Package ‘mvout’

May 30, 2025

Type Package

Title Robust Multivariate Outlier Detection

Version 1.2

Date 2025-05-28

Description Detection of multivariate outliers using robust estimates of location and scale. The Minimum Covariance Determinant (MCD) estimator is used to calculate robust estimates of the mean vector and covariance matrix. Outliers are determined based on robust Mahalanobis distances using either an unstructured covariance matrix, a principal components structured covariance matrix, or a factor analysis structured covariance matrix. Includes options for specifying the direction of interest for outlier detection for each variable.

License GPL (>= 2)

Encoding UTF-8

Depends R (>= 3.5.0), robustbase

NeedsCompilation no

Author Jesus E. Delgado [aut],
Jed T. Elison [ctb],
Nathaniel E. Helwig [aut, cre]

Maintainer Nathaniel E. Helwig <helwig@umn.edu>

Repository CRAN

Date/Publication 2025-05-30 09:20:05 UTC

Contents

`mvout` *Robust Multivariate Outlier Detection*

Description

Detection of multivariate outliers using robust estimates of location and scale.

Usage

```
mvout(x, method = c("none", "princomp", "factanal"), standardize = TRUE,
      robust = TRUE, direction = rep("two.sided", ncol(x)), thresh = 0.01,
      keepx = TRUE, factors = 2, scores = c("regression", "Bartlett"),
      rotation = c("none", "varimax", "promax"), ...)
```

Arguments

`x` Data matrix (n x p)

`method` Character specifying the factorization method used to define the covariance matrix: "none" uses the unfactorized (robust) covariance matrix, "princomp" uses the (robust) principal components analysis (PCA) implied covariance matrix, and "factanal" uses the (robust) factor analysis (FA) implied covariance matrix.

`standardize` Logical specifying whether to apply PCA to the correlation (default) or covariance matrix. Ignored if `method = "none"` or `method = "factanal"`.

`robust` If `TRUE` (default), robust estimates of the mean vector and covariance matrix are obtained using the `covMcd` function. Otherwise standard estimators are obtained using the `colMeans` and `cov` functions.

`direction` Direction defining "outlier" for each variable (character). Three options are available: "two.sided" considers large postive and negative deviations from the mean as outliers, "less" only considers large negative deviations as outliers, and "greater" only considers large positve deviations as outliers. Accepts a single character giving the common direction for each variable, or a character vector of length p.

`thresh` Scalar specifying the threshold for flagging outliers (0 < thresh < 1). See Note.

`keepx` Logical indicating if input `x` should be saved and returned as part of the output.

`factors` Integer giving the number of factors for PCA or FA model. Ignored if `method = "none"`.

`scores` Method used to compute factor scores (only used if `method = "factanal"`).

`rotation` Factor rotation method aapplied to PCA or FA loadings. Ignored if `method = "none"`.

`...` Additional arguments passed to the `covMcd` function, e.g., `alpha`, `nsamp`, etc. Note that the `cor` argument should not be used, as this is controlled by the `standardize` argument.

Details

Outliers are determined using a (squared) Mahalanobis distance calculated using either the Minimum Covariance Determinant (MCD) estimator for the mean vector and covariance matrix (default) or the standard unbiased sample estimators. The MCD is computed using the `covMcd` function. Includes options for specifying the direction of interest for outlier detection, as well as options for using bilinear models (PCA and FA) to define the covariance matrix used for the Mahalanobis distance.

Value

An object of class `mvout` which is a list with the following components:

`distance` Numeric vector of (squared) Mahalanobis distances for the n observations.

`outlier` Logical vector indicating whether or not each of the n observations is an outlier.

`mcd` Object of class `mcd` that is output from the `covMcd` function.

`args` List of input arguments (e.g., x, method, standardize, etc.)

`scores` Factor or principal component scores (will be `NULL` if `method = "none"`).

`loadings` Factor or principal component loadings (will be `NULL` if `method = "none"`).

`uniquenesses` Variables uniquenesses (will be `NULL` if `method = "none"`).

`invrot` Inverse of the matrix that was used to rotate the loadings (will be `NULL` if `method = "none"`).

`cormat` Factor or principal component score correlation matrix (will be `NULL` if `method = "none"`).

Warning

The default behavior of the `covMcd` function (and, consequently, the `mvout` function) is for the MCD estimator to be computed from a random sample of 500 observations. The `nsamp` argument of the `covMcd` function can be used to control the number of samples or request a different method (e.g., nsamp = "deterministic").

Note

For observations included in the (robust) covariance calculation, the critical value that designates an observation as an outlier is defined as `qchisq(1 - thresh, df = p)`.

For the excluded observations, the critical value is defined as `qf(1 - thresh, df1 = p, df2 = n - p) * ((n - 1) * p / (n - p))`.

Author(s)

Jesus E. Delgado <delga220@umn.edu> Nathaniel E. Helwig <helwig@umn.edu>

References

Todorov, V., & Filzmoser, F. (2009). An Object-Oriented Framework for Robust Multivariate Analysis. Journal of Statistical Software, 32(3), 1-47.

See Also

predict.mvout for obtaining predictions from mvout objects.

Examples

```
# generate some data
n <- 200
p <- 2
set.seed(0)
x <- matrix(rnorm(n * p), n, p)

# thresh = 0.01
set.seed(1)    # for reproducible MCD estimate
out1 <- mvout(x)
plot(out1)

# thresh = 0.05
set.seed(1)    # for reproducible MCD estimate
out5 <- mvout(x, thresh = 0.05)
plot(out5)

# direction = "greater"
set.seed(1)    # for reproducible MCD estimate
out <- mvout(x, direction = "greater", thresh = 0.05)
plot(out)

# direction = c("greater", "less")
set.seed(1)    # for reproducible MCD estimate
out <- mvout(x, direction = c("greater", "less"), thresh = 0.05)
plot(out)
```

pheno *Phenotypic Risk Profiles of Toddlers*

Description

The pheno data frame has 1570 rows and 12 columns. Each row contains data from one toddler, which consists of multiple parent-report measures aimed at early identification of phenotypes at-risk for neurodevelopmental disorders such as autism. The data were originally analyzed by Doyle et al. (2021).

Usage

```
data("pheno")
```

Format

A data frame with observations on the following 12 variables.

`age` a numeric vector representing the child's age in months. Ranges from 17 to 26 months.

`sex` a factor vector representing the child's assigned sex at birth.

`RM` an integer vector representing the child's repetitive movement score. Ranges from 0 to 36.

`SDSI` an integer vector representing the child's self-directed/self-injurious score. Ranges from 0 to 22.

`RR` an integer vector representing the child's ritual and routine score. Ranges from 0 to 28.

`RI` an integer vector representing the child's restricted interests score. Ranges 0 to 35.

`WP` an integer vector representing the child's number of words-produced score. Ranges from 0 to 396.

`TG` an integer vector representing the child's total number of gestures used. Ranges from 0 to 63.

`RSB` an integer vector representing the child's reciprocal social behavior score. Ranges from 3 to 65.

`Ext` an integer vector representing the child's externalizing behavior score. Ranges from 0 to 24. Only availabe for subsample of 107 toddlers.

`Int` an integer vector representing the child's internalizing behavior score. Ranges from 4 to 33. Only availabe for subsample of 107 toddlers.

`Dys` an integer vector representing the child's dysregulation behavior score. Ranges from 3 to 43. Only availabe for subsample of 107 toddlers.

Details

These data consist of parents' responses to questionnaires concerning their toddler's behavior. The `RM`, `SDSI`, `RR`, and `RI` scores were obtained using the Repetitive Behavior Scale for Early Childhood (RBS-EC; Wolff et al., 2016). The `RSB` score was obtained using the Video-Referenced Rating of Reciprocal Social Behavior scale (vrRSB; Marrus et al., 2015). `WP` and `TG` scores were obtained via the MacArthur-Bates Communicative Developmental Inventories (MCDI: Fenson et al., 2007). Scores for `Ext`, `Int`, and `Dys` were obtained via the Infant Toddler Social Emotional Assessment (ITSEA; Carter et al., 2003). For a detailed description of these data see Doyle et al. (2021).

Source

Doyle, C.M., Lasch, C., Vollman, E.P., Desjardins, C.D., Helwig, N.E., Jacob, S., Wolff, J.J. and Elison, J.T. (2021), Phenoscreening: a developmental approach to research domain criteria-motivated sampling. Journal of Child Psychology and Psychiatry, 62: 884-894. doi:10.1111/jcpp.13341

References

Carter, A. S., Briggs-Gowan, M. J., Jones, S. M., & Little, T. D. (2003). The Infant-Toddler Social and Emotional Assessment (ITSEA): Factor structure, reliability, and validity. Journal of Abnormal Child Psychology, 31(5), 495–514. doi:10.1023/A:1025449031360

Fenson, L., Bates, E., Dale, P.S., Marchman, V.A., Reznick,J.S., & Thal, D. (2007). MacArthur-Bates communicative development inventories user's guide and technical manual (2nd ed.). Baltimore, MD: Paul H. Brookes Publishing Company.

Marrus, N., Glowinski, A. L., Jacob, T., Klin, A., Jones, W., Drain, C. E., Holzhauer, K. E., Hariprasad, V., Fitzgerald, R. T., Mortenson, E. L., Sant, S. M., Cole, L., Siegel, S. A., Zhang, Y., Agrawal, A., Heath, A. C., & Constantino, J. N. (2015). Rapid video-referenced ratings of reciprocal social behavior in toddlers: a twin study. Journal of child psychology and psychiatry, and allied disciplines, 56(12), 1338–1346. doi:10.1111/jcpp.12391

Wolff, J. J., Boyd, B. A., & Elison, J. T. (2016). A quantitative measure of restricted and repetitive behaviors for early childhood. Journal of Neurodevelopmental Disorders, 8(1), 27–27. doi:10.1186/s116890169161x

Examples

```
# load data
library(mvout)
data(pheno)
X <- pheno[,3:9]

# method = "princomp" and direction = "two-sided"
set.seed(1)    # for reproducible MCD estimate
pc2sided <- mvout(X, method = "princomp", rotation = "promax")
plot(pc2sided)

# method = "princomp" and directional
dir <- c(rep("greater", 4), rep("less", 2), "greater")
set.seed(1)    # for reproducible MCD estimate
pc1sided <- mvout(X, method = "princomp", rotation = "promax", direction = dir)
plot(pc1sided)

# method = "factanal" and direction = "two.sided"
set.seed(1)    # for reproducible MCD estimate
fa2sided <- mvout(X, method = "factanal", rotation = "promax")
plot(fa2sided)

# method = "factanal" and directional
dir <- c(rep("greater", 4), rep("less", 2), "greater")
set.seed(1)    # for reproducible MCD estimate
fa1sided <- mvout(X, method = "factanal", rotation = "promax", direction = dir)
plot(fa1sided)
```

plot.mvout *Plot Method for mvout Objects*

Description

Default S3 plot method for objects of class "mvout".

Usage

```
## S3 method for class 'mvout'
plot(x, outcol = "red", incol = "black", outpch = 0, inpch = 1,
     xlab = "PC1", ylab = "PC2", xresign = FALSE, yresign = FALSE, ...)
```

Arguments

`x`	Object of class `mvout`.
`outcol`	Color used for cases labeled as outliers.
`incol`	Color used for cases not labeled as outliers.
`outpch`	Plotting character used for cases labeled as outliers.
`inpch`	Plotting character used for cases not labeled as outliers.
`xlab`	Label for the x-axis.
`ylab`	Label for the y-axis.
`xresign`	Logical argument. If `TRUE`, values on x-axis have their signs reversed. This is primarlity for improved visualization.
`yresign`	Logical argument. If `TRUE`, values on y-axis have their signs reversed. This is primarlity for improved visualization.
`...`	Additional arguments passed to the `plot` function, e.g., `cex`, `main`, etc.

Details

Produces a bivariate plot highlighting cases that have been flagged as outliers. If `method = "princomp"` or `method = "factanal"` was used, then the `scores` component of `x` is plotted. Otherwise the data are projected onto the first two principal components for visualization.

Value

A plot is produced and nothing is returned.

Author(s)

Jesus E. Delgado <delga220@umn.edu> Nathaniel E. Helwig <helwig@umn.edu>

See Also

`mvout`

Examples

```
# load package and data
library(mvout)
data(pheno)
X <- pheno[,3:9]

# Example using pheno dataset
dir <- c(rep("greater", 4), rep("less", 2), "greater")
set.seed(1)    # for reproducible MCD estimate
out <- mvout(X, method = "princomp", rotation = "promax", direction = dir)
plot(out, outpch = 4)
```

`predict.mvout` *Predict method for Robust Multivariate Outlier Detection*

Description

`predict` method for class "mvout".

Usage

```
## S3 method for class 'mvout'
predict(object,
        x,
        type = c(”distance”, ”outlier”, ”scores”),
        thresh = 0.01,
        ...)
```

Arguments

`object` Object of class `mvout`

`x` Optional matrix of new data used for the predictions. If omitted, the original data are used if `keepx = TRUE` (and error is produced otherwise).

`type` Type of prediction to return: "distance" returns the predicted Mahalanobis distance, "outlier" returns the predicted outlier status (T/F) using the specified `thresh`, and "scores" returns the predicted principal component or factor scores (if applicable).

`thresh` Scalar specifying the threshold for flagging outliers (0 < thresh < 1). See `mvout` for details.

`...` Additional arguments (ignored)

Details

Produces predictions from the input new data `x` using the robust parameter estimates (of location and scatter) from the input "mvout" `object`.

Value

Returns a vector of numerics ("distance" or "scores") or logicals ("outlier").

Note

If you input the same `x` that was used to estimate the location and scale parameters you will obtain:

(1) the same "distance" and "scores" as output by the `mvout` function

(2) a potentially different "outlier" result than what is output by the `mvout` function

The discrepancy in (2) is because all of the observations are considered to have been excluded from the location/scatter estimation when the `x` argument is provided. This results in a different critical value being used for the observations that were included in the MCD estimate. For boarderline cases, this slight change in the critical value could result in a change of outlier status.

Author(s)

Jesus E. Delgado <delga220@umn.edu> Nathaniel E. Helwig <helwig@umn.edu>

See Also

mvout for estimation of (robust) location/scatter.

Examples

```
# generate some data
n <- 200
p <- 2
set.seed(0)
x <- matrix(rnorm(n * p), n, p)

# thresh = 0.01
set.seed(1)    # for reproducible MCD estimate
out1 <- mvout(x)

# predicted distance (same as before)
fit1 <- predict(out1, x = x)
max(abs(fit1 - out1$distance))

# predicted outlier (differs from before)
fit1 <- predict(out1, x = x, type = "outlier")
mean(abs(fit1 == out1$outlier))
```

---

print.mvout *Print method for Robust Multivariate Outlier Detection*

Description

print method for class "mvout".

Usage

```
## S3 method for class 'mvout'
print(x, ...)
```

Arguments

x	Object of class mvout
...	Additional arguments (ignored).

Details

Prints the percentage of observations flagged as outliers, five quantiles of the robust Mahalanobis distances, and basic information about the options used for the outlier detection.

Value

Nothing returned (just prints to console).

Author(s)

Jesus E. Delgado <delga220@umn.edu> Nathaniel E. Helwig <helwig@umn.edu>

See Also

mvout for estimation of (robust) location/scatter.

predict.mvout for obtaining predictions from mvout objects.

Examples

```
# generate some data
n <- 200
p <- 2
set.seed(0)
x <- matrix(rnorm(n * p), n, p)

# detect outliers
set.seed(1)    # for reproducible MCD estimate
out <- mvout(x)

# print results
out
```

Index