

Package ‘GiniDistance’

June 15, 2019

Type Package

Title A New Gini Correlation Between Quantitative and Qualitative Variables

Version 0.1.0

Author Dao Nguyen and Xin Dang

Maintainer Dao Nguyen <dxnguyen@go.olemiss.edu>

Description An implementation of a new Gini covariance and correlation to measure dependence between a categorical and numerical variables.

Depends R(>= 3.0.0)

Imports Rcpp (>= 1.0.0), energy, readxl, randomForest

LinkingTo Rcpp, RcppArmadillo

License GPL (>= 2)

Encoding UTF-8

LazyData true

Collate GiniDistance-package.R RcppExports.R gmd.R gCov.R gCor.R
dCov.R dCor.R Kgmd.R KgCov.R KgCor.R KdCov.R KdCor.R
ConfidenceInterval.R PermutationTest.R utils.R

RoxygenNote 6.1.0

NeedsCompilation yes

R topics documented:

GiniDistance-package	2
ConfidenceInterval	3
CriticalValue	4
dCor	5
dCov	6
gCor	8
gCov	9
gmd	10
KdCor	12

KdCov	13
KgCor	15
KgCov	16
Kgmd	17
PermutationTest	18
RcppgCor	19
RcppgCov	20
RcppGmd	21
RcppKgCor	22
RcppKgCov	23
RcppKGmd	24
Index	25

GiniDistance-package *GiniDistance*

Description

A new Gini correlation to measure dependence between categorical and numerical variables are implemented. Analogous to Pearson in ANOVA model, the Gini correlation is interpreted as the ratio of the between-group variation and the total variation, but it characterizes independence (zero Gini correlation mutually implies independence). Closely related to the distance correlation, the Gini correlation is of the simple formulation by considering the nature of the categorical variable. As a result, the Gini correlation has a lower computational cost than the distance correlation and is more straightforward to perform inference. The dependence test and confidence interval are implemented. Also, the corresponding kernelized dependence measures are also implemented.

Details

The details are described in the following papers "A new Gini correlation between quantitative and qualitative variables" and "Estimating Feature-Label Dependence Using Gini Distance Statistics"

Author(s)

Dao Nguyen <dxnguyen@olemiss.edu> and Xin Dang <xdang@olemiss.edu>

References

- Dang, X., Nguyen, D., Chen, Y. and Zhang, J., (2019). A new Gini correlation between quantitative and qualitative variables, *Journal of the American Statistical Association* (submitted), <https://arxiv.org/pdf/1809.09793.pdf>
- Zhang, S., Dang, X., Nguyen, D. and Chen, Y. (2019). Estimating feature - label dependence using Gini distance statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (submitted), <https://arXiv.org/pdf/1906.02171.pdf>

ConfidenceInterval	<i>Confidence Interval of Dependence measure</i>
--------------------	--

Description

Find confidence intervals for dependence measures in which Xs are quantitative, Y are categorical using jack-knife method.

Usage

```
ConfidenceInterval(x, y, sigma, alpha, level, method)
```

Arguments

x	data
y	label of data or univariate response variable
sigma	kernel parameter
alpha	exponent on Euclidean distance, in (0,2]
level	level of confidence, in [0,1]
method	name of dependence measure which can chosen from "gCor", "gCov", "dCor", "dCov", "KgCor", "KgCov", "KdCor" and "KdCov"

Details

ConfidenceInterval compute the confidence interval of the distance correlation statistics. It is a self-contained R function returning a variance of the measure of dependence statistics.

The sample size (number of rows) of the data must agree with the length of the label vector, and samples must not contain missing values. Arguments x, y are treated as data and labels. alpha if missing by default is 1, otherwise it is exponent on the Euclidean distance.

Suppose a sample data $\mathcal{D} = \{(x_i, y_i)\}$ for $i = 1, \dots, n$ available. The confidence interval is built upon the asymptotic normality of sample dependence statistic. The asymptotic variance is estimated by the Jackknife method. More details refer to Shao and Tu (1996).

Value

ConfidenceInterval returns the confidence interval of distance correlation

References

- Dang, X., Nguyen, D., Chen, Y. and Zhang, J. (2019). A new Gini correlation between quantitative and qualitative variables. Submitted.
- Shao, J. and Tu, D. (1996). The Jackknife and Bootstrap. Springer, New York.

Examples

```
x <- iris[,1:4]
y <- unclass(iris[,5])
ConfidenceInterval(x, y, alpha=1, level=0.95, method='gCor')
```

CriticalValue	<i>Find a critical value by permutation test of dependence between X and Y using kernel (Gini) distance covariance or correlation statistics</i>
---------------	--

Description

Find a critical value by permutation test using variance of kernel (Gini) distance covariance or correlation statistics, in which Xs are quantitative, Y are categorical, sigma is kernel standard deviation, alpha is an exponent on Euclidean distance and returns the critical value of the measures of dependence.

Usage

```
CriticalValue(x, y, sigma, alpha, level, M = 1000, method)
```

Arguments

x	data
y	label of data or univariate response variable
sigma	kernel standard deviation
alpha	exponent on Euclidean distance, in (0,2]
level	significance level of the test, the default value = 0.05
M	number of permutations
method	string name of the method for permutation test, e.g. gCov

Details

CriticalValue compute the critical value of a dependence test of a kernel (Gini) distance covariance or correlation statistics. It is a self-contained R function returning the critical value of the measure of dependence statistics.

The critical value of the test of significance level γ , however, is obtained by a permutation procedure. Let $\nu = 1 : n$ be the vector of original sample indices of the sample for Y labels and $\hat{\rho}_g(\alpha) = \hat{\rho}(\nu; \alpha)$. Let $\pi(\nu)$ denote a permutation of the elements of ν and the corresponding $\hat{\rho}_g(\pi; \alpha)$ is computed. Under the \mathcal{H}_0 , $\hat{\rho}_g(\nu)$ and $\hat{\rho}_g(\pi; \alpha)$ are identically distributed for every permutation π of ν . Hence, based on M permutations, the critical value q_γ is estimated by the $(1 - \gamma)100\%$ sample quantile of $\hat{\rho}_g(\pi_m; \alpha)$, $m = 1, \dots, M$. Usually $100 \leq M \leq 1000$ is sufficient for a good estimation on the critical value.

See [PermutationTest](#) for a test of multivariate independence based on the (Gini) distance statistic.

Value

CriticalValue returns return the critical value of the measures of the dependence of the permutation test of a specified function

See Also

[PermutationTest](#)

Examples

```
n = 50
x <- runif(n)
y <- c(rep(1,n/2),rep(2,n/2))
CriticalValue(x, y, sigma=1, alpha=2, level=0.04, M = 1000, method='KgCov')
```

dCor

Distance Covariance and Correlation Statistics

Description

Computes distance covariance and correlation statistics, in which Xs are quantitative and Ys are categorical and return the measures of dependence.

Usage

```
dCor(x, y, alpha)
```

Arguments

x	data
y	label of data or univariate response variable
alpha	exponent on Euclidean distance, in (0,2]

Details

The sample size (number of rows) of the data must agree with the length of the label vector, and samples must not contain missing values. Arguments x, y are treated as data and labels.

dCor calls dcor function from energy package which computes the distance correlation between X and Y where both are numerical variables. If Y is categorical, the set difference metric on the support of Y is used. That is, $d(y, y') = |y - y'| := I(y \neq y')$, where $I(\cdot)$ is the indicator function. Then the sample distance correlation between data and labels is computed as follows.

Let $A = (a_{ij})$ be a symmetric, $n \times n$, centered distance matrix of sample $\mathbf{x}_1, \dots, \mathbf{x}_n$. The (i, j) -th entry of A is $a_{ij} - \frac{1}{n-2}a_{i.} - \frac{1}{n-2}a_{.j} + \frac{1}{(n-1)(n-2)}a_{..}$ if $i \neq j$ and 0 if $i = j$, where $a_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^\alpha$, $a_{i.} = \sum_{j=1}^n a_{ij}$, $a_{.j} = \sum_{i=1}^n a_{ij}$, and $a_{..} = \sum_{i,j=1}^n a_{ij}$. Similarly, using the set difference metric, a symmetric, $n \times n$, centered distance matrix is calculated for samples y_1, \dots, y_n and denoted by $B = (b_{ij})$. Unbiased estimators of $\text{dCov}(\mathbf{X}, \mathbf{Y}; \alpha)$, $\text{dCov}(\mathbf{X}, \mathbf{X}; \alpha)$ and $\text{dCov}(\mathbf{Y}, \mathbf{Y}; \alpha)$ are given

respectively as, $\frac{1}{n(n-3)} \sum_{i \neq j} A_{ij} B_{ij}$, $\frac{1}{n(n-3)} \sum_{i \neq j} A_{ij}^2$ and $\frac{1}{n(n-3)} \sum_{i \neq j} B_{ij}^2$. Then the distance correlation is

$$dCor(\mathbf{X}, Y; \alpha) = \frac{dCov(\mathbf{X}, Y, \alpha)}{\sqrt{dCov(\mathbf{X}, \mathbf{X}; \alpha)} \sqrt{dCov(Y, Y)}}.$$

Value

dCor returns the sample distance variance of x, distance variance of y, distance covariance of x and y and distance correlation of x, y.

References

Lyons, R. (2013). Distance covariance in metric spaces. The Annals of Probability, 41 (5), 3284-3305.

Szekely, G. J., Rizzo, M. L. and Bakirov, N. (2007). Measuring and testing dependence by correlation of distances. Annals of Statistics, 35 (6), 2769-2794.

Rizzo, M.L. and Szekely, G.J., (2017). Energy: E-Statistics: Multivariate Inference via the Energy of Data (R Package), Version 1.7-0.

See Also

[dCov](#) [KdCov](#) [KdCor](#)

Examples

```
x <- iris[,1:4]
y <- unclass(iris[,5])
dCor(x, y, alpha = 1)
```

dCov	<i>Distance Covariance Statistic</i>
------	--------------------------------------

Description

Computes distance covariance statistic, in which Xs are quantitative and Y are categorical and return the measures of dependence.

Usage

```
dCov(x, y, alpha)
```

Arguments

- | | |
|-------|--|
| x | data |
| y | label of data or response variable |
| alpha | exponent on Euclidean distance, in (0,2] |

Details

dCov calls `dcov` function from `energy` package to compute distance covariance statistic. The sample size (number of rows) of the data must agree with the length of the label vector, and samples must not contain missing values. Arguments `x`, `y` are treated as data and labels.

The distance covariance (Szekley07) is extended from Euclidean space to general metric spaces by Lyons (2013). Based on that idea, we define the discrete metric

$$d(y, y') = |y - y'| := I(y \neq y'),$$

where $I(\cdot)$ is the indicator function. Equipped with this set difference metric on the support of Y and Euclidean distance on the support of \mathbf{X} , the corresponding distance covariance and distance correlation for numerical \mathbf{X} and categorical Y variables are as follows.

Let $A = (a_{ij})$ be a symmetric, $n \times n$, centered distance matrix of sample $\mathbf{x}_1, \dots, \mathbf{x}_n$. The (i, j) -th entry of A is $a_{ij} - \frac{1}{n-2}a_{i.} - \frac{1}{n-2}a_{.j} + \frac{1}{(n-1)(n-2)}a_{..}$ if $i \neq j$ and 0 if $i = j$, where $a_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^\alpha$, $a_{i.} = \sum_{j=1}^n a_{ij}$, $a_{.j} = \sum_{i=1}^n a_{ij}$, and $a_{..} = \sum_{i,j=1}^n a_{ij}$. Similarly, using the set difference metric, a symmetric, $n \times n$, centered distance matrix is calculated for samples y_1, \dots, y_n and denoted by $B = (b_{ij})$. Unbiased estimators of $\text{dCov}(\mathbf{X}, \mathbf{Y}; \alpha)$ is

$$\frac{1}{n(n-3)} \sum_{i \neq j} A_{ij} B_{ij}.$$

Value

dCov returns the sample distance covariance between data `x` and label `y`.

References

- Lyons, R. (2013). Distance covariance in metric spaces. *The Annals of Probability*, 41 (5), 3284-3305.
- Rizzo, M.L. and Szekely, G.J., (2017). *Energy: E-Statistics: Multivariate Inference via the Energy of Data* (R Package), Version 1.7-0.
- Szekely, G. J., Rizzo, M. L. and Bakirov, N. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35 (6), 2769-2794.

See Also

[dCor](#) [KdCov](#) [KdCor](#)

Examples

```
x <- iris[,1:4]
y <- unclass(iris[,5])
dCov(x, y, alpha = 1)
```

gCor

*Gini Distance Covariance and Correlation Statistics***Description**

Computes Gini distance covariance and correlation statistics, in which Xs are quantitative, Y are categorical, alpha is exponent on the Euclidean distance and returns the measures of dependence.

Usage

```
gCor(x, y, alpha)
```

Arguments

x	data
y	label of data or univariate response variable
alpha	exponent on Euclidean distance, in (0,2)

Details

gCor compute Gini distance correlation statistics. It is a self-contained R function returning a measure of dependence statistics.

The sample size (number of rows) of the data must agree with the length of the label vector, and samples must not contain missing values. Arguments x, y are treated as data and labels. alpha if missing by default is 1, otherwise it is exponent on the Euclidean distance.

Suppose a sample data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$ for $i = 1, \dots, n$ available. The sample counterparts can be easily computed. Let \mathcal{I}_k be the index set of sample points with $y_i = L_k$, then p_k is estimated by the sample proportion of that category, that is, $\hat{p}_k = \frac{n_k}{n}$ where n_k is the number of elements in \mathcal{I}_k . With a given $\alpha \in (0, 2)$, a point estimator of $\rho_g(\alpha)$ is given as follows.

$$\hat{\Delta}_k(\alpha) = \binom{n_k}{2}^{-1} \sum_{i < j \in \mathcal{I}_k} \|\mathbf{x}_i - \mathbf{x}_j\|^\alpha,$$

$$\hat{\Delta}(\alpha) = \binom{n}{2}^{-1} \sum_{1 \leq i < j = n} \|\mathbf{x}_i - \mathbf{x}_j\|^\alpha,$$

$$gCor = \hat{\rho}_g(\alpha) = 1 - \frac{\sum_{k=1}^K \hat{p}_k \hat{\Delta}_k(\alpha)}{\hat{\Delta}(\alpha)}.$$

Value

gCor returns the sample Gini distance covariance and correlation between x and y.

References

Dang, X., Nguyen, D., Chen, Y. and Zhang, J. (2019). A new Gini correlation between quantitative and qualitative variables. Submitted to Journal of American Statistics Association.

See Also

[gmd](#) [gCov](#) [KgCov](#) [KgCor](#)

Examples

```
x <- iris[,1:4]
y <- unclass(iris[,5])
gCor(x, y, alpha = 1)
```

gCov

Gini Distance Covariance Statistics

Description

Computes Gini distance covariance statistics, in which X s are quantitative, Y are categorical, α is an exponent on Euclidean distance and returns the measures of dependence.

Usage

```
gCov(x, y, alpha)
```

Arguments

x	data
y	label of data or univariate response variable
α	exponent on Euclidean distance, in $(0,2]$

Details

`gCov` compute Gini distance covariance statistics. It is a self-contained R function returning a measure of dependence statistics.

The sample size (number of rows) of the data must agree with the length of the label vector, and samples must not contain missing values. Arguments x , y are treated as data and labels. α if missing by default is 1, otherwise it is exponent on the Euclidean distance.

Gini distance covariance is a new measure of dependence between random vectors and its labels. For all distributions with finite first moments, Gini distance correlation `gCov` has the following fundamental properties:

- (1) $\text{gCov}(X,Y)$ is defined for X in arbitrary dimension quantitative variable and Y a univariate categorical variable.
- (2) $\text{gCov}(X,Y)=0$ characterizes independence of X and Y .

Gini distance covariance satisfies $0 \leq \text{gCov}(X,Y)$, and $\text{gCov} = 0$ only if X and Y are independent. Gini distance covariance `gCov` provides a new approach to the problem of testing the joint independence of random vectors. The formal definitions of the population coefficients `gCov` is given in (DNCZ 2018). The empirical Gini distance covariance $\text{gCov}_n(X,Y;\alpha)$ is the nonnegative number computed as follows.

Suppose a sample data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$ for $i = 1, \dots, n$ available. The sample counterparts can be easily computed. Let \mathcal{I}_k be the index set of sample points with $y_i = L_k$, then p_k is estimated by the sample proportion of that category, that is, $\hat{p}_k = \frac{n_k}{n}$ where n_k is the number of elements in \mathcal{I}_k . With a given $\alpha \in (0, 2)$, a point estimator of $\rho_g(\alpha)$ is given as follows.

$$\hat{\Delta}_k(\alpha) = \binom{n_k}{2}^{-1} \sum_{i < j \in \mathcal{I}_k} \|\mathbf{x}_i - \mathbf{x}_j\|^\alpha,$$

$$\hat{\Delta}(\alpha) = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \|\mathbf{x}_i - \mathbf{x}_j\|^\alpha,$$

$$gCov = \hat{\Delta}(\alpha) - \sum_{k=1}^K \hat{p}_k \hat{\Delta}_k(\alpha).$$

Value

gCov returns the sample Gini distance covariance

References

Dang, X., Nguyen, D., Chen, Y. and Zhang, J., (2019). A new Gini correlation between quantitative and qualitative variables, *Journal of the American Statistical Association* (submitted), <https://arxiv.org/pdf/1809.09793.pdf>

See Also

[gCor](#) [gmd](#) [KgCov](#) [KgCor](#)

Examples

```
x <- iris[,1:4]
y <- unclass(iris[,5])
gCov(x, y, alpha = 1)
```

gmd

Gini Mean Difference

Description

Computes Gini mean difference of x, where alpha is an exponent on the Euclidean distance and return the Gini mean difference. The default value for alpha is 1.

Usage

```
gmd(x, alpha)
```

Arguments

x	data
alpha	exponent on Euclidean distance, in (0,2)

Details

gmd compute Gini mean difference of data. It is a self-contained R function dealing with both univariate and multivariate data.

The samples must not contain missing values. alpha if missing by default is 1, otherwise it is exponent on the Euclidean distance.

Gini mean difference (GMD) was originally introduced as an alternative measure of variability to the usual standard deviation (*Gini14*, *Yitzhaki13*). Let X and X' be independent random variables from a univariate distribution F with finite first moment in R . The GMD of F is

$$\Delta = \Delta(X) = \Delta(F) = E|X - X'|,$$

the expected distance between two independent random variables. If the sample data $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ is available, the sample Gini mean difference is calculated by

$$\hat{\Delta} = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} |x_i - x_j| = \binom{n}{2}^{-1} \sum_{i=1}^n (2i - n - 1)x_{(i)},$$

where $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ are the order statistics of \mathbf{x} (*Schechtman87*). The computation complexity for univariate Gini Mean difference is $O(n \log n)$.

Gini mean difference has been generalized for multivariate distributions (*Koshvoy97*) That is, the Gini mean difference of a distribution F in \mathbf{R}^d is $\Delta = E\|\mathbf{X} - \mathbf{X}'\|$, or even more generally for some $\alpha \in (0, 2)$,

$$\Delta(\alpha) = E\|\mathbf{X} - \mathbf{X}'\|^\alpha,$$

where $\|\mathbf{x}\|$ is the Euclidean norm. The sample Gini mean difference is computed by

$$\hat{\Delta}(\alpha) = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \|x_i - x_j\|^\alpha.$$

Its computation complexity is $O(n^2)$.

Value

gmd returns the sample Gini mean distance.

References

- Gini, C. (1914). Sulla misura della concentrazione e della variabilit  dei caratteri. Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti, 62, 1203-1248. English Translation: On the measurement of concentration and variability of characters (2005). *Metron*, LXIII(1), 3-38.
- Koshevoy, G. and Mosler, K. (1997). Multivariate Gini indices. *Journal of Multivariate Analysis*, 60, 252-276.
- Schechtman, E. and Yitzhaki, S. (1987). A measure of association based on Gini's mean difference. *Communication in Statistics-Theory and Methods*, 16 (1), 207-231.
- Yitzhaki, S. and Schechtman, E. (2013). *The Gini Methodology*, Springer, New York.

See Also

[RcppGmd](#) [gCov](#) [gCor](#)

Examples

```
n = 100
x <- runif(n)

t0 = proc.time()
gmd(x, alpha=1)
proc.time()- t0

t1 = proc.time()
gmd(x, alpha=0.5)
proc.time()- t1

x <- matrix(runif(n), n/2, 2)
gmd(x,alpha=1)
```

KdCor

Kernel Distance Correlation Statistics

Description

Computes Kernel distance correlation statistics, in which Xs are quantitative, Y are categorical, sigma is kernel standard deviation and returns the measures of dependence.

Usage

```
KdCor(x, y, sigma)
```

Arguments

x	data
y	label of data or univariate response variable
sigma	kernel standard deviation

Details

KdCor compute distance correlation statistics. The sample size (number of rows) of the data must agree with the length of the label vector, and samples must not contain missing values. Arguments x, y are treated as data and labels.

The kernel distance correlation is defined as follow.

$$dCor_{\kappa_X, \kappa_Y}(\mathbf{X}, Y) = \frac{dCov_{\kappa_X, \kappa_Y}(\mathbf{X}, Y)}{\sqrt{dCov_{\kappa_X, \kappa_X}(\mathbf{X}, \mathbf{X})} \sqrt{dCov_{\kappa_Y, \kappa_Y}(Y, Y)}}$$

where

$$\text{dCov}_{\kappa_X, \kappa_Y}(X, Y) = Ed_{\kappa_X}(X, X')d_{\kappa_Y}(Y, Y') + Ed_{\kappa_X}(X, X')Ed_{\kappa_Y}(Y, Y') - 2E[E_{X'}d_{\kappa_X}(X, X')E_{Y'}d_{\kappa_Y}(Y, Y')].$$

Value

KdCor returns the sample kernel distance correlation

References

Sejdinovic, D., Sriperumbudur, B., Gretton, A. and Fukumizu, K. (2013). Equivalence of Distance-based and RKHS-based Statistics in Hypothesis Testing, *The Annals of Statistics*, 41 (5), 2263-2291.

Zhang, S., Dang, X., Nguyen, D. and Chen, Y. (2019). Estimating feature - label dependence using Gini distance statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence (submitted)*.

See Also

[KgCov](#) [KgCor](#) [dCor](#)

Examples

```
x<-iris[,1:4]
y<-unclass(iris[,5])
KdCor(x, y, sigma=1)
```

KdCov

Kernel Distance Covariance Statistics

Description

Computes Kernel distance covariance statistics, in which Xs are quantitative, Y are categorical, sigma is kernel standard deviation and returns the measures of dependence.

Usage

```
KdCov(x, y, sigma)
```

Arguments

x	data
y	label of data or univariate response variable
sigma	kernel standard deviation

Details

KdCov compute distance correlation statistics. The sample size (number of rows) of the data must agree with the length of the label vector, and samples must not contain missing values. Arguments `x`, `y` are treated as data and labels.

Distance covariance was introduced in (Szekely07) as a dependence measure between random variables $X \in R^p$ and $Y \in R^q$. If X and Y are embedded into RKHS's induced by κ_X and κ_Y , respectively, the generalized distance covariance of X and Y is (Sejdinovic13):

$$\text{dCov}_{\kappa_X, \kappa_Y}(X, Y) = Ed_{\kappa_X}(X, X')d_{\kappa_Y}(Y, Y') + Ed_{\kappa_X}(X, X')Ed_{\kappa_Y}(Y, Y') - 2E[E_{X'}d_{\kappa_X}(X, X')E_{Y'}d_{\kappa_Y}(Y, Y')].$$

In the case of Y being categorical, one may embed it using a set difference kernel κ_Y ,

$$\kappa_Y(y, y') = \begin{cases} \frac{1}{2} & \text{if } y = y', \\ 0 & \text{otherwise.} \end{cases}$$

This is equivalent to embedding Y as a simplex with edges of unit length (Lyons13), i.e., L_k is represented by a K dimensional vector of all zeros except its k -th dimension, which has the value $\frac{\sqrt{2}}{2}$. The distance induced by κ_Y is called the set distance, i.e., $d_{\kappa_Y}(y, y') = 0$ if $y = y'$ and 1 otherwise. Using the set distance, we have the following results on the generalized distance covariance between a numerical and a categorical random variable.

$$\text{dCov}_{\kappa_X, \kappa_Y}(X, Y) := \text{dCov}_{\kappa_X}(X, Y) = \sum_{k=1}^K p_k^2 [2Ed_{\kappa_X}(X_k, X) - Ed_{\kappa_X}(X_k, X_k') - Ed_{\kappa_X}(X, X')].$$

Value

KdCov returns the sample kernel distance correlation

References

- Sejdinovic, D., Sriperumbudur, B., Gretton, A. and Fukumizu, K. (2013). Equivalence of Distance-based and RKHS-based Statistics in Hypothesis Testing, *The Annals of Statistics*, 41 (5), 2263-2291.
- Zhang, S., Dang, X., Nguyen, D. and Chen, Y. (2019). Estimating feature - label dependence using Gini distance statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (submitted).

See Also

[KgCov](#) [KgCor](#) [dCov](#)

Examples

```
x<-iris[,1:4]
y<-unclass(iris[,5])
KdCov(x, y, sigma=1)
```

KgCor

*Kernel Gini Distance Correlation Statistics***Description**

Computes Kernel Gini distance correlation statistics, in which Xs are quantitative, Y are categorical, sigma is kernel standard deviation, alpha is an exponent on the Euclidean distance and returns the kernel Gini mean difference.

Usage

```
KgCor(x, y, sigma)
```

Arguments

x	data
y	label of data or univariate response variable
sigma	kernel standard deviation

Details

Kgcor compute kernel Gini distance correlation statistics for data. It is a self-contained R function dealing with both univariate and multivariate data. The sample size (number of rows) of the data must agree with the length of the label vector, and samples must not contain missing values. Arguments x, y are treated as data and labels.

Gini distance correlation are generalized to RKHS, \mathcal{H}_κ , as

$$\text{gCor}_\kappa(X, Y) = \frac{\sum_{k=1}^K p_k [2Ed_\kappa(X_k, X) - Ed_\kappa(X_k, X'_k) - Ed_\kappa(X, X')]}{Ed_\kappa(X, X')}.$$

In this case, we use the default Gaussian distance function

$$d_\kappa(x, x') = \sqrt{1 - e^{-\frac{|x-x'|_q^2}{\sigma^2}}},$$

induced by a weighted Gaussian kernel, $\kappa(x, x') = \frac{1}{2}e^{-\frac{|x-x'|_q^2}{\sigma^2}}$.

Value

KgCor returns the sample Kernel Gini distance correlation between x and y.

References

Zhang, S., Dang, X., Nguyen, D. and Chen, Y. (2019). Estimating feature - label dependence using Gini distance statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (submitted), <https://arXiv.org/pdf/1906.02171.pdf>

See Also

[gCov](#) [gCor](#) [dCor](#)

Examples

```
x<-iris[,1:4]
y<-unclass(iris[,5])
KgCor(x, y, sigma=1)
```

KgCov

Kernel Gini Distance Covariance Statistics

Description

Computes Kernel Gini distance covariance statistics, in which Xs are quantitative, Y are categorical, sigma is kernel standard deviation and returns the kernel Gini covariance.

Usage

```
KgCov(x, y, sigma)
```

Arguments

x	data
y	label of data or univariate response variable
sigma	kernel standard deviation

Details

Kgcov compute kernel Gini distance covariance statistics for data. It is a self-contained R function dealing with both univariate and multivariate data. The sample size (number of rows) of the data must agree with the length of the label vector, and samples must not contain missing values. Arguments x, y are treated as data and labels.

Gini distance covariance are generalized to reproducing kernel Hilbert space (RKHS), \mathcal{H}_κ , as

$$\text{gCov}_\kappa(X, Y) = \sum_{k=1}^K p_k [2Ed_\kappa(X_k, X) - Ed_\kappa(X_k, X_k') - Ed_\kappa(X, X')],$$

In this case, we use the default Gaussian distance function

$$d_\kappa(x, x') = \sqrt{1 - e^{-\frac{|x-x'|_q^2}{\sigma^2}}},$$

induced by a weighted Gaussian kernel, $\kappa(x, x') = \frac{1}{2}e^{-\frac{|x-x'|_q^2}{\sigma^2}}$.

Value

KgCov returns the sample Kernel Gini distance covariance of x and y.

References

Zhang, S., Dang, X., Nguyen, D. and Chen, Y. (2019). Estimating feature - label dependence using Gini distance statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (submitted), <https://arXiv.org/pdf/1906.02171.pdf>

See Also

[gCov](#) [gCor](#) [dCor](#)

Examples

```
x<-iris[,1:4]
y<-unclass(iris[,5])
KgCov(x, y, sigma=1)
```

Kgmd

Kernel Gini Mean Difference Statistics

Description

Computes Kernel Gini mean difference statistics, in which Xs are quantitative, sigma is kernel standard deviation, alpha is an exponent on the Euclidean distance and returns the kernel Gini mean difference.

Usage

```
Kgmd(x, sigma)
```

Arguments

x	data
sigma	kernel standard deviation

Details

Kgmd compute kernel Gini mean difference statistics for data. It is a self-contained R function dealing with both univariate and multivariate data.

The sample size (number of rows) of the data must agree with the length of the label vector, and samples must not contain missing values. Argument x, is treated as data.

Energy distance based statistics naturally generalizes from a Euclidean space to metric spaces (Lyons13). By using a positive definite kernel (Mercer kernel) (Mercer1909), distributions are mapped into a RKHS (Smola07) with a kernel induced distance. Hence one can extend energy distances to a much richer family of statistics defined in RKHS (Sejdinovic13). Let $\kappa : R^q \times R^q \rightarrow R$

be a Mercer kernel (Mercer1909). There is an associated RKHS H_κ of real functions on R^q with reproducing kernel κ , where the function $d : R^q \times R^q \rightarrow R$ defines a distance in \mathcal{H}_κ ,

$$d_\kappa(x, x') = \sqrt{\kappa(x, x) + \kappa(x', x') - 2\kappa(x, x')}.$$

Here KgCov is defined as Gini distance covariance between x and $\text{rank}(x)$.

Value

Kgmd returns the sample Kernel Gini distance

References

Lyons, R. (2013). Distance covariance in metric spaces. The Annals of Probability, 41 (5), 3284-3305.

See Also

[gCov](#) [gCor](#) [dCor](#)

Examples

```
x<-iris[,1]
Kgmd(x, sigma=1)
```

PermutationTest	<i>Permutation test of dependence between X and Y using (Gini) distance covariance or correlation statistics</i>
-----------------	--

Description

Perform permutation test using various dependence measures, in which Xs are quantitative, Y are categorical, alpha is an exponent on Euclidean distance, sigma is kernel parameter in kernel methods and return the test statistic, critical value, p-value and decision of the test.

Usage

```
PermutationTest(x, y, method, sigma, alpha, M = 200, level = 0.05)
```

Arguments

x	data
y	label of data or univariate response variable
method	name of permutation test method and is chosen from one of the method list: dCov, dCor, KdCov, KdCor, gCov, gCor, KgCov, Kgcov
sigma	kernel parameter for kernel methods
alpha	exponent on Euclidean distance, in (0,2), the default value = 1
M	number of permutations
level	significance level of the test, the default value = 0.05

Details

$H_0 : X \text{ and } Y \text{ are independent} \iff H_0 : F(x|y = 1) = F(x|Y = 2) = \dots = F(x|Y = K)$

PermutationTest compute the p-value value of a permutation test of a (Gini) distance covariance or correlation statistics. It is a self-contained R function the measure of dependence statistics.

The p-value is obtained by a permutation procedure. Let $\hat{\rho}(\nu)$ be the sample dependence measure based on the original sample indexed by $\nu = \{1, 2, \dots, n\}$. Let $\pi(\nu)$ denote a permutation of the elements of ν and the corresponding $\hat{\rho}(\pi)$ is computed for the permuted data on y labels. Under the \mathcal{H}_0 , $\hat{\rho}(\nu)$ and $\hat{\rho}(\pi)$ are identically distributed for every permutation π of ν . Hence, based on M permutations, the critical value q_γ is estimated by the $(1 - \gamma)100\%$ sample quantile of $\hat{\rho}(\pi_m)$, $m = 1, \dots, M$ and the p-value is estimated by the proportion of $\hat{\rho}(\pi_m)$ greater than $\hat{\rho}(\nu)$. Usually $100 \leq M \leq 1000$ is sufficient for a good estimation on the critical value or p-value. The default value is $M = 200$.

Value

PermutationTest returns the p-value, critical value and decision of the permutation test of a specified method.

See Also

[gCor](#) [gCov](#) [dCor](#) [dCov](#) [KgCov](#) [KgCov](#) [KdCov](#)

Examples

```
n = 50
x <- runif(n)
y <- c(rep(1,n/2),rep(2,n/2))
PermutationTest(x, y, method = "gCor", alpha = 2, M = 50 )
```

RcppgCor

Gini Distance Correlation Statistics

Description

Computes Gini distance correlation statistics, in which Xs are quantitative, Y are categorical, alpha is exponent on the Euclidean distance and returns the measures of dependence.

Usage

```
RcppgCor(x, y, alpha)
```

Arguments

x	data
y	label of data or univariate response variable
alpha	exponent on Euclidean distance, in (0,2]

Details

RcppgCor compute Gini distance correlation statistic between x and y. It is a Rcpp version of [gCor](#).

Value

RcppgCor returns the sample Gini distance correlation

See Also

[RcppKgCov](#) [RcppKgCor](#) [RcppgCov](#)

Examples

```
x<-iris[,1:4]
y<-unclass(iris[,5])
RcppgCor(x, y, alpha=2)
```

RcppgCov

Gini Distance Covariance Statistics

Description

Computes Gini distance covariance statistics, in which Xs are quantitative, Y are categorical, alpha is an exponent on Euclidean distance and returns the measures of dependence.

Usage

```
RcppgCov(x, y, alpha)
```

Arguments

x	data
y	label of data or univariate response variable
alpha	exponent on Euclidean distance, in (0,2]

Details

RcppgCov compute Gini distance covariance statistics. It is Rcpp version of [gCov](#).

Value

RcppgCov returns the sample Gini distance covariance

See Also

[RcppgCor](#) [RcppKgCov](#) [RcppKgCor](#)

Examples

```
x<-iris[,1:4]
y<-unclass(iris[,5])
RcppgCov(x, y, alpha=2)
```

RcppGmd*Gini Mean Difference Statistics*

Description

Computes Gini mean difference of x , where α is an exponent on the Euclidean distance and return the Gini mean difference. The default value for α is 1.

Usage

```
RcppGmd(x, alpha)
```

Arguments

<code>x</code>	data
<code>alpha</code>	exponent on Euclidean distance, in (0,2]

Details

RcppGmd compute Gini mean difference statistics for data. It is a Rcpp version of [gmd](#).

Value

RcppGmd returns the sample Gini mean difference of x .

See Also

[RcppKgCov](#) [RcppgCor](#) [gCov](#) [gCor](#)

Examples

```
n=1000
x<-runif(n)
RcppGmd(x, alpha=1)
```

`RcppKgCor`*Kernel Gini Distance Correlation Statistics*

Description

Computes Kernel Gini distance correlation statistics, in which Xs are quantitative, Y are categorical, sigma is kernel standard deviation and return the kernel Gini mean difference.

Usage

```
RcppKgCor(x, y, sigma)
```

Arguments

<code>x</code>	data
<code>y</code>	label of data or univariate response variable
<code>sigma</code>	kernel standard deviation

Details

`RcppKgCor` compute kernel Gini distance correlation statistics for data. It is Rcpp version of [KgCor](#).

Value

`RcppKgCor` returns the sample Kernel Gini distance covariance

See Also

[gCov](#) [gCor](#) [dCor](#)

Examples

```
n=100
x<-runif(n)
y<-c(rep(1,n/2),rep(2,n/2))
RcppKgCor(x, y, sigma=1)
```

`RcppKgCov`*Kernel Gini Distance Covariance Statistics*

Description

Computes Kernel Gini distance covariance statistics, in which Xs are quantitative, Y are categorical, sigma is kernel standard deviation and return the kernel Gini mean difference.

Usage

```
RcppKgCov(x, y, sigma)
```

Arguments

<code>x</code>	data
<code>y</code>	label of data or univariate response variable
<code>sigma</code>	kernel standard deviation

Details

`RcppKgCov` compute kernel Gini distance covariance statistics for data. It is Rcpp version of [KgCov](#).

Value

`RcppKgCov` returns the sample Kernel Gini distance covariance

See Also

[gCov](#) [gCor](#) [dCor](#)

Examples

```
n=100
x<-runif(n)
y<-c(rep(1,n/2),rep(2,n/2))
RcppKgCov(x, y, sigma=1)
```

`RcppKGmd`*Kernel Gini Mean Difference Statistics*

Description

Computes Kernel Gini mean difference of X, sigma is the kernel parameter and returns the kernel Gini mean difference.

Usage

```
RcppKGmd(x, sigma)
```

Arguments

x	data
sigma	kernel parameter for Gaussian kernel

Details

RcppKGmd compute kernel Gini mean difference for data It is Rcpp version of [Kgmd](#).

Value

RcppKGmd returns the sample Kernel Gini distance

See Also

[gmd](#) [Kgmd](#)

Examples

```
x<-iris[,1]
RcppKGmd(x, sigma=1)
```


Index

*Topic **multivariate**

- ConfidenceInterval, 3
- CriticalValue, 4
- dCor, 5
- dCov, 6
- gCor, 8
- gCov, 9
- gmd, 10
- KdCor, 12
- KdCov, 13
- KgCor, 15
- KgCov, 16
- Kgmd, 17
- PermutationTest, 18
- RcppgCor, 19
- RcppgCov, 20
- RcppGmd, 21
- RcppKgCor, 22
- RcppKgCov, 23
- RcppKGmd, 24

*Topic **package**

- GiniDistance-package, 2

ConfidenceInterval, 3

CriticalValue, 4

dCor, 5, 7, 13, 16–19, 22, 23

dCov, 6, 6, 14, 19

gCor, 8, 10, 12, 16–23

gCov, 9, 9, 12, 16–23

GiniDistance (GiniDistance-package), 2

GiniDistance-package, 2

gmd, 9, 10, 10, 21, 24

KdCor, 6, 7, 12

KdCov, 6, 7, 13, 19

KgCor, 9, 10, 13, 14, 15, 22

KgCov, 9, 10, 13, 14, 16, 19, 23

Kgmd, 17, 24

PermutationTest, 4, 5, 18

RcppgCor, 19, 20, 21

RcppgCov, 20, 20

RcppGmd, 12, 21

RcppKgCor, 20, 22

RcppKgCov, 20, 21, 23

RcppKGmd, 24