Robust Statistics
Part 1: Introduction and univariate data

Peter Rousseeuw

LARS-IASC School, May 2019

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

- General notions of robustness
- Robustness for univariate data
- Multivariate location and scatter
- Linear regression
- Principal component analysis
- Advanced topics

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.

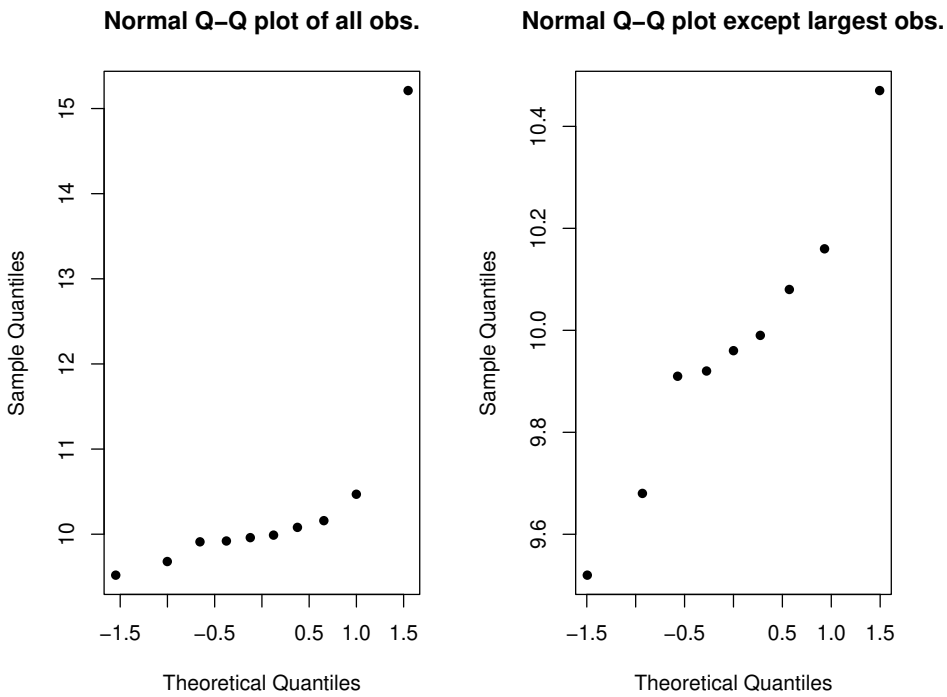
Example. Location-scale model: $x_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, n$.

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

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

Classical versus robust estimators

For the data of the example we obtain:

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

Robust estimators aim to combine high robustness with high efficiency

Outlier detection

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

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

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:

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

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

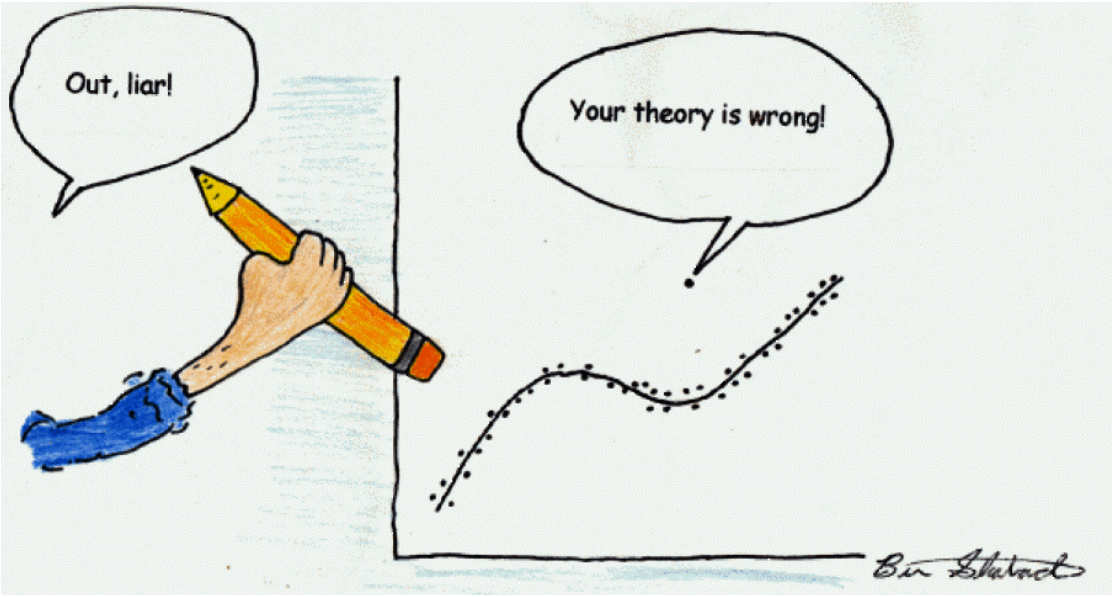
In this example the classical and the robust fits are quite different, and from the robust residuals we see that one of the observations deviates strongly from the others. For the remaining 9 observations a normal model seems appropriate.

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

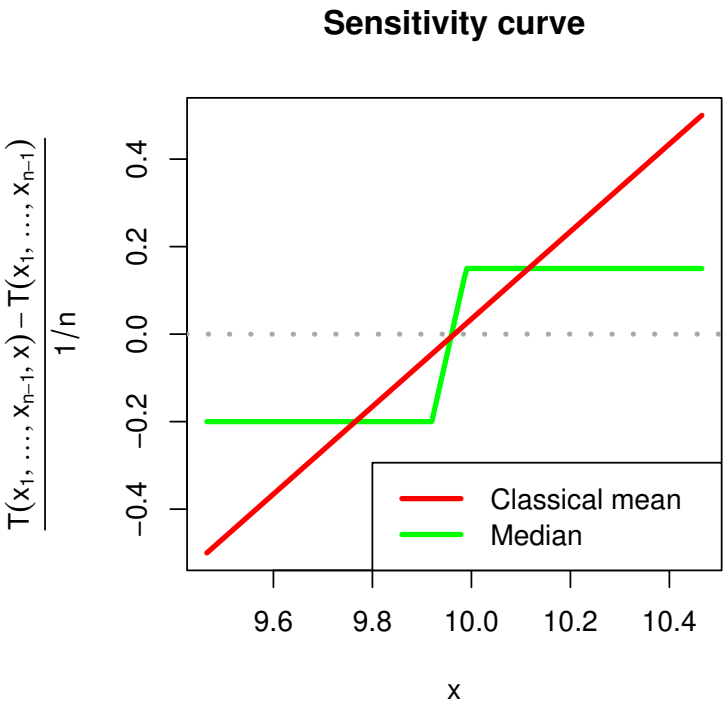
$$\mathrm{SC}(x, T_n, X_{n-1}) = \frac{T_n(x_1, \ldots, x_{n-1}, x) - T_{n-1}(x_1, \ldots, x_{n-1})}{1/n}$$

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.



Mechanical analogy

How do the concepts of breakdown value and sensitivity curve differ?
From Galilei (1638):

Effect of a small weight is linear: Hooke's law (sensitivity).
Effect of a large weight is nonlinear (breakdown).

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

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

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

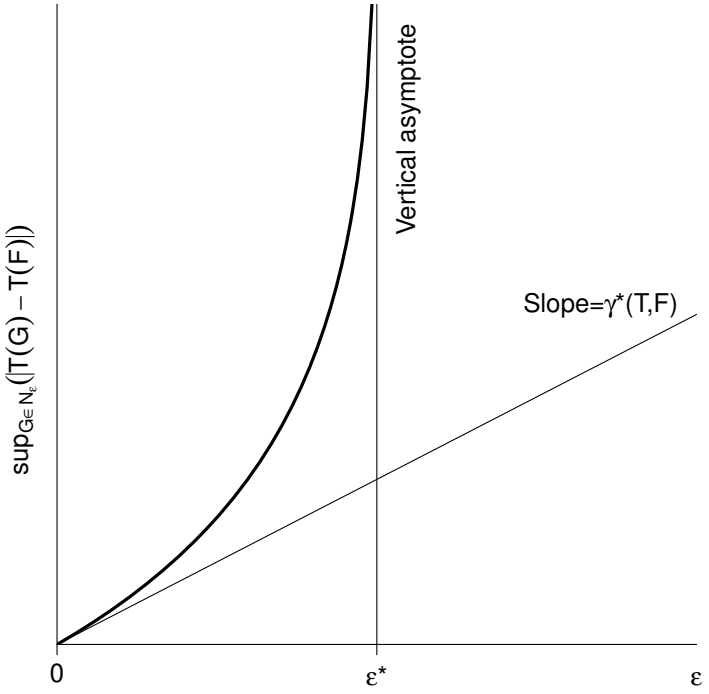
$$\text{maxbias}(\varepsilon,T,F) = \sup_{G\in N_\varepsilon} |T(G) - T(F)|$$

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.



Robustness for univariate data: Outline

1. Location only:
   - explicit location estimators
   - M-estimators of location
2. Scale only:
   - explicit scale estimators
   - M-estimators of scale
3. Location and scale combined
4. Measures of skewness

The pure location model

Assume that $x_1, \ldots, x_n$ are independent and identically distributed (i.i.d.) as

$$\boxed{F_\mu(x) = F(x - \mu)}$$

where $-\infty < \mu < +\infty$ is the unknown location parameter and F is a continuous distribution with density f, hence $f_\mu(x) = F'_\mu(x) = f(x - \mu)$.

Often f is assumed to be symmetric. A typical example is the standard normal (gaussian) distribution Φ with density ϕ.

We say that a location estimator T is **Fisher-consistent** at this model iff

$$\boxed{T(F_\mu) = \mu \qquad \text{for all } \mu.}$$

Note that F_μ is only a model for the uncontaminated data. We do not model outliers.

Some explicit location estimators

1. **Median**
2. **Trimmed mean**: ignore the m smallest and the m largest observations and just take the average of the observations in between:

$$\hat{\mu}_{TM} = \frac{1}{n - 2m} \sum_{i=m+1}^{n-m} x_{(i)}$$

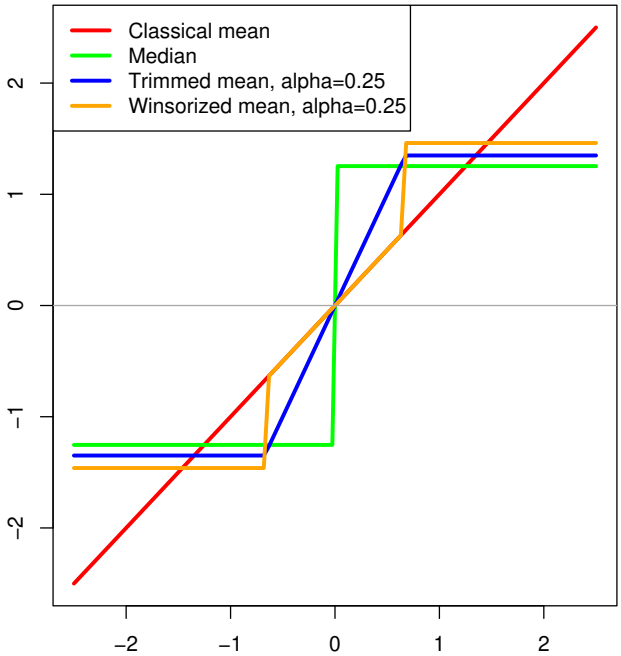
   with $m = [(n-1)\alpha]$ and $0 \leqslant \alpha < 0.5$.
   For $\alpha = 0$ this is the mean, and for $\alpha \to 0.5$ this becomes the median.

3. **Winsorized mean**: replace the m smallest observations by $x_{(m+1)}$ and the m largest observations by $x_{(n-m)}$. Then take the average:

$$\hat{\mu}_{WM} = \frac{1}{n} \left(m x_{(m+1)} + \sum_{i=m+1}^{n-m} x_{(i)} + m x_{(n-m)} \right)$$

Robustness properties

Breakdown value: $\varepsilon_n^*(\text{med}) \to 0.5$; $\varepsilon_n^*(\hat{\mu}_{TM}) = \varepsilon_n^*(\hat{\mu}_{WM}) = (m+1)/n \to \alpha$.
Maxbias: For any ε, the median achieves the smallest maxbias among all location equivariant estimators.
Influence function at the normal model:



Implicit location estimators

The location model says that $F_\mu(x) = F(x-\mu)$ with unknown μ.

The **maximum likelihood estimator (MLE)** therefore satisfies

$$\begin{aligned}\hat{\mu}_{\text{MLE}} &= \underset{\mu}{\text{argmax}} \prod_{i=1}^{n} f(x_i - \mu) \\ &= \underset{\mu}{\text{argmax}} \sum_{i=1}^{n} \log f(x_i - \mu) \\ &= \underset{\mu}{\text{argmin}} \sum_{i=1}^{n} -\log f(x_i - \mu)\end{aligned}$$

For $f = \phi$ (standard normal), this yields $\hat{\mu}_{\text{MLE}} = \bar{x}_n$.

For $f(x) = \frac{1}{2}e^{-|x|}$ (Laplace distribution), this yields $\hat{\mu}_{\text{MLE}} = \text{med}(X_n)$.

For most f the MLE has no explicit formula.

M-estimators of location

Let $\rho(x)$ be an even function, weakly increasing in $|x|$, with $\rho(0) = 0$.

M-estimator of location

$$\hat{\mu}_M = \underset{\mu}{\text{argmin}} \sum_{i=1}^{n} \rho(x_i - \mu)$$

If ρ is differentiable with $\psi = \rho'$, then $\hat{\mu}_M$ satisfies:

$$\boxed{\sum_{i=1}^{n} \psi(x_i - \hat{\mu}_M) = 0} \tag{1}$$

If ψ is discontinuous, we take $\hat{\mu}_M$ as the μ where $\sum_{i=1}^{n} \psi(x_i - \mu)$ changes sign.

Note that the MLE is an M-estimator, with $\rho(x) = -\log f(x)$ and $\psi(x) = \rho'(x) = -f'(x)/f(x)$. For $F = \Phi$, $\psi(x) = -\phi'(x)/\phi(x) = x$.

Some often used ρ functions

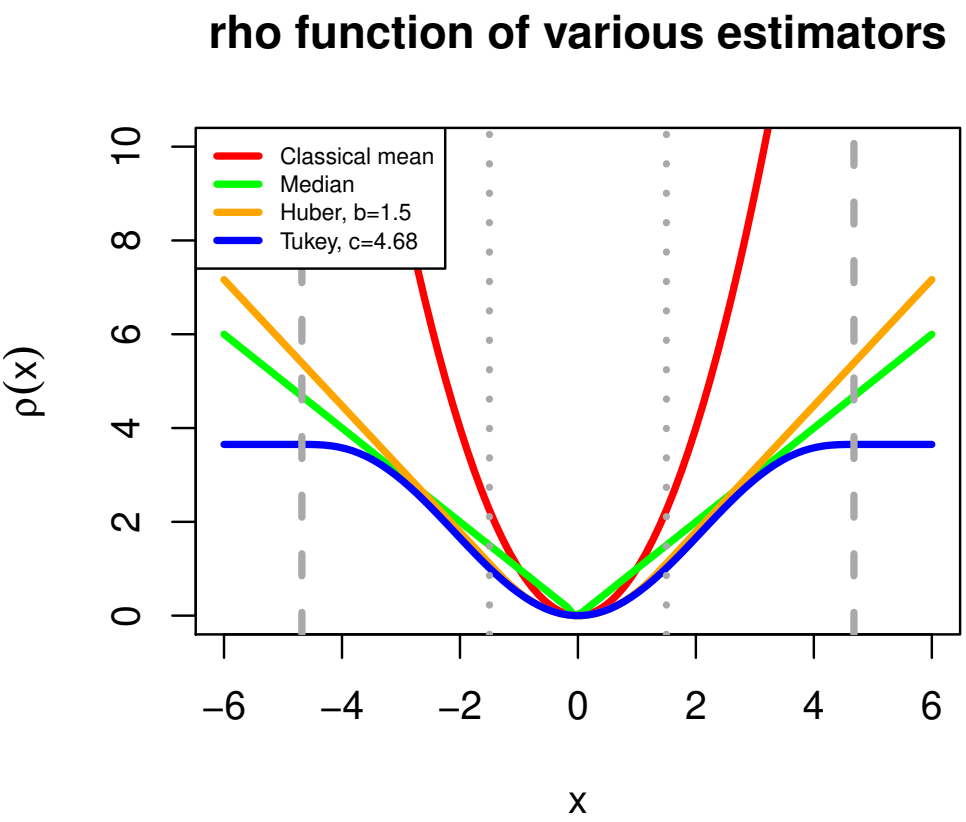
- **Mean**: $\rho(x) = x^2/2$
- **Median**: $\rho(x) = |x|$
- **Huber**:

$$\rho_b(x) = \begin{cases} x^2/2 & \text{if} \quad |x| \leqslant b \\ b|x| - b^2/2 & \text{if} \quad |x| > b \end{cases}$$

- **Tukey's bisquare**:

$$\rho_c(x) = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4} & \text{if} \quad |x| \leqslant c \\ \frac{c^2}{6} & \text{if} \quad |x| > c \end{cases} \tag{2}$$

Some often used ρ functions



The corresponding score functions $\psi = \rho'$

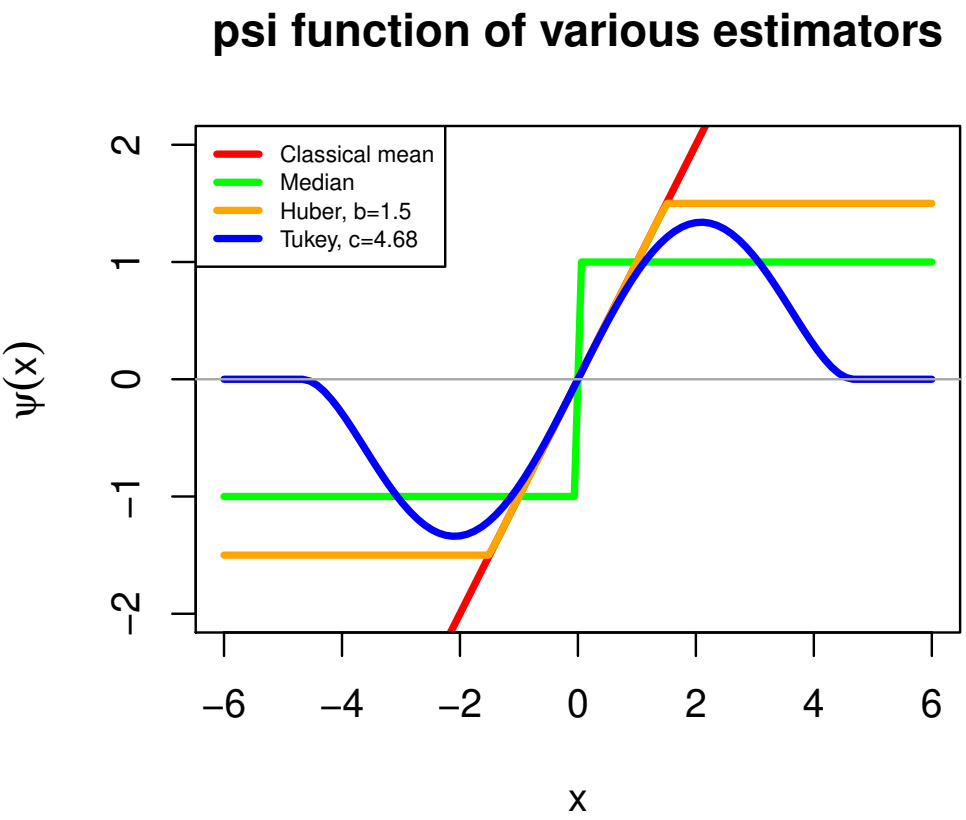
- **Mean**: $\psi(x) = x$
- **Median**: $\psi(x) = \text{sign}(x)$
- **Huber**:

$$\psi_b(x) = \begin{cases} x & \text{if} \quad |x| \leqslant b \\ b\ \text{sign}(x) & \text{if} \quad |x| > b \end{cases}$$

- **Tukey's bisquare**:

$$\psi_c(x) = \begin{cases} x\left(1 - \frac{x^2}{c^2}\right)^2 & \text{if} \quad |x| \leqslant c \\ 0 & \text{if} \quad |x| > c \end{cases}$$

The corresponding score functions $\psi = \rho'$



Properties of location M-estimators

- Fisher-consistent iff $\int \psi(x) dF(x) = 0$.
- Influence function:

$$\mathrm{IF}(x, T, F) = \frac{\psi(x)}{\int \psi'(y) dF(y)}$$

  The influence function of an M-estimator is proportional to its ψ-function. A bounded ψ-function thus leads to a bounded IF.

- Asymptotically normal with asymptotic variance

$$V(T, F) = \int \mathrm{IF}(x, T, F)^2 dF(x) = \frac{\int \psi^2(x) dF(x)}{(\int \psi'(y) dF(y))^2}$$

- By the information inequality, the asymptotic variance satisfies

$$V(T, F) \geqslant \frac{1}{I(F)}$$

  where $I(F) = \int (-f'(x)/f(x))^2 dF(x)$ is the Fisher information of the model.

Properties of location M-estimators

- The asymptotic efficiency of an estimator T at the model distribution F is defined as
$$\text{eff} = \frac{1}{V(T,F)I(F)}$$
so by the information inequality it lies between 0 and 1.

- The Fisher information of the normal location model is 1, so the asymptotic efficiency is eff $= 1/V(T,F)$. For different choices of the tuning constants we obtain the following efficiencies:

Huber: $b = 1.345$ gives eff $= 95\%$
$b = 1.5$ gives eff $= 96.5\%$
$b \to 0$ (median) gives eff $= 64\%$

Bisquare: $c = 4.68$ gives eff $= 95\%$
$c = 3.14$ gives eff $= 80\%$

Properties of location M-estimators

- Breakdown value: 50% if ψ is bounded.
Note that it does not depend on the tuning parameter (b or c).

- Maxbias curve: does grow with the tuning parameter.

- The Huber M-estimator has a **monotone** ψ-function, hence:
  - unique solution for (1)
  - large outliers still affect the estimate, but the effect remains bounded.

- The bisquare M-estimator has a **redescending** ψ-function, hence:
  - no unique solution for (1)
  - the effect of large outliers on the estimate reduces to zero.

Remarks

- The trimmed mean and the Huber M-estimator have the same IF, and thus the same asymptotic efficiency, when

$$b = \frac{F^{-1}(1-\alpha)}{1-2\alpha}$$

  For instance, for $\alpha = 0.25$ we obtain $b = 1.349$ and eff $= 95\%$.

  But the Huber M-estimator has a 50% breakdown value, whereas the 25%-trimmed mean only has a 25% breakdown value.

- M-estimators of location are NOT scale equivariant. We will see later that we can make them scale equivariant by incorporating a scale estimate as well.

The pure scale model

The scale model assumes that the data are i.i.d. according to:

$$\boxed{F_\sigma(x) = F\left(\frac{x}{\sigma}\right)}$$

where $\sigma > 0$ is the unknown scale parameter. As before F is a continuous distribution with density f, but now

$$f_\sigma(x) = F'_\sigma(x) = \frac{1}{\sigma} f(\frac{x}{\sigma}) \ .$$

We say that a scale estimator S is Fisher-consistent at this model iff

$$S(F_\sigma) = \sigma \qquad \text{for all } \sigma > 0 \ .$$

Robustness measures of scale estimators

- The influence function is defined as for any other estimator.
- The breakdown value of a scale estimator is defined as the minimum of the *explosion breakdown value* and the *implosion breakdown value*.

  Explosion is when the scale estimate is inflated ($\hat{\sigma} \to \infty$).
  The classical standard deviation can explode due to a single far outlier.

  Implosion is when the scale estimate becomes arbitrarily small ($\hat{\sigma} \to 0$), which would be a problem because scale estimates often occur in the denominator of a statistic (such as the z-score).

  For equivariant scale estimators the breakdown value is at most 50%:

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

- Analogously, we can compute two maxbias curves: one for implosion, and one for explosion.

Explicit scale estimators

Some explicit scale estimators:

1. **Standard deviation (Stdev)** Not robust.
2. **Interquartile range**

$$\text{IQR}(X_n) = x_{(n-[n/4]+1)} - x_{([n/4])}$$

However, at $F_\sigma = N(0, \sigma^2)$ it holds that $IQR(F_\sigma) = 2\Phi^{-1}(0.75)\sigma \neq \sigma$.

Normalized IQR:

$$\text{IQRN}(X_n) = \frac{1}{2\Phi^{-1}(0.75)}\,\text{IQR}(X_n) \ .$$

The constant $1/2\Phi^{-1}(0.75) = 0.7413$ is a *consistency factor*.

When using software, it should be checked whether the consistency factor is included or not!

Explicit scale estimators

Estimators with 50% breakdown value:

3. **Median absolute deviation**

$$\text{MAD}(X_n) = \underset{i}{\text{med}}(|x_i - \text{med}(X_n)|)$$

At any symmetric sample it holds that IQR $= 2\,$MAD.

At the normal model we use the normalized version:

$$\text{MADN}(X_n) = \frac{1}{\Phi^{-1}(0.75)}\,\text{MAD}(X_n) = 1.4826\,\text{MAD}(X_n)$$

Explicit scale estimators

4. Q_n **estimator** (Rousseeuw and Croux, 1993)

$$Q_n = 2.219\{|x_i - x_j|; i < j\}_{(k)}$$

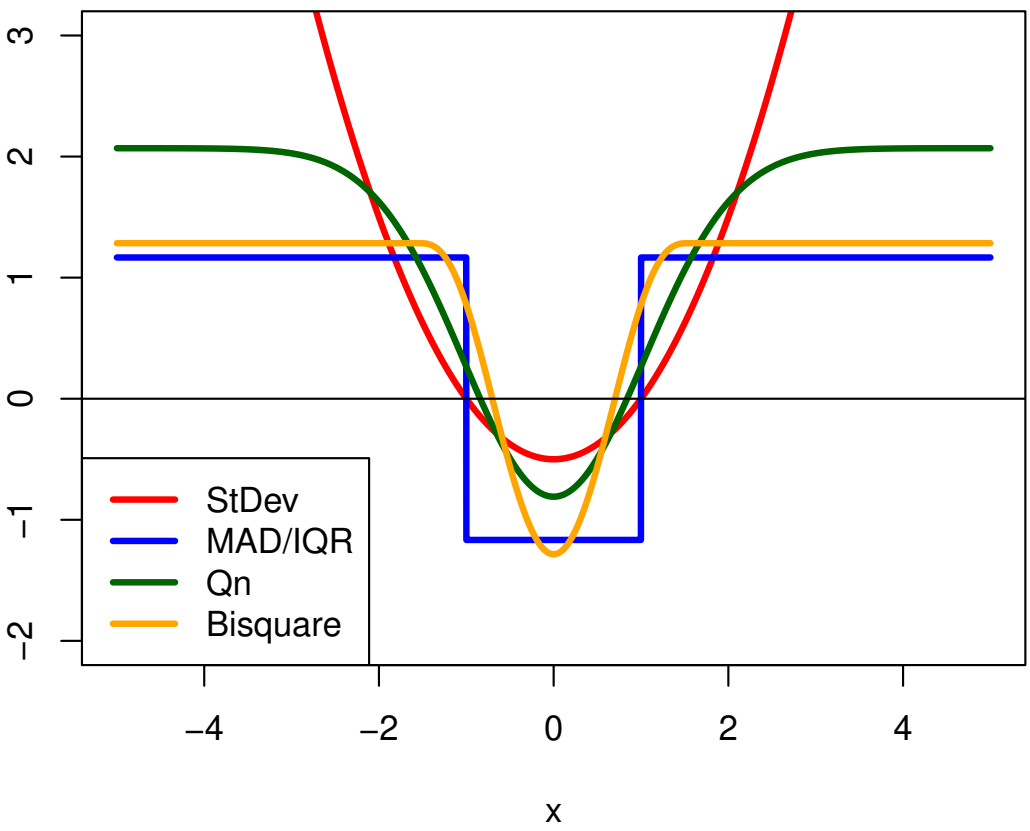
with $k = \binom{h}{2} \approx \binom{n}{2}/4$ and $h = [\frac{n}{2}] + 1$.

Q_n does not depend on an initial location estimate!

Its breakdown value is 50% .

Despite appearances, Q_n can be computed in O$(n \log n)$ time.

Influence function of various scale estimators



Explicit scale estimators

Robustness and efficiency at the normal model:

	ε^*	γ^*	eff
Stdev	0%	∞	100%
IQRN	25%	1.167	37%
MADN	50%	1.167	37%
Q_n	50%	2.069	82%

Note that IQRN and MADN have the same influence function, but that the breakdown value of MADN is twice as high as that of IQRN. We thus prefer MADN over IQRN. Also note that Q_n has a higher efficiency.

MLE estimator of scale

The maximum likelihood estimator (MLE) of σ satisfies

$$\hat{\sigma}_{\text{MLE}} = \underset{\sigma}{\text{argmax}} \prod_{i=1}^{n} \frac{1}{\sigma} f(\frac{x_i}{\sigma})$$
$$= \underset{\sigma}{\text{argmax}} \sum_{i=1}^{n} \left\{ -\log(\sigma) + \log f(\frac{x_i}{\sigma}) \right\}$$

Zeroing the derivative with respect to σ yields:

$$\sum_{i=1}^{n} \left\{ -\frac{1}{\sigma} + \frac{f'(\frac{x_i}{\sigma})}{f(\frac{x_i}{\sigma})} \frac{-x_i}{\sigma^2} \right\} = 0$$

$$\sum_{i=1}^{n} -\frac{f'(\frac{x_i}{\sigma})}{f(\frac{x_i}{\sigma})} \frac{x_i}{\sigma} = n$$

$$\frac{1}{n} \sum_{i=1}^{n} -\frac{x_i}{\hat{\sigma}} \frac{f'(x_i/\hat{\sigma})}{f(x_i/\hat{\sigma})} = 1 \ .$$

MLE estimator of scale

We can rewrite this last expression as

$$\frac{1}{n} \sum_{i=1}^{n} \rho \left(\frac{x_i}{\hat{\sigma}} \right) = 1$$

if we put

$$\rho(t) = -t \frac{f'(t)}{f(t)} \ .$$

If $f = \phi$, then $\rho(t) = t^2$ and $\hat{\sigma}_{MLE} = \sqrt{\sum_{i=1}^{n} x_i^2 / n}$ (the root mean square).
If $f = \frac{1}{2} e^{-|x|}$ (Laplace), then $\rho(t) = |t|$ and $\hat{\sigma}_{MLE} = \sum_{i=1}^{n} |x_i| / n$.

For most other densities f there is no explicit formula for $\hat{\sigma}_{MLE}$.

We can now generalize the formula above to a function ρ that was not obtained from the density of a model distribution.

M-estimators of scale

Let $\rho(x)$ be an even function, weakly increasing in $|x|$, with $\rho(0) = 0$.

M-estimator of scale

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{x_i}{\hat{\sigma}_M}\right) = \delta$$

The constant δ is usually taken as

$$\delta = \int \rho(t)dF(t)$$

to obtain Fisher-consistency at the model F_σ .

The breakdown value of an M-estimator of scale is

$$\varepsilon^*(\hat{\sigma}_M) = \min\left(\varepsilon^*_{expl}, \varepsilon^*_{impl}\right) = \min\left(\frac{\delta}{\rho(\infty)}, 1 - \frac{\delta}{\rho(\infty)}\right)$$

so it is 0% for unbounded ρ and 50% for a bounded ρ with $\delta = \rho(\infty)/2$.

Properties of M-estimators of scale

- At the model distribution F we have $\hat{\sigma} = 1$ by Fisher-consistency, and

  $$\mathrm{IF}(x, T, F) = \frac{\rho(x) - \delta}{\int y\rho'(y)dF(y)}$$

  The influence function of an M-estimator is proportional to $\rho(x) - \delta$.
  A bounded ρ-function thus leads to a bounded IF.

- Asymptotically normal with asymptotic variance

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

- By the information inequality, the asymptotic variance satisfies

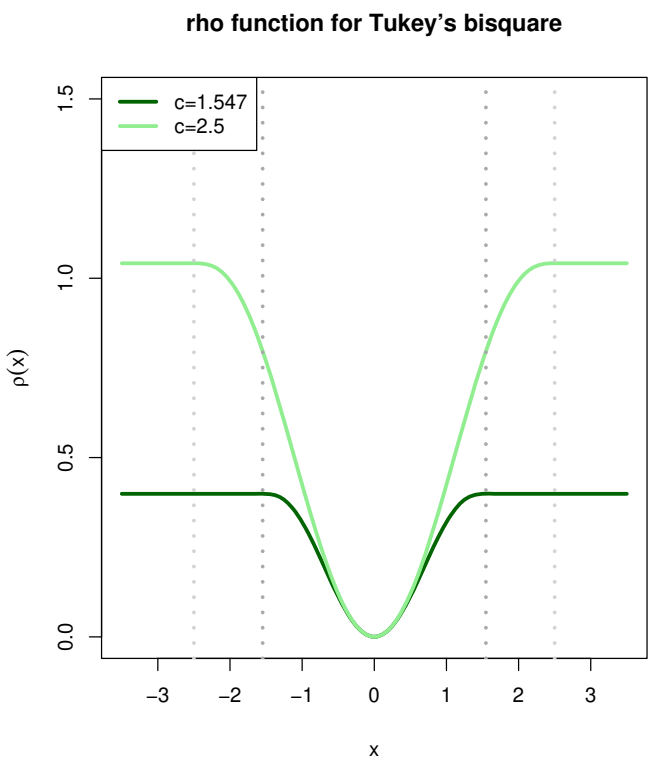
  $$V(T, F) \geqslant \frac{1}{I(F)}$$

  where $I(F) = \int(-1 - \frac{xf'(x)}{f(x)})^2 dF(x)$ is the Fisher information of the scale model. For $F = \Phi$ we find $I(F) = 2$ and $\mathrm{IF}(x; \mathrm{MLE}, \Phi) = (x^2 - 1)/2$.

From standard deviation to MAD

Bisquare M-estimator of scale

A popular choice for ρ is the bisquare function (2).
The maximal breakdown value of 50% is achieved at $c = 1.547$.



Model with both location and scale unknown

The general location-scale model assumes that the x_i are i.i.d. according to

$$\boxed{F_{(\mu,\sigma)}(x) = F\left(\frac{x-\mu}{\sigma}\right)}$$

where $-\infty < \mu < +\infty$ is the location parameter and $\sigma > 0$ is the scale parameter. In this general model, both μ and σ are assumed to be unknown which is realistic. The density is now

$$f_{(\mu,\sigma)}(x) = F'_{(\mu,\sigma)}(x) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) .$$

In this general situation we can still estimate location and scale by means of the explicit estimators we saw for the pure location model (Median, trimmed mean, and winsorized mean) and the pure scale model (IQRN, MADN, and Q_n).

Model with both location and scale unknown

Note that the location M-estimators we saw before, given by

$$\hat{\mu}_M = \underset{\mu}{\text{argmin}} \sum_{i=1}^{n} \rho(x_i - \mu)$$

are not scale equivariant. But we can define a scale equivariant version by

$$\hat{\mu}_M = \underset{\mu}{\text{argmin}} \sum_{i=1}^{n} \rho\left(\frac{x_i - \mu}{\hat{\sigma}}\right)$$

where $\hat{\sigma}$ is a robust scale estimate that we compute beforehand. The robustness of the end result depends on how robust $\hat{\sigma}$ is, so it is best to use a scale estimator with high breakdown value such as Q_n.

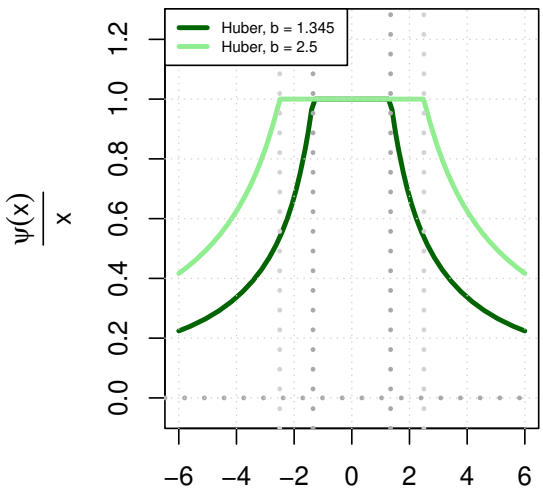
For instance, a location M-estimator with monotone and bounded ψ-function (say, the Huber ψ with $b = 1.5$) and with $\hat{\sigma} = Q_n$ attains a 50% breakdown value, which is the highest possible.

An algorithm for location M-estimators

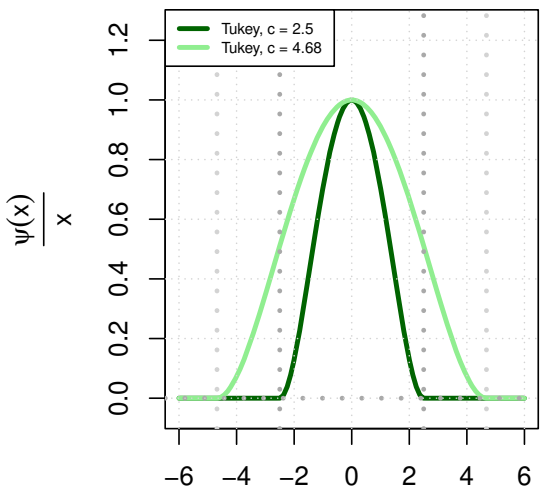
Based on $\psi = \rho'$ we define the weight function

$$W(x) = \begin{cases} \psi(x)/x & \text{if} \quad x \neq 0 \\ \psi'(0) & \text{if} \quad x = 0. \end{cases}$$

weight functions for Huber



weight functions for Tukey's bisquare



An algorithm for location M-estimators

Using this function $W(x) = \psi(x)/x$, the estimating equation $\sum_{i=1}^n \psi\left(\frac{x_i - \hat{\mu}_M}{\hat{\sigma}}\right) = 0$ can be rewritten as

$$\sum_{i=1}^n \frac{x_i - \hat{\mu}_M}{\hat{\sigma}} \, W\left(\frac{x_i - \hat{\mu}_M}{\hat{\sigma}}\right) = 0$$

or equivalently

$$\hat{\mu}_M = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

with weights $w_i = W((x_i - \hat{\mu}_M)/\hat{\sigma})$, so we can see the location M-estimator $\hat{\mu}_M$ as a weighted mean of the observations.

But this is still an implicit equation, as the w_i depend on $\hat{\mu}_M$ itself.

An algorithm for location M-estimators

Iterative algorithm:

1. Start with an initial estimate, typically $\hat{\mu}_0 = \text{med}(X_n)$
2. For $k = 0, 1, 2, \ldots,$ set
$$w_{k,i} = W\left(\frac{x_i - \hat{\mu}_k}{\hat{\sigma}}\right)$$
and then compute
$$\hat{\mu}_{k+1} = \frac{\sum_{i=1}^n w_{k,i} x_i}{\sum_{i=1}^n w_{k,i}}$$
3. Stop when $|\hat{\mu}_{k+1} - \hat{\mu}_k| < \epsilon\hat{\sigma}$.

Since each step is a weighted mean, which is a special case of weighted least squares, this algorithm is called **iteratively reweighted least squares (IRLS)**.

For monotone M-estimators, this algorithm is guaranteed to converge to the (unique) solution of the estimating equation.

Algorithms for M-estimators

Remarks:

- IRLS is not the only algorithm for computing M-estimators. One can also use Newton-Raphson steps. Taking a single Newton-Raphson step starting from $\text{med}(X_n)$ yields an estimator by itself, which has good properties.
- Similar algorithms also exist for M-estimators of scale.
- An alternative approach to M-estimation in the location-scale model would be to consider a system of two estimating equations:
$$\sum_{i=1}^n \psi\left(\frac{x_i - \mu}{\sigma}\right) = 0 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^n \rho\left(\frac{x_i - \mu}{\sigma}\right) = \delta$$
and to search for a pair $(\hat{\mu}, \hat{\sigma})$ that solves both equations simultaneously. But we don't do this because it yields less robust estimates!

Example

Applying all these location estimators to the annual income data set yields:

	regular obs.	all obs.
$\bar{x}_n$	9.97	10.49
med	9.96	9.98
trimmed mean, $\alpha = 0.25$	9.97	10.00
Winsorized mean, $\alpha = 0.25$	9.98	10.01
Huber, $b = 1.5$	9.97	10.00
Bisquare, $c = 4.68$	9.96	9.96

Example

Applying the scale estimators to these data:

	regular obs.	all obs.
Stdev	0.27	1.68
IQRN	0.13	0.17
MADN	0.18	0.22
Q_n	0.31	0.37
Huber, $b = 1.5$	0.17	0.19
Bisquare, $c = 4.68$	0.23	0.29

Robust measures of skewness

We know that the third moment is not robust. The **quartile skewness** measure is defined as

$$\frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$$

where Q_1, $Q_2 = \text{med}(X_n)$, and Q_3 are the quartiles of the data. This skewness measure has a 25% breakdown value but is not very 'efficient' in that deviations from symmetry may not be detected well.

Medcouple (MC) (Brys et al., 2004)

$$\text{MC}(X_n) = \text{med}\left(\{h(x_i, x_j);\ x_i < Q_2 < x_j\}\right)$$

with

$$h(x_i, x_j) = \frac{(x_j - Q_2) - (Q_2 - x_i)}{x_j - x_i} \quad .$$

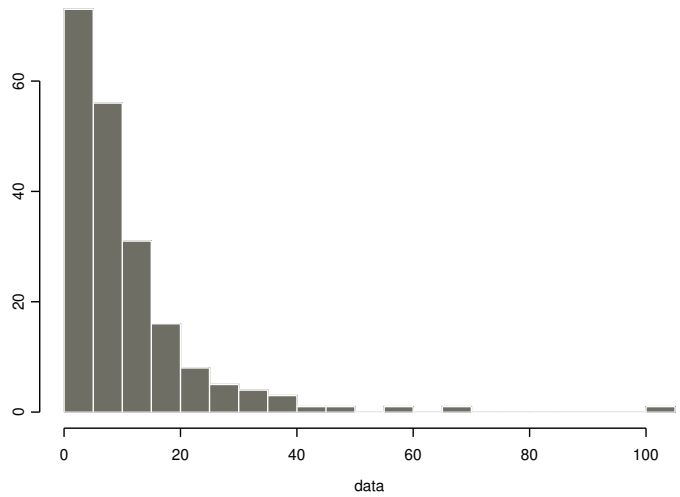
This measure also has $\varepsilon^* = 25\%$ and is more sensitive to asymmetry .

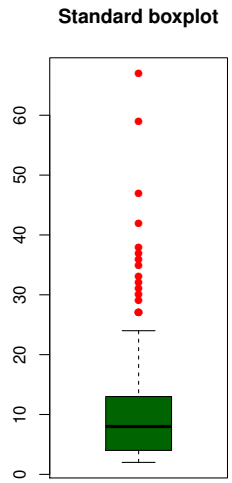
Standard boxplot

The boxplot is a tool of exploratory data analysis. It flags as outliers all points outside the 'fence'

$$[Q_1 - 1.5\ \text{IQR}, Q_3 + 1.5\ \text{IQR}]$$

Example: Length of stay in hospital (in days):





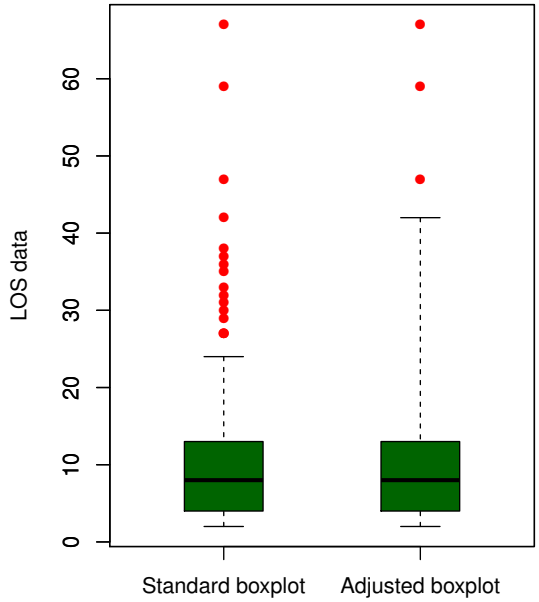
This outlier detection rule is not very accurate at asymmetric data.

Adjusted boxplot

For right-skewed distributions, the fence is now defined as

$$[Q_1 - 1.5\, e^{-4\,\mathrm{MC}}\,\mathrm{IQR}\,,\, Q_3 + 1.5\, e^{3\,\mathrm{MC}}\,\mathrm{IQR}]$$

(Hubert and Vandervieren, 2008).



Software

In the freeware package R:

- Mean, Median: `mean, median`
- trimmed mean: `mean(x,trim=0.25)`
- Winsorized mean: `winsor.mean(x,trim=0.25)` in package *psych*
- Huber's M: `huberM` in package *robustbase*, `hubers` in package *MASS*, `rlm` in package *MASS* [rlm(X ∼ 1,psi=psi.huber)]
- Tukey Bisquare: `rlm` in package *MASS* [rlm(X ∼ 1,psi=psi.bisquare)]
- MADN, IQR: `mad` and `IQR`
- Q_n: function `Qn` in package *robustbase*
- Medcouple: `mc` in package *robustbase*
- adjusted boxplot: `adjbox` in package *robustbase*

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.