Covariance-Based Outlier Detection for Compositional Data with Structural Zeros: Application to Italian Survey of Household Income and Wealth Data

Gianna S. Monti, Karel Hron, Peter Filzmoser and Matthias Templ

Abstract Outlier detection is an important task for the statistical analysis of multivariate data, because often the outliers contain important information about the data structure. In compositional data, represented usually as proportions subject to a unit sum constraint, the ratios between the parts (variables) contain the essential information. This inherent property is, however, incompatible with the presence of zeros in compositions. Here we consider structural zeros, i.e., zeros that are truly observed, and not zeros related to measurement errors (rounded zeros). In order to identify possible outliers in compositional data with structural zeros, we apply the Mahalanobis distance approach, where the key task is a robust estimation of the covariance matrix. This resulting outlier detection procedure is applied to the Italian Survey of Household Income and Wealth (SHIW) data, collected by the Bank of Italy.

Key words: compositional data, missing values, Mahalanobis distance, robust statistics, SHIW data

Gianna S. Monti
Department of Economics, Management and Statistics, University of Milano-Bicocca, e-mail: gianna.monti@unimib.it

Karel Hron
Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, and Department of Geoinformatics, Faculty of Science, Palacký University, tř. Svobody 26, Olomouc, Czech Republic, e-mail: hronk@seznam.cz

Peter Filzmoser and Matthias Templ
Department of Statistics and Probability Theory, Vienna University of Technology, e-mail: p.filzmoser@tuwien.ac.at, templ@statistik.tuwien.ac.at
1 Introduction

The sample space of compositional data (or compositions, for short) is the simplex

$$\mathcal{S}^D = \{ x = (x_1, \ldots, x_D), x_i > 0, \sum_{i=1}^{D} x_i = \kappa \},$$

where $\kappa$ is a constant, usually taken to be one. The simplex has a Euclidean vector space structure of dimension $D - 1$, represented by the Aitchison geometry (Egozcue and Pawlowsky-Glahn, 2006). Standard statistical analysis techniques for standard (unconstrained) data are not appropriate for compositions, since they do not take into account their specific properties (scale invariance, relative scale). For this reason, the log-ratio transformations from the simplex to the real space were introduced (Aitchison, 1986). In most cases, the isometric log-ratio (ilr) transformation from $\mathcal{S}^D$ to $\mathbb{R}^{D-1}$ (Egozcue et al., 2003) is appropriate. It results in orthonormal coordinates $z = (z_1, \ldots, z_{D-1})$, which can be defined by

$$z_j = \sqrt{\frac{D - j}{D - j + 1}} \ln \sqrt{\frac{x_j}{\prod_{k=j+1}^{D} x_k}}, \quad j = 1, \ldots, D - 1. \quad (1)$$

The log-ratio approach is incompatible with the presence of zeros in the compositions. There are essentially two types of zeros in a composition: the rounded zeros and the structural zeros. While the rounded zeros are related to measurement errors, the latter correspond to parts that are completely absent in the composition. 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.

This contribution is focused on outlier detection in presence of structural zeros in compositional data. In Section 2 we introduce the Mahalanobis distance approach to outlier detection, proposed in Hron et al. (2013). In Section 3, an application of the theoretical developments to the Italian Survey of Household Income and Wealth (SHIW) data is presented.

2 Outlier detection methods in presence of structural zeros in compositional data

There are essentially two ways of dealing with structural zeros in compositional data (Martín-Fernández et al., 2012): amalgamation of the parts that contain zeros, with the consequence of reducing the dimensionality of the compositional data, or the replacement technique, which consists in a non-zero replacement of zero parts. However, amalgamation is a non-linear operation with respect to the Aitchison ge-
ometry, i.e. it does not preserve the distances. The replacement technique is suitable for rounded zeros, but it is inappropriate in case of structural zeros.

Outlier detection is commonly based on the Mahalanobis distance (MD), computed for the ilr transformed data (Filzmoser and Hron, 2008). For a sample of ilr transformed compositions \( z_1, \ldots, z_n \) the MD is defined as

\[
\text{MD}(z_i) = [(z_i - t)'C^{-1}(z_i - t)]^{1/2},
\]

where the \((D - 1)\)-dimensional vector \( t \) is the location estimator, and the \((D - 1) \times (D - 1)\) matrix \( C \) stands for the covariance estimator. Robust versions of \( t \) and \( C \) are recommended in order to protect the resulting estimates from the influence of outlying observations. A popular choice for a robust location and covariance estimator is the Minimum Covariance Determinant (MCD) estimator, which is defined by those \( h \) observations (typically \( h \approx 3n/4 \)) that results in the smallest determinant of their sample covariance matrix. The location estimator \( t \) is the arithmetic mean of these \( h \) observations, and the covariance estimator \( C \) is given by their sample covariance matrix, multiplied by a factor for consistency at normal distribution (Rousseeuw and von Driessen, 1999). If \( \text{MD}^2(z_i) \) 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 Filzmoser and Hron, 2008, for details). In presence of structural zeros in compositional data, the MD cannot be computed. In order to avoid this failure, three outlier detection methods are considered: zero replacement (the standard approach), and the novel imputation and the variation matrix methods, briefly described in the following.

### 2.1 Zero replacement

Fry et al. (2000) suggested a zero replacement method for dealing with structural zeros by modifying the original approach of Aitchison (1986). Here the zeros are replaced by small numbers, the non-zero parts are modified by adding an error component, and then outlier detection is performed. For details, see Fry et al. (2000); Hron et al. (2013).

### 2.2 Imputation method

This method extracts the information contained in the non-zero parts of the compositional data to identify the outlying observations. For this purpose, robust estimates of \( t \) and \( C \), based on the MCD estimator, are derived using a complex algorithm, and the ilr variables of the non-zero information of each compositional observation is involved for outlier detection. Details are described in Hron et al. (2013).
2.3 Variation matrix method

This approach is based on location and covariance estimation computed using only the available positive data entries. For this purpose, the geometric mean and a robust version of the variation matrix (Aitchison, 1986) are employed, see again Hron et al. (2013) for details.

3 Analysis of the Italian SHIW data set

The above introduced outlier detection methods are applied to the Italian Survey of Household Income and Wealth (SHIW) data, collected by the Bank of Italy (Bank of Italy, 2009). The sample size of the data set is about 8,000 households and consists of repeated cross-sections. We examine the 2008 survey year, referred to the collection period January to September 2009. Here we analyze the relative structure of household income that consists of different sources of the income: payroll and self-employment income, pensions, transfers, and property income of household members. Since the income components contain a lot of zeros, we amalgamate the parts according to Table 1 to obtain the following four compositional parts: pay (payroll income), pens (pensions and net transfers), self (net self-employment income), and prop (property income). The structure of zero parts is plotted in Figure 1 (Templ et al., 2012).

Table 1  Amalgamation of the income components.

<table>
<thead>
<tr>
<th>pay</th>
<th>=YL1</th>
<th>=YL2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Net wages and salaries)</td>
<td>(Fringe benefits)</td>
</tr>
<tr>
<td>pens</td>
<td>=YTP</td>
<td>=YTA</td>
</tr>
<tr>
<td></td>
<td>(Pensions and arrears)</td>
<td>(Other transfers)</td>
</tr>
<tr>
<td>self</td>
<td>=YMA1</td>
<td>=YMA2</td>
</tr>
<tr>
<td></td>
<td>(Self-employment income)</td>
<td>(Entrepreneurial income)</td>
</tr>
<tr>
<td>prop</td>
<td>=YCA</td>
<td>=YCF</td>
</tr>
</tbody>
</table>
|           | (Income from real-estate) | (Income from financial as-
|           |                           | sets)                     |

Notes: Variable names according to SHIW definition.

The three outlier detection methods are applied to the income amalgamated components of the SHIW data set. Note that compositions with less than two parts without zeros are omitted from the analysis. In Table 2, the numbers and percentages of identified outliers are presented, according to different levels of the external variable employment status of the head of the household. The different groups (levels) are: 1=blue-collar worker, 2=office worker or school teacher, 3=cadre or manager,
Fig. 1 Zero structure of the income component in the Italian SHIW data. LEFT: the proportions of zeros for Net wages and salaries (pay), Pensions and arrears (pens), Self-employment income (self), Income from real-estate (prop). RIGHT: combination of zeros belonging to these four parts.

4=sole proprietor/member of the arts or professions, 5=other self-employed, 6=pensioner, 7=other not-employed. The method of Fry et al. (2000) identifies more than 40% of the observations as outliers that tends to be rather unrealistic. In the group 5 there are even 100% outliers. The other two methods show less differences in the identification of outliers. The rather high numbers of identified outliers can be explained by possible sub-populations in the data, which are not distinguished by this approach. Moreover, outliers do not automatically contain “wrong” data values, but they rather refer to inconsistencies in the overall multivariate data structure.

Acknowledgements The authors gratefully acknowledge the support by the Operational Program Education for Competitiveness - European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic).
Table 2 Number and percentage of identified outliers in different groups of employment status of the head of the household.

<table>
<thead>
<tr>
<th>Employment status, group</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of observations</td>
<td>7247</td>
<td>1093</td>
<td>1017</td>
<td>265</td>
<td>287</td>
<td>399</td>
<td>3567</td>
</tr>
<tr>
<td>Rows with zeros</td>
<td>100</td>
<td>15.08</td>
<td>14.03</td>
<td>3.66</td>
<td>3.96</td>
<td>5.51</td>
<td>49.22</td>
</tr>
<tr>
<td>Rows as percentages</td>
<td>7085</td>
<td>1078</td>
<td>998</td>
<td>259</td>
<td>261</td>
<td>371</td>
<td>3508</td>
</tr>
<tr>
<td>Fry</td>
<td>3037</td>
<td>527</td>
<td>368</td>
<td>85</td>
<td>286</td>
<td>399</td>
<td>1061</td>
</tr>
<tr>
<td>Rows with zeros</td>
<td>41.91</td>
<td>48.22</td>
<td>36.18</td>
<td>32.08</td>
<td>99.65</td>
<td>100.00</td>
<td>29.74</td>
</tr>
<tr>
<td>Rows as percentages</td>
<td>1689</td>
<td>427</td>
<td>269</td>
<td>65</td>
<td>55</td>
<td>99</td>
<td>574</td>
</tr>
<tr>
<td>Variation matrix</td>
<td>1694</td>
<td>381</td>
<td>221</td>
<td>53</td>
<td>59</td>
<td>91</td>
<td>686</td>
</tr>
<tr>
<td>Rows with zeros</td>
<td>23.31</td>
<td>34.86</td>
<td>21.73</td>
<td>20.00</td>
<td>20.56</td>
<td>22.81</td>
<td>19.23</td>
</tr>
<tr>
<td>Rows as percentages</td>
<td>1694</td>
<td>381</td>
<td>221</td>
<td>53</td>
<td>59</td>
<td>91</td>
<td>686</td>
</tr>
</tbody>
</table>

References


