**Next:** Least Median of Squares **Up:** Robust Estimation **Previous:** Regression Diagnostics

# M-estimators

One popular robust technique is the so-called *M-estimators*. Let $r_i$ be the *residual* of the $i^{\text{th}}$ datum, the difference between the $i^{\text{th}}$ observation and its fitted value. The standard least-squares method tries to minimize $\sum_i r_i^2$, which is unstable if there are outliers present in the data. Outlying data give an effect so strong in the minimization that the parameters thus estimated are distorted. The M-estimators try to reduce the effect of outliers by replacing the squared residuals $r_i^2$ by another function of the residuals, yielding

$$\min \sum_i \rho(r_i)\,, \tag{28}$$

where $\rho$ is a symmetric, positive-definite function with a unique minimum at zero, and is chosen to be less increasing than square. Instead of solving directly this problem, we can implement it as an iterated reweighted least-squares one. Now let us see how.

Let $\mathbf{p} = [p_1, \ldots, p_m]^T$ be the parameter vector to be estimated. The M-estimator of $\mathbf{p}$ based on the function $\rho(r_i)$ is the vector $\mathbf{p}$ which is the solution of the following *m* equations:

$$\sum_i \psi(r_i) \frac{\partial r_i}{\partial p_j} = 0\,, \quad \text{for } j = 1, \ldots, m, \tag{29}$$

where the derivative $\psi(x) = \mathrm{d}\rho(x)/\mathrm{d}x$ is called the *influence function*. If now we define a *weight function*

$$w(x) = \frac{\psi(x)}{x}\,, \tag{30}$$

then Equation (29) becomes

$$\sum_i w(r_i) r_i \frac{\partial r_i}{\partial p_j} = 0\,, \quad \text{for } j = 1, \ldots, m. \tag{31}$$

This is exactly the system of equations that we obtain if we solve the following iterated reweighted least-squares problem

$$\min \sum_i w(r_i^{(k-1)}) r_i^2 , \qquad (32)$$

where the superscript $(k)$ indicates the iteration number. The weight $w(r_i^{(k-1)})$ should be recomputed after each iteration in order to be used in the next iteration.

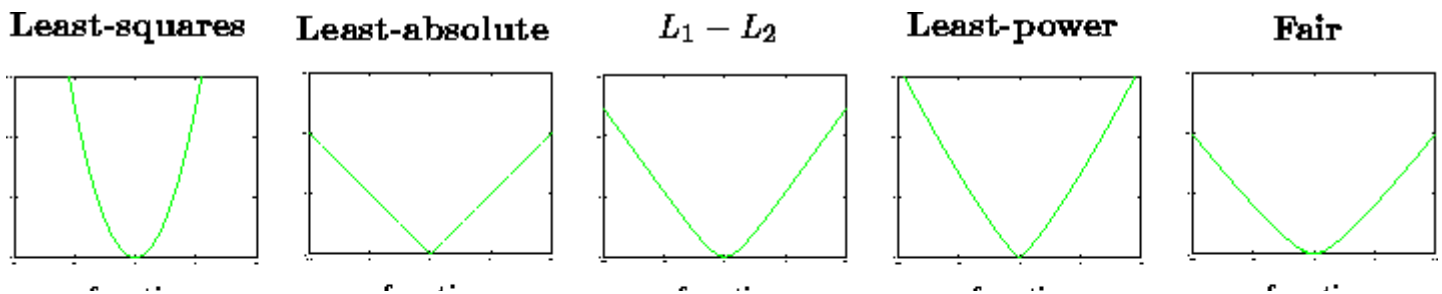
The influence function $\psi(x)$ measures the influence of a datum on the value of the parameter estimate. For example, for the least-squares with $\rho(x) = x^2/2$, the influence function is $\psi(x) = x$, that is, the influence of a datum on the estimate increases linearly with the size of its error, which confirms the non-robusteness of the least-squares estimate. When an estimator is robust, it may be inferred that the influence of any single observation (datum) is insufficient to yield any significant offset [18]. There are several constraints that a robust *M*-estimator should meet:

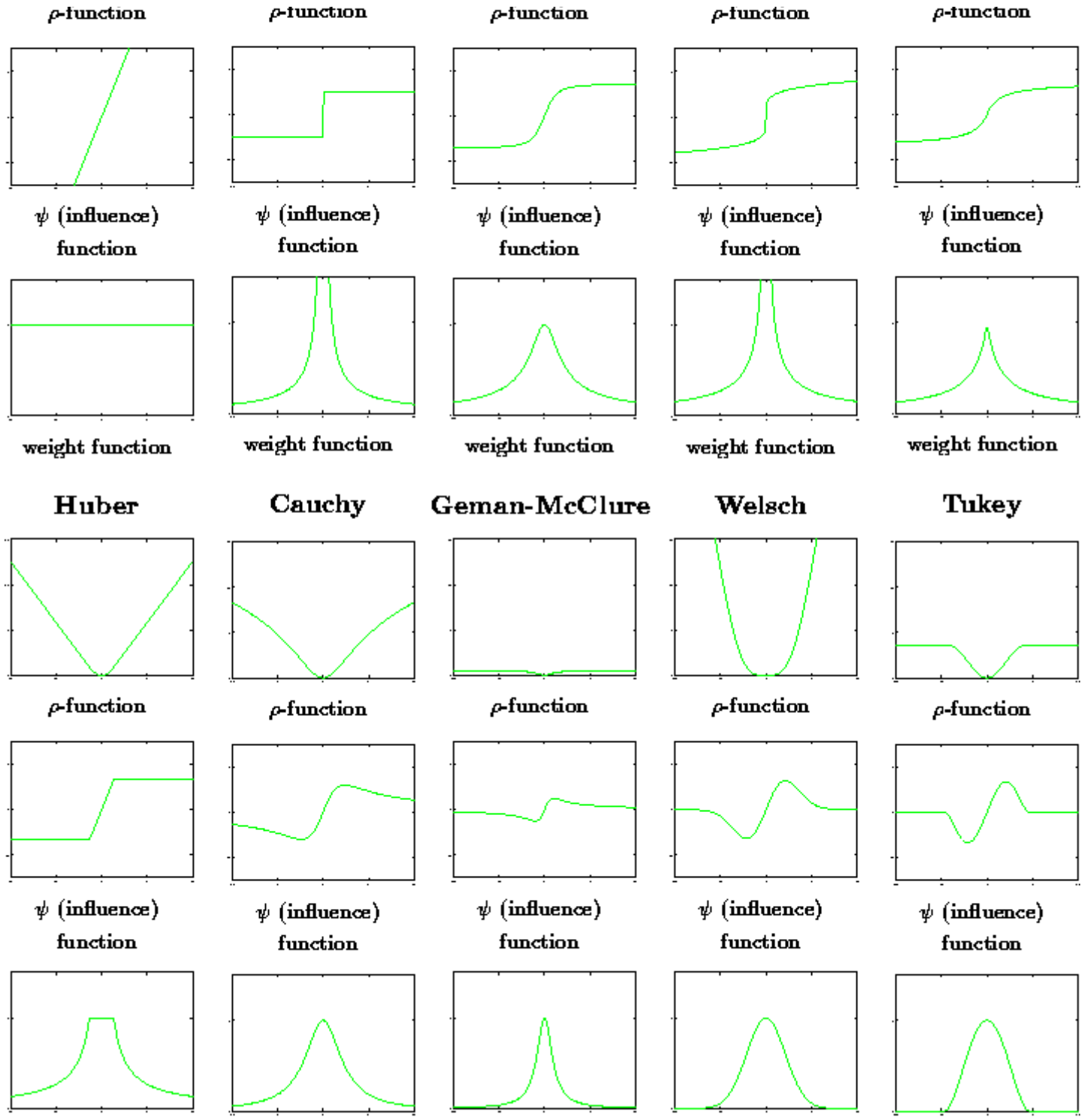
- The first is of course to have a bounded influence function.
- The second is naturally the requirement of the robust estimator to be unique. This implies that the objective function of parameter vector $\mathbf{p}$ to be minimized should have a unique minimum. This requires that *the individual $\rho$-function is convex in variable* $\mathbf{p}$. This is necessary because only requiring a $\rho$-function to have a unique minimum is not sufficient. This is the case with maxima when considering mixture distribution; the sum of unimodal probability distributions is very often multi-modal. The convexity constraint is equivalent to imposing that $\frac{\partial^2 \rho(.)}{\partial \mathbf{p}^2}$ is non-negative definite.
- The third one is a practical requirement. Whenever $\frac{\partial^2 \rho(.)}{\partial \mathbf{p}^2}$ is singular, the objective should have a gradient, $\frac{\partial \rho(.)}{\partial \mathbf{p}} \neq \mathbf{0}$. This avoids having to search through the complete parameter space.

Table 1 lists a few commonly used influence functions. They are graphically depicted in Fig. 4. Note that not all these functions satisfy the above requirements.

| type | $\rho(x)$ | $\psi(x)$ | $w(x)$ |
|---|---|---|---|
| $L_2$ | $x^2/2$ | $x$ | $1$ |
| $L_1$ | $\lvert x\rvert$ | $\operatorname{sgn}(x)$ | $\frac{1}{\lvert x\rvert}$ |
| $L_1-L_2$ | $2(\sqrt{1+x^2/2}-1)$ | $\frac{x}{\sqrt{1+x^2/2}}$ | $\frac{1}{\sqrt{1+x^2/2}}$ |
| $L_p$ | $\frac{\lvert x\rvert^\nu}{\nu}$ | $\operatorname{sgn}(x)\lvert x\rvert^{\nu-1}$ | $\lvert x\rvert^{\nu-2}$ |
| "Fair" | $c^2[\frac{\lvert x\rvert}{c}-\log(1+\frac{\lvert x\rvert}{c})]$ | $\frac{x}{1+\lvert x\rvert/c}$ | $\frac{1}{1+\lvert x\rvert/c}$ |
| Huber $\begin{cases}\text{if } \lvert x\rvert\le k\\ \text{if } \lvert x\rvert\ge k\end{cases}$ | $\begin{cases}x^2/2\\ k(\lvert x\rvert-k/2)\end{cases}$ | $\begin{cases}x\\ k\operatorname{sgn}(x)\end{cases}$ | $\begin{cases}1\\ k/\lvert x\rvert\end{cases}$ |
| Cauchy | $\frac{c^2}{2}\log(1+(x/c)^2)$ | $\frac{x}{1+(x/c)^2}$ | $\frac{1}{1+(x/c)^2}$ |
| Geman-McClure | $\frac{x^2/2}{1+x^2}$ | $\frac{x}{(1+x^2)^2}$ | $\frac{1}{(1+x^2)^2}$ |
| Welsch | $\frac{c^2}{2}[1-\exp(-(x/c)^2)]$ | $x\exp(-(x/c)^2)$ | $\exp(-(x/c)^2))$ |
| Tukey $\begin{cases}\text{if } \lvert x\rvert\le c\\ \text{if } \lvert x\rvert> c\end{cases}$ | $\begin{cases}\frac{c^2}{6}\left(1-[1-(x/c)^2]^3\right)\\ (c^2/6)\end{cases}$ | $\begin{cases}x[1-(x/c)^2]^2\\ 0\end{cases}$ | $\begin{cases}[1-(x/c)^2]^2\\ 0\end{cases}$ |

**Table 1:** A few commonly used M-estimators




ρ-function
ψ (influence) function
weight function
Huber
Cauchy
Geman-McClure
Welsch
Tukey

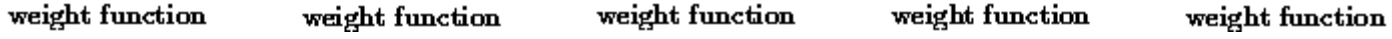



**Figure 4:** Graphic representations of a few common M-estimators

Briefly we give a few indications of these functions:

- $L_2$ (least-squares) estimators are not robust because their influence function is not bounded.
- $L_1$ (absolute value) estimators are not stable because the $\rho$-function $|x|$ is not strictly convex in $x$. Indeed, the second derivative at $x=0$ is unbounded, and an indeterminant solution may result.
- $L_1$ estimators reduce the influence of large errors, but they still have an influence because the influence function has no cut off point.
- $L_1 - L_2$ estimators take both the advantage of the $L_1$ estimators to reduce the influence of large errors and that of $L_2$ estimators to be convex.
- The $L_p$ (*least-powers*) function represents a family of functions. It is $L_2$ with $\nu = 2$ and $L_1$ with $\nu = 1$. The smaller $\nu$, the smaller is the incidence of large errors in the estimate $\mathbf{p}$. It appears that $\nu$ must be fairly moderate to provide a relatively robust estimator or, in other words, to provide an estimator scarcely perturbed by outlying data. The selection of an optimal $\nu$ has been investigated, and for $\nu$ around 1.2, a good estimate may be expected [18]. However, many difficulties are encountered in the computation when parameter $\nu$ is in the range of interest $1 < \nu < 2$, because zero residuals are troublesome.
- The function ``Fair'' is among the possibilities offered by the Roepack package (see [18]). It has everywhere defined continuous derivatives of first three orders, and yields a unique solution. The 95% asymptotic efficiency on the standard normal distribution is obtained with the tuning constant $c=1.3998$.
- Huber's function [7] is a parabola in the vicinity of zero, and increases linearly at a given level $|x| > k$. The 95% asymptotic efficiency on the standard normal distribution is obtained with the tuning constant $k = 1.345$. This estimator is so satisfactory that it has been recommended for almost all situations; very rarely it has been found to be inferior to some other $\rho$-function. However, from time to time, difficulties are encountered, which may be due to the lack of stability in the gradient values of the $\rho$-function because of its *discontinuous second derivative*:

$$\frac{\mathrm{d}^2\rho(x)}{\mathrm{d}x^2} = \begin{cases} 1 & \text{if } |x| \leq k, \\ 0 & \text{if } |x| \geq k. \end{cases}$$

  The modification proposed in [18] is the following

  The 95% asymptotic efficiency on the standard normal distribution is obtained with the tuning constant $c=1.2107$.

- Cauchy's function, also known as the Lorentzian function, does not guarantee a unique solution. With a descending first derivative, such a function has a tendency to yield erroneous solutions in a way which cannot be observed. The 95% asymptotic efficiency on the standard normal distribution is obtained with the tuning constant $c=2.3849$.
- The other remaining functions have the same problem as the Cauchy function. As can be seen from the influence function, the influence of large errors only decreases linearly with their size. The Geman-McClure and Welsh functions try to further reduce the effect of large errors, and

the Tukey's biweight function even suppress the outliers. The 95% asymptotic efficiency on the standard normal distribution of the Tukey's biweight function is obtained with the tuning constant *c=4.6851*; that of the Welsch function, with *c=2.9846*.

There still exist many other $\rho$-functions, such as Andrew's cosine wave function. Another commonly used function is the following tri-weight one:

$$w_i = \begin{cases} 1 & |r_i| \leq \sigma \\ \sigma/|r_i| & \sigma < |r_i| \leq 3\sigma \\ 0 & 3\sigma < |r_i| \end{cases},$$

where $\sigma$ is some estimated standard deviation of errors.

It seems difficult to select a $\rho$-function for general use without being rather arbitrary. Following Rey [18], for the location (or regression) problems, the best choice is the $L_p$ in spite of its theoretical non-robustness: they are quasi-robust. However, it suffers from its computational difficulties. The second best function is ``Fair'', which can yield nicely converging computational procedures. Eventually comes the Huber's function (either original or modified form). All these functions do not eliminate completely the influence of large gross errors.

The four last functions do not guarantee unicity, but reduce considerably, or even eliminate completely, the influence of large gross errors. As proposed by Huber [7], one can start the iteration process with a convex $\rho$-function, iterate until convergence, and then apply a few iterations with one of those non-convex functions to eliminate the effect of large errors.

---

**Next:** Least Median of Squares **Up:** Robust Estimation **Previous:** Regression Diagnostics

*Zhengyou Zhang*
*Thu Feb 8 11:42:20 MET 1996*