Robust Statistics

So far we have mainly had a working assumption about the form of the underlying distribution f e.g. that it belongs to some parametric family. In robust inference we are concerned with reducing the dependence of the conclusions on the specification of this f.

We seek an inference method that does almost as well as possible if the assumptions we make about f are true **but** does not perform that much worse within a range of alternatives to that assumption.

Outliers

Outliers are sample values that cause surprise in relation to the majority of the sample. It should be noted that outlying observations may be correct but should always be checked for transcription errors.

Outliers are not always easy to spot and can hugely effect some estimators e.g. sample mean and variance.

Robust and resistant statistical methods aim to do as good a job as possible in the presence of departures from the assumptions on the form of distribution f on which they are based. In particular, long tailed distributions which give rise to outliers.

Breakdown points

Given any estimator, we can try and measure its global reliability (up to what distance from the model distribution the estimator gives some relevant information) by the *breakdown point* $\epsilon^\star$.

Informally, $\epsilon^\star$ is the largest fraction of the data that can be moved arbitrarily without perturbing the estimator to the boundary of the parameter space. the higher the breakdown point, the more robust the estimator is against extreme outliers.

Estimator	Breakdown Point
Sample Mean	0
Sample Median	$\frac{1}{2}$
α-Trimmed mean	α

where for α-trimmed mean we discard a proportion α of the observations and average the remainder.

M-estimators

The MLE is defined as the value $T_n = T_n(X_1, \ldots, X_n)$ where

$$\sum_{i=1}^{n} -\log f_{T_n}(X_i) = \min_{T_n}!$$

Here f_θ denotes the density corresponding to the distribution function F_θ.

In the hope of making this more robust to miss-specification of the distribution, we can generalize this to

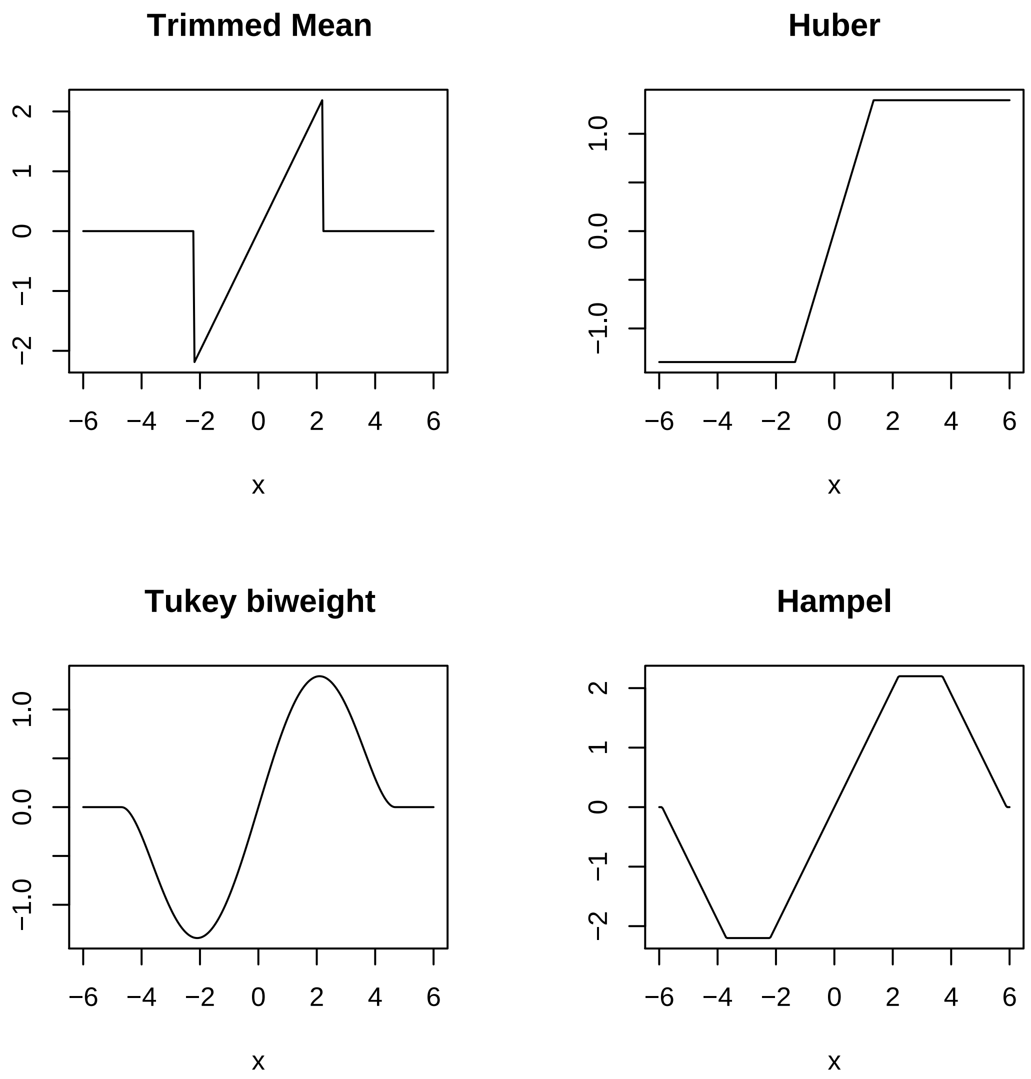
$$\sum_{i=1}^{n} \rho(X_i, T_n) = \min_{T_n}!$$

If ρ has a derivative $\psi(x, \theta) = \partial/\partial\theta\rho(x, \theta)$, then we can find T_n by solving the equation

$$\sum_{i=1} \psi(X_i, T_n) = 0.$$

Any estimator defined by either of these two routes is called an M-estimator.

We now give some examples of the ψ-functions for a location family (this includes the normal dist) where $\rho(X_i, T_n) = \rho(X_i - T_n)$


Trimmed Mean
Huber
Tukey biweight
Hampel
x
−6 −4 −2 0 2 4 6
−2 −1 0 1 2
−1.0 0.0 1.0