Exploring outliers in compositional data with structural zeros

K. Hron[1], M. Templ[2], P. Filzmoser[3]

[1]Department of Mathematical Analysis and Applications of Mathematics - Palacký University, Czech Republic
[2]Institute of Data Analysis and Process Design - Zurich University of Applied Sciences, Switzerland
[3]Institute of Statistics and Mathematical Methods in Economics - Vienna University of Technology, Austria

ERCIM 2016, December 10, 2016

Compositional data

= *D-part vectors, describing quantitatively the parts of some whole, which carry exclusively relative information between the parts* (Aitchison, 1986; Pawlowsky-Glahn et al., 2015)

Compositional data

= *D-part vectors, describing quantitatively the parts of some whole, which carry exclusively relative information between the parts* (Aitchison, 1986; Pawlowsky-Glahn et al., 2015)

- **usual units of measurement**: percentages, mg/kg (*constant sum constraint*), mg/l (*constant sum does not occur*)

Compositional data

= *D-part vectors, describing quantitatively the parts of some whole, which carry exclusively relative information between the parts* (Aitchison, 1986; Pawlowsky-Glahn et al., 2015)

- **usual units of measurement**: percentages, mg/kg (*constant sum constraint*), mg/l (*constant sum does not occur*)
- **examples**: **(a)** vegetation compositions of various plant species in different survey areas, **(b)** election results of political parties in different regions of a country, **(c)** household expenditures on various costs (housing, foodstuff, etc.) for a sample of households

Compositional data

= *D-part vectors, describing quantitatively the parts of some whole, which carry exclusively relative information between the parts* (Aitchison, 1986; Pawlowsky-Glahn et al., 2015)

- **usual units of measurement**: percentages, mg/kg (*constant sum constraint*), mg/l (*constant sum does not occur*)
- **examples**: **(a)** vegetation compositions of various plant species in different survey areas, **(b)** election results of political parties in different regions of a country, **(c)** household expenditures on various costs (housing, foodstuff, etc.) for a sample of households
- the constant sum of parts (1, 100) = *proper representation of compositions*

Geometric aspects of compositional data analysis

- assumptions of a relevant analysis of compositions: *scale invariance*, *subcompositional coherence*, *relative scale preserving* $\Rightarrow$ the **Aitchison geometry** (AG; EVS of dimension $D-1$)

- most of statistical methods rely on assumption of Euclidean geometry (Eaton, 1983)

Geometric aspects of compositional data analysis

- assumptions of a relevant analysis of compositions: *scale invariance*, *subcompositional coherence*, *relative scale preserving* $\Rightarrow$ the **Aitchison geometry** (AG; EVS of dimension $D-1$)

- most of statistical methods rely on assumption of Euclidean geometry (Eaton, 1983)

$\Rightarrow$ express compositional data in coordinates with respect to an orthonormal basis on the simplex (Egozcue et al., 2003) $\rightarrow$ statistical analysis, interpretation (*balances, lack of standard/Carthesian coordinates*)

- **log-ratio analysis** of compositional data (Aitchison, 1986)

Structural zeros are not welcome

- scale invariance principle → all relevant information in compositional data is contained in **ratios** between parts

Structural zeros are not welcome

- scale invariance principle → all relevant information in compositional data is contained in **ratios** between parts

⇒ **the logratio methodology cannot cope with zero values in parts** (Martín-Fernández et al., 2011)

Structural zeros are not welcome

- scale invariance principle → all relevant information in compositional data is contained in **ratios** between parts

⇒ **the logratio methodology cannot cope with zero values in parts** (Martín-Fernández et al., 2011)

- *rounded zeros* – caused by rounding errors (replacement strategies are used ☺)

Structural zeros are not welcome

- scale invariance principle → all relevant information in compositional data is contained in **ratios** between parts

⇒ **the logratio methodology cannot cope with zero values in parts** (Martín-Fernández et al., 2011)

- *rounded zeros* – caused by rounding errors (replacement strategies are used ☺)

- ***structural zeros*** – resulting from structural processes (replacement is not meaningful ☹)

- **examples**: **(a)** plant species that are not able to survive in a given soil type or climate, **(b)** a political party that has no candidates in a region, **(c)** teetotal households that do not have expenditures on alcohol and tobacco

Strategies for dealing with structural zeros

- **amalgamation** of compositional parts (Aitchison, 1986) (tobacco and alcohol parts amalgamated into a new part representing expenditures for both commodities)

Strategies for dealing with structural zeros

- **amalgamation** of compositional parts (Aitchison, 1986) (tobacco and alcohol parts amalgamated into a new part representing expenditures for both commodities)

× non-linear operation w.r.t. the Aitchison geometry, information on ratios between the corresponding compositional parts gets lost

Strategies for dealing with structural zeros

- **amalgamation** of compositional parts (Aitchison, 1986) (tobacco and alcohol parts amalgamated into a new part representing expenditures for both commodities)

× non-linear operation w.r.t. the Aitchison geometry, information on ratios between the corresponding compositional parts gets lost

- **parametric approach**: **(1)** determine where the zero entries occur in the data set (zero pattern structure), **(2)** model the distribution of the unit available from the non-zero parts using a binomial conditional logistic normal model

Strategies for dealing with structural zeros

- **amalgamation** of compositional parts (Aitchison, 1986) (tobacco and alcohol parts amalgamated into a new part representing expenditures for both commodities)

× non-linear operation w.r.t. the Aitchison geometry, information on ratios between the corresponding compositional parts gets lost

- **parametric approach**: **(1)** determine where the zero entries occur in the data set (zero pattern structure), **(2)** model the distribution of the unit available from the non-zero parts using a binomial conditional logistic normal model

× derivation of the likelihood assumes the usual Euclidean geometry, not followed by the original compositions

Strategies for dealing with structural zeros

- **zero patterns as indicators** of different subgroups of interest (teetotal households are forming a different household budget pattern)

Strategies for dealing with structural zeros

- **zero patterns as indicators** of different subgroups of interest (teetotal households are forming a different household budget pattern)

× small sample sizes of the resulting subsets of observations, necessary to get relevant estimates in a statistical model; zero patterns don't necessarily induce a different data structure

Strategies for dealing with structural zeros

- **zero patterns as indicators** of different subgroups of interest (teetotal households are forming a different household budget pattern)

× small sample sizes of the resulting subsets of observations, necessary to get relevant estimates in a statistical model; zero patterns don't necessarily induce a different data structure

⇒ use a **reasonable imputation** of zero parts as an **auxiliary step** to get estimates of parameters (e.g. covariance) from the overall data set (no new information is added to the data structure)

Strategies for dealing with structural zeros

- **zero patterns as indicators** of different subgroups of interest (teetotal households are forming a different household budget pattern)

× small sample sizes of the resulting subsets of observations, necessary to get relevant estimates in a statistical model; zero patterns don't necessarily induce a different data structure

⇒ use a **reasonable imputation** of zero parts as an **auxiliary step** to get estimates of parameters (e.g. covariance) from the overall data set (no new information is added to the data structure)

- the resulting estimates are used for an **analysis in the subcompositions resulting from the single zero patterns**

Mahalanobis distances for outlier detection

- the most widely used methods for multivariate outlier detection are those based on covariance estimates and Mahalanobis distances (MDs)

Mahalanobis distances for outlier detection

- the most widely used methods for multivariate outlier detection are those based on covariance estimates and Mahalanobis distances (MDs)
- given a sample of coordinates $\mathbf{z}_1, \ldots, \mathbf{z}_n \in \mathbf{R}^{D-1}$, the MD is defined as

$$\mathrm{MD}(\mathbf{z}_i) = \left[(\mathbf{z}_i - \mathbf{t})' \mathbf{C}^{-1} (\mathbf{z}_i - \mathbf{t})\right]^{1/2}, i = 1, \ldots, n; \qquad (1)$$

  $\mathbf{t}$ and $\mathbf{C}$ stand for (robust → MCD) location and covariance estimators

- if a certain threshold value is exceeded ($\chi^2_{D-1;0.975}$), the observation is flagged as potential outlier

Mahalanobis distances for outlier detection

- the most widely used methods for multivariate outlier detection are those based on covariance estimates and Mahalanobis distances (MDs)

- given a sample of coordinates $\mathbf{z}_1, \ldots, \mathbf{z}_n \in \mathbf{R}^{D-1}$, the MD is defined as

$$\mathrm{MD}(\mathbf{z}_i) = \left[(\mathbf{z}_i - \mathbf{t})' \mathbf{C}^{-1} (\mathbf{z}_i - \mathbf{t})\right]^{1/2}, i = 1, \ldots, n; \qquad (1)$$

  $\mathbf{t}$ and $\mathbf{C}$ stand for (robust $\rightarrow$ MCD) location and covariance estimators

- if a certain threshold value is exceeded $(\chi^2_{D-1;0.975})$, the observation is flagged as potential outlier

- **MDs are not directly applicable to compositional data with structural zeros**

Imputation approach to outlier detection

- the (auxiliary) **imputation strategy** is used to detect outliers in single zero patterns (Templ et al., 2016)

Imputation approach to outlier detection

- the (auxiliary) **imputation strategy** is used to detect outliers in single zero patterns (Templ et al., 2016)

- **orthonormal** (pivot) **coordinates** $\mathbf{z} = (z_1, \ldots, z_{D-1})'$,

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D-i]{\prod_{k=i+1}^{D} x_k}}, \; i = 1, \ldots, D-1$$

(Fišerová and Hron, 2011), guarantee that the subcomposition $(x_i, \ldots, x_D)'$ is represented by the last $i-1$ coordinates

Imputation approach to outlier detection

- the (auxiliary) **imputation strategy** is used to detect outliers in single zero patterns (Templ et al., 2016)

- **orthonormal** (pivot) **coordinates** $\mathbf{z} = (z_1, \ldots, z_{D-1})'$,

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D-i]{\prod_{k=i+1}^{D} x_k}}, \quad i = 1, \ldots, D-1$$

  (Fišerová and Hron, 2011), guarantee that the subcomposition $(x_i, \ldots, x_D)'$ is represented by the last $i-1$ coordinates

- **permutation of parts** and **affine equivariance** of the MCD estimator are used to perform **outlier detection in any subcomposition resulting from the zero patterns**

Outliers according to zero patterns

- **MDs used to reveal outliers** resulting just **from non-zero parts** of compositions → in the second step **outlying zero patterns** are of interest

Outliers according to zero patterns

- **MDs used to reveal outliers** resulting just **from non-zero parts** of compositions → in the second step **outlying zero patterns** are of interest
- the data are recoded into a binary matrix (non-zeros . . . 1)
- outliers 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

Outliers according to zero patterns

- **MDs used to reveal outliers** resulting just **from non-zero parts** of compositions → in the second step **outlying zero patterns** are of interest
- the data are recoded into a binary matrix (non-zeros . . . 1)
- outliers 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
- **the multivariate structure and outlyingness of the zero patterns** are analyzed using **principal component analysis** (PCA) for **binary data** (Leeuw, 2006) → loadings and scores
- **results from the previous steps are merged together**

Austrian EU-SILC data set

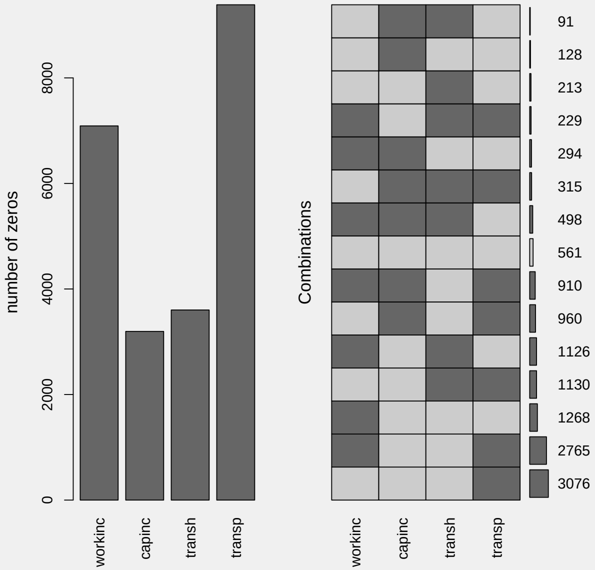
- European Union Statistics on Income and Living Conditions (EU-SILC) is an annual panel household survey conducted in most of European countries, data basis for measuring risk-of-poverty and social cohesion in Europe

- the Austrian EU-SILC 2006 data set is considered, the data set is simulated from the original (confidential) data with the R package `simPopulation`

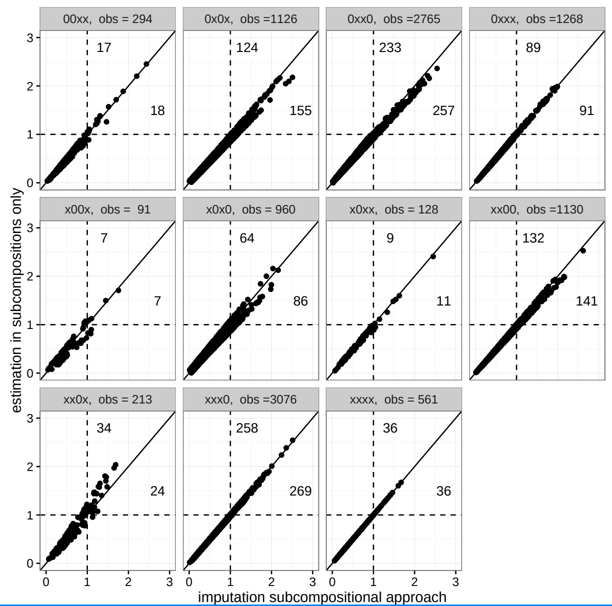
- 14,827 observations from 6,000 households and 28 variables are obtained (data `eusilc` from the R package `laeken`)

- the income components contain (too) many zeros → the parts are amalgamated to obtain the four compositional parts *workinc* (work income), *capinc* (capital income), *transh* (household transfers), and *transp* (personal transfers)

Austrian EU-SILC data: zero structure



Austrian EU-SILC data: Mahalanobis distances



Austrian EU-SILC data: findings

- MDs results from all patterns are similar → **zero patterns do not cause significant changes in covariance structure**

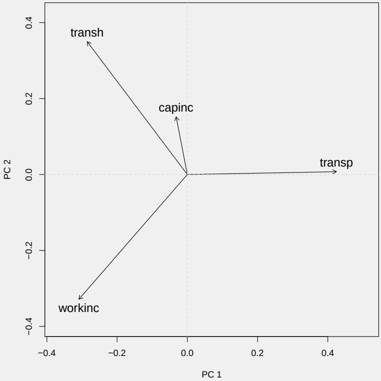
- the imputation approach guarantees that enough sample size is used for robust estimation of MDs in single zero patterns

- **PCA biplot**: patterns with observed values in a specific variable (indicated by x) are located in direction of the respective arrow

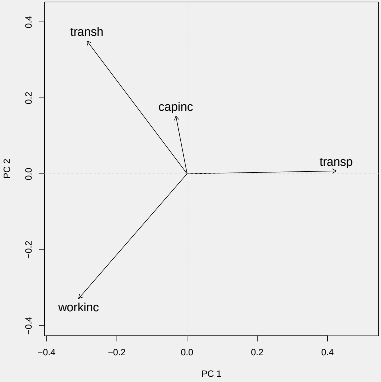
- no clear outlier visible in the scores plot, i.e. none of the zero patterns shows extreme behavior

- though, some atypical patterns, located further from the origin, are present, like *x00x* (occurs only 91 times)

Austrian EU-SILC data: PCA for binary data



Austrian EU-SILC data: PCA for binary data



- the respective R functions (`zeroOut`, `zeroPatterns`) from the package `robCompositions` soon available at CRAN

Austrian EU-SILC data: findings

- MDs results from all patterns are similar → **zero patterns do not cause significant changes in covariance structure**
- the imputation approach guarantees that enough sample size is used for robust estimation of MDs in single zero patterns
- **PCA biplot**: patterns with observed values in a specific variable (indicated by x) are located in direction of the respective arrow

Austrian EU-SILC data: findings

- MDs results from all patterns are similar → **zero patterns do not cause significant changes in covariance structure**
- the imputation approach guarantees that enough sample size is used for robust estimation of MDs in single zero patterns
- **PCA biplot**: patterns with observed values in a specific variable (indicated by x) are located in direction of the respective arrow
- no clear outlier visible in the scores plot, i.e. none of the zero patterns shows extreme behavior
- though, some atypical patterns, located further from the origin, are present, like *x00x* (occurs only 91 times)

Conclusions

- **outlier detection** is daily routine in statistical offices when specific data sets are checked for **plausibility**; the usual procedure then is to 'correct' implausible data values, or to reduce the effect of outliers in statistical estimation

Conclusions

- **outlier detection** is daily routine in statistical offices when specific data sets are checked for **plausibility**; the usual procedure then is to 'correct' implausible data values, or to reduce the effect of outliers in statistical estimation

- for statistical estimation, the **compositional nature of the data needs to be taken into account** $\times$ the logratio methodology of compositional data can cope with structural zeros just indirectly as demonstrated also with the proposed procedure

Conclusions

- **outlier detection** is daily routine in statistical offices when specific data sets are checked for **plausibility**; the usual procedure then is to 'correct' implausible data values, or to reduce the effect of outliers in statistical estimation

- for statistical estimation, the **compositional nature of the data needs to be taken into account** $\times$ the logratio methodology of compositional data can cope with structural zeros just indirectly as demonstrated also with the proposed procedure

- 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

References

- Aitchison, J. : *The statistical analysis of compositional data*. Chapman and Hall, London, 1986.
- Eaton, M.L.: *Multivariate statistics: A vector space approach*. Wiley, New York, 1983.
- Egozcue, J.J., Pawlowsky-Glahn, V.: *Groups of parts and their balances in compositional data analysis*. Mathematical Geology 37, 795-828, 2005.
- Fišerová, E., Hron, K.: *On interpretation of orthonormal coordinates for compositional data*. Mathematical Geosciences 43, 455-468, 2011.
- Leeuw, J. de: *Principal component analysis of binary data by iterated singular value decomposition*. Computational Statistics and Data Analysis 50, 21-39, 2006.
- Martín-Fernández, J.A., Palarea-Albaladejo, J., Olea, R.A.: *Dealing with zeros*. In Pawlowsky-Glahn, V., Buccianti, A., editors, *Compositional data analysis: Theory and applications*. Wiley, Chichester, 2011.
- Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R.: *Modeling and analysis of compositional data*. Wiley, Chichester, 2015.
- **Templ, M., Hron, K., Filzmoser, P.: *Exploratory tools for outlier detection in compositional data with structural zeros*. Journal of Applied Statistics, DOI: 10.1080/02664763.2016.1182135.**