

Introduction to Robust Statistics

Klaus Nordhausen

Department of Mathematics and Statistics
University of Turku

Autumn 2015

Rejection point

Another interesting feature of a influence function is the **rejection point** if it exists.

Let F be a symmetric distributions around the origin and consider only location functionals now, then the rejection point ρ^* is defined as

$$\rho^* = \inf \{r > 0; IF(x^*, T, F) = 0 \text{ when } |x^*| > r\}.$$

If there exists no such point, then $\rho^* = \infty$.

Hence the main idea is that IF vanishes after certain points and therefore points outside $[-r, r]$ do not contribute to the estimate.

Some further comments on the IF

- The existence of the Gâteaux derivative is actually not needed and even weaker conditions can be made.
- If F is replaced by F_n and $\epsilon = 1/n$ then the IF is approximately measuring n times the change of the functional T when adding an additional observation.

Some further comments on the IF II

- Considering a von Mises expansion (derived from a Tyler series) one can obtain

$$T(F_n) \approx T(F) + \int IF(x^*, T, F) dF_n(x^*) + \text{remainder}$$

Evaluating the integral and rearranging the terms this can be written

$$\sqrt{n}(T(F_n) - T(F)) \approx \frac{1}{\sqrt{n}} \sum_{i=1}^n IF(x_i, T, F) + \text{remainder}$$

As the first term on the right side is asymptotically normal by the CLT and the remainder becomes under very general conditions negligible also T_n must be asymptotically normal

$$\sqrt{n}(T(F_n) - T(F)) \rightarrow N(0, \int IF(x, T, F)^2 dF(x))$$

Some further comments on the IF III

- Hence under general conditions one can also get the asymptotic relative efficiencies using influence functions. Let T and S be two competing functionals, then

$$ARE(S, T) = \frac{\int IF(x, T, F)^2 dF(x)}{\int IF(x, S, F)^2 dF(x)}.$$

Local and global measures

The influence function is the first measure of robustness presented here. It describes the infinitesimal stability of an functional.

It is however a completely local concept and therefore also a **global** measure of robustness is required, which describes up to which “distance” from the target distribution F the estimate still provides “useful” information.

Breakdown point

The most popular global measure of robustness is the **breakdown point (BP)**.

Consider again the general contamination model

$$F_\epsilon = (1 - \epsilon)F + \epsilon H.$$

Let T be the functional for the parameter $\theta \in \Theta$.

Then the **breakdown point** $\epsilon^*(T, F)$ is defined as the largest value $\epsilon \in [0, 1]$ such that $T(F_\epsilon)$ as a function of H stays bounded and also bounded away from the boundaries of Θ .

Comments on breakdown point

- If the functional T is not unique, the BP has to be considered for every possible “solution”.
- The BP has to be considered differently for each type of estimate, for example:
 - A location functional breaks down if it can be “pushed” to $\pm\infty$.
 - A scale functional breaks down if it can be “pushed” to either 0 or $+\infty$ (this is known as “implosion” and “explosion”).
- It is easy to construct estimates with high BP. For example the estimate of location $\mu \equiv 0$ has $\epsilon^*(\mu, F) = 1$.
- For “reasonable” estimates the BP usually is in $[0, 0.5]$. For example for location estimates which are shift equivariant it can be shown that the BP can not exceed 0.5.
- There are many other, similar, definitions of BP.

BP for some location estimates

- Mean: $T(F) = E_F(X) \Rightarrow \epsilon^*(T, F) = 0$
- Median: $T(F) = F^{-1}(0.5) \Rightarrow \epsilon^*(T, F) = 0.5$
- Trimmed mean: $T(F) = T_{TM}(F) \Rightarrow \epsilon^*(T, F) = \alpha$
- Winsorized mean: $T(F) = T_{WM}(F) \Rightarrow \epsilon^*(T, F) = \alpha$

Finite sample breakdown point

Of more practical interest is the **finite sample breakdown point (FBP)**.

Let $\mathbf{x} = (x_1, \dots, x_n)$ be a sample of size n and $T(F_{\mathbf{x}})$ be the estimate of the parameter $\theta \in \Theta$.

The **replacement finite sample BP** at \mathbf{x} is the largest proportion $\epsilon_n^*(T, \mathbf{x})$ of data points that can be replaced by arbitrary outliers without $T(F_{\mathbf{x}})$ leaving a bounded set and also bounded away from the boundaries of Θ .

Finite sample breakdown point II

- In most cases the FBP does not depend on \mathbf{x} .
- The FBP often converges to the BP.
- For affine equivariant location estimates it can be shown that for the FBP holds

$$\epsilon_n^* \leq \frac{1}{n} \left[\frac{n-1}{2} \right]$$

Finite sample breakdown point III

Let \mathbf{y} be the set of n data points having $n - m$ in common with \mathbf{x} :

$$\mathcal{X}_m = \{\mathbf{y} : \#(\mathbf{y}) = n, \#(\mathbf{x} \cap \mathbf{y}) = n - m\}$$

Then

$$\epsilon_n^*(T, \mathbf{x}) = \frac{m^*}{n},$$

with

$$m^* = \max \{m \leq 0 : T(F_{\mathbf{y}}) \text{ bound and also bounded away from boundaries of } \Theta \forall \mathbf{y} \in \mathcal{X}_m\}.$$

Finite sample breakdown point IV

Another possible finite sample breakdown point is the **addition finite sample BP**.

$$\mathcal{X}_m = \{\mathbf{y} : \#(\mathbf{y}) = n + m, \mathbf{x} \subset \mathbf{y}\}$$

Then

$$\epsilon_n^{**}(T, \mathbf{x}) = \frac{m^*}{n},$$

with

$$m^* = \max \{m \leq 0 : T(F_{\mathbf{y}}) \text{ bound and also bounded away from boundaries of } \Theta \forall \mathbf{y} \in \mathcal{X}_m\}.$$

For large n usually $\epsilon_n^*(T, \mathbf{x})$ and $\epsilon_n^{**}(T, \mathbf{x})$ are quite similar. Usually $\epsilon_n^*(T, \mathbf{x})$ is the preferred version.

Maximum bias

IF and BP are now two measures of robustness which kind of measure different extreme situations.

Given a BP ϵ^* it is then often of interest to measure what is the worst that can happen if ϵ in the contamination model is smaller than ϵ^* .

The **asymptotic bias** of a functional T is defined as

$$b_T(F, \theta) = T(F) - \theta.$$

The **maximum bias (MB)** of a functional T is defined as

$$\text{MB}_T(\epsilon, \theta) = \max \{|b_T(F, \theta)| : F \in F_\epsilon\}.$$

Note that two functionals with the same BP might have different MBs.

Classes of Estimates

We have defined now the fundamental measures of robustness and have met already some examples for different location and scale estimates.

It is however quite cumbersome to go through all possible estimates and we will now consider classes of estimates.

We will focus here then mainly on location estimates.

L-statistics

All four location estimates presented here so far can actually be seen as member of **L-statistics**, which are **linear combinations of order statistics**.

$$T(F_n) = \sum_{i=1}^n a_{i,n} X_{(i)},$$

where $a_{i,n}$ are constant weights.

For example the estimates mean and median have then the weights:

- mean: $a_{i,n} = 1/n$
- median:
 - if n is odd:

$$a_{i,n} = \begin{cases} 1 & i = (n+1)/2 \\ 0 & i \neq (n+1)/2 \end{cases}$$

- if n is even:

$$a_{i,n} = \begin{cases} 0.5 & i = n/2, n/2 + 1 \\ 0 & \text{otherwise} \end{cases}$$

Properties of L-statistics

- The general form of the influence function exists.
- The IF can obtain any desirable monotonic shape by an appropriate choice of weights. This means however the rejection point is always ∞ .
- L-statistics are difficult to generalize to general statistical concepts.

Maximum likelihood estimation

The most popular estimation method in statistics is **maximum likelihood estimation (MLE)**. The general idea is that based on a parametric model to choose the parameters which make the obtained sample the most **likely**.

Let x_1, \dots, x_n be a iid sample coming from a distribution F with density f and fixed parameter $\theta \in \Theta$.

The joint density of the data is then

$$f(x_1, \dots, x_n; \theta) = f(\mathbf{x}; \theta) = \prod_{i=1}^n f(x_i; \theta).$$

The **likelihood function** reverses then the roles of the observations and θ by taking the data as fixed and θ as the “input”.

$$L(\theta; \mathbf{x}) = \prod_{i=1}^n f(x_i; \theta)$$

Maximum likelihood estimation

The **maximum likelihood estimate** $\hat{\theta}$ is then defined as the value which maximizes the likelihood function.

$$\hat{\theta} = \operatorname{argmax}_{\theta} L(\theta; \mathbf{x})$$

Since the “product” of the densities in the likelihood function is usually not so easy to treat and contains many irrelevant constants and since optimizers usually rather minimize, it is often practise to rather solve the following equation:

$$\hat{\theta} = \operatorname{argmin}_{\theta} l(\theta; \mathbf{x}),$$

where $l(\theta; \mathbf{x}) = -\log(L(\theta; \mathbf{x}))$ (and irrelevant constants are dropped). Or we could search the root of $l'(\theta; \mathbf{x}) = \frac{\partial l(\theta; \mathbf{x})}{\partial \theta}$.

MLE for the normal distribution and known variance

Consider the simple case where f is the normal density with known variance 1.

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2}(x - \mu)^2 \right\}$$

Hence

$$l(\mu; \mathbf{x}) = \sum_{i=1}^n (x_i - \mu)^2$$

or

$$l'(\mu; \mathbf{x}) = \sum_{i=1}^n (x_i - \mu) \quad (\text{dropping again irrelevant constants})$$

And to get $\hat{\mu}$ we either have to minimize $l(\mu; \mathbf{x})$ or solve $l'(\mu; \mathbf{x}) = 0$. In both cases it is easy to see that the MLE for the location in the normal model is the mean \bar{x} .

MLE for the double exponential distribution

Consider the case where f is the double exponential density with known scale 1.

$$f(x) = \frac{1}{2} \exp \left\{ -\frac{1}{2} |x - \mu| \right\}$$

Hence

$$l(\mu; \mathbf{x}) = \sum_{i=1}^n |x_i - \mu|$$

or

$$l'(\mu; \mathbf{x}) = \sum_{i=1}^n \text{sign}(x_i - \mu) \quad (\text{dropping again irrelevant constants})$$

And to get $\hat{\mu}$ we either have to minimize $l(\mu; \mathbf{x})$ or solve $l'(\mu; \mathbf{x}) = 0$. In both cases it is easy to see that the MLE for the location in the double exponential model is the median $\text{med}(\mathbf{x})$.

MLE form another point of view

So from the previous slides it is obvious that MLE can be seen from two different sides:

- 1 Objective function approach: an objective function needs to be minimized

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^n \rho(x_i, \theta),$$

where for the MLE $\rho(x, \theta) = -\log(f(x, \theta))$.

- 2 M-equation approach: an equation needs to be solved

$$\hat{\theta} \text{ solves } \sum_{i=1}^n \psi(x_i, \theta) = 0$$

where for the MLE $\psi(x, \theta) = \frac{\partial(-\log(f(x, \theta)))}{\partial \theta}$.

M-estimation

So how about defining the estimates based on either of the two points of view from the previous slide?

The functions ρ and ψ would not even be required to be related to any distribution (density) or to each other!

This is known as **M-estimation**, which roughly could be called maximum-likelihood type of estimation.

M-estimation for location

Consider the again the location-scale model. Now we are only interested in estimating the location and assume that the distribution is symmetric.

A reasonable assumption is then to require that the estimate is shift equivariant.

Hence we have the logical requirements:

- Equivariance:

$$x_i \rightarrow x_i + c \Rightarrow T_n \rightarrow T_n + c$$

which translates to natural conditions on ρ and ψ

$$\rho(x, \mu) = \rho(x - \mu) \text{ and } \psi(x, \mu) = \psi(x - \mu)$$

- Symmetry:

$$x_i \rightarrow -x_i \Rightarrow T_n \rightarrow -T_n$$

which translates to natural conditions on ρ and ψ

$$\rho(-r) = \rho(r) \text{ and } \psi(-r) = -\psi(r)$$

M-estimation for location II

Another point of view in this setup would be that in general the density must be a function of $|x - \mu|$ and we would like to replace now the density function with some function g .

Hence ρ and ψ would be derived from

$$g(|x - \mu|).$$

M-estimation for location III

For the mean we have then the following functions:

$$\rho(x) = \frac{(x - \mu)^2}{2} \quad \text{and} \quad \psi(x) = x - \mu.$$

For the median the functions are:

$$\rho(x) = |x - \mu| \quad \text{and} \quad \psi(x) = \text{sign}(x - \mu).$$

Since however the relation with μ is “fixed” by the assumptions it is often simply stated that $\rho(r) = r^2$, or $\rho(r) = |r|$ and similarly $\psi(r) = r$ or $\psi(r) = \text{sign}(r)$.

M-estimation for location IV

So, besides these natural requirements listed above, how to choose ρ and/or ψ ?

As we are doing robust statistics, the choice is naturally such to obtain an estimate which is

- Efficient at the target distribution F .
- Efficient also in a neighborhood of F .

M-functional for location

An **M-functional** of location $T(F)$ is the solution to

$$E_F [\psi(x, T(F))] = 0$$

It has the influence functional

$$IF(x^*, T, F) = c(T, F)\psi(x^*, T(F)),$$

where $c(T, F) = -\frac{1}{E_F \left[\frac{\partial \psi(x, \mu)}{\partial \mu} \right]}$ evaluated at $\mu = T(F)$.

Note that $E_F[IF(x, T, F)] = 0$.

Hence one can think of a desired influence function and then construct the appropriate M-estimate.

Huber's M-functional

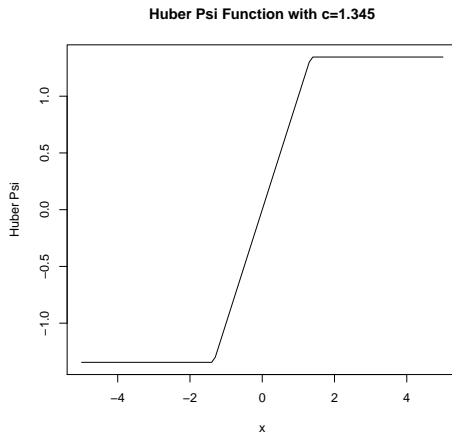
Huber suggested for example

$$\psi(r) = \begin{cases} c & r \geq c \\ r & |r| < c \\ -c & r \leq -c \end{cases}$$

where c is a tuning constant.

This is basically a adaptively trimmed mean, where given the tuning constant the proportion trimmed depends on the data.

Huber's ψ



Interpretation of M-functionals for location

Recall that we have two conditions on the ψ -function to be reasonable location functionals.

Express then $\psi(r) = ru(r)$ and let $w_i = u(x_i - \mu)$, then

$$0 = \sum_{i=1}^n \psi(x_i - \mu) = \sum_{i=1}^n (x_i - \mu) u(x_i - \mu) = \sum_{i=1}^n w_i (x_i - \mu)$$

And hence

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

Therefore M-estimates of location can be seen as **adaptively weighted means** where the weights are chosen by the data.


Computation of M-functionals of location

The main algorithm for M-estimates of location is the **iterated reweighted least squares (IRLS)** algorithm.

- 1 Choose an initial value for μ , i.e. $\hat{\mu}_0$.
- 2 The weights at iteration k are then given by

$$w_{i,k} = u(x_i - \hat{\mu}_k).$$

- 3 The updating step is


$$\hat{\mu}_{k+1} = \frac{\sum_{i=1}^n w_{i,k} x_i}{\sum_{i=1}^n w_{i,k}}$$

- 4 Repeat steps 2 and 3 until $|\hat{\mu}_{k+1} - \hat{\mu}_k| < \epsilon$ for some prechosen tolerance limit ϵ .

One-step M-functionals

It is sometimes already a quite good strategy not to iterate the IRLS algorithm until convergence but to choose instead as initial value $\hat{\mu}_0$ a “real” and robust location estimate, like for example the median. And then to do only one “iteration”.

This is known as **one-step M estimation**. And is hence just a reweighted mean.