UNIVERZITA PALACKÉHO V OLOMOUCI
Přírodovědecká fakulta
Katedra matematické analýzy a aplikací matematiky

DIZERTAČNÍ PRÁCE

Složité struktury v kompozičních datech

Vedoucí práce:
Doc. RNDr. Karel Hron, Ph.D.
Olomouc 2016

Vypracovala:
Mgr. Kamila Fačevicová
P1104 Aplikovaná matematika

PALACKÝ UNIVERSITY OLOMOUC
Faculty of Science
Department of Mathematical Analysis and Applications of Mathematics

DOCTORAL THESIS

Complex Structures of Compositional Data

Supervised by:
Doc. RNDr. Karel Hron, Ph.D.
Olomouc 2016

Written by:
Mgr. Kamila Fačevicová
P1104 Applied Mathematics

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Mgr. Kamila Fačevicová

Název práce: Složité struktury v kompozičních datech

Typ práce: Dizertační práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

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

Rok obhajoby: 2016

Abstrakt:

Dizertační práce je zaměřena na analýzu kompozičních tabulek, které představují přímé zobecnění D–složkových (vektorových) kompozičních dat. Kompoziční tabulky mohou být navíc chápány jako spojitá alternativa kontingenčních tabulek, také totiž zachycují vztah mezi dvěma faktory, založený na informaci o poměrech mezi prvky tabulky. Kvůli této relativní povaze se kompoziční tabulky (stejně jako kompoziční data obecně) řídí tzv. Aitchisonovou geometrií. Aby bylo možné použít standardní analytické metody, je potřeba tento typ dat převést prostřednictvím ortonormálních souřadnic do prostoru se standardní euklidovskou metrikou. Vyjádření v ortonormálních souřadnicích je běžně prováděno prostřednictvím tzv. postupného binárního dělení, takto získané souřadnice (bilance) však nerespektují dvojrozměrnou povahu dat obsažených v kompozičních tabulkách. Kvůli zachování informace o vztahu mezi faktory je proto v práci navržena metoda, která bilance doplňuje o souřadnice, jejichž interpretace je úzce spjatá s poměry šancí mezi skupinami prvků. Právě konstrukci těchto souřadnic a jejich interpretaci je věnována hlavní část práce. Uveden je také speciální případ těchto souřadnic (pivotové souřadnice), jehož použití je vhodné v situaci, kdy nemáme žádnou znalost o povaze řádkového a sloupcového faktoru. Představení souřadnic jako takových je doplněno o jejich varianční strukturu, která umožní lepší pochopení jejich interpretace. Teoretické aspekty problematiky jsou demonstrované na několika příkladech a pomocí ilustrací.

Klíčová slova: analýza nezávislosti, bilance, kompoziční tabulky, ortonormální souřadnice

Počet stran: 74

Počet příloh: 0

Jazyk: anglický

BIBLIOGRAPHICAL IDENTIFICATION

Author: Mgr. Kamila Fačevicová

Title: Complex Structures of Compositional Data

Type of thesis: Doctoral thesis

Department: Dept. of Mathematical Analysis and Applications of Mathematics

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

The year of presentation: 2016

Abstract:

Compositional tables can be considered as a continuous counterpart to the well-known contingency tables. Accordingly, their cells, containing in general positive real numbers rather than just counts, carry relative information about relationships between two factors. As a consequence, compositional tables can be considered as a generalization of (vector) compositional data. Due to relative character of these observations, compositions are popularly expressed in orthonormal coordinates using sequential binary partition prior to further processing using standard statistical tools. Even though the resulting coordinates (balances) are well interpretable in sense of logratio between two groups of parts, they do not respect the two-dimensional nature of compositional tables and the information about relationship between factors is thus not well captured. The main aim of the thesis is to present a general system of orthonormal coordinates with respect to the Aitchison geometry of compositional data, which enables to analyze interactions between factors in a compositional table. This is realized in sense of logarithms of odds ratios, which are popular also in context of contingency tables. Moreover, the pivot coordinate system is presented, which is useful particularly in case, when no a priori knowledge about row and column factors is available. For the sake of completeness, a part of thesis also concerns covariance structure of the coordinates that enables to understand better their interpretation. All proposed coordinate systems are illustrated by examples and graphical representations.

Key words: analysis of independence, balances, compositional tables, orthonormal coordinates

Number of pages: 74

Number of appendices: 0

Language: English

Prohlášení:

Prohlašuji, že jsem dizertační práci zpracovala samostatně pod vedením doc. RNDr. Karla Hrona, Ph.D. a všechny použité zdroje jsem uvedla v seznamu literatury.

Olomouc, 25. dubna 2016

Poděkování

Na tomto místě bych ráda poděkovala Karlu Hronovi za veškeré rady a podnětné připomínky, díky nimž mohla tato práce vzniknout. Zároveň ale děkuji za jeho důslednost, která byla hnacím motorem pro překonávání všech překážek, spojených se studiem.

Poděkování patří také mé matce a Petrovi, kteří mi jsou velkou oporou a sdíleli se mnou veškeré radosti i trápení, které s sebou studium přineslo.

V neposlední řadě bych ráda poděkovala Bětce a Kláře, s nimiž jsme před čtyřmi lety cestu k doktorskému titulu započaly a věřím, že ji spolu i zdárně dokončíme. Jejich podpora a možnost spolupráce pro mě byly velmi důležité.

Contents

Introduction

In many practical situations, the object of statistical analysis is a table representing distribution of a variable of interest, according to two (row and column) factors. If relative contributions of cells on the overall distribution are of primary interest rather than concrete absolute values which they convey, it is referred to compositional tables [7, 8]. From this perspective, compositional tables form a generalization of vector compositional data, where only ratios between parts contain all relevant information [2, 24]. Compositional tables can be thus considered as a complex structure of compositional data, whose specific nature is captured by the Aitchison geometry with the structure of finite-dimensional Euclidean vector space. Contrary to contingency tables, representing result of a multinomial sampling with cell probabilities $p_{ij} > 0, \sum_i \sum_j p_{ij} = 1$, a compositional table itself represents one observation of distribution-valued variables with some continuous multivariate distribution (e.g. relative structure of population according to social and economic status). On the other hand, compositional and contingency tables are closely linked, since the probability table with entries p_{ij}, corresponding to given contingency table, forms just a proportional representation (and thus one particular case) of compositional table, see [8] for details. Statistical analysis of contingency tables is characterized by using Pearson χ^2 statistic or log-linear models for independence testing. As these methods strongly rely on the assumption of Euclidean geometry [8] (similarly as most of standard statistical methods [3]), they are not suitable for compositional tables that are driven by the Aitchison geometry. Moreover, similarly as for compositional data, it is also natural to consider a sample of compositional tables with a possibility of their processing using popular multivariate statistical methods (like principal component analysis, clustering, classification, etc.). This is a particular difference to the case of contingency tables, where such issues are usually not of primary interest. Although one possible approach to treat a sample of contingency tables statistically is to consider three-way contingency tables [1], where the third factor would be used to construct the sample of tables, this approach does not inherently contain the case of tables with continuous origin of entries as well as a possibility of a random sample of tables. Another approach to analysis of contingency tables is represented by correspondence analysis, see, e.g. [19], for details. But again this method is not primarily designed for a sample of tables.

Taking into account the relative character and the specific geometry of compositional tables (together with replacing the arithmetic marginals by the geometric ones), the analysis of independence between factors can be performed advantageously through a decomposition of the original table into its independent and interactive parts [7, 8]. In particular, the interaction table conveys the key information for understanding the sources of association between both factors. The key point in statistical analysis of compositional tables is then (as in the case of vector compositional data) to express them in orthonormal coordinates with respect to the Aitchison geometry, where rules of the standard Euclidean geometry apply. As there is no standard canonical basis with respect to the Aitchison geometry, the main aim of this thesis is to derive interpretable coordinate representation for compositional tables. For the case of vector compositional data, it is possible to construct coordinates in sense of balances between groups of compositional parts [6]. Nevertheless, balances are not satisfactory from the perspective of compositional tables as they do not follow two-factor nature of compositional tables and their possible decomposition into independent and interactive parts.

The first part of the thesis introduces the concept of compositional data. Besides the definition of D-part compositional data, this section summarizes the basic principles of their analysis, structure of the Aitchison geometry and, finally, several coordinate systems, which allow to process them statistically in the real space.

The main part of the thesis is formed by Section 2, dealing with the compositional tables. Here the new coordinate system is proposed, which completes the balances between whole rows or columns with another group of coordinates, closely connected to odds ratios between groups of parts. Firstly, a general system of coordinates is provided, which allows to respect the nature of row and column factors, then its special case is proposed, called pivot coordinates in the following, whose construction is easier and, finally, coordinate representation of 2×2 tables as popular special case follows. All proposed methods are accompanied by several examples and illustrations, which allow their better understanding. The second part of this section completes the theory with covariance structure of all proposed coordinate systems, where the general features are more specified in the case of pivot coordinates.

The final Section 3 discusses options of analysing relationship between factors from one or a sample of compositional tables. Also this section is accompanied by examples.

The thesis summarizes results of the following papers (except the first one, all of them were proposed during my Ph.D. studies):

a) Egozcue, J. J, Díaz-Barrero, J. L. and Pawlowsky-Glahn, V. (2008). Compositional analysis of bivariate discrete probabilities. In *Proceedings of CODAWORK08, The 3rd Compositional Data Analysis Workshop.* (eds. Daunis-i-Estadella, J. and Martín-Fernández.) University of Girona, Spain.

b) Fačevicová, K., Hron, K., Todorov, V., Guo, D. and Templ, M. (2014) Logratio approach to statistical analysis of 2×2 compositional tables. *Journal of Applied Statistics*, **41**, 944–958.

c) Fačevicová, K. and Hron, K. (2015) Covariance structure of compositional tables. *Austrian Journal of Statistics*, **44**, 31–44.

d) Fačevicová, K., Hron, K., Todorov, V. and Templ, M. (2016) Compositional tables analysis in coordinates. *Scandinavian Journal of Statistics*. DOI: 10:1111/sjos.12223.

e) Fačevicová, K., Hron, K., Todorov, V. and Templ, M. (2016) General approach to coordinate representation of compositional tables. In progress.

The first manuscript establishes decomposition of a compositional table onto its independent and interactive parts. This concept is used in b), d) and e) for construction of 2×2, pivot and general coordinates. Covariance structure of coordinates is discussed in paper c).

1 Compositional data

Since $I \times J$ compositional tables represent a direct generalisation of vector compositional data, the concepts of the logratio approach to compositional data analysis can be easily adapted for compositional tables and used to derive the corresponding specific issues. Accordingly the vector compositional data are introduced first. This type of multivariate observations differs from the standard one by their relative nature, as ratios between parts are of the main interest rather than their absolute values. Compositional data frequently occur e.g. in geochemistry and the logratio methodology represents quite young and still growing statistical discipline (first analytical methods were proposed in [2]). The main principles of (vector) compositional data analysis are summarized in the following two sections.

1.1 Basic principles of compositional data

A (random) D-part composition is defined as a row vector

$$\mathbf{x} = (x_1, x_2, \ldots, x_D) \quad , \tag{1}$$

where all components (parts) describe quantitatively their relative contributions to the whole [2, 24]. Thus absolute values of parts are not of the main interest, since all the relevant information in the composition is contained in the ratios between its parts. Consequently, the composition could be rescaled (closed) to a prescribed constant sum representation $\kappa > 0$ (i.e. to 1 in case of proportions and 100 for percentages) without any loss of information; formally, we refer to a closure operation and denote

$$\mathcal{C}(\mathbf{x}) = \left(\frac{\kappa \cdot x_1}{\sum_{i=1}^{D} x_i}, \frac{\kappa \cdot x_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa \cdot x_D}{\sum_{i=1}^{D} x_i} \right) \quad . \tag{2}$$

This closed representation is useful, e.g., for a first brief comparison of two compositional vectors. The sample space of representations of D-part compositional data with an arbitrary, but fixed κ is a subset of $\mathbf{R}^D$, called D-part simplex,

$$\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, x_2, \ldots, x_D) |\ x_i > 0,\ i = 1, 2, \ldots, D;\ \sum_{i=1}^{D} x_i = \kappa \right\} \quad . \tag{3}$$

The constant sum constraint reduces the dimension of $\mathcal{S}^D$ to $D - 1$, i.e. one less than actual number of parts of the composition.

The assumption that only ratios between components carry relevant information about the composition leads to the following principles of compositional data analysis [24]. The first of them is the *scale invariance*, which means that the results of the analysis should not depend on the particular sum κ of compositional parts. Thus application of closure operation $\mathcal{C}(\mathbf{x})$ should not alter results of the analysis. Scale invariance is also related to the property of *relative scale* of compositions, since ratios should express the differences between observations rather than Euclidean distances based on absolute values of components. Next principle is called *subcompositional coherence*. As in standard statistics the results obtained from a composition with D parts should not be in contradiction with results that are obtained from a subcomposition containing d parts, $d < D$ and subcompositions should behave like orthogonal projections in real space. 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 noninformative part is removed, results should not change. The final basic principle of compositional analysis is *permutation invariance*, output of the analysis cannot depend on the order of parts in the composition.

Due to relative nature of compositional data and the above principles, the standard Euclidean geometry should be replaced by the Aitchison geometry, endowed with the Euclidean vector space structure. Accordingly, operations of perturbation and power transformation (powering) for D-part compositional vectors $\mathbf{x}$ and $\mathbf{y}$ and a real constant α are defined as

$$\mathbf{x} \oplus \mathbf{y} = (x_1 y_1, \ldots, x_D y_D) \quad \text{and} \quad \alpha \odot \mathbf{x} = (x_1^\alpha, \ldots, x_D^\alpha) \quad , \tag{4}$$

respectively. Consequently, $\mathbf{n} = \mathcal{C}\,(1, \ldots, 1)$ represents the neutral element in the $(D-1)$-dimensional vector space $(\mathcal{S}^D, \oplus, \odot)$. To complete the Euclidean vector space structure, the Aitchison inner product of two compositional vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \frac{1}{2D} \sum_{i,j} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j} \tag{5}$$

and the Aitchison norm and distance as

$$\|\mathbf{x}\|_A = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_A} \quad \text{and} \quad d_A(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_A \quad , \tag{6}$$

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

1.2 Coordinate representation of compositional data

Due to specific nature of compositional data, represented by the above principles, standard statistical methods are not suitable for their analysis. Instead of developing their counterparts within the Aitchison geometry, it seems much more intuitive to express compositions isometrically in real coordinates with respect to the Aitchison geometry and then proceed with usual statistical processing there

[24]. 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

$$\mathrm{clr}(\mathbf{x}) = \left(\ln \frac{x_1}{g(\mathbf{x})}, \ln \frac{x_2}{g(\mathbf{x})}, \ldots, \ln \frac{x_D}{g(\mathbf{x})} \right) \quad , \tag{7}$$

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, treat compositional parts symmetrically and have quite intuitive interpretation, they lead to singular covariance matrix (note that sum of clr coordinates is zero). Apart from purely geometrical disadvantages (like ambiguity of coordinate representation), this fact restricts seriously usability of clr coefficients in many statistical methods. A way out is to apply isometric logratio (ilr) coordinates [5, 6], i.e., coordinates with respect to an orthonormal basis on the simplex. According to basic algebraic-geometrical rules and dimensionality of the Aitchison geometry, the real vector $\mathbf{z} \in \mathbf{R}^{D-1}$ of ilr coordinates is defined as

$$\mathbf{z} = \mathrm{ilr}(\mathbf{x}) = \left(\left\langle \mathbf{x}, \mathbf{e}^1 \right\rangle_A, \left\langle \mathbf{x}, \mathbf{e}^2 \right\rangle_A, \ldots, \left\langle \mathbf{x}, \mathbf{e}^{D-1} \right\rangle_A \right) = (z_1, z_2, \ldots, z_{D-1}) \quad , \tag{8}$$

where $\mathbf{e}^i = \mathcal{C}\left(e_1^i, e_2^i, \ldots, e_D^i\right), i = 1, 2, \ldots, D-1$ form an orthonormal basis on the simplex. Due to isometric isomorphism of ilr coordinates it immediately follows

$$\mathrm{ilr}\left((\alpha \odot \mathbf{x}) \oplus (\beta \odot \mathbf{y})\right) = \alpha \cdot \mathrm{ilr}(\mathbf{x}) + \beta \cdot \mathrm{ilr}(\mathbf{y}), \qquad \langle \mathbf{x}, \mathbf{y} \rangle_A = \langle \mathrm{ilr}(\mathbf{x}), \mathrm{ilr}(\mathbf{y}) \rangle \quad , \tag{9}$$

$$\|\mathbf{x}\|_A = \|\mathrm{ilr}(\mathbf{x})\| \qquad \text{and} \qquad \mathrm{d}_A(\mathbf{x}, \mathbf{y}) = \mathrm{d}(\mathrm{ilr}(\mathbf{x}), \mathrm{ilr}(\mathbf{y})) \quad . \tag{10}$$

It could be also shown that different ilr coordinate systems are linked through an orthogonal transformation [5].

Clearly, it is not possible to assign an orthonormal coordinate to each of compositional parts simultaneously, like it was in the case of clr coefficients. Therefore, interpretable orthonormal coordinates are of primary interest. Since coordinates $\mathbf{z}$ correspond to a particular choice of basis vectors (compositions) $\mathbf{e}^i, i = 1, \ldots, D-1$, they can be chosen according to aim of the analysis and possible a priori knowledge about compositional parts. One popular option for construction of interpretable orthonormal coordinates is to apply sequential binary partition (SBP) procedure [6], based on stepwise division of parts into non-overlapping groups. This method represents a crucial point for the next chapter and is thus described in a detail in the following. Accordingly, in the first step of SBP, the whole composition is divided into two subcompositions. For the next step only one of subcompositions from the previous step is taken and further divided into two groups. This process continues until all groups of parts consist

Table 1: Example of sequential binary partition 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	$\frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2}$

of only one single component. The SBP is done in $D-1$ steps; in each step one basis vector $\mathbf{e}^i$ with parts

$$\begin{aligned}
e^i_j &= \exp\left(\sqrt{\tfrac{v}{u(u+v)}}\right) && \text{for} && j = j_1, \ldots, j_u, \\
e^i_k &= \exp\left(-\sqrt{\tfrac{u}{v(u+v)}}\right) && && k = k_1, \ldots, k_v, \\
e^i_{\cdot} &= \exp(0) && \text{otherwise} && ,
\end{aligned} \tag{11}$$

is obtained. 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. These basis vectors induce the final ilr coordinates

$$z_i = \sqrt{\frac{uv}{u+v}} \ln \frac{(x_{j_1} x_{j_2} \cdots x_{j_u})^{1/u}}{(x_{k_1} x_{k_2} \cdots x_{k_v})^{1/v}}, \; i = 1, \ldots, D-1 \quad . \tag{12}$$

When parts assigned to the first group are marked by $+$, parts in the second group by $-$ and parts not included in any of both groups in the i-th step of the partition by 0, SBP can be represented also graphically. Table 1 results from one possible SBP for five-part compositional data.

Orthonormal coordinates resulting from SBP (12) 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 to capture the most relevant information contained in ratios between compositional parts and their groups. For example, geochemical data are formed by major and minor elements, further divided according to concrete composition of the analyzed rock/soil. Because of this flexibility, balances form the most popular class of orthonormal coordinates that was recently successfully applied in a number of real-world studies [23].

If there are no patterns determinating the SBP, balances can be constructed as proposed in [17],

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{(x_{i+1} \cdot \cdots \cdot x_D)^{1/(D-i)}} \quad , i = 1, \ldots, D-1 \quad . \tag{13}$$

Here each step of SBP separates the i-th part of composition and coordinate z_i represents relative amount of the i-th part compared to rest of parts in the given step of SBP. This coordinate system can be thus considered as a recommended choice when no a priori information about grouping of compositional parts is known. Moreover, all relative information about x_1 is contained in the first coordinate z_1.

Inverse transformation

Since ilr coordinates $\mathbf{z}$ of compositional data result from a one-to-one mapping, it is also possible to transform them back to the D-part simplex using basis vectors $\mathbf{e}^1, \ldots, \mathbf{e}^{D-1}$ from (11). For this purpose, consider the $(D-1, D)$ dimensional matrix $\boldsymbol{\Psi}$ with rows equal to $\mathrm{clr}(\mathbf{e}^i)$. Since $\mathbf{e}^i$, $i = 1, \ldots, D-1$ form an orthonormal basis of $\mathcal{S}^D$, matrix $\boldsymbol{\Psi}$ satisfies $\boldsymbol{\Psi}\boldsymbol{\Psi}' = \mathbf{I}_{D-1}$. The inverse transformation from the $(D-1)$-dimensional real space to $\mathcal{S}^D$ is given as

$$\mathbf{x} = \mathcal{C}\left(\exp(\mathbf{z}\boldsymbol{\Psi})\right) \quad . \tag{14}$$

Consequently, the back-transformed compositional vector can be represented with an arbitrary sum of parts using the closure operation.

2 Compositional tables

Even though the theory of compositional data analysis is already well developed, it is primarily designed for vector compositional data, which carry information about relative structure according to only one factor. In cases, when compositional data carry information about distribution according to two factors (e.g. population structure according to age and BMI index), it seems to be appropriate to work with two dimensional data, which besides the relative structure contain inherently also information about relationships between these factors.

An $I \times J$ table

$$\mathbf{x} = \begin{pmatrix} x_{11} & \cdots & x_{1J} \\ \vdots & \ddots & \vdots \\ x_{I1} & \cdots & x_{IJ} \end{pmatrix} \quad , \tag{15}$$

whose cells $x_{ij} > 0$, for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots J$ convey relative contributions on a whole (probability, overall output, etc.) can be considered as a natural extension of vector compositional data and is called compositional table. From this point on, $\mathbf{x}$ will denote a $I \times J$ compositional table instead of compositional vector, unless otherwise stated. As it was mentioned above, this type of observations basically conveys relative information on relationship between two factors with I and J values, respectively. But also the other way around, by vectorization of compositional tables vector compositional data would be obtained. Therefore, any reasonable analysis of compositional tables should follow the same assumptions as analysis of compositional vectors, which were introduced in Section 1.1, just with specific (two-factor) interpretation of their parts; here a subcomposition of compositional table arises by omitting the whole row(s) and/or column(s) and it is called subtable or partial table. Note here, that on the contrary to contingency tables, containing n independent realisations of random variable from multinomial distribution, a compositional table is considered to be one realisation from a multivariate continuous distribution. On the other hand, there is quite close connection between both types of tables, since probability table, that corresponds to the contingency table, can be considered as one particular representation of compositional table. And finally, even the contingency table itself can be considered as a compositional table, if the total number of counts is high enough that its role as a source of uncertainty for estimation of the underlying probabilities is negligible.

2.1 Basic definitions

Since compositional tables (15) represent a direct extension of vector compositional data (1), all operations defined in Section 1.1 can be easily accommodated for this case. Proportional representation of a compositional table can be reached by application of closure operation with $\kappa = 1$,

$$\mathcal{C}(\mathbf{x}) = \begin{pmatrix} \frac{\kappa x_{11}}{\sum_{ij} x_{ij}} & \cdots & \frac{\kappa x_{1J}}{\sum_{ij} x_{ij}} \\ \vdots & \ddots & \vdots \\ \frac{\kappa x_{I1}}{\sum_{ij} x_{ij}} & \cdots & \frac{\kappa x_{IJ}}{\sum_{ij} x_{ij}} \end{pmatrix} \quad , \tag{16}$$

and by varying $\kappa > 0$, any other constant sum representation can be obtained. The sample space of compositional tables is again $(IJ-1)$-dimensional simplex

$$\mathcal{S}^{IJ} = \left\{ \mathbf{x} = (x_1, x_2, \ldots, x_{IJ}) |\ x_i > 0,\ i = 1, 2, \ldots, IJ;\ \sum_{i=1}^{IJ} x_i = \kappa \right\} \quad , \tag{17}$$

since each IJ-part compositional vector can be re-ordered into the form of table with I rows and J columns. On the other hand, note that the table form is appropriate only for such data, which carry information about distribution of some total with respect to two factors. Also basic operations of the Aitchison geometry should 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

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C} \begin{pmatrix} x_{11}y_{11} & \cdots & x_{1J}y_{1J} \\ \vdots & \ddots & \vdots \\ x_{I1}y_{I1} & \cdots & x_{IJ}y_{IJ} \end{pmatrix} \quad ; \tag{18}$$

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

$$\alpha \odot \mathbf{x} = \mathcal{C} \begin{pmatrix} x_{11}^{\alpha} & \cdots & x_{1J}^{\alpha} \\ \vdots & \ddots & \vdots \\ x_{I1}^{\alpha} & \cdots & x_{IJ}^{\alpha} \end{pmatrix} \tag{19}$$

is obtained. The Aitchison inner product modifies to

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \frac{1}{2IJ} \sum_{i,j} \sum_{k,l} \ln \frac{x_{ij}}{x_{kl}} \ln \frac{y_{ij}}{y_{kl}} \tag{20}$$

and the Aitchison norm and distance should be restated as follows,

$$\|\mathbf{x}\|_A = \sqrt{\frac{1}{2IJ} \sum_{i,j} \sum_{k,l} \left(\ln \frac{x_{ij}}{x_{kl}} \right)^2} \tag{21}$$

and

$$d_A(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{2IJ} \sum_{i,j} \sum_{k,l} \left(\ln \frac{x_{ij} y_{kl}}{x_{kl} y_{ij}} \right)^2} \quad . \tag{22}$$

2.2 Decomposition of compositional tables

Since the analysis of compositional tables is based on projections of the table onto subspaces with specific interpretation [7], such projections shall be introduced before we proceed to main part of the thesis, construction of orthonormal coordinates of compositional tables.

Various projections are used for different purposes in the case of compositional tables. At first, projections of a compositional table $\mathbf{x}$ onto row subspaces $\mathcal{S}^{IJ}(\text{row}_i)$, for $i = 1, \dots, I$, each with dimension $J-1$, are considered. In order to construct these projections, an orthonormal basis in $\mathcal{S}^J$ should be defined. According to [7] this basis is formed by vectors $\mathbf{e}_k = \mathcal{C}(\exp[\xi_{k1}, \dots, \xi_{kJ}]), k = 1, \dots, J-1$, where $\xi_{kj} = \ln\left(e_{kj}/g(\mathbf{e}_k)\right), j = 1, \dots, J$. Note that the row index i is suppressed because this basis remains the same for all rows. Moreover, this basis can be reached e.g. by SBP applied to levels of the column factor. Consequently, the basis of subspace $\mathcal{S}^{IJ}(\text{row}_i)$ is formed by tables

$$\mathbf{E}_{ik} = \mathcal{C}\exp\begin{pmatrix} 0 & \cdots & 0 \\ \cdots & \cdots & \cdots \\ \xi_{k1} & \cdots & \xi_{kJ} \\ \cdots & \cdots & \cdots \\ 0 & \cdots & 0 \end{pmatrix}, \quad k = 1, \dots, J-1 \quad , \tag{23}$$

where the only nonzero row is the i-th row. Finally, the projection of the compositional table $\mathbf{x}$ onto subspace $\mathcal{S}^{IJ}(\text{row}_i)$, denoted by $\text{row}_i(\mathbf{x})$ is according to [7] defined as

$$\text{row}_i(\mathbf{x}) = \bigoplus_{k=1}^{I-1} \langle \mathbf{x}, \mathbf{E}_{ik}\rangle_A \odot \mathbf{E}_{ik} \quad i = 1, \dots, I \tag{24}$$

and equals

$$\text{row}_i(\mathbf{x}) = \mathcal{C}\begin{pmatrix} g(\text{row}_i[\mathbf{x}]) & \cdots & g(\text{row}_i[\mathbf{x}]) \\ \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{iJ} \\ \cdots & \cdots & \cdots \\ g(\text{row}_i[\mathbf{x}]) & \cdots & g(\text{row}_i[\mathbf{x}]) \end{pmatrix} \quad , \tag{25}$$

where $g(\text{row}_i[\mathbf{x}])$ denotes the geometric mean of elements in the i-th row of $\mathbf{x}$. The projection onto the subspace, formed by the i-th row of the compositional table $\mathbf{x}$, $\text{row}_i[\mathbf{x}] = \mathcal{C}(x_{i1}, \dots, x_{iJ}) \in \mathcal{S}^J, i = 1, \dots, I$, is thus still a $I \times J$ compositional table $\text{row}_i(\mathbf{x})$ whose entries consist of the i-th row itself and the rest elements are equal to geometric mean of $\text{row}_i[\mathbf{x}]$.

Analogously, also projections of the compositional table $\mathbf{x}$ onto its columns, $\text{col}_j[\mathbf{x}] = \mathcal{C}(x_{1j}, \dots, x_{Ij}) \in \mathcal{S}^I, j = 1, \dots, J$, forming subspaces $\mathcal{S}^{IJ}(\text{col}_j)$ with dimension $I-1$, can be constructed. Similarly, to the case of projections onto rows, the resulting projected compositional tables $\mathrm{col}_j(\mathbf{x})$ are given by the j-th column of $\mathbf{x}$ and its geometric mean in the other parts of the table.

Orthogonality between $\mathrm{row}_i(\mathbf{x})$ and $\mathrm{row}_{i'}(\mathbf{x})$, $i \neq i'$, or between $\mathrm{col}_j(\mathbf{x})$ and $\mathrm{col}_{j'}(\mathbf{x})$, $j \neq j'$, can be proven directly using the Aitchison inner product (20) or the isometric properties of the ilr coordinates [7].

Projection onto the subspace of the i-th row results in a compositional table $\mathrm{row}_i(\mathbf{x})$ that explains the relative information (ratios) exclusively for this row. In order to complete the information about the original compositional table $\mathbf{x}$, it is necessary to introduce a projection that explains the remaining ratios between parts in different rows [17]. In other words, a projection onto the subspace of dimension $I-1$ that forms the orthogonal complement to row subspaces $\mathcal{S}^{IJ}(\mathrm{row}_i)$, $i = 1, \ldots, I$, needs to be constructed. This subspace will be denoted as $\mathcal{S}^{IJ}(\mathrm{row}^{\perp})$ and projection onto this subspace as $\mathrm{row}^{\perp}$. For this purpose consider a basis of $\mathcal{S}^{IJ}(\mathrm{row}^{\perp})$ in form

$$\mathbf{F}_k = \mathcal{C}\exp\begin{pmatrix} \nu_{1k} & \cdots & \nu_{1k} \\ \cdots & \cdots & \cdots \\ \nu_{I1} & \cdots & \nu_{Ik} \end{pmatrix}, \quad k = 1, \ldots, I-1 \quad , \tag{26}$$

where the vectors $(\nu_{1k}, \ldots, \nu_{Ik})$, for $k = 1, \ldots, I-1$ form an orthonormal basis in $\mathbf{R}^{I-1}$. The resulting projection according to [7, 8] is a compositional table

$$\mathrm{row}^{\perp}(\mathbf{x}) = \bigoplus_{k=1}^{I-1} \langle \mathbf{x}, \mathbf{F}_k \rangle_A \odot \mathbf{F}_k = \mathcal{C}\begin{pmatrix} g(\mathrm{row}_1[\mathbf{x}]) & \ldots & g(\mathrm{row}_1[\mathbf{x}]) \\ g(\mathrm{row}_2[\mathbf{x}]) & \ldots & g(\mathrm{row}_2[\mathbf{x}]) \\ \cdots & \cdots & \cdots \\ g(\mathrm{row}_I[\mathbf{x}]) & \ldots & g(\mathrm{row}_I[\mathbf{x}]) \end{pmatrix} \quad , \tag{27}$$

formed by row geometric means of the original table. Similarly, projection of $\mathbf{x}$ onto subspace orthogonal to column subspaces, $\mathcal{S}^{IJ}(\mathrm{col}^{\perp})$, of dimension $J-1$ that carries information about ratios between different columns of the original compositional table, results in

$$\mathrm{col}^{\perp}(\mathbf{x}) = \mathcal{C}\begin{pmatrix} g(\mathrm{col}_1[\mathbf{x}]) & \ldots & g(\mathrm{col}_J[\mathbf{x}]) \\ g(\mathrm{col}_1[\mathbf{x}]) & \ldots & g(\mathrm{col}_J[\mathbf{x}]) \\ \cdots & \cdots & \cdots \\ g(\mathrm{col}_1[\mathbf{x}]) & \ldots & g(\mathrm{col}_J[\mathbf{x}]) \end{pmatrix} \quad . \tag{28}$$

From their construction, projections $\mathrm{row}^{\perp}(\mathbf{x})$ and $\mathrm{col}^{\perp}(\mathbf{x})$ are orthogonal to all row or column projections, respectively, and even to each other (see [7] for proof). This fact is crucial for compositional tables analysis as it will be shown later.

Orthogonality of all row/column subspaces allows to reconstruct the original compositional table $\mathbf{x}$ using decompositions

$$\mathbf{x} = \mathrm{row}^{\perp}(\mathbf{x}) \oplus \left(\bigoplus_{i=1}^{I} \mathrm{row}_i(\mathbf{x}) \right) = \mathrm{col}^{\perp}(\mathbf{x}) \oplus \left(\bigoplus_{j=1}^{J} \mathrm{col}_j(\mathbf{x}) \right) \quad . \tag{29}$$

As mentioned above, projections $\mathrm{row}^{\perp}(\mathbf{x})$ and $\mathrm{col}^{\perp}(\mathbf{x})$ carry information exclusively about ratios between parts of different rows and columns, respectively. This information is sufficient for the reconstruction of the compositional table, when row and column factors are independent (motivated by the probabilistic sense of the formulation). This corresponds to the case when the original table can be expressed as a product of row and column (geometric) marginals of $\mathbf{x}$ [7, 8], similarly as for contingency tables [1], where arithmetic marginals are considered instead. The resulting $I \times J$ compositional table $\mathbf{x}_{ind} = \mathrm{row}^{\perp}(\mathbf{x}) \oplus \mathrm{col}^{\perp}(\mathbf{x})$, obtained as a perturbation of these two projections, is called *independence table* with entries

$$x_{ij}^{ind} = \left(\prod_{k=1}^{I} \prod_{l=1}^{J} x_{kj} x_{il} \right)^{\frac{1}{IJ}} \quad , \tag{30}$$

x_{ij} denote parts of the original compositional table $\mathbf{x}$. Since the dimensions of subspaces $\mathcal{S}^{IJ}(\mathrm{row}^{\perp})$ and $\mathcal{S}^{IJ}(\mathrm{col}^{\perp})$ are $I-1$ and $J-1$, respectively, dimension of the subspace of independence tables $\mathcal{S}^{IJ}_{\mathrm{ind}}$ equals $I+J-2$. The remaining information about the original table, i.e. about the relations between row and column factors, is contained in the *interaction table* $\mathbf{x}_{int}$, which is orthogonal to $\mathbf{x}_{ind}$ and results from the decomposition

$$\mathbf{x} = \mathbf{x}_{ind} \oplus \mathbf{x}_{int} \quad . \tag{31}$$

The interaction table can be obtained from (31) as $\mathbf{x}_{int} = \mathbf{x} \ominus \mathbf{x}_{ind}$. It also forms an $I \times J$ compositional table and its parts can be computed from the original table $\mathbf{x}$ by

$$x_{ij}^{int} = \left(\prod_{k=1}^{I} \prod_{l=1}^{J} \frac{x_{ij}}{x_{kj} x_{il}} \right)^{\frac{1}{IJ}} \quad . \tag{32}$$

From Equation (31) and orthogonality between $\mathbf{x}_{ind}$ and $\mathbf{x}_{int}$ it follows that the dimension of the subspace of interaction tables, $\mathcal{S}^{IJ}_{\mathrm{int}}$, equals $I \cdot J - 1 - (I+J-2) = (I-1)(J-1)$. In the following section, interpretable orthonormal coordinates for interaction tables will be of particular interest.

Decomposition of a 2×2 compositional table

$$\mathbf{x} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} \tag{33}$$

is discussed in detail in paper [11]. The projection onto $S^4(\mathrm{row}_1)$, is formed by the orthonormal basis composition $\mathbf{e} = \mathcal{C}\exp\left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right) \in \mathcal{S}^2$ and the orthonormal basis $\mathbf{E}_1$ in $S^4(\mathrm{row}_1)$ defined by equation (23)

$$\mathbf{E}_1 = C\mathrm{exp}\begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 0 & 0 \end{pmatrix} \quad . \tag{34}$$

By considering (25) the resulting projection equals

$$\mathrm{row}_1(\mathbf{x}) = \mathcal{C}\begin{pmatrix} \sqrt{\frac{x_{11}}{x_{12}}} & \sqrt{\frac{x_{12}}{x_{11}}} \\ 1 & 1 \end{pmatrix} = \mathcal{C}\begin{pmatrix} x_{11} & x_{12} \\ \sqrt{x_{11}x_{12}} & \sqrt{x_{11}x_{12}} \end{pmatrix} \quad . \tag{35}$$

The orthonormal basis vector

$$\mathbf{E}_2 = C\mathrm{exp}\begin{pmatrix} 0 & 0 \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \tag{36}$$

is used to construct projection of $\mathbf{x}$ onto $S^4(\mathrm{row}_2)$,

$$\mathrm{row}_2(\mathbf{x}) = \mathcal{C}\begin{pmatrix} \sqrt{x_{21}x_{22}} & \sqrt{x_{21}x_{22}} \\ x_{21} & x_{22} \end{pmatrix} \quad ; \tag{37}$$

analogously, we obtain the remaining projections

$$\mathrm{col}_1(\mathbf{x}) = \mathcal{C}\begin{pmatrix} x_{11} & \sqrt{x_{11}x_{21}} \\ x_{21} & \sqrt{x_{11}x_{21}} \end{pmatrix}, \quad \mathrm{col}_2(\mathbf{x}) = \mathcal{C}\begin{pmatrix} \sqrt{x_{12}x_{22}} & x_{12} \\ \sqrt{x_{12}x_{22}} & x_{22} \end{pmatrix} \quad . \tag{38}$$

The projection of $\mathbf{x}$ onto the complementary subspace $\mathcal{S}^4(\mathrm{row}^\perp)$, orthogonal to both $S^4(\mathrm{row}_1)$ and $S^4(\mathrm{row}_2)$, can be formed analogously. From the orthonormal basis vector

$$\mathbf{F} = \mathcal{C}\mathrm{exp}\begin{pmatrix} \frac{1}{\sqrt[4]{2}} & \frac{1}{\sqrt[4]{2}} \\ -\frac{1}{\sqrt[4]{2}} & -\frac{1}{\sqrt[4]{2}} \end{pmatrix} \tag{39}$$

(orthogonal to $\mathbf{E}_1$ and $\mathbf{E}_2$) and $\langle \mathbf{x}, \mathbf{F}\rangle_A = \frac{1}{\sqrt{2}}\ln\frac{x_{11}x_{12}}{x_{21}x_{22}}$ we immediately obtain

$$\mathrm{row}^\perp(\mathbf{x}) = \mathcal{C}\begin{pmatrix} \sqrt[4]{\frac{x_{11}x_{12}}{x_{21}x_{22}}} & \sqrt[4]{\frac{x_{11}x_{12}}{x_{21}x_{22}}} \\ \sqrt[4]{\frac{x_{21}x_{22}}{x_{11}x_{12}}} & \sqrt[4]{\frac{x_{21}x_{22}}{x_{11}x_{12}}} \end{pmatrix} = \mathcal{C}\begin{pmatrix} \sqrt{x_{11}x_{12}} & \sqrt{x_{11}x_{12}} \\ \sqrt{x_{21}x_{22}} & \sqrt{x_{21}x_{22}} \end{pmatrix} \quad . \tag{40}$$

From (28), we can also obtain the projection onto the complementary subspace to column subspaces $\mathcal{S}^4(\mathrm{col}^\perp)$

$$\mathrm{col}^\perp(\mathbf{x}) = \mathcal{C}\begin{pmatrix} \sqrt{x_{11}x_{21}} & \sqrt{x_{12}x_{22}} \\ \sqrt{x_{11}x_{21}} & \sqrt{x_{12}x_{22}} \end{pmatrix} \tag{41}$$

that carries information about relative information between columns.

Finally, the independence table corresponding to a 2×2 compositional table is obtained as

$$\mathbf{x}_{ind} = \mathcal{C}\begin{pmatrix} x_{11}\sqrt{x_{12}x_{21}} & x_{12}\sqrt{x_{11}x_{22}} \\ x_{21}\sqrt{x_{11}x_{22}} & x_{22}\sqrt{x_{12}x_{21}} \end{pmatrix} \tag{42}$$

and the interaction table results in

$$\mathbf{x}_{int} = \mathcal{C}\begin{pmatrix} \frac{1}{\sqrt{x_{12}x_{21}}} & \frac{1}{\sqrt{x_{11}x_{22}}} \\ \frac{1}{\sqrt{x_{11}x_{22}}} & \frac{1}{\sqrt{x_{12}x_{21}}} \end{pmatrix} = \mathcal{C}\begin{pmatrix} \sqrt{x_{11}x_{22}} & \sqrt{x_{12}x_{21}} \\ \sqrt{x_{12}x_{21}} & \sqrt{x_{11}x_{22}} \end{pmatrix} \quad . \tag{43}$$

2.3 Coordinate representation of compositional tables

As explained in Section 1.2, standard analytical methods cannot be applied directly for vector compositional data, they need to be expressed first in orthonormal coordinates (8). For this purpose sequential binary partition and the balance coordinates (12) were introduced. Even though in the case of vector compositional data balances have a simple interpretation in the sense of logratio between two groups of parts represented by its geometric means, logratio interpretation seems not to be appropriate for compositional tables with parts representing relationship between two factors, because it does not respect the two dimensional nature of the data. It this section, an alternative coordinate system will be introduced. The main idea of this system is to complete balances between whole rows or columns by those dealing with odds ratios between four groups of parts [1], which represent a natural extension of balances for the case of compositional tables. This has quite an intuitive motivation. Balances can be used to capture (log-)ratios within row and column factors, respectively, while odds ratios naturally link relative information between both factors. Accordingly, it is also useful to have such coordinate system, which respects the possibility of decomposition of a compositional table $\mathbf{x}$ as described in Section 2.2.

2.3.1 General coordinates

For construction of the general coordinates of $I \times J$ compositional table, consider first SBP of the whole rows (columns) of compositional table $\mathbf{x}$, denoted in the following by SBPr (SBPc). This partition is constructed with respect to nature of levels of row (column) factor and similarly as for the usual SBP, in each of $I-1$ $(J-1)$ steps, levels with some common property are separated from the others. Thus the first $I+J-2$ coordinates $\mathbf{z}^r$ and $\mathbf{z}^c$ of $I \times J$ compositional table $\mathbf{x}$ result in

$$z_i^r = \sqrt{\frac{stJ}{s+t}} \ln \frac{[g(\mathbf{x}_{j_1\cdot})\cdots g(\mathbf{x}_{j_s\cdot})]^{1/s}}{[g(\mathbf{x}_{k_1\cdot})\cdots g(\mathbf{x}_{k_t\cdot})]^{1/t}}, \quad \text{for} \quad i = 1,2,\ldots,I-1 \tag{44}$$

Table 2: Example of sequential binary partition applied to whole rows (SBPr, left table) and whole columns (SBPc, right table) of $I \times J$ 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

and

$$z_j^c = \sqrt{\frac{uvI}{u+v}} \ln \frac{[g(\mathbf{x}_{\cdot l_1}) \cdots g(\mathbf{x}_{\cdot l_u})]^{1/u}}{[g(\mathbf{x}_{\cdot m_1}) \cdots g(\mathbf{x}_{\cdot m_v})]^{1/v}}, \quad \text{for} \quad j = 1, 2, \ldots, J-1 \quad , \tag{45}$$

where s, t (u, v) are numbers of rows (columns) separated in the i-th (j-th) step of SBP, 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 $(\cdot m_1, \ldots, \cdot m_v)$ denote rows/columns involved in this step and $g(.)$ stands for the geometric mean. Steps of SBPr are denoted by Roman numerals, while those of SBPc are denoted by Arabic numerals. As an example consider a 3×5 compositional table. The corresponding six coordinates could follow SBPs from Table 2, represented also graphically in Figure 1,

$$z_1^r = \sqrt{\frac{10}{3}} \ln \frac{g(x_{1.})}{(g(x_{2.})g(x_{3.}))^{1/2}} \quad , \tag{46}$$

$$z_2^r = \sqrt{\frac{5}{2}} \ln \frac{g(x_{2.})}{g(x_{3.})} \quad , \tag{47}$$

$$z_1^c = \sqrt{\frac{18}{5}} \ln \frac{(g(x_{.1})g(x_{.2}))^{1/2}}{(g(x_{.3})g(x_{.4})g(x_{.5}))^{1/3}} \quad , \tag{48}$$

$$z_2^c = \sqrt{\frac{3}{2}} \ln \frac{g(x_{.1})}{g(x_{.2})} \quad , \tag{49}$$

$$z_3^c = \sqrt{\frac{6}{3}} \ln \frac{g(x_{.3})}{(g(x_{.4})g(x_{.5}))^{1/2}} \quad , \tag{50}$$

$$z_4^c = \sqrt{\frac{3}{2}} \ln \frac{g(x_{.4})}{g(x_{.5})} \quad . \tag{51}$$

The remaining coordinates should be orthogonal to these first $I+J-2$ ones and for their construction some generalization of SBP needs to be introduced. This generalization is based on separation of parts of the compositional table into four groups (blocks) in a systematic manner and computation of coordinates in form of logarithm of odds ratio between these four groups (marked as A (upper

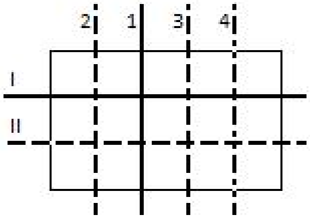


Figure 1: Graphical representation of sequential binary partitions SBPr and SBPc, applied on 3×5 compositional table.

left block), B (upper right block), C (lower left block) and D (lower right block)) using following formula

$$z^{OR} = \sqrt{\frac{a \cdot d}{a+b+c+d}} \ln \frac{\left(x_{i_1} \cdots x_{i_a}\right)^{1/a} \left(x_{l_1} \cdots x_{l_d}\right)^{1/d}}{\left(x_{j_1} \cdots x_{j_b}\right)^{1/b} \left(x_{k_1} \cdots x_{k_c}\right)^{1/c}}, \tag{52}$$

where a, b, c, d are numbers of parts in each of groups A, B, C and D, respectively and $i_., j_., k_., l_.$ are possitions of these parts. In the following steps this separation proceeds within smaller subtables according to row/column SBPs.

The separation into subgroups (A–D) and construction of the partial tables should respect the row and column grouping defined in SBPr and SBPc. Thus the first four groups are formed by steps I of SBPr and 1 of SBPc and determine the first coordinate. If the compositional table has more than four parts, the partition should continue with the next step. Firstly, a proper subtable should be identified, when the only possible partial tables are formed by pairs of groups (A,B), (C,D), (A,C) and (B,D), which should be successively analysed. If (A,B) has more than one row, the next coordinate is related to parts of this subtable, where the four groups are again determined by steps of the SBPr and SBPc of the lowest possible order. The next possible subtable is firstly searched within the current partial table, but if this one is formed by only 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 subtables with only one row or column, or subtables, which were already analysed in some of previous steps of partition, are skipped. The process continues, until all possible subtables formed by pairs of groups (A,B), (C,D), (A,C) and (B,D) of each proper partial tables are analysed. It results in $(I-1)(J-1)$ coordinates, each with interpretation in terms of log-odds ratios among groups of entries within the respective partial table. Alternatively, each coordinate could be also interpreted as a sum of log-odds ratios among four parts. There are $\binom{I}{2}\binom{J}{2}$ of them together in the whole table, each contained in one of these new coordinates. In [1] it is stated that the whole information about relations in $I \times J$ (not necessarily compositional) table is contained in $(I-1)(J-1)$ simple odds ratios of type

$$OR^1_{ij} = \frac{x_{ij}x_{i+1,j+1}}{x_{i,j+1}x_{i+1,j}}, \quad i = 1,\dots,I-1 \quad \text{and} \quad j = 1,\dots,J-1 \quad , \tag{53}$$

among neighbouring parts or of type

$$OR^2_{ij} = \frac{x_{ij}x_{IJ}}{x_{iJ}x_{Ij}}, \quad i = 1,\dots,I-1 \quad \text{and} \quad j = 1,\dots,J-1 \quad , \tag{54}$$

with a reference part x_{IJ} (both types of odds ratios are graphically illustrated in Figure 2). These basic systems of odds ratios could not be used to construct orthonormal coordinates with respect to the Aitchison geometry. In our case they are replaced by the system of $(I-1)(J-1)$ coordinates $\mathbf{z}^{OR}$, whose idea of aggregating the information into odds ratio among four groups of parts (not just four parts) seems to be similar to the concept of cumulative odds ratio as proposed in [1], page 276.

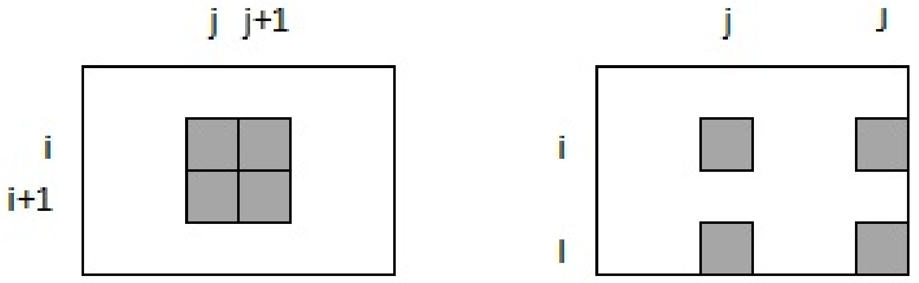


Figure 2: Graphical representation of basic odds ratio systems proposed in [1]. System of type (53) left and system of type (54) right.

Construction of partial tables and coordinates can be also considered simply as a result of combination of row and column SBPs. Although the above description shows how the coordinates are naturally derived, the output can be summarized as follows. For the first step of SBPr to rows of the table, all $J-1$ steps of SBPc to columns are performed and according to (52) the first $J-1$ coordinates are obtained. The next $J-1$ coordinates are obtained from application of the second step of SBPr to rows and all steps of SBPc to columns, and so on until $I-1$ steps of SBPr are run out. All $(I-1)(J-1)$ coordinates of $\mathbf{z}^{OR}$ thus result from successive application of all steps of SBPr combined with repeated use of all steps of SBPc, or conversely.

For the sake of completeness, the basis vectors from (8) corresponding to proposed coordinates are

$$\mathbf{e}_i^r \quad \text{with parts} \quad \begin{cases} \exp\left(\sqrt{\frac{t}{Js(s+t)}}\right) & \text{for rows} \qquad j_1,\dots,j_s, \\ \exp\left(-\sqrt{\frac{s}{Jt(s+t)}}\right) & \qquad\qquad\quad k_1,\dots,k_t, \\ \exp(0) & \text{otherwise} \quad , \end{cases} \tag{55}$$

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

$$\mathbf{e}_j^c \quad \text{with parts} \quad \begin{cases} \exp\left(\sqrt{\frac{v}{Iu(u+v)}}\right) & \text{for columns} \quad l_1,\dots,l_u, \\ \exp\left(-\sqrt{\frac{u}{Iv(u+v)}}\right) & \qquad\qquad\quad m_1,\dots,m_v, \\ \exp(0) & \text{otherwise} \quad , \end{cases} \tag{56}$$

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

$$\mathbf{e}_k^{OR} \quad \text{with parts} \quad \begin{cases} \exp\left(\sqrt{\frac{d}{a(a+b+c+d)}}\right) & \text{for group} \qquad A, \\ \exp\left(-\frac{1}{b}\sqrt{\frac{ad}{a+b+c+d}}\right) & \qquad\qquad\qquad B, \\ \exp\left(-\frac{1}{c}\sqrt{\frac{ad}{a+b+c+d}}\right) & \qquad\qquad\qquad C, \\ \exp\left(\sqrt{\frac{a}{d(a+b+c+d)}}\right) & \qquad\qquad\qquad D, \\ \exp(0) & \text{otherwise} \quad , \end{cases} \tag{57}$$

for $k = 1,\dots,(I-1)(J-1)$, 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 was described above.

Beside the advantageous interpretation, there is another useful feature of this coordinate system. When the coordinate representation $\mathbf{z}^r = (z_1^r,\dots,z_{I-1}^r)$, $\mathbf{z}^c = (z_1^c,\dots,z_{J-1}^c)$, $\mathbf{z}^{OR} = (z_1^{OR},\dots,z_{(I-1)(J-1)}^{OR})$ is applied to the independence table $\mathbf{x}_{ind}$, the only nonzero coordinates are z_i^r, z_j^c for $i = 1,\dots,I-1$, $j = 1,\dots,J-1$, and their values are the same as for the original table $\mathbf{x}$. Moreover, the number of these nonzero coordinates equals to dimension of subspace of independence tables (see e.g. [11] for details). Analogous feature holds also for the interaction table and coordinates $\mathbf{z}^{OR}$. Accordingly, the vector of coordinates $(\mathbf{z}^r, \mathbf{z}^c, \mathbf{0}_{(I-1)(J-1)})$ of the independence table can be denoted as $\mathbf{z}_{ind}$ and coordinates $(\mathbf{0}_{I+J-2}, \mathbf{z}^{OR})$ of the interaction table as $\mathbf{z}_{int}$. Finally, the vector of coordinates of the original compositional table $\mathbf{x}$ can be written as $\mathbf{z} =$

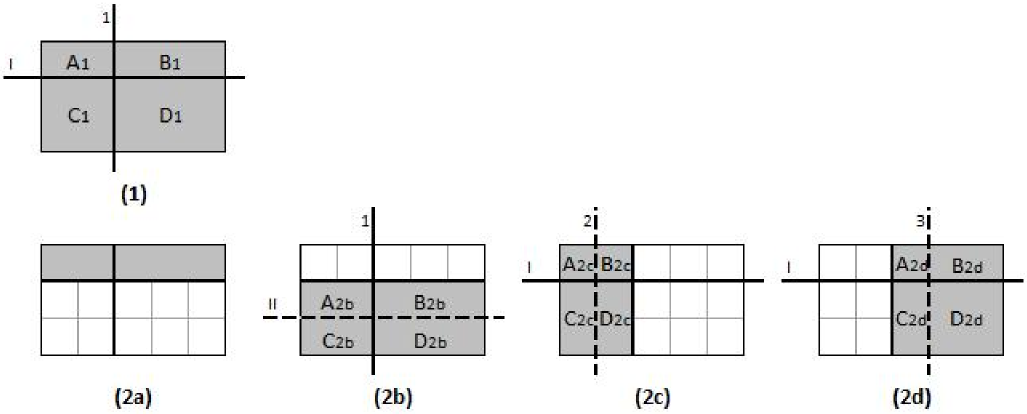


Figure 3: Graphical representation of group separation in the 3 × 5 table **(1)**. Lower grey tables **(2a-d)** illustrate construction of possible subtables. New coordinates could be computed only from tables **(2b-d)**.

$\mathrm{ilr}(\mathbf{x}_{ind}) + \mathrm{ilr}(\mathbf{x}_{int}) = \mathbf{z}_{ind} + \mathbf{z}_{int} = (\mathbf{z}^r, \mathbf{z}^c, \mathbf{z}^{OR})$. This feature will be utilized especially in the final Section 3 about analysis of relationship between two factors.

In the following a simple example for the case of 3 × 5 compositional table is presented, which should illustrate the algorithm of construction of the partial tables and the respective coordinates.

Example 1 (Coordinate representation of 3 × 5 compositional table) The first step of coordinate representation is to define SBP of the whole rows and columns. As an example, SBPs from Table 2 and Figure 1 and the respective coordinates are used. The second step is to define the subtables and compute the remaining eight coordinates according to formula (52). Firstly, the whole table is 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, is illustrated in Figure 3 (table **(1)**). According to this separation, the first coordinate is computed as

$$z_1^{OR} = \frac{2\sqrt{5}}{5} \ln \frac{(x_{11}x_{12})^{1/2}\,(x_{23}x_{24}x_{25}x_{33}x_{34}x_{35})^{1/6}}{(x_{13}x_{14}x_{15})^{1/3}\,(x_{21}x_{22}x_{31}x_{32})^{1/4}} \quad . \tag{58}$$

Next partial tables are formed by parts from 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)** is formed by an only single row, it cannot be further divided and thus we skip it and start to analyse the next possible partial table **(2b)**. This subtable has already more than one row and column, thus it represents the first partial table generating one of the coordinates. Column separation within this table still constitutes the step 1 from SBPc. In SBPr, the second and the third row of the compositional tables were separated by step II, thus the four groups in this

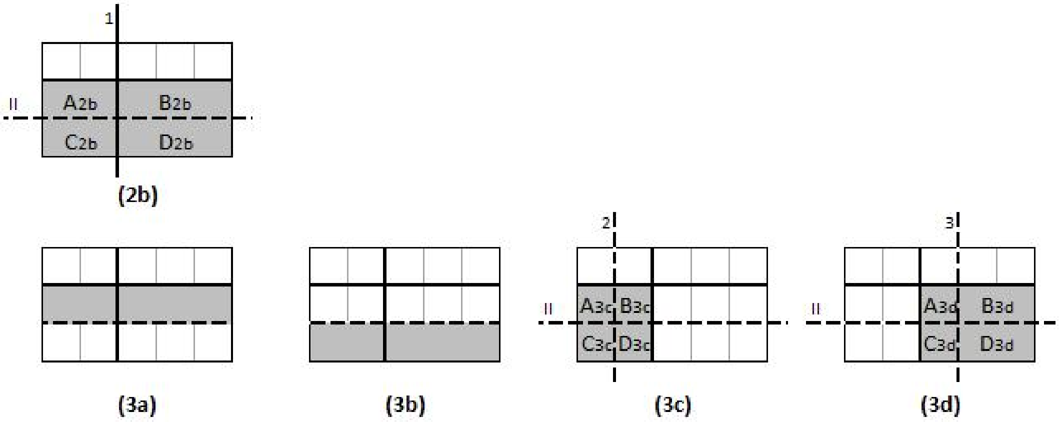


Figure 4: Graphical representation of group separation in the subtable **(2b)**. Lower grey tables **(3a-d)** illustrate construction of possible subtables. New coordinates could be computed only from tables **(3c)** and **(3d)**.

partial table are separated by steps 1 and II and according to this separation, the next coordinate could be computed in form

$$z_2^{OR} = \sqrt{\frac{3}{5}} \ln \frac{\left(x_{21}x_{22}\right)^{1/2} \left(x_{33}x_{34}x_{35}\right)^{1/3}}{\left(x_{23}x_{24}x_{25}\right)^{1/3} \left(x_{31}x_{32}\right)^{1/2}} \quad . \tag{59}$$

This table could be further separated and the next two coordinates are related to subtables **(3c)** (formed by groups A_{2b} and C_{2b}) and **(3d)** (groups B_{2b} and D_{2b}), since tables **(3a)** and **(3b)** are formed by an only single row, as is evident from Figure 4. In table **(3c)** is partition of rows already set by step II of SBPr, furthermore, columns are separated by step 2 of SBPc. Now, the next coordinate

$$z_3^{OR} = \frac{1}{2} \ln \frac{x_{21}x_{32}}{x_{22}x_{31}} \tag{60}$$

could be computed. Since each group in this table is formed by only one single part x_{21}, x_{22}, x_{31} and x_{32}, this table cannot be further partitioned and we can focus on the partial table **(3d)**. Also in this table is row separation determined by step II of SBPr and the columns are here separated by step 3 of SBPc. After assessment of the next coordinate

$$z_4^{OR} = \frac{\sqrt{3}}{3} \ln \frac{x_{23}\left(x_{34}x_{35}\right)^{1/2}}{\left(x_{24}x_{25}\right)^{1/2} x_{33}} \quad , \tag{61}$$

the subtable **(3d)** can be further divided. Figure 5 typifies all possible subtables.

Only subtable **(4d)** has more than one row and column and, consequently, could be used for construction of the next coordinate. With respect to steps II

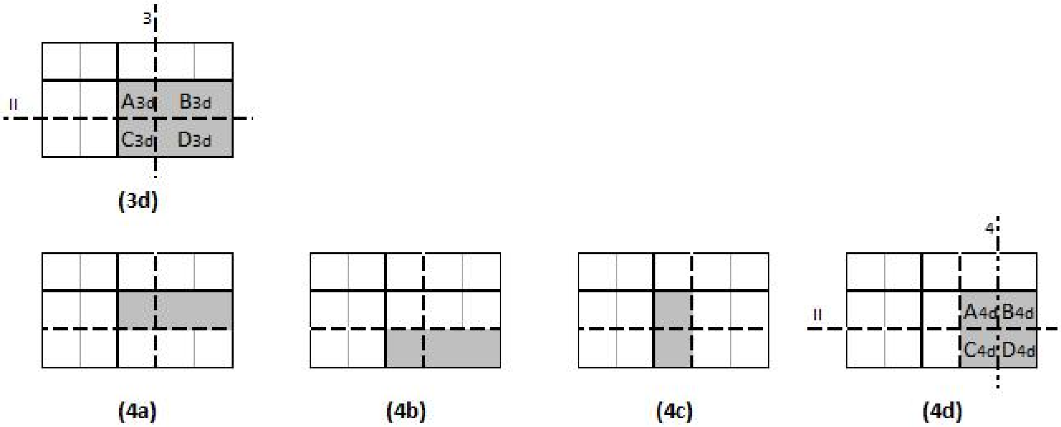


Figure 5: Graphical representation of group separation in the subtable **(3d)**. Lower grey tables **(4a-d)** illustrate construction of possible subtables. New coordinates could be computed only from table **(4d)**.

and 4 of SBPr and SBPc this coordinate is

$$z_5^{OR} = \frac{1}{2} \ln \frac{x_{24}x_{35}}{x_{25}x_{34}} \quad . \tag{62}$$

Hereby is finished the partition of subtable **(2b)** and we can return to partial table **(2c)** (Figures 3 and 6). This table, separated by steps I and 2, determines the next coordinate

$$z_6^{OR} = \frac{\sqrt{3}}{3} \ln \frac{x_{11}\left(x_{22}x_{32}\right)^{1/2}}{x_{12}\left(x_{21}x_{31}\right)^{1/2}} \tag{63}$$

and the only regular partial table contained within it is **(5b)**. But since this table is identical with table **(3c)** and was already analyzed, we can immediately skip to partial table **(2d)**. This subtable is divided by steps I and 3 of SBPr and SBPc, thus the next coordinate is

$$z_7^{OR} = \frac{2}{3} \ln \frac{x_{13}\left(x_{24}x_{25}x_{34}x_{35}\right)^{1/4}}{\left(x_{14}x_{15}\right)^{1/2}\left(x_{23}x_{33}\right)^{1/2}} \quad . \tag{64}$$

The consequent possible partial tables are illustrated in Figure (7), which clearly shows, that the only regular tables are **(6b)** and **(6d)**. Since **(6b)** is similar to **(3b)**, the last coordinate is based on subtable **(6d)** and steps I and 4.

$$z_8^{OR} = \frac{\sqrt{3}}{3} \ln \frac{x_{14}\left(x_{25}x_{35}\right)^{1/2}}{x_{15}\left(x_{24}x_{34}\right)^{1/2}} \quad . \tag{65}$$

For completeness, Figure (8) illustrates partition of this table, which leads to partial table **(7b)**, similar to **(4d)**.

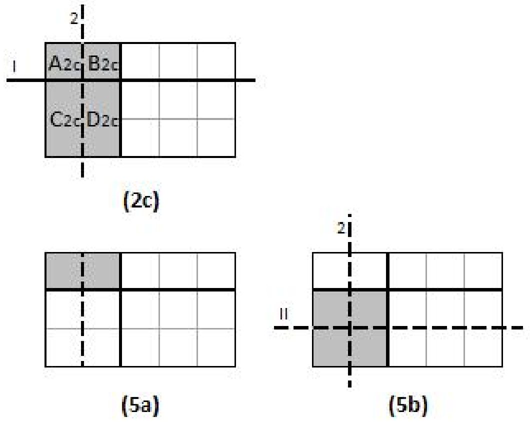


Figure 6: Graphical representation of group separation in the subtable **(2c)**. Lower grey tables **(5a-d)** illustrate construction of possible subtables. The only regular subtable is table **(5b)**, which was already analyzed (table **(3c)**).

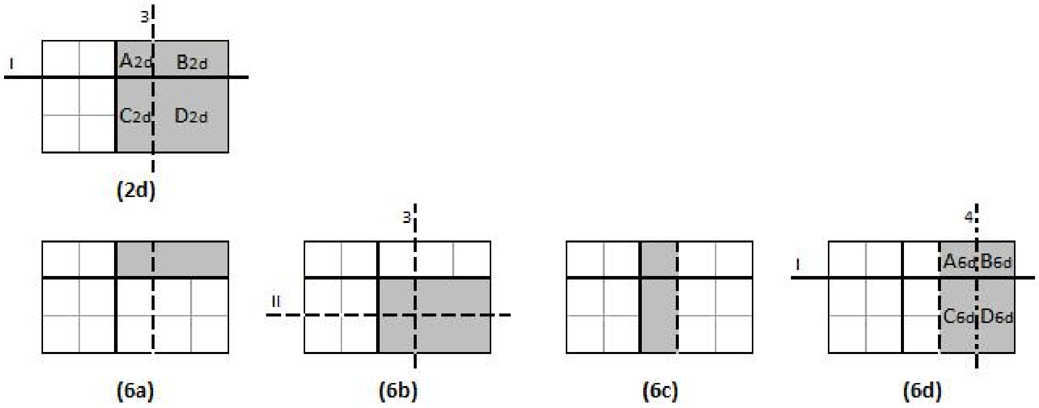


Figure 7: Graphical representation of group separation in the subtable **(2d)**. Lower grey tables **(6a-d)** illustrate construction of possible subtables. The only regular subtable is table **(6b)**, which was already analyzed as table **(3b)** and table **(6d)**, which forms the last coordinate z_8^{OR}.

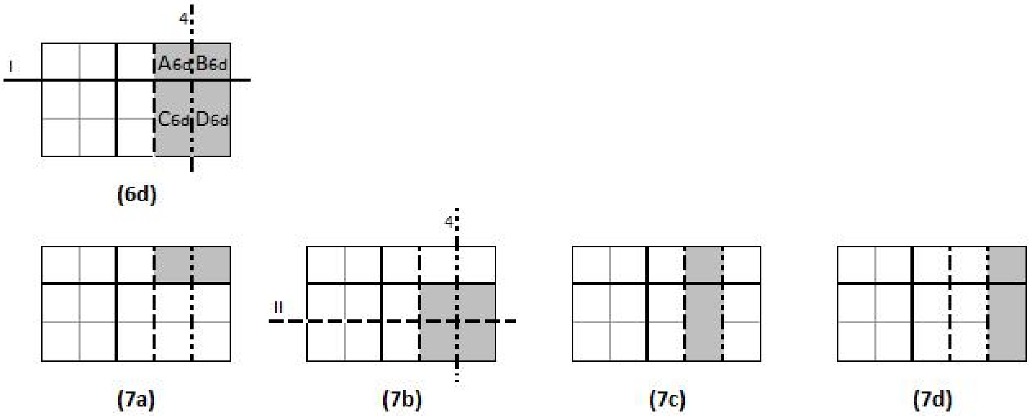


Figure 8: Graphical representation of group separation in the subtable **(6d)**. Lower grey tables **(7a-d)** illustrate construction of possible subtables. The only regular subtable is table **(7b)**, which was already analyzed as table **(4d)**.

Table 3: Houston criminal cases in January 2014, distributed according to type of offense and locality.

	Ag. assault and rape	Robbery	Burglary	Auto theft	Theft
Centre	146	229	467	243	1380
Outskirts	425	610	1078	691	3001
Peripheral	193	183	399	251	2032

Example 2 (Houston criminality) For better understanding of the interpretation of the proposed coordinates, consider Table 4. This 3×5 compositional table represents the distribution of all criminal cases (except murders) which happened in Houston, Texas, in January 2014 structured according to the type of offense and locality. The values of the type of offense are aggravated assault and rape, robbery, burglary, auto theft and theft. The second factor is defined by the distance from the city centre. The offenses which perpetrated in the inner part of the loop formed by the road 610 (see Figure 9) are classified in the first category Centre. The second category Outskirts contains offenses which happened between 610 and beltway 8 and the remaining cases are collected in the category Peripheral. The original data are available in the database of Houston Government [20].

In this example only the relative structure of the criminality is of the interest. Consequently, Table 4 can be treated as compositional table with proportional representation and with respect to Figure 1 expressed in coordinates, which values are collected in Table 5.

The first two values represent balances between rows of the table. The high negative value of z_1^r means that there is much more criminality cases out of the city

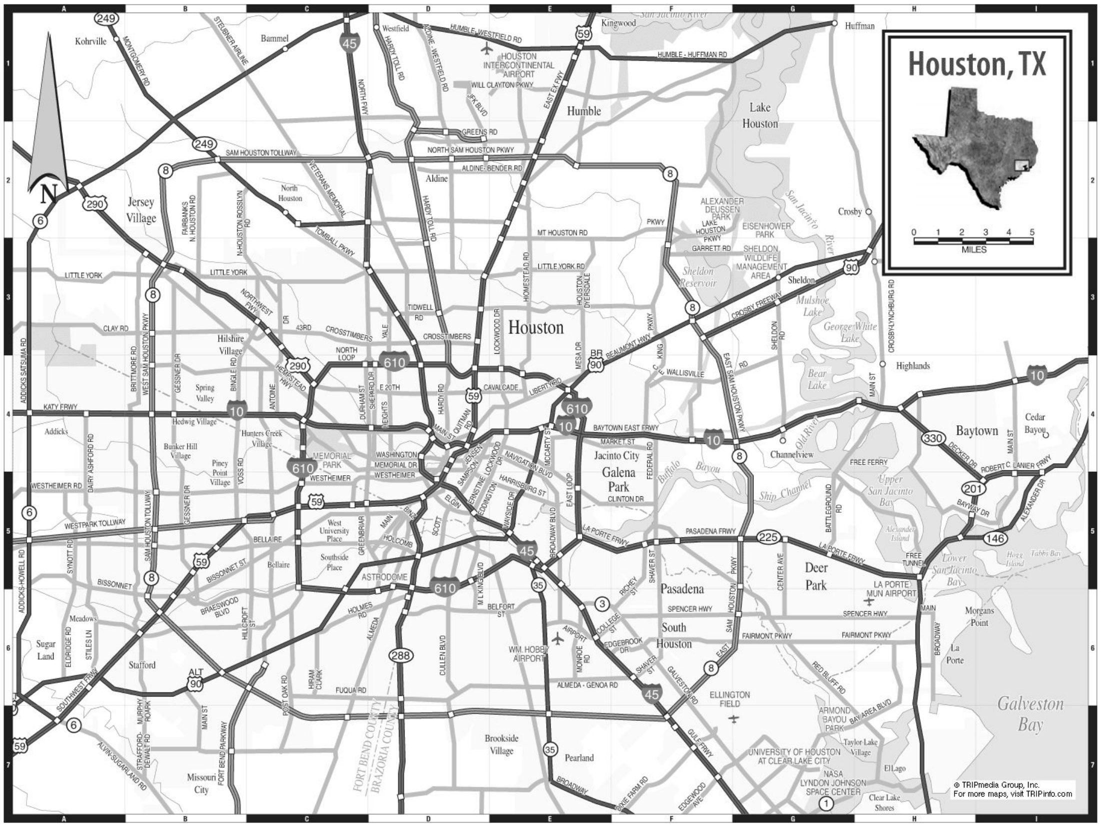


Figure 9: Road network in Houston.

centre. On the other hand, positive value of z_2^r captures higher occurrence of criminality in outskirts compared to peripheral. Coordinates $z_1^c - z_4^c$ compare different types of offenses. Coordinate z_1^c is balance between violent (aggravated assault and rape, robbery) and other crimes (burglary, auto theft, theft), when, according to its high negative value, non-violent crimes outbalance. Because of the negative value of the next coordinate, we can also conclude that there is slightly more robberies than aggravated assaults. Proportion of burglaries and thefts compares coordinate z_3^c, which is also negative and thus the amount of thefts dominates the amount of burglaries. Finally, the coordinate z_4^c captures that there is more cases of thefts then auto thefts.

Table 4: Relative structure of Houston criminal cases in January 2014, distributed according to type of offense and locality.

	Ag. assault and rape	Robbery	Burglary	Auto theft	Theft
Centre	0.0129	0.0202	0.0412	0.0215	0.1218
Outskirts	0.0375	0.0538	0.0952	0.0610	0.2649
Peripheral	0.0170	0.0162	0.0352	0.0222	0.1794

Table 5: Coordinate representation of Table 4.

z_1^r	z_2^r	z_1^c	z_2^c	z_3^c	z_4^c	z_1^{OR}
-12.53	1.39	-16.55	-0.31	-10.04	-2.16	-0.03
z_2^{OR}	z_3^{OR}	z_4^{OR}	z_5^{OR}	z_6^{OR}	z_7^{OR}	z_8^{OR}
0.15	-0.21	0.17	0.31	-0.17	0.15	0.02

The interactions between location and type of offense are covered with coordinates $z_1^{OR} - z_8^{OR}$. The most complex is the first coordinate, which is formed by logarithm of odds ratio between violent and non-violent crimes in and outside the city centre. Because of its negative value, the ratio between violent and non-violent crimes is slightly higher outside the city centre. On the other hand, the simple odds ratio between these four groups is 0.97 and thus the difference between locations is really small. The second possible interpretation of this coordinate is in sense of comparison of ratios between cases appearing in and outside the city centre according to type of crime. Now the negative value means, that the ratio between cases in and outside the city centre is higher in case of non-violent crimes.

More detailed comparison offer the other odds ratio coordinates. The second one again compares violent and non-violent crimes but only in outskirts and peripheral area. This coordinate is positive and thus the ratio between violent and non-violent crimes is slightly higher in outskirts. Aggravated assaults and robberies in outskirts and peripheral area compares coordinate z_3^{OR}, which is again negative and thus the ratio between ag. assaults and robberies in outskirts dominates the same in peripheral area. It could be easily calculated that the ratio in peripheral area is about 1.5 times higher then in outskirts. Interpretation of the remaining coordinates is similar. They are formed respectively by logarithms of odds ratio between burglary and (auto and normal) thefts in outskirts and peripheral area (z_4^{OR}), auto and normal thefts in these two areas (z_5^{OR}), ag. assault and robbery in and outside the city centre (z_6^{OR}), burglary and thefts in and outside the city centre (z_7^{OR}) and, finally, auto and normal theft in and outside the city centre (z_8^{OR}).

2.3.2 Pivot coordinates

In the case, when there are no clues, how to form groups within the row and column factor, a special case of the general coordinates can be considered, which was introduced in [13]. This coordinate system can be applied to each compositional table almost automatically in the situation, when interpretation of the coordinates is not the main goal of the analysis (like outlier detection or classification of observations). Due to this feature, it represents a two-factorial alternative to balances (13). On the other hand, such coordinates obviously still follow the decomposition (31).

The main idea by construction of these coordinates is that in each partial table the group D is formed by only single part (pivot), which is denoted as x_{rs} and which gave the note to these coordinates. In order to construct such coordinates, the basis vectors of coordinates $z_i^r, i = 1, \dots, I-1$ and $z_j^c, j = 1, \dots, J-1$ must be defined as follows,

$$\mathbf{e}_i^r \quad \text{with parts} \quad \begin{cases} \exp\left(\sqrt{\frac{I-i}{J(I-i+1)}}\right) & \text{for rows} \qquad I-i+1, \\ \exp\left(-\sqrt{\frac{1}{(I-i+1)(IJ-iJ)}}\right) & \qquad\qquad\quad 1, \dots, I-i, \\ \exp(0) & \text{otherwise} \quad , \end{cases} \tag{66}$$

and

$$\mathbf{e}_j^c \quad \text{with parts} \quad \begin{cases} \exp\left(\sqrt{\frac{J-j}{I(J-j+1)}}\right) & \text{for columns} \quad J-j+1, \\ \exp\left(-\sqrt{\frac{1}{(J-j+1)(IJ-jI)}}\right) & \qquad\qquad\quad 1, \dots, J-j, \\ \exp(0) & \text{otherwise} \quad . \end{cases} \tag{67}$$

Consequently, the first $I+J-2$ coordinates are

$$z_i^r = \sqrt{\frac{(I-i)J}{I-i+1}} \ln \frac{g(\mathbf{x}_{I-i+1.})}{[g(\mathbf{x}_{1.}) \cdots g(\mathbf{x}_{I-i.})]^{1/(I-i)}}, \quad \text{for} \quad i = 1, \dots, I-1 \tag{68}$$

(for rows), and

$$z_j^c = \sqrt{\frac{I(J-j)}{J-j+1}} \ln \frac{g(\mathbf{x}_{J-j+1.})}{[g(\mathbf{x}_{.1}) \cdots g(\mathbf{x}_{.J-j})]^{1/(J-j)}}, \quad \text{for} \quad j = 1, \dots, J-1 \tag{69}$$

(for columns), respectively. These orthonormal coordinates form again nonzero coordinate representation for the independence table and their number reflects the dimension of $\mathcal{S}_{\text{ind}}^{IJ}$. Because of mutual orthogonality of the subspaces corresponding to tables $\text{row}^{\perp}(\mathbf{x})$, $\text{col}^{\perp}(\mathbf{x})$ and $\mathbf{x}_{int}$, and decomposition (31), the remaining $(I-1)(J-1)$ coordinates of $\mathbf{x}_{ind}$ are equal to zero. Conversely, coordinate representation of the interaction table results in zero coordinates of the corresponding independence table.

In contrast to the general method, it is easier to start construction of partial tables from the smallest one in the upper left corner of the table $\mathbf{x}$. Each consequent table is then formed by the current one expanded by one row, or column. The first two steps of this stepwise procedure are as follows. The method firstly assigns a basis compositional vector to the table given only by parts x_{11}, x_{12}, x_{21} and x_{22}. This basis element compares parts on the main diagonal x_{11}, x_{22} with parts on the minor diagonal x_{12}, x_{21} of the 2×2 partial table and thus the first basis composition has the form

$$\mathbf{e}^{22} = \exp\left(\frac{1}{2}, -\frac{1}{2}, 0, \ldots, -\frac{1}{2}, \frac{1}{2}, 0, \ldots\right) \quad , \tag{70}$$

where the upper index expresses the dimension of the current partial table as well as position of the pivot part forming group D. This notation is thus used in the following instead of e_k^{OR} taken for the general case. Obviously, an odds-ratio interpretation of the resulting coordinate is again possible. In the next step the third column is added to the previous partial table and the basis vector $\mathbf{e}^{23}$ deals with the new partial table with $r = 2$ rows and $s = 3$ columns and parts $x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{23}$. The corresponding basis element compares again parts on the main diagonal of a virtual 2×2 table with parts on the minor diagonal, when these diagonals are formed by geometric mean of x_{11} and x_{12} (that thus merges information on the employed components together), and part x_{23}, and by geometric mean of x_{21} and x_{22}, and part x_{13}, respectively. This results in

$$\mathbf{e}^{23} = \exp\left(\frac{1}{2\sqrt{3}}, \frac{1}{2\sqrt{3}}, -\frac{1}{\sqrt{3}}, 0, \ldots, -\frac{1}{2\sqrt{3}}, -\frac{1}{2\sqrt{3}}, \frac{1}{\sqrt{3}}, 0, \ldots\right) \quad . \tag{71}$$

In general, the basis composition $\mathbf{e}^{rs}$ compares parts on the main diagonal (formed by geometric mean of all parts at rows of order smaller than r and column of order smaller than s and by pivot part x_{rs}) and parts on the minor diagonal (formed by geometric mean of the first $s-1$ parts of the r-th row and by geometric mean of the first $r-1$ parts of the s-th column). This resulting basis vector is

$$\mathbf{e}^{rs} \quad \text{with parts} \quad \begin{cases} \exp\left(\sqrt{\frac{1}{rs(r-1)(s-1)}}\right) & \text{for positions} \quad i = 1, \ldots, r-1, \\ & \qquad\qquad\qquad j = 1, \ldots, s-1, \\ \exp\left(-\sqrt{\frac{r-1}{rs(s-1)}}\right) & \qquad\qquad\qquad i = r, \\ & \qquad\qquad\qquad j = 1, \ldots, s-1, \\ \exp\left(-\sqrt{\frac{s-1}{rs(r-1)}}\right) & \qquad\qquad\qquad i = 1, \ldots, r-1, \\ & \qquad\qquad\qquad j = s, \\ \exp\left(\sqrt{\frac{(r-1)(s-1)}{rs}}\right) & \qquad\qquad\qquad i = r, \\ & \qquad\qquad\qquad j = s, \\ \exp(0) & \text{otherwise} \quad , \end{cases} \tag{72}$$

where the upper index represents the particular choice of $r = 2, 3, \ldots, I$ and $s = 2, 3, \ldots, J$ and the parts of $\mathbf{e}^{rs}$ are arranged as follows,

$$\mathbf{e}^{rs} = (e_{11}^{rs}, \ldots, e_{1J}^{rs}, \ldots, e_{I1}^{rs}, \ldots, e_{IJ}^{rs}) \quad . \tag{73}$$

This procedure continues until $r = I$ and $s = J$, accordingly a system of $(I-1)(J-1)$ basis vectors is obtained.

For example, the basis of 2×3 compositional tables contains compositions

$$\begin{aligned}
\mathbf{e}^{22} &= \exp\left(1/2, -1/2, 0, -1/2, 1/2, 0\right) \quad , & (74)\\
\mathbf{e}^{23} &= \exp\left(1/2\sqrt{3}, 1/2\sqrt{3}, -1/\sqrt{3}, -1/2\sqrt{3}, -1/2\sqrt{3}, 1/\sqrt{3}\right) \quad , & (75)\\
\mathbf{e}^{\mathrm{r}}_1 &= \exp\left(-1/\sqrt{6}, -1/\sqrt{6}, -1/\sqrt{6}, 1/\sqrt{6}, 1/\sqrt{6}, 1/\sqrt{6}\right) \quad , & (76)\\
\mathbf{e}^{\mathrm{c}}_1 &= \exp\left(-1/2\sqrt{3}, -1/2\sqrt{3}, 1/\sqrt{3}, -1/2\sqrt{3}, -1/2\sqrt{3}, 1/\sqrt{3}\right) \quad , & (77)\\
\mathbf{e}^{\mathrm{c}}_2 &= \exp\left(-1/2, 1/2, 0, -1/2, 1/2, 0\right) \quad . & (78)
\end{aligned}$$

Finally, the basis vectors $\mathbf{e}^{rs}$ lead to nonzero coordinates of the interaction table (out of $IJ-1$) and thus to the remaining group of coordinates of the compositional table $\mathbf{x}$

$$z_{rs} = \frac{1}{\sqrt{r \cdot s \cdot (r-1) \cdot (s-1)}} \ln \prod_{i=1}^{r-1} \prod_{j=1}^{s-1} \frac{x_{ij} x_{rs}}{x_{is} x_{rj}} \quad , \tag{79}$$

for $r = 2, 3, \ldots, I$ and $s = 2, 3, \ldots, J$. Although the above formula is advantageous for interpretation purposes, in practice it is easier to compute coordinates of the interaction table from the following modified formula with expanded products

$$\frac{1}{\sqrt{r \cdot s \cdot (r-1) \cdot (s-1)}} \ln \frac{x_{11} x_{12} \cdots x_{1,s-1} \cdots x_{r-1,1} \cdots x_{r-1,s-1} x_{rs}^{(r-1)(s-1)}}{x_{r1}^{r-1} \cdots x_{r,s-1}^{r-1} x_{1s}^{s-1} \cdots x_{r-1,s}^{s-1}} \tag{80}$$

for $r = 2, 3, \ldots, I$ and $s = 2, 3, \ldots, J$. Note that, even though x_{ij}'s in both formulas stand for parts of the original table $\mathbf{x}$, the result would not change if they are replaced by parts of the interaction table $\mathbf{x}_{int}$.

Another useful property of these coordinates is that they contain also the nonzero coordinates of the interaction tables of all tables with sizes smaller than the considered $I \times J$ table. For example, the set of four nonzero coordinates of 3×3 interaction table contains two nonzero coordinates of the 2×3 table as well as of the 3×2 table and in turn both (as well as 3×3 table) contain the only nonzero ilr coordinate of the 2×2 interaction table.

Moreover, the interpretability of these coordinates is still supported by their relation to odds ratios of parts in the original table ([1], p. 44). This fact is obvious directly from the form of (79) since each coordinate is formed by the sum of logarithms of odds ratios which compares cell of the original table in position (r, s) with all cells that are north-west from the r-th row and s-th column - group A (this feature will be thoroughly analyzed in the next example).

Table 6: Structure of population in the Czech Republic in 2008 according to age and BMI index (in proportions).

CZE	under	normal	over	obesity
$25-44$	0.0144	0.2196	0.1410	0.0554
$45-64$	0.0022	0.1014	0.1792	0.0988
$65-84$	0.0014	0.0473	0.0900	0.0493

Although the pivot coordinate system is proposed particularly for the cases, when the interpretation of single coordinates is not the main goal of the analysis, a new set of coordinates (with different interpretation) can be reached by permutation of rows and/or columns in the original compositional table. Accordingly, e.g., orthonormal coordinates that contain log odds ratio of a given 2×2 table can be easily constructed. They also enable to extract the only coordinate with log odds ratio interpretation that contains a given entry x_{rs}.

Example 3 (Relationship between age and BMI index - part 1) The pivot coordinates and their interpretation are illustrated with an example analyzing the relationship between age and BMI index in 18 European countries [9, 10]. For this purpose a sample of 3×4 compositional tables was collected. Each of the tables records the population structure of a country in 2008 according to age and BMI index ((weight in kg)/(height in m)^2). The two factors to be considered correspond to the age classes $25-44, 45-64, 65-84$ and their BMI index in categories underweight, normal, overweight and obesity respectively. Note that finer categories of age are available, but the chosen classes lead to better interpretability. Table 6 shows an example of a compositional table from the sample from Czech Republic.

Applying Equation (30), the values of the independence table are

$$\mathbf{x}_{ind} = \begin{pmatrix} 0.0061 & 0.1716 & 0.2218 & 0.1090 \\ 0.0039 & 0.1090 & 0.1409 & 0.0692 \\ 0.0020 & 0.0569 & 0.0736 & 0.0361 \end{pmatrix} . \tag{81}$$

Using Equation (32) the interaction table can be obtained,

$$\mathbf{x}_{int} = \begin{pmatrix} 0.1813 & 0.0973 & 0.0483 & 0.0387 \\ 0.0444 & 0.0707 & 0.0967 & 0.1085 \\ 0.0541 & 0.0632 & 0.0930 & 0.1037 \end{pmatrix} . \tag{82}$$

Note that these tables follow the condition $\mathbf{x}_{ind} \oplus \mathbf{x}_{int} = \mathbf{x}$.

In order to express the independence table in coordinates, two SBPs according to Table 7 were introduced.

The steps of SBP1 result in the first two nonzero coordinates of the independence table that contain relative information (ratios) between different rows of $\mathbf{x}$.

Table 7: Sequential binary partitions used for expression of independence tables in coordinates

SBP1	x_{11}	x_{12}	x_{13}	x_{14}	x_{21}	x_{22}	x_{23}	x_{24}	x_{31}	x_{32}	x_{33}	x_{34}	s	t
Step 1	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$+$	$+$	$+$	$+$	1	2
Step 2	$-$	$-$	$-$	$-$	$+$	$+$	$+$	$+$					1	1
SBP2	x_{11}	x_{12}	x_{13}	x_{14}	x_{21}	x_{22}	x_{23}	x_{24}	x_{31}	x_{32}	x_{33}	x_{34}	u	v
Step 1	$-$	$-$	$-$	$+$	$-$	$-$	$-$	$+$	$-$	$-$	$-$	$+$	1	3
Step 2	$-$	$-$	$+$		$-$	$-$	$+$		$-$	$-$	$+$		1	2
Step 3	$-$	$+$			$-$	$+$			$-$	$+$			1	1

The next three coordinates result from SBP2 and separate columns of the original compositional table. For example, the coordinates of the independence table in case of the Czech Republic equal to

$$\mathbf{z}_{ind} = (-1.4315, -0.6415, 0.8621, 2.7267, 4.0947, 0, 0, 0, 0, 0, 0) \quad . \tag{83}$$

Note that, when both SBPs from Table 7 are applied to $\mathbf{x}_{int}$, the resulting coordinates are equal to zero, as well as coordinates of SBP1 and SBP2 applied to $\text{col}^{\perp}(\mathbf{x})$ and $\text{row}^{\perp}(\mathbf{x})$, respectively. Thus, because of decomposition (31), the same coordinates would be obtained if SBPs from Table 7 were applied directly to the independence table $\mathbf{x}_{ind}$ from (30), or if SBP1 was applied to $\text{row}^{\perp}(\mathbf{x})$ and SBP2 to $\text{col}^{\perp}(\mathbf{x})$, respectively. As a consequence of coordinate isomorphism and (31), the coordinates of the independence table also form coordinates of the original table $\mathbf{x}$. The remaining coordinates of $\mathbf{x}$ equal to $(I-1)(J-1) = 6$ nonzero coordinates of the interaction table, and can be expressed using formula (79). In case of the Czech Republic, these coordinates are

$$\mathbf{z}_{int} = (0, 0, 0, 0, 0, 0.5439, 0.8988, 0.8428, 0.1354, 0.4648, 0.4441) \quad , \tag{84}$$

where the first five zero coordinates refer to SBP1 and SBP2 applied to $\mathbf{x}_{int}$. The relation of the coordinates of the interaction table to the partial tables and odds ratios within them is illustrated in Figure 10. The basic descriptive statistics of all coordinates for the given data set are summarized in Table 8.

The first coordinate z_1^r compares age category $65 - 84$ with the younger categories. Relatively high negative value of the mean of this coordinate, compared to its standard deviation, gives an evidence that the younger population categories dominate in average here. Similar statement holds also for the next coordinate z_2^r, which compares categories $25 - 44$ and $45 - 64$, and thus it can be concluded that the youngest generation ($25 - 44$ years) dominates.

Relation between the BMI index categories are described by coordinates z_1^c, z_2^c, z_3^c. High values of the mean for the second and third coordinates, compared to their standard deviations, indicate that there is a tendency to have

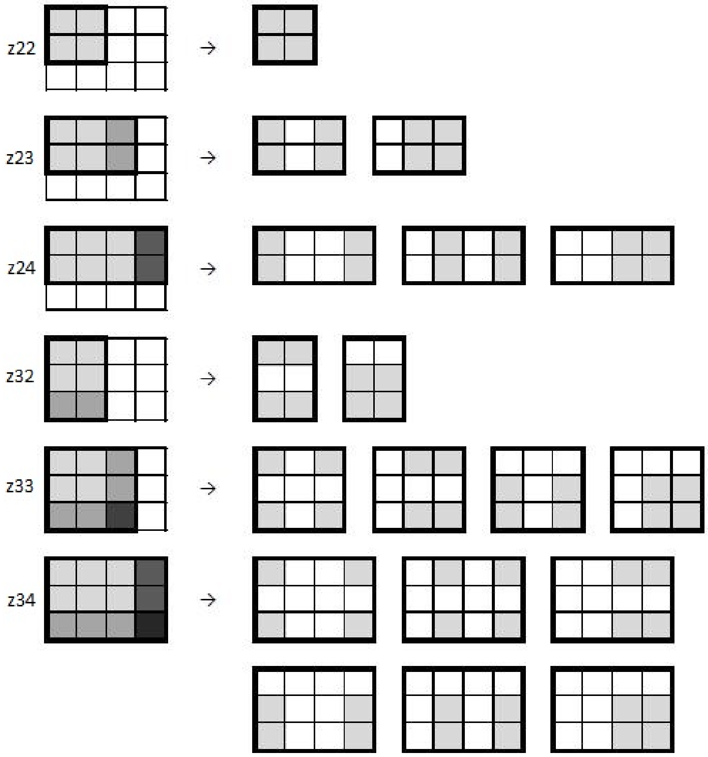


Figure 10: Relation of coordinates of the interaction table to the partial tables of **x** and odds ratios within them. In the first column the shades of grey denote the parts of the compositional table **x** used for computation of each coordinate; darker shade grades represent higher power of the corresponding parts in formula (80). The right part of the figure represents the odds ratios contained in each coordinate. This gives a visual interpretation of Equation (79) in case of 3×4 tables of age groups and BMI index in Example 3. Specifically, the second coordinate z_{23} is computed only from parts $x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{23}$ and this coordinate could be interpreted as $1/2\sqrt{3}$ of logarithm of two multiplied odds ratios, $(x_{11}x_{23})/(x_{13}x_{21})$ and $(x_{12}x_{23})/(x_{13}x_{22})$.

higher weight in our dataset (z_2^c compares overweight with underweight aggregated with normal weight and z_3^c normal weight with underweight). Coordinate z_1^c compares obesity with other weight categories. Even though its mean is positive, the standard deviation is relatively high, thus any conclusion about tendency to obesity compared to averaged other categories is not very relevant.

The first interaction coordinate z_{22} is computed for $r = s = 2$ for all 18 European countries. From Table 8 it could be seen that the sample mean equals to 0.3674, and the standard deviation is 0.1488. This coordinate could be interpreted as a ratio of the chance that people with age between 25 and 44 years will be

Table 8: Sample means and standard deviations (according to the analyzed 18 European countries) of coordinates.

Coordinate	z_1^r	z_2^r	z_1^c	z_2^c	z_3^c	
Sample mean	-1.2301	-0.4452	0.4604	2.4267	4.1192	
Sample st. dev.	0.3705	0.2413	0.4500	0.3477	0.3394	

Coordinate	z_{22}	z_{23}	z_{24}	z_{32}	z_{33}	z_{34}
Sample mean	0.3674	0.6096	0.6494	0.1057	0.3624	0.3783
Sample st. dev.	0.1488	0.1357	0.1426	0.2412	0.2175	0.1945

underweight rather than normal weight and the same chance for people between 45 and 64. From (79) the mean odds ratio $\mathrm{e}^{2\cdot 0.3674} \approx 2$ is obtained. Consequently, the chance that younger people are underweight is about twice as high as for people with age between 45 and 64.

The next coordinate z_{23} corresponds to a table for people aged between $25-44$ or $45-64$ being under-, normal, or overweight, respectively. This coordinate could be also interpreted as sum of logarithms of two odds ratios, divided by $2\sqrt{3}$. The first odds ratio compares chances that people underweight against overweight for age ranges $25-44$ and $45-64$ years, respectively. The second odds ratio compares almost the same with the only difference of taking normal weight instead of underweight. The sum of logarithms of these odds ratios is $2\sqrt{3}\cdot 0.6096 = 2.1117 > 0$ on average. Consequently, at least one of the chances that one is underweight against overweight, or the normal weight against overweight, respectively, is higher for people between 25 and 44 years. Coordinate z_{24}, which adds the column for obese people, has almost the same interpretation. The fourth coordinate z_{32} corresponds to a partial table with three age ranges ($25-44$, $45-64$ and $65-84$) and two weight possibilities (underweight and normal weight) and interpretation of this coordinate also analogous to the previous cases.

Since the remaining coordinates of the interaction table could be interpreted analogously as in the previous cases, they are only described using Figure 10 and Table 8. To sum it up, the first three nonzero coordinates of the interaction table carry information about odds ratios, which compare chances of lower weight ranges to a higher one for age group between 25 and 44 years and group between 45 and 64 years. The first coordinate compares underweight with normal weight. In the next coordinate, these two groups are both compared with overweight. Finally, the third coordinate compares groups with underweight, normal weight and overweight with the group of obese people. The last three coordinates compare the same chances, but now the first age group contains age ranges $25-44$ and $45-64$ together and the second group covers exclusively age range $65-84$ years. Quite interesting is the absence of negative values in the sample means of all coordinates, lower weight categories are thus typical for younger population.

The relationship between age and BMI index will be further analysed in Section 3.

2.3.3 Coordinate representation of 2×2 tables

Let us now consider 2×2 compositional table (33). Although it is possible to apply general coordinate representation of compositional tables, they equal to pivot coordinates in this particular case, and have form

$$z^r = \frac{1}{2}\ln\frac{x_{11}x_{12}}{x_{21}x_{22}}, \quad z^c = \frac{1}{2}\ln\frac{x_{11}x_{12}}{x_{21}x_{22}}, \quad z^{OR} = \frac{1}{2}\ln\frac{x_{11}x_{22}}{x_{12}x_{21}} \quad . \tag{85}$$

According to [11] there is also other coordinate system appropriate for this table that follows the decomposition $\mathbf{x} = \mathbf{x}_{ind} \oplus \mathbf{x}_{int}$. Moreover, this system can be reached by sequential binary partition presented in Table 9. This SBP was constructed by separating the parts on a diagonal of $\mathbf{x}$ from the remaining two parts in the first step, and dividing the remaining parts into separate groups in the following two steps.

SBP	x_{11}	x_{12}	x_{21}	x_{22}	u	v
Step 1	+	−	−	+	2	2
Step 2		+	−		1	1
Step 3	+			−	1	1

Table 9: Tabular representation of SBP for 2×2 compositional table.

This sequential binary partition also results in two (in general) nonzero coordinates of the independence table and one coordinate of the interaction table,

$$z^{int} = \frac{1}{2}\ln\frac{x_{11}x_{22}}{x_{12}x_{21}}, \quad z_1^{ind} = \frac{1}{2}\ln\frac{x_{12}}{x_{21}}, \quad z_2^{ind} = \frac{1}{2}\ln\frac{x_{11}}{x_{22}} \quad , \tag{86}$$

when all the remaining coordinates are always zero. And these three coordinates together form coordinate representation of a 2×2 compositional table $\mathbf{x}$. Besides the description of relations within the table, which will be discussed in Section 3, this coordinate system can be used as a compositional alternative to standard test of symmetry in contingency tables.

2.4 Covariance structure of coordinate representation

In the following covariance structure of the above mentioned coordinate representations will be expressed as linear combinations of variances of pairwise logratios. For this purpose, $I \times J$ compositional table is transformed in the vector form

$$\mathbf{x}_{vec} = \text{vec}(\mathbf{x}) = (x_{11}, \ldots, x_{1J}, x_{21}, \ldots, x_{IJ}) \quad . \tag{87}$$

Variances of pairwise logratios form the elemental information on variability in compositional tables and are summarized in $IJ \times IJ$ variation matrix

$$\mathbf{T} = \begin{pmatrix} \mathrm{var}\left(\ln \frac{x_{11}}{x_{11}}\right) & \mathrm{var}\left(\ln \frac{x_{11}}{x_{12}}\right) & \cdots & \mathrm{var}\left(\ln \frac{x_{11}}{x_{IJ}}\right) \\ \mathrm{var}\left(\ln \frac{x_{12}}{x_{11}}\right) & \mathrm{var}\left(\ln \frac{x_{12}}{x_{12}}\right) & \cdots & \mathrm{var}\left(\ln \frac{x_{12}}{x_{IJ}}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{var}\left(\ln \frac{x_{IJ}}{x_{11}}\right) & \mathrm{var}\left(\ln \frac{x_{IJ}}{x_{12}}\right) & \cdots & \mathrm{var}\left(\ln \frac{x_{IJ}}{x_{IJ}}\right) \end{pmatrix} . \tag{88}$$

As it is usual within the logratio methodology all coordinates are logcontrasts, i.e. they can be expressed in form

$$z = \sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij} \ln x_{ij} = \mathbf{a}' \ln \mathbf{x}_{vec}, \text{ where } \sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij} = 0 \tag{89}$$

and $\mathbf{a}$ is vector with elements $a_{11}, \ldots, a_{1J}, a_{21}, \ldots, a_{IJ}$. Also the covariance structure can be derived accordingly [2].

Proposition 2.1. *Variances and covariances for logcontrasts* $\mathbf{a}' \ln \mathbf{x}_{vec}$ *and* $\mathbf{b}' \ln \mathbf{x}_{vec}$ *of a* IJ*-part compositional table* $\mathbf{x}$ *are*

$$\mathrm{var}(\mathbf{a}' \ln \mathbf{x}_{vec}) = -\frac{1}{2}\mathbf{a}'\mathbf{T}\mathbf{a} \quad , \tag{90}$$

$$\mathrm{cov}(\mathbf{a}' \ln \mathbf{x}_{vec}, \mathbf{b}' \ln \mathbf{x}_{vec}) = -\frac{1}{2}\mathbf{a}'\mathbf{T}\mathbf{b} \quad . \tag{91}$$

Due to the possible logcontrast representation of coordinates from Sections 2.3.1, 2.3.2 and 2.3.3, Equations (90) and (91) are crucial for derivation of their covariance structure. At first, covariance structure of the general coordinates is considered. Consequently, results are adapted for the case of pivot coordinates together with a closer look at simplifications for interpretation of coordinates. Finally, the simplest case of 2 × 2 compositional tables follows.

2.4.1 General coordinates

Consider first the general coordinate system of $I \times J$ compositional table, whose construction was described in Section 2.3.1. If not otherwise stated, the following theorems were proved by applying directly Proposition 2.1.

Theorem 2.2. *Consider an arbitrary coordinate* z^{OR} *constructed with respect to equation (52). Its variance is formed by three parts,*

$$\mathrm{var}(z^{OR}) = A_1 - B_1 - C_1. \tag{92}$$

The first part, increasing the variance, is

$$\begin{aligned}A_1 &= \frac{d}{b(a+b+c+d)}\sum_{(i,j)\in I_A}\sum_{(i',j')\in I_B}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+\\ &+\frac{d}{c(a+b+c+d)}\sum_{(i,j)\in I_A}\sum_{(i',j')\in I_C}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+\\ &+\frac{a}{b(a+b+c+d)}\sum_{(i,j)\in I_B}\sum_{(i',j')\in I_D}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+\\ &+\frac{a}{c(a+b+c+d)}\sum_{(i,j)\in I_C}\sum_{(i',j')\in I_D}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right) \quad . \end{aligned} \tag{93}$$

The variance of the coordinate is reduced by parts

$$\begin{aligned}B_1 &= \frac{1}{2}\frac{d}{a(a+b+c+d)}\sum_{(i,j)\in I_A}\sum_{(i',j')\in I_A}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+\\ &+\frac{1}{2}\frac{a}{d(a+b+c+d)}\sum_{(i,j)\in I_D}\sum_{(i',j')\in I_D}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+\\ &+\frac{1}{(a+b+c+d)}\sum_{(i,j)\in I_A}\sum_{(i',j')\in I_D}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)\end{aligned} \tag{94}$$

and

$$\begin{aligned}C_1 &= \frac{1}{2}\frac{ad}{b^2(a+b+c+d)}\sum_{(i,j)\in I_B}\sum_{(i',j')\in I_B}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+\\ &+\frac{1}{2}\frac{ad}{c^2(a+b+c+d)}\sum_{(i,j)\in I_C}\sum_{(i',j')\in I_C}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+\\ &+\frac{ad}{bc(a+b+c+d)}\sum_{(i,j)\in I_B}\sum_{(i',j')\in I_C}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right) \quad . \end{aligned} \tag{95}$$

Here I_A, I_B, I_C and I_D are sets of indices of parts from groups A, B, C and D as defined in Sections 2.3.1 and a, b, c and d are numbers of parts in these groups.

From Equation (92) results that the variance of coordinate z^{OR} is increased by variances of logratios between parts from groups A and B, A and C, B and D and C and D. On the other hand, the overall variance of the coordinate is reduced by variances of logratios between parts from the same group or between parts from groups A and D or B and C. Graphical representation of this feature is provided by Figure 11.

The next theorem defines covariance between two odds ratio coordinates.

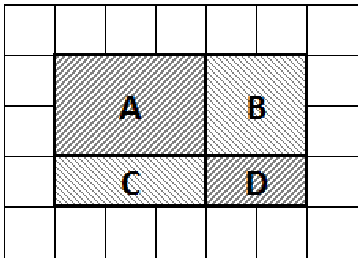


Figure 11: Graphical representation of variance of coordinate z^{OR}, which is increased by variances of logratios between parts from blocks highlighted by (/) and (\) - A_1 and reduced by variances of logratios, where both parts are from blocks highlighted by (/) - B_1 or by (\) - C_1.

Theorem 2.3. *Consider two coordinates z_k^{OR}, z_l^{OR}, for $k, l = 1, \ldots, (I-1)(J-1), k \neq l$, constructed according to (52). Then for their covariance the following holds,*

$$\operatorname{cov}(z_k^{OR}, z_l^{OR}) = A_2 + B_2 - C_2 - D_2, \tag{96}$$

where

$$\begin{aligned} A_2 \quad = \quad & \frac{1}{2}\sqrt{\frac{d_k}{a_k(a_k+b_k+c_k+d_k)}}\sqrt{\frac{a_l d_l}{b_l^2(a_l+b_l+c_l+d_l)}}\sum_{(i,j)\in I_{A_k}}\sum_{(i',j')\in I_{B_l}}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+ \\ & +\frac{1}{2}\sqrt{\frac{d_k}{a_k(a_k+b_k+c_k+d_k)}}\sqrt{\frac{a_l d_l}{c_l^2(a_l+b_l+c_l+d_l)}}\sum_{(i,j)\in I_{A_k}}\sum_{(i',j')\in I_{C_l}}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+ \\ & +\frac{1}{2}\sqrt{\frac{a_k}{d_k(a_k+b_k+c_k+d_k)}}\sqrt{\frac{a_l d_l}{b_l^2(a_l+b_l+c_l+d_l)}}\sum_{(i,j)\in I_{D_k}}\sum_{(i',j')\in I_{B_l}}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+ \\ & +\frac{1}{2}\sqrt{\frac{a_k}{d_k(a_k+b_k+c_k+d_k)}}\sqrt{\frac{a_l d_l}{c_l^2(a_l+b_l+c_l+d_l)}}\sum_{(i,j)\in I_{D_k}}\sum_{(i',j')\in I_{C_l}}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)\quad , \end{aligned} \tag{97}$$

$$\begin{aligned} B_2 \quad = \quad & \frac{1}{2}\sqrt{\frac{a_k d_k}{b_k^2(a_k+b_k+c_k+d_k)}}\sqrt{\frac{d_l}{a_l(a_l+b_l+c_l+d_l)}}\sum_{(i,j)\in I_{B_k}}\sum_{(i',j')\in I_{A_l}}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+ \\ & +\frac{1}{2}\sqrt{\frac{a_k d_k}{b_k^2(a_k+b_k+c_k+d_k)}}\sqrt{\frac{a_l}{d_l(a_l+b_l+c_l+d_l)}}\sum_{(i,j)\in I_{B_k}}\sum_{(i',j')\in I_{D_l}}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+ \\ & +\frac{1}{2}\sqrt{\frac{a_k d_k}{c_k^2(a_k+b_k+c_k+d_k)}}\sqrt{\frac{d_l}{a_l(a_l+b_l+c_l+d_l)}}\sum_{(i,j)\in I_{C_k}}\sum_{(i',j')\in I_{A_l}}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+ \\ & +\frac{1}{2}\sqrt{\frac{a_k d_k}{c_k^2(a_k+b_k+c_k+d_k)}}\sqrt{\frac{a_l}{d_l(a_l+b_l+c_l+d_l)}}\sum_{(i,j)\in I_{C_k}}\sum_{(i',j')\in I_{D_l}}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)\quad , \end{aligned} \tag{98}$$

$$\begin{aligned} C_2 \;=\;& \frac{1}{2}\sqrt{\frac{d_k}{a_k(a_k+b_k+c_k+d_k)}}\sqrt{\frac{d_l}{a_l(a_l+b_l+c_l+d_l)}}\sum_{(i,j)\in I_{A_k}}\sum_{(i',j')\in I_{A_l}}\mathrm{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+\\ &+\frac{1}{2}\sqrt{\frac{d_k}{a_k(a_k+b_k+c_k+d_k)}}\sqrt{\frac{a_l}{d_l(a_l+b_l+c_l+d_l)}}\sum_{(i,j)\in I_{A_k}}\sum_{(i',j')\in I_{D_l}}\mathrm{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+\\ &+\frac{1}{2}\sqrt{\frac{a_k}{d_k(a_k+b_k+c_k+d_k)}}\sqrt{\frac{d_l}{a_l(a_l+b_l+c_l+d_l)}}\sum_{(i,j)\in I_{D_k}}\sum_{(i',j')\in I_{A_l}}\mathrm{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+\\ &+\frac{1}{2}\sqrt{\frac{a_k}{d_k(a_k+b_k+c_k+d_k)}}\sqrt{\frac{a_l}{d_l(a_l+b_l+c_l+d_l)}}\sum_{(i,j)\in I_{D_k}}\sum_{(i',j')\in I_{D_l}}\mathrm{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)\quad, \end{aligned}\tag{99}$$

$$\begin{aligned} D_2 \;=\;& \frac{1}{2}\sqrt{\frac{a_kd_k}{b_k^2(a_k+b_k+c_k+d_k)}}\sqrt{\frac{a_ld_l}{b_l^2(a_l+b_l+c_l+d_l)}}\sum_{(i,j)\in I_{B_k}}\sum_{(i',j')\in I_{B_l}}\mathrm{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+\\ &+\frac{1}{2}\sqrt{\frac{a_kd_k}{b_k^2(a_k+b_k+c_k+d_k)}}\sqrt{\frac{a_ld_l}{c_l^2(a_l+b_l+c_l+d_l)}}\sum_{(i,j)\in I_{B_k}}\sum_{(i',j')\in I_{C_l}}\mathrm{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+\\ &+\frac{1}{2}\sqrt{\frac{a_kd_k}{c_k^2(a_k+b_k+c_k+d_k)}}\sqrt{\frac{a_ld_l}{b_l^2(a_l+b_l+c_l+d_l)}}\sum_{(i,j)\in I_{C_k}}\sum_{(i',j')\in I_{B_l}}\mathrm{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+\\ &+\frac{1}{2}\sqrt{\frac{a_kd_k}{c_k^2(a_k+b_k+c_k+d_k)}}\sqrt{\frac{a_ld_l}{c_l^2(a_l+b_l+c_l+d_l)}}\sum_{(i,j)\in I_{C_k}}\sum_{(i',j')\in I_{C_l}}\mathrm{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)\quad. \end{aligned}\tag{100}$$

Here $I_{A_k}, I_{B_k}, I_{C_k}$ and $I_{D_k}/I_{A_l}, I_{B_l}, I_{C_l}$ and I_{D_l} are sets of indices of parts from groups A_k, B_k, C_k and $D_k/A_l, B_l, C_l$ and D_l and a_k, b_k, c_k and $d_k/a_l, b_l, c_l$ and d_l are numbers of parts in these groups.

Similarly as in the case of variance, the covariance between two odds ratio coordinates is increased by variances of logratios between parts from groups A_k and B_l, A_k and C_l, D_k and B_l and D_k and C_l, and conversely (see (97) and (98)). The covariance is reduced by variances of logratios between parts from blocks (A_k, A_l), (B_k, B_l), (C_k, C_l) and (D_k, D_l) or (A_k, D_l), (B_k, C_l), and conversely (see (99) and (100)). These groups are displayed in Figure 12.

The second group of coordinates of a compositional table is formed by balances between whole rows or columns, each represented by the respective geometrical means. The following Theorems 2.4 and 2.5 are direct consequences of results from paper [17]:

Theorem 2.4. *Consider row balance z_i^r for $i = 1, \dots, I-1$ and column balance z_j^c for $j = 1, \dots, J-1$, computed with respect to (44) and (45). Their variances*

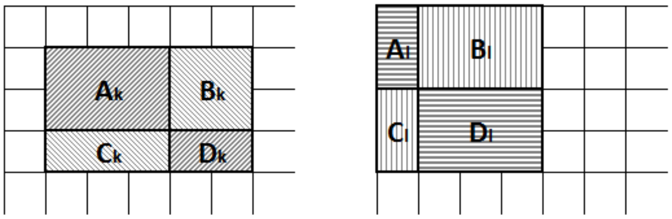


Figure 12: Graphical representation of covariance between two odds ratio coordinates z_k^{OR} and z_l^{OR}, which is increased by variances of logratios between parts from blocks (/) and (|) - A_2 and variances of logratios between parts from (\) and (−) - B_2. The covariance is reduced by variances of logratios between parts from blocks (/) and (−) - C_2 or (\) and (|) - D_2.

are

$$\begin{aligned}\operatorname{var}(z_i^r) &= \tfrac{1}{2J(s+t)} \quad \Bigg[-\frac{t}{s}\sum_{(i,j)\in I_s}\sum_{(i',j')\in I_s}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+ \\ &\qquad +2\sum_{(i,j)\in I_s}\sum_{(i',j')\in I_t}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)- \\ &\qquad -\frac{s}{t}\sum_{(i,j)\in I_t}\sum_{(i',j')\in I_t}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)\Bigg] \quad ; \qquad (101)\end{aligned}$$

where I_s, I_t are sets of indices of parts from the rows from the first and second group of the i-th step of row SBP, respectively. And

$$\begin{aligned}\operatorname{var}(z_j^c) &= \tfrac{1}{2I(u+v)} \quad \Bigg[-\frac{v}{u}\sum_{(i,j)\in I_u}\sum_{(i',j')\in I_u}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+ \\ &\qquad +2\sum_{(i,j)\in I_u}\sum_{(i',j')\in I_v}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)- \\ &\qquad -\frac{u}{v}\sum_{(i,j)\in I_v}\sum_{(i',j')\in I_v}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)\Bigg] \quad , \qquad (102)\end{aligned}$$

here I_u, I_v are positions of parts from the columns from the first and second group of the j-th step of column SBP.

According to this theorem the variance of row coordinates is increased by variances of logratios between parts from different groups of rows, which were defined in the respective step of row SBP. On the other hand, the variance is reduced by variances of logratios between parts of rows, which belong to the same group. Analogous interpretation holds also for variance of the column coordinate z_j^c. Graphical representation of these features is provided by Figure 13.

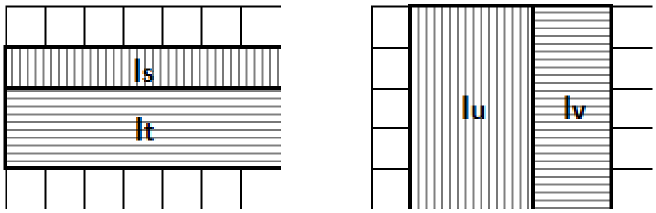


Figure 13: Graphical representation of variance of row balance z_i^r (left) and column balance z_j^c (right). Variances of both coordinates are increased by variances of logratios between a part from block highlighted by | and a part from block −. Variances of logratios between parts which are both from the same block decrease the resulting variance. Covariance between both balances is increased by variances of logratios between a part of the left table highlighted by | and part of the right table highlighted by −, or conversely. Variances of logratios between parts from blocks with the same marking decrease the resultant covariance.

Theorem 2.5. *Consider three row balances $z_{k_1}^r, z_{k_2}^r$ and z_k^r for $k_1, k_2, k = 1, \dots, I-1$, $k_1 \neq k_2$ and three column balances $z_{l_1}^c, z_{l_2}^c$ and z_l^c for $l_1, l_2, l = 1, \dots, J-1$, $l_1 \neq l_2$ from (44) and (45), respectively. Their covariances can be expressed as*

$$\begin{aligned}\operatorname{cov}(z_{k_1}^r, z_{k_2}^r) \quad = K \quad & \left[-\sqrt{\frac{t_1 t_2}{s_1 s_2}} \sum_{(i,j)\in I_{s_1}} \sum_{(i',j')\in I_{s_2}} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{i'j'}}\right) + \right. \\ & + \sqrt{\frac{t_1 s_2}{s_1 t_2}} \sum_{(i,j)\in I_{s_1}} \sum_{(i',j')\in I_{t_2}} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{i'j'}}\right) + \\ & + \sqrt{\frac{s_1 t_2}{t_1 s_2}} \sum_{(i,j)\in I_{t_1}} \sum_{(i',j')\in I_{s_2}} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{i'j'}}\right) - \\ & \left. - \sqrt{\frac{s_1 s_2}{t_1 t_2}} \sum_{(i,j)\in I_{t_1}} \sum_{(i',j')\in I_{t_2}} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{i'j'}}\right)\right] \quad , \qquad (103)\end{aligned}$$

with $K = \sqrt{\frac{1}{4J^2(s_1+t_1)(s_2+t_2)}}$, where I_{s_1}, I_{t_1} and I_{s_2}, I_{t_2} are sets of indices of parts from the rows from the first and second group of the k_1-th and k_2-th step of row

SBP, respectively. Covariance between column balances is obtained as

$$\begin{aligned}\operatorname{cov}(z^c_{l_1}, z^c_{l_2}) \;&= K \;\;\Bigg[-\sqrt{\frac{v_1v_2}{u_1u_2}}\sum_{(i,j)\in I_{u_1}}\sum_{(i',j')\in I_{u_2}}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+\\ &+\sqrt{\frac{v_1u_2}{u_1v_2}}\sum_{(i,j)\in I_{u_1}}\sum_{(i',j')\in I_{v_2}}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+\\ &+\sqrt{\frac{u_1v_2}{v_1u_2}}\sum_{(i,j)\in I_{v_1}}\sum_{(i',j')\in I_{u_2}}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)-\\ &-\sqrt{\frac{u_1u_2}{v_1v_2}}\sum_{(i,j)\in I_{v_1}}\sum_{(i',j')\in I_{v_2}}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)\Bigg] \quad, \qquad (104)\end{aligned}$$

with $K=\sqrt{\frac{1}{4I^2(u_1+v_1)(u_2+v_2)}}$, I_{u_1}, I_{v_1} and I_{u_2}, I_{v_2} are sets of indices of parts from the columns from the first and second group of the j_1-th and j_2-th step of column SBP, respectively. Finally, covariance between row and column balances is obtained as

$$\begin{aligned}\operatorname{cov}(z^r_k, z^c_l) \;&= K \;\;\Bigg[-\sqrt{\frac{tv}{su}}\sum_{(i,j)\in I_s}\sum_{(i',j')\in I_u}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+\\ &+\sqrt{\frac{tu}{sv}}\sum_{(i,j)\in I_s}\sum_{(i',j')\in I_v}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+\\ &+\sqrt{\frac{sv}{tu}}\sum_{(i,j)\in I_t}\sum_{(i',j')\in I_u}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)-\\ &-\sqrt{\frac{su}{tv}}\sum_{(i,j)\in I_t}\sum_{(i',j')\in I_v}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)\Bigg] \quad, \qquad (105)\end{aligned}$$

with $K=\sqrt{\frac{1}{4IJ(s+t)(u+v)}}$, I_s, I_t and I_u, I_v are sets of indices of parts from the rows and columns from the first and second group of the k-th step of row SBP and l-th step of column SBP, respectively.

Covariance between different row/column balances is increased by variances of logratios between a part from row/column, which was in the k_1-th/l_1-th step of row/column SBP included in the first group and a part of row/column from the second group, according to k_2-th/l_2-th step of row/column SBP or conversely. The covariance is reduced by variances of logratios between parts, which are both from the first groups of the respective steps of row/column SBP, or both from the second groups. Also this interpretation can be supported graphically (Figure 14, 15). Covariance between row and column balances can be interpreted analogously with groups defined according to k-th step of row SBP and l-th step of column SBP. Graphical representation is also provided by Figure 13.

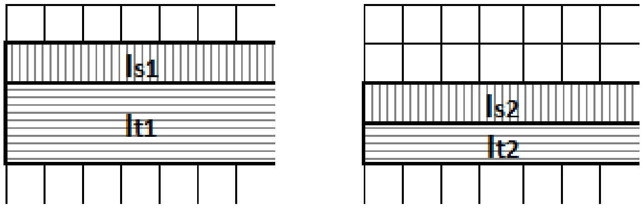


Figure 14: Graphical representation of covariance between row balances $z^r_{k_1}$ (left) and $z^r_{k_2}$ (right), which is increased by variances of logratios between a part of the left table highlighted by $|$ and a part of the right table highlighted by $-$, or conversely. Variances of logratios between parts from blocks with the same marking reduce the resulting covariance.

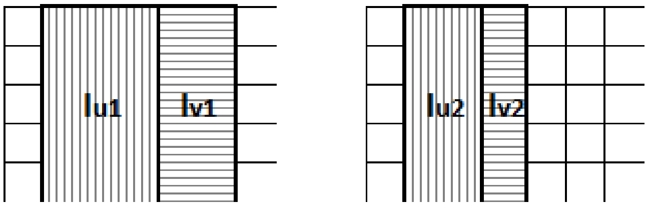


Figure 15: Graphical representation of covariance between column balances $z^c_{l_1}$ (left) and $z^c_{l_2}$ (right), which is increased by variances of logratios between a part of the left table highlighted by $|$ and a part of the right table highlighted by $-$, or conversely. Variances of logratios between parts from blocks with the same marking reduce the resulting covariance.

Finally, the last theorem of this section derives covariances between odds ratio coordinates and row or column balances, respectively.

Theorem 2.6. *Consider coordinates z^{OR}_i, z^r_k and z^c_l for $i = 1, \ldots, (I-1)(J-1)$, $k = 1, \ldots, I-1$ and $l = 1, \ldots, J-1$, then for covariances between odds ratio coordinate and row or column balance, respectively, the following holds. For* *the first case,*

$$\begin{aligned}\mathrm{cov}(z_i^{OR}, z_k^r) = \quad & \tfrac{1}{2}\sqrt{\tfrac{ad}{a+b+c+d}}\sqrt{\tfrac{t}{sJ(s+t)}} \quad \left[-\frac{1}{a}\sum_{(i,j)\in I_A}\sum_{(i',j')\in I_s}\mathrm{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+\right.\\ & \qquad +\frac{1}{b}\sum_{(i,j)\in I_B}\sum_{(i',j')\in I_s}\mathrm{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+\\ & \qquad +\frac{1}{c}\sum_{(i,j)\in I_C}\sum_{(i',j')\in I_s}\mathrm{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)-\\ & \qquad \left.-\frac{1}{d}\sum_{(i,j)\in I_D}\sum_{(i',j')\in I_s}\mathrm{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)\right] -\\ & -\tfrac{1}{2}\sqrt{\tfrac{ad}{a+b+c+d}}\sqrt{\tfrac{s}{tJ(s+t)}} \quad \left[-\frac{1}{a}\sum_{(i,j)\in I_A}\sum_{(i',j')\in I_t}\mathrm{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+\right.\\ & \qquad +\frac{1}{b}\sum_{(i,j)\in I_B}\sum_{(i',j')\in I_t}\mathrm{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+\\ & \qquad +\frac{1}{c}\sum_{(i,j)\in I_C}\sum_{(i',j')\in I_t}\mathrm{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)-\\ & \qquad \left.-\frac{1}{d}\sum_{(i,j)\in I_D}\sum_{(i',j')\in I_t}\mathrm{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)\right] \quad ,\end{aligned} \tag{106}$$

where I_A, I_B, I_C, I_D are sets of indices of parts from groups A, B, C, D and I_s, I_t are sets of indices of parts from the first and second group from k-th step of row SBP. Similarly, with I_u, I_v defining sets of indices of parts from the first and second group of column SBP, covariance between odds ratio coordinate and *column balance is given as*

$$\begin{aligned}
\operatorname{cov}(z_i^{OR}, z_l^c) = \quad & \tfrac{1}{2}\sqrt{\tfrac{ad}{a+b+c+d}}\sqrt{\tfrac{v}{uI(u+v)}} \quad \left[-\frac{1}{a}\sum_{(i,j)\in I_A}\sum_{(i',j')\in I_u} \operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right) + \right. \\
& +\frac{1}{b}\sum_{(i,j)\in I_B}\sum_{(i',j')\in I_u} \operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right) + \\
& +\frac{1}{c}\sum_{(i,j)\in I_C}\sum_{(i',j')\in I_u} \operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right) - \\
& \left. -\frac{1}{d}\sum_{(i,j)\in I_D}\sum_{(i',j')\in I_u} \operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right) \right] - \\
& -\tfrac{1}{2}\sqrt{\tfrac{ad}{a+b+c+d}}\sqrt{\tfrac{u}{vI(u+v)}} \quad \left[-\frac{1}{a}\sum_{(i,j)\in I_A}\sum_{(i',j')\in I_v} \operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right) + \right. \\
& +\frac{1}{b}\sum_{(i,j)\in I_B}\sum_{(i',j')\in I_v} \operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right) + \\
& +\frac{1}{c}\sum_{(i,j)\in I_C}\sum_{(i',j')\in I_v} \operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right) - \\
& \left. -\frac{1}{d}\sum_{(i,j)\in I_D}\sum_{(i',j')\in I_v} \operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right) \right] \quad .
\end{aligned} \tag{107}$$

Also covariance between coordinate z_i^{OR} and balance z_k^r or z_l^c is increased by variances of logratios between parts from the opposite side of fractions, defining both respective coordinates. Concretely, between a part from group A or D and a part from the second group of the respective step of row or column SBP, or a part from area B or C and first group of the respective step of the SBPs. The remaining variances reduce the resulting covariance. Figure 16 provides graphical representation of this feature.

All the introduced theorems are more specified in the next section, concerning the pivot coordinates.

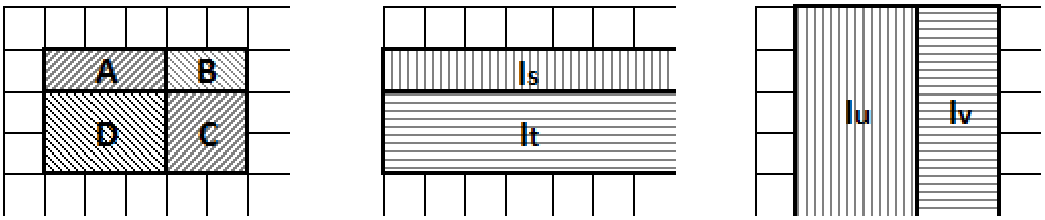


Figure 16: Graphical representation of covariance between coordinate z_i^{OR} (left) and row balance z_k^r (center) or column balance z_l^c (right). The covariance is increased by variances of logratios between a part from block highlighted by / and part of the middle or right table highlighted by $-$, or parts from areas \ and |. Variances of logratios between parts from areas / and |, or \ and $-$ reduce the resulting covariance.

2.4.2 Pivot coordinates

Results from the previous section represent a direct generalization of those from paper [12], which focuses on covariance structure of pivot coordinates of compositional tables. The corresponding theorems together with a detailed interpretation are provided in this section.

Theorem 2.7. *Consider an arbitrary coordinate z_{rs} constructed with respect to (79), for $r = 2, \dots, I$ and $s = 2, \dots, J$. Its variance is formed by three parts,*

$$\operatorname{var}(z_{rs}) = A_3 - B_3 - C_3 \quad . \tag{108}$$

The first part, increasing the variance, is

$$\begin{aligned} A_3 \quad = \quad & \frac{1}{rs(s-1)} \sum_{i=1}^{r-1} \sum_{j,j'=1}^{s-1} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{rj'}}\right) + \\ & + \frac{1}{rs(r-1)} \sum_{i,i'=1}^{r-1} \sum_{j=1}^{s-1} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{i's}}\right) + \\ & + \frac{r-1}{rs} \sum_{j=1}^{s-1} \operatorname{var}\left(\ln \frac{x_{rj}}{x_{rs}}\right) + \\ & + \frac{s-1}{rs} \sum_{i=1}^{r-1} \operatorname{var}\left(\ln \frac{x_{is}}{x_{rs}}\right) \quad . \end{aligned} \tag{109}$$

The variance of the coordinate is reduced by parts

$$\begin{aligned} B_3 &= \frac{1}{2}\frac{1}{rs(r-1)(s-1)}\sum_{i,i'=1}^{r-1}\sum_{j,j'=1}^{s-1}\mathrm{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+ \\ &\quad +\frac{1}{rs}\sum_{i=1}^{r-1}\sum_{j=1}^{s-1}\mathrm{var}\left(\ln\frac{x_{ij}}{x_{rs}}\right) \end{aligned} \tag{110}$$

and

$$\begin{aligned} C_3 &= \frac{1}{2}\frac{r-1}{rs(s-1)}\sum_{j,j'=1}^{s-1}\mathrm{var}\left(\ln\frac{x_{rj}}{x_{rj'}}\right)+ \\ &\quad +\frac{1}{2}\frac{s-1}{rs(r-1)}\sum_{i,i'=1}^{r-1}\mathrm{var}\left(\ln\frac{x_{is}}{x_{i's}}\right)+ \\ &\quad +\frac{1}{rs}\sum_{i=1}^{r-1}\sum_{j'=1}^{s-1}\mathrm{var}\left(\ln\frac{x_{is}}{x_{rj'}}\right) \quad . \end{aligned} \tag{111}$$

Proof: *When parts of the compositional table* $\mathbf{x}$ *are rearranged in form of composition* $\mathbf{x}_{vec} = (x_{11}, x_{12}, \ldots, x_{1J}, x_{21}, \ldots, x_{IJ})$*, coordinate* z_{rs} *of the interaction table can be expressed as* $z_{rs} = \mathbf{a}' \ln \mathbf{x}_{vec}$*, where for elements of the coefficient vector* $\mathbf{a} = (a_{11}, a_{12}, \ldots, a_{1J}, a_{21}, \ldots, a_{IJ})$ *the following relations hold,*

$$\begin{array}{llll} a_{ij} = 1/\sqrt{rs(r-1)(s-1)} & \text{for} & i = 1,\ldots,r-1 & j = 1,\ldots,s-1 \\ a_{ij} = -(r-1)/\sqrt{rs(r-1)(s-1)} & \text{for} & i = r & j = 1,\ldots,s-1 \\ a_{ij} = -(s-1)/\sqrt{rs(r-1)(s-1)} & \text{for} & i = 1,\ldots,r-1 & j = s \\ a_{ij} = (r-1)(s-1)/\sqrt{rs(r-1)(s-1)} & \text{for} & i = r & j = s \\ a_{ij} = 0 & & \textit{otherwise.} & \end{array}$$

Equation (108) is then consequence of Proposition 2.1.

◇

From Theorem 2.7 it is clear that variance of the coordinate z_{rs} is formed by nine groups of logratio variances. Four of them increase the overall variability and the other five reduce it. The first four groups are represented by A_3, which is formed by logratios of “inner” parts of the partial table or part x_{rs}, with its last row and column (i.e. r-th row and s-th column of the original table $\mathbf{x}$) except of the part x_{rs} itself:

- variances of logratios between an inner part of the partial table and a part from its last row (except of x_{rs}),

- variances of logratios between an inner part of the partial table and a part from its last column (except of x_{rs}),
- variances of logratios between a part from the last row (except of x_{rs}) and x_{rs} itself,
- variances of logratios between a part from the last column (except of x_{rs}) and x_{rs} itself.

The variance of z_{rs} is reduced by B_3 and C_3, formed by variances of logratios corresponding to remaining possible relations between parts of the above defined groups (inner parts, last row/column without x_{rs}, part x_{rs} itself). Concretely, B_3 consists of

- variances of logratios between inner parts of the partial table,
- variances of logratios between an inner part and x_{rs}.

Similarly, C_3 is formed by

- variances of logratios between parts from the last row (except of x_{rs}),
- variances of logratios between parts from the last column (except of x_{rs}),
- variances of logratios between parts from the last row and the last column (except of x_{rs}).

The above relations can be expressed also graphically, as shown in Figure 17.

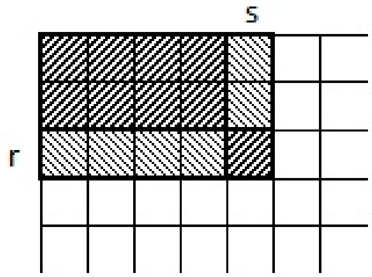


Figure 17: Variance of coordinate z_{rs} is increased by variances of logratios between a part from area highlighted by (/) and a part from the second area highlighted by (\) – A_3. The variance of z_{rs} is reduced by variances of logratios between two parts from area (/) – B_3 or two parts from (\) – C_3.

Covariances between coordinates of the interaction table are derived in the next theorem.

Theorem 2.8. *Consider two coordinates $z_{r_1s_1}, z_{r_2s_2}$, for $r_1, r_2 = 2, \dots, I$ and $s_1, s_2 = 2, \dots, J$, constructed according to (79). Then for their covariance the following holds,*

$$\operatorname{cov}(z_{r_1s_1}, z_{r_2s_2}) = K(A_4 + B_4 - C_4 - D_4) \quad , \tag{112}$$

where

$$\begin{aligned}
A_4 &= (s_2-1)\sum_{i_1=1}^{r_1-1}\sum_{i_2=1}^{r_2-1}\sum_{j_1=1}^{s_1-1}\operatorname{var}\left(\ln\frac{x_{i_1j_1}}{x_{i_2s_2}}\right) + \\
&+(r_2-1)\sum_{i_1=1}^{r_1-1}\sum_{j_1=1}^{s_1-1}\sum_{j_2=1}^{s_2-1}\operatorname{var}\left(\ln\frac{x_{i_1j_1}}{x_{r_2j_2}}\right) + \\
&+(r_1-1)(s_1-1)(s_2-1)\sum_{i_2=1}^{r_2-1}\operatorname{var}\left(\ln\frac{x_{r_1s_1}}{x_{i_2s_2}}\right) + \\
&+(r_1-1)(r_2-1)(s_1-1)\sum_{j_2=1}^{s_2-1}\operatorname{var}\left(\ln\frac{x_{r_1s_1}}{x_{r_2j_2}}\right) \quad ,
\end{aligned} \tag{113}$$

$$\begin{aligned}
B_4 &= (s_1-1)\sum_{i_1=1}^{r_1-1}\sum_{i_2=1}^{r_2-1}\sum_{j_2=1}^{s_2-1}\operatorname{var}\left(\ln\frac{x_{i_1s_1}}{x_{i_2j_2}}\right) + \\
&+(s_1-1)(s_2-1)(r_2-1)\sum_{i_1=1}^{r_1-1}\operatorname{var}\left(\ln\frac{x_{i_1s_1}}{x_{r_2s_2}}\right) + \\
&+(r_1-1)\sum_{j_1=1}^{s_1-1}\sum_{i_2=1}^{r_2-1}\sum_{j_2=1}^{s_2-1}\operatorname{var}\left(\ln\frac{x_{r_1j_1}}{x_{i_2j_2}}\right) + \\
&+(r_1-1)(r_2-1)(s_2-1)\sum_{j_1=1}^{s_1-1}\operatorname{var}\left(\ln\frac{x_{r_1j_1}}{x_{r_2s_2}}\right) \quad ,
\end{aligned} \tag{114}$$

$$\begin{aligned}
C_4 &= \sum_{i_1=1}^{r_1-1}\sum_{i_2=1}^{r_2-1}\sum_{j_1=1}^{s_1-1}\sum_{j_2=1}^{s_2-1}\operatorname{var}\left(\ln\frac{x_{i_1j_1}}{x_{i_2j_2}}\right) + \\
&+(r_2-1)(s_2-1)\sum_{i_1=1}^{r_1-1}\sum_{j_1=1}^{s_1-1}\operatorname{var}\left(\ln\frac{x_{i_1j_1}}{x_{r_2s_2}}\right) + \\
&+(r_1-1)(s_1-1)\sum_{i_2=1}^{r_2-1}\sum_{j_2=1}^{s_2-1}\operatorname{var}\left(\ln\frac{x_{r_1s_1}}{x_{i_2j_2}}\right) + \\
&+(r_1-1)(r_2-1)(s_1-1)(s_2-1)\operatorname{var}\left(\ln\frac{x_{r_1s_1}}{x_{r_2s_2}}\right) \quad ,
\end{aligned} \tag{115}$$

$$
\begin{aligned}
D_4 &= (s_1-1)(s_2-1)\sum_{i_1=1}^{r_1-1}\sum_{i_2=1}^{r_2-1}\mathrm{var}\left(\ln\frac{x_{i_1s_1}}{x_{i_2s_2}}\right)+\\
&\quad+(s_1-1)(r_2-1)\sum_{i_1=1}^{r_1-1}\sum_{j_2=1}^{s_2-1}\mathrm{var}\left(\ln\frac{x_{i_1s_1}}{x_{r_2j_2}}\right)+\\
&\quad+(r_1-1)(s_2-1)\sum_{j_1=1}^{s_1-1}\sum_{i_2=1}^{r_2-1}\mathrm{var}\left(\ln\frac{x_{r_1j_1}}{x_{i_2s_2}}\right)+\\
&\quad+(r_1-1)(r_2-1)\sum_{j_1=1}^{s_1-1}\sum_{j_2=1}^{s_2-1}\mathrm{var}\left(\ln\frac{x_{r_1j_1}}{x_{r_2j_2}}\right) \qquad (116)
\end{aligned}
$$

and $K=\frac{1}{2}\frac{1}{\sqrt{r_1r_2s_1s_2(r_1-1)(r_2-1)(s_1-1)(s_2-1)}}$.

Proof: *The covariances are obtained using the general formula (91), where the corresponding coefficient vectors* $\mathbf{a}^1$ *and* $\mathbf{a}^2$ *have elements*

$$
a_{ij}^k=\begin{cases}
1/\sqrt{r_ks_k(r_k-1)(s_k-1)} & \text{for} \quad i=1,\dots,r_k-1 \quad j=1,\dots,s_k-1\\
-(r_k-1)/\sqrt{r_ks_k(r_k-1)(s_k-1)} & \text{for} \quad i=r_k \quad j=1,\dots,s_k-1\\
-(s_k-1)/\sqrt{r_ks_k(r_k-1)(s_k-1)} & \text{for} \quad i=1,\dots,r_k-1 \quad j=s_k\\
(r_k-1)(s_k-1)/\sqrt{r_ks_k(r_k-1)(s_k-1)} & \text{for} \quad i=r_k \quad j=s_k\\
0 & \quad \text{otherwise,}
\end{cases} \qquad (117)
$$

and $k=1,2$.

$\diamond$

Similarly as for the case of variances, there is a group of logratio variances that increases the overall covariance between coordinates (A_4 and B_4) and the remaining variances reduce it (C_4 and D_4). Specifically, for construction of logratios in A_4 the following parts are employed,

- an inner part of the first partial table and a part from the last column of the second partial table (except of $x_{r_2s_2}$),
- an inner part of the first partial table and a part from the last row of the second partial table (except of $x_{r_2s_2}$),
- the part $x_{r_1s_1}$ and a part from the last column of the second partial table (except of $x_{r_2s_2}$),
- the part $x_{r_1s_1}$ and a part from the last row of the second partial table (except of $x_{r_2s_2}$),

where we always deal with two "virtual" tables corresponding to the coordinates of interest. Similarly, B_4 is formed by variances of logratios of

- a part from the last column of the first partial table (except of $x_{r_1s_1}$) and an inner part of the second partial table,
- a part from the last column of the first partial table (except of $x_{r_1s_1}$) and the part $x_{r_2s_2}$,
- a part from the last row of the first partial table (except of $x_{r_1s_1}$) and an inner part of the second partial table,
- a part from the last row of the first partial table (except of $x_{r_1s_1}$) and the part $x_{r_2s_2}$.

On the other hand, the covariance is reduced by C_4 involving logratios between

- an inner part of the first partial table and an inner part of the second partial table,
- an inner part of the first table and the part part $x_{r_2s_2}$,
- the part $x_{r_1s_1}$ and an inner part of the second partial table,
- parts $x_{r_1s_1}$ and $x_{r_2s_2}$,

and by D_4 consisting of logratios formed by

- a part from the last column of the first partial table (except of $x_{r_1s_1}$) and a part from the last column of the second partial table (except of $x_{r_1s_1}$),
- a part from the last column of the first partial table (except of $x_{r_1s_1}$) and a part from the last row of the second partial table (except of $x_{r_2s_2}$),
- a part from the last row of the first partial table (except of $x_{r_1s_1}$) and a part from the last column of the second partial table (except of $x_{r_2s_2}$),
- a part from the last row of the first partial table (except of $x_{r_1s_1}$) and a part from the last row of the second partial table (except of $x_{r_2s_2}$).

Also covariance between two coordinates of the interaction table could supported by its graphical representation, see Figure 18.

Since coordinates of the independence table (68), (69) are balances obtained from sequential binary partitions, dividing rows and columns of the original table, respectively, their variances and covariances are obtained as direct consequence of [17].

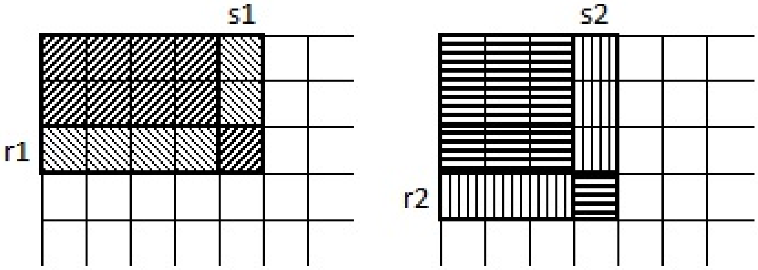


Figure 18: Covariance between coordinates $z_{r_1 s_1}$ and $z_{r_2 s_2}$ is increased by variances of logratios between a part of the first partial table from area highlighted by (/) and a part of the second partial table from area highlighted by (|) – A_4. The second group of variances increasing the covariance between coordinates are connected to logratios between parts from (\) and (−) areas – B_4. The covariance is reduced by variances of logratios between parts from (/) and (−) area – C_4 or two parts from (\) and (|) – D_4.

Theorem 2.9. *Consider coordinates of the independence table z_k^r for $k = 1, \dots, I-1$ and z_l^c for $l = 1, \dots, J-1$, then their variances are*

$$\begin{aligned}\operatorname{var}(z_k^r) &= K \sum_{i'=1}^{I-k} \sum_{j,j'=1}^{J} \operatorname{var}\left(\ln \frac{x_{I-k+1,j}}{x_{i'j'}}\right) - \frac{K}{2}(I-k) \sum_{j,j'=1}^{J} \operatorname{var}\left(\ln \frac{x_{I-k+1,j}}{x_{I-k+1,j'}}\right) - \\ &\quad - \frac{K}{2(I-k)} \sum_{i,i'=1}^{I-k} \sum_{j,j'=1}^{J} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{i'j'}}\right), \qquad (118)\end{aligned}$$

where $K = \frac{1}{J(I-k+1)}$, for balances between rows, and

$$\begin{aligned}\operatorname{var}(z_l^c) &= K \sum_{i,i'=I}^{I} \sum_{j'=1}^{J-l} \operatorname{var}\left(\ln \frac{x_{i,J-j+1}}{x_{i'j'}}\right) - \frac{K}{2}(J-l) \sum_{i,i'=1}^{I} \operatorname{var}\left(\ln \frac{x_{i,J-j+1}}{x_{i',J-j+1}}\right) - \\ &\quad - \frac{K}{2(J-l)} \sum_{i,i'=1}^{I} \sum_{j,j'=1}^{J-l} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{i'j'}}\right), \qquad (119)\end{aligned}$$

where $K = \frac{1}{I(J-l+1)}$, for balances between columns.

The variances of these coordinates are increased by variances of logratios between a part from the $(I-k+1)$-th row/$(J-l+1)$-th column and any part from the previous rows/columns. On the other hand, the variances of z_k^r and z_l^c are reduced by variances of logratios between parts from the same row/column.

According to relation (91) there are three main options how to get covariance between coordinates of the independence table, depending on concrete balances of interest. All these possible covariances are summarized in the following theorem.

Theorem 2.10. *Consider three row balances $z^r_{k_1}$, $z^r_{k_2}$ and z^r_k, for $k_1, k_2, k = 1, \ldots, I-1$, $k_1 \neq k_2$, computed using expression (68), and three column balances $z^c_{l_1}$, $z^c_{l_2}$ and z^c_l, for $l_1, l_2, l = 1, \ldots, J-1$, $l_1 \neq l_2$, computed from (69). Then*

$$\begin{aligned}
\operatorname{cov}(z^r_{k_1}, z^r_{k_2}) &= \frac{K}{(I-k_2)} \sum_{i'=1}^{I-k_2} \sum_{j,j'=1}^{J} \operatorname{var}\left(\ln \frac{x_{I-k_1+1,j}}{x_{i'j'}}\right) + \\
&\quad + \frac{K}{(I-k_1)} \sum_{i=1}^{I-k_1} \sum_{j,j'=1}^{J} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{I-k_2+1,j'}}\right) - \\
&\quad - K \sum_{j,j'=1}^{J} \operatorname{var}\left(\ln \frac{x_{I-k_1+1,j}}{x_{I-k_2+1,j'}}\right) - \\
&\quad - \frac{K}{(I-k_1)(I-k_2)} \sum_{i=1}^{I-k_1} \sum_{i'=1}^{I-k_2} \sum_{j,j'=1}^{J} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{i'j'}}\right) \quad , \qquad (120)
\end{aligned}$$

where $K = \frac{1}{2J}\sqrt{\frac{(I-k_1)(I-k_2)}{(I-k_1+1)(I-k_2+1)}}$, for row balances,

$$\begin{aligned}
\operatorname{cov}(z^c_{l_1}, z^c_{l_2}) &= \frac{K}{(J-l_2)} \sum_{i,i'=1}^{I} \sum_{j'=1}^{J-l_2} \operatorname{var}\left(\ln \frac{x_{i,J-l_1+1}}{x_{i'j'}}\right) + \\
&\quad + \frac{K}{(J-l_1)} \sum_{i,i'=1}^{I} \sum_{j=1}^{J-l_1} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{i',J-l_2+1}}\right) - \\
&\quad - K \sum_{i,i'=1}^{I} \operatorname{var}\left(\ln \frac{x_{i,J-l_1+1}}{x_{i',J-l_2+1}}\right) - \\
&\quad - \frac{K}{(J-l_1)(J-l_2)} \sum_{i,i'=1}^{I} \sum_{j=1}^{J-l_1} \sum_{j'=1}^{J-l_2} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{i'j'}}\right) \quad , \qquad (121)
\end{aligned}$$

where $K = \frac{1}{2I}\sqrt{\frac{(J-l_1)(J-l_2)}{(J-l_1+1)(J-l_2+1)}}$, for column balances, and

$$\begin{aligned}
\operatorname{cov}(z^r_k, z^c_l) &= \frac{K}{(J-l)} \sum_{i'=1}^{I} \sum_{j=1}^{J} \sum_{j'=1}^{J-l} \operatorname{var}\left(\ln \frac{x_{I-k+1,j}}{x_{i'j'}}\right) + \\
&\quad + \frac{K}{(I-k)} \sum_{i=1}^{I} \sum_{i'=1}^{I-k} \sum_{j'=1}^{J} \operatorname{var}\left(\ln \frac{x_{i,J-l+1}}{x_{i'j'}}\right) - \\
&\quad - K \sum_{i'=1}^{I} \sum_{j=1}^{J} \operatorname{var}\left(\ln \frac{x_{I-k+1,j}}{x_{i',J-l+1}}\right) - \\
&\quad - \frac{K}{(I-k)(J-l)} \sum_{i=1}^{I-k} \sum_{i'=1}^{I} \sum_{j=1}^{J} \sum_{j'=1}^{J-l} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{i'j'}}\right) \quad , \qquad (122)
\end{aligned}$$

where $K=\frac{1}{2}\sqrt{\frac{(I-k)(J-l)}{IJ(I-k+1)(J-l+1)}}$, *between row and column balances.*

To complete the covariance structure of coordinates of the compositional table $\mathbf{x}$, covariances between coordinates of the interaction and independence tables z_{rs}, z_k^r and z_l^c, respectively, are necessary. They are provided in the last theorem.

Theorem 2.11. *Consider coordinate of the interaction table z_{rs}, for $r=2,\dots,I$ and $s=2,\dots,J$, and two coordinates of the independence table, z_k^r, for $k=1,\dots,I-1$, and z_l^c, for $l=1,\dots,J-1$. Then for covariances between coordinates of the interaction and independence tables the following hold,*

$$\operatorname{cov}(z_{rs},z_k^r)=K\cdot(A_5-B_5)\quad, \tag{123}$$

where

$$\begin{aligned} A_5 &= \frac{1}{J(I-k)}\sum_{i=1}^{r-1}\sum_{i'=1}^{I-k}\sum_{j=1}^{s-1}\sum_{j'=1}^{J}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right)+\\ &+\frac{s-1}{J}\sum_{i=1}^{r-1}\sum_{j'=1}^{J}\operatorname{var}\left(\ln\frac{x_{is}}{x_{I-k+1,j'}}\right)+\\ &+\frac{r-1}{J}\sum_{j=1}^{s-1}\sum_{j'=1}^{J}\operatorname{var}\left(\ln\frac{x_{rj}}{x_{I-k+1,j'}}\right)+\\ &+\frac{(r-1)(s-1)}{J(I-k)}\sum_{i'=1}^{I-k}\sum_{j'=1}^{J}\operatorname{var}\left(\ln\frac{x_{rs}}{x_{i'j'}}\right)\quad, \end{aligned} \tag{124}$$

$$\begin{aligned} B_5 &= \frac{1}{J}\sum_{i=1}^{r-1}\sum_{j=1}^{s-1}\sum_{j'=1}^{J}\operatorname{var}\left(\ln\frac{x_{ij}}{x_{I-k+1,j'}}\right)+\\ &+\frac{s-1}{J(I-k)}\sum_{i=1}^{r-1}\sum_{i'=1}^{I-k}\sum_{j'=1}^{J}\operatorname{var}\left(\ln\frac{x_{is}}{x_{i'j'}}\right)+\\ &+\frac{r-1}{J(I-k)}\sum_{i'=1}^{I-k}\sum_{j=1}^{s-1}\sum_{j'=1}^{J}\operatorname{var}\left(\ln\frac{x_{rj}}{x_{i'j'}}\right)+\\ &+\frac{(r-1)(s-1)}{J}\sum_{j'=1}^{J}\operatorname{var}\left(\ln\frac{x_{rs}}{x_{I-k+1,j'}}\right)\quad, \end{aligned} \tag{125}$$

for $K=\frac{1}{2}\frac{1}{\sqrt{rs(r-1)(s-1)}}\sqrt{\frac{J(I-k)}{I-k+1}}$, *and*

$$\operatorname{cov}(z_{rs},z_l^c)=K\cdot(A_6-B_6)\quad, \tag{126}$$

where

$$
\begin{aligned}
A_6 \quad = \quad & \frac{1}{I(J-l)} \sum_{i=1}^{r-1} \sum_{i'=1}^{I} \sum_{j=1}^{s-1} \sum_{j'=1}^{J-l} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{i'j'}}\right) + \\
& + \frac{s-1}{I} \sum_{i=1}^{r-1} \sum_{i'=1}^{I} \operatorname{var}\left(\ln \frac{x_{is}}{x_{i',J-l+1}}\right) + \\
& + \frac{r-1}{I} \sum_{i'=1}^{I} \sum_{j=1}^{s-1} \operatorname{var}\left(\ln \frac{x_{rj}}{x_{i',J-l+1}}\right) + \\
& + \frac{(r-1)(s-1)}{I(J-l)} \sum_{i'=1}^{I} \sum_{j'=1}^{J-l} \operatorname{var}\left(\ln \frac{x_{rs}}{x_{i'j'}}\right) \quad , \qquad (127) \\
B_6 \quad = \quad & \frac{1}{I} \sum_{i=1}^{r-1} \sum_{i'=1}^{I} \sum_{j=1}^{s-1} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{i',J-l+1}}\right) + \\
& + \frac{s-1}{I(J-l)} \sum_{i=1}^{r-1} \sum_{i'=1}^{I} \sum_{j'=1}^{J-l} \operatorname{var}\left(\ln \frac{x_{is}}{x_{i'j'}}\right) + \\
& + \frac{r-1}{I(J-l)} \sum_{i'=1}^{I} \sum_{j=1}^{s-1} \sum_{j'=1}^{J-l} \operatorname{var}\left(\ln \frac{x_{rj}}{x_{i'j'}}\right) + \\
& + \frac{(r-1)(s-1)}{I} \sum_{i'=1}^{I} \operatorname{var}\left(\ln \frac{x_{rs}}{x_{i',J-l+1}}\right) \quad , \qquad (128)
\end{aligned}
$$

for $K = \frac{1}{2}\frac{1}{\sqrt{rs(r-1)(s-1)}}\sqrt{\frac{I(J-l)}{J-l+1}}$.

Proof: *The assertion of the theorem is a direct consequence of Proposition 2.1 and Equations (68), (69) and (79).*

◇

Similarly as for the case of interaction table, also the above results can be interpreted graphically. Because Theorems 2.9 and 2.10 represent a special case of balances, that were in detail analyzed in [17], in Figure 19 we focus just on covariances, resulting from Theorem 2.11.

Example 4 To illustrate the presented theoretical outputs, let us consider the sample of eighteen 2×3 compositional tables, each containing population structure in a given European country according to age and BMI index ((weight in kg)/(height in m)^2), with values $25-44, 45-64, 65-84$ and under- or normal weight and

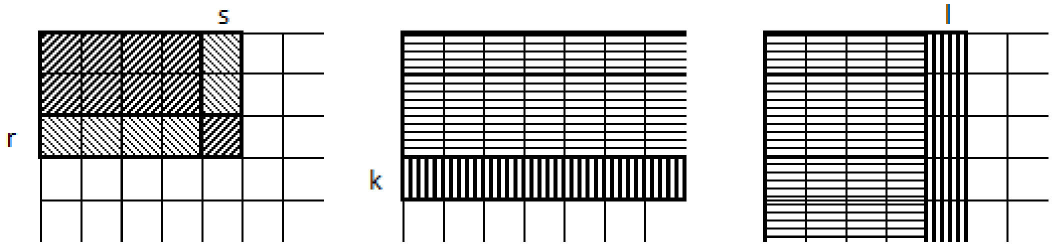


Figure 19: Covariance between a coordinate of the interaction table, z_{rs} (left), and coordinates of the independence table, z_k^r (middle) or z_l^c (right), is increased by variances of logratios between parts from areas (/) and (−), or (\) and (|), respectively, and reduced by variances of logratios between parts from areas (/) and (|), or (\) and (−), respectively.

overweight or obesity, respectively. This data set is an aggregated version of data from Example 3. Table 10 shows an example of compositional table from the sample.

AUT	25 − 44	45 − 64	65 − 84
under or normal	0.249	0.144	0.074
over or obesity	0.171	0.221	0.140

Table 10: Structure of population in Austria in 2008 according to age and BMI index (in proportions).

Firstly, each table from the sample has been expressed in pivot coordinates (79) and, consequently, their descriptive statistics were calculated. Note that the sample mean is

$$\bar{\mathbf{z}} = (0.409, 0.294, -0.450, 0.578, 0.637) \quad , \tag{129}$$

but for our purposes the covariance structure of the sample is of primary interest. The variation matrix (88), as a source of elemental information in compositional tables, equals

$$\mathbf{T} = \begin{pmatrix} 0 & 0.037 & 0.083 & 0.024 & 0.030 & 0.069 \\ 0.037 & 0 & 0.030 & 0.077 & 0.050 & 0.051 \\ 0.083 & 0.030 & 0 & 0.127 & 0.098 & 0.065 \\ 0.024 & 0.077 & 0.127 & 0 & 0.019 & 0.078 \\ 0.030 & 0.050 & 0.098 & 0.019 & 0 & 0.040 \\ 0.069 & 0.051 & 0.065 & 0.078 & 0.040 & 0 \end{pmatrix} \quad . \tag{130}$$

For example, using this matrix and equation (108), variance of the first coordinate of the interaction table, z_{22}, can be obtained as

$$\begin{aligned} \mathrm{var}(z_{22}) &= \tfrac{1}{4}\mathrm{var}\left(\ln\tfrac{x_{11}}{x_{21}}\right)+\tfrac{1}{4}\mathrm{var}\left(\ln\tfrac{x_{11}}{x_{12}}\right)+\tfrac{1}{4}\mathrm{var}\left(\ln\tfrac{x_{21}}{x_{22}}\right)+\tfrac{1}{4}\mathrm{var}\left(\ln\tfrac{x_{12}}{x_{22}}\right)\\ &\quad -\tfrac{1}{8}\mathrm{var}\left(\ln\tfrac{x_{11}}{x_{11}}\right)-\tfrac{1}{4}\mathrm{var}\left(\ln\tfrac{x_{11}}{x_{22}}\right)-\tfrac{1}{8}\mathrm{var}\left(\ln\tfrac{x_{21}}{x_{21}}\right)-\tfrac{1}{8}\mathrm{var}\left(\ln\tfrac{x_{12}}{x_{12}}\right)\\ &\quad -\tfrac{1}{4}\mathrm{var}\left(\ln\tfrac{x_{12}}{x_{21}}\right)\\ &= \tfrac{1}{4}t_{14}+\tfrac{1}{4}t_{12}+\tfrac{1}{4}t_{45}+\tfrac{1}{4}t_{25}-\tfrac{1}{8}t_{11}-\tfrac{1}{4}t_{15}-\tfrac{1}{8}t_{44}-\tfrac{1}{8}t_{22}-\tfrac{1}{4}t_{24}\\ &= 0.0057 \quad . \end{aligned} \tag{131}$$

By comparing with the corresponding elements of the variation matrix we can conclude that none of logratios contributes exceptionally (in the positive sense) to variability of the coordinate. In the negative sense, the logratio ln(*underweight or normal weight in age 45-64/overweight or obesity in age 25-44*) shows a dominant effect. Similarly, also other variances and covariances can be derived (and further analysed for structural patterns), resulting in a covariance matrix

$$\mathrm{var}(\mathbf{z}) = \begin{pmatrix} 0.006 & 0.003 & 0.010 & -0.007 & -0.004\\ 0.003 & 0.013 & 0.007 & 0.002 & 0.003\\ 0.010 & 0.007 & 0.051 & -0.021 & -0.012\\ -0.007 & 0.002 & -0.021 & 0.055 & 0.024\\ -0.004 & 0.003 & -0.012 & 0.024 & 0.022 \end{pmatrix} \quad . \tag{132}$$

2.4.3 Coordinates of 2×2 tables

An interesting interpretation results from covariance structure of coordinates of the smallest possible table with 2 rows and 2 columns. Also this case is discussed in [12]. By applying the above theorems to coordinates

$$z^r = \frac{1}{2}\ln\frac{x_{11}x_{12}}{x_{21}x_{22}}, \quad z^c = \frac{1}{2}\ln\frac{x_{11}x_{21}}{x_{12}x_{22}} \quad \text{and} \quad z^{OR} = \frac{1}{2}\ln\frac{x_{11}x_{22}}{x_{12}x_{21}} \quad , \tag{133}$$

their covariance structure can be easily derived,

$$\begin{aligned}
\mathrm{var}(z_{22}) &= \frac{1}{4}\left[\mathrm{var}\left(\ln\frac{x_{11}}{x_{21}}\right)+\mathrm{var}\left(\ln\frac{x_{11}}{x_{12}}\right)+\mathrm{var}\left(\ln\frac{x_{21}}{x_{22}}\right)\right.\\
&\left.+\mathrm{var}\left(\ln\frac{x_{12}}{x_{22}}\right)-\mathrm{var}\left(\ln\frac{x_{11}}{x_{22}}\right)-\mathrm{var}\left(\ln\frac{x_{12}}{x_{21}}\right)\right] \quad , \qquad (134)\\
\mathrm{var}(z_1^r) &= \frac{1}{4}\left[\mathrm{var}\left(\ln\frac{x_{11}}{x_{21}}\right)+\mathrm{var}\left(\ln\frac{x_{11}}{x_{22}}\right)+\mathrm{var}\left(\ln\frac{x_{12}}{x_{21}}\right)\right.\\
&\left.+\mathrm{var}\left(\ln\frac{x_{12}}{x_{22}}\right)-\mathrm{var}\left(\ln\frac{x_{11}}{x_{12}}\right)-\mathrm{var}\left(\ln\frac{x_{21}}{x_{22}}\right)\right] \quad , \qquad (135)\\
\mathrm{var}(z_1^c) &= \frac{1}{4}\left[\mathrm{var}\left(\ln\frac{x_{11}}{x_{12}}\right)+\mathrm{var}\left(\ln\frac{x_{11}}{x_{22}}\right)+\mathrm{var}\left(\ln\frac{x_{21}}{x_{12}}\right)\right.\\
&\left.+\mathrm{var}\left(\ln\frac{x_{21}}{x_{22}}\right)-\mathrm{var}\left(\ln\frac{x_{11}}{x_{21}}\right)-\mathrm{var}\left(\ln\frac{x_{12}}{x_{22}}\right)\right] \quad , \qquad (136)\\
\mathrm{cov}(z_{22},z_1^r) &= \frac{1}{4}\left[\mathrm{var}\left(\ln\frac{x_{11}}{x_{21}}\right)-\mathrm{var}\left(\ln\frac{x_{22}}{x_{12}}\right)\right] \quad , \qquad (137)\\
\mathrm{cov}(z_{22},z_1^c) &= \frac{1}{4}\left[\mathrm{var}\left(\ln\frac{x_{11}}{x_{12}}\right)-\mathrm{var}\left(\ln\frac{x_{21}}{x_{22}}\right)\right] \quad , \qquad (138)\\
\mathrm{cov}(z_1^r,z_1^c) &= \frac{1}{4}\left[\mathrm{var}\left(\ln\frac{x_{11}}{x_{22}}\right)-\mathrm{var}\left(\ln\frac{x_{12}}{x_{21}}\right)\right] \quad . \qquad (139)
\end{aligned}$$

Moreover, from the above covariance structure it is also interesting to see that coordinates (133) are uncorrelated (or even independent under the assumption of normality) if, and only if

$$\mathrm{var}\left(\ln\frac{x_{11}}{x_{21}}\right)=\mathrm{var}\left(\ln\frac{x_{12}}{x_{22}}\right),\quad \mathrm{var}\left(\ln\frac{x_{11}}{x_{12}}\right)=\mathrm{var}\left(\ln\frac{x_{21}}{x_{22}}\right) \quad ,$$

$$\text{and}\quad \mathrm{var}\left(\ln\frac{x_{11}}{x_{22}}\right)=\mathrm{var}\left(\ln\frac{x_{12}}{x_{21}}\right) \quad . \qquad (140)$$

In other words, it means that zero covariances can be easily expressed in terms of logratio variances. Consequently, the above relations could be used, e.g., by designing simulation settings for 2 × 2 compositional tables using elements of the variation matrix as a source of elemental information in covariance structure of compositional tables.

Following [11], it is possible to assign also another system of orthonormal coordinates to a 2 × 2 compositional table (see Section 2.3.3). Specifically, we get

$$z_1^{ind}=\frac{1}{\sqrt{2}}\ln\frac{x_{12}}{x_{21}},\quad z_2^{ind}=\frac{1}{\sqrt{2}}\ln\frac{x_{11}}{x_{22}},\quad z^{int}=\frac{1}{2}\ln\frac{x_{11}x_{22}}{x_{12}x_{21}} \quad , \qquad (141)$$

for the interaction and independent tables, respectively, and the covariance structure changes as follows,

$$\begin{aligned}
\mathrm{var}(z^{int}) &= \frac{1}{4}\left[\mathrm{var}\left(\ln\frac{x_{11}}{x_{12}}\right)+\mathrm{var}\left(\ln\frac{x_{11}}{x_{21}}\right)+\mathrm{var}\left(\ln\frac{x_{12}}{x_{22}}\right)\right.\\
&\quad\left.+\mathrm{var}\left(\ln\frac{x_{21}}{x_{22}}\right)-\mathrm{var}\left(\ln\frac{x_{11}}{x_{22}}\right)-\mathrm{var}\left(\ln\frac{x_{12}}{x_{21}}\right)\right] \quad , &(142)\\
\mathrm{var}(z_1^{ind}) &= \frac{1}{2}\mathrm{var}\left(\ln\frac{x_{12}}{x_{21}}\right) \quad , &(143)\\
\mathrm{var}(z_2^{ind}) &= \frac{1}{2}\mathrm{var}\left(\ln\frac{x_{11}}{x_{22}}\right) \quad , &(144)\\
\mathrm{cov}(z^{int},z_1^{ind}) &= \frac{1}{4\sqrt{2}}\left[\mathrm{var}\left(\ln\frac{x_{11}}{x_{12}}\right)+\mathrm{var}\left(\ln\frac{x_{11}}{x_{21}}\right)-\mathrm{var}\left(\ln\frac{x_{12}}{x_{22}}\right)\right.\\
&\quad\left.-\mathrm{var}\left(\ln\frac{x_{21}}{x_{22}}\right)\right] \quad , &(145)\\
\mathrm{cov}(z^{int},z_2^{ind}) &= \frac{1}{4\sqrt{2}}\left[\mathrm{var}\left(\ln\frac{x_{11}}{x_{21}}\right)+\mathrm{var}\left(\ln\frac{x_{21}}{x_{22}}\right)-\mathrm{var}\left(\ln\frac{x_{11}}{x_{12}}\right)\right.\\
&\quad\left.-\mathrm{var}\left(\ln\frac{x_{12}}{x_{22}}\right)\right] \quad , &(146)\\
\mathrm{cov}(z_1^{ind},z_2^{ind}) &= \frac{1}{4}\left[\mathrm{var}\left(\ln\frac{x_{11}}{x_{21}}\right)+\mathrm{var}\left(\ln\frac{x_{12}}{x_{22}}\right)-\mathrm{var}\left(\ln\frac{x_{11}}{x_{12}}\right)\right.\\
&\quad\left.-\mathrm{var}\left(\ln\frac{x_{21}}{x_{22}}\right)\right] \quad . &(147)
\end{aligned}$$

Now, although coordinates of the independent table are formed just by (scaled) logratios, the covariance structure becomes more complex than before. For example, coordinates (141) are *mutually* uncorrelated (independent) if, and only if

$$\mathrm{var}\left(\ln\frac{x_{11}}{x_{12}}\right)=\mathrm{var}\left(\ln\frac{x_{11}}{x_{21}}\right)=\mathrm{var}\left(\ln\frac{x_{12}}{x_{22}}\right)=\mathrm{var}\left(\ln\frac{x_{21}}{x_{22}}\right) \quad . \tag{148}$$

In other words, it means that $\mathrm{var}\left(\ln\frac{x_{12}}{x_{21}}\right)$ and $\mathrm{var}\left(\ln\frac{x_{11}}{x_{22}}\right)$ are influential just for variances of coordinates z_1^{ind}, z_2^{ind}, z^{int}, forming also natural constraints for their possible values.

3 Analysis of relationships between two factors

A natural aim of the analysis of compositional tables is to study relationships between its row and column factors. In [7] it was proposed to measure distance between the original compositional table $\mathbf{x}$ and its independent part $\mathbf{x}_{ind}$ using squared distance

$$\Delta^2(\mathbf{x}) = ||\mathbf{x}_{int}||_A^2 = ||\mathbf{x}||_A^2 - ||\mathbf{x}_{ind}||_A^2 \quad , \tag{149}$$

or relative squared distance

$$R_\Delta^2(\mathbf{x}) = \frac{\Delta^2(\mathbf{x})}{||\mathbf{x}||_A^2}, \quad 0 \leq R_\Delta^2 \leq 1 \quad , \tag{150}$$

which suppresses the influence of dimensions of compositional table $\mathbf{x}$ on squared distance. Values of relative squared distance, which are near to 1, are typical for tables with strong interactions between factors. On the other hand, low values give an evidence about independence between row and column factors. Moreover, due to decomposition

$$\mathbf{x} = \mathbf{x}_{ind} \oplus \left(\bigoplus_{i=1}^{I} \mathrm{row}_i(\mathbf{x}_{int}) \right) = \mathbf{x}_{ind} \oplus \left(\bigoplus_{j=1}^{J} \mathrm{col}_j(\mathbf{x}_{int}) \right) \quad , \tag{151}$$

the contribution of the i-th row to the squared norm is $||\mathrm{row}_i(\mathbf{x}_{int})||_A^2$ and similarly contribution of the j-th column is $||\mathrm{col}_j(\mathbf{x}_{int})||_A^2$.

Since orthonormal coordinates accounting for interactions between rows and columns were not available in [7], these features were measured using cross-contrasts and so called cell-interactions. The cross-contrast is defined as a simple balance of the part of interaction table at position (i, j) against the other parts in the same row or column,

$$I_{cross}(i,j) = \sqrt{\frac{I+J-2}{I+J-1}} \ln \frac{x_{ij}^{int}}{\left(\prod_{r \neq i} x_{rj}^{int} \prod_{s \neq j} x_{is}^{int} \right)^{1/(I+J-2)}} \quad . \tag{152}$$

The problem of these balances is that they are not orthogonal. On the other hand, their sum is closely connected to the square norm of $\mathbf{x}_{int}$ through relation

$$\sum_i \sum_j \left(I_{cross}(i,j) \right)^2 = \frac{(I+J)^2}{(I+J-1)(I+J-2)} ||\mathbf{x}_{int}||_A^2 \quad . \tag{153}$$

The cell-interaction is defined as balance between part of interaction table at position (i, j) and the rest of parts

$$I_{cell}(i,j) = \sqrt{\frac{IJ-1}{IJ}} \ln \frac{x_{ij}^{int}}{\left(\prod_{(k,l) \neq (i,j)} x_{kl}^{int} \right)^{1/(IJ-1)}} \quad , \tag{154}$$

and is also connected to squared norm of the interaction table

$$\sum_i \sum_j \left(I_{cell}(i,j)\right)^2 = \frac{IJ}{I+J-1} ||\mathbf{x}_{int}||_A^2 \quad . \qquad (155)$$

Furthermore, it can be shown, that in the case of 2×2 table all the cell-interactions are the same (up to its sign) and proportional to the interaction coordinate (52),

$$I_{cell}(1,1) = I_{cell}(2,2) = -I_{cell}(1,2) = -I_{cell}(2,1) = \frac{1}{2\sqrt{3}} \ln \frac{x_{11}x_{22}}{x_{12}x_{21}} \quad . \qquad (156)$$

However, coordinate system proposed in Section 2.3.1 or its special case from Sections 2.3.2 and 2.3.3 enable a deeper insight into the source of interactions between both factors, by considering interpretation of the odds ratio coordinates of the interaction table. Particularly if row and column factors are independent ($\mathbf{x} = \mathbf{x}_{ind}$), the interaction table equals to a neutral element of perturbation, all its parts are the same and the vector of odds ratio coordinates (52) $\mathbf{z}_{int}$ equals to a zero vector. On the other hand, high absolute values of this vector indicate presence of interactions between factors. Consequently, in the situation when a random sample of compositional tables is available, the analysis of independence reduces to multivariate test on zero mean value of the vector of interaction coordinates $\mathbf{z}_{int}$. The structural approach to the analysis of independence between factors is also supported by the interpretation of these coordinates. As it was described in Section 2.3.1, coordinate z_i^{OR} can be interpreted as logarithm of odds ratio among groups of parts. Since in the independence case the odds ratio equals one, zero values of coordinates give an evidence against the presence of interactions between factors.

The information about relations within the table can be completed using the remaining coordinates. Row and column balances $\mathbf{z}^r, \mathbf{z}^c$ focus exclusively on relations between levels of row or column factor, respectively. The proposed coordinate system thus forms also a compositional alternative to log-linear model approach, since parameters of two-factor log-linear model with interactions can be also interpreted in the sense of logarithm of ratio between different levels of row or column factors, or logarithm of odds ratio. Specifically, in the case of binary factors (2 × 2 contingency table) the respective log-linear model is

$$\ln x_{ij} = \beta_0 + \beta_1 I_{2nd\_row} + \beta_2 I_{2nd\_column} + \beta_3 I_{2nd\_row} I_{2nd\_column}, \quad i,j = 1,2 \quad , \qquad (157)$$

where I_{2nd_row} and I_{2nd_column} are identifiers of the second row and column, respectively. Parameter β_1 indicates value of the logratio between parts x_{11} and x_{21} and analogously parameter β_2 describes the logratio between parts in the first row ($\ln \frac{x_{11}}{x_{12}}$). Finally, the last parameter β_3 defines log-odds ratio within this table and this parameter vanishes when row and column factors are independent.

As was already described in Section 2.3.3, coordinates of a 2 × 2 compositional table have form

$$z^r = \frac{1}{2}\ln\frac{x_{11}x_{12}}{x_{21}x_{22}}, \quad z^c = \frac{1}{2}\ln\frac{x_{11}x_{21}}{x_{12}x_{22}}, \quad z^{OR} = \frac{1}{2}\ln\frac{x_{11}x_{22}}{x_{12}x_{21}} \quad . \tag{158}$$

Consequently, between parameters of the log-linear model and orthonormal coordinates the following relations hold

$$z^r = -\beta_1 - \frac{\beta_3}{2}, \quad z^c = -\beta_2 - \frac{\beta_3}{2}, \quad z^{OR} = \frac{\beta_3}{2} \quad . \tag{159}$$

In terms of log-linear models, zero value of parameter β_3 means, that there are no interactions between row and column factor, similarly as zero value of coordinate z^{OR}.

The applicability of the proposed coordinate representation of compositional tables to analysis of relationship between factors is illustrated in the following with two real-world examples.

Example 5 (Relationship between age and BMI index - part 2) Let's continue with the Example 3 from Section 2.3.2. Description of data as well as interpretation of coordinates were already provided there. We focus now on relations between age and BMI index. If the factors were independent, the interaction table would equal to the neutral element on the simplex, i.e. all parts would be approximately $1/(IJ) = 1/12 = 0.0833$. In case of the Czech Republic it is easy to see that this condition does not hold as well as in the case of the other countries. This feature is clearly visible also from the mean interaction table (in sense of the Aitchison geometry)

$$\overline{\mathbf{x}}_{int} = \frac{1}{n} \odot \bigoplus_{k=1}^{n} \mathbf{x}_{int,k} = \begin{pmatrix} 0.1483 & 0.0967 & 0.0589 & 0.0465 \\ 0.0554 & 0.0753 & 0.0917 & 0.1031 \\ 0.0604 & 0.0682 & 0.0922 & 0.1035 \end{pmatrix} \quad . \tag{160}$$

Despite relatively high standard deviations of some coordinates with respect to the corresponding means, the above findings lead to a preliminary conclusion that age and BMI index are not independent.

In order to extend the univariate conclusions to a multivariate one, the coordinates of the interaction table as well as of the original compositional table and the independence table are also analyzed using the well-known biplot [18] of the first two principal components of the corresponding coordinates. In Figure 20 biplots of the original, independence and interaction tables are collected. The biplot of the original compositional table seems to be dominated by high variability of the coordinates of the independence table, thus here mainly the data structure (with Romania and Slovakia as outlying observations) can be observed. The other two biplots provide further information on the relations leading to independence and interaction between the age and BMI factors.

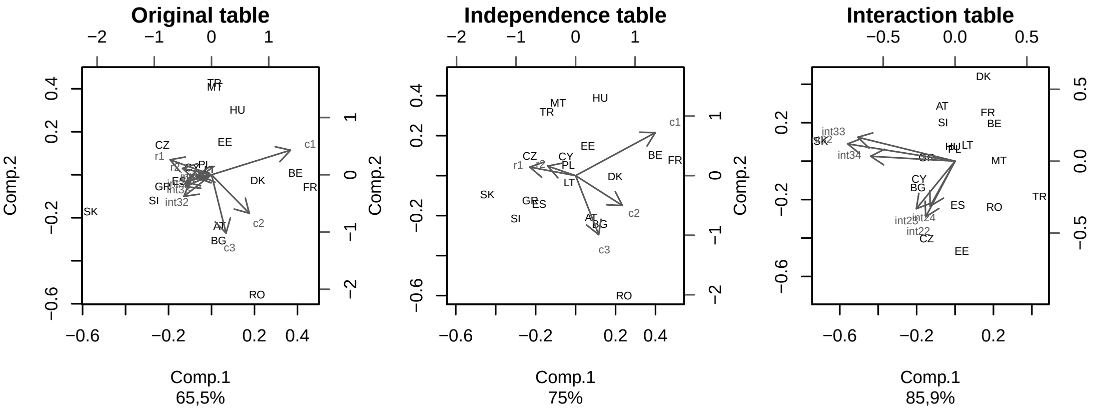


Figure 20: Biplots of coordinates of original, independence, and interaction tables.

The biplot of the independence table shows that its first two nonzero coordinates (that carry relative information on relations between the rows of the original table, i.e. age ranges) are strongly positively correlated, and also negatively correlated with the remaining three nonzero coordinates (explaining the relations between the columns representing BMI categories). From the directions of arrows (loadings) we can observe that moving from the left to the right side of the biplot, the values of the first two coordinates decrease and the next three coordinates increase. Also the locations of countries resulting from the principal component scores provide interesting information about the data structure, like cluster of countries Belgium, Denmark and France with quite high values of coordinates z_1^c, z_2^c and z_3^c. It means that these countries contribute substantially to the independence between age and BMI index, in particular due to the high relative contributions of underweight people over all other age categories. Similarly, we can observe Romania as an outlying observation with particular importance of the positive ratio between overweight and obese people. Contrary, Poland and Lithuania lay in the centre of the biplot. The centre represents an average behaviour of both factors.

The interaction biplot shows some interesting features as well. In particular, the first three nonzero coordinates are strongly correlated and the last three ones as well, but no correlation between these two groups is visible. This means that odds ratios with the third row of the interaction table (age range $65-84$ years) yield results different from those within the younger categories. Also in this biplot, Belgium, Denmark and France are placed quite near to each other and these western European countries together with Switzerland and Austria represent states with lower values of all coordinates, thus with rather smaller BMI (weight) growth for increasing age. On the other hand, countries like Czech Republic and Estonia with high values of coordinates z_{22}, z_{23} and z_{24} indicate a substantial weight growth from the younger to middle age generation, and thus contribute substantially to interaction between the factors. An interesting outlier is represented by Turkey with small values of coordinates z_{32}, z_{33} and z_{34}. This testifies that the weight growth tends to be particularly small from $25-44$ and $45-64$ to $65-84$ age group, just conversely to Slovakia. Nearest to the origin are placed Poland and Lithuania again, i.e. these countries do not contribute neither to interaction nor independence between the age and BMI factors.

Interestingly, small correlation between coordinates z_{22}, z_{23}, z_{24} and z_{32}, z_{33}, z_{34} remains unaltered when rows of the original compositional tables are permuted, see Figure 21. This result indicates an independence behaviour of single row factor values (age groups) with respect to BMI categories.

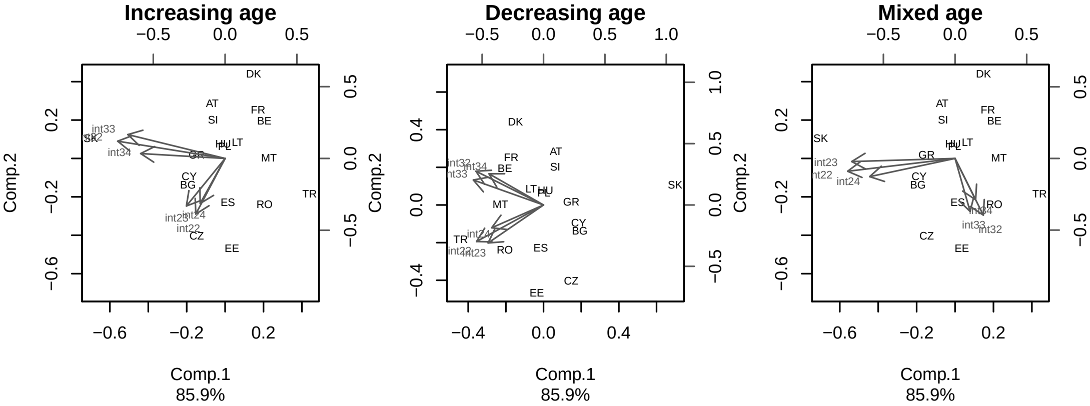


Figure 21: Biplots of coordinates of the interaction table with rows in increasing ($25-44$, $45-64$ and $65-84$), decreasing ($65-84$, $45-64$ and $25-44$) and mixed ($25-44$, $65-84$ and $45-64$) order of age categories.

Example 6 (Distribution of manufacturing output) Also this application discusses the possibility of independence analysis between two factors using a sample of compositional tables. For this purpose, the sample of 42 3×5 compositional tables, each representing distribution of manufacturing output in a given country in years 2007–2009, is available. This example is taken from [14]. Tables to be analyzed focus on food and beverages production only, distributed according to manufacturing categories, which were obtained using the 3-digit level of the International Standard Industrial Classification of All Economic Activities ISIC (Revision 3) [25]. Values of this first factor are thus as follows (numbers correspond to ISIC coding):

- 151 Processed meat, fish, fruit, vegetables, fats.
- 152 Dairy products.

Table 11: Distribution of USA food and beverages production in 2008 according to ISIC category and source of production (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

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

The second factor is formed by components of the output with categories Labour, Surplus and Input. Since the interest is devoted to relative structure of the output, the compositional approach is preferred. Percentage representation for one table from the sample is provided in Table 11.

In order to express each table in coordinates, the SBPs of row and column factors should be defined first. In the case of manufacturing categories, it seems to be logical to separate the production of beverages from all food products in the first step. The next steps could be based on separation of the category with not well specified types of food products (154 Other food products) from the remaining three, followed by the separation of supplementary products (153 Grain mill products, starches, animal feeds). Finally, the last step separates categories 151 (Processed meat, fish, fruit, vegetables, fats) and 152 (Dairy products). Similarly, the components of production should be firstly divided onto Input and value added (Labour and Surplus) components, which are further divided in the second step. These two SBPs, visualized graphically in Table 12, determine uniquely coordinate representation of the compositional tables in the sample. This means, the whole set of coordinates $\mathbf{z}$ can be immediately computed for each table from the sample. Since only one category was separated in each step of the SBPs, the resulting set of coordinates corresponds to those proposed and extensively described in Section 2.3.2 and [13]. Due to easy construction and interpretability, this coordinate representation can be also considered as a basic option for compositional tables. Accordingly, both Beverages and Input categories have an exceptional position as there are coordinates that capture their relative contributions with respect to the other categories in rows (z_1^r), columns (z_1^c) of the tables and by considering interactions between both factors (z_1^{OR}). It is thus a natural generalization of the approach to interpretable balances for compositional data as introduced in [17] and recently employed in a range of applications [16, 22, 21].

By following the presented methodology and proposed SBPs, each table

Table 12: Sequential binary partition of manufacturing categories (left) and sources of output (right).

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

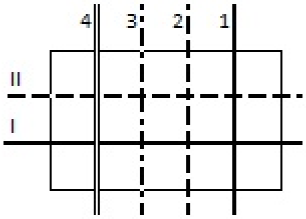


Figure 22: Graphical representation of sequential binary partitions SBPr and SBPc, defined on Table 12.

from the sample has been expressed in coordinates. For example, coordinate representation of the model table, distribution of output in USA, results in

$$\mathbf{z}^r_{USA} = (2.52, 2.39) \quad , \tag{161}$$

$$\mathbf{z}^c_{USA} = (-0.68, 0.92, -0.89, -1.34) \quad , \tag{162}$$

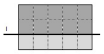
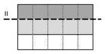
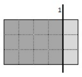
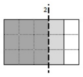
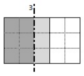
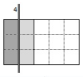
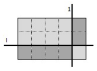
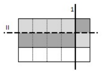
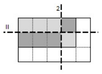
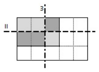
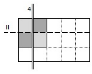
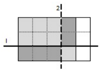
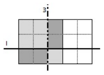
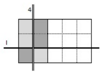
$$\mathbf{z}^{OR}_{USA} = (-0.33, 0.25, -0.09, 0.49, 0.09, -0.67, -0.09, 0.11) \quad , \tag{163}$$

$$\mathbf{z}_{USA} = \left(\mathbf{z}^r_{USA}, \mathbf{z}^c_{USA}, \mathbf{z}^{OR}_{USA}\right) \quad . \tag{164}$$

The positive values of $\mathbf{z}^r_{USA}$ indicate that Input dominates the value added components and, further, Surplus dominates Labour across all (averaged) food and beverage branches of the US economics. Production of beverages is slightly dominated by average production of food components; this feature is captured by the first coordinate of $\mathbf{z}^c_{USA}$, which equals -0.68. Relationships between production sources and manufacturing categories are described by vector of coordinates $\mathbf{z}^{OR}_{USA}$. Because most of its values are not far from zero, it suggests near independence between the factors.

These very preliminary observations for the case of USA are followed by detailed inspection of the whole data structure. As a result of dimension reduction using principal component analysis, biplot of row and column balances and

Table 13: List of coordinates in the second example together with their graphical representations.

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

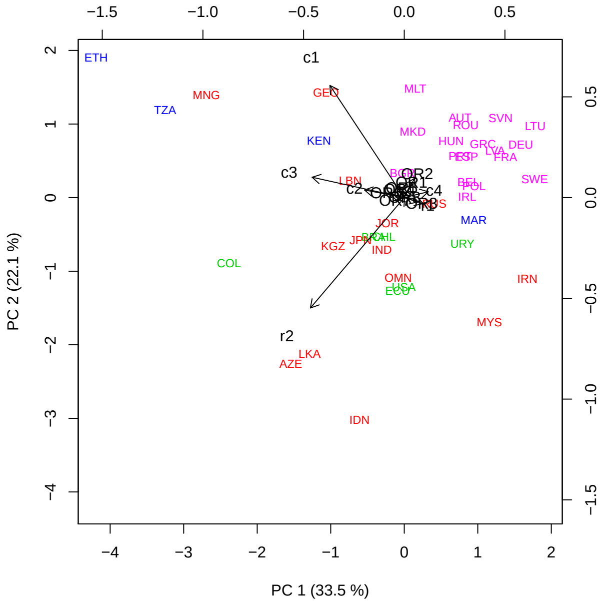


Figure 23: Biplot of compositional tables in coordinates with countries divided according to continent pertinence (Europe - purple, Africa - blue, America - green, Asia - red).

odds ratio coordinates is displayed in Figure 23. All coordinates are just centred prior to further processing as it is common so in compositional data analysis [24]. While balances represent information within both factors, odds ratios capture relations between them. The preliminary expectations about independence between factors were confirmed as odds-ratio variables play marginal role for capturing multivariate variability. The concrete choice of SBP for columns of compositional tables shows its relevance here, the coordinate z_1^c that separates beverages from the other branches belongs to one of three main marker variables. In the right upper corner of the biplot a compact cluster of European countries is observed; they are predominantly characterized by low values of coordinate z_2^r, i.e., by dominance of Labour over Surplus across manufacturing categories. On contrary, high values of this variable occur for developing Asian countries (Azerbaijan, Indonesia, Sri Lanka). A certain level of grouping can be observed also for African and American countries, while Asian continent shows a higher diversity. Coordinates z_1^c and z_3^c (the latter one being strongly correlated with z_2^c) stand for beverage and food production specifics of countries. Particularly, it is interesting to see beverage dominance over aggregated food production across output sources for European and African countries that might correspond to specifics of their drinking culture. Note that, in contrast to analyzing standard multivariate (or even compositional) data, variables with different interpretations are considered together - those in sense of row/column factors (balances) as well as odds ratios that connect both of them. This is necessary to take into account, when any conclusion from the biplot is derived.

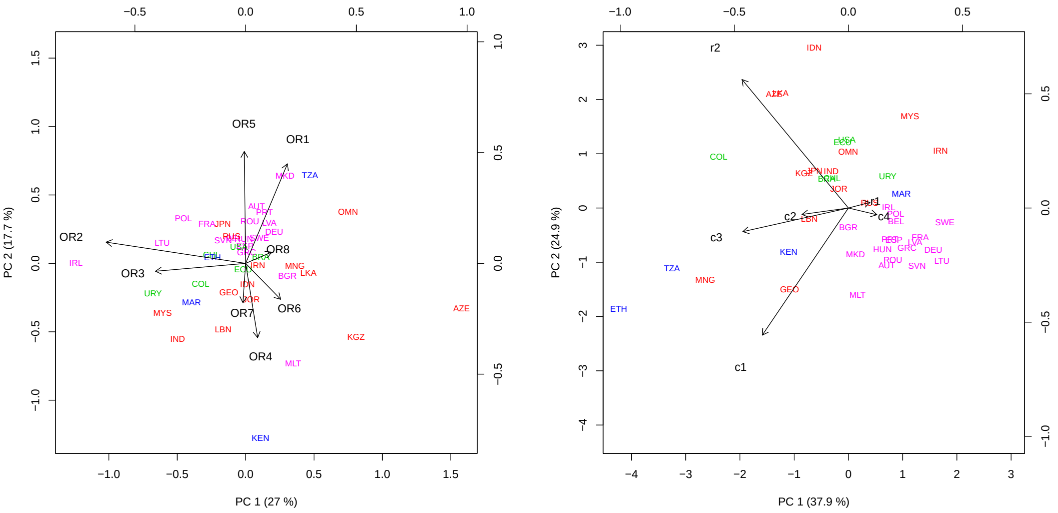


Figure 24: Biplots of odds ratio coordinates (left) and balances (right) with countries divided according to continent pertinence (Europe - purple, Africa - blue, America - green, Asia - red).

In order to see more detailed patterns, biplots were constructed also for the above two main groups of variables that form the coordinate system of compositional tables, balances and odds ratios, separately, see Figure 3. While, as expected, for balances the structure of loadings and scores remains almost unchanged (Figure 24, right) comparing to Figure 23, biplot for odds ratio coordinates (Figure 24, left) enables to reveal further interesting features about sources of relations between both factors. Accordingly, grouping of American countries abound the origin shows that relations between both factors are suppressed there. While from Figure 24 (right) it is clear that Labour part of output sources dominates Surplus over averaged manufacturing categories in European economics, provided by z_2^r, coordinate z_5^{OR} indicates that this ratio is higher for category 152 than 151. Similarly, from z_1^{OR} it is easy to see that dominance of beverages over other food branches is higher for Input than for Labour & Surplus output sources. Finally, coordinate z_2^{OR} provides a more detailed insight than z_5^{OR} to relation between value added sources: for countries like Ireland, Lithuania and Poland dominance of Labour over Surplus is much stronger for beverage category over the others. As both "marker variables" z_2^r and z_1^c form source of interpretation for z_2^{OR} this might be also the main reason for border position of these countries in Figure 23.

Conclusion

Compositional tables as observations carrying relative information about relationships between two factors represent a direct generalization of vector compositional data. Consequently, possibility of their appropriate orthonormal coordinate representation forms an important step for coordinate representation of multifactorial compositional data. The thesis presents a general coordinate system for compositional data, which respects their two-factorial character. The resulting coordinates form a natural generalization of the concept of balances as introduced in [6], that 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 vector compositional data, proper coordinate representation of compositional tables is necessary to enable statistical processing using standard multivariate statistical tools. The proposed coordinate system (in both general and pivot versions) contains both balances and coordinates with log odds ratio interpretation and forms the main contribution of the thesis. These coordinates respect the possibility of decomposition of a compositional table into its independent and interactive parts. Consequently, it allows to study tables from the decomposition also separately and analyze, e.g. possible interactions between both factors only from the interactive part of coordinates. Accordingly, the general orthonormal coordinate system respects the nature of row and column factors and thus allows for their better interpretability. On the other hand, the pivot coordinates as their special case seem to be easier to handle and provide an automated version of the coordinate representation. Construction of the coordinate systems was described in a detail and endowed with examples and graphical illustrations for better understanding. The theoretical part of the thesis is completed with the covariance structure of the proposed coordinates. Finally, the possibility of structural analysis of relationships between factors was discussed in the last section.

Beside the new coordinates a promising result comes from comparison of coordinates of 2×2 compositional table with parameters of log-linear model, since development of a compositional alternative to standard methods of analysis of independence between two variables (factors) represents one possible direction of our further research. The new coordinates thus seem to have great potential for compositional data analysis itself (statistical analysis of compositional tables, multifactorial compositional data), but open also its new challenging prospectives.

Bibliography

[1] Agresti, A. (2002) *Categorical data analysis (2 ed.).* New York: John Wiley & Sons.

[2] Aitchison, J. (1986) *The statistical analysis of compositional data.* London: Chapman & Hall.

[3] Eaton, M. L. (1983) *Multivariate statistics. A vector space approach.* New York: John Wiley & Sons.

[4] Egozcue, J.J. (2009) Reply to "On the Harker Variation Diagrams" by J.A. Cortés. *Mathematical Geosciences, **41**, 829–834.*

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

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

[7] Egozcue, J.J., Díaz-Barrero, J.L. and 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-Estadella and J. A. Martín-Fernández. University of Girona, Spain.

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

[9] EUROSTAT (2013) Body mass index (BMI) by sex, age and educational level - collection round 2008. (Available from http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=hlth_ehis_de1& lang=en. Accessed September 12, 2013.)

[10] EUROSTAT (2013) Population on 1 January: Structure indicators. (Available from http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=demo_pjanind& lang=en. Accessed September 12, 2013.)

[11] Fačevicová, K., Hron, K., Todorov, V., Guo, D. and Templ, M. (2014) Logratio approach to statistical analysis of 2×2 compositional tables. *Journal of Applied Statistics, **41**, 944–958.*

[12] Fačevicová K. and Hron, K. (2015) Covariance structure of compositional tables. *Austrian Journal of Statistics, **44**, 31–44.*

[13] Fačevicová, K., Hron, K., Todorov, V. and Templ, M. (2016) Compositional tables analysis in coordinates. *Scandinavian Journal of Statistics.*. DOI: 10:1111/sjos.12223.

[14] Fačevicová K., Hron, K., Todorov, V. and Templ, M. (2016) General approach to coordinate representation of compositional tables. In progress.

[15] Filzmoser, P., Hron, K. and Reimann, C. (2009) Univariate analysis of environmental (compositional) data: Problems and possibilities. *Science of the Total Environment, **407**, 6100–6108.*

[16] Filzmoser, P., Hron, K. and Reimann, C. (2012) Interpretation of multivariate outliers for compositional data. *Computers & Geosciences, **39**, 77–85.*

[17] Fišerová, E. and Hron, K. (2011) On interpretation of orthonormal coordinates for compositional data. *Mathematical Geology, **43**, 455–468.*

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

[19] Greenacre, M. (2007) *Correspondence analysis in practice*, 2nd edition. London: Chapman and Hall/CRC Press.

[20] Houston crime statistics (2015) (Available from http://www.houstontx.gov/police/cs/stats2.htm).

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

[22] Martín-Fernández, J.A., Hron, K., Templ, M., Filzmoser, P. and Palarea-Albaladejo, J. (2012) Model-based replacement of rounded zeros in compositional data: Classical and robust approaches. *Computational Statistics and Data Analysis, **56**, 9, 2688–2704.*

[23] Pawlowsky-Glahn, V. and Buccianti, A., eds. (2011) *Compositional data analysis: Theory and applications.* Chichester: John Wiley & Sons.

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

[25] UN (2002) International Standard Industrial Classification of All Economic Activities (ISIC) Rev. 3.1. (Available from http://unstats.un.org/unsd/cr/registry/regdnld.asp?Lg=1.)

PALACKÝ UNIVERSITY OLOMOUC
Faculty of Science
Department of Mathematical Analysis and Applications of Mathematics

DOCTORAL THESIS SUMMARY

Complex Structures of Compositional Data

Supervised by:
Doc. RNDr. Karel Hron, Ph.D.
Olomouc 2016

Written by:
Mgr. Kamila Fačevicová
P1104 Applied Mathematics

The doctoral thesis was carried out under the full-time postgradual programme Mathematics, field P1104 Applied Mathematics, in the Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University Olomouc.

Applicant: **Mgr. Kamila Fačevicová**
Dept. of Mathematical Analysis and Applications of Mathematics
Faculty of Science
Palacký University Olomouc

Supervisor: **doc. RNDr. Karel Hron, Ph.D.**
Dept. of Mathematical Analysis and Applications of Mathematics
Faculty of Science
Palacký University Olomouc

Reviewers: **Prof. Dr. Juan José Egozcue**
Dept. of Applied Mathematics III
Technical University of Catalonia

Prof. RNDr. Jaromír Antoch, CSc.
Dept. of Probability and Mathematical Statistics
Faculty of Mathematics and Physics
Charles University Prague

Doctoral thesis summary was sent to distribution on ....

Oral defence of doctoral thesis will be performed on .... at Department of Mathematical Analysis and Application of Mathematics in front of the committee for Ph.D. study programme Applied Mathematics, Faculty of Science, Palacký University Olomouc, room ......, 17. listopadu 12, Olomouc.

Full text of the doctoral thesis is available at Study Department of Faculty of Science, Palacký University Olomouc.

Contents

1 Abstract

Compositional tables can be considered as a continuous counterpart to the well-known contingency tables. Accordingly, their cells, containing in general positive real numbers rather than just counts, carry relative information about relationships between two factors. As a consequence, compositional tables can be considered as a generalization of (vector) compositional data. Due to relative character of these observations, compositions are popularly expressed in orthonormal coordinates using sequential binary partition prior to further processing using standard statistical tools. Even though the resulting coordinates (balances) are well interpretable in sense of logratio between two groups of parts, they do not respect the two-dimensional nature of compositional tables and the information about relationship between factors is thus not well captured. The main aim of the thesis is to present a general system of orthonormal coordinates with respect to the Aitchison geometry of compositional data, which enables to analyze interactions between factors in a compositional table. This is realized in sense of logarithms of odds ratios, which are popular also in context of contingency tables. Moreover, the pivot coordinate system is presented, which is useful particularly in case, when no a priori knowledge about row and column factors is available. For the sake of completeness, a part of thesis also concerns its covariance structure of the coordinates that enables to understand better their interpretation. All proposed coordinate systems are illustrated by examples and graphical representations.

Key words: analysis of independence, balances, compositional tables, orthonormal coordinates

2 Abstrakt v českém jazyce

Dizertační práce je zaměřena na analýzu kompozičních tabulek, které představují přímé zobecnění D–složkových (vektorových) kompozičních dat. Kompoziční tabulky mohou být navíc chápány jako spojitá alternativa kontingenčních tabulek, také totiž zachycují vztah mezi dvěma faktory, založený na informaci o poměrech mezi prvky tabulky. Kvůli této relativní povaze se kompoziční tabulky (stejně jako kompoziční data obecně) řídí tzv. Aitchisonovou geometrií. Aby bylo možné použít standardní analytické metody, je potřeba tento typ dat převést prostřednictvím ortonormálních souřadnic do prostoru se standardní euklidovskou metrikou. Vyjádření v ortonormálních souřadnicích je běžně prováděno prostřednictvím tzv. postupného binárního dělení, takto získané souřadnice (bilance) však nerespektují dvojrozměrnou povahu dat obsažených v kompozičních tabulkách. Kvůli zachování informace o vztahu mezi faktory je proto v práci navržena metoda, která bilance doplňuje o souřadnice, jejichž interpretace je úzce spjatá s poměry šancí mezi skupinami prvků. Právě konstrukci těchto souřadnic a jejich interpretaci je věnována hlavní část práce. Uveden je také speciální případ těchto souřadnic (pivotové souřadnice), jehož použití je vhodné v situaci, kdy nemáme žádnou znalost o povaze řádkového a sloupcového faktoru. Představení souřadnic jako takových je doplněno o jejich varianční strukturu, která umožní lepší pochopení jejich interpretace. Teoretické aspekty problematiky jsou demonstrované na několika příkladech a pomocí ilustrací.

Klíčová slova: analýza nezávislosti, bilance, kompoziční tabulky, ortonormální souřadnice

3 Introduction

In many practical situations, the object of statistical analysis is a table representing distribution of a variable of interest, according to two (row and column) factors. If relative contributions of cells on the overall distribution are of primary interest rather than concrete absolute values which they convey, it is referred to compositional tables [16, 17]. From this perspective, compositional tables form a generalization of vector compositional data, where only ratios between parts contain all relevant information [12, 20]. Compositional tables can be thus considered as a complex structure of compositional data, whose specific nature is captured by the Aitchison geometry with the structure of finite-dimensional Euclidean vector space. Contrary to contingency tables, representing result of a multinomial sampling with cell probabilities $p_{ij} > 0, \sum_i \sum_j p_{ij} = 1$, a compositional table itself represents one observation of distribution-valued variables with some continuous multivariate distribution (e.g. relative structure of population according to social and economic status). On the other hand, compositional and contingency tables are closely linked, since the probability table with entries p_{ij}, corresponding to given contingency table, forms just a proportional representation (and thus one particular case) of compositional table, see [17] for details. Statistical analysis of contingency tables is characterized by using Pearson χ^2 statistic or log-linear models for independence testing. As these methods strongly rely on the assumption of Euclidean geometry [17] (similarly as most of standard statistical methods [13]), they are not suitable for compositional tables that are driven by the Aitchison geometry. Moreover, similarly as for compositional data, it is also natural to consider a sample of compositional tables with a possibility of their processing using popular multivariate statistical methods (like principal component analysis, clustering, classification, etc.). This is a particular difference to the case of contingency tables, where such issues are usually not of primary interest.

Taking into account the relative character and the specific geometry of compositional tables (together with replacing the arithmetic marginals by the geometric ones), the analysis of independence between factors can be performed advantageously through a decomposition of the original table into its independent and interactive parts [16, 17]. In particular, the interaction table conveys the key information for understanding the sources of association between both factors. The key point in statistical analysis of compositional tables is then (as in the case of vector compositional data) to express them in orthonormal coordinates with respect to the Aitchison geometry, where rules of the standard Euclidean geometry apply. As there is no standard canonical basis with respect to the Aitchison geometry, the main aim of this thesis is to derive interpretable coordinate representation for compositional tables.

4 Recent state summary

Since $I \times J$ compositional tables represent a direct generalisation of vector compositional data, the concepts of the logratio approach to compositional data analysis can be easily adapted for compositional tables and used to derive the corresponding specific issues.

4.1 Compositional data

The vector compositional data need to be introduced first. This type of multivariate observations differs from the standard one by their relative nature, as ratios between parts are of the main interest rather than their absolute values. Compositional data frequently occur e.g. in geochemistry and the logratio methodology represents quite young and still growing statistical discipline (first analytical methods were proposed in [12]). The main principles of (vector) compositional data analysis are as follows.

Basic definitions

A (random) D-part composition is defined as a row vector

$$\mathbf{x} = (x_1, x_2, \dots, x_D) \quad , \tag{1}$$

where all components (parts) describe quantitatively their relative contributions to the whole [12, 20]. Thus absolute values of parts are not of the main interest, since all the relevant information in the composition is contained in the ratios between its parts. Consequently, the composition could be rescaled (closed) to a prescribed constant sum representation $\kappa > 0$ (i.e. to 1 in case of proportions and 100 for percentages) without any loss of information; formally, we refer to a closure operation and denote

$$\mathcal{C}(\mathbf{x}) = \left(\frac{\kappa \cdot x_1}{\sum_{i=1}^{D} x_i}, \frac{\kappa \cdot x_2}{\sum_{i=1}^{D} x_i}, \dots, \frac{\kappa \cdot x_D}{\sum_{i=1}^{D} x_i} \right) \quad . \tag{2}$$

This closed representation is useful, e.g., for a first brief comparison of two compositional vectors. The sample space of representations of D-part compositional data with an arbitrary, but fixed κ is a subset of $\mathbf{R}^D$, called D-part simplex,

$$\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, x_2, \dots, x_D) |\ x_i > 0,\ i = 1, 2, \dots, D;\ \sum_{i=1}^{D} x_i = \kappa \right\} \quad . \tag{3}$$

The constant sum constraint reduces the dimension of $\mathcal{S}^D$ to $D - 1$, i.e., one less than actual number of parts of the composition.

The assumption that only ratios between components carry relevant information about the composition leads to the following principles of compositional data analysis [20]. The first of them is the *scale invariance*, which means that the results of the analysis should not depend on the particular sum κ of compositional parts. Thus application of closure operation $\mathcal{C}(\mathbf{x})$ should not alter results of the analysis. Scale invariance is also related to the property of *relative scale* of compositions, since ratios should express the differences between observations rather than Euclidean distances based on absolute values of components. Next principle is called *subcompositional coherence.* As in standard statistics the results obtained from a composition with D parts should not be in contradiction with results that are obtained from a subcomposition containing d parts, $d < D$ and subcompositions should behave like orthogonal projections in real space. 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 noninformative part is removed, results should not change. The final basic principle of compositional analysis is *permutation invariance*, output of the analysis cannot depend on the order of parts in the composition.

Due to relative nature of compositional data and the above principles, the standard Euclidean geometry should be replaced by the Aitchison geometry, endowed with the Euclidean vector space structure. Accordingly, operations of perturbation and power transformation (powering) for D-part compositional vectors $\mathbf{x}$ and $\mathbf{y}$ and a real constant α are defined as

$$\mathbf{x} \oplus \mathbf{y} = (x_1 y_1, \ldots, x_D y_D) \quad \text{and} \quad \alpha \odot \mathbf{x} = (x_1^{\alpha}, \ldots, x_D^{\alpha}) \quad , \tag{4}$$

respectively. Consequently, $\mathbf{n} = \mathcal{C}\,(1, \ldots, 1)$ represents the neutral element in the $(D-1)$-dimensional vector space $(\mathcal{S}^D, \oplus, \odot)$. To complete the Euclidean vector space structure, the Aitchison inner product of two compositional vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \frac{1}{2D} \sum_{i,j} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j} \tag{5}$$

and the Aitchison norm and distance as

$$\|\mathbf{x}\|_A = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_A} \quad \text{and} \quad d_A(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_A \quad , \tag{6}$$

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

Coordinate representation

Due to specific nature of compositional data, represented by the above principles, standard statistical methods are not suitable for their analysis. Instead of developing their counterparts within the Aitchison geometry, it seems much more intuitive to express compositions isometrically in real coordinates with respect to the Aitchison geometry and then proceed with the usual statistical processing there [20]. The most popular coordinate system is represented by isometric logratio (ilr) coordinates [14, 15], i.e., coordinates with respect to an orthonormal basis on the simplex. According to basic algebraic-geometrical rules and dimensionality of the Aitchison geometry, the real vector $\mathbf{z} \in \mathbf{R}^{D-1}$ of ilr coordinates is defined as

$$\mathbf{z} = \mathrm{ilr}(\mathbf{x}) = \left(\left\langle \mathbf{x}, \mathbf{e}^1 \right\rangle_A, \left\langle \mathbf{x}, \mathbf{e}^2 \right\rangle_A, \ldots, \left\langle \mathbf{x}, \mathbf{e}^{D-1} \right\rangle_A\right) = (z_1, z_2, \ldots, z_{D-1}) \quad , \tag{7}$$

where $\mathbf{e}^i = \mathcal{C}\left(e_1^i, e_2^i, \ldots, e_D^i\right), i = 1, 2, \ldots, D-1$ form an orthonormal basis on the simplex. Due to isometric isomorphism of ilr coordinates it immediately follows

$$\mathrm{ilr}\left((\alpha \odot \mathbf{x}) \oplus (\beta \odot \mathbf{y})\right) = \alpha \cdot \mathrm{ilr}(\mathbf{x}) + \beta \cdot \mathrm{ilr}(\mathbf{y}), \qquad \langle \mathbf{x}, \mathbf{y} \rangle_A = \langle \mathrm{ilr}(\mathbf{x}), \mathrm{ilr}(\mathbf{y}) \rangle \quad , \tag{8}$$

$$\|\mathbf{x}\|_A = \|\mathrm{ilr}(\mathbf{x})\| \qquad \text{and} \qquad \mathrm{d}_A(\mathbf{x}, \mathbf{y}) = \mathrm{d}(\mathrm{ilr}(\mathbf{x}), \mathrm{ilr}(\mathbf{y})) \quad . \tag{9}$$

It could be also shown that different ilr coordinate systems are linked through an orthogonal transformation [14].

Clearly, it is not possible to assign an orthonormal coordinate to each of compositional parts simultaneously. Therefore, interpretable orthonormal coordinates are of primary interest. Since coordinates $\mathbf{z}$ correspond to a particular choice of basis vectors (compositions) $\mathbf{e}^i, i = 1, \ldots, D-1$, they can be chosen according to aim of the analysis and possible a priori knowledge about compositional parts. These coordinates are usually reached by sequential binary partition (SBP) procedure [15], based on stepwise division of parts into non-overlapping groups. Accordingly, in the first step of SBP, the whole composition is divided into two subcompositions. For the next step only one of subcompositions from the previous step is taken and further divided into two groups. This process continues until all groups of parts consist of only one single component. The SBP is done in $D-1$ steps; in each step one coordinate

$$z_i = \sqrt{\frac{uv}{u+v}} \ln \frac{(x_{j_1} x_{j_2} \cdots x_{j_u})^{1/u}}{(x_{k_1} x_{k_2} \cdots x_{k_v})^{1/v}}, \; i = 1, \ldots, D-1 \tag{10}$$

is obtained. 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 parts assigned to the first group are marked by $+$, parts in the second group by $-$ and parts not included in any of both groups in the i-th step of the partition by 0, SBP can be represented also graphically. Table 1 results from one possible SBP for five-part compositional data.

Orthonormal coordinates resulting from SBP (10) 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

Table 1: Example of sequential binary partition for five-part compositional data.

i	x_1	x_2	x_3	x_4	x_5	u	v	z_i
1	$+$	$+$	$-$	$-$	$-$	2	3	$\sqrt{\frac{6}{5}}\ln\frac{\sqrt{x_1x_2}}{\sqrt[3]{x_3x_4x_5}}$
2	$+$	$-$	0	0	0	1	1	$\frac{1}{\sqrt{2}}\ln\frac{x_1}{x_2}$
3	0	0	$+$	$-$	$-$	1	2	$\sqrt{\frac{2}{3}}\ln\frac{x_3}{\sqrt{x_4x_5}}$
4	0	0	0	$+$	$-$	1	1	$\frac{1}{\sqrt{2}}\ln\frac{x_1}{x_2}$

the aim to capture the most relevant information contained in ratios between compositional parts and their groups. For example, geochemical data are formed by major and minor elements, further divided according to concrete composition of the analyzed rock/soil. Because of this flexibility, balances form the most popular class of orthonormal coordinates that was recently successfully applied in a number of real-world studies [19]. Unfortunately, balances do not respect the two-dimensional nature of compositional tables and are not appropriate for their analysis.

4.2 Compositional tables

Even though the theory of compositional data analysis is already well developed, it is primarily designed for vector compositional data, which carry information about relative structure according to only one factor. In cases, when compositional data carry information about distribution according to two factors (e.g. population structure according to age and BMI index), it seems to be appropriate to work with two dimensional data, which besides the relative structure contain inherently also information about relationship between these factors.

An $I \times J$ table

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

whose cells $x_{ij} > 0$, for $i = 1, 2, \dots, I$ and $j = 1, 2, \dots J$ convey relative contributions on a whole (probability, overall output, etc.) can be considered as a natural extension of vector compositional data and is called compositional table. From this point on, $\mathbf{x}$ will denote a $I \times J$ compositional table instead of compositional vector, unless otherwise stated. As it was mentioned above, this type of observations basically conveys relative information on relationship between two factors with I and J values, respectively. But also the other way around, by vectorization of compositional tables vector compositional data would be obtained. Therefore, any reasonable analysis of compositional tables should follow the same assumptions as analysis of compositional vectors, which were introduced in Section 4.1, just with specific (two-factor) interpretation of their parts; here a subcomposition of compositional table arises by omitting the whole row(s) and/or column(s) and it is called subtable or partial table. Note here, that on the contrary to contingency tables, containing n independent realisations of random variable from multinomial distribution, compositional table is considered to be one realisation from a multivariate continuous distribution. On the other hand, there is quite close connection between both types of tables, since probability table, that corresponds to the contingency table, can be considered as one particular representation of compositional table. And finally, even the contingency table itself can be considered as a compositional table, if the total number of counts is high enough that its role as a source of uncertainty for estimation of the underlying probabilities is negligible.

Basic definitions

Since compositional tables (11) represent a direct extension of vector compositional data (1), all operations defined in Section 4.1 can be easily accommodated for this case. Proportional representation of a compositional table can be reached by application of closure operation with $\kappa = 1$,

$$\mathcal{C}(\mathbf{x}) = \begin{pmatrix} \frac{\kappa x_{11}}{\sum_{i,j} x_{ij}} & \cdots & \frac{\kappa x_{1J}}{\sum_{i,j} x_{ij}} \\ \vdots & \ddots & \vdots \\ \frac{\kappa x_{I1}}{\sum_{i,j} x_{ij}} & \cdots & \frac{\kappa x_{IJ}}{\sum_{i,j} x_{ij}} \end{pmatrix} , \qquad (12)$$

and by varying $\kappa > 0$, any other constant sum representation can be obtained. The sample space of compositional tables is again $(IJ-1)$-dimensional simplex

$$\mathcal{S}^{IJ} = \left\{ \mathbf{x} = (x_1, x_2, \dots, x_{IJ}) |\ x_i > 0,\ i = 1, 2, \dots, IJ;\ \sum_{i=1}^{IJ} x_i = \kappa \right\} , \qquad (13)$$

since each IJ-part compositional vector can be re-ordered into the form of table with I rows and J columns. On the other hand, note that the table form is appropriate only for such data, which carry information about distribution of some total with respect to two factors. Also basic operations of the Aitchison geometry should 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

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C} \begin{pmatrix} x_{11}y_{11} & \cdots & x_{1J}y_{1J} \\ \vdots & \ddots & \vdots \\ x_{I1}y_{I1} & \cdots & x_{IJ}y_{IJ} \end{pmatrix} ; \qquad (14)$$

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

$$\alpha \odot \mathbf{x} = \mathcal{C}\begin{pmatrix} x_{11}^{\alpha} & \cdots & x_{1J}^{\alpha} \\ \vdots & \ddots & \vdots \\ x_{I1}^{\alpha} & \cdots & x_{IJ}^{\alpha} \end{pmatrix} \tag{15}$$

is obtained. The Aitchison inner product modifies to

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \frac{1}{2IJ} \sum_{i,j} \sum_{k,l} \ln \frac{x_{ij}}{x_{kl}} \ln \frac{y_{ij}}{y_{kl}} \tag{16}$$

and the Aitchison norm and distance should be restated as follows,

$$\|\mathbf{x}\|_A = \sqrt{\frac{1}{2IJ} \sum_{i,j} \sum_{k,l} \left(\ln \frac{x_{ij}}{x_{kl}} \right)^2} \tag{17}$$

and

$$d_A(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{2IJ} \sum_{i,j} \sum_{k,l} \left(\ln \frac{x_{ij} y_{kl}}{x_{kl} y_{ij}} \right)^2} \quad . \tag{18}$$

Decomposition of compositional tables

The construction of coordinates of compositional tables is based on projections of the table onto subspaces with specific interpretation [16].

At first, projections of a compositional table $\mathbf{x}$ onto row subspaces $\mathcal{S}^{IJ}(\mathrm{row}_i)$, for $i = 1, \dots, I$, each with dimension $J - 1$, are considered. According to [16], this projection denoted by $\mathrm{row}_i(\mathbf{x})$ equals

$$\mathrm{row}_i(\mathbf{x}) = \mathcal{C}\begin{pmatrix} g(\mathrm{row}_i[\mathbf{x}]) & \cdots & g(\mathrm{row}_i[\mathbf{x}]) \\ \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{iJ} \\ \cdots & \cdots & \cdots \\ g(\mathrm{row}_i[\mathbf{x}]) & \cdots & g(\mathrm{row}_i[\mathbf{x}]) \end{pmatrix} \quad , \tag{19}$$

where $g(\mathrm{row}_i[\mathbf{x}])$ denotes the geometric mean of elements in the i-th row of $\mathbf{x}$. The projection onto the subspace, formed by the i-th row of the compositional table $\mathbf{x}$, $\mathrm{row}_i[\mathbf{x}] = \mathcal{C}(x_{i1}, \dots, x_{iJ}) \in \mathcal{S}^J, i = 1, \dots, I$, is thus still a $I \times J$ compositional table $\mathrm{row}_i(\mathbf{x})$ whose entries consist of the i-th row itself and the rest elements are equal to geometric mean of $\mathrm{row}_i[\mathbf{x}]$.

Analogously, also projections of the compositional table $\mathbf{x}$ onto its columns, $\mathrm{col}_j[\mathbf{x}] = \mathcal{C}(x_{1j}, \dots, x_{Ij}) \in \mathcal{S}^I, j = 1, \dots, J$, forming subspaces $\mathcal{S}^{IJ}(\mathrm{col}_j)$ with dimension $I - 1$, can be constructed. Similarly to the case of projections onto rows, the resulting projected compositional tables $\mathrm{col}_j(\mathbf{x})$ are given by the j-th column of $\mathbf{x}$ and its geometric mean in the other parts of the table.

Projection onto the subspace of the i-th row results in a compositional table $\mathrm{row}_i(\mathbf{x})$ that explains the relative information (ratios) exclusively for this row. In order to complete the information about the original compositional table $\mathbf{x}$, it is necessary to introduce a projection that explains the remaining ratios between parts in different rows [18]. In other words, a projection onto the subspace of dimension $I-1$ that forms the orthogonal complement to row subspaces $\mathcal{S}^{IJ}(\mathrm{row}_i)$, $i=1,\dots,I$, needs to be constructed. This subspace will be denoted as $\mathcal{S}^{IJ}(\mathrm{row}^{\perp})$ and projection onto this subspace is a compositional table

$$\mathrm{row}^{\perp}(\mathbf{x}) = \mathcal{C}\begin{pmatrix} g(\mathrm{row}_1[\mathbf{x}]) & \dots & g(\mathrm{row}_1[\mathbf{x}]) \\ g(\mathrm{row}_2[\mathbf{x}]) & \dots & g(\mathrm{row}_2[\mathbf{x}]) \\ \dots & \dots & \dots \\ g(\mathrm{row}_I[\mathbf{x}]) & \dots & g(\mathrm{row}_I[\mathbf{x}]) \end{pmatrix} \quad , \tag{20}$$

formed by row geometric means of the original table. Similarly, projection of $\mathbf{x}$ onto subspace orthogonal to column subspaces, $\mathcal{S}^{IJ}(\mathrm{col}^{\perp})$, of dimension $J-1$ that carries information about ratios between different columns of the original compositional table, results in

$$\mathrm{col}^{\perp}(\mathbf{x}) = \mathcal{C}\begin{pmatrix} g(\mathrm{col}_1[\mathbf{x}]) & \dots & g(\mathrm{col}_J[\mathbf{x}]) \\ g(\mathrm{col}_1[\mathbf{x}]) & \dots & g(\mathrm{col}_J[\mathbf{x}]) \\ \dots & \dots & \dots \\ g(\mathrm{col}_1[\mathbf{x}]) & \dots & g(\mathrm{col}_J[\mathbf{x}]) \end{pmatrix} \quad . \tag{21}$$

From their construction, projections $\mathrm{row}^{\perp}(\mathbf{x})$ and $\mathrm{col}^{\perp}(\mathbf{x})$ are orthogonal to all row or column projections, respectively, and even to each other (see [16] for proof). This fact is crucial for compositional tables analysis as it will be shown later.

As mentioned above, projections $\mathrm{row}^{\perp}(\mathbf{x})$ and $\mathrm{col}^{\perp}(\mathbf{x})$ carry information exclusively about ratios between parts from different rows and columns, respectively. This information is sufficient for the reconstruction of the compositional table, when row and column factors are independent (motivated by the probabilistic sense of the formulation). This corresponds to the case when the original table can be expressed as a product of row and column (geometric) marginals of $\mathbf{x}$ [16, 17], similarly as for contingency tables [11], where arithmetic marginals are considered instead. The resulting $I \times J$ compositional table $\mathbf{x}_{ind} = \mathrm{row}^{\perp}(\mathbf{x}) \oplus \mathrm{col}^{\perp}(\mathbf{x})$, obtained as a perturbation of these two projections, is called *independence table* with entries

$$x_{ij}^{ind} = \left(\prod_{k=1}^{I}\prod_{l=1}^{J} x_{kj}x_{il}\right)^{\frac{1}{IJ}} \quad , \tag{22}$$

x_{ij} denote parts of the original compositional table $\mathbf{x}$. Since the dimensions of subspaces $\mathcal{S}^{IJ}(\mathrm{row}^{\perp})$ and $\mathcal{S}^{IJ}(\mathrm{col}^{\perp})$ are $I-1$ and $J-1$, respectively, dimension of the subspace of independence tables $\mathcal{S}^{IJ}_{\text{ind}}$ equals $I + J - 2$. The remaining information about the original table, i.e. about the relations between row and column factors, is contained in the *interaction table* $\mathbf{x}_{int}$, which is orthogonal to $\mathbf{x}_{ind}$ and results from the decomposition

$$\mathbf{x} = \mathbf{x}_{ind} \oplus \mathbf{x}_{int} \quad . \tag{23}$$

The interaction table can be obtained from (23) as $\mathbf{x}_{int} = \mathbf{x} \ominus \mathbf{x}_{ind}$. It also forms an $I \times J$ compositional table and its parts can be computed from the original table $\mathbf{x}$ by

$$x^{int}_{ij} = \left(\prod_{k=1}^{I} \prod_{l=1}^{J} \frac{x_{ij}}{x_{kj} x_{il}} \right)^{\frac{1}{IJ}} \quad . \tag{24}$$

From Equation (23) and orthogonality between $\mathbf{x}_{ind}$ and $\mathbf{x}_{int}$ it follows that the dimension of the subspace of interaction tables, $\mathcal{S}^{IJ}_{\text{int}}$, equals $I \cdot J - 1 - (I + J - 2) = (I - 1)(J - 1)$.

5 Thesis objectives

The main aim of the thesis is to construct interpretable coordinates of compositional tables, which will also respect their two-dimensional nature. Consequently, such a coordinate system allows to describe relations within the table using standard statistical methods. Moreover, statistical processing of a sample of tables is also possible. In its second part, the thesis also focuses on covariance structure of proposed coordinates (not shown here), which supports their interpretability.

6 Theoretical framework

The principal aim of the thesis is to propose a new coordinate system for compositional tables, which respects the possibility of decomposition of a compositional table $\mathbf{x}$ as described in Section 5. The main idea of this system is to complete balances between whole rows or columns by those dealing with odds ratios between four groups of parts [11], which represent a natural extension of balances for the case of compositional tables. This has quite an intuitive motivation. Balances can be used to capture (log-)ratios within row and column factors, respectively, while odds ratios naturally link relative information between both factors.

6.1 General coordinates

For construction of the general coordinates of $I \times J$ compositional table, consider first SBP of the whole rows (columns) of compositional table $\mathbf{x}$, denoted in the following by SBPr (SBPc). This partition is constructed with respect to nature of levels of row (column) factor and similarly as for the usual SBP, in each of $I-1$ $(J-1)$ steps, levels with some common property are separated from the others. Thus the first $I+J-2$ coordinates $\mathbf{z}^r$ and $\mathbf{z}^c$ of $I \times J$ compositional table $\mathbf{x}$ result in

$$z_i^r = \sqrt{\frac{stJ}{s+t}} \ln \frac{\left[g(\mathbf{x}_{j_1 \cdot}) \cdots g(\mathbf{x}_{j_s \cdot})\right]^{1/s}}{\left[g(\mathbf{x}_{k_1 \cdot}) \cdots g(\mathbf{x}_{k_t \cdot})\right]^{1/t}}, \quad \text{for} \quad i = 1, 2, \dots, I-1 \tag{25}$$

and

$$z_j^c = \sqrt{\frac{uvI}{u+v}} \ln \frac{\left[g(\mathbf{x}_{\cdot l_1}) \cdots g(\mathbf{x}_{\cdot l_u})\right]^{1/u}}{\left[g(\mathbf{x}_{\cdot m_1}) \cdots g(\mathbf{x}_{\cdot m_v})\right]^{1/v}}, \quad \text{for} \quad j = 1, 2, \dots, J-1 \quad , \tag{26}$$

where s, t (u, v) are numbers of rows (columns) separated in the i-th (j-th) step of SBP, indices $(j_1\cdot, \dots, j_s\cdot)$ and $(k_1\cdot, \dots, k_t\cdot)$ or $(\cdot l_1, \dots, \cdot l_u)$ and $(\cdot m_1, \dots, \cdot m_v)$ denote rows/columns involved in this step and $g(.)$ stands for the geometric mean. Steps of SBPr are denoted by Roman numerals, while those of SBPc are denoted by Arabic numerals. As an example consider a 3×5 compositional table. The corresponding six coordinates could follow SBPs from Table 2, represented also graphically in Figure 1,

$$z_1^r = \sqrt{\frac{10}{3}} \ln \frac{g(x_{1.})}{\left(g(x_{2.})g(x_{3.})\right)^{1/2}} \quad , \tag{27}$$

$$z_2^r = \sqrt{\frac{5}{2}} \ln \frac{g(x_{2.})}{g(x_{3.})} \quad , \tag{28}$$

$$z_1^c = \sqrt{\frac{18}{5}} \ln \frac{\left(g(x_{.1})g(x_{.2})\right)^{1/2}}{\left(g(x_{.3})g(x_{.4})g(x_{.5})\right)^{1/3}} \quad , \tag{29}$$

$$z_2^c = \sqrt{\frac{3}{2}} \ln \frac{g(x_{.1})}{g(x_{.2})} \quad , \tag{30}$$

$$z_3^c = \sqrt{\frac{6}{3}} \ln \frac{g(x_{.3})}{\left(g(x_{.4})g(x_{.5})\right)^{1/2}} \quad , \tag{31}$$

$$z_4^c = \sqrt{\frac{3}{2}} \ln \frac{g(x_{.4})}{g(x_{.5})} \quad . \tag{32}$$

The remaining coordinates should be orthogonal to these first $I+J-2$ ones and for their construction some generalization of SBP needs to be introduced.

Table 2: Example of sequential binary partition applied to whole rows (SBPr, left table) and whole columns (SBPc, right table) of $I \times J$ 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

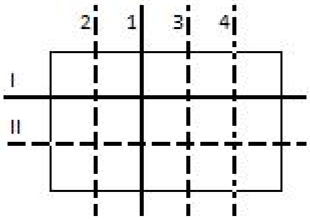


Figure 1: Graphical representation of sequential binary partitions SBPr and SBPc, applied to 3×5 compositional table.

This generalization is based on separation of parts of the compositional table into four groups (blocks) in a systematic manner and computation of coordinates in form of logarithm of odds ratio between these four groups (marked as A (upper left block), B (upper right block), C (lower left block) and D (lower right block)) using the following formula,

$$z^{OR} = \sqrt{\frac{a \cdot d}{a+b+c+d}} \ln \frac{(x_{i_1} \cdots x_{i_a})^{1/a} (x_{l_1} \cdots x_{l_d})^{1/d}}{(x_{j_1} \cdots x_{j_b})^{1/b} (x_{k_1} \cdots x_{k_c})^{1/c}}, \tag{33}$$

where a, b, c, d are numbers of parts in each of groups A, B, C and D, respectively and $i_., j_., k_., l_.$ are possitions of these parts. In the following steps this separation proceeds within smaller subtables according to row/column SBPs.

The separation into subgroups (A–D) and construction of the partial tables should respect the row and column grouping defined in SBPr and SBPc. Thus the first four groups are formed by steps I of SBPr and 1 of SBPc and determine the first coordinate. If the compositional table has more than four parts, the partition should continues with the next step. Firstly, a proper subtable should be identified, when the only possible partial tables are formed by pairs of groups (A,B), (C,D), (A,C) and (B,D), which should be successively analysed. If (A,B) has more than one row, the next coordinate is related to parts of this subtable, where the four groups are again determined by steps of the SBPr and SBPc of the lowest possible order. The next possible subtable is firstly searched within the current partial table, but if this one is formed by only 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 subtables with only one row or column, or subtables, which were already analysed in some of previous steps of partition, are skipped. The process continues, until all possible subtables formed by pairs of groups (A,B), (C,D), (A,C) and (B,D) of each proper partial tables are analysed. It results in $(I-1)(J-1)$ coordinates, each with interpretation in terms of log-odds ratios among groups of entries within the respective partial table. For the SBPs from Table 2 this process results in eight new coordinates

$$\begin{aligned}
z_1^{OR} &= \frac{2\sqrt{5}}{5}\ln\frac{\left(x_{11}x_{12}\right)^{1/2}\left(x_{23}x_{24}x_{25}x_{33}x_{34}x_{35}\right)^{1/6}}{\left(x_{13}x_{14}x_{15}\right)^{1/3}\left(x_{21}x_{22}x_{31}x_{32}\right)^{1/4}} \quad , && (34)\\
z_2^{OR} &= \sqrt{\frac{3}{5}}\ln\frac{\left(x_{21}x_{22}\right)^{1/2}\left(x_{33}x_{34}x_{35}\right)^{1/3}}{\left(x_{23}x_{24}x_{25}\right)^{1/3}\left(x_{31}x_{32}\right)^{1/2}} \quad , && (35)\\
z_3^{OR} &= \frac{1}{2}\ln\frac{x_{21}x_{32}}{x_{22}x_{31}} \quad , && (36)\\
z_4^{OR} &= \frac{\sqrt{3}}{3}\ln\frac{x_{23}\left(x_{34}x_{35}\right)^{1/2}}{\left(x_{24}x_{25}\right)^{1/2}x_{33}} \quad , && (37)\\
z_5^{OR} &= \frac{1}{2}\ln\frac{x_{24}x_{35}}{x_{25}x_{34}} \quad , && (38)\\
z_6^{OR} &= \frac{\sqrt{3}}{3}\ln\frac{x_{11}\left(x_{22}x_{32}\right)^{1/2}}{x_{12}\left(x_{21}x_{31}\right)^{1/2}} \quad , && (39)\\
z_7^{OR} &= \frac{2}{3}\ln\frac{x_{13}\left(x_{24}x_{25}x_{34}x_{35}\right)^{1/4}}{\left(x_{14}x_{15}\right)^{1/2}\left(x_{23}x_{33}\right)^{1/2}} \quad , && (40)\\
z_8^{OR} &= \frac{\sqrt{3}}{3}\ln\frac{x_{14}\left(x_{25}x_{35}\right)^{1/2}}{x_{15}\left(x_{24}x_{34}\right)^{1/2}} \quad , && (41)
\end{aligned}$$

whose construction is described in detail in the thesis.

Alternatively, each coordinate could be also interpreted as a sum of log-odds ratios among four parts. There are $\binom{I}{2}\binom{J}{2}$ of them together in the whole table, each contained in one of these new coordinates. In [11] it is stated that the whole information about relations in $I \times J$ (not necessarily compositional) table is contained in $(I-1)(J-1)$ simple odds ratios of type

$$OR_{ij}^1 = \frac{x_{ij}x_{i+1,j+1}}{x_{i,j+1}x_{i+1,j}}, \quad i=1,\dots,I-1 \quad \text{and} \quad j=1,\dots,J-1 \quad , \qquad (42)$$

among neighbouring parts or of type

$$OR_{ij}^2 = \frac{x_{ij}x_{IJ}}{x_{iJ}x_{Ij}}, \quad i=1,\dots,I-1 \quad \text{and} \quad j=1,\dots,J-1 \quad , \qquad (43)$$

with a reference part x_{IJ}. These basic systems of odds ratios could not be used to construct orthonormal coordinates with respect to the Aitchison geometry. In our case they are replaced by the system of $(I-1)(J-1)$ coordinates $\mathbf{z}^{OR}$, whose idea of aggregating the information into odds ratio among four groups of parts (not just four parts) seems to be similar to the concept of cumulative odds ratio as proposed in [11], page 276.

Beside the advantageous interpretation, there is another useful feature of this coordinate system. When the coordinate representation $\mathbf{z}^r = (z_1^r, \dots, z_{I-1}^r)$, $\mathbf{z}^c = (z_1^c, \dots, z_{J-1}^c)$, $\mathbf{z}^{OR} = (z_1^{OR}, \dots, z_{(I-1)(J-1)}^{OR})$ is applied to the independence table $\mathbf{x}_{ind}$, the only nonzero coordinates are z_i^r, z_j^c for $i = 1, \dots, I-1$, $j = 1, \dots, J-1$, and their values are the same as for the original table $\mathbf{x}$. Moreover, the number of these nonzero coordinates equals to dimension of subspace of independence tables (see e.g. [1] for details). Analogous feature holds also for the interaction table and coordinates $\mathbf{z}^{OR}$. Accordingly, the vector of coordinates $(\mathbf{z}^r, \mathbf{z}^c, \mathbf{0}_{(I-1)(J-1)})$ of the independence table can be denoted as $\mathbf{z}_{ind}$ and coordinates $(\mathbf{0}_{I+J-2}, \mathbf{z}^{OR})$ of the interaction table as $\mathbf{z}_{int}$. Finally, the vector of coordinates of the original compositional table $\mathbf{x}$ can be written as $\mathbf{z} = \mathrm{ilr}(\mathbf{x}_{ind}) + \mathrm{ilr}(\mathbf{x}_{int}) = \mathbf{z}_{ind} + \mathbf{z}_{int} = (\mathbf{z}^r, \mathbf{z}^c, \mathbf{z}^{OR})$. This feature is utilized especially in case of analysis of relationships between two factors, which will be described in Section 7.

6.2 Pivot coordinates

In the case, when there are no clues, how to form groups within the row and column factor, a special case of the general coordinates can be considered, which was introduced in [4]. This coordinate system can be applied to each compositional table almost automatically in the situation, when interpretation of the coordinates is not the main goal of the analysis (like outlier detection or classification of observations). On the other hand, such coordinates obviously still follow the decomposition (23).

The main idea by construction of these coordinates is that in each partial table the group D if formed by only single part (pivot), which is denoted as x_{rs} and which gave the name to these coordinates. Consequently, the first $I+J-2$ coordinates are

$$z_i^r = \sqrt{\frac{(I-i)J}{I-i+1}} \ln \frac{g(\mathbf{x}_{I-i+1.})}{[g(\mathbf{x}_{1.}) \cdots g(\mathbf{x}_{I-i.})]^{1/(I-i)}}, \quad \text{for} \quad i = 1, \dots, I-1 \qquad (44)$$

(for rows), and

$$z_j^c = \sqrt{\frac{I(J-j)}{J-j+1}} \ln \frac{g(\mathbf{x}_{J-j+1.})}{[g(\mathbf{x}_{.1}) \cdots g(\mathbf{x}_{.J-j})]^{1/(J-j)}}, \quad \text{for} \quad j = 1, \dots, J-1 \qquad (45)$$

(for columns), respectively. These orthonormal coordinates form again nonzero coordinate representation for the independence table and their number reflects the dimension of $\mathcal{S}^{IJ}_{\text{ind}}$. Because of mutual orthogonality of the subspaces corresponding to tables $\text{row}^{\perp}(\mathbf{x})$, $\text{col}^{\perp}(\mathbf{x})$ and $\mathbf{x}_{int}$, and decomposition (23), the remaining $(I-1)(J-1)$ coordinates of $\mathbf{x}_{ind}$ are equal to zero. Conversely, coordinate representation of the interaction table results in zero coordinates of the corresponding independence table.

In contrast to the general method, it is easier to start construction of partial tables from the smallest one in the upper left corner of the table $\mathbf{x}$. Each consequent table is then formed by the current one expanded by one row, or column. The first two steps of this stepwise procedure are as follows. The method firstly assigns a basis compositional vector to the table given only by parts x_{11}, x_{12}, x_{21} and x_{22}. The first coordinate than compares parts on the main diagonal with those on the minor one. In the next step the third column is added to the previous partial table and the basis vector $\mathbf{e}^{23}$ deals with the new partial table with $r=2$ rows and $s=3$ columns and parts $x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{23}$. The corresponding basis element compares again parts on the main diagonal of a virtual 2×2 table with parts on the minor diagonal, when these diagonals are formed by geometric mean of x_{11} and x_{12} (that thus merges information on the employed components together) and part x_{23}, and by geometric mean of x_{21} and x_{22}, and part x_{13}, respectively. In general, the coordinate z_{rs} compares parts on the main diagonal (formed by geometric mean of all parts at rows of order smaller than r and column of order smaller than s and by pivot part x_{rs}) and parts on the minor diagonal (formed by geometric mean of the first $s-1$ parts of the r-th row and by geometric mean of the first $r-1$ parts of the s-th column). This procedure continues until $r=I$ and $s=J$, accordingly a system of $(I-1)(J-1)$ nonzero coordinates of the interaction table (out of $IJ-1$) is obtained,

$$z_{rs} = \frac{1}{\sqrt{r\cdot s\cdot(r-1)\cdot(s-1)}} \ln \prod_{i=1}^{r-1}\prod_{j=1}^{s-1} \frac{x_{ij}x_{rs}}{x_{is}x_{rj}} \quad , \tag{46}$$

for $r=2,3,\ldots,I$ and $s=2,3,\ldots,J$. Note that, even though x_{ij}'s in both formulas stand for parts of the original table $\mathbf{x}$, the result would not change if they are replaced by parts of the interaction table $\mathbf{x}_{int}$.

Another useful property of these coordinates is that they contain also nonzero coordinates of the interaction tables of all tables with sizes smaller than the considered $I\times J$ table. For example, the set of four nonzero coordinates of 3×3 interaction table contains two nonzero coordinates of the 2×3 table as well as of the 3×2 table, and in turn both (as well as 3×3 table) contain the only nonzero ilr coordinate of the 2×2 interaction table.

Moreover, the interpretability of these coordinates is still supported by their relation to odds ratios of parts in the original table ([11], p. 44). This fact is obvious directly from the form of (46) since each coordinate is formed by the sum of logarithms of odds ratios which compare cell of the original table in position (r, s) with all cells that are north-west from the r-th row and s-th column - group A.

Although, the pivot coordinate system is proposed particularly for the cases, when the interpretation of single coordinates is not the main goal of the analysis, a new set of coordinates (with different interpretation) can be reached by permutation of rows and/or columns in the original compositional table. Accordingly, e.g., orthonormal coordinates that contain log odds ratio of a given 2 × 2 table can be easily constructed. They also enable to extract the only coordinate with log odds ratio interpretation that contains a given entry x_{rs}.

7 Applied methods

A natural aim of the analysis of compositional tables is to study relationship between its row and column factors. In [16] it was proposed to measure distance between the original compositional table $\mathbf{x}$ and its independent part $\mathbf{x}_{ind}$ using squared distance

$$\Delta^2(\mathbf{x}) = ||\mathbf{x}_{int}||_A^2 = ||\mathbf{x}||_A^2 - ||\mathbf{x}_{ind}||_A^2 \quad , \tag{47}$$

or relative squared distance

$$R_\Delta^2(\mathbf{x}) = \frac{\Delta^2(\mathbf{x})}{||\mathbf{x}||_A^2}, \quad 0 \leq R_\Delta^2 \leq 1 \quad , \tag{48}$$

which suppresses the influence of dimensions of compositional table $\mathbf{x}$ on squared distance. Values of relative squared distance, which are near to 1, are typical for tables with strong interactions between factors. On the other hand, low values give an evidence about independence between row and column factors. Moreover, due to decomposition

$$\mathbf{x} = \mathbf{x}_{ind} \oplus \left(\bigoplus_{i=1}^{I} \mathrm{row}_i(\mathbf{x}_{int}) \right) = \mathbf{x}_{ind} \oplus \left(\bigoplus_{j=1}^{J} \mathrm{col}_j(\mathbf{x}_{int}) \right) \quad , \tag{49}$$

the contribution of the i-th row to the squared norm is $||\mathrm{row}_i(\mathbf{x}_{int})||_A^2$ and similarly contribution of the j-th column is $||\mathrm{col}_j(\mathbf{x}_{int})||_A^2$.

Since orthonormal coordinates accounting for interactions between rows and columns were not available in [16], these features were measured using cross-contrasts and so called cell-interactions. The cross-contrast is defined as a simple balance of the part of interaction table at position (i, j) against the other parts in the same row or column,

$$I_{cross}(i,j) = \sqrt{\frac{I+J-2}{I+J-1}} \ln \frac{x_{ij}^{int}}{\left(\prod_{r \neq i} x_{rj}^{int} \prod_{s \neq j} x_{is}^{int}\right)^{1/(I+J-2)}} \quad . \tag{50}$$

The problem of these balances is that they are not orthogonal. On the other hand, their sum is closely connected to the square norm of $\mathbf{x}_{int}$ through relation

$$\sum_i \sum_j \left(I_{cross}(i,j)\right)^2 = \frac{(I+J)^2}{(I+J-1)(I+J-2)} ||\mathbf{x}_{int}||_A^2 \quad . \tag{51}$$

The cell-interaction is defined as balance between part of interaction table at position (i, j) and the rest of parts,

$$I_{cell}(i,j) = \sqrt{\frac{IJ-1}{IJ}} \ln \frac{x_{ij}^{int}}{\left(\prod_{(k,l) \neq (i,j)} x_{kl}^{int}\right)^{1/(IJ-1)}} \quad , \tag{52}$$

and is also connected to squared norm of the interaction table

$$\sum_i \sum_j \left(I_{cell}(i,j)\right)^2 = \frac{IJ}{I+J-1} ||\mathbf{x}_{int}||_A^2 \quad . \tag{53}$$

Furthermore, it can be shown, that in the case of 2×2 table all the cell-interactions are the same (up to its sign) and proportional to the interaction coordinate (33),

$$I_{cell}(1,1) = I_{cell}(2,2) = -I_{cell}(1,2) = -I_{cell}(2,1) = \frac{1}{2\sqrt{3}} \ln \frac{x_{11}x_{22}}{x_{12}x_{21}} \quad . \tag{54}$$

However, coordinate system proposed in Section 6.1 or its special case from Section 6.2 enables a deeper insight into the source of interactions between both factors, by considering interpretation of the odds ratio coordinates of the interaction table. Particularly if row and column factors are independent ($\mathbf{x} = \mathbf{x}_{ind}$), the interaction table equals to a neutral element of perturbation, all its parts are the same and the vector of odds ratio coordinates (33) $\mathbf{z}_{int}$ equals to a zero vector. On the other hand, high absolute values of this vector indicate presence of interactions between factors. Consequently, in the situation, when a random sample of compositional tables is available, the analysis of independence reduces to multivariate test on zero mean value of the vector of interaction coordinates $\mathbf{z}_{int}$. The structural approach to the analysis of independence between factors is also supported by the interpretation of these coordinates. As it was described in Section 6.1, coordinate z_i^{OR} can be interpreted as logarithm of odds ratio among groups of parts. Since in the independence case the odds ratio equals one, zero values of coordinates give an evidence against the presence of interactions between factors.

The thesis provides two real-world examples, where the logratio approach to analysis of independence between factors is employed. The first example analyzes relationship between age and BMI index and the second one studies distribution of the manufacturing output.

8 Original results

Compositional tables as observations carrying relative information about relationship between two factors represent a direct generalization of vector compositional data. Consequently, possibility of their appropriate orthonormal coordinate representation forms an important step for coordinate representation of multifactorial compositional data. The thesis presents a general coordinate system for compositional tables, which respects their two-factorial character. The resulting coordinates form a natural generalization of the concept of balances as introduced in [15], that have already proven their practical usefulness in a wide range of applications, and open a variety of perspectives for their further development. Moreover, a special case of this system is provided (pivot coordinates), which seems to be easier to handle. The interpretation of both coordinate system is in the thesis supported by detailed inspection of their covariance structure. And, finally, the logratio approach to analysis of independence between factors using orthonormal coordinates is proposed.

9 Summary of results

Similarly as for vector compositional data, proper coordinate representation of compositional tables is necessary to enable statistical processing using standard multivariate statistical tools. The proposed coordinate system (in both general and pivot versions) contains both balances and coordinates with log odds ratio interpretation and forms the main contribution of the thesis. These coordinates respect the possibility of decomposition of a compositional table into its independent and interactive parts. Consequently, it allows to study tables from the decomposition also separately and analyze, e.g., possible interactions between both factors only from the interactive part of coordinates. Accordingly, the general orthonormal coordinate system respects the nature of row and column factors and thus allows for their better interpretability. On the other hand, the pivot coordinates as their special case seem to be easier to handle and provide an automated version of the coordinate representation. Construction of the coordinate systems was described in a detail and endowed in the thesis with examples and graphical illustrations for better understanding. The theoretical part of the thesis is completed with the covariance structure of the proposed coordinates. Finally, the possibility of structural analysis of relationship between factors is discussed in its last section.

Beside the new coordinates, a promising result comes from comparison of coordinates of 2×2 compositional table with parameters of log-linear model, since development of a compositional alternative to standard methods of analysis of independence between two variables (factors) represents one possible direction of our further research. The new coordinates thus seem to have great potential for compositional data analysis itself (statistical analysis of compositional tables, multifactorial compositional data), but open also its new challenging prospectives.

List of publications

[1] Fačevicová, K., Hron, K., Todorov, V., Guo, D. and Templ, M. (2014) Logratio approach to statistical analysis of 2 × 2 compositional tables. *Journal of Applied Statistics, **41**, 944–958.*

Cited in:

- Fačevicová, K., Todorov, V. and Hron, K. (2014) Compositional tables analysis with application to manufacturing. *Mathematical methods in economics*, (eds. Talašová, J., Stoklasa, J. and Talášek, T.). Palacký University Olomouc. ISBN: 978-80-244-4209-9.
- Fačevicová, K. and Hron, K. (2015) Covariance structure of compositional tables. *Austrian Journal of Statistics, **44**, 31–44.*
- Templ, K. and Todorov, V. (2015) The software environment R for official statistics and survey methodology. *Austrian Journal of Statistics, **45**, 97–124.*

[2] Fačevicová, K. (2015) Použití logistické regrese pro diagnostiku výskytu rakoviny prostaty. *Informační bulettin České statistické společnosti, **26**(1-2), 10–17.*

[3] Fačevicová, K. and Hron, K. (2015) Covariance structure of compositional tables. *Austrian Journal of Statistics, **44**, 31–44.*

[4] Fačevicová, K., Hron, K., Todorov, V. and Templ, M. (2016) Compositional tables analysis in coordinates. *Scandinavian Journal of Statistics.* DOI: 10:1111/sjos.12223.

[5] Fačevicová, K., Bábek, O., Hron, K. and Kumpan, T. (2016) Geochemical signature of Devonian/Carboniferous boundary - a compositional approach. (Submitted.)

[6] Fačevicová, K., Hron, K., Todorov, V. and Templ, M. (2016) General approach to coordinate representation of compositional tables. (In progress.)

[7] Vencálek, O., Fačevicová, K., Fürst, T. and Grepl, M. (2013) When less is more: A simple predictive model for repeated prostate biopsy outcomes. *Cancer Epidemiology, **37**(6), 864–869.*

Proceedings

[8] Fačevicová, K. and Hron, K. (2013) Statistical analysis of compositional 2 × 2 tables. *Proceedings of the 5th international workshop on compositional*

data analysis., (eds Hron, K., Filzmoser, P. and Templ, M.). Technische Universität Wien. ISBN: 978-3-200-03103-6.

[9] Fačevicová, K., Todorov, V. and Hron, K. (2014) Compositional tables analysis with application to manufacturing. *Mathematical methods in economics*, (eds. Talašová, J., Stoklasa, J. and Talášek, T.). Palacký University Olomouc. ISBN: 978-80-244-4209-9.

Other references

[11] Agresti, A. (2002) *Categorical data analysis (2 ed.).* New York: Wiley.

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

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

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

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

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

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

[18] Fišerová, E. and Hron, K. (2011) On interpretation of orthonormal coordinates for compositional data. *Mathematical Geology, **43**, 455–468.*

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

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

List of conferences

- ROBUST, 9.-14.9.2012, Němčičky (CZ): Použití logistické regrese pro diagnostiku výskytu rakoviny prostaty (poster, in czech)
- CoDaWork 2013, 3.-7.6.2013, Vorau (AT): Statistical analysis of 2×2 compositional tables (poster)
- ODAM 2013, 12.-14.6.2013, Olomouc (CZ): Compositional tables analysis in coordinates (presentation)
- ROBUST, 19.-24.1.2014, Jetřichovice (CZ): Statistická analýza kompozičních tabulek (poster+presentation, in czech)
- StatGeo 2014, 17.-20.6.2014, Olomouc (CZ): Časové řady kompozičních tabulek (presentation, in czech)
- LinStat, 24.-28.8.2014, Linköping (SW): Coordinate representation of compositional tables (presentation)
- Mathematical Methods in Economics 2014, 10.-12.9.2014, Olomouc (CZ): Compositional tables analysis with application to manufacturing (presentation)
- ERCIM, 6.-8.12.2014, Pisa (IT): Robust exploratory data analysis of compositional tables (presentation)
- CoDaWork 2015, 1.-5.6.2015, L´Escala (ES): Orthonormal coordinate representation of compositional tables (presentation)
- IAMG, 5.-13.9.2015, Freiberg (DE): Geochemical signature of the Devonian/Carboniferous boundary - a compositional approach (presentation)
- AMISTAT, 10.-13.11.2015, Prague (CZ): Exponential families as affine subspaces of the Bayes space (poster)
- ERCIM, 12.-14.12.2015, London (GB): Orthonormal coordinate representation of compositional tables - the general approach (presentation)