Session 2: Robustness for univariate data

Winter course, CMStatistics 2016

Mia Hubert, Peter Rousseeuw, Stefan Van Aelst

Department of Mathematics
KU Leuven, Belgium

December 6–7, 2016

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

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

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

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:

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

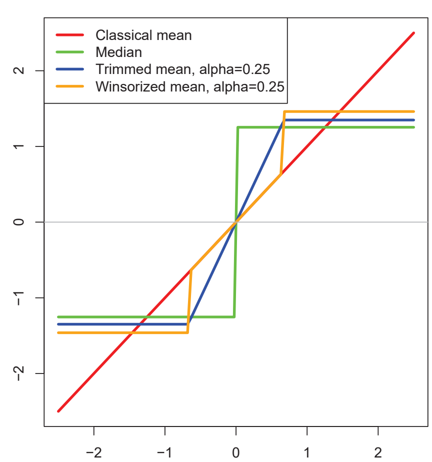
with $m = [(n-1)\alpha]$ and $0 \leqslant \alpha < 0.5$.
For $\alpha = 0$ this is the mean, and for $\alpha \to 0.5$ this becomes the median.

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

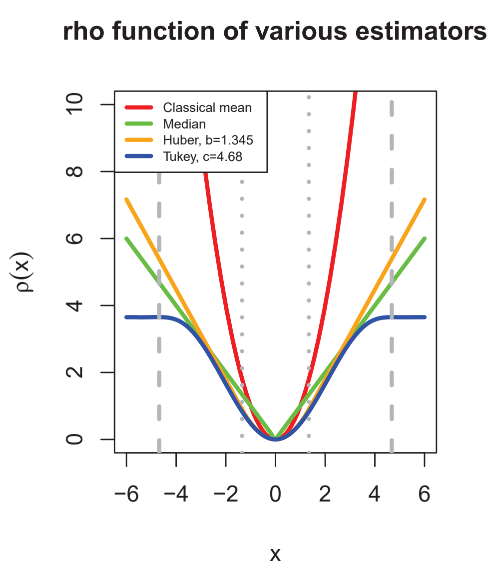
- **Mean**: $\rho(x) = x^2/2$
- **Median**: $\rho(x) = |x|$
- **Huber**:

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

- **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



Score functions

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

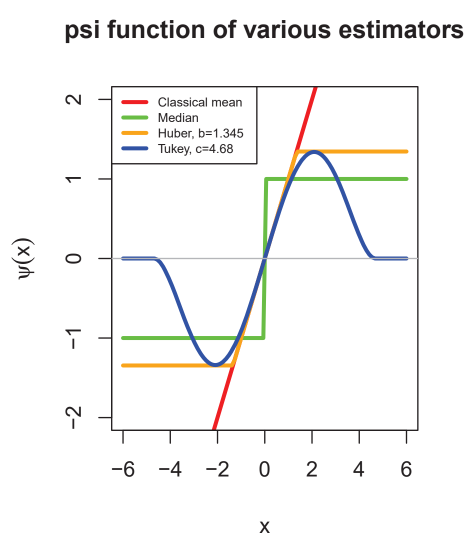
$$\psi_b(x) = \begin{cases} x & \text{if} \quad |x| \leqslant b \\ b\ \text{sign}(x) & \text{if} \quad |x| > b \end{cases}$$

- **Tukey's bisquare**:

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

Score functions

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



Properties of location M-estimators

- Fisher-consistent iff $\int \psi(x)dF(x) = 0$.
- Influence function:
$$\text{IF}(x, T, F) = \frac{\psi(x)}{\int \psi'(y)dF(y)}$$
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
$$V(T, F) = \int \text{IF}(x, T, F)^2 dF(x) = \frac{\int \psi^2(x)dF(x)}{(\int \psi'(y)dF(y))^2}$$
- 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

Two estimators which do not depend on an initial location estimate (Rousseeuw and Croux, 1993):

4. **Q_n estimator**

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

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

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

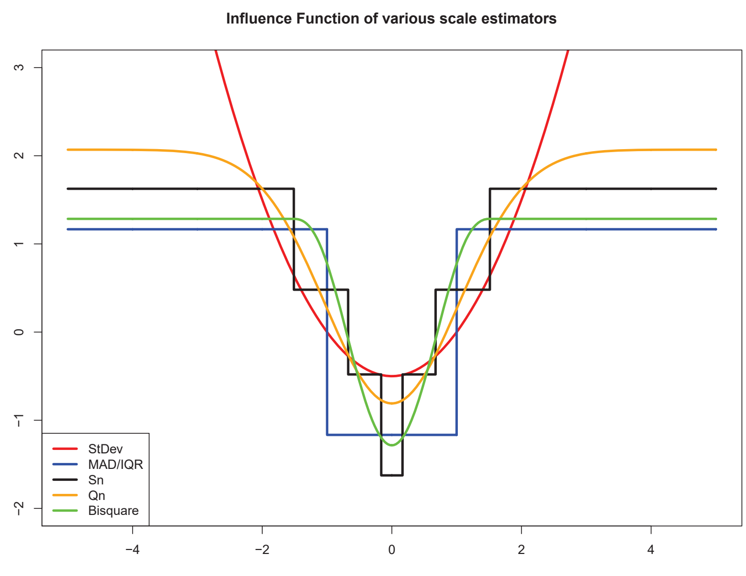
5. **S_n estimator**

$$S_n = 1.1926 \operatorname*{med}_i \operatorname*{med}_j \{|x_i - x_j|\}$$

For each i we compute the median of $\{|x_i - x_j|; j = 1, \ldots, n\}$. The median of these n numbers is then the estimate S_n.

Also S_n can be computed in O($n \log n$) time.

Explicit 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%
S_n	50%	1.625	58%
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.

MLE estimator of scale

The maximum likelihood estimator (MLE) of σ satisfies

$$\begin{aligned}\hat{\sigma}_{\text{MLE}} &= \operatorname*{argmax}_{\sigma} \prod_{i=1}^{n} \frac{1}{\sigma} f(\frac{x_i}{\sigma}) \\ &= \operatorname*{argmax}_{\sigma} \sum_{i=1}^{n} \left\{ -\log(\sigma) + \log f(\frac{x_i}{\sigma}) \right\}\end{aligned}$$

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

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

  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

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

- By the information inequality, the asymptotic variance satisfies

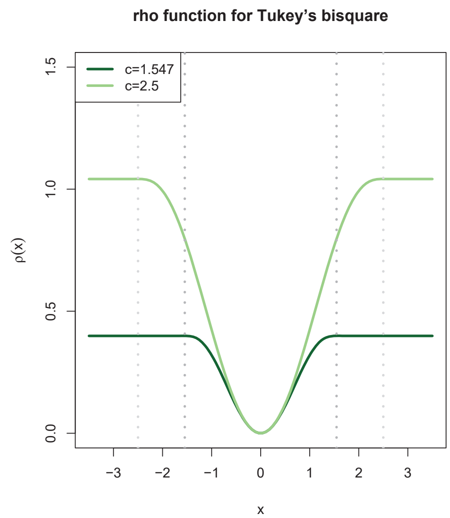
$$V(T, F) \geqslant \frac{1}{I(F)}$$

  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 $\text{IF}(x; \text{MLE}, \Phi) = (x^2 - 1)/2$.

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

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

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, S_n, 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 MADN.

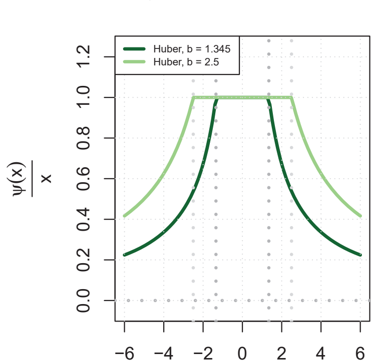
For instance, a location M-estimator with monotone and bounded ψ-function (say, the Huber ψ with $b = 1.5$) and with $\hat{\sigma}$=MADN attains a 50% breakdown value, which is the highest possible.

An algorithm for location M-estimators

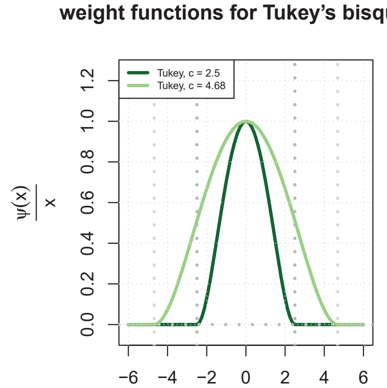
Based on $\psi = \rho'$ we define the weight function

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

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 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. However, this 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
S_n	0.23	0.29
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)

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

with

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

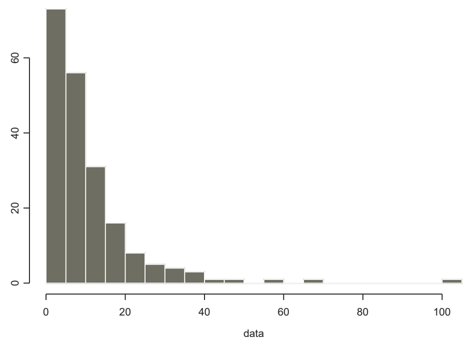
This measure also has $\varepsilon^* = 25\%$ and is more sensitive to asymmetry .

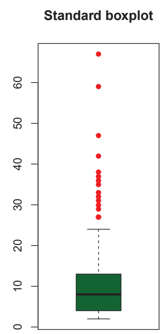
Standard boxplot

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):





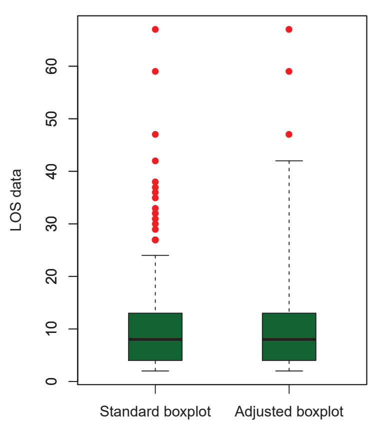
This outlier detection rule is not very accurate at asymmetric data.

Adjusted boxplot

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

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

(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(data 1,psi=psi.huber,scale.est="MAD"))
- Tukey Bisquare: `rlm` in package *MASS* (rlm(data 1,psi=psi.bisquare,scale.est="MAD"))
- MADN, IQR: `mad` and `IQR`
- Q_n, S_n: `Qn` and `Sn` in package *robustbase*
- Medcouple: `mc` in package *robustbase*
- adjusted boxplot: `adjbox` in package *robustbase*