Robust Statistical Methods for Outlier Detection with Application to Household Expenditure Data

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

1Vienna University of Technology
2World Bank

October 27, 2015
Outlier in household expenditure data

- Household expenditure data usually provided through household surveys
  - Data subject to human error
  - participants don’t want to share 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

- Outliers may have a huge impact on non-robust estimators
- Ranking between countries may completely change only because of outliers
- From then on World Bank used simple univariate outlier detection and replacement of outliers
- Projekt with World Bank to improve outlier detection and replacement
Provided data and data structure


- Product of large household surveys containing value of goods or services consumed in local currency for each household over a period of time

- World Bank started to harmonize the resulting data into a common framework

- Household consumption categorized by
  - ICP basic headings / ICP class / ICP group / ICP category
Robust statistical methods

- Use robust statistical methods to detect potential outliers
- Due to the structure of the data 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
    - \[ [\text{med} - c \cdot S_{IQR}, \text{med} + c \cdot S_{IQR}] \]
    - \[ [\text{med} - c \cdot S_{MAD}, \text{med} + c \cdot S_{MAD}] \]
  - 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
Replace univariate potential outliers

- Place potential outliers onto 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 recalibrate for the rest of the data
Applying univariate methods to multivariate data

Figure: Simulated data from multivariate standard normal distribution including one outlier (point A)
Mahalanobis distance

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

$$MD_i^2 = (x_i - \bar{x})^t S^{-1} (x_i - \bar{x})$$

- 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 $\chi_p^2;0.975$
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
    - Incorporates multivariate measure of outlyingness
  - OGK estimate
    - $\text{Cov}(X, Y) = \frac{1}{4}(\text{Var}(X + Y) - \text{Var}(X - Y))$
Multivariate methods

- **BACON-EEM**
  - Combines 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

- Apply univariate outlier detection methods on total annual household expenditures
  - Exclude missing values/zeros from calculations

- Apply multivariate outlier detection methods by
  - Log transforming the data
  - Impute 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

Gini Coefficient

<table>
<thead>
<tr>
<th>Method</th>
<th>Original</th>
<th>IQR</th>
<th>bc.IQR</th>
<th>bcrob.IQR</th>
<th>MAD</th>
<th>bc.MAD</th>
<th>bcrob.MAD</th>
<th>box</th>
<th>adjbox</th>
<th>Pareto.cn</th>
<th>Pareto.rn</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Variance of Gini

<table>
<thead>
<tr>
<th>Method</th>
<th>Original</th>
<th>IQR</th>
<th>bc.IQR</th>
<th>bcrob.IQR</th>
<th>MAD</th>
<th>bc.MAD</th>
<th>bcrob.MAD</th>
<th>box</th>
<th>adjbox</th>
<th>Pareto.cn</th>
<th>Pareto.rn</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

% of potential Outliers

<table>
<thead>
<tr>
<th>Method</th>
<th>Original</th>
<th>IQR</th>
<th>bc.IQR</th>
<th>bcrob.IQR</th>
<th>MAD</th>
<th>bc.MAD</th>
<th>bcrob.MAD</th>
<th>box</th>
<th>adjbox</th>
<th>Pareto.cn</th>
<th>Pareto.rn</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Potential outlier

- lower
- upper
Results for Albanian data set

- Gini Coefficient
- Variance of Gini
- % of potential Outliers

Variables:
- Original
- Mcd
- Mest
- MMest
- Mve
- Ogk
- Sde
- Sest
- BEM
- EA

Graphs showing the distribution of values across different measures and the presence of potential outliers.
Simulation

- Simulate such kind of data sets which can be comparable, regarding the data on household expenditure, with the ones provided by the World Bank

- Know the number and position of "true" outliers beforehand
Simulation setup

- Use Albanian data set and
  - split data into "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 way
    ⇒ \((\mu_{cl}, \Sigma_{cl}), (\mu_{co}, \Sigma_{co})\)
  - Simulate data from \(MVN(\mu_{cl}, \Sigma_{cl})\)
Simulation setup

- swap observations with contaminated values generated from $MVN(\mu_{co}, \Sigma_{co})$
  - swap only a single cell for share of contaminated data

- Simulated data set $X$ follows the following distribution

\[
X \sim (1 - \epsilon) MVN(\mu_{cl}, \Sigma_{cl}) + \epsilon MVN(\mu_{co}, \Sigma_{co}),
\]

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 Gini coefficient of the total sum of each observation
### Estimates of Gini for the 5 different countries

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

A. Alfons and M. Templ.
Estimation of social exclusion indicators from complex surveys: The R package laeken.

C. Béguin and B. Hulliger.
The BACON-EEM algorithm for multivariate outlier detection in incomplete survey data.

BACON: Blocked adaptative computationally-efficient outlier nominators.

Robust methodology for laeken indicators.

T. Todorov and P. Filzmoser.
An object oriented framework for robust multivariate analysis.

This work has been funded by the World Bank under the project "Improving the quality of sample household expenditure data and the reliability of poverty and inequality measures", selection number 1157976.