JOURNAL OF APPLIED STATISTICS, 2017
VOL. 44, NO. 4, 734–752
http://dx.doi.org/10.1080/02664763.2016.1182135

Taylor & Francis
Taylor & Francis Group

# Exploratory tools for outlier detection in compositional data with structural zeros

M. Templ[a], K. Hron[b] and P. Filzmoser[a]

[a]Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Wiedner Hauptstraße 8-10, A-1040 Vienna, Austria; [b]Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, 17. listopadu 12, CZ-77146 Olomouc, Czech Republic

**ABSTRACT**

The analysis of compositional data using the log-ratio approach is based on ratios between the compositional parts. Zeros in the parts thus cause serious difficulties for the analysis. This is a particular problem in case of structural zeros, which cannot be simply replaced by a non-zero value as it is done, e.g. for values below detection limit or missing values. Instead, zeros to be incorporated into further statistical processing. The focus is on exploratory tools for identifying outliers in compositional data sets with structural zeros. For this purpose, Mahalanobis distances are estimated, computed either directly for subcompositions determined by their zero patterns, or by using imputation to improve the efficiency of the estimates, and then proceed to the subcompositional and subgroup level. For this approach, new theory is formulated that allows to estimate covariances for imputed compositional data and to apply estimations on subgroups using parts of this covariance matrix. Moreover, the zero pattern structure is analyzed using principal component analysis for binary data to achieve a comprehensive view of the overall multivariate data structure. The proposed tools are applied to larger compositional data sets from official statistics, where the need for an appropriate treatment of zeros is obvious.

**ARTICLE HISTORY**
Received 11 August 2015
Accepted 20 April 2016

**KEYWORDS**
Structural zeros; Aitchison geometry on the simplex; covariance estimation; Mahalanobis distance; principal component analysis

**AMS SUBJECT CLASSIFICATION**
97K80; 97K70; 97K40

## 1. Introduction

Compositional data are observations that describe quantitatively relative contributions of parts on a whole. Such data occur frequently in many practical situations [1,31,32]. Typical examples are vegetation compositions of various plant species in different survey areas, election results of political parties in different regions of a country, or household expenditures on various costs such as housing, foodstuff, alcohol and tobacco, furnishings, health and transportation for a sample of households. The sum of these parts is not necessarily 1 (or 100%), but since the relevant information is contained in the ratios between the parts, a constant sum constraint 1 or 100 can be achieved without any loss of information.

Compositional data induce an own sample space; they are represented in the simplex, with the Aitchison geometry [14,15] that is substantially different from the usual Euclidean

**CONTACT** M. Templ matthias.templ@tuwien.ac.at

geometry. Thus, standard statistical methods designed for the usual Euclidean geometry cannot be directly applied to compositions (see, e.g. [31]). As a way out, a family of log-ratio coordinates that enables to express compositional data from the simplex in the Euclidean real space was introduced [1]. Nowadays, the isometric log-ratio (ilr) coordinates [15] are preferred due to advantageous theoretical properties like isometry and non-singularity. The latter property is not fulfilled for the second popular coordinate system, the centered log-ratio (clr) coordinates [1] that are often used for theoretical considerations. Other (non-isometric) log-ratio coordinates that occur in practice are the additive log-ratio (alr) and multiplicative log-ratio (mlr) coordinates; the latter can even be used in the context of dealing with structural zeros in compositional data [35].

The clr coordinates are defined for a $D$-part composition $\mathbf{x} = (x_1, \ldots, x_D)'$ as

$$clr(\mathbf{x}) = \mathbf{y} = (y_1, \ldots, y_D)' = \left( \ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right)' \tag{1}$$

and map the composition $\mathbf{x}$ to a hyperplane $\mathcal{H} : y_1 + \cdots + y_D = 0$, i.e. to a subspace of $\mathbf{R}^D$. The definition of the hyperplane explains the singularity issue, and the coefficients of an orthonormal basis of $\mathcal{H}$ form ilr coordinates. A popular choice for this basis [15,24] leads to $ilr(\mathbf{x}) = \mathbf{z} = (z_1, \ldots, z_{D-1})'$, where

$$z_j = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{x_j}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_k}}, \quad j = 1, \ldots, D-1. \tag{2}$$

If the orthonormal basis vectors, here

$$\mathbf{v}_{.j} = \sqrt{\frac{D-j}{D-j+1}} \left( 0, \ldots, 0, 1, -\frac{1}{D-j}, \ldots, -\frac{1}{D-j} \right)' \quad \text{for } j = 1, \ldots, D-1, \tag{3}$$

with $j-1$ zero entries, are collected as columns in a $D \times (D-1)$ matrix $\mathbf{V} = (\mathbf{v}_{.1}, \ldots, \mathbf{v}_{.D-1})$, then the relations $\mathbf{y} = \mathbf{V}\mathbf{z}$ and $\mathbf{z} = \mathbf{V}'\mathbf{y}$ between both log-ratio coordinates can be easily derived [15].

In all the above-mentioned examples, zeros might naturally occur in the data set. Here, we do not consider zeros caused by any rounding errors (this refers to so-called *rounded zeros*), but rather to the result of structural processes (*structural zeros*). Examples for structural zeros are plant species that are not able to survive in a given soil type or climate, a political party that has no candidates in a region, or teetotal households that do not have expenditures on alcohol and tobacco. Zero values are in contradiction with the definition of compositions as data with *positive* entries. This is quite a natural requirement, because a multivariate observation is a composition if and only if all the relevant information is contained in the ratios between the compositional parts [12]. However, as a severe consequence, the log-ratio coordinates, where logarithms of ratios of compositional parts are taken, cannot be applied to compositional data with zeros. Also some alternative transformations for compositional data were proposed that avoid the problem of dealing with zero compositional parts, like the square root and the hyperspherical transformations [7,34,35,43], resulting from considering a fixed constant sum constraint 1 of compositional

parts instead of scale invariance as it is the case in the log-ratio approach. Although these transformations represent concepts of dealing with compositional data that allow for zero parts, they fail (from the perspective of the log-ratio approach) in other important features like incorporating relative scale of compositions, or their subcompositional coherence [12].

A problem similar to rounded zeros are *count zeros*, which occur when dealing with discrete compositions [6,27,29]. Discrete compositions are formed by counts, and the zeros occur in parts that correspond to rare events. Here, the problem with zeros is usually solved by estimating parameters of the underlying multinomial distribution, e.g. using generalized linear models [6] or Bayesian estimation based on Dirichlet prior distributions [27,29]. In contrast to count zeros, structural zeros result from continuous compositions (where the parts change continuously) and a zero stands for a complete absence of the corresponding part in the composition.

This paper is devoted to an exploratory analysis of compositional data with structural zeros, focused on the detection of outliers and extreme values. As existing methods for identifying outliers in compositional data assume that all data entries are strictly positive [18,20], they need to be adapted in order to involve also the zero pattern structure. Structural zeros are a peculiar problem when using the log-ratio approach to compositional data analysis. This is also the main reason why there is still not a concise methodology for this purpose. The aim of the paper is not to develop a specific new method, but rather to combine existing tools by following the nature of compositional data and the log-ratio methodology for their statistical processing in the presence of structural zeros. The main innovation is formed by splitting the data analysis into two steps. In the first step, the overall covariance structure (by suppressing the zero patterns via imputation) is constructed and used for outlier detection within subcompositions of non-zero parts. In the second step, the zero patterns are analyzed with appropriate statistical tools, particularly with principal component analysis for binary data.

In the next section, we provide an overview of the recent developments concerning structural zeros in compositional data. In Section 3, we introduce a new methodology for outlier detection in compositions with structural zeros, where the information on non-zero components as well as the information on structural zeros is employed. The presented methods can be seen as a first (explorative) step for research on compositional data analysis including structural zeros. Two real data examples are used in Section 4 to discuss the different approaches. The final Section 5 concludes.

## 2. Review of methods dealing with structural zeros

The problem of structural zeros was already of special interest at the time when the log-ratio approach for compositional data was introduced. Possible solutions were discussed in [1], where amalgamation and a non-zero replacement of zeros was proposed. Amalgamation of a $D$-part composition $\mathbf{x} = (x_1, \ldots, x_D)'$ leads in general to an $r$-part composition, $r < D$, whose parts are formed by amalgamized (combined, unified) parts of the original composition $\mathbf{x}$. For example, in case of household expenditures, the tobacco and alcohol parts can be amalgamated into a new part representing expenditures for both commodities. In that way, possible zero values in alcohol (for teetotal households) may disappear in the amalgamized part. Although amalgamation can reduce the amount of zero values, it is a nonlinear operation with respect to the Aitchison geometry, and also the information

on ratios between the corresponding compositional parts gets lost [13]. For this reason, such an operation can be carried out only in special cases with a clear interpretation of the amalgamated results.

The original approach to structural zeros replacement as described in [1, p. 269] (also called additive replacement), was accommodated by Fry *et al.* [21]. However, complications arise when the structural zeros are replaced by non-zero elements since small values lay on the boarder of the simplex; consequently, they result in outliers that have great impact to any distance calculation (Aitchison distance) and to log-ratio coordinates, and therefore they can influence the estimation of parameters. Non-zero replacement is fully reasonable in case of rounded zeros (e.g. correction of a previously imprecise small value) [26,28]. Structural zeros are considered as relevant correct values, and thus their replacement can lead to meaningless results. Typical examples are distance-based analyses of compositions, where the Aitchison distance

$$d_A(\mathbf{x},\mathbf{y}) = \sqrt{\frac{1}{D}\sum_{i=1}^{D}\sum_{j=i+1}^{D}\left(\ln\frac{x_i}{x_j} - \ln\frac{y_i}{y_j}\right)^2} \tag{4}$$

between compositions $\mathbf{x} = (x_1, \ldots, x_D)'$ and $\mathbf{y} = (y_1, \ldots, y_D)'$ is used. If a structural zero in a component is replaced by a small value, the resulting distance may become arbitrarily large. Consequently, zero replacement might lead to severe (artificial) outliers which become a serious problem for the analysis.

In [1], and in particular in [3,7,35], also parametric approaches to dealing with structural zeros were introduced. The principle is to build up a model in two stages; the first to determine where the zero entries occur in the data set (zero pattern structure), and the second to model the distribution of the unit available from the non-zero parts using a binomial conditional logistic normal model. Nevertheless, for such a modeling, the likelihood needs to be expressed explicitly which can lead to computational problems. Moreover, the derivation of the likelihood assumes the usual Euclidean geometry, that is not followed by the original compositions.

Finally, in many cases it is possible to interpret structural zeros in a certain component as indicators of two different subgroups of interest: observations with a zero value in the particular component versus observations taking on a positive value instead [27]. Aitchison and Kay [3] justify this with an example where households with non-smoking and non-(alcohol-) drinking members are forming a different household budget pattern, because they know in advance that they will spend nothing on these commodity groups and will allocate their expenditures over the remaining parts. Moreover, also the ratios among the parts can change: for teetotal households one may expect smaller relative expenditures on health than on drinking and smoking. From this perspective, it might seem that the only reasonable way is to split the data set according to the zero patterns and then analyze directly the resulting subcompositions. Although this is reasonable, in practical applications such an approach frequently leads to small sample sizes for the obtained subsets of observations, necessary to obtain relevant estimates for the parameters of the distributions. There is also a related question, whether differences among non-zero parts in different zero patterns are indeed so fundamental in practice, or whether they reflect rather purely the methodological viewpoint. Recent literature [27] states that non-zero data structure must

necessarily differ for different zero patterns. This was also introduced in [1]. However, in Section 4 we show that this assumption is not necessarily valid in practice.

Consequently, a natural question arises, whether a reasonable imputation of zero parts could be used just as an auxiliary step to get estimates of parameters (e.g. covariance) from the overall data set, i.e. by ignoring possible differences induced from the zero patterns. In the next step, the resulting estimates could then be used for an analysis in the subcompositions resulting from the single zero patterns. Such imputation should not add new information to the data structure (at least theoretically) and just aims to provisionally complete the data matrix in order to enable the estimation of parameters. Consequently, such requirement needs to be reflected by the corresponding imputation algorithm. Of course, the analysis itself (like outlier detection) should then be performed exclusively in subcompositions of non-zero parts. This seems to be in line also with [42, p. 225], who state that although imputation of structural zeros itself does not make sense, nevertheless,

> ... if we relax our aim and say that we would like to infer a value that would be observed if it could be defined, we implicitly assume the observations with structural zeros are similar to those fully observed, and thus the imputed value would follow the assumptions of MCAR (missing completely at random) .

The above possibility will be considered in the following as possible alternative to working directly in subcompositions.

## 3. Outlier detection with structural zeros

The most widely used methods for multivariate outlier detection are based on covariance estimates and Mahalanobis distances (MD). As possible alternatives not discussed in this paper, also other outlyingness measures could be considered [8,9]. Detected outliers are candidates for aberrant data that may otherwise adversely lead to model misspecification, biased parameter estimation and incorrect results. It is therefore important to identify them prior to modeling and analysis. Even in case of non-compositional data, outlier detection with data containing zeros is non-trivial, see [40].

Mahalanobis distances are computed for compositional data in ilr coordinates [18]. For a sample of coordinates $\mathbf{z}_1, \ldots, \mathbf{z}_n$, the MD is defined as

$$\mathrm{MD}(\mathbf{z}_i) = [(\mathbf{z}_i - \mathbf{t})' \mathbf{C}^{-1} (\mathbf{z}_i - \mathbf{t})]^{1/2}, \tag{5}$$

for $i = 1, \ldots, n$, where the $(D-1)$-dimensional vector $\mathbf{t}$ and the $(D-1) \times (D-1)$ matrix $\mathbf{C}$ stand for location and covariance estimators, respectively.

Mahalanobis distances cannot be directly computed from compositional data including structural zeros, because the ilr coordinates would result in data values of $\pm\infty$ which in turn cause problems for the (robust) estimation of $\mathbf{t}$ and $\mathbf{C}$. Based on the above considerations, there are two possible ways to avoid these difficulties: (a) compute Mahalanobis distances directly for subcompositions according to the zero patterns; (b) impute the structural zeros based on the available multivariate data information, estimate $\mathbf{t}$ and $\mathbf{C}$ based on the full (imputed) information and compute Mahalanobis distances with the respective blocks of these estimates for subcompositions with originally non-zero parts. Then it is possible to perform outlier detection (first step). In the second step, outliers related to the zero structure are analyzed using principal component analysis, adapted for binary data.

Finally, the information from the previous two steps is merged together by considering an appropriate interpretation of the results. The second option is described in detail in the following two sections, because (a) is just its trivial modification.

### 3.1. Imputation and zero structure for outlier detection

At first, we focus on outliers that result from the covariance structure of compositional data with non-zero parts. Following the above discussion, in order to enhance the computation of Mahalanobis distances, the zero-parts will be provisionally imputed in a way that no (substantial) information is added that would otherwise lead to an alteration of the overall multivariate structure of the non-zero compositions.

For computing the Mahalanobis distances (5), two key ingredients are required, (a) robust estimates of $\mathbf{t}$ and $\mathbf{C}$, and (b) appropriate ilr coordinates. The latter need to be based on the subcompositions corresponding to the particular zero structures of the data set (see Section 3.1.2 for details). The Mahalanobis distances for the data with the zero patterns make use of the particular ilr coordinates and of the corresponding blocks of $\mathbf{t}$ and $\mathbf{C}$.

#### 3.1.1. Robust estimation of location and covariance

As mentioned before, a standard location and covariance estimation is based only on the non-zero data information, while the information of the structural zeros is not used for the estimation.

Because the Mahalanobis distance given in Equation (5) itself can be influenced by outlying observations, robust versions of $\mathbf{t}$ and $\mathbf{C}$ are recommended. A popular choice is the minimum covariance determinant (MCD) estimator, which is defined by those $h$ observations (typically, $h \approx 3n/4$) that result in the smallest determinant of their sample covariance matrix. The location estimator $\mathbf{t}$ is the arithmetic mean of these $h$ observations, and the covariance estimator $\mathbf{C}$ is given by their sample covariance matrix, multiplied by a factor for consistency at normal distribution [33]. If $\mathrm{MD}^2(\mathbf{z}_i)$ exceeds a certain quantile $q$ of $\chi^2$-distribution with $D-1$ degrees of freedom, $\chi^2_{D-1;q}$ (usually, the quantile $q = 0.975$ is taken), the corresponding observation is flagged as potential outlier (see [18], for details).

The ilr data matrix, however, needs to be complete, which implies that the original data cannot have zero entries. All structural zeros thus first need to be replaced by positive entries. Here, we do not use zero replacement as e.g. given in [21], because we want to avoid that the replacement adds new multivariate data information. Therefore, we will use $k$ nearest neighbor imputation as described in [24], which uses the available multivariate data information for imputation. The idea of the method, originally introduced in [41], is to use the Aitchison distance (4), computed from non-zero components, for finding the $k$ most similar observations to a composition containing zeros, and to replace the zeros by using the available variable information of the neighbors. The $k$ nearest neighbor imputation is numerically stable (no iterative scheme is required), but it has also some limitations [24]. Particularly, the optimal number $k$ of nearest neighbors has to be determined based on simulation or resampling. Data with small sample size is a further limitation, since too few 'good' neighbors might be available. Nevertheless, practical experiences confirm that except of extreme cases with very low numbers of observations the multivariate data structure after $k$ nearest neighbor imputation will not be changed, and we will not create or hide multivariate outliers. Accordingly, in the following the default choice of $k = 3$ is taken.

With the completed data matrix, we can proceed with the robust estimation of location $\mathbf{t}$ and covariance $\mathbf{C}$ based on the ilr transformed data, by using for example the ilr coordinates (2), see [20].

### 3.1.2. Appropriate ilr coordinates

We consider a $D$-part composition $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD})'$, for $i = 1, \ldots, n$, where possibly some of the entries are structural zeros. Suppose that $\mathbf{x}_i$ has $D - K(i)$ structural zeros, $2 \leq K(i) \leq D - 1$. Then we will put the zeros on the first positions, resulting in $\tilde{\mathbf{x}}_i = (0, \ldots, 0, x_{ij_1}, \ldots, x_{ij_{K(i)}})'$. The cell $x_{ij_k}$ corresponds to the $k$th non-zero position in the vector $\mathbf{x}_i$, for $k \in \{1, \ldots, K(i)\}$. It is straightforward to find a permutation matrix $\tilde{\mathbf{P}}_i$ of dimension $D \times D$ with 0/1 entries, such that $\tilde{\mathbf{x}}_i = \tilde{\mathbf{P}}_i \mathbf{x}_i$, for $i = 1, \ldots, n$.

The idea behind the re-arrangement of the parts is to construct an ilr representation of the non-zero parts. For that purpose we can use Equation (2): an ilr coordinate $z_{ij_k}$ will describe all the relative information about the part $x_{ij_k}$ with respect to all 'subsequent' parts $x_{ij_l}$, with $l > k$. Therefore, the corresponding ilr coordinates $z_{ij_1}, \ldots, z_{ij_{K(i)-1}}$ contain all the relative information of $x_{ij_1}, \ldots, x_{ij_{K(i)}}$, for $i = 1, \ldots, n$. We will use these ilr coordinates in the following. Note that it would also be possible to use any other orthonormal log-ratio coordinates for the representation of $x_{ij_1}, \ldots, x_{ij_{K(i)}}$. The result of outlier detection based on Mahalanobis distances would be the same, because different ilr coordinates are just rotations of each other [15,18]. The present choice is followed with the aim not to further complicate the algorithm by introducing an unnecessarily general setting. Finally, note that this procedure coincides with the computation of the ilr coordinates on the subcompositions containing just the non-zero part.

### 3.1.3. Computing robust Mahalanobis distances on non-zero parts

In the previous Section 3.1.2, we have re-arranged the parts to construct an ilr representation of the non-zero information. Generally, when we consider a composition $\mathbf{x}$ and its ilr transformation $\mathbf{z}$ using Equation (2), then the permuted composition $\tilde{\mathbf{x}} = \tilde{\mathbf{P}}\mathbf{x}$ with a permutation matrix $\tilde{\mathbf{P}}$ has an ilr representation $\tilde{\mathbf{z}}$, where

$$\tilde{\mathbf{z}} = \mathbf{Q}'\mathbf{z} \tag{6}$$

and $\mathbf{Q}$ is an orthonormal matrix with $\mathbf{Q} = \mathbf{V}'\tilde{\mathbf{V}}$. The matrix $\mathbf{V}$ is defined in Equation (3), and $\tilde{\mathbf{V}} = \tilde{\mathbf{P}}\mathbf{V}$, see [16]. Of course, Equation (6) makes sense only when Equation (2) makes sense, i.e. with compositional data with positive entries. Moreover, if estimates $\mathbf{t}$ for location and $\mathbf{C}$ for covariance have been derived via ilr coordinates (2), then the corresponding estimates $\tilde{\mathbf{t}}$ and $\tilde{\mathbf{C}}$ for the new ilr representation are linked with the previous ones by $\tilde{\mathbf{t}} = \mathbf{Q}'\mathbf{t}$ and $\tilde{\mathbf{C}} = \mathbf{Q}'\mathbf{C}\mathbf{Q}$. These relations only hold if the corresponding location and covariance estimators are affine equivariant, which is the case for the MCD estimator, see [18].

The above relations are important for computing the Mahalanobis distances, because we want to extract only the information corresponding to the non-zero parts. Considering Section 3.1.2, we obtain for the $i$th composition the permutation matrix $\tilde{\mathbf{P}}_i$ and further a matrix $\mathbf{Q}_i = \mathbf{V}'\tilde{\mathbf{V}}_i = \mathbf{V}'\tilde{\mathbf{P}}_i\mathbf{V}$. Now we can use the relation (6) to define $\tilde{\mathbf{z}}_i = \mathbf{Q}_i'\mathbf{z}_i$, where $\mathbf{z}_i$ would be the ilr representation of $\mathbf{x}_i$ using the coordinates (2). According to Section 3.1.2, the last $K(i) - 1$ components of $\tilde{\mathbf{z}}_i$ include all the relative information about the non-zero entries of $\mathbf{x}_i$. We denote these components by $\tilde{\mathbf{z}}_i^*$, computed solely from the original (non-imputed) compositions. Similarly, we define $\tilde{\mathbf{t}}_i = \mathbf{Q}_i'\mathbf{t}$ and $\tilde{\mathbf{C}}_i = \mathbf{Q}_i'\mathbf{C}\mathbf{Q}_i$, and denote with $\tilde{\mathbf{t}}_i^*$

the last $(K(i)-1)$ components of $\tilde{\mathbf{t}}_i$, and with $\tilde{\mathbf{C}}_i^*$ the lower right $(K(i)-1)\times(K(i)-1)$ block of the matrix $\tilde{\mathbf{C}}_i$. Thus, $\tilde{\mathbf{t}}_i^*$ and $\tilde{\mathbf{C}}_i^*$ are robust location and covariance estimations of the non-zero parts of $\mathbf{x}_i$. According to Equation (5), the robust Mahalanobis distance used for outlier detection of the $i$th observation ($i=1,\ldots,n$) is then

$$\mathrm{MD}(\tilde{\mathbf{z}}_i^*) = [(\tilde{\mathbf{z}}_i^* - \tilde{\mathbf{t}}_i^*)'\tilde{\mathbf{C}}_i^{*-1}(\tilde{\mathbf{z}}_i^* - \tilde{\mathbf{t}}_i^*)]^{1/2}. \tag{7}$$

The composition $\mathbf{x}_i$ is flagged as potential outlier if $\mathrm{MD}^2(\tilde{\mathbf{z}}_i^*) > \chi^2_{K(i)-1;0.975}$.

### 3.2. Outliers related to the zero structure

To analyze the zero structure, the data are recoded into a binary matrix [3]; while values being non-zero are replaced with 1, the zeros remain untouched. The task is to analyze the zero pattern structure, i.e. all different arrangements of zeros occurring in (sets of) observations of the compositional data set at hand. Outliers then refer to atypical phenomena that occur rarely in the binary matrix of the zero patterns together with frequencies, arising from their occurrence in the data set. It means that also the information is taken into account how frequent each pattern is contained in the data set.

Two approaches for analyzing atypical patterns are briefly discussed.

Firstly, basic summary statistics of the zero patterns (basically counts) give an indication about the structure. The number of zeros in the variables and the combinations of zeros (zero patterns) and their frequencies can be graphically visualized (an example is shown in Figure 1).

Secondly, the multivariate structure and outlyingness of the zero patterns is analyzed. There are several possibilities for this task. One could think of computing Mahalanobis distances for binary data describing the zero patterns. The required 'covariance matrix' could be obtained by the classical Pearson-correlation as an equivalent to the binary correlation 'phi' [23]. Observations belonging to a specific zero pattern will have the same Mahalanobis distance, while outliers in the binary (zero) structure will have a large value of the Mahalanobis distance. However, here it is unclear which outlier threshold should be taken. Moreover, singularity of the covariance matrix occurs whenever two variables in the binary pattern have the same 0/1 values, or if two variables are fully observed. For these reasons, we do not follow this approach.

The above limitations can be overcome by using principal component analysis (PCA) for binary data [10,25,36]. PCA even allows to get deeper insight into the multivariate structure of the binary pattern, as it is useful generally in case of non-compositional or compositional data [2,19]. We employ the method proposed by de Leeuw [10], which builds disjoint convex regions. Each region corresponds to a profile, i.e. a vector of zeros and ones, and is separated by hyperplanes that have to be estimated. For the corresponding algorithm, a complex optimization problem must be solved. Distances between binary vectors and a predicted matrix defining the hyperplanes are given by a loss function; see [10] for details. Basically, majorization algorithms that iterate a sequence of weighted or unweighted singular value decompositions are employed to obtain loadings and scores, similar as in standard PCA. The interpretation of the loadings, scores and the resulting biplot is done in the same manner as in standard PCA. In the biplot, similar zero patterns will result in scores that

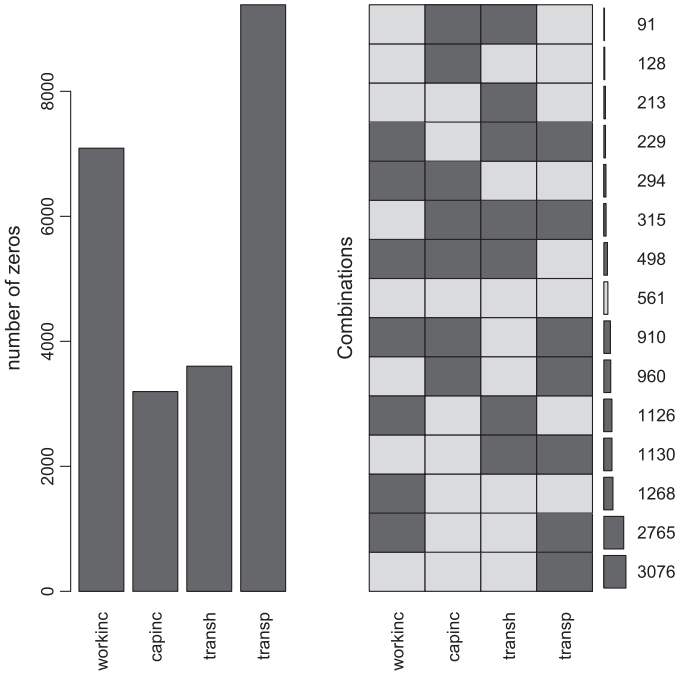


**Figure 1.** Zero structure of the Austrian EU-SILC data. LEFT: the amount of zeros for work income, capital income, household transfers and personal transfers. RIGHT: combination of zeros belonging to these four parts.

are located next to each other, and variables with similar patterns lead to similar loadings. The practical use of this method is further discussed in the next section.

## 4. Numerical examples

### 4.1. Austrian EU-SILC data set

The *European Union Statistics on Income and Living Conditions* (EU-SILC) is an annual panel household survey conducted in EU member states and most other European countries, and it serves as data basis for measuring risk-of-poverty and social cohesion in Europe [17]. Here, we consider the Austrian EU-SILC 2006 data set. However, since the original data from this survey are confidential, we use data simulated according to the methodology described in [5] and implemented in the R package `simPopulation` [30]. The data set with 14, 827 observations from 6 000 households and 28 variables (household information and various income components) is available in the R package `laeken` [4] as data set *eusilc*. Since the income components contain (too) many zeros, we amalgamate the parts according to Table 1 to obtain the four compositional parts *workinc* (work income), *capinc* (capital income), *transh* (household transfers), and *transp* (personal transfers).

For a complex data set, it is also interesting to analyze the structure of the zeros in the parts. Figure 1 is a first step in this direction: in the left plot the number of zeros in the parts is shown by barplots, and the right plot shows all combinations of zeros sorted

**Table 1.** Amalgamation of the income components.

| | | | | | | |
|---|---|---|---|---|---|---|
| workinc | = | PY010n *(employee cash or near cash income)* | + | PY050n *(cash benefit or losses from self-employment)* | | |
| capinc | = | [HY040n *(income from rental of a property or land)* | + | HY090n]/hhsize *(interests, dividends, profit from capital investments in unincorporated business)* | | |
| transh | = | [HY050n *(family/ children related allowances)* | + | HY110n *(income received by people aged under 16)* | + | HY070n + *(housing allowances)* |
| | + | HY080n *(inter-household cash transfers received)* | − | HY130n *(inter-household cash transfers paid)* | − | HY145n]/hhsize *(payments/receipts for tax adjustments)* |
| transp | = | PY090n *(unemployment benefits)* | + | PY110n *(survivor benefits)* | + | PY130n *(disability benefits)* |
| | + | PY100n *(old-age benefits)* | + | PY120n *(sickness benefits)* | + | PY140n *(education related allowances)* |

Notes: Variable names according to EUROSTAT definition.

according to their frequency in the data. A dark gray rectangle indicates zeros in the corresponding parts, a light gray rectangle represents non-zero data. In addition, the frequencies of the different combinations are represented by a small bar plot and by numbers. These plots are adapted from the methodology described in [37] to visualize missing values. For example, the bottom row in Figure 1 (right) represents compositions with only zeros in the last part (*transp*), in this case the majority of the observations. Also a lot of compositions have zeros in both parts, *workinc* and *transp*. The least frequent combination is displayed in the top row: zeros that are present in both parts *capinc* and *transh*. These findings can be further helpful when considering results of principal component analysis of binary data.

Note that outlier detection based on the covariance structure of the compositions can only be performed if at least two parts do not contain structural zeros. Otherwise, the whole to which the composition refers would be degenerated (the case of all zero parts), or trivial (given by the only one non-zero part). There is even a conceptual conflict: If we assume that the (log-)ratios form the relevant information of the compositional data, and if (log-)ratios cannot be computed, then the whole approach would be inappropriate. According to Figure 1, we would have to exclude $229+315+498+910=1952$ compositions, approximately 17% of the observations that containing two or more non-zero values. This would result in a substantial loss of information. In contrast, for the analysis of the zero pattern structure using PCA it is not necessary to exclude any observation.

The first step of the analysis consists of measuring the influence of imputation to the covariance estimation in the subcompositions formed by the different zero patterns. This is evaluated by comparing the Mahalanobis distances obtained by (a) estimating the covariance matrix of the whole (imputed) data and subsequent outlier detection by considering the non-zero parts only (see Section 3.1.3), to (b) the Mahalanobis distances estimated from subcompositions/subsets of variables and observations (no common covariance). If both types of Mahalanobis distances are the same for the observations corresponding to

a zero pattern, we can assume that their covariance structure corresponds to the overall covariance structure of the data set – otherwise the zero pattern has a different multivariate data structure. It should be mentioned that a difference in both approaches might also occur due to small sample sizes in approach (b). In that case one could expect high variability of the estimates of **t** and **C**, causing instabilities in the corresponding Mahalanobis distances.

Figure 2 compares the two approaches of estimating Mahalanobis distances. On the *x*-axis the approach in Section 3.1.3 is mapped, while for the *y*-axis the covariances and resulting Mahalanobis distances are estimated only on the subcompositional level. The zero structure is reported in the headers where the corresponding zero patterns are indicated by 0 and x; e.g. *00xx* means that all the observations that have zeros in the first two components are considered. The Mahalanobis distances are normalized using the corresponding cut-off values of the $\chi^2$ distribution (or from a standard normal in the one-dimensional case). Thus outliers are visible in the plot when they exceed the threshold 1 (dashed lines).

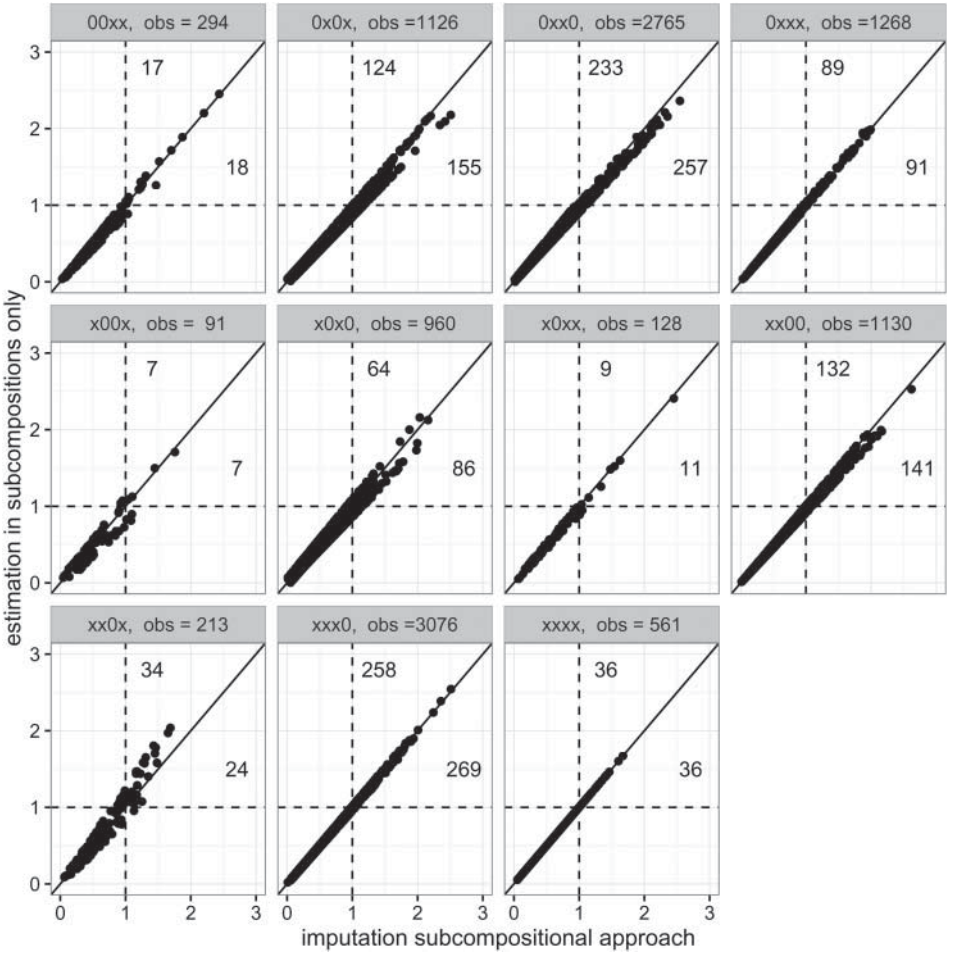


**Figure 2.** Comparing Mahalanobis distances obtained from the imputation approach and from the estimation in subcompositions applied to the EU-SILC data.

The numbers of outliers are reported for each of the two methods. E.g. for zero pattern *00xx*, 15 outliers are reported for the subcompositional estimation and 17 for the imputation approach. The results are quite comparable – the Mahalanobis distances are almost the same since they lay on the bisector. Using the imputation approach and a common covariance matrix for all observations has the advantage that the sample size is much larger and thus the covariance estimation is more stable than considering the subcompositions only.

Figure 2 also shows that results from all patterns are similar. The largest deviations can be seen in patterns '`x00x`' (zeros in *capinc* and *transh*) and '`xx0x`' (zeros in *transh* only) but also these results reveal only slightly larger deviations for both methods. An explanation is the relatively small sample size in these groups.

Figure 3(a) presents the results of PCA applied to the zero patterns in form of a scores and loadings plot of the first two PCs. These plots can be jointly interpreted in the sense of the standard covariance biplot [22], just by considering the specific structure of the observations that are processed by the PCA. For example, the variable household transfer (*transh*) points to the upper left of the (loadings) plot, which means that patterns with observed values in this variable (indicated by *x*) are located in the upper left, and patterns with zeros in *transh* are in the opposite direction. According to the configuration in the loadings plot, the patterns referring to *transh* and capital income (*capinc*) show similar behavior, which is also visible in Figure 1. In contrast, *transh* and *transp* (personal transfer) point at very different directions in the loadings plot, also the direction for capital income (*capinc*) is very different, and thus the occurrence of zeros in these variables is rather independent from each other.

There is no clear outlier visible in the scores plot (Figure 3(a), left), i.e. none of the zero patterns shows extreme behavior. There are just some atypical patterns that tend to be located further from the origin. For example, the pattern *x00x* – the pattern expressed by observed positive values in the first and fourth variable, zeros in the second and third variable – is in the bottom right of the plot, further away from the origin, and this is the pattern which occurs only 91 times, see Figure 1.

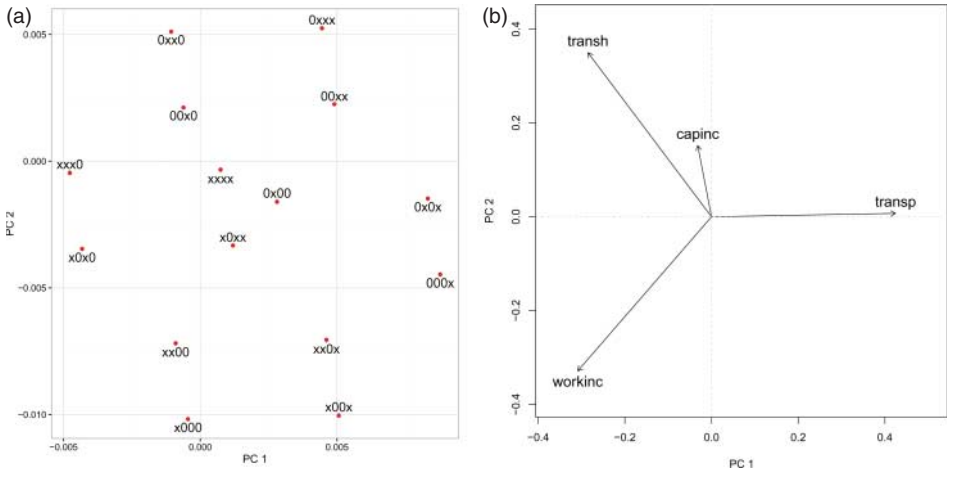


**Figure 3.** Results from the binary PCA for the EU-SILC data: (a) Scores plot. Scores are printed with their group level (zero pattern) whereas *x* stands for observed. (b) Loadings plot.

### 4.2. Consumption data from Albania

The household expenditures from the year 2008 of Albania were provided by the World Bank (see also http://datatopics.worldbank.org/consumption/). The data were obtained through a survey, and the participants were asked about their household consumptions over a given time horizon in various spending categories. These categories range from different kinds of food-products over general living expenditures such as gas, electricity or water to expenses for education, health and others. The number and type of categories differ from each survey but have in common that the combined categories reflect the whole consumption of a household for this given time horizon. The Albanian household survey consists of 3600 households, including 14,785 individuals. Here, we analyze the seven major parts, namely household consumptions on

- food and non-alcoholic beverages (*food*)
- alcoholic beverages, tobacco and narcotics (*alcohol*)
- clothing and footwear (*clothing*)
- housing, water, electricity, gas and other fuels (*housing*)
- communication (*commun.*)
- education (*education*)
- miscellaneous goods and services (*misc*).

The amount of zeros varies between almost 0% until almost 10% per variable, see Figure 4. Out of the total amount of households, 2903 observations include not any zero, 260 observations have zeros in variable alcohol only. There are several zero patterns which

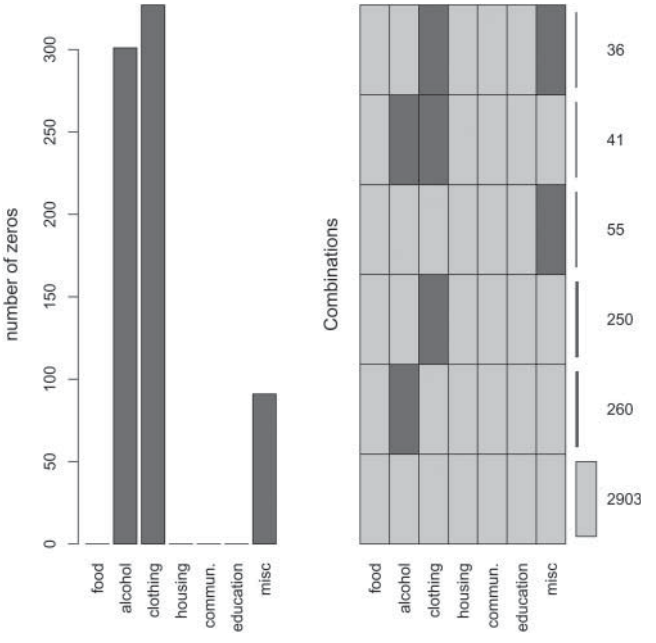


**Figure 4.** Amount of zeros per variable and zero patterns for the Albanian consumption data.

occur only once or a few times. For these patterns, it will not be possible to do outlier detection based on their individual covariance estimation in the subcomposition, but rather we rely on imputation and on a joint covariance estimation.

For the comparison of the Mahalanobis distances, we therefore use only groups of reasonable size, e.g. the six largest groups. We see from the resulting Figure 5 that the Mahalanobis distances from the imputation approach are comparable with those from the direct estimation of the covariance in the subcompositions. Also the number of identified outliers is very similar in both approaches, although not necessarily the same outliers are detected. A more pronounced difference is only in the pattern *x00xxxx*, referring to zeros in *alcohol* and *clothing*. These might be atypical households in any case, with a different data structure also in the observed variables. The occurrence of the pattern could also be caused by data collection problems.

Finally, Figure 6(a) and 6(b) shows the scores and loadings plots from a PCA on the zero patterns. The loadings reveal that *housing* and *education*, as well as *misc* and *clothing* are strongly connected in the zero patterns. The scores plot shows some deviating patterns which are far from the origin of the plot. For example, the pattern *x0xx0xx* on the top left occurs only once, and it is the only pattern with zero in communication. The patterns on

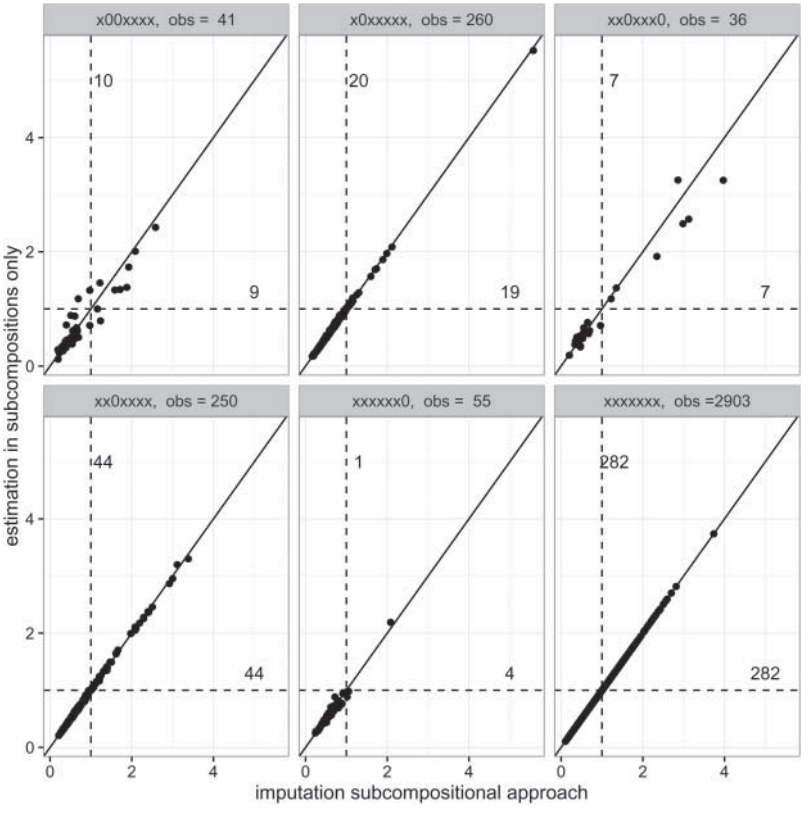


**Figure 5.** Comparing Mahalanobis distances obtained from the imputation approach with the direct estimation in subcompositions applied to the household consumption data.

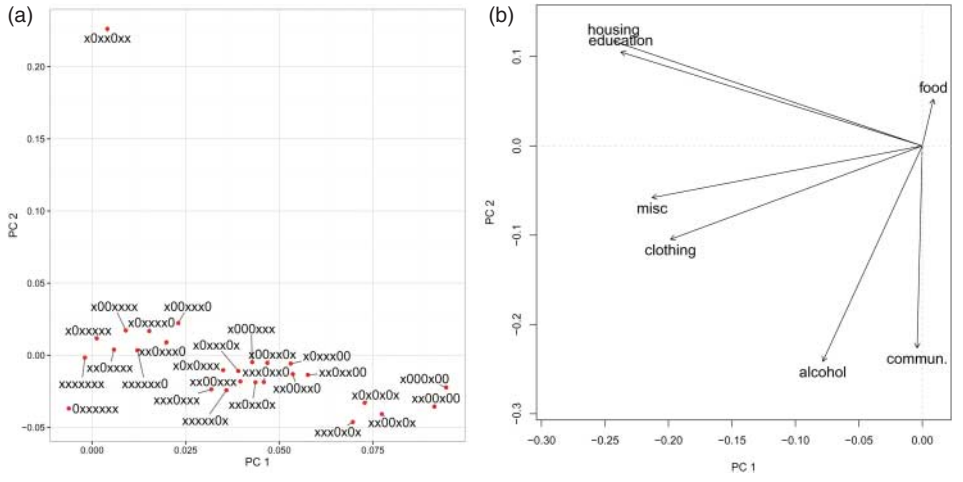


**Figure 6.** Results from the binary PCA for the household consumption data: (a) Scores plot. Scores are printed with their group level (zero pattern). (b) Loadings plot.

the bottom right are also rare in the data set, and they all have zeros in *housing*. One must be aware that these 'outlying' patterns are rather related to data artifacts from the survey, and the zeros should better be replaced by sensible values (or the observation should be deleted in order to avoid biased estimates). In this sense, the analysis of the zero patterns is a valuable part of the overall analysis.

## 5. Summary and outlook

In this paper we focused on outlier detection for compositions with structural zeros. In fact, outlier detection is a quite relevant task which is daily routine in statistical offices when specific data sets are checked for plausibility. Moreover, data sets as used in our examples are the basis for political decisions, and outliers might have an effect on derived indicators supporting these decisions (see [11] and the outliers in consumption data project of the International Household Survey Network, http://www.ihsn.org/home/projects/outliers). The usual procedure then is to 'correct' implausible data values, or to reduce the effect of an outlier in the statistical estimation procedure. For statistical estimation, however, the compositional nature of the data needs to be accounted for. Structural zeros create difficulties for compositional data analysis already in the definition of compositional data, since they are supposed to consist of strictly positive entries [12]. Standard transformations and methods from a log-ratio approach to compositional data analysis thus cannot be used.

However, they could be applied to the sub-populations formed by the observations of the different zero patterns. While this might be a solution for rather simple data sets, more realistic data sets as those presented in this paper prevent this approach, simply because of the large number of zero patterns, and the possibly small sample sizes for some patterns. For this reason, we focused on exploratory tools to treat the problem of structural zeros in compositional data sets. The proposed heuristic procedure introduced tools that allow

to identify atypical observations, or atypical patterns of occurrences of zeros. It can be summarized by the following four steps:

(1) The overall covariance structure of the compositional data set is estimated by provisional imputation of structural zeros. Since the non-zero parts are unchanged, their covariance structure should be well preserved.
(2) Location and covariance estimates based on the full (imputed) information are used to compute Mahalanobis distances with the respective blocks of these estimates for subcompositions with originally non-zero parts. If the sample size allows, the results are compared with those based solely on subcompositions from single non-zero patterns.
(3) The zero structure is analyzed by recoding the data into a binary matrix and employing appropriate statistical tools (particularly, PCA for binary data).
(4) Observations flagged as outliers in points 2. or 3. are considered as deviating data points with respect to the covariance and/or zero structure, respectively.

Standard tools for outlier detection are based on robust Mahalanobis distances, where robust estimates of location and covariance are plugged in. In the proposed exploratory plot, we compare robust Mahalanobis distances computed for (a) the subcompositions and data subsets which have no zeros, with those computed for (b) the overall covariance structure, where the zeros are imputed first. For (a), every zero pattern needs to be considered separately, and it can happen that (robust) covariance estimation is not feasible due to too small sample size. For example, when using the MCD estimator, at least twice as many observations than variables are required [33]. In case of (b), the imputation step is done robustly [28], and since the available data information is employed when replacing the zero values, the multivariate data structure of the available information is not changed with the replacement step. This imputation is thus only a convenience in order to be able to estimate the joint covariance matrix. In our examples it turned out that both types of Mahalanobis distances are very similar for most zero patterns, and also the flagged outliers are essentially the same. If differences occur in a zero pattern, we conclude that the multivariate data structure of this pattern is different from the overall data structure. Differences, however, could also occur due to small sample sizes of the data groups forming those zero patterns, since the covariance estimation may be instable in that case. There is a further limitation for outlier detection based on Mahalanobis distances, concerning the necessity of having at least two non-zero parts in a composition. This is a direct consequence of using the log-ratio methodology and the concept of compositional data in general for structural zeros.

The exploratory analysis of the zero patterns is another important step to get an overall picture. The matrix for this analysis is constructed by replacing the non-zero information in the data set by values of one. As a result, binary information of the different arrangements of zeros in the compositional parts of the data set is retrieved, and these 'patterns' are analyzed. There are not too many methods available for analyzing this binary information, and we decided to employ a rather complex nonlinear PCA methodology. This analysis is focused on a different aspect as before: here, not the outlyingness of single observations, but the structure of the zero patterns and information about their 'outlyingness' is revealed. This analysis, together with the information of the frequencies of the individual patterns in the data, allows to identify systematic occurrences of zeros in different parts, and also

patterns where the zeros are probably not structural zeros but simply data errors. In that case, imputation would indeed make sense, especially if the data structure of the non-zero information is consistent with the overall data structure. This is in fact investigated in the first step of the analysis.

One goal of the paper was to provide possibilities to deal with compositions containing structural zeros, as they frequently occur in practice. Besides the suggested exploratory tools for outlier detection, we also provide R code for the computations included in the R package robCompositions [38,39]; see https://github.com/matthias-da/robCompositions for the newest version.

Outlier detection may be of primary interest to the investigator, but also further analysis may be desirable. Since outlier detection already involves the (robust) pattern-individual and joint covariance estimation, it is straightforward to continue with other multivariate analysis methods which are based on the estimated covariance matrices.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

This work was supported by the COST Action CRoNoS IC1408, the Internal Grant Agency of the Palacky University in Olomouc [IGA_PrF_2015_013, IGA_PrF_2016_025], the Austrian Science Fund (FWF) [I 1910-N26], and by the K-project DEXHELPP through COMET - Competence Centers for Excellent Technologies supported by BMVIT, BMWFI and the province Vienna, administrated by the Austrian Research Promotion Agency (FFG).

## References

[1] J. Aitchison, *The Statistical Analysis of Compositional Data*, Chapman & Hall, London, 1986.
[2] J. Aitchison and M. Greenacre, *Biplots of compositional data*, J. Appl. Stat. 51 (2002), pp. 375–392.
[3] J. Aitchison and J. Kay, Possible solutions of some essential zero problems in compositional data analysis. pp. 1–6. Available at http://ima.udg.edu/Activitats/CoDaWork03/paper_Aitchison_and_Kay.pdf.
[4] A. Alfons and M. Templ, *Estimation of social exclusion indicators from complex surveys: The R package laeken*, J. Statist. Softw. 54 (2013), pp. 1–25.
[5] A. Alfons, S. Kraft, M. Templ, and P. Filzmoser, *Simulation of close-to-reality population data for household surveys with application to EU-SILC*, Statist. Methods Appl. 20 (2011), pp. 383–407.
[6] J. Bacon-Shone, *Discrete and continuous compositions*, in *CoDaWork'08*, Universitat de Girona. Departament d'Informática i Matemática Aplicada, 2008, p. 11.
[7] A. Butler and C. Glasbey, *A latent Gaussian model for compositional data with zeros*, J. Appl. Stat. 57 (2008), pp. 505–520.
[8] F. Chebana and T. Ouarda, *Depth-based multivariate descriptive statistics with hydrological applications*, J. Geophys. Res: Atmos. 116 (2011), pp. 1–19.
[9] X. Dang and R. Serfling, *Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties*, J. Stat. Plan. Inference 140 (2010), pp. 198–213.
[10] J. de Leeuw, *Principal component analysis of binary data by iterated singular value decomposition*, Comput. Stat. Data Anal. 50 (2006), pp. 21–39.
[11] O. Dupriez, *Building a household consumption database for the calculation of poverty ppps*, Technical note, World Bank, 2007, Available at http://siteresources.worldbank.org/ICPINT/Resources/270056-1195253046582/Dupriez_BuildingaHHCdatabasefortheCalculationofPovertyPPPs_Mar07.pdf.

[12] JJ. Egozcue, *Reply to 'On the Harker variation diagrams; . . .' by J.A. Cortés*, Math. Geosci. 41 (2009), pp. 829–834.
[13] JJ. Egozcue and V. Pawlowsky-Glahn, *Groups of parts and their balances in compositional data analysis*, Math. Geol. 37 (2005), pp. 795–828.
[14] J. Egozcue and V. Pawlowsky-Glahn, *Compositional Data Analysis in the Geosciences: From theory to Practice*, chap. Simplicial geometry for compositional data, Geological Society, London, 2006, pp. 145–160, special Publications 264.
[15] JJ. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal, *Isometric logratio transformations for compositional data analysis*, Math. Geol. 35 (2003), pp. 279–300.
[16] J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal, *Compositional Data Analysis: Theory and Applications*, Elem. Simplicial Linear Algebra Geometry. Wiley, Chichester, 2011, 139–145.
[17] Eurostat, *Description of target variables: Cross-sectional and longitudinal*, EU-SILC 065/04, Unit E-2: Living conditions, Directorate E: Social and regional statistics and geographical information system, Eurostat, Luxembourg, 2004.
[18] P. Filzmoser and K. Hron, *Outlier detection for compositional data using robust methods*, Math. Geosci. 40 (2008), pp. 233–248.
[19] P. Filzmoser, K. Hron, and C. Reimann, *Principal component analysis for compositional data with outliers*, Environmetrics 20 (2009), pp. 621–632.
[20] P. Filzmoser, K. Hron, and C. Reimann, *Interpretation of multivariate outliers for compositional data*, Comput. Geosci. 39 (2012), pp. 77–85.
[21] JM. Fry, TR. Fry, and KR. McLaren, *Compositional data analysis and zeros in micro data*, Appl. Econom. 32 (2000), pp. 953–959, Available at http://www.tandfonline.com/doi/abs/10.1080/000368400322002.
[22] K.R. Gabriel, *The biplot – graphic display of matrices with application to principal component analysis*, Biometrika 58 (1971), pp. 453–467.
[23] J. Guilford, *Psychometric Methods*, McGraw-Hill series in psychology, McGraw-Hill, New York City, 1954.
[24] K. Hron, M. Templ, and P. Filzmoser, *Imputation of missing values for compositional data using classical and robust methods*, Comput. Statist. Data Anal. 54 (2010), pp. 3095–3107.
[25] S. Lee, JZ. Huang, and J. Hu, *Sparse logistic principal components analysis for binary data*, Ann. Appl. Stat. 4 (2010), pp. 1579–1601, Available at http://dx.doi.org/10.1214/10-AOAS327.
[26] JA. Martín-Fernández, C. Barceló-Vidal, and V. Pawlowsky-Glahn, *Dealing with zeros and missing values in compositional data sets using nonparametric imputation*, Math. Geol. 35 (2003), pp. 253–278.
[27] J. Martín-Fernández, J. Palarea-Albaladejo, and R. Olea, *Compositional Data Analysis: Theory and Applications*, Dealing with Zeros, Wiley, Chichester, 2011, 43–58.
[28] JA. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser, and J. Palarea-Albaladejo, *Model-based replacement of rounded zeros in compositional data: Classical and robust approaches*, Comput. Statist. Data Anal. C 56 (2012), pp. 2688–2704.
[29] J. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser, and J. Palarea-Albaladejo, *Bayesian-multiplicative treatment of count zeros in compositional data sets*, Stat. Model. 15 (2015), doi:10.1177/1471082X14535524.
[30] B. Meindl, M. Templ, A. Alfons, and A. Kowarik, *simPop: Simulation of Synthetic Populations for Survey Data Considering Auxiliary Information*, 2015, Available at http://CRAN.R-project.org/package = simPop, R package version 0.2.9.
[31] V. Pawlowsky-Glahn and A. Buccianti, *Compositional Data Analysis: Theory and Applications*, Wiley, Chichester, 2011.
[32] V. Pawlowsky-Glahn, J. Egozcue, and R. Tolosana-Delgado, *Modeling and Analysis of Compositional Data*, Wiley, Chichester, 2015.
[33] P. Rousseeuw and K. von Driessen, *A fast algorithm for the minimum covariance determinant estimator*, Technometrics 41 (1999), pp. 212–223.
[34] JL. Scealy and AH. Welsh, *Regression for compositional data by using distributions defined on the hypersphere*, J. R. Stat. Soc. Ser. B. Stat. Methodol. 73 (2011), pp. 351–375.

[35] C. Stewart and C. Field, *Managing the essential zeros in quantitative fatty acid signature analysis*, J. Agric. Biol. Environ. Stat. 16 (2010), pp. 45–69.
[36] F. Tang and H. Tao, *Binary principal component analysis*, *In Proc. British Machine Vision Conference, Volume I*, 2006, pp. 377–386.
[37] M. Templ, A. Alfons, and P. Filzmoser, *Exploring incomplete data using visualization techniques*, Adv. Data Anal. Classif. 6 (2012), pp. 29–47.
[38] M. Templ, K. Hron, and P. Filzmoser, *robCompositions: An R-package for robust statistical analysis of compositional data, in Compositional Data Analysis: Theory and Applications*, V. Pawlowsky-Glahn and A. Buccianti, eds., Wiley, Chichester, 2011, pp. 341–355.
[39] M. Templ, K. Hron, and P. Filzmoser, *Robust Estimation for Compositional Data*, 2015, Available at https://github.com/matthias-da/robCompositions, R package version 1.9.2.
[40] V. Todorov, M. Templ, and P. Filzmoser, *Detection of multivariate outliers in business survey data with incomplete information*, Adv. Data Anal. Classif. 5 (2011), pp. 37–56.
[41] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and RB. Altman, *Missing value estimation methods for dna microarrays*, Bioinformatics 17 (2001), pp. 520–525.
[42] K. van den Boogaart and R. Tolosana-Delgado, *Analyzing Compositional Data with R*, Springer, Heidelberg, 2013.
[43] H. Wang, Q. Liu, HMK. Mok, L. Fu, and W. Man Tse, *A hyperspherical transformation forecasting model for compositional data*, Eur. J. Oper. Res. 179 (2007), pp. 459–468.