and $\mathbf{x}_{\mathrm{int}}$ fulfil the following relation,

$$\mathbf{x} = \mathbf{x}_{\mathrm{ind}} \oplus \mathbf{x}_{\mathrm{int}} \tag{17}$$

and their entries are given as

$$x_{ij}^{\text{ind}} = \left(\prod_{k=1}^{I} \prod_{l=1}^{J} x_{kj} x_{il} \right)^{\frac{1}{IJ}}, \tag{18}$$

$$x_{ij}^{\text{int}} = \left(\prod_{k=1}^{I} \prod_{l=1}^{J} \frac{x_{ij}}{x_{kj} x_{il}} \right)^{\frac{1}{IJ}}, \tag{19}$$

respectively. Note that in case of independence (in the above sense), the interaction table equals to the neutral element, i.e. a table with all the same entries. When the coordinate representation $\mathbf{z}^{\text{r}} = (z_1^{\text{r}}, \ldots, z_{I-1}^{\text{r}})$, $\mathbf{z}^{\text{c}} = (z_1^{\text{c}}, \ldots, z_{J-1}^{\text{c}})$, $\mathbf{z}^{\text{OR}} = (z_1^{\text{OR}}, \ldots, z_{(I-1)(J-1)}^{\text{OR}})$ is applied to the independence table $\mathbf{x}_{\text{ind}}$, the only non-zero coordinates are $z_i^{\text{r}}, z_j^{\text{c}}$ for $i = 1, \ldots, I-1$, $j = 1, \ldots, J-1$, and their values are the same as for the original table $\mathbf{x}$. Moreover, the number of these non-zero coordinates equals the dimension of the subspace of independence tables (see, e.g., Fačevicová *et al.* (2016) for details). An analogous feature also holds for the interaction table and the coordinates $\mathbf{z}^{\text{OR}}$. Accordingly, the vector of coordinates $(\mathbf{z}^{\text{r}}, \mathbf{z}^{\text{c}}, \mathbf{0}_{(I-1)(J-1)})$ of the independence table can be denoted as $\mathbf{z}_{\text{ind}}$ and the coordinates $(\mathbf{0}_{I+J-2}, \mathbf{z}^{\text{OR}})$ of the interaction table as $\mathbf{z}_{\text{int}}$. Finally, the vector of coordinates of the original compositional table $\mathbf{x}$ can be written as $\mathbf{z} = \text{ilr}(\mathbf{x}_{\text{ind}}) + \text{ilr}(\mathbf{x}_{\text{int}}) = \mathbf{z}_{\text{ind}} + \mathbf{z}_{\text{int}} = (\mathbf{z}^{\text{r}}, \mathbf{z}^{\text{c}}, \mathbf{z}^{\text{OR}})$.

3.3 Inverse transformation

Since the coordinates of compositional tables result from a one-to-one mapping, it is also possible to transform them back into the IJ-part simplex using generating vectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_{IJ-1}\} = \left\{\mathbf{e}^{\text{r}}, \mathbf{e}^{\text{c}}, \mathbf{e}^{\text{OR}}\right\}$ from (8), (9) and (10). For this purpose, the compositional table $\mathbf{x}$ is vectorized, $\text{vec}(\mathbf{x}) =$

$(x_{11}, \ldots, x_{1J}, \ldots, x_{I1}, \ldots, x_{IJ})$, thus the basis compositions have a length of IJ. Consider the $(IJ-1, IJ)$ dimensional matrix $\mathbf{\Psi}$, with rows equal to $\mathrm{clr}(\mathbf{e}_i)$. Since $\mathbf{e}_i$, $i = 1, \ldots, IJ-1$ forms an orthonormal basis of $\mathcal{S}^{IJ}$, matrix $\mathbf{\Psi}$ satisfies $\mathbf{\Psi}\mathbf{\Psi}' = \mathbf{I}_{IJ-1}$. The inverse transformation from the $(IJ-1)$-dimensional real space to $\mathcal{S}^{IJ}$ is given as

$$\mathrm{vec}(\mathbf{x}) = \mathcal{C}\left(\exp(\mathbf{z}\mathbf{\Psi})\right). \tag{20}$$

Now the back-transformed $I \times J$ compositional table can be easily reconstructed by rearranging the parts into a matrix with I rows and J columns.

4 Distribution of manufacturing output - analysis of independence

The aim of this application is to discuss possibility of an independence analysis between two factors using a sample of tables. For this purpose, a sample of 42 3×5 compositional tables, each representing the distribution of manufacturing output in a given country in 2009, is used. For countries for which the 2009 data were incomplete, data from 2008 or 2007 are used. The list of all countries in the sample, accompanied with the year of data origin, is provided in Table 3. The tables cover the category "Manufacture of food products and beverages", classified according to the 3-digit level of the International Standard Industrial Classification of All Economic Activities ISIC (Revision 3) (UN, 2002). Thus, the values of this first factor are as follows (numbers correspond to ISIC codes):

- 151 Processed meat, fish, fruit, vegetables, fats

- 152 Dairy products
- 153 Grain mill products, starches, animal feeds
- 154 Other food products
- 155 Beverages.

The second factor consists of components of the output with the categories Labour cost (LAB), Operational surplus (SUR) and Input (INP). Since we are interested in the relative structure of manufacturing output, we use the compositional approach. Table 4 provides the percentage representation of one table from the sample, specifically, distribution of manufacturing in the U.S. in 2008.

Table 3 about here

Table 4 about here

To express the tables in coordinates, we start by defining the SBPs of the row and column factors. Obviously, partitions discussed in Section 3.1 and their respective coordinates are applicable here and could represent (after a change of the columns order), e.g., situation when 154 Other food products and 155 Beverages form a separate group. Nevertheless, in the case of manufacturing industries it seems more logical to separate the production of beverages from that of food products in the first step and we will follow this strategy now. Accordingly, in a next step we can separate industries that produce food products that are not well-specified (154 Other food products) from the remaining three, followed by the separation of supplementary

products (153 Grain mill products, starches, animal feeds). Finally, in the last step, industries 151 (Processed meat, fish, fruit, vegetables, fats) and 152 (Dairy products) are separated. Similarly, the components of output should first be divided into Input and Value added (Value added = Labour cost + Operational surplus), and Value added is further divided into Labour cost and Operational surplus in the second step. These two SBPs, visualized graphically in Table 5 and Figure 8, provide a unique coordinate representation of the compositional tables in the sample. This implies that the entire set of coordinates $\mathbf{z}$ can be immediately computed for each table of the sample (the complete list of coordinates, together with their graphical representations is provided by Table 6). Since only one category was split in each step of the SBPs, the resulting set of coordinates corresponds to pivot coordinates, proposed and extensively described in Fačevicová *et al.* (2016). Due to its easy construction and interpretability, such coordinate representation can also be considered as a basic option for compositional tables. Accordingly, the categories Beverages and Input assume an exceptional position as there are coordinates that capture their relative contribution with respect to the other categories in the rows (z_1^{r}) and columns (z_1^{c}) of the tables and by considering interactions between the two factors (z_1^{OR}). This is thus a natural generalization of the approach to interpretable balances for compositional data as introduced in Fišerová & Hron (2011) and recently applied in a range of applications (Filzmoser *et al.*, 2012; Martín-Fernández *et al.*, 2012; Kalivodová *et al.*, 2015).

Table 5 about here

Figure 8 about here

Each table from the sample is expressed in coordinates based on the presented methodology and proposed SBPs. For example, the coordinate representation of the example table, distribution of manufacturing output in the U.S., results in

$$\mathbf{z}^{\mathrm{r}}_{\mathrm{USA}} = (2.52, 2.39)\,,$$

$$\mathbf{z}^{\mathrm{c}}_{\mathrm{USA}} = (-0.68, 0.92, -0.89, -1.34)\,,$$

$$\mathbf{z}^{\mathrm{OR}}_{\mathrm{USA}} = (-0.33, 0.25, -0.09, 0.49, 0.09, -0.67, -0.09, 0.11)\,,$$

$$\mathbf{z}_{\mathrm{USA}} = \left(\mathbf{z}^{\mathrm{r}}_{\mathrm{USA}}, \mathbf{z}^{\mathrm{c}}_{\mathrm{USA}}, \mathbf{z}^{\mathrm{OR}}_{\mathrm{USA}}\right).$$

The positive values of $\mathbf{z}^{\mathrm{r}}_{\mathrm{USA}}$ indicate that the Input component is higher than the Value added component and, further, that the Operational surplus component is higher than the Labour cost component across all (averaged) food and beverage industries in the U.S. economy. The average production of food is slightly higher than the production of beverages; this feature is captured by the first coordinate of $\mathbf{z}^{\mathrm{c}}_{\mathrm{USA}}$, which equals -0.68. The relationship between the output components and the manufacturing industries are described by the vector of coordinates $\mathbf{z}^{\mathrm{OR}}_{\mathrm{USA}}$. Because most of the values of this vector are close to zero, it suggests near independence between the two factors.

Table 6 about here

These very preliminary observations for the case of the U.S. are followed by a detailed inspection of the complete data structure. In order to visualize both the observations (the countries) and the variables (the row and column balances and the odds ratio coordinates), we use principal component analysis (PCA) to reduce the dimensionality and then present the data

as a covariance biplot (Gabriel, 1971) in Figure 9. All coordinates are centred prior to further processing as is common in compositional data analysis (Pawlowsky-Glahn *et al.*, 2015). While the balances represent information within both factors, odds ratios capture the relations between them. The preliminary expectation about independence between factors is confirmed as the odds-ratio variables play a marginal role in capturing the multivariate variability. The concrete choice of SBP for the columns of the compositional tables demonstrates its relevance here, the coordinate z_1^{c} that separates beverages from the other industries belongs to one of the three main marker variables. In the right upper corner of the biplot, a compact cluster of industrialized countries emerges; they are predominantly characterized by low values of the coordinate z_2^{r}, i.e. by the dominance of Labour cost over Operational surplus across manufacturing industries. Within this cluster, several other European countries are represented (Poland, Romania, Greece, Latvia, The Former Republic of Macedonia; most of them EU member states) which seems to follow the same pattern. However, several industrialized but non-European countries (Japan, USA, Malaysia) have higher values for this coordinate. By contrast, high values of this variable occur for developing and emerging industrial economies (Azerbaijan, Indonesia, Sri Lanka, Colombia). The two least developed countries included in the sample (Ethiopia and Tanzania) have high values on z_1^{c} and z_3^{c}, not far away from them are Mongolia, Georgia and Kenya. The coordinates z_1^{c} and z_3^{c} (the latter being strongly correlated with z_2^{c}) denote countries' beverage and food production specifics. Particularly, it is interesting that beverages dominate aggregated food production across output components for (industrialized) European and (least

developed) African countries. Note that in contrast to analysing standard multivariate (or even compositional) data, variables with different interpretations are considered together, namely row/column factors (balances) as well as odds ratios that connect both of them. This must be taken into account when deriving any conclusion from the biplot.

Figure 9 about here

To identify more detailed patterns, separate covariance biplots are constructed for the two main groups of variables that form the coordinate system of compositional tables, balances and odds ratios, see Figure 10. While, as expected, the structure of loadings and scores remains almost unchanged for balances (Figure 10, right) compared to Figure 9, the biplot for odds ratio coordinates (Figure 10, left) reveals additional interesting features about sources of relationships between the two factors. The grouping of several countries around the origin shows that no relationship between the two factors exists for these countries (Greece, Brazil, Chile, Iran, Ecuador). Figure 10 (right) clearly indicates that the Labor cost part of the output dominates the Operating surplus over averaged manufacturing industries in European (industrialized and emerging) economies, provided by z_2^{r}. Coordinate z_5^{OR} in Figure 10 (left) indicates that this ratio is higher for category 152 than 151. Similarly, coordinate z_1^{OR} shows that the dominance of beverages over other food producing industries is higher for Input than for Labour cost and Operating surplus. Finally, coordinate z_2^{OR} provides a more detailed insight into the relationship between the value added components than z_5^{OR}: for countries like Ireland, Lithuania and Poland, the dominance of Labour cost

over Operating surplus is much stronger for the beverages industry than for others. As both "marker variables" z_2^{r} and z_1^{c} are a source of interpretation for z_2^{OR}, this might also be the main reason for the border position of these countries in Figure 9.

Figure 10 about here

5 Conclusion

The general approach to orthonormal coordinates for compositional tables, as introduced in the paper, represents an important step for coordinate representation of multifactorial compositional data. The coordinates presented here represent a natural generalization of the concept of balances as introduced in Egozcue & Pawlowsky-Glahn (2005), which have already proven their practical usefulness in a wide range of applications and open a variety of perspectives for their further development.

Similarly as for compositional data, proper coordinate representation of compositional tables is necessary to enable statistical processing using standard multivariate statistical tools. The proposed coordinate system is partly formed by balances and partly by coordinates with log odds ratio interpretation. This choice takes into account the possibility to decompose compositional tables into its independent and interactive parts. Consequently, it allows to study tables from the decomposition also separately and analyse, e.g., the possible independence of both factors only from the interactive part of coordinates. Accordingly, the presented orthonormal coordinate system respects the nature of row and column factors and thus allows for their bet-

ter interpretability. It is also important to emphasize the complementarity of both balance and odds ratio coordinates - as demonstrated in the application part of this paper, the first are inherently contained in the interpretation of the latter ones.

In addition to the standard statistical processing of individual compositional tables and their samples (clustering, classification, time series), a solid geometrical background of the new coordinates enables proceeding to other tasks related to the analysis of compositional (probability) tables, e.g. the compositional counterpart of log-linear models and related methods. Together with the mentioned possibility of further generalization of compositional structures with more than two factors, the presented orthonormal coordinates aim to become key reference for further development of compositional data analysis.

Disclaimer The views expressed herein are those of the authors and do not necessarily reflect the views of the United Nations Industrial Development Organization.

Acknowledgments The paper was supported by the grant IGA_PdF_2017_007 of the Internal Grant Agency of the Palacký University in Olomouc.

References

Agresti, A. (2002). *Categorical data analysis,* 2nd ed. Wiley, New York.

Aitchison, J. (1986). *The statistical analysis of compositional data.* Chapman and Hall, London.

Billheimer, D., Guttorp, P. & Fagan, W. (2001). Statistical interpretation of species composition. *J. Amer. Statist. Assoc.*. **96**:456, 1205–1214.

Eaton, M. L. (1983). *Multivariate statistics. A vector space approach.* Wiley, New York.

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

Egozcue, J.J., Díaz-Barrero, J.L. & Pawlowsky-Glahn, V. (2008). Compositional analysis of bivariate discrete probabilities. In *Proceedings of CODAWORK'08, The 3rd Compositional Data Analysis Workshop* (eds J. Daunis-i-Estadela & J. A. Martín-Fernández.). University of Girona, Spain.

Egozcue, J.J., Pawlowsky-Glahn, V., Templ, M. & Hron, K. (2015). Independence in contingency tables using simplicial geometry. *Comm. Statist. Theory Methods*. **44**:18, 3978–3996.

Fačevicová, K., Hron, K., Todorov, V. & Templ, M. (2016). Compositional tables analysis in coordinates. *Scand. J. Stat.*. **43**, 962-977.

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**:4, 455–468.

Gabriel, K.R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*. **58**:3, 453–467.

Greenacre, M. (2007). *Correspondence analysis in practice*, 2nd ed. Chapman & Hall/CRC Press.

Greenacre, M. (2011). Compositional data and correspondence analysis. In *Compositional data analysis: Theory and practice*, (eds V. Pawlowsky-Glahn & A. Buccianti), 104–113. Wiley, Chichester.

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

Martín-Fernández, J.A., Hron, K., Templ, M., Filzmoser, P. & Palarea-Albaladejo, J. (2012). Model-based replacement of rounded zeros in compositional data: Classical and robust approaches. *Comput. Statist. Data Anal.*. **56**:9, 2688–2704.

Ortego, M.I. & Egozcue, J.J. (2016). Bayesian estimation of the orthogonal decomposition of a contingency table. *Austrian Journal of Statistics.*. **45**, 45–56.

Pawlowsky-Glahn, V. & Egozcue, J.J. (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment.*. **15**:5, 384–398.

Pawlowsky-Glahn, V. & Buccianti, A. (2011). *Compositional data analysis: Theory and applications.* Wiley, Chichester.

Pawlowsky-Glahn, V., Egozcue, J.J. & Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data.* Wiley, Chichester.

UN (2002). *International Standard Industrial Classification of All Economic Activities (ISIC) Rev. 3.1.* (Available from http://unstats.un.org/). [Accessed November 23, 2017.]

UNIDO (2017). *International Yearbook of Industrial Statistics 2017*, p. 7. Vienna.

Address Department of Mathematics, Faculty of Education, Palacký University, Žižkovo nám. 5, CZ-77140 Olomouc, Czech Republic.
E-mail kamila.facevicova@gmail.com

Table 1: Example of sequential binary partition and the corresponding orthonormal coordinates for five-part compositional data

i	x_1	x_2	x_3	x_4	x_5	u	v	z_i
1	+	+	−	−	−	2	3	$\sqrt{\frac{6}{5}} \ln \frac{\sqrt{x_1 x_2}}{\sqrt[3]{x_3 x_4 x_5}}$
2	+	−	0	0	0	1	1	$\frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2}$
3	0	0	+	−	−	1	2	$\sqrt{\frac{2}{3}} \ln \frac{x_3}{\sqrt{x_4 x_5}}$
4	0	0	0	+	−	1	1	$\frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2}$

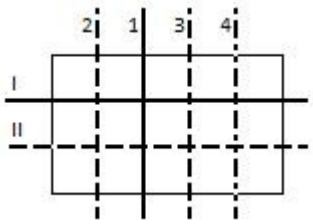


Figure 1: Graphical representation of sequential binary partitions SBPr and SBPc, applied to a 3×5 compositional table.

Table 2: Example of sequential binary partition applied to complete rows (SBPr, upper table) and entire columns (SBPc, lower table) of a 3×5 compositional table $\mathbf{x}$

i	$x_{1.}$	$x_{2.}$	$x_{3.}$	s	t
I	$+$	$-$	$-$	1	2
II	0	$+$	$-$	1	1

j	$x_{.1}$	$x_{.2}$	$x_{.3}$	$x_{.4}$	$x_{.5}$	u	v
1	$+$	$+$	$-$	$-$	$-$	2	3
2	$+$	$-$	0	0	0	1	1
3	0	0	$+$	$-$	$-$	1	2
4	0	0	0	$+$	$-$	1	1

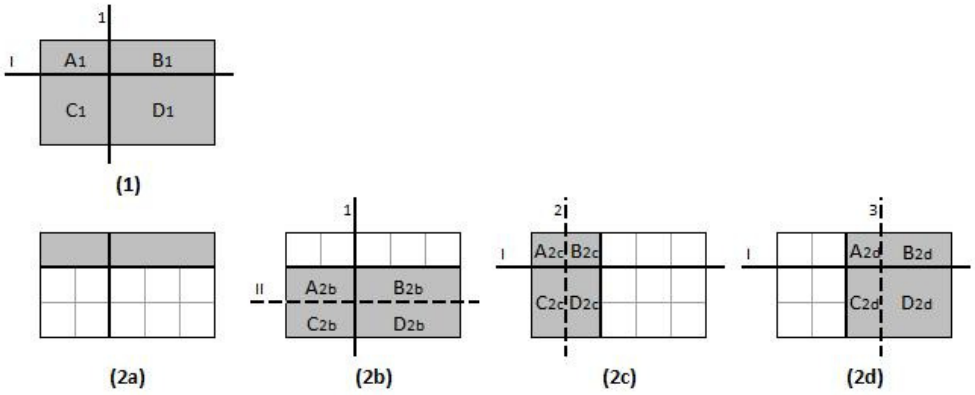


Figure 2: Graphical representation of group separation in the 3×5 table **(1)**. Lower grey tables **(2a-d)** illustrate construction of possible partial tables. New coordinates can be computed only from tables **(2b-d)**.

Table 3: List of countries analyzed in Section 4.1 (Distribution of manufacturing output), their abbreviations, years of data origin and classification to one of the categories Industrialized (1), Emerging industrial (2), Other developing economies (3) and Least developed countries (4) (UNIDO, 2017).

Country	Abbreviation	Year	Category
Azerbaijan	AZE	2009	3
Austria	AUT	2008	1
Belgium	BEL	2008	1
Brazil	BRA	2007	2
Bulgaria	BGR	2007	2
Sri Lanka	LKA	2009	3
Chile	CHL	2008	2
Colombia	COL	2009	2
Ecuador	ECU	2008	3
Ethiopia	ETH	2009	4
France	FRA	2008	1
Georgia	GEO	2009	3
Germany	DEU	2008	1
Greece	GRC	2007	2
Hungary	HUN	2008	1
India	IND	2007	2
Indonesia	IDN	2009	2
Iran (Islamic Republic of)	IRN	2009	3
Ireland	IRL	2008	1
Japan	JPN	2007	1
Jordan	JOR	2009	3
Kenya	KEN	2009	3
Kyrgyzstan	KGZ	2009	3
Lebanon	LBN	2007	3
Latvia	LVA	2007	2
Lithuania	LTU	2008	1
Malaysia	MYS	2008	1
Malta	MLT	2008	1
Mongolia	MNG	2009	3
Morocco	MAR	2009	3
Oman	OMN	2009	2
Poland	POL	2008	2
Portugal	PRT	2008	1
Romania	ROU	2008	2
Russian Federation	RUS	2009	1
Slovenia	SVN	2007	1
Spain	ESP	2008	1
Sweden	SWE	2008	1
The f. Yugosl. Rep of Macedonia	MKD	2009	2
United Republic of Tanzania	TZA	2007	4

Table 4: Distribution of food and beverages production in the U.S. in 2008 according to the 3-digit ISIC category and components of output (in %)

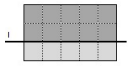
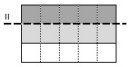
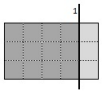
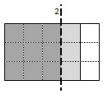
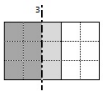
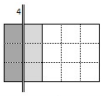
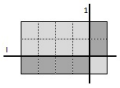
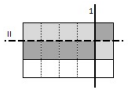
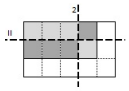
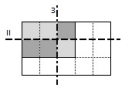
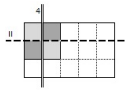
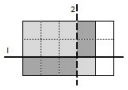
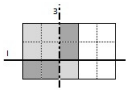
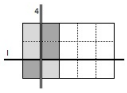
USA	151	152	153	154	155
Labour	2.78	0.80	0.55	2.90	0.84
Surplus	8.37	2.88	4.19	10.94	5.24
Input	25.32	9.65	7.62	12.05	5.87

Table 5: Sequential binary partition of manufacturing industries (upper table) and components of output (lower table)

SBPc	151	152	153	154	155	u	v
1	$-$	$-$	$-$	$-$	$+$	1	4
2	$-$	$-$	$-$	$+$	0	1	3
3	$-$	$-$	$+$	0	0	1	2
4	$-$	$+$	0	0	0	1	1

SBPr	Labour	Surplus	Input	s	t
I	$-$	$-$	$+$	1	2
II	$-$	$+$	0	1	1

Table 6: List of coordinates in the example (Distribution of manufacturing output) together with their graphical representations

$z_1^r = \sqrt{\frac{10}{3}} \ln \frac{g(x_{3.})}{(g(x_{1.})g(x_{2.}))^{1/2}}$ 	$z_2^r = \sqrt{\frac{5}{2}} \ln \frac{g(x_{2.})}{g(x_{1.})}$ 
$z_1^c = \sqrt{\frac{12}{5}} \ln \frac{g(x_{.5})}{(g(x_{.1})g(x_{.2})g(x_{.3})g(x_{.4}))^{1/4}}$ 	$z_2^c = \sqrt{\frac{9}{4}} \ln \frac{g(x_{.4})}{(g(x_{.1})g(x_{.2})g(x_{.3}))^{1/3}}$ 
$z_3^c = \sqrt{\frac{6}{3}} \ln \frac{g(x_{.3})}{(g(x_{.1})g(x_{.2}))^{1/2}}$ 	$z_4^c = \sqrt{\frac{3}{2}} \ln \frac{g(x_{.2})}{g(x_{.1})}$ 
$z_1^{OR} = \sqrt{\frac{8}{15}} \ln \frac{(x_{11}...x_{14}x_{21}...x_{24})^{1/8}x_{35}}{(x_{15}x_{25})^{1/2}(x_{31}...x_{34})^{1/4}}$ 	$z_2^{OR} = \sqrt{\frac{4}{10}} \ln \frac{(x_{11}...x_{14})^{1/4}x_{25}}{x_{15}(x_{21}...x_{24})^{1/4}}$ 
$z_3^{OR} = \sqrt{\frac{3}{8}} \ln \frac{(x_{11}x_{12}x_{13})^{1/3}x_{24}}{x_{14}(x_{21}x_{22}x_{23})^{1/3}}$ 	$z_4^{OR} = \sqrt{\frac{2}{6}} \ln \frac{(x_{11}x_{12})^{1/2}x_{23}}{x_{13}(x_{21}x_{22})^{1/2}}$ 
$z_5^{OR} = \frac{1}{2} \ln \frac{x_{11}x_{22}}{x_{12}x_{21}}$ 	$z_6^{OR} = \sqrt{\frac{6}{12}} \ln \frac{(x_{11}x_{12}x_{13}x_{21}x_{22}x_{23})^{1/6}x_{34}}{(x_{14}x_{24})^{1/2}(x_{31}x_{32}x_{33})^{1/3}}$ 
$z_7^{OR} = \sqrt{\frac{4}{9}} \ln \frac{(x_{11}x_{12}x_{21}x_{22})^{1/4}x_{33}}{(x_{13}x_{23})^{1/2}(x_{31}x_{32})^{1/2}}$ 	$z_8^{OR} = \sqrt{\frac{2}{6}} \ln \frac{(x_{11}x_{21})^{1/2}x_{32}}{(x_{12}x_{22})^{1/2}x_{13}}$ 

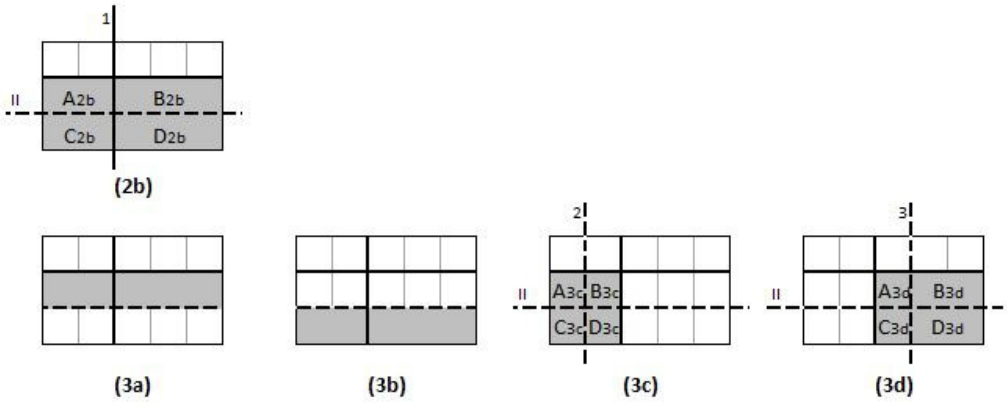


Figure 3: Graphical representation of group separation in the partial table **(2b)**. Lower grey tables **(3a-d)** illustrate the construction of possible partial tables. New coordinates can only be computed from tables **(3c)** and **(3d)**.

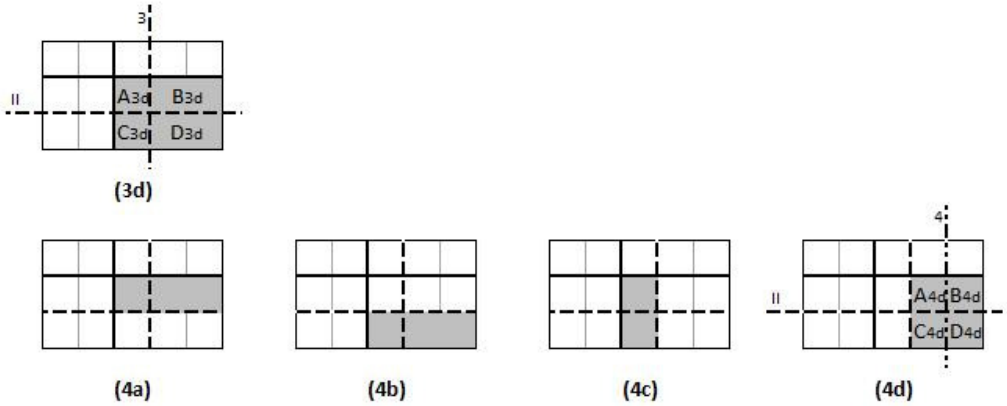


Figure 4: Graphical representation of group separation in the partial table **(3d)**. Lower grey tables **(4a-d)** illustrate the construction of possible partial tables. New coordinates can only be computed from table **(4d)**.

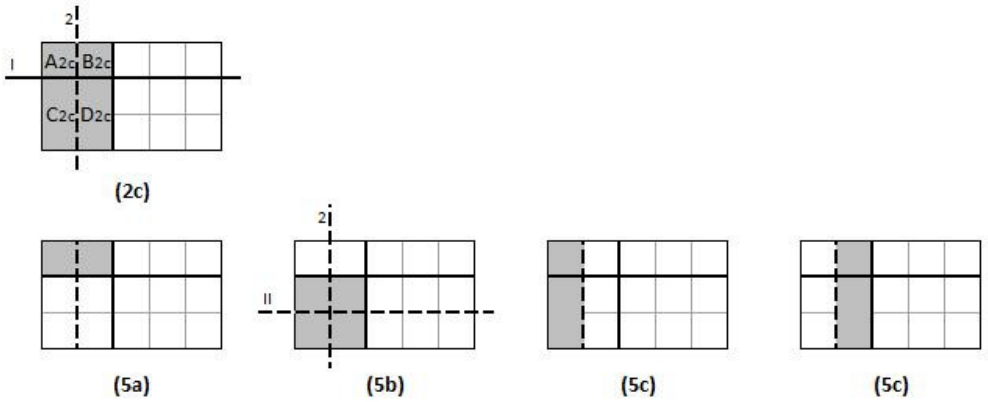


Figure 5: Graphical representation of group separation in partial table **(2c)**. Lower grey tables **(5a-d)** illustrate the construction of possible partial tables. The only regular partial table is table **(5b)**, which has already been considered (table **(3c)**).

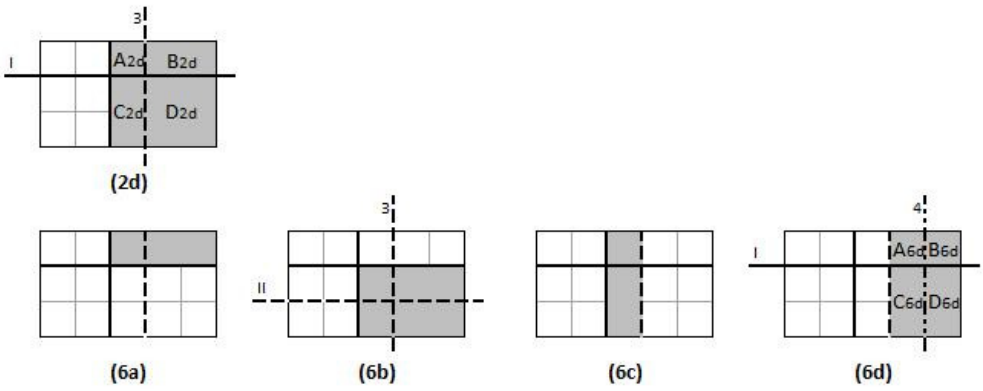


Figure 6: Graphical representation of group separation in partial table **(2d)**. Lower grey tables **(6a-d)** illustrate the construction of possible partial tables. The only regular partial table is table **(6b)**, which has already been analysed as table **(3d)**, and table **(6d)**, which represents the last coordinate z_8^{OR}

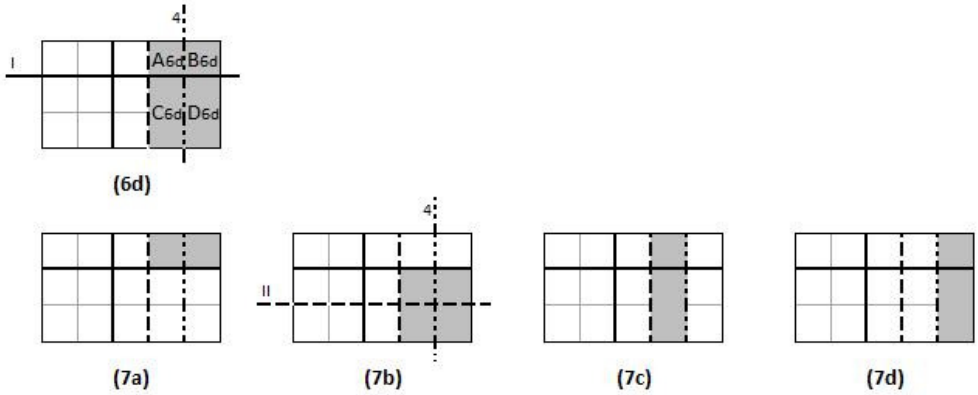


Figure 7: Graphical representation of group separation in the partial table **(6d)**. Lower grey tables **(7a-d)** illustrate construction of possible partial tables. The only regular partial table is table **(7b)**, which was already analysed as table **(4d)**.

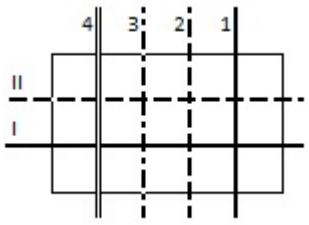


Figure 8: Graphical representation of sequential binary partitions SBPr and SBPc, defined in Table 5

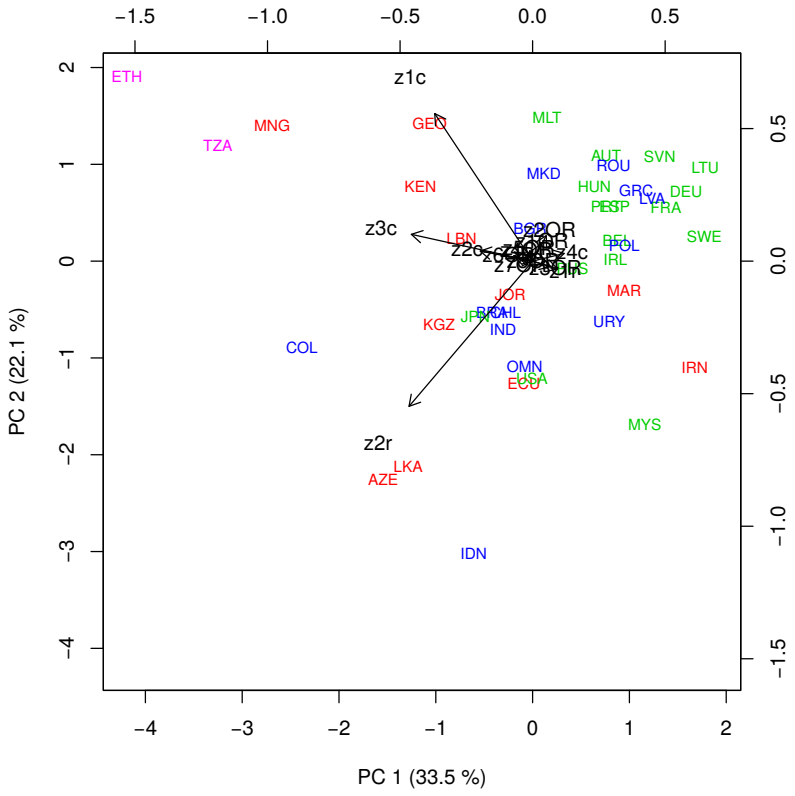


Figure 9: Covariance biplot of compositional tables in coordinates with countries classified according to level of development (Industrialized economies - green, Emerging industrial economies - blue, Developing economies - red, Least developed countries - purple).

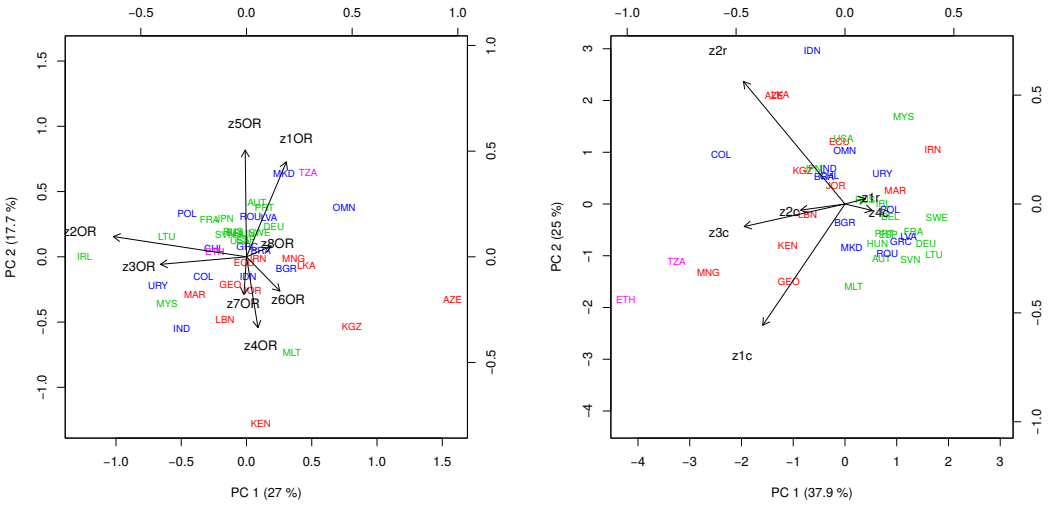


Figure 10: Covariance biplots of odds ratio coordinates (left) and balances (right) of countries classified according to level of development (Industrialized countries - green, Emerging industrial economies - blue, Developing countries - red, Least developed countries - purple)