3.4 M-estimators of location

Let ρ be an even and nondecreasing function of $|x|$, with $\rho(0) = 0$. If ρ is differentiable, we define $\psi = \rho'$. We then have that ψ is odd and $\psi(x) \geqslant 0$ for $x \geqslant 0$. For the location model, an M-estimator [26] is defined as

$$\hat{\mu}_n = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} \rho(x_i - \mu). \tag{3.9}$$

If ρ is differentiable, differentiating (3.9) with respect to μ yields

$$\sum_{i=1}^{n} \psi(x_i - \hat{\mu}_n) = 0. \tag{3.10}$$

Examples:

- $\rho(x) = -\ln f(x)$ hence $\psi(x) = -f'(x)/f(x)$. Then it follows from (3.9) that

  $$\hat{\mu}_n = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} -\ln(f(x_i - \mu)) = \underset{\mu}{\operatorname{argmax}} \ln \prod_{i=1}^{n} f(x_i - \mu)$$

  and thus $\hat{\mu}_n$ equals the MLE of μ. M-estimators are thus generalizations of MLE-estimators.

- $\rho(x) = \frac{x^2}{2}$. Then $\psi(x) = x$ and we have that

  $$\sum_{i=1}^{n} (x_i - \hat{\mu}_n) = 0$$

  which means that $\hat{\mu}_n = \bar{x}$. Thus the sample mean is an M-estimator. At the normal model, $f(x) = \phi(x)$, it is the MLE.

- $\rho(x) = |x|$. Then it follows from (3.9) that

  $$\hat{\mu}_n = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} |x_i - \mu|.$$

  The sample median attains this minimum, so it is an M-estimator with $\psi(x) = \operatorname{sign}(x)$. It is the MLE at the double exponential distribution $f(x) = \frac{1}{2} e^{-|x|}$ (also called the Laplace distribution).

- *Huber M-estimator* with parameter $b > 0$:

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

and

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

Note that the limiting cases $b \to \infty$ and $b \to 0$ correspond to the mean and the median.

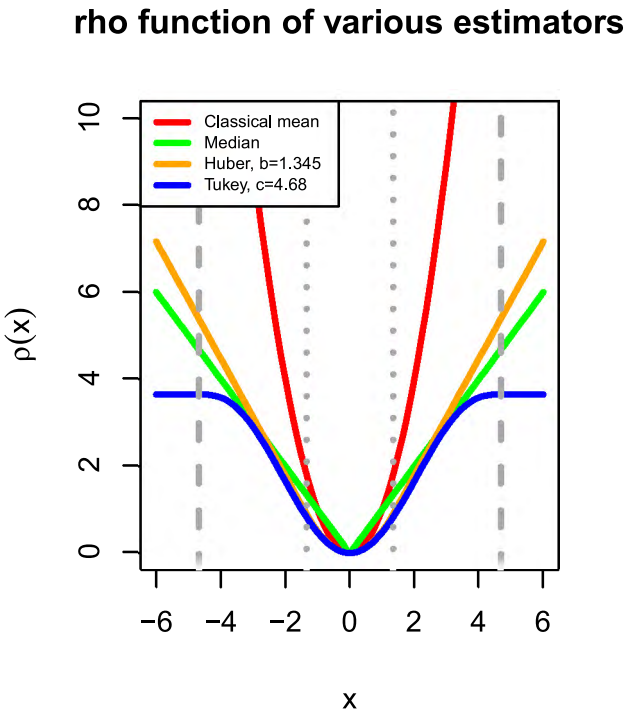
- *Tukey's bisquare M-estimator* with parameter $c > 0$:

$$\rho(x) = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4} & \text{if} \quad |x| \leqslant c \\ c^2/6 & \text{if} \quad |x| > c \end{cases} \tag{3.11}$$

and

$$\psi(x) = x \left(1 - \frac{x^2}{c^2}\right)^2 I(|x| \leqslant c).$$

The graphs of these ρ and ψ are depicted in Figure 3.1. For the mean both functions are unbounded. For the median and the Huber M-estimator ψ is increasing and bounded and ρ is unbounded. Such M-estimators are called *monotone*. One can prove that the optimization problem in (3.9) is then a convex optimization problem and no local minima exist. This is a big computational advantage. For the bisquare M-estimator both ρ and ψ are bounded. Moreover ψ is not increasing: for large x the function $\psi(x)$ decreases. Such M-estimators are called *redescending*. Computationally this yields a more difficult optimization since local minima can arise. This means that some solutions of (3.10) may not correspond to the global minimum of (3.9).



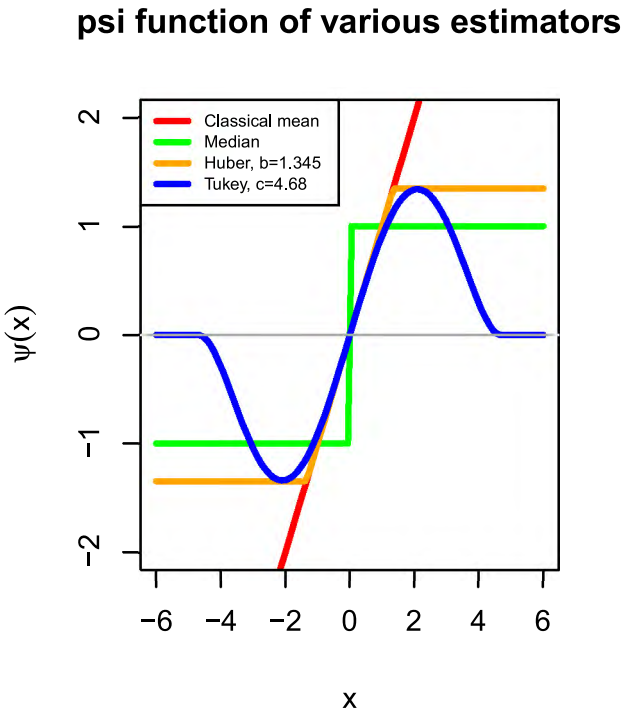


Figure 3.1 The ρ and ψ functions of the mean, the median, the Huber estimator with $b = 1.345$, and the bisquare M-estimator with $c = 4.68$.

Note that M-estimators of location can achieve a high breakdown value. If ψ is odd, bounded and nondecreasing, they attain the maximal breakdown value (3.8). This implies that the Huber location M-estimator has maximal breakdown value for all $b < \infty$.

3.5 M-estimators of scale

Assume the scale model from (3.2). An M-estimator of scale satisfies an equation of the form

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{x_i}{\hat{\sigma}_n}\right)=\delta \tag{3.12}$$

for some $0<\delta<\rho(\infty)$. Usually we set $\delta=\int\rho(t)f(t)dt$ which makes $\hat{\sigma}_n$ consistent.

Examples:

- If $\rho(x)=-xf'(x)/f(x)$ and $\delta=1$, then $\hat{\sigma}_n$ equals the MLE estimator of σ. (Note that $\rho(x)=x\psi(x)$ where $\psi(x)=-f'(x)/f(x)$ corresponds to the MLE estimator for location we saw before.)
- $\rho(x)=x^2$. If $\delta=1$, then $\hat{\sigma}_n=\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2}$ which is called the root mean square. This is the MLE of σ at the normal scale model with known $\mu=0$.
- $\rho(x)=|x|$ and $\delta=1$. Then $\hat{\sigma}_n=\frac{1}{n}\sum_{i=1}^{n}|x_i|$. This is the MLE at the double exponential (Laplace) distribution.
- $\rho(x)$ is the bisquare function (3.11) and $\delta=\frac{1}{2}\rho(\infty)$. The resulting $\hat{\sigma}_n$ is called the bisquare scale M-estimator.
- $\rho(x)=I(|x|>c)$ with $c>0$ and $\delta=0.5$, then $\hat{\sigma}_n=\frac{1}{c}\,\text{median}_i(|x_i|)$.
- As in the case of the MLE, we can also take the ψ function of a location M-estimator and construct an M-estimator of scale with the ρ function $\rho(x)=x\psi(x)$.

3.6 M-estimators of location with unknown dispersion

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

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

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

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

In this general situation we can still estimate location by the median and scale by MADN which possess the required equivariance and invariance properties. But the location M-estimators defined by (3.9) are not scale equivariant, so they depend on the measurement units. If σ were known, the natural solution would be to divide $x_i - \mu$ in the formulas by this σ:

$$\hat{\mu} = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(\frac{x_i-\mu}{\sigma}\right) \tag{3.13}$$

and thus

$$\sum_{i=1}^{n} \psi\left(\frac{x_i-\hat{\mu}}{\sigma}\right) = 0. \tag{3.14}$$

Since σ is unknown, two strategies can be followed:

1. A simple solution is to estimate σ beforehand and to plug this estimate $\hat{\sigma}$ into expression (3.13):

$$\hat{\mu} = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(\frac{x_i-\mu}{\hat{\sigma}}\right) \tag{3.15}$$

which yields the equation

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right) = 0. \tag{3.16}$$

(Note that we have dropped the subscript n from $\hat{\mu}$ and $\hat{\sigma}$ for simplicity.)

Naturally, $\hat{\sigma}$ should be a robust estimator of σ (e.g. the MADN) if one wants $\hat{\mu}$ to be a robust estimator of μ. For monotone M-estimators with a bounded and odd ψ, the breakdown value of the location estimate $\hat{\mu}$ can be shown to be equal to the breakdown value of the scale estimate $\hat{\sigma}$, thus resulting in a maximal breakdown value when the MADN is chosen. Using the standard deviation as scale estimate would result in a zero breakdown value. For redescending M-estimators, the breakdown value depends on the data. For the bisquare estimator with the MADN as preliminary scale estimate, the breakdown value is $1/2$ for all practical purposes.

2. An alternative approach is to simultaneously estimate both parameters μ and σ by a system of equations combining (3.10) and (3.12):

$$\sum_{i=1}^{n} \psi_{\text{location}}\left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right) = 0 \tag{3.17}$$

$$\frac{1}{n}\sum_{i=1}^{n} \rho_{\text{scale}}\left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right) = \delta \tag{3.18}$$

where we may choose ρ_{scale} differently from ρ_{location}. This approach is more complicated than the first, and unfortunately it yields less robust estimators, so the first approach is preferable.

3.7 Computation of M-estimators

3.7.1 Computing an M-estimator of location

Here we follow the first approach above so we first compute a scale estimate $\hat{\sigma}$, typically by MADN. To solve (3.16) an iterative reweighting algorithm can be used. Based on the ψ function for location, define

$$W(x) = \begin{cases} \psi(x)/x & \text{if} \quad x \neq 0 \\ \psi'(0) & \text{if} \quad x = 0 \end{cases} \tag{3.19}$$

where the second line is just L'Hopital's rule for $x \to 0$. Then (3.16) can be written as

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

It follows that $\hat{\mu}$ can be interpreted as a weighted mean:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{3.20}$$

with $w_i = W((x_i - \hat{\mu})/\hat{\sigma})$. Note that this is still an implicit equation, as the w_i depend on $\hat{\mu}$ itself. This suggests the following algorithm:

1. Compute $\hat{\sigma}$. Set $\mu_0 = \text{Med}(\mathbf{x})$.

2. For $k = 0, 1, \ldots$ compute the weights

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

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

4. Stop when $|\hat{\mu}_{k+1} - \hat{\mu}_k| < \varepsilon\hat{\sigma}$.

Because a weighted mean like (3.20) is a special case of weighted least squares, this algorithm is called *iteratively reweighted least squares* (IRLS). If $W(x)$ is bounded and nonincreasing for $x > 0$, the sequence $\hat{\mu}_k$ is guaranteed to converge to a solution of (3.16).

The weight functions for the Huber estimator and the Tukey bisquare estimator are shown in Figure 3.2. We see that the weights for the Huber estimator converge to zero at the rate $1/x$, whereas the Tukey estimator assigns zero weight to large outliers.

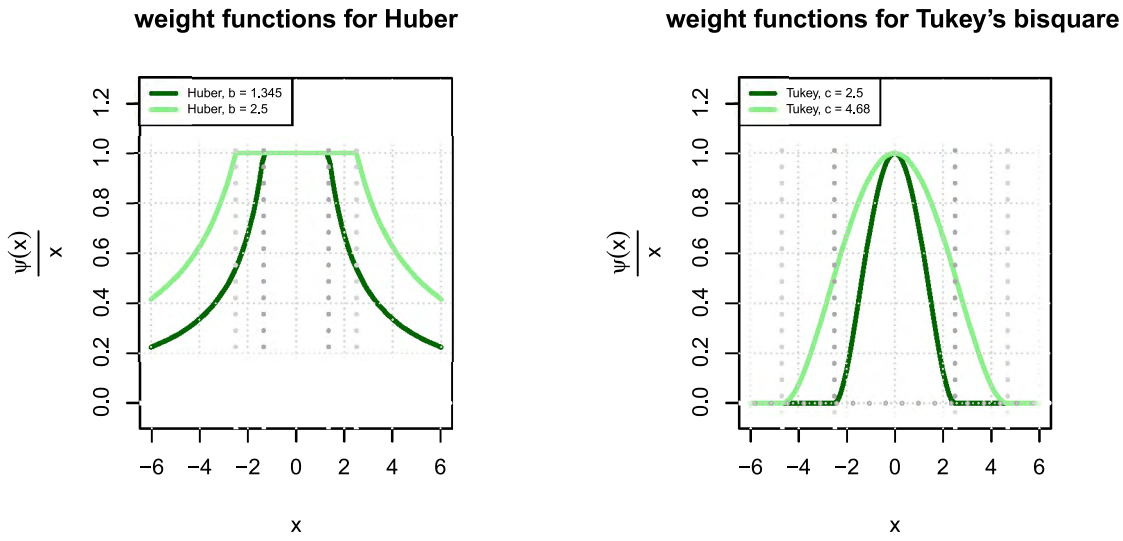


Figure 3.2 Weight functions for the Huber and Tukey M-estimators

3.7.2 Computing an M-estimator of scale

Based on ρ_{scale} we now define the weight function

$$W(x) = \begin{cases} \rho(x)/x^2 & \text{if} \quad x \neq 0 \\ \rho''(0) & \text{if} \quad x = 0. \end{cases} \tag{3.21}$$

Then (3.12) is equivalent to

$$\hat{\sigma}^2 = \frac{1}{n\delta} \sum_{i=1}^{n} W\left(\frac{x_i}{\hat{\sigma}}\right) x_i^2$$

which suggests the following iterative reweighting algorithm:

1. For $k = 0, 1, \ldots$ compute the weights
$$w_{k,i} = W\left(\frac{x_i}{\hat{\sigma}_k}\right).$$
2. Set $\hat{\sigma}_{k+1} = \sqrt{\frac{1}{n\delta} \sum_{i=1}^{n} w_{k,i} x_i^2}$.
3. Stop when $|\hat{\sigma}_{k+1}/\hat{\sigma}_k - 1| < \varepsilon$.

3.7.3 Computing simultaneous M-estimates

In this case the previous algorithms can be combined as follows:

1. Compute starting values $\hat{\mu}_0$ and $\hat{\sigma}_0$.
2. For $k = 0, 1, \ldots$ compute the weights
$$w_{1k,i} = W_1((x_i - \hat{\mu}_k)/\hat{\sigma}_k) \quad \text{and} \quad w_{2k,i} = W_2((x_i - \hat{\mu}_k)/\hat{\sigma}_k)$$
with W_1 and W_2 as in (3.19) and (3.21).
3. Set
$$\hat{\mu}_{k+1} = \frac{\sum_{i=1}^{n} w_{1k,i} x_i}{\sum_{i=1}^{n} w_{k,i}} \quad \text{and} \quad \hat{\sigma}_{k+1} = \sqrt{\frac{1}{n\delta} \sum_{i=1}^{n} w_{k,i} (x_i - \hat{\mu}_k)^2}.$$
4. Stop when $|\hat{\mu}_{k+1} - \hat{\mu}_k| < \varepsilon \hat{\sigma}_{k+1}$ and $|\hat{\sigma}_{k+1}/\hat{\sigma}_k - 1| < \varepsilon$.

3.8 Other robust estimators of location and scale

The *trimmed mean* with parameter $0 \leqslant \alpha < 0.5$ and $m = \lfloor (n-1)\alpha \rfloor$ is defined as

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

Its breakdown value is $(m+1)/n$.

The *Hodges-Lehmann* estimator [22]:

$$\hat{\mu} = \text{med}\, \frac{x_i + x_j}{2}$$

has breakdown value $1 - \sqrt{0.5} = 0.293$ (for $n \to \infty$).

The *least median of squares* (LMS) location estimator [39] is defined as

$$\hat{\mu} = \underset{\mu}{\text{argmin}}\, \text{med}(x_i - \mu)^2.$$

It equals the midpoint of the shortest 'half'. This is the subsample containing $h = \lfloor n/2 \rfloor + 1$ observations that possesses the shortest range. Its breakdown value is maximal. The length of the shortest half yields a robust estimator of scale.

The *least trimmed squares* (LTS) location estimator [39] is defined as

$$\hat{\mu} = \underset{\mu}{\text{argmin}} \sum_{i=1}^{h} (x-\mu)^2_{(i)}.$$

where $(x-\mu)^2_{(i)}$ is the i-th order statistic of the set $\{(x_i - \mu)^2; i = 1, \ldots, n\}$, so the differences are first squared and then ordered. The LTS is the mean of the subsample with h observations that has the smallest standard deviation. Its breakdown value is maximal for $h = \lfloor n/2 \rfloor + 1$. The standard deviation of the 'best' subsample yields a robust scale estimator.

Many scale estimators such as Stdev and MAD require the estimation of a center as a first step. The IQR does not use centering. A more robust scale estimator that does not require centering is the Q_n estimator [40]. This scale estimator is essentially the first quartile of all distances between two data points. More precisely, the estimator is defined as

$$Q_n = 2.219\{|x_i - x_j|; i < j\}_{(k)} \tag{3.22}$$

with $k = \binom{h}{2} \approx \binom{n}{2}/4$ and $h = \lfloor \frac{n}{2} \rfloor + 1$. The breakdown value of Q_n is 50% whereas its statistical efficiency at the normal model is 82%.