Package 'didec'

August 26, 2024

Type Package

Title Directed Dependence Coefficient

Version 0.1.0

Maintainer Yuping Wang <yuping.wang@plus.ac.at>

Description Directed Dependence Coefficient (didec) is a measure of directed dependence. Multivariate Feature Ordering by Conditional Independence (MFOCI) is a variable selection algorithm based on didec. Hierarchical Variable Clustering (VarClustPartition) is a variable clustering method based on didec. For more information, see the paper by Ansari and Fuchs (2024, <doi:10.48550/arXiv.2212.01621>), and the paper by Fuchs and Wang (2024, <doi:10.1016/j.ijar.2024.109185>).

License MIT + file LICENSE

Encoding UTF-8

LazyData true

RoxygenNote 7.3.0

Imports copBasic (>= 2.2.3), cowplot (>= 1.1.2), dendextend (>= 1.17.1), factoextra (>= 1.0.7), FOCI (>= 0.1.3), ggplot2 (>= 3.4.4), graphics (>= 4.3.0), grDevices (>= 0.5-1), gtools (>= 3.9.5), phylogram (>= 2.1.0), rlang (>= 1.1.4), stats (>= 4.3.0)

Depends R (>= 2.10)

NeedsCompilation no

Author Yuping Wang [aut, cre], Sebastian Fuchs [aut], Jonathan Ansari [aut]

Repository CRAN

Date/Publication 2024-08-26 15:40:05 UTC

Contents

bioclimatic 2
didec 2
mfoci 4
VarClustPartition 5

Index **8**

`bioclimatic` *Bioclimatic variables*

Description

A data set of bioclimatic variables for $n = 1,862$ locations homogeneously distributed over the global landmass from CHELSA("Climatologies at high resolution for the earth's land surface areas").

Usage

```
bioclimatic
```

Format

An object of class `data.frame` with 1862 rows and 14 columns.

References

D.N. Karger, O. Conrad, J. Böhner, T. Kawohl, H. Kreft, R.W. Soria-Auza, N.E. Zimmermann, H.P. Linder, M. Kessler, Climatologies at high resolution for the Earth's land surface areas, Sci. Data 4(1), 2017.

Examples

```
head(bioclimatic)
```

`didec` *Computes the directed dependence coefficient.*

Description

The directed dependence coefficient (`didec`) estimates the degree of directed dependence of a random vector Y on a random vector X, based on an i.i.d. sample of (X,Y).

Usage

```
didec(X, Y, perm = FALSE, perm.method = c("decreasing"))
```

Arguments

X	A numeric matrix or data.frame/data.table. Contains the predictor vector X.
Y	A numeric matrix or data.frame/data.table. Contains the response vector Y.
perm	A logical. If True a version of didec is computed that takes into account the permutations (specified by perm.method) of the response variables.
perm.method	An optional character string specifying a method for permuting the response variables. This must be one of the strings "sample", "increasing", "decreasing" (default) or "full". The version "full" is invariant with respect to permutations of the response variables.

Details

The directed dependence coefficient (didec) is an extension of Azadkia & Chatterjee's measure of directed dependence (Azadkia & Chatterjee, 2021) to a vector of response variables introduced in (Ansari & Fuchs, 2023). Its calculation is based on the function codec which estimates Azadkia & Chatterjee's measure of directed dependence and is provided in the R package FOCI.

By definition, didec is invariant with respect to permutations of the variables within the predictor vector X. Invariance with respect to permutations within the response vector Y is achieved by computing the arithmetic mean over all possible (or chosen) permutations. In addition to the option "full" of running all $q!$ permutations of $(1, ..., q)$, less computationally intensive options are also available (here, q denotes the number of response variables): a random selection of q permutations "sample", cyclic permutations such as $(1, 2, ..., q)$, $(2, ..., q, 1)$ either "increasing" or "decreasing". Note that when the number of variables q is large, choosing "full" may result in long computation times.

Value

The degree of directed dependence of the random vector Y on the random vector X.

Author(s)

Yuping Wang, Sebastian Fuchs, Jonathan Ansari

References

M. Azadkia, S. Chatterjee, A simple measure of conditional dependence, Ann. Stat. 49 (6), 2021.

J. Ansari, S. Fuchs, A simple extension of Azadkia & Chatterjee's rank correlation to multi-response vectors, Available at https://arxiv.org/abs/2212.01621, 2024.

mfoci *Multivariate feature ordering by conditional independence.*

Description

A variable selection algorithm based on the directed dependence coefficient (didec).

Usage

```
mfoci(
  X,
  Y,
  pre.selected = NULL,
  perm = FALSE,
  perm.method = c("decreasing"),
  autostop = TRUE
)
```

Arguments

X	A numeric matrix or data.frame/data.table. Contains the predictor vector X.
Y	A numeric matrix or data.frame/data.table. Contains the response vector Y.
pre.selected	An integer vector for indexing pre-selected predictor variables from X.
perm	A logical. If True a version of didec is computed that takes into account the permutations (specified by perm.method) of the response variables.
perm.method	An optional character string specifying a method in didec for permuting the response variables. This must be one of the strings "sample", "increasing", "decreasing" (default) or "full". The version "full" is invariant with respect to permutations of the response variables.
autostop	A logical. If True the algorithm stops at the first non-increasing value of didec.

Details

mfoci is a forward feature selection algorithm for multiple-outcome data that employs the directed dependence coefficient (didec) at each step. mfoci is proved to be consistent in the sense that the subset of predictor variables selected via mfoci is sufficient with high probability.

If autostop == TRUE the algorithm stops at the first non-increasing value of didec, thereby selecting a subset of variables. Otherwise, all predictor variables are ordered according to their predictive strength measured by didec.

Value

A data.frame listing the selected variables.

Author(s)

Sebastian Fuchs, Jonathan Ansari, Yuping Wang

References

J. Ansari, S. Fuchs, A simple extension of Azadkia & Chatterjee's rank correlation to multi-response vectors, Available at https://arxiv.org/abs/2212.01621, 2024.

Examples

```
library(didec)
data("bioclimatic")
X <- bioclimatic[, c(9:12)]
Y <- bioclimatic[, c(1,8)]
mfoci(X, Y, pre.selected = c(1, 3))
```

`VarClustPartition` *Hierarchical variable clustering.*

Description

`VarClustPartition` is a hierarchical variable clustering algorithm based on the directed dependence coefficient (`didec`) or a concordance measure (Kendall tau τ or Spearman's footrule) according to a pre-selected number of clusters or an optimality criterion (Adiam&Msplit or Silhouette coefficient).

Usage

```
VarClustPartition(
  X,
  dist.method = c("PD"),
  linkage = FALSE,
  link.method = c("complete"),
  part.method = c("optimal"),
  criterion = c("Adiam&Msplit"),
  num.cluster = NULL,
  plot = FALSE
)
```

Arguments

`X`	A numeric matrix or data.frame/data.table. Contains the variables to be clustered.
`dist.method`	An optional character string computing a distance function for clustering. This must be one of the strings `"PD"` (default), `"MPD"`, `"kendall"` or `"footrule"`.
`linkage`	A logical. If `TRUE` a linkage method is used.

link.method	An optional character string selecting a linkage method. This must be one of the strings "complete" (default), "average" or "single".
part.method	An optional character string selecting a partitioning method. This must be one of the strings "optimal" (default) or "selected".
criterion	An optional character string selecting a criterion for the optimal partition, if part.method = "optimal". This must be one of the strings "Adiam&Msplit" (default) or "Silhouette".
num.cluster	An integer value for the selected number of clusters, if part.method = "selected".
plot	A logical. If TRUE a dendrogram is plotted with colored branches according to the corresponding partitioning method.

Details

VarClustPartition performs a hierarchical variable clustering based on the directed dependence coefficient (didec) and provides a partition of the set of variables.

If dist.method =="PD" or dist.method =="MPD", the clustering is performed using didec either as a directed ("PD") or as a symmetric ("MPD") dependence coefficient. If dist.method =="kendall" or dist.method =="footrule", clustering is performed using either multivariate Kendall's tau ("kendall") or multivariate Spearman's footrule ("footrule").

Instead of using one of the above-mentioned four multivariate measures for the clustering, the option linkage == TRUE enables the use of bivariate linkage methods, including complete linkage (link.method == "complete"), average linkage (link.method == "average") and single linkage (link.method == "single"). Note that the multivariate distance methods are computationally demanding because higher-dimensional dependencies are included in the calculation, in contrast to linkage methods which only incorporate pairwise dependencies.

A pre-selected number of clusters num.cluster can be realized with the option part.method == "selected". Otherwise (part.method == "optimal"), the number of clusters is determined by maximizing the intra-cluster similarity (similarity within the same cluster) and minimizing the inter-cluster similarity (similarity among the clusters). Two optimality criteria are available:

"Adiam&Msplit": *Adiam* measures the intra-cluster similarity and *Msplit* measures the inter-cluster similarity.

"Silhouette": A mixed coefficient incorporating the intra-cluster similarity and the inter-cluster similarity. The optimal number of clusters corresponds to the maximum Silhouette coefficient.

Value

A list containing a dendrogram without colored branches (**dendrogram**), an integer value determining the number of clusters after partitioning (**num.cluster**), and a list containing the clusters after partitioning (**clusters**).

Author(s)

Yuping Wang, Sebastian Fuchs

References

S. Fuchs, Y. Wang, Hierarchical variable clustering based on the predictive strength between random vectors, Int. J. Approx. Reason. 170, Article ID 109185, 2024.

P. Hansen, B. Jaumard, Cluster analysis and mathematical programming, Math. Program. 79 (1) 191–215, 1997.

L. Kaufman, Finding Groups in Data, John Wiley & Sons, 1990.

Examples

```
library(didec)
n  <- 50
X1 <- rnorm(n,0,1)
X2 <- X1
X3 <- rnorm(n,0,1)
X4 <- X3 + X2
X  <- data.frame(X1=X1,X2=X2,X3=X3,X4=X4)
vcp <- VarClustPartition(X,
                           dist.method = c("PD"),
                           part.method = c("optimal"),
                           criterion   = c("Silhouette"),
                           plot        = TRUE)
vcp$clusters

data("bioclimatic")
X   <- bioclimatic[c(2:4,9)]
vcp1 <- VarClustPartition(X,
                          linkage     = TRUE,
                          link.method = c("complete"),
                          dist.method = "PD",
                          part.method = "optimal",
                          criterion   = "Silhouette",
                          plot        = TRUE)
vcp1$clusters
vcp2 <- VarClustPartition(X,
                          linkage     = TRUE,
                          link.method = c("complete"),
                          dist.method = "footrule",
                          part.method = "optimal",
                          criterion   = "Adiam&Msplit",
                          plot        = TRUE)
vcp2$clusters
```

Index

∗ **datasets**
bioclimatic, 2

bioclimatic, 2

didec, 2, *4–6*

mfoci, 4

VarClustPartition, 5