Journal of Experimental Social Psychology 74 (2018) 150–156

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance

Christophe Leys[a,*], Olivier Klein[a], Yves Dominicy[b,1], Christophe Ley[c]

[a] Université libre de Bruxelles, Centre de Recherche en Psychologie Sociale et Interculturelle, Belgium
[b] Université libre de Bruxelles, Solvay Brussels School of Economics and Management, ECARES, Belgium
[c] Ghent University, Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium

4. The Minimum Covariance Determinant estimators

The Minimum Covariance Determinant approach was proposed by Rousseeuw (1984, 1985).

Minimum Covariance Determinant (MCD). The MCD estimators of location and scatter, denoted $\widehat{\mu}_{MCD}$ and $\widehat{\Sigma}_{MCD}$

From a practical point of view, the MCD is computationally demanding. However there exists an algorithm called FAST-MCD (Rousseeuw & Van Driessen, 1999), which renders the computation of the MCD faster. The MCD is implemented in R (see Fauconnier & Haesbroeck, 2009, for practical details). Rousseeuw and Van Driessen (1999) proposed the FAST-MCD command on R.

5. The Mahalanobis-MCD distance

In view of what precedes, the robust criterion for multivariate outlier detection we shall propose corresponds to

$$\sqrt{(X_i - \widehat{\mu}_{MCD})^T (\widehat{\Sigma}_{MCD})^{-1} (X_i - \widehat{\mu}_{MCD})} > c_k,$$

where c_k remains to be determined. Note that as the MCD estimator is affine equivariant, the robust Mahalanobis distances are affine invariant. Theoretically, the squared Mahalanobis-MCD distance (in abbreviation MMCD distance) can be approximated by a $\chi_k{}^2$ distribution (Rousseeuw & Van Zomeren, 1990), hence we can use $c_k = \sqrt{\chi^2_{k;1-\alpha}}$, which is the square-root of the upper-α quantile of the chi-square distribution with k degrees of freedom. Natural choices for 1-α are 90%, 95%, 97.5%, 99% and 99.9%, the latter being the most conservative choice. This criterion is a natural extension of the median plus or minus a coefficient times the MAD method (Leys et al., 2013).

Statistical Methodology 6 (2009) 363–379

Contents lists available at ScienceDirect

Statistical Methodology

journal homepage: www.elsevier.com/locate/stamet

Outliers detection with the minimum covariance determinant estimator in practice

C. Fauconnier, G. Haesbroeck *

University of Liège, Belgium

FASTMCD algorithm constructed by Rousseeuw and Van Driessen [13] which has been now implemented in R (function `covMcd(rrcov)`)

Package 'rrcov'

July 23, 2025

Version 1.7-7

VersionNote Released 1.7-6 on 2024-08-19 on CRAN

Title Scalable Robust Estimators with High Breakdown Point

Description Robust Location and Scatter Estimation and Robust
Multivariate Analysis with High Breakdown Point:
principal component analysis (Filzmoser and Todorov (2013), <doi:10.1016/j.ins.2012.10.017>),
linear and quadratic discriminant analysis (Todorov and Pires (2007)),
multivariate tests (Todorov and Filzmoser (2010) <doi:10.1016/j.csda.2009.08.015>),
outlier detection (Todorov et al. (2010) <doi:10.1007/s11634-010-0075-2>).
See also Todorov and Filzmoser (2009) <urn:isbn:978-3838108148>,
Todorov and Filzmoser (2010) <doi:10.18637/jss.v032.i03> and
Boudt et al. (2019) <doi:10.1007/s11222-019-09869-x>.

Maintainer Valentin Todorov <valentin.todorov@chello.at>

Depends R (>= 2.10), robustbase (>= 0.92.1), methods

Imports stats, stats4, mvtnorm, lattice, pcaPP

Suggests grid, MASS

LazyLoad yes

License GPL (>= 3)

URL https://github.com/valentint/rrcov

BugReports https://github.com/valentint/rrcov/issues

Repository CRAN

NeedsCompilation yes

Author Valentin Todorov [aut, cre] (ORCID:
<https://orcid.org/0000-0003-4215-0245>)

Date/Publication 2025-04-21 21:50:02 UTC

RoxygenNote 7.3.2

e

CovMcd *Robust Location and Scatter Estimation via MCD*

Description

Computes a robust multivariate location and scatter estimate with a high breakdown point, using the 'Fast MCD' (Minimum Covariance Determinant) estimator.

Usage

```
CovMcd(x,
       raw.only=FALSE, alpha=control@alpha, nsamp=control@nsamp,
       scalefn=control@scalefn, maxcsteps=control@maxcsteps,
       initHsets=NULL, save.hsets=FALSE,
       seed=control@seed, trace=control@trace,
       use.correction=control@use.correction,
       control=CovControlMcd(), ...)
```

Arguments

x	a matrix or data frame.
raw.only	should only the "raw" estimate be returned.
alpha	numeric parameter controlling the size of the subsets over which the determinant is minimized, i.e., alpha*n observations are used for computing the determinant. Allowed values are between 0.5 and 1 and the default is 0.5.
nsamp	number of subsets used for initial estimates or "best", "exact" or "deterministic". Default is nsamp = 500. For nsamp="best" exhaustive enumeration is done, as long as the number of trials does not exceed 5000. For "exact", exhaustive enumeration will be attempted however many samples are needed. In this case a warning message will be displayed saying that the computation can take a very long time. For "deterministic", the *deterministic* MCD is computed; as proposed by Hubert et al. (2012) it starts from the h most central observations of *six* (deterministic) estimators.
scalefn	function to compute a robust scale estimate or character string specifying a rule determining such a function, see rrcov.control.
maxcsteps	maximal number of concentration steps in the deterministic MCD; should not be reached.
initHsets	NULL or a Kxh integer matrix of initial subsets of observations of size h (specified by the indices in 1:n).
save.hsets	(for deterministic MCD) logical indicating if the initial subsets should be returned as initHsets.
seed	starting value for random generator. Default is seed = NULL
trace	whether to print intermediate results. Default is trace = FALSE
use.correction	whether to use finite sample correction factors. Default is use.correction=TRUE
control	a control object (S4) of class CovControlMcd-class containing estimation options - same as these provided in the function specification. If the control object is supplied, the parameters from it will be used. If parameters are passed also in the invocation statement, they will override the corresponding elements of the control object.
...	potential further arguments passed to **robustbase**'s covMcd.

Details

This function computes the minimum covariance determinant estimator of location and scatter and returns an S4 object of class CovMcd-class containing the estimates. The implementation of the function is similar to the existing R function covMcd() which returns an S3 object. The MCD method looks for the $h(> n/2)$ observations (out of n) whose classical covariance matrix has the lowest possible determinant. The raw MCD estimate of location is then the average of these h points, whereas the raw MCD estimate of scatter is their covariance matrix, multiplied by a consistency factor and a finite sample correction factor (to make it consistent at the normal model and unbiased at small samples). Both rescaling factors are returned also in the vector raw.cnp2 of length 2. Based on these raw MCD estimates, a reweighting step is performed which increases the finite-sample efficiency considerably - see Pison et al. (2002). The rescaling factors for the reweighted estimates are returned in the vector cnp2 of length 2. Details for the computation of the finite sample correction factors can be found in Pison et al. (2002). The finite sample corrections can be suppressed by setting use.correction=FALSE. The implementation in rrcov uses the Fast MCD algorithm of Rousseeuw and Van Driessen (1999) to approximate the minimum covariance determinant estimator.

Value

An S4 object of class CovMcd-class which is a subclass of the virtual class CovRobust-class.

Author(s)

Valentin Todorov <valentin.todorov@chello.at>

References

P. J. Rousseeuw and A. M. Leroy (1987) *Robust Regression and Outlier Detection.* Wiley.

P. J. Rousseeuw and K. van Driessen (1999) A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41**, 212–223.

M. Hubert, P. Rousseeuw and T. Verdonck (2012) A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics* **21**(3), 618–637.

Pison, G., Van Aelst, S., and Willems, G. (2002), Small Sample Corrections for LTS and MCD, *Metrika*, **55**, 111-123.

Todorov V & Filzmoser P (2009), An Object Oriented Framework for Robust Multivariate Analysis. *Journal of Statistical Software*, **32**(3), 1–47. doi:10.18637/jss.v032.i03.

See Also

cov.rob from package **MASS**

Examples

```
data(hbk)
hbk.x <- data.matrix(hbk[, 1:3])
CovMcd(hbk.x)
cD <- CovMcd(hbk.x, nsamp = "deterministic")
summary(cD)
```

```
## the following three statements are equivalent
c1 <- CovMcd(hbk.x, alpha = 0.75)
c2 <- CovMcd(hbk.x, control = CovControlMcd(alpha = 0.75))
## direct specification overrides control one:
c3 <- CovMcd(hbk.x, alpha = 0.75,
             control = CovControlMcd(alpha=0.95))
c1
```