Session 3: Robust multivariate methods
Winter course, CMStatistics 2016

Mia Hubert, Peter Rousseeuw, Stefan Van Aelst

Department of Mathematics
KU Leuven, Belgium

December 6–7, 2016

Outline of the course

1. General notions of robustness
2. Robustness for univariate data
3. Robust multivariate methods
4. Robust regression
5. Robust principal component analysis
6. Inference
7. Multivariate and functional depth
8. High dimensional data and sparsity
9. Cellwise outliers

Multivariate location and scatter: Outline

Multivariate location and scatter

Data: $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ where the observations $\boldsymbol{x}_i$ are p-variate column vectors.

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

$$X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)' = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

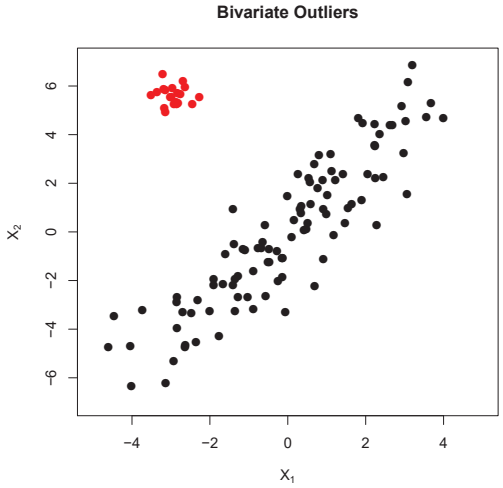
Model for the observations:

$$\boxed{\boldsymbol{x}_i \sim N_p(\boldsymbol{\mu}, \Sigma)}$$

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

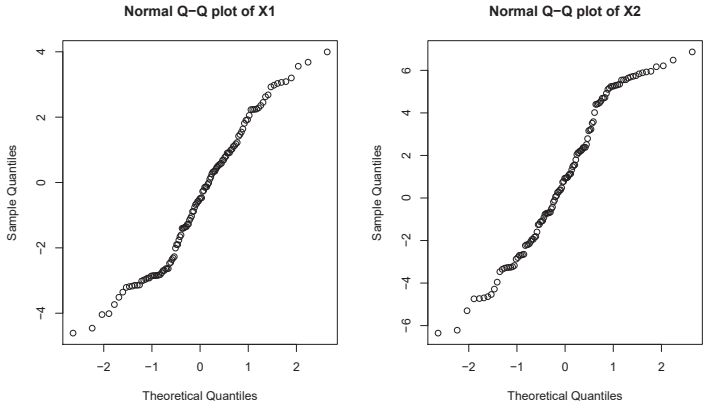
Outlier detection

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 $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ that are affine equivariant.

Affine equivariance

$$\hat{\boldsymbol{\mu}}(\{A\boldsymbol{x}_1 + \boldsymbol{b}, \ldots, A\boldsymbol{x}_n + \boldsymbol{b}\}) = A\hat{\boldsymbol{\mu}}(\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}) + \boldsymbol{b}$$
$$\hat{\Sigma}(\{A\boldsymbol{x}_1 + \boldsymbol{b}, \ldots, A\boldsymbol{x}_n + \boldsymbol{b}\}) = A\hat{\Sigma}(\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\})A'$$

for any nonsingular matrix A and any vector $\boldsymbol{b}$.

Affine equivariance implies that the estimator transforms well under any non-singular reparametrization of the space of the $\boldsymbol{x}_i$.

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**

$$\hat{\boldsymbol{\mu}}(\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}) = (\underset{i=1}{\overset{n}{\text{med}}}\, x_{i1}\ ,\ \ldots\ ,\ \underset{i=1}{\overset{n}{\text{med}}}\, x_{ip})'$$

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 $\{\boldsymbol{x}_1 + \boldsymbol{b}, \ldots, \boldsymbol{x}_n + \boldsymbol{b}\}$ 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 $\hat{\boldsymbol{\mu}}$ breaks down when it can be carried outside any bounded set.
Every affine equivariant location estimator satisfies

$$\varepsilon_n^*(\hat{\boldsymbol{\mu}}, X_n) \leqslant \frac{1}{n}\left\lfloor \frac{n+1}{2} \right\rfloor .$$

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

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

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:

	affine equivariant	non affine equivariant
Low BV	Classical mean, covariance M-estimators Convex peeling Tukey median Simplicial median Oja median	
High BV	Stahel-Donoho estimator MCD, MVE S-estimators MM-estimators	coordinatewise median spatial median, sign covariance OGK DetMCD

Classical estimators

	affine equivariant	non affine equivariant
Low BV	Classical mean, covariance M-estimators Convex peeling Tukey median Simplicial median Oja median	
High BV	Stahel-Donoho estimator MCD, MVE S-estimators MM-estimators	coordinatewise median spatial median, sign covariance OGK DetMCD

Classical estimators

The classical estimators for $\boldsymbol{\mu}$ and Σ are the empirical mean and covariance matrix:

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

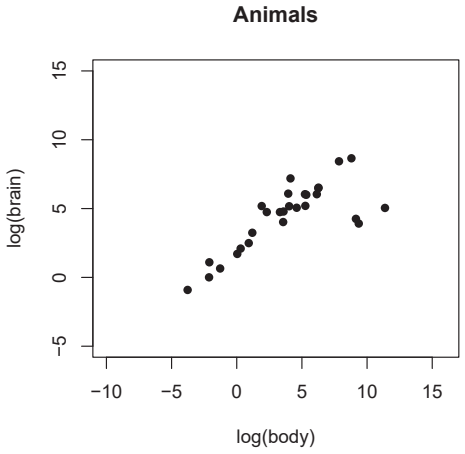
$$S_n = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})'.$$

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:



Tolerance ellipsoid

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

Mahalanobis distance

$$\mathrm{MD}(\boldsymbol{x}) = \sqrt{(\boldsymbol{x} - \bar{\boldsymbol{x}}_n)' S_n^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}}_n)}$$

Classical tolerance ellipsoid

$$\{\boldsymbol{x}; \mathrm{MD}(\boldsymbol{x}) \leqslant \sqrt{\chi^2_{p,0.975}}\}$$

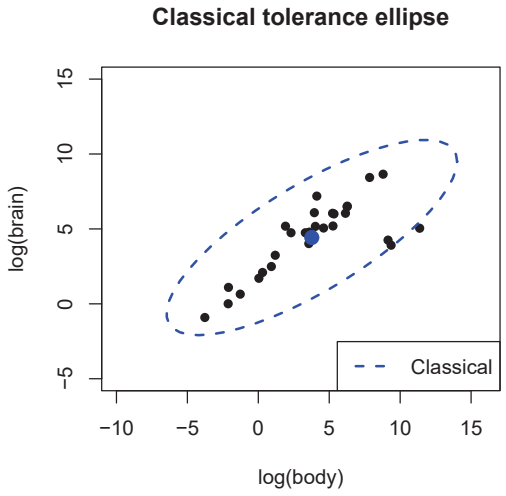
with $\chi^2_{p,0.975}$ the 97.5% quantile of the χ^2-distribution with p degrees of freedom.

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

We could flag observation $\boldsymbol{x}_i$ 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:



Point estimates

On all data points:

$$\bar{\boldsymbol{x}}_{28} = (3.77 \;\; 4.425)'$$

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

This yields an estimated correlation of $r = 7.05/\sqrt{14.22 * 5.76} = 0.78$.
On the reduced data set (without observations 6, 16 and 26):

$$\bar{\boldsymbol{x}}_{25} = (3.03 \;\; 4.428)'$$

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

which yields an estimated correlation of $r = 0.96$!

M-estimators of location and scatter

	affine equivariant	non affine equivariant
Low BV	Classical mean, covariance	
	M-estimators	
	Convex peeling	
	Tukey median	
	Simplicial median	
	Oja median	
High BV	Stahel-Donoho estimator	coordinatewise median
	MCD, MVE	spatial median, sign covariance
	S-estimators	OGK
	MM-estimators	DetMCD

M-estimators of location and scatter

At the normal model, the MLE estimators of $\boldsymbol{\mu}$ and Σ are given by:

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

M-estimators of location and scatter

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

$$\sum_{i=1}^{n} W_1(d_i^2)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}) = \mathbf{0} \tag{1}$$

$$\frac{1}{n}\sum_{i=1}^{n} W_2(d_i^2)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})' = \hat{\Sigma} \tag{2}$$

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

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 $\hat{\boldsymbol{\mu}}_0$ 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 $d_{ki} = \sqrt{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)'\hat{\Sigma}_k^{-1}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)}$ and

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

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

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

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

  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 larger breakdown value, but the algorithm may converge to a wrong solution.

Affine equivariant estimators with high breakdown value

	affine equivariant	non affine equivariant
Low BV	Classical mean, covariance M-estimators Convex peeling Tukey median Simplicial median Oja median	
High BV	Stahel-Donoho estimator MCD, MVE S-estimators MM-estimators	coordinatewise median spatial median, sign covariance OGK DetMCD

The Stahel-Donoho estimator

The **Stahel-Donoho estimator** was the first affine equivariant estimator of location and scatter with 50% breakdown value (Stahel, 1981; Donoho, 1982).

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.

1. Data are projected on a direction $\boldsymbol{a}$.
2. For each data point $\boldsymbol{x}_i$ its absolute residual is computed, where the residual is defined as the robustly standardized distance of its projection $\boldsymbol{a}'\boldsymbol{x}_i$ to the median of the projected observations.
3. For each data point, the largest absolute residual over all directions $\boldsymbol{a}$ is retained. This is called the *outlyingness* of $\boldsymbol{x}_i$.
4. The Stahel-Donoho estimate of location and scatter is a weighted mean and covariance matrix, where the weight function $W(t)$ is a strictly positive and weakly decreasing function of the outlyingnesses of the $\boldsymbol{x}_i$.

The Stahel-Donoho estimator: definition

Stahel-Donoho estimator

The Stahel-Donoho outlyingness of a point $\boldsymbol{x}_i$ is given by

$$\mathrm{SDO}_i = \sup_{\boldsymbol{a}\in\mathbb{R}^p} \frac{|\boldsymbol{a}'\boldsymbol{x}_i - \mathrm{med}_j(\boldsymbol{a}'\boldsymbol{x}_j)|}{\mathrm{MAD}_j(\boldsymbol{a}'\boldsymbol{x}_j)} \ . \tag{3}$$

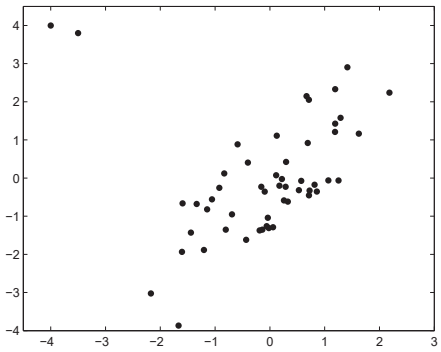
A typical weight function is

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

The Stahel-Donoho estimator is then defined as the weighted mean and covariance matrix of the $\boldsymbol{x}_i$ with weights $w_i = W(\mathrm{SDO}_i)$.

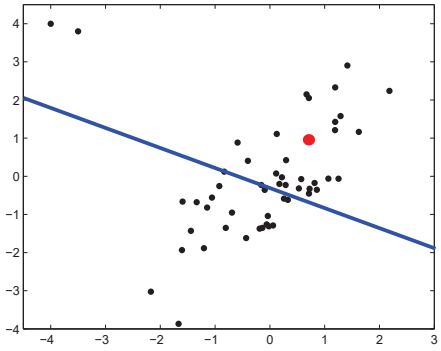
The Stahel-Donoho estimator: example

Consider the following two-dimensional dataset, with 50 observations and two outliers:



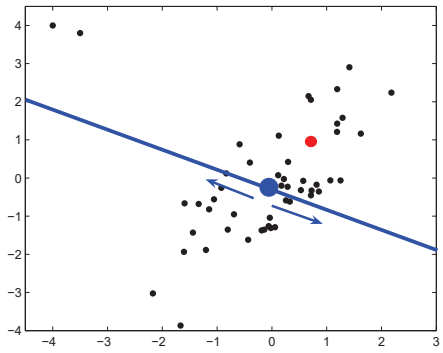
The Stahel-Donoho estimator: example

Consider the observation marked in red:



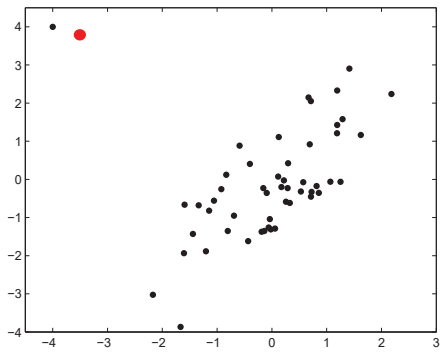
The Stahel-Donoho estimator: example

In every direction it has a small outlyingness:



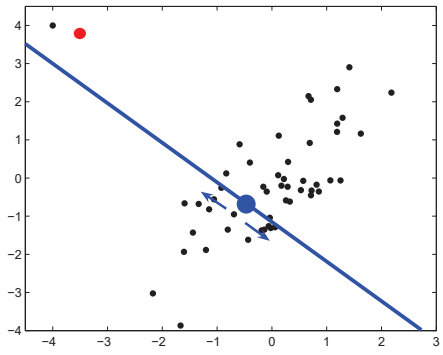
The Stahel-Donoho estimator: example

Now consider one of the outlying observations:



The Stahel-Donoho estimator: example

In at least one direction it has a large outlyingness:



The Stahel-Donoho estimator: properties

- If $tW(t)$ and $t^2W(t)$ are bounded, then the breakdown value of the SD-estimator is 50%.
- In (3) 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 ψ and ρ functions.
- To compute the Stahel-Donoho estimator, the number of directions $\boldsymbol{a}$ 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.
- For many outliers or in high dimensions it can happen that fewer than $p+1$ observations receive a weight $w_i > 0$, leading to a singular $\hat{\Sigma}$. We can then replace the w_i by 0–1 weights that are set to 1 for the $[n+p+1]/2$ points with lowest outlyingness. We call this the MSDE estimator.

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 $[n+p+1]/2 \leqslant h \leqslant n$,

1. $\hat{\boldsymbol{\mu}}_0$ is the mean of the h observations for which the determinant of the sample covariance matrix is minimal;
2. $\hat{\Sigma}_0$ 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 $n \geqslant 2p$. It is however recommended that $n > 5p$.

Robustness of the MCD

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

  At samples in general position,

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

  The maximal breakdown value is achieved by taking $h = [n+p+1]/2$.

  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 α.

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**

$$\text{RD}_i = \sqrt{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_0)' \hat{\Sigma}_0^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_0)}$$

Reweighted MCD

Then put

$$w_i = \begin{cases} 1 & \text{if} \quad \text{RD}_i \leqslant \sqrt{\chi^2_{p,0.975}} \\ 0 & \text{otherwise.} \end{cases}$$

Reweighted MCD (RMCD)

$$\hat{\boldsymbol{\mu}}_{RMCD} = \frac{\sum_{i=1}^n w_i \boldsymbol{x}_i}{\sum_{i=1}^n w_i}$$

$$\hat{\Sigma}_{RMCD} = \frac{1}{\sum_{i=1}^n w_i - 1} \sum_{i=1}^n w_i (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{RMCD})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{RMCD})'$$

- 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 $p = 10$.

Reweighted MCD: example

Example: RMCD with $\alpha = 0.5$ yields:

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

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
```


Reweighted MCD: example

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

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

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  1.0000000 0.9816633
brain 0.9816633 1.0000000
```

Outlier detection

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

$$\text{RD}_i = \sqrt{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{RMCD})' \hat{\Sigma}^{-1}_{RMCD} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{RMCD})}$$

Flag observation $\boldsymbol{x}_i$ as an outlier if $\text{RD}_i > \sqrt{\chi^2_{p,0.975}}$.

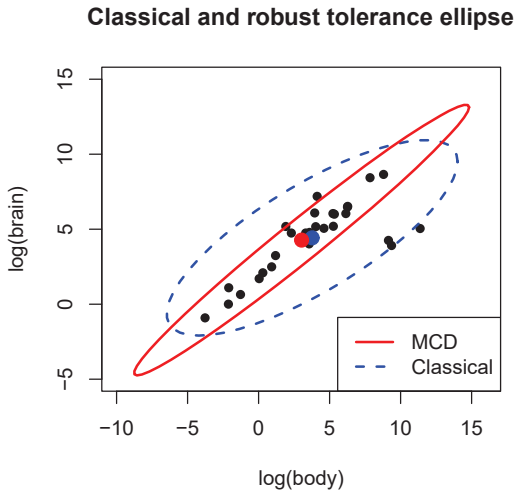
This is equivalent to flagging the observations that do not belong to the robust tolerance ellipsoid:

Robust tolerance ellipsoid

$$\{\boldsymbol{x}; \text{RD}(\boldsymbol{x}) \leqslant \sqrt{\chi^2_{p,0.975}}\}$$

Outlier detection

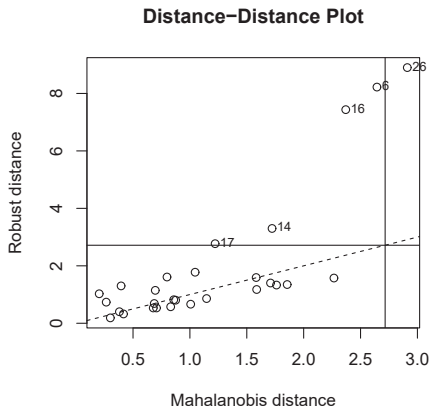
Outlier detection based on RMCD correctly flags the outliers in the animals data:



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 $(\mathrm{MD}_i, \mathrm{RD}_i)$:



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.

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 $n \leqslant 600$:

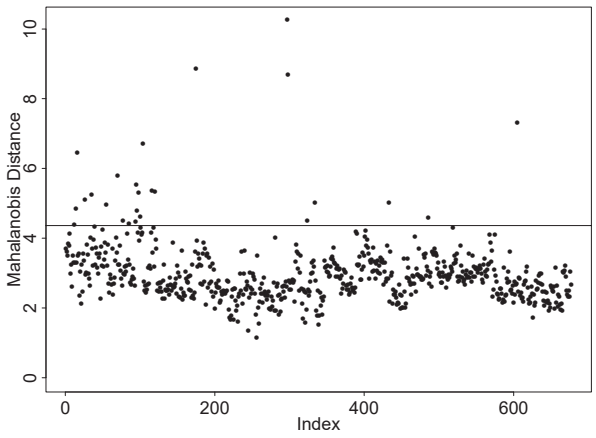
1. For $m = 1$ to 500:
   - Draw a random subset of size $p+1$ and compute its mean and covariance matrix.
   - Apply a C-step:
     1. Compute robust distances RD_i based on the most recent mean and covariance estimate.
     2. Take the h observations with smallest robust distance.
     3. Compute mean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

FAST-MCD

- 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.

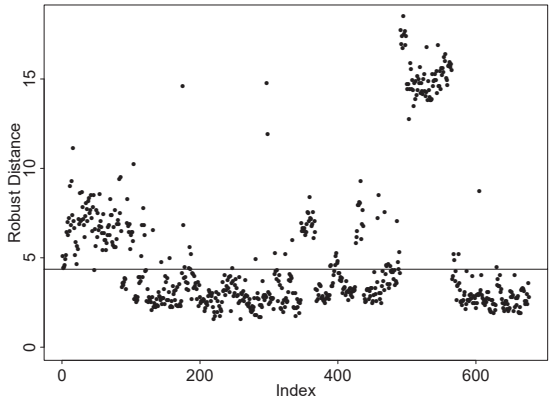
FAST-MCD: Philips example

Data from Philips Mecoma about the production of thin metal plates, with $n = 677$ and $p = 9$ characteristics, for statistical process control. The classical Mahalanobis distances (and their chi-squared QQ-plot) indicate a few outliers:



FAST-MCD: Philips example

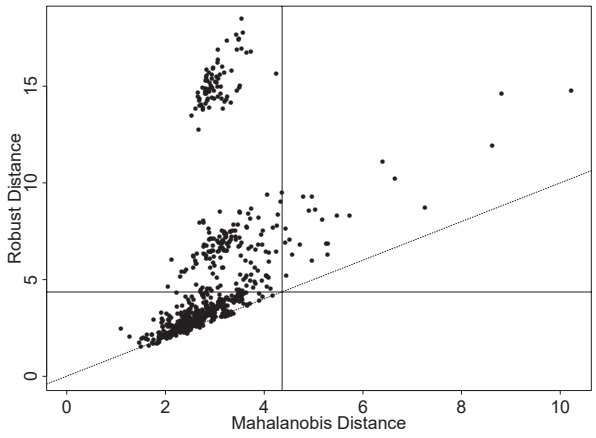
The robust distances from FAST-MCD give a different picture:



The process changed after the first 100 points, and between index 491–565 it was out of control.

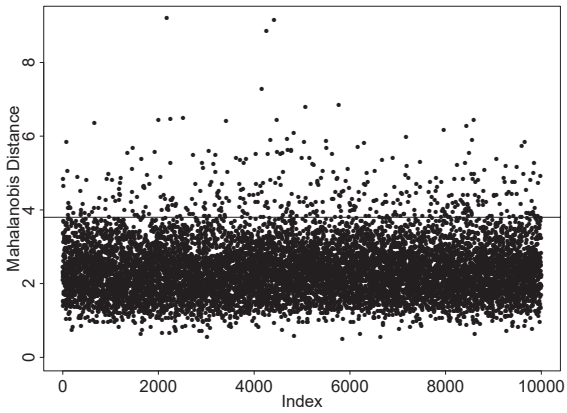
FAST-MCD: Philips example

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



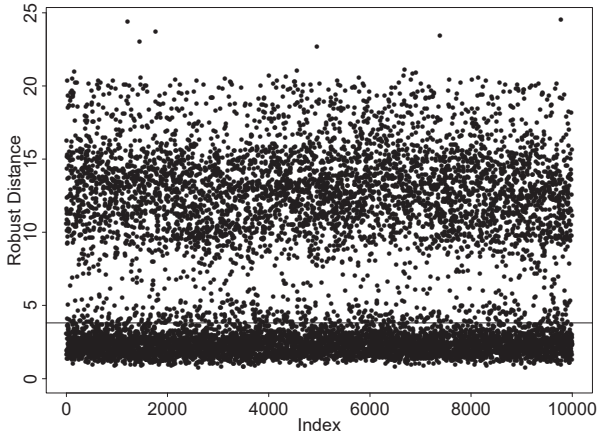
FAST-MCD: Digital sky survey

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:



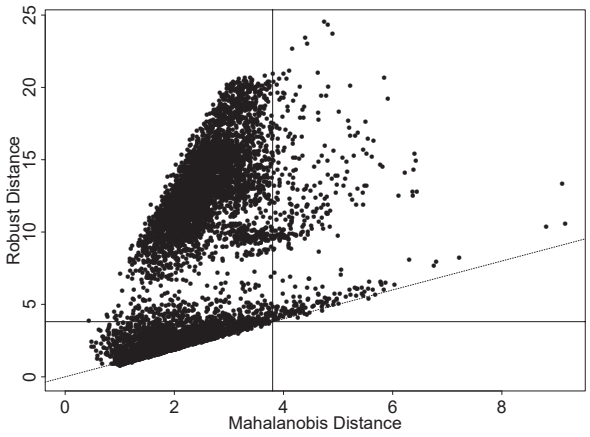
FAST-MCD: Digital sky survey

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



FAST-MCD: Digital sky survey

The distance-distance plot makes a clear distinction between stars 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 $[n+p+1]/2 \leqslant h \leqslant n$,

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

over all real $\boldsymbol{\mu}$ and symmetric positive definite Σ that satisfy

$$\#\{i; d_i = \sqrt{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})' \hat{\Sigma}^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})} \leqslant c^2\} \geqslant h\}.$$

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

Its breakdown value is optimal (50%) when $h = [(n+p+1)/2]$, 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

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

over all real $\boldsymbol{\mu}$ and symmetric positive definite Σ that satisfy

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

with $d_i = \sqrt{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})' \hat{\Sigma}^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})}$ and ρ a smooth *bounded* ρ-function (Rousseeuw and Leroy, 1987).

Efficiency of S-estimators

- To obtain (Fisher-)consistency at normal distributions, we set δ 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 ρ 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 ρ_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.

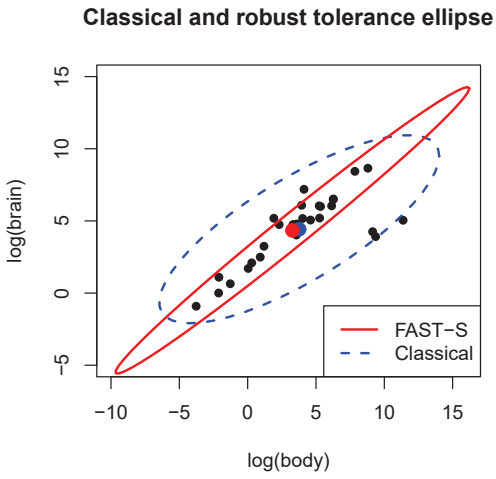
- To compute an S-estimator, the FAST-S algorithm can be used (Salibian-Barrera and Yohai, 2006). It is similar to FAST-MCD.

S-estimators: example

```
> resultS=CovSest(log(Animals))
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

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



MM-estimators of location and scatter

MM-estimators combine **high robustness** with **high efficiency** (Tatsuoka and Tyler, 2000).

They are based on two rho functions ρ_0 and ρ_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 Γ the **shape matrix**.

MM-estimators of location and scatter

MM-estimator of location and scatter

1. Let $(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma})$ be an S-estimator with rho function ρ_0. Denote $\hat{\sigma}^2 = |\tilde{\Sigma}|^{1/p}$.
2. The MM-estimator for location and shape $(\hat{\boldsymbol{\mu}}, \hat{\Gamma})$ minimizes

$$\frac{1}{n}\sum_{i=1}^{n} \rho_1\left(\frac{\sqrt{(\boldsymbol{x}_i - \boldsymbol{\mu})'\Gamma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})}}{\hat{\sigma}}\right) \tag{4}$$

among all real $\boldsymbol{\mu}$ and symmetric positive definite Γ with $|\Gamma| = 1$.

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

MM-estimators of location and scatter

- 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 ρ_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 ρ_1.

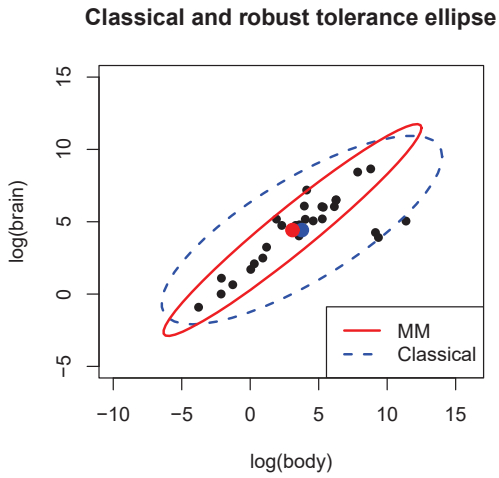
  For bisquare ρ_1, $c = 4.685$ yields 95% efficiency (at the normal model).

- The FAST-MM algorithm starts with FAST-S and then applies IRLS steps to minimize (4).

MM-estimators: example

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

Robust Estimate of Location:
[1] 3.086  4.427

Robust Estimate of Covariance:
       body    brain
body   12.036   9.021
brain   9.021   7.272

> covMM=getCov(resultMM)
> cov2cor(covMM)
           body     brain
body  1.0000000 0.9642449
brain 0.9642449 1.0000000
```

MM-estimators: example



Some non affine equivariant estimators

	affine equivariant	non affine equivariant
Low BV	Classical mean, covariance M-estimators Convex peeling Tukey median Simplicial median Oja median	
High BV	Stahel-Donoho estimator MCD, MVE S-estimators MM-estimators	coordinatewise median spatial median, sign covariance OGK DetMCD

The coordinatewise median

Coordinatewise median:

$$\hat{\boldsymbol{\mu}} = (\underset{i=1}{\overset{n}{\text{med}}}\, x_{i1}\ ,\ \underset{i=1}{\overset{n}{\text{med}}}\, x_{i2}\ ,\ \ldots\ ,\ \underset{i=1}{\overset{n}{\text{med}}}\, x_{ip})'\ .$$

- Easy to compute and to interpret
- 50% breakdown value!
- not affine equivariant, and not even orthogonally equivariant
- $\hat{\boldsymbol{\mu}}$ does not have to lie in the convex hull of the sample when $p \geqslant 3$. 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

Spatial median

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

$$\hat{\boldsymbol{\mu}} = \underset{\boldsymbol{\mu}}{\text{argmin}} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\mu}\|.$$

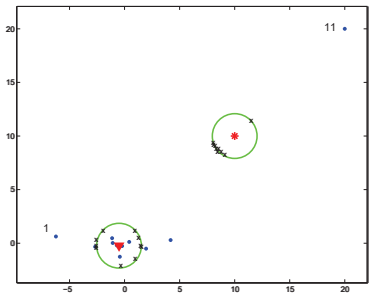
This is equivalent to

$$\sum_{i=1}^{n} \frac{\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}}{\|\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}\|} = 0\ . \tag{5}$$

- 50% breakdown value, bounded influence function
- not affine equivariant, but orthogonal equivariant
- Computation: Equation (5) corresponds to equation (1) of M-estimators, with $W_1(t) = 1/\sqrt{t}$. We can thus use the iterative algorithm with $\Sigma = I$. Other algorithms are discussed in Fritz et al. (2012).

The spatial median

Geometric interpretation: take a point $\boldsymbol{\mu}$ in $\mathbb{R}^p$ and project all observations onto a sphere around $\boldsymbol{\mu}$. If the mean of these projections equals $\boldsymbol{\mu}$, then $\boldsymbol{\mu}$ 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. By definition, the triangle equals the spatial median. Note the moderate influence of observation 11.

The sign covariance matrix

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

Sign covariance estimator

$$\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})}{\|\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}\|} \frac{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})'}{\|\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}\|}$$

with $\hat{\boldsymbol{\mu}}$ the spatial median.

- 50% breakdown value, bounded influence function
- not affine equivariant, but orthogonally equivariant.

The orthogonalized Gnanadesikan-Kettenring estimator

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

It is inspired by the fact that the classical variance σ^2 and the classical covariance σ_{jk} between two variables Y_j and Y_k satisfy:

$$\sigma_{jk} = \frac{1}{4}\left(\sigma(Y_j + Y_k)^2 - \sigma(Y_j - Y_k)^2\right)$$

Gnanadesikan and Kettenring (1972) had proposed to compute a robust covariance measure between 2 variables by replacing the σ's on the right hand side by a robust scale estimator. However, the resulting scatter matrix need not be PSD.

The OGK estimator: definition

OGK = **orthogonalized** Gnanadesikan-Kettenring estimator:

1. Let $m(.)$ and $s(.)$ be robust univariate estimators of location and scale.
2. Construct $\boldsymbol{y}_i = D^{-1}\boldsymbol{x}_i$ 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 $u_{jk} = \frac{1}{4}(s(Y_j + Y_k)^2 - s(Y_j - Y_k)^2)$. This matrix is symmetric but not necessarily PSD.
4. Put the eigenvectors of U as columns in a matrix E and
   1. project the data on these eigenvectors, i.e. $V = YE$;
   2. compute 'robust variances' of $V = (V_1, \ldots, V_p)$, i.e. $\Lambda = \text{diag}(s^2(V_1), \ldots, s^2(V_p))$;
   3. Set the $p \times 1$ vector $\hat{\boldsymbol{\mu}}(Y) = E\boldsymbol{m}$ where $\boldsymbol{m} = (m(V_1), \ldots, m(V_p))'$ and compute the positive definite matrix $\hat{\Sigma}(Y) = E\Lambda E'$.
5. Transform back to X, i.e. $\hat{\boldsymbol{\mu}}(X) = D\hat{\boldsymbol{\mu}}(Y)$ and $\hat{\Sigma} = D\hat{\Sigma}(Y)D'$.

The OGK estimator: properties

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

- Makes the estimates location and scale equivariant.
- Standardized data: Z with rows $\boldsymbol{z}_i'$ and columns Z_j.

The DetMCD algorithm

- Construct six initial estimates $\hat{\boldsymbol{\mu}}_k(Z)$ 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 $L = \text{diag}\left(Q_n(B_1)^2, \ldots, Q_n(B_p)^2\right)$.
  - Estimate the center: $\hat{\boldsymbol{\mu}}_k(Z) = \hat{\Sigma}_k^{1/2}(\text{med}(Z\hat{\Sigma}_k^{-1/2}))$.
- For each initial estimate do:
  - Compute statistical distances $d_{ik} = d(\boldsymbol{z}_i, \hat{\boldsymbol{\mu}}_k(Z), \hat{\Sigma}_k(Z))$.
  - 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. Select the h observations 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 (a sigmoid) of the standardized data:

   $$Y_j = \tanh(Z_j) \qquad \forall j = 1, \ldots, p.$$

   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 $S_3 = \text{corr}(T)$.

DetMCD: Preliminary estimates

4. Related to sign covariance matrix:
   Define $\boldsymbol{k}_i = \frac{\boldsymbol{z}_i}{\|\boldsymbol{z}_i\|}$ and let

$$S_4 = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{k}_i \boldsymbol{k}_i'$$

   (Here the coordinatewise median is used instead of the spatial median to estimate the center.)

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

6. The raw OGK estimator of location and scatter.

DetMCD: Properties

- Faster than FAST-MCD and equally robust in moderate dimensions (say, $p \leqslant 10$)
- Faster than FAST-MCD and more robust in higher dimensions, 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.

When to use DetMCD

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

- When $p \leqslant 10$ 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"` .
- Bagplot: function `bagplot` in R package *aplpack*.

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%.
- FAST-S: the function `fastsloc.m` from Christophe Croux's webpage (`www.econ.kuleuven.be/public/NDBAE06/programs/`)
- MM: the function `MMrse.m` from Christophe Croux's webpage, and the function `multimm.m` (containing `multiS` as auxiliary function) from Stefan Van Aelst's webpage.
- Also the FSDA toolbox, available at `www.riani.it/MATLAB.htm`, contains implementations of S and MM-estimators.
- DetMCD: available in LIBRA. It has OGK as a subroutine.
- Bagplot: available in LIBRA.