Package ‘cdcsis’

August 23, 2024

Type Package

Title Conditional Distance Correlation Based Feature Screening and Conditional Independence Inference

Description Conditional distance correlation <doi:10.1080/01621459.2014.993081> is a novel conditional dependence measurement of two multivariate random variables given a confounding variable. This package provides conditional distance correlation, performs the conditional distance correlation sure independence screening procedure for ultrahigh dimensional data <https://www3.stat.sinica.edu.tw/statistica/J28N1/J28N114/J28N114.html>, and conducts conditional distance covariance test for conditional independence assumption of two multivariate variable.

Version 2.0.5

Date 2024-08-22

Depends R(>= 3.0.1)

Imports ks (>= 1.8.0), mvtnorm, utils, Rcpp

Suggests testthat

Maintainer Canhong Wen <wench@ustc.edu.cn>

License GPL (>= 2)

NeedsCompilation yes

Encoding UTF-8

RoxygenNote 7.2.3

LinkingTo Rcpp

URL https://github.com/Mamba413/cdcsis

BugReports https://github.com/Mamba413/cdcsis/issues

Author Wenhao Hu [aut],
Mian Huang [aut],
Wenliang Pan [aut],
Xueqin Wang [aut] (<https://orcid.org/0000-0001-5205-9950>),
Canhong Wen [aut, cre],
Yuan Tian [aut],
Heping Zhang [aut] (<https://orcid.org/0000-0002-0688-4076>),
Jin Zhu [aut] (<https://orcid.org/0000-0001-8550-5822>)

Repository CRAN

Date/Publication 2024-08-23 12:00:02 UTC

Contents

`cdcsis-package` *Conditional Distance Correlation Based Feature Screening and Conditional Independence Inference*

Description

Conditional distance correlation <doi:10.1080/01621459.2014.993081> is a novel conditional dependence measurement of two multivariate random variables given a confounding variable. This package provides conditional distance correlation, performs the conditional distance correlation sure independence screening procedure for ultrahigh dimensional data <doi:10.5705/ss.202014.0117>, and conducts conditional distance covariance test for conditional independence assumption of two multivariate variable.

Author(s)

Wenhao Hu, Mian Huang, Wenliang Pan, Xueqin Wang, Canhong Wen, Yuan Tian, Heping Zhang, Jin Zhu Maintainer: Canhong Wen <wench@ustc.edu.cn>

References

Wang, X., Pan, W., Hu, W., Tian, Y. and Zhang, H., 2015. Conditional distance correlation. Journal of the American Statistical Association, 110(512), pp.1726-1734.

Wen, C., Pan, W., Huang, M. and Wang, X., 2018. Sure independence screening adjusted for confounding covariates with ultrahigh-dimensional data. Statistica Sinica, 28, pp.293-317. URL http://www3.stat.sinica.edu.tw/statistica/J28N1/28-1.html

cdcov *Conditional Distance Covariance/Correlation Statistics*

Description

Computes conditional distance covariance and conditional distance correlation statistics, which are multivariate measures of conditional dependence.

Usage

```
cdcov(x, y, z, width, index = 1, distance = FALSE)

cdcor(x, y, z, width, index = 1, distance = FALSE)
```

Arguments

`x`	a numeric vector, matrix, or `dist` object
`y`	a numeric vector, matrix, or `dist` object
`z`	`z` is a numeric vector or matrix. It is the variable being conditioned.
`width`	a user-specified positive value (univariate conditional variable) or vector (multivariate conditional variable) for gaussian kernel bandwidth. Its default value is relies on `stats::bw.nrd0`.
`index`	exponent on Euclidean distance, in $(0, 2]$
`distance`	if `distance = TRUE`, `x` and `y` will be considered as distance matrices. Default: `distance = FALSE`.

Details

`cdcov` and `cdcor` compute conditional distance covariance and conditional distance correlation statistics. The sample sizes (number of rows or length of the vector) of the two variables must agree, and samples must not contain missing values. If we set `distance = TRUE`, arguments `x`, `y` can be a `dist` object recording distance between samples; otherwise, these arguments are treated as multivariate data.

Value

`cdcov`	conditional distance covariance test statistic.
`cdcor`	conditional distance correlation statistic.
`cdc`	conditional distance covariance/correlation vector.

Author(s)

Canhong Wen, Wenliang Pan, and Xueqin Wang

References

Wang, X., Pan, W., Hu, W., Tian, Y. and Zhang, H., 2015. Conditional distance correlation. Journal of the American Statistical Association, 110(512), pp.1726-1734.

See Also

cdcor

Examples

```
library(cdcsis)

############# Conditional Distance Covariance #############
set.seed(1)
x <- rnorm(25)
y <- rnorm(25)
z <- rnorm(25)
cdcov(x, y, z)
############# Conditional Distance Correlation #############
num <- 25
set.seed(1)
x <- rnorm(num)
y <- rnorm(num)
z <- rnorm(num)
cdcor(x, y, z)
```

---

cdcov.test *Conditional Distance Covariance Independence Test*

Description

Performs the nonparametric conditional distance covariance test for conditional independence assumption

Usage

```
cdcov.test(
  x,
  y,
  z,
  num.bootstrap = 99,
  width,
  distance = FALSE,
  index = 1,
  seed = 1,
  num.threads = 1
)
```

Arguments

x	a numeric vector, matrix, or dist object
y	a numeric vector, matrix, or dist object
z	z is a numeric vector or matrix. It is the variable being conditioned.
num.bootstrap	the number of local bootstrap procedure replications. Default: num.bootstrap = 99.
width	a user-specified positive value (univariate conditional variable) or vector (multivariate conditional variable) for gaussian kernel bandwidth. Its default value is relies on stats::bw.nrd0 function when conditional variable is univariate, ks::Hpi.diag when conditional variable with at most trivariate, and stats::bw.nrd on the other cases.
distance	if distance = TRUE, x and y will be considered as distance matrices. Default: distance = FALSE.
index	exponent on Euclidean distance, in $(0, 2]$
seed	the random seed
num.threads	number of threads. Default num.threads = 1.

Value

cdcov.test returns a list with class "htest" containing the following components:

statistic	conditional distance covariance statistic.
p.value	the p-value for the test.
replicates	the number of local bootstrap procedure replications.
size	sample sizes.
alternative	a character string describing the alternative hypothesis.
method	a character string indicating what type of test was performed.
data.name	description of data.

References

Wang, X., Pan, W., Hu, W., Tian, Y. and Zhang, H., 2015. Conditional distance correlation. Journal of the American Statistical Association, 110(512), pp.1726-1734.

See Also

cdcov

Examples

```
library(cdcsis)
set.seed(1)
num <- 50
################# Conditional Independent #################
```

```
## Case 1:
cov_mat <- matrix(c(1, 0.36, 0.6, 0.36, 1, 0.6, 0.6, 0.6, 1), nrow = 3)
dat <- mvtnorm::rmvnorm(n = num, sigma = cov_mat)
x <- dat[, 1]
y <- dat[, 2]
z <- dat[, 3]
cdcov.test(x, y, z)
## Case 2:
z <- rnorm(num)
x <- 0.5 * (z^3 / 7 + z / 2) + tanh(rnorm(num))
x <- x + x^3 / 3
y <- (z^3 + z) / 3 + rnorm(num)
y <- y + tanh(y / 3)
cdcov.test(x, y, z, num.bootstrap = 99)

################# Conditional Dependent #################
## Case 3:
cov_mat <- matrix(c(1, 0.7, 0.6, 0.7, 1, 0.6, 0.6, 0.6, 1), nrow = 3)
dat <- mvtnorm::rmvnorm(n = num, sigma = cov_mat)
x <- dat[, 1]
y <- dat[, 2]
z <- dat[, 3]
cdcov.test(x, y, z, width = 0.5)
## Case 4:
z <- matrix(rt(num * 4, df = 2), nrow = num)
x <- z
y <- cbind(sin(z[, 1]) + cos(z[, 2]) + (z[, 3])^2 + (z[, 4])^2,
           (z[, 1])^2 + (z[, 2])^2 + z[, 3] + z[, 4])
z <- z[, 1:2]
cdcov.test(x, y, z, seed = 2)

################# Distance Matrix Input #################
x <- dist(x)
y <- dist(y)
cdcov.test(x, y, z, seed = 2, distance = TRUE)
```

`cdcsis` *Conditional Distance Correlation Sure Independence Screening (CDC-SIS)*

Description

Performs conditional distance correlation sure independence screening (CDC-SIS).

Usage

```
cdcsis(
  x,
  y,
```

```
  z = NULL,
  width,
  threshold = nrow(y),
  distance = FALSE,
  index = 1,
  num.threads = 1
)
```

Arguments

`x`	a numeric matrix, or a list which contains multiple numeric matrix
`y`	a numeric vector, matrix, or `dist` object
`z`	`z` is a numeric vector or matrix. It is the variable being conditioned.
`width`	a user-specified positive value (univariate conditional variable) or vector (multivariate conditional variable) for gaussian kernel bandwidth. Its default value is relies on `stats::bw.nrd0` function when conditional variable is univariate, `ks::Hpi.diag` when conditional variable with at most trivariate, and `stats::bw.nrd` on the other cases.
`threshold`	the threshold of the number of predictors recuited by CDC-SIS. Should be less than or equal than the number of column of `x`. Default value `threshold` is sample size.
`distance`	if `distance = TRUE`, only `y` will be considered as distance matrices. Default: `distance = FALSE`
`index`	exponent on Euclidean distance, in $(0, 2]$
`num.threads`	number of threads. Default `num.threads = 1`.

Value

`ix`	the vector of indices selected by CDC-SIS
`cdcor`	the conditional distance correlation for each univariate/multivariate variable in `x`

Author(s)

Canhong Wen, Wenliang Pan, Mian Huang, and Xueqin Wang

References

Wen, C., Pan, W., Huang, M. and Wang, X., 2018. Sure independence screening adjusted for confounding covariates with ultrahigh-dimensional data. Statistica Sinica, 28, pp.293-317. URL http://www3.stat.sinica.edu.tw/statistica/J28N1/J28N114/J28N114.html

See Also

`cdcor`

Examples

```
## Not run:

library(cdcsis)

########## univariate explanative variables ##########
set.seed(1)
num <- 100
p <- 150
x <- matrix(rnorm(num * p), nrow = num)
z <- rnorm(num)
y <- 3 * x[, 1] + 1.5 * x[, 2] + 4 * z * x[, 5] + rnorm(num)
res <- cdcsis(x, y, z)
head(res[["ix"]], n = 10)

########## multivariate explanative variables ##########
x <- as.list(as.data.frame(x))
x <- lapply(x, as.matrix)
x[[1]] <- cbind(x[[1]], x[[2]])
x[[2]] <- NULL
res <- cdcsis(x, y, z)
head(res[["ix"]], n = 10)

########## multivariate response variables ##########
num <- 100
p <- 150
x <- matrix(rnorm(num * p), nrow = num)
z <- rnorm(num)
y1 <- 3 * x[, 1] + 5 * z * x[, 4] + rnorm(num)
y2 <- 3 * x[, 2] + 5 * x[, 3] + 2 * z + rnorm(num)
y <- cbind(y1, y2)
res <- cdcsis(x, y, z)
head(res[["ix"]], n = 10)

## End(Not run)
```

Index