Package 'PCIdep'

August 22, 2025

Type Package

Title Post-clustering inference for general matrix normal models.

Version 0.1.0

Author Javier González-Delgado

Maintainer Javier González-Delgado <javier.gonzalez-delgado@ensai.fr>

Description PCIdep extends the framework introduced in the R packages clusterpval and KMeansInference for arbitrary dependence structures between observations and features.

Encoding UTF-8

Imports matrixNormal, Matrix, stats, KmeansInference, clusterpval, intervals, fastcluster, RColorBrewer, graphics

Remotes lucylgao/clusterpval, yiqunchen/KmeansInference

RoxygenNote 7.2.3

NeedsCompilation no

License MIT + file LICENSE

R topics documented:

`test.clusters.hc` *Test for the difference of two cluster means after hierarchical clustering, for matrix normal model with arbitrary scale matrices. Supported linkages (as in Gao et al. 2022) are "single", "average", "centroid", "ward.D", "median", "mcquitty" and "complete".*

Description

Test for the difference of two cluster means after hierarchical clustering, for matrix normal model with arbitrary scale matrices. Supported linkages (as in Gao et al. 2022) are "single", "average", "centroid", "ward.D", "median", "mcquitty" and "complete".

Usage

```
test.clusters.hc(
  X,
  U = NULL,
  Sigma = NULL,
  Y = NULL,
  UY = NULL,
  precUY = NULL,
  NC,
  clusters,
  linkage = "average",
  ndraws = 2000
)
```

Arguments

X	A $n \times p$ matrix drawn from a $n \times p$ matrix normal distribution $\mathcal{MN}(\texttt{M}, \texttt{U}, \texttt{Sigma})$. X must have n rows and p columns.
U	A $n \times n$ positive-definite matrix describing the dependence structure between the rows in X. If NULL, observations are considered independent and U is set to the $n \times n$ identity matrix.
Sigma	A $p \times p$ positive-definite matrix describing the dependence structure between the columns in X. If NULL, Sigma is over-estimated (in the sense of the Loewner partial order).
Y	If Sigma is NULL, an i.i.d. copy of X allowing its estimation. Y must have the same number of columns as X.
UY	If Sigma is NULL, a positive-definite matrix describing the dependence structure between the rows in Y. If NULL and its inverse is not provided, set to the identity matrix by default.
precUY	The inverse matrix of UY, that can be provided to increase computational efficiency. If UY is not NULL and precUY is NULL, precUY is obtained by inverting UY.
NC	The number of clusters to choose.
clusters	A vector of two integers from 1 to NC indicating the pair of clusters whose means have to be compared.
linkage	The type of linkage for hierarchical clustering. Must be either single, average, centroid, ward.D, median, mcquitty or complete.
ndraws	If linkage is complete, the number of Monte Carlo iterations.

Value

- pvalue - The p-value for the difference of cluster means.
- stat - The test statistic.
- stdrr - If linkage is `complete`, the Monte Carlo standard error.
- hcl - The partition of the `n` observations retrieved by the clustering algorithm.

References

[1] L. L. Gao, J. Bien, and D. Witten. Selective inference for hierarchical clustering. Journal of the American Statistical Association, 0(0):1–11, 2022.

Examples

```
n <- 50
p <- 20
M <- Matrix::Matrix(0, nrow = n , ncol = p) # Mean matrix
Sigma <- stats::toeplitz(seq(1, 0.1, length = p)) # Sigma: dependence between features
U <- matrixNormal::I(n) # U: dependence between observations
X <- matrixNormal::rmatnorm(s = 1, M, U, Sigma)
Y <- matrixNormal::rmatnorm(s = 1, M, U, Sigma) # i.i.d. copy of X

# HAC with average linkage under the global null hypothesis
test.clusters.hc(X, U, Sigma, NC = 3, clusters = sample(1:3, 2), linkage = "average")
# HAC with complete linkage under the global null hypothesis and over-estimation of Sigma
test.clusters.hc(X, U, Sigma = NULL, Y = Y, NC = 3, clusters = sample(1:3, 2), linkage = "complete")
```

`test.clusters.km` *Test for the difference of two cluster means after k-means clustering, for matrix normal model with arbitrary scale matrices.*

Description

Test for the difference of two cluster means after k-means clustering, for matrix normal model with arbitrary scale matrices.

Usage

```
test.clusters.km(
  X,
  U = NULL,
  Sigma = NULL,
  Y = NULL,
  UY = NULL,
  precUY = NULL,
  NC,
  clusters,
```

```
  itermax = 10,
  tol = 1e-06
)
```

Arguments

X	A $n \times p$ matrix drawn from a $n \times p$ matrix normal distribution $\mathcal{MN}$(M, U, Sigma). X must have n rows and p columns.
U	A $n \times n$ positive-definite matrix describing the dependence structure between the rows in X. If NULL, observations are considered independent and U is set to the $n \times n$ identity matrix.
Sigma	A $p \times p$ positive-definite matrix describing the dependence structure between the columns in X. If NULL, Sigma is over-estimated (in the sense of the Loewner partial order).
Y	If Sigma is NULL, an i.i.d. copy of X allowing its estimation. Y must have the same number of columns as X.
UY	If Sigma is NULL, a positive-definite matrix describing the dependence structure between the rows in Y. If NULL and its inverse is not provided, set to the identity matrix by default.
precUY	The inverse matrix of UY, that can be provided to increase computational efficiency. If UY is not NULL and precUY is NULL, precUY is obtained by inverting UY.
NC	The number of clusters to choose.
clusters	A vector of two integers from 1 to NC indicating the pair of clusters whose means have to be compared.
itermax	The iter.max parameter of the k-means algorithm in kmeans_estimation function of KmeansInference package.
tol	The tol_eps parameter of the k-means algorithm in kmeans_estimation function of KmeansInference package.

Value

- pvalue - The p-value for the difference of cluster means.
- stat - The test statistic.
- km - The partition of the n observations retrieved by the clustering algorithm.

References

[1] L. L. Gao, J. Bien, and D. Witten. Selective inference for hierarchical clustering. Journal of the American Statistical Association, 0(0):1–11, 2022. [2] Y. T. Chen and D. M. Witten. Selective inference for k-means clustering, 2022. arXiv:2203.15267.

Examples

```
n <- 50
p <- 20
M <- Matrix::Matrix(0, nrow = n , ncol = p) # Mean matrix
Sigma <- stats::toeplitz(seq(1, 0.1, length = p)) # Sigma: dependence between features
U <- matrixNormal::I(n) # U: dependence between observations
X <- matrixNormal::rmatnorm(s = 1, M, U, Sigma)
Y <- matrixNormal::rmatnorm(s = 1, M, U, Sigma) # i.i.d. copy of X

# k-means under the global null hypothesis
test.clusters.km(X, U, Sigma, NC = 3, clusters = sample(1:3, 2))
# k-means under the global null hypothesis and over-estimation of Sigma
test.clusters.km(X, U, Sigma = NULL, Y = Y, NC = 3, clusters = sample(1:3, 2))
```

`test.clusters.MC` *Test for the difference of two cluster means after any clustering algorithm, for matrix normal model with arbitrary scale matrices.*

Description

Test for the difference of two cluster means after any clustering algorithm, for matrix normal model with arbitrary scale matrices.

Usage

```
test.clusters.MC(
  X,
  U = NULL,
  Sigma = NULL,
  Y = NULL,
  UY = NULL,
  precUY = NULL,
  clusters,
  cl_fun,
  NC = NULL,
  cl = NULL,
  ndraws = 2000
)
```

Arguments

`X`	A $n \times p$ matrix drawn from a $n \times p$ matrix normal distribution $\mathcal{MN}(\texttt{M}, \texttt{U}, \texttt{Sigma})$. `X` must have n rows and p columns.
`U`	A $n \times n$ positive-definite matrix describing the dependence structure between the rows in `X`. If `NULL`, observations are considered independent and `U` is set to the $n \times n$ identity matrix.

Sigma	A $p \times p$ positive-definite matrix describing the dependence structure between the columns in X. If NULL, Sigma is over-estimated (in the sense of the Loewner partial order).
Y	If Sigma is NULL, an i.i.d. copy of X allowing its estimation. Y must have the same number of columns as X.
UY	If Sigma is NULL, a positive-definite matrix describing the dependence structure between the rows in Y. If NULL and its inverse is not provided, set to the identity matrix by default.
precUY	The inverse matrix of UY, that can be provided to increase computational efficiency. If UY is not NULL and precUY is NULL, precUY is obtained by inverting UY.
clusters	A vector of two integers from 1 to NC indicating the pair of clusters whose means have to be compared.
cl_fun	A function returning assignments to clusters. The function must take as input the data matrix X and the number of clusters NC.
NC	The number of clusters to choose, that will be passed as argument to cl_fun. Must be set to NULL if not required by cl_fun.
cl	The result of clustering X using cl_fun. It can useful to precompute this quantity before choosing clusters.
ndraws	The number of Monte Carlo iterations.

Value

- pvalue - The p-value for the difference of cluster means.
- stat - The test statistic.
- stdrr - The Monte Carlo standard error.
- clusters - The partition of the n observations retrieved by the clustering algorithm.

References

[1] L. L. Gao, J. Bien, and D. Witten. Selective inference for hierarchical clustering. Journal of the American Statistical Association, 0(0):1–11, 2022.

Examples

```
n <- 50
p <- 20
M <- Matrix::Matrix(0, nrow = n , ncol = p) # Mean matrix
Sigma <- stats::toeplitz(seq(1, 0.1, length = p)) # Sigma: dependence between features
U <- matrixNormal::I(n) # U: dependence between observations
X <- matrixNormal::rmatnorm(s = 1, M, U, Sigma)
Y <- matrixNormal::rmatnorm(s = 1, M, U, Sigma) # i.i.d. copy of X

# Using HDBSCAN clustering from dbscan package. This algorithm selects
# automatically the number of clusters NC.
# Additional clustering parameters must be set as default values
# when defining cl_fun.
```

```
# install.packages('dbscan')

hdbscan.clustering <- function(X, NC = NULL, min.occupancy = 5){

 X.clus <- dbscan::hdbscan(X, minPts = min.occupancy)
 return(X.clus$cluster + 1)

}

# We start by clustering the data
clusters_X <- hdbscan.clustering(X)
# We test for the equality of clusters 3 and 1
test.clusters.MC(X, U = U, Sigma = Sigma, clusters = c(3,1),
 cl = clusters_X, cl_fun = hdbscan.clustering, NC = NULL, ndraws = 500)
```

Index