Noname ADAC manuscript No.
(will be inserted by the editor)

Exploratory data analysis for interval compositional data

Karel Hron · Paula Brito · Peter Filzmoser

The date of receipt and acceptance will be inserted by the editor

Abstract Compositional data are considered as data where relative contributions of parts on a whole, conveyed by (log-)ratios between them, are essential for the analysis. In Symbolic Data Analysis (SDA), we are in the framework of interval data when elements are characterized by variables whose values are intervals on $\mathbb{R}$ representing inherent variability. In this paper, we address the special problem of the analysis of interval compositions, i.e., when the interval data are obtained by the aggregation of compositions. It is assumed that the interval information is represented by the respective midpoints and ranges, and both sources of information are considered as compositions. In this context, we introduce the representation of interval data as three-way data. In the framework of the log-ratio approach from compositional data analysis, it is outlined how interval compositions can be treated in an exploratory context. The goal of the analysis is to represent the compositions by coordinates which are interpretable in terms of the original compositional parts. This is achieved by summarizing all relative information (logratios) about each part into one coordinate from the coordinate system. Based on an example from the European Union Statistics on Income and Living Conditions (EU-SILC), several possibilities for an exploratory data analysis approach for interval compositions are outlined and investigated.

Karel Hron
Palacký University, Faculty of Science, 17. listopadu 12, CZ-77146, CZECH REPUBLIC (Corresponding author) E-mail: hronk@seznam.cz

Paula Brito
FEP & LIAAD INESC TEC, Universidade do Porto, Rua Dr. Roberto Frias, 4200-464 Porto, PORTUGAL E-mail: mpbrito@fep.up.pt

Peter Filzmoser
Vienna University of Technology, Institute of Statistics and Mathematical Methods in Economics, Wiedner Hauptstrasse 8-10, A-1040 Vienna, AUSTRIA E-mail: p.filzmoser@tuwien.ac.at

Keywords interval data · symbolic data analysis · Aitchison geometry on the simplex · orthonormal coordinates · outlier detection · principal component analysis

Subject classification 62H25, 62H99.

1 Introduction

It is often the case that analysts have huge sets of data, recorded in very large databases, but that the elements of interest are not the individual records but rather some second-level entities. For instance, in a database of individual expenses in different items, we are surely more interested in describing the general behavior of a person (or some pre-defined class or group of persons) rather than each of the expenses itself. The analysis then requires that the data for each person (or group) be somehow aggregated to obtain the information of interest. However, the observed variability for each person or within each group, which cannot be kept by summary statistics, should not be disregarded, so that data can no longer be properly described by the usual numerical and categorical variables without an unacceptable loss of information. Symbolic Data Analysis (henceforth SDA) (Billard and Diday, 2003; Bock and Diday, 2000; Diday and Noirhomme-Fraiture, 2008) provides a framework where the variability observed may effectively be considered in the data representation, and methods developed that take it into account. To describe groups of individuals or concepts, new variable types have been introduced, which may now assume other forms of realizations, to take into account the data intrinsic variability, by assuming multiple, possibly weighted, values for each case (see also Noirhomme-Fraiture and Brito, 2011). Let $S = \{s_1, \ldots, s_n\}$ be the set of entities under analysis - the considered groups or concepts - which are now the statistical units, and $Y_j, j = 1, \ldots, D$, the variables describing them. Then each $s_i \in S$ is represented by a symbolic description $d_i = (d_{i1}, \ldots, d_{iD})$ where $d_{ij} = Y_j(s_i)$ is a (finite) set of values or categories, an interval, or a distribution over a given set of sub-intervals or categories.

1.1 Interval compositional variables

One case of particular interest is when individual (first-level) numerical data are aggregated in the form of intervals, and represented by interval-valued variables, i.e., where $d_{ij} = Y_j(s_i) = [l_{ij}, u_{ij}], j = 1, \ldots, D, i = 1, \ldots, n$. However, it is well-known that using the standard interval arithmetic (Moore, 1966) for the statistical analysis of interval data quickly leads to wide intervals in the resulting quantities, useless in practice. In different works within the field of SDA, an alternative approach has been considered, based on the representation of multivariate interval data by the corresponding midpoints $c_{ij} = (l_{ij} + u_{ij})/2$ and ranges $r_{ij} = u_{ij} - l_{ij}$. This has been successfully applied to statistical analysis of interval data using known multivariate statistical methods (Brito

and Duarte Silva, 2011; Lauro and Palumbo, 2005; Neto and De Carvalho, 2008, 2010; Teles and Brito, 2013).

Considering further inherent properties of the individual observations, a natural question arises, of how to proceed with observations carrying relative information. In practice, the decision whether absolute information or rather the relative structure should be extracted from the data at hand, relies strongly on the purpose of the analysis. If relative information is in focus, the ratios between the variables should be kept for the statistical processing (Aitchison, 1986; Pawlowsky-Glahn et al, 2015a). In the standard case of individual observations, such data can be rescaled to an arbitrary sum of the components without any loss of information (although the most used ones are 1 and 100, resulting in proportional and percentage representations, respectively) – we refer to the principle of scale invariance of compositional data analysis (Pawlowsky-Glahn et al, 2015a). Obviously, such type of observations induce different geometrical properties from the standard multivariate observations (driven by the Euclidean geometry in the real space), represented by the Aitchison geometry on the simplex (Billheimer et al, 2001; Pawlowsky-Glahn and Egozcue, 2001; Egozcue and Pawlowsky-Glahn, 2006). The statistical analysis of compositional data is then performed in new real variables, constructed with respect to the Aitchison geometry. Because of their interpretation in terms of (log-)ratios of the original compositional parts, we refer to logratio analysis (Aitchison, 1986). If also the original scale of the data is important for interpretation purposes, the sum of the variable's values can be stored in an additional variable and analyzed as a part of multivariate information (Pawlowsky-Glahn et al, 2015b).

In our case, the question is then about how to extract the relevant information from the data matrix,

	X_1	$\dots$	X_j	$\dots$	X_D
s_1	$[l_{11}, u_{11}]$	$\dots$	$[l_{1j}, u_{1j}]$	$\dots$	$[l_{1D}, u_{1D}]$
$\dots$	$\dots$		$\dots$		$\dots$
s_i	$[l_{i1}, u_{i1}]$	$\dots$	$[l_{ij}, u_{ij}]$	$\dots$	$[l_{iD}, u_{iD}]$
$\dots$	$\dots$		$\dots$		$\dots$
s_n	$[l_{n1}, u_{n1}]$	$\dots$	$[l_{nj}, u_{nj}]$	$\dots$	$[l_{nD}, u_{nD}]$

where the single interval-valued variables represent quantitative descriptions of relative contributions on the whole as well, we refer to *interval compositional data*. Such data arise in many fields, where an aggregation is a natural consequence of the huge data collection process. For example, when relative structure of household expenditures is of primary interest, contributions of single variables (foodstuff, housing, clothing, etc.) can be merged according to regions or any other relevant key to intervals, formed usually by quantile characteristics (lower and upper quartiles, 0.1- and 0.9-quantiles, or even some standard deviation based interval, respectively). Of course, the aggregation step of data processing is thus always scale dependent.

1.2 Representations of interval compositional data

In the following, two approaches for a representation of interval compositional data are discussed. Although just the first one is then further considered in the paper, the aim of mentioning both is to point out that more options exist for the purpose. Once again, the standard interval arithmetic fails in the context of interval compositional data as it is not scale invariant, and further problems arise when even a fixed constant sum representation is required (Pavlačka, 2013). A way out thus seems to represent each single interval $[l_{ij}, u_{ij}]$, $i = 1, \ldots, n$, $j = 1, \ldots, D$, by the corresponding midpoint and range.

A natural first choice seems to be to take c_{ij} for the midpoint and r_{ij} for the range, as it has been done in interval data analysis in Brito and Duarte Silva (2011). Considering both midpoints and ranges together, we get the starting data for further statistical processing. Accordingly, this results in two $n \times D$ compositional data matrices $\mathbf{C} = (c_{ij})$ and $\mathbf{R} = (r_{ij})$. In both data sets just the ratios between the variables are informative, although it needs to be considered from a more general perspective that the ranges are seen as relative contributions on a "whole range", as it follows from the definition of compositional data. In any case, rescaling the original interval values, represented by midpoints and ranges, by any positive constant (for example, to get proportional representations of midpoints for an easier interpretation) should not affect their statistical analysis. Moreover, from the perspective of the Aitchison geometry, ratios between the corresponding midpoint and range components, that might have an attractive interpretation (like in the sense of a coefficient of variation), can be considered, up to a possible scaling constant, as shifting operation on the simplex, commonly known as perturbation between two compositional vectors (Aitchison and Ng, 2005).

One could also think about another possibility to represent the interval data, namely with $c_{ij} = \sqrt{l_{ij} u_{ij}}$ and $r_{ij} = \ln(u_{ij}/l_{ij})$. With this choice, the resulting ranges are already scale invariant by definition, so the matrix $\mathbf{R}$ would contain standard (positive) multivariate observations instead. Apparently, this approach might seem to be methodologically most consistent. Although taking the geometric mean instead of the arithmetic mean for the midpoints can be indeed considered as a relevant alternative, different origins of both representations would cause further interpretational problems. Particularly, the ranges r_{ij} are now logratios, forced to be positive by the definition. As most standard statistical methods are designed for the real sample space, taking a logarithmic transformation before their further processing would be recommendable. As a consequence, values of such preprocessed ranges would be quite far away from the original values and even hardly comparable with outputs of the logratio methodology, proposed below for the midpoints. For this reason, we prefer the previously formulated standard choice for the interval representation instead.

The rest of the paper is organized as follows. In the next section we recall the coordinate system for compositional data and derive a coordinate representation for interval compositions. Section 3 presents the analysis of interval compositional data. In Section 4, an application to a real data set is presented,

comparing results with those obtained using different approaches. Section 5 concludes the paper, pointing out directions for future work.

2 Coordinate representation of interval compositions

Compositional data are defined as D-part positive observations $\mathbf{x} = (x_1, \ldots, x_D)'$ carrying relative information. Compositions are characterized by the Aitchison geometry on the simplex, the sample space of their constant sum representations (Egozcue and Pawlowsky-Glahn, 2006; Pawlowsky-Glahn et al, 2015a), with Euclidean vector space structure of dimension $D-1$. Since the standard multivariate statistical analysis relies on the Euclidean geometry in real space (Eaton, 1983), $D-1$ orthonormal coordinates (with respect to the Aitchison geometry) are required that allow to proceed in a reasonable way. Unfortunately, it is not possible to derive $D-1$ orthonormal coordinates which are interpretable in terms of the original compositional parts $x_1, \ldots, x_D$ simultaneously. However, it is possible to employ isometric log-ratio (ilr) coordinates (Egozcue et al, 2003) which allow for an interpretation. Among other (more general) options (Egozcue and Pawlowsky-Glahn, 2005), one possible choice is to consider D coordinate systems, where always just one of the coordinates captures all the relative information about one of the compositional parts (x_k, $k = 1, \ldots, D$), which is then of main interest for the interpretation. Consequently, the remaining $D-2$ coordinates represent the resulting subcomposition by omitting the part x_k (Fišerová and Hron, 2011; Filzmoser et al, 2012). Without loss of generality, let the first orthonormal coordinate have such a property. Due to the Aitchison geometry and the nature of compositions, this coordinate will have the form of a log-ratio of the chosen part to the remaining parts in the composition, represented by their geometric mean. Concretely, for a composition $\mathbf{x}$ and a chosen part x_k we get $(D-1)$-dimensional real vectors $\mathbf{z}^{(k)} = (z_1^{(k)}, \ldots, z_{D-1}^{(k)})'$, $k = 1, \ldots, D$,

$$z_i^{(k)} = \sqrt{\frac{D-i}{D-i+1}}\, \ln \frac{x_i^{(k)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(k)}}},\ i = 1, \ldots, D-1, \tag{1}$$

where $(x_1^{(k)}, x_2^{(k)}, \ldots, x_k^{(k)}, x_{k+1}^{(k)}, \ldots, x_D^{(k)})$ stands for such a permutation of the parts $(x_1, \ldots, x_D)$ that always the k-th compositional part fills the first position, $(x_k, x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_D)$. In such a configuration, the first ilr variable $z_1^{(k)}$ explains all the relative information (logratios) about the original compositional part x_k, the coordinates $z_2^{(k)}, \ldots, z_{D-1}^{(k)}$ then explain the remaining logratios in the composition (Fišerová and Hron, 2011). Note that the only important position is that of $x_1^{(k)}$ (that is interpretable through $z_1^{(k)}$), the other parts can be chosen arbitrarily from the perspective of x_k, because different orthonormal coordinate systems are orthogonal rotations of each other (Egozcue et al, 2003). It is worth realizing that if all relative information concerning part

x_k in a given composition should be merged into one coordinate, then all pairwise logratios $\ln(x_k/x_1), \ldots, \ln(x_k/x_{k-1}), \ln(x_k/x_{k+1}), \ldots, \ln(x_k/x_D)$ need to be aggregated. So we arrive at

$$\begin{aligned}\ln(x_k/x_1) + \cdots + \ln(x_k/x_{k-1}) + \ln(x_k/x_{k+1}) + \cdots + \ln(x_k/x_D) \\ = (D-1)\ln \frac{x_1^{(k)}}{\sqrt[D-1]{\prod_{j=2}^{D} x_j^{(k)}}}.\end{aligned}$$

Up to a scaling constant, this is nothing else than the coordinate $z_1^{(k)}$ from (1). Of course, $z_1^{(k)}$ cannot be identified with the compositional part x_k, as the other parts are also naturally involved through the corresponding logratios. This coordinate is formed by a logratio between the part x_k and an "average part", resulting from the geometric mean of the remaining parts in the composition. Therefore, values of $z_1^{(k)}$ represent a measure of dominance of the part x_k with respect to the other parts.

Thus, when any statistical inference concerning all compositional parts (through the corresponding coordinates $z_1^{(k)}$) is of interest, D multivariate statistical models need to be constructed. Note that by merging and analyzing all coordinates $z_1^{(1)}, \ldots, z_1^{(D)}$ together, we would get (up to a scaling constant $\sqrt{\frac{D-1}{D}}$) the well-known centered log-ratio (clr) transformation (Aitchison, 1986), defined as

$$\mathbf{y} = \left(\frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \ldots, \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right)', \tag{2}$$

that yield coordinates with respect to a generating system. It means that there is one composition more than needed to form a basis with respect to the Aitchison geometry; for concrete formulas, see Pawlowsky-Glahn et al (2015a) (Chapter 4). However, since the covariance matrix of clr-transformed data is singular, this coordinate representation is not suitable for multivariate statistical analyses that are based on regular covariance matrices.

As mentioned in Section 1, midpoint and range representation of interval compositional data yield for each interval composition $(X_1, \ldots, X_D)'$ two assigned compositional vectors $\mathbf{c} = (c_1, \ldots, c_D)'$ and $\mathbf{r} = (r_1, \ldots, r_D)'$, that should be expressed in coordinates in order to enable their analysis using standard statistical methodology. Considering the clr transformation, as in (2), we obtain $\mathbf{y^c}$ for midpoints and $\mathbf{y^r}$ for ranges. However, the aim here is to deal with orthonormal coordinates $\mathbf{c}^{(k)}$ and $\mathbf{r}^{(l)}$, $k, l \in \{1, \ldots, D\}$ obtained as in (1) . Taking their interpretation into account, we primarily focus on combinations $\mathbf{c}^{(k)}$, $\mathbf{r}^{(k)}$, $k = 1, \ldots, D$, that merge (relative) information about both midpoint and range of the same (interval) part $X_k = [l_k, u_k]$ with respect to the remaining midpoint and range parts, respectively. Nevertheless, also the other combinations, when $k \neq l$, i.e., considering the relation between the midpoint of an interval part and the range of a another one, might be interesting in some special cases.

When analyzing standard interval data, it is proposed in Brito and Duarte Silva (2011) to consider various structures of the block covariance matrix of the joint vector of midpoints and ranges. The reason is that some of them (particularly for those that imply independence of midpoints and ranges under the assumption of normality) correspond to more parsimonious models, which may provide a good fit and result in less parameters to be estimated - which may be important in presence of small samples. In Brito and Duarte Silva (2011) the vector $[\mathbf{c}', (\mathbf{r}^*)']'$ is considered, where $\mathbf{r}^*$ represents the logarithmic transformation of the interval ranges $\mathbf{r}$. In the compositional case, we are forced to consider vectors $[(\mathbf{c}^{(k)})', (\mathbf{r}^{(l)})']'$ instead, i.e., use the ilr coordinates as defined in (1). For the purpose of further simplifications, the covariance matrix can be rewritten as a $2(D-1) \times 2(D-1)$ block matrix $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{c}} & \boldsymbol{\Sigma}_{\mathbf{cr}} \\ \boldsymbol{\Sigma}_{\mathbf{rc}} & \boldsymbol{\Sigma}_{\mathbf{r}} \end{pmatrix}$, where $\boldsymbol{\Sigma}_{\mathbf{c}}$ and $\boldsymbol{\Sigma}_{\mathbf{r}}$ are the covariance matrices of midpoints and ranges in orthonormal coordinates $\mathbf{c}^{(k)}$ and $\mathbf{r}^{(l)}$ for any chosen $k, l \in \{1, \ldots, D\}$, respectively, and $\boldsymbol{\Sigma}_{\mathbf{cr}} = \boldsymbol{\Sigma}'_{\mathbf{rc}}$ is the covariance matrix between midpoints and ranges. The special structure of $\boldsymbol{\Sigma}$ arises in practice usually directly from the design of the experiment. Nevertheless, in case of interval compositional data, with a restricted interpretation of the orthonormal coordinates, always just coordinates representing one midpoint ($c_1^{(k)}$) and one range ($r_1^{(l)}$) can be considered simultaneously for the analysis. Nevertheless, note that the covariance matrix cannot be simply built using the first coordinate of all the permutations in (1); in such a case the resulting $2D \times 2D$ covariance matrix would have at least two null eigenvalues due to the relation of these coordinates to clr variables. This would be very inconvenient for considering normal distribution (mentioned below) as it becomes singular.

Obviously, the most frequent case is that of an unrestricted covariance matrix. In some situations, it seems to be reasonable to consider uncorrelated midpoints and ranges, i.e. $\boldsymbol{\Sigma}_{\mathbf{cr}} = \boldsymbol{\Sigma}_{\mathbf{rc}} = \mathbf{0}$. A possible further extension of the latter case could also result in no correlation of the coordinates $c_1^{(k)}$ and $r_1^{(l)}$ with the remaining variables in $\mathbf{c}^{(k)}$ and $\mathbf{r}^{(l)}$, representing the corresponding remaining subcompositions.

For all the above mentioned covariance structures it holds that elements of the joint covariance matrix $\boldsymbol{\Sigma}$ can be reordered in form of a block-diagonal covariance matrix that is advantageous for maximum likelihood estimation of $\boldsymbol{\Sigma}$ from a random sample in case of normal distribution of both random compositions $\mathbf{c}^{(k)}$, $\mathbf{r}^{(l)}$, i.e. normal distribution of any of their coordinate representations (Brito and Duarte Silva, 2011; Mateu-Figueras and Pawlowsky-Glahn, 2008). A concrete selected covariance structure can also be tested from the sample using a range of tests on independence between random vectors (Kojadinovic and Holmes, 2009; Morrison, 1990; Seber, 1984).

3 Exploratory data analysis of interval compositions in coordinates

Although the nature of interval compositions differs substantially from the case of standard multivariate observations, the same problems concerning their statistical treatment arise. In particular, we are interested in the visualization of patterns in data, including groups and outlying observations. The coordinate representation of both midpoints and ranges of interval compositions enables proceeding to standard exploratory statistical analysis. For most object-oriented methods like cluster analysis or discriminant analysis, it is sufficient to consider simply any coordinates $\mathbf{c}^{(k)}$, $\mathbf{r}^{(l)}$, where $k, l \in \{1, \ldots, D\}$ must not even necessarily be the same (Filzmoser et al, 2012; Palarea-Albaladejo and Martín-Fernández, 2012). The reason is that, due to orthogonal relations between different ilr coordinates, results of such methods will be invariant to the choice of k and l. Nevertheless, when also the original interval variables, or patterns of midpoints/ranges are of primary interest, restrictions resulting from the interpretation of the orthonormal coordinates should be taken into account to accommodate the statistical procedures accordingly. Moreover, due to the same dimension of the resulting data sets for both midpoints and ranges, the coordinate representation of interval compositions can also be considered as a special case of three-way data, where the observations form the first mode, the compositional parts the second mode, and interactions between midpoints and ranges are represented in the third mode.

Visualization of outlying observations in multivariate data sets belongs to an initial task of any reasonable statistical processing. Outliers can strongly influence descriptive characteristics like the arithmetic mean and the sample covariance matrix that are used to construct, e.g., the well-known principal component analysis for dimension reduction, thus they can destroy the overall view on the multivariate data structure. Obviously, this is also the case in the statistical analysis of compositional data (Filzmoser and Hron, 2008; Filzmoser et al, 2009), although due to the Aitchison geometry other sources of outlyingness naturally arise (Filzmoser and Hron, 2011). With the coordinate representation at hand, we introduce a generalization of the approach from Filzmoser et al (2012), to see the effect of single interval variables X_k on forming multivariate outliers. Concretely, it is possible to visualize both coordinates $c_1^{(k)}$ and $r_1^{(k)}$ together using the bagplot to show location, spread and skewness of the joint data distribution (Rousseeuw et al, 1999). Note that the bagplot consists of three nested polygons, called the bag (includes regular observations), the loop (observations out of the bag, but still regular), and the fence (observations outside are flagged as outliers). The center of the data set is represented by the depth median, constructed using the Tukey depth. In order to evaluate the degree of outlyingness, to each observation a number corresponding to the actual data depth is assigned and the resulting output can be visualized graphically in form of univariate scatterplots. In the context of interval compositional data, outlier detection of the corresponding coordinates carrying relative information on midpoints and ranges of the com-

positional parts using the bagplot helps to reveal anomalous observations with respect to both mentioned characteristics.

A further step is to evaluate the relation between midpoints and ranges. Obviously, the data structure of midpoints and ranges, respectively, can in general differ, what can already be revealed from the bivariate analysis using the bagplot. From the point of view of *correlation analysis*, in addition to looking for pairwise (classical and robust) correlations between variables $c_1^{(k)}$ and $r_1^{(k)}$, $k = 1, \ldots, D$, we want to analyze an overall relation between both vectors $\mathbf{c}^{(k)}$ and $\mathbf{r}^{(k)}$ to consider eventually a particular covariance structure of the joint vector $[(\mathbf{c}^{(k)})', (\mathbf{r}^{(k)})']'$ as well as patterns, deviating from the main trend. One possibility to do that is to perform canonical correlation analysis in orthonormal coordinates (Filzmoser and Hron, 2009). Due to orthogonal relations between different ilr coordinates of midpoints/ranges, it turns out that for any choice of coordinates $\mathbf{c}^{(k)}$, $\mathbf{r}^{(l)}$ the canonical correlations will be exactly the same. Plotting the first canonical variables together then helps to reveal which observations do not follow the dominant data structure.

Other popular exploratory tool which will be of main interest in the following, are *dimension reduction methods*, based on singular value decomposition of midpoints and ranges in coordinates. They result in the well-known Principal Component Analysis (PCA), but also in the Parallel Factor Analysis (PARAFAC) method, applied to three-way data by considering midpoints and ranges as different layers. PCA of interval data has first been addressed in Chouakria et al (2000) and Cazes et al (1997), representing the observed intervals by their midpoints - the "centers method" - or by considering all the vertices of the hypercube representing each of the n entities in a p-dimensional space - the "vertices method". In Lauro and Palumbo (2005) a different approach is followed, where each variable is represented by the midpoints and ranges of its interval values. Three methods for principal component analysis of fuzzy interval data are investigated in Giordani and Kiers (2006). Zuccolotto (2007) uses a symbolic data approach for PCA of data described by the estimated means of a p-dimensional variable. In Wang et al (2012) a new method CIPCA is proposed , which takes a hypercube view with infinitely dense points uniformly distributed, defines the inner product of interval-valued variables, and transforms the PCA modeling into the computation of some inner products in the covariance matrix. Having the information on both midpoints and ranges available, we can either proceed to apply PCA separately, or in form of unfolded PCA, where common variables (coordinates) of midpoints and ranges are considered for the analysis of the block data matrix. Due to relations between midpoints and ranges that can be expected for most practical examples, the latter case seems to yield more reasonable results.

In case of compositional data, usually the clr coordinates are applied to perform PCA as they lead to an intuitive interpretation of the biplot of loadings and scores (represented as vertices and points, respectively) (Aitchison and Greenacre, 2002; Filzmoser et al, 2009). In particular, the links (distances) between vertices approximate the standard deviation of $\ln \frac{x_i}{x_j}$ for $i, j \in$

$\{1, \ldots, D\}$, which stands for a measure of strength of the relation between these compositional parts. In other words, when the vertices coincide, or nearly so, the ratio between the parts x_i and x_j is a constant, or nearly so. In case of interval compositional data, biplots for both midpoints and ranges can be constructed using the corresponding clr coordinates $\mathbf{y^c}$ and $\mathbf{y^r}$ (see (2)), respectively, to see the effect of the interval representations separately.

As mentioned above, it could also be interesting to merge both midpoints and ranges into one model to visualize the overall information on the input observations. Although unfolded PCA enables displaying information on both midpoints and ranges simultaneously in one common compositional biplot, we can even proceed to consider the three-way analysis as midpoints and ranges follow the coordinate representations of the same objects and variables. While the third mode, corresponding to different interval representations, can be used just to evaluate dissimilarity of midpoints and ranges, the first two modes (observations and compositional parts) can be used to get an overall picture of the multivariate data structure. For this purpose, basically two models are available, both to be preferably performed in clr coordinates (Engle et al, 2014; Di Palma et al, 2015). Let us denote the $n \times D \times 2$ clr data matrix $\mathbf{Y} = (y_{ijk})$ of midpoints (first layer, $k = 1$) and ranges (second layer, $k = 2$). One possibility is to apply a Tucker3 model, where a decomposition of the array into sets of scores and loadings is performed that should describe the data in a more condensed form than the input data array. For the above case of interval compositional data in clr coordinates, the Tucker3 model is defined as

$$y_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} a_{ip} b_{jq} c_{kr} g_{pqr} + e_{ijk},$$

$i = 1, \ldots, n$, $j = 1, \ldots, D$, $k = 1, 2$, where a_{ip}, b_{jq}, c_{kr} are loading elements, g_{pqr} are elements of the core array that stands for interactions between the three modes, and e_{ijk} is an element of the error term. For the usual setting of the number of factors, $P = Q = R = 2$, that enables to display the loadings graphically (except of the third mode that is not interesting for graphical visualization here), it is possible to proceed to a Tucker2 model that reduces only two of the three modes to components. The core array then represents the interactions among the elements of the midpoint-range mode and the components of the reduced modes (capturing information on observations and compositional parts, respectively) (Kroonenberg, 1983; Kroonenberg and De Leeuw, 1980). Formally, the Tucker2 model for this concrete case can be written as

$$y_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} a_{ip} b_{jq} g_{pqk} + e_{ijk}.$$

Moreover, the decision not to reduce the third mode has also methodological background; similarly one would proceed, e.g., also with longitudinal data, for which one typically may not want to reduce the time mode (Kroonenberg,

1983). Note that another modification of the Tucker3 model is provided by the well-known PARAFAC/Candecomp (CP) model (Bro, 1997), defined as

$$y_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + e_{ijk},$$

i.e. the core array takes the form of a superidentity array. Nevertheless, as it is not easily possible to avoid reducing the midpoint-range mode within the framework of CP, we consider just the Tucker2 model in the following.

Finally, note that for dimension reduction methods, the special covariance structure of midpoints and ranges in coordinates, as discussed in Section 2, could be applied as well. We can just remark that only under the null hypothesis $\boldsymbol{\Sigma}_{\mathbf{cr}} = \mathbf{0}$ it would be meaningful to perform PCA for midpoints and ranges separately, otherwise we have to proceed necessarily to any kind of three-way analysis (unfolded PCA, Tucker2 model).

4 Application

To illustrate the methodological considerations on an exploratory statistical analysis of interval compositions, we employ a data set that was synthetically generated from real Austrian EU-SILC (European Union Statistics on Income and Living Conditions) data, quoted as `eusilc` in the R-library `laeken` (Alfons and Templ, 2013), where almost 15 000 observations were collected and analyzed for a range of variables on living conditions in Austria. For the purpose of our study, just four of them, *employee* (employee net income), *self* (income from self-employment), *unemploy* (unemployment benefits) and *old-age* (old-age benefits) were taken into account. Although the original data are provided in Euro, we are indeed faced with compositional data, because the primary interest is devoted to relative contributions of single compounds to the overall income.

Here, we are interested in analysing and comparing the situation for both genders in the different Austrian regions, and not on the information at individual level. Therefore, the nonzero income data are aggregated according to the nine Austrian states - Burgenland (Bu), Carinthia (Ca), Lower Austria (LA), Salzburg (Sa), Styria (St), Tyrol (Ty), Upper Austria (UA), Vienna (Vi) and Vorarlberg (Vo) - and gender information (m, f). The resulting data matrix has 18 rows (observations), both genders for each Austrian state, and four variables. The aggregation is done in form of interquartile range and its center, playing the role of ranges and midpoints, respectively. For example, for the females from Styria, there are 458 non-zero values of the variable *employee* available, and this information is aggregated in form of interquartile range and its center in order to reduce the impact of data outliers. Thus, the large-scale input data is reduced to the description of the 18 demographic groups of interest (see Table 1). The symbolic data representation allows for a remarkable reduction of the dataset, bringing it to a more manageable size, by aggregating

data at the user's chosen degree of granularity while keeping the information on the intrinsic variability.

	employee	*self*	*unemploy*	*old-age*
Bu.f	11580 (*11757*)	8599 (*8461*)	5021 (*5973*)	10379 (*7318*)
Bu.m	17851 (*10800*)	14721 (*17009*)	3697 (*4376*)	13851 (*9061*)
Ca.f	13842 (*11490*)	9059 (*8872*)	2710 (*2682*)	11433 (*7014*)
Ca.m	20484 (*11104*)	15010 (*18256*)	3970 (*3007*)	15953 (*10383*)
⋮	⋮	⋮	⋮	⋮
Vo.f	11666 (*11868*)	6135 (*4267*)	2968 (*3063*)	12425 (*8430*)
Vo.m	21166 (*12870*)	15828 (*201230*)	4060 (*4607*)	20753 (*12591*)

Table 1 Midpoints (antique) and interquartile ranges (*italic*) for the aggregated Austrian EU-SILC data (in EUR).

By following the above reasoning, the log-ratio analysis of midpoints and ranges according to Sections 2 and 3 needs to be applied, because we focus on the relative structure of income variables (rather than their absolute values).

As a first step, it might be useful to visualize the original aggregated data structure, i.e., midpoints and ranges compositions. As compositional data can be expressed as observations with one dimension less than the actual number of parts (proportions, percentages), it is possible to display the three-part midpoints and ranges in form of a ternary diagram. A ternary diagram is an equilateral triangle $X_1X_2X_3$ such that a composition $\mathbf{x} = (x_1, x_2, x_3)'$ is plotted at a distance x_1 from the opposite side of vertex X_1, at a distance x_2 from the opposite side of vertex X_2, and at a distance x_3 from the opposite side of vertex X_3 (see, e.g. Aitchison, 1986, for further details). Since two compositional vectors correspond to each group under analysis, both need to be displayed simultaneously to see possible systematic patterns between midpoints and ranges. Because our data set has four parts, just those parts related to active age respondents were considered. Accordingly, midpoints and ranges for subcompositions with parts x_1 =*employee*, x_2 =*self* and x_3 =*unemploy* are visualized in Figure 1 (left), joined by segments with respect to the Aitchison geometry (Egozcue and Pawlowsky-Glahn, 2006). The ternary diagram reveals essentially the same pattern of midpoints and ranges for most of the observations. This is even easier to see when both midpoints and ranges are expressed in orthonormal coordinates $z_1^{(1)}$, $z_2^{(1)}$ which stand in favor of the *employee* variable (any other choice would lead just to a rotation of the data structure). In other words, the variable *employee* is taken in the nominator of formula (1), and thus all relative information about *employee* is expressed in the coordinate $z_1^{(1)}$. The coordinate $z_2^{(1)}$ expresses the (scaled) log-ratio of *self* to *unemploy*; explicitly

$$z_1^{(1)} = \frac{\sqrt{2}}{\sqrt{3}} \ln \frac{x_1}{\sqrt{x_2 x_3}}, \quad z_2^{(1)} = \frac{1}{\sqrt{2}} \ln \frac{x_2}{x_3}.$$

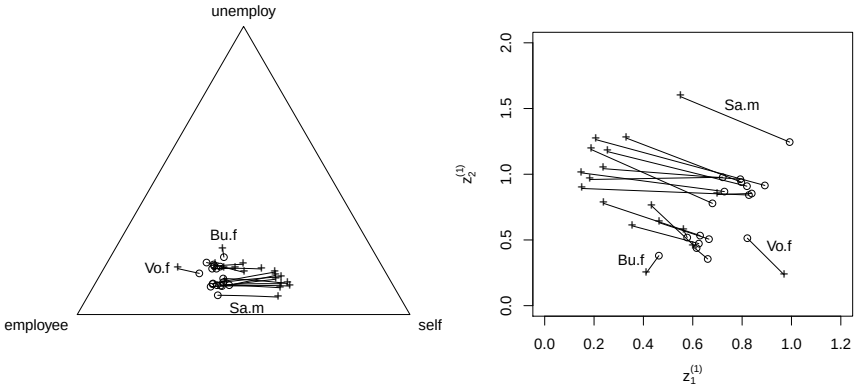

Fig. 1 Interval compositional observations (EU-SILC data set) in the ternary diagram (left) and in orthonormal coordinates (right). Midpoints are shown by ∘, ranges are marked by +.

In the plots of Figure 1, the midpoints are shown by the symbol ∘, and the ranges by +. Midpoints and ranges are connected for the same gender-state group. From both figures one can conclude that the part *employee* dominates for the midpoints (higher values for $z_1^{(1)}$), and that the variability in ranges is conveyed mostly by the variable *self*. There are also some subgroups and outliers visible. For example, the female group of Vorarlberg (Vo.f) has a much higher midpoint and range for the variable *employee*, relative to the other variables. For the group Bu.f, on the other hand, the unemployment benefits are much more dominating than for the other groups, in terms of both midpoint and range.

In order to see relations between coordinates that correspond to midpoints and ranges of single compositional parts (in the sense of a coordinate representation (1)), four bagplots are shown in Figure 2. It is easy to see that except of *employee* there is quite strong positive relation between the corresponding midpoints and ranges. This is a kind of justification for the rather systematic pattern between midpoints and ranges in Figure 1, even when the variable *old-age* was not employed there. We also find back the two outliers Sa.m and Bu.f in the unemployment benefits, that were visible already in Figure 1.

Before we proceed to dimension reduction methods, it is interesting to see whether the covariance matrix of the random vector $[(\mathbf{c}^{(k)})', (\mathbf{r}^{(l)})']'$ for any $k, l = 1, \ldots, 4$ can be expressed in block diagonal form, i.e. we perform an independence test between $(\mathbf{c}^{(k)})'$ and $(\mathbf{r}^{(l)})'$. As we cannot strictly refer to a random sample in case of symbolic data analysis, an empirical test is preferable. From a range of possibilities, we have chosen an independence test among continuous random vectors based on the empirical copula process (Kojadinovic and Holmes, 2009), where the corresponding p-values are obtained

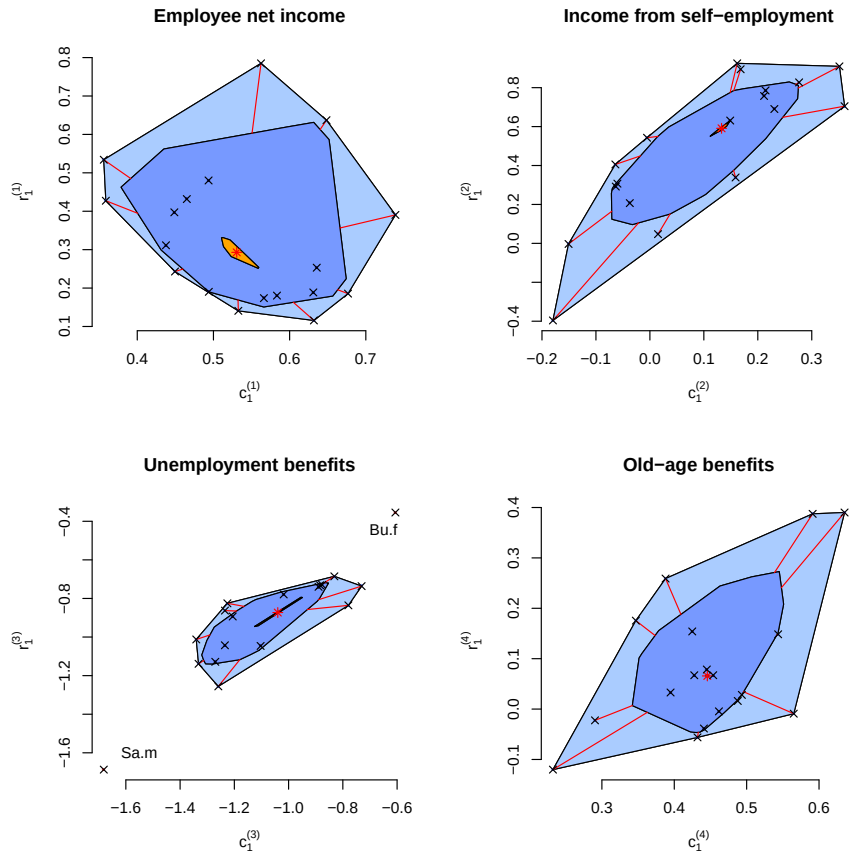

Fig. 2 Bivariate analysis of coordinates of midpoints and ranges for single parts in the EU-SILC data set: *employee* (upper left), *self* (upper right), *unemploy* (lower left) and *old-age* (lower right).

through the bootstrap/permutation methodology. Obviously, the result of the test will in principle not depend on the particular input coordinates. For a particular bootstrap realization, we obtained $p = 0.089$, so that the independence would be nearly rejected on the usual significance level $\alpha = 0.05$.

As a consequence of independence testing, we start also with PCA performed on both midpoints and ranges separately before we proceed to three-way analysis. The resulting biplots are displayed in Figure 3. Both biplots explain about 97% of the total variability in the respective data sets, thus they reflect the multivariate compositional data structure very well. Although there are some differences between both biplots, basic features are common for both midpoints and ranges. In both cases, the first principal component separates well male from female information, with clearly defined outliers in

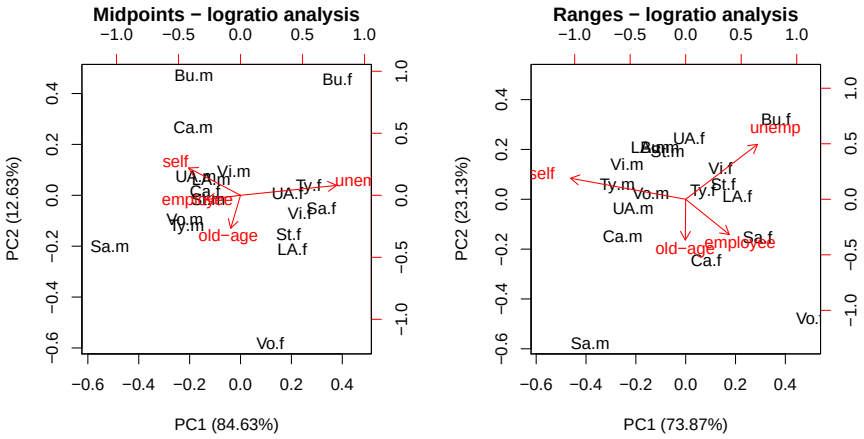

Fig. 3 Compositional biplots for both midpoints (left) and ranges (right).

both gender groups (Sa.m, Vo.f). The relative importance and also the corresponding relative variability, represented in the range information, of the unemployment benefits is higher for female income structure. Men tend to prefer self-employment. As expected, there is no relation of the relative amount of old-age benefits to any of both groups.

The picture is completely changed if we move to biplots of the original midpoints and logarithmized ranges. This approach was proposed in Brito and Duarte Silva (2011) for standard interval data and for completeness' sake, it is mentioned here as well. Although taking the logarithmic transformation of ranges was motivated in particular to enable using the normal distribution for both representations, it has also purely compositional consequences. In addition to preserve the compositional information it also keeps information of the geometric mean of ranges (Pawlowsky-Glahn et al, 2015a). Nevertheless, by taking the original midpoints the relative character of the data set is ignored, also considering two different scales for the analysis (multiplicative for midpoints and additive for log-ranges) can cause further interpretational problems. Thus, it cannot be generally recommended, even when the amount of explained variability by the first two principal components is quite comparable (97% for midpoints and 92% for log-ranges). While the basic data configuration (differences between male and female income structure) is quite comparable (except of that previous outliers vanished for the original ranges), the interpretation of variables (see loadings) is completely different, in case of midpoints the covariance structure is apparently distorted. The reason for that seems to be different scales, as mentioned above, but also inappropriate sample spaces of both compositional data sets. In fact, an analysis in terms of Brito and Duarte Silva (2011) would be appropriate if the absolute values of

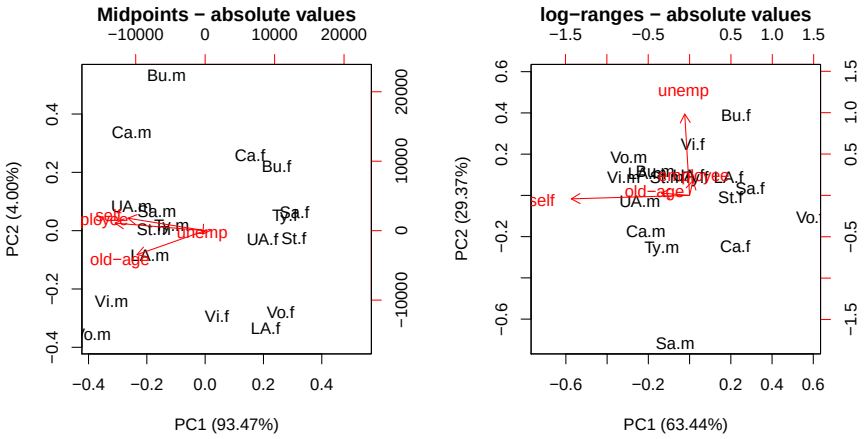

Fig. 4 Standard biplots for both midpoints (left) and log-ranges (right).

allowances would be of primary interest, but not their relative structure as it is the case here.

For this reason, for the three-way analysis we employ just the log-ratio analysis (described in Section 3). At first the effect of merging midpoints and ranges together can be analyzed using unfolded PCA and the respective compositional biplot (Figure 5), where the partial information corresponding to the same demographic groups are joined by a segment. We can see that the basic information on observations, visible already for the separate analysis (data structure, outliers) is well reflected in the biplot. Also the relative effect of self-employment for the male income structure is clearly visible, female income structure is now driven by more parts simultaneously, following rather regional patterns. Moreover, there is quite a strong relation between midpoints and ranges, reflected by almost the same direction and orientation of the segment lines. Note that in the default setting, the compositional biplot is represented by the covariance biplot (Aitchison and Greenacre, 2002) that favours the display of variables (or links between vertices). The points for the individuals, scores, are scaled, thus distorting the projected Aitchison distances between points. Therefore, as an alternative, also the form biplot can be inspected for which the interpretation in terms of points is more relevant. However, in this case both biplot versions are very similar to each other.

Finally, we proceed to the Tucker2 model for three-way log-ratio analysis of midpoints and ranges. Here the information on midpoints and ranges is merged into single objects, represented in the first mode, while the original parts (in form of their corresponding clr variables) are displayed in the second mode. As in the previous analyses, again quite a similar structure among the observations can be observed in the first mode (Figure 6, left). This supports

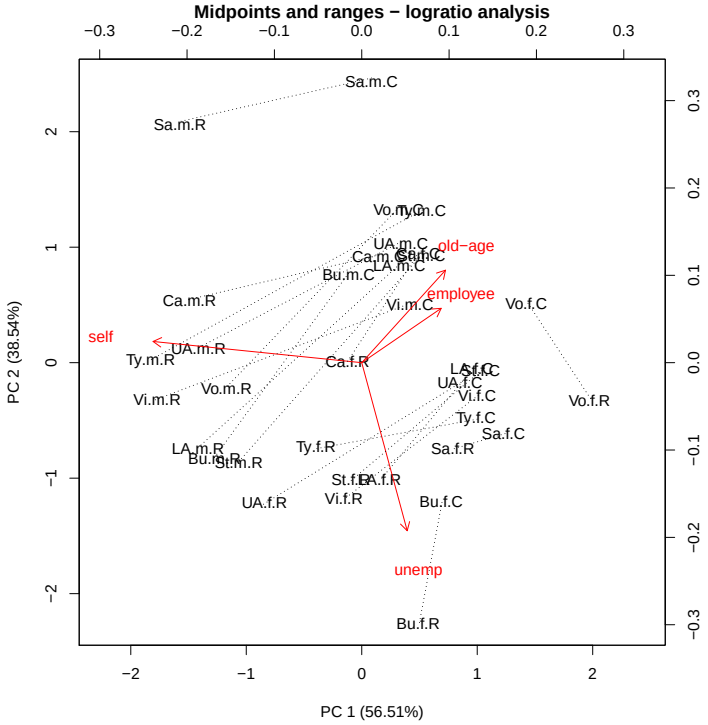

Fig. 5 Compositional biplot (covariance biplot) resulting from unfolded PCA for both midpoints (.C) and ranges (.R).

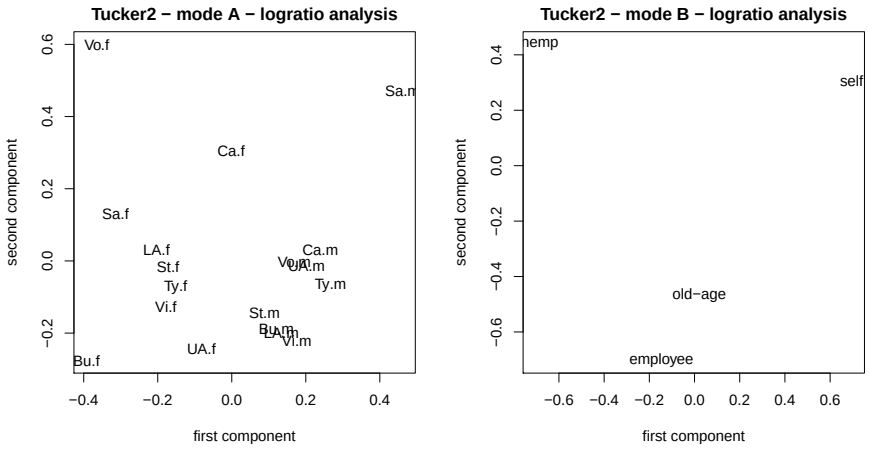

Fig. 6 Tucker2 log-ratio analysis: first mode (left) and second mode (right).

the consistency of the performed analysis based on the log-ratio approach. Also the relations between the clr variables reflect the case of unfolded PCA (and even the case of separate compositional biplots for midpoints and ranges), see Figure 6 (right). A close relation between employee net income and old-age benefits corresponds well to the expected reality. Namely, both kinds of income are rather state dependent (old age benefits almost exclusively), so the proportion between employee net income and old-age benefits should be more or less the same among the regions of Austria and the gender groups. On the other hand, unemployment and self-employment can differ much more due to specific regional effects.

5 Summary

Symbolic data are often the result of aggregating larger amounts of numerical information to intervals, which convey the variability intrinsic to the groups of interest. The intervals can then be represented in form of midpoints and ranges. We introduced a three-way representation of interval data, which allows for a wide range of multivariate analyses – and may open the way to novel approaches. If the multivariate data are compositions, only the pairwise proportions between the compositional parts carry the essential information that needs to be analyzed. This concept applies also to symbolic data, where relative information of midpoints and ranges is of interest.

In this paper we used the log-ratio approach to analyze the relative information (Aitchison, 1986). In particular, ilr-coordinates were employed (Egozcue et al, 2003), which is a standard procedure nowadays in many applications (Pawlowsky-Glahn et al, 2015a). This approach has the appeal that one can work in the usual Euclidean geometry rather than in the simplex, and that it allows to construct coordinates with an interpretation in terms of the original parts.

We addressed several possibilities for an exploratory data analysis in this context. In essence, one is interested in a similar data inspection as in an analysis of standard data: groups and outliers among the observations, relations between the variables, relations between variables and observations. For this type of interval data with midpoint and range representation, it is also of interest to analyze the relation between midpoints and ranges. This is because this relation needs to be taken into account for a joint modeling (Brito and Duarte Silva, 2011). The exploratory tools listed here are definitely not comprehensive, and they can be extended to the specific needs of the analysis.

References

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

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

Aitchison, J. and Ng, K.W. (2005). The role of perturbation in compositional data analysis. Statistical Modelling 5: 173–185.

Alfons, A. and Templ, M. (2013). Estimation of social exclusion indicators from complex surveys: The R package laeken, Journal of Statistical Software, 54 (15): 1–25.

Billheimer, D. Guttorp, P. and Fagan, W. (2001). Statistical interpretation of species composition, Journal of the American Statistical Association, 96: 1205–1214.

Billard, L. and Diday, E. (2003). From the statistics of data to the statistics of knowledge: Symbolic Data Analysis, Journal of the American Statistical Association, 98 (462): 470–487.

Bock, H.-H. and Diday, E. (editors) (2000). Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data, Springer, Heidelberg.

Brito, P. and Duarte Silva, A.P. (2011). Modelling interval data with Normal and Skew-Normal distributions, Journal of Applied Statistics, Vol. 39, Issue 1: 3-20.

Bro, R. (1997). PARAFAC. Tutorial and applications, Chemometrics and Intelligent Laboratory Systems 38: 149–171.

Cazes, P., Chouakria, A., Diday, E. and Schektman, Y. (1997). Extensions de l'Analyse en Composantes Principales à des données de type intervalle, Revue de Statistique Appliquée, 24: 5–24.

Chouakria, A., Cazes, P. and Diday, E. (2000). Symbolic Principal Component Analysis. In: Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data, H.-H. Bock and E. Diday (Eds.),Springer, Heidelberg: pp. 200–212.

Diday, E. and Noirhomme-Fraiture, M. (editors) (2008). Symbolic Data Analysis and the SODAS Software. Wiley, Chichester.

Di Palma, A.M., Filzmoser, P., Gallo, M. and Hron, K. (2015). A robust CP model for compositional data. Submitted.

Eaton, M.L. (1983). Multivariate Statistics. A Vector Space Approach. John Wiley & Sons, New York.

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

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

Egozcue J.J., Pawlowsky-Glahn, V. (2006). Simplicial geometry for compositional data. In: *Compositional data analysis in the geosciences: From theory to practice*, A. Buccianti, G. Mateu-Figueras and V. Pawlowsky-Glahn (Eds.) Geological Society, London, Special Publications 264: 145–160.

Filzmoser, P. and Hron, K. (2008). Outlier detection for compositional data using robust methods, Mathematical Geosciences, 40 (3): 233–248.

Filzmoser, P., Hron, K. and Reimann, C. (2009). Principal component analysis for compositional data with outliers, Environmetrics, 20 (6): 621–632.

Filzmoser, P. and Hron, K. (2009). Correlation analysis for compositional data, Mathematical Geosciences, 41 (8): 905–919.

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

Filzmoser, P. and Hron, K. (2011). Robust statistical analysis. In: V. Pawlowsky-Glahn, A. Buccianti (Eds.), Compositional data analysis: Theory and applications, Wiley, Chichester: 59–72.

Fišerová, E. and Hron, K. (2011). On interpretation of orthonormal coordinates for compositional data, Mathematical Geosciences, 43: 455–468.

Engle, M.A., Gallo, M., Schroeder, K.T., Geboy, N.J. and Zupancic, J.W. (2014). Three-way compositional analysis of water quality monitoring data, Environmental and Ecological Statistics, 21 (3): 565–581.

Giordani, P. and Kiers, H.A.L. (2006). A comparison of three methods for Principal Component Analysis of fuzzy interval data, Computational Statistics & Data Analysis, special issue "The Fuzzy Approach to Statistical Analysis", 51 (1): 379–397.

Kojadinovic, I. and Holmes, M. (2009). Tests of independence among continuous random vectors based on Cramér-von Mises functionals of the empirical copula process. Journal of Multivariate Analysis, 100: 1137–1154.

Kroonenberg, E.M. (1983). Three-mode principal component analysis: Theory and applications, DSWO, Leiden.

Kroonenberg, E.M., and De Leeuw, J. (1980). Principal component analysis of three-mode data by means of alternating least squares algorithms. Psychometrika, 45: 69–97.

Lauro, C. and Palumbo, F. (2005). Principal component analysis for non-precise data. In: New Developments in Classification and Data Analysis, M. Vichi *et al* (Eds.), Springer, Berlin, Heidelberg: 173–184.

Mateu-Figueras, G. and Pawlowsky-Glahn V. (2008). A critical approach to probability laws in geochemistry, Mathematical Geosciences 40: 489–502.

Moore, R.E. (1966). Interval Analysis, Prentice Hall, New Jersey.

Morrison, D.F. (1990). Multivariate Statistical Methods, 3rd ed., McGraw-Hill, New York.

Neto, E.A.L. and De Carvalho, F.A.T. (2008). Centre and range method for fitting a linear regression model to symbolic intervalar data, Computational Statistics & Data Analysis, 52, 3: 1500–1515.

Neto, E.A.L. and De Carvalho, F.A.T. (2010). Constrained linear regression models for symbolic interval-valued variables, Computational Statistics & Data Analysis, 54, 2: 333-347.

Noirhomme-Fraiture, M. and Brito, P. (2011). Far Beyond the Classical Data Models: Symbolic Data Analysis, Statistical Analysis and Data Mining, Vol. 4, Issue 2: 157-170.

Palarea-Albaladejo, J. and Martín-Fernández, J.A. (2012). Dealing with distances and transformations for fuzzy c-means clustering of compositional data, Journal of Classification 29: 144–169.

Pavlačka, O. (2013). Note on the lack of equality between fuzzy weighted average and fuzzy convex sum, Fuzzy Sets and Systems 213: 102–105.

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

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

Pawlowsky-Glahn, V., Egozcue, J.J. and Lovell, D. (2015b) Tools for compositional data with a total, Statistical Modelling, 15: 175–190.

Rousseeuw, P.J., Ruts, I. and Tukey, J.W. (1999). The bagplot: A bivariate boxplot, The American Statistician 53 (4): 382–387.

Seber, G.A.F. (1984). Multivariate Observations, Wiley.

Teles, P. and Brito, P. (2013). Modeling interval time series with space-time processes, Communications in Statistics - Theory and Methods, 44 (17): 3599–3627.

Wang, H., Guan, R. and Wu, J. (2012). CIPCA: Complete-Information-based Principal Component Analysis for interval-valued data, Neurocomputing, 86: 158–169.

Zuccolotto, P. (2007). Principal components of sample estimates: an approach through symbolic data analysis, Statistical Methods and Applications, 16 (2): 173–192.