# Detecting outliers in weighted univariate survey data

Anna Pauliina Sandqvist, KOF ETH Zurich

The 7th Joint EC-OECD Workshop on recent developments in Business and Consumer Surveys, Paris

30 November 2015

# Outline

Motivation

Overview of outlier detection methods

Setting and approaches

Empirical influence function analysis

Simulation analysis

Conclusions

# What is this paper about? Why relevant?

- Survey data often skewed, most outlier detection approaches not optimal for asymmetric or long-tailed data
- We propose two alternative outlier detection approaches
- In general, lot of subjectivity is involved in outlier detection
- Size weights
- Data Generating Process unknown
- Let $y_i$ be the (year-on-year) growth rate of a company $i$ and $w_i$ be its size weight. Then the simple point estimator for the weighted mean of the sample with $n$ observations is as follows

$$\hat{\mu}_n = \frac{\sum w_i y_i}{\sum w_i}$$

# Overview of outlier detection methods

- Rule based methods:
  - Relative distance (a location estimate $l$ divided by a scale estimate $s$)

$$rd_i = \frac{|y_i - l|}{s}$$

  - Tolerance interval: $l - k_l s, l + k_h s$
- Method of Hidiroglou and Berthelot (1986)
- Parametric approaches

# Method of Hidiroglou and Berthelot (1986) (HB-method) I

- First step: transformation

$$s_i = \begin{cases} 1 - q_{0.5}/r_i & \text{if } 0 < r_i < q_{0.5} \\ r_i/q_{0.5} - 1 & \text{if } r_i \geq q_{0.5} \end{cases}$$

  where $r_i = y_{i,t+1}/y_{i,t}$ and $q_{0.5}$ is the median of $r$ values.

- Second step:

$$E_i = s_i\{max([y_i(t), y_i(t+1)]\}^V$$

  where $0 \leq V \leq 1$. The parameter $V$ controls the importance of the magnitude of the data.

# Method of Hidiroglou and Berthelot (1986) (HB-method) II

- Acceptance interval

$$\{E_{med} - CD_{Q1}, E_{med} + CD_{Q3}\}$$

  where $D_{Q1} = max\{E_{med} - E_{Q1}, |aE_{med}|\}$ and $D_{Q3} = max\{E_{Q3} - E_{median}, |aE_{med}|\}$ and the $a$ parameter takes usually the value of 0.05.

- The idea of the term $|aE_{med}|$ is to hinder that the interval becomes very small in the case that the observations are clustered around some value. The parameter $C$ defines the width of the acceptance interval.
- We set $a$ = 0.05 and $C$ = 7

# Right-left scale approach I

- Hubert and Van Der Veeken (2008) propose *adjusted outlyingness* (AO) measure based on different scale measures on both sides of the median:

$$AO_i = \begin{cases} \frac{y_i - q_{0.5}}{w_u - q_{0.5}} & \text{if } y_i > q_{0.5} \\ \frac{q_{0.5} - y_i}{q_{0.5} - w_l} & \text{if } y_i < q_{0.5} \end{cases}$$

where $w_l$ and $w_u$ are the lower and upper whisker of the adjusted boxplot as in Hubert and Vandervieren (2008):

$$w_l = q_{0.25} - 1.5^{-4MC} IQR; w_u = q_{0.75} + 1.5^{3MC} IQR \quad \text{for } MC \geq 0$$

$$w_l = q_{0.25} - 1.5^{-3MC} IQR; w_u = q_{0.75} + 1.5^{4MC} IQR \quad \text{for } MC < 0$$

where MC equals to the *medcouple*, a measure of skewness, introduced by Brys et al. (2004).

# Right-left scale approach II

- We adjust the denominator to account for the cases where it could become relatively low or high as follows:

$$rl_i = \begin{cases} \dfrac{y_i - q_{0.5}}{min(max(w_u - q_{0.5}, range(y)/n_y^{0.6}), IQR(y))} & \text{if } y_i \geq q_{0.5} \\ \dfrac{y_i - q_{0.5}}{min(max(q_{0.5} - w_l, range(y)/n_y^{0.6}), IQR(y))} & \text{if } y_i < q_{0.5} \end{cases}$$

- In the second step, this measure is also multiplied by the size weight:

$$rl_i^w = rl_i * w_i^V$$

  where the parameter $0 \leq V \leq 1$ again controls the importance of the weights.

- The interquartile range interval $q_{0.25} - 3IQR, q_{0.75} + 3IQR$ is applied to identify the outliers.

# $Q_n$-approach

- We apply the $Q_n$ estimator of Rousseeuw and Croux (1993) as scale estimator:

  $$Q_n = d\{|x_i - x_j|; i < j\}_{(k)} \tag{1}$$

  where $d$ is a constant and $k = \binom{h}{2} \approx \binom{n}{2}/4$ and $h = [n/2] + 1$.
- First step: the distance to median divided by $Q_n$ is calculated for unweighted values:

  $$Q_i = \frac{y_i - median(y)}{Q_n(y)}$$

- Second step

  $$Q_i^w = Q_i * w_i^V$$

- We apply the following interquartile interval for $Q_i^w$: $q_{0.25} - 3IQR, q_{0.75} + 3IQR$.

# Empirical influence function analysis

- With *empirical influence function* (EIF) the influence of a single observation on the estimator can be studied
- We define the EIF as follows:

  $$EIF_{\hat{\mu}}(w_0 y_0) = \hat{\mu}(w_1 y_1, ..., w_{n-1} y_{n-1}, w_0 y_0)$$

  where $y_0$ is the additional observation which takes values between -40 and 40 and $w_0 = 5$.
- Data: the weights of the observations $w_i$ are drawn from exponential distribution with $\lambda = 1/6$ and truncated at 20 and round up to make sure no zero values are possible. Therefore, 1 to 20 is the range of the size weight classes we take into account. Thereafter, the actual observations $y_1, ..., y_{n-1}$, which are in this case growth rates respectively ratios, are randomly generated from normal distribution with mean = 0 and standard deviation = 3.

# Empirical influence function analysis

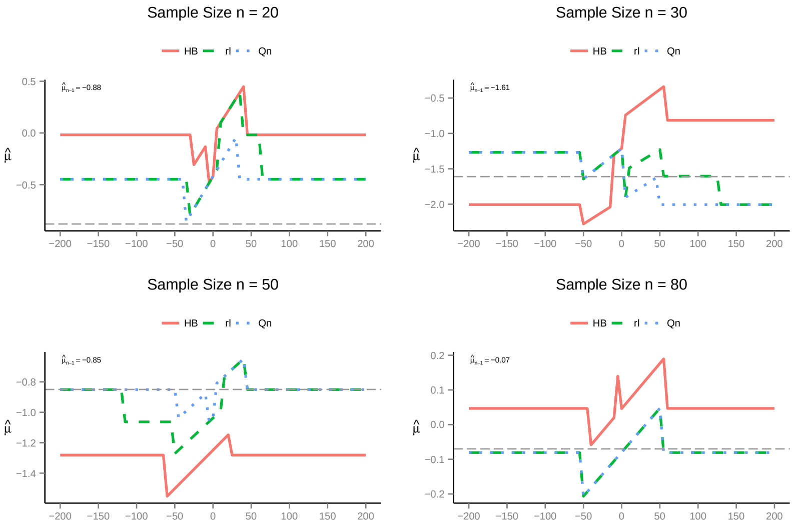

Figure 1 : Influence of one additional observation on the estimators

# Empirical influence function analysis

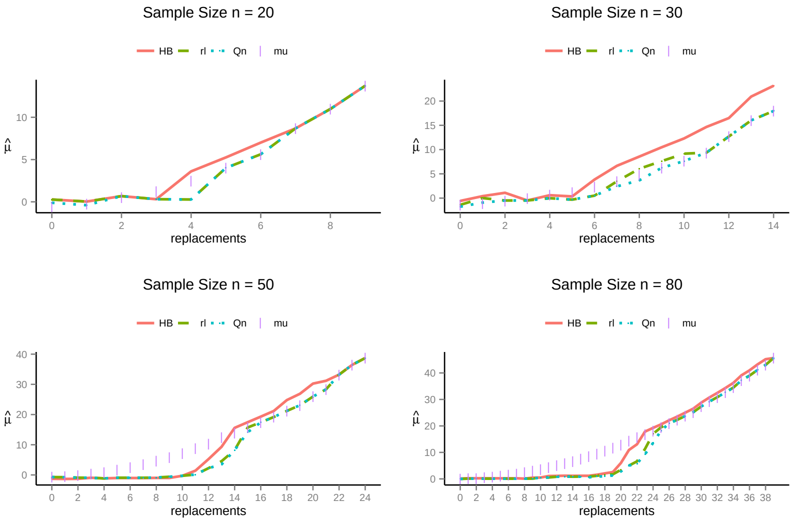

Figure 2 : Effect of multiple increasing (positive) replacements on the estimators

# Simulation analysis

- Within each simulation, we generate 1000 samples of sizes 15, 30, 50 and 100 observations
- The weights of the observations $w_i$ are drawn from exponential distribution with $\lambda = 1/6$ truncated at 20
- Asymmetric data, we generated data as a sum of two random variables drawn from at zero truncated normal (sd equal to 0.01 for big companies, and 0.02 for smaller companies) and exponential distribution where as the mean of the normal distribution was set to between 0.6 and 1.0 and the mean of the exponential distribution was adjusted so that the total mean stays constant at 1.05 or respectively 5%.
- Fat-tailed data, the actual observations $y_i$ are generated from truncated non-standard *t*-distribution distribution with varying values for degrees of freedom. For observation with smaller weight higher standard deviation is applied.

# Simulation analysis - asymmetric data

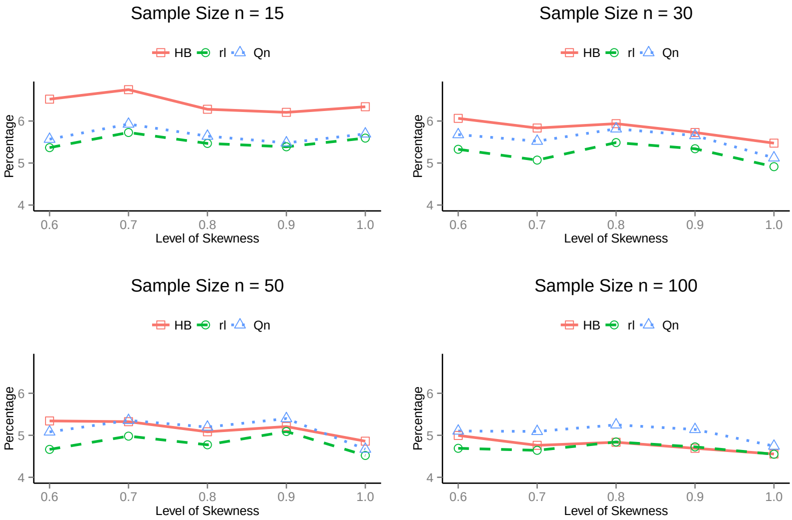

Figure 3 : Average percentage of observation declared as outliers depending on the skewness of the data

# Simulation analysis - asymmetric data

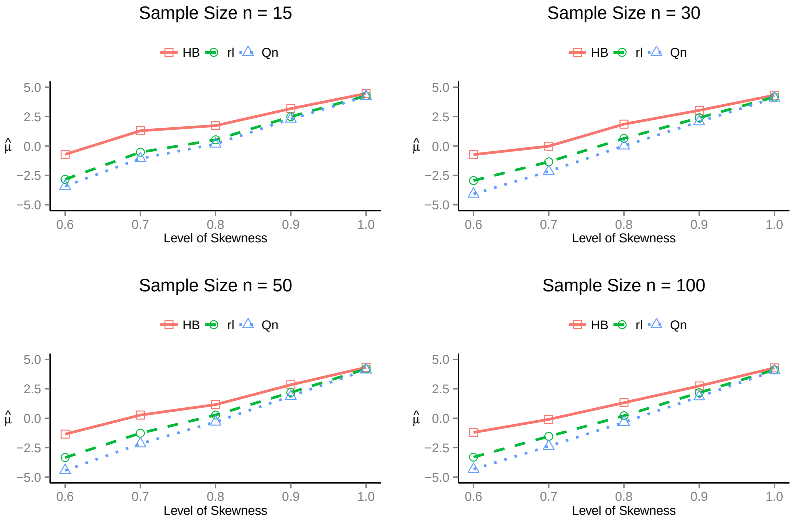

Figure 4 : Average estimates of weighted growth rate depending on the skewness of data

# Simulation analysis - asymmetric data

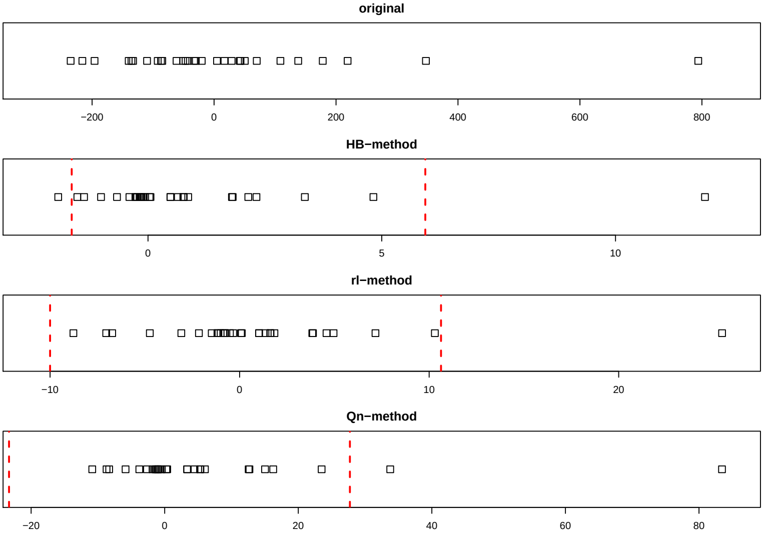

Figure 5 : Original weighted growth rates and test statistics of each method with acceptance boundaries for an example data set $n = 30$

# Simulation analysis - long-tailed data

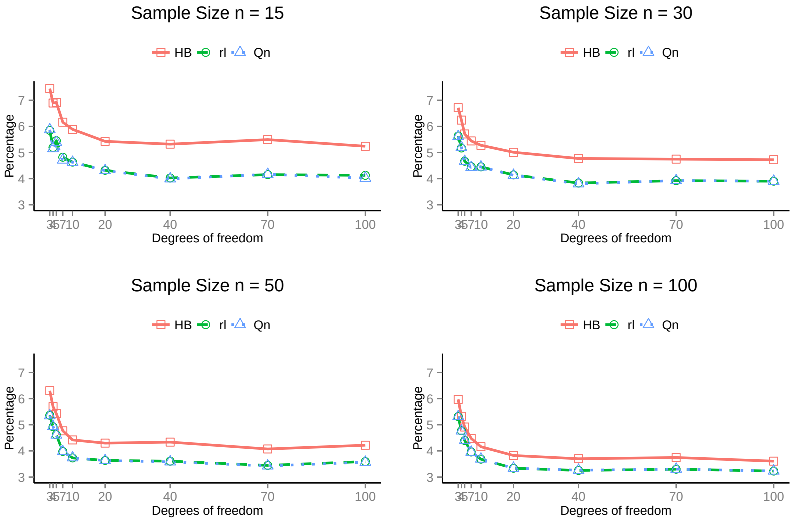

Figure 6 : Average percentage of observation declared as outliers depending on the heavy-tailness of data

# Simulation analysis - long-tailed data

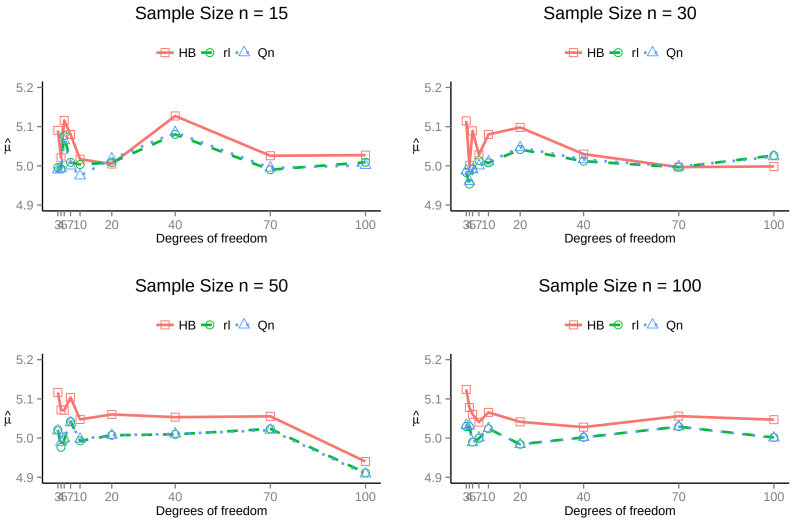

Figure 7 : Average estimates of weighted growth rates depending on the heavy-tailness of data

# Simulation analysis - long-tailed data

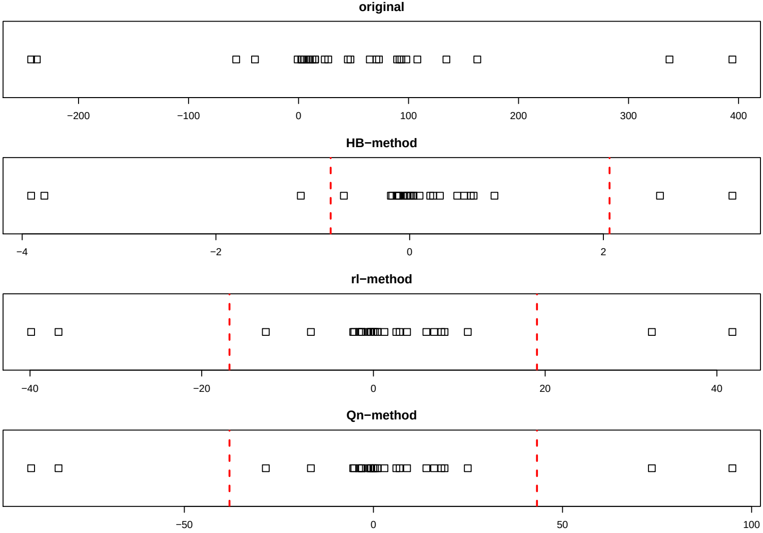

Figure 8 : Original weighted growth rates and test statistics of each method with acceptance boundaries for an example data set $n = 30$

# Conclusions

- We show that the HB-method (with fixed parameter $C$ value) can be too sensitive in case of asymmetric respectively right-skewed data and detecting too many outliers in the left tail of data.
- The other two newly introduced approaches (rl- and Qn-approaches) seem to be able deal better with right-skewed data. As the Qn-approach is easier to understand and simpler to calculate, we prefer this approach as outlier detection method for asymmetric data.
- The HB-method could surely also work better in case of asymmetric data if the parameter $C$ is always optimally determined for each data set depending on the characteristics of the data distribution.
- However, for the Qn-approach there is no need to adjust any parameters, expect the weighting parameter $V$