Robust Statistics: Foundations and Recent Developments

Winter course, CMStatistics 2016

Mia Hubert, Peter Rousseeuw, Stefan Van Aelst

Department of Mathematics
KU Leuven, Belgium

December 6–7, 2016

General references

- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A. *Robust Statistics: the Approach based on Influence Functions.* Wiley Series in Probability and Mathematical Statistics. Wiley, John Wiley and Sons, New York, 1986.
- Rousseeuw, P.J., Leroy, A. *Robust Regression and Outlier Detection*. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, New York, 1987.
- Maronna, R.A., Martin, R.D., Yohai, V.J. *Robust Statistics: Theory and Methods*. Wiley Series in Probability and Statistics. John Wiley and Sons, Chichester, 2006.
- Hubert, M., Rousseeuw, P.J., Van Aelst, S. (2008), High-breakdown robust multivariate methods, *Statistical Science*, 23, 92–119.

wis.kuleuven.be/stat/robust

Outline of the course

- 1. General notions of robustness
- 2. Robustness for univariate data
- 3. Robust multivariate methods
- 4. Robust regression
- 5. Robust principal component analysis
- 6. Inference
- 7. Multivariate and functional depth
- 8. High dimensional data and sparsity
- 9. Cellwise outliers

Session 1: General notions of robustness

Outline:

1. Introduction: outliers and their effect on classical estimators
2. Measures of robustness: breakdown value, sensitivity curve, influence function, gross-error sensitivity, maxbias curve.

What is robust statistics?

Real data often contain outliers. Most classical methods are highly influenced by these outliers.

Robust statistical methods try to fit the model imposed by the **majority** of the data. They aim to find a 'robust' fit, which is similar to the fit we would have found without the outliers.

This allows for **outlier detection**: flag those observations deviating from the robust fit.

What is an outlier? How much is the majority?

Assumptions

- We assume that the majority of the observations satisfy a **parametric** model and we want to estimate the parameters of this model.

  E.g. $x_i \sim N(\mu, \sigma^2)$
  $\boldsymbol{x}_i \sim N_p(\boldsymbol{\mu}, \Sigma)$
  $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma^2)$

- Moreover, we assume that some of the observations might not satisfy this model.
- We do NOT *model* the outlier generating process.
- We do NOT know the *proportion* of outliers in advance.

Example

The classical methods for estimating the parameters of the model may be affected by outliers.

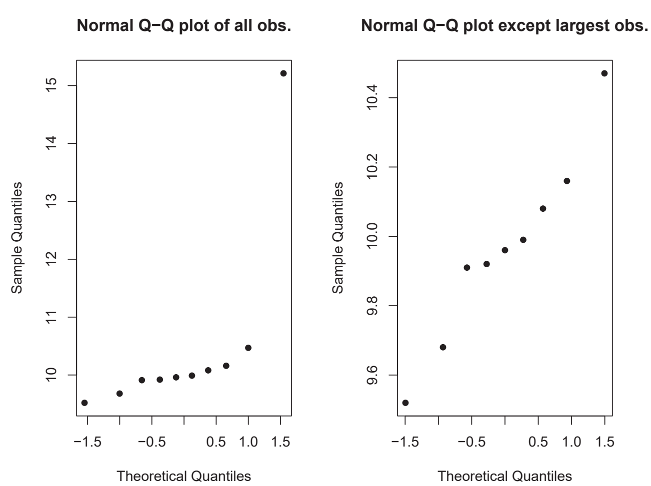
Example. Location-scale model: $x_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, n$.

Data: $X_n = \{x_1, \ldots, x_{10}\}$ are the natural logarithms of the annual incomes (in US dollars) of 10 people.

9.52	9.68	10.16	9.96	10.08
9.99	10.47	9.91	9.92	15.21

Example

The income of person 10 is much larger than the other values.
Normality cannot be rejected for the remaining ('regular') observations:



Classical versus robust estimators

Location:

Classical estimator: arithmetic mean

$$\hat{\mu} = \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Robust estimator: sample median

$$\hat{\mu} = \text{med}(X_n) = \begin{cases} x_{(\frac{n+1}{2})} & \text{if} \quad n \text{ is odd} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & \text{if} \quad n \text{ is even} \end{cases}$$

with $x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$ the ordered observations.

Classical versus robust estimators

Scale:

Classical estimator: sample standard deviation

$$\hat{\sigma} = \text{Stdev}_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}$$

Robust estimator: interquartile range

$$\hat{\sigma} = \text{IQRN}(X_n) = \frac{1}{2\Phi^{-1}(0.75)}(x_{(n-[n/4]+1)} - x_{([n/4])})$$

Classical versus robust estimators

For the data of the example we obtain:

	the 9 regular observations	all 10 observations
$\bar{x}_n$ med	9.97 9.96	10.49 9.98
Stdev_n IQRN	0.27 0.13	1.68 0.17

1. The classical estimators are highly influenced by the outlier
2. The robust estimators are less influenced by the outlier
3. The robust estimate computed from all observations is comparable with the classical estimate applied to the non-outlying data.

Classical versus robust estimators

Robustness: being less influenced by outliers

Efficiency: being precise at uncontaminated data

$\boxed{\text{Robust estimators aim to combine high robustness with high efficiency}}$

Outlier detection

The usual standardized values (z-scores, standardized residuals) are:

$$r_i = \frac{x_i - \bar{x}_n}{\mathsf{Stdev}_n}$$

Classical rule: if $|r_i| > 3$, then observation x_i is flagged as an outlier.

Here: $|r_{10}| = 2.8 \rightarrow$?

Outlier detection based on robust estimates:

$$r_i = \frac{x_i - \mathsf{med}(X_n)}{\mathsf{IQRN}(X_n)}$$

Here: $|r_{10}| = 31.0 \rightarrow$ very pronounced outlier!

MASKING is when actual outliers are not detected.

SWAMPING is when regular observations are flagged as outliers.

Remark

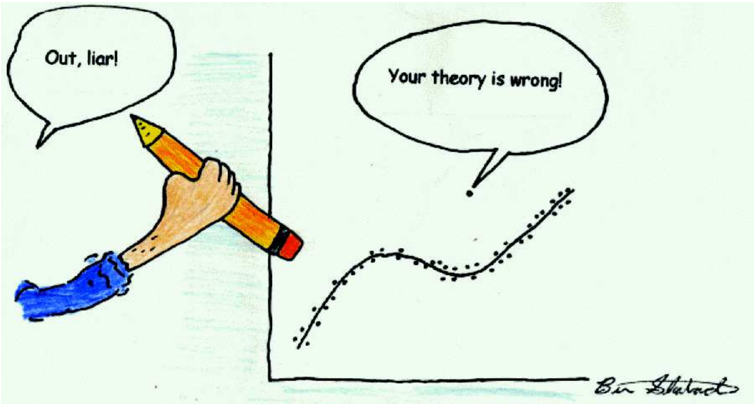
In this example the classical and the robust fits are quite different, and from the robust residuals we see that one of the observations deviates strongly from the others. For the remaining 9 observations a normal model seems appropriate.

It could also be argued that the normal model may not be appropriate itself, and that all 10 observations could have been generated from a single long-tailed or skewed distribution.

We could try to decide which of the two models is more appropriate if we had a much bigger sample. Then we could fit a long-tailed distribution and apply a goodness-of-fit test of that model, and compare it with the goodness-of-fit of the normal model on the non-outlying data.

What is an outlier?

An outlier is an observation that deviates from the fit suggested by the majority of the observations.



How much is the majority?

Some estimators (e.g. the median) already work reasonably well when **50%** or more of the observations are uncontaminated. They thus allow for almost 50% of outliers.

Other estimators (e.g. the IQRN) require that at least **75%** of the observations are uncontaminated. They thus allow for almost 25% of outliers.

This can be measured in general.

Measures of robustness: Breakdown value

Breakdown value (breakdown point) of a location estimator

A data set with n observations is given. If the estimator stays in a fixed *bounded* set even if we replace any $m-1$ of the observations by any outliers, and this is no longer true for replacing any m observations by outliers, then we say that:

the breakdown value of the estimator at that data set is m/n

Notation:

$$\varepsilon_n^*(T_n, X_n) = m/n$$

Typically the breakdown value does not depend much on the data set.
Often it is a fixed constant as long as the original data set satisfies some weak condition, such as the absence of ties.

Breakdown value

Example: $X_n = \{x_1, \ldots, x_n\}$ univariate data, $T_n(X_n) = \text{med}(X_n)$.

Assume n odd, then $T_n = x_{((n+1)/2)}$.

- Replace $\frac{n-1}{2}$ observations by any value, yielding a set X_n^*
  $\Rightarrow$ $T_n(X_n^*)$ always belongs to $[x_{(1)}, x_{(n)}]$, hence $T_n(X_n^*)$ is bounded.
- Replace $\frac{n+1}{2}$ observations by $+\infty$, then $T_n(X_n^*) = +\infty$.
- More precisely, if we replace $\frac{n+1}{2}$ observations by $x_{(n)} + a$, where a is any positive real number, then $T_n(X_n^*) = x_{(n)} + a$.
  Since we can choose a arbitrarily large, $T_n(X_n^*)$ cannot be bounded.

For n odd or even, the (finite-sample) breakdown value ε_n^* of T_n is

$$\varepsilon_n^*(T_n, X_n) = \frac{1}{n}\left[\frac{n+1}{2}\right] \approx 50\% \; .$$

Note that for $n \to \infty$ the finite-sample breakdown value tends to $\varepsilon^* = 50\%$ (which we call the asymptotic breakdown value).
For instance, the arithmetic mean satisfies $\varepsilon_n^*(T_n, X_n) = \frac{1}{n} \to \varepsilon^* = 0\%$.

Breakdown value

A location estimator $\hat{\mu}$ is called **location equivariant** and **scale equivariant** iff

$$\hat{\mu}(aX_n + b) = a\hat{\mu}(X_n) + b$$

for all samples X_n and all $a \neq 0$ and $b \in \mathbb{R}$.

A scale estimator $\hat{\sigma}$ is called **location invariant** and **scale equivariant** iff

$$\hat{\sigma}(aX_n + b) = |a|\hat{\sigma}(X_n) \; .$$

For equivariant location estimators the breakdown value can be at most 50%:

$$\epsilon_n^*(\hat{\mu}, X_n) \leqslant \frac{1}{n}\left[\frac{n+1}{2}\right] \approx 50\% \; .$$

Intuitively: with more than 50% of outliers, the estimator cannot distinguish between the outliers and the regular observations.

Sensitivity curve

The **sensitivity curve** measures the effect of a single outlier on the estimator.

Assume we have $n-1$ fixed observations $X_{n-1} = \{x_1, x_2, \ldots, x_{n-1}\}$.
Now let us see what happens if we add an additional observation equal to x, where x can be any real number.

Sensitivity curve

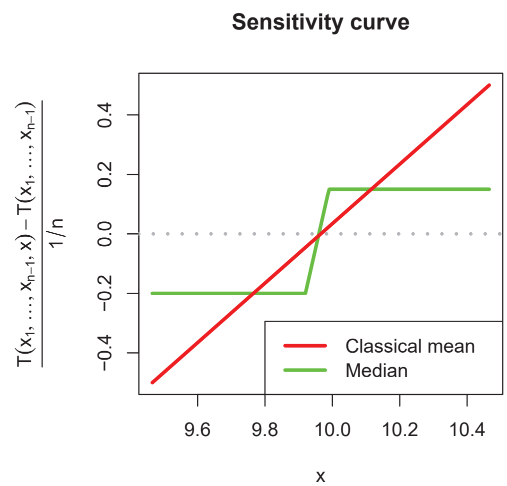
$$\mathrm{SC}(x, T_n, X_{n-1}) = \frac{T_n(x_1, \ldots, x_{n-1}, x) - T_{n-1}(x_1, \ldots, x_{n-1})}{1/n}$$

Example: for the arithmetic mean $T_n = \bar{X}_n$ we find $\mathrm{SC}(x, T_n, X_{n-1}) = x - \bar{x}_{n-1}$.

Note that the sensitivity curve depends strongly on the data set X_{n-1} .

Sensitivity curve: example

Annual income data: let X_9 consist of the 9 'regular' observations.



Influence function

- The influence function is the asymptotic version of the sensitivity curve. It is computed for an estimator T at a certain distribution F, and does not depend on a specific data set.
- For this purpose, the estimator should be written as a function of a distribution F. For example, $T(F) = E_F[X]$ is the functional version of the sample mean, and $T(F) = F^{-1}(0.5)$ is the functional version of the sample median.
- The influence function measures how $T(F)$ changes when contamination is added in x. The contaminated distribution is written as

$$F_{\varepsilon,x} = (1-\varepsilon)F + \varepsilon\Delta_x$$

for $\varepsilon > 0$, where Δ_x is the distribution that puts all its mass in x.

Influence function

Influence function

$$\mathsf{IF}(x,T,F)=\lim_{\varepsilon\to 0}\frac{T(F_{\varepsilon,x})-T(F)}{\varepsilon}=\frac{\partial}{\partial\varepsilon}T(F_{\varepsilon,x})\,|_{\varepsilon=0}$$

Example: for the arithmetic mean $T(F)=E_F[X]$ at a distribution F with finite first moment:

$$\mathsf{IF}(x,T,F)=\frac{\partial}{\partial\varepsilon}E[(1-\varepsilon)F+\varepsilon\Delta_x]\,|_{\varepsilon=0}$$
$$=\frac{\partial}{\partial\varepsilon}[\varepsilon x+(1-\varepsilon)T(F)]\,|_{\varepsilon=0}=x-T(F)$$

At the standard normal distribution $F=\Phi$ we find $\mathsf{IF}(x,T,\Phi)=x$.

We prefer estimators that have a *bounded* influence function.

Gross-error sensitivity

Gross-error sensitivity

$$\gamma^*(T,F)=\sup_x|\mathsf{IF}(x,T,F)|$$

We prefer estimators with a fairly small sensitivity (not just finite).

Asymptotic variance

For asymptotically normal estimators, the asymptotic variance is given by

$$V(T,F)=\int\mathsf{IF}(x,T,F)^2dF(x)$$

under some regularity conditions.

We would like estimators with a small $\gamma^*(T,F)$ but at the same time a small $V(T,F)$, i.e., a high statistical efficiency.

Maxbias curve

The influence function measures the effect of a single outlier, whereas the breakdown value says how many outliers are needed to completely destroy the estimator. These tools thus reflect opposite extremes.

We would also like to know what happens in between, i.e. when there is more than one outlier but not enough to break down the estimator. For any fraction ε of outliers, we consider the maximal bias that can be attained.

Maxbias curve

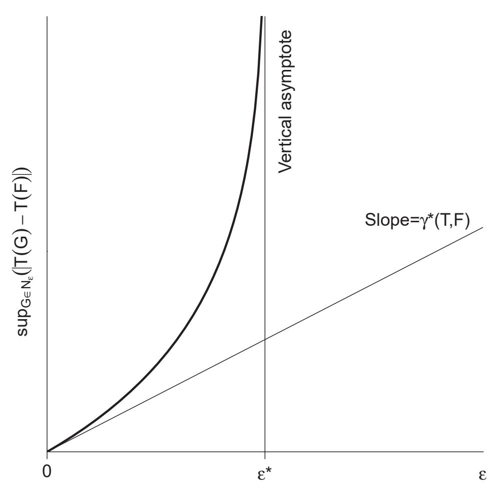
$$\text{maxbias}(\varepsilon, T, F) = \sup_{G \in N_\varepsilon} |T(G) - T(F)|$$

with the 'neighborhood' $N_\varepsilon = \{(1-\varepsilon)F + \varepsilon H;\ H \text{ is any distribution}\}$.

The maxbias curve is useful to compare estimators with the same breakdown value. For the median at the standard normal distribution we obtain $\text{maxbias}(\varepsilon, \text{med}, \Phi) = \Phi^{-1}(1/(2-2\varepsilon))$ which is plotted on the next slide.

Maxbias curve

This graph combines the maxbias curve, the gross-error sensitivity and the breakdown value.



References (for the entire course)

- Billor, N., Hadi, A., Velleman, P. (2000). Bacon: blocked adaptive computationally efficient outlier nominators, *Computational Statistics & Data Analysis,* 34, 279–298.
- Brys, G., Hubert, M., Struyf, A. (2004). A robust measure of skewness, *Journal of Computational and Graphical Statistics,* 13, 996–1017.
- Croux, C., Haesbroeck, G. (2000). Principal components analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies, *Biometrika*, 87, 603–618.
- Croux, C., Ruiz-Gazen, A. (2005). High breakdown estimators for principal components: the projection-pursuit approach revisited, *Journal of Multivariate Analysis,* 95, 206–226.
- Devlin, S.J., Gnanadesikan, R., Kettenring, J.R. (1981). Robust estimation of dispersion matrices and principal components, *Journal of the American Statistical Association,* 76, 354–362.
- Donoho, D.L. (1982). Breakdown properties of multivariate location estimators, Ph.D. thesis, Harvard University.
- Fritz, H., Filzmoser, P., Croux, C. (2012). A comparison of algorithms for the multivariate L_1-median, *Computational Statistics,* 27, 393–410.

References

- Hubert, M., Rousseeuw, P.J. (1996). Robust regression with both continuous and binary regressors, *Journal of Statistical Planning and Inference,* 57, 153–163.
- Hubert, M., Rousseeuw, P.J., Vakili, K. (2014). Shape bias of robust covariance estimators: an empirical study. *Statistical Papers,* 55, 15-28.
- Hubert, M., Rousseeuw, P.J., Vanden Branden, K. (2005). ROBPCA: a new approach to robust principal components analysis, *Technometrics*, 47, 64–79.
- Hubert, M., Rousseeuw, P.J., Verboven, S. (2002). A fast robust method for principal components with applications to chemometrics, *Chemometrics and Intelligent Laboratory Systems,* 60, 101–111.
- Hubert, M., Rousseeuw, P.J., Verdonck, T. (2012). A deterministic algorithm for robust location and scatter, *Journal of Computational and Graphical Statistics,* 21, 618–637.
- Hubert, M., Vandervieren, E. (2008). An adjusted boxplot for skewed distributions, *Computational Statistics and Data Analysis,* 52, 5186–5201.
- Liu, R. (1990). On a notion of data depth based on random simplices, *The Annals of Statistics,* 18, 405–414.

References

- Locantore, N., Marron, J.S., Simpson, D.G., Tripoli, N., Zhang, J.T., Cohen, K.L. (1999). Robust principal component analysis for functional data, *Test*, 8, 1–73.
- Maronna, R.A. and Yohai, V.J. (2000). Robust regression with both continuous and categorical predictors, *Journal of Statistical Planning and Inference,* 89, 197–214.
- Maronna, R.A., Zamar, R.H. (2002). Robust estimates of location and dispersion for high-dimensional data sets, *Technometrics,* 44, 307–317.
- Oja, H. (1983). Descriptive statistics for multivariate distributions, *Statistics and Probability Letters,* 1, 327–332.
- Rousseeuw, P.J. (1984). Least median of squares regression, *Journal of the American Statistical Association,* 79, 871–880.
- Rousseeuw, P.J., Croux, C. (1993). Alternatives to the median absolute deviation, *Journal of the American Statistical Association,* 88, 1273–1283.
- Rousseeuw, P.J., Van Driessen, K. (1999). A fast algorithm for the Minimum Covariance Determinant estimator, *Technometrics,* 41, 212–223.
- Rousseeuw, P.J., van Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association,* 85, 633–651.

References

- Rousseeuw, P.J., Yohai, V.J. (1984). Robust regression by means of S-estimators, in *Robust and Nonlinear Time Series Analysis,* edited by J. Franke, W. Härdle and R.D. Martin. Lecture Notes in Statistics No. 26, Springer, New York, 256–272.
- Salibian-Barrera, M., Yohai, V.J. (2006). A fast algorithm for S-regression estimates, *Journal of Computational and Graphical Statistics,* 15, 414–427.
- Stahel, W.A. (1981). Robuste Schätzungen: infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen, Ph.D. thesis, ETH Zürich.
- Tatsuoka, K.S., Tyler, D.E. (2000). On the uniqueness of S-functionals and M-functionals under nonelliptical distributions, *The Annals of Statistics,* 28, 1219–1243.
- Tukey, J.W. (1975). Mathematics and the Picturing of Data, *Proceedings of the International Congress of Mathematicians,* Vancouver,2, 523–531.
- Visuri, S., Koivunen, V., Oja, H. (2000). Sign and rank covariance matrices, *Journal of Statistical Planning and Inference,* 91, 557–575.
- Yohai, V.J. (1987). High breakdown point and high efficiency robust estimates for regression, *The Annals of Statistics,* 15, 642–656.
- Yohai, V.J., Maronna, R.A. (1990). The maximum bias of robust covariances, *Communications in Statistics–Theory and Methods,* 19, 3925–2933.