Robust Statistics
Part 2: Multivariate location and scatter

Peter Rousseeuw
LARS-IASC School, May 2019

Multivariate location and scatter: Outline

- Classical estimators and outlier detection
- M-estimators
- The Stahel-Donoho estimator
- The MCD estimator
- The MVE estimator
- S-estimators
- MM-estimators
- Some non affine equivariant estimators
- Software availability
Multivariate location and scatter

Data: \( x_1, \ldots, x_n \) where the observations \( x_i \) are \( p \)-variate column vectors.

We often combine the coordinates of the observations in an \( n \times p \) matrix:

\[
X = (x_1, \ldots, x_n)' = \begin{pmatrix}
x_{11} & x_{12} & \cdots & x_{1p} \\
\vdots & \vdots & & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{np}
\end{pmatrix}
\]

Model for the observations:

\( x_i \sim N_p(\mu, \Sigma) \)

More generally we can assume that the data were generated from an elliptical distribution, whose density contours are ellipses too.

Outlier detection

In the multivariate setting, outliers cannot always be detected by simply applying outlier detection rules to each variable separately:
Outlier detection

These points are not outlying in either variable:

We can only detect such outliers by correctly estimating the covariance structure!

Affine equivariance

We usually want estimators $\hat{\mu}$ and $\hat{\Sigma}$ that are affine equivariant.

Affine equivariance

$$\hat{\mu}(\{Ax_1 + b, \ldots, Ax_n + b\}) = A\hat{\mu}(\{x_1, \ldots, x_n\}) + b$$

$$\hat{\Sigma}(\{Ax_1 + b, \ldots, Ax_n + b\}) = A\hat{\Sigma}(\{x_1, \ldots, x_n\})A'$$

for any nonsingular matrix $A$ and any vector $b$.

Affine equivariance implies that the estimator transforms well under any non-singular reparametrization of the space of the $x_i$.

Consequently, the data might be rotated, translated or rescaled (for example through a change of measurement units) without affecting the outlier detection diagnostics.
Affine equivariance

A counterexample to affine equivariance is the coordinatewise median

$$\hat{\mu}(\{x_1, \ldots, x_n\}) = (\text{med}_{i=1}^{n} x_{i1}, \ldots, \text{med}_{i=1}^{n} x_{ip})'$$

which is very easy to compute.

It is not affine equivariant, and not even orthogonally equivariant since it does not transform well under rotations.

What we can do is shift the data like $\{x_1 + b, \ldots, x_n + b\}$ and rescale by a diagonal matrix $A$ (that is, change the measurement units of the original variables).

We will study the robustness of the coordinatewise median later.

Breakdown value

We say that a multivariate location estimator $\hat{\mu}$ breaks down when it can be carried outside any bounded set.

Every affine equivariant location estimator satisfies

$$\varepsilon^*_n(\hat{\mu}, X_n) \leq \frac{1}{n} \left[ \frac{n + 1}{2} \right].$$

The breakdown value of a scatter estimator $\hat{\Sigma}$ is defined as the minimum of the explosion breakdown value and the implosion breakdown value.

Explosion occurs when the largest eigenvalue becomes arbitrarily large. Implosion occurs when the smallest eigenvalue becomes arbitrarily small.
Breakdown value

Any affine equivariant scatter estimator $\hat{\Sigma}$ satisfies

$$\varepsilon^*_n(\hat{\Sigma}, X_n) \leq \frac{1}{n} \left\lfloor \frac{n - p + 1}{2} \right\rfloor$$

if the sample $X_n$ is in general position:

General position

A multivariate data set of dimension $p$ is said to be in general position if at most $p$ observations lie in a $(p - 1)$-dimensional hyperplane.

For example, at most 2 observations lie on a line, at most 3 on a plane, etc.

Overview

Estimators of multivariate location and scatter can be divided into those that are affine equivariant or not, and those with low or high breakdown value:

<table>
<thead>
<tr>
<th></th>
<th>affine equivariant</th>
<th>non affine equivariant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low BV</td>
<td>Classical mean &amp; covariance M-estimators</td>
<td></td>
</tr>
<tr>
<td>High BV</td>
<td>Stahel-Donoho estimator MCD, MVE S-estimators MM-estimators</td>
<td>coordinatewise median spatial median, sign covariance OGK DetMCD</td>
</tr>
</tbody>
</table>
The classical estimators for $\mu$ and $\Sigma$ are the empirical mean and covariance matrix:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$S_n = \frac{1}{n - 1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'.$$

Both are affine equivariant but highly sensitive to outliers, as they have:

- zero breakdown value
- unbounded influence function.
Classical estimators

Consider the **Animals** data set containing the logarithm of the body and brain weight of 28 animals:

\[
\begin{align*}
\text{log}(\text{body}) & \quad \text{log(brain)} \\
-10 & \quad -5 \\
-5 & \quad 0 \\
0 & \quad 5 \\
5 & \quad 10 \\
10 & \quad 15 \\
15 & \quad -5 \\
\end{align*}
\]

**Tolerance ellipsoid**

On this plot we can add the 97.5% tolerance ellipsoid. Its boundary contains those \(x\)-values with constant Mahalanobis distance to the mean.

**Mahalanobis distance**

\[
\text{MD}(x) = \sqrt{(x - \bar{x}_n)'S_n^{-1}(x - \bar{x}_n)}
\]

**Classical tolerance ellipsoid**

\[
\{x; \text{MD}(x) \leq \sqrt{\chi^2_{p,0.975}}\}
\]

with \(\chi^2_{p,0.975}\) the 97.5% quantile of the \(\chi^2\)-distribution with \(p\) degrees of freedom.

We expect (for large \(n\)) that about 97.5% of the observations belong to this ellipsoid.

We could flag observation \(x_i\) as an outlier if it does not belong to the classical tolerance ellipsoid, but...
Tolerance ellipsoid

Based on the classical mean and covariance matrix, the outliers do not stand out:

![Classical tolerance ellipse and Mahalanobis distances](image)

The classical Mahalanobis distances do not flag all the outliers!

Point estimates

On all data points:

\[
\bar{x}_{28} = (3.77 \quad 4.425)'
\]

\[
S_{28} = \begin{pmatrix}
14.22 & 7.05 \\
7.05 & 5.76
\end{pmatrix}
\]

This yields an estimated correlation of \( r = 7.05 / \sqrt{14.22 \times 5.76} = 0.78 \).

On the reduced data set (without observations 6, 16 and 26):

\[
\bar{x}_{25} = (3.03 \quad 4.428)'
\]

\[
S_{25} = \begin{pmatrix}
10.50 & 7.90 \\
7.90 & 6.45
\end{pmatrix}
\]

which yields an estimated correlation of \( r = 0.96 \)!
### M-estimators of location and scatter

At the normal model, the MLE estimators of $\mu$ and $\Sigma$ are given by:

$$\sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \quad \text{together with} \quad \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})' = \hat{\Sigma}$$

An M-estimator $(\hat{\mu}, \hat{\Sigma})$ is defined as the solution of

$$\sum_{i=1}^{n} W_1(d_i^2)(x_i - \hat{\mu}) = 0 \quad (1)$$

$$\frac{1}{n} \sum_{i=1}^{n} W_2(d_i^2)(x_i - \hat{\mu})(x_i - \hat{\mu})' = \hat{\Sigma} \quad (2)$$

where $d_i = \sqrt{(x_i - \hat{\mu})'\hat{\Sigma}^{-1}(x_i - \hat{\mu})}$ depends on the $\hat{\mu}$ and $\hat{\Sigma}$ themselves.

Estimating $\Sigma$ is the most difficult part by far! No ‘easy’ solution like MAD or $Q_n$.

<table>
<thead>
<tr>
<th>Low BV</th>
<th>affine equivariant</th>
<th>non affine equivariant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical mean &amp; covariance M-estimators</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High BV</td>
<td>Stahel-Donoho estimator MCD, MVE S-estimators MM-estimators</td>
<td>coordinatewise median spatial median, sign covariance OGK DetMCD</td>
</tr>
</tbody>
</table>
M-estimators of location and scatter

- There are conditions on $W_1$ and $W_2$ that ensure the existence, uniqueness and consistency of the estimators. Important conditions are that $\sqrt{t}W_1(t)$ and $tW_2(t)$ are bounded. An M-estimator for which $tW_2(t)$ is weakly increasing is called monotone, otherwise it is called redescending.

- M-estimators can be computed with an iterative algorithm.
  1. Start with initial choices $\hat{\mu}_0$ and $\hat{\Sigma}_0$, e.g. the coordinatewise median and the diagonal matrix with the squared coordinatewise MAD at the diagonal.
  2. At iteration $k$ we compute $d_{k1} = \sqrt{(x_i - \hat{\mu}_k)'\hat{\Sigma}_k^{-1}(x_i - \hat{\mu}_k)}$ and

$$\hat{\mu}_{k+1} = \frac{\sum_{i=1}^n W_1(d_{ki}^2)x_i}{\sum_{i=1}^n W_1(d_{ki}^2)},$$

$$\hat{\Sigma}_{k+1} = \frac{1}{n} \sum_{i=1}^n W_2(d_{ki}^2)(x_i - \hat{\mu}_{k+1})(x_i - \hat{\mu}_{k+1})'.$$

For a monotone M-estimator this algorithm always converges to the unique solution, no matter the choice of the initial values. For a redescending M-estimator the algorithm can converge to a bad solution.

Efficiency and robustness of M-estimators

Properties of M-estimators:
- Under some regularity conditions on $W_1$ and $W_2$, M-estimators are asymptotically normal.
- The influence function is bounded if $\sqrt{t}W_1(t)$ and $tW_2(t)$ are bounded.
- The asymptotic breakdown value of a monotone M-estimator satisfies

$$\epsilon^* \leq \frac{1}{p+1}.$$

Although monotone M-estimators attain the optimal value of 0.5 in the univariate case, this is no longer true in higher dimensions!

A monotone M-estimator is thus computationally attractive, but at the cost of a rather low breakdown value. Redescending M-estimators can have a higher breakdown value, but the algorithm may converge to a wrong solution.
Affine equivariant estimators with high breakdown value

<table>
<thead>
<tr>
<th>Low BV</th>
<th>Classical mean &amp; covariance M-estimators</th>
<th>non affine equivariant</th>
</tr>
</thead>
<tbody>
<tr>
<td>High BV</td>
<td>Stahel-Donoho estimator MCD, MVE S-estimators MM-estimators</td>
<td>coordinatewise median spatial median, sign covariance OGK DetMCD</td>
</tr>
</tbody>
</table>

The Stahel-Donoho outlyingness

Stahel (1981) and Donoho (1982) measured how outlying a point is.

It is based on the projection pursuit principle: a multivariate outlier should be outlying in at least one direction, but not necessarily the directions of the coordinate axes.

The Stahel-Donoho outlyingness of a point $\mathbf{x}$ relative to the data set $\{x_1, \ldots, x_n\}$ is given by

$$SDO_i = \sup_{a \in \mathbb{R}^p} \frac{|a'x - \text{med}_j(a'x_j)|}{\text{MAD}_j(a'x_j)}.$$

This projects the data in many directions $a$.

The projected data are univariate, so we can compute the outlyingness of $a'x$ as its ‘absolute robust z-score’ relative to $\{a'x_1, \ldots, a'x_n\}$.

The final outlyingness is the maximum of the univariate one over all directions.
The Stahel-Donoho outlyingness: example

Consider the following bivariate dataset with $n = 50$, with two outliers:

![Scatter plot with outliers](image)

Consider the observation marked in red:
The Stahel-Donoho outlyingness: example

In every direction it has a small outlyingness:

Now consider one of the outlying observations:
The Stahel-Donoho outlyingness: example

In at least one direction it has a large outlyingness:

$$\begin{array}{c}
-4 & -3 & -2 & -1 & 0 & 1 & 2 & 3 \\
-4 & -3 & -2 & -1 & 0 & 1 & 2 & 3 \\
4 & 3 & 2 & 1 & 0 & 1 & 2 & 3 \\n\end{array}$$

The Stahel-Donoho estimator: definition

The Stahel-Donoho estimator is defined as the weighted mean and covariance matrix of the $x_i$ with weights $w_i = W(SDO_i)$ where the weight function $W$ is a bounded, strictly positive and weakly decreasing function.

A typical weight function is

$$W(t) = \min \left(1, \frac{\chi^2_{p,0.95}}{t^2} \right).$$

If $t^2 W(t)$ is bounded (like here) the breakdown value of the Stahel-Donoho estimator is 50%. It was the first affine equivariant estimator of location and scatter with maximal breakdown value.
The Stahel-Donoho estimator: properties

- In the formula of the outlyingness $SDO_i$, also other estimators of univariate location and scale can be used, such as M-estimators of location and scale.

- The IF is bounded when using M-estimators of location and scale with bounded and monotone $\psi$ and $\rho$ functions.

- To compute the Stahel-Donoho estimator, the number of directions $a$ needs to be restricted to a finite set. These can be obtained by subsampling: take the directions orthogonal to hyperplanes spanned by random subsamples of size $p$. This yields an affine equivariant algorithm.

The MCD estimator

The MCD estimator (Rousseeuw, 1984) is an often used high-breakdown and affine equivariant estimator of location and scatter:

**Minimum Covariance Determinant estimator**

For fixed $h$, with $[n+p+1]/2 \leq h \leq n$,

- $\hat{\mu}_0$ is the mean of the $h$ observations for which the determinant of the sample covariance matrix is minimal;

- $\hat{\Sigma}_0$ is that covariance matrix (multiplied by a consistency factor).

The MCD estimator can only be computed when $h > p$, otherwise the covariance matrix of any $h$-subset will be singular. This condition is certainly satisfied when $n \geq 2p$. It is however recommended that $n \geq 5p$. 
Robustness of the MCD

- The influence function is bounded.
- The value of $h$ determines the breakdown value.

At samples in general position,

$$e^*_n = \min \left( \frac{n - h + 1}{n}, \frac{h - p}{n} \right)$$

The maximal breakdown value is achieved by taking $h = \lceil n + p + 1 \rceil / 2$.

Typical choices are $\alpha = h/n = 0.5$ or $\alpha = 0.75$, yielding a breakdown value of 50% and 25% respectively.

Efficiency of the MCD

The MCD is asymptotically normal, but it has a low efficiency. The efficiency increases with increasing $\alpha$.

For example, with $\alpha = 0.5$, the asymptotic relative efficiency of the diagonal elements of the MCD scatter matrix with respect to the sample covariance matrix, at the normal model, is only 6% when $p = 2$, and 20.5% when $p = 10$.

With $\alpha = 0.75$ the relative efficiencies are 26.2% for $p = 2$ and 45.9% for $p = 10$.

The efficiency of the MCD can be increased by applying a reweighting step:

First, compute the robust distances

$$\text{RD}_i = \sqrt{(x_i - \hat{\mu}_0)'\hat{\Sigma}_0^{-1}(x_i - \hat{\mu}_0)}$$
Reweighted MCD

Then put

\[ w_i = \begin{cases} 
1 & \text{if } RD_i \leq \sqrt{\chi^2_{p,0.975}} \\
0 & \text{otherwise.}
\end{cases} \]

Reweighted MCD (RMCD)

\[
\hat{\mu}_{RMCD} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}
\]

\[
\hat{\Sigma}_{RMCD} = \frac{1}{\sum_{i=1}^{n} w_i - 1} \sum_{i=1}^{n} w_i (x_i - \hat{\mu}_{RMCD}) (x_i - \hat{\mu}_{RMCD})'
\]

- The reweighting step does not decrease the breakdown value.
- It increases the efficiency: when \( \alpha = 0.5 \) the efficiency goes up to 45.5% for \( p = 2 \) and 82% for \( p = 10 \).

MCD example with \( \alpha = 0.5 \)

```r
> library(rrcov)
> resultMCD=CovMcd(x = log(Animals)); resultMCD

Robust Estimate of Location:
  body    brain
3.029    4.276

Robust Estimate of Covariance:
  body    brain
  body 18.86 14.16
  brain 14.16 11.03

> covMCD=getCov(resultMCD)
> cov2cor(covMCD)

  body    brain
  body 1.0000000 0.9816633
  brain 0.9816633 1.0000000
```
MCD example with $\alpha = 0.5$

We can also use the function `covMcd` from the `robustbase` library:

```r
> library(robustbase)
> resultMCD=covMcd(x=log(Animals)); resultMCD

Robust Estimate of Location:
  body   brain
   3.029  4.276

Robust Estimate of Covariance:
  body   brain
  body 18.86 14.16
  brain 14.16 11.03

> resultMCD$cor
      body  brain
  body 1.000000 0.981663
  brain 0.981663 1.000000
```

Outlier detection

For outlier detection, recompute the robust distances (this time based on the reweighted MCD):

$$RD_i = \sqrt{(x_i - \hat{\mu}_{RMCD})' \tilde{\Sigma}_{RMCD}^{-1} (x_i - \hat{\mu}_{RMCD})}$$

Flag observation $x_i$ as an outlier if $RD_i > \sqrt{\chi^2_{p,0.975}}$.

This is equivalent to flagging the observations that do not belong to the robust tolerance ellipsoid:

**Robust tolerance ellipsoid**

$$\{x; RD(x) \leq \sqrt{\chi^2_{p,0.975}}\}$$
Outlier detection

The MCD ellipse correctly flags the outliers in the animals data:

![Classical and MCD tolerance ellipse](image1)

![Robust distances based on MCD](image2)

The MCD-based robust distances do flag all the outliers!

Distance-distance plot

In dimensions $p > 2$, we cannot draw a scatterplot or a tolerance ellipsoid.

To explore the differences between a classical and a robust analysis we can draw a **distance-distance plot**, which plots the points $(MD_i; RD_i)$:
The univariate MCD estimator

In the special case of univariate data \((p = 1)\) the MCD becomes:

1. \(\hat{\mu}_0\) is the mean of the \(h\) observations for which the classical standard deviation is minimal;
2. \(\hat{\sigma}_0\) is that standard deviation (multiplied by a consistency factor).

Note that the optimal \(h\)-subset has to be contiguous, i.e. it must consist of successive ordered observations (no 'gaps').

So, in order to compute the univariate MCD we only have to loop over \(n - h + 1\) contiguous subsets. If we use an update formula for the variance the time complexity is only \(O(n \log(n))\).

However, as an estimator for univariate location and scale the MCD is outperformed by other methods (in terms of robustness and efficiency). Therefore the MCD is mainly useful for higher-dimensional data.

Computation of the MCD

Exact algorithm:
- Consider all \(h\)-subsets.
- Compute the mean and covariance matrix of each.
- Retain the subset with smallest covariance determinant.

But: infeasible for large \(n\) or \(p\)...

Approximate algorithms:
- Consider a selected set of \(h\)-subsets, starting from random subsets of size \(p + 1\). The most often used algorithm is FAST-MCD (Rousseeuw and Van Driessen, 1999).
- A faster, but not fully affine equivariant alternative is DetMCD (Hubert et al., 2012). We will describe this later.
FAST-MCD

Computation of the raw estimates for small to moderate data sizes \( n \leq 600 \):

1. For \( m = 1 \) to 500:
   - Draw a random subset of size \( p + 1 \) and compute its mean and covariance matrix.
   - Apply a C-step:
     1. Compute robust distances \( RD_i \) based on the most recent mean and covariance estimate.
     2. Take the \( h \) observations with smallest robust distance.
     3. Compute mean and covariance matrix of this \( h \)-subset.
   - Apply a second C-step.

2. Retain the 10 \( h \)-subsets with smallest covariance determinant.

3. Apply C-steps on these subsets until convergence.

4. Retain the \( h \)-subset with smallest covariance determinant.

C-steps always decrease the determinant of the covariance matrix!

As there are only a finite number of \( h \)-subsets, convergence to a (local) minimum is guaranteed.

The algorithm is not guaranteed to yield the global minimum. The fixed number of initial \((p + 1)\)-subsets (500) is a compromise between robustness and computation time.

At larger data sets \((n > 600)\), the algorithm randomly splits the data set in disjoint subsets. First, C-steps are applied within the subsets, and next in the full data set.
The process changed after the first 100 points, and between index 491–565 it was out of control.
FAST-MCD: Philips example

Also the distance-distance plot highlights the out-of-control period:

FAST-MCD: Digital sky survey

The Digital Palomar Sky Survey (DPOSS) contains data about celestial objects (light sources). After removing physically impossible data, we have \( n = 132402 \) objects with \( p = 6 \) variables. The classical Mahalanobis distances (and their chi-squared QQ-plot) look homogeneous:
FAST-MCD: Digital sky survey

The robust distances from FAST-MCD give a different picture:

The DD plot makes a clear distinction between stars (lower group) and galaxies:
Software for MCD

Implementations of the FAST-MCD algorithm are widely available:

- R: as the function `CovMcd` in the package `rrcov`, and as the function `covMcd` in the package `robustbase`
- S-PLUS: as the built-in function `cov.mcd`
- Matlab: as the function `mcdcov` in the toolbox LIBRA (wis.kuleuven.be/stat/robust), and the PLS toolbox of Eigenvector Research (www.eigenvector.com)
- in SAS/IML Version 7+, and in PROC ROBUSTREG in SAS Version 9+
- STATA, see http://ideas.repec.org/a/tsj/stataj/v10y2010i2p259-266.html

Note that some functions use $\alpha = 0.5$ as default, yielding a breakdown value of 50%, whereas other implementations use the default $\alpha = 0.75$.

The MVE estimator

The MVE (Rousseeuw, 1985) is one of the oldest robust covariance estimators that is affine equivariant and has a positive breakdown value.

Minimum Volume Ellipsoid

For fixed $h$, with $\lfloor (n + p + 1)/2 \rfloor \leq h \leq n$,

$$(\hat{\mu}, \hat{\Sigma}) = \arg\min_{\mu, \Sigma} |\hat{\Sigma}|$$

over all real $\mu$ and symmetric positive definite $\Sigma$ that satisfy

$$\#\{i; d_i = \sqrt{(x_i - \hat{\mu})'\hat{\Sigma}^{-1}(x_i - \hat{\mu})} \leq c^2 \} \geq h \}.$$

The estimator is thus defined by the ellipsoid with minimal volume which contains (at least) $h$ observations.

Its breakdown value is optimal (50%) when $h = \lfloor (n + p + 1)/2 \rfloor$, but the MVE lacks asymptotic normality.
S-estimators of location and scatter

Remember the definition of an M-estimator $\hat{\sigma}_M$ of univariate scale:

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

S-estimator of location and scatter (Rousseeuw and Leroy 1987)

$$(\hat{\mu}, \hat{\Sigma}) = \text{argmin}_{\mu, \Sigma} |\hat{\Sigma}|$$

over all real $\mu$ and symmetric positive definite $\Sigma$ that satisfy

$$\frac{1}{n} \sum_{i=1}^{n} \rho(d_i) = \delta$$

with $d_i = \sqrt{(x_i - \hat{\mu})' \hat{\Sigma}^{-1} (x_i - \hat{\mu})}$ and $\rho$ a smooth bounded $\rho$-function.

Efficiency of S-estimators

- To obtain (Fisher-)consistency at normal distributions, we set $\delta$ to

  $$\delta = E_{N_p(0, I)}(\rho(\|X\|))$$

- S-estimators are asymptotically normal. Their efficiency at the gaussian model is somewhat better than the efficiency of the RMCD, especially in higher dimensions.

  For example, the diagonal element of the bisquare S scatter matrix with 50% breakdown value has an asymptotic relative efficiency of 50.2% for $p = 2$, and 92% for $p = 10$. (RMCD: 45.5% for $p = 2$ and 82% for $p = 10$).

- S-estimators are smoothed versions of the MVE, which corresponds to a function $\rho$ that only takes on the values 0 and 1.
Robustness of S-estimators

- The breakdown value of both the location and scatter estimator is:
  \[ \varepsilon^* = \min \left( \frac{\delta}{\rho(\infty)}, 1 - \frac{\delta}{\rho(\infty)} \right) \]

  if the data are in general position.

  The tuning parameter in \( \rho_c \) thus determines the robustness, as well as the efficiency.

- To obtain a bounded influence function, it is required that \( \psi'(x) \) and \( \psi(x)/x \) are bounded and continuous. The influence function of S-estimators can then be seen as a smoothed version of the MCD’s influence function.

- To compute an S-estimator, the FAST-S algorithm can be used (Salibian-Barrera and Yohai, 2006). It is similar to FAST-MCD.

S-estimators: example

```r
> resultS=CovSest(log(Animals)); resultS
Call: CovSest(x = log(Animals))
  -> Method: S estimation: S-FAST

Robust Estimate of Location:
[1] 3.271 4.345

Robust Estimate of Covariance:
     body   brain
body 22.72  17.24
brain 17.24  13.36

> covS=getCov(resultS)
> cov2cor(covS)
     body   brain
body 1.0000000 0.9898186
brain 0.9898186 1.0000000
```
S-estimators: example

```r
> plot(resultS, which="tolEllipse", classic=TRUE)
```

Classical and S tolerance ellipse

```
log(body)
```

```
log(brain)
```

MM-estimators of location and scatter

MM-estimators combine **high robustness** with **high efficiency**.

They are based on two rho functions $\rho_0$ and $\rho_1$. The first rho function is chosen to obtain a high breakdown value. The second rho function is chosen to achieve a high efficiency.

To construct an MM-estimator, note that a scatter matrix can be separated into a scale estimate and a shape matrix:

Put $\Gamma := |\Sigma|^{-1/p} \Sigma$, then

$$|\Gamma| = 1 \quad \text{and} \quad \Sigma = |\Sigma|^{1/p} \Gamma.$$  

We call $|\Sigma|^{1/2p}$ the **scale** estimate, and $\Gamma$ the **shape matrix**.
Let \((\tilde{\mu}, \tilde{\Sigma})\) be an S-estimator with rho function \(\rho_0\). Denote \(\hat{\sigma}^2 = |\tilde{\Sigma}|^{1/p}\).

The MM-estimator for location and shape \(\left(\hat{\mu}, \hat{\Gamma}\right)\) minimizes

\[
\frac{1}{n} \sum_{i=1}^{n} \rho_1 \left( \frac{\sqrt{(x_i - \mu)\Gamma^{-1}(x_i - \mu)}}{\hat{\sigma}} \right)
\]

among all real \(\mu\) and symmetric positive definite \(\Gamma\) with \(|\Gamma| = 1\).

The MM-estimator for the covariance matrix is then \(\hat{\Sigma} = \hat{\sigma}^2 \hat{\Gamma}\).

The location and shape estimates inherit the breakdown value of the auxiliary scale. Thus one typically chooses an S-estimator with 50% breakdown value.

For a bisquare \(\rho_0\), \(c = 1.547\) yields a 50% breakdown value.

The influence functions (and thus asymptotic variance) of MM-estimators for location and scatter equal those of M-estimators of location and scatter that use the function \(\rho_1\).

For a bisquare \(\rho_1\), \(c = 4.685\) yields 95% efficiency (at the normal model).

However, MM-estimators with high efficiency are less robust.

In particular, they tend to give too much weight to ‘fairly nearby’ outliers, unlike methods with a ‘hard’ objective function like MCD and MVE.

The FAST-MM algorithm starts with FAST-S and then applies IRLS steps to minimize (3).
MM-estimators: example

> resultMM=CovMMest(log(Animals)); resultMM
Call: CovMMest(x = log(Animals))
--> Method: MM-estimates

Robust Estimate of Location:
[1] 3.086 4.427

Robust Estimate of Covariance:

<table>
<thead>
<tr>
<th></th>
<th>body</th>
<th>brain</th>
</tr>
</thead>
<tbody>
<tr>
<td>body</td>
<td>12.036</td>
<td>9.021</td>
</tr>
<tr>
<td>brain</td>
<td>9.021</td>
<td>7.272</td>
</tr>
</tbody>
</table>

> covMM=getCov(resultMM)
> cov2cor(covMM)

<table>
<thead>
<tr>
<th></th>
<th>body</th>
<th>brain</th>
</tr>
</thead>
<tbody>
<tr>
<td>body</td>
<td>1.00000000</td>
<td>0.9642449</td>
</tr>
<tr>
<td>brain</td>
<td>0.9642449</td>
<td>1.0000000</td>
</tr>
</tbody>
</table>

Classical and MM tolerance ellipse
The coordinatewise median

Coordinatewise median:

\[ \hat{\mu} = \left( \text{med}_{i=1}^{n} x_{i1}, \text{med}_{i=1}^{n} x_{i2}, \ldots, \text{med}_{i=1}^{n} x_{ip} \right)'. \]

- Easy to compute and to interpret
- 50% breakdown value!
- not affine equivariant, and not even orthogonally equivariant
- \( \hat{\mu} \) does not have to lie in the convex hull of the sample when \( p \geq 3 \). As an example, consider the set \( \{(1, 0, 0)', (0, 1, 0)', (0, 0, 1)\}' \) whose convex hull does not contain the coordinatewise median \( (0, 0, 0)' \).
The spatial median

The spatial median, also known as the $L^1$ location estimator, is defined as

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \| x_i - \mu \| .$$

This is equivalent to

$$\sum_{i=1}^{n} \frac{x_i - \hat{\mu}}{\| x_i - \hat{\mu} \|} = 0 . \quad (4)$$

- 50% breakdown value, bounded influence function
- not affine equivariant, but orthogonal equivariant
- Computation: Equation (4) corresponds to equation (1) of M-estimators, with $W_1(t) = 1/\sqrt{t}$. We can thus use the iterative algorithm with $\Sigma = I$ fixed. Other algorithms are discussed in Fritz et al. (2012).

Geometric interpretation: take a point $\mu$ in $\mathbb{R}^p$ and project all observations onto a sphere around $\mu$. If the mean of these projections equals $\mu$, then $\mu$ is the spatial median.

When projecting all data points on a sphere around the star, the mean of these projections (depicted as crosses) does not equal the center of the sphere. For the triangle it does, so it is the spatial median. Observation 11 only has a small effect.
The spatial sign covariance matrix

The spatial sign covariance matrix (SSCM) is the classical covariance matrix computed on the projected data points (Visuri et al., 2000).

**Spatial sign covariance estimator**

\[
\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(x_i - \hat{\mu})(x_i - \hat{\mu})'}{\|x_i - \hat{\mu}\| \|x_i - \hat{\mu}\|}
\]

with \(\hat{\mu}\) the spatial median.

- 50% breakdown value, bounded influence function
- not affine equivariant, only orthogonally equivariant.
- the resulting scatter matrix can give a very poor fit, even if the data is uncontaminated. In that case its eigenvectors are consistent, but its eigenvalues are not.

The orthogonalized Gnanadesikan-Kettenring estimator

- Introduced by Maronna and Zamar (2002)
- Fast to compute, also in rather high dimensions
- Not affine or orthogonal equivariant, only scale equivariant

It is inspired by the fact that the classical correlation between two variables \(Y_j\) and \(Y_k\) satisfies:

\[
\text{Cor}(Y_j, Y_k) = \frac{\text{Var}(z(Y_j) + z(Y_k)) - \text{Var}(z(Y_j) - z(Y_k))}{\text{Var}(z(Y_j) + z(Y_k)) + \text{Var}(z(Y_j) - z(Y_k))}
\]

where \(z(Y_j) = (Y_j - \text{ave}(Y_j))/\text{Std}(Y_j)\) contains the classical z-scores of \(Y_j\).

Gnanadesikan and Kettenring (1972) proposed to compute a robust correlation between 2 variables by replacing the classical z-scores and variances on the right hand side by robust versions.

This works, but the resulting correlation matrix need not be PSD...
The OGK estimator: definition

OGK = \textbf{orthogonalized} Gnanadesikan-Kettenring estimator:

1. Let \( m(.) \) and \( s(.) \) be robust univariate estimators of location and scale.
2. Construct \( y_i = D^{-1} x_i \) for \( i = 1, \ldots, n \) with \( D = \text{diag}(s(X_1), \ldots, s(X_p)) \).
3. Compute the ‘correlation matrix’ \( U \) of the variables of \( Y = (Y_1, \ldots, Y_p) \), given by
   \[
   u_{jk} = \frac{(s(Y_j + Y_k)^2 - s(Y_j - Y_k)^2)}{(s(Y_j + Y_k)^2 + s(Y_j - Y_k)^2)}.
   \]
   This matrix is symmetric but not necessarily PSD.
4. Put the eigenvectors of \( U \) as columns in a matrix \( E \) and
   \begin{itemize}
   \item project the data on these eigenvectors, i.e. \( V = YE \);
   \item compute ‘robust variances’ of \( V = (V_1, \ldots, V_p) \), i.e.
     \[
     \Lambda = \text{diag}(s^2(V_1), \ldots, s^2(V_p));
     \]
   \item Set the \( p \times 1 \) vector \( \hat{\mu}(Y) = Em \) where \( m = (m(V_1), \ldots, m(V_p))' \)
     and compute the positive definite matrix \( \hat{\Sigma}(Y) = E\Lambda E' \).
   \end{itemize}
5. Transform back to \( X \), i.e. \( \hat{\mu}(X) = D\hat{\mu}(Y) \) and \( \hat{\Sigma} = D\hat{\Sigma}(Y)D' \).

Step 4 of the method (the ‘orthogonalization’) uses the fact that the eigenvalues of the covariance matrix equal the variances of the data projected on the eigenvectors. Here the eigenvalues are estimated by a robust univariate scale estimator. As these estimates are nonnegative, the new scatter matrix \( E\Lambda E' \) is positive semidefinite.

When high-breakdown estimators are chosen for \( m \) and \( s \), then the breakdown value of the OGK estimator is 50%.

Also a reweighting step can be added, which increases the efficiency. The proposed cutoff for the robust distances is

\[
c = \frac{\chi^2_{p,0.9}}{\chi^2_{p,0.5}} \text{med}(d_1, \ldots, d_n)
\]

with \( d_i \) the robust distances from the raw OGK estimates.

The reweighted estimators are ‘approximately’ affine equivariant.
The DetMCD algorithm

Deterministic algorithm for MCD (Hubert et al., 2012).

Overall idea:

- Compute several 'promising' $h$-subsets, based on
  - transformations of variables
  - easy-to-compute robust estimators of location and scatter.
- Apply C-steps until convergence.

This yields a fast algorithm which is at least as robust as FAST-MCD, but not fully affine equivariant.

Preprocessing: standardize $X$ by subtracting the columnwise median and dividing by the columnwise $Q_n$ scale estimator:

- this makes the final estimates location and scale equivariant.
- yields the standardized dataset $Z$ with rows $z'_i$ and columns $Z_j$.

Construct six initial estimates $\hat{\mu}_k(Z)$ and $\hat{\Sigma}_k(Z)$ for center and scatter:

- Obtain six preliminary estimates $S_k$ for covariance/correlation matrix of $Z$.
- Compute eigenvectors $E$ of $S_k$ and put $B = ZE$.
- Estimate covariance of $Z$ by $\hat{\Sigma}_k(Z) = ELE'$ with $L = \text{diag}(Q_n(B_1)^2, \ldots, Q_n(B_p)^2)$.
- Estimate the center: $\hat{\mu}_k(Z) = \hat{\Sigma}_k^{1/2}(\text{med}(Z\hat{\Sigma}_k^{-1/2}))$.

For each initial estimate do:

- Compute statistical distances $d_{ik} = d(z_i, \hat{\mu}_k(Z), \hat{\Sigma}_k(Z))$.
- Initial $h_0$-subset: $h_0 = \lceil n/2 \rceil$ observations with smallest $d_{ik}$.
- Compute the statistical distances $d_{ik}^*$ based on these $h_0$ observations. Take the $h$ points with smallest $d_{ik}^*$ and apply C-steps until convergence.

- Retain the $h$-subset with smallest covariance determinant.
**DetMCD: Preliminary estimates**

1. Take hyperbolic tangent ('sigmoid') of the standardized data:

   \[ Y_j = \tanh(Z_j) \quad \forall j = 1, \ldots, p. \]

   Take Pearson correlation matrix of \( Y \)
   \[ S_1 = \text{corr}(Y). \]

2. Consider the Spearman correlation matrix:
   \[ S_2 = \text{corr}(R) \]
   where \( R_j \) is the rank of \( Z_j \).

3. Compute normal scores \( T_j \) from the ranks \( R_j \):
   \[ T_j = \Phi^{-1} \left( \frac{R_j - \frac{1}{3}}{n + \frac{1}{3}} \right) \]
   where \( \Phi(.) \) is the standard normal cdf, and put \( S_3 = \text{corr}(T) \).

4. Related to spatial sign covariance matrix:
   Define \( k_i = \frac{z_i}{\|z_i\|} \) and let
   \[ S_4 = \frac{1}{n} \sum_{i=1}^{n} k_i k_i' \]
   (Here the center is estimated by the coordinatewise median instead of the spatial median.)

5. First step of the BACON algorithm (Billor et al., 2000):
   Consider the \( \lceil n/2 \rceil \) standardized observations \( z_i \) with smallest norm, and compute their mean and covariance matrix.

6. The raw OGK estimator of location and scatter.
**DetMCD: Properties**

- In moderate dimensions (say, $p \leq 10$): faster than FAST-MCD and equally robust.
- In higher dimensions: faster than FAST-MCD and more robust, especially when there is much contamination.
- Deterministic: does not depend on any random selection.
- Permutation invariant.
- Nearly affine equivariant.
- Initial estimates do not yet depend on the value $h$ which determines the breakdown value. This makes it easy to compute DetMCD for several $h$-values, and to see whether at some $h$ there is a substantial change in the objective function or the estimates ("monitoring").

**When to use DetMCD**

When should we use FAST-MCD and when DetMCD? Recommendation:

- When $p \leq 10$ run FAST-MCD.
- When $p$ is larger than this it becomes harder or even infeasible to draw enough initial subsets, and then it is better to run DetMCD.

DetMCD is useful as a building block for multivariate analysis (multivariate regression, exponential smoothing, calibration, ...).
Robust Covariance Estimation: R

- FAST-MCD: the function `CovMcd` in the package `rrcov`, and the function `covMcd` in the package `robustbase`.
- MVE, FAST-S: the package `rrcov` contains implementations of the MVE (`CovMve`) and S-estimators (`CovSest`), as well as several other robust estimators of location and scatter (MM-estimators, Stahel-Donoho, OGK).
- DetMCD: use the function `covMcd` in the package `robustbase` with optional argument `nsamp = "deterministic"`.

Robust Covariance Estimation: Matlab

- FAST-MCD: the function `mcdcov` in the toolbox LIBRA (`wis.kuleuven.be/stat/robust`), and the PLS toolbox of Eigenvector Research (`www.eigenvector.com`). Default: $\alpha = 0.75$, yielding a breakdown value of 25%.
- Also the FSDA toolbox available at `www.riani.it/MATLAB.htm` contains implementations of FastMCD, S, and MM-estimators.
- DetMCD: available in LIBRA. It has OGK as a subroutine.