Outlier detection in complex survey data including semi-continuous components and missing values

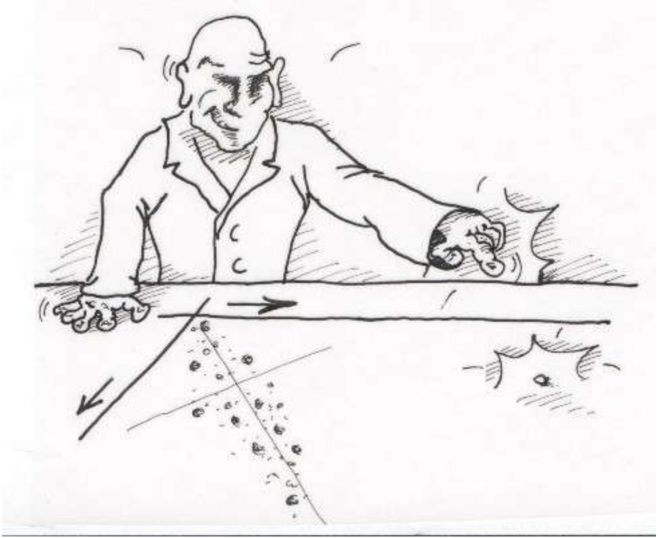
Matthias Templ [1] Johannes Gussenbauer [1]
Peter Filzmoser [1] Oliver Dupriez [2]

[1]Vienna University of Technology

[2]World Bank
This work was funded by the World Bank (project: "Improving the quality of sample household expenditure data and the reliability of poverty and inequality measures", selection no. 1157976.

December 13, 2015

We are happy doing this...



source: http://www.vias.org/science_cartoons/outlier.html

Outlier in household expenditure data

- household expenditure information is usually gathered through complex household surveys
  - data are subject to human error
  - participants don't want to share or know every information
- the Gini coefficient plays an important role in connection with household expenditure data
  - measures the inequality of the household spendings among the surveyed households

Impact of outliers

- huge impact on non-robust estimators
- ranking between countries may completely change
- World Bank have used simple univariate outlier detection and replacement of outliers
- projekt with World Bank to improve outlier detection and replacement

Provided data and data structure

- household expenditure data from Albania(2008), Mexico(2010), India(2009), Malawi(2010) and Tajikistan(2007)
- containing value of goods or services for each household over a period of time
- World Bank started to harmonize the resulting data
- household consumption categorized by
  - ICP basic headings / ICP class / ICP group / ICP category

Data preparation & missing values

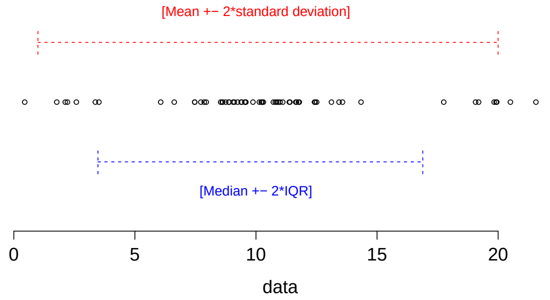
- household consumption of good or service only listed if greater than zero
- not possible to differentiate if those are real zeros or missing values
- number of zeros/missing values is very high when using the ICP classification (many categories)
- amalgamation of components is thus necessary
  - combine variables with comparably large household expenditures
  - combine variables to efficiently reduce zeros/missings

Category	Zeros/Missing entries
Food and non-alcoholic beverages	2
Alcoholic beverages, tobacco and narcotic	1476
Clothing and footwear	347
Furnishings, household equipment, household maintenance	2
Health	1264
Transport	1468
Communication	407
Recreation and culture	19
Education	3278
Restaurants and hotels	1814
Miscellaneous goods and services	114
Net purchases abroad	3600

Table: Number of missing entries per category for the Albanian household survey, which contains 3600 households

Robust statistical methods

- we use robust statistical methods to detect potential outliers
- univariate and multivariate methods were tested



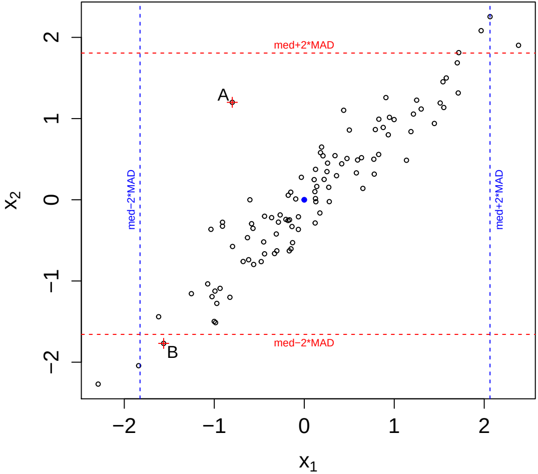
Univariate methods

- data points which are "far enough" away from the main bulk of the data
- the following methods were used:
  - estimate location and scale in a robust way to determine interval for "good" observations
    - $[med - c \cdot S_{IQR} , med + c \cdot S_{IQR}]$
    - $[med - c \cdot S_{MAD} , med + c \cdot S_{MAD}]$
    - boxplot
  - expenditure data usually skewed to the right
    - use Box-Cox transformation $\Rightarrow$ estimate interval $\Rightarrow$ transform back interval boundaries
    - use skewness-adjusted Boxplot
  - Pareto tail modeling using robust methods that can deal with sampling weights (Alfons, Templ, Filzmoser, 2013)

Replacement of univariate potential outliers

- potential outliers are winsorized to the lower/upper ends of the calculated intervals
- for Pareto tail modeling, values larger than a certain quantile of the fitted distribution
  - are replaced by values drawn from the fitted distribution
  - their sample weights are set to 1 and the rest of the data are re-calibrated

Applying univariate methods to multivariate data



Mahalanobis distance

- use distance measure which takes into account the multidimensional structure of the data $\Rightarrow$ squared Mahalanobis distance MD_i^2

$$MD_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})^t S^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \quad ,$$

- estimate center and covariance in a robust way to gain squared robust distances, RD_i^2
- if data follows a multivariate normal distribution $\Rightarrow$ $MD_i^2 \sim \chi_p^2$
- declare data points as potential outliers if they exceed $\chi^2_{p;0.975}$

Multivariate methods

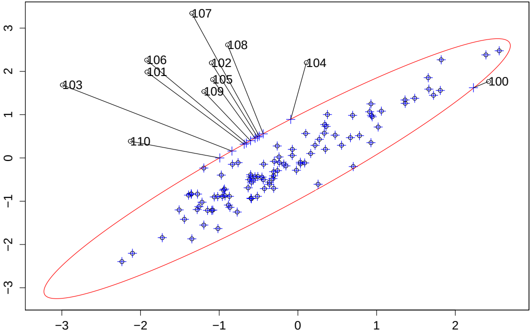
- robust methods to estimate center and covariance
  - M-estimate
    - generalization of Maximum Likelihood estimate
  - S-estimate
  - MM-estimate
    - uses high breakdown preliminary S-estimate
  - MCD-& MVE-estimate
    - Minimum covariance determinant estimate
    - Minimum volume ellipsoid estimate
  - Stahel-Donoho estimate
  - OGK estimate
    - $Cov(X, Y) = \frac{1}{4}(Var(X + Y) - Var(X - Y))$

Multivariate methods

- BACON-EEM
  - combines the BACON algorithm and EEM algorithm
  - uses EEM-algorithm to estimate center and covariance during BACON-procedure
  - EEM-algorithm able to handle missing values in the data
- Epidemic Algorithm
  - simulate an epidemic, starting from the center of the data
  - data points with high infection times are declared potential outliers

Replace potential outliers

- multivariate potential outliers are winsorised onto the boundaries of the 97.5% tolerance ellipse.



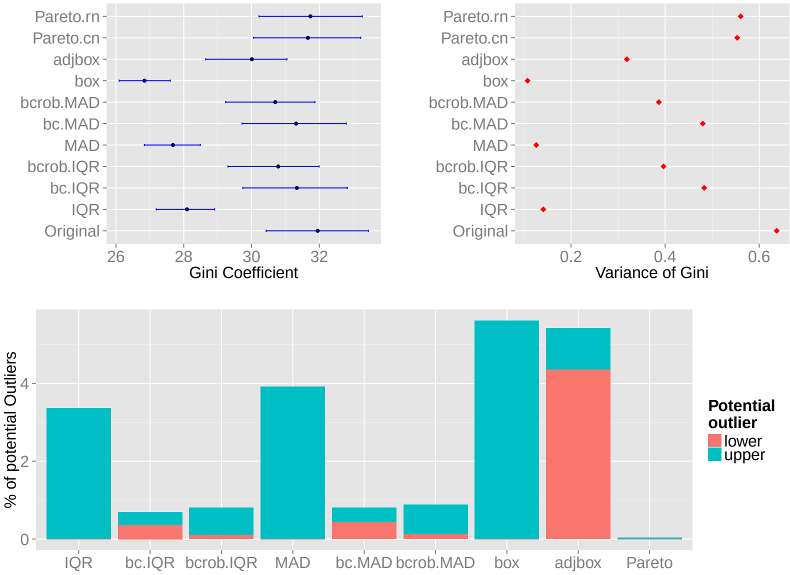
Applying outlier detection methods

- univariate outlier detection methods were applied on the total annual household expenditures
  - exclude missing values/zeros from calculations

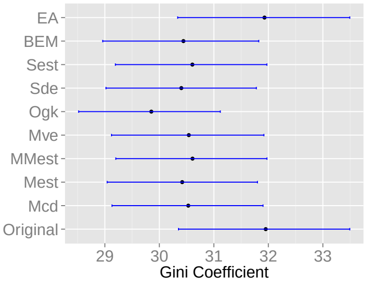
- multivariate outlier detection methods after
  - log transforming the data
  - imputation of zeros/missing values if necessary with kNN algorithm
    - BACON-EEM & EA have an internal imputation mechanism

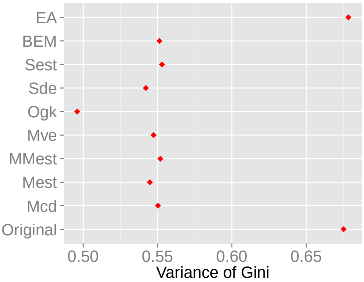
- estimate weighted Gini coefficient of total annual expenditures

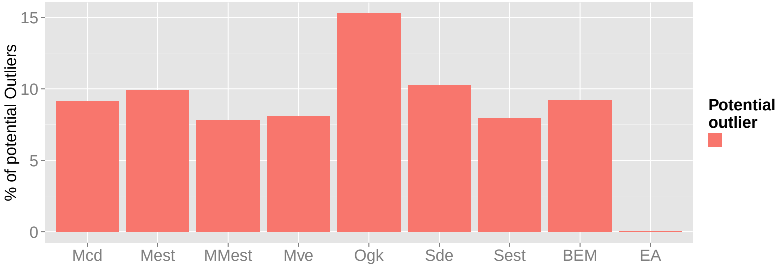
Results for Albanian data set



Results for Albanian data set







Simulation setup

- To know the number of “true” outliers.
  - split Albanian data into a ”clean” and ”contaminated” data set
    - data point never flagged ⇒ ”clean” data
    - data point flagged by at least 5 univariate outlier detection methods OR at least 6 multivariate outlier detection methods ⇒ ”contaminated” data
- estimate location and covariance for ”clean” and ”contaminated” data set in a classical manner ⇒ $(\boldsymbol{\mu}_{cl}, \boldsymbol{\Sigma}_{cl}),(\boldsymbol{\mu}_{co}, \boldsymbol{\Sigma}_{co})$
- simulate data from $MVN(\boldsymbol{\mu}_{cl}, \boldsymbol{\Sigma}_{cl})$

Simulation setup

- swap observations with contaminated values generated from $MVN(\boldsymbol{\mu}_{co}, \boldsymbol{\Sigma}_{co})$
  - swap only a single cell for share of contaminated data
- simulated data set X follows the following distribution

$$\mathbf{X} \;\sim\; (1-\epsilon)MVN(\boldsymbol{\mu}_{cl}, \boldsymbol{\Sigma}_{cl}) + \epsilon MVN(\boldsymbol{\mu}_{co}, \boldsymbol{\Sigma}_{co}) \quad ,$$

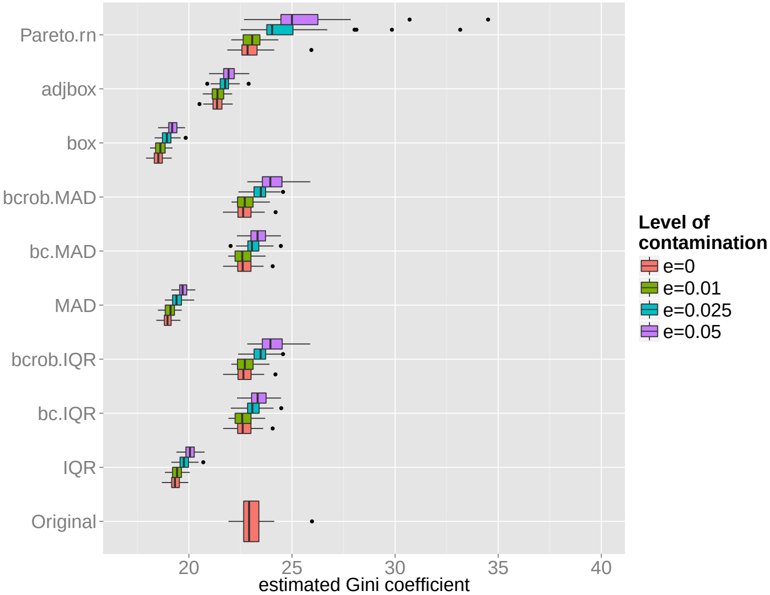
  with $\epsilon \in (0,1)$ determining the share of contaminated data points.
- include missing values and sample weights from the Albanian data set

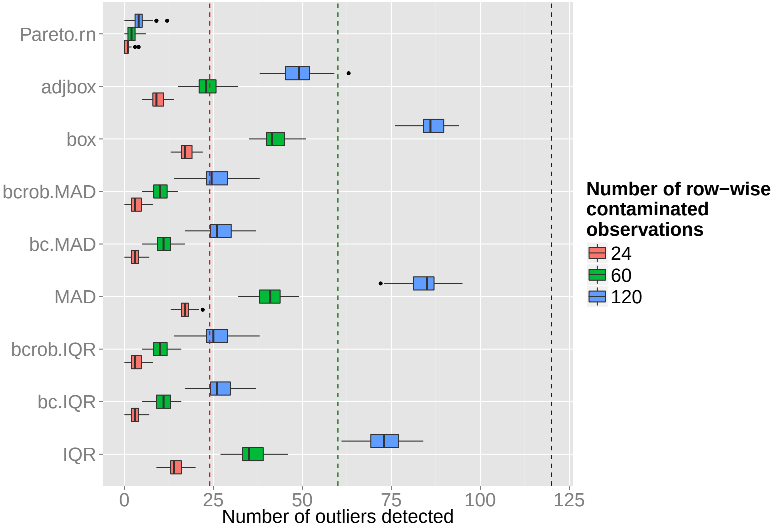
Simulation parameters

- simulation and application of univariate and multivariate outlier detection methods is repeated 50 times
- $\epsilon \in \{0; 0.01; 0.025; 0.05\}$
- $1/3$ of the contamination is cell-wise

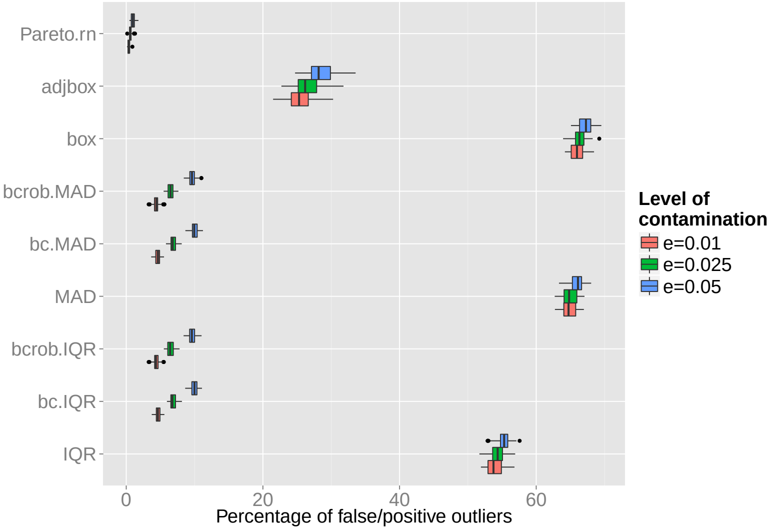
Application of outlier detection methods methods

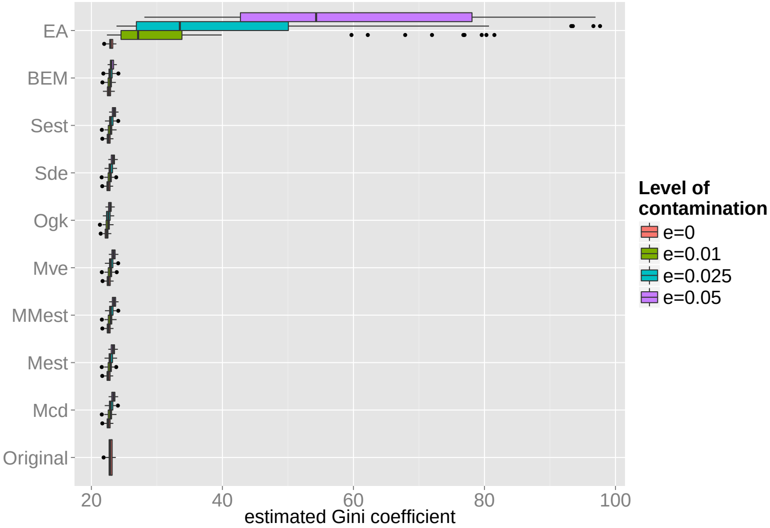
- simulate data
- apply outlier detection methods
  - apply univariate methods on each of the columns of the generated data $\Rightarrow$ results more comparable to multivariate case
- detect and impute potential outliers
- count correctly identified outliers and false positive outliers
- estimate the Gini coefficient of the total sum of each observation

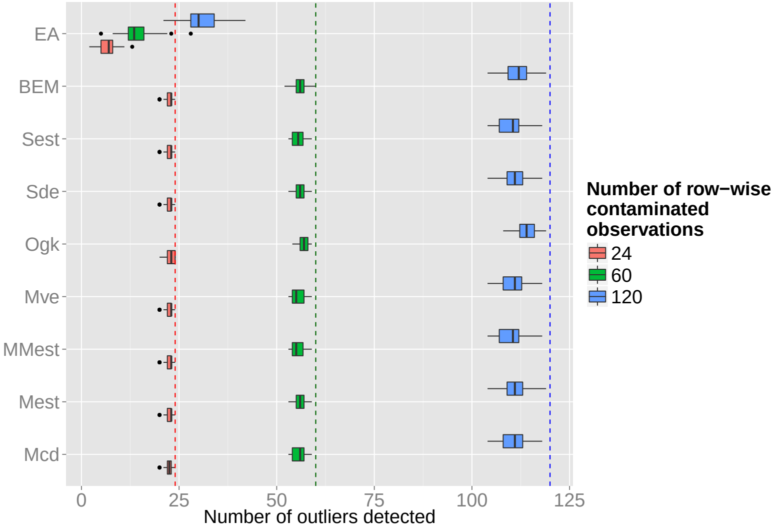

Pareto.rn
adjbox
box
bcrob.MAD
bc.MAD
MAD
bcrob.IQR
bc.IQR
IQR
Original
20
25
30
35
40
estimated Gini coefficient
Level of contamination
e=0
e=0.01
e=0.025
e=0.05

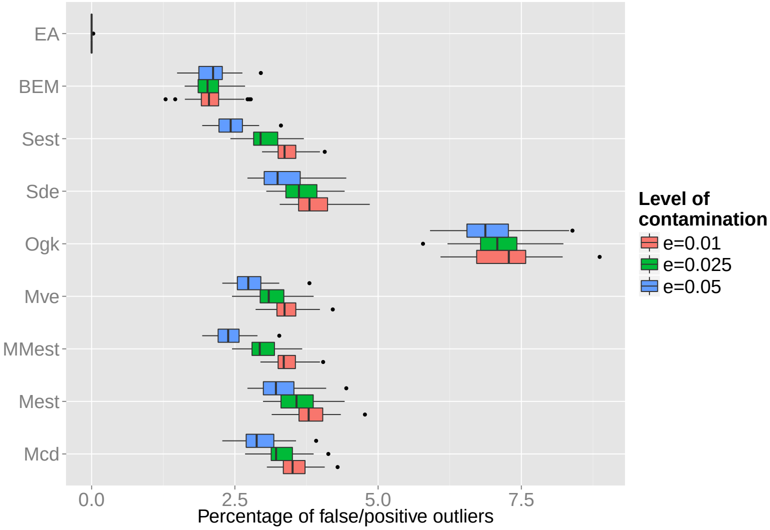


Pareto.rn
adjbox
box
bcrob.MAD
bc.MAD
MAD
bcrob.IQR
bc.IQR
IQR
0
25
50
75
100
125
Number of outliers detected
Number of row−wise contaminated observations
24
60
120


Pareto.rn
adjbox
box
bcrob.MAD
bc.MAD
MAD
bcrob.IQR
bc.IQR
IQR
0
20
40
60
Percentage of false/positive outliers
Level of contamination
e=0.01
e=0.025
e=0.05


EA
BEM
Sest
Sde
Ogk
Mve
MMest
Mest
Mcd
Original
20
40
60
80
100
estimated Gini coefficient
Level of contamination
e=0
e=0.01
e=0.025
e=0.05


EA
BEM
Sest
Sde
Ogk
Mve
MMest
Mest
Mcd
0
25
50
75
100
125
Number of outliers detected
Number of row−wise contaminated observations
24
60
120



EA
BEM
Sest
Sde
Ogk
Mve
MMest
Mest
Mcd
0.0
2.5
5.0
7.5
Percentage of false/positive outliers
Level of contamination
e=0.01
e=0.025
e=0.05

Estimates of Gini for the 5 different countries

Country	Number of households		Original	IQR	BACON EEM
Albania(2008)	3600	Gini	31.95	28.10	30.44
		Number outlier	–	121	332
India(2009)	100852	Gini	39.82	33.56	37.44
		Number outlier	–	9131	9404
Mexico(2010)	27655	Gini	44.20	37.62	42.75
		Number outlier	–	1669	2429
Malawi(2010)	12096	Gini	48.52	36.13	41.22
		Number outlier	–	1003	796
Tajikistan(2007)	4860	Gini	33.11	28.59	30.32
		Number outlier	–	244	505

Summary

- Simulation study necessary to determine performance of outlier detection methods on household expenditure data
- The simulation study presented in this work favored the BACON-EEM to be the most suitable method, but
  - simulation study favored multivariate methods in contrast to univariate methods
  - did not take into account sociodemographic criteria or household specific information
  - $\rightarrow$ cell-wise outlier detection methods using regression on compositional parts are just tested. First results are promising.

References

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

A. Alfons, M. Templ, and P. Filzmoser.
Robust estimation of economic indicators from survey samples based on Pareto tail modeling.
Journal of the Royal Statistical Society, Series C, 62(2):271â–286, 2013.

C. Béguin and B. Hulliger.
The BACON-EEM algorithm for multivariate outlier detection in incomplete survey data.
Survey Methodology, 34(1):91–103, 2008.

N. Billor, A. S. Hadi, and P. F. Vellemann.
BACON: Blocked adaptative computationally-efficient outlier nominators.
Computational Statistics and Data Analysis, 34(3):279–298, 2000.

B. Hulliger, A. Alfons, P. Filzmoser, A. Meraner, T. Schoch, and M. Templ.
Robust methodology for laeken indicators.
Research Project Report WP4 – D4.2, FP7-SSH-2007-217322 AMELI, 2011.

T. Todorov and P. Filzmoser.
An object oriented framework for robust multivariate analysis.
Journal of Statistical Software, 32(3):1–47, 2009.