# Package 'rrcovHD'

July 23, 2025

**Title** Robust Multivariate Methods for High Dimensional Data

**Version** 0.3-1

**VersionNote** Released 0.3-0 on 2024-02-04 on CRAN

**Description** Robust multivariate methods for high dimensional data including
outlier detection (Filzmoser and Todorov (2013) <doi:10.1016/j.ins.2012.10.017>),
robust sparse PCA (Croux et al. (2013) <doi:10.1080/00401706.2012.727746>, Todorov and Filzmoser (2013) <doi:10.1007/978-3-642-33042-1_31>),
robust PLS (Todorov and Filzmoser (2014) <doi:10.17713/ajs.v43i4.44>),
and robust sparse classification (Ortner et al. (2020) <doi:10.1007/s10618-019-00666-8>).

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

**Depends** rrcov (>= 1.3-7), robustbase (>= 0.92-1), methods

**Imports** stats4, pls, spls, pcaPP, robustHD, Rcpp

**LinkingTo** Rcpp

**Suggests** MASS

**LazyLoad** yes

**LazyData** yes

**License** GPL (>= 3)

**URL** https://github.com/valentint/rrcovHD

**BugReports** https://github.com/valentint/rrcovHD/issues

**NeedsCompilation** yes

**Repository** CRAN

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

**Date/Publication** 2024-08-17 21:40:02 UTC

## Contents

---

`CSimca` *Classification in high dimensions based on the (classical) SIMCA method*

---

## Description

CSimca performs the (classical) SIMCA method. This method classifies a data matrix x with a known group structure. To reduce the dimension on each group a PCA analysis is performed. Afterwards a classification rule is developped to determine the assignment of new observations.

## Usage

```
CSimca(x, ...)
## Default S3 method:
CSimca(x, grouping, prior=proportions, k, kmax = ncol(x),
    tol = 1.0e-4, trace=FALSE, ...)
## S3 method for class 'formula'
CSimca(formula, data = NULL, ..., subset, na.action)
```

## Arguments

| | |
|---|---|
| formula | a formula of the form y~x, it describes the response and the predictors. The formula can be more complicated, such as y~log(x)+z etc (see formula for more details). The response should be a factor representing the response variable, or any vector that can be coerced to such (such as a logical variable). |
| data | an optional data frame (or similar: see model.frame) containing the variables in the formula formula. |
| subset | an optional vector used to select rows (observations) of the data matrix x. |
| na.action | a function which indicates what should happen when the data contain NAs. The default is set by the na.action setting of options, and is na.fail if that is unset. The default is na.omit. |
| x | a matrix or data frame containing the explanatory variables (training set). |
| grouping | grouping variable: a factor specifying the class for each observation. |
| prior | prior probabilities, default to the class proportions for the training set. |
| tol | tolerance |
| k | number of principal components to compute. If k is missing, or k = 0, the algorithm itself will determine the number of components by finding such k that $l_k/l_1 >= 10.E-3$ and $\Sigma_{j=1}^{k} l_j/\Sigma_{j=1}^{r} l_j >= 0.8$. It is preferable to investigate the scree plot in order to choose the number of components and then run again. Default is k=0. |
| kmax | maximal number of principal components to compute. Default is kmax=10. If k is provided, kmax does not need to be specified, unless k is larger than 10. |
| trace | whether to print intermediate results. Default is trace = FALSE |
| ... | arguments passed to or from other methods. |

## Details

CSimca, serving as a constructor for objects of class CSimca-class is a generic function with "formula" and "default" methods.

SIMCA is a two phase procedure consisting of PCA performed on each group separately for dimension reduction followed by classification rules built in the lower dimensional space (note that the dimension in each group can be different). In original SIMCA new observations are classified by means of their deviations from the different PCA models. Here (and also in the robust versions implemented in this package) the classification rules will be obtained using two popular distances arising from PCA - orthogonal distances (OD) and score distances (SD). For the definition of these distances, the definition of the cutoff values and the standartization of the distances see Vanden Branden K, Hubert M (2005) and Todorov and Filzmoser (2009).

## Value

An S4 object of class CSimca-class which is a subclass of of the virtual class Simca-class.

## Author(s)

Valentin Todorov <valentin.todorov@chello.at>

## References

Vanden Branden K, Hubert M (2005) Robust classification in high dimensions based on the SIMCA method. Chemometrics and Intellegent Laboratory Systems 79:10–21

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.

Todorov V & Filzmoser P (2014), Software Tools for Robust Analysis of High-Dimensional Data. *Austrian Journal of Statistics*, **43**(4), 255–266, doi:10.17713/ajs.v43i4.44.

## Examples

```
data(pottery)
dim(pottery)        # 27 observations in 2 classes, 6 variables
head(pottery)

## Build the SIMCA model. Use RSimca for a robust version
cs <- CSimca(origin~., data=pottery)
cs
summary(cs)


## generate a sample from the pottery data set -
##  this will be the "new" data to be predicted
smpl <- sample(1:nrow(pottery), 5)
test <- pottery[smpl, -7]       # extract the test sample. Remove the last (grouping) variable
print(test)


## predict new data
pr <- predict(cs, newdata=test)

pr@classification
```

---

`CSimca-class` *Class "CSimca" - classification in high dimensions based on the (classical) SIMCA method*

---

## Description

The class `CSimca` represents the SIMCA algorithm for classification in high dimensions. The objects of class `CSImca` contain the results of the SIMCA method.

## Objects from the Class

Objects can be created by calls of the form `new("CSImca", ...)` but the usual way of creating `CSimca` objects is a call to the function `CSimca()` which serves as a constructor.

### Slots

`call`: the (matched) function call.

`prior`: prior probabilities used, default to group proportions

`counts`: number of observations in each class

`pcaobj`: A list of Pca objects - one for each group

`k`: Object of class `"numeric"` number of (choosen) principal components

`flag`: Object of class `"Uvector"` The observations whose score distance is larger than cutoff.sd or whose orthogonal distance is larger than cutoff.od can be considered as outliers and receive a flag equal to zero. The regular observations receive a flag 1

`X`: the training data set (same as the input parameter x of the constructor function)

`grp`: grouping variable: a factor specifying the class for each observation.

### Extends

Class `"Simca"`, directly.

### Methods

No methods defined with class "CSimca" in the signature.

### Author(s)

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

### References

Vanden Branden K, Hubert M (2005) Robust classification in high dimensions based on the SIMCA method. Chemometrics and Intellegent Laboratory Systems 79:10–21

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.

Todorov V & Filzmoser P (2014), Software Tools for Robust Analysis of High-Dimensional Data. *Austrian Journal of Statistics*, **43**(4), 255–266, doi:10.17713/ajs.v43i4.44.

### Examples

```
showClass("CSimca")
```

---

## `getWeight-methods` *Accessor methods to the essential slots of `Outlier` and its subclasses*

---

### Description

Accessor methods to the essential slots of `Outlier` and its subclasses

### Methods

**obj = "Outlier"** generic functions - see `getWeight`, `getOutliers`, `getClassLabels`, `getCutoff`

kibler *1985 Auto Imports Database*

## Description

The original data set kibler.orig consists of three types of entities: (a) the specification of an auto in terms of various characteristics, (b) its assigned insurance risk rating and (c) its normalized losses in use as compared to other cars.

The second rating corresponds to the degree to which the auto is more risky than its price indicates. Cars are initially assigned a risk factor symbol associated with its price. Then, if it is more risky (or less), this symbol is adjusted by moving it up (or down) the scale. Actuarians call this process "symboling". A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.

The third factor is the relative average loss payment per insured vehicle year. This value is normalized for all autos within a particular size classification (two-door small, station wagons, sports/speciality, etc...), and represents the average loss per car per year.

## Usage

```
data(kibler)
```

## Format

A data frame with 195 observations on the following 14 variables. The original data set (also available as kibler.orig) contains 205 cases and 26 variables of which 15 continuous, 1 integer and 10 nominal. The non-numeric variables and variables with many missing values were removed. Cases with missing values were removed too.

symboling a numeric vector

wheel-base a numeric vector

length a numeric vector

width a numeric vector

height a numeric vector

curb-weight a numeric vector

bore a numeric vector

stroke a numeric vector

compression-ratio a numeric vector

horsepower a numeric vector

peak-rpm a numeric vector

city-mpg a numeric vector

highway-mpg a numeric vector

price a numeric vector

## Details

The original data set contains 205 cases and 26 variables of which 15 continuous, 1 integer and 10 nominal. The non-numeric variables and variables with many missing values were removed. Cases with missing values were removed too. Thus the data set remains with 195 cases and 14 variables.

## Source

www.cs.umb.edu/~rickb/files/UCI/

## References

Kibler, D., Aha, D.W. and Albert, M. (1989). Instance-based prediction of real-valued attributes. *Computational Intelligence*, Vo.l 5, 51-57.

## Examples

```
data(kibler)
x.sd <- apply(kibler,2,sd)
xsd <- sweep(kibler, 2, x.sd, "/", check.margin = FALSE)
apply(xsd, 2, sd)

x.mad <- apply(kibler, 2, mad)
xmad <- sweep(kibler, 2, x.mad, "/", check.margin = FALSE)
apply(xmad, 2, mad)

x.qn <- apply(kibler, 2, Qn)
xqn <- sweep(kibler, 2, x.qn, "/", check.margin = FALSE)
apply(xqn, 2, Qn)


## Display the scree plot of the classical and robust PCA
screeplot(PcaClassic(xsd))
screeplot(PcaGrid(xqn))

########################################
##
## DD-plots
##
## Not run:
usr <- par(mfrow=c(2,2))
plot(SPcaGrid(xsd, lambda=0, method="sd", k=4), main="Standard PCA")    # standard
plot(SPcaGrid(xqn, lambda=0, method="Qn", k=4))                      # robust, non-sparse

plot(SPcaGrid(xqn, lambda=1,43, method="sd", k=4), main="Stdandard sparse PCA")  # sparse
plot(SPcaGrid(xqn, lambda=2.36, method="Qn", k=4), main="Robust sparse PCA")    # robust sparse
par(usr)

########################################
##  Table 2 in Croux et al
##  - to compute EV=Explained variance and Cumulative EV we
##      need to get all 14 eigenvalues
##
```

```
rpca <- SPcaGrid(xqn, lambda=0, k=14)
srpca <- SPcaGrid(xqn, lambda=2.36, k=14)
tab <- cbind(round(getLoadings(rpca)[,1:4], 2), round(getLoadings(srpca)[,1:4], 2))

vars1 <- getEigenvalues(rpca);  vars1 <- vars1/sum(vars1)
vars2 <- getEigenvalues(srpca); vars2 <- vars2/sum(vars2)
cvars1 <- cumsum(vars1)
cvars2 <- cumsum(vars2)
ev <- round(c(vars1[1:4], vars2[1:4]),2)
cev <- round(c(cvars1[1:4], cvars2[1:4]),2)
rbind(tab, ev, cev)

## End(Not run)
```

---

`Outlier-class` *Class* "`Outlier`" *– a base class for outlier identification*

---

## Description

The class `Outlier` represents the results of outlier identification.

## Objects from the Class

A virtual Class: No objects may be created from it.

## Slots

`call`: Object of class "`language`"

`counts`: Number of observations in each class

`grp`: Grouping variable

`wt`: Vector of weights

`flag`: 0/1 flags identifying the outliers

`method`: A character string specifying the method used to identify the outliers. In case of `OutlierMahdist` class this is the name of the robust estimator of multivariate location and covariance matrix used

`singularity`: a list with singularity information for the covariance matrix (or `NULL` if not singular)

## Methods

**getClassLabels** Returns a vector with indices for a given class

**getDistance** Returns a vector containing the computed distances

**getFlag** Returns the flags identifying the outliers

**getOutliers** Returns a vector with the indices of the identified outliers

**getWeight** Returns a vector of weights

**plot**

**show**

## Author(s)

Valentin Todorov <valentin.todorov@chello.at>

## References

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.

Filzmoser P & Todorov V (2013). Robust tools for the imperfect world, *Information Sciences* **245**, 4–20. doi:10.1016/j.ins.2012.10.017.

## Examples

```
showClass("Outlier")
```

---

`OutlierMahdist` *Outlier identification using robust (mahalanobis) distances based on robust multivariate location and covariance matrix*

---

## Description

This function uses the Mahalanobis distance as a basis for multivariate outlier detection. The standard method for multivariate outlier detection is robust estimation of the parameters in the Mahalanobis distance and the comparison with a critical value of the Chi2 distribution (Rousseeuw and Van Zomeren, 1990).

## Usage

```
    OutlierMahdist(x, ...)
    ## Default S3 method:
OutlierMahdist(x, grouping, control, trace=FALSE, ...)
    ## S3 method for class 'formula'
OutlierMahdist(formula, data, ..., subset, na.action)
```

## Arguments

| | |
|---|---|
| `formula` | a formula with no response variable, referring only to numeric variables. |
| `data` | an optional data frame (or similar: see `model.frame`) containing the variables in the formula `formula`. |
| `subset` | an optional vector used to select rows (observations) of the data matrix `x`. |
| `na.action` | a function which indicates what should happen when the data contain `NA`s. The default is set by the `na.action` setting of `options`, and is `na.fail` if that is unset. The default is `na.omit`. |
| `...` | arguments passed to or from other methods. |
| `x` | a matrix or data frame. |
| `grouping` | grouping variable: a factor specifying the class for each observation. |

| | |
|---|---|
| control | a control object (S4) for one of the available control classes, e.g. CovControlMcd-class, CovControlOgk-class, CovControlSest-class, etc., containing estimation options. The class of this object defines which estimator will be used. Alternatively a character string can be specified which names the estimator - one of auto, sde, mcd, ogk, m, mve, sfast, surreal, bisquare, rocke. If 'auto' is specified or the argument is missing, the function will select the estimator (see below for details) |
| trace | whether to print intermediate results. Default is trace = FALSE |

## Details

If the data set consists of two or more classes (specified by the grouping variable grouping) the proposed method iterates through the classes present in the data, separates each class from the rest and identifies the outliers relative to this class, thus treating both types of outliers, the mislabeled and the abnormal samples in a homogenous way.

The estimation method is selected by the control object control. If a character string naming an estimator is specified, a new control object will be created and used (with default estimation options). If this argument is missing or a character string 'auto' is specified, the function will select the robust estimator according to the size of the dataset - for details see CovRobust.

## Value

An S4 object of class OutlierMahdist which is a subclass of the virtual class Outlier.

## Author(s)

Valentin Todorov <valentin.todorov@chello.at>

## References

P. J. Rousseeuw and B. C. Van Zomeren (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*. Vol. 85(411), pp. 633-651.

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.

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.

Filzmoser P & Todorov V (2013). Robust tools for the imperfect world, *Information Sciences* **245**, 4–20. doi:10.1016/j.ins.2012.10.017.

## Examples

```
data(hemophilia)
obj <- OutlierMahdist(gr~.,data=hemophilia)
obj

getDistance(obj)            # returns an array of distances
getClassLabels(obj, 1)      # returns an array of indices for a given class
```

```
getCutoff(obj)          # returns an array of cutoff values (for each class, usually equal)
getFlag(obj)                #  returns an 0/1 array of flags
plot(obj, class=2)          # standard plot function
```

---

`OutlierMahdist-class` *Class* `OutlierMahdist` - *Outlier identification using robust (mahalanobis) distances based on robust multivariate location and covariance matrix*

---

## Description

Holds the results of outlier identification using robust mahalanobis distances computed by robust multivarite location and covarince matrix.

## Objects from the Class

Objects can be created by calls of the form `new("OutlierMahdist", ...)` but the usual way of creating `OutlierMahdist` objects is a call to the function `OutlierMahdist()` which serves as a constructor.

## Slots

`covobj`: A list containing the robust estimates of multivariate location and covariance matrix for each class

`call`: Object of class `"language"`

`counts`: Number of observations in each class

`grp`: Grouping variable

`wt`: Weights

`flag`: 0/1 flags identifying the outliers

`method`: Method used to compute the robust estimates of multivariate location and covariance matrix

`singularity`: a list with singularity information for the covariance matrix (or `NULL` of not singular)

## Extends

Class `"Outlier"`, directly.

## Methods

**getCutoff** Return the cutoff value used to identify outliers

**getDistance** Return a vector containing the computed distances

## Author(s)

Valentin Todorov <valentin.todorov@chello.at>

## References

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.

Filzmoser P & Todorov V (2013). Robust tools for the imperfect world, *Information Sciences* **245**, 4–20. doi:10.1016/j.ins.2012.10.017.

## See Also

OutlierMahdist, Outlier-class

## Examples

```
showClass("OutlierMahdist")
```

---

`OutlierPCDist` *Outlier identification in high dimensions using the PCDIST algorithm*

---

## Description

The function implements a simple, automatic outlier detection method suitable for high dimensional data that treats each class independently and uses a statistically principled threshold for outliers. The algorithm can detect both mislabeled and abnormal samples without reference to other classes.

## Usage

```
    OutlierPCDist(x, ...)
    ## Default S3 method:
OutlierPCDist(x, grouping, control, k, explvar, trace=FALSE, ...)
    ## S3 method for class 'formula'
OutlierPCDist(formula, data, ..., subset, na.action)
```

## Arguments

| | |
|---|---|
| `formula` | a formula with no response variable, referring only to numeric variables. |
| `data` | an optional data frame (or similar: see `model.frame`) containing the variables in the formula `formula`. |
| `subset` | an optional vector used to select rows (observations) of the data matrix `x`. |
| `na.action` | a function which indicates what should happen when the data contain `NA`s. The default is set by the `na.action` setting of `options`, and is `na.fail` if that is unset. The default is `na.omit`. |
| `...` | arguments passed to or from other methods. |
| `x` | a matrix or data frame. |
| `grouping` | grouping variable: a factor specifying the class for each observation. |

| | |
|---|---|
| control | a control object (S4) for one of the available control classes, e.g. CovControlMcd-class, CovControlOgk-class, CovControlSest-class, etc., containing estimation options. The class of this object defines which estimator will be used. Alternatively a character string can be specified which names the estimator - one of auto, sde, mcd, ogk, m, mve, sfast, surreal, bisquare, rocke. If 'auto' is specified or the argument is missing, the function will select the estimator (see below for details) |
| k | Number of components to select for PCA. If missing, the number of components will be calculated automatically |
| explvar | Minimal explained variance to be used for calculation of the number of components in PCA. If explvar is not provided, automatic dimensionality selection using profile likelihood, as proposed by Zhu and Ghodsi will be used. |
| trace | whether to print intermediate results. Default is trace = FALSE |

## Details

If the data set consists of two or more classes (specified by the grouping variable grouping) the proposed method iterates through the classes present in the data, separates each class from the rest and identifies the outliers relative to this class, thus treating both types of outliers, the mislabeled and the abnormal samples in a homogenous way.

The first step of the algorithm is dimensionality reduction using (classical) PCA. The number of components to select can be provided by the user but if missing, the number of components will be calculated either using the provided minimal explained variance or by the automatic dimensionality selection using profile likelihood, as proposed by Zhu and Ghodsi.

## Value

An S4 object of class OutlierPCDist which is a subclass of the virtual class Outlier.

## Author(s)

Valentin Todorov <valentin.todorov@chello.at>

## References

A.D. Shieh and Y.S. Hung (2009). Detecting Outlier Samples in Microarray Data, *Statistical Applications in Genetics and Molecular Biology* **8**.

M. Zhu, and A. Ghodsi (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, **51**, pp. 918–930.

Filzmoser P & Todorov V (2013). Robust tools for the imperfect world, *Information Sciences* **245**, 4–20. doi:10.1016/j.ins.2012.10.017.

## See Also

OutlierPCDist, Outlier

## Examples

```
data(hemophilia)
obj <- OutlierPCDist(gr~.,data=hemophilia)
obj

getDistance(obj)            # returns an array of distances
getClassLabels(obj, 1)      # returns an array of indices for a given class
getCutoff(obj)          # returns an array of cutoff values (for each class, usually equal)
getFlag(obj)                #  returns an 0/1 array of flags
plot(obj, class=2)          # standard plot function
```

---

`OutlierPCDist-class` *Class* "`OutlierPCDist`" - *Outlier identification in high dimensions using using the PCDIST algorithm*

---

## Description

The function implements a simple, automatic outlier detection method suitable for high dimensional data that treats each class independently and uses a statistically principled threshold for outliers. The algorithm can detect both mislabeled and abnormal samples without reference to other classes.

## Objects from the Class

Objects can be created by calls of the form `new("OutlierPCDist", ...)` but the usual way of creating `OutlierPCDist` objects is a call to the function `OutlierPCDist()` which serves as a constructor.

## Slots

`covobj`: A list containing intermediate results of the PCDIST algorithm for each class

`k`: Number of selected PC

`call`, `counts`, `grp`, `wt`, `flag`, `method`, `singularity`: from the "`Outlier`" class.

## Extends

Class "`Outlier`", directly.

## Methods

**getCutoff** Return the cutoff value used to identify outliers

## Author(s)

Valentin Todorov <valentin.todorov@chello.at>

## References

A.D. Shieh and Y.S. Hung (2009). Detecting Outlier Samples in Microarray Data, *Statistical Applications in Genetics and Molecular Biology* Vol. 8.

Filzmoser P & Todorov V (2013). Robust tools for the imperfect world, *Information Sciences* **245**, 4–20. doi:10.1016/j.ins.2012.10.017.

## See Also

OutlierPCDist, Outlier

## Examples

```
showClass("OutlierPCDist")
```

---

# OutlierPCOut *Outlier identification in high dimensions using the PCOUT algorithm*

---

## Description

The function implements a computationally fast procedure for identifying outliers that is particularly effective in high dimensions. This algorithm utilizes simple properties of principal components to identify outliers in the transformed space, leading to significant computational advantages for high-dimensional data. This approach requires considerably less computational time than existing methods for outlier detection, and is suitable for use on very large data sets. It is also capable of analyzing the data situation commonly found in certain biological applications in which the number of dimensions is several orders of magnitude larger than the number of observations.

## Usage

```
    OutlierPCOut(x, ...)
    ## Default S3 method:
OutlierPCOut(x, grouping, explvar=0.99, trace=FALSE, ...)
    ## S3 method for class 'formula'
OutlierPCOut(formula, data, ..., subset, na.action)
```

## Arguments

| | |
|---|---|
| formula | a formula with no response variable, referring only to numeric variables. |
| data | an optional data frame (or similar: see model.frame) containing the variables in the formula formula. |
| subset | an optional vector used to select rows (observations) of the data matrix x. |
| na.action | a function which indicates what should happen when the data contain NAs. The default is set by the na.action setting of options, and is na.fail if that is unset. The default is na.omit. |
| ... | arguments passed to or from other methods. |

| | |
|---|---|
| x | a matrix or data frame. |
| grouping | grouping variable: a factor specifying the class for each observation. |
| explvar | a numeric value between 0 and 1 indicating how much variance should be covered by the robust PCs (default to 0.99) |
| trace | whether to print intermediate results. Default is trace = FALSE |

## Details

If the data set consists of two or more classes (specified by the grouping variable grouping) the proposed method iterates through the classes present in the data, separates each class from the rest and identifies the outliers relative to this class, thus treating both types of outliers, the mislabeled and the abnormal samples in a homogenous way.

## Value

An S4 object of class OutlierPCOut which is a subclass of the virtual class Outlier.

## Author(s)

Valentin Todorov <valentin.todorov@chello.at>

## References

P. Filzmoser, R. Maronna and M. Werner (2008). Outlier identification in high dimensions, *Computational Statistics & Data Analysis*, Vol. 52 1694–1711.

Filzmoser P & Todorov V (2013). Robust tools for the imperfect world, *Information Sciences* **245**, 4–20. doi:10.1016/j.ins.2012.10.017.

## See Also

OutlierPCOut, Outlier

## Examples

```
data(hemophilia)
obj <- OutlierPCOut(gr~.,data=hemophilia)
obj

getDistance(obj)            # returns an array of distances
getClassLabels(obj, 1)      # returns an array of indices for a given class
getCutoff(obj)          # returns an array of cutoff values (for each class, usually equal)
getFlag(obj)                #  returns an 0/1 array of flags
plot(obj, class=2)          # standard plot function
```

`OutlierPCOut-class` *Class* "OutlierPCOut" - *Outlier identification in high dimensions using using the PCOUT algorithm*

## Description

Holds the results of outlier identification using the PCOUT algorithm.

## Objects from the Class

Objects can be created by calls of the form `new("OutlierPCOut", ...)` but the usual way of creating `OutlierPCOut` objects is a call to the function `OutlierPCOut()` which serves as a constructor.

## Slots

`covobj`: A list containing intermediate results of the PCOUT algorithm for each class

`call`, `counts`, `grp`, `wt`, `flag`, `method`, `singularity`: from the "Outlier" class.

## Extends

Class "Outlier", directly.

## Methods

**getCutoff** Return the cutoff value used to identify outliers

**getDistance** Return a vector containing the computed distances

**plot** Plot the results of the outlier detection process

## Author(s)

Valentin Todorov <valentin.todorov@chello.at>

## References

P. Filzmoser, R. Maronna and M. Werner (2008). Outlier identification in high dimensions, *Computational Statistics & Data Analysis*, Vol. 52 1694–1711.

Filzmoser P & Todorov V (2013). Robust tools for the imperfect world, *Information Sciences* **245**, 4–20. doi:10.1016/j.ins.2012.10.017.

## See Also

OutlierPCOut, "Outlier"

## Examples

```
showClass("OutlierMahdist")
```

`OutlierSign1` *Outlier identification in high dimensions using the SIGN1 algorithm*

## Description

Fast algorithm for identifying multivariate outliers in high-dimensional and/or large datasets, using spatial signs, see Filzmoser, Maronna, and Werner (CSDA, 2007). The computation of the distances is based on Mahalanobis distances.

## Usage

```
    OutlierSign1(x, ...)
    ## Default S3 method:
OutlierSign1(x, grouping, qcrit = 0.975, trace=FALSE, ...)
    ## S3 method for class 'formula'
OutlierSign1(formula, data, ..., subset, na.action)
```

## Arguments

| | |
|---|---|
| `formula` | a formula with no response variable, referring only to numeric variables. |
| `data` | an optional data frame (or similar: see `model.frame`) containing the variables in the formula `formula`. |
| `subset` | an optional vector used to select rows (observations) of the data matrix `x`. |
| `na.action` | a function which indicates what should happen when the data contain `NA`s. The default is set by the `na.action` setting of `options`, and is `na.fail` if that is unset. The default is `na.omit`. |
| `...` | arguments passed to or from other methods. |
| `x` | a matrix or data frame. |
| `grouping` | grouping variable: a factor specifying the class for each observation. |
| `qcrit` | a numeric value between 0 and 1 indicating the quantile to be used as critical value for outlier detection (default to 0.975). |
| `trace` | whether to print intermediate results. Default is `trace = FALSE` |

## Details

Based on the robustly sphered and normed data, robust principal components are computed. These are used for computing the covariance matrix which is the basis for Mahalanobis distances. A critical value from the chi-square distribution is then used as outlier cutoff.

## Value

An S4 object of class `OutlierSign1` which is a subclass of the virtual class `Outlier`.

## Author(s)

Valentin Todorov <valentin.todorov@chello.at>

## References

P. Filzmoser, R. Maronna and M. Werner (2008). Outlier identification in high dimensions, *Computational Statistics & Data Analysis*, Vol. 52 1694–1711.

Filzmoser P & Todorov V (2013). Robust tools for the imperfect world, *Information Sciences* **245**, 4–20. doi:10.1016/j.ins.2012.10.017.

## See Also

OutlierSign1, OutlierSign2, Outlier

## Examples

```
data(hemophilia)
obj <- OutlierSign1(gr~.,data=hemophilia)
obj

getDistance(obj)            # returns an array of distances
getClassLabels(obj, 1)      # returns an array of indices for a given class
getCutoff(obj)          # returns an array of cutoff values (for each class, usually equal)
getFlag(obj)                #  returns an 0/1 array of flags
plot(obj, class=2)          # standard plot function
```

---

# OutlierSign1-class *Class "OutlierSign1" - Outlier identification in high dimensions using the SIGN1 algorithm*

---

## Description

Fast algorithm for identifying multivariate outliers in high-dimensional and/or large datasets, using spatial signs, see Filzmoser, Maronna, and Werner (CSDA, 2007). The computation of the distances is based on Mahalanobis distances.

## Objects from the Class

Objects can be created by calls of the form `new("OutlierSign1", ...)` but the usual way of creating `OutlierSign1` objects is a call to the function `OutlierSign1()` which serves as a constructor.

## Slots

**`covobj`:** A list containing intermediate results of the SIGN1 algorithm for each class

**`call`, `counts`, `grp`, `wt`, `flag`, `method`, `singularity`:** from the "Outlier" class.

## Extends

Class "Outlier", directly.

## Methods

**getCutoff** Return the cutoff value used to identify outliers

**getDistance** Return a vector containing the computed distances

## Author(s)

Valentin Todorov <valentin.todorov@chello.at>

## References

P. Filzmoser, R. Maronna and M. Werner (2008). Outlier identification in high dimensions, *Computational Statistics & Data Analysis*, Vol. 52 1694–1711.

Filzmoser P & Todorov V (2013). Robust tools for the imperfect world, *Information Sciences* **245**, 4–20. doi:10.1016/j.ins.2012.10.017.

## See Also

OutlierSign1, OutlierSign2, Outlier

## Examples

```
showClass("OutlierSign1")
```

---

`OutlierSign2` *Outlier identification in high dimensions using the SIGN2 algorithm*

---

## Description

Fast algorithm for identifying multivariate outliers in high-dimensional and/or large datasets, using spatial signs, see Filzmoser, Maronna, and Werner (CSDA, 2007). The computation of the distances is based on principal components.

## Usage

```
    OutlierSign2(x, ...)
    ## Default S3 method:
OutlierSign2(x, grouping, qcrit = 0.975, explvar=0.99, trace=FALSE, ...)
    ## S3 method for class 'formula'
OutlierSign2(formula, data, ..., subset, na.action)
```

## Arguments

| | |
|---|---|
| formula | a formula with no response variable, referring only to numeric variables. |
| data | an optional data frame (or similar: see model.frame) containing the variables in the formula formula. |
| subset | an optional vector used to select rows (observations) of the data matrix x. |
| na.action | a function which indicates what should happen when the data contain NAs. The default is set by the na.action setting of options, and is na.fail if that is unset. The default is na.omit. |
| ... | arguments passed to or from other methods. |
| x | a matrix or data frame. |
| grouping | grouping variable: a factor specifying the class for each observation. |
| explvar | a numeric value between 0 and 1 indicating how much variance should be covered by the robust PCs. Default is 0.99. |
| qcrit | a numeric value between 0 and 1 indicating the quantile to be used as critical value for outlier detection. Default is 0.975. |
| trace | whether to print intermediate results. Default is trace = FALSE |

## Details

Based on the robustly sphered and normed data, robust principal components are computed which are needed for determining distances for each observation. The distances are transformed to approach chi-square distribution, and a critical value is then used as outlier cutoff.

## Value

An S4 object of class OutlierSign2 which is a subclass of the virtual class Outlier.

## Author(s)

Valentin Todorov <valentin.todorov@chello.at>

## References

P. Filzmoser, R. Maronna and M. Werner (2008). Outlier identification in high dimensions, *Computational Statistics & Data Analysis*, Vol. 52 1694–1711.

Filzmoser P & Todorov V (2013). Robust tools for the imperfect world, *Information Sciences* **245**, 4–20. doi:10.1016/j.ins.2012.10.017.

## See Also

OutlierSign2, OutlierSign1, Outlier

## Examples

```
data(hemophilia)
obj <- OutlierSign2(gr~.,data=hemophilia)
obj

getDistance(obj)            # returns an array of distances
getClassLabels(obj, 1)      # returns an array of indices for a given class
getCutoff(obj)          # returns an array of cutoff values (for each class, usually equal)
getFlag(obj)                # returns an 0/1 array of flags
plot(obj, class=2)          # standard plot function
```

---

`OutlierSign2-class` *Class* "`OutlierSign2`" - *Outlier identification in high dimensions using the SIGN2 algorithm*

---

## Description

Fast algorithm for identifying multivariate outliers in high-dimensional and/or large datasets, using spatial signs, see Filzmoser, Maronna, and Werner (CSDA, 2007). The computation of the distances is based on principal components.

## Objects from the Class

Objects can be created by calls of the form `new("OutlierSign2", ...)` but the usual way of creating `OutlierSign2` objects is a call to the function `OutlierSign2()` which serves as a constructor.

## Slots

`covobj`: A list containing intermediate results of the SIGN2 algorithm for each class

`call`, `counts`, `grp`, `wt`, `flag`, `method`, `singularity`: from the "`Outlier`" class.

## Extends

Class "`Outlier`", directly.

## Methods

**getCutoff** Return the cutoff value used to identify outliers

**getDistance** Return a vector containing the computed distances

## Author(s)

Valentin Todorov <valentin.todorov@chello.at>

## References

P. Filzmoser, R. Maronna and M. Werner (2008). Outlier identification in high dimensions, *Computational Statistics & Data Analysis*, Vol. 52 1694–1711.

Filzmoser P & Todorov V (2013). Robust tools for the imperfect world, *Information Sciences* **245**, 4–20. doi:10.1016/j.ins.2012.10.017.

## See Also

`OutlierSign2`, `OutlierSign1`, `Outlier`

## Examples

```
showClass("OutlierSign2")
```

---

# PredictSimca-class *Class* "PredictSimca" - *prediction of* "Simca" *objects*

---

## Description

The prediction of a "Simca" object

## Objects from the Class

Objects can be created by calls of the form `new("PredictSimca", ...)` but most often by invoking `predict()` on a "Simca" object. They contain values meant for printing by `show()`

## Slots

**`classification`:** Object of class "factor" ~~

**`odsc`:** A "matrix" containing the standartized orthogonal distances for each group

**`sdsc`:** A "matrix" containing the standartized score distances for each group

**`ct`:** re-classification table of the training sample

## Methods

**show** `signature(object = "PredictSimca")`: Prints the results..

## Author(s)

Valentin Todorov <valentin.todorov@chello.at>

## References

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.

Todorov V & Filzmoser P (2014), Software Tools for Robust Analysis of High-Dimensional Data. *Austrian Journal of Statistics*, **43**(4), 255–266, doi:10.17713/ajs.v43i4.44.

## See Also

Simca-class

## Examples

```
showClass("PredictSimca")
```

---

PredictSosDisc-class *Class* "PredictSosDisc" - *prediction of* "SosDisc" *objects*

---

## Description

The prediction of a "SosDisc" object

## Objects from the Class

Objects can be created by calls of the form `new("PredictSosDisc", ...)` but most often by invoking `predict()` on a "SosDisc" object. They contain values meant for printing by `show()`

## Slots

`classification`: Object of class "factor" representing the predicted classification

`mahadist2`: A "matrix" containing the squared robust Mahalanobis distances to each group center in the subspace (see Details).

`w`: A "vector" containing the weights derived from robust Mahalanobis distances to the closest group center (see Details).

## Details

For the prediction of the class membership a two step approach is taken. First, the `newdata` are scaled and centered (by obj@scale and obj@center) and multiplied by `obj@beta` for dimension reduction. Then the classification of the transformed data is obtained by prediction with the `Linda` object obj@fit. The Mahalanobis distances to the closest group center in this subspace is used to derive case weights `w`. Observations where the squared robust mahalanobis distance is larger than the 0.975 quantile of the chi-square distribution with Q degrees of freedom receive weight zero, all others weight one.

## Methods

**show** `signature(object = "PredictSosDisc")`: Prints the results.

## Author(s)

Irene Ortner <irene.ortner@applied-statistics.at> and Valentin Todorov <valentin.todorov@chello.at>

## References

Clemmensen L, Hastie T, Witten D & Ersboll B (2012), Sparse discriminant analysis. *Technometrics*, **53**(4), 406–413.

Ortner I, Filzmoser P & Croux C (2020), Robust and sparse multigroup classification by the optimal scoring approach. Data Mining and Knowledge Discovery **34**, 723–741. doi:10.1007/s10618019-006668.

## See Also

SosDisc-class

## Examples

```
showClass("PredictSosDisc")
```

---

RSimca *Robust classification in high dimensions based on the SIMCA method*

---

## Description

RSimca performs a robust version of the SIMCA method. This method classifies a data matrix x with a known group structure. To reduce the dimension on each group a robust PCA analysis is performed. Afterwards a classification rule is developped to determine the assignment of new observations.

## Usage

```
RSimca(x, ...)
## Default S3 method:
RSimca(x, grouping, prior=proportions, k, kmax = ncol(x),
    control="hubert", alpha, tol = 1.0e-4, trace=FALSE, ...)
## S3 method for class 'formula'
RSimca(formula, data = NULL, ..., subset, na.action)
```

## Arguments

| | |
|---|---|
| formula | a formula of the form y~x, it describes the response and the predictors. The formula can be more complicated, such as y~log(x)+z etc (see formula for more details). The response should be a factor representing the response variable, or any vector that can be coerced to such (such as a logical variable). |
| data | an optional data frame (or similar: see model.frame) containing the variables in the formula formula. |
| subset | an optional vector used to select rows (observations) of the data matrix x. |
| na.action | a function which indicates what should happen when the data contain NAs. The default is set by the na.action setting of options, and is na.fail if that is unset. The default is na.omit. |

| | |
|---|---|
| x | a matrix or data frame containing the explanatory variables (training set). |
| grouping | grouping variable: a factor specifying the class for each observation. |
| prior | prior probabilities, default to the class proportions for the training set. |
| tol | tolerance |
| control | a control object (S4) for specifying one of the available PCA estimation methods and containing estimation options. The class of this object defines which estimator will be used. Alternatively a character string can be specified which names the estimator - one of auto, hubert, locantore, grid, proj. If 'auto' is specified or the argument is missing, the function will select the estimator (see below for details) |
| alpha | this parameter measures the fraction of outliers the algorithm should resist. In MCD alpha controls 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. |
| k | number of principal components to compute. If k is missing, or k = 0, the algorithm itself will determine the number of components by finding such k that $l_k/l_1 >= 10.E-3$ and $\Sigma_{j=1}^{k} l_j / \Sigma_{j=1}^{r} l_j >= 0.8$. It is preferable to investigate the scree plot in order to choose the number of components and then run again. Default is k=0. |
| kmax | maximal number of principal components to compute. Default is kmax=10. If k is provided, kmax does not need to be specified, unless k is larger than 10. |
| trace | whether to print intermediate results. Default is trace = FALSE |
| ... | arguments passed to or from other methods. |

## Details

RSimca, serving as a constructor for objects of class RSimca-class is a generic function with "formula" and "default" methods.

SIMCA is a two phase procedure consisting of PCA performed on each group separately for dimension reduction followed by classification rules built in the lower dimensional space (note that the dimension in each group can be different). Instead of classical PCA robust alternatives will be used. Any of the robust PCA methods available in package Pca-class can be used through the argument control. In original SIMCA new observations are classified by means of their deviations from the different PCA models. Here the classification rules will be obtained using two popular distances arising from PCA - orthogonal distances (OD) and score distances (SD). For the definition of these distances, the definition of the cutoff values and the standartization of the distances see Vanden Branden K, Hubert M (2005) and Todorov and Filzmoser (2009).

## Value

An S4 object of class RSimca-class which is a subclass of of the virtual class Simca-class.

## Author(s)

Valentin Todorov <valentin.todorov@chello.at>

## References

Vanden Branden K, Hubert M (2005) Robust classification in high dimensions based on the SIMCA method. Chemometrics and Intellegent Laboratory Systems 79:10–21

Todorov V & Filzmoser P (2014), Software Tools for Robust Analysis of High-Dimensional Data. *Austrian Journal of Statistics*, **43**(4), 255–266, doi:10.17713/ajs.v43i4.44.

## Examples

```
data(pottery)
dim(pottery)        # 27 observations in 2 classes, 6 variables
head(pottery)

## Build the SIMCA model. Use RSimca for a robust version
rs <- RSimca(origin~., data=pottery)
rs
summary(rs)


## generate a sample from the pottery data set -
##  this will be the "new" data to be predicted
smpl <- sample(1:nrow(pottery), 5)
test <- pottery[smpl, -7]       # extract the test sample. Remove the last (grouping) variable
print(test)


## predict new data
pr <- predict(rs, newdata=test)

pr@classification
```

---

RSimca-class *Class* "RSimca" - robust classification in high dimensions based on the SIMCA method

---

## Description

The class RSimca represents robust version of the SIMCA algorithm for classification in high dimensions. The objects of class RSImca contain the results of the robust SIMCA method.

## Objects from the Class

Objects can be created by calls of the form new("RSImca", ...) but the usual way of creating RSimca objects is a call to the function RSimca() which serves as a constructor.

## Slots

**call:** the (matched) function call.

**prior:** prior probabilities used, default to group proportions

`counts`: number of observations in each class

`pcaobj`: A list of Pca objects - one for each group

`k`: Object of class `"numeric"` number of (choosen) principal components

`flag`: Object of class `"Uvector"` The observations whose score distance is larger than cutoff.sd or whose orthogonal distance is larger than cutoff.od can be considered as outliers and receive a flag equal to zero. The regular observations receive a flag 1

`X`: the training data set (same as the input parameter x of the constructor function)

`grp`: grouping variable: a factor specifying the class for each observation.

## Extends

Class `"Simca"`, directly.

## Methods

No methods defined with class "RSimca" in the signature.

## Author(s)

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

## References

Vanden Branden K, Hubert M (2005) Robust classification in high dimensions based on the SIMCA method. Chemometrics and Intellegent Laboratory Systems 79:10–21

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.

Todorov V & Filzmoser P (2014), Software Tools for Robust Analysis of High-Dimensional Data. *Austrian Journal of Statistics*, **43**(4), 255–266, doi:10.17713/ajs.v43i4.44.

## Examples

```
showClass("RSimca")
```

---

# Simca-class *Class "Simca" - virtual base class for all classic and robust SIMCA classes representing classification in high dimensions based on the SIMCA method*

---

## Description

The class `Simca` searves as a base class for deriving all other classes representing the results of the classical and robust SIMCA methods

## Objects from the Class

A virtual Class: No objects may be created from it.

## Slots

`call`: the (matched) function call.

`prior`: prior probabilities used, default to group proportions

`counts`: number of observations in each class

`pcaobj`: A list of Pca objects - one for each group

`k`: Object of class `"numeric"` number of (choosen) principal components

`flag`: Object of class `"Uvector"` The observations whose score distance is larger than cutoff.sd or whose orthogonal distance is larger than cutoff.od can be considered as outliers and receive a flag equal to zero. The regular observations receive a flag 1

`X`: the training data set (same as the input parameter x of the constructor function)

`grp`: grouping variable: a factor specifying the class for each observation.

## Methods

**predict** `signature(object = "Simca")`: calculates prediction using the results in `object`. An optional data frame or matrix in which to look for variables with which to predict. If omitted, the training data set is used. If the original fit used a formula or a data frame or a matrix with column names, newdata must contain columns with the same names. Otherwise it must contain the same number of columns, to be used in the same order.

**show** `signature(object = "Simca")`: prints the results

**summary** `signature(object = "Simca")`: prints summary information

## Author(s)

Valentin Todorov <valentin.todorov@chello.at>

## References

Vanden Branden K, Hubert M (2005) Robust classification in high dimensions based on the SIMCA method. Chemometrics and Intellegent Laboratory Systems 79:10–21

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.

Todorov V & Filzmoser P (2014), Software Tools for Robust Analysis of High-Dimensional Data. *Austrian Journal of Statistics*, **43**(4), 255–266, doi:10.17713/ajs.v43i4.44.

## Examples

```
showClass("Simca")
```

SosDisc-class — *Class "SosDisc" - virtual base class for all classic and robust SosDisc classes representing the results of the robust and sparse multigroup classification by the optimal scoring approach*

## Description

Robust and sparse multigroup classification by the optimal scoring approach. The class SosDisc searves as a base class for deriving all other classes representing the results of the robust and sparse multigroup classification by the optimal scoring approach.

## Details

The sparse optimal scoring problem (Clemmensen et al, 2011): for $h = 1, ...., Q$

$$\min_{\beta_h, \theta_h} \frac{1}{n} \|Y\theta_h - X\beta_h\|_2^2 + \lambda \|\beta_h\|_1$$

subject to

$$\frac{1}{n}\theta_h^T Y^T Y \theta_h = 1, \quad \theta_h^T Y^T Y \theta_l = 0 \quad \forall l < h.$$

where $X$ deontes the robustly centered and scaled input matrix x (or alternativly the predictors from formular) and $Y$ is an dummy matrix coding die classmemberships from grouping.

For each $h$ this problem can be solved interatively for $\beta_h$ and $\theta_h$. In order to obtain robust estimates, $\beta_h$ is estimated with reweighted sparse least trimmed squares regression (Alfons et al, 2013) and $\theta_h$ with least absolut deviation regression in the first two iterations. To speed up the following repetitions an iterative down-weighting of observations with large residuals is combined with the iterative estimation of the optimal scoring coefficients with their classical estimates.

The classification model is estimated on the low dimensional sparse subspace $X[\beta_1, ..., \beta_Q]$ with robust LDA (Linda).

## Objects from the Class

A virtual Class: No objects may be created from it.

## Slots

**call:** The (matched) function call.

**prior:** Prior probabilities; same as input parameter.

**counts:** Number of observations in each class.

**beta:** Object of class "matrix": Q coefficient vectors of the predictor matrix from optimal scoring (see Details); rows corespond to variables listed in varnames.

**theta:** Object of class "matrix": Q coefficient vectors of the dummy matrix for class coding from optimal scoring (see Details).

**lambda:** Non-negative tuning paramer from L1 norm penaly; same as input parameter

`varnames`: Character vector: Names of included predictor variables (variables where at least one beta coefficient is non-zero).

`center`: Centering vector of the input predictors (coordinate wise median).

`scale`: Scaling vector of the input predictors (mad).

`fit`: Object of class `"Linda"`: Linda model (robust LDA model) estimated in the low dimensional subspace $X[\beta_1, ..., \beta_Q]$ (see Details)

`mahadist2`: These will go later to Linda object: squared robust Mahalanobis distance (calculated with estimates from Linda, with common covariance structure of all groups) of each observation to its group center in the low dimensional subspace $X[\beta_1, ..., \beta_Q]$ (see Details).

`wlinda`: These will go later to Linda object: 0-1 weights derived from `mahadist2`; observations where the squred robust Mahalanobis distance is larger than the 0.975 quantile of the chi-square distribution with Q degrees of freedom resive weight zero.

`X`: The training data set (same as the input parameter `x` of the constructor function)

`grp`: Grouping variable: a factor specifying the class for each observation (same as the input parameter `grouping`)

## Methods

**predict** `signature(object = "SosDisc")`: calculates prediction using the results in `object`. An optional data frame or matrix in which to look for variables with which to predict. If omitted, the training data set is used. If the original fit used a formula or a data frame or a matrix with column names, newdata must contain columns with the same names.

**show** `signature(object = "SosDisc")`: prints the results

**summary** `signature(object = "SosDisc")`: prints summary information

## Author(s)

Irene Ortner <irene.ortner@applied-statistics.at> and Valentin Todorov <valentin.todorov@chello.at>

## References

Clemmensen L, Hastie T, Witten D & Ersboll B (2012), Sparse discriminant analysis. *Technometrics*, **53**(4), 406–413.

Ortner I, Filzmoser P & Croux C (2020), Robust and sparse multigroup classification by the optimal scoring approach. Data Mining and Knowledge Discovery **34**, 723–741. doi:10.1007/s10618019-006668.

## Examples

```
showClass("SosDisc")
```

SosDiscClassic-class *Class* SosDiscClassic - *sparse multigroup classification by the optimal scoring approach*

## Description

Sparse multigroup classification by the optimal scoring approach.

## Objects from the Class

Objects can be created by calls of the form `new("SosDiscClassic", ...)` but the usual way of creating `SosDiscClassic` objects is a call to the function `SosDiscRobust()` which serves as a constructor.

## Slots

`call`: The (matched) function call.

`prior`: Prior probabilities; same as input parameter.

`counts`: Number of observations in each class.

`beta`: Object of class `"matrix"`: Q coefficient vectors of the predictor matrix from optimal scoring (see Details); rows corespond to variables listed in `varnames`.

`theta`: Object of class `"matrix"`: Q coefficient vectors of the dummy matrix for class coding from optimal scoring (see Details).

`lambda`: Non-negative tuning paramer from L1 norm penaly; same as input parameter

`varnames`: Character vector: Names of included predictor variables (variables where at least one beta coefficient is non-zero).

`center`: Centering vector of the input predictors (coordinate wise median).

`scale`: Scaling vector of the input predictors (mad).

`fit`: Object of class `"Linda"`: Linda model (robust LDA model) estimated in the low dimensional subspace $X[\beta_1, ..., \beta_Q]$ (see Details)

`mahadist2`: These will go later to Linda object: squared robust Mahalanobis distance (calculated with estimates from Linda, with common covariance structure of all groups) of each observation to its group center in the low dimensional subspace $X[\beta_1, ..., \beta_Q]$ (see Details).

`wlinda`: These will go later to Linda object: 0-1 weights derived from `mahadist2`; observations where the squred robust Mahalanobis distance is larger than the 0.975 quantile of the chi-square distribution with Q degrees of freedom resive weight zero.

`X`: The training data set (same as the input parameter `x` of the constructor function)

`grp`: Grouping variable: a factor specifying the class for each observation (same as the input parameter `grouping`)

## Extends

Class `"SosDisc"`, directly.

## Methods

No methods defined with class "SosDiscClassic" in the signature.

## Author(s)

Irene Ortner <irene.ortner@applied-statistics.at> and Valentin Todorov <valentin.todorov@chello.at>

## References

Clemmensen L, Hastie T, Witten D & Ersboll B (2012), Sparse discriminant analysis. *Technometrics*, **53**(4), 406–413.

Ortner I, Filzmoser P & Croux C (2020), Robust and sparse multigroup classification by the optimal scoring approach. Data Mining and Knowledge Discovery **34**, 723–741. doi:10.1007/s10618019-006668.

## Examples

```
showClass("SosDiscClassic")
```

---

# SosDiscRobust *Robust and sparse multigroup classification by the optimal scoring approach*

---

## Description

Robust and sparse multigroup classification by the optimal scoring approach is robust against outliers, provides a low-dimensional and sparse representation of the predictors and is also applicable if the number of variables exeeds the number of observations.

## Usage

```
SosDiscRobust(x, ...)
## Default S3 method:
SosDiscRobust(x, grouping, prior=proportions,
    lambda, Q=length(unique(grouping))-1, alpha=0.5, maxit=100,
    tol = 1.0e-4, trace=FALSE, ...)
## S3 method for class 'formula'
SosDiscRobust(formula, data = NULL, ..., subset, na.action)
```

## Arguments

| | |
|---|---|
| formula | A formula of the form y~x, it describes the response and the predictors. The formula can be more complicated, such as y~log(x)+z etc (see formula for more details). The response should be a factor representing the response variable, or any vector that can be coerced to such (such as a logical variable). |
| data | An optional data frame (or similar: see model.frame) containing the variables in the formula formula. |

| | |
|---|---|
| subset | An optional vector used to select rows (observations) of the data matrix x. |
| na.action | A function which indicates what should happen when the data contain NAs. The default is set by the na.action setting of options, and is na.fail if that is unset. The default is na.omit. |
| x | A matrix or data frame containing the explanatory variables (training set); colnames of x have to be provided. |
| grouping | Grouping variable: a factor specifying the class for each observation. |
| prior | Prior probabilities, a vector of positive numbers that sum up to 1; default to the class proportions for the training set. |
| lambda | A non-negative tuning parameter for L1 norm penalty introducing sparsity on the optimal scoring coefficients $\boldsymbol{\beta}_h$ (see Details). If the number of variables exceeds the number of observations lambda has to be positive. |
| Q | Number of optimal scoring coefficient vectors; Q has to be smaller than the number of groups. Defaults to number of groups - 1. |
| alpha | Robustness parameter used in sparseLTS (for initial estimation, see Details). Default alpha=0.5. |
| maxit | Number of iterations for the estimation of optimal scoring coefficients and case weights. Default maxit=100. |
| tol | Tolerance for convergence of the normed weighted change in the residual sum of squares for the estiamtion of optimal scoring coefficeints. Default is tol=1.0e-4. |
| trace | Whether to print intermediate results. Default is trace = FALSE. |
| ... | Arguments passed to or from other methods. |

## Details

The sparse optimal scoring problem (Clemmensen et al, 2011): for $h = 1, ...., Q$

$$\min_{\beta_h, \theta_h} \frac{1}{n} \|Y\theta_h - X\beta_h\|_2^2 + \lambda \|\beta_h\|_1$$

subject to

$$\frac{1}{n}\theta_h^T Y^T Y \theta_h = 1, \quad \theta_h^T Y^T Y \theta_l = 0 \quad \forall l < h,$$

where $X$ deontes the robustly centered and scaled input matrix x (or alternativly the predictors from formular) and $Y$ is an dummy matrix coding die classmemberships from grouping.

For each $h$ this problem can be solved interatively for $\beta_h$ and $\theta_h$. In order to obtain robust estimates, $\beta_h$ is estimated with reweighted sparse least trimmed squares regression (Alfons et al, 2013) and $\theta_h$ with least absolut deviation regression in the first two iterations. To speed up the following repetitions an iterative down-weighting of observations with large residuals is combined with the iterative estimation of the optimal scoring coefficients with their classical estimates.

The classification model is estimated on the low dimensional sparse subspace $X[\beta_1, ..., \beta_Q]$ with robust LDA (Linda).

## Value

An S4 object of class SosDiscRobust-class which is a subclass of of the virtual class SosDisc-class.

## Author(s)

Irene Ortner <irene.ortner@applied-statistics.at> and Valentin Todorov <valentin.todorov@chello.at>

## References

Clemmensen L, Hastie T, Witten D & Ersboll B (2011), Sparse discriminant analysis. *Technometrics*, **53**(4), 406–413.

Alfons A, Croux C & Gelper S (2013), Sparse least trimmed squares regression for analysing high-dimensional large data sets. *The Annals of Applied Statistics*, **7**(1), 226–248.

Ortner I, Filzmoser P & Croux C (2020), Robust and sparse multigroup classification by the optimal scoring approach. Data Mining and Knowledge Discovery **34**, 723–741. doi:10.1007/s10618019-006668.

## Examples

```
## EXAMPLE 1 ####################################
data(olitos)
grind <- which(colnames(olitos)=="grp")

set.seed(5008642)
mod <- SosDiscRobust(grp~., data=olitos, lambda=0.3, maxIte=30, Q=3, tol=1e-2)

pred <- predict(mod, newdata=olitos[,-grind])

summary(mod)
plot(mod, ind=c(1:3))


## EXAMPLE 2 ####################################
##

## Not run:
library(sparseLDA)
data(penicilliumYES)

## for demonstration only:
set.seed(5008642)
X <- penicilliumYES$X[, sample(1:ncol(penicilliumYES$X), 100)]

## takes a subsample of the variables
## to have quicker computation time

colnames(X) <- paste0("V",1:ncol(X))
y <- as.factor(c(rep(1,12), rep(2,12), rep(3,12)))

set.seed(5008642)
mod <- SosDiscRobust(X, y, lambda=1, maxit=5, Q=2, tol=1e-2)

summary(mod)
plot(mod)
```

```
## End(Not run)
```

---

`SosDiscRobust-class` *Class* `SosDiscRobust` - *robust and sparse multigroup classification by the optimal scoring approach*

---

## Description

Robust and sparse multigroup classification by the optimal scoring approach.

## Objects from the Class

Objects can be created by calls of the form `new("SosDiscRobust", ...)` but the usual way of creating `SosDiscRobust` objects is a call to the function `SosDiscRobust()` which serves as a constructor.

## Slots

`call`: The (matched) function call.

`prior`: Prior probabilities; same as input parameter.

`counts`: Number of observations in each class.

`beta`: Object of class `"matrix"`: Q coefficient vectors of the predictor matrix from optimal scoring (see Details); rows corespond to variables listed in `varnames`.

`theta`: Object of class `"matrix"`: Q coefficient vectors of the dummy matrix for class coding from optimal scoring (see Details).

`lambda`: Non-negative tuning paramer from L1 norm penaly; same as input parameter

`varnames`: Character vector: Names of included predictor variables (variables where at least one beta coefficient is non-zero).

`center`: Centering vector of the input predictors (coordinate wise median).

`scale`: Scaling vector of the input predictors (mad).

`fit`: Object of class `"Linda"`: Linda model (robust LDA model) estimated in the low dimensional subspace $X[\beta_1, ..., \beta_Q]$ (see Details)

`mahadist2`: These will go later to Linda object: squared robust Mahalanobis distance (calculated with estimates from Linda, with common covariance structure of all groups) of each observation to its group center in the low dimensional subspace $X[\beta_1, ..., \beta_Q]$ (see Details).

`wlinda`: These will go later to Linda object: 0-1 weights derived from `mahadist2`; observations where the squred robust Mahalanobis distance is larger than the 0.975 quantile of the chi-square distribution with Q degrees of freedom resive weight zero.

`X`: The training data set (same as the input parameter `x` of the constructor function)

`grp`: Grouping variable: a factor specifying the class for each observation (same as the input parameter `grouping`)

## Extends

Class "SosDisc", directly.

## Methods

No methods defined with class "SosDiscRobust" in the signature.

## Author(s)

Irene Ortner <irene.ortner@applied-statistics.at> and Valentin Todorov <valentin.todorov@chello.at>

## References

Clemmensen L, Hastie T, Witten D & Ersboll B (2012), Sparse discriminant analysis. *Technometrics*, **53**(4), 406–413.

Ortner I, Filzmoser P & Croux C (2020), Robust and sparse multigroup classification by the optimal scoring approach. Data Mining and Knowledge Discovery **34**, 723–741. doi:10.1007/s10618019-006668.

## Examples

```
showClass("SosDiscRobust")
```

---

# SPcaGrid — *Sparse Robust Principal Components based on Projection Pursuit (PP): GRID search Algorithm*

---

## Description

Computes an approximation of the PP-estimators for sparse and robust PCA using the grid search algorithm in the plane.

## Usage

```
    SPcaGrid(x, ...)
    ## Default S3 method:
SPcaGrid(x, k = 0, kmax = ncol(x), method = c ("mad", "sd", "qn", "Qn"),
    lambda = 1, scale=FALSE, na.action = na.fail, trace=FALSE, ...)
    ## S3 method for class 'formula'
SPcaGrid(formula, data = NULL, subset, na.action, ...)
```

## Arguments

| | |
|---|---|
| formula | a formula with no response variable, referring only to numeric variables. |
| data | an optional data frame (or similar: see model.frame) containing the variables in the formula formula. |
| subset | an optional vector used to select rows (observations) of the data matrix x. |
| na.action | a function which indicates what should happen when the data contain NAs. The default is set by the na.action setting of options, and is na.fail if that is unset. The default is na.omit. |
| ... | arguments passed to or from other methods. |
| x | a numeric matrix (or data frame) which provides the data for the principal components analysis. |
| k | number of principal components to compute. If k is missing, or k = 0, the algorithm itself will determine the number of components by finding such k that $l_k/l_1 >= 10.E-3$ and $\Sigma_{j=1}^{k} l_j/\Sigma_{j=1}^{r} l_j >= 0.8$. It is preferable to investigate the scree plot in order to choose the number of components and then run again. Default is k=0. |
| kmax | maximal number of principal components to compute. Default is kmax=10. If k is provided, kmax does not need to be specified, unless k is larger than 10. |
| method | the scale estimator used to detect the direction with the largest variance. Possible values are "sd", "mad" and "Qn". "mad" is the default value. |
| lambda | the sparseness constraint's strength(sPCAgrid only). A single value for all components, or a vector of length k with different values for each component can be specified. See opt.TPO for the choice of this argument. |
| scale | a value indicating whether and how the variables should be scaled. If scale = FALSE (default) or scale = NULL no scaling is performed (a vector of 1s is returned in the scale slot). If scale = TRUE the data are scaled to have unit variance. Alternatively it can be a function like sd or mad or a vector of length equal the number of columns of x. The value is passed to the underlying function and the result returned is stored in the scale slot. Default is scale = FALSE |
| trace | whether to print intermediate results. Default is trace = FALSE |

## Details

SPcaGrid, serving as a constructor for objects of class SPcaGrid-class is a generic function with "formula" and "default" methods. For details see sPCAgrid and the relevant references.

## Value

An S4 object of class SPcaGrid-class which is a subclass of PcaGrid-class which in turn is a subclass of the virtual class PcaRobust-class.

## Author(s)

Valentin Todorov <valentin.todorov@chello.at>

## References

C. Croux, P. Filzmoser, M. Oliveira, (2007). Algorithms for Projection-Pursuit Robust Principal Component Analysis, *Chemometrics and Intelligent Laboratory Systems*, Vol. 87, pp. 218-225.

C. Croux, P. Filzmoser, H. Fritz (2013). Robust Sparse Principal Component Analysis, *Technometrics* **55**(2), pp. 202–2014, doi:10.1080/00401706.2012.727746.

V. Todorov, P. Filzmoser (2013). Comparing classical and robust sparse PCA. In R Kruse, M Berthold, C Moewes, M Gil, P Grzegorzewski, O Hryniewicz (eds.), *Synergies of Soft Computing and Statistics for Intelligent Data Analysis*, volume 190 of *Advances in Intelligent Systems and Computing*, pp. 283–291. Springer, Berlin; New York. ISBN 978-3-642-33041-4, doi:10.1007/9783642330421_31.

## Examples

```
data(bus)
bus <- as.matrix(bus)

## calculate MADN for each variable
xmad <- apply(bus, 2, mad)
cat("\nMin, Max of MADN: ", min(xmad), max(xmad), "\n")

## calculate MADN for each variable
xqn <- apply(bus, 2, Qn)
cat("\nMin, Max of Qn: ", min(xqn), max(xqn), "\n")


## MADN vary between 0 (for variable 9) and 34. Therefore exclude
##  variable 9 and divide the remaining variables by their MADNs.
bus1 <- bus[, -c(9)]
p <- ncol(bus1)

madbus <- apply(bus1, 2, mad)
bus2 <- sweep(bus1, 2, madbus, "/", check.margin = FALSE)

xsd <- apply(bus1, 2, sd)
bus.sd <- sweep(bus1, 2, xsd, "/", check.margin = FALSE)

xqn <- apply(bus1, 2, Qn)
bus.qn <- sweep(bus1, 2, xqn, "/", check.margin = FALSE)

## Not run:
spc <- SPcaGrid(bus2, lambda=0, method="sd", k=p, kmax=p)
rspc <- SPcaGrid(bus2, lambda=0, method="Qn", k=p, kmax=p)
summary(spc)
summary(rspc)
screeplot(spc, type="line", main="Classical PCA", sub="PC", cex.main=2)
screeplot(rspc, type="line", main="Robust PCA", sub="PC", cex.main=2)

##  find lambda

K <- 4
lambda.sd <- 1.64
```

```
    to.sd <- .tradeoff(bus2, k=K, lambda.max=2.5, lambda.n=100, method="sd")
    plot(to.sd, type="b", xlab="lambda", ylab="Explained Variance (percent)")
    abline(v=lambda.sd, lty="dotted")

spc.sd.p <- SPcaGrid(bus2, lambda=lambda.sd, method="sd", k=p)
.CPEV(spc.sd.p, k=K)
spc.sd <- SPcaGrid(bus2, lambda=lambda.sd, method="sd", k=K)
getLoadings(spc.sd)[,1:K]
plot(spc.sd)

lambda.qn <- 2.06
    to.qn <- .tradeoff(bus2, k=K, lambda.max=2.5, lambda.n=100, method="Qn")
    plot(to.qn, type="b", xlab="lambda", ylab="Explained Variance (percent)")
    abline(v=lambda.qn, lty="dotted")

spc.qn.p <- SPcaGrid(bus2, lambda=lambda.qn, method="Qn", k=p)
.CPEV(spc.qn.p, k=K)
spc.qn <- SPcaGrid(bus2, lambda=lambda.qn, method="Qn", k=K)
getLoadings(spc.qn)[,1:K]
plot(spc.qn)

## End(Not run)

## DD-plots
##
## Not run:
## Not run:
usr <- par(mfrow=c(2,2))
plot(SPcaGrid(bus2, lambda=0, method="sd", k=4), id.n.sd=0, main="Standard PCA")
plot(SPcaGrid(bus2, lambda=0, method="Qn", k=4), id.n.sd=0, ylim=c(0,20))

plot(SPcaGrid(bus2, lambda=1.64, method="sd", k=4), id.n.sd=0, main="Stdandard sparse PCA")
plot(SPcaGrid(bus2, lambda=3.07, method="Qn", k=4), id.n.sd=0, main="Robust sparse PCA")

par(usr)
## End (Not run)

## End(Not run)
```

---

`SPcaGrid-class` *Class* `SPcaGrid` - *Sparse Robust PCA using PP - GRID search Algorithm*

---

## Description

Holds the results of an approximation of the PP-estimators for sparse and robust PCA using the grid search algorithm in the plane.

## Objects from the Class

Objects can be created by calls of the form `new("SPcaGrid", ...)` but the usual way of creating `SPcaGrid` objects is a call to the function `SPcaGrid()` which serves as a constructor.

## Slots

**`call`, `center`, `scale`, `loadings`, `eigenvalues`, `scores`, `k`, `sd`, `od`, `cutoff.sd`, `cutoff.od`, `flag`, `n.obs`:** from the "`Pca-class`" class.

## Extends

Class "`PcaGrid-class`", directly. Class "`PcaRobust-class`", by class "`PcaGrid-class`", distance 2. Class "`Pca-class`", by class "`PcaGrid-class`", distance 3.

## Methods

**getQuan** `signature(obj = "SPcaGrid")`: ...

## Author(s)

Valentin Todorov <valentin.todorov@chello.at>

## References

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.

C. Croux, P. Filzmoser, H. Fritz (2013). Robust Sparse Principal Component Analysis, *Technometrics* **55**(2), pp. 202–2014, doi:10.1080/00401706.2012.727746.

V. Todorov, P. Filzmoser (2013). Comparing classical and robust sparse PCA. In R Kruse, M Berthold, C Moewes, M Gil, P Grzegorzewski, O Hryniewicz (eds.), *Synergies of Soft Computing and Statistics for Intelligent Data Analysis*, volume 190 of *Advances in Intelligent Systems and Computing*, pp. 283–291. Springer, Berlin; New York. ISBN 978-3-642-33041-4, doi:10.1007/9783642330421_31.

## See Also

`SPcaGrid`, `PcaGrid-class`, `PcaRobust-class`, `Pca-class`, `PcaClassic`, `PcaClassic-class`

## Examples

```
showClass("SPcaGrid")
```

SummarySimca-class *Class* "SummarySimca" - *summary of* "Simca" *objects*

## Description

Contains summary information about a Simca object - classification in high dimensions based on the SIMCA method

## Objects from the Class

Objects can be created by calls of the form new("SummarySimca", ...), but most often by invoking summary() on an "Simca" object. They contain values meant for printing by show().

## Slots

simcaobj: Object of class "Simca"

## Methods

**show** signature(object = "SummarySimca"): display the object

## Author(s)

Valentin Todorov <valentin.todorov@chello.at>

## References

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.

Todorov V & Filzmoser P (2014), Software Tools for Robust Analysis of High-Dimensional Data. *Austrian Journal of Statistics*, **43**(4), 255–266, doi:10.17713/ajs.v43i4.44.

## See Also

Simca-class

## Examples

```
showClass("SummarySimca")
```

SummarySosDisc-class *Class* "SummarySosDisc" - *summary of* "SosDisc" *objects*

## Description

Contains summary information about a SosDisc object representing the results of the robust and sparse multigroup classification by the optimal scoring approach.

## Objects from the Class

Objects can be created by calls of the form new("SummarySosDisc", ...), but most often by invoking summary() on an "SosDisc" object. They contain values meant for printing by show().

## Slots

obj: Object of class "SosDisc"

## Methods

**show** signature(object = "SummarySosDisc"): display the object

## Author(s)

Irene Ortner <irene.ortner@applied-statistics.at> and Valentin Todorov <valentin.todorov@chello.at>

## References

Clemmensen L, Hastie T, Witten D & Ersboll B (2012), Sparse discriminant analysis. *Technometrics*, **53**(4), 406–413.

Ortner I, Filzmoser P & Croux C (2020), Robust and sparse multigroup classification by the optimal scoring approach. Data Mining and Knowledge Discovery **34**, 723–741. doi:10.1007/s10618019-006668.

## See Also

SosDisc-class

## Examples

```
showClass("SummarySosDisc")
```

# Index