Measuring robustness

1 Introduction

While in the classical approach to statistics one aims at estimates which have desirable properties at an exactly specified model, the aim of robust methods is loosely speaking to develop estimates which have a "good" behavior in a "neighborhood" of a model.

This notion will now be made precise.

To gain some insight, we use an artificial dataset x by generating $n = 20$ random N(0,1) numbers.

To measure the effect of different locations of an outlier, we add an extra data point x_0 which is allowed to range on the whole line.

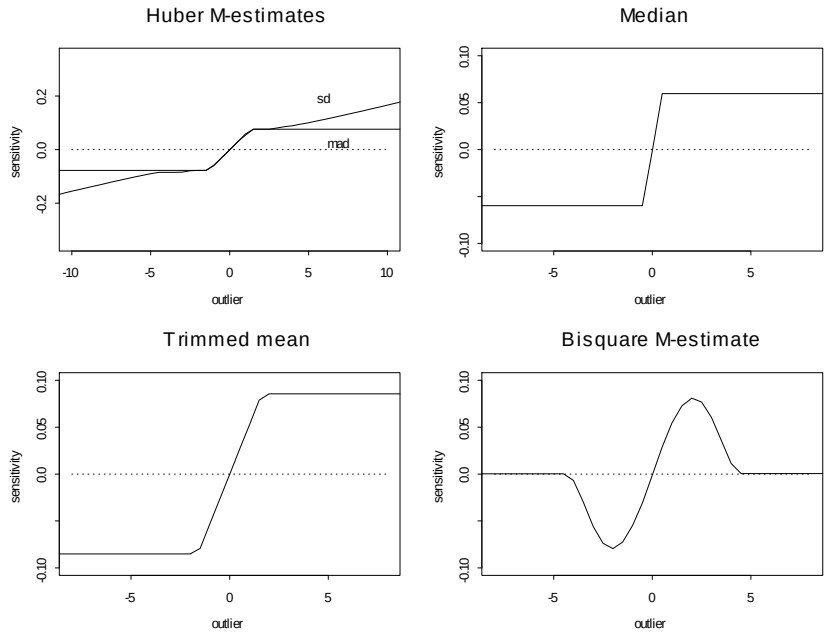
The *sensitivity curve* of the estimate $\widehat{\mu}$ for the sample $x_1, ..., x_n$ is the difference

$$\widehat{\mu}(x_1, ., x_n, x_0) - \widehat{\mu}(x_1, ., x_n)$$

as a function of the location x_0 of the outlier.

The next Figure plots the sensitivity curves of:

- the median,
- the 25% trimmed mean $\overline{x}_{0.25}$,
- the Huber M estimates with k =1.37 using both the SD and the MADN as previously-computed dispersion estimates, and
- the bisquare M-estimate with k =4.68 using the MADN as dispersion.



Sensitivity curves of location estimates

We see that all curves are bounded, except the one corresponding to the Huber estimate with the SD as dispersion, which grows without bound with x_0.

This shows the importance of a robust previous dispersion.

All curves are nondecreasing for positive x_0, except the one for the bisquare estimate.

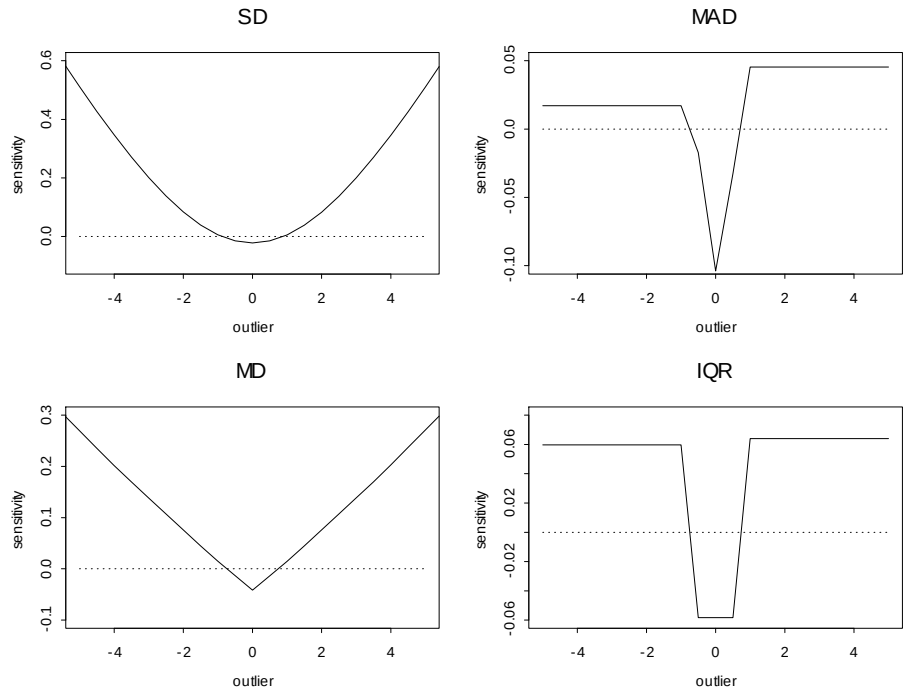
Loosely speaking, we say that the bisquare M-estimate *rejects* extreme values, while the others don't.

The curve for the trimmed mean shows that it does not reject large observations, but just limits their influence.

The curve for the median is very steep at the origin.

The next Figure shows the sensitivity curves of the SD along with the normalized MD, MAD and IQR.

The SD and MD have unbounded sensitivity curves while those of the normalized MAD and IQR are bounded.



Sensitivity curves of dispersion estimates

Imagine now that instead of adding a single point at a variable location, we replace m points by a fixed value $x_0 = 1000$.

The next Table shows the resulting "biases"

$$\widehat{\mu}(x_0, x_0, .., x_0, x_{m+1}, ., x_n) - \widehat{\mu}(x_1, ., x_n)$$

as a function of m for following location estimates:

- the median,
- the Huber estimate with k =1.37 and previously estimated MAD (denoted by H-MADp),
- the Huber estimate with k =1.37 and simultaneous MAD ("H-MADs"),

- the Huber estimate with k =1.37 and previous SD ("H-SD"),

- the trimmed mean with $\alpha = 0.085$; and

- the bisquare estimate.

Besides we provide the biases for the normalized MAD and IQR dispersion estimates.

The choice of k and α was made in order that both the Huber estimates and the trimmed mean have the same asymptotic variance at the normal distribution.

m	mean	Med	H-MADp	H-MADs	H-SD	$\overline{x}_\alpha$	M-Bisq	MAD	IQR
1	50	0.00	0.03	0.04	16.06	0.04	-0.02	0.12	0.08
2	100	0.01	0.10	0.11	46.78	55.59	0.04	0.22	0.14
4	200	0.21	0.36	0.37	140.5	166.7	0.10	0.46	0.41
5	250	0.34	0.62	0.95	202.9	222.3	0.15	0.56	370.3
7	350	0.48	1.43	42.66	350.0	333.4	0.21	1.29	740.3
9	450	0.76	3.23	450.0	450.0	444.5	0.40	2.16	740.2
10	500	500.5	500.0	500.0	500.0	500.0	500.0	739.3	740.2

Table 1: The effect of increasing contamination on a sample of size 20

The mean deteriorates immediately when $m = 1$ as expected, and since $[\alpha n] = [.085 \times 20] = 1$ the trimmed mean $\overline{x}_\alpha$ deteriorates when $m = 2$, as could be expected.

The H(MADs) deteriorates rapidly starting at $m = 8$ while H(SD) is already quite bad at $m = 1$.

By contrast

- the median, H(MADp) and M-Bisq deteriorate only when $m = n/2$,
- M-Bisq has smaller bias than H(MADp)
- the median (Med) has small biases comparable to those of the M-Bisq (only slightly higher bias than M-Bisq at $m = 4, 5, 7, 9$).

To formalize these notions, it will be easier to study the behavior of estimates when the sample size tends to infinity ("asymptotic behavior").

Consider an estimate $\widehat{\theta}_n = \widehat{\theta}_n(x)$ depending on a sample $\mathbf{x} = \{x_1, ..., x_n\}$ of size n of i.i.d. variables with distribution F.

In all cases of practical interest, there is a value depending on F, $\widehat{\theta}_\infty = \widehat{\theta}_\infty(F)$, such that

$$\widehat{\theta}_n \to_p \widehat{\theta}_\infty(F).$$

$\widehat{\theta}_\infty(F)$ is the *asymptotic value* of the estimate at F.

If $\widehat{\theta}_n = \overline{x}$ (the sample mean) then $\widehat{\theta}_\infty(F) = \mathsf{E}_F x$ (the distribution mean).

If $\widehat{\theta}_n(\mathbf{x}) = \mathsf{Med}(\mathbf{x})$ (the sample median) then $\widehat{\theta}_\infty(F) = F^{-1}(0.5)$ (the distribution median).

If $\widehat{\theta}_n$ is a location M-estimator with ψ monotonic, it was stated that $\widehat{\theta}_\infty(F)$ is the solution of

$$\mathsf{E}_F \psi\left(x - \theta\right) = 0.$$

The same reasoning shows that if $\widehat{\theta}_n$ is a scale M-estimate, then $\widehat{\theta}_\infty(F)$ is the solution of

$$\mathsf{E}_F \rho\left(\frac{x}{\theta}\right) = \delta.$$

It can also be shown that if $\widehat{\theta}_n$ is a location M-estimator, then $\widehat{\theta}_\infty(F)$ is the solution of

$$\mathsf{E}_F \rho\left(x - \theta\right) = \mathsf{min}\,.$$

Asymptotic values exist also for the trimmed mean.

The typical distribution of data depends on one or more unknown parameters.

Thus in the location model the data have distribution function

$$F_\mu(x) = F_0(x - \mu),$$

and in the location-dispersion model the distribution is

$$F_\theta(x) = F_0\left(\frac{x-\mu}{\sigma}\right) \quad \text{with} \quad \theta = (\mu, \sigma).$$

These are called *parametric models*.

In the location model we have seen that if the data are symmetric about μ and $\widehat{\mu}$ is an M-estimate, then $\widehat{\mu} \to_p \mu$ and so $\widehat{\mu}_\infty(F_\mu) = \mu$.

An estimator $\widehat{\theta}$ of the parameter(s) of a parametric family F_θ will be called *consistent* if

$$\widehat{\theta}_\infty(F_\theta) = \theta.$$

Since we assume F to be only approximately known, we are interested in the behavior of $\widehat{\theta}_\infty(F)$ when F ranges over a "neighborhood" of a distribution F_0.

There are several ways to characterize neighborhoods.

The easiest to deal with are *contamination neighborhoods*:

$$\mathcal{F}(F, \varepsilon) = \{(1-\varepsilon)F + \varepsilon G : G \in \mathcal{G}\}$$

where $\mathcal{G}$ is a suitable set of distributions, often the set of all distributions.

But in some cases $\mathcal{G}$ is the set of point mass distributions, where the "point mass" δ_{x_0} is the distribution such that $\mathsf{P}(x = x_0) = 1$.

2 The influence function

The *influence function* (IF) of an estimator (Hampel, 1974) is an asymptotic version of its sensitivity curve.

It is an approximation to the behavior of $\widehat{\theta}_\infty$ when the sample contains a small fraction ε of identical outliers. It is defined as

$$\begin{aligned} \mathsf{IF}_{\widehat{\theta}}(x_0, F) &= \lim_{\varepsilon \downarrow 0} \frac{\widehat{\theta}_\infty\left((1-\varepsilon)\,F + \varepsilon \delta_{x_0}\right) - \widehat{\theta}_\infty(F)}{\varepsilon} \\ &= \frac{\partial}{\partial \varepsilon}\, \widehat{\theta}_\infty\left[(1-\varepsilon)\,F + \varepsilon \delta_0\right]\Big|_{\varepsilon \downarrow 0}, \end{aligned}$$

where δ_{x_0} is the point-mass at x_0 and "$\downarrow$" stands for "limit from the right".

If there are p unknown parameters, then $\widehat{\theta}_\infty$ is a p-dimensional vector and so is its IF.

The argument of $\widehat{\theta}_\infty(F)$ will be dropped if there is no ambiguity.

The quantity $\widehat{\theta}_\infty\left((1-\varepsilon)\,F+\varepsilon\delta_{x_0}\right)$ is the asymptotic value of the estimate when the underlying distribution is F and a fraction ε of outliers are equal to x_0.

Thus if ε is small this value can be approximated by

$$\widehat{\theta}_\infty\left((1-\varepsilon)\,F+\varepsilon\delta_{x_0}\right)\approx\widehat{\theta}_\infty(F)+\varepsilon\mathsf{IF}_{\widehat{\theta}}(x_0,F)$$

and the *bias*

$$\widehat{\theta}_\infty\left((1-\varepsilon)\,F+\varepsilon\delta_{x_0}\right)-\widehat{\theta}_\infty(F)$$

is approximated by $\varepsilon\mathsf{IF}_{\widehat{\theta}}(x_0,F)$.

The IF may be considered as a "limit version" of the sensitivity curve, in the following sense.

When we add the new observation x_0 to the sample $x_1, ..., x_n$ the fraction of contamination is $1/(n+1)$, and so we define the *standardized sensitivity curve* as

$$\begin{aligned} \mathsf{SC}_n(x_0) &= \frac{\widehat{\theta}_{n+1}(x_1, ..., x_n, x_0) - \widehat{\theta}_n(x_1, ..., x_n)}{1/(n+1)}, \\ &= (n+1)\left(\widehat{\theta}_{n+1}(x_1, ..., x_n, x_0) - \widehat{\theta}_n(x_1, ..., x_n)\right) \end{aligned}$$

which is similar to the $\mathsf{I}F$ with $\varepsilon = 1/(n+1)$.

One would expect that if the x_i's are i.i.d. with distribution F, then $\mathsf{SC}_n(x_0) \approx \mathsf{IF}(x_0, F)$ for large n.

Actually we have for each x_0

$$\mathsf{SC}_n(x_0) \to_{a.s.} \mathsf{IF}_{\widehat{\theta}}(x_0, F),$$

where "a.s." denotes convergence with probability one ("almost sure convergence").

For a location M-estimate $\widehat{\mu}$ we have

$$\mathsf{IF}_{\widehat{\mu}}(x_0, F) = \frac{\psi(x_0 - \widehat{\mu}_\infty)}{\mathsf{E}\psi'(x - \widehat{\mu}_\infty)}.$$

The similarity between the IF and the SC of a given estimator can be seen comparing the former Figures.

We see above that the IF of an M-estimate is proportional to its ψ-function (or an off-set ρ-function in the case of the scale estimate), and this behavior holds in general for M-estimates (details omitted).

Consider now an M-estimate $\widehat{\mu}$ of location with known dispersion σ, where the asymptotic value $\widehat{\mu}_\infty$ satisfies

$$\mathsf{E}_F \psi\left(\frac{x - \widehat{\mu}_\infty}{\sigma}\right) = 0.$$

Then the influence function of $\widehat{\mu}$ is

$$\mathsf{IF}_{\widehat{\mu}}(x_0, F) = \sigma \frac{\psi\left(\left(x_0 - \widehat{\mu}_\infty\right)/\sigma\right)}{\mathsf{E}_F \psi'\left(\left(x_0 - \widehat{\mu}_\infty\right)/\sigma\right)}.$$

Now consider location estimation with a previously-computed dispersion estimate $\widehat{\sigma}$.

In this case the influence function is much more complicated than the one above, and depends on the influence function of $\widehat{\sigma}$.

But it can be proved that if F is symmetric, the influence function simplifies to:

$$\mathsf{IF}_{\widehat{\mu}}(x_0, F) = \widehat{\sigma}_\infty \frac{\psi\left(\left(x_0 - \widehat{\mu}_\infty\right)/\widehat{\sigma}_\infty\right)}{\mathsf{E}_F \psi'\left(\left(x_0 - \widehat{\mu}_\infty\right)/\widehat{\sigma}_\infty\right)}.$$

It can be shown that the IF of an $\alpha-$trimmed mean $\widehat{\mu}$ at a symmetric F is proportional to Huber's ψ function:

$$\mathsf{IF}_{\widehat{\mu}}(x_0, F) = \frac{\psi_k(x - \widehat{\mu}_\infty)}{1 - 2\alpha}$$

with $k = F^{-1}(1 - \alpha)$.

Hence the trimmed mean and the Huber estimate in the example at the beginning of the chapter have not only the same asymptotic variances, but also the same influence function.

However the last Table showed that they have very different degrees of robustness.

We can verify that the asymptotic variance v of location M-estimates satisfies

$$v = \mathsf{E}_F\left(\mathsf{IF}(x, F)^2\right)$$

and this property holds for a wide family (but not for all!) of "smooth" estimates.

3 The breakdown point

We have seen the effects of replacing several data values by outliers.

Roughly speaking, the breakdown point (BP) of an estimate $\widehat{\theta}$ of the parameter θ is the largest amount of contamination (proportion of atypical points) that the data may contain
such that $\widehat{\theta}$ still gives some information about θ, i.e., about the distribution of the "typical" points.

More precisely, let θ range over a set Θ.

For example, $\Theta = R$ for a location parameter, $\Theta = [0, \infty)$ for a scale or dispersion parameter, and $\Theta = [-1, 1]$ for a correlation coefficient.

In order for $\widehat{\theta}$ to give some information about θ it must at least stay away from the boundary of Θ.

This means being bounded away from $\pm\infty$ in the first case, from 0 and ∞ in the second, and from ± 1 in the third.

DEFINITION

The asymptotic contamination BP of the estimate $\widehat{\theta}$ at F, denoted by $\varepsilon^*(\widehat{\theta}, F)$ is the largest $\varepsilon^* \in (0,1)$ such that for $\varepsilon < \varepsilon^*$, we have that

$$\widehat{\theta}_\infty \left[(1-\varepsilon)\, F + \varepsilon G\right]$$

remains bounded away from the boundary of Θ for all G.

The definition means that there exists a set $K \subset \Theta$ such that $K \cap \partial\Theta = \emptyset$, (where $\partial\Theta$ denotes the boundary of Θ) such that

$$\widehat{\theta}_\infty\left((1-\varepsilon)\,F + \varepsilon G\right) \in K \;\forall\; \varepsilon < \varepsilon^* \text{ and } \forall\; G.$$

When the estimate is not uniquely defined, e.g. when it is the solution of an equation that may have multiple roots, the boundedness of the estimate means that *all* solutions remain in a bounded set.

Note that it is easy to find estimates with high BP. For instance, the "estimate" identically equal to 0 has $\varepsilon^* = 1$!

However, for "reasonable" (e.g., equivariant) estimates it is intuitively clear that there must be more "typical" than "atypical" points and so $\varepsilon^* \leq 1/2$.

Actually, it can be proved that all shift equivariant location estimates have $\varepsilon^* \leq 1/2$.

3.1 Location and scale M-estimates, and dispersion estimates

For an M-estimate with known scale and odd ψ, we have

$$\varepsilon^* = 0.5$$

3.1.1 Scale estimates

Note that while a high proportion of atypical points with large values (outliers) may cause the estimate $\widehat{\sigma}$ to overestimate the true scale, while a high proportion of data near zero ("inliers") may result in under estimation of the true scale.

Thus it is desirable that the estimate remains bounded away from zero ("implosion") as well as away from infinity ("explosion").

This is equivalent to keeping the *logarithm* of $\widehat{\sigma}$ bounded.

For a scale M-estimate

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

we have

$$\varepsilon^* = \mathbf{min}(\delta, 1-\delta).$$

3.1.2 Dispersion estimates

It is easy to show that the breakdown points of the SD, the MAD and the IQR are 0, 1/2 and 1/4, respectively

In general, the BP of an equivariant dispersion estimate is ≤ 0.5.

3.2 Location with previously-computed dispersion estimate

In the Table above we have seen the bad consequences of using an M-estimate $\widehat{\mu}$ with the SD as previously-computed dispersion estimate $\widehat{\sigma}$**.**

The reason is that the outliers inflate this dispersion estimate, and hence outliers do not appear as such in the "standardized" residuals $(x_i - \widehat{\mu})/\widehat{\sigma}$.

Hence the robustness of $\widehat{\sigma}$ is essential for that of $\widehat{\mu}$.

For monotone M-estimates with a bounded and odd ψ, it can be shown that $\varepsilon^*(\widehat{\mu}) = \varepsilon^*(\widehat{\sigma})$.

Thus if $\widehat{\sigma}$ is the MAD then $\varepsilon^*(\widehat{\mu})$ =0.5, but if $\widehat{\sigma}$ is the SD then $\varepsilon^*(\widehat{\mu}) = 0$.

Our former results imply that the location estimates using the SD and the MAD as previous dispersion have the *same IF*, but quite *different BP's.*

By the way, this is an example of an estimate with a *bounded IF* but a *zero BP*.

For redescending M-estimates with a bounded ρ the situation is more complex, but roughly speaking we can say that for the bisquare ρ with MAD scale, the BP is 0.5 for all practical purposes.

3.3 Simultaneous estimation

The BP for simultaneous estimates of location and dispersion is much more complicated.

In general, the BP of $\widehat{\mu}$ is less than 0.5.

In particular, using Huber's ψ_k with $\widehat{\sigma}$=MAD yields that with $k = 1.37$ we have $\varepsilon^* = 0.33$.

This is clearly lower than the BP=0.5 which corresponds to using a previously computed dispersion estimate treated above.

3.4 Finite-sample breakdown point

Although the asymptotic BP is an important theoretical concept, it may be more useful to define the notion of BP for a finite sample.

Let $\widehat{\theta}_n = \widehat{\theta}_n(x)$ be an estimate defined for samples $\mathbf{x} = \{x_1, ..., x_n\}$.

The *replacement finite sample breakdown point* (FBP) of $\widehat{\theta}_n$ at x is the largest proportion $\varepsilon_n^*(\widehat{\theta}_n, \mathbf{x})$ of data points that can be arbitrarily replaced by outliers without $\widehat{\theta}_n$ leaving a set bounded away from the boundary of Θ (Huber and Donoho, 1983).

More formally, call $\mathcal{X}_m$ the set of all datasets $\mathbf{y}$ of size n having $n-m$ elements in common with $\mathbf{x}$:

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

Then

$$\varepsilon_n^*(\widehat{\theta}_n, \mathbf{x}) = \frac{m^*}{n},$$

where

$$m^* = \max\left\{m \geq 0 : \widehat{\theta}_n(\mathbf{y}) \text{ bounded away from } \partial\Theta \ \forall\, \mathbf{y} \in \mathcal{X}_m\right\}.$$

In most cases of interest, ε_n^* does not depend on $\mathbf{x}$, and tends to the asymptotic BP when $n \to \infty$.

For equivariant location estimates we have

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

and that this bound is attained by M-estimates with an odd and bounded ψ.

For the trimmed mean, it is easy to verify that $m^* = [n\alpha]$, so that for large n we have

$$\varepsilon_n^* \approx \alpha.$$

4 Maximum asymptotic bias

The IF and the BP consider extreme situations in the study of contamination.

The first deals with "infinitesimal" values of ε, while the second deals with the largest ε an estimate can tolerate.

Note that an estimate having a high BP means that $\widehat{\theta}_\infty(F)$ will remain in a bounded set when F ranges in an ε-neighborhood with $\varepsilon \leq \varepsilon^*$, but this set may be very large.

What we want to do now is, roughly speaking, to measure the worst behavior of the estimate for each given $\varepsilon < \varepsilon^*$.

We again consider F ranging in the ε-neighborhood

$$\mathcal{F}_{\varepsilon,\theta} = \{(1-\varepsilon)\, F_\theta + \varepsilon G : G \in \mathcal{G}\}$$

of the an assumed parametric distribution F_θ with $\mathcal{G}$ is a family of distribution functions.

Unless otherwise specified, $\mathcal{G}$ will be the family of all distribution functions, but in some cases we will choose a the family of point-mass distributions.

The *asymptotic bias* of $\widehat{\theta}$ at any $F \in \mathcal{F}_{\varepsilon,\theta}$ is

$$\mathsf{b}_{\widehat{\theta}}(F,\theta) = \widehat{\theta}_{\infty}(F) - \widehat{\theta}_{\infty}(F_{\theta})$$

and the maximum bias is

$$\mathsf{MB}_{\widehat{\theta}}(\varepsilon,\theta) = \mathsf{max}\left\{\left|b_{\widehat{\theta}}(F,\theta)\right| : F \in \mathcal{F}_{\varepsilon,\theta}\right\}.$$

In the case that the parameter space is the whole set of real numbers, the relationship between maximum bias and BP is

$$\varepsilon^{*}(\widehat{\theta}, F_{\theta}) = \mathsf{max}\left\{\varepsilon \geq 0 : \mathsf{MB}_{\widehat{\theta}}(\varepsilon,\theta) < \infty\right\}.$$

Note that two estimates may have the same BP but different maximum biases.

The *contamination sensitivity* of $\widehat{\theta}$ at θ is defined as

$$\gamma_c(\widehat{\theta}, \theta) = \left[\frac{d}{d\varepsilon}\mathrm{MB}_{\widehat{\theta}}(\varepsilon, \theta)\right]_{\varepsilon=0}.$$

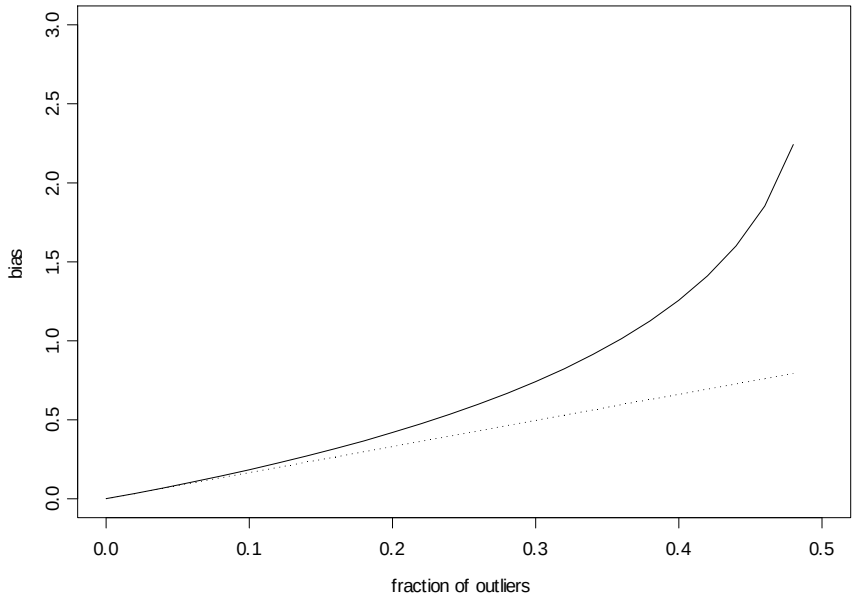
Since $\mathrm{MB}_{\widehat{\theta}}(0, \theta) = b_{\widehat{\theta}}(0, \theta) = 0$, γ_c gives an approximation to the maximum bias for small ε :

$$\mathrm{MB}_{\widehat{\theta}}(\varepsilon, \theta) \approx \varepsilon\gamma_c(\widehat{\theta}, \theta).$$

Note however that $\mathrm{MB}_{\widehat{\theta}}(\varepsilon^*, \theta) = \infty$, while the right-hand side above always yields a finite result.

Therefore this approximation will be quite unreliable for sufficiently large values of ε.

The next Figure shows the maximum bias $\mathrm{MB}_{\widehat{\theta}}(\varepsilon, \theta)$ at $F_\theta = \mathsf{N}(\theta, 1)$ and its approximation for the Huber location estimate with $k = 1.37$
(note that the bias does not depend on θ due to the estimate's shift equivariance).



Maximum bias of Huber estimate (—) and its linear aproximation (.....) as a function of ε

The *gross-error sensitivity* (GES) of $\widehat{\theta}$ at θ is

$$\gamma^*(\widehat{\theta}, \theta) = \max_{x_0} \left| \mathsf{IF}_{\widehat{\theta}}(x_0, F_\theta) \right| .$$

Since $(1-\varepsilon)\,F_\theta + \varepsilon\delta_{x_0} \in \mathcal{F}_{\varepsilon,\theta}$, we have for all x_0

$$\left|\widehat{\theta}_\infty\left((1-\varepsilon)\,F_\theta + \varepsilon\delta_{x_0}\right) - \widehat{\theta}_\infty(F_\theta)\right| \leq \mathsf{MB}_{\widehat{\theta}}(\varepsilon,\theta).$$

So dividing by ε and taking the limit we get

$$\gamma^* \leq \gamma_c.$$

Equality above holds for M-estimates with bounded ψ-functions, but not in general.

For instance we have seen that the IF of the Huber estimate with the SD as previous dispersion is bounded, but since $\varepsilon^* = 0$ we have

$$\mathsf{MB}_{\widehat{\theta}}(\varepsilon,\theta) = \infty \quad \forall\ \varepsilon > 0$$

and so γ_c is infinite.

For location M-estimates $\widehat{\mu}$ with odd ψ and assuming a location model $F_\mu(x) = F_0(x-\mu)$, we have

$$\gamma^*(\widehat{\mu}, \mu) = \frac{\psi(\infty)}{\mathsf{E}_{F_\mu}\psi'(x - \widehat{\mu}_\infty)} = \frac{\psi(\infty)}{\mathsf{E}_{F_0}\psi'(x)}$$

so that $\gamma^*(\widehat{\mu}, \mu)$ does not depend on μ.

In general for equivariant estimates $\mathsf{MB}_{\widehat{\theta}}(\varepsilon, \theta)$ does not depend on θ.

It is shown that the median minimizes the MB for M-estimates at symmetric models.

5 Balancing robustness and efficiency

In this section we consider a parametric model F_θ and an estimate $\widehat{\theta}$ which is consistent for θ and such that the distribution of $\sqrt{n}(\widehat{\theta}_n - \theta)$ under F_θ tends to a normal distribution with mean 0 and variance $v = v(\widehat{\theta}, \theta)$.

Under the preceding assumptions $\widehat{\theta}$ has no asymptotic bias and we care only about its variability.

Let $v_{\mathsf{min}} = v_{\mathsf{min}}(\theta)$ be the smallest possible asymptotic variance within a "reasonable" class of estimates (e.g. equivariant).

Under reasonable regularity conditions v_{min} is the asymptotic variance of the MLE for the model.

Then the *asymptotic efficiency* of $\widehat{\theta}$ at θ is defined as

$$\mathsf{aeff} = \frac{v_{\mathsf{min}}(\theta)}{v(\widehat{\theta}, \theta)}.$$

If instead F does not belong to the family F_θ but is in a neighborhood of F_θ, the squared bias will dominate the variance component of mean-squared error for all sufficiently large n.

To see this let $b = \widehat{\theta}_\infty(F) - \theta$ and note that in general under F we have

$$\mathcal{D}\left(\sqrt{n}(\widehat{\theta}_n - \widehat{\theta}_\infty)\right) \approx \mathsf{N}\left(0, v\right).$$

Then

$$\mathcal{D}\left(\widehat{\theta}_n - \theta\right) \approx \mathsf{N}\left(b, \frac{v}{n}\right),$$

so that the variance tends to zero while the bias does not.

Thus we must balance the efficiency of $\widehat{\theta}$ at the model F_θ with the bias in a neighborhood of it.

We have seen that location M-estimates with a bounded ψ and previously computed dispersion estimate with BP $= 1/2$ attain the maximum breakdown point of $1/2$.

To choose among them we must compare their biases for a given efficiency.

We consider the Huber and bisquare estimates with previously-computed MAD dispersion and efficiency 0.95.

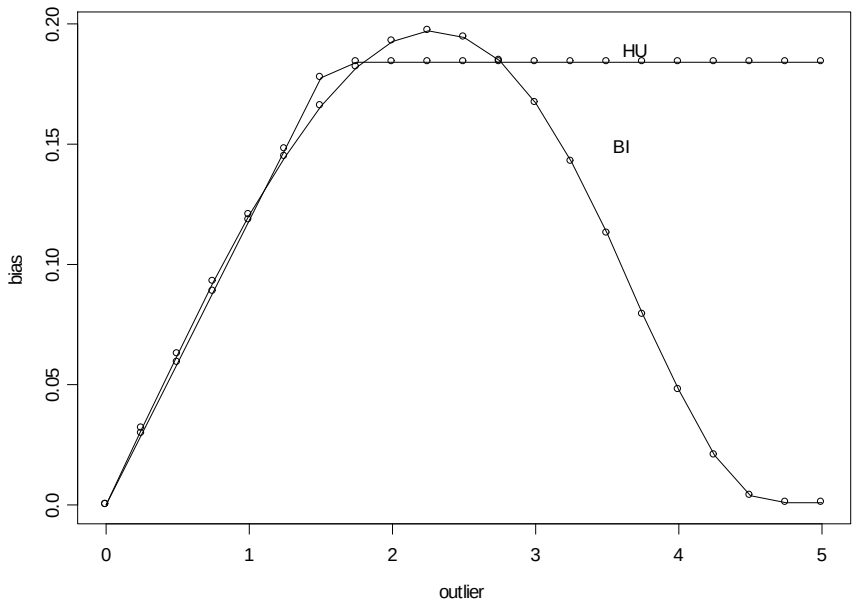
Their maximum biases for the model

$$F_{\varepsilon,\theta} = \{(1-\varepsilon)\, F_\theta + \varepsilon G : G \in \mathcal{G}\} \text{ with } F_\theta = N(0,1)$$

and a few values of ε are:

ε	0.05	0.10	0.20
Huber	0.087	0.184	0.419
Bisq.	0.093	0.197	0.450

The next Figure shows the respective biases for point contamination at K with $\varepsilon = 0.1$, as a function of the outlier location K.



Asymptotic biases of Huber and Bisquare estimates for 10% contamination as functions of the outlier location K.

It is seen that although the maximum bias of the bisquare is higher, the difference is very small and its bias remains below that of the Huber estimate for the majority of the values.

This shows that, although the maximum bias contains much more information than the BP, it is not informative enough to discriminate among estimates and that one should look at the whole bias behavior when possible.

Now, to study the behavior of the estimates under symmetric heavy-tailed distributions, we computed the asymptotic variances of the Huber and bisquare estimates,
and of the location/dispersion MLE for the Cauchy distribution ("CMLE"),
at both the Normal and Cauchy distributions,
the latter of which can be considered an extreme case of heavy-tailed behavior.

The efficiencies are given below.

	Huber	Bisq.	CMLE
Normal	0.95	0.95	0.60
Cauchy	0.57	0.72	1.00

It is seen that the bisquare estimate yields the best trade-off between the efficiencies at the two distributions.

For all the above reasons we recommend for estimating location the bisquare M-estimate with previously computed MAD.

6 "Optimal" robustness

In this section we consider different way in which an "optimal" estimate may be defined.

6.1 Bias- and variance-optimality of location estimates

6.1.1 Minimax bias

If we pay attention only to bias, the quest for an "optimal" location estimate is simple: Huber (1964) has shown that the median has the smallest maximum bias ("minimax bias") among *all* shift equivariant estimates if the underlying distribution is symmetric and unimodal.

6.1.2 Minimax variance

Huber (1964) studied location M-estimates in neighborhoods of a symmetric F with symmetric contamination, so that there is no bias problem.

The dispersion is assumed known.

Call $v(\widehat{\theta}, H)$ the asymptotic variance of the estimate $\widehat{\theta}$ at the distribution H, and

$$v_\varepsilon(\widehat{\theta}) = \sup_{H \in \mathcal{F}(F,\varepsilon)} v(\widehat{\theta}, H),$$

where $\mathcal{F}(F, \varepsilon)$ is a contamination neighborhood with G ranging over all *symmetric* distributions.

Assume that F has a density f and that $\psi_0 = -f'/f$ is nondecreasing.

Then the M-estimate minimizing $v_\varepsilon(\widehat{\theta})$ has

$$\psi(x) = \begin{cases} \psi_0(x) & \text{if} \quad |\psi_0(x)| \le k \\ k\,\text{sgn(x)} & \text{else} \end{cases}$$

where k depends on F and ε. For normal F, this is the Huber ψ_k.

Since ψ_0 corresponds to the MLE for f, the result may be described as a truncated MLE.

6.2 The Hampel approach

Hampel (1974) stated the balance problem between bias and efficiency for general estimates as minimizing the asymptotic variance under a bound on the gross-error sensitivity.

For a symmetric location model, his result coincides with Huber's.

It is remarkable that both approaches coincide at the location problem, and furthermore the result has a high breakdown point.

The case of multidimensional parameters is treated likewise but the details are more complex.